Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.fw001

40 Years of Chemometrics – From Bruce Kowalski to the Future

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.fw001

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

ACS SYMPOSIUM SERIES **1199**

40 Years of Chemometrics – From Bruce Kowalski to the Future

Barry K. Lavine, Editor
Oklahoma State University
Stillwater, Oklahoma

Steven D. Brown, Editor
University of Delaware
Newark, Delaware

Karl S. Booksh, Editor
University of Delaware
Newark, Delaware

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.fw001

Sponsored by the
ACS Division of Computers in Chemistry

American Chemical Society, Washington, DC

Distributed in print by Oxford University Press

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Library of Congress Cataloging-in-Publication Data

40 years of chemometrics : from Bruce Kowalski to the future / Barry K. Lavine, editor, Oklahoma State University, Stillwater, Oklahoma, Karl S. Booksh, editor, University f Delaware, Newark, Delaware, Steven D. Brown, editor, University of Delaware, Newark, elaware ; sponsored by the ACS Division of Computers in Chemistry.
pages cm. -- (ACS symposium series ; 1199)
Includes bibliographical references and index.
ISBN 978-0-8412-3098-9 -- ISBN 978-0-8412-3097-2 1. Chemometrics. 2. Chemistry, Analytic. I. Lavine, Barry K., 1955- editor. II. Booksh, Karl S., editor. III. Brown, Steven D., 950- IV. American Chemical Society. Division of Computers in Chemistry. V. Title: Forty ears of chemometrics.
QD75.4.C45F67 2015
543.01'5195--dc23

2015033032

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI Z39.48n1984.

Distributed in print by Oxford University Press

PRINTED IN THE UNITED STATES OF AMERICA

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.fw001

Foreword

The ACS Symposium Series was first published in 1974 to provide a mechanism for publishing symposia quickly in book form. The purpose of the series is to publish timely, comprehensive books developed from the ACS sponsored symposia based on current scientific research. Occasionally, books are developed from symposia sponsored by other organizations when the topic is of keen interest to the chemistry audience.

Before agreeing to publish a book, the proposed table of contents is reviewed for appropriate and comprehensive coverage and for interest to the audience. Some papers may be excluded to better focus the book; others may be added to provide comprehensiveness. When appropriate, overview or introductory chapters are added. Drafts of chapters are peer-reviewed prior to final acceptance or rejection, and manuscripts are prepared in camera-ready format.

As a rule, only original research papers and original review papers are included in the volumes. Verbatim reproductions of previous published papers are not accepted.

ACS Books Department

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.fw001

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.pr001

Preface

This American Chemical Society (ACS) Symposium series text is based on the full day symposium entitled, "The Birth of Chemometrics - In Honor and Memory of Bruce Kowalski," that was held at FACSS in Milwaukee, WI (October 2013) and cosponsored by the Division of Computers and Chemistry (COMP) of the American Chemical Society. Bruce Kowalski is recognized by the scientific community as the founder of the field of chemometrics. This symposium Series text is a follow up to the Symposium Series Volume 52 (Chemometrics: Theory and Application), edited by Bruce Kowalski and is based on the symposium organized by Bruce at the National ACS meeting in San Francisco in 1976, which was also cosponsored by COMP.

The 14 contributors to the current volume (see Table of Contents) are all leaders in the field of chemometrics and have strong personal recollections of Bruce as a man who was a catalyst able to bring together creative minds. All major areas in the field are well represented in this collection: pattern recognition, library searching, multivariate calibration, multivariate curve resolution, variable selection, data fusion, calibration transfer, environmental chemometrics, forensics, and biological and mixture analysis. Many chapters have a link to previous work done by Bruce and will serve as a retrospective to the career of Bruce Kowalski, who believed that a rational approach was needed to improve both the quality of measurements and to extract information from them. Bruce believed that chemometrics would serve as a guiding theory for analytical chemistry and believed that it would be used both to optimize existing analytical methodology and to direct researchers attempting to construct better tools. Each chapter in this text demonstrates the progress that has been made in the field towards the realization of Bruce Kowalski's goal.

The first chapter of the text entitled, "Chemometrics and Bruce: Some Fond Memories," is written by Svante Wold and describes the history of empirical and semi-empirical "data driven, soft, analogy" models for the design of experiments and the analysis of the resulting data. This history is marked by a number of influential events inspired and encouraged by Bruce and is illustrated by examples of method development driven by necessity to solve specific problems and leading to data driven soft models, which have been shown to be superior to the classical first principles approaches to the same problems.

Although Bruce is recognized for his accomplishments in attracting the talents of chemists and engineers, he also showed enormous vision in his efforts to assimilate statisticians and mathematicians into the field of chemometrics. Bill Rayens, a mathematician turned statistician, describes his journey from statistician to chemometrician as a result of his interactions with Bruce Kowalski

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

in Chapter Two. Chapter Three, written by Peter Wentzell, describes the evolution of maximum likelihood principal component analysis and related techniques from a personal perspective highlighting the author's collaboration with Bruce Kowalski.

Chapter Four focuses on Bruce Kowalski as a mentor, innovator and pioneer through the solution to a problem involving dioxin involving dioxin concentrations in excess of background levels in the harbor of Port Angeles in Western Washington. A mixture analysis study undertaken by Scott Ramos using pattern recognition and multivariate curve resolution methods to understand the nature of the contamination indicated several characteristic patterns that could be associated with identifiable source materials. The work in this study was performed at Infometrix, a software company founded by Bruce Kowalski in 1978.

Chapters Five and Six focus on multivariate curve resolution. In Chapter Five, Roma Tauler provides an exhaustive review of multivariate curve resolution, which is the generic denomination for a family of methods used to solve the ubiquitous problem of mixture analysis. Tauler became interested in curve resolution while visiting Bruce in the late 1980's. Phil Hopke in Chapter Six describes some recent developments in multivariate curve resolution related to the problem of source apportionment in air monitoring of atmospheric particulates.

Recent developments in the field of pattern recognition are delineated by Steve Brown and Barry Lavine in Chapters Seven and Eight. In Chapter Seven, hierarchical class modeling approach in which samples receive more than one class label are compared to traditional "flat" classification for the modeling of hierarchical geospatial data, a problem that relates to those studied by Brown and Kowalski in the late 1970's. In Chapter Eight, pattern recognition techniques are applied to the problem of searching the infrared spectral libraries of the Paint Data Query (PDQ) automotive paint database to differentiate between similar IR spectra and to determine the assembly plant, model, and line of an automotive vehicle from a clear coat paint smear recovered at a crime scene where damage to a vehicle and/or injury or death to a pedestrian has occurred. This, too, echoes work started by Brue in early studies of pattern recognition, but his focus was on paper.

Chapters Nine, Ten, and Eleven focus on recent developments in multivariate calibration, another area where Bruce contributed significantly. Model selection is usually limited to the evaluation of cross validation prediction errors. However, there are advantages of using multiple criteria for model selection, which is discussed by John Kalivas in Chapter Nine. Chapter Ten focuses on the solution to the variable selection problem in PLS and PCR using adaptive regression subspace elimination approach pioneered by Karl Booksh. The essentials of multivariate calibration transfer are discussed by Jerry Workman in Chapter Eleven.

The remaining three chapters of this text focus on biological applications of chemometrics. The field of proteomics and metabolomics from the standpoint of chemometrics is reviewed by Jeff Cramer in Chapter Twelve. Rene Jiji in Chapter Thirteen explores the application of data fusion for spectroscopic data to improve predictions of protein secondary structure. The remaining chapter

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.pr001

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

by Frank Vogt summarizes nonlinear modeling of microalgal biomasses for the purpose of exploring the impact of pollutants on our environment.

This text will be of interest to individuals who are interested in modeling data. Interest in modeling data continues to grow with the emergence of new areas such as computational statistics, business intelligence, big data, and analytics. In chemistry, modeling of data has taken a different path as it has become integrated into the field of analytical chemistry. Because chemometrics is not well understood by chemists, this text should prove beneficial and be of great interest to researchers who need to take advantage of techniques such as principal component analysis, partial least squares, linear discriminant analysis and outlier analysis in their work.

This book allows the reader quick access to different areas of current research in chemometrics featured in the literature by providing key references and viewpoints not found elsewhere. This text also highlights changes that have occurred in the field since its origins in the mid-1970's and will serve as a report on the current state of the art of the field of chemometrics. The editors of this text believe that it will be of interest not only to physical scientists and engineers but also to statisticians and informatics types who have come to the realization that chemometrics is worth a second look.

Barry K. Lavine
Oklahoma State University
Department of Chemistry
Stillwater, Oklahoma 74078

Steven D. Brown
University of Delaware
Department of Chemistry and Biochemistry
Newark, Delaware 19716

Karl S. Booksh
University of Delaware
Department of Chemistry and Biochemistry
Newark, Delaware 19716

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.pr001

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ot001

Editors' Biographies

Barry K. Lavine

Barry K. Lavine received his PhD from Pennsylvania State University in 1986. His thesis advisor was Peter C. Jurs. In the same year, Lavine became a faculty member in the Chemistry Department at Clarkson University where he taught and performed research in analytical chemistry and chemometrics for 18 years. In 2004, Lavine moved to Oklahoma State University (OSU) where he continues to be active in both teaching and research. Lavine's publications include some 150 publications, chapters, and review articles as well as three books. His research encompasses applications of multivariate data analysis to a wide range of problems in the areas of analytical and forensic chemistry including automotive paint analysis, vibrational spectroscopy (attenuated total reflection, infrared and Raman imaging), library searching, chemical fingerprinting, and biomarker identification. Lavine is a member of the editorial board of several journals including *Analytical Letters*, *Journal of Chemometrics*, and the *Microchemical Journal* and served as the author of the fundamental review of chemometrics which was published biennially by Analytical Chemistry. Lavine has served as Chair of the Northern New York Section and Oklahoma Section of the ACS and as Program Chair for SCiX in 1992 and the Northeast Regional Meeting of the ACS in 1999.

Steven D. Brown

Steven D. Brown (PhD, University of Washington) began research as an inorganic chemist, earning an MS for work in fluorine chemistry. When he enrolled in the PhD program at the University of Washington, he took up the new field of chemometrics and earned a PhD from Bruce Kowalski in 1978. He has served as Assistant Professor at UC Berkeley, as Associate Professor at Washington State University, and is now Willis F. Harrington Professor at the University of Delaware, where he teaches analytical chemistry and multivariate statistical methods in chemistry.

Dr. Brown's publications include some 200 publications, chapters and reports, as well as two books, including the four-volume treatise *Comprehensive Chemometrics* (2009, Elsevier). His research comprises application of multivariate data analysis to a wide range of problems relying on chemical measurements, including approaches to data fusion, transfer of calibration Bayesian analysis, and multivariate classification.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Karl S. Booksh

Karl S. Booksh is a professor of Chemistry and Biochemistry at the University of Delaware. He earned his Doctorate in Analytical Chemistry working with Prof. Bruce R. Kowalski in the Center for Process Analytical Chemistry at the University of Washington. He was an NSF Postdoctoral Fellow at the University of South Carolina before joining the faculty at Arizona State University in 1996. He has received a NSF CAREER Award, Camille and Henry Dreyfus New Faculty Fellowship and Elseiver Chemometics Award. He served as the North American Editor for the Journal of Chemometrics. Booksh is a Fellow of the Society for Applied Spectroscopy and a Fellow of the American Chemical Society.

Booksh's research interests revolve around sensor design and calibration. Booksh's graduate work was on multi-way calibration. As a postdoc he began designing sensors to become more compatible with multi-way calibration methods. Recently Booksh has been working in multivariate image analysis and developing multivariate calibration and classification strategies that are robust to uncalibrated interferents. Booksh is also active in broadening participation in chemistry, particularly for students with disabilities.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ot001

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch001

Chapter 1

Chemometrics and Bruce: Some Fond Memories

Svante Wold*

Institute of Chemistry, Umeå University, Sweden
***Phone: 603-465-2622, e-mail: sbwold@gmail.com**

This chapter describes the transformation of a young physical organic chemist (SW, 1964), from a believer in first principles models to a middle-aged chemometrician (SW, 1974) promoting empirical and semiempirical "data driven, soft, analogy" models for the design of experiments and the analysis of the resulting data. This transformation was marked by a number of influential events, each tipping the balance towards the data driven, soft, analogy models until the point of no return in 1974. On June 10, 1974, Bruce and I together with our research groups joined forces formed the Chemometrics Society (later renamed to the International Chemometrics Society), and we took off into multidimensional space. This review of my personal scientific history, inspired and encouraged by Bruce, is illustrated by examples of method development driven by necessity to solve specific problems and leading to data driven soft models, which, at least in my own eyes, were superior to the classical first principles approaches to the same problems. Bruce and I met at numerous conferences between 1975 and 1990, but after that, Bruce and I gradually slid out of the academic world, and now Bruce has taken his final step.

Introduction

I first met Bruce in 1973 during my stay as Statistician in Residence (Department of Statistics) at the University of Wisconsin at Madison (UW). This one year appointment was a direct result of attending the 1972 Gordon

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Research Confence in Statistics and meeting George Box and Bill Hunter, eminent statisticians at UW. In the previous year, I had received my PhD in physical organic chemistry at Umeå University in Sweden, and I then coined the term chemometrics in a grant application for support of my research. Prior to assuming my position as Statistician in Residence at Wisconsin, Kowalski and Bender had published two seminal papers on pattern recognition in chemistry (*1*, *2*). Their work was a revelation to me, and I was anxious to meet Bruce Kowalski as well as receive an education of sorts about statistics during my year with Box and Hunter at UW.

The opportunity to meet Bruce was the result of a telephone conversation betwen George Box and Rudi Marcus. A symposium involving chemistry faculty who use statistics and computers in their research was organized by ONR in Tucson, AZ in the fall of 1973. Box was asked by Marcus to be the expert in statistics and computers and serve as a referee for the symposium. Box instead volunteered my services as a chemist knowledgeable in computers, and I flewthe night beforeto Tucson. Unfortunately, the airlines lost my luggage during the trip, and I arrived on Monday morning at the lecture hall wearing dirty dungarees and a black shirt. Faculty giving presentations were all dressed in suits and ties. As I entered the lecture hall, a gentlemen approached me, introduced himself as Dr. Marcus and asked if he could help me. After explaining that I was the referee from Wisconsin and apologizing for my appearance, I was warmly greeted by Dr. Marcus and brought over to meet the faculty participants. They formed a line in the lecture hall, and I greeted each of them with a hand-shake. The second individual in the line was Bruce Kowalski, and I was stunned to see a young man of my own age, but I held a straight face and expressed my appreciation for his interesting work. Later, when listening to Bruces's presentation I concluded that he was the only individual at this conference who understood how computers should be used in chemistry. After the meeting, I shared this observation with ONR in my report.

During the Tucson conference, I had the opportunity to speak with Bruce at great length. I told Bruce that he and I were the only individuals attending the conference who were active in the field of chemometrics. Bruce was not familiar with the term chemometrics and initially exhibited some resistance to this term. He nonetheless accepted it as a description also of his own research and quickly became both a strong proponent and spokesman for this new field. Bruce in our conversations at the ONR meeting also expressed interest in learning more about SIMCA (*3*, *4*) and invited me to visit him at the University of Washington (UW) in June 1974 to demonstrate SIMCA's capabilities relative to his own software package ARTHUR (*5*), which emphasized the linear learning machine, PCA, K-nearest neighbor classification, hierarichical clustering, and graphics. During the head-to-head competition between SIMCA and ARTHUR which involved the analysis of 20 standard data sets, we concluded that SIMCA was superior to the linear learning machine and K-nearest neighbor and Bruce subsequently incorporated a version of SIMCA into ARTHUR. While celebrating the success of SIMCA at a Seattle bar and grille on the strip boarding UW with Bruce and his research group, the International Chemometrics Society was formed after ten shots of taquila. My June trip to UW was followed by attending the 1974 Gordon

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch001

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Research Conference on Statistics in Chemistry and Chemical Engineering in July where Bruce presented a very impressive talk on chemometrics.

Bruce accepted my invitation to visit Umeå in 1978 (where Bruce at his farewell party consumed more bodycakes, aka dumplings, than any other participant). Then we both were invited to Herman Wold's PLS Conference in Cartigny, Switzerland in 1979. At Cartingny, both Bruce and I were scolded by Herman Wold for arriving late to a presentation because we had spent the afternoon in Geneva.

The first international meeting in chemometrics, well organized by Bruce, was held in 1983 at Cozenza. At this meeting both PLS discriminant analysis and PLS regression were discussed at great length. Although this meeting was rich in intellectual content, the living conditions for this meeting were somewhat spartan (e.g., no toilet seats or paper). The label "Cozenza survivors" is often affixed to meeting participants and both Bruce and I considered our participation in this meeting as a badge of honor. Other international meetings attended by both Bruce and I included MULDAST 84 (Umeå) and Gothenburg 90 where Bruce and I were awarded Bergman medals by the Swedish Chemical Society.

Bruce's retirement from UW was celebrated at FACSS 99 as was the 25th anniversity of the formation of the International Chemometrics Society. At this meeting, it was evident that both Bruce and I were sliding out of the academic world. Our paths were diverging. Bruce's was stepping down as Head of the Center for Process Analytical Chemistry while my involvement in Umetrics, a company that I had founded with Rolf Carlson in 1987, was burgeoning. Bruce and I saw each other less and less often, and the last time was SSC 11 in Loen, Norway 2009 (see Figure 1).

Figure 1. My last conference with Bruce at SSC11 in Loen, Norway

Strolling down memory lane, I will attempt in the remaining part of this chapter to explain what I have gained from Bruce over the years by way of projects performed alone or by my own research group that I view as historically important. It is easy to become entangled and distracted by sentimental memories

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch001

of Bruce, as well as those funny moments and unexpected revelations. However, the remainder of is chapter are devoted to chemometrics, a field that Bruce and I have both helped to nurture and develop from its infancy.

Principal Commponents Analysis (PCA)

Chemical data often occur in the form of tables with, say, K variables (properrties, meaurements, ...) made on N samples (objects, cases, ...). A typical example consists of N analytical samples of, say, seawater, with K GC-MS variables measured on each sample. Another example is provided by N samples of tumor cells with K GC variables measured in the cell walls of each tumor sample.

Around 1965-70 the chemical labs were invaded by new measuring devices such as infrared and NMR spectrometers, and mass spectrometers combined with separation methods such as gas- and liquid chromatography. Many of these instruments gave data with more variables than samples. Statistical folklore stated that these data could not be analyzed as such, but first the number of variables K must be decreased to be subsantially smaller than the number of samples, N. Luckily, my father Herman was working with multivariate (K > 1) economics data and found to his satisfaction that certain types of data analysis examplified by PCA worked well also with data matrices with K > N. We of course tried PCA on chemical data tables with K > N. And PCA worked well there too. It was just that nobody had really tried -- the N > K dogma was too strong.

Hence, PCA has become a cornerstone of chemometrics. It is closely related to factor analysis (FA) and often called principal factor analysis. PCA provides a decomposition of a (N x K)matrix in terms of a set of pairs of score column vectors (t_a) times loading row vectors (p_a'). This separates the information in the data table into one part concerned with the samples times one part concerned with the variables. See eqn.1 below.

It is easy to understand how PCA applies to a spectral data matrix where each row is a sample spectrum, and each column is the absorbtion of radiation at a certain frequency. According to Lambert-Beer's law, each row spectrum is the sum of the sample constituent spectra times the corresponding concentrations. This was shown early by Bruce, and it led to a way to determine the constituent concentrations in new sample on the basis of a training set of samples with known constituents.

PCA was further clarified by seing the data matrix represented as N points in a K dimensional space, and the components (scores times loadings) as a hyperplane in this space with as many dimentions as the number of significant constituents in the actual data.

Linear Free Energy Relationships

Being an organic chemist I observed that a number of PC-like models – LFERs -- with one or two components were used by physical organic chemists to understand the reactivity of sets of similar molecules in different reactions.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch001

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch001

Examples are the Brönsted and the Hammett "equations". At a visit to Otto Exner in Prague in spring 1971, he helped me understand how PCA was a rather natural way to model such reactivity data, and after returning to Sweden I managed to derive these models as Taylor-expansions of unknown relationships with a certain plausible structure.

Hence, chemometrics can trace its origins also to the field of physical organic chemistry where linear free energy relationships (LFERs) have been used since the late nineteenth and early twentieth centuries to describe the variation in reactivity among similar compounds. Early examples include the relationship between the narcotic effect of a series of drugs and their partition coefficient (*6*) and the chemical reactivity of substituted benzene derivatives as a function of their substitutents, the so-called Hammett equation (*7*). When the model errors are too large, one can often decrease their size either by limiting the domain of the model or by including more terms in the equation. During my dissertation studies in physical organic chemistry, I came to the conclusion that an interpretation of LFERs in terms of expressing combinations of fundamental effects was simply too restrictive. A better approach would be to treatthe LFERs as empirical models of similarity. Using PCA (*8*), it was easy to demonstrate that LFERs could be derived from a table of measured data. When comparing a dual substituent parameter model in a modified Hammett equation (see eqn. 2) with the equation used for a PC model (see eqn.1), LFERs are seen to be mathematically and statistically equivalent to few-components PC models.

One advantage of using PCA to develop LFERs is that this provides information about the number of product terms necessary to give the model its optimal predictive properties. The substituent scales obtained directly from PCA of the datainclude sensitivity parameters (i.e., the loadings), the influence of the substituent on the phenomena investigated (i.e., the scores) and the reference point corresponding to the substituent having no influence on the properties or reactivity of the compounds (i.e., the mean of the measured data). Residuals in both the modified Hammett equation and the PC model describe the nonmodeled part of the data which can be attributed to measurement errors and the inherent limitations of the model being simplifications of reality.

$$y_{ik} = \alpha_i + \sum_{a=1}^{A} \rho_{ia}\sigma_{ak} + \varepsilon_{ik} = \alpha_i + \sum_{a=1}^{A} p_{ia}\, t_{ak} + e_{ik} = \mathbf{a'} + \mathbf{T}\,\mathbf{P'} + \mathbf{E} \quad (1)$$

$$logk_{ik} = logk_{i0} + \rho_{i1}\sigma_{1k} + \rho_{i2}\sigma_{2k} + \varepsilon_{ik} \quad (2)$$

A direct consequence of this realization is that LFERs can be treated as locally valid linearizations of complicated functional relationships. By equating the measured data with a continuous function in two vector variables followed by a differentiation of the function and a grouping of the terms in the resulting Taylor expansion, it is shown that PC models can approximate data measured on an ensemble of similar objects whether these objects are complex biological samples, chemical reactions or equilibria (*9*, *10*). Although mathematically this argument is straightforward, its consequences are profound. PC models can be used to describe any data of a class of similar objects whether the data are kinetic, thermodynamic, spectroscopic, or express product distributions or

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

constituent concentrations in the objects. Furthermore, any variable measured on an ensemble of sufficiently similar objects is correlated to any other variable measured on the same objects. The closer the similarity between objects, the stronger the correlations are. From the standpoint of a philosophy to investigate chemistry, empirical models that are locally valid can be constructed for similar objects or processes but the variables that are measured often have no fundamental meaning other than serving as indicators of similarity. This forms the basis of the SIMCA method for classification (*3*, *4*). Here each class is modelled by a separate PC model which is derived from the training set, and new observations (cases, samples, ...) are classified according to their similarity – proximity in the multivariate data space – to the class models.This provides a quantification of the concepts of similarity and analogy, cornerstones in the understanding of complex systems.

To demonstrate both the efficacy and value of this approach, Albano and Wold (*11*) employed PC models to investigate the variation in the rate of solvolysis reactions involving both exo-and endo-2-norbornyl compounds using published but nonanalyzed kinetic data. Winstein in 1949 had proposed the existence of so-called nonclassical carbonium ions to explain the abnormally fast solvolysis of exo- 2-norbornyl compounds in comparison to the corresponding endo compounds. H. C. Brown proposed an alternative explanation invoking steric strain release and rapid equilibria between classical ions.

The data set obtained from the literature consisted of 26 compounds and 14 rate constants for 7 solvents of different polarities ranging from methanol to trifluoro-acetic acid at two different temperatures. A reaction where the charge is localized in the transition state would be more affected by a change in solvent polarity than a reaction where the charge is delocalized in the transition state. Therefore, these data could provide a solution to the controversy arising from the different interpretations for the abnormally fast solvolysis of exo-2-norbornyl compounds. Of the 26 compounds in the data set, two were exo- norbornyl and two were endo-norbornyl compounds. It was the consensus of workers in the field that two of the 26 compounds reacted via a classical ion transition state with delocalized charge, one compound reacted through a nonclassical ion transition state where delocalization occurred, and 16 compounds reacted via a classical ion transition state with localized charge. The remaining three compounds were labeled as interesting compounds as workers in the field could not come to an agreement regarding charge delocalization in their transition state.

A cross validation analysis (*12*, *13*) shows that a two components PC model adequately describes this matrix. This corresponds to using a model consisting of two phenomenological factors to describe the 26 solvolysis reactions. Figure 2 shows a score plot of the two largest PCs of the data. An examination of this plot revealed four potential compounds clusters. Two clusters (labeled as primary and secondary which refers to the substituents varied in the parent structure) are comprised of compounds that react through a classical ion transition state with localized charge. Both endo norbornyl compounds fall in the cluster containing compounds that form classical ion transition states. The other two clusters contain compoundsexhibiting charge delocalization in their transition state. Since the exo- 2-norbornyl compounds fall in a cluster associated with known charge

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch001

delocalization in the transition state, Winstein's interpretation would appear to be correct as these results would be difficult to interpret using Brown's formalism.

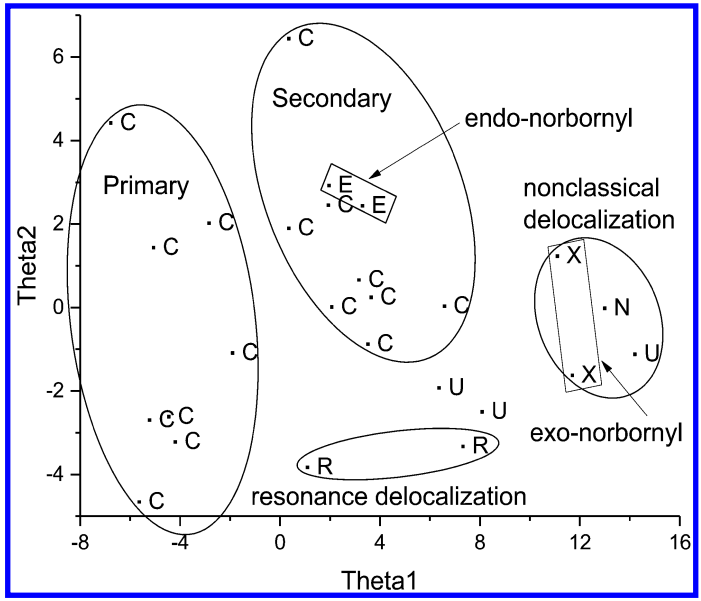

Figure 2. Plot of the scores of the two largest PC.s of the 26 compounds and the 14 specific rate constants comprising the data set. C = Classical, E = Endo-Norbonyl, N = Non-Classical, R = Resonance, U = Unknown, X = Exo-Norbonyl.

The clear similarity between the endo-norbornyl substrates and the ordinary cyclic secondary substrates such as cyclohexyl and cyclopentyl and between exo-norbornyl and methylcyclopropyl constitutes additional evidence for the presence of nonclassical charge delocalization in exo-norbornyl solvolyses. The approach used here, structuring the problem as one of empirical similarities which can be tackled by multivariate methods, is more straightforward than theories that rely upon the detailed behavior of solvolytic transition states.A fundamental model relating the degree of charge delocalization to measured rate constants would be problematic to construct. The importance of using model systems with known behavior, in this case transition states with and without charge delocalization, the need for data related to the problem of interest and the application of the appropriate multivariate analysis method cannot be emphasized enough.

PCA Extensions; PLS, PLS-DA, OPLS and OPLS-DA

PCA is an excellent modelling tool for data where all variables are of the same type. However, a very common data-analytical problem in chemistry and

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch001

elsewhere is given by that of multiple regression. Here to main problem is to model and predict one or several response variables, y or Y, from a set of predictor variables collected in the matrix X. We note that in the typical "chemometrics" situation, the X-variables are numerous and collinear, and hence cannot be called "independent".

Between 1975 and 1980 my father Herman spent much of his time thinking about this problem. He developed something he called multivariate path models estimated by partial least squares (PLS). Together we showed that the simplest of these PLS models with only two blocks could handle the multiple regression problem also for data where the predictor matrix X had many variables and relatively few samples. This turned out to solve also the discriminant analysis (classification) problem using the response matrix Y with binary (1/0) variables expressing the class structure. We naturally called this PLS-DA for PLS discriminant analysis.

PLS models can also be seen as extensions of PCA where a PC model is derived to (a) approximate the predictor matrix X, and (b) form a linear relation between the X-scores and the response variable(s) y or Y. Thus, two-block PLS addresses the same problem as multiple linear regression (MLR) with the difference that PLS also forms a bilinear model of X. The latter makes PLS capable of modeling the relationship between X and y (or Y) also when the number of X-variables, K, greatly exceeds the number of cases (samples, ...), N.

Like PCA, PLS can be derived as a truncated Taylor expansion for a set of similar samples. Hence PLS applies to any data measured on a set of similar objects (cases, samples, ...). This has made the PLS approach widely used with data from modern instrumental techniques providing many variables such as GC and LC/MS, NIR, NMR, as well as genomic and other bioanalytical data.

To facilitate the interpretation of PLS models (including PLS-DA), different rotations of the models have been tried. We founda way to capture all information in X about y (or Y) in a single score vector, t_{OPLS} for the case with a single y. This was given the name OPLS for Orthogonal PLS (*14*). Below we see an example of OPLS-DA (OPLS discriminant analysis) applied to a biological data set with many variables.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch001

Multivariate Calibration

In an intensive and exciting collaboration between Herman, Harald Martens, and myself, we formulated around 1980 what we called multivariate calibration. Here we wish to predict the concentration of one or several analytes in "new" samples from the spectra and known composition of a training set of N samples. And the spectra may have many more variables K than the number of training samples, N.

PLS worked like a charm for this type of data, creating great excitement in analytical chemistry, Harald Martens and Tormod Naes wrote an excellent book about the subject, and the rest is history.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Detection of Ovarian Cancer Using Chemometric Analysis of Proteomic Profiles

The early diagnosis of ovarian cancer could significantly reduce mortality rates among women. The disease usually presents few symptoms until it spreads beyond the ovaries. The five year survival rate for late stage presentation is 35%, whereas a diagnosis in the early stage is associated with a five-year survival rate of over 90% (*15*). In a landmark paper, Petricoin and coworkers (*16*) discriminated between known ovarian cancer patients and normal controls using a data analytical tool based on a combination of genetic algorithms and cluster analysis applied to low resolution SELDI-TOF mass spectra of blood serum samples. Petricoin's study has raised the possibility of developing a rapid and inexpensive technique for the accurate and timely diagnosis of ovarian cancer.

Petrocoin's data has been analyzed previously by other workers (*15*, *17–19*) using univariate methods for variable selection with the selected variables entered into a stepwise linear discriminant analysis routine for discriminant development. Here, we reanalyzed Petrocoin's data (*20*) using a more straight forward approach –OPLS-DA (*14*). Unlike the methods used by Petrocoin and other workers, OPLS-DA is a single step approach that analyzes all 15,154 m/z values (which ranged from 0 to 20,000) simultaneously without the need for prior variable selection and without the rigid constraints of having more samples than variables.

Like PLS, OPLS-DA is a scale dependent method. When applying these methods to optical spectra, the variables (e.g., absorbance values at each wavelength) are routinely centered but not scaled prior to the analysis to ensure that wavelength regions with the largest signal amplitude variation exhibit the most influence on the data analysis. The alternative is to apply autoscaling where each variable is centered and then scaled by its standard deviation. The drawback of autoscaling is that noisy wavelength regions in the spectra can become inflated which may mask the effects of interests.

However, with NMR and MS data, Pareto scaling has become popular. Pareto scaling is a compromise between mean centering and autoscaling and involves dividing each spectral variable by the square root of its standard deviation after first centering the data. Pareto scaling was applied to the SELDI-TOF mass spectral data prior to OPLS-DA.

To optimize OPLS-DA, samples comprisng the training set were selected from the control and ovarian cancer groups according to a statistical design. First, PCA was used to calculate a number of scores (known as principal properties) and then the principles of design of experiments were applied to these scores. The use of design of experiments ensured the selection of a diverse and representative training set that embraces the validation set. The full data set of 91 controls and 100 cancer patients was divided into a training set of 97 samples and a validation set of 94 samples.

PCA performed for the controls and the cancer group together did not reveal any strong outliers. Three PCs provided a good summary of the data explaining 77% and 75% of the total cumulative variance of the data respectively. A 4^3 full factorial experimental design defining 64 combinations of the first three PC.swere used to select the training set. Each PC score vector was divided into 4 levels and

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch001

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch001

with three PC.s this created 64 combinations. Selection of samples for the training set was limited to individuals corresponding to design points. This resulted in a training set of 43 controls and 54 ovarian cancer patients. The remaining 48 controls and 46 cancer patients formed the validation set.

OPLS-DA was applied to the training set. The number of PLS-DAcomponents necessary to describe the training set was estimated using cross validation with 1/7 of the data being excluded during each round. This yielded six PLS components. However, the OPLS-DA concentrated all discriminating information into the first component. In this case, only 10% of the original mass spectral variation is responsible for the class separation. As shown in Figure 3 (training set) and Figure 4 (validation set), the classes are completely separated by the first component. A second component is shown in each plot for visualization purposes only as it offers no additional discriminatory power. There is no risk of of overfitting because there is only one predictive component and the data set was split into a training set and validation set. For both sets, 100% selectivity and specificity were obtained.

The results of the chemometric analysis reported here are transparent and interpretable using a few intuitive plots. The degree of class separation is readily apparent from the score plots. In contrast to the multi-step approaches reported previously, OPLS is a single step technique that requires no variable selection as the entire spectrum is utilized. Variable selection should be undertaken with great care as there is a serious risk of throwing away crucial diagnostic information. The inherent danger of univariate t-tests and related nonparametric techniques for variable selection is that such tests do not take into account how variables can combine to form informative and diagnostic patterns. As pointed out by Alexe (*19*), the combined discriminatory power of a group of biomarkers cannot be inferred from their individual p-values.

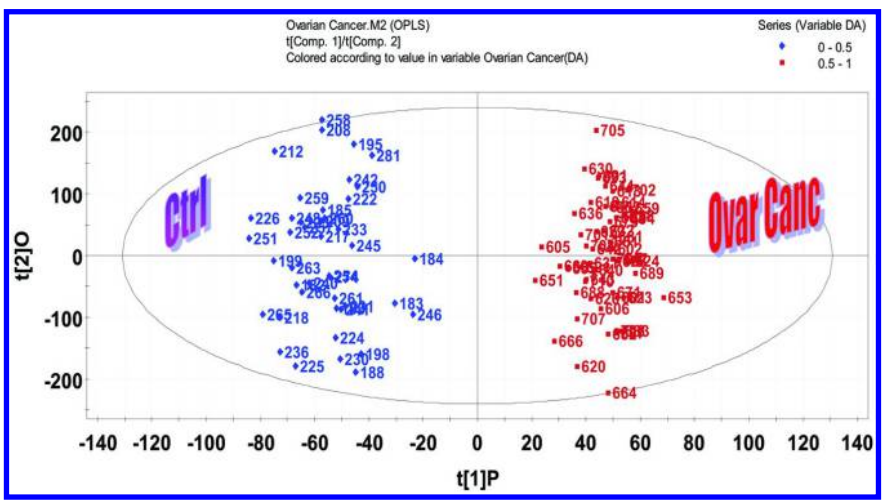

Figure 3. OPLS plot of the 43 controls and the 54 ovarian cancer samples comprising the training set

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch001

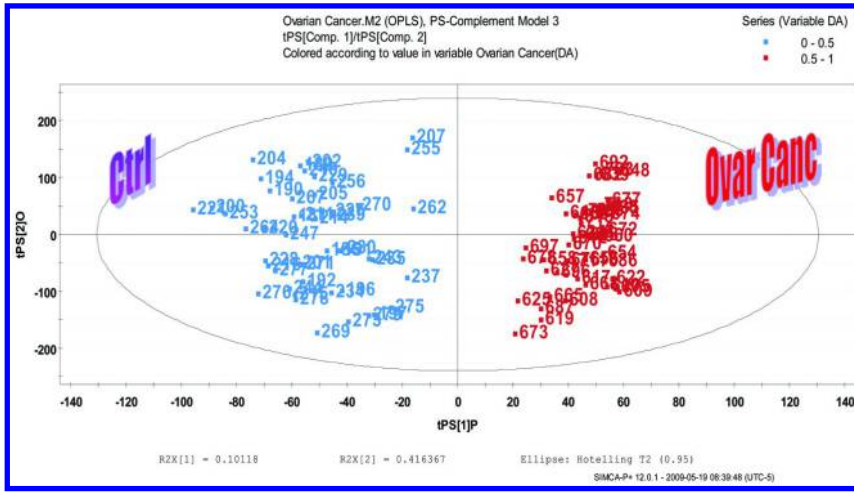

Figure 4. Projection of the the blind validation samples onto the OPLS map of the training set data

Conclusions

Data analytical approaches are practical and understandable when applied to good and relevant data. When utilized in tandem with design of experiments and predictive diagnostics (as well as insight) the results of the analysis are reliable. Truly, Bruce has been a beacon leading us to these conclusions.

However, more work needs to be done also in this area. Semi-automatic procedures need to be developed for data transformations, scaling and centering, and outlier detection. This may require the use of robust estimators. Variable selection is often crucial for a successful data analysis but it must be carried out with great care and insight. Better ways to look at multivariate data in data space to conceptualize and capture the data structure are also needed. Chemometrics in the early days was synonymous with pattern recognition, classification, and linear and nonlinear mapping, and this situation remains today.

Using the principles of experimental design, investigations can be designed to explore this multivariate space more efficiently and informationally than what is presently done with the traditional one factor at a time approach. Experimental design can be viewed as an indirect approach to the study of the joint effects of many factors. Multivariate analysis such as PCA and PLS is the indirect observation of intrinsic latent factors. Teaching and further developing these principles is an important task. Nevertheless, after many analyses of data sets in a variety of projects with the expectation that an important secret will be revealed by the analysis, I have come to the realization that data sets often contain very little information. This absence of information is best summarized by the observation by George Box, “The low quality of a data set is not revealed until it is properly analyzed.”

For chemometrics to prosper, contact between chemists (not only analytical chemists) and chemometricians must be strengthened. It is crucial that relevant and interesting problems be tackled in chemometrics as the success of any field is ultimately defined by the societal and scientific problems that it has tackled and solved. Fortunately, many problems in chemistry and biochemistry (proteomics and metabolomics) can be structured in a form that can be expressed as a mathematical relation; the related mathematical problems are often quite straightforward. Collaborations and research teams will play an important role in the future of chemometrics due to the increasing interdiciplinary nature of problems tackled by chemists. For this reason, publishing chemometrics in chemical or biochemical journals and not being limited to chemometric journals should be the goal of every researcher in this field. But, remember what Einstein said, "Make things as simple as possible but not simpler."

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch001

References

1. Kowalski, B. R.; Bender, C. F. *J. Am. Chem. Soc.* **1972**, *94*, 5632–5639.
2. Kowalski, B. R.; Bender, C. F. *J. Am. Chem. Soc.* **1973**, *95*, 686–93.
3. Wold, S.; Sjostrom, M. SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy. In *Chemometrics: Theory and Application*; ACS Symposium Series 52; Kowalski, B. R., Ed.; American Chemical Society: Washington, DC, 1977; pp 243−282.
4. Wold, S. *Pattern Recogit.* **1976**, *8*, 127–139.
5. Harper, A. M.; Duewar, D. L.; Kowalski, B. R.; Fasching, J. L. ARTHUR and Experimental Data Analysis: The Hueristic Use of a Polyalgorithm In *Chemometrics: Theory and Application*; ACS Symposium Series 52; Kowalski, B. R., Ed.; American Chemical Society: Washington, DC, 1977; pp 14−52.
6. Overton, C. E. *Viertel. Naturf. Ges. Zuerich.* **1899**, *44*, 88–135.
7. Hammett, L. P. *Chem. Rev.* **1935**, *17*, 125–136.
8. Jolliffe, I. T. *Principal Component Analysis*; Springer Verlag: New York, 1986.
9. Wold, S. *Chem. Scr.* **1974**, *5*, 97–106.
10. Wold, S.; Sjostrom, M. *Acta Chem. Scand.* **1998**, *52*, 517–523.
11. Albano, C.; Wold, S. *J. Chem. Soc., Perkins Trans.* **1980**, *2*, 1447–1451.
12. Wold, S. *Technometrics* **1978**, *20*, 397–405.
13. Stahle, L.; Wold, S. *J. Chemom.* **1987**, *1*, 185–196.
14. Trygg, J.; Wold, S. *J. Chemom.* **2002**, *16*, 119–128.
15. Zhu, W.; Wang, X.; Ma, Y.; Rao, M. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 14666–14671.
16. Petrocian, E. F., III; Ardekani, A. M.; Hitt, B. A.; Levine, P. *J. Lancet* **2002**, *359*, 572–577.
17. Sorace, J. M.; Zhan, M. *BMC Bioinform.* **2003**, *4*, 24–43.
18. Baggerly, K. A.; Morris, J. S.; Coombers, K. R. *Bioinformatics* **2004**, *20*, 777–785.

19. Alexe, G.; Alexe, S.; Liotta, L. A.; Petrocian, E. *Proteomics* **2004**, *4*, 766–783.
20. Whelehan, O. P.; Earll, M.; Johansson, E.; Toft, M.; Eriksson, L. *Chemom. Intell. Lab. Syst.* **2006**, *84*, 82–87.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch001

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch002

Chapter 2

Kowalski's Vision on Strength through Diversity: One Researcher's Story

William S. Rayens*

Department of Statistics, University of Kentucky, 725 Rose Street, Lexington, Kentucky 40536-0082
***E-mail: rayens@uky.edu**

Bruce R. Kowalski, founder of CPAC and co-founder of the field of chemometrics, died on December 1, 2012. Although he is often remembered as bringing together the academic talents of chemists and engineers, he also showed enormous vision in his effort to assimilate statisticians and mathematicians into the field. This chapter recalls the journey of one such mathematician-turned-statistician-turned-chemometrician, and is offered in honor of Dr. Kowalski's vision. The piece is written largely in a non-mathematical style and, as such, should be accessible to anyone, regardless of mathematical background.

Introduction

When Dr. Barry Lavine asked me to write this chapter as a summary of my work in chemometrics, my first thought was simply that no one would be interested. While I have worked in chemometrics for over 25 years and have seen the field grow in mathematical and technical complexity during that time, I am just one of many who contributed from the perspective of a non-laboratory scientist. There is nothing particularly special about my story. However, when I reviewed an early proposal for this book I was struck by how that non-scientist role, typically taken up by statisticians and mathematicians, was at risk of being ignored. So I agreed to write a summary of my contributions, as a representation of the many contributions that non-bench scientists have made to the field.

Finding a theme or some unifying principle around which to write was difficult. Initially, I was tempted to write about the meaning of truth through

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

research differed noticeably among authors in the early days of chemometrics. Those with more mathematical training tended toward mathematical proofs and axiomatic reasoning, while those with bench expertise often leaned toward the empirical truth offered by a well-designed, replicable experiment. But that distinction is not only awkwardly abstract, it does not serve well to distinguish chemometric research over time. As the field has evolved we have seen classically-trained chemists, engineers, physicists, psychologists, and computer scientists publish rigorous (enough) statistical arguments in the chemometrics literature, just as we have seen statisticians and mathematicians in the field develop an (at least passable) appreciation for how algorithms and empirical work may already speak directly to the problem they are working on. We have all been influenced for the better by this diversity. What I would argue, and want to remain as an important subtext throughout, is that what we now have as an academic field, was not at all what we started with, and, indeed, the field might not have survived at all if not for this early vision of Kowalski.

In the end, I had to admit that I did not have the philosophical qualifications to pull off a chapter on the nature of truth, so I decided to stay within my comfort area, and offer a summary of my own research - as I had been asked to do in the first place. I have adopted an informal, first-person style with which I am most comfortable. I have also been careful to limit my discussion to just three broad research areas where I have had the opportunity to contribute in the chemometrics literature. We will stick to just those three: mixture surfaces, compositional data analysis, and discriminant analysis.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch002

Mixture Surfaces

My exposure to chemometrics began in the spring of 1986 as I was preparing to defend my dissertation in the mathematics department at Duke University. I was studying under Dr. Donald Burdick, a first-rate scholar and a student of the late, great statistician Dr. John Tukey from Princeton. Dr. Burdick is a rare combination of mathematician, exploratory statistician, and cross-disciplinary researcher, and was the only statistician in the department at that time. Dr. Burdick had a history of partnering with chemists, and his latest venture included me in an ongoing grant he had with the Research Triangle Institute. This grant, which brought me back to graduate school out of a low-brow, high-volume industry job as a quality control engineer, was directed by Dr. Edo D. Pellizzari, a biochemist out of Purdue and long-time researcher at RTI. Our assignment – which paid my graduate student stipend and produced my dissertation topic – was to construct a statistically defensible way of identifying the presence and relative proportions of the nine PCB Aroclors (© Monsanto Corporation) - first in laboratory samples, then in real adipose tissue.

Dr. Burdick already had some ideas about how we would approach this problem. He shared those with me and turned me loose to see what i could come up with. For data we had six chromatograms on each of the nine (pure) PCB Aroclor samples, measured on 93 isomer/congener pairs. And we had 38 laboratory mixtures of these Aroclor classes, with mixing proportions unknown to us, but known by Pellizzari and his staff. We viewed each of those chromatograms as a multivariate observation in 93-dimensional space. That kind of multivariate thinking has long become second nature in chemometrics, but it was not at all common at the outset.

Burdick and I approached this as a problem in higher-dimensional convex geometry. We used the mean of each Aroclor group as a vertex of an eight-dimensional simplex (think higher dimensional triangle or pyramid) in 93-dimensional space. The unknowns also were represented as multivariate observations in 93-dimensional space that, owing to laboratory construction and variability, would not necessarily lie on the surface of, or interior to the simplex. Ideally we wanted to best separate the means in a meaningful way before attempting to classify the unknowns. Fisher's linear (canonical) discriminant analysis (LDA) was the most appropriate tool we had at our disposal at the time, but the basic Fisher optimization problem – and there is still confusion on this point in the chemometrics literature today - can only be solved in practice if the number of variables (93 in this case) is less than the number of total observations (54 here) minus the number of groups (9 for us). Otherwise the pooled within-groups sums-of-squares and cross-products matrix is not invertible. So some initial dimension reduction was necessary and we naively chose to use principal components analysis for this step. We typically used 25 component scores and replaced the original 54 x 93 data set with a 54 x 25 matrix of scores. It would be a decade before the superiority of PLS to PCA for this dimension reduction would be understood (*1*). After this first step the original simplex in R^{93} was replaced by a proxy simplex in R^{25} (Figure 1).

While the idea of what to do next was clear, the devil was in the mathematical details. Once we had the 54 observations on 25 principal component scores, we applied Fisher's LDA to best separate the 9 class means in 8-dimensional space. Thus, after the LDA step, we had 54 observations in 8-dimensional space and our original simplex in R^{93} was now transformed into a nine-faceted simplex in R^{8}. Recall, this came about as the result of two transformations, the initial PCA transformation and then the LDA transformation. It is worth noting that this brand of "discriminant analysis" cannot be easily mapped to any claims of minimized "misclassification probabilities", aside from the two-group case. This confusion is still apparent in many of the papers I referee for the chemometrics literature. To be fair, the language does not really affect how well the procedure performs in practice, but it does confuse the sense in which we think we have done something "optimal." This was, and may be still, a much more natural way for a statistician or mathematician to think, than a chemist.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch002

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch002

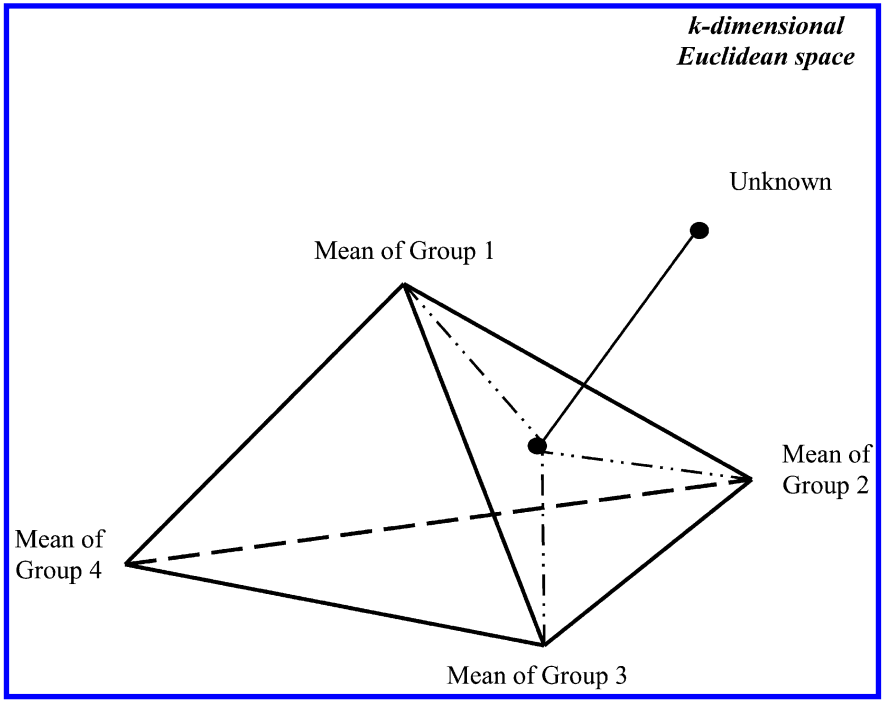

Figure 1. Classical Mixture Surface

The practical estimation problem was then solved as follows. An unknown observation was first encountered in 93-dimensional space, just as the original data were. Then it was twice transformed by the two transformations mentioned above, appearing then in R^8, either interior to, exterior to, or on a facet of the above-mentioned simplex. This is where some of the fun mathematics (convex geometry) came into play in those early days. An observation would be on the simplex if and only if it could be written as a non-negative convex combination of the vertices (group means); that is as a weighted average of the mean vectors, with all non-negative weights that sum to 1. These "barycentric coordinates" were perfect candidates for formal estimates of the underlying mixing proportions. If the observation fell on a facet of the simplex, then there were several possibilities. It could be on a 7-dimensional face proper, or on any of the lower-dimensional facets of that 7-dimensional face. Again, the barycentric coordinates computed for that observation would serve to locate it. It was expected that most of the test observations would fall outside the simplex, however. So it became immediately necessary to be able to identify (and proclaim as unique) the point on the simplex that was closest to that observation, use that as a proxy for the observation, and use the barycentric coordinates of that closest point as estimates of the mixing proportions for the original observation. For young researchers in the area, take note that this part of the problem was challenging. Adopting the right metric for "closest" and solving this closest point problem practically involved some non-

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

trivial programming (remember this is pre-1986) and a clear understanding of the geometry of a k-dimensional simplex. In that sense, it was an ideal problem for a mathematical scientist to work on. In any case, this is how we produced our estimates of the pseudo-unknown mixing proportions.

When Pellizzari and his team revealed the true proportions, we learned that our model had performed exceptionally well. Acting on Burdick's advice, I wrote up the modeling procedure and the results, and submitted a paper to *Technometrics*, which at the time was about the only mainstream statistics journal that would review this kind of hybrid theory-application work. A lot has changed since then. A letter came back from *Technometrics* on July 23, 1986. Too soon to be good news! Indeed, the Editor-elect wrote:

> *You seem to be unaware of the important literature that is available for solving problems of this type. See the list of references given in the Associate Editor's report. Perhaps your methods offer some advantages over the standard methods, but his need to be demonstrated.*

This was followed with the requisite apologies for not being able to accept the paper. I was particularly stymied because here I was finishing a degree in a mathematics department, with only two or three statistics courses in my repertoire, trying to write a dissertation under a well-known statistician, with a primary application in analytical chemistry, and having little feeling for the statistical references the Associate Editor had suggested were missing. However, when Dr. Burdick looked at the references, chief among these being the work by D.M. Titterington on mixtures of distributions, he concluded immediately that the Associate Editor had confused the physical mixtures problem we were working on with problems of a similar name, that the statistical community had been working on for over a decade. That seminal work by Titterington bore little resemblance to our physical mixtures problem. This is when my association with chemometrics began, quite by accident.

Dr. Burdick had, at some point, though I do not know that I ever knew why, contacted Dr. Kowalski and told him about what we were doing. Not long after that I received a letter from Dr. Kowalski that explained a little about the new field of chemometrics, and a little more about the journal he had just launched. He then proceeded to invite me to submit our work there, noting that in his view the field would only grow the way he hoped it would grow if he could get mathematicians and statisticians involved. It was a short letter, but one that clearly spelled out his vision. Dr. Steven Brown acknowledged receipt of our paper on August 18, 1986. The reviews, which were completed by mid-fall, were generally positive and the paper was accepted and ultimately published in the first volume of the *Journal of Chemometrics* (*2*). It is interesting to look back at one of the referee's reports. In her report she starts a process of terminology mapping that I think continues in the journal today, and is almost an inevitable part of an environment that brings together people with such different backgrounds. She wrote:

> *In this article, Burdick and Rayens use principle (sic) components analysis (dimensionality reduction) and discriminant analysis (noise*

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch002

filtering and mean separation) to find the g-1 dimensional space and consequent simplex of a constrained mixture problem.

She went on to rightly take us to task for not understanding the complexity of real PCB data and the limitations of taking data "on the vertices only." She ends up concluding, however that:

... (I) find the thinking involved to be an elegant and well written description of the chemist's intuitive feel for mixture surfaces.

There is little doubt in my mind that when Dr. Kowalski extended an invitation to statisticians and mathematicians, he set in motion a powerful intellectual synergy. While this interaction may have occasionally frustrated chemists and statisticians alike, it forced a collaboration of perspectives. Some of us were forced to appreciate the laboratory environment, even if we did not know which end of a chromatogram the sample went in. Others began to accept the usefulness of mathematical models and the generality provide by "proofs", even if they might still have quietly preferred to think of a well-designed experiment as providing all the generality needed.

We were never asked to apply our model to real adipose data, by the way, and I am not sure why, though I do recall the RTI chemists talking at the time about worries of liability. The real adipose data were associated with a larger problem of lawsuits in the workplace. At issue was whether the PCB contamination found in workers was a result of workplace exposure, or simply a result of exposure through common sources such as the food chain. Knowing which Aroclors were present and in what proportion were critical pieces of that assessment, but as far as I know our model was never put to that final test. I should also point out that before the paper was submitted to *J. Chemom*., Dr. Burdick and I were careful to say in what sense the final barycentric estimates had a formal, statistical interpretation as maximum likelihood estimators under a typical multivariate normal model in the original p-dimensional space. After all, this had to ultimately be part of a dissertation that was approved by a committee of mathematics faculty! A few years down the road I eventually published the mathematical details associated with the convex geometry and the closest-point projection in *The Journal of Mathematical Chemistry* (*3*).

This initial work led to many other opportunities to publish in the chemometrics literature as well as the more traditional statistical literature. Young researchers interested in chemometrics should note that in those early days most statistics departments would not consider chemometrics work as worthy. I have heard many chemists say that they had the same problems within their chemistry departments. While these misconceptions are not completely gone now, they are far less an issue than 25 years ago. The sustained high mathematical level of most (not all) chemometrics research has silenced many of the critics on the statistical side. This is just another positive consequence of a dynamic that Dr. Kowalski set in motion, one I would argue is quite profound and for which Kowalski should receive credit. While the specific arguments defending the quality of specific work had to be made by those of us scattered about the chemometrics world, those

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch002

arguments would not have been successful if not for the generally high statistical quality of papers appearing in the field. For completeness, I end this section with a brief listing of some of my work related to this initial mixture surface model:

- **Error Checking**: As part of that original dissertation, but published separately, we developed a rather formal, but nonetheless useful method of flagging barycentric estimates that were unreliable (*4*). We did this by capturing information in the residual vectors associated with each of the transformations mentioned above: the identification of the original chromatogram in R^{93} with a proxy in R^{25}, then with a best-separated proxy in R^{8}, and the subsequent identification of this proxy in R^{8} with the closest point on the simplex (mixture surface). We developed a "total" residual measure that proved to be surprisingly good (given how non-unique it was) at flagging unreliable observations, and even flagging particular variables in those unusual chromatograms that might be creating a problem. The chemists we were working with at the time traced some of the anomalies we identified back to original laboratory records, and even found evidence of columns that were not correctly cleaned before runs were initiated.
- **Refined Discriminant Analysis Step**: About the time this work was appearing in chemometrics, Dr. Tom Greene and I were working on some purely statistical problems related to the estimation of the between-groups sums-of-squares and cross-products matrix critical to the LDA discriminant transformation step. At issue was the linear pooling assumption nested in the Fisher's LDA transformation and, in particular, how reasonable or unreasonable this might be if the group sizes were close to or even smaller than the number of variables (as was the case in the mixture surface model we introduced above). We submitted the first of what eventually became two papers on this work just before I received a paper to referee, authored by Dr. Jerome Friedman, on "Regularized Discriminant Analysis" (*5*). I contacted the editors, and we ultimately agreed that there was no conflict of interest with the coincidental overlap of submissions and refereeing tasks. Indeed, Friedman's paper was in many ways more elegant than ours because he had wisely employed cross-validation (an idea that went on to become common in the chemometrics community) to produce his pooled estimates, while Greene and Rayens employed a theory-heavy generalized Bayesian context (*6*, *7*). As we all know, RDA became a well-known and important fixture in several chemometics papers in the 1990s. The theory-heavy generalized Bayesian approach did not!
  Several papers submitted to the chemometrics literature in the last decade purport to have this or that new method which does discriminant analysis in a better way. In several of the ones that I have refereed it was clear that the presence of inadvertent pooling ("regularization") was creating the improvement, so, in that sense, there was not really a new discovery. This kind of larger perspective on the statistics literature is a contribution that statisticians have been able to bring to the field. In any case, I was

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch002

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

able to close an intellectual loop here by revisiting the classical mixture surface mentioned above and incorporating some of this new information on covariance pooling. There are many things about that work that will appeal to an applied statistician or mathematician, more than a chemist probably, including the need to develop and employ efficient updating strategies for a matrix inversion that had to be done a very large number of times owing to the cross-validation estimation strategy (*8*). In the end, substantial evidence was produced that the type of ridging suggested by these regularization methods would improve the simplex model that we used to introduce this section.

Compositional Data

In my opinion, the proper analysis of compositional data is still one of the most overlooked areas in chemometrics. Compositional data are non-negative multivariate data on p-variables, whose values sum to 1 across those variables. I encountered them quite organically when thinking about the stochastic properties of barycentric coordinates (see above), as well as in the chromatograms we used for that original work on mixture surfaces. Indeed, it is still simply ordinary in chromatography to adjust for the amount of the sample used to generate a chromatogram by scaling it by the sum of all the peak heights for that observation. This is simple, rational and very intuitive. Mathematically what you end up with is an N by p data matrix (N observations on p features) that sums to the Nx1 **1** vector across columns. What is interesting about these kinds of data is where that scaling now forces them to reside. An observation that was originally free to appear anywhere in R^p (or at least in the positive orthant of R^p) has now been confined to the positive-orthant facet of a simplex in R^p – a simplex that has the elementary vectors as vertices. See Figure 2. Such data are encountered with sobering frequency anytime it seems natural and just intuitive to focus on vectors of relative proportions and not the original data.

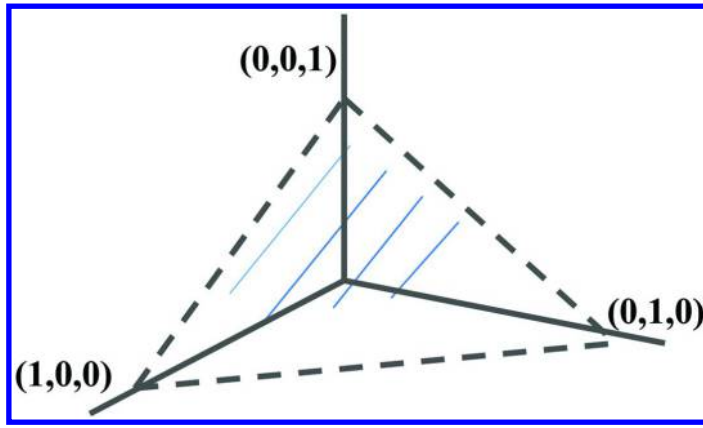

Figure 2. Domain for Compositional Data

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch002

Unfortunately, multivariate data that reside in such a confined space cannot be analyzed using the same multivariate techniques that one uses on unconstrained data. That means no PCA, LDA, PLS, etc.! The statistician John Atchison spent a lot of time trying to educate the statistical community about this. In spite of being brilliant and a wonderful communicator with a huge body of work, he only had limited success changing common practice. This is likely owing to the relative few statisticians who did (and do) non-distributional exploratory multivariate analysis. And even among those, the bar may have been high because the problem created by these constraints is so fundamental that even the most basic statistical constructs – e.g. "correlation" and "covariance" – have to be rethought and redefined. And since nearly all of our common (exploratory) multivariate techniques (e.g. PCA, LDA, PLS) depend on some version of correlation and covariance arrays, the challenge is great.

I will not remake the case for why this is so important, but leave that to papers referenced below or to the excellent, seminal work by Aitchison (*9*). Some intuition might be useful though. Look again at the simplex below, and remember that the data are confined to the facet in the positive orthant. Somewhat naturally it stands to reason that data scatters are not likely to be elliptical or even symmetric in that space, but reshaped (perhaps as banana shapes) because of the geometry of the boundaries. It is like trying to stand up in a small triangular-shaped room. You are likely to have to bend your back to fit. This constraint on how variables can jointly change together requires a completely different way of thinking about correlation and covariance. One simple, but profound, illustration of this is in the so-called negative bias that the usual covariance matrix has to have when computed for compositional data. That is, any covariance matrix computed on closed data has to have at least one negative value in every row. You can see the "bending" that this causes, as the data have to accommodate the geometry. Aitchison proposed many elegant ways out of this dilemma, but his basic plan was based on the application of logarithmic transformations, not on the original data, but on well-chosen ratios formed from the original data. He went on to redevelop in this context a host of the standard multivariate techniques, including log-contrast principal components analysis and log-contrast discriminant analysis.

I became interested in the problem for a variety of reasons. First, the PCA transformation that Burdick and Rayens and used for their original simplex model was clearly no longer ideal, since it was performed on the closed chromatograms that formed our data. But I could not simply apply Aitchison's theory directly since many of those chromatograms had real zeros among the 93 features. The log-contrast theory that Aitchison developed was not able to handle zeros effectively and this became more than a little problem from a practical point of view. While practitioners would do lots of ad hoc things, including replacing 0's with very small values just so transformations were defined, there was no coherent theory to support such practice. Eventually, Rayens and Srinivasan addressed this in a general way in the chemometrics literature by constructing a theory that used the Box-Cox family of transformations to do what the log-contrast was being asked to do by Aitchison (*10*). This family was attractive because it was relatively rich and could be indexed by a single parameter. Further, as that parameter became small, this family would converge to the usual logarithmic

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch002

transformation, so the log-contrast theory of Aitchison existed as a limit point of this larger perspective. We developed a pseudo-likelihood approach and, in essence, chose the indexing parameter to maximize that pseudo-likelihood by way of cross-validation. In a follow-up paper Rayens and Srinivasan developed a much more statistically-detailed set of non-parametric results that employed, among other things, minimum distance estimators and multivariate bootstrap techniques (*11*). These were used to develop point and interval estimates of some conceptualized true set of constrained mixing proportions, around which observations (e.g. scaled chromatograms) were generated in the presence of a specific kind of modeled error. A handful of chemometricians were dealing with the same kind of statistical models and statistical thinking in the literature by this time.

Later, I worked with a student of mine to develop a log-contrast partial least squares (PLS) model, with the refinements that Rayens and Srinivasan suggested in the papers mentioned above (*12*). Other authors have published on compositional data in the chemometrics literature as well, certainly. Still, my distinct impression (based on papers I referee in the field every year) is that the message never really got through to the larger audience. It is not completely clear why the chemometics community has not been more open to the magnitude of possible mistakes that can be made when analyzing compositional data using the standard (multivariate) methods. I think part of the reason, as eluded to above, is that it forces a kind of return to first principles, a rethinking of what correlation and covariance even mean. Another part of the reason may be that the community was never given enough actual evidence that simply applying usual multivariate methods to closed data causes any seriously wrong-headed results from a practical point of view. That is, there has not been enough attention focused on producing empirical evidence that this is a substantive practical problem and in the absence of that empirical evidence the effort involved in recreating one's entire multivariate toolbox is simply not going to be entertained.

There have also been some misunderstandings owing to the diversity of perspectives, even as we celebrate that diversity. I remember a stinging series of email exchanges with a very well-known and brilliant chemometrician who dismissed some of our compositional ideas as unimportant. His primary encounter with compositional data was in an experimental design setting, where he basically chose certain points on that positive-orthant facet of the simplex in an appropriately patterned way. When I tried to explain that designed variability and observed variability were two different things, it became clear he thought the discussion had degenerated to statistical semantics. In fact, they are two quite different things. A good design plan will place points on that facet in a particularly organized way – obviously – since that is why you are designing the points. When one is generating chromatograms and scaling them by total peak height, that observed variability across samples is far less likely to have a nice scatter in that confined space. Indeed, they are much more likely to form those banana-shaped patterns that Aitchison discusses in an introduction to the field. It is those kinds of common compositional scatters that can lead to analysis mistakes. This is still very much an open area (within chemometrics) and I would encourage young statisticians working in the area to consider the potential that is

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch002

still here to educate, and develop better techniques, particularly techniques that take better advantage of the computing that we have now, but did not have twenty years ago.

While working with these compositional data problems it became clear to me that the negative bias issues mentioned above were a part of the much larger problem of how to model rich dependence structures on a simplex. I worked on these problems with a colleague, Dr. Cidambi Srinivasan, and we derived the dependence structures of a family of simplex distributions known as the Liouville family (*13*). I then partnered with another student (Dr. Brian Smith) and we developed other classes of parametric distributions on the simplex, derived from the Liouville family, but having richer dependence profiles (*14*, *15*). In fact, there is still work here to be done in statistics. As most statisticians will know, the so-called Dirichlet distribution is still being used everywhere in Bayesian analysis as a prior (closed of course) because of its nice conjugate properties. It has long been known, though it seems not widely so, that the Dirichlet distribution embodies a kind of "maximum" independence (in the abstract language of that area). So, in that sense it is perhaps a rather odd choice for a prior in many of the instances it is being used.

It should be noted that the adverse effects of closure on higher dimensional data may or may not be as dramatic as with lower dimensional data. This is research yet to be done, however. A proper investigation of this would require one to vary many things, including the number of features, the number of samples, the degree of separation among the groups (for grouped problems), and the measures used to evaluate the quality of the results, to name just a few.

Partial Least Squares and Discriminant Analysis

I end this chapter by revisiting some of my work in partial least squares, and in dimension reduction for the purposes of discrimination. This is probably the area I am most identified with in the chemometrics literature and it has occupied most of my work over the last 10 years. Before summarizing that work, it will be useful to briefly discuss a fundamental difference that sometimes surfaces between how classically trained statisticians and other chemometricians think about exploratory multivariate techniques. When a statistician works with an exploratory technique such as principal components, or canonical (linear) discriminant analysis, or partial least squares, we think of these techniques as having been produced, and hence, validated by some form of constrained optimization problem. And the language and precise statement of that problem matters. For example, PCA results from solving the problem of finding the optimal set of weights that can be used on the original variables to define a new linear compound (score). As we all know, that first new variable is chosen in such a way as to have the maximum variance possible within the paradigm that it has to be formed as a linear combination of the original features. What about the second score? The weights defining that new variable are chosen in such a way that its variance is maximized - subject to the constraint that this new score has to be uncorrelated with the first one. Or is it that the second score is chosen in such a way that its variance is maximized subject to that second vector of weights being orthogonal to the first vector of weights in

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch002

the original variable space? And does it really even matter? For PCA of course we know it does not matter. Both constrained optimization problems produce the same results, and hence, any algorithm to produce one should also produce the other. Unfortunately, PCA is nearly unique in that regard. Neither LDA nor PLS enjoy that property for example.

The issue of what the underlying optimization problem really is – and if that really matters - may perhaps still be one of the most misunderstood multivariate issues in chemometrics research today. If you are a statistician planning on working in this area you are going to find a lot of PCA-like, PLS-like methodologies in the literature. You will need to grow to appreciate just how well some of those work in practice. And you will need to learn to deal with your frustration that they may not really technically be PCA or PLS at all, at least not in the sense you want to think about those methods. I still routinely referee papers for the area that will use a common algorithm for PLS but then make optimality claims that would really only apply if a different set of constraints had been imposed on the original optimization problem, constraints that are not addressed by that algorithm. I bring this up mostly so my brief description of some of my work in this area, which may on the surface seem to just be focused on changing a constraint set, can be better understood.

Let us look at a couple examples in the realm of PLS in particular. In some of my earlier work with Dr. Anders Andersen, an oriented version of PLS was constructed and applied (*16–18*). Intuitively, the goal of the orientation was to allow PLS to be directed away from certain types of variability - with the hope of allowing it to focus more on the (co)-variability of interest. For instance, in functional magnetic resonance imaging (fMRI) information in scans are often confused by signals that come from boundaries, such as the between the brain and skull, and open spaces, such as the nasal cavity. If the variability of the signal in those areas can be well enough understood, then in theory an application of PLS to those data could be orientated away from this undesirable signal and left to focus more on the co-variability seen in the actual brain region being scanned. Mathematically, this orientation arises from the way in which the original PLS optimization problem is (further) constrained. This assumes of course that the new problem, once posed, indeed has a mathematical solution. In this case, actually, the solution was very familiar to statisticians. But to get there, we had to be able to understand the sense in which PLS was already a constrained optimization problem, infuse the additional constraints, and then derive the solution.

Likewise, when I partnered with another student, Dr. Kjell Johnson, to develop formal statistical influence functions for PLS, we quickly realized that we were going to have to develop very different methodologies depending on whether we started with the intention of producing uncorrelated scores or producing orthogonal directions (*19*, *20*). Those problems, at their core, are not just trivially different. They are fundamentally different. Even the well-known solutions to Fishers LDA optimization problem - maximizing among-groups variability relative to pooled within-groups variability - are only "correct" if one is invoking (usually unsaid and often unknown) a kind of within-groups uncorrelated scores constraint. There is no claim to orthogonal directions in the original feature space (like you get for free with PCA). Indeed, much of my work

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch002

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch002

with PLS and variants of PLS are firmly grounded in trying to understand the role of various constraint sets. Granted this may have created some headaches for chemometrics reviewers over the last couple of decades, but the headaches were mutual as I had to also learn to appreciate the gains from chemometric techniques that were PLS (or LDA or whatever) in spirit if not in mathematical detail - even as I mourned the loss in clarity of structure. I suspect this is exactly the kind of headache that Dr. Kowalski wanted to create.

With this introduction as backdrop, I want to briefly summarize some of my work that is directly related to PLS and dimension reduction for purposes of discrimination. In the early 1990s it was clear to chemometrics researchers that PLS in general did a better job of separating groups than did PCA. Using PLS was an option when you had the luxury of knowing your group structure in advance and you simply coded that structure in what seemed to be reasonable way. No one seemed to know why, or at least no one had articulated mathematically why PLS seemed so effective at this task. Making this even more confusing was the fact that some papers in the field used PLS for classification problems directly, even when (as statisticians well know) Fisher's LDA was both defined (on a low-dimensional problem) and optimal. Worse, of course, are those papers that still appear using PCA for classification when a group structure is known. But that is a topic for another time! In any case, I set out to look into the relationship between PLS and discriminant analysis in the early 2000s with another student, (now) Dr. Matt Barker (*1*). It was clear from the outset that any connection would be between PLS and Fisher's LDA (canonical discriminant analysis) since LDA also arose from a constrained optimization problem, and at least employed matrix constructs that were part of the same vernacular as did PLS. Dr. Barker and I ended up establishing a fundamental connection between the two (Fisher's LDA and PLS), showing in essence that the engine that makes PLS go (when coded for classification) is essentially the among-groups sums-of-squares and cross-products matrix from Fisher's LDA! Of course, the details of this connection depended on how the PLS problem was coded for classification and on the set of constraints attending the PLS problem. I believe this established the first mathematically crystal clear understanding of the separating potential inherent in the definition of PLS. I do not want to over-make this point since it is one ripe for misinterpretation and, even worse, can be seen as too self-serving, but I do think that this way of thinking (e.g. dependence on coding, dependence on constraint set) is not necessarily a natural way for a chemist to think, though it may now be a natural way for a chemometrian to think. Dr. Kowalski's purposeful blending of the field, like it or not, is probably responsible for that kind of thinking now being common in our literature.

The real import of my work with Barker is sometimes still misunderstood though. If one has a known group structure and is intent on doing linear discrimination, then LDA is what you should use if it is defined. It is optimal in a sense that one can articulate, and statisticians like to be able to say in what sense a methodology is optimal. It does not really make sense to use PLS for that kind of problem, because PLS is going to be suboptimal. In fact, Barker and Rayens have offered examples of what can happen in those situations when the "other part" of Fisher's optimization problem (the inverse of the pooled

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

within-groups sums-of-squares and cross-products matrix) is ignored, as it is in PLS. This is still an on-going misunderstanding I see somewhat routinely in chemometrics submissions. On the other hand, if one needs to do dimension reduction first, prior to being able to actually apply Fisher's LDA, or perhaps any type of discrimination procedure, ad hoc or well-known, then it becomes undeniably clear that PLS is how you want to do that dimension reduction, and not PCA. When an author argues "it does not matter which one you use", that kind of statement can usually be mapped to a situation where the original total variability was dominated by the among-groups, so that PCA was, in effect, also focusing on discrimination at the dimension reduction stage. Otherwise it has to matter. The mathematics of the underlying optimization problems guarantee that it will. These are precisely the type of constructive conflicts that Dr. Kowalski enabled all those years ago.

I more recently partnered with another student (Liu) and others to extend what Barker and Rayens started (*21–23*). In those papers we both showed in what sense the full Fisher's LDA is a special case of a kind or "oriented" PLS (mentioned above), really just the original PLS optimization problem subjected to a different kind of constraint set. Perhaps more importantly we extended these ideas to the situation where the original covariance arrays are heterogeneous, hence the situation that classical statisticians would recognize as more appropriate for quadratic discrimination than for linear discrimination. Of course, that statement alone can be very confusing since quadratic discrimination was developed in the realm of misclassification probabilities and Mahalanobis distance. There was no sense of "quadratic" discrimination in Fisher's original canonical discrimination problem. But the fact remains that it is not always optimal to use a pooled form of the within-groups sums- of-squares and cross-products matrix to do discrimination. The question Liu and I asked was what if you step outside of the Fisher LDA context and classify the transformed results using a standard misclassification probabilities paradigm. What then is the best PLS-type transformation available to do that initial dimension reduction step? It turns out, of course, that those arguments had to be made not as constrained optimization arguments, but necessarily as misclassification rate arguments. Neither of these last two works appeared in chemometrics journals so they may still be largely unknown to that community. However, there is a fair amount there that would be immediately useful if redeployed within chemometrics.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch002

Summary

In this brief chapter I have tried to offer a coherent cross-section of my 25-plus years of work in chemometrics. I have left out many interesting collaborations, including those in the neurosciences, and particularly those with Dr. Barry Lavine, with whom I have enjoyed a 25-year professional friendship. We have visited each other's institutions and homes, and generated more ideas together - as a chemist and a statistician - than we could ever possibly work on. I want to thank Dr. Lavine for asking me to write this chapter. But most of all I want to acknowledge Dr. Kowalski and the legacy he created, one that was maintained for decades thanks to

Dr. Steven Brown, and continues today, a legacy of robustness through academic diversity.

I am now largely an administrator, having served at both the departmental and university levels, but I have immensely enjoyed being a part of the chemometrics community. I have seen both statisticians and chemists in the field defy the odds and win promotions at all levels at their institutions, even as they have had to sometimes educate their less outward-looking colleagues. I have no idea if Dr. Kowalski thought he would be setting so much in motion with his open invitation to non-chemists back in the 1980s. Indeed, aside from a short visit with him in the very early 1990s I did not know him personally. I cannot say what kind of person he was or even what kind of a scientist he was. But I do know and want to acknowledge that careers were made possible, families were fed, kids were sent to college, all because of a dynamic that Kowalski set in motion nearly thirty years ago. I would even venture to argue that the reason chemometrics has survived for thirty years is because of the strength the field developed as a result of the cross-disciplinary diversity it fostered and then embraced.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch002

References

1. Barker, M.; Rayens, W. S. *J. Chemom.* **2003**, *17*, 166–173.
2. Burdick, D. S.; Rayens, W. S. *J. Chemom.* **1987**, *1*, 157–173.
3. Rayens, W. S. *J. Math. Chem.* **1992**, *9*, 147–160.
4. Rayens, W. S. *J. Chemom.* **1988**, *2*, 121–136.
5. Friedman, J. *J. Am. Stat. Assoc.* **1989**, *84*, 165–175.
6. Greene, T.; Rayens, W. S. *Comm. Stat.* **1989**, *18*, 3679–3702.
7. Rayens, W. S.; Greene, T. *Comp. Stat. Data Anal.* **1991**, *11*, 17–42.
8. Rayens, W. S. *J. Chemom.* **1990**, *4*, 159–170.
9. Aitchison, J. *The Statistical Analysis of Compositional Data*; Chapman and Hall: New York, 1986.
10. Rayens, W. S.; Srinivasan, C. *J. Chemom.* **1991**, *5*, 227–239.
11. Rayens, W. S.; Srinivasan, C. *J. Chemom.* **1991**, *5*, 361–374.
12. Hinkle, J.; Rayens, W. S. *Chemom. Intell. Lab. Syst.* **1995**, *30*, 159–172.
13. Rayens, W. S.; Srinivasan, C. *J. Am. Stat. Assoc.* **1994**, *89*, 1465–1470.
14. Smith, B.; Rayens, W. S. *Statistics* **2002**, *36*, 75–78.
15. Smith, B.; Rayens, W. S. *Statistics* **2002**, *36*, 185–194.
16. Rayens, W. S.; Andersen, A. *Ital. J. Appl. Stat.* **2003**, *15*, 367–388.
17. Rayens, W. S.; Andersen, A. *Chemom. Intell. Lab. Syst.* **2004**, *71*, 121–127.
18. Andersen, A.; Rayens, W. S. *NeuroImage* **2004**, *22*, 728–739.
19. Johnson, K.; Rayens, W. S. *Statistics* **2006**, *40*, 65–93.
20. Johnson, K.; Rayens, W. S. *Comp. Stat.* **2007**, *22*, 293–306.
21. Liu, Y.; Rayens, W. S. *Comp. Stat.* **2007**, *22*, 189–208.
22. Liu, Y.; Rayens, W. S.; Andersen, A.; Smith, C. *J. Chemom.* **2011**, *25*, 109–115.
23. Andersen, A.; Rayens, W. S.; Liu, Y; Smith, C. *Magn. Reson. Imaging* **2012**, *30*, 446–52.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

Chapter 3

The Errors of My Ways: Maximum Likelihood PCA Seventeen Years after Bruce

Peter D. Wentzell*

Department of Chemistry, Dalhousie University, PO Box 15000, Halifax, Nova Scotia B3H 4R2, Canada
***E-mail: peter.wentzell@dal.ca**

The evolution of maximum likelihood principal components (MLPCA) and related techniques is described from a personal perspective, highlighting the author's collaboration with Bruce Kowalski and others. Topics include the motivation for the development of MLPCA, error structures in analytical measurements, and the theoretical principles behind MLPCA. The developments in the field are reviewed and future challenges are outlined.

Introduction

To say that Bruce Kowalski was a leader in the field of chemometrics is certainly an understatement. All of the participants in the symposium from which these chapters are drawn have particular memories of the man who was the motivation for the gathering. For me, Bruce was a catalyst who was able to bring together creative minds in a symbiotic environment, giving birth to new ideas that extended like the spokes on a wheel. The purpose of the present undertaking is reflect on the genesis and growth one of those spokes from a personal perspective, and Bruce's role in this process. The challenge is to weave the right balance of science and personal narrative so as to make this article both useful and insightful. To this end, I will take the rather unscientific approach of chronicling the evolution of ideas that led to the development of maximum likelihood principal components analysis (MLPCA) and the parts that Bruce and others played in this. I will follow this with a brief discussion of how this line of research has grown since then, and the challenges for the future. Given the context of this work, it will no doubt have less rigor than some other treatments

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

and will be rife with the first person. However, it is hoped that this perspective may give new insight into the human element of the scientific subject matter that may be lost in other work.

Errors in Analytical Measurements

Univariate Measurement Errors

From the very beginning of chemical metrology, the estimate of uncertainty has been recognized as an integral part of the analytical measurement, reflecting its information content and therefore its value. The concepts of precision and accuracy are introduced early in our chemical education (*1*), albeit usually in relatively simplistic ways because the complex statistical framework from which they are derived is beyond our appreciation at that point. Even for the simple case of an analytical concentration derived from a univariate measurement, a sophisticated lexicon has emerged to describe the estimation of uncertainty that incorporates experimental design, error budget assessment and the evaluation of figures of merit (*2*). Although such topics typically provoke less enthusiasm in most than the analytical measurement itself, they are nonetheless critical in the practical reporting of analytical results.

In the estimation of uncertainty of analytical concentrations based on univariate measurements, approaches are typically based on replication, error propagation, or a combination of both. Complete replication of analytical procedures through nested designs at multiple levels has the advantage of dissecting all of the sources of variance, but is often impractical. Moreover, without the use of error propagation methods, such approaches may not offer insights into limiting sources of error or allow the estimation of figures of merit such as the limit of detection (LOD). As analytical instrumentation developed increasing levels of sophistication in the latter half of the twentieth century, there was a greater interest in understanding the origins of measurement errors and how these affected concentration estimates, with a view toward improving the quality of results through better design of instruments, experiments, and data treatment methods. By the time I entered graduate school to work with Stan Crouch at Michigan State University in 1982, the groundwork for this had already been established by my academic forebears in this and other laboratories. One of these was Jim Ingle, who had carried out rigorous error analyses on a number of widely used analytical systems (*3–5*). Many of these eventually appeared in the book *Spectrochemical Analysis* (*6*), which stands out for its treatment of the concepts. Thus, there was an emerging systems view of analytical science, further inspired by influences that ranged from the philosophical writings of Robert Pirsig (*7*) to the growing field of chemometrics (*8–10*).

The terms “error”, “uncertainty” and “noise” are sometimes used interchangeably in the discussion of analytical precision, but the distinction is important. The first, of course, refers to the difference between a measured value and a “true” value, typically represented by a defined population mean, while the second is a statistical characterization of the first, usually quantified either directly or indirectly through the error variance. Finally, use of the term “noise”

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

generally implies an ordered series of errors with particular characteristics (*e.g.* photomultiplier noise, drift noise) even though a univariate measurement is a single sampling of this sequence.

For univariate analytical measurements, there is generally an assumption of normality in the distribution of errors, reasonably supported by the Central Limit Theorem. Likewise, for a group of measurements (*e.g.* calibration samples) an assumption of independence in the errors is not unreasonable. Finally, a condition of uniformity in the error variance for a series of measurements (*i.e.* homoscedastic errors) is often assumed, leading to a generic characterization of the errors as "*iid* normal", or independent and identically distributed with normal distribution. Coupled with an assumption of negligible errors in the reference concentrations for calibration samples (a reasonable inference for carefully prepared standards), this scenario fits the well-established framework for linear least squares regression widely used to model instrument response functions. Under the prescribed conditions, least squares produces the so-called maximum likelihood solution, meaning the model for which the observed data lead to the highest value for the joint probability density function. The well-described statistical properties of these estimators can then be extended to establish prediction intervals for unknown samples and evaluate figures of merit such as the LOD. When the assumption of homoscedastic measurement errors is violated and measurement error variances are non-uniform (*i.e.* heteroscedastic errors), these models can be extended by weighted least-squares methods assuming that an appropriate variance estimation function can be found (*11*). These retain the principles of maximum likelihood estimation, as illustrated in Figure 1.

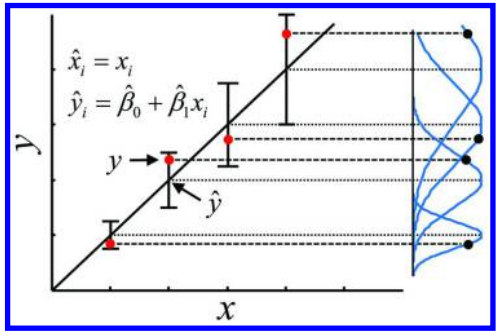

Figure 1. The principle of maximum likelihood estimation illustrated through weighted regression. The fit parameters maximize the product of the probability density functions of the observed points, as shown at the right of the figure.

Multivariate Measurement Errors

The rapid development of analytical instruments and computer hardware in the 1970s and 1980s led to the routine acquisition of vectors of measurements such as spectra and chromatograms. When I entered graduate school (a few years prior to the widespread commercial availability of personal computers), the first project undertaken by many incoming graduate students was the construction of

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

a microcomputer that would acquire the data from the instrument they would ultimately design and use. As the capabilities of instruments expanded, analytical scientists began to discover that they were no longer bound by the traditional univariate calibration curve in terms of the kind of information that could be extracted from measurements. This data-driven approach was pioneered by a number of visionary groups, mainly in the USA and Europe, and intial applications were focused on problems in classification, multivariate calibration and modeling (*10*). Software tools were not yet widely available, but one of the first was the Arthur software package from the Kowalski group (*12*). This was distributed on magnetic tape (see Figure 2) and designed to run on mainframe computers.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

Figure 2. Prof. Roy Bruns (Universidade Estadual de Campinas, Brazil) displays a copy of the ARTHUR program from 1981 at a meeting in 2013.

The initial applications of multivariate analysis to chemical measurements often relied on well-established statistical methods that had found widespread application in other fields, sometimes with adaptations to the peculiarities of chemical systems. However, little attention was paid to the structure of measurement errors in these applications and generally there was an implicit presumption that the *iid* normal error structure assumed for univariate measurements extended to the multivariate case. For the most part, this seemed to work well, but the deficiencies in the assumptions became evident through the variety of data preprocessing tools that were found necessary for different types of data. In some cases, this required scaling of the data in certain ways, while in others, the application of derivative filters or adjustment algorithms such as multiplicative signal correction (MSC) were required to obtain satisfactory results. This was a consequence of data-driven methods that were not also error-driven.

Multivariate chemical data differ substantially from univariate data in the assumptions that can be made about the error structure (*13*). These differences

are reflected in the heteroscedasticity of the measurement error variances and the correlation of measurement errors. For example, while univariate measurements made on a single instrument may be mildly heteroscedastic due to shot noise (*e.g.* fluorescence), this will not have a dramatic effect on univariate calibration. However, a vector of data comprised of analytical measurements with different ranges or units (*e.g.* trace element concentrations) can exhibit grossly heteroscedastic behavior (on an absolute scale) that will impact variance-based multivariate methods. Likewise, univariate measurements made on separate samples are likely to exhibit independent errors, but the errors at adjacent channels in the spectral or temporal domain are likely to have a statistical correlation due to instrument characteristics (*e.g.* cross-talk, drift, filtering). To consider multivariate measurement errors in a comprehensive way, it is necessary to admit that it is possible for the error in each measurement to be correlated to some extent with that for every other measurement. Since many multivariate data analysis methods exploit correlations among the variables, the presence of correlation in the error structure can affect the results.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

Characterization of Multivariate Measurement Errors

There were two principal impediments to the adoption of "error-driven" methodologies in the development of multivariate methods for chemical data. First, there were no tools available to incorporate the measurement error information into the data analysis. A second barrier was the practical difficulties associated with characterizing the measurement errors. In the case of univariate methods when errors can be assumed to be independent, a single variance can be associated with each measurement, but the picture becomes more complex for multivariate methods, where measurement error covariance needs to be considered. These complications relate to the experimental challenges of estimating the error structure, as well as the computational challenges of storing this information.

There are two commonly used approaches to the evaluation of the error structure in first-order (vectorial) measurements. The first is the Fourier transform (FT) which has long been used to examine signals and noise in the frequency domain (or, more generally, in the Fourier domain, since not all signals are collected as a function of time). This a very useful tool in the broad classification of stationary error sources, such as white noise or pink noise (1/*f* noise) through the noise power spectrum (*14–17*). The latter classification is widely observed for contiguous signal vectors in chemistry (*e.g.* spectra, chromatograms) and is often described as drift noise or source flicker noise. Figure 3 shows examples of white noise and pink noise sequences along with their corresponding amplitude spectra in the Fourier domain. While the differences between the two types of noise is evident, it is also noted that the amplitude spectra are quite noisy themselves since they are subject to the stochastic variations of a single time sequence. In general, accurate characterization of measurement errors requires many replicates, which is often a significant experimental impediment.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

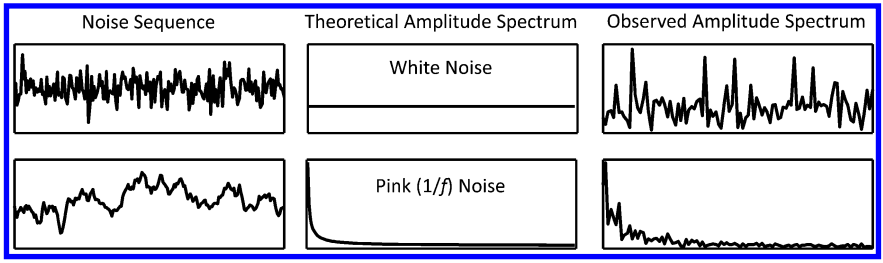

Figure 3. Examples of white noise and pink noise and their amplitude spectra in the Fourier domain.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

The role of low frequency noise sources, such as 1/*f* noise, is important in data analysis since such sources often represent a dominant source of error variance that is not easily eliminated. A good example of this is near-infrared (NIR) reflectance spectroscopy, which became a more powerful tool with the introduction of multivariate calibration to chemistry. Early work reported that an advantage of this method was its high signal-to-noise ratio (S/N) (*18*). While this is true when high frequency noise is examined, NIR methods often suffer from multiplicative and offset noise effects which are not immediately evident and require mitigation through various types of preprocessing.

Fourier analysis can provide some useful diagnostics, but it is limited in the information it can provide for multivariate analysis. While many signals exhibit a correlated error structure amenable to description in the Fourier domain *e.g.* spectra), other data sets are not structured in this way (*e.g.* gene expression levels in a microarray experiment). In addition, the amplitude spectrum does not make apparent localized structures in the original domain, such as changes in error variance (heteroscedasticity) that occur in the signal. Finally, the information in the FT lacks the compatability for integration into most common multivariate analysis methods.

An alternative to the Fourier domain description of multivariate measurement errors is the error covariance matrix, $\mathbf{\Sigma}$, which describes the variance and covariance of all measurement errors in the vector $\mathbf{x}$. If $\mathbf{x}$ is a row vector ($1\times n$), then the $n\times n$ error covariance matrix is defined by Equation 1.

$$\mathbf{\Sigma} = E(\mathbf{e}^{\mathrm{T}}\mathbf{e}) = E[(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu})] \tag{1}$$

In this equation, E designates the expectation operator and $\mathbf{e}$ ($1\times n$) is the vector of measurement errors, defined as the difference between the measurement vector $\mathbf{x}$ and its population mean, $\boldsymbol{\mu}$ ($1\times n$). For this equation, $\mathbf{x}$ has been defined as a row vector rather than a column vector (the more common statistical representation) since matrices of chemical data are typical presented with the measured variables along the columns and samples as the rows. A more practical definition of the sample error covariance matrix based on N replicate measurement vectors, $\mathbf{x}_i$, is given by Equation 2, where $\bar{\mathbf{x}}$ represents the sample mean.

$$\mathbf{S} = \frac{1}{(N-1)}\sum_{i=1}^{N}[(\mathbf{x}_i-\bar{\mathbf{x}})^{\mathrm{T}}(\mathbf{x}_i-\bar{\mathbf{x}})] \tag{2}$$

The error covariance matrix is to multivariate measurements what the error variance is to univariate measurements, but it incorporates an additional level of complexity. As is the case for univariate measurements, the characterization of uncertainty in this manner is critically dependent on the definition of the population, *i.e.* on what constitutes a replicate. For example, a replicate for a spectroscopic measurement might constitute a repeated scan, or a scan after replacement of the sample cell, or a scan after a complete work-up of a replicate sample. For univariate measurements, changes in this definition will only change the magnitude of the variance, but for multivariate measurements, this can result in a complete alteration of the topology of the error covariance matrix. For example, replacement of the sample cell between scans may reveal a large contribution from offset error, a highly correlated error source, which may not be evident for simple replicate scans. Therefore, it is important to define what constitutes an error in the measurement and choose the replication accordingly.

In cases where the measurement vector has a natural order (*e.g.* wavelength, time), visual examination of the error covariance matrix can reveal some qualitative information about the error structure in the data. This is illustrated in Figure 4, where the error covariance matrices that result from some commonly encountered correlated error structures are shown as surface plots. The x and y axes in these plots are the measurement channels and the z axis represents the error variance (diagonal elements) or covariance (off-diagonal elements) for the corresponding measurement channels. The plots are therefore symmetric across the main diagonal. For the case of independent errors (not shown) these plots would appear (ideally) as a diagonal ridge, with uniform values along the diagonal in the case of *iid* normal errors. For correlated errors, however, the off-diagonal elements are not zero. In generating these figures, a small amount of *iid* noise was also added for computational reasons, but the dominant error sources are as indicated. Offset noise is commony observed in chemical measurements, for example when changes in cell position or baseline offset give rise to random vertical shifts of the signal between replicates. This results in errors that are completely correlated, yielding a flat error covariance matrix. Pink noise, or $1/f$ noise, is commonly the result of so-called source flicker (*e.g.* low frequency variations in a light source in optical spectroscopy) or more generally designated as detector drift. The dominant low frequency components of such noise sources results in covariance values that fall off with the distance between channels, independent of the signal shape. Such noise can also exhibit a dependence on signal intensity, as in the case of proportional $1/f$ noise. Such noise might dominate in cases where the signal intensity is proportional to the source intensity, such as in fluorescence, atomic emission or even mass spectrometry. In these cases, the error covariance increases with the magnitude of the signals, as shown in the figure. Another commonly observed error structure is multiplicative noise, sometimes referred to as multiplicative offset noise, which is especially common in IR and NIR spectroscopic methods based on diffuse reflectance. The changes in the signal profile can be interpreted in a simple way as changes in the path length, and therefore result in shifts that are proportional to the signal intensity. Although fundamentally different from proportional $1/f$ noise, this results in a very similar error covariance structure that depends on the magnitudes of the signals involved.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

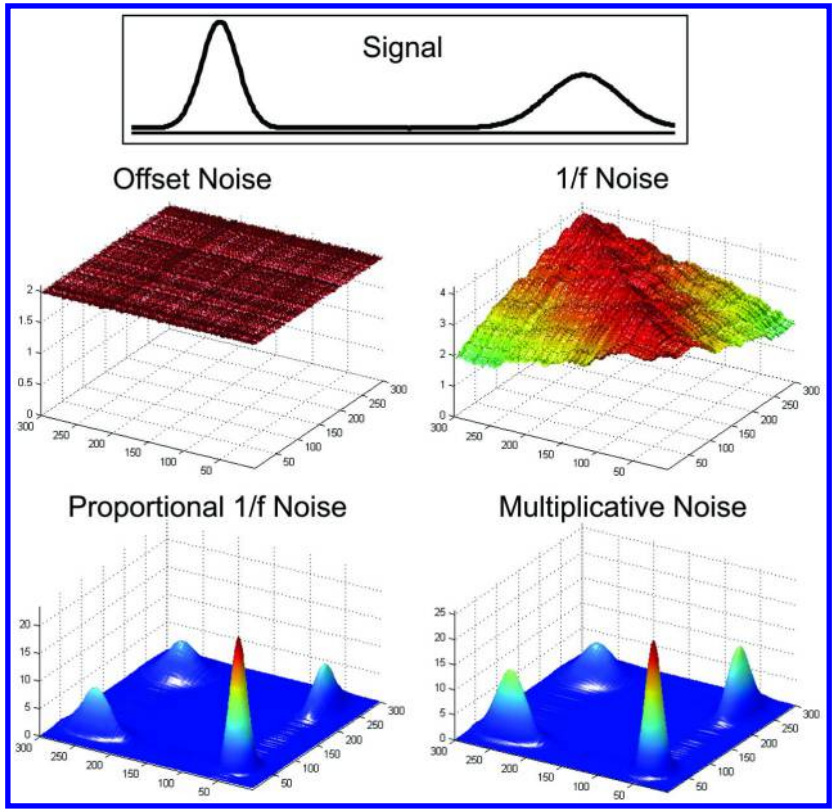

Figure 4. Error covariance matrices for various types of error structures (as indicated) applied to the signal profile shown at the top of the figure.

The error covariance matrix can give insight into the magnitude of the interdependence of measurement errors, but the nature of the interdependence (*i.e.* the degree of correlation) may be obscured in cases where the variance of measurements is signal dependent. To get a clearer understanding of this, the correlation matrix can be examined, where each element is given by Equation 3.

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} \tag{3}$$

Here σ_{ij} is the corresponding element of $\mathbf{\Sigma}$, and σ_i and σ_j are the standard deviations at channels i and j (extracted from the diagonal of $\mathbf{\Sigma}$). By definition, the diagonal of the error correlation matrix will be unity. The off-diagonal elements will indicate the extent of correlation between the errors in the corresponding channels, irrespective of the magnitude of the variance. The error correlation matrices for the examples in Figure 4 are shown in Figure 5. For cases of offset noise and 1/*f* noise, the correlation matrices appear the same as the covariance matrices, since the errors are homoscedastic (same variance at all channels). For the other two cases, some differences are observed. It should be noted that for

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

these simulations, in which the error standard deviation is proportional to the signal, a small amount of *iid* normal noise was added to make the data more realistic in the limiting case where the signal goes to zero. For both sources of proportional errors, the peaks in the error covariance surface become plateaus in the error correlation map, reflecting the degree of correlation. In the case of the multiplicative errors, the surface would be flat like that for the offset errors except for the limiting effect of the small *iid* errors, which dominate when the multiplicative errors become even smaller. This gives rise to the regions near zero between the plateaus (since *iid* errors are uncorrelated) and the diagonal ridge. The error correlation map for proportional 1/*f* noise has a similar appearance, but the plateaus are not as flat and the the off-diagonal plateaus are not as high, reflecting the diminishing correlation with channel number.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

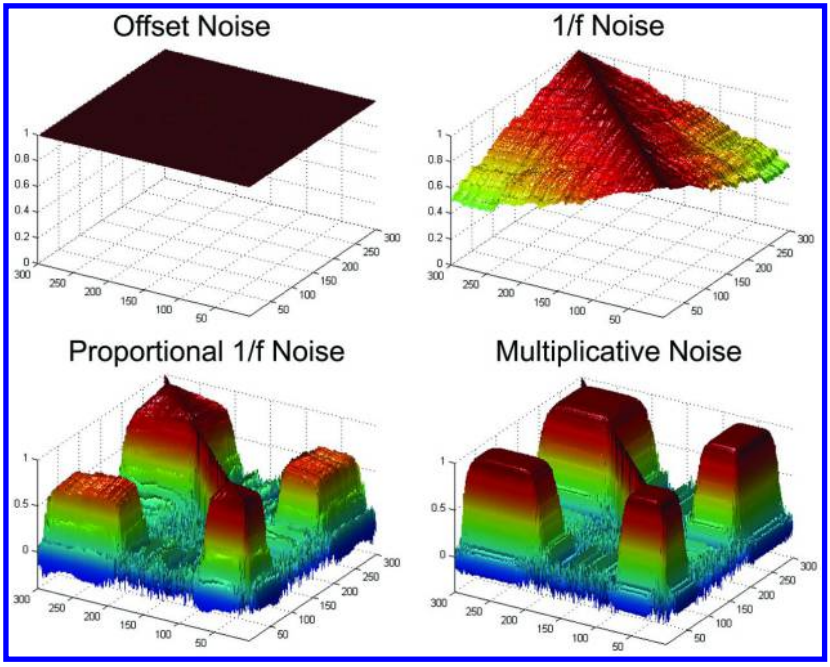

Figure 5. Error correlation matrices corresponding to the covariance surfaces in Figure 4.

Although this kind of qualitative interpretation is useful, it is limited in reality by a number of practical considerations. First, experimental error structures typically result from a combination of error sources, each of which may dominate in different regions. For example, a fluorescence signal may have contributions from proportional 1/*f* noise due to source fluctuations, as well as shot noise (uncorrelated, heteroscedastic) arising from the photomultiplier and offset noise originating from shifts in the baseline. A second consideration is the reliable estimation of the covariance matrix. Estimates of variance and covariance are inherently uncertain in the absence of a large number of replicates, resulting in a

substantial amount of "noise in the noise". Moreover, the noise in the covariance matrix is heteroscedastic, further complicating a clear interpretation. For the simulated examples presented here, 100 replicates were generated, but this is clearly impractical for most experimental measurements. Nevertheless, the error covariance matrix is at the core of the theoretical framework for the maximum likelihood methods described here, so the issue of its estimation will be set aside for the moment.

Maximum Likelihood Principal Components Analysis

A Pilgrimage to Seattle

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

My own less-than-formal introduction to the discipline of chemometrics began in graduate school, some of which was provided by my mentor, Stan Crouch, who dabbled at the edges of the field. However, much of the credit for my initiation has to go to a fellow student, Tom Doherty, who had a voracious appetite for scientific literature and would often engage me in extended discussions on works that he had discovered. Two of these stand out in my mind: Richard Hamming's *Digital Filters* (*19*), which set into motion a fascination of signals and noise, and Edmund Malinowski's *Factor Analysis in Chemistry* (*20*), which opened up the world of multivariate modeling with its clear descriptions of a complex subject. Around that time I also began following Steve Brown's work on Kalman filtering, which integrated the concepts of modeling and noise (*21*). Of course, I was aware of the work coming out of Bruce Kowalski's group, but for some reason it seemed too "hard core" and above my level of comprehension.

Following my post-doctoral research at the University of British Columbia with Adrian Wade, whose work in chemometrics had more of an applied flavor, I moved to my current position in 1989. I continued to work in Kalman filtering and tried to make connections with principal components analysis (PCA) (*22*). I was conscious of the limitations of PCA, particularly with regard to the lack of a consistent approach to preprocessing. Many strategies for pretreating data seemed to be mainly trial-and-error or based on *ad hoc* rules. I became convinced that the key to understanding this resided in the error structure of the measurements. I discussed this with Steve Brown at a conference in Utah in 1994 and he pointed me in the direction of the errors-in-variables literature, which was helpful but somehow disconnected from the PCA framework around which so much of chemometrics was built.

I first met Bruce Kowalski when I presented a seminar at the University of Washington early in 1995. One of my former students, Steve Vanslyke, was doing post-doctoral work with Bruce and had engineered my visit. I was rather intimidated as I did not regard myself as a serious chemometrician at the time and felt that Bruce could easily expose my ignorance. He was a figure who loomed larger than life in my perception but, as I was to find out, he was also someone who saw the potential in new ideas. We made arrangements for me to spend six months of my sabbatical in Seattle, starting in January of 1996. For me, it felt like spending time in Bruce's lab would somehow legitimize me in the field of

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

chemometrics. I can't say whether my visit actually achieved that goal, but it was an entirely memorable and productive sabbatical.

Figure 6 is a photograph of the Kowalski group members at the time I departed in June of 1996. Notably absent from the group picture is Klaas Faber who was doing post-doctoral work with Bruce at the time. Klaas's role in the work that I carried out in Seattle was critical, and it is safe to say that it would not have come to fruition in that time period without his help. This is illustrative of one of Bruce Kowalski's greatest talents, which was the ability to bring together creative minds and inspire them in creative synergy. Klaas and I ate lunch together most days and I soon came to appreciate his encyclopedic knowledge of the literature. He would often point me towards research that was unknown to me and through him I became more familiar with the statistical literature and the text of Magnus and Neudecker on matrix differentials (*23*). At key points in the development of the algorithm where I became stuck, it was often Klaas who was able to push me in the right direction. What follows is a description of the evolution of these ideas.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

Figure 6. Kowalski group members in June of 1996. Left-to-right: Cliona Fleming, Stephen Vanslyke, Chris Stork, Paul Mobley, Astrid Perez-Clark, Bruce Kowalski, Dave Veltkamp, Peter Wentzell.

The Many Faces of PCA

Perhaps one of the most challenging aspects of learning chemometrics for the novice is understanding the principles of PCA. Since PCA (and the concept of latent variables in general) is at the core of many methods for exploratory data analysis, classification. multivariate regression, and mixture analysis, a clear understanding is essential for establishing a foundation of knowledge. However, this can be complicated by the various motivations and interpretations of the method in different contexts, which obscure the differences between what a technique is and what it is intended to do. This extends back to what are generally

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

attributed to be the early origins of PCA. In 1878, Adcock provided what is arguably the first description of the extraction of the first principal component in a two-dimensional space when he developed a method for orthogonal least squares (*24*). In 1901, Pearson independently extended this concept to fitting planes (or hyperplanes) in a multidimensional space (*25*). Hotelling's description in 1933, also independently conceived, was intended to describe the multivariate normal distribution of independent factors for correlated variables (*26*). Further confusion arises from the subtle but important distinction between PCA and factor analysis (FA) and the implementation of PCA through singular value decomposition (SVD).

To be clear, PCA itself defines a method for describing multivariate data in a new space where the new axes result from a rotation of the original axes such that each successively generated basis vector (principal component) accounts for the largest amount of variance in the data not accounted for by the preceding basis vectors, while maintaining orthogonality of the axes in the new space. As such, the new variables (principal components, eigenvectors, loadings) and their values (scores) conform to certain mathematical properties (*e.g.* orthogonality) that are often convenient. SVD describes an algorithm for extracting these principal components and is therefore essentially synonomous with PCA aside from some notational differences. PCA is one way of performing FA, but the latter encompasses a broader range of methods used to model multivariate data.

There are essentially three motivations in the application of PCA to multivariate chemical data. The first, consistent with Hotelling's description, is to provide the parameters associated with the multivariate normal distribution of latent variables describing the samples. This finds application in areas such as multivariate statistical process control (MSPC) and defining class boundaries in classification by SIMCA (soft independent modeling based on class analogy). However, many applications in chemistry do not involve single multivariate normal populations (*e.g.* designed experiments, multiclass separation, kinetic studies). A second motivation is dimensionality reduction or variable compression for exploratory data analysis (visualization), classification, or multivariate calibration. In such cases, no model is necessarily implied, but the minimization of residual variance is a convenient (even if sub-optimal) method to preserve information in the data. Finally, consistent with Pearson's description, PCA can be used for subspace modeling in cases where there is an underlying model defining a hyperplane. This is the case for chemical mixture analysis methods, such as multivariate curve resolution (MCR) and certain instances of multivariate calibration. In these cases, there is a well-defined rank for the underlying model (pseudorank). In MCR, this rank will be equal to the number of independently observable chemical components. For multivariate calibration, this interpretation of PCA is only applicable in cases where there is a defined number of mixture components (*e.g.* in a pharmaceutical preparation). PCA is still applied for calibration of more complex mixtures (*e.g.* foods, petroleum products), but in these cases the primary objective is dimensionality reduction rather than subspace modeling.

Of these interpretations, only the first and the third have defined models that lend themselves to optimization in a maximum likelihood context. The

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

second interpretation of PCA as a general tool for dimensionality reduction does not impose a pre-existing model on the system and application in this manner may have a variety of objectives. While PCA may serve these ends well, other approaches such as variable selection or projection pursuit (PP) may provide superior results for a particular application.

In the modeling of multivariate normal distributions (*i.e.* the first interpretation), PCA will provide the optimal estimation of the underlying latent variables when there are no measurement errors, but when such errors are present, the situation becomes more complex. In such instances, the distributions of the variables are convolved with the distributions of the measurement errors and, since PCA only considers total variance, it will not give the most accurate representation of the latent variables. To solve this problem, Lawley and Maxwell devised maximum likelihood common factor analysis (MLCFA), or simply common factor analysis, to model multivariate distributions in the presence of measurement errors (*27*, *28*). In addition to assumptions of multivariate normality in the variables, their model assumes *iid* normal measurement errors and provides an estimate of the variance of those errors. While there are a few applications of this method in chemistry (*29*, *30*), for the most part chemical data do not satisfy the underlying assumptions of multivariate normality and this approach is not likely to perform better than PCA in the majority of applications.

The principal focus of our work in this area has been in the domain of subspace modeling (*i.e.* the third interpretation of PCA), specifically how to obtain the best estimate of the hyperplane containing the underlying model given known characteristics of the measurement errors. At the time I travelled to Seattle, it had been well known that PCA provides the optimal subspace estimation (in a maximum likelihood sense) for cases of *iid* normal errors, but its deficiencies for more complex error structures had not been addressed. When I arrived at the University of Washington, it was my objective to formulate a more generalized form of PCA that would provide optimal subspace estimation for diverse types of error structures.

It is important to note, in hindsight, that in spite of its name, MLPCA is neither "maximum likelihood" nor is it PCA. Since PCA, by definition, describes a particular decomposition of the data based on variance, any method that does otherwise cannot be PCA, nor will it share all of the properties of a PCA decomposition. Moreover, PCA provides a full set of basis vectors describing the data, even if these are subsequently truncated, whereas MLPCA only makes sense when the number of basis vectors extracted is equal to the rank of the underlying model. For these basis vectors to provide a maximum likelihood estimate of the subspace, the model must be linear with a known dimensionality and the error covariance structure should be exactly known. While the first condition may be met, the second is virtually impossible to achieve and we must generally rely on estimates of the measurement uncertainty rather than population information. For this reason (which has been pointed out to me by numerous statisticians), MLPCA cannot be considered to be a true maximum likelihood method. At the time, however, the name seemed to fit and embodied at least the goals of the method, which were to provide (almost) optimal subspace estimation in a PCA framework.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

Maximum Likelihood Subspace Estimation

The underlying bilinear model for chemical mixture problems can be expressed in a number of equivalent forms. Here, for reasons of mathematical convenience, the general representation is presented as Equation 4 in a manner consistent with SVD notation for a truncated model of rank p.

$$\mathbf{D} = \mathbf{T}_p\mathbf{V}_p^{\mathrm{T}} + \mathbf{E} = \mathbf{U}_p\mathbf{S}_p\mathbf{V}_p^{\mathrm{T}} + \mathbf{E} \tag{4}$$

Here, $\mathbf{D}$ ($m\times n$) is the data matrix, $\mathbf{V}_p$ ($n\times p$) is a set of p ($p<m,n$) orthonormal loading or basis vectors in the column space of $\mathbf{D}$, $\mathbf{T}$ ($m\times p$) is a set of orthogonal score vectors describing the coordinates of the objects (rows of $\mathbf{D}$) in the subspace, and $\mathbf{E}$ ($m\times n$) is a matrix of residuals. The matrix $\mathbf{T}_p$ can be further decomposed into the product of an orthonormal matrix, $\mathbf{U}_p$ ($m\times p$), and a scaling matrix, $\mathbf{S}_p$ ($p\times p$). Although the notation for the latter decomposition is borrowed from SVD, it is not meant to imply that SVD has been carried out; it is simply a convenient way to describe the subspace.

The rotational and scale ambiguities of this decomposition are well-known and, without imposing constraints, there are an infinite number of basis vectors that can define the subspace, even with orthogonality imposed. PCA uses the amount of variance captured to uniquely define the loadings. MCR relaxes the orthogonality constraint, but applies other constraints (*e.g.* non-negativity) to restrict the solution. For the moment, however, no such constraints are imposed here.

Regardless of the representation of the bilinear model, it is important to distinguish between the (true) underlying model and the method used to estimate this space (*e.g.* PCA, SVD, MCR, etc.). To obtain, the "best" model, principles of maximum likelihood estimation are often imposed. As with regression, this approach maximizes the probability density function (PDF) for the observed values. In the general case, there there are no presumed constraints on the distribution of scores or loadings, so the PDF of interest is that of the residuals. For simplicity, it will initially be assumed that there is no correlation of the errors between the objects (rows of $\mathbf{D}$, $\mathbf{d}_{i\bullet}$), but that errors within a row vector may be correlated and/or heteroscedastic, characterized by the row error covariance matrix $\mathbf{\Sigma}_i$ ($n\times n$). This situation is consistent with many analytical measurements in which the rows represent different samples and columns represent a set of variables (*e.g.* spectral intensities at different wavelength channels). In this case, the PDF for the residual vector is given by

$$PDF_i = (2\pi)^{-n/2}\left|\mathbf{\Sigma}_i\right|^{-1/2}\exp\left(-\frac{1}{2}\mathbf{e}_{i\bullet}\mathbf{\Sigma}_i^{-1}\mathbf{e}_{i\bullet}^{\mathrm{T}}\right) \tag{5}$$

where $\mathbf{e}_{i\bullet}$ is a row vector of $\mathbf{E}$ and $\mathbf{\Sigma}_i$ is the corresponding known error covariance matrix for object i. The residual vector is defined as

$$\mathbf{e}_{i\bullet} = \mathbf{d}_{i\bullet} - \hat{\mathbf{d}}_{i\bullet} \tag{6}$$

where $\hat{\mathbf{d}}_{i\bullet}$ is the measurement vector estimated by the model.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Maximum likelihood estimation of the model can be viewed as a two step procedure (a so-called expectation-maximization, or EM, algorithm). In the first step, the model (*i.e.* the basis vectors in $\mathbf{V}_p$ defining the subspace) is assumed to be known, and the question posed is: "What is the value of $\hat{\mathbf{d}}_{i\bullet}$ that gives the largest PDF in Equation 5 for each row of **D**?". In other words, we want to know the best projection of $\mathbf{d}_{i\bullet}$ onto the assumed model. It can be shown through the application of matrix calculus that this is defined by the so-called maximum likelihood projection, given by Equation 7.

$$\hat{\mathbf{d}}_{i\bullet} = \mathbf{d}_{i\bullet}\boldsymbol{\Sigma}_i^{-1}\mathbf{V}_p(\mathbf{V}_p^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\mathbf{V}_p)^{-1}\mathbf{V}_p^{\mathrm{T}} \tag{7}$$

Notwithstanding potential complicating issues such as knowledge of $\boldsymbol{\Sigma}_i$ or its possible singularity, this equation gives the "best" projection of the objects into the assumed subspace. Under conditions where $\boldsymbol{\Sigma}_i$ is a multiple of the identity matrix (*iid* normal errors), this turns into the orthogonal projection used by PCA to generate estimates of the measurements.

The definition of the maximum likelihood projection is important because it generalizes the optimal projection of data onto the model and establishes the principle that this projection is not necessarily the same for all measurement vectors. In this way, it uses prior knowledge about the errors in each measurement vector to make optimal use of the individual elements of the vector, exploiting redundant information in the data structure by emphasizing those dimensions of the error covariance matrix with the smallest uncertainty. In other words, the modeling is error-driven as well as data-driven.

Once the optimal estimates of the measurement vectors have been obtained, a likelihood function can be defined based on the PDF given in Equation 5.

$$L = (2\pi)^{-mn/2}\prod_{i=1}^{m}\left|\boldsymbol{\Sigma}_i\right|^{-1/2}\cdot\prod_{i=1}^{m}\exp\left[-\frac{1}{2}(\mathbf{d}_{i\bullet}-\hat{\mathbf{d}}_{i\bullet})\boldsymbol{\Sigma}_i^{-1}(\mathbf{d}_{i\bullet}-\hat{\mathbf{d}}_{i\bullet})^{\mathrm{T}}\right] \tag{8}$$

The second part of the algorithm involves maximizing this likelihood function subject to the selection of basis vectors in $\mathbf{V}_p$. As is usual in such cases, the negative log-likelihood function is minimized instead, leading to the objective function given in Equation 9.

$$S^2 = \sum_{i=1}^{m}(\mathbf{d}_{i\bullet}-\hat{\mathbf{d}}_{i\bullet})\boldsymbol{\Sigma}_i^{-1}(\mathbf{d}_{i\bullet}-\hat{\mathbf{d}}_{i\bullet})^{\mathrm{T}} \tag{9}$$

This is a generalization of the usual sum of squares to the case of correlated and heteroscedastic errors. In the case of *iid* errors, this minimizes the sum of squares of orthogonal residuals (truncated PCA solution). In the case of uncorrelated errors, Equation 9 reduces to Equation 10, where each residual is weighted by the inverse of the measurement variance.

$$S^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(d_{ij}-\hat{d}_{ij})^2}{\sigma_{ij}^2} \tag{10}$$

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

While the derivation of Equations 7 and 9 are relatively straightforward and allow the problem to be defined, the search for the optimum subspace model is a more challenging problem.

Alternating Least Squares

When I first arrived in Seattle in 1996, much of my initial effort was directed towards finding ways to optimize the subspace model for heteroscedastic independent errors, subject to Equations 7 and 10. This was complicated by the large number of variables to be optimized, my lack of familiarity with the practicalities of nonlinear optimization methods, and the rotational degeneracy in the solution space. Increasingly complex implementations of conjugate gradient methods and other strategies were slow, non-convergent and entirely frustrating. After many weeks of this, the idea of implementing alternating least squares (ALS) emerged. I can't remember how this came about, but it was likely suggested by Klaas Faber in one of our many discussions and perhaps inspired by the residual spirit of Roma Tauler, who had visited Bruce's lab a few years earlier and developed the constrained ALS approach to MCR (*31*).

Although somewhat different from its implementation in MCR, in hindsight the application of ALS to MLPCA, which is described in the original publication and elsewhere (*32*, *33*), was a natural strategy. The basic principle underlying the application, as in the case of MCR, is the symmetry of the bilinear decomposition. If the data matrix in Equation 4 is transposed, the resulting equation is

$$\mathbf{D}^{\mathrm{T}} = \mathbf{V}_p \mathbf{S}_p \mathbf{U}_p^{\mathrm{T}} + \mathbf{E}^{\mathrm{T}} \tag{11}$$

For PCA/SVD the projected data can be obtained either by projection of the rows into the column space or the columns in the row space, since both projections are orthogonal. However, for maximum likelihood subspace estimation, the projection matrix will generally be different in each subspace, since the row error covariance matrix will be different from the column error covariance matrix. For the solution to be valid, the row and column projections must lead to a consistent estimation of **D**, and it is this constraint that is applied to the ALS implementation of the MLPCA algorithm.

The MLPCA algorithm is implemented as illustrated in Figure 7. An initial estimate of the column space, $\mathbf{V}_p$, is obtained, typically using SVD (although other methods could also be used). Note that this requires that the dimensionality of the subspace (pseudorank), p, be known in advance. The measurement rows are then projected into the column space using the maximum likelihood projection in Equation 7 (or alternative equations, as discussed below) and the known (or estimated) error covariance matrix for each row. PCA (SVD) is then carried out on the projected data matrix, $\hat{\mathbf{D}}$, extracting new estimates for $\mathbf{U}_p$, $\mathbf{S}_p$ and $\mathbf{V}_p$. In the next step, a maximum likelihood projection of the columns of the original data are projected into the row space defined by $\mathbf{U}_p$. To do this, an analog to Equation 7 can be used in conjunction with the column error covariance matrices, $\mathbf{\Psi}_j$, as presented in Equation 12.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

$$\hat{\mathbf{d}}_{\bullet j} = \mathbf{U}_p(\mathbf{U}_p^{\mathrm{T}}\boldsymbol{\Psi}_j^{-1}\mathbf{U}_p)^{-1}\mathbf{U}_p^{\mathrm{T}}\boldsymbol{\Psi}_j^{-1}\mathbf{d}_{\bullet j} \tag{12}$$

The column error covariance matrix, $\boldsymbol{\Psi}_j$, is defined in a manner analogous to the row error covariance, $\boldsymbol{\Sigma}_i$, except that the column vectors of **D** are used as indicated in Equations 13 and 14.

$$\boldsymbol{\Sigma}_i = E(\mathbf{e}_{i\bullet}^{\mathrm{T}}\mathbf{e}_{i\bullet}) = E[(\mathbf{d}_{i\bullet} - \boldsymbol{\mu}_i)^{\mathrm{T}}(\mathbf{d}_{i\bullet} - \boldsymbol{\mu}_i)] \tag{13}$$

$$\boldsymbol{\Psi}_i = E(\mathbf{e}_{\bullet j}\mathbf{e}_{\bullet j}^{\mathrm{T}}) = E[(\mathbf{d}_{\bullet j} - \boldsymbol{\mu}_j)(\mathbf{d}_{\bullet j} - \boldsymbol{\mu}_j)^{\mathrm{T}}] \tag{14}$$

Here, $\mathbf{d}_{i\bullet}$ indicates row vector i of the data matrix **D**, $\mathbf{d}_{\bullet j}$ indicates column vector j, and $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$ are the population mean vectors for row i and column j, respectively.

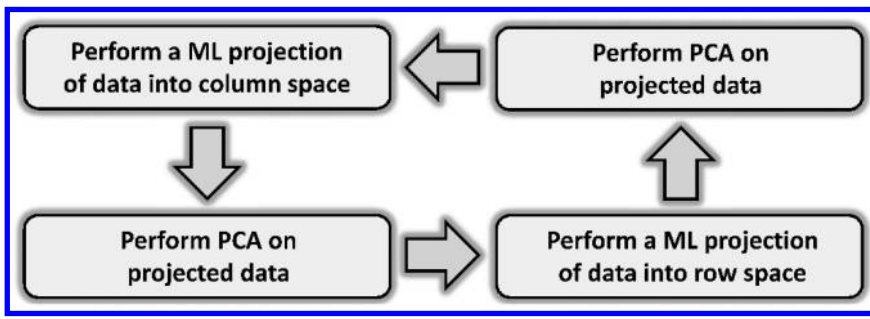

Figure 7. Alternating least squares algorithm for maximum likelihood principal components analysis, in which the data are alternately projected into the row and column spaces.

Following the maximum likelihood projection of the data into the row space defined by $\mathbf{U}_p$, SVD is carried out on the projected data $\hat{\mathbf{D}}$ to generate a new estimate of the column space $\mathbf{V}_p$, at which point the process is repeated, using Equation 7 to project the data into the new column space. The principle behind the ALS algorithm is that, upon convergence, the maximum likelihood projections into the row and column spaces should result in the same projected data matrix and the objective function given in Equation 9 (or analogous equations) should be minimized.

The ALS method described by the above procedure proved to be the simplest and most straightforward method to implement MLPCA. However, problems with this basic implementation soon became apparent. Although the algorithm worked well with independent, heteroscedastic errors and simple correlated error structures, more complex correlated errors were problematic. To understand this, a more detailed examination of common error structures was required.

Classification of Measurement Error Structures

The measurement error covariance structures commonly encountered for two-way data are represented in Figure 8, which shows a cartoon characterization of six general situations. The connectivity of the individual elements (boxes) in each

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

case indicates their independence or correlation, and the color patterns indicate the relationships among the errors in the different rows of the data matrix. Although six cases are shown, in reality there are nine, since there can be an analog to case B in which there is a common pattern of heteroscedasticity down the columns, or analogs to cases D and E in which the errors are correlated only along the columns. In these cases, however, the matrix can always be transposed with no loss of generality in the solution, so they are not considered separately.

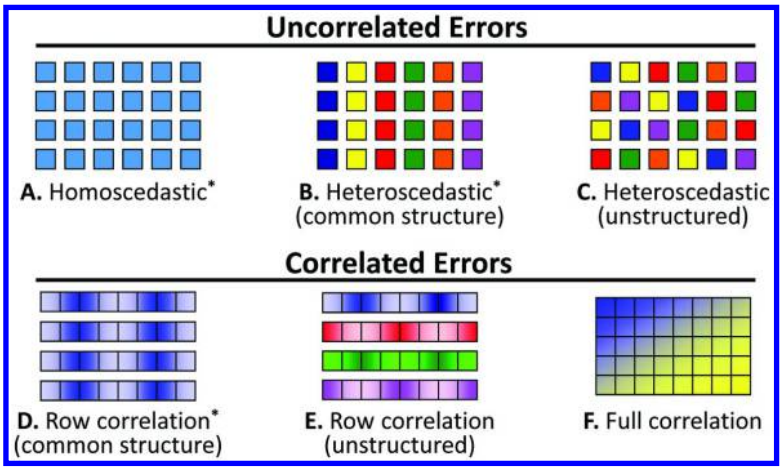

Figure 8. Common error structures observed for two-way data sets. (Adapted with permission from reference (33). Copyright 2009 Elsevier.)

Three of the six cases shown involve independent errors and three involve correlated errors. In addition, the MLPCA model for three of the cases (marked with an asterisk) can be obtained directly without resorting to the ALS algorithm. Case A represents the trivial case of *iid* normal errors, for which the MLPCA solution is the same as the PCA solution, the latter being a special case of the former. Case B is a special case of independent errors in which all of the measurements within a column (or a row) have a common error variance. This occurs fairly often in practice, at least to a first approximation. For example, this might be the case when the columns represent distinct univariate measurements with different units (*e.g.* trace element concentrations), but with a uniform uncertainty for measurements of a given type. Under these conditions, case B can be reduced to case A by simply scaling each column by the inverse of its measurement standard deviation. Traditional PCA can then be applied to obtain the maximum likelihood model in the scaled space. The MLPCA model in the original space can be obtained by generating the PCA projection truncated to rank p, rescaling the projected data to the original space, and applying PCA to generate the rank p estimates of scores and loadings. In traditional applications of PCA, case B is typically dealt with through the use of range scaling or autoscaling (scaling by the standard deviation of each column) in the data preprocessing

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

regimen. This is a rather crude approach, since it essentially assumes that the measurement errors represent a constant proportion of the signal range.

Case D is another situation that can be handled through appropriate pretreatment of the data, although it is less obvious. In this case, the errors are correlated within the rows, but the samples are independent. Additionally, the error covariance matrix is the same for each row. This situation is frequently encountered in cases where sample spectra show a high degree of similarity and are dominated by common sources of correlated errors, such as in IR or NIR spectroscopy. Conceptually, the data are first rotated in the original space to align the common error covariance matrix with the new axes. This effectively removes the correlation and reduces the data to case B. Scaling of the columns then reduces this to case A, at which point traditional PCA can be applied and the rank p projected data can be obtained. After reversing the scaling and rotation for the projected data, the MLPCA solution in the original space can be obtained in the same way as for case B. The mathematics for this procedure has been described elsewhere (*33–35*).

Just as case B has been traditionally dealt with by autoscaling, a variety of preprocessing methods have evolved over the years to treat case C data, particularly in the areas of IR and NIR reflectance spectroscopy. One approach is to apply derivative filters to the spectra, which has the effect of reducing correlated errors and making the data more consistent with case A. While this is somewhat effective, it also has some undesirable characteristics and has been shown to be sub-optimal in the treatment of errors (*35*).

It is important to note that the MLPCA solution gives the maximum likelihood estimate of the rank p subspace that describes the data, but (except in case A), it cannot be interpreted in the same way as the PCA solution in the original space. For example, the variance captured by the MLPCA loadings will not be the same as for PCA, since the error variance has been segregated from the chemical variance, so traditional diagnostics based on variance do not apply. Also, estimates of the data can be generated from the product of scores and loadings ($\hat{\mathbf{D}} = \mathbf{U}_p\mathbf{S}_p\mathbf{V}_p^{\mathrm{T}}$), but cannot be generated by an orthogonal projection of the data onto the loadings ($\hat{\mathbf{D}} \neq \mathbf{D}\mathbf{V}_p\mathbf{V}_p^{\mathrm{T}}$). Instead a maximum likelihood projection is needed (Equation 7).

The remaining cases in Figure 8 (cases C, E and F) can be addressed with the ALS algorithm. Case C represents general heteroscedastic independent errors. Often, the heteroscedasticity observed in chemical data sets is mild and approximations to case A or case B are adequate. Increasingly, however, data are obtained that exhibit more dramatic and unsystematic variations in uncertainty. For example, DNA microarray data are often based on fluorescence ratios and the uncertainty of a given measurement is highly dependent on spot quality, which can vary across a microarray (*36*). In such situations, a companion matrix for $\mathbf{D}$ ($m \times n$) can be created which contains the measurement error variances for each measurement. Under these conditions, the row error covariance matrix, $\mathbf{\Sigma}_i$, will be a diagonal matrix consisting of the elements from row i of the companion matrix. Likewise, $\mathbf{\Psi}_j$, will consist of a diagonal matrix of the variances from column j. With these definitions, the ALS algorithm described in the previous section can be applied to obtain the MLPCA solution.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

In the original development of MLPCA, cases E and F were problematic. At first glance, the application of the ALS algorithm previously described would seem straightforward. However, the algorithm relies on alternating projections into the row and column spaces, and equivalent error information has to be available in each set of projections. For case C (independent errors), the row and column error covariance matrices carry all of the error information, even though it is arranged differently. In case E, the row error covariance matrices carry all of the information, but the information on measurement error correlation is lost when the data are projected into the row space. Since the column error covariance matrices are diagonal, they carry only the variance information and cannot convey the relationships among the measurement channels. Case F is even worse, since neither dimension can carry all of the error information.

I wrestled with this problem for some time, and it was ultimately Klaas Faber (Figure 9) who came up with a solution that was so obvious that I was amazed that I hadn't seen it. I remember quite distinctly that I was presenting the problem in a group meeting in Seattle when Klaas casually asked "Why don't you vectorize it?". I immediately saw the logic and rushed back to the lab to test it. All of the situations in Figure 8 represent special cases of case F, which is the general case where the errors among all of the measurements are correlated. When errors are correlated within the rows and within the colums, the only way to describe the relationships is to unfold the data matrix into a vector and define an error covariance matrix for the vectorized matrix. Depending on which way the data are unfolded, this leads to two forms of the full error covariance matrix, designated as $\boldsymbol{\Xi}$ or $\boldsymbol{\Omega}$ (both $mn \times mn$), analogous to the row and column error covariance matrices. The relationships among the various error matrices is shown in Figure 10 for a simple 2×3 data matrix. With these definitions a vectorized form of the ALS algorithm could be derived and projection equations were obtained for all six cases. These equations are summarized in Table 1. Although the projection equations are needed for the ALS algorithm in only three of the cases, they are also needed in cases B-F to project new data onto the model. In addition to the expanded error covariance matrices, cases E and F also require the definition of matrices $\mathbf{U}$ ($mn \times n$) and $\mathbf{V}$ ($mn \times m$), which are block diagonal forms of $\mathbf{U}_p$ and $\mathbf{V}_p$.

Figure 9. Bruce Kowalski (left) and Klaas Faber in 2006. (Photo courtesy of Susana Navea.)

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

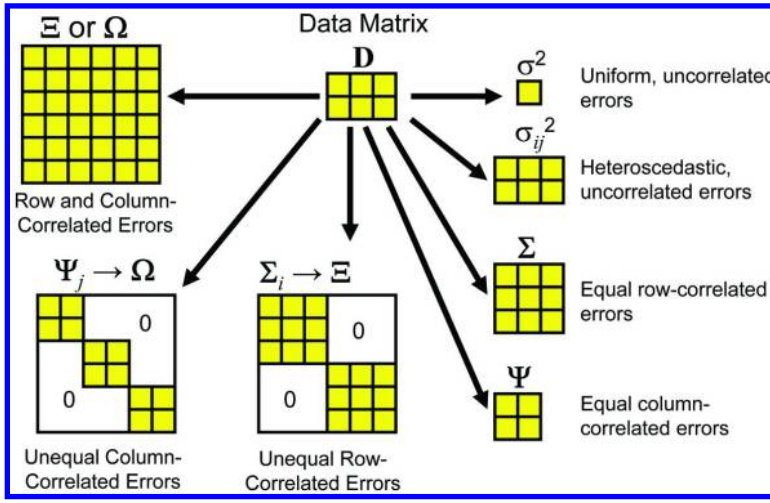

Figure 10. Pictorial representation of different error matrices used in MLPCA. (Adapted with permission from reference (89). Copyright 2005 Elsevier.)

Table 1. Projection equations for MLPCA

Case A	$\hat{\mathbf{D}} = \mathbf{D}\mathbf{V}_p\mathbf{V}_p^{\mathrm{T}}$ $\hat{\mathbf{D}} = \mathbf{U}_p\mathbf{U}_p^{\mathrm{T}}\mathbf{D}$
Case B	$\hat{\mathbf{D}} = \mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{V}_p(\mathbf{V}_p^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{V}_p)^{-1}\mathbf{V}_p^{\mathrm{T}}$ $\hat{\mathbf{D}} = \mathbf{U}_p(\mathbf{U}_p^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\mathbf{U}_p)^{-1}\mathbf{U}_p^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\mathbf{D}$
Case C	$\hat{\mathbf{d}}_{i\bullet} = \mathbf{d}_{i\bullet}\boldsymbol{\Sigma}_i^{-1}\mathbf{V}_p(\mathbf{V}_p^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\mathbf{V}_p)^{-1}\mathbf{V}_p^{\mathrm{T}}$ $\hat{\mathbf{d}}_{\bullet j} = \mathbf{U}_p(\mathbf{U}_p^{\mathrm{T}}\boldsymbol{\Psi}_j^{-1}\mathbf{U}_p)^{-1}\mathbf{U}_p^{\mathrm{T}}\boldsymbol{\Psi}_j^{-1}\mathbf{d}_{\bullet j}$
Case D	$\hat{\mathbf{D}} = \mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{V}_p(\mathbf{V}_p^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{V}_p)^{-1}\mathbf{V}_p^{\mathrm{T}}$ $\hat{\mathbf{D}} = \mathbf{U}_p\mathbf{U}_p^{\mathrm{T}}\mathbf{D}$
Case E	$\hat{\mathbf{d}}_{i\bullet} = \mathbf{d}_{i\bullet}\boldsymbol{\Sigma}_i^{-1}\mathbf{V}_p(\mathbf{V}_p^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\mathbf{V}_p)^{-1}\mathbf{V}_p^{\mathrm{T}}$ $vec(\hat{\mathbf{D}}) = \mathbf{U}(\mathbf{U}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{U})^{-1}\mathbf{U}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}vec(\mathbf{D})$
Case F	$vec(\hat{\mathbf{D}}^{\mathrm{T}}) = \mathbf{V}(\mathbf{V}^{\mathrm{T}}\boldsymbol{\Xi}^{-1}\mathbf{V})^{-1}\mathbf{V}^{\mathrm{T}}\boldsymbol{\Xi}^{-1}vec(\mathbf{D}^{\mathrm{T}})$ $vec(\hat{\mathbf{D}}) = \mathbf{U}(\mathbf{U}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{U})^{-1}\mathbf{U}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}vec(\mathbf{D})$

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

The solutions obtained for cases E and F, while theoretically and practically relevant, can be difficult to implement. Cases where measurement errors are correlated in both the row and column directions can occur, for example, in the case of fluorescence excitation-emission matrices (EEMs). However, the practical realities of trying to obtain reliable estimates for the error covariance matrices and dealing with the expanded matrices in the ALS algorithm can be challenging. Nevertheless these challenges have been met.

MLPCA: The Present

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

MLPCA is not unique in its ability to generate subspace models based on errors-in-variables principles. I discovered this on my return trip from Seattle to Halifax when I stopped in Potsdam, New York to meet with Pentti Paatero of the University of Helsinki, who was visiting with Phil Hopke in the Chemical Engineering Department at Clarkson University. Pentti is well-known for his work on positive matrix factorization (PMF) which has been widely applied, especially in environmental source-receptor modeling problems (*37*–*40*). At the time, the PMF algorithm had been recently expanded to deal with the case of heteroscedastic measurement errors (*40*). Although the algorithms used were different and the error models were not extended as broadly as MLPCA because of the context of the application, the same optimization criteria were being used and it has since been demonstrated that the two approaches produce equivalent results for independent heteroscedastic noise (*41*, *42*). Another technique which had been developed at this time was total least squares (TLS) (*43*, *44*). At the time MLPCA was being developed in Seattle, I was aware of TLS, but it wasn't clear to me how it related to MLPCA since it was formulated as a regression problem. It was some years later, in 2003, when I met with Sabine van Huffel at a Gordon Conference that the parallels between the two methods became apparent. The two methods, although presented differently and incorporating different optimization methods, are essentially equivalent in the objective functions optimized and the error structures employed. A comparison of the two methods was later reported (*45*). Following the publication of our initial work, other algorithmic modifications also appeared. Bro *et al* described the application of the MILES algorithm (Maximum likelihood via Iterative Least Squares EStimation) for MLPCA (*46*), and Nounou *et al* generalized the MLPCA concept to Bayesian PCA (BPCA) (*47*).

Despite these (and likely more) replications of scientific discovery, an advantage of MLPCA is that it is formulated in the traditional PCA framework, allowing it to be readily adapted to a number of chemometric applications that are based on this type of latent variable representation. In this section, some of the developments around MLPCA since 1996 are summarized. While the discussion is mainly focused on applications relevant to chemistry, the subject of maximum likelihood subspace modeling is applicable to virtually all areas of science. A survey of citations of the original MLPCA paper (*32*) reveals interest and relevance from a wide variety of fields, as shown in Figure 11. While many

of these will no doubt be incidental citations, the figure indicates the relevance of the problem.

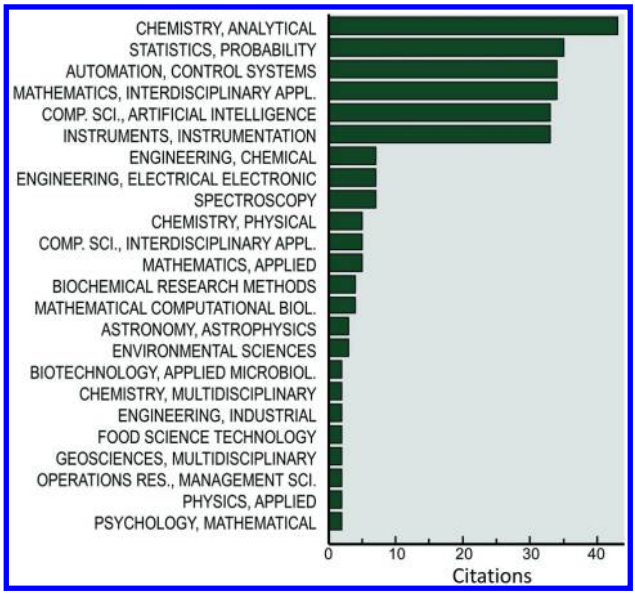

Figure 11. Statistical summary of research areas of articles citing the original work on MLPCA (1997-2015, Web of Science™ results).

Multivariate Calibration and Data Preprocessing

Multivariate calibration was perhaps the most obvious application of MLPCA and consequently was the first that was explored. This resulted in two methods, maximum likelihood principal components regression (MLPCR) and maximum likelihood latent root regression (MLLRR) (*48*). The former was a straightforward extension of principal components regression (PCR) in which the subspace was determined by MLPCA instead of PCA and orthogonal projections were replaced by ML projections. The MLLRR implementation, however, arose from a lunchtime conversation with Bruce Kowalski. As we were discussing ways to incorporate the measurement errors in the reference values, Bruce described the technique of latent root regression (LRR), a technique unfamiliar to me, in which the matrix of predictors, **X**, is augmented with the predictand vector, **y**, and common subspace is found for both. A weakness of LRR is that is that the common subspace is sensitive to the scaling of the variables, so this seemed a natural fit for ML methods. This is yet another example of how Bruce's casual insights revealed a broad and comprehensive understanding of the field.

Both of these methods represented a significant departure from traditional regression methods in that they were error-driven. Unlike most chemometric methods that sought to derive a single universal regression vector, such a vector could only be obtained for these methods if all future samples could be assumed to have a common error structure. Otherwise, the error covariance matrix for the

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

unknown sample had to first be used to project it into the regression subspace before the prediction equation could be applied to the scores. In this way, the prediction was designed to be optimal for the measurement error characteristics of a particular sample.

Since the introduction of these methods, there have been a variety of applications, investigations and improvements of the techniques (*49–56*), but they have not been widely adopted for problems in multivariate calibration. There are a number of reasons likely contributing to this. One of these is the well-established dominance of methods such as partial least squares (PLS) regression, PCR, classical least squares (CLS) regression, and ridge regression (RR), which have long-standing protocols and application histories. This can make it difficult for many new methods to establish a foothold, especially if they involve additional levels of complexity, unless dramatic improvements are demonstrated. Another important restriction on the use of ML methods is the requirement of measurement error information in the form of error covariance matrices. Since this information is not routinely available and cannot generally be obtained retroactively, it is difficult to assess the full extent of improvements that can be achieved across various applications. However, methods have been proposed to estimate error structures in an iterative fashion and these warrant further attention since they have the potential to greatly expand applications (*55*, *56*).

It is also important to recognize that, in spite of the appeal of optimal subspace estimation, ML calibration methods are not necessarily well-suited to all problems, and improvements in prediction ability may be marginal. These improvements arise from two main sources: increased sensitivity through better subspace estimation and reduced errors through ML projections. It is known that maximum sensitivity is achieved when the subspace of the latent variables is aligned with the subspace of the pure component profiles (*e.g.* spectra) (*57*, *58*), but such alignment is not required for a functional calibration model. Near optimal sensitivity may be obtained even if the subspace is suboptimal. A more important factor is likely to be the noise reduction that occurs through the ML projection (as opposed to orthogonal projection) of new samples into the subspace. Since this projection makes optimal use of the measurements with the smallest errors, substantial improvements can be achieved, but only if there is a significant deviation from *iid* normal errors. Another condition that needs to be met to realize these theoretical improvements is a subspace with a well-defined rank. Unlike conventional calibration methods that primarily seek variable compression without the strict imposition of a subspace model, the optimality of ML methods is predicated on the assumption of a model with a known rank. While this is reasonable in the analysis of pre-formulated or well-defined mixtures, it is a difficult presumption to make in the case of complex mixtures such petroleum products or food samples, so the advantages of ML methods are less clear.

An important contribution of calibration methods based on MLPCA to multivariate calibration is an improved understanding of the role of data preprocessing. In the context of ML estimation, many methods can be viewed as strategies to modify the error structure of the data to be closer to the *iid* normal characteristics assumed by conventional methods such as PCR. These strategies range from simple scaling of the data to more complex operations such

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

as derivative filtering (*35*). A generalized approach to whitening of the data matrix was proposed by Martens *et al* (*59*). This reduces to variants of MLPCA in certain instances, although it is limited to particular types of error covariance structures. It is also significant that simple calculations allow the prediction of the post-transformation error covariance matrix after linear operations such as digital filtering or the application of wavelet transforms (*33*, *60*, *61*). This enables the effect of preprocessing methods on the error structure to be evaluated and simplifies the application of MLPCA in alternate domains. Finally, because of its noise filtering capabilities, Hoefsloot *et al* have proposed MLPCA as part of a general form of data preprocessing referred to as maximum likelihood scaling (MALS) (*62*).

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

Exploratory Data Analysis

The principal goal in the application of PCA to exploratory data analysis is the visualization of the relationships among objects or variables in a subspace of lower dimensionality, with the intent of assessing correlations and clustering. PCA is often well-suited to this objective when the between-class variance is larger than the within-class variance since it is a variance-based method. However, especially when these conditions are not met, other methods may provide superior visualization. Methods based on distance metrics, such as hierarchical cluster analysis (HCA), or those based on other measures of projection utility, such as projection pursuit (PP) or independent component analysis (ICA), are also used for this purpose. A key aspect of such exploratory visualization methods is that there is no well-defined objective function for optimal projection other than its utility to the analyst.

Unlike PCA, which is variance-based, MLPCA is model-based, requiring bilinear data with a defined pseudorank and prior knowledge of the error structure. These conditions may be uncertain in a purely exploratory study, and therefore the application of MLPCA for such purposes may not be ideal. Moreover, the application of an MLPCA model with an appropriate dimensionality for visualization (*i.e.* two or three) which is less than the pseudorank of the data can result in erratic projections of the measurements into the subspace and confound visualization efforts. When the pseudorank of the data and the error structure are known, the application of MLPCA for visualization falls into two categories. In the first case, the visualization dimension matches the pseudorank and MLPCA can be an effective visualization tool because it ensures optimal projection of the data into the subspace. In the second case, when the pseudorank of the data is greater than the visualization dimension, MLPCA essentially acts as a noise filtering tool prior to the application of PCA (which is applied to the projected data in the final step of the MLPCA algorithm), and in that way can lead to a more useful visualization of the data. The complex issues associated with the application of exploratory analysis to noisy data have been addressed in the literature (*63*), and this work introduced new definitions of rank associated with data visualization.

While MLPCA itself may not be ideally suited to early stage exploratory data analysis, where issues of rank and error structure may be unresolved, the

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

theoretical developments associated with measurement errors and projections of the data into the subspace have relevance. Given a defined projection method such as PCA, it is straightforward to define how the measurement error covariance matrix will project into the subspace (*33*, *63*). This provides an estimate of the uncertainties in the projected points and is particularly useful in cases of measurement errors with a high level of heteroscedasticity. In these situations, a few measurements with large errors can confound the visual interpretation when the corresponding objects are projected into the wrong regions of the subspace. The extent to which this occurs depends on the orientation of the subspace relative to the original variables and the nature of the projection, but this can be quantified through propagation of error and used by the analyst for a more accurate assessment of the results. One method proposed to do this is has been referred to as a partial transparency projection (PTP) in which the transparency of a projected object is related to its spatial uncertainty (*63*). This means that the visual interpretation is biased towards those objects that are projected with the greatest reliability. This is illustrated in Figure 12, which shows the projected objects (in this case genes from a microarray experiment) before and after the application of the PTP.

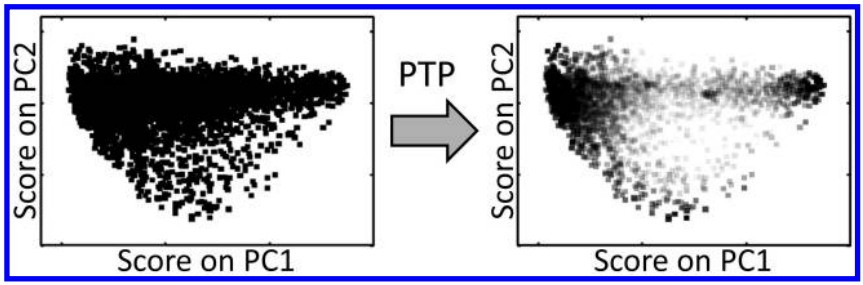

Figure 12. Example of partial transparency transform (PTP) applied to projections of microarray data with heteroscedastic noise. (Adapted with permission from reference (63). Copyright 2012 John Wiley & Sons.)

Multivariate Curve Resolution and Multiway Analysis

MCR represents perhaps the most natural application of MLPCA to problems of multivariate analysis in chemistry. The criterion of a well-defined pseudorank follows directly from the bilinear model imposed for MCR and, unlike calibration or exploratory analysis, accurate subspace estimation is of paramount importance in MCR since this defines the space of the pure component profiles. In spite of this, it was almost ten years before algorithms were developed to unite the two (*64*). There were a number of reasons for this delay that included algorithmic development, access to suitably characterized test data, and the absence of a pressing need. Although heteroscedastic errors were a concern in applications of receptor modeling, these cases were handled by Paatero's PMF software (*40*).

The need to apply ML principles to curve resolution arose fortuitously when I spent four months in the Biology Department at the University of New Mexico, where I was involved in a project studying longitudinal gene expression data for

yeast exiting stationary phase (*65*). These time course experiments employed spotted dual-color microarray data, where measurements of expression data were based on two-channel fluorescence ratios that were highly heteroscedastic due to variations in spot quality (*36*). We were interested in applying MCR to the microarray data, something which hadn't been done before, but the heteroscedasticity necessitated a new approach.

The extension of MLPCA strategies to MCR is quite natural since it simply involves a redefinition of the basis vectors employed (*64*). There is no prerequisite within the MLPCA framework that requires that the vectors defining the subspace be orthogonal. This is done simply for convenience so that the basis vectors are uniquely defined and consistent with PCA, but the projection equations remain the same whether or not the vectors are orthogonal. Moreover, the ALS algorithm that is used in MLPCA is very similar to that used in the most popular implementation of MCR (MCR-ALS), so the coupling is natural. In MCR, the relevant model can be represented as:

$$\mathbf{D} = \mathbf{CP} + \mathbf{E} \tag{15}$$

Here, $\mathbf{D}$ ($m \times n$) is the data matrix consisting of m response profiles at n channels, $\mathbf{C}$ ($m \times p$) is the contribution matrix (typically concentrations) for p components, and $\mathbf{P}$ ($p \times n$) is the matrix containing the pure component profiles (typically spectra). For MCR-ALS, the alternating least squares equations are:

$$\hat{\mathbf{C}} = \mathbf{D}\hat{\mathbf{P}}^{\mathrm{T}}(\hat{\mathbf{P}}\hat{\mathbf{P}}^{\mathrm{T}})^{-1} \tag{16}$$

$$\hat{\mathbf{P}} = (\hat{\mathbf{C}}^{\mathrm{T}}\hat{\mathbf{C}})^{-1}\hat{\mathbf{C}}^{\mathrm{T}}\mathbf{D} \tag{17}$$

with suitable constraints applied on each cycle. For the modified method, the same equations are applied, but the original data matrix is replaced with the maximum likelihood projections before each solution is obtained. Assuming independent, heteroscedastic errors, these are:

$$\hat{\mathbf{d}}_{i\bullet} = \mathbf{d}_{i\bullet}\boldsymbol{\Sigma}_i^{-1}\hat{\mathbf{P}}^{\mathrm{T}}(\hat{\mathbf{P}}\boldsymbol{\Sigma}_i^{-1}\hat{\mathbf{P}}^{\mathrm{T}})^{-1}\hat{\mathbf{P}} \tag{18}$$

$$\hat{\mathbf{d}}_{\bullet j} = \hat{\mathbf{C}}(\hat{\mathbf{C}}^{\mathrm{T}}\boldsymbol{\Psi}_j^{-1}\hat{\mathbf{C}})^{-1}\hat{\mathbf{C}}^{\mathrm{T}}\boldsymbol{\Psi}_j^{-1}\mathbf{d}_{\bullet j} \tag{19}$$

with the notation and definitions as given earlier. Alternative projection equations can be used for different error structures, making the implementation completely general. The new method is referred to as MCR-WALS for "weighted alternating least squares". Although the name "maximum likelihood MCR" was proposed, it was abandoned because the name might imply some statistical advantage of the solution that removed the rotational ambiguity, which is not the case.

An alternative implementation of the MCR-WALS approach is to carry out MLPCA on the **D** matrix first and then apply the standard MCR-ALS to the projected data. This method, referred to as MLPCA-MCR-ALS, has been shown to produce results that are essentially the same, but has the distinct advantage the MLPCA and MCR algorithms can be applied separately without the need to integrate them (*66*, *67*). It has also been shown that MCR-WALS produces

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

the same results as PMF when applied to receptor modeling data (*41*, *42*), and PMF has likewise been applied to other types of data (*39*). This equivalence of final results in spite of some algorithmic differences is not surprising given that the same objective functions are optimized. Although most of the applications of MCR-WALS since the original description of the algorithm (*41*, *42*, *66–75*) have been to Case C error structures (heteroscedastic, independent), it can in principle be applied to any of the noise structures in Figure 8, which may present an advantage over PMF.

The maximum likelihood treatment of measurement errors has also been extended to multiway methods by its inclusion in parallel factor analysis (PARAFAC) models. This extension is natural given that the ALS algorithm is already used for PARAFAC, but is complicated by the introduction of additional orders, which expand the possible error covariance structures given in Figure 8 and necessitate increased memory usage to deal with error covariance matrices of the unfolded data. In 1997, Paatero presented an efficient modification of the PMF method for three-way data, referred to as PMF3 (*76*), although this was limited to independent heteroscedastic errors. In 2002, Bro *et al* applied the MILES technique to PARAFAC models in addition to MLPCA (*46*). Two examples were used; one was a small scale simulated data set employing a fixed error correlation structure along one order, and the other imposed an *iid* error structure to remove artefacts from fluorescence excitation-emission matrices (EEMs). These special cases were generalized to broader error covariance structures by Vega-Montoto in 2003 with the introduction of MLPARAFAC (*77*). This original work was limited to relatively small scale simulations to validate the algorithms, but the theoretical aspects were later expanded (*78*), and the experimental application to fluorescence EEMs in carefully designed experiments was used to compare methods and examine the complexities of evaluating error covariance matrices for three-way data (*79*). Practical difficulties have so far limited applications of MLPARAFAC, but one demonstrated application has been to the direct exponential curve resolution algorithm (DECRA) (*80*, *81*). Because this method converts bilinear data into trilinear data by a shifting procedure, a predictable error correlation structure is introduced, and this can be addressed through MLPARAFAC to achieve improved results (*82*).

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

Modeling of Measurement Noise

One of the biggest barriers to the implementation of the ML methods described in this work is a lack of information about the measurement error structure for most data sets. As noted earlier, the error covariance matrix can be estimated from a series of replicates, but these are often unavailable. Even when replication is carried out with appropriate design considerations, however, the variance and covariance estimates are subject to high variability unless the number of samples is quite large. Theoretical modeling of errors is an alternative in certain cases, for example when counting statistics dominate the uncertainty, but generally the contributions of multiple sources to the overall error makes this impossible. A third option which shows promise is a compromise between these two approaches, where experimental data is used to develop an empirical model that describes the

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

error covariance matrix for a given system in a manner which is consistent with known measurement characteristics.

The importance of measurement noise in analytical measurements has been recognized from the beginning and continues today, but most studies have focused on noise variance estimation and/or frequency characteristics in the context of univariate methods, with little attention paid to correlation properties or the implications for multivariate analysis (*83–88*). Since the development of multivariate methods that make use of measurement error information, the need for a more complete characterization of errors has become more widely appreciated among chemometricians and strategies to characterize measurement errors in the context of multivariate analysis have slowly begun to appear (*36*, *89–98*). Of particular interest are those methods that attempt to obtain empirical models of the error covariance matrix for various systems (*89–91*). These have several advantages that include improved quality of variance/covariance estimates, general applicability in the absence of replicates, possible extension to similar systems, and enhanced insights into the nature and origin of errors. Presently, however, the scope of understanding of measurement errors in most systems has fallen behind the ability to incorporate this information.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

MLPCA: The Future

Having explored the evolution and current state of the art of MLPCA-related methods, it naturally remains to speculate on the future trajectory of these techniques. This is a dangerous undertaking, best left to fortune tellers and economists, but some general comments can be made regarding areas in need of further research.

The first area of need is the rather practical aspect of algorithmic development and dissemination. To my knowledge, there are no commercial software packages that support MLPCA-related software, although some code is freely available from individual researchers. To be fair, impediments to this incorporation include the implementation complexity associated with the different variants of the code for different error structures, the potentially demanding memory requirements, and the slow convergence of some of the ALS procedures. These are all issues that can be addressed through appropriate engineering of the software, however. There is not a large demand for the software at present, in part because it is not readily available, but also because many routine applications continue to be served well by traditional multivariate methods, especially where error strucures do not deviate appreciably from *iid* assumptions or can be treated to approximate this assumption. However, to extend the reach of these tools into ever more challenging domains, more refined methods will be needed to separate the information from the noise.

Research also needs to be pursued to develop better diagnostics associated with MLPCA, especially with regard to pseudorank estimation. Standard practices for the estimation of pseudorank using PCA are well-established and ingrained into most practitioners. While not foolproof, methods such as scree plots, F-tests and residual analysis, among others, offer reasonable arguments for rank estimation in cases where the *iid* approximation is valid. These same methods cannot be reliably

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

employed for heteroscedastic and correlated error structures. In the ideal case, the MLPCA objective function should follow a χ^2 distribution, but this is no longer valid when estimates of the covariance matrix are obtained from replicate data or empirical equations and may therefore exhibit variable levels of uncertainty or bias. Currently, there is no reliable statistic that has been proposed to estimate pseudorank for MLPCA. Even the definition of pseudorank is unclear in certain cases. For example, spectra from three-component mixtures that are characterized by *iid* noise with a random baseline offset could be considered as a four-component system with *iid* errors (baseline offset as one component) or as a three-component system with a correlated noise component. These basic statistical issues remain to be addressed.

Finally, to fully assess the implications of MLPCA-based methods across all potential domains of application requires both an understanding of the error structures and how those errors impact both the traditional and ML methods. In the case of multivariate calibration, for example, estimation equations for prediction intervals for traditional methods based on *iid* assumptions have been available for some time (*99*), but there is a need to extend these to non-*iid* cases and new calibration methods so that the implications of new approaches can be evaluated theoretically as well as empirically. Before this can be done, a better understanding of relevant error structures needs to be obtained. Although this seems like a daunting task, it is becoming more apparent that the measurement errors associated with certain types of analytical methods follow particular patterns of behavior that can be effectively modeled with a limited number of parameters. A number of well-designed investigations into commonly employed instrumental systems is likely highlight this commonality and provide error models that can be used to fully explore the limits of multivariate methods. It may even be possible to generate these models in the absence of replicated data by imposing iterative (*55*, *56*) or localized (*96*) estimation procedures. Ultimately, incorporating the constraints imposed by these models, it may be possible to approach a true ML method, where the measurement error parameters are extracted along with the model parameters.

No one who engaged with Bruce Kowalski for any significant period of time could help but be impacted by his broad philosophical views of science and research, often expressed as maxims, that he conveyed openly and that pervaded the attitudes of those who worked with him. In concluding this work, I will share three of these that have remained with me in the context of the development of MLPCA. One of Bruce's core principles was simplicity ("keep it simple, stupid"- U.S. Navy) and at the root of MLPCA is the goal of removing the confounding ambiguities associated with data preprocessing and method selection for multivariate methods in analytical chemistry. Despite the apparent mathematical complexities, the guiding philosophy behind MLPCA was simply the optimal integration of measurement error information into data analysis. Bruce also believed that chemometrics was the foundation of analytical chemistry, providing the theoretical engine behind a field often regarded as lacking fundamental unifying principles. I got the sense that, of all of his publications, he was most proud of the one entitled "Theory of Analytical Chemistry" since it embodied the core ideas of chemical measurement in a coherent way (*57*).

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Certainly this was one of the most influential publications in my career, and I like to think that the work on MLPCA fits into this scaffold nicely, since it establishes a general framework for the incorporation of errors, which are an inherent part of any measurement. Finally, as a corollary to this view, Bruce believed that it is the role of chemometrics to guide the design of new analytical instruments, rather than the other way around ("theory guides, experiment decides"- I.M. Kolthoff). From that perspective, I like to hope that some of the principles embodied in the work that we have done will go at least a small way to expanding the capabilities of analytical measurements in the future.

Acknowledgments

The author gratefully acknowledges the support of the Natural Sciences and Engineering Research Council (NSERC) of Canada.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

References

1. Harris, D. C. *Quantitative Chemical Analysis*, 8th ed.; W.H. Freeman: New York, 2010.
2. Currie, L. A. *Pure Appl. Chem.* **1995**, *67*, 1699–1723.
3. Ingle, J. D.; Crouch, S. R. *Anal. Chem.* **1972**, *44*, 1375–1386.
4. Ingle, J. D. *Anal. Chem.* **1974**, *46*, 2161–2171.
5. Ingle, J. D. *Appl. Spectrosc.* **1982**, *36*, 588–589.
6. Ingle, J. D.; Crouch, S. R. *Spectrochemical Analysis*; Prentice Hall: Englewood Cliffs, NJ, 1988.
7. Pirsig, R. M. *Zen and the Art of Motorcycle Maintenance*; Wiliam Morrow and Co.: New York, 1974.
8. Kowalski, B. R. *J. Chem. Inf. Comp. Sci.* **1975**, *15*, 201–203.
9. *Chemometrics: Theory and Application*; Kowalski, B. R., Ed.; ACS Symposium Series 52; American Chemical Society: Washington, DC, 1977.
10. Kowalski, B. R. *Anal. Chem.* **1980**, *52*, 112R–122R.
11. Tellinghuisen, J. *Analyst* **2007**, *132*, 536–543.
12. Kowalski, B. R. *Anal. Chem.* **1975**, *47*, 1152A–1162A.
13. Wentzell, P. D. *J. Braz. Chem. Soc.* **2014**, *25*, 183–196.
14. Hieftje, G. M. *Anal. Chem.* **1972**, *44*, 81A–88A.
15. Steyaert, M.; Lambrechts, M.; Sansen, W. *Sens. Actuators* **1987**, *12*, 185–192.
16. Crain, J. S.; Houk, R. S.; Eckels, D. E. *Anal. Chem.* **1989**, *61*, 606–612.
17. Hayashi, Y.; Matsuda, R. *Anal. Chem.* **1994**, *66*, 2874–2881.
18. Wetzel, D. L. *Anal. Chem.* **1983**, *55*, 1165A–1176A.
19. Hamming, R. W. *Digital Filters*, 3rd ed.; Prentice-Hall: Englewood Cliffs, NJ, 1989.
20. Malinowski, E. R. *Factor Analysis in Chemistry*, 3rd ed.; Wiley: Hoboken, NJ, 2002.
21. Brown, S. D. *Anal. Chim. Acta* **1986**, *181*, 1–29.
22. Vanslyke, S. J.; Wentzell, P. D. *Anal. Chem.* **1991**, *63*, 2512–2519.

23. Magnus, J. R.; Neudecker, H. *Matrix Differential Calculus with Applications in Statistics and Econometrics*; Wiley: Chichester, U.K., 1988.
24. Adcock, R. J. *Analyst* **1878**, *5* (2), 53–54.
25. Pearson, K. *Philos. Mag.* **1901**, *2* (11), 559–572.
26. Hotelling, H. *J. Educ. Psychol.* **1933**, *24*, 417–441, 498–520.
27. Lawley, D. N. *Proc. R. Soc. Edinb. (A)* **1940**, *60*, 64–82.
28. Lawley, D. N.; Maxwell, A. E. *Factor Analysis as a Statistical Method*; Macmillan: New York, 1971.
29. De Volder, P.; Hoogewijs, R.; De Gryse, R.; Fiermans, L.; Vennik, J. *Surf. Interface Anal.* **1991**, *17*, 363–372.
30. Moens, P.; De Volder, P.; Hoogewijs, R.; Callens, F.; Verbeeck, R. *J. Magn. Reson. A* **1993**, *101*, 1–15.
31. Tauler, R.; Kowalski, B.; Fleming, S. *Anal. Chem.* **1993**, *65*, 2040–2047.
32. Wentzell, P. D.; Andrews, D. T.; Hamilton, D. C.; Faber, K; Kowalski, B. R. *J. Chemom.* **1997**, *11*, 339–366.
33. Wentzell, P. D. In *Comprehensive Chemometrics*; Brown, S. D.; Tauler, R.; Walczak, B., Eds.; Elsevier: Amsterdam, The Netherlands, 2009; vol 2, pp 507−558.
34. Wentzell, P. D.; Lohnes, M. T. *Chemom. Intell. Lab. Syst.* **1999**, *45*, 65–85.
35. Brown, C. D.; Vega-Montoto, L.; Wentzell, P. D. *Appl. Spectrosc.* **2000**, *54*, 1055–1068.
36. Karakach, T. K; Flight, R. M.; Wentzell, P. D. *Anal. Bioanal. Chem.* **2007**, *389*, 2125–2141.
37. Taiwo, A. M.; Harrison, R. M.; Shi, Z. B. *Atmos. Environ.* **2014**, *97*, 109–120.
38. Brown, S. G.; Eberly, S.; Paatero, P.; Norris, G. A. *Sci. Total Environ.* **2015**, *518-519*, 626–635.
39. Xie, Y.-L.; Hopke, P. K.; Paatero, P. *J. Chemom.* **1998**, *12*, 357–364.
40. Paatero, P.; Tapper, U. *Environmetrics* **1994**, *5*, 111–126.
41. Stanimirova, I.; Tauler, R.; Walczak, B. *Environ. Sci. Technol.* **2011**, *45*, 10102–10110.
42. Tauler, R.; Viana, M.; Querol, X.; Alastuey, A.; Flight, R. M.; Wentzell, P. D.; Hopke, P. K. *Atmos. Environ.* **2009**, *43*, 3989–3997.
43. Van Huffel, S.; Vandewalle, J. *The Total Least Squares Problem: Computational Aspects and Analysis*; SIAM: Philadelphia, PA, 1991.
44. Markovsky, I.; Van Huffel, S. *Signal Process.* **2007**, *87*, 2283–2302.
45. Schuermans, M.; Markovsky, I.; Wentzell, P. D.; Van Huffel, S. *Anal. Chim. Acta* **2005**, *544*, 254–267.
46. Bro, R.; Sidiropoulos, N. D.; Smilde, A. K. *J. Chemom.* **2002**, *16*, 387–400.
47. Nounou, M. M.; Bakshi, B. R.; Goel, P. K.; Shen, X. T. *J. Chemom.* **2002**, *16*, 576–595.
48. Wentzell, P. D.; Andrews, D. T.; Kowalski, B. R. *Anal. Chem.* **1997**, *69*, 2299–2311.
49. Burnham, A. J.; MacGregor, J. F.; Viveros, R. *J. Chemom.* **1999**, *13*, 49–65.
50. Schreyer, S. K.; Bidinosti, M.; Wentzell, P. D. *Appl. Spectrosc.* **2002**, *56*, 789–796.
51. Martinez, A.; Riu, J.; Rius, F. X. *J. Chemom.* **2002**, *16*, 189–197.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

52. Reis, M. S.; Saraiva, P. M. *J. Chemom.* **2004**, *18*, 526–536.
53. Reis, M. S.; Saraiva, P. M. *AIChE J.* **2005**, *51*, 3007–3019.
54. Aboonajmi, M.; Najafabadi, T. A. *Int. J. Food Prop.* **2014**, *17*, 2166–2176.
55. Bhatt, N. P.; Mitna, A.; Narasimhan, S. *Chemom. Intell. Lab. Syst.* **2007**, *85*, 70–81.
56. Bhatt, N.; Narasimhan, S. *Chemom. Intell. Lab. Syst.* **2009**, *98*, 182–194.
57. Booksh, K. S.; Kowalski, B. R. *Anal. Chem.* **1994**, *66*, A782–A791.
58. Faber, K.; Lorber, A.; Kowalski, B. R. *J. Chemom.* **1997**, *11*, 419–461.
59. Martens, H.; Hoy, M.; Wise, B. M.; Bro, R.; Brockhoff, P. B. *J. Chemom.* **2003**, *17*, 153–165.
60. Brown, C. D.; Wentzell, P. D. *J. Chemomet.* **1999**, *13*, 133–152.
61. Leger, M. N.; Wentzell, P. D. *Appl. Spectrosc.* **2004**, *58*, 855–862.
62. Hoefsloot, H. C. J.; Verouden, M. P. H.; Westerhuis, J. A.; Smilde, A. K. *J. Chemom.* **2006**, *20*, 120–127.
63. Wentzell, P. D.; Hou, S. *J. Chemom.* **2012**, *26*, 264–281.
64. Wentzell, P. D.; Karakach, T. K.; Roy, S.; Martinez, M. J.; Allen, C. P.; Werner-Washburne, M. *BMC Bioinf.* **2006**, *7*, 343.
65. Martinez, M. J.; Roy, S.; Archletta, A. B.; Wentzell, P. D.; Anna-Arriola, S. S.; Rodriguez, A. L.; Aragon, A. D.; Quinones, G.; Allen, C; Werener-Washburne, M. *Mol. Biol. Cell* **2004**, *15*, 5295–5305.
66. Dadashi, M.; Abdollahi, H; Tauler, R. *Chemom. Intell. Lab. Syst.* **2012**, *118*, 33–40.
67. Dadashi, M.; Abdollahi, H; Tauler, R. *J. Chemom.* **2013**, *27*, 34–41.
68. Karakach, T. K.; Knight, R.; Lenz, E. M.; Viant, M. R.; Walter, J. A. *Magn. Reson. Chem.* **2009**, *47*, S105–S117.
69. Jaumot, J.; Pina, B.; Tauler, R. *Chemom. Intell. Lab. Syst.* **2010**, *104*, 53–64.
70. Soanes, K. H.; Achenbach, J. C.; Burton, I. W.; Hui, J. P. M.; Penny, S. L.; Karakach, T. K. *J. Proteome Res.* **2011**, *10*, 5102–5117.
71. Decesari, S.; Finessi, E.; Rinaldi, M.; Paglione, M.; Fuzzi, S.; Stephanou, E. G.; Tziaras, T.; Spyros, A.; Ceburnis, D.; O'Dowd, C.; Dall'Osto, M.; Harrison, R. M.; Allan, J.; Coe, H.; Facchini, M. C. *J. Geophys. Res.: Atmos.* **2011**, *116*, D22210.
72. Li, N.; Hopke, P. K.; Kumar, P.; Cliff, S. S.; Zhao, Y. J.; Navasca, C. *Chemom. Intell. Lab. Syst.* **2013**, *129*, 15–20.
73. Paglione, M.; Kiendler-Scharr, A.; Mensah, A. A.; Finessi, E.; Giulianelli, L.; Sandrini, S.; Facchini, M. C.; Fuzzi, S.; Schlag, P.; Piazzalunga, A.; Tagliavini, E.; Henzing, J. S.; Decesari, S. *Atmos. Chem. Phys.* **2014**, *14*, 24–45.
74. Paglione, M.; Saarikoski, S.; Carbone, S.; Hillamo, R.; Facchini, M. C.; Finessi, E.; Giulianelli, L.; Carbone, C.; Fuzzi, S.; Moretti, F.; Tagliavini, E.; Swietlicki, E.; Stenstrom, K. E.; Prevot, A. S. H.; Massoli, P.; Canaragatna, M.; Worsnop, D.; Decesari, S. *Atmos. Chem., Phys.* **2014**, *14*, 5089–5110.
75. Malik, A.; Tauler, R. *Chemom. Intell. Lab. Syst.* **2014**, *135*, 223–234.
76. Paatero, P. *Chemom. Intell. Lab. Syst.* **1997**, *38*, 223–242.
77. Vega-Montoto, L.; Wentzell, P. D. *J. Chemom.* **2003**, *17*, 237–253.
78. Vega-Montoto, L.; Gu, H.; Wentzell, P. D. *J. Chemom.* **2005**, *19*, 216–235.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

79. Vega-Montoto, L.; Wentzell, P. D. *J. Chemom.* **2005**, *19*, 236–252.
80. Antalek, B; Windig, W. *J. Am. Chem. Soc.* **1996**, *118*, 10331–10332.
81. Windig, W.; Antalek, B. *Chemom. Intell. Lab. Syst.* **1997**, *37*, 241–254.
82. Vega-Montoto, L.; Wentzell, P. D. *Anal. Chim. Acta* **2006**, *556*, 383–399.
83. Viciani, S.; Marin, F.; De Natale, P. *Rev. Sci. Instrum.* **1998**, *69*, 372–376.
84. Sperline, R. P.; Knight, A. K.; Gresham, C. A.; Koppenaal, D. W.; Hieftje, G. M.; Denton, M. B. *Appl. Spectrosc.* **2005**, *59*, 1315–1323.
85. Kitajima, A.; Kashirajima, T.; Minamizawa, T.; Sato, H.; Iwaki, K.; Ueda, T.; Kimura, Y.; Toyo'oka, T.; Maitani, T.; Matsuda, R.; Hayashi, Y. *Anal. Sci.* **2007**, *23*, 1077–1080.
86. Du, P. C.; Stolovitzky, G.; Horvatovich, P.; Bischoff, R.; Lim, J.; Suits, F. *Bioinformatics* **2008**, *24*, 1070–1077.
87. Cappadona, S.; Levander, F.; Jansson, M.; James, P.; Cerutti, S.; Pattini, L. *Anal. Chem.* **2008**, *80*, 4960–4968.
88. Dinitto, J. M.; Kenney, J. M. *Appl. Spectrosc.* **2012**, *66*, 180–187.
89. Leger, M. N.; Vega-Montoto, L.; Wentzell, P. D. *Chemom. Intell. Lab. Syst.* **2005**, *77*, 181–205.
90. Karakach, T. K.; Wentzell, P. D.; Walter, J. A. *Anal. Chim. Acta* **2009**, *636*, 163–174.
91. Blanchet, L.; Rehault, J.; Ruckebusch, C.; Huvenne, J. P.; Tauler, R.; de Juan, A. *Anal. Chim. Acta* **2009**, *642*, 19–26.
92. Chen, H.; Bakshi, B. R.; Goel, P. K. *AIChE J.* **2009**, *55*, 2883–2895.
93. Thomas, E. V.; Stork, C. L.; Mattingly, J. K. *IEEE Nucl. Sci. Symp. Conf. Rec.* **2010**, 902–907.
94. Schneider, H.; Reich, G. *Anal. Chem.* **2011**, *83*, 2172–2178.
95. Jones, H. D. T.; Haaland, D. M.; Sinclair, M. B.; Melgaard, D. K.; Collins, A. M.; Timlin, J. A. *Chemom. Intell. Lab. Syst.* **2012**, *117*, 149–158.
96. Wentzell, P. D.; Tarasuk, A. C. *Anal. Chim. Acta* **2014**, *847*, 16–28.
97. Feital, T.; Prata, D. M.; Pinto, J. C. *Can. J. Chem. Eng.* **2014**, *92*, 2228–2245.
98. Bloemsma, M. R.; Weltje, G. J. *Chemom. Intell. Lab. Syst.* **2015**, *142*, 206–218.
99. Olivieri, A. C. *Chem. Rev.* **2014**, *114*, 5358–5378.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch003

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

Chapter 4

Inferring Dioxin Sources in Sediments from a Coastal Harbor Using Multivariate Analysis

L. Scott Ramos,*[,1] Jon Nuwer,[2] and Gregory L. Glass[3]

[1]Infometrix, 11807 N Creek Parkway S, Ste B-111, Bothell, Washington 98011
[2]NewFields, 115 2nd Ave N, Suite 100, Edmonds, Washington 98020
[3]Gregory Glass Consulting, Seattle, Washington 98115
***E-mail: scott_ramos@infometrix.com**

The harbor of Port Angeles in western Washington has been shown to have dioxin concentrations in excess of background levels. A multivariate study was undertaken to understand the nature of contamination. Mixture analysis methods indicated several characteristic patterns that could be associated with identifiable source materials. Principal among these were patterns indicative of burning of salt-laden wood and of pentachlorophenol-related profiles. Spatial interpolation allowed identification of potential source locations upland from the harbor.

Introduction

Harbors and waterways have long histories of industrial contamination. Among environmental pollutants of significant concern are the classes of persistent organic chemicals known as polychlorinated dibenzo-*p*-dioxins and dibenzofurans, commonly referred to as dioxins and furans, respectively. Dioxins and furans have a common origin and enter the environment as by-products of chemical manufacturing and from combustion of materials with chlorine present. The concern derives from the potentially significant toxicity to wildlife and humans of certain isomers of these chemicals and from their persistence in the ecosystem.

The coastal harbor of Port Angeles, WA, has been identified by the Washington State Department of Ecology (DOE) as a priority cleanup and

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

restoration site. Recent investigations have shown dioxin and furan congener toxic equivalency (TEQ) concentrations of surface sediments in excess of background levels across the entire harbor. DOE initiated a study of the occurrence of dioxins and furans in this harbor to understand the nature of the contamination.

Background

Dioxins fall into two main classes of compounds: polychlorinated dibenzo-*p*-dioxins (PCDD) and polychlorinated dibenzofurans (PCDF), the group of both classes generally referred to as 'Dioxins'. The structures shown in Figure 1 describe the possible isomers of the two classes; individual isomers are generally referred to as dioxin congeners.

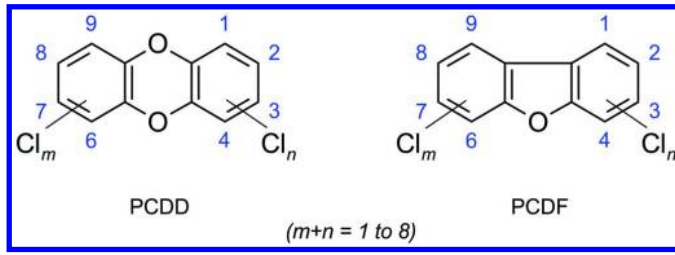

Figure 1. Structures of polychlorinated dibenzo-p-dioxins and furans

Dioxins are formed as by-products of several chemical processes (*1*), including during the chlorine bleaching process at pulp and paper mills. They can occur as contaminants in the manufacture of certain organic chemicals, for example, in the herbicide 2,4,5-T (Silvex) and in Agent Orange. Dioxins are released into the air in emissions from municipal solid waste and industrial incinerators, from smelting and refining operations, as well as from residential, backyard barrel burning. A historical source of dioxins derives from leaded fuel emissions.

The EPA has established methodology (*2*) for the analysis of dioxins using isotope dilution, high resolution capillary column gas chromatography coupled to high resolution mass spectrometry, specifically for the determination of tetra- through octa-chloro dioxins and furans.

Of the 210 possible chlorinations on the dioxin and furan ring systems (dioxins: 75; furans: 135), 17 congeners share the base chlorine substitution pattern of the most toxic compound, 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (2,3,7,8-TCDD), as well as the same mode of toxicity.

Toxic equivalence factors (TEF; see Table 1) have been derived for these 17 congeners in which the TEF (*3*) for 2,3,7,8-TCDD is defined as 1. The TEQ for a sample is defined as the sum of the TEF-adjusted concentration values. Sediment TEQ values for dioxins/furans are generally used by regulatory agencies for decision making purposes, including establishment of sediment cleanup criteria and assessment of both human health and ecological risks.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Table 1. Toxic equivalence factors of PCDD and PCDF congeners (*3*)

Congener	*Abbreviation*	*TEF*
2,3,7,8-tetrachlorodibenzo-*p*-dioxin	2,3,7,8-TCDD	1
1,2,3,7,8-pentachlorodibenzo- *p* -dioxin	1,2,3,7,8-PECDD	1
1,2,3,4,7,8-hexachlorodibenzo- *p* -dioxin	1,2,3,4,7,8-HXCDD	0.1
1,2,3,6,7,8-hexachlorodibenzo- *p* -dioxin	1,2,3,6,7,8-HXCDD	0.1
1,2,3,7,8,9-hexachlorodibenzo- *p* -dioxin	1,2,3,7,8,9-HXCDD	0.1
1,2,3,4,6,7,8-heptachlorodibenzo- *p* -dioxin	1,2,3,4,6,7,8-HPCDD	0.01
Octachlorodibenzo- *p* -dioxin	OCDD	0.0003
2,3,7,8-tetrachlorodibenzofuran	2,3,7,8-TCDF	0.1
1,2,3,7,8-pentachlorodibenzofuran	1,2,3,7,8-PECDF	0.03
2,3,4,7,8-pentachlorodibenzofuran	2,3,4,7,8-PECDF	0.3
1,2,3,4,7,8-hexachlorodibenzofuran	1,2,3,4,7,8-HXCDF	0.1
1,2,3,6,7,8-hexachlorodibenzofuran	1,2,3,6,7,8-HXCDF	0.1
1,2,3,7,8,9-hexachlorodibenzofuran	1,2,3,7,8,9-HXCDF	0.1
2,3,4,6,7,8-hexachlorodibenzofuran	2,3,4,6,7,8-HXCDF	0.1
1,2,3,4,6,7,8-heptachlorodibenzofuran	1,2,3,4,6,7,8-HPCDF	0.01
1,2,3,4,7,8,9-heptachlorodibenzofuran	1,2,3,4,7,8,9-HPCDF	0.01
Octachlorodibenzofuran	OCDF	0.0003

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

Within Port Angeles harbor and surrounding areas, several potential sources of dioxins have been identified. These include paper mills under a series of ownership dating back more than 60 years, combined sewer overflows (CSO) within the city of Port Angeles, hog fuel boilers operated by Rayonier and other mills, as well as deepwater outfalls from the city wastewater treatment facility and the Rayonier Mill.

Earlier attempts at characterizing the source and distribution of dioxins in Port Angeles harbor have largely focused on evaluation of the TEQ values for surface and subsurface sediment samples (*4*). A single measure cannot differentiate disparate sources, therefore, a multivariate approach was considered in which all 17 dioxin and furan congeners would be evaluated together. Such an approach was previously applied to the analysis of dioxins in soils in the Port Angeles vicinity (*5*).

Among the multivariate tools used in the environmental field, factor-based and mixture analysis methods are among the most common. Factor and principal component analysis (PCA) are exploratory methods (*6*, *7*) that seek to find and understand relationships among samples, locate potential outliers or

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

aberrant samples, and describe differences and similarities among measurements. Particularly suited to studies of source apportionment, mixture analysis algorithms (*8*) such as target factor analysis, polytopic vector analysis (PVA) and multivariate curve resolution-alternating least squares (MCR-ALS) can reveal underlying patterns of chemical constituents and then assign contributions of these patterns to sample mixtures.

Multivariate statistics have been used frequently to study dioxins and furans and other types of persistent chemical contamination in the environment. For example, PCA was used in a study of air and soil pollution from municipal solid waste incineration in a large Taiwan city to characterize trends in dioxin patterns across the region (*9*). Differentiation of dioxin sources among residential and industrial locations was facilitated in another study in Taiwan via the use of PCA (*10*).

The distribution of PCDD/F profiles in soils impacted by waste incineration in Madrid was compared using PCA (*11*). Studies in British Columbia combined dioxin congeners with PCBs, then applied PCA to provide insight into source distributions near a pulp mill (*12*). PCA was also used in Korea to demonstrate a progression of the dioxin patterns from that characteristic of an incinerator to that of less polluted sites (*13*) and to evaluate the influence of a municipal waste facility on the contamination of both air and soil in the vicinity (*14*).

A study of Baltic Sea surface sediments used PCA to help identify PCDD/F patterns and associate them with known patterns originating from industrial processes and atmospheric deposition (*15*). Comparisons of dioxin patterns in a North Sea port to those emanating from combustion sources was facilitated by cluster analysis and neural networks (*16*). PCA was also used in a study in coastal British Columbia to examine trends in PCDD/F contamination in sediments and crab (*17*).

Although PCA and associated techniques can help identify likely patterns of dioxins present in a suite of samples, mixture analysis algorithms can, in addition, quantify the contributions of each proposed pattern to the composition of the samples. For example, assessments of dioxins contamination in abandoned military sites in northern Canada were done using PCA and PVA (*18*). In a study of sediment cores in Newark Bay, PVA was used to propose sources of dioxins as originating from combustion, sewage sludge and sources associated with PCBs (*19*). Another study in Tokyo Bay used positive matrix factorization (PMF) to demonstrate the presence of pentachlorophenol in ocean sediments (*20*).

In this study, MCR-ALS was applied to dioxin patterns of a large collection of sediment samples obtained in Port Angeles harbor and considered three aspects:

- Identify distinct PCDD/F congener source signatures present in harbor sediments;
- Determine relative contributions of identified PCDD/F sources to harbor-wide contamination; and
- Use spatial distributions of sediment PCDD/F sources to identify potential upland point source locations.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

Methods

Data Sources

Study Data

Port Angeles Harbor sediment dioxin/furan congener data from several individual data sets were combined for use in the multivariate analysis, including:

- Port Angeles Harbor Sediment Characterization Study (*21*)
- National Park Service Sediment Sampling for Nippon Paper Industries Outfall 002 Replacement (*22*)
- Nippon Paper Industries USA Pulp and Paper Mill Environmental Baseline Investigation (*23*)
- Remedial Investigation for the Marine Environment Near the Former Rayonier Mill Site (*24*)
- Phase 2 Addendum Remedial Investigation for the Marine Environment Near the Former Rayonier Mill Site (*25*)
- Summary of the Log Pond Survey Scoping Effort for the Remedial Investigation (*26*)

Comparison Data

Several past studies of dioxins in environmental and industrial settings were obtained, and the patterns of dioxins ascribed to different sources were extracted and combined into a comparison database. These studies included:

- EPA Inventory of Sources and Environmental Releases (*27*)
- Studies on Canadian hog fuel boilers (*28*, *29*)
- Effluent samples from Rayonier (*24*)
- Stack samples from Rayonier (*30*)
- EPA mill studies (*31*)
- New Zealand soil studies (*32*)
- Denver Front Range residential soils (*33*)
- Chimney soot from home heating systems (*34*)
- PCDDs and PCDFs in PCB Aroclors (*35*)
- Study of treated utility poles (*36*)
- Polychlorophenols in industrial preparations (*37*)

Multivariate Methods

In multivariate analysis, we assume that the patterns associated with each source input add linearly to form the observed pattern. In mathematical terms, a matrix of data (whose dimension is number of samples down the rows by number of measurements across the columns) derived from a single material can be built by multiplying the vector (or list) containing the amounts of this material in

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

the different samples by the vector that represents the pattern of measurements for that material. If there are two materials, then the data would be formed by multiplying a table of compositions (of size number of samples by the two columns of compositions of the two materials) by the table containing the two patterns (one row of numbers for each material). Data originating from more than two sources would come from similarly larger composition and profile tables.

PCA is a multivariate projection technique which decomposes such a matrix (X) into the product of two underlying matrices—the scores matrix T and the transpose of the loadings matrix L:

$$X = T * L^T \tag{1}$$

The PCA decomposition orders the scores and loadings so that each successive combination explains a decreasing amount of the variance in the data. Thus, later columns of the T and L matrices contain mostly noise. If only the first k columns of the scores and loadings matrices are considered relevant and retained, then

$$X \approx \hat{X} = T_k * L_k^T \tag{2}$$

where $\hat{X}$ is an estimate of X, and the dimensionality of the data matrix is said to have been reduced. The columns of L are the loadings, or principal components, the new factors which are linear combinations of the original variables; they are also the eigenvectors of X^TX, where ^{T} is the transpose operator. The first loading, the m elements of the first column of L, indicates how much each original variable contributes to the first principal component, PC1. The scores matrix T is the projection of the sample vector onto the axes defined by the eigenvectors. Each sample has a coordinate on each new axis; the columns of T contain these coordinates.

Although PCA may indicate the presence of multiple relevant loading vectors, they are abstract vectors and do not in general correlate to real patterns. A mixture analysis algorithm, on the other hand, attempts to discover the true underlying patterns and the contributions of these patterns to the samples in the data matrix. For example, in the MCR-ALS algorithm (*38*), the process begins by estimating the source patterns S, and, using matrix algebra, to estimate the compositions C.

$$X = S * C^T \tag{3}$$

To assure that the processing converges to a meaningful solution, after each estimation step, constraints are applied to the newly estimated data. There are many forms the constraints can take; the most common are to apply non-negativity to both matrices: we assume that the intensities in the measurements cannot be less than zero and we also assume that the proportions of the patterns that make up the compositions must also be zero or positive. By applying these constraints, the iterations through the steps of estimating first the patterns matrix and then the compositions matrix will eventually converge to a solution where the patterns and compositions should allow meaningful interpretation of the data.

Most of the various mixture analysis algorithms produce similar results (*8*). For this study, the MCR-ALS algorithm was used.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

Results

The distribution of dioxin/furan TEQ in surface sediments of the harbor suggests two main upland source regions: industrial properties of the western harbor and the former Rayonier Mill, located along the southern shoreline at the mouth of the harbor. The highest TEQs are found along the western harbor shoreline, with concentrations decreasing with distance into the central and outer harbor (Figure 2). TEQs in the surface sediment samples collected within the western harbor lagoon were among the highest observed. Upland facilities along the western harbor shoreline potentially responsible for this dioxin include a sawmill and two pulp and paper mills.

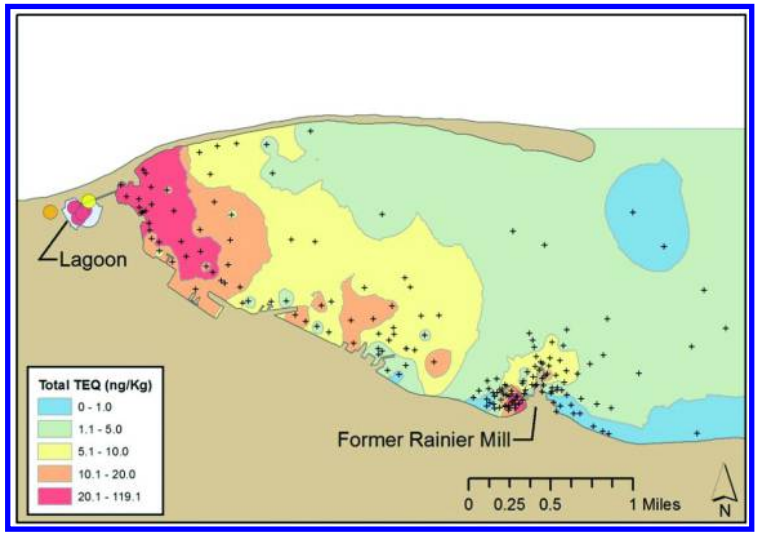

Figure 2. Total Dioxin TEQ for Port Angeles Sediments

Dioxin/furan TEQs in the former Rayonier Mill log pond and dock area are greater than those of the central and outer harbor, suggesting the former mill property as a potential dioxin source (Figure 2). Unlike the well-protected western harbor, the mouth of the harbor where the former Rayonier Mill resides is subjected to waves and currents that have the potential to cause resuspension and dispersion of sediments.

Despite the utility of dioxin TEQ spatial patterns for identifying potential upland source regions, the footprint of individual sources cannot be deciphered without further understanding of unique source profiles. Multivariate analyses provide these means.

Data Screening

Dioxin levels in the collected samples varied considerably; in some samples certain dioxin congeners were not detected (the non-detect level varies with congener). Before any processing was done, data were censored to exclude those

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

samples for which there were too many non-detect congener measurements. Excluding samples with non-detect values should not be done lightly as this can create an upward bias in statistics based on the retained data (*39*). On the other hand, substituting a value for the non-detect value can also lead to biases, and this fabrication of data may lead to misleading conclusions about structure in the data. For this study, a middle ground was sought: exclude samples for which a large number of non-detect values were present (see below) while for samples that were included, substitute the non-detect values with half the detection limit.

Of the 279 sediment samples for which 17 congeners were measured, there were a total of 597 non-detect values, approximately 13%. However, these non-detect values were not uniformly distributed; for example, the two congeners for which non-detect values were the highest included 1,2,3,7,8,9-HxCDF (127 samples; 46%) and 1,2,3,4,7,8,9-HpCDF (72 samples; 26%). Nine congeners exceeded 10% non-detect values. The distribution of non-detect values is shown in Figure 3.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

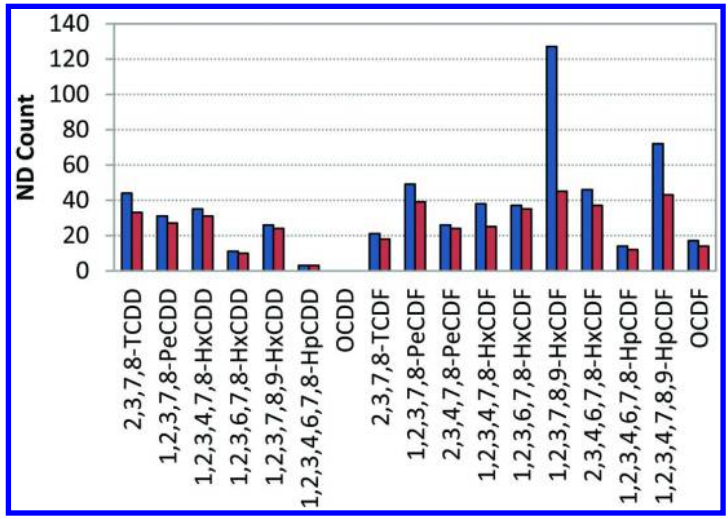

Figure 3. Distribution of non-detect values across congeners before (blue) and after (red) exclusions

The frequency of non-detect values within samples is skewed in an expected way, as shown in Figure 4. The frequency curve flattens after about 4 non-detect values. Using this as a threshold, that is, excluding samples for which more than 4 congeners were non-detect values, 234 samples (84%) were retained. Of the 45 samples excluded in this manner, none had a TEQ over 8 and most were below 2 ppt, compared to the range of values in the data set of 0 to 119 ppt.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

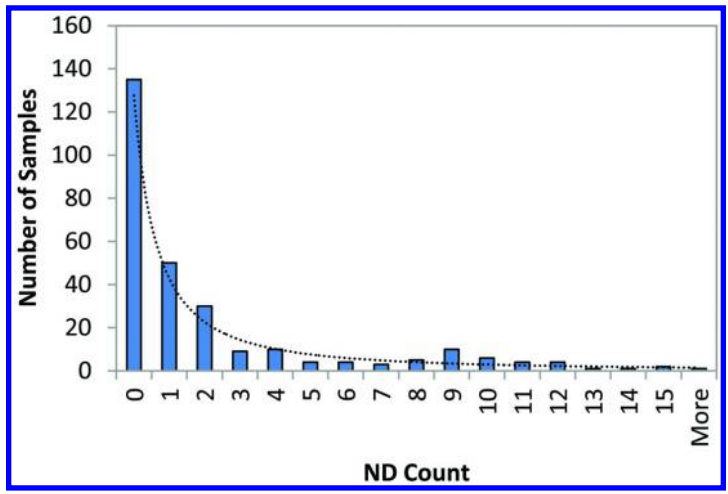

Figure 4. Frequency of non-detect values in study samples

After excluding samples, the distribution of non-detect values in the remaining included samples is more uniform. The highest number of remaining non-detect values was 45 (19%), as shown in Figure 3.

Data Pretreatments

Preliminary evaluation of the appropriateness of the data was conducted by examining line plots and by principal component analysis (PCA). For example, all bulk congener data can be shown overlaid in a line plot (Figure 5), without any scaling of the response values.

Plotted in this way, it is clear that the overwhelming contribution to overall intensity comes from the octa-chloro dioxin congener, while the lesser-chlorinated dioxins and furans contribute relatively little intensity. Thus, before proceeding with any analyses, it is important to scale the different variables such that they are all roughly in the same order of magnitude. There are different approaches to accomplish variable scaling. For this study, two approaches were applied to investigate whether conclusions would differ: scaling by the TEF and scaling by the standard deviation (variance scaling).

One method frequently used in studies of dioxins is to scale by a toxic equivalency factor (*40–42*), based on toxicities relative to 2,3,7,8-TCDD. An advantage of TEF scaling not shared by other scaling methods is that the result is not dependent on the particular data set because scaling is done for each sample independently.

In Figure 6, the data from Figure 5 have been scaled by the TEFs. There remains variation in magnitude for the different congeners but patterns are discernible and the intensities are more directly correlated to risk assessment.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

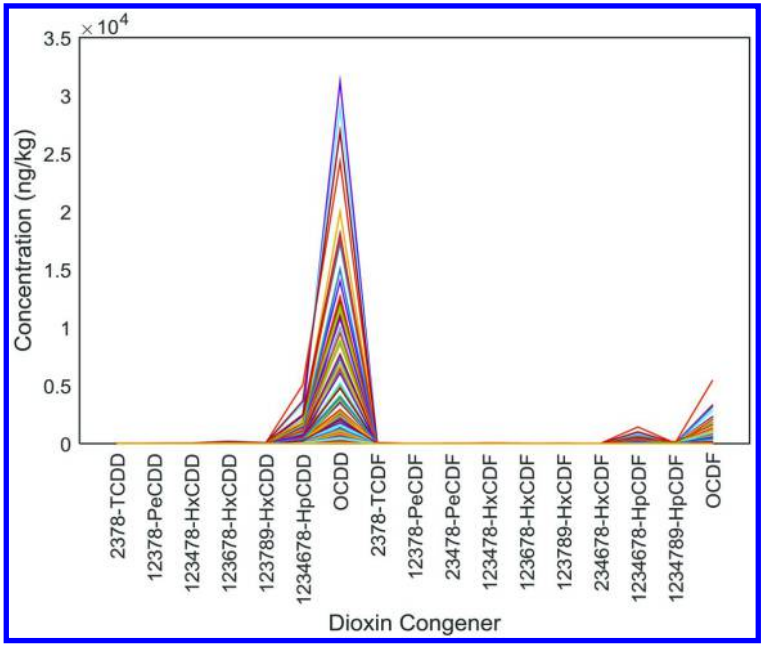

Figure 5. Bulk congener profiles of all study samples, showing chromatographic peak intensities

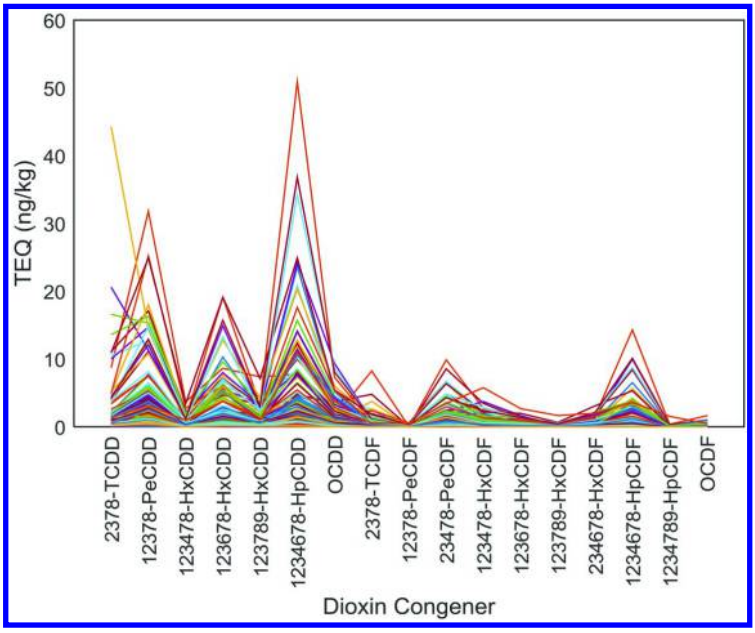

Figure 6. TEF-scaled profiles of all study samples

In Figure 7, each variable was scaled by the standard deviation across the set of samples. This form of scaling results in a variance of 1 for each variable but runs the risk of inflating noisy variables to the same importance of other variables.

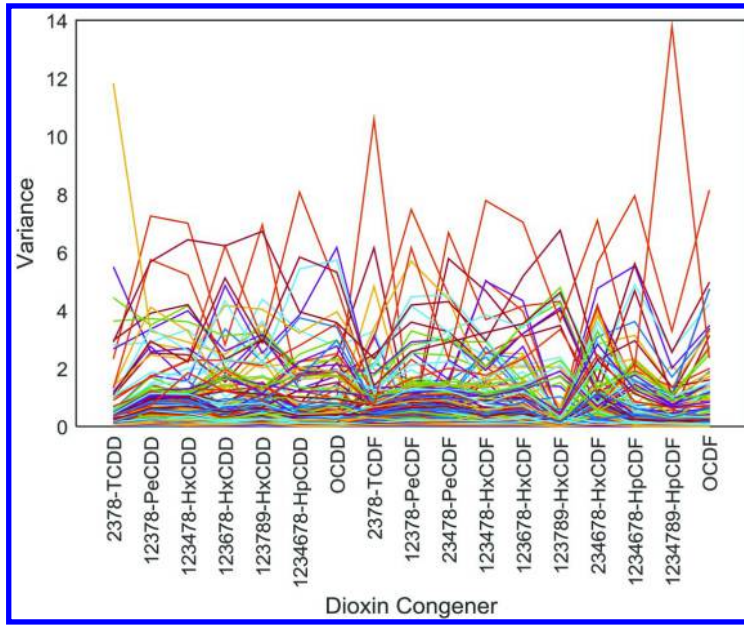

Figure 7. Variance-scaled profiles of study samples

We chose to use TEF-scaling as the method of pretreatment for the following reasons:

- Scaling factors (congener-specific) are independent of which samples are in the data set, whereas variance scaling factors are a function of the specific samples included in the calculation and would therefore change if different samples were treated.
- Because TEF scaling factors can be applied universally to dioxin/furan congener data, analysis results can be compared to profile libraries scaled by the same means. Variance scaling factors are specific to the data set being scaled, thus the resulting congener profiles cannot be directly compared to profiles outside of the data set, such as a comparison library.
- Chemometric analysis of TEF-scaled data identifies dioxin/furan profiles that contribute to a significant portion of sample TEQ. This is useful for decision making, as human health risk, ecological risk, and cleanup criteria are all based on TEQ.

It is customary to further normalize the data to account for different sample sizes thereby minimizing variation in absolute concentration. Although various methods of normalization are used in the multivariate field, area % normalization

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

is typical for chromatography data and was used in this study. Figure 8 shows the area % normalized, TEF-scaled data. All subsequent multivariate analyses were performed on these data.

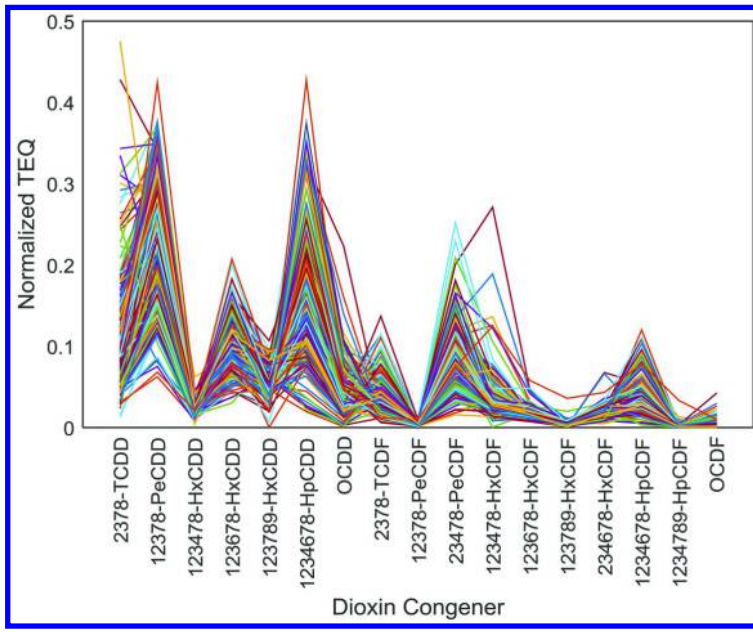

Figure 8. Normalized TEF-scaled profiles

PCA Analysis of Dioxins/Furans

After censoring and pre-treating, data were processed by PCA. Cross-validation (*43*) was used during the processing. The resulting prediction residual error sum of squares as a function of the number of factors (PRESS; see Figure 9) can be used to help understand how many underlying components may be present in the data set and indicated 4-6 factors to be optimal.

The PCA scores of the normalized, TEF-scaled data are shown in Figure 10. In this view of scores in the first 3 factor directions, which represents more than 97% of the information in this data set, samples are distributed in multiple directions. Each direction indicates a likely underlying source material; to get at what these source profiles might be, we turned to the MCR-ALS mixture analysis algorithm.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

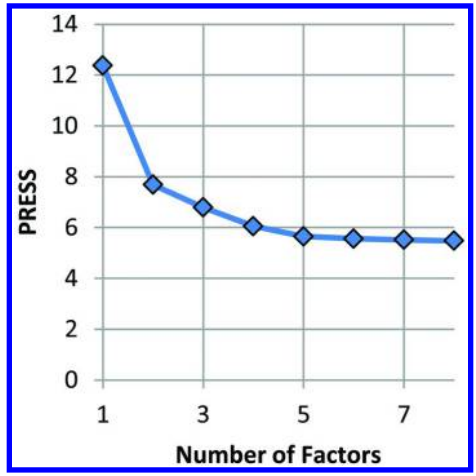

Figure 9. PRESS plot from PCA on normalized, TEF-scaled data

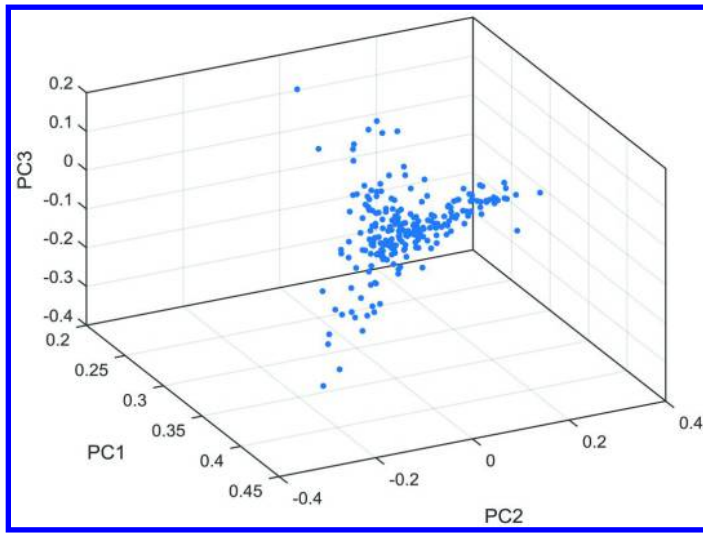

Figure 10. PCA scores of normalized, TEF-scaled data

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

Mixture Analysis of Dioxins/Furans

Mixture analysis was run using the MCR-ALS algorithm, in which up to 8 possible sources were considered. Constraints included non-negativity of both the source contributions and profiles. Based on the non-random structure in the first six PCA factors and on a quality of fit diagnostic for the ALS results, it was decided that 6 sources would be a good estimate. The estimated dioxin profiles for 6 sources are plotted together in Figure 11.

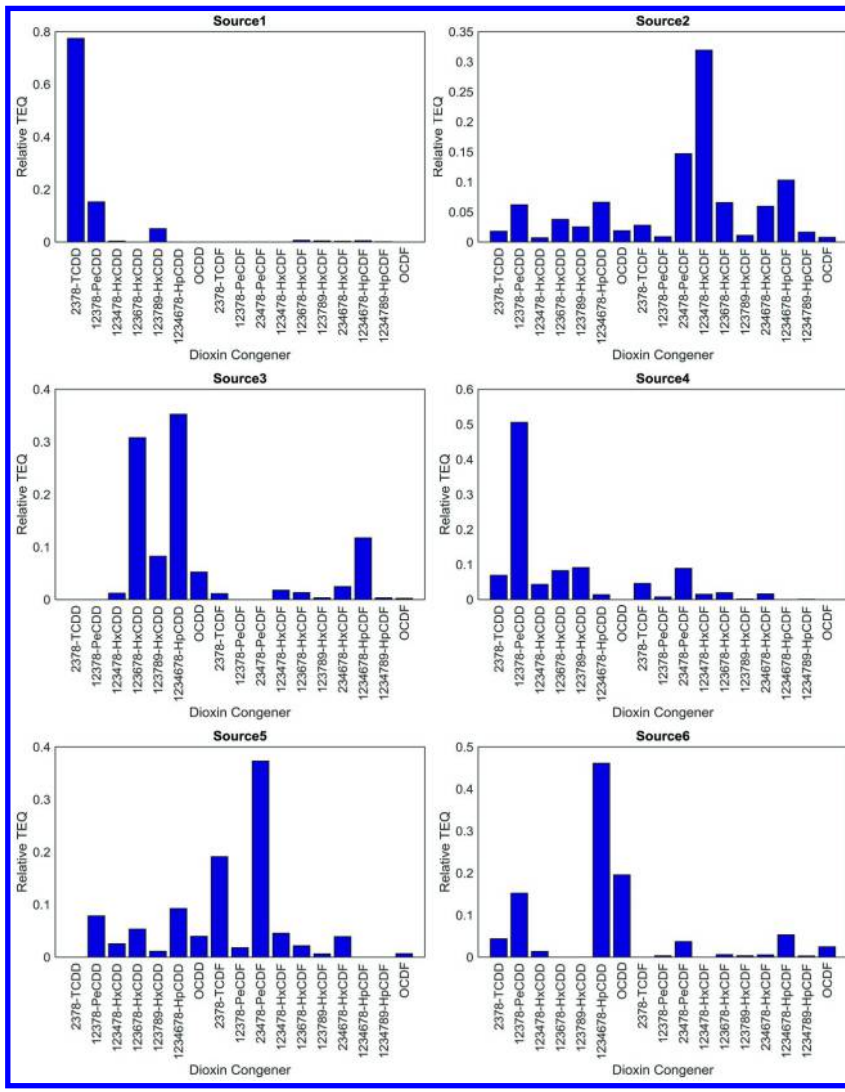

Figure 11. Source profiles following mixture analysis; 6 source solution

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

The patterns of dioxins in these source profiles are estimates of the patterns of materials that were deposited in the harbor sediments. However, in such a mixed environment, it is likely that every sample location has some contribution from most, if not all, of these underlying source materials. In Figure 12, the relative contributions from each source are displayed graphically.

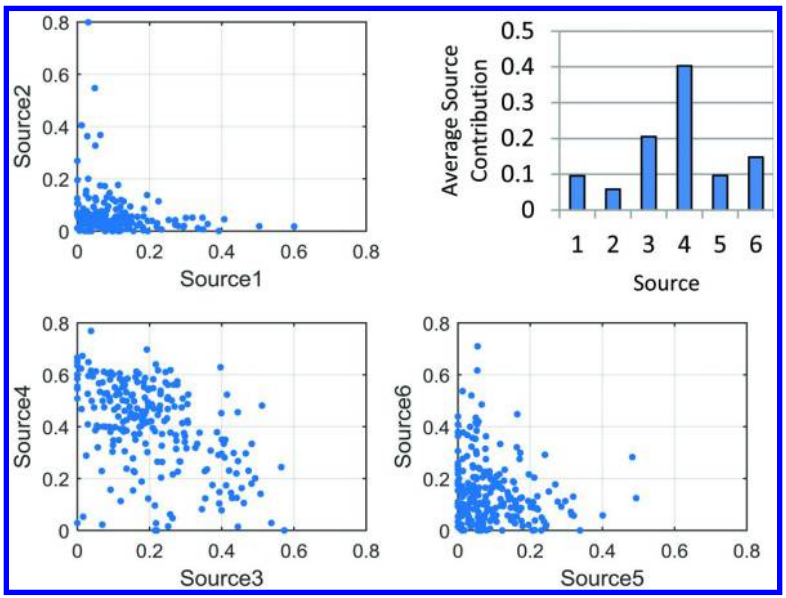

Figure 12. Source contributions to sediment samples; 6-source model

From these plots, it appears that major contributions to most of the sample locations come from sources 3 and 4, with somewhat lesser contributions from sources 1 and 6. Sources 2 and 5 seem to contribute to the fewest samples.

Two of the six sources appear to be related to dioxins that originate from pentachlorophenol and two others appear related to patterns that are similar to those from degradation of PCBs (see below). Thus, it was considered interesting to restrict the number of sources in the ALS model to only 4. When this is done, the residual profiles, which show features not contained in the ALS model, indicate that only a few samples are not well modeled. In particular, in Figure 13, three samples show more deviation than other samples and are highlighted in the plot with thicker lines.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

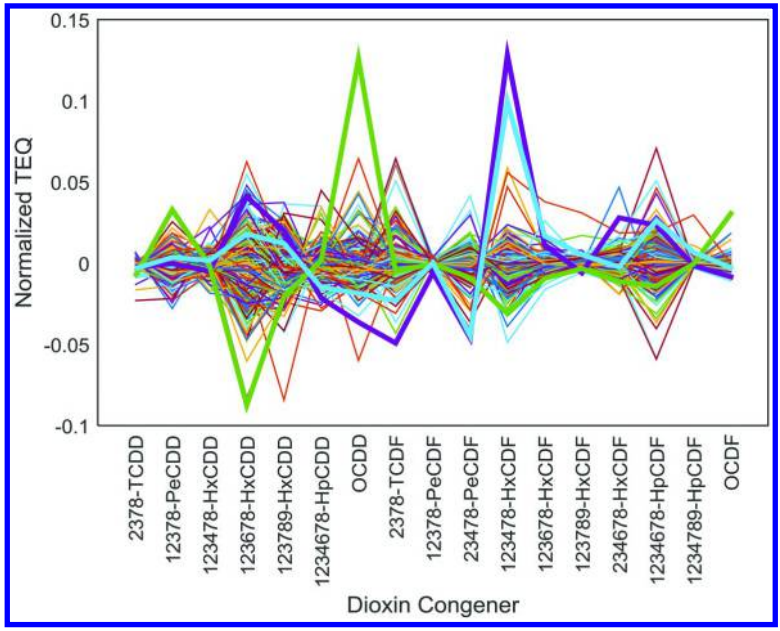

Figure 13. X-residuals for 4-source mixture analysis model; thick traces denote aberrant samples

By making a model with more sources, these residuals would diminish. On the other hand, it is these three samples that largely drive the 6-source ALS model. In fact, observing the normalized profiles of Figure 8 again, where the traces for the aberrant samples are highlighted (Figure 14), we can see that:

- the extra intensity of the first highlighted trace of Figure 14 occurs in the OCDD congener and is represented by Source 6 of Figure 11
- the extra intensity in the second and third highlighted traces occurs in the 1,2,3,4,7,8-HxCDF congener and is represented by Source 2 of Figure 11

With only three outlier samples, it is not clear if they represent true sources that need to be incorporated into the evaluation or if they represent inconsistencies in either the sampling or in the instrumental analysis. The fact that two of the samples show much the same pattern (in particular the relatively high 1,2,3,4,7,8 HXCDF content) implies that these samples cannot be completely dismissed from consideration. It is instructive, however, to look at a 4-source model and compare results.

When a 4-source mixture analysis is computed (see Figure 15), sources 1 and 4 change only a little from their shapes in the 6-source model (recall Figure 11).

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

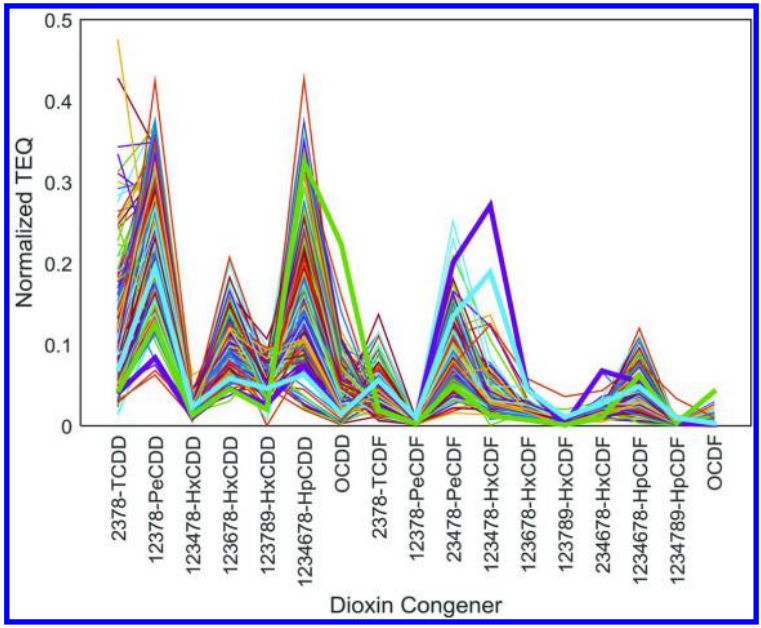

Figure 14. Normalized TEF-scaled profiles; thick traces denote aberrant samples

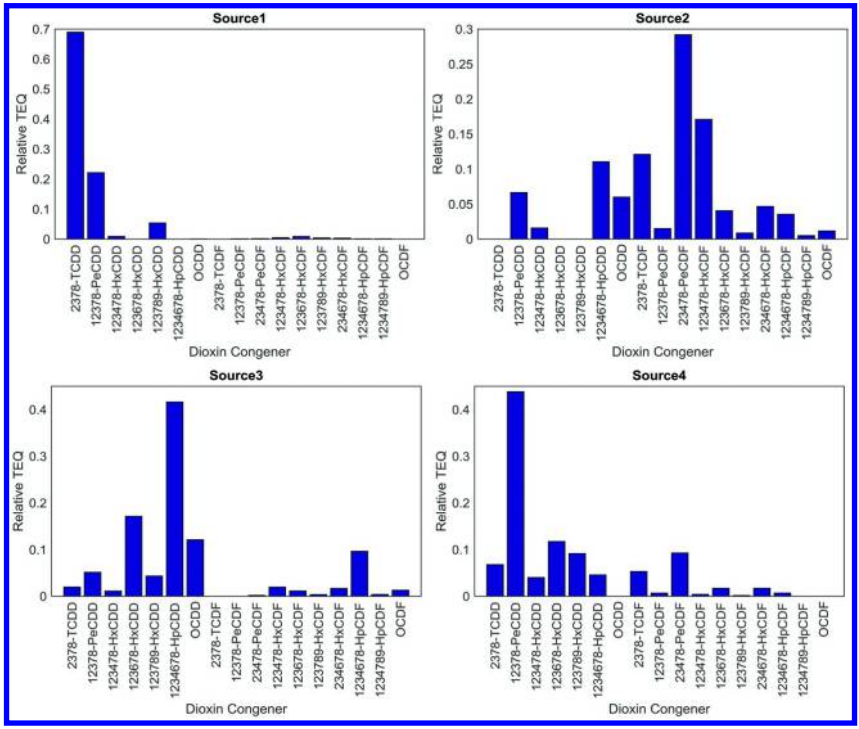

Figure 15. Source profiles following mixture analysis; 4 source solution

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

In addition, Source 2 in the 4-source model appears to be a composite of sources 2 and 5 in the 6-source model, and Source 3 in the 4-source model appears to be a composite of sources 3 and 6 in the 6-source model.

In the 4-source model, the contributions from each source to the samples are shown in Figure 16. The major contributions derive from sources 3 and 4, as shown by the greater cluster of points in the upper end of their axes.

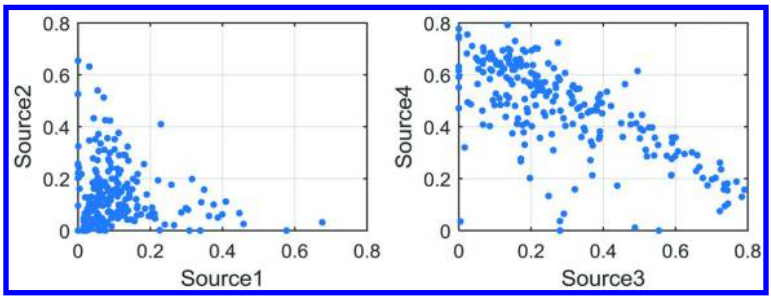

Figure 16. Source contributions to sediment samples; 4-source model

A comparison of the source profiles in the 6- and 4-source unmixing models shows a strong degree of correspondence between appropriately matched profiles. While some differences in source profiles and source amounts can be identified between the two unmixing models, a comparative evaluation indicates that these differences are relatively small. The two unmixing models lead to results that are not markedly different with respect to source profiles and spatial patterns of source contributions. Thus, the 4-source model was used as the basis for all further evaluations.

Source Interpretation

To understand the nature of the source patterns determined via mixture analysis, a database of comparison patterns was constructed. The patterns for comparison were drawn from multiple sources, listed in the Methods section.

After removing duplicate patterns, the comparison set comprised 154 patterns of the 17 dioxin/furan congeners. These patterns were compiled into a spreadsheet, TEF-scaled, then merged with the 4 source profiles from mixture analysis on the TEF-scaled sample set. The combined data were analyzed in two complementary manners: first by hierarchical cluster analysis (HCA), then by a tabulation of correlation coefficients. These approaches are discussed below.

The results from the cluster analysis are best viewed in the form of a dendrogram (Figure 17), which shows samples in clusters according to their relative similarity. In the figure, the leaf nodes for the 4 sources are shown as points with corresponding labels. Thus, we can see in the dendrogram that the source patterns are distinguishable from each other and that there are groups of comparison patterns similar to each source.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

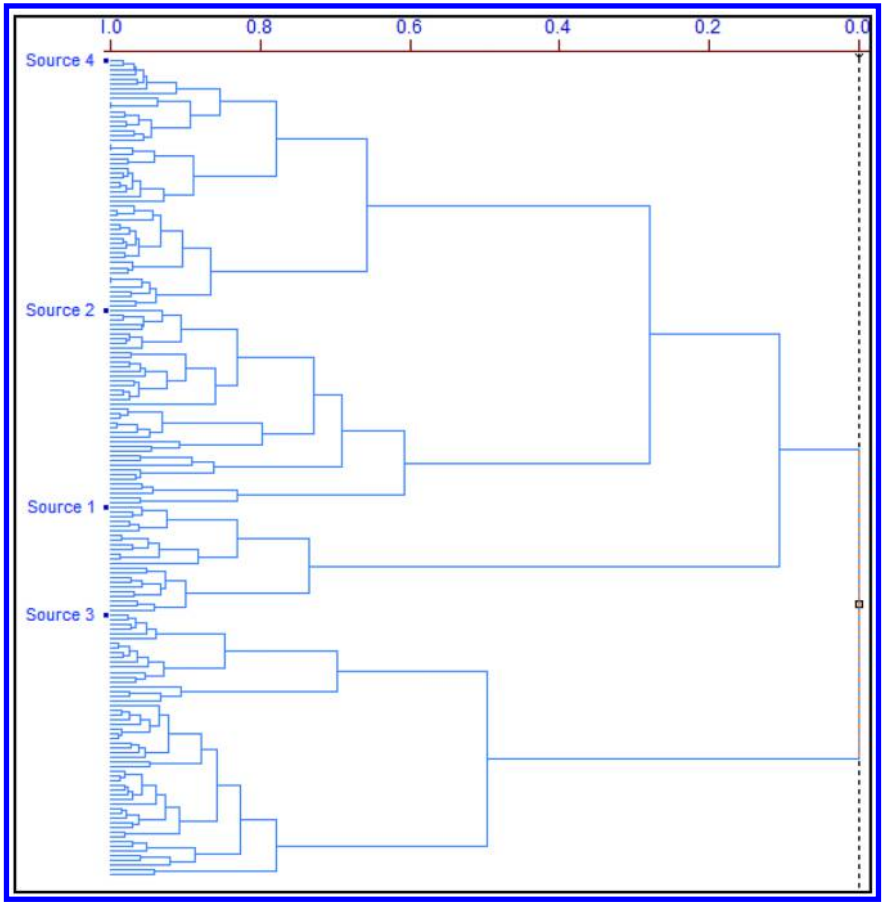

Figure 17. HCA dendrogram comparing source to comparison patterns

Based on the nearest neighbors in the HCA dendrogram and on the comparison patterns with highest correlations to the source patterns, conclusions can be drawn about the nature of possible forms of dioxin contamination. It appears that four types of input are present.

The comparison pattern that most resembles that of Source 1 was from a New Zealand study of Silvex-contaminated sediments (see Figure 18). Other comparison patterns that demonstrated an acceptable match came from samples of wood and fly ash.

Source 2 did not have a strong match from any one comparison pattern, although the major furan peak distributions were similar to the patterns seen among PCB sources (for example, Figure 19). Comparisons from industrial and residential soot also exhibited patterns that were similar.

A very strong similarity to pentachlorophenol was demonstrated for the Source 3 pattern (Figure 20). PCP-treated utility poles and wood forms show the characteristic intense hepta-chloro dioxin congener.

The fourth source profile matched well the source patterns from several other hog-fuel boiler effluent examples from Canadian sources (see Figure 21).

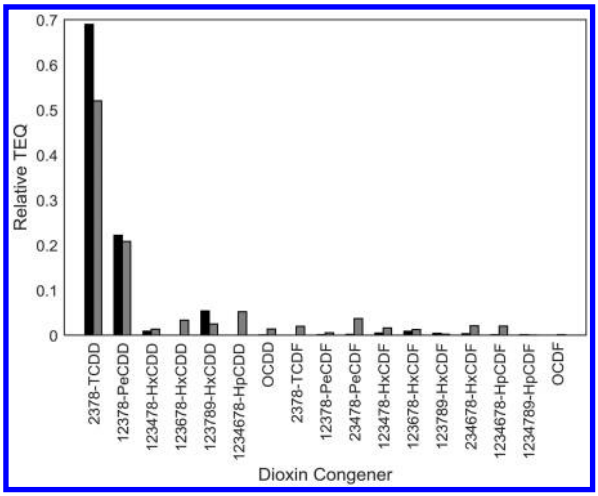

Figure 18. Best match to Source 1 (black): New Zealand sediment (gray)

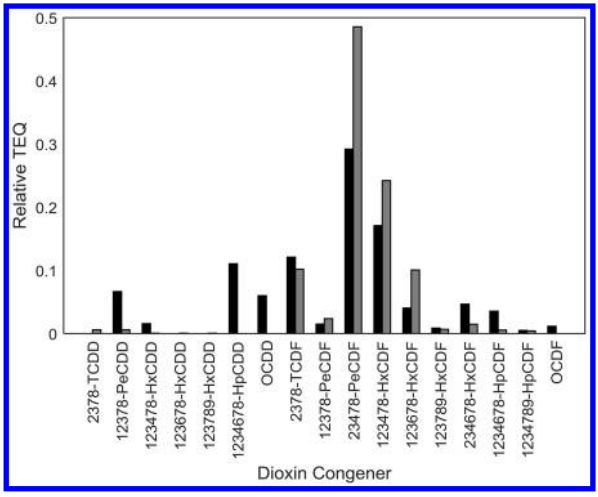

Figure 19. Best match to Source 2 (black): Aroclor 1254 (gray)

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

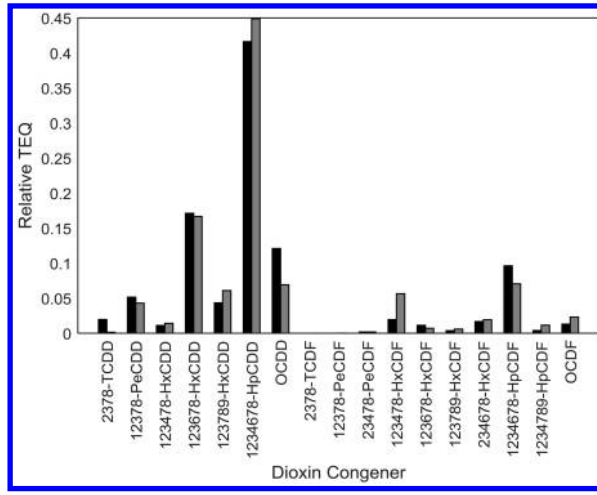

Figure 20. Best match to Source 3 (black): PCP-treated utility poles (gray)

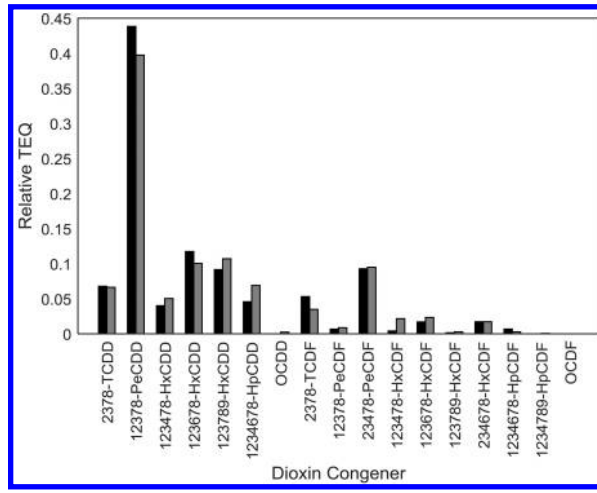

Figure 21. Best match to Source 4 (black): Canadian hog fuel boiler (gray)

Spatial Interpretation

When source amounts are re-scaled to their respective TEQ values, the result is called a source increment. Spatial interpolation of the source increments can show the relative importance of a source as a contribution to each sample as well as the relative importance of a source to harbor-wide sediment contamination. TEQ increments are shown below using the same scale among the four sources such that the relative magnitude of sources can be visually compared.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

The majority of sample locations for which Source 1 is the dominant contribution occur in the lagoon (see Figure 22). This source does not have a significant contribution to the remainder of the harbor.

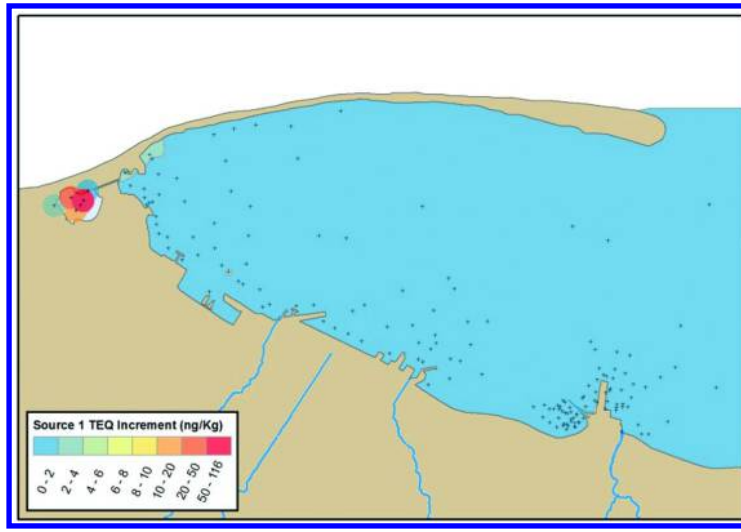

Figure 22. Source increment map--Source 1

Source 2 is a low contributor to locations in the lagoon, in the western harbor, and adjacent the Rayonier facility (Figure 23). It is not important elsewhere in the harbor.

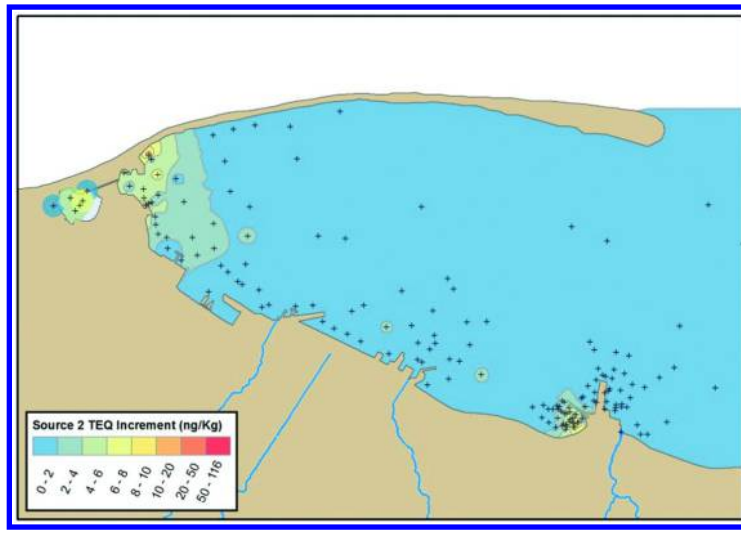

Figure 23. Source increment map--Source 2

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

Contributions from Source 3 are high in the western harbor, the lagoon and adjacent the Rayonier site (Figure 24); many near-shore locations in the western and middle harbor show moderate contributions as well.

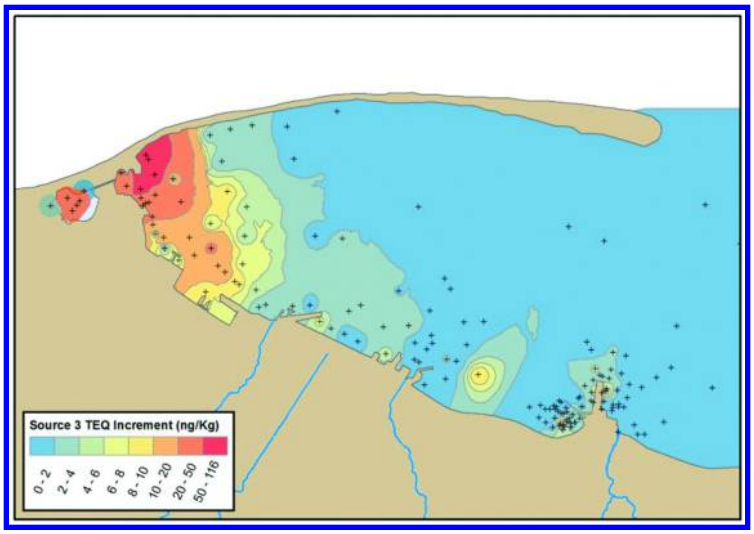

Figure 24. Source increment map--Source 3

Source 4 contributions are highest in the lagoon and at the Rayonier site. Moderate contributions can be seen from this source in the near and off shore regions of the western and middle harbor (see Figure 25).

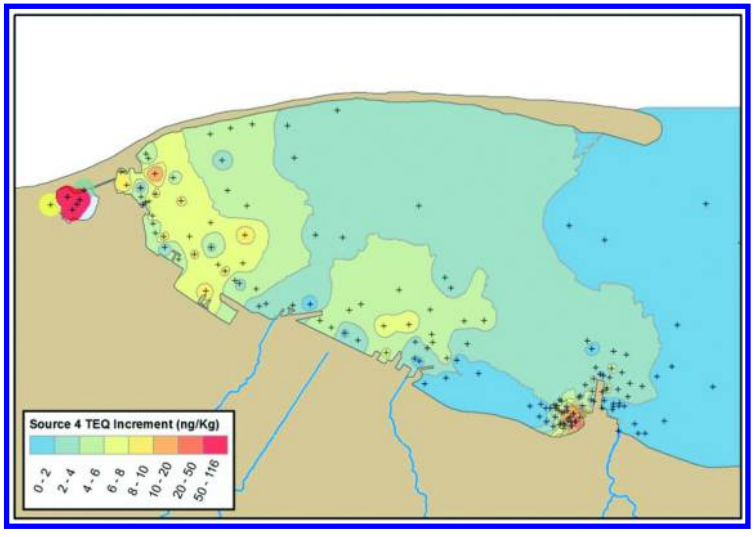

Figure 25. Source increment map--Source 4

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

Discussion

In this section each of the four dioxin sources to Port Angeles Harbor sediments identified through mixture analysis are discussed separately, including an evaluation of their contribution to harbor-wide sediment contamination, spatial patterns and potential sources.

Source 1

The Source 1 dioxin profile is most similar to that of Silvex and wood ash. It is the most minimal contributor to harbor-wide dioxin contamination of the considered sources, comprising approximately 6 percent of surface sediment TEQ. The Source 1 spatial pattern is unlike the other sources in that it is almost entirely restricted to the western harbor lagoon (Figure 22). This pattern suggests direct discharge of Source 1 to the lagoon from the nearby upland and relatively little input of Source 1 to the remaining harbor from other upland locations.

Possible mechanisms by which Source 1 dioxin became deposited in the lagoon are apparent based on its industrial history. The lagoon is a natural feature formed in accordance with Ediz Hook, enclosing the northern harbor. Since the early 1900s the lagoon has served as a log storage area for the adjacent paper mill prior to the pulping process. The herbicide 2,4,5-T may have been applied to the surrounding upland for weed control or directly to the lagoon as an algaecide. Algae control in the lagoon would have been particularly important, as the shallow, stagnant conditions of the lagoon promote biological fouling of logs prior to processing.

The lagoon also served as a disposal site for the adjacent mill. Before the practice of log storage in the lagoon was abandoned in the mid-1970s, approximately 12 acres of the lagoon were filled with ash from the mill's wood-fired boiler. Regardless of what the exact source material is for Source 1 dioxin (2,4,5-T and/or wood ash), historic industrial uses of the lagoon suggest input of Source 1 dioxin from adjacent upland activities.

Source 2

The Source 2 dioxin profile is most similar to that of a number of manufactured PCBs. In contrast to Source 1, the spatial pattern of Source 2 dioxin suggests multiple physically separated point source locations. Interpolation of Source 2 increments (Figure 23) shows a distinct pattern, with clusters of high-TEQ within the lagoon, close to the western harbor shoreline, and within the former Rayonier log pond. Although present in multiple areas of the harbor, Source 2 dioxin only contributes approximately 10 percent to total TEQ harbor-wide.

Ancillary sediment data support the theory that Source 2 dioxin is at least partially derived from PCBs. PCB Aroclors were generally not detected in sediments of the central and outer harbor where Source 2 increments are the lowest. When detected, the greatest total PCB Aroclor concentrations were confined to locations of highest Source 2 increments. The co-occurrence of PCBs and Source 2 in Port Angeles Harbor sediments suggests that Source 2 dioxin is a

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

chemical component of PCBs and they are transported and deposited in a similar manner.

The spatial distribution of Source 2, as well as PCBs, suggest the former Rayonier Mill property and western harbor industries are the primary sources of this form of dioxin to harbor sediment. There is a high likelihood that PCBs were extensively used at these facilities. PCBs have historically been used as coolants and lubricants in electrical equipment such as transformers and capacitors, and they are found in older fluorescent lighting fixtures and electrical appliances, paints, pesticide additives, sealants, building materials, and hydraulic oils (*44*). PCBs were identified as contaminants of concern for the marine environment near the former Rayonier Mill property because of their possible presence in process wastewater effluent and from incidents of leaking transformers (*45*). Similar concerns regarding PCB discharge also exist for other pulp and paper mills and wood treatment facilities of the western harbor.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

Source 3

The Source 3 dioxin profile is most similar to that of manufactured PCP (both oil and water soluble forms) and wood treated with PCP as a preservative. The spatial pattern of Source 3 in harbor sediments is dominated by a likely point source in the western harbor, in the vicinity of a paper mill that has been in operation since the early 1900s (Figure 24). While Source 3 was found to contribute 40 percent to total TEQ harbor-wide, virtually all Source 3 dioxin is found in the western harbor and lagoon. Lower levels of Source 3 dioxin also exist in small, isolated pockets in close proximity to the former Rayonier Mill property.

Almost all PCP production in the United States has been used for commercial wood treatment and slime control in pulp and paper production (*46*). It is expected that sediments of Port Angeles Harbor may have a dioxin component derived from PCP because of the long-term existence of both lumber and pulp and paper mills along the harbor waterfront. Despite PCP being a likely source of dioxin in harbor sediments, PCP itself has not recently been detected in any surface sediments of the harbor (*21*). This absence is likely due to the rapid degradation rate of PCP (*47*) compared to its associated dioxin. Additionally, the lack of PCP in harbor sediments may indicate that Source 3 dioxin is relatively old and not associated with modern upland activities.

The co-association of Source 3 dioxin and mercury in Port Angeles Harbor surface sediments (Figure 26) implies that these chemicals may be derived from a common upland source in the western harbor (Figure 2). Because of their synergistic value, the combined use of PCP and mercury played an important role in slime control for the pulp and paper industry between 1940 and 1970 (*48*). These slimicides prevent the uncontrolled growth of microorganisms that can result in slime deposits. When unchecked, slime can clog filters, screens, and pipelines, and result in spots and breaks in the paper sheet. The chemicals used as slimicides have varied over the past century and have often been implicated as highly toxic components of mill effluent and major sources of aquatic pollution (*49*).

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

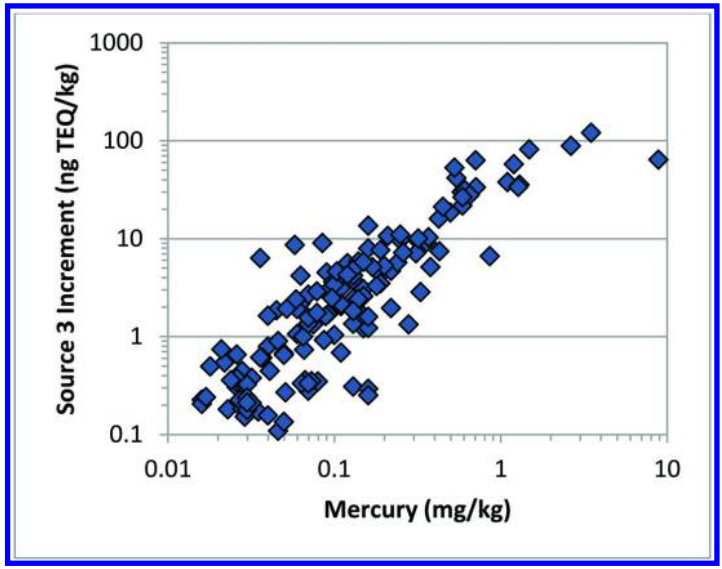

Figure 26. Correlation between Source 3 increments and mercury concentrations

Source 4

The Source 4 dioxin profile is most similar to that of air emissions, ash, and effluent from facilities with hog fuel boilers burning salt-laden wood. The spatial pattern of Source 4 is much more dispersed than the other dioxin sources previously discussed (Figure 25). Source 4 contributes 44 percent to harbor-wide TEQ, making it the most abundant source of dioxin to sediments of the harbor. Additionally, Source 4 is the dominant contributor to total TEQ of the southern and central harbor, regions where the other sources generally have much lower contributions.

Due to the waterfront location of facilities on Port Angeles Harbor and the abundance of wood as a source of fuel for onsite burners, burning wood chips and wood wastes coming from logs floated in the harbor was a common practice. This salt-laden wood was utilized as boiler fuel at four industrial properties along the harbor waterfront, including at the former Rayonier Mill and adjacent to the western harbor lagoon. Burning salt-laden wood in an industrial boiler can result in significantly higher emissions of dioxins/furans than can burning salt-free wood (*28*, *50*–*53*).

Source 4 dioxin in sediment may reflect contributions from more than one hog fuel boiler source and also mixtures of boiler ash and stack emissions. Additionally, Source 4 may be introduced to the harbor through a variety of transport pathways that are physically disconnected from the boilers themselves. Prior to deposition in the harbor, transport of Source 4 dioxin may involve:

- Aerial deposition of boiler emissions onto the harbor surface;
- Aerial deposition of boiler emissions in the uplands and subsequent delivery to the harbor in stormwater runoff and municipal effluent;

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

- Erosion/runoff of boiler ash from industrial properties and disposal sites;
- Incorporation of boiler ash into industrial process water and effluent; and
- Direct disposal of boiler ash into the harbor.

The complexity of possible transport mechanisms prior to deposition in sediments make the partitioning of Source 4 between the different industrial locations challenging. Based on their proximity to the highest Source 4 increments, both the former Rayonier Mill and western harbor properties are likely sources of Source 4 dioxin to harbor sediments. The dominant role of Source 4 in the southern and central harbor, as well as overall dominance of Source 4 in surface sediments harbor-wide, suggests that delivery from stormwater runoff is a more important transport mechanism for Source 4 than other dioxin sources.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

Conclusions

The chemometric evaluation of Port Angeles Harbor PCDD/F congeners identified four source patterns that provide a good model for measured TEQ values. Each of the four proposed source patterns has an analog in known dioxin-producing materials:

- Source 1 – 2,4,5-T or wood ash
- Source 2 – PCBs
- Source 3 – Pentachlorophenol
- Source 4 – Emissions and effluent related to burning of salt-laden wood in hog-fuel boilers

Spatial interpolation of dioxin source TEQ increments allowed for the determination of the relative contribution of each source to harbor-wide dioxin contamination.

Each of the dioxin sources has a unique spatial pattern in harbor sediments, which, along with supplemental data, can be used to identify potential upland source locations and to understand mechanisms by which the dioxin became deposited.

Acknowledgments

LSR would like to acknowledge Prof. Bruce R. Kowalski for revealing the field of curve resolution and mixture analysis and suggesting its use in environmental studies. Bruce's insight and perspicacity, combined with a well-tuned sense of humor, provided a perfect environment for scientific development.

References

1. U.S. DHHS. *Chlorinated Dibenzo-p-dioxins (CDDs)*; Agency for Toxic Substances and Disease Registry, Division of Toxicology and Environmental

Medicine ToxFAQs, U.S. Department of Health and Human Services: Atlanta, GA, 1999; pp 1−2.
2. U.S. EPA. *Method 1613: Tetra- through Octa-Chlorinated Dioxins and Furans by Isotope Dilution HRGC/HRMS*; U.S. Environmental Protection Agency, Office of Water, Engineering and Analysis Division: Washington, DC, 1994; pp 1−89.
3. Van den Berg, M.; Birnbaum, L. S.; Denison, M.; De Vito, M.; Farland, W.; Feeley, M.; Fiedler, H.; Hakansson, H.; Hanberg, A.; Haws, L.; Rose, M.; Safe, S.; Schrenk, D.; Tohyama, C.; Tritscher, A.; Tuomisto, J.; Tysklind, M.; Walker, N.; Peterson, R. E. *Toxicol. Sci.* **2006**, *93*, 223–241.
4. *Newfields, Port Angeles Harbor Supplemental Data Evaluation to the Sediment Investigation Report*; Washington State Department of Ecology Toxics Cleanup Program: Lacey, WA, 2012.
5. Glass, G. L. *Ecology and Environment; Rayonier Mill Off-Property Soil Dioxin Study; Port Angeles, WA, Public Review Draft*; Washington State Department of Ecology Toxics Cleanup Program: Lacey, WA, 2011.
6. Massart, D. L.; Vandeginste, B. G. M.; Deming, S. N.; Michotte, Y.; Kaufman, L. *Chemometrics: a textbook*; Elsevier: New York, NY, 1988; Vol. 2.
7. Beebe, K. R.; Pell, R. J.; Seasholtz, M. B. *Chemometrics; a practical guide*; John Wiley & Sons, Inc.: New York, 1998.
8. Johnson, G. W.; Ehrlich, R.; Full, W. E.; Ramos, L. S. Principal Components Analysis and Receptor Models in Environmental Forensics. In *Introduction to Environmental Forensics*, 2nd ed.; Murphy, B. L., Morrison, R. D., Eds.; Elsevier Academic Press: Amsterdam, 2007; pp 207−272.
9. Cheng, P.-S.; Hsu, M.-S.; Ma, E.; Chou, U.; Ling, Y.-C. *Chemosphere* **2003**, *52*, 1389–1396.
10. Lee, W.-S.; Chang-Chien, G.-P.; Wang, L.-C.; Lee, W.-J.; Tsai, P.-J.; Wu, K.-Y.; Lin, C. *Environ. Sci. Technol.* **2004**, *38*, 4937–4944.
11. Jiménez, B.; Eljarrat, E.; Hernández, L. M.; Rivera, J.; González, M. J. *Chemosphere* **1996**, *32*, 1327–1348.
12. Macdonald, R. W.; Ikonomou, M. G.; Paton, D. W. *Environ. Sci. Technol.* **1998**, *32*, 331–337.
13. Park, S.; Kim, S.-J.; Kim, K. S.; Lee, D. S.; Kim, J. G. *Environ. Sci. Technol.* **2004**, *38*, 3820–3826.
14. Oh, J.-E.; Choi, S.-D.; Lee, S.-J.; Chang, Y.-S. *Chemosphere* **2006**, *64*, 579–587.
15. Sundqvist, K. L.; Tysklind, M.; Geladi, P.; Cato, I.; Wiberg, K. *Chemosphere* **2009**, *77*, 612–20.
16. Götz, R.; Lauer, R. *Environ. Sci. Technol.* **2003**, *37*, 5559–5565.
17. Yunker, M. B.; Cretney, W. J.; Ikonomou, M. G. *Environ. Sci. Technol.* **2002**, *36*, 1869–1878.
18. Grundy, S. L.; Bright, D. A.; Dushenko, W. T.; Dodd, M.; Englander, S.; Johnston, K.; Pier, D.; Reimer, K. J. *Chemosphere* **1997**, *34*, 1203–1219.
19. Huntley, S. L.; Carlson-Lynch, H.; Johnson, G. W.; Paustenbach, D. J.; Finley, B. L. *Chemosphere* **1998**, *36*, 1167–1185.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

20. Uchimiya, M.; Arai, M.; Masunaga, S. *Environ. Sci. Technol.* **2007**, *41*, 3864–3870.
21. *Ecology and Environment, Port Angeles Harbor Sediment Characterization Study, Port Angeles, WA: Public Review Draft*; Washington State Department of Ecology Toxics Cleanup Program: Lacey, WA, 2012.
22. National Park Service. *Sampling for the Nippon Paper Industries Outfall 002 Replacement*; 2010.
23. Exponent. *Environmental Baseline Investigation DNR Lease 22-077766*; Prepared for Nippon Paper Industries by Exponent: 2008.
24. Malcolm Pirnie. *Remedial Investigation for the Marine Environment Near the Former Rayonier Mill Site, Proposed Public Review Draft*; Prepared for Rayonier: Seattle, WA, 2007.
25. Malcolm Pirnie. *Phase 2 Addendum Remedial Investigation for the Marine Environment Near the Former Rayonier Mill Site, Proposed Public Review Draft*; Prepared for Rayonier: Seattle, WA 2007.
26. Foster Wheeler Environmental Corporation. *Summary of the Log Pond Survey Scoping Effort for the Remedial Investigation*; Prepared for Rayonier: Port Angeles, WA, 2001.
27. Environmental Protection Agency. *An Inventory of Sources and Environmental Releases of Dioxin-like Compounds in the United States for the Years 1987, 1995, and 2000*; National Center for Environmental Assessment, Office of Research and Development: Washington, DC, 2006.
28. Duo, W.; Leclerc, D. *Organohalogen Compd.* **2004**, *66*, 992–1000.
29. Uloth, V.; Duo, W.; Leclerc, D.; Karidio, I.; Kish, J.; Singbeil, D. Investigations into the variability in and control of dioxins formation and emissions from coastal power boilers. *Proceedings of the 2005 Engineering, Pulping, and Environmental Conference*; Pulp and Paper Research Institute of Canada (PAPRICAN): 2005.
30. Foster Wheeler Environmental Corporation. *Current Situation/Site Conceptual Model Report for Rayonier, Port Angeles Mill Site, Mt Pleasant Road Landfill and 13th and M Street Landfill*; Prepared for Rayonier: Port Angeles, WA, 1997.
31. Whittemore, R. USEPA/paper industry cooperative dioxin study: the 104 mill study. *NCASI technical bulletin (USA)*; National Council of the Paper Industry for Air and Stream Improvement, Inc.: Washington, DC, 1990.
32. Pattle Delamore Partners. *Dioxin Concentrations in Residential Soil, Paritutu, New Plymouth*; The Ministry for the Environment: Wellington, New Zealand, 2002.
33. Rocky Mountain Center for Occupational and Environmental Health. *A Comparison of Dioxin levels found in Residential Soils of Davis County Utah with those found in Residential Soils in the Denver Front Range*; Prepared for the Utah Division of Air Quality by Rocky Mountain Center for Occupational and Environmental Health, Department of Family and Preventative Medicine: Salt Lake City, UT, 2003.
34. Thoma, H. *Chemosphere* **1988**, *17*, 1369–1379.
35. Johnson, G. W.; Hansen, L. G.; Hamilton, M. C.; Fowler, B.; Hermanson, M. H. *Environ. Toxicol. Pharmacol.* **2008**, *25*, 156–63.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

36. Lorber, M. N.; Barton, R. G.; Winters, D. L.; Bauer, K. M.; Davis, M.; Palausky, J. *Sci. Total Environ.* **2002**, *290*, 15–39.
37. Troyanskaya, A. F.; Moseeva, D. P.; Rubtsova, N. A. *Chem. Sustainable Dev.* **2004**, *12*, 225–231.
38. Tauler, R.; Kowalski, B.; Fleming, S. *Anal. Chem.* **1993**, *65*, 2040–2047.
39. Helsel, D. R. *Environ. Sci. Technol.* **1990**, *24*, 1766–1774.
40. Lohmann, R.; Jones, K. C. *Sci. Total Environ.* **1998**, *219*, 53–81.
41. Alcock, R. E.; Sweetman, A. J.; Anderson, D. R.; Fisher, R.; Jennings, R. A.; Jones, K. C. *Chemosphere* **2002**, *46*, 383–391.
42. Hilscherova, K.; Kannan, K.; Nakata, H.; Hanari, N.; Yamashita, N.; Bradley, P. W.; McCabe, J. M.; Taylor, A. B.; Giesy, J. P. *Environ. Sci. Technol.* **2003**, *37*, 468–474.
43. Eastment, H. T.; Krzanowski, W. J. *Technometrics* **1982**, *24*, 73–77.
44. Agency for Toxic Substances and Disease Registry. *Toxicological Profile for Polychlorinated Biphenyls (PCBs)*; U.S. Department of Health and Human Services, Public Health Service: Atlanta, GA, 2000.
45. *Ecology and Environment, Port Angeles Harbor Final Summary of Existing Information and Identification of Data Gaps Report, Port Angeles, WA*; Washington State Department of Ecology Toxics Cleanup Program: Lacey, WA, 2008.
46. Borysiewicz, M. *Pentachlorophenol, Dossier prepared in support of a proposal of pentachlorophenol to be considered as a candidate for inclusion in the Annex I to the Protocol to the 1979 Convention on Long-Range Transboundary Air Pollution on Persistent Organic Pollutants (LRTAP Protocol on POPs)*. Institute of Environmental Protection: Warszaw, 2008.
47. Kao, C.; Chai, C.; Liu, J.; Yeh, T.; Chen, K.; Chen, S. *Water Res.* **2004**, *38*, 663–672.
48. Sherbin, I. *Mercury in the canadian environment. Economic and technical review report*; Environmental Protection Service, Environment Canada: Ottawa, 1979.
49. Ali, M.; Sreekrishnan, T. *Adv. Environ. Res.* **2001**, *5*, 175–196.
50. Lavric, E. D.; Konnov, A. A.; De Ruyck, J. *Biomass Bioenergy* **2004**, *26*, 115–145.
51. Luthe, C.; Karidio, I.; Uloth, V. *Chemosphere* **1998**, *36*, 231–249.
52. Pandompatam, B.; Kumar, Y.; Guo, I.; Liem, A. J. *Chemosphere* **1997**, *34*, 1065–1073.
53. Preto, F.; McCleave, R.; McLaughlin, D.; Wang, J. *Chemosphere* **2005**, *58*, 935–941.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch004

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

Chapter 5

Multivariate Curve Resolution: A Different Way To Examine Chemical Data

Amrita Malik,[1] Anna de Juan,[2] and Roma Tauler[1,*]

[1]IDAEA-CSIC, Jordi Girona 18-26, Barcelona 08034, Spain
[2]Universitat de Barcelona, Diagonal 647, Barcelona 08028, Spain
[*]E-mail: Roma.Tauler@idaea.csic.es

During the last 40 years, Multivariate Curve Resolution (MCR) has emerged as a powerful tool to investigate (bio) chemical data sets. MCR has evolved from the analysis of a single data set to multiset and multiway data analysis. MCR has been extended to the investigation of new application domains and of more challenging problems. MCR is achieving a mature state and it includes different ways to ascertain and validate the reliability of its solutions

1. History, Concept and Problem To Solve

Multivariate curve resolution (MCR) is already more than 40 years old. Nowadays, Multivariate Curve resolution is the generic denomination for a family of methods used to solve the ubiquitous problem of mixture analysis. This is named differently in other scientific fields, like blind source separation in telecommunications, source apportionment in atmospheric studies, factor analysis in psychometrics, or end member mixing analysis in geosciences, or endmember signatures extraction in remote sensing hyperspectral imaging tele detection. In all these cases the goal of the data analysis is to provide a bilinear decomposition of mixed raw data into meaningful pure component profiles.

The first description of the MCR method in chemistry was given in 1971 by Lawton and Sylvester (*1*). Their data consisted of a few two dye mixtures at different concentrations measured at 30 wavelengths, from 410 to 700 nm. Their goal was to estimate the UV-visible spectra of the two dye molecules, in the absence of other information. They termed their approach self-modeling curve resolution (SMCR). In this simple data example, the basic aspects of

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

multivariate curve resolution were already defined, including the concepts of ambiguity, feasible solutions and constraints (non-negativity). The region of feasible solutions in the subspace of the two first eigenvectors was obtained and graphically displayed. A similar approach was later extended by Borgen and Kowalski (*2*) to mixtures of three components. Extension to more than three components was difficult and new perspectives to the problem were postponed until more recently (*3*). Early developments of multivariate curve resolution methods can be encountered in the review about Mixture Analysis by Hamilton and Gemperline (*4*). Basically, during this initial period of time, two types of approaches were emerging to solve the multivariate curve resolution problem. On one side there was the proposal of methods attempting to find solutions of the multivariate curve resolution problem using a direct (non-iterative) strategy to calculate algebraically a set of feasible solutions. In these non-iterative approaches, the concepts of rank annihilation (*5*) and of local rank (*6*) were usually used. Examples of this type of approaches were rank annihilation evolving factor analysis, RAFA (*7*), window factor analysis, WFA (*8*), and the heuristic evolving latent projection, HELP (*9*), methods. There were also methods which try to get feasible solutions directly from the data using concepts like 'purest variables', such as the Simple-to-use Interactive Self-Modeling Mixture Analysis, SIMPLISMA (*10*), orthogonal projections, such as Orthogonal Projection Approach, OPA (*11*), or 'key set' variables, such as Key Set Factor Analysis, KSFA (*12*).

On the other side, other approaches were developed to improve initial estimates by means of iterative approaches like Iterative Target Factor Analysis method, ITTFA (*13*, *14*), or from the Evolving Factor Analysis, EFA, and Alternating Least Squares, ALS (*15*), approaches. Following the seminal work of Lawton and Sylvestre (*1*), most of these methods used natural constraints such as non-negativity for the concentration and spectra profiles. From the beginning, it was clear that the application of constraints such as non-negativity and other more powerful constraints was clearly facilitated alternating least squares (*16*) approaches, and this promoted the fast progress of this type of approach.

From early 90's and on, efforts in the development of multivariate curve resolution methods were in two directions. One was in searching unique solutions by means of implementation of constraints during the resolution process, eliminating or minimizing ambiguities. The other trend was the extension of MCR to handle more complex data structures, from the analysis of a single data set ordered in a two-way data table or data matrix, to the simultaneous analysis of multiple related data sets and multiway data structures. Decreasing ambiguities by using constraints can be achieved in multiple circumstances, and playsa central role for in curve resolution (see section 2 of this chapter). According to the resolution theorems formulated by Rolf Manne in 1995 (*16*), the detection and use of local rank and selectivity constraints (*17*) allow for the recovery of the true profiles without ambiguities, whenever the data structure has a favorable structure and conditions. Therefore, resolving a particular multicomponent system without ambiguities depends on many circumstances on the proper estimation of its local rank and selectivity regions and on the proper and active use of these properties. For instance, the presence of selective regions of one particular component

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

in one of the two modes (concentration or spectra), allows for the recovery without ambiguities of the profile of this component in the other mode. Complete resolution of a system will depend on how the profiles of the different components are overlapped and on the fulfillment of the conditions stated by the resolution theorems.

The other main trend sought in MCR development from the 90's was the extension of MCR methods to multiway/multiset data analysis. This was in parallel to the introduction and consolidation of methods like Generalized Rank Annihilation Method, GRAM (*18*), or Parallel Factor Analysis, PARAFAC (*19*), in chemometrics. In section 4 of this chapter, multiset and multiway extensions of the MCR-ALS method are described (*17*), including quadrilinear and mixed multilinear models recently introduced (*20–22*). Another interesting aspect in the development of MCR methods for multiset/multiway data analysis has been the possibility of solving rank deficiency problems present in the analysis of single data matrices. Rank deficiencies occur for instance in chemical reaction systems where concentration profiles of the different constituents show linear dependencies among them. Simultaneous analysis of the reaction system at different initial conditions in matrix augmentation strategies can eliminate rank deficiency problems associated with the analysis of a single data set (*23*, *24*).

A step forward in chemical modelling from the late 90s has been the extension of MCR methods to hybrid hard-soft type of modeling (*25*, *26*). This is especially interesting for systems involving chemical reactions where kinetic or equilibria models can be imposed in the concentration profiles of some of the system constituents. Hybrid approaches are superior to traditional deterministic approaches since they allow for the presence of multiple disturbing effects of the system apart from deterministic (hard) physically modelled part of the measured data variance. It is clear that ambiguities are eliminated from those profiles responding to the applied fundamental laws, which are fitted according to the postulated model parameters, like equilibrium or kinetic constants (and related species stoichiometries).

After all these years of continuous and steady development of MCR methods, the field of multivariate curve resolution has achieved a mature state. In the next sections we will show this situation in more detail, in particular for one of the more extensively applied MCR methods, which is the Multivariate Curve Resolution Alternating Least Squares, MCR-ALS (*17*, *27*), method. In the last sections we will describe some recent application domains of this method, like in environmental studies, in hyperspectral imaging, or in high throughput omics analytical methods which ensure a further and wide spread of this method in the near future.

2. MCR-ALS and Constraints

From a mathematical point of view the mixture analysis problem solved by MCR methods can be described by a bilinear model. In this model, experimental data are arranged in a table or matrix, **D**, where a number of spectra (i=1,...I, or of any other multivariate instrumental response) from a set of samples (e.g.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

chemical mixtures formed by multiple constituents at different concentrations or compositions) are arranged as a row vectors of this data matrix, having their common wavelengths (j=1,...,J, instrumental channels) in the columns of the matrix. The MCR bilinear factor decomposition model can be written using linear algebra notation as:

$$\underset{(I,J)}{\mathbf{D}} = \underset{(I,N)}{\mathbf{C}}\,\underset{(N,J)}{\mathbf{S^T}} + \underset{(I,J)}{\mathbf{E}} \qquad \text{Equation 1}$$

where **C** (concentration profiles) and $\mathbf{S^T}$ (spectra) are the factor matrices obtained by the bilinear decomposition of the experimental data matrix **D**. This bilinear decomposition is performed for a number of components (n=1,..,N), which are contributing to the observed data variance in matrix **D**. In MCR methods, this bilinear decomposition implies that the measured experimental spectra are the linear combinatiosn of the pure spectra of the components present in the analyzed mixtures weighted by their respective concentrations. The same bilinear decomposition in matrix form can be described also by the following element-wise equation

$$\mathbf{D} = \{d_{i,j} + e_{i,j}\} = \sum_{n=1}^{N} c_n s_n + \mathbf{E} \qquad \text{Equation 2}$$

which implies that every element of the data matrix (every individual measurement), $d_{i,j}$, for sample i=1,...,I, wavelength j = 1,..,J , is the sum of different constituent contributions (n=1,..,N), defined by the product of the concentration of these constituents in this sample, $c_{i,n}$, by their signal contribution (intensity) at this wavelength j, $s_{n,j}$.. In these two Equations 1 and 2, **E** and $e_{i,j}$ refer to the non-modelled noise/error/residual contributions. Stated in this way, the bilinear model is expressing the generalization of Beer's law in molecular spectroscopy, at multiple wavelengths and for multiple mixture samples. Although MCR is not restricted to the analysis of spectroscopic data, and without loss of generality, the expressions used in previous equations are by far the most frequently used to express the bilinear MCR model.

It is worth emphasizing that the bilinear decomposition expressed in previous equation is similar to the one used in other chemometrics methods, such as in Principal Component Analysis, PCA (*28*). However, the goals and the way the bilinear matrix decomposition is performed are totally different. In PCA, the bilinear decomposition is performed under the constraint of orthogonality (non-overlapping variance) and with the goal of explaining maximum variance for every component obtained sequentially during the decomposition. PCA goals are usually explorative and interpretative, whereas MCR goals try defining the contributions of the constituents (components) with physically meaningful profiles (concentration and spectra profiles).

As mentioned in the previous section, a typical way of solving the MCR problem is by means of an Alternating Least Squares optimization, using the so-called MCR-ALS method (*17*, *27*, *29*, *30*). Equation 1 is solved iteratively in two constrained least squares steps:

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

$$\min_{\hat{C},\text{constraints}} \left\| \hat{\mathbf{D}}_{\mathbf{PCA}} - \hat{\mathbf{C}}\hat{\mathbf{S}}^{\mathbf{T}} \right\|$$

Equations 3 and 4

$$\min_{S^T,\text{constraints}} \left\| \hat{\mathbf{D}}_{\mathbf{PCA}} - \hat{\mathbf{C}}\hat{\mathbf{S}}^{\mathbf{T}} \right\|$$

In the first least squares optimization step, Equation 3, spectra matrix $\mathbf{S^T}$ is unknown and it is estimated by projection of the PCA filtered data matrix, $\mathbf{D_{PCA}}$, onto the subspace spanned by the current estimation of the concentration matrix, **C**. And in the second least squares optimization step, Equation 4, the same is performed by exchanging the concentration matrix **C**, which is now unknown, by the spectra matrix, $\mathbf{S^T}$ which is now considered known, i.e. by projection of the PCA reproduced data matrix, $\mathbf{D_{PCA}}$, onto the subspace spanned by the current estimation of the spectra matrix, $\mathbf{S^T}$**.** Both equations are solved iteratively by linear least squares under constraints which should be defined previously (see below). The use of PCA filtered matrix instead of the experimental matrix during ALS stabilizes the calculations and filters non-embedded noise in the components.

Some additional remarks about the ALS procedure should be mentioned. First, the initial number of components (constituents) used in the bilinear decomposition and ALS can be estimated as in PCA. This should be equal to the number of components needed to explain 'sufficiently' the systematic changes (no noise) observed in the data variance. To start the ALS optimization, initial estimations of the profiles for the selected number of components are required, either for **C** or for $\mathbf{S^T}$. They can be obtained in different manners, for instance using any of the direct methods previously mentioned in section 2 (*10–12*, *15*) or simply by direct selection of uncorrelated rows or columns of the analyzed data matrix. And, finally, convergence and finalization tests of ALS optimization should be tested by means of relative changes in data fitting parameters (such as lack of fit), between consecutive iterations, maximum number of iterations, and maximum number of divergence steps. MCR-ALS has been described in detail in several publications (*7*, *27*) and its implementation in a MATLAB user friendly graphical interface (*29*, *30*) is freely available (*31*).

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

Description of Constraints

Constraints are the corner stone of MCR iterative methods and can be defined as the systematic properties used to bring the iterative resolution process to get the optimal and chemically meaningful solution for the concentration and/or response profiles (*17*, *32*, *33*). After selection of appropriate initial estimates, the alternating least squares optimization of concentration profiles and pure responses under constraints can start. Implementation of constraints converts the chemical or mathematical knowledge about the profiles into a mathematical condition, which can be set in two ways: forcing the profile (or some elements in a profile) to be equal to some pre-set values or to be higher or lower than them. These two options define the so-called equality and inequality constraints, respectively (*33*, *34*). Inequality constraints are generally preferred since they modify the profiles more gently, and minimize the disturbance to the convergence of the resolution

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

process. However, the sound application of an equality condition in some instances (e.g., accurate knowledge of a pure spectrum) significantly decreases the ambiguity of the final resolution results. The kinds of knowledge that can be incorporated into a constraint are very diverse. The main distinction is between constraints linked to chemical properties of the concentration or response profiles and mathematical properties dependent on the inner structure of the data set (graphical summary for some of the constraints is provided in Figure 1. Typical chemical constraints are as follows:

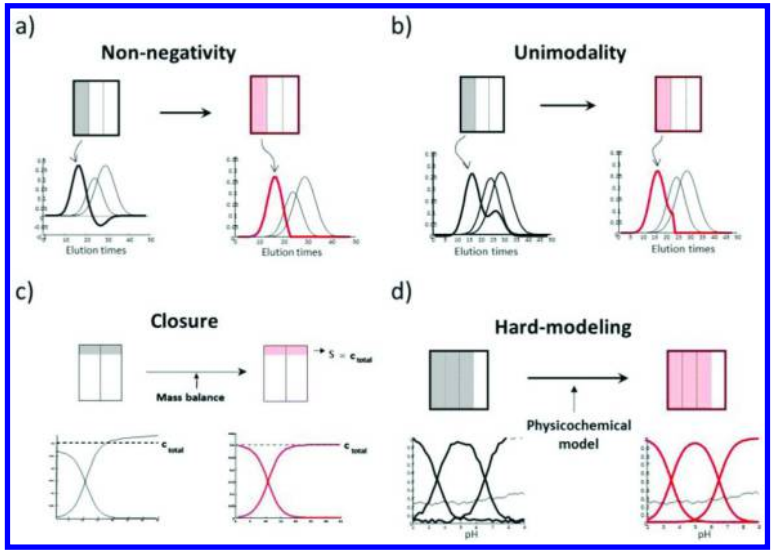

*Figure 1. Examples of common constraints used in MCR. a) non-negativity, b) unimodality, c) closure, d) hard-modeling. Bold black profiles on the left plots are unconstrained. Red profiles on the right profiles have been constrained. (reproduced from de Juan A.; Tauler R. Crit. Rev. Anal. Chem., **2006**, 36, 163)*

Non-Negativity

Forces the profiles to be positive and it can be implemented replacing negative values by zeros or with softer algorithms, such as non-negative least-squares or fast non-negative least-squares. It applies to all concentration profiles and to many instrumental responses which, by nature, should be positive, for example ultraviolet (UV) absorption spectra, mass spectra, etc.

Unimodality

This constraint allows the presence of only a single maximum per profile. This specific condition is fulfilled by peak-shaped signals, for example, elution profiles,

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

some voltametric signals, and monotonic reaction profiles (always increasing or decreasing).

Closure

This is a mathematical expression for the mass balance condition in chemistry. It applies to some concentration profiles of reaction systems, forcing the concentration profiles within a closed system to add up to a certain constant value (the closure constant).

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

Known Pure Spectra /Concentration Profiles

This is a kind of equality constraint, which makes the concentration profile and/or spectrum of a component to be equal to a certain known predefined shape according to a mathematical function. In the concentration directions, these functions are, usually, physicochemical models (kinetic or equilibrium).

Mathematical constraints usually refer to properties linked to subspaces of data sets, for example, concentration windows with a certain rank value or zero-concentration windows for particular components. Constraints related to mathematical conditions are:

Local Rank/Selectivity

This constraint defines the zones of absence of components in some regions (windows) , usually, in the concentration profiles. The selectivity expresses the presence of a single component in a specific concentration window, and the local rank information is about the absence of some components in the concentration window. Incorporation of local rank information suppresses the ambiguity in the profiles retrieved by MCR (*16*, *27*).

Correspondence of Species

This condition is only applicable to the augmented data matrices (with augmented concentration direction) and expresses the correspondence and presence/absence of components in the analyzed samples (*17*, *27*, *35*, *36*). It is a strong constraint to suppress ambiguity, similarly to that of local rank. The known composition of standard samples (presence/absence of analytes and interferences), makes it a desirable constraint in calibration problems.

Model Constraints

These are constraints applicable to multiset and only to the augmented matrices that may be present in the decomposition model. Although MCR gives by default a bilinear model, the conditions of trilinearity (*20*, *36*), multilinearity (*37*) or factor interaction (like in Tucker models) (*38*, *39*) can be implemented in a component-wise way. Therefore, completely bilinear, completely trilinear or hybrid models can be set to be obtained in the final MCR results.

Hard-Modeling

Forces the concentration profiles to be fitted by a parametric physicochemical model and the parameters of the model are obtained as an additional output. This constraint implies a model fitting task be performed during the iterative optimization process (*36*, *40*). When this constraint is applied, the related concentration profiles do not present ambiguity. It can also be applied to pure analytical responses when the shape of the pure signal can be defined by a parametric equation.

Correlation Constraint

In this case, internal univariate calibration models can be applied to particular concentration profiles of components in the system. The model is established between concentration values obtained with MCR (in arbitrary units) and real concentrations in calibration samples. The model is used to predict real concentrations in unknown samples (*35*, *41–43*)

All these constraints are applied optionally and can be implemented differently to the concentration and response profiles of the data set, to the different components in the data set and to the different subsets in a multiset structure. The application of constraints can also be modulated according to tolerance criteria. The flexibility in the application of the constraints explains the versatility of MCR algorithms, which can adapt to very diverse scenarios through the proper selection of these restricting conditions.

In MCR, the constraints can be implemented using two approaches: the first approach directly incorporates the constraints into the least-squares procedure (*44–47*) via a penalty function (*48*, *49*) or related method, whereas the second one constrains the profiles in separate steps from the least-squares fitting (external constraints). The direct methods have the advantage of guaranteed convergence, but implementations are not available for all of the constraints described above, and many of the available constraints are computationally expensive. The external methods can sometimes have more convergence problems; however, their implementation is generally much faster and their application more flexible than the least-squares optimal constraints. Notably, these external implementations allow for a profile-wise selection in the application of constraints, which is not feasible in more strict least-squares procedures. These methods are often

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

employed in a manner that allows for some deviations from the strict application of the constraint condition (*17*, *50*). Whatever is the approach, used to implement the constraints, the problem to be solved should always involve the application of constraints fulfilled by the profiles (concentration, **C**, and spectra, $\mathbf{S}^T$, in Equation 1) describing the true data variance sources and, in no case, an alternative set of profiles fitting better the data and not fulfilling the applied constraints (apart from noise and loss of degrees of freedom) is possible. Therefore, the MCR solution fulfilling the constraints should be the best solution also from a least-squares point of view. In this sense, constraints are mostly used to guide the optimization to the most physically meaningful solution and to reduce rotation ambiguities.

The application of constraints should be fully justified from a chemical or a mathematical point of view. Misapplying a constraint can produce much worse results than performing a completely unconstrained optimization. In case of doubts about the appropriateness of introducing a particular constraint to resolve a data set, some guidelines can be followed, such as gradually introducing constraints in the resolution process, following an increasing order of strength, and checking the effect of the introduction of each constraint on the variation of the fit quality in the reproduction of the raw data set. A significant decrease in the fit and the emergence of residuals with non-random trends linked to the introduction of a particular constraint may imply that the data set does not really fulfil the selected condition or that, at least, some deviations from the ideal behaviour should be permitted.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

3. Extended MCR to Multiset and Multiway Data Analysis

MCR analysis is enhanced significantly when multiple data sets or multiset data are simultaneously analyzed using the extension of this method (*17*, *27*, *51*). The multiset data are analyzed by MCR via matrix augmentation schemes, and the typical application of MCR bilinear models to augmented data matrices can be extended to impose multiway structures during the resolution process in the form of constraints (*35*). The most commonly used data arrangement is column-wise augmented matrices, $\mathbf{D}_{aug}$, keeping the same number of columns, which, in MATLAB notations is written as $\mathbf{[D_1;D_2;D_3;...;D_K]}$ for different data matrices $\mathbf{D}_k$, k=1,..K. For this type of augmented data matrix, the MCR bilinear model can be written as:

$$\mathbf{[D_1;D_2;D_3;...;D_K] = [C_1;C_2;C_3;...;C_K]\ S^T + [E_1;E_2;E_3;...;E_K]}$$

$$\begin{pmatrix} \mathbf{D}_1 \\ \mathbf{D}_2 \\ \mathbf{D}_3 \\ \cdots \\ \mathbf{D}_K \end{pmatrix} = \begin{pmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \\ \mathbf{C}_3 \\ \cdots \\ \mathbf{C}_K \end{pmatrix} \mathbf{S}^T + \begin{pmatrix} \mathbf{E}_1 \\ \mathbf{E}_2 \\ \mathbf{E}_3 \\ \cdots \\ \mathbf{E}_K \end{pmatrix} = \mathbf{C}_{aug}\mathbf{S}^T + \mathbf{E}_{aug} \qquad \text{Equation 5}$$

Or in a more compact form

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

$$\mathbf{D_{aug}} = \mathbf{C_{aug}}\,\mathbf{S^T} + \mathbf{E_{aug}}, \qquad \text{Equation 6}$$

The bilinear model described above is formulated using a single spectra matrix $\mathbf{S^T}$ with the pure spectra of the different components present in all the considered $\mathbf{D_k}$ data matrices. The augmented concentration matrix $\mathbf{[C_1;C_2;C_3;...;C_K]}$ describes freely the concentration changes of the resolved components in each related $\mathbf{D_k}$ data matrix. The multiset analysis facilitates the possible submatrix-by-submatrix application of new constraints. One of these constraints is the correspondence among the components in the different simultaneously analyzed matrices. This constraint fixes the sequence and the presence or absence of components in a particular $\mathbf{C_k}$ and/or $\mathbf{S^T_l}$ matrices. This type of constraint contributes significantly to the elimination of rotation ambiguities, which facilitates the achievement of unique resolution conditions (*16*, *17*, *27*). In case of multiway data (also called multimode data), the bilinear model previously described for augmented data matrices can also be extended to the so-called multilinear models, like the PARAFAC (*19*, *52*) or the Tucker models (*39*, *52*). In MCR-ALS, these type of multilinear models are implemented as constraints during the alternating least squares optimization as explained in detail in (*17*, *27*, *36*, *51*). Next, a brief description of the implementation of these model constraints can be found.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

The Trilinearity Constraint in MCR-ALS

The expression of the trilinear model to describe decomposition of a three-way data set, is given element-wise as:

$$d_{ijk} = \sum_{n=1}^{N} c_{in} s_{jn} z_{kn} + e_{ijk} \qquad \text{Equation 7}$$

and the reproduction of each slice data matrix (slice k) is carried out as follows:

$$\mathbf{D_k} = \mathbf{C}\,\mathbf{Z_k}\,\mathbf{S^T} + \mathbf{E_k} \qquad \text{Equation 8}$$

where in Equation 7, d_{ijk} represents the ijk^{th} element in the three-way data set (I = 1,..I, j = 1,...J and k = 1,...K slices), n is the number of components (chemical rank) common to the three modes (n = 1,…,N), c_{in}, s_{jn} and z_{kn} are the elements of $\mathbf{C}$, $\mathbf{S^T}$ and $\mathbf{Z}$ factor matrices (component profiles in the three modes) for component n and e_{ijk} is the residual term (part of the data not explained by the model). The chemical meaning of the factor matrices **(C** and $\mathbf{S^T}$**)** is the same as for the analogous matrices considered in the description of the bilinear model decomposition in MCR analysis, and the factor matrix $\mathbf{Z_k}$ of loadings in the third direction or mode is a diagonal matrix giving the relative amounts of every component in each considered data matrix $\mathbf{D_k}$. A trilinear model forces decompositions to give unique solutions for the three factor matrices (apart from scale and trivial permutation rotation ambiguities), and avoids the presence of rotational ambiguities associated with lower structured bilinear models (*53*). Because of the inherent freedom in the modeling of the profiles of the augmented

$\mathbf{C_{aug}}$ matrix, the so-called trilinear structure is incorporated as an optional constraint during the ALS optimization of the $\mathbf{C_{aug}}$ profiles (see Figure 2) (*20*).

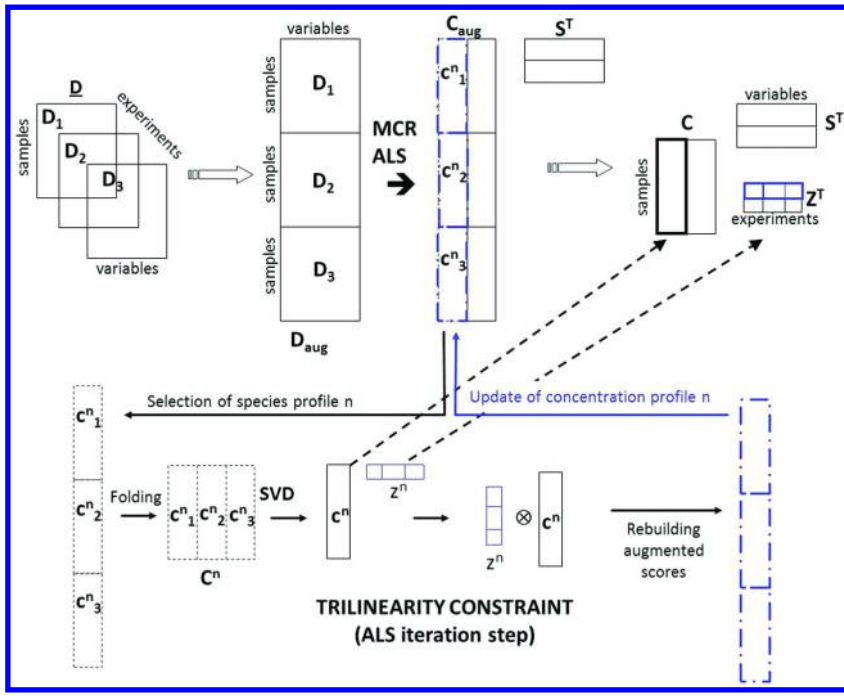

Figure 2. Implementation of the trilinearity constraint in the MCR-ALS algorithm (Ä stands for the Kronecker product between two vectors, see 54 and text for details of the procedure), reproduced from Comprehensive Chemometrics, S. Brown, R. Tauler, and R. Walczak, Eds. Elsevier, Oxford, 2009 Vol. 2, pp 473-505.

As shown in Figure 2, when the trilinear constraint is applied during each iteration of the ALS optimization, the concentration profiles of the same component in the different matrices, $\mathbf{c^n_k}$ (k=1,2,3, n=1,2 components in this case), are forced to be invariant in shape. For a full trilinear model, every component of the matrix, $\mathbf{C^n}$, is approximated by its one-component bilinear decomposition (using for instance PCA or SVD) as follows:

$$\mathbf{C^n \approx c^n z^{nT}} \qquad \text{Equation 9}$$

where $\mathbf{c^n}$ is one column vector (I rows) which contains the common (average) concentration profile of this *n* component in the different *K* matrices and $\mathbf{z^{nT}}$ is a row vector (N columns) with the relative amounts of this concentration profile in the different *K* matrices. The appropriate Kronecker product of these two vectors rebuilds the augmented concentration vector of this component (*54*). $\mathbf{C_{aug}}$ augmented matrix is finally obtained and updated by appropriate rearrangement of each of its columns corresponding to the concentration of different individual components in the different matrices. $\mathbf{c^n}$ and $\mathbf{z^{nT}}$ vector profiles give the

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

current estimation of the first and third mode loading profiles of the considered component. When MCR-ALS optimization finishes, only the recovered spectral information in $\mathbf{S^T}$ matches with the PARAFAC profiles. However, the matrix $\mathbf{C_{aug}}$ contains implicitly the information related to matrices **C** and **Z,** which can be recovered using a similar procedure as in the application of the trilinearity constraint in Figure 2.

The Quadrilinearity Constraint in MCR-ALS

The procedure previously described for the implementation of the trilinearity constraint for three-way data can be extended to the implementation of the quadrilinearity constraint for four-way data. The expression of the quadrilinear model to describe the decomposition of a four-way data set '$\mathbf{D}_{IJK,L}$', is given element-wise, as:

$$d_{ijkl} = \sum_{n=1}^{N} c_{in} s_{jn} z_{kn} y_{ln} + e_{ijkl} \qquad \text{Equation 10}$$

where d_{ijkl} represents the ijklth element in the four-way data set (i = 1,..I; *j* = 1,...J; k = 1,...K; and L=1,…L), *n* is the number of components (chemical rank) common to the four modes (*n* = 1,…,N), c_{in}, s_{jn}, z_{kn} and y_{ln} are the elements of **C**, $\mathbf{S^T}$, **Z**, and **Y** factor matrices (component profiles in the four modes) and e_{ijkl} is the residual term (part of the data not explained by the model). The chemical meaning of the factor matrices **(C** and $\mathbf{S^T}$**)** is the same as for the analogous matrices considered in the description of the bilinear model decomposition in MCR analysis, and the factor matrices **Z**, and **Y** belong to the loadings in the third and fourth directions or modes, respectively.

A four-way dataset '$\mathbf{D}_{IJK,L}$', of dimensions I,J,K, and L in the 1st, 2nd, 3rd, and 4th mode respectively, can be arranged in an augmented column-wise manner ($\mathbf{D_{aug}}$) where three of the modes (I, K and L) are concatenated and intermixed in the column direction of the data matrix, and the other mode J is left invariant in the row direction (Figure 3). MCR-ALS can be applied to this augmented data matrix as described in Equation 6, $\mathbf{D_{aug}} = \mathbf{C_{aug}}\mathbf{S}^T + \mathbf{E_{aug}}$, where $\mathbf{C_{aug}}$ is the augmented concentration matrix containing the first, second and third mode profiles and $\mathbf{S^T}$ is the spectra (loadings) matrix for the second mode, and $\mathbf{E_{aug}}$ is the error term.

Schematic representation of the quadrilinearity constraint implemented in MCR-ALS model (MCR-ALS$_Q$) to analyse a four-way dataset, $\underline{\mathbf{D}}_{\mathbf{I,J,K,L}}$, is provided in Figure 3. In the same figure, the relation with bilinear (MCR-ALS$_B$) and trilinear modelling (MCR-ALS$_T$) and the application of corresponding modelling constraints is also displayed. In brief, the quadrilinear constraint is applied during the ALS optimization of the augmented concentration matrix ($\mathbf{C_{aug}}$). This quadrilinear constraint can be applied independently and optionally to each component of the data set, giving more flexibility to the whole data analysis and allowing testing full and partial quadrilinear models. This approach can be extended algorithmically to any multilinear model. See references (*22*, *37*) for preliminary applications of this constraint to environmental data sets. Following these ideas, a generalization of the implementation of trilinearity and

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

quadrilinearity constraints to any type of multilinear model in MCR-ALS is possible and under development.

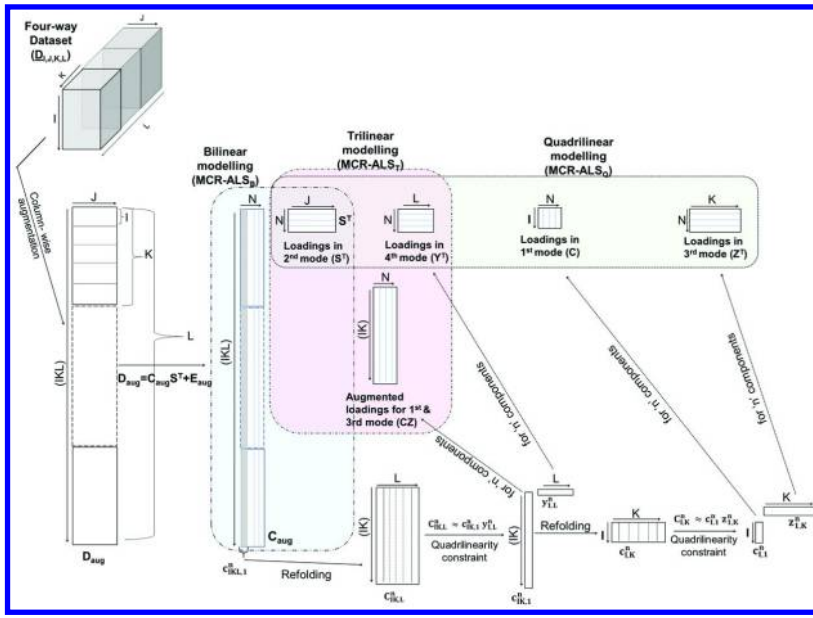

Figure 3. Implementation of the quadrilinearity constraint in the MCR-ALS algorithm (see text for details of the procedure); reproduced from Malik, A.; Tauler, R. Anal. Chim. Acta, ***2013****, 794, 20-28.*

Interaction between the Profiles of Different Components

The general expression to describe the decomposition of three-way data sets where interaction between components is possible is given in Equation 11:

$$d_{ijk} = \sum_{p=1}^{Np} \sum_{q=1}^{Nq} \sum_{r=1}^{Nr} g_{pqr} c_{ip} s_{jq} z_{kr} + e_{ijk} \qquad \text{Equation 11}$$

where d_{ijk} represents the ijk^{th} element in the three-way data set, c_{ip}, s_{jq} and z_{kr} are the elements in **C**(I,Np), **S**(J,Nq**)** and **Z**(K,Nr) factor matrices (loadings in the three modes) used to reconstruct the d_{ijk} element of $\underline{\mathbf{D}}$(I,J,K) and e_{ijk} is the residual term in $\underline{\mathbf{E}}$(I,J,K). *Np, Nq* and *Nr* are the number of components considered in each of the three modes, not necessarily equal as in the trilinear model, in which *Np=Nq=Nr=N* (Equation 7). g_{pqr} is the pqr^{th} element of the core array $\underline{\mathbf{G}}$(Np,Nq,Nr), where the non-null elements are spread out in different manners depending on each particular data set. The decomposition of a three-way array $\underline{\mathbf{D}}$ according to Equation 11 is called the Tucker3 model (*39*, *52*) in multiway literature. Restricted Tucker3 models are a simpler type of Tucker3 models, where only a small number of possible selected interactions (triads) is allowed. To do so,

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

the elements of the core matrix **G** unrelated to the selected triads are set equal to zero.

MCR-ALS can be adapted to implement situations where the components interact, with different number of profiles (loadings) in the different modes. Figure 4 shows graphically how this can be achieved for an example where two components have the same shape for their concentration profiles. In the example shown in this Figure, the two suitable column profiles on the augmented $\mathbf{C_{aug}}$ matrix are first grouped to give the row-wise augmented concentration matrix $[\mathbf{C^1,C^2}]$. This $[\mathbf{C^1,C^2}]$ matrix has a number of rows equal to the number of rows in the individual matrices and a number of columns equal to twice the number of matrices simultaneously analyzed because two components have the same profile shape (2xK). This folded matrix containing the profiles of the two components with common shape for the concentration profile is then approximated by their bilinear decomposition (using for instance the first component of PCA or SVD) as:

$$[\mathbf{C^1,C^2}] \approx \mathbf{c}\,\mathbf{z}^T \qquad \text{Equation 12}$$

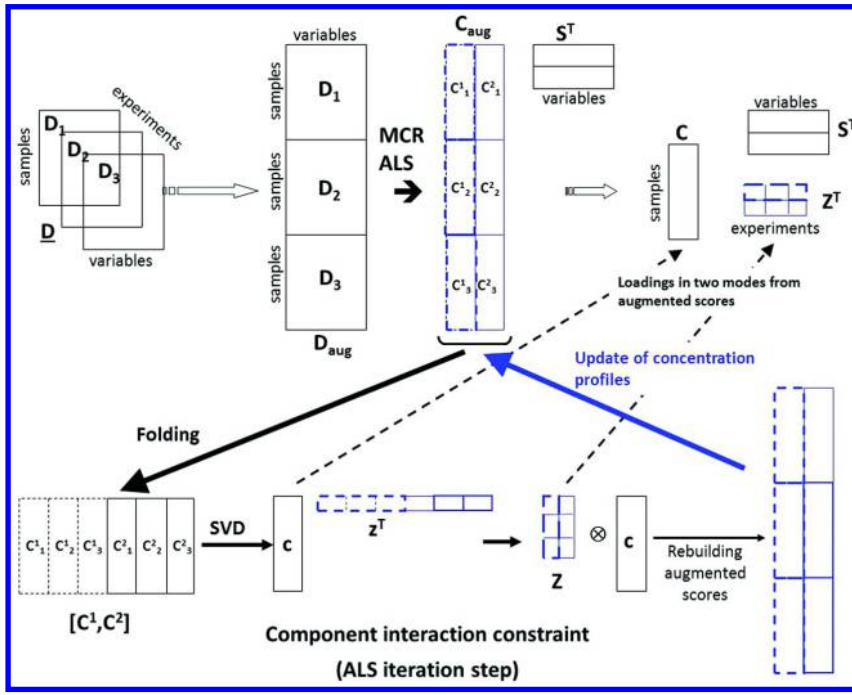

Figure 4. Implementation of the Tucker3 model constraint in the MCR-ALS algorithm (Ä stands for the Kronecker product between two vectors, see 54 and text for details of the procedure); reproduced from Peré-Trepat, E.; Ginebreda, A.; Tauler, R. Chemom. Intell. Lab. Syst. ***2007****, 88, 69–83.*

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

This first component bilinear decomposition gives directly the loadings (the shape of the concentration profile) in the first mode **c**, and the loadings (scaling factor) in the third mode $\mathbf{z}^T$, which after adequate component-wise rearrangement gives matrix **Z** (see Figure 4).

The appropriate Kronecker product (*39*, *54*) of them gives the two new augmented profile vectors that substitute (update) the corresponding two columns of the $\mathbf{C}_{aug}$ matrix. Observe that in this case only one profile (one loading vector) will be recovered in the first mode for the matrix **C**. On the contrary, two profiles (two loading vectors) are recovered for the third mode matrix $\mathbf{Z}^T$, as well as for the second mode, in the $\mathbf{S}^T$ matrix. When these component interaction constraints are inserted during each step of the ALS iterative optimization procedure, the results obtained are analogous to those obtained for the restricted Tucker3 model with one component in the first mode and two components in the two other modes. This procedure can be generalized to other cases and it has been tested in some data examples (*21*, *36*, *38*).

The augmented data matrix decompositions using bilinear model MCR-ALS, give directly the loadings in the second mode ($\mathbf{S}^T$ matrix) only, whereas the loadings in the other modes are confounded in the augmented mode ($\mathbf{C}_{aug}$ matrix). Using a proper profile rearrangement and SVD analysis on the suitable augmented concentration profiles, as proposed in Figures 2-4, the loadings in the different modes can be recovered, irrespective of the application of the multilinearity constraints. An advantage of the multi-linearity constraint as it is implemented in MCR-ALS is that it is applied independently and optionally to each component of the data set, giving more flexibility to the whole data analysis and allowing for full multilinear and for partial multiilinear models.

4. Reliability of MCR Solutions

As described earlier, MCR methods decompose the data matrix, **D,** into the product of a concentration matrix **C** and of a spectral matrix $\mathbf{S}^T$ using a bilinear model (Equations 1-2). The contribution of each component to the whole measured signal is the rank one matrix obtained by the vector product of its concentration profile by its pure spectrum, i.e. for component *n*, $\mathbf{c_n s_n}^T$. **E** matrix in Equation 1 gives the part of data matrix **D** that is not described by the matrix product $\mathbf{CS}^T$. Although MCR solutions have more physical meaning and an easier interpretation than those obtained by PCA or SVD, they are not unique in the general case. There are three types of ambiguities which usually occur in the MCR solutions: permutation ambiguity, intensity (or scale) ambiguities, and rotation ambiguities (*17*, *55*). Permutation ambiguity means that there is no sorting order on the MCR components. Therefore, they can be shuffled in the **C** and $\mathbf{S}^T$ matrix keeping the appropriate dyad correspondence of components and obtain identical results. Under scalar ambiguity it is understood that only the shapes of concentration profiles and spectra can be determined. Multiplication of any concentration profile with a scalar and the corresponding spectrum with the inverse of the scalar has no net effect. For any component or species, n = 1...N, this can be illustrated as:

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

$$\mathbf{c_n}\,\mathbf{s_n}^T = (\mathbf{c_n}\,k)\;\;(k^{-1}\,\mathbf{s_n}^T) \qquad \text{Equation 13}$$

which states that there is a scale indeterminacy in the product of concentration by spectra profiles. If the scale of one of both, $\mathbf{c_n}$ or $\mathbf{s_n}^T$, is known for one or several components, their scalar ambiguity problem is solved. Or if, as it is often done, the scale is fixed arbitrarily, then the problem is also solved for this arbitrary selected scale. Due to the unavoidable scale ambiguity, only the shapes of the concentration and spectra profiles can be determined and no absolute scale (quantitative) information can be obtained in general, directly from MCR methods unless external quantitative calibration information is provided during the resolution process, like it is done for instance, in multivariate calibration methods. However, relative quantitative information can be obtained directly by curve resolution methods, when they are applied simultaneously to multiple data sets or data matrices (*56–58*), like in other multiway data analysis methods (*52*).

The more critical and difficult type of ambiguity to be avoided in MCR solutions is the so called rotation ambiguity. In absence of any constraint, Equation 1 has an infinite number of solutions, since there are an infinite number of matrices **C** and $\mathbf{S}^T$, providing the same result, the data matrix **D**. This indeterminacy can be described mathematically as:

$$\mathbf{D} = \mathbf{C}\,\mathbf{S}^T = (\mathbf{C}\,\mathbf{T}^{-1})\,(\mathbf{T}\,\mathbf{S}^T) = \mathbf{C_{new}}\,\mathbf{S_{new}}^T \qquad \text{Equation 14}$$

According to Equation 14, any invertible matrix **T**(*N*,*N*) gives a new set of equivalent solutions of the MCR model. Or said in other words, any linear combination of **C** and $\mathbf{S}^T$ solutions will produce new solutions of the bilinear model which will be equivalent from a mathematical point of view. The application of appropriate constraints in MCR methods can limit the rotation ambiguities. Apart from natural constraints like non-negativity, the more powerful strategies to avoid rotation ambiguities in MCR methods are the use of local rank and selective constraints (*16*, *17*), the extension to simultaneous analysis of multiple data sets (*17*, *20*, *27*) and the use of hard (deterministic) modeling (*26*). Different methods have been proposed in the literature for the evaluation of rotational ambiguities, including the development of methods for the calculation of the boundaries of the so called feasible bands (*1–3*, *53*, *59–63*)

Calculation of the Extent of Rotation Ambiguities Using the MCR-BANDS Method (*53*, *64*)

Feasible MCR solutions include the whole range of linear combinations of a particular MCR solution that fit the experimental data equally well and fulfil the constraints of the system, as defined by appropriate rotation matrices **T** (Equation 14). For every component rotation matrices, **T**, define the range or band of feasible solutions. Consider the possibility to define maximum and minimum values of these rotation matrices, $\mathbf{T_{max}}$ and $\mathbf{T_{min}}$, which should fulfill the following equation:

$$\mathbf{D} = \mathbf{C_{init}}\,\mathbf{S_{init}}^T = \mathbf{C_{init}}\,\mathbf{T_{min}}\;\;\mathbf{T_{min}}^{-1}\,\mathbf{S_{init}}^T = \mathbf{C_{min}}\,\mathbf{S_{min}}^T =$$
$$= \mathbf{C_{init}}\,\mathbf{T_{max}}\,\mathbf{T_{max}}^{-1}\,\mathbf{S_{init}}^T = \mathbf{C_{max}}\,\mathbf{S_{max}}^T \qquad \text{Equation 15}$$

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

In the previous equation, initial values of **C** and $\mathbf{S^T}$ matrices, $\mathbf{C_{init}}$ and $\mathbf{S^T_{init}}$, are known and $\mathbf{C_{min}}$, $\mathbf{S^T_{min}}$ and $\mathbf{C_{max}}$, $\mathbf{S^T_{max}}$ correspond to $\mathbf{T_{min}}$ and $\mathbf{T_{max}}$ values. A possible algorithm to determine the extent of rotation ambiguities is proposed based on the definition of an objective function which should be maximized and minimized as a function of **T** in order to find $\mathbf{T_{max}}$ and $\mathbf{T_{min}}$ values for every resolved component. This objective function should be a scalar function of the variables and should have well defined boundaries (maximum and minimum). For a good performance of the optimization algorithm, this optimization function is scaled, for instance between 0 and 1. The proposed optimization function is defined as follows (*53*, *60*):

$$f_n(\mathbf{T}) = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}(c_{i,n}(\mathbf{T})s_{j,n}(\mathbf{T}))^2}{\sum_{n=1}^{N}\sum_{i=1}^{I}\sum_{j=1}^{J}(c_{i,n}(\mathbf{T})s_{j,n}(\mathbf{T}))^2} = \frac{\left\|\mathbf{c}_n(\mathbf{T})\mathbf{s}_n^{\mathbf{T}}(\mathbf{T})\right\|}{\left\|\mathbf{CS}^{\mathbf{T}}\right\|} \qquad \text{Equation 16}$$

This function gives the ratio between the contribution of a particular species n (numerator of Equation 16) with respect to the total contribution for all the components of the mixture (denominator of Equation 16). The optimization of this objective function under constraints (see below) for each component n=1,..N, either maximized or minimized, will give respectively an estimate of its maximum and minimum solutions ($f_n(\mathbf{T})$ max and min values), from which the corresponding $\mathbf{T_{max}}$ and $\mathbf{T_{min}}$ matrices will be obtained as well as the corresponding $\mathbf{c_{n,max}}$, $\mathbf{s^T_{n,max}}$ and $\mathbf{c_{n,min}}$ and $\mathbf{s^T_{n,min}}$ profile dyads, for every one of the resolved component, n=1,...N. These extreme solutions should fulfill the constraints of the problem and give the relative maximum and minimum signal contribution of every component according to the function defined by $f_n(\mathbf{T})$ in Equation 16, i.e. the ratio of the norm of $\mathbf{c_n s_n^T}$ over the norm of the whole signal contribution from all components, $\mathbf{CS^T}$. The variables to optimize are the elements of the rotation matrix **T,** As pointed out previously, the number of variables in matrices **T**, and therefore the complexity of the optimization, increases with the number of components of the system. Since this is a non-linear optimization, initial values of the variables (**T)** and of the component profiles ($\mathbf{C_{inic}}$ and $\mathbf{S^T_{inic}}$) are required. Details of the implementation of the algorithm, called MCR-BANDS, are given in (*53*, *64*). The optimization of the function given by Equation 16 under constraints is performed using a non-linear constrained non-linear optimization problem which is solved using a Sequential Quadratic Programming (SQP) algorithm implemented in the MATLAB Optimization Toolbox *fmincon* function (*65*).

Figure 5 shows an example of the extent of rotation ambiguities, calculated for concentration/elution profiles, **C**, and spectra, $\mathbf{S^T}$, in a chromatographic separation of three coeluting components obtained by using multi-wavelength UV diode array detection. The considered data matrix **D** has strongly overlapped chromatographic elution profiles. The detailed description and discussion about the dataset and results can be found in (*53*). An indicator of the extent of rotation ambiguity for component n can be the difference between maximum and minimum values of the $f_n(T)$ function, which will be zero in case of unique solutions and will increase with

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

extent of ambiguity. Plotting the profiles related to the boundaries, as in Figure 5, will indicate the location and magnitude of the rotation ambiguity, and which profiles are most affected by this phenomenon.

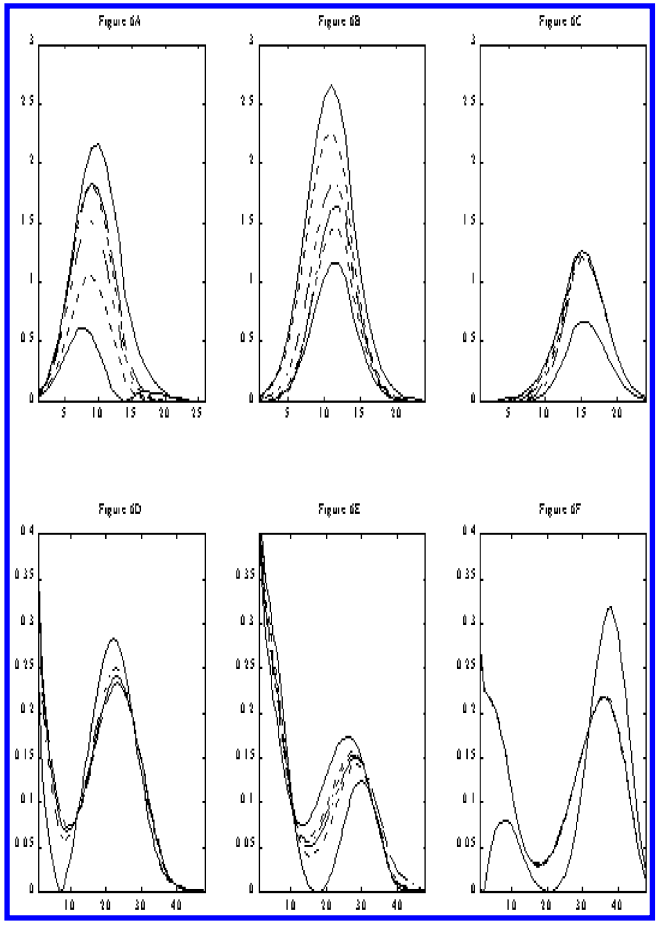

Figure 5. Extent of rotation ambiguities calculated for the component profiles of data matrix ***D*** *consisting of overlapped chromatographic elution profiles. Solid lines give the extreme solutions (Equation 16) obtained by MCR-BANDS using the non-negativity and spectra normalization constraints. Dotted lines are the MCR-BANDS solutions using non-negativity, spectra normalization and selectivity/local rank constraints. Dash-dotted lines are the 'true' profiles. Dashed lines are the profiles obtained using MCR-ALS. Figures 5A, 5B and 5C give the elution profiles of species 1, 2 and 3, respectively, and Figures 5D, 5E and 5F the corresponding spectra profiles; reproduced from R. Tauler J. Chemom.* ***2001****, 15, 627.*

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

Error Propagation and MCR Solutions

Evaluation of error estimates due to propagation of experimental errors is also important in the quality assessment of MCR results. Analytical propagation expressions can be proposed when working with linear methods and known experimental uncertainties (*66*). Unfortunately, these expressions are unknown for MCR methods. In order to obtain error estimations and confidence limits of parameters in these cases, it is rather common nowadays to use resampling methods (*67*). In MCR methods, it should be considered that rotation ambiguities are simultaneously present and that both problems cannot be separated. For brevity only one example of estimation of error propagation effects using Monte Carlo simulations will be presented for the case of MCR solutions obtained using the MCR-ALS method. See previous works for more details (*67–69*).

The data example chosen here for illustration is the spectrophotometric titration of the polynucleotide poly(I)-poly(C). in the pH range between 2.0 and 8.1. UV absorption spectra were recorded from 240 to 320 nm, at every 1 nm. In the pH range studied, three acid-base species were identified. From the concentration, $\mathbf{C}$, and spectra profiles, $\mathbf{S^T}$, of the three species obtained in the investigation of this system, the data matrix $\mathbf{D_{sim}}$ was generated, $\mathbf{D_{sim} = C\ S^T}$. Due to the selectivity of the system at the beginning and end of the titration, using appropriate initial estimates, MCR-ALS analysis of the data matrix, $\mathbf{D_{sim}}$, with non-negativity and closure constraints always converged to the same solution without ambiguities. Random homoscedastic white noise matrices $\mathbf{N_n}$, were generated with zero mean and with values of their relative standard deviation at four different levels (0.1%, 1%, 2%, and 5%) of the maximum intensity of the measured signal in the data matrix $\mathbf{D_{sim}}$. These noise matrices were then added to the theoretical simulated data matrix $\mathbf{D_{sim}}$ giving Monte Carlo simulated data matrices $\mathbf{M_n}$, with a known amount of noise $\mathbf{M_n = D_{sim} + N_n}$, where n indicates 0.1%, 1%, 2% and 5% noise levels. At each noise level (0.1%, 1%, 2% and 5%), 250 replicates were generated and analyzed by MCR-ALS under non-negativity and closure constraints and, new estimates of the pure spectra and concentration profiles of the resolved species for each one of the $\mathbf{M_n}$ replicate matrices were obtained. From these resolved profiles, error estimates were obtained.

Concentration and spectra profiles resolved for the Monte Carlo data sets, $\mathbf{M_n}$, are shown in Figures 6 and 7, respectively. At the lower error levels, profiles were not much distorted and error bands were narrow. In Figure 6, concentration profiles obtained at the higher error levels of 5% already showed distorted shapes, which were rather different to the original profiles. At this error level of 5% also, only the spectrum for the more acidic species was correctly recovered with a narrow error band. The other two species spectra showed a much wider error band and distorted shape (Figure 7).

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

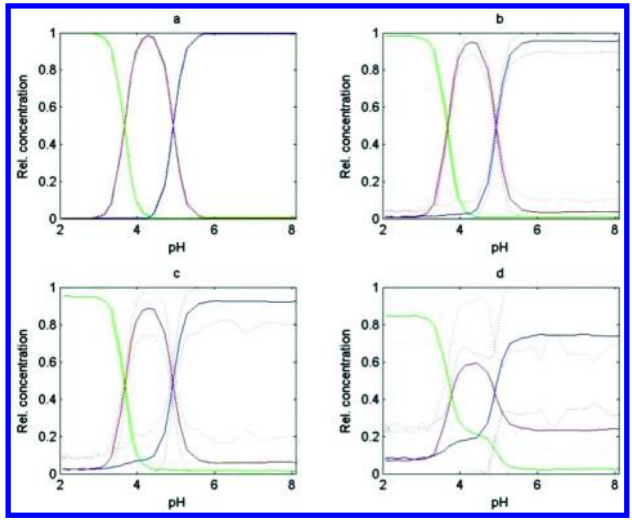

Figure 6. Results obtained at several error levels by Monte Carlo simulations. Mean (solid line), maximum and minimum (dashed lines) concentration band profiles (a) at 0.1%; (b) at 1.0%; (c) at 2.0%; (d) at 5.0% (reproduced from Jaumot, J.; Gargallo, R.; Tauler, R. J. Chemom. ***2004****, 18, 327–340).*

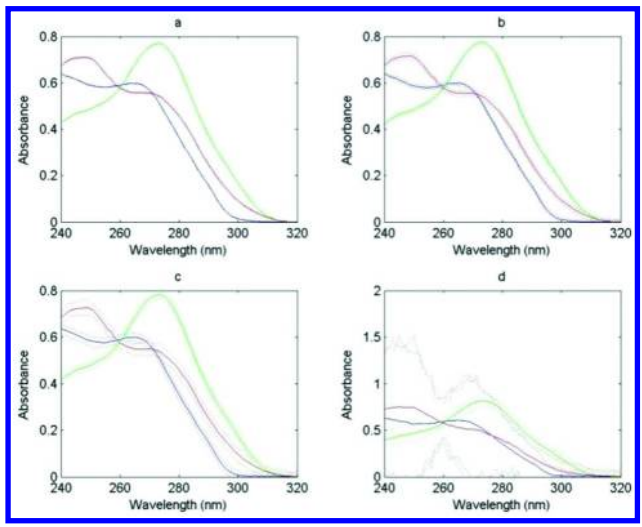

Figure 7. Results obtained at several error levels by Monte Carlo simulations. Mean (solid line), maximum and minimum (dashed lines) spectra band profiles (a) at 0.1%; (b) at 1.0%; (c) at 2.0%; (d) at 5.0% (reproduced from Jaumot, J.; Gargallo, R.; Tauler, R. J. Chemom. ***2004****, 18, 327–340).*

At high noise levels, due to the high overlapping of the species spectra, and to the fact that selectivity and local rank constraints were not applied explicitly during MCR-ALS resolution, noise propagation effects and rotation ambiguity

effects appeared intermixed. Increasing rotation ambiguity band happens with increasing noise levels. In order to further check for noise propagation and rotation ambiguity effects and to discern among them at different noise levels, the accuracy and recovery of the profiles can be evaluated by means of the calculation of their dissimilarity with the true ones (known in this simulated data case). Dissimilarity is measured by the sine of the angle between two profiles evaluated as:

$$\text{dissimilarity} = \sin\alpha = \sqrt{1 - r^2} \qquad \text{Equation 17}$$

where r^2 is the squared correlation coefficient between these two vectors. When the two profiles are different (correlation coefficient close to zero), the dissimilarity is high. This magnitude evaluates the correct recovery of profiles, and therefore, evaluates the effect of rotational ambiguities. Figure 8 shows the dissimilarities calculated for the three resolved spectra profiles using Equation 16 at the different investigated noise levels, except at 5% error, since they would obscure completely the rest of the plot. At 0% and 0.1%, the accuracy in the profiles recovery is very good, with dissimilarity values equal or below 0.001 in average, which represent correlation coefficients close to the unity, i.e. no significant rotational ambiguity was detected. At 1% noise level, spectra dissimilarities start being somewhat different from zero, around 0.01 units (correlation coefficients still at the order of 0.9999). Rotation ambiguities are still very low if present and they do not produce any significant distortion of profiles. In the case of 2% noise level, dissimilarities are around 0.02 units (correlation coefficients around 0.9998). Results are still very good and acceptable although rotation ambiguity effects start showing up. It is at the 5% level, where rotation ambiguities appeared more importantly and dissimilarity values showed a clear disagreement between estimated ALS values and theoretical ones used for the data simulation.

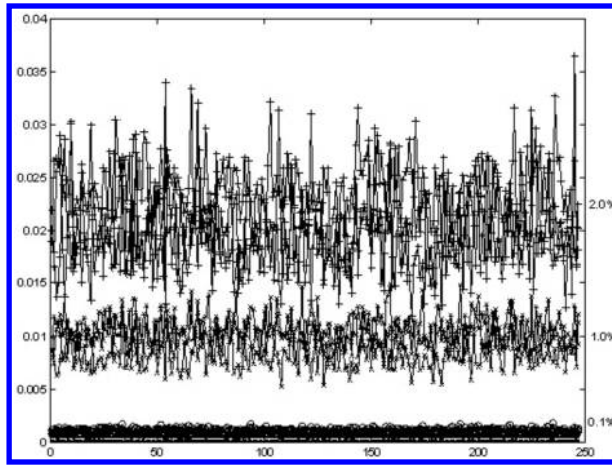

Figure 8. Effect of noise propagation and rotational ambiguities on MCR estimations at 0.1%, 1% and 2% noise levels. y left axis gives dissimilarities (Equation 17). Y right axis gives the noise level of the simulations. x axis gives simulation number (reproduced from Jaumot, J.; Gargallo, R.; Tauler, R. J. Chemom. ***2004****, 18, 327–340).*

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

Taking into Account Error Propagation Uncertainties in MCR Estimations

MCR methods assume that the dataset error structure is independent and identically distributed, i.i.d, noise. Although this assumption is not general, since heteroscedastic and correlated error structures exist and they can influence significantly the results obtained (*70*), MCR-ALS results are usually considered good approximations of the true results if the noise structure is unknown and error sources are not very high and disparate. Recently, Multivariate Curve Resolution Weighted Alternating Least Squares (MCR-WALS), and the use of Maximum Likelihood Principal Component Analysis MLPCA as initial projection step before ALS, have been proposed for the analysis of datasets where noise structure is known and relatively high non-homoscedastic error sources are present in the investigated datasets (*71–75*). In all these cases, MCR-WALS or MLPCA+MCR-ALS results were better than those obtained when applying ordinary MCR-ALS (ALS without considering the data error structure). The main aim of MCR-WALS is to work on the space of solutions having a minimum error contribution. MLPCA also attempts to solve the same problem, although the use of these algorithms requires a previous good knowledge of data error structure, which unfortunately is seldom possible. The use of MLPCA preliminary subspace projection of the data matrix has the advantage versus MCR-WALS of an easier application and generalization of previously developed Multivariate Curve Resolution methods. See previous references (*72–75*) for more details.

5. New Application Domains

With the new advancements in MCR-ALS method, its application arena has increased enormously covering datasets of different structures and complexities as diverse as , -omics data (*76–78*), environmental data (*79–82*), or hyperspectral imaging data, to mention a few. Next, some of these examples are briefly described.

Metabolomics Studies (*76*, *77*)

In the first work (*76*), liquid chromatography mass spectrometry (LC–MS) profiling experiments were performed to investigate metabolic changes in *S. cerevisiae* yeast culture samples at different temperatures (30° and 42 ºC), to differentiate between them and to find possible biomarkers of temperature stress. MCR-ALS was applied to full scan LC–MS preprocessed data multisets, arranged in selected time window augmented column-wise data matrices which include control (30°C) and temperature stressed (42°C) yeast samples. The working workflow is displayed in Figure 9. MCR-ALS resolved high-resolution accurate MS spectra were used to identify possible metabolite structures by comparison with MS spectra entries in metabolite databases. At the same time, statistically significant changes in MCR-ALS chromatographic peak areas of metabolites in control and stressed yeast samples, were used to detect possible markers of metabolism changes caused by temperature. PLS-DA-VIP scores analysis (*86*) was used for this purpose. The proposed strategy allowed simplifying

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

considerably biochemical interpretation of LC–MS metabolomics detected changes and allowed the uncovering of new targets for discovery (biomarkers).

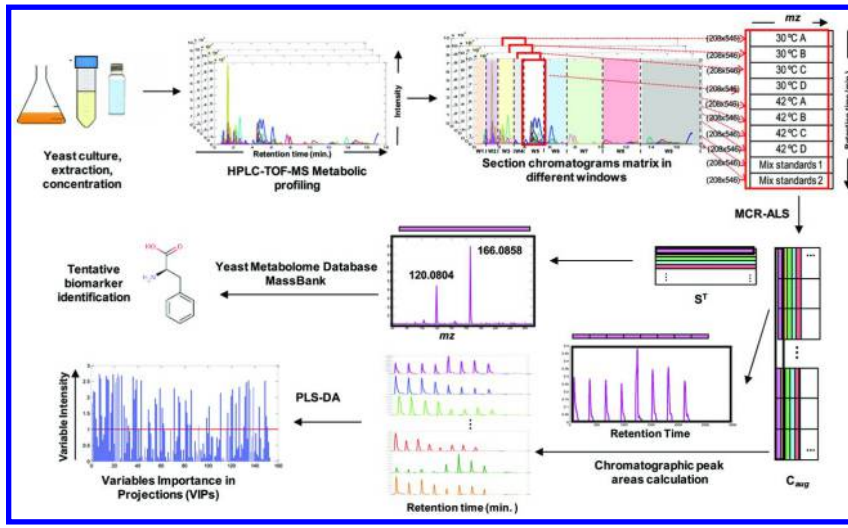

Figure 9. Schematic representation of the workflow following untargeted (LC-MS) data generation. The workflow involved experimental analysis, data pre-processing and data analysis in order to identify possible biomarkers (yeast metabolites), reproduced from Farres M.; Piña B.; Tauler R. Metabolomics ***2015,*** *11, 210–224.*

In another recent work (*77*), the application of MCR-ALS to untargeted UHPLC– TOF-MS lipid profiles analysis of a human placental choriocarcinoma cell line (JEG-3) exposed to different xenobiotics is described. MCR-ALS was applied to the column-wise augmented data matrices of 20 distinct chromatographic windows of UHPLC–TOF-MS data of the different cell samples, including treated and control samples. Application of MCR-ALS on this UHPLC–TOF-MS lipidomic augmented data matrices allowed the resolution of a large number of coeluted chromatographic peaks, the calculation of their respective peak areas and the resolution of their corresponding pure mass spectra. A total number of 86 MCR-ALS components were successfully resolved, and the peak areas of the elution profiles of untreated (control) and treated cell line samples were statistically compared.

Peak areas of the resolved elution (concentration) profiles showed distinct responses for the lipids of exposed versus control cells, evidencing a lipidome disruption attributed to the presence of the investigated xenobiotics. Results from one-way ANOVA followed by a multiple comparisons test and from partial least squares discriminant analysis (PLS-DA) were compared as usual strategies for the determination of potential biomarkers. The combination of the results of both strategies gave rise to a total of 33 lipids showing significant differences between control and treated sampless. Identification of 24 out of the 33 potential biomarkers was positively achieved, using the resolved pure MS

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

spectra from MCR-ALS analysis together with the high mass accuracy of the TOF analyzer. The untargeted methodology proposed in this study noticeably simplifies the interpretation of the lipidome, exclusively focusing the attention on lipids showing important differences among normal (control samples) and stressing (xenobiotic treated samples) conditions. Overall, this study proposes an innovative untargeted LC–MS MCR-ALS approach (Figure 10) valid for most of the MS -omic sciences, and being extended at present with new research work.

Environmental Studies (*81*, *82*)

A recent example showing the capabilities of MCR-ALS method with non-negativity and a new recently implemented quadrilinear constraint is shown in reference (*82*). In this work a very large multidimensional dataset is summarized and resolved with four way/mode component profiles (see Figure 11). The four-way/mode data set was obtained in a long term environmental monitoring study (15 sampling sites × 9 variables × 12 months × 7 years) belonging to the polluted Yamuna river of India. MCR-ALS resolved pollution profiles described appropriately the major observed changes on pH, organic pollution, bacteriological pollution and temperature, along with their spatial and temporal distribution patterns for the studied stretch of Yamuna River. Results obtained by MCR-ALS were also compared with those obtained by other multi-way methods, like PARAFAC. The implemented method and strategy are completely general and can be used for the analysis of other multi-way data sets obtained in extensive environmental monitoring studies of different type and compartments (air, water, solid, etc.), over large geographical areas and during different time periods (daily, weekly, monthly, yearly), as well as and in other similar high dimensional (multiway, multimode) mixture analysis problems. See reference (*82*) for more details about this work.

The implementation of the multilinearity constraints and interactions for multiway data sets in the MCR-ALS method was proposed for the investigation of the temporal distribution of the pollution by nitric oxide (NO) and ozone (O_3) in a city sampling station (urban center of Barcelona. Catalonia, Spain), during the years 2000–2006 (*81*). Different specific studies were performed considering the annual and pluriannual contamination by these two contaminants, individually or in combination using different data matrix augmentation strategies and multiway and multiset data analysis models (Figure 12). The MCR-ALS method with appropriate constraints could successfully extract the different patterns of daily, hourly and annual profiles summarizing the main contamination processes. Interpretation of these patterns describe in detail the ozone-nitric oxide atmospheric contamination time evolution and situation in the specific city site under investigation. MCR-ALS with different constraints like trilinearity and component interaction produced results analogous to well-established methods like PARAFAC and restricted TUCKER3 model-based methods in the analysis of environmental multiway data sets. The extension of MCR-ALS method to multiset data analysis using different constraints like non-negativity, trilinearity and interaction among components is shown to provide a powerful method to

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

improve the interpretability of the different contamination patterns in atmospheric and other environmental contamination problems.

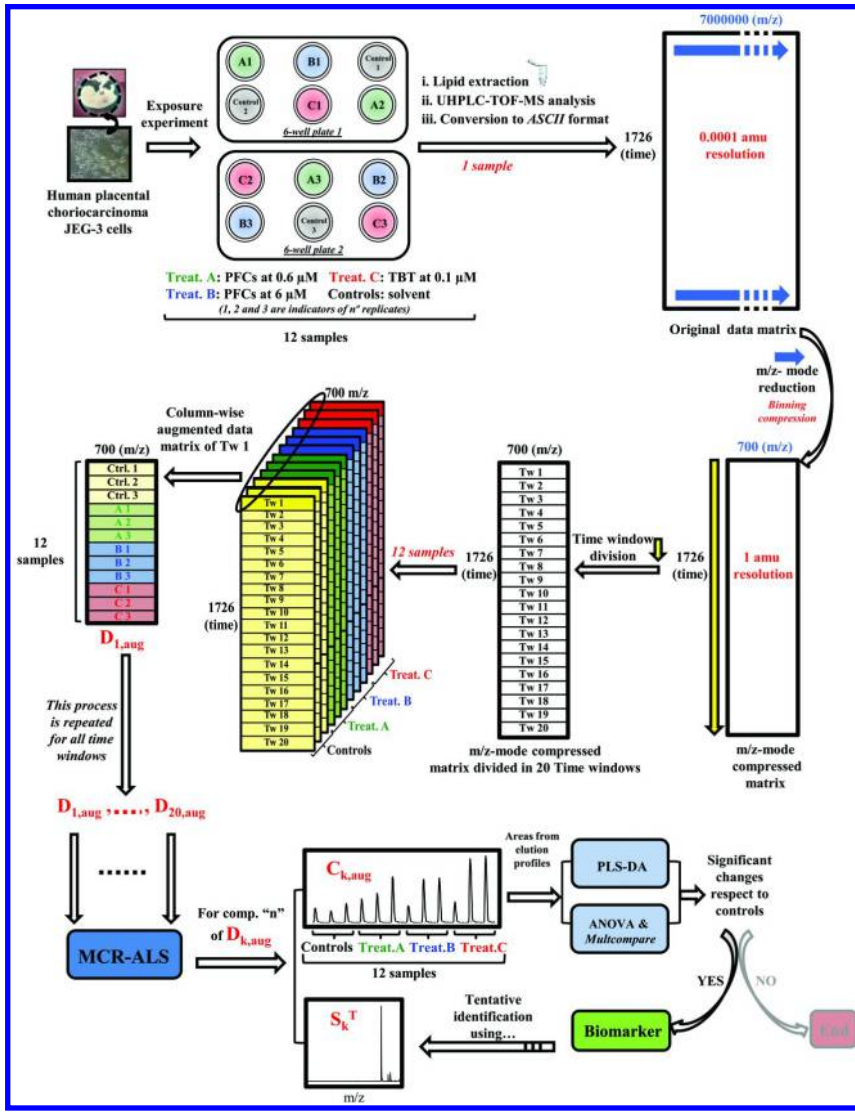

Figure 10. Scheme of the steps of the untargeted LC-MS MCR-ALS strategy proposed for the analysis of lipid profiles and determination of potential biomarkers of lipid disruption by different xenobiotic compounds. Reproduced from Gorrochategui E., Casas J.; Porte C.; Lacorte S.; Tauler R. Anal. Chim Acta ***2015****, 854 20–33.*

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

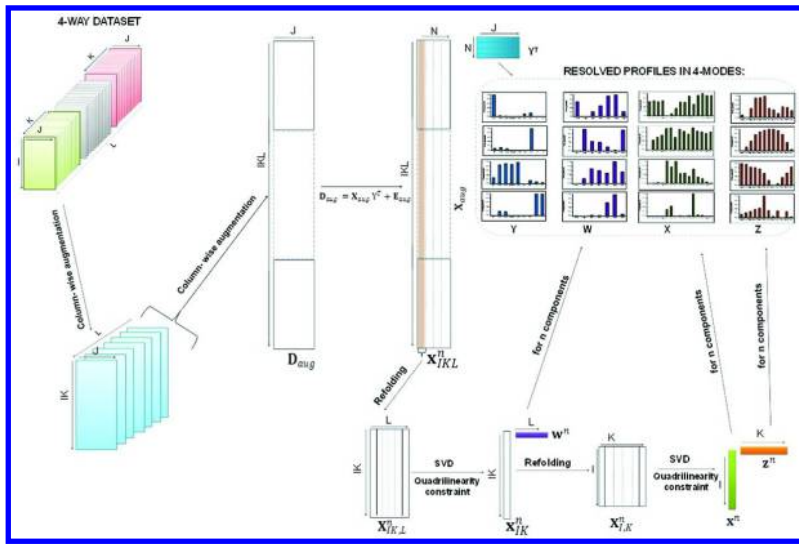

Figure 11. Implementation of the quadrilinear model constraint in MCR-ALS; and profiles resolved by 4 component MCR-ALS with quadrilinearity and non-negativity constraints: (Y) variable mode; (W) year mode; (X) sites mode; and (Z) month mode. Reproduced from Malik A.; Tauler R. Anal. Chim. Acta, ***2013****, 794, 20–28.*

Hyperspectral Imaging (*83–85*)

Hyperspectral image data sets are an extension of traditional spectral data adding the spatial dimension. Although traditionally hyperspectral images of a sample are visualized as three-dimensional data cubes (2D pixels by spectral wavelength), the cube structure does not reflect the mathematical behavior of these data sets. In fact, two of the data dimensions are only markers of the spatial position of the pixels under consideration and to be treated adequately, the image cube should be unfolded to provide a data table containing the spectra of all pixels each as a separate row. Therefore, it should be considered a two-dimensional (two- way or two-mode) data set. In this context, it has no sense to consider the image data set as a cube following a trilinear model and the ordinary MCR bilinear model should be used instead. As recent examples of application of the MCR-ALS method to hyperspectral images, two examples will be briefly shown taken from recently published works (*83*, *84*).

In the first work (*83*), a new data processing strategy is proposed to increase the natural spatial detail present in the acquired raw hyperspectral images provided by the image acquisition systems. The strategy proposed consists of a proper design in the acquisition of series of hyperspectral images with a small motion step among them, as small as the pixel size desired, combined with the appropriate MCR-ALS image multiset analysis and a super-resolution post-processing

strategy. The data treatment includes the application of multivariate curve resolution (unmixing) multiset analysis to a set of multiple collected images to obtain distribution maps and spectral signatures of the sample constituents. These sets of maps are noise-filtered and compound-specific representations of all the relevant information in the pixel space and decrease the dimensionality of the original image from hundreds of spectral channels to few sets of maps, one per sample constituent or element. The information in each compound-specific set of maps is combined via a super-resolution post-processing algorithm, which takes into account the shifting, decimation, and point spread function of the instrument to reconstruct a single map per sample constituent with much higher spatial detail than that of the original image measurement. This strategy overcomes the problem of computation time that could arise if sets of raw hyperspectral images at hundreds or thousands of spectral channels had to be processed individually and provides better results, expressed as detailed noise-filtered maps with constituent-specific information. This approach was tested on IR images collected on a HeLa cell (Figure 13).

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

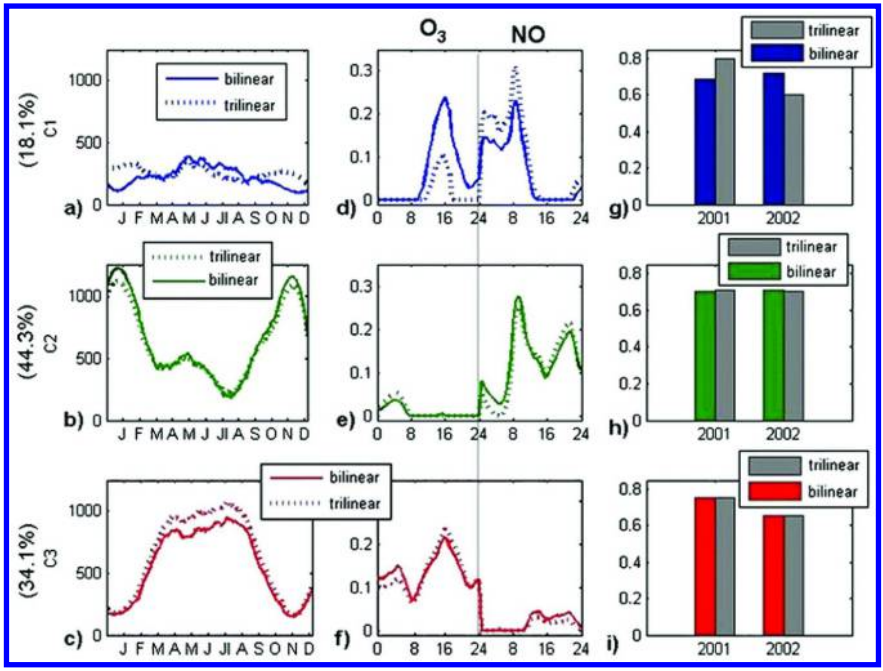

*Figure 12. MCR-ALS profiles for three components using non-negativity (continuous line) and non-negativity and trilinearity constraints (dotted line) for column-row-wise augmented data matrices of NO and O_3: **a**–**c** daily (within a year); **d**–**f** hourly (within a day) and **g**–**h** yearly profiles (between years). Reproduced from Alier M.; Felipe M.; Hernández I.: Tauler R. Anal Bioanal Chem **2011** 399, 2015-29.*

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

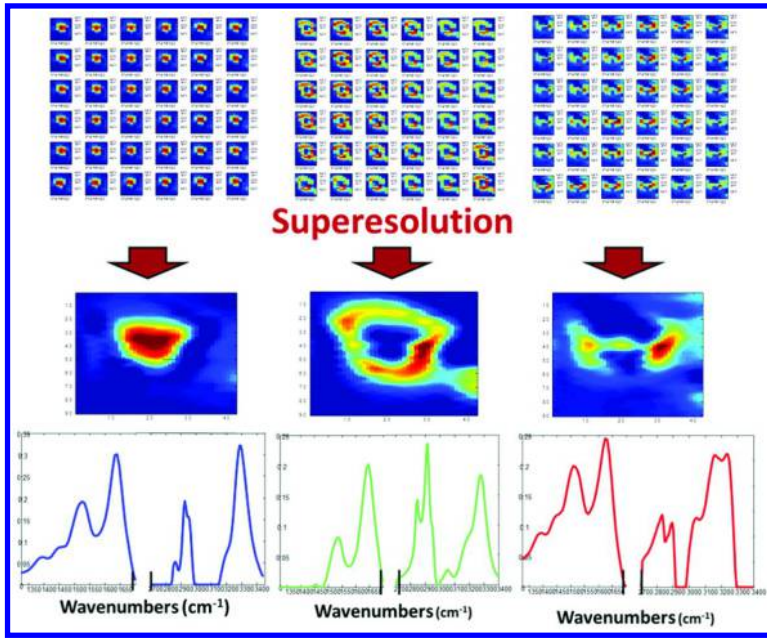

Figure 13. MCR-ALS results of a multiset analysis of low-spatial-resolution HeLa cell images combined with a postprocessing super-resolution step. Top plots: sets of 36 low resolution distribution maps of the nucleus, membrane, and cytoplasm. Central plots: superesolved maps obtained by postprocessing combining sets of low-resolution distribution maps. Bottom plots: resolved spectral signatures for each cell component. Adapted from Piqueras S.; Duponchel L.; Offroy M.; Jamme F.; Tauler R.; de Juan A. Anal Chem ***2013****, 85, 6303-6311.*

After acquiring the raw spectra of 36 low-resolution images of a single HeLa cell (with a pixel size of 3.6 μm and shifted 0.6 μm in the x- and/or y- direction from one another), the Asymmetric Least Squares (AsLS (*87*)) was applied to the raw spectra to remove the Mie scattering effect. After Mie scattering correction by AsLS, the spectral ranges between 1300 and 1700 cm–^{1} and 2800–3400 cm–^{1} were selected for resolution analysis. SVD analysis indicated the presence of three contributions on the Mie corrected data for the multiset of low-spatial-resolution images of the HeLa cell. MCR-ALS analysis using SIMPLISMA (10) to obtain initial estimates of pure spectra was performed under the constraints of non-negativity in concentration profiles and spectra and with spectra normalization in matrix $\mathbf{S}^{\mathbf{T}}$. The final results of distribution maps and pure spectra from the multiset analysis of low spatial resolution images of HeLa cell clearly represent the different cell regions: nucleus, cellular membrane, and cytoplasm. Spectral signatures were obtained directly from MCR-ALS analysis. The super-resolution postprocessing applied to each of the three sets of 36 low-resolution distribution maps (with pixel size of 3.5 μm) provided three superesolved maps with a pixel size equal to the motion step among images, 0.6 μm.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

Another interesting type of hyperspectral imaging datasets are obtained by remote sensing satellites or airplanes. A last example of application is shown (*84*) where the spectra signatures of the constituents present in the remote sensing spectroscopic images, obtained by the Airborn Visible/Infrared Imaging Spectrometer (AVIRIS), and their concentration distribution at a pixel level were estimated by MCR-ALS. Results obtained by MCR-ALS were in general similar to those obtained by other methods used in this field like Minimum Volume Simplex Analysis (MVSA) and Vertex Component Analysis (VCA) methods (*88*), except for cases where the latter method produce spectra and concentration profiles with negative values, which were not feasible from a physical point of view and according to the desired constraints of the sought solutions. MCR-ALS results were evaluated for the presence of rotational ambiguities using the MCR-BANDS method. The obtained results confirmed that the MCR-ALS method can be successfully used for remote sensing hyperspectral image resolution purposes (Figure 14). However, the amount of rotation ambiguity still present in the solutions obtained by this and other resolution methods (like VCA or MVSA) still are large and it should be evaluated with care, trying to reduce its effects by selecting the more appropriate constraints. MCR-BANDS (*53*, *64*) results suggest that the extent of rotation ambiguity associated with the MCR-ALS resolved profiles can be rather high and that the correct solutions can only be guaranteed if additional constraints are applied, such as those providing information about the local rank properties of the image, i.e., about the presence or absence of the different components in the image pixels.

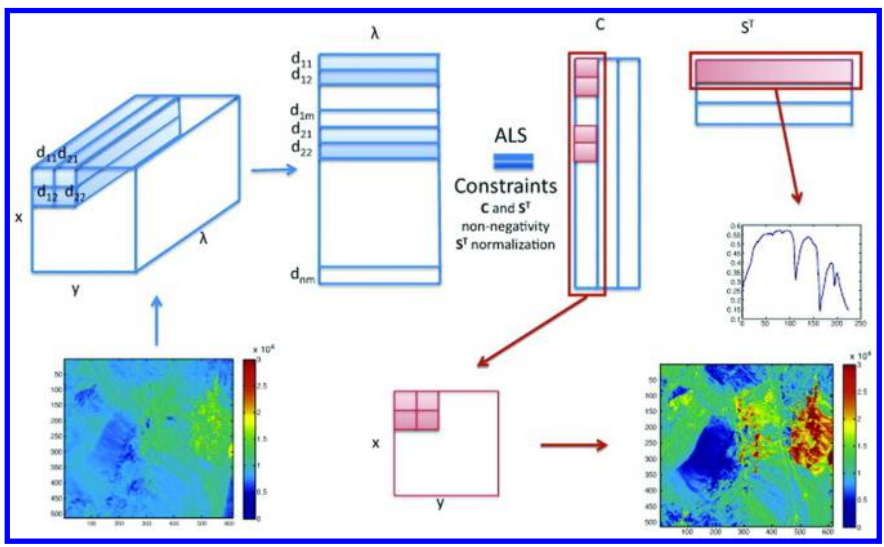

Figure 14. MCR-ALS strategy for the resolution of spectra (signatures) and of 2D image concentrations of the pure components. In the remote sensing image obtained by Airborne Visible Infrared Imaging Spectrometer (AVIRIS). Reproduced from Zhang X.; Tauler R. Anal Chim. Acta, ***2013****, 762, 25–38.*

6. Bruce Kowalski and MCR (Figure 15)

Bruce Kowalski was always interested in curve resolution methods; in fact the term 'multivariate curve resolution' was already used in his publications of 1984 with Osten D.W. (*88*) and in 1985 with Borgen O.S. (*2*). Until then, the more used term had been 'self-modelling curve resolution' initially proposed by Lawton and Sylvestre (*1*, *13*). There were of course other terms used for analogous purposes such as mixture analysis (*4*, *10*) or factor analysis (*8*, *12*, *89*). Bruce was recommending the use of the term 'multivariate curve resolution' in analogy to other family of methods in chemometrics such as 'multivariate calibration' methods. Bruce was always well aware of the ubiquitous rotation ambiguity problems associated to all bilinear data decompositions, such as those performed in curve resolution of two-way data sets (data tables, data matrices). Bruce was pushing to higher-order data (data cubes, multiway data), where all these ambiguities could be eliminated, for instance using generalized rank annihilation (*5*) or multilinear model based methods. In 1992, during a research visit to the Center for Process Analytical Chemistry of the University of Washington, in Seattle, I was fortunate to work with Bruce, his PhD students, and other visiting scientists of his Chemometrics Laboratory. In the frequent research group meetings we had, the main topic of discussion was the extension of multivariate calibration methods to multiway data (to higher order data, in Bruce's preferred notation (*90*)). As a result of these new developments, and as a consequence of my previous research activity in the study of multiequilibria and multispeciation systems using spectroscopic methods using (evolving) factor analysis soft modelling methods (*91*), we extended curve resolution methods to the simultaneous analysis of multiple spectroscopic titrations of chemical reaction systems (*92*), to the analysis of multiple runs of an industrial chemical process monitored by spectroscopic probes (*56*), and to the development of spectroscopic chemical sensors (*93*). All this preliminary work finished with the publication of the seminal paper in 1995 about local rank, selectivity and multiway analysis, using multivariate curve resolution methods (*19*), written in cooperation with Age Smilde and Bruce Kowalski. This paper received more than 500 citations, and it is still receiving a lot of attention at present. We established the basis of modern curve resolution methods, describing how rotation ambiguities can be solved, either using selectivity/local rank constraints, (which was simultaneously proposed by Rolf Manne (*16*) in his resolution theorems), or with the extension of MCR methods to multiset and multiway data, and implementation of trilinearity and other multilinearity type of constraints (*20*, *21*, *27*, *29*, *30*, *35–38*, *57*, *75*).

As we mentioned at the beginning of this chapter, and after a rather unsteady development, MCR is reaching its mature state. It is clear from our point of view, that a fundamental milestone in this development was due to the interaction of the work performed by Bruce and co-workers, and of our research effort to find new ways of analyzing chemical data, in particular by means of spectroscopic methods. These synergies catapulted the development of MCR methods to the investigation and application of new problems and challenges in Analytical Chemistry and related fields. We have entitled this book chapter 'Multivariate Curve Resolution: a different way to examine chemical data', because this is

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

indeed the philosophy behind the development and application of MCR methods, and this is the reason for its widespread use at present (*94*), as we have tried to summarize in this chapter.

We want to finish this chapter with our sincere and vivid acknowledgment to the work done by Bruce Kowalski in the promotion of the Chemometrics field worldwide, and we are specially honored of having contributed under his initial guidance, to the consolidation of one of its main subfields at present: Multivariate Curve Resolution.

Figure 15. Bruce Kowalski and Romà Tauler (one of the authors of this Chapter) in one of the breaks of the Xth Chemometrics in Analytical Chemistry meeting, celebrated in Campinas, Brazil, in 2006 (Photograph taken by Susana Navea, PhD student).

References

1. Lawton, W. H.; Sylvestre, E. A. *Technometrics* **1971**, *13*, 617–633.
2. Borgen, O. S; Kowalski, B. R *Anal. Chim. Acta* **1985**, *174*, 1–26.
3. Rajko, R.; Istvan, K. *J. Chemom.* **2005**, *19*, 448–463.
4. Hamilton, J. C.; Gemperline, P. J. *J. Chemom.* **1990**, *4*, 1–13.
5. Sanchez, E.; Kowalski, B. R. *Anal. Chem.* **1986**, *58*, 496–499.
6. Geladi, P.; Wold, S. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 273.
7. Maeder, M.; Zilian, A. *Chemomet. Intell. Lab. Syst.* **1988**, *3*, 205–213.
8. Malinowski, E. R. *J. Chemom.* **1992**, *6*, 29–40.
9. Kvalheim, O. M.; Liang, Y. Z. *Anal. Chem.* **1992**, *64*, 936–946.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

10. Windig, W.; Guilment, J. *Anal. Chem.* **1991**, *63*, 1425.
11. Cuesta Sánchez, F.; Rutan, S. C.; Gil García, M. D.; Massart, D. L. *Chemomet. Intell. Lab. Syst.* **1997**, *36*, 153–164.
12. Malinowski, E. R. *Anal. Chim. Acta* **1982**, *134*, 129–137.
13. Vandeginste, B. G. M.; Derks, W.; Kateman, G. *Anal. Chim. Acta* **1985**, *173*, 253–264.
14. Gemperline, P. J. A *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 206–212.
15. Maeder, M.; Zuberbühler, A. D. *Anal. Chim. Acta* **1986**, *181*, 287–291.
16. Manne, R. *Chemom. Intell. Lab. Syst.* **1995**, *27*, 89–94.
17. Tauler, R.; Smilde, A. K.; Kowalski, B. R. *J. Chemom.* **1995**, *9*, 31–58.
18. Faber, N. M.; Buydens, L. M. C.; Kateman, G. *J. Chemom.* **1994**, *8*, 147–154.
19. Bro, R. *Chemom. Intell. Lab. Syst.* **1997**, *38*, 149–171.
20. Tauler, R.; Marques, I.; Casassas, E. *J. Chemom.* **1998**, *12*, 55–75.
21. Peré-Trepat, E.; Ginebreda, A.; Tauler, R. *Chemom. Intell. Lab. Syst.* **2007**, *88*, 69–83.
22. Malik, A.; Tauler, R. *Anal. Chim. Acta* **2013**, *794*, 20–28.
23. Izquierdo-Ridorsa, A.; Saurina, J.; Hernández-Cassou, S.; Tauler, R. *Chemomet. Lab. Syst.* **1997**, *38*, 183–196.
24. Saurina, J.; Hernández-Cassou, S.; Tauler, R.; Izquierdo-Ridorsa, A. *J. Chemom.* **1998**, *12*, 183–203.
25. Tauler, R.; de Juan, A. Multivariate Curve Resolution. In *Practical Guide to Chemometrics*, 2nd ed.; Gemperline, P., Ed.; CRC Taylor & Francis: New York, 2006; Chapter 11.
26. de Juan, A.; Maeder, M.; Martínez, M.; Tauler, R. *Chemom. Intell. Lab. Syst.* **2000**, *54*, 123–141.
27. Tauler, R. *Chemom. Intell. Lab. Syst.* **1995**, *30*, 133–146.
28. Jolliffe I. T. *Principal Component Analysis*, 2nd ed.; Springer: Berlin, 2002.
29. Jaumot, J.; Gargallo, R.; de Juan, A.; Tauler, R. *Chemom. Intell. Lab. Syst.* **2005**, *76*, 101–110.
30. Jaumot, J.; de Juan, A.; Tauler, R. *Chemomet. Intell. Lab. Syst.* **2015**, *140*, 1–12.
31. http://mcrals.wordpress.com/download/mcr-als-toolbox/ (accessed May 7, 2015).
32. de Juan, A.; Rutan, S. C.; Tauler, R. Two-way data analysis: Multivariate Curve Resolution: Iterative resolution methods. In *Comprehensive Chemometrics*; Elsevier: Amsterdam, 2009; Vol. 2, Chapter 2.19.
33. Tauler, R.; de Juan, A. Multivariate Curve Resolution. In *Practical Guide to Chemometrics*; Gemperline, P., Ed.; CRC Press: Boca Raton, FL, 2006; pp 417–474
34. Tauler, R. *J. Chemom.* **2001**, *15*, 627–646.
35. de Juan, A.; Tauler, R. *Crit. Rev. Anal. Chem.* **2006**, *36*, 163–176.
36. Tauler, R.; Maeder, M.; de Juan, A. Two-Way Extended Curve Resolution. In *Comprehensive Chemometrics*; Brown, S., Tauler, R., Walczak, R., Eds.; Elsevier: Oxford, 2009; Vol. 2, pp 473−505.
37. Malik, A.; Tauler, R. *Chemom. Intell. Lab. Syst.* **2014**, *135*, 223–234.
38. Alier, M.; Felipe, M.; Hernández, I.; Tauler, R. *Anal. Bioanal. Chem.* **2011**, *399*, 2015–29.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

39. Kroonenberg, P. *Three Mode Principal Component Analysis*; DSWO Press: Leiden, 1983.
40. de Juan, A.; Maeder, M.; Martínez, M.; Tauler, R. *Anal. Chim. Acta* **2001**, *442*, 337–350.
41. Antunes, M. C.; Simão, J. E. J.; Duarte, A. C.; Tauler, R. *Analyst* **2002**, *127*, 809–817.
42. de Oliveira, R. R.; de Lima, K. M. G.; Tauler, R.; de Juan, A. *Talanta* **2014**, *125*, 233–241.
43. Ahmadi, G.; Tauler, R.; Abdollahi, H. *Chemomet. Intell. Lab. Syst.* **2015**, *142*, 143–150.
44. Haskell, K. H.; Hanson, R. J. *Math. Program.* **1981**, *21*, 98–118.
45. Hanson, R. J.; Haskell, K. H. *ACM Trans. Math. Softw.* **1982**, *8*, 323–333.
46. Bro, R.; de Jong, S. *J. Chemom.* **1997**, *11*, 393–401.
47. Bro, R.; Sidiropoulos, N. D. *J. Chemom.* **1998**, *12*, 223–247.
48. Gemperline, P. J.; Cash, E. *Anal. Chem.* **2003**, *75*, 4236–4243.
49. Van Benthem, M. H.; Keenan, M. R.; Haaland, D. H. *J. Chemom.* **2002**, *16*, 613–622.
50. de Juan, A.; Vander Heyden, Y.; Tauler, R.; Massart, D. L. *Anal. Chim. Acta* **1997**, *346*, 307–318.
51. de Juan, A.; Tauler, R. *Anal. Chim. Acta* **2003**, *500*, 195–210.
52. Smilde, A.; Bro, R.; Geladi, P. *Multi-way Analysis. Applications in the Chemical Sciences*; John Wiley & Sons, Ltd.: West Sussex, U.K., 2004.
53. Tauler, R. *J. Chemom.* **2001**, *15*, 627–646.
54. Burdick, D. S. *Chemom. Intell. Lab. Syst.* **1995**, *28*, 229–237.
55. Spjotvoll, E.; Martens, H.; Volden, R. *Technometrics* **1982**, *24*, 173–180.
56. Tauler, R.; Kowalski, B.; Fleming, S. *Anal. Chem.* **1993**, *65*, 2040–2047.
57. Tauler, R.; Barcelo, D. *TrAC-Trends Anal. Chem.* **1993**, *12*, 319–327.
58. Peré-Trepat, E.; Lacorte, S.; Tauler, R. *Anal. Chim. Acta* **2007**, *595*, 228–237.
59. Wentzell, P. D.; Wang, J.; Loucks, L. F.; Miller, K. M. *Can. J. Chem.* **1998**, *76*, 1144–1155.
60. Gemperline, P. J. *Anal. Chem.* **1999**, *71*, 5398–5404.
61. Rajko, R. *Anal. Chim. Acta* **2009**, *645*, 18–24.
62. Golshnan, A.; Abdollahi, H.; Maeder, M. *Anal. Chem.* **2011**, *83*, 836–841.
63. Sawall, M.; Kubis, C.; Neymeyer, K. *J. Chemom.* **2013**, *27*, 106–116.
64. Jaumot, J.; Tauler, R. *Chemomet. Intell. Lab. Syst.* **2010**, *103*, 96–107.
65. *Optimization Toolbox*, Version 2.0; The Mathworks: Natick, MA, U.S.A., 1998.
66. Faber, N. M. *J. Chemom.* **2001**, *15*, 169–188.
67. Bijlsma, S.; Smilde, A. K. *J. Chemom.* **2000**, *14*, 541–560.
68. Jaumot, J.; Gargallo, R.; Tauler, R. *J. Chemom.* **2004**, *18*, 327–340.
69. Jaumot, J.; Menezes, J. C.; Tauler, R. *J.Chemom.* **2006**, *20*, 54–67.
70. Kiers, H. *Psychometrika* **1997**, *62*, 251–266.
71. Wentzell1, P. D.; Karakach, T. K.; Roy, S.; Martinez, J.; Allen, C. P.; Werner-Washburne, M. *BMC Bioinformatics* **2006**, *7*, 343.
72. Tauler, R.; Viana, M.; Querol, X.; Alastuey, A.; Flight, R. M.; Wentzell, P. D.; Hopke, P. K. *Atmos. Environ.* **2009**, *43*, 3989–3997.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

73. Stanimirova, I.; Tauler, R.; Walczak, B. *Env. Sci. Technol.* **2011**, *45*, 10102–10110.
74. Dadashi, M.; Abdollahi, H.; Tauler, R. *Chemom. Intell. Lab. Syst.* **2012**, *118*, 33–40.
75. Malik, A.; Tauler, R. *Chemom. Intell. Lab. Syst.* **2014**, *135*, 223–234.
76. Farres, M.; Pina, B.; Tauler, R. *Metabolomics* **2015**, *11*, 210–224.
77. Gorrochategui, E.; Casas, J.; Porte, C.; Lacorte, S.; Tauler, R. *Anal. Chim. Acta* **2015**, *854*, 20–33.
78. Lima, K. M. G.; Bedia, C.; Tauler, R. *Microchem. J.* **2014**, *117*, 255–261.
79. Terrado, M.; Kuster, M.; Raldúa, D.; Lopez de Alda, M.; Barceló, D.; Tauler, R. *Anal. Bioanal. Chem.* **2007**, *387*, 1479–1488.
80. Terrado, M.; Barceló, D.; Tauler, R. *Environ. Sci. Technol.* **2009**, *43*, 5321–5326.
81. Alier, M.; Felipe, M.; Hernández, I.; Tauler, R. *Anal. Bioanal. Chem.* **2011**, *399*, 2015–29.
82. Malik, A.; Tauler, R. *Anal. Chim. Acta* **2014**, *794*, 20–28.
83. Piqueras, S.; Duponchel, L.; Offroy, M.; Jamme, F.; Tauler, R.; de Juan, A. *Anal Chem* **2013**, *85*, 6303–6311.
84. Zhang, X.; Tauler, R. *Anal. Chim. Acta* **2013**, *762*, 25–38.
85. Felten, J.; Hall, H.; Jaumot, J.; Tauler, R.; de Juan, A.; Gorzsás, A. *Nat. Protocols* **2015**, *10*, 217–240.
86. Andersen, C. M.; Bro, R. *J. Chemom.* **2010**, *24*, 728–737.
87. Eilers, P. H. C *Anal. Chem.* **2003**, *75*, 3631–3636.
88. Osten, D. W.; Kowalski, B. R. *Anal. Chem.* **1984**, *56*, 991–5.
89. Malinowski E. R. *Factor Analysis in Chemistry*, 3rd ed.; Wiley: New York, 2002.
90. Booksh, K. S.; Kowalski, B. R. *Anal. Chem.* **1994**, *66*, 782A–791A.
91. Tauler, R.; Casassas, E.; Izquierdo-Ridorsa, A. *Anal. Chim. Acta* **1991**, *248*, 447–458.
92. Tauler, R.; Izquierdo-Ridorsa, A.; Casassas, E. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 293–300.
93. Tauler, R.; Smilde, A. K.; Henshaw, J. M.; Burgess, L. W.; Kowalski, B. R. *Anal. Chem.* **1994**, *66*, 3337–3344.
94. Lavine, B. K.; Workman, J., Jr. *Anal. Chem.* **2013**, *85*, 705–714.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch005

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

Chapter 6

Applying Multivariate Curve Resolution to Source Apportionment of the Atmospheric Aerosol

Philip K. Hopke*

Institute for a Sustainable Environment and Department of Chemical and Biomolecular Engineering, Clarkson University, Potsdam, New York 13699
***E-mail: phopke@clarkson.edu**

A major application of chemometrics is the mixture resolution problem. Ideally in a chemical analysis, all of the constituents in a complex mixture are fully separated from one another so that their identification and quantification are relatively simple to achieve. However, in spite of advances in separation science, such resolution of complex mixtures is often not possible. Thus, it becomes necessary to separate overlapping components using mathematical methods. A similar problem exists in atmospheric science where it is useful to identify the sources giving rise to the observed concentrations of chemical constituents in the ambient aerosol and to quantitatively apportion the measured particulate mass concentrations to those identified sources. This process has come to be called Receptor Modeling and various methods have been developed and applied over the past 40 years to provide source apportionments. This chapter will outline these methods and their application to ambient particle composition data.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

Introduction

One of the major applications of chemometrics is the mixture resolution problem. Ideally in an analysis, all of the various constituents in a mixture are fully separated from one another so that their identification and quantification are relatively simple to achieve. However, in spite of significant advances in separation science, such resolution of complex mixtures is often not possible. Thus, it becomes necessary to separate overlapping components using mathematical methods. The approach used to perform these analysis is called Self-Modeling Curve Resolution (SMCR) (*1*, *2*) and has been in use since around 1960. This approach has been widely applied to a variety of spectrochemical data including vibrational spectroscopy (*3*) and a variety of complex physical chemistry problems (*4*).

A similar problem exists in atmospheric science where it is useful to identify the sources giving rise to the observed concentrations of chemical constituents in the ambient aerosol and to quantitatively apportion the measured particulate mass concentrations to those identified sources. This process has come to be called Receptor Modeling and various methods have been developed and applied over the past 40 years to provide source apportionments (*5–10*). This chapter will outline these methods and their application to ambient particle composition data.

Airborne particulate matter is a complex mixture of materials from a variety of sources including natural and anthropogenic. Some particles are emitted directly into the atmosphere (primary) while others are formed through atmospheric oxidative processes (secondary). Thus, sources need to be considered in the context of a 2x2 matrix of human and natural versus primary and secondary. Figure 1 provides a useful framework for considering particle origins (*11*). Most of the coarse mode particles are primary in origin whereas most of the fine mode particles are secondary. Particles larger than about 1 μm in aerodynamic diameter are produced through mechanical processes such as tires rolling over wet pavement or waves breaking on the shore and in the open ocean and throwing droplets of water into the air. When the water evaporates, the material dissolved in the water produces a particle. Fine particles (<1 μm in aerodynamic diameter) come from chemical processes including combustion and atmospheric oxidation such as the reaction of ozone with the terpenes emitted by coniferous trees to form particles.

To effectively manage air quality, it is essential to identify the sources that contribute pollutants to the observed concentrations and to apportion the contributions of those sources to the observed values. Then air quality management strategies can focus on those sources that are most important to the air quality problems and effective and efficient control plans can be devised.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

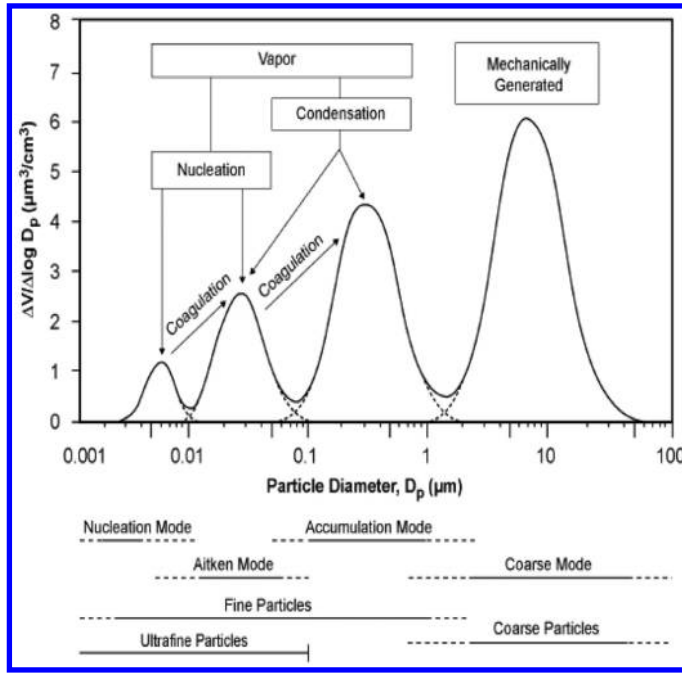

Figure 1. Volume size distribution measured in traffic showing fine and coarse particles and the nuclei and accumulation modes of the particles. (USEPA, 2004).

Mass Balance Principle

The fundamental principle of receptor modeling is that mass conservation can be assumed and a mass balance analysis can be used to identify and apportion sources of contaminants in the atmosphere. The approach to obtaining a data set for receptor modeling is to determine a large number of chemical constituents such as elemental concentrations in a number of samples. Alternatively, automated electron microscopy can be used to characterize the composition and shape of particles in a series of particle samples. In either case, a mass balance equation can be written to account for all m chemical species in the n samples as contributions from p independent sources.

$$x_{ij} = \sum_{k=1}^{p} g_{ij} f_{kj} \tag{1}$$

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

where x_{ij} is the j^{th} chemical species concentration measured in the i^{th} sample, f_{kj} is the gravimetric concentration of the j^{th} species in material from the k^{th} source, and g_{ik} is the airborne mass concentration of material from the k^{th} source contributing to the i^{th} sample.

This concept of a linear superposition of independent components is similar to spectral deconvolutions based on independent contributions to the sample absorbance in UV-Visible spectra of mixtures or the mass spectra of a gas chromatographic peak of unresolved components. However, in the receptor model problem, there are additional complications that make the resolution of the components more difficult. In spectrochemical problems, the uncertainty in the measurements are typically less than 1% whereas in the receptor modeling problem, the errors are 5% to 10% for well measured variables and can range to much higher values. Of more importance is the variability in the source profiles. The spectrum of a particular compound is the same no matter where and when it is measured. The only difference in the quality of that spectrum is the quality of the spectrometer. However, the composition of particles emitted by a coal-fired power plant depends on the nature of the mineral matter in that coal. Different coals from different locations will have different mineral species laid down with the carbonaceous material from which the coal has formed. Thus, as a plant burns through a supply of coal, there will be a variable input of the various mineral phases (*12*). In receptor modeling, it is thus necessary to account for both the measurement error and the variability of the profile compositions. This problem is exacerbated when the profile includes reactive species whose concentrations may change as a result of oxidative processes in the atmosphere. These differences are discussed in more detail by Hopke (*13*).

There exist a set of natural physical constraints on the system that must be considered in developing any model for identifying and apportioning the sources of airborne particle mass (*14*). The fundamental, natural physical constraints that must be obeyed are:

1) The original data must be reproduced by the model; the model must explain the observations.
2) The predicted source compositions must be non-negative; a source cannot have a negative percentage of an element.
3) The predicted source contributions to the aerosol must all be non-negative; a source cannot emit negative mass.
4) The sum of the predicted elemental mass contributions for each source must be less than or equal to total measured mass for each element; the whole is greater than or equal to the sum of its parts.

While developing and applying these models, it is necessary to keep these constraints in mind in order to be certain of obtaining physically realistic solutions.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

Conceptual Framework

To solve the mass balance problem outlined in equation 1, there are several approaches that can be taken. The simplest approach is if the profiles of the major sources are known, and thus, the values of source profile matrix, *F,* are available. Then equation 1 can be rewritten as

$$x_j = \sum_{k=1}^{p} g_j f_{kj} + e_j \qquad (2)$$

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

This equation now pertains to a single sample and since we are fitting a model to the data, we have to consider the residual values, e_j. The equation has now become an ordinary least squares (OLS) problem with the vector **x** and matrix **F** known and the vector of *g* values as the unknown coefficients to be estimated. This basic framework for solving the mass balance problem was initially proposed by Winchester and Nifong (*15*) and Miller et al. (*16*). Friedlander (*17*) introduced an ordinary least-squares regression analysis but based on very few species and called it a Chemical Element Balance (CEB).

Kowalczyk et al. (*18*) recognized that since there were not equal errors in all of the dependent variables, OLS was inappropriate and an ordinary weighted least squares (OWLS) fit was required to take the varying variances into account. However, In 1979, both John Watson and Alan Dunker independently recognized that the use of ordinary regression analysis was incorrect because source profiles are measured with error. Thus, OWLS does not take into account the errors in the independent variables. There are a number of ways to incorporate the errors in the independent variables into the analysis (*19*). One approach termed effective variance least squares (EVLS) incorporates the measurement error in the objective function to solve the chemical mass balance (CMB) problem. The approach was described by Cooper et al. (*20*) and was developed into software provided to the receptor modeling community by the U.S. Environmental Protection Agency (*21*). The key issue in the application of the CMB model is knowing the profiles. It is difficult and expensive to perform emissions sampling and very few sources other than motor vehicles have been examined in the past 15 years. Thus, many of the profiles of stationary sources may be out of date. Very little is known with respect to the variability in composition in the profiles for a given source type. There continue to be measurements of emissions from mobile sources, but even then there are relatively few measurements relative to the total number of motor vehicles in various weight classes and engine types that are on the road. An assessment of the utility of existing profiles relative to the ambient concentrations of compounds that can serve as tracers for spark- and compression ignition vehicles was presented by Subramanian et al. (*22*).

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Thus, alternative approaches that only utilize the ambient concentration data have been developed in terms of multiple forms of factor analysis that are actually trying to solve the self-modeling curve resolution or mixture resolution problem. To solve this problem, it is necessary to solve equation (1) using multiple sample data so that the model being fit is now given as:

$$x_{ij} = \sum_{k=1}^{p} g_{ij} f_{kj} + e_{ij} \tag{3}$$

where the residuals, e_{ij}, account for the part of the variation in the data than cannot be fit to the model.

The first receptor modeling analyses reported in the literature were factor analysis using eigenvector methods that had been developed in the social sciences for interpreting large data sets. Blifford and Meeker (*23*) used a principal component analysis with several types of axis rotations to examine particle composition data collected by the National Air Sampling Network (NASN) during 1957-61 in 30 U.S. cities. Prinz and Stratmann (*24*) examined both the aromatic hydrocarbon content of the air in 12 West German cities and data on the air quality of Detroit using factor analysis methods. In both cases, they found solutions that yielded readily interpretable results. There was no further use of factor analysis until it was reintroduced in the mid-1970's by Hopke et al. (*25*) and Gaarenstroom et al. (*26*) in their analyses of particle composition data from Boston, MA and Tucson, AZ, respectively. A problem that exists with these forms of factor analysis is that they do not permit quantitative source apportionment of particle mass or of specific elemental concentrations. In an effort to find alternative methods that would provide information on source contributions when only the ambient particulate analytical results are available, other approaches were employed. Hopke and coworkers used Target Transformation Factor Analysis (*27*) originally developed by Malinowski (*4*). Henry and coworkers (*28–31*) have developed alternative methods based on eigenvector methods. The initial model was SAFER and it has evolved into Unmix (*21*). These concepts provide the basis for a geometrical interpretation using "edges" as outlined by Henry (*32*).

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

Edges or End Members

The essence of SMCR or quantitative factor resolutions is the idea of edges or end members (*33*). In the spectrochemical problem, they are often referred to as pure spectra. They represent the relationships among the measured variables that are characteristic of the specific source type being resolved. To illustrate the idea, Figure 2 shows a plot of simulated data for a mixture of two crustal materials with different iron to silicon ratios. The critical idea here is not to look at a regression line showing the relationship between Fe and Si because there are two such relationships represented by the two solid lines. Those lines bound all of the measured values and then represent two new axes for the plot that represent the amounts of each of the two types of materials in each mixture represented by a point in the plot. The points show the relative amounts of Fe and Si relative to the orthogonal x and y-axes, but the non-orthogonal axes represent the amounts

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

of the source materials present when lines are dropped from a point to the two solid lines. There are points along those two solid lines that represent samples that contain only one of the two types of materials present in the samples. Thus, the defining lines in the plot are those where the concentration of one source material is zero.

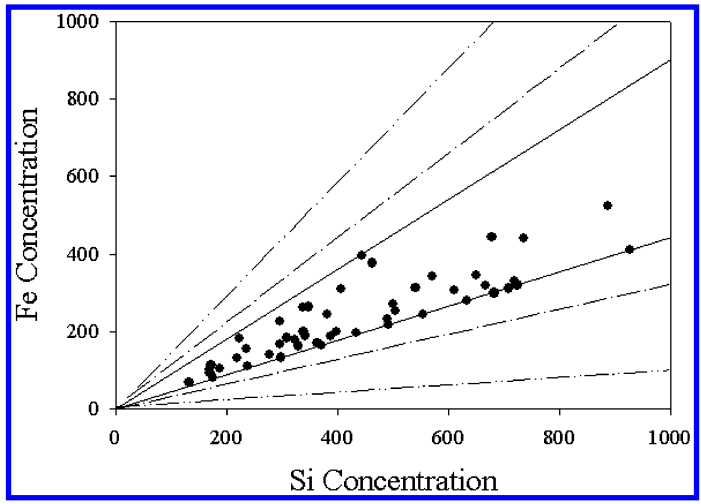

Figure 2. Plot of simulated data for mixtures of two crustal materials with different Fe/Si ratios.

These edge points or end member points then define the source profiles. For spectra, it is possible to be sure we have a pure compound that we put into the instrument to make a measurement and there are likely to be locations in the spectra where the absorbance of all of compounds in the mixture are zero. It is them possible to extract the pure spectra from the data (*1*, *34*). However, in the receptor modeling problem, there may be sources for which their contributions are never zero. For example, is there a day when motor vehicle traffic is truly zero? Thus, the solid lines in Figure 2 could represent minimum values within the data set, but the true profiles could be the dashed lines. Using those lines as the new axes will permit fitting the data as well as the solid lines. This uncertainty in the true relationships among the variables is the *rotational ambiguity* that is a problem for all forms of factor analysis. There are conditions that ensure a unique solution (*35*), but these are rarely attained in real environmental data.

The critical question is whether there are a sufficient number of points in the data set where the contributions of each source are sufficiently close to zero that the edges are defined. In an attempt to have a higher probability of having a sufficient number of edge points, large data sets are preferable. Analyses can be performed on smaller data sets, but typically few sources can be resolved and often the profiles are clearly a mixture of different sources. Thus, a key task for any factor analysis approach is to accurately identify these edges and the ability to achieve this end depends on the quality of the data set.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

Unmix

Introduction

Henry (*32*) outlined the conceptual framework of Unmix. It is designed to identify the major sources contributing to a set of samples. It recognizes that it will not resolve minor sources. Unmix uses the higher dimensional edges in the data to define the additional constraints needed to find a unique solution to the receptor model problem. However, it is then essential that the data set have a sufficient number of edge points that the algorithm can find it. The starting point for the analysis is a singular value decomposition of the uncentered but scaled data matrix to reduce the dimensionality of the problem to the number of causal factors that created the data. It is expected that these factors would be independent sources or source types that have well defined source profiles. Unmix finds the edges and uses them to calculate the vertices of the simplex, which are then converted back to source compositions and contributions. The edge-finding algorithm works in an arbitrary dimensional space and is described in Reference (*32*). It has been developed into software that is available from the U.S. EPA (*21*) and has been applied to a number of air quality data sets including airborne PM (*36–39*), particle number size distribution (*40*) and semivolatile compounds associated with airborne PM (*41–45*). Until the most recent version (V6), there were limitations to the number of factors that could be extracted and it often could not find an acceptable solution so it has not been widely used.

Illustrative Example

Lewis et al. (*37*) analyzed particulate matter with aerodynamic diameters less than 2.5 μm ($PM_{2.5}$) composition data collected in Phoenix, AZ. Daily, integrated 24-h samples were collected on 37mm diameter Teflon and quartz filter media for fine particle mass and species measurements using a dual fine particle sequential sampler (DFPSS). The samples were collected during the time period from March 1995 through June 1998. A total of 981 samples was finally obtained. Two energy dispersive X-ray spectrometers were used to produce the chemical elemental concentration data; a custom-made machine from Lawrence Berkeley Laboratories (LBL) and a commercially available one from Kevex (KEV). Both XRF instruments employed multiple choices for secondary excitation and utilized a helium atmosphere rather than vacuum in order to preserve volatile species. The quartz filters collected with the DFPSS were analyzed by Sunset Laboratory, Forest Grove, OR, USA using the thermal optical transmission technique (*46*). This technique measured both OC and EC. Each sample was characterized by the measured concentrations of the following4 6 chemical elements: Na, Mg, Al, Si, P, S, Cl, K, Ca, Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Ga, Ge, As, Se, Br, Rb, Sr, Y, Zr, Mo, Rh, Pd, Ag, Cd, Sn, Sb, Te, I, Cs, Ba, La, W, Au, Hg, Pb, organic carbon (OC), and elemental carbon (EC). In addition, water soluble potassium,

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Kw, was determined in water extracts of the samples as a marker species for biomass combustion. Unmix cannot accept negative concentration data, all such occurrences were replaced by half the minimum detection limit for that species. While this was necessary for species with low concentrations or poor detection limits, no replacements were needed for any of the species that were ultimately used in the Unmix analysis. After data screening, 789 samples were used in the Unmix analysis. These data were also used in the subsequent intercomparison of receptor modeling methods described by Hopke et al. (*5*).

Figure 3 shows the plots for a subset of the measured constituents. In the upper left pane, an edge (red dotted line) can be observed showing the relationship between one factor (source) that is rich in Si and the amount of $PM_{2.5}$. There are also clearly sources of $PM_{2.5}$ that are unrelated to Si. The upper right panel shows that there is a single source of Si and Al that would typically be assigned to be "soil." In the lower panels, it can be seen that there are likely two sources of material that include both Si and Ca and Fe, respectively, in their compositions. The Fe relationship is strong enough that an edge can be assigned, but that is not the case with Ca.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

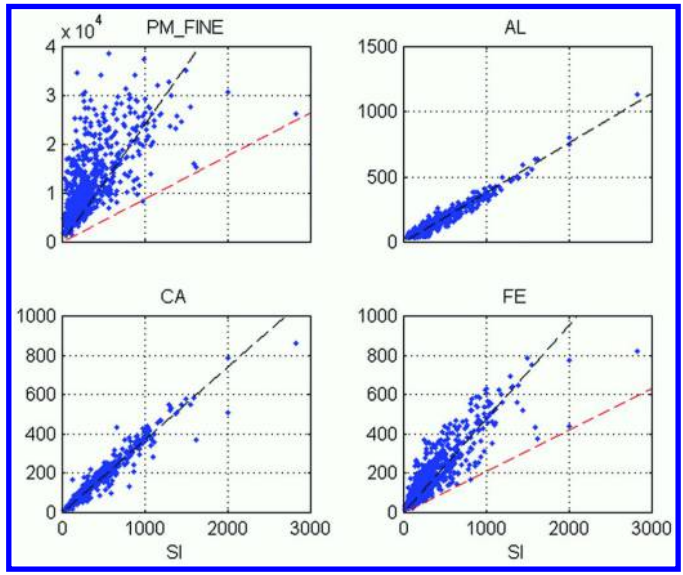

Figure 3. Plots of PM_Fine, Al, Ca, and Fe against Si for the Phoenix, AZ data set. (see color insert)

Five profiles were derived for these data and they are presented in Table 1. Since Unmix nominally provides a unique solution, error estimates can be obtained using a bootstrapping technique that is built into the software.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

Table 1. Unmix-derived source profiles (weight-fraction) and 1-sigma uncertainties (*37*)

	Crustal/Soil	*Gasoline*	*Secondary*	*Vegetative Burning*	*Diesel*
Si	0.0092 ± 0.1066	0.0075 ± 0.0113	0.0027 ± 0.0129	0.0087 ± 0.0269	0.0067 ± 0.0435
S	0.0056 ± 0.0316	0.0059 ± 0.0008	0.0077 ± 0.1361	0.0096 ± 0.0219	0.0075 ± 0.0118
K	0.0013 ± 0.0172	0.0013 ± 0.0004	0.0007 ± 0.0018	0.0078 ± 0.0386	0.0013 ± 0.0086
KW	0.0006 ± 0.0012	0.0005 ± 0.0021	0.0006 ± 0.0001	0.0078 ± 0.0345	0.0007 ± 0.002
Ca	0.0033 ± 0.0393	0.0026 ± 0.0033	0.001 ± 0.0037	0.0032 ± 0.0118	0.0025 ± 0.0186
Mn	0.0001 ± 0.0007	0.0001 ± 0	0.0000 ± 0.0001	0.0001 ± 0.0001	0.0003 ± 0.0021
Fe	0.0023 ± 0.0327	0.0015 ± 0.0072	0.0010 ± 0.0014	0.0033 ± 0.0048	0.0028 ± 0.032
OC	0.025 ± 0.214	0.028 ± 0.553	0.013 ± 0.326	0.037 ± 0.445	0.022 ± 0.309
EC	0.01 ± 0.034	0.017 ± 0.179	0.0091 ± 0.0072	0.016 ± 0.075	0.023 ± 0.204

Crustal material accounted for 22 ± 2% of the average mass concentration with gasoline, diesel exhaust, secondary $SO_4^=$. and vegetative burning contributing 33 ± 4%, 16 ± 2%, 19 ± 2%, and 10 ± 2%, respectively. The source strengths behave as expected with the diesel contributions being lower on weekends compared to weekdays and the ground level sources, diesel, gasoline, and agricultural burning, being higher in winter compared to summer since dispersion conditions are poorer in the winter. Sulfate was higher in the summer when there is more photochemical activity to convert the emitted SO_2 into $SO_4^=$. Nitrate was not measured in these samples, and its absence will result in the mass partitioning into the resolved sources being an overestimate of the true source contributions. Since particulate nitrate is highest in the winter, this artifact will likely more strongly affect the winter PM mass apportionment.

From these profiles, the contributions can be estimated for each sample. The sum of the contributions can be compared to the measured mass concentrations and an r^2 value of 0.97 was obtained. Thus, the mass concentrations were well described by the resulting 5 factor model, and the overall results were physically realistic.

Limited SEM examination of filter samples from the field study indicated the presence of additional sources (sea salt, copper smelter, iron foundry, fly ash) that presumably did not contribute large amounts of PM mass to the samples (*37*). As previously noted, the objective of Unmix is to identify and quantify the major sources and these weaker sources would not be expected to be resolved.

Unmix and many self-modeling curve resolution methods have used eigenvalue or singular value decompositions to obtain results. However, it is important to note that eigenvector decomposition is actually an implicit least squares fit that is minimizing the objective function as follows (*4*):

$$Q = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(e_{ij}\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \sum_{k=1}^{p} g_{ik} f_{kj})^2 \qquad (4)$$

For virtually all environmental data, the uncertainties in the measured variables are not uniform. In general, the errors scale with the measured values and thus, an unweighted least squares for the factor analysis approaches is as inappropriate as it was for the CMB approach.

Positive Matrix Factorization (PMF)

Introduction

An alternative approach is to explore the factor analysis problem as an explicit least-squares problem. This concept applied to the receptor modeling problem was first presented by Paatero and Tapper (*47*, *48*) and given that the solutions were constrained to be non-negative, it was termed positive matrix factoriztion. It solves the receptor modeling program outlined in Eq. 3 by minimizing the modified objective function given by:

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

$$Q = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{e_{ij}}{s_{ij}}\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\left[\frac{x_{ij} - \sum_{k=1}^{p} g_{ik} f_{kj}}{s_{ij}}\right]^2 \quad (5)$$

where s_{ij} is an estimate of the uncertainty for the jth species in the ith sample. This uncertainty is a combination of measurement error and the variability in the source profile value. As has been discussed above, there is natural variability in the source profile values given that it is not a fundamental property of the activity in the same way that a spectrum is a fundamental property of a chemical compound.

The advantage of this approach is that it makes it easy to handle values that fall below the detection limit (BDL) or that are missing. Eigenvalue methods cannot easily handle BDL values. If the BDL values are replaced with a fixed fraction of the detection limit (DL), then artificial correlations can be induced between variables that both have a number of BDL values in the same samples. In PMF, it is not a problem since the uncertainty can be made large enough to ensure that there is no artifact correlations. BDL values are often considered "missing," but it is not the case. Something is known about the values in that they are small. The exact value is not known but it is likely to fall between 0 and the DL. Thus, a fixed fraction of the DL can be substituted for the BDL point and a sufficiently large uncertainty can be assigned to allow it to properly influence the fit. Polissar et al. (*49*) showed that even with up to 86% of the Cl values being BDL at one site in Alaska, a good profile for sea salt could be obtained from the data since the Cl was well measured on those occasions when there was a strong intrusion of marine aerosol to the measurement sites.

For missing values, the ability to put in values with low weights in the fits allows the replacement of some missing data. If there are no values available for a given sampling period, then there is no bases to estimate them However, for many monitoring programs, multiple filters are collected and analyzed to provide the complete suite of chemical species that will be used in the source apportionment. Suppose one filter is lost, but the others have been properly analyzed. Then the missing values can be replaced with some central estimate of the distribution of that variable, but at the same time, a very large uncertainty is assigned to the value since nothing is known regarding the magnitude of the missing determination.

Based on these concepts, Polissar et al. (*49*) empirically explored many ways to estimate the uncertainties and proposed one that has come to be widely used. The approach has no underlying statistical theory, but rather is a practical set of rules that has been found to work well in many applications. For well determined values where there are reported measurement errors, the uncertainly is the sum of the measurement error plus 0.5 times the DL. For BDL values, the species concentration used in the PMF analysis is DL/2 and the uncertainty is assigned as 5/6 of the DL. For missing values, the median or geometric mean value is assigned to the variable value and the uncertainty is 4 times that value (400% error).

Another advantage of the explicit least-square formulation is that it permits easy incorporation of the natural constraints as well as any other constraints that are known a priori. The application of such constraints has been the subject of recent

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

work (*50*–*52*) and now has been incorporated into the latest version of EPA PMF (V5.0.14) (*21*). These constraints help to reduce the degree of rotational ambiguity. Extensive discussions of the problem of rotations are provided by Paatero et al. (*53*) and Paatero and Hopke (*54*).

Finally, the least-squares formulation allows the development of other models beyond the simple mass balance model outlined in Eq. 3. Examples of these more complex models will be discussed below.

Illustrative Examples

PMF has also been applied to the Phoenix data set (*5*, *55*, *56*). These analyses identified more sources including biomass burning, motor vehicles (with higher contribution in winter), coal-fired power plants (secondary particles with higher contributions in summer), soil, solid waste incinerator, and nonferrous smelting processes so that PMF was able to resolve minor sources that Unmix could not, but that were observed in the SEM analyses (*37*). For the sources that were identified by both methods, the mass apportionments were quite similar and both analyses suffer from the lack of nitrate data.

To illustrate PMF, data from St. Louis, MO will be examined. St. Louis is one of the few cities in the United States that still has large point sources of particulate emissions. It has been the subject of receptor modeling studies going back to the analyses of the Regional Air Pollution Study data collected in the mid-1970s (*57*, *58*). Subsequent studies were conducted (*59*), but the area remained in non-attainment of the ambient air quality standards for particulate matter. In 2001, the St. Louis – Midwest Supersite was established in East St. Louis, IL to conduct intensive studies of the ambient aerosol in the area (*60*).

Daily $PM_{2.5}$ samples were collected from June 2001 until May 2003 using multiple sampling devices and analyzed for a suite of composition variables including elements by XRF, ions by ion chromatography, and elemental and organic carbon (EC/OC) using the IMPROVE protocol (*46*). Details of the sampling and analyses are provided by Lee et al. (*61*). A total of 709 samples and 33 species were used in the PMF analysis that produced 10 identified sources including secondary sulfate, secondary nitrate, carbon-rich secondary sulfate, soil, gasoline vehicles, diesel vehicles, and the 4 major point sources, a steel mill, a primary lead smelter, a primary zinc smelter, and a copper products plant. Figure 4 shows the source profiles derived from the data while Figure 5 presents the time series of mass contributions of those identified sources.

A critical aspect of source apportionment analyses is to build evidence to support the assignment of source names to the various derived factors and to assess the quality of the results. There are several opportunities in these results. For the soil profile, it can be seen that there is a very strong peak in early July 2002. Other peaks are seen in the summers of 2001 and 2003. Soil is usually a weak source for $PM_{2.5}$ mass because wind blown or traffic resuspended soil typically has particle sizes greater than 2.5 μm. To explore the nature of this 2002 peak, it is possible to estimate the location of the air parcel backward in time using Lagrangian air parcel back trajectory models like HYSPLIT (*62*). A trajectory arriving on July 1,

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

2002 is shown in Figure 6. It can be seen that this event is related to the transport of dust from the Sahara Desert that occurs during the summer (*63*).

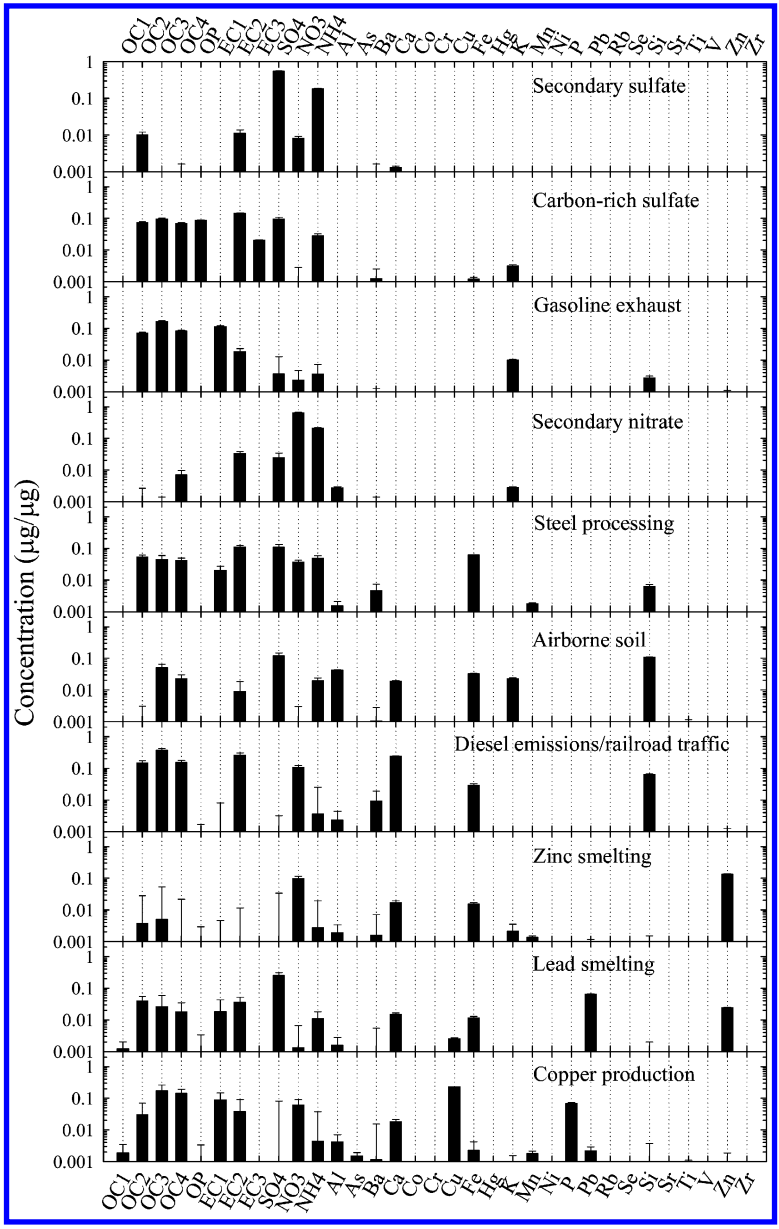

Figure 4. Source profiles derived from the daily $PM_{2.5}$ composition data from the St. Louis-Midwest Supersite. Figure taken from Lee et al. (61).

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

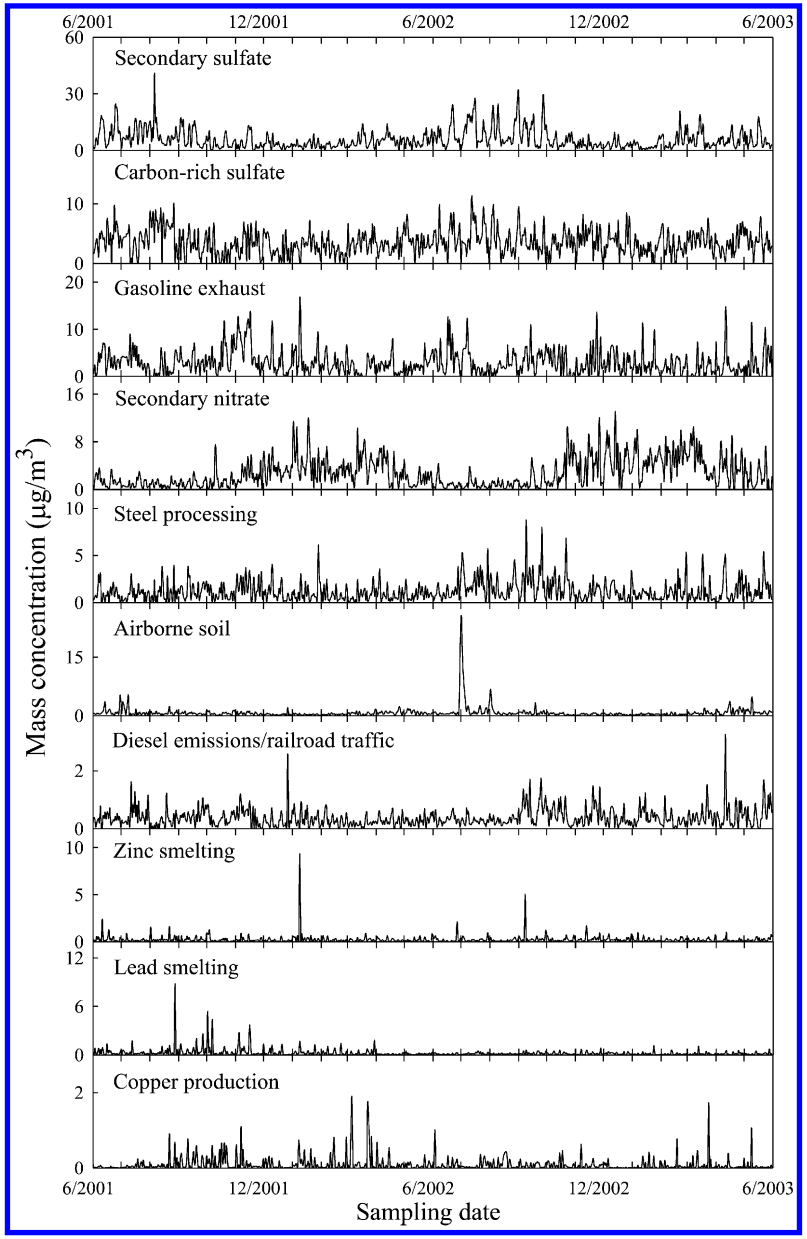

Figure 5. Source contributions derived from the daily PM2.5 composition data from the St. Louis-Midwest Supersite. Figure taken from Lee et al. (61).

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

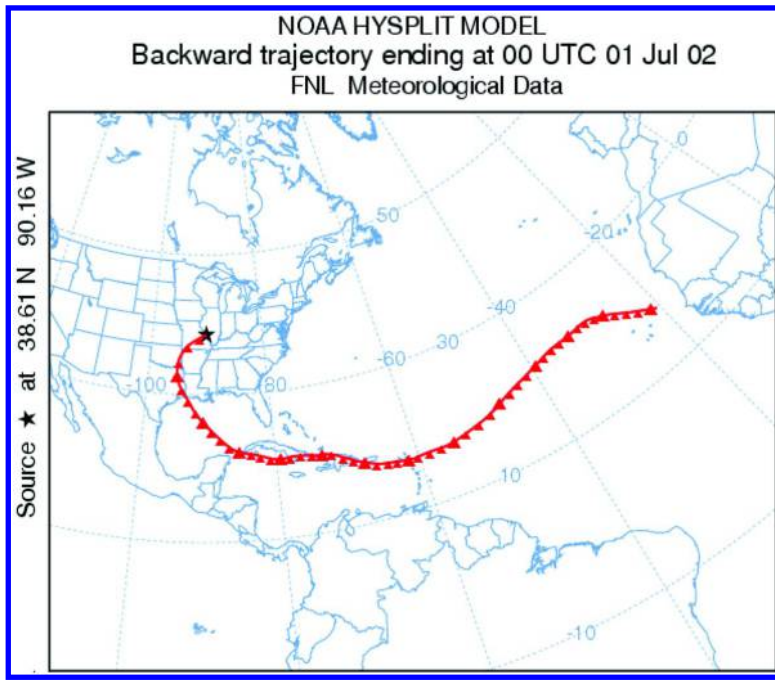

Figure 6. Air parcel back trajectory from St Louis on July 1, 2002. (see color insert)

PMF has now been used as a source apportionment tool in many airborne PM composition studies from a variety of locations around the world [see (*64*) and references therein]. It has been applied to VOC data (*65*), to Aerosol Mass Spectrometry (AMS) data [(*66*, *67*) and many others], and to particle number size distributions [e.g., (*68*, *69*)]. With the distribution of a version of PMF by the US EPA (*21*), PMF is now being routinely applied to many air pollution data sets.

Another interesting feature is in the time series of contributions ascribed to the lead smelter. It is the only primary lead smelter that was operating in the United States at that time. It can be seen that there are higher contributions in 2001 than for the rest of the measurements with smaller values during the first 4 months of 2002, followed by relatively low contributions from the rest of the measurement period. The smelter had been required to install additional control technology to be operational at the beginning of 2002. There were clearly some initial issues with its operation, but by May 2002, it was operating properly and continued to do so. Thus, the external information regarding the emissions from this facility agree well with the receptor modeling results and support the analysis.

There are two additional monitoring sites in St. Louis, MO operated as part of the US Environmental Protection Agency's Chemical Speciation Network. One of the sites is Blair Street site (38.6555 N, 90.1983W) in St. Louis City. The Blair Street site is surrounded by several major interstate highways. The other site, Arnold Street (38.4377 N, 90.3613W, 154m elevation) is in a suburban residential area in Jefferson County. Samples at these sites were collected only every third

day and were analyzed for the same set of chemical species but by a different set of analytical facilities. These data were also analyzed with PMF (*70*).

Seven and nine sources were identified at the Blair and Arnold sites, respectively, as compared with the ten identified at the Supersite. Several common sources were identified at all three sites (Table 2). However, several of the major sources were not resolved such as the zinc smelter and copper plan at Arnold St., and lead smelting and motor vehicles at Blair St. Part of these difference may be attributable to the differences in the analytical data base as well as the greater number of samples at the Supersite.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

Table 2. Mean contributions to $PM_{2.5}$ (Initial[a] vs constrained[b]) presented as µg/m³ (%)

	Blair St. (BS)		*Arnold St. (AS)*		*St. Louis Supersite (SS)*	
	Constrained	*Initial*	*Constrained*	*Initial*	*Constrained*	*Initial*
Sulfate	6.2 (38)	6.5 (40)	6.1 (39)	5.5 (36)	6.6 (37)	5.8 (33)
Nitrate	2.8 (17)	3.2 (20)	2.0 (13)	2.0 (13)	2.4 (13)	2.7 (15)
Zinc smelting	1.4 (9)	0.5 (3)	0.6 (4)		1.4 (8)	0.2 (1.3)
Copper products	0.1 (1)	0.6 (3)	0.03 (0.2)		0.2 (1)	0.1 (0.5)
Diesel	0.7 (5)		0.4 (3)	0.8 (5)	0.7 (4)	0.4 (2)
Lead smelting	0.4 (3)		0.4 (3)	0.5 (3)	0.6 (3)	0.2 (1)
Gasoline	2.6 (16)		2.9 (19)	3.2 (21)	1.3 (7)	2.9 (16)
Ca-rich			1.9 (12)	1.8 (12)		
Soil	1.6 (10)	2.5 (15)	0.4 (3)	0.5 (3)	0.5 (3)	0.8 (4)
Biomass			0.5 (3)	0.4 (2)		
Metal (Fe, Cu, Zn)				0.4 (3)		
Steel processing	0.3 (2)	0.05 (0. 3)			0.6 (3)	1.2 (7)
Motor vehicles (total)		2.8 (17)				
C-rich sulfate					3.4 (19)	3.5 (20)

[a] From Lee and Hopke (*70*); Lee et al. (*61*). [b] From Amato and Hopke (*52*).

Constrained Models

As mentioned above, adding constraints to the least square fitting process can reduce the extent of rotational ambiguity. With PMF being applied through

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

the use of the multilinear engine (*71*), it is possible to build constraints into the model as was done in two studies where there was an effort to separate multiple sources of similar composition (*50*, *51*). Amato et al. (*50*) applied the multilinear engine to data from an urban background site in Barcelona (Spain) to quantify the contribution of road dust resuspension to PM_{10} and $PM_{2.5}$ concentrations. A recent emission profile of local resuspended road dust had been previously obtained (*72*). This a priori information was introduced into the model as auxiliary terms in the object function to be minimized by the implementation of so-called "pulling equations" (*54*).

The multilinear engine permitted an enhanced solution when compared to the basic unconstrained PMF results. The enhanced analysis identified road dust source which accounted for 6.9 μg/m^3 (17%) of PM_{10}, 2.2 μg/m^3 (8%) of $PM_{2.5}$ and 0.3 μg/m^3 (2%) of PM_1 in addition to the other sources identified by in the initial analysis. These results reveal that resuspension was responsible of the 37%, 15% and 3% of total traffic emissions, respectively, of PM_{10}, $PM_{2.5,}$ and PM_1. Therefore, the overall traffic contribution resulted in 18 μg/m^3 (46%) of PM_{10}, 14 μg/m^3 (51%) of $PM_{2.5}$ and 8 μg/m^3 (48%) in PM_1. In the onconstrained solution, this mass explained by road dust resuspension was re-distributed among the rest of sources, increasing mostly the mineral, secondary nitrate and aged sea salt contributions.

Escrig et al. (*51*) applied a similar approach to speciated PM10 data obtained at three air quality monitoring sites between 2002 and 2007 in a highly industrialized area in Spain. The source apportionment of PM in this area is an especially difficult task. There are industrial mineral dust emissions that need to be separately quantified from the natural sources of mineral PM. On the other hand, the diversity of industrial processes in the area results in a puzzling industrial emissions scenario. The availability of specific source profiles for particular major industrial emissions permitted the resolution of the industrial emissions from other sources providing an opportunity to quantitatively evaluate the effectiveness of abatement programs for regional air quality improvement.

Amato and Hopke (*52*) have applied constraints to combine the analysis of the three sites in the St. Louis area into a single analysis such that known source profiles could be worked into the analysis. To obtain good target profiles for major sources derived from data independent from the particle composition data collected at each of the three sites, additional high time-resolution data collected as part of the St. Louis - Midwest Supersite study was employed. Organic and elemental carbon concentrations were measured hourly using a Sunset field OC/EC system (*73*). Elements were measured using a semi-continuous elements in aerosol system (SEAS) described by Kidwell and Ondov (*74*, *75*) (2001, 2004). Briefly, the method uses condensational growth by direct steam injection to grow particles as small as 0.084 mm thereby delivering a particle slurry that is suitable for analysis by multi-element graphite furnace atomic absorption spectrometry. The SEAS samples were analyzed for eleven elements (Al, As, Cd, Cr, Cu, Fe, Mn, Ni, Pb, Se, and Zn) by graphite furnace atomic absorption spectrometry. A particle-into-liquid system (PILS) (*76*) was used to collect $PM_{2.5}$ samples and analyze them for sulfate, nitrate, sodium, potassium, and ammonium ions. Applying PMF to these data permitted identifying factor profiles of the copper products plant, the zinc smelter

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

and the steel mill factor. Average tailpipe emissions profiles were available from Schauer et al. (*77*). These profiles were taken as targets and introduced in the ME continuation run with the aim of extend the number of sources found.

For the cases where the metallurgic factors were found by in the original studies (*61*, *70*), the pulling equations permitted improvements in the source profiles and therefore better impact estimates of their contributions. These new contributions were in better agreement with the location of point source and receptors. For example, at Blair St, zinc and copper metallurgy contributed in proportion relative to the resolved contributions at the Supersite (1.4 $\mu g/m^3$ and 0.1 $\mu g/m^3$, respectively). The initial analysis attributed higher zinc emissions at Blair St than at the Supersite, which is located closer to and more downwind of the zinc smelter. Copper emissions were also overestimated in the initial analyses (*70*), exceeding those at the Supersite by a factor of 6. In addition, contributions from steel mill and lead smelter could be estimated at those sites where the previous studies could not. Gasoline and diesel emissions were separated at Blair St, with the diesel contributions being higher (similarly to the the Supersite results) than at Arnold St, while initially the opposite was observed, with no estimate of diesel emission impacts at Blair St. Additional features could be observed with respect to the sources that were now identified. The metallurgic factors at the Arnold and Blair St sites are in good agreement with the position of the point sources. Gasoline and diesel factors were obtained at Blair St. A better resolved factor (Ca-rich factor at Arnold St) shows directionality to the southeast where a cement plant is located in the same direction as the lead smelter.

Constraints have shown to be of sufficient value that they have now been incorporated into the US EPA version of PMF in version 5.0.14 (*21*). Using them requires a multiple step analysis in which an initial solution is obtained and then constraints applied to the continuation run. Details of how to perform such analyses are provided in the user's manual (*78*).

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

Complex Models

Expanded Model: Other approaches to reduce the rotational ambiguity and increase the number of sources resolved have been developed. Paatero and Hopke (*79*) first introduced the concept of solving parallel equations in which the second equation can take into account other drivers of variation such as wind speed and direction, day of week, seasonality, etc. Airborne concentrations due to specific sources may display a sharp directional pattern with respect to wind wind directions. In these cases, concentrations are high when the air arrives from certain direction(s) while concentrations associated with other directions are low or nil. Such non-linear dependency cannot be directly modeled so that wind information would be included in a factor analytic model as one or a few special variables, used in parallel with the ordinary variables, the concentrations. There may be other similar kinds of effects such as weekend/weekday activity patterns, time of day, time during the year, etc. that significantly affect the observed elemental concentrations. The non-linear variables can be included in the model

as independent or free variables. This incorporation of a secondary equation has been terms an expanded factor analysis model (*80*).

In the expanded ME analysis, the bilinear model shown in Eq. (3) is augmented by additional complex equation that contains modeling information. The most basic form of the this equation is

$$x_{ij} = \sum_{k=1}^{p} D(\delta_i, k) V(\nu_i, k) f_{kj} + e'_{ij} \tag{6}$$

where **D** and **V** represent matrices, consisting of unknown values to be estimated during the model fitting process. The known index value δ_i and v_i indicate wind direction and wind speed of the ith day for the kth source, respectively. The indices are shown in parentheses, not as subscripts for the typographic reasons. In this model, the index value δ_i is obtained from the classification of the wind direction on the ith day into a set of indices. For example, 18 indices could used to represent 20º wind direction sectors, Then if the actual wind direction was 170º, the value of δ_i would be 7 (*80*).

In Eq. (6), other information on the sources of variation in the concentration that might aid the separation of the sources can be incorporated. For example, Kim et al. (*80*) used wind direction, wind speed, time of day, time of year, and weekend/weekday were used. For the wind direction and wind speed, hourly averaged values were used. The complete expanded model consists of the basic bilinear equation and a multilinear equation specifying the physical model:

$$x_{ij} = \sum_{k=1}^{p} g_{ij} f_{kj} + e_{ij} \tag{7}$$

$$x_{ij} = \sum_{k=1}^{p} \mathbf{S}(\eta_i, k)\mathbf{W}(\omega_i, k) \sum_{h=1}^{24} \mathbf{D}(\delta_{ih}, k)\mathbf{V}(\nu_{ih}, k)\mathbf{R}(\varepsilon_{ih}, k)\mathbf{T}(\lambda_{ih}, k) f_{kj} + e'_{ij}$$

where $\mathbf{S}(\eta_i,k)$ is the element of matrix **S** with the index values η_i corresponding to the time-of-year classification of the i^{th} day for the k^{th} source. Time-of-year is classified into six two-month periods (or seasons). $\mathbf{W}(\omega_i,k)$ is an element of the matrix **W** with the index values ω_i corresponding to weekend/weekday factor of the ith day for the kth source. The weekend effect matrix **W** has dimension 1 by p. Often, the weekday coefficients have been fixed at unity so that only the weekend coefficients are variable. The elements of matrix **W** specify the average strength of each factor on weekend relative to the strength in weekday. $D(\delta_{ih},k)$ is the element of matrix **D** with the index values δ_{ih} for the wind direction during hour h of the ith day for the kth source. $V(\nu_{ih},k)$ is the element of matrix **V** with the index values ν_{ih} for the wind speed during hour h of the ith day for the kth source. $R(\varepsilon_{ih},k)$ is the element of matrix **R** with the index values ε_{ih} for the calm wind (< 1 m/sec) during hour h of the ith day for the kth source. Because of isotropic wind direction, calm wind was separated as an separate matrix **R** in this analysis instead of being included in the wind speed index matrix **V**. Also, the wind direction of calm wind was not incorporated in the wind direction index matrix **D**. $T(\lambda_{ih},k)$ is the element of matrix **T** with the index values λ_{ih} for the time-of-day during hour h of the ith day for the kth source. The matrices, **S**, **W**, **D**, **V**, **R**, and **T** contain unknown values to be

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

estimated in the fitting process. The specific factor elements used to fit a particular data point are selected based on the hourly or daily values of the corresponding auxiliary variables. Therefore, these auxiliary variables are not fitted, but served to determine the indices of the values to be fitted.

ME provides a solution that minimizes the value of Q, based upon uncertainty estimates for each observation [] while the values of the unknown matrices **G**, **F**, **S, W, D**, **V**, **R,** and **T** are to be determined so that the model fits the data as well as possible. The Q value is defined as:

$$Q = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{e_{ij}}{\sigma_{ij}}\right)^2 + \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{e'_{ij}}{\sigma'_{ij}}\right)^2 \quad (8)$$

where σ_{ij} is an uncertainty estimate for the bilinear model and σ'_{ij} is an uncertainty estimate for the multilinear model in Equation 6 in the jth element measured in the ith day.

Equation 7 is one of multiple possible models depending on the understanding of the system under study while the mass balance in the bilinear equation should always be applicable. Because the variability of the index factors is restricted by the model, Eq. 7 will produce a significantly poorer fit to the data than the bilinear equation (Eq. 3). Therefore, the uncertainty estimates corresponding to the multilinear equation must be larger than those corresponding to the bilinear equation to decrease the weight of the multilinear equation in the solution. In prior studies, experiments have been performed with different uncertainties resulting in estimated uncertainties in the multilinear equation being set at nine times the estimated uncertainties of the bilinear equation (*80*).

This approach has been applied to data from Atlanta (*80*), Washington, DC (*81*), New York City (*82*), and Cleveland (*83*) with mixed results. In Atlanta and Washingon, the expanded model clearly permitted more sources to be better resolved. However, in NYC and Cleveland, there were little differences between the conventional PMF solution and the expanded model results. It is not yet clear why some locations seem to be more amenable to the expanded model and some are not. Further work is required to better understand its applicability.

Other Complex Models: An advantage of the explicit least squares formulations such as PMF is that conceptual models can be built and tested based on the nature of the processes underlying the creation of the data set. A number of such models have been developed to maximize the information recovery from the collected data sets.

Multiple Sample Type Data: In many panel studies of the effects of airborne particles on health, measurements are made in multiple environments. For example, Hopke et al. (*84*) report on the analysis of elderly subjects living in a single multifamily residence. Measurements were made at a central outdoor site, an unoccupied room in the building and using personal samplers on specific individuals. Thus, different sources will affect different sample types. Only "external" sources of ambient particles will affect the outdoor samples. However, ambient particles will penetrate into indoor air and add to the exposure observed in the indoor and personal samples. Indoor sources such as cooking and the use of personal care products will not affect the outdoor samples.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

The expanded receptor model for this study can be expressed as:

$$x_{ijdt} = \sum_{p=1}^{N} g_{ipdt} f_{jp} + \sum_{p=N+1}^{N+H} g_{ipdt} f_{jp} \quad (t=1/2\text{: personal/indoor}) \quad (9)$$

$$x_{jdt} = \sum_{p=1}^{N} g_{pdt} f_{jp} \quad (t=3\text{: outdoor}) \quad (10)$$

where i is the individual (subject or participant) index, j is the species index, d is the sampling date index, t is the type index, N is the number of external sources, H is the number of internal sources. x_{ijdt} denotes the concentration of species j in the sample of type t collected by subject i on date d, g_{ipdt} denotes the contribution of source p to the sample of type t collected by subject i on date d, f_{jp} denotes the relative concentration of species j in source p.

This model has been used to analyze data for cardiac patients in the Raleigh-Chapel Hill area of North Carolina (*85*) and of data for asthmatic children attending a special school for moderate to severe asthmatics in Denver (*86*). In the case of the Denver study, four external sources and three internal sources were resolved from the $PM_{2.5}$ data for the three different environments. Secondary nitrate and motor vehicle emissions were the two largest external sources in this study. Cooking was the largest internal source. A significant influence of indoor tobacco smoking on daily personal exposures to particles was observed for those houses in which smokers reside and the environmental tobacco smoke contribution correlated with urinary cotinine levels in these urban schoolchildren. The influence of the high traffic flow outside the school on the indoor air quality was also observed.

Time Synchronization Model: One of the major developments in atmospheric monitoring over the past 15 years has been the deployment of more real-time and near real-time instruments. However, these instruments collect data at different frequencies ranging from a few minutes to a few hours. Higher frequency data have the advantage that transient events can be observed that can often provide edge points that would otherwise be averaged out of a longer interval sample (*87*). Thus, it is not desirable to average the higher frequency data to the longer time interval instrument data in the suite of data. There is no way to split the longer integration time data down to the shorter time intervals so it is necessary to have models that permit each set of data to be included within its own measurement frequency. Such models have been applied to several of the sets of data from the US EPA's Supersite program. Zhou et al. (*88*) analyzed data from Pittsburgh, PA while Ogulei et al. (*89*) used the same model for data from Baltimore, MD. The model has been examined further using simulated data (*90*) and found that the model performed well.

Multiway Data: The vast majority of applications of SMCR are to matrices that provide information on chemical properties of a series of samples. However, there is also the potential for data with increased dimensionality. For example, if particles are segregated by aerodynamic diameter into multiple samples collected during a given time interval that are then analyzed for their chemical composition, the data set is then a 3-way array or tensor consisting of size, composition, and

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

time period. Data for a single variable like $PM_{2.5}$ mass concentrations could be collected from multiple sites across an area so that the three-ways would be latitude, longitude, and concentration. If those samples were then analyzed for composition, there would then be a 4-way array. Various such applications of PMF have been made and demonstrate how conceptual models can be built to fit the data rather than all data be fit to the same bilinear model.

Spatially Distributed Data: Paatero et al. (*91*) examined a spatial data set of $PM_{2.5}$ mass concentrations measured every third day at over 300 locations in the eastern United States during 2000. The basic PMF model was enhanced by modeling the dependence of $PM_{2.5}$ concentrations on temperature, humidity, pressure, ozone concentrations, and wind velocity vectors. The model comprises 12 general factors, augmented by 5 urban-only factors intended to represent excess concentration present in urban locations only. The flux density maps showed the major transport patterns of $PM_{2.5}$. For example, they show the increase in particle mass as the air moves from the regions of the gaseous precursor (SO2) and is converted in sulfate. Recognition of this combination of transport and transformation is necessary in order that control procedures can be targeted to significant causes of high $PM_{2.5}$ concentrations.

A different spatial model was developed by Chuienta et al. (*92*) for the analysis of the spatial patterns and possible sources affecting haze and its visual effects in the southwestern United States. The data are from the Measurement of Haze and Visual Effects (MOHAVE) project thate were collected during the late winter and midsummer of 1992 at the monitoring sites in four states (i.e., California, Arizona, Nevada and Utah). The resulting three-way data array was analyzed by a four product-term model. This study makes a direct effort to include wind patterns as a component in the model in order to obtain the information of the spatial patterns of source contributions. The solution is computed using the conjugate gradient algorithm with applied non-negativity constraints. For the winter data set, reasonable solutions contained six sources and six wind patterns. The analysis of summer data required seven sources and seven wind patterns.

Size-Composition-Time Data: There are a number of devices that can separate particles by size such that samples can be collected that represent a relatively limited particle size range. The most common of these systems is a cascade impactor in which particles are sequentially separated and collected for analysis. Most of these systems are manually operated so there is considerable effort involved in collecting a series of samples. However, there have been several systems developed for collecting a time-series of time- and size-resolved samples that can then be analyzed. One of these systems is the rotating DRUM impactor sampler (*93*) that collects the particles on Mylar films placed on a rotating drum under the nozzle that determines the aerodynamic behavior of the particles. The resulting samples can be analyzed using synchrotron XRF (*94*) to provide the 3-way data set.

Different sources have different size-composition profiles in their emissions (*95*). Thus, a source profile for size segregated data is a matrix of composition as a function of size and therefore, a special model is required to properly account for the processes by which the particles are formed and emitted into the atmosphere. The main equation of the model is as follows:

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

$$\vec{X} = A \otimes \vec{B} + \vec{E} \tag{11}$$

where $\vec{X}(I,J,K)$ is the three-way array of observed data, ⊗ represents a Kroneker product (*96*, *97*) of the source profile array $\vec{B}(I,J,K)$ with the contribution matrix, A(I,P), P is the number of factors, and $\vec{E}(I,J,K)$ is the three-way array of residuals.

This model has been applied to several data sets including three-stage DRUM impactor data from Detroit, MI with the samples collected between February and April 2002 (*98*) and eight-stage DRUM impactor data from the Washington-Dulles International Airport (*99*). For the Detroit data (*98*), nine factors were identified: road salt, industrial (Fe+Zn), cloud processed sulfate, two types of metal works, road dust, local sulfate source, sulfur with dust, and homogeneously formed sulfate. Road salt had high concentrations of Na and Cl. Mixed industrial emissions are characterized by Fe and Zn. The cloud processed sulfate had a high concentration of S in the intermediate size mode. The first metal works represented by Fe in all three size modes and by Zn, Ti, Cu, and Mn. The second included a high concentration of small size particle sulfur with intermediate size Fe, Zn, Al, Si, and Ca. Road dust contained Na, Al, Si, S, K, and Fe in the large size mode. The local and homogeneous sulfate factors show high concentrations of S in the smallest size mode, but different time series behavior in their contributions. Sulfur with dust is characterized by S and a mix of Na, Mg, Al, Si, K, Ca, Ti, and Fe from the medium and large size modes. The analysis utilized light absorption measurements at 4 wavelengths, 350, 450, 550, and 650 nm, to provide limited information on the carbonaceous components in the samples.

At Dulles International Airport, five major emission sources: soil, road salt, aircraft landings, transported secondary sulfate, and local sulfate/construction were identified (*99*). Aircraft landing was notable for it had not previously been identified as a significant source of PM2.5. Its pattern showed small particles of sulfur, zinc, bromine, zirconium and molybdenum. This factor is assigned to particles that are emitted during landings. The sulfur and zinc come from tire wear. These elements are key constituents in tires. Often a visible puff of smoke is observed at touchdown. There is considerable frictional heat produced at this instant and particles are generated across the particle size range. Both zirconium and molybdenum are used in high temperature greases as might be used to lubricate bearings that would undergo significant heat stress. The energy deposited in the bearings can be expected to liberate particles from the lubricants. The study shows that time- and size-resolved DRUM data can assist in the identification of the airport emission sources and atmospheric processes leading to the observed ambient concentrations.

Conclusions

For more than 45 years, data analysis tools like factor analysis have been applied to atmospheric chemical species data to help understand the nature of the sources of pollutants and the relative contributions those sources make to the observed ambient concentrations. Many of these tools have been forms of

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

self-modeling curve resolution and two of them, Positive Matrix Factorization (PMF) and Unmix, have come to dominate the field of receptor modeling. It is particularly important to recognize that atmospheric compositional data are qualitatively different from many SMCR problems given the higher measurement errors, but more importantly the variability in their chemical characteristics (profile). This variability makes the receptor modeling challenging. There is also the problem of rotational ambiguity that is present for lack of adequate numbers of true edge points in most data sets. Unmix has not been as widely used as PMF. PMF has proven to be very useful and now is widely used by a large number of investigators that are able to take advantage of an easy-to-use tool that now has greatly improve approaches for estimating the uncertainties in the solutions (*100*). The key step forward for source apportionment in the future will be developing new measurement tools that will provide more chemical species to better define and separate sources and higher precision to reduce the level of noise in the data. These improvements would permit even better source resolutions to be performed.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

References

1. Lawton, W. H.; Sylvestre, E. A. *Technometrics* **1973**, *13*, 617–633.
2. Jiang, J. H.; Liang, Y.; Ozaki, Y. *Chemom. Intell. Lab. Syst.* **2004**, *71*, 1–12.
3. Rasmussen, G. T.; Isenhour, T. L.; Lowry, S. R.; Ritter, G. L. *Anal. Chim. Acta* **1978**, *103*, 213–221.
4. Malinowski, E. *Factor Analysis in Chemistry*, 2nd ed.; Wiley: New York, 1991.
5. Hopke, P. K.; Ito, K.; Mar, T.; Christensen, W. F.; Eatough, D. J.; Henry, R. C.; Kim, E.; Laden, F.; Lall, R.; Larson, T. V.; Liu, H.; Neas, L.; Pinto, J.; Stölzel, M.; Suh, H.; Paatero, P.; Thurston, G. D. *J. Exp. Anal. Environ. Epidemiol.* **2006**, *16*, 275–286.
6. Engel-Cox, J. A.; Weber, S. A. *J. Air Waste Manage. Assoc.* **2007**, *57*, 1307–1316.
7. Viana, M.; Kuhlbusch, T. A. J.; Querol, X.; Alastuey, A.; Harrison, R. M.; Hopke, P. K.; Winiwarter, W.; Vallius, M.; Szidat, S.; Prévôt, A. S. H.; Hueglin, C.; Bloemen, H.; Wåhlin, P.; Vecchi, R.; Miranda, A. I.; Kasper-Giebl, A.; Maenhaut, W.; Hitzenberger, R. *J. Aerosol Sci.* **2008**, *39*, 827–849.
8. Belis, C. A.; Karagulian, F.; Larsen, B. R.; Hopke, P. K. *Atmos. Environ.* **2013**, *69*, 94–108.
9. Hopke, P. K. *Receptor Modeling in Environmental Chemistry*; John Wiley & Sons, Inc.: New York, 1985.
10. Hopke, P. K., Ed. *Receptor Modeling for Air Quality Management*; Elsevier Science Publishers: Amsterdam, 1991.
11. U.S. Environmental Protection Agency. *Air Quality Criteria for Particulate Matter*, Volume I of II; Report no. EPA/600/P-99/002aF; Office of Research, National Center for Environmental Assessment: Research Triangle Park, NC, 2004.

12. Roscoe, B. A.; Chen, C. Y.; Hopke, P. K. *Anal. Chim. Acta* **1984**, *160*, 121–134.
13. Hopke, P. K. *Chemometrics in Environmental Chemistry*; Springer-Verlag: Heidelberg, 1995; pp 47−86.
14. Henry, R. C. Multivariate Receptor Models. In *Receptor Modeling for Air Quality Management*; Hopke, P. K., Ed.; Elsevier: Amsterdam, 1991; pp 117−147.
15. Winchester, J. W.; Nifong, G. D. *Water Air Soil Pollut.* **1971**, *1*, 50–64.
16. Miller, M. S.; Friedlander, S. K.; Hidy, G. M. *J. Colloid Interface Sci.* **1972**, *39*, 65–176.
17. Friedlander, S. K. *Environ. Sci. Technol.* **1973**, *7*, 235–240.
18. Kowalczyk, G. S.; Choquette, C. E.; Gordon, G. E. *Atmos. Environ.* **1978**, *12*, 1143–1153.
19. Fuller, W. A., *Neasurement Error Models*; J. Wiley & Sons, Inc.: New York, 1987.
20. Watson, J. G.; Cooper, J. A.; Huntzicker, J. J. *Atmos. Environ.* **1984**, *18*, 1347–1355.
21. U.S. Environmental Protection Agency. Receptor Modeling. http://www.epa.gov/ttn/scram/receptorindex.htm.
22. Subramanian, R.; Donahue, N. M.; Bernardo-Bricker, A.; Rogge, W. F.; Robinson, A. L. *Atmos. Environ.* **2006**, *40*, 8002–8019.
23. Blifford, I. H.; Meeker, G. O. *Atmos. Environ.* **1967**, *1*, 147–157.
24. Prinz, B.; Stratmann, H. *Staub-Reinhalt Luft* **1968**, *28*, 33–39.
25. Hopke, P. K.; Gladney, E. S.; Gordon, G. E.; Zoller, W. H.; Jones, A. G. *Atmos. Environ.* **1976**, *10*, 1015–1025.
26. Gaarenstroom, P. D.; Perone, S. P.; Moyers, J. P. *Environ. Sci. Technol.* **1977**, *11*, 795–800.
27. Hopke, P. K. *Atmos. Environ.* **1988**, *22*, 1777–1792.
28. Henry, R. C.; Kim, B.-M. *Chemom. Intell. Lab. Syst.* **1989**, *8*, 205–216.
29. Henry, R. C. *Chemom. Intell. Lab. Syst.* **1997**, *37*, 37–42.
30. Kim, B.-M.; Henry, R. C. *Chemom. Intell. Lab. Syst.* **1999**, *49*, 67–77.
31. Kim, B.-M.; Henry, R. C. *Atmos. Environ.* **2000**, *34*, 1747–1759.
32. Henry, R. C. *Chemom. Intell. Lab. Syst.* **2003**, *65*, 179–189.
33. Imbrie, J. Technical Report No. 6, ONR Task No. 389-135; Northwestern University: Evanston, IL, 1963.
34. Windig, W.; Guilment, J. *Anal. Chem.* **1991**, *63*, 1425–1432.
35. Anderson T. W. *An Introduction to Multivariate Statistical Analysis*, 2nd ed.; Wiley: New York, 1984.
36. Chen, L.-W. A.; Doddridge, B. G.; Dickerson, R. R.; Chow, J. C.; Henry, R. C. *Atmos. Environ.* **2002**, *36*, 4541–4554.
37. Lewis, C. W.; Norris, G. A.; Conner, T. L.; Henry, R. C. *J. Air Waste Manage. Assoc.* **2003**, *53*, 325–338.
38. Hu, S. H.; McDonald, R.; Martuzevicius, D.; Biswas, P.; Grinshpun, S. A.; Kelley, A.; Reponen, T.; Lockey, J.; LeMasters, G. *Atmos. Environ.* **2006**, *40*, S378–S395.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

39. Li, C.; Wen, T. X.; Li, Z. Q.; Dickerson, R. R.; Yang, Y. J.; Zhao, Y. A.; Wang, Y. S.; Tsay, S. C. *J. Geophys. Sci.-Atmos.* **2010**, D00K23; DOI: 10.1029/2009JD013639.
40. Kim, E.; Hopke, P. K.; Larson, T. V.; Covert, D. S. *Environ. Sci. Technol.* **2004**, *38*, 202–209.
41. Miller, S. L.; Anderson, M. J.; Daly, E. P.; Milford, J. B. *Atmos. Environ.* **2002**, *36*, 3629–3641.
42. Mukerjee, S.; Norris, G. A.; Smith, L. A.; Noble, C. A.; Neas, L. M.; Ozkaynak, A. H.; Gonzales, M. *Environ. Sci. Technol.* **2004**, *38*, 2317–2327.
43. Song, Y.; Dai, W.; Shao, M.; Liu, Y.; Lu, S. H.; Kuster, W.; Goldan, P. *Environ. Pollut.* **2008**, *156*, 174–183.
44. Khairy, M. A.; Lohmann, R. *Chemosphere* **2013**, *91*, 895–903.
45. Patokoski, J.; Ruuskanen, T. M.; Hellen, H.; Taipale, R.; Gronholm, T.; Kajos, M. K.; Petaja, T.; Hakola, H.; Kulmala, M.; Rinne, J. *Boreal Environ. Res.* **2014**, *19*, 79–103.
46. Chow, J. C.; Watson, J. G.; Pritchett, L. C.; Pierson, W. R.; Frazier, C. A.; Purcell, R. G. *Atmos. Environ.* **1993**, *27A*, 1185–1201.
47. Paatero, P.; Tapper, U. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 183–194.
48. Paatero, P.; Tapper, U. *Environmetrics* **1994**, *5*, 111–126.
49. Polissar, A. V.; Hopke, P. K.; Paatero, P.; Malm, W. C.; Sisler, J. F. *J. Geophys. Res.* **1998**, *103*, 19045–19057.
50. Amato, F.; Pandolfi, M.; Escrig, A.; Querol, X.; Alastuey, A.; Pey, J.; Perez, N.; Hopke, P. K. *Atmos. Environ.* **2009**, *43*, 2770–2780.
51. Escrig, A.; Monfort, E.; Celades, I.; Querol, X.; Amato, F.; Minguillion, M. C.; Hopke, P. K. *J. Air Waste Manage. Assoc.* **2009**, *59*, 1296–1307.
52. Amato, F.; Hopke, P. K. *Atmos. Environ.* **2012**, *46*, 329–337.
53. Paatero, P.; Hopke, P. K.; Song, X.; Ramadan, Z. *Chemom. Intell. Lab. Syst.* **2002**, *60*, 253–264.
54. Paatero, P.; Hopke, P. K. *J. Chemom.* **2009**, *23*, 91–100.
55. Ramadan, Z.; Song, X.-H.; Hopke, P. K. *J. Air Waste Manage. Assoc.* **2000**, *50*, 1308–1320.
56. Ramadan, Z.; Eickhout, B.; Song, X.-H.; Buydens, L. M. C.; Hopke, P. K. *Chemom. Intell. Lab. Syst.* **2003**, *66*, 15–28.
57. Alpert, D. J.; Hopke, P. K. *Atmos. Environ.* **1981**, *15*, 675–687.
58. Kim, E.; Hopke, P. K.; Pinto, J. P.; Wilson, W. E. *Environ. Sci. Technol.* **2005**, *39*, 4172–4179.
59. Glover, D. M.; Hopke, P. K.; Vermette, S. J.; Landsberger, S.; D'Auben, D. R. *J. Air Waste Manage. Assoc.* **1991**, *41*, 294–305.
60. Turner, J. R. *St. Louis – Midwest Fine Particulate Matter Supersite, Final Report to the U.S. Environmental Protection Agency*; March 2007, available at http://www.epa.gov/ttnamti1/files/ambient/super/STL-SS_FinalReport_Rev02_March2007.pdf.
61. Lee, J. H.; Hopke, P. K.; Turner, J. R. *J. Geophys. Res.* **2006**, *111*, D10S10; DOI: 10.1029/2005JD006329.
62. Draxler, R. R.; Rolph, G. D. *HYSPLIT (HYbrid Single-Particle Lagrangian Integrated Trajectory) Model*; NOAA Air Resources Laboratory:

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

Silver Spring, MD, 2010; access via NOAA ARL READY Website, http://ready.arl.noaa.gov/HYSPLIT.php.
63. Gatz, D. F.; Prospero, J. M. *Atmos. Environ.* **1996**, *30*, 3789–3799.
64. Hopke, P. K. *Pollution Atmosphérique* **2010** (September), 91–109 (Special Issue).
65. Kim, E.; Brown, S. G.; Hafner, H. R.; Hopke, P. K. *Atmos. Environ.* **2005**, *39*, 5934–5946.
66. Lanz, V. A.; Alfarra, M. R.; Baltensperger, U.; Buchmann, B.; Hueglin, C.; Prevot, A. S. H. *Atmos. Chem. Phys.* **2007**, *7*, 1503–1522.
67. Ulbrich, I. M.; Canagaratna, M. R.; Zhang, Q.; Worsnop, D. R.; Jimenez, J. L. *Atmos. Chem. Phys.* **2009**, *9*, 2891–2918.
68. Kim, E.; Hopke, P. K.; Larson, T. V.; Covert, D. S. *Environ. Sci. Technol.* **2004**, *38*, 202–209.
69. Kasumba, J.; Hopke, P. K.; Chalupa, D. C.; Utell, M. J. *Sci. Total Environ.* **2009**, *407*, 5071–5084.
70. Lee, J. H.; Hopke, P. K. *Atmos. Environ.* **2006**, *40* (Suppl. 2), S360–S377.
71. Paatero, P. *Comput. Graphic. Stats.* **1999**, *8*, 1–35.
72. Amato, F; Pandolfi, M; Viana, M; Querol, X; Alastuey, A; Moreno, T. *Atmos. Environ.* **2009**, *43*, 1650–1659.
73. Jeong, C. H.; Hopke, P. K.; Kim, E.; Lee, D. W. *Atmos. Environ.* **2004**, *38*, 5193–5204.
74. Kidwell, C. B.; Ondov, J. M. *Aerosol Sci. Technol.* **2001**, *35*, 596–601.
75. Kidwell, C. B.; Ondov, J. M. *Aerosol Sci. Technol.* **2004**, *38*, 205–218.
76. Weber, R. J.; Orsini, D.; Daun, Y.; Lee, Y.-N.; Klotz, P. J.; Brechtel, F. *Aerosol Sci. Technol.* **2001**, *35*, 718–727.
77. Schauer, J. J.; Lough, G. C.; Shafer, M. M.; Christensen, W. F.; Arndt, M. F.; DeMinter, J. T.; Park, J.-S. *Characterization of metals emitted from motor vehicles*; Health Effects Institute: Boston, MA, 2006.
78. U.S. Environmental Protection Agency. *EPA Positive Matrix Factorization (PMF) 5.0 Fundamentals and User Guide*; available at http://www.epa.gov/heasd/documents/EPA%20PMF%205.0%20User%20Guide.pdf.
79. Paatero, P.; Hopke, P. K. *Chemom. Intell. Lab. Syst.* **2002**, *60*, 25–41.
80. Kim, E.; Hopke, P. K.; Edgerton, E. S. *Atmos. Environ.* **2003**, *37*, 5009–5021.
81. Begum, B. A.; Hopke, P. K.; Zhao, W. *Environ. Sci. Technol.* **2005**, *55*, 227–240.
82. Qin, Y; Kim, E; Hopke, P. K. *Atmos. Environ.* **2006**, *40*, S312–S332.
83. Zhou, L.; Hopke, P. K.; Zhao, W. *J. Air Waste Manage. Assoc.* **2009**, *59*, 321–331.
84. Hopke, P. K.; Ramadan, Z.; Paatero, P.; Norris, G.; Landis, M.; Williams, R.; Lewis, C. W. *Atmos. Environ.* **2003**, *37*, 3289–3302.
85. Zhao, W.; Hopke, P. K.; Norris, G.; Williams, R.; P. Paatero, P. *Atmos. Environ.* **2006**, *40*, 3788–3801.
86. Zhao, W.; Hopke, P. K.; Gelfand, E. W.; Rabinovitch, N. *Atmos. Environ.* **2007**, *41*, 4084–4096.
87. Lioy, P. J.; Zelenka, M. P.; Cheng, M. D.; Reiss, N. M.; Wilson, W. E. *Atmos. Environ.* **1989**, *23*, 239–254.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

88. Zhou, L.; Hopke, P. K.; Paatero, P.; Ondov, J. M.; Pancras, J. P.; Penney, N. J.; Davidson, C. I. *Atmos. Environ.* **2004**, *38*, 4909–4920.
89. Ogulei, D.; Hopke, P. K.; Paatero, P.; Park, S.-S.; Ondov, J. M. *Atmos. Environ.* **2005**, *39*, 3751–3762.
90. Liao, H.-T.; Kuo, C.-P.; Hopke, P. K.; Wu, C.-F. *Aerosol Air Qual. Res.* **2013**, *13*, 1253–1262.
91. Paatero, P. A.; Hopke, P. K.; Hoppenstock, J.; Eberly, S. *Environ. Sci. Technol.* **2003**, *37*, 2460–2476.
92. Chueinta, W.; Hopke, P. K.; Paatero, P. *Environ. Sci. Technol.* **2004**, *38*, 544–554.
93. Raabe, O. G.; Braaten, D. A.; Axelbaum, R. L.; Teague, S. V.; Cahill, T. A. *J. Aerosol Sci.* **1988**, *19*, 183–195.
94. Knochel, A. *Basic principles of XRF with synchrotron radiation, 2nd International Workshop on XRF and PIXE Applications in Life Science, Capri, Italy*; World Scientific Publishing Co.: Singapore, 29–30 June, 1989.
95. Dodd, J. A.; Ondov, J. M.; Tuncel, G.; Dzubay, T. G.; Stevens, R. K. *Environ. Sci. Technol.* **1991**, *25*, 890–903.
96. Burdick, D. S. *Chemom. Intell. Lab. Syst.* **1995**, *28*, 229–237.
97. Kiers, H. A. L. *J. Chemom.* **2000**, *14*, 105–122.
98. Pere-Trepat, E.; Hopke, P. K.; P. Paatero, P. *Atmos. Environ.* **2007**, *41*, 5921–5933.
99. Li, N.; Hopke, P. K.; Kumar, P.; Cliff, S. S.; Zhao, Y.; Navasca, C. *Chemom. Intell. Lab. Syst.* **2013**, *129*, 15–20.
100. Paatero, P.; Eberly, S.; Brown, S. G.; Norris, G. A. *Atmos. Meas. Tech.* **2014**, *7*, 781–797.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch006

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

Chapter 7

Hierarchical Classification Modeling of Watershed Data by Chemical Signatures

Steven D. Brown* **and Liyuan Chen**

Department of Chemistry and Biochemistry, University of Delaware, Brown Laboratory, 163 The Green, Newark, Delaware 19716, United States
***E-mail: sdb@udel.edu**

Complex data with many objects and classes may benefit from the use of a hierarchical class modeling approach in which samples receive more than one class label. A hierarchical model employing multiple class labels is better suited to making use of class relationships in the data as compared to traditional "flat" modeling methods. However, hierarchical modeling requires a number of choices that make the data analysis much more complex than a traditional classification. This chapter introduces concepts from hierarchical modeling of complex data with multi-label class ontologies, and considers what choices must be made to establish class labels in complex data and build a hierarchical classifier. An example is provided to show the details of the methodology in modeling hierarchical geospatial data.

1. Introduction

Traditional classification, in chemometrics or elsewhere, consists of developing a rule for assignment of class labels to two classes. The fundamental assumption is that each sample belongs to only a single class characterizing its semantics, so a classification step is used to discover each new sample's semantic meaning. The classification is focused on discovery of the single rule that separates the two classes. Subsequent to that assignment, the unique semantics of new samples are assigned by application of the rule. This approach has been

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

heavily used since Fisher's early work and was an area where Bruce Kowalski published extensively early in his career. The field of chemometrics became strongly associated with what was then called "pattern recognition," in large part because of early machine-learning papers by Kowalski and Bender (*1–3*), by Kowalski and his students (*4*, *5*), and by Isenhour and his students (*6*, *7*).

More recent work in classification in chemometrics has often been conducted with partial-least squares discriminant analysis, a version of the two-class Fisher discriminant in which the boundary is established by partial least squares regression (*8*) which permits work with data having more variables than samples and having more than two class labels. This partial least squares discriminant (PLSD) classifier is often used in one form or another in multi-label classification in chemometrics (*9–12*).

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

As the size and complexity of data has increased, it is necessary to perform classifications where the number of variables is much larger than the number of samples *and* where there are complex relationships among a large number of class labels. Simply assigning one class label to each group of data, as is done in both traditional classification and in the more modern PLSD analyses, is often no longer adequate, as it ignores the relationships that exist among classes and discards an important source of information about the set of measurements being modeled; it is increasingly common that real-world samples may be regarded as having multiple semantic meanings and therefore may be defined by multiple class labels, better reflecting their place in a complex data set.

Multi-label classification, in which each sample is associated with *multiple* class labels, each with separate semantic meaning, is new to the field of chemometrics but has been an active area of research in machine learning for several years (*13–15*). Multi-label modeling involving multiple labels using an hierarchical class taxonomy should offer advantages when close relationships between classes exist, e.g., in classifying gene or protein function, or in environmental studies where the classes define regions of space or time. However, the nature of the class taxonomy and the multi-label nature of the data require a modeling approach that differs from that used in traditional classification analysis used in chemometrics. Studies on complex data show that treating complex data as having multiple class labels has real advantages in performance sufficient to justify the additional work and complexity of the data analysis (*16*).

This chapter considers some issues involved in multi-label modeling and overviews some results from multi-label modeling applied to classification of water from different regions of the United States by using chemical and isotopic signatures of the water samples.

2. Structure in Class Labels

2.1. Hierarchical Taxonomy and Relationships in Class Labels

Establishing multiple class labels for a set of data requires a system by which the structure of class labels is specified. This system, known as the taxonomy of the labels, reflects interrelations in classes, in particular the correlation of class label

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

information. Much of the work in machine learning suggests that large amounts of related class information are often better structured in a hierarchical taxonomy. This taxonomy is defined over the partially ordered set (C, $\prec$), where C is a finite set of class labels that enumerates all class concepts in the data domain, and the relation $\prec$ represents an "is-a" relationship, as defined by the following rules:

1. The largest element $R \in C$ is the root of the hierarchical taxonomy.
2. $\forall\ c_i, c_j, \in C$, if $c_i \prec c_j$ then $c_j \not\prec c_i$
3. $\forall\ c_i \in C$, $c_i \not\prec c_i$
4. $\forall\ c_i, c_j, c_k \in C$, $c_i \prec c_j$ and $c_j \prec c_k$ imply $c_i \prec c_k$

These rules define a tree-structured class taxonomy, but they are also suited to directed acyclic graph (DAG) class taxonomies as well. Note that the rules defining the "is-a" relation create an *asymmetric* (by rule 3: e.g., all cats have 4 feet, but not all animals with 4 feet are cats) and *transitive* (by rule 4: e.g., all cats are mammals, and all mammals are animals, therefore all cats are animals) hierarchical relationship among classes. Such a relationship among classes is usually expressed as a phylogenic tree. With the "is-a" relationship, a classification done using this class taxonomy is inherently multi-class, because there are more than two class labels defining the data, and it also becomes inherently multi-label, because each object belongs not only to a single class but also to all of that class's ancestor classes, and any sample defined by one of these tree-structured taxonomies will naturally be associated with more than one class label to indicate its place in the taxonomy.

2.2. Multi-Label Hierarchical Classification

Because the taxonomy defined above is graphical, it is convenient to represent a class by a node. The class node a graphical representation of a single class, defined by a node label. The set of four rules given above in Section 2.1 permit class taxonomies in which the class nodes each have either a single parent or multiple parents; that is, the class represented by the node has a relationship with classes immediately above it in the tree. Taxonomies with class nodes having only single parents have an inverted tree structure, with a "root" node at the top and terminal "leaf" nodes at the bottom of the graph. Taxonomies where some or all nodes have multiple parents are described by a directed acyclic graph (DAG). Examples of a tree-structured taxonomy and a DAG-structured taxonomy are shown in Figure 1 below. The connections highlighted in blue form loops in the hierarchy that distinguish a DAG ontology; for example, there are 2 parents (A_{12} and A_{22}) for one of the nodes (A_{222}) there.

Class nodes are specified by taxonomy, by level, and by label. Most often, the structure of the class taxonomy is made clear by a diagram of the class nodes showing the ancestry of each class node.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

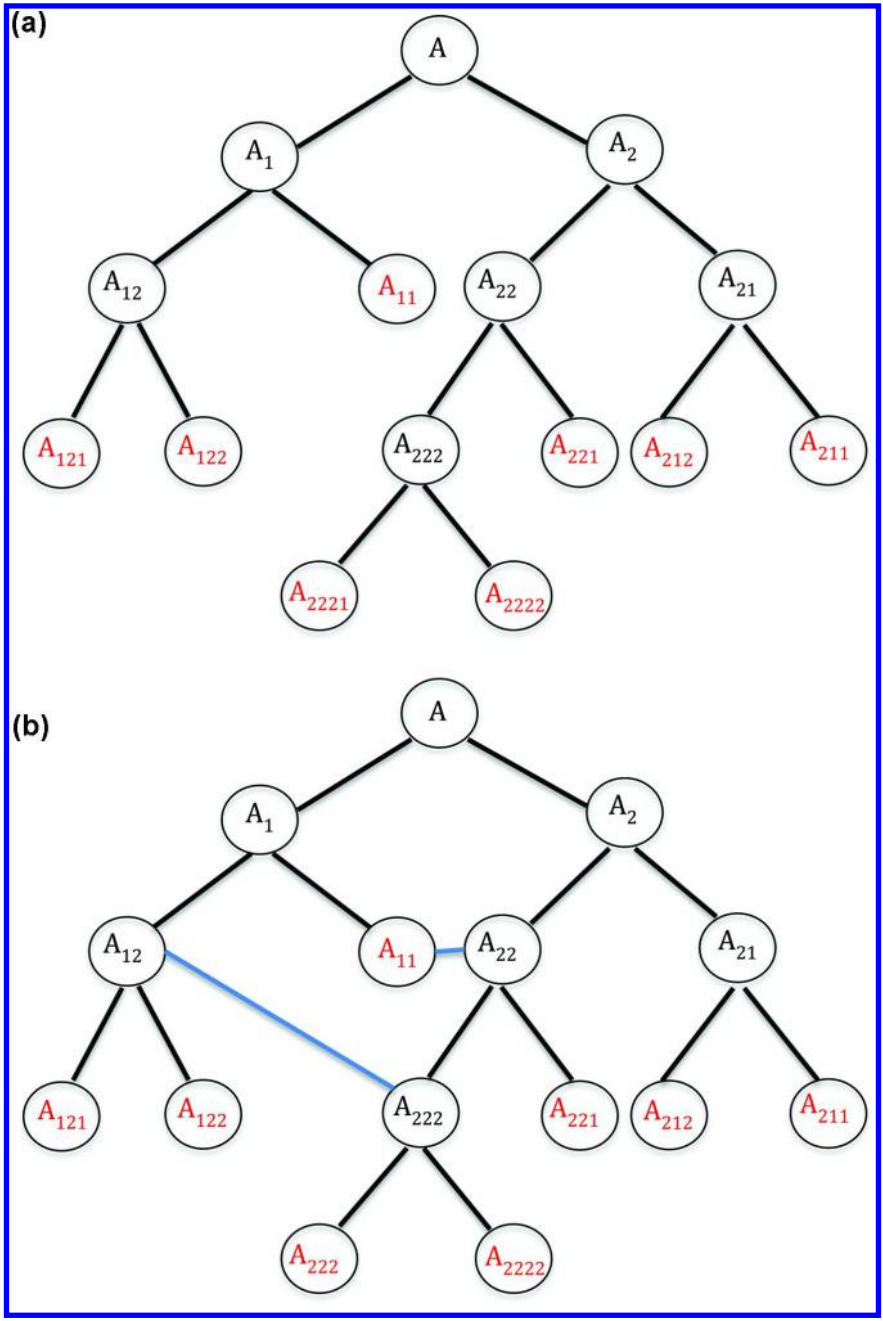

Figure 1. a: Tree-structured taxonomy. Nodes labeled in red are terminal nodes. b: Directed acyclic graph (DAG) taxonomy. Terminal nodes are labeled in red. The two connections shown in blue make this an acyclic graph.

Next, the nature of the class labels must be specified. The specification must distinguish between multi-label data modeled by a set of class nodes all located at one level (the top level) and multi-label data modeled by a hierarchical class node structure like that of Figure 1. A class taxonomy with all nodes at the same level in the hierarchy is called "flat" because in that structure, any class node hierarchy is collapsed to a single level, and all node ancestry is ignored. Assignments of class labels making use of this class taxonomy are therefore called "flat" classifications. A flat class structure has a single class identity as a label: e.g. 1, 2, 3, …, and while there may be multiple class labels describing the data, class node identities are each assigned to only one class label. These flat, multi-class classifications can be regarded as a special case of multi-level, multi-label classifications. Similarly, a two-class, flat classification is also a special case of the hierarchical, multi-class, multi-label case (*13*).

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

2.3. Algorithms for Hierarchical Classification

Almost all classifications performed in chemometrics have been done with flat data class taxonomy, using two or more class labels. Learning the classification rules for the 2-class, special case that encompasses the bulk of the applications of classification is by now well-established, and a wide range of algorithms are available (*17*). Similarly, learning classification rules for multi-label data assigned a flat taxonomy is also well-studied (*18*).

2.3.1. Algorithmic Strategy

In contrast to algorithms suited to flat classification, however, algorithms for learning the rules for a multi-label, multi-class classification are still being actively explored (*14*, *15*). When a class taxonomy defines multiple layers of nodes, it is often useful to set the class labels to reflect that structure. The flexibility of the structure poses a challenge to developing approaches to multi-label learning, but even more difficult is the fact that the possible number of possible label sets grows exponentially as the number of possible class labels increases. To deal with the huge number of possible labels, it is necessary to exploit dependencies and correlations among *labels* in the data if possible to discover the underlying taxonomy of the class labels (*14*).

One way of taking advantage of these dependencies is to deal with the classifications in a label - by- label fashion, ignoring the other labels, in effect decomposing the classification problem into a series of independent binary classifications, one per label. This "first-order" strategy (*14*, *19*) is conceptually simple, but it has the disadvantage of ignoring any label-label correlations present in the data, potentially biasing the classification model. A way of capturing these label relations is to deal with pairs of labels, possibly by ranking relevant and irrelevant labels or by examining interactions between pairs of labels. This "second-order" approach (*14*, *20*) generalizes well, but other interactions in the data may not be captured by this strategy. An extension of this approach might consider all other labels' influence on a label or might look for correlations and

dependencies among random subsets of the labels. These "higher-level" strategies (*14*, *21*) are much more computationally demanding, not to mention increasingly problematic to implement as the number of labels increases. A last option is to encode the relations between labels by setting up a particular taxonomy based on the semantics of the label set in the data. This option can produce strong classifier performance, since by specifying the taxonomy, all interactions are specified, but using a higher-level strategy of this sort usually requires *a priori* knowledge of the data structure beyond that provided by the features and any incomplete knowledge of the structure results in a bias.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

2.3.2. Class Labels

Assigning class labels in a hierarchy involves assignment of level as well as class within that level. In Figure 1 the nodes are labeled by level; thus, a node at the first (top) level has a single label (e.g., 1) while a node at the 3rd level has three labels (e.g., 1.1.2). In the tree-structured hierarchy, this multiple label identifies the ancestors of the node: the node labeled 1.1.2 indicates that this is a node (2) at the third level, but the label also tells us that this node is associated with node 1 at second the level and node 1 at the first level, so this node can be regarded as belonging to class 2 at the third level, but also to class 1 at the second level and class 1 at the first level, giving each sample of this class 3 labels, and reflecting its place in the hierarchy defining the data and the classes in the data. Multiple labels are needed because objects assigned to this node can be regarded as belonging simultaneously to multiple classes present in the data. In a multi-label classification, the aim is to assign all of the labels of the data, either sequentially or, if possible, *simultaneously*. A multi-label classification requires that the labeling capture the classes and all class hierarchy. It has the benefit of using all data for all labeling, but both hierarchy and labels for the entire data set must be determined from those data.

2.4. Classification Metrics

The hierarchy of classes is ultimately set by the performance of the multi-label classifier on a set of data with known class labels. In traditional, "flat" supervised learning, the usual metrics for deciding on the location of the class boundaries is often based on the overall classification accuracy, or sometimes on other common classification metrics such as the F-measure or the area under the Receiver Operating Characteristic (ROC) curve (AUC). For a multi-label classification, these metrics are not especially well-suited, and other evaluation metrics have been proposed (*14*, *15*). These fall into two groups: example-based metrics and label-based metrics. Example-based metrics use the results from each test sample to return a mean value of the metric over the test set. The common example-based metrics for multi-label classification are:

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

$$Accuracy = \frac{1}{p}\sum_{i=1}^{p}\frac{\left|y_i \cap \hat{y}_i\right|}{\left|y_i \cup \hat{y}_i\right|} \quad (1)$$

$$Precision = \frac{1}{p}\sum_{i=1}^{p}\frac{\left|y_i \cap \hat{y}_i\right|}{\left|\hat{y}_i\right|} \quad (2)$$

$$Recall = \frac{1}{p}\sum_{i=1}^{p}\frac{\left|y_i \cap \hat{y}_i\right|}{\left|y_i\right|} \quad (3)$$

where there are p test samples and where y_i and $\hat{y}_i$ are the true and predicted labels for sample i, respectively.

A label-based metric assesses the classifier performance on each of the q class labels separately, then returns the metric averaged over all class labels. For the j^{th} class label y_j, the binary performance of the classifier on this label is

$$\begin{aligned} TP_j &= \left|\left\{x_i y_j \in Y_i \wedge y_j \in \hat{y}_i, 1 \le i \le p\right\}\right| \\ FP_j &= \left|\left\{x_i y_j \notin Y_i \wedge y_j \in \hat{y}_i, 1 \le i \le p\right\}\right| \\ TN_j &= \left|\left\{x_i y_j \notin Y_i \wedge y_j \notin \hat{y}_i, 1 \le i \le p\right\}\right| \\ FN_j &= \left|\left\{x_i y_j \in Y_i \wedge y_j \notin \hat{y}_i, 1 \le i \le p\right\}\right| \end{aligned} \quad (4)$$

where TP_j are true positives, FP_j are false positives, TN_j are true negatives, and FN_j are false negatives for data with label j with respect to the true value y_i. If we take M as some binary classification metric (e.g., accuracy), the label-based classification metrics are defined as

$$\begin{aligned} M_{macro} &= \frac{1}{q}\sum_{j=1}^{q} M(TP_j, FP_j, TN_j, FN_j) \\ M_{micro} &= M\left(\sum_{j=1}^{q} TP_j, \sum_{j=1}^{q} FP_j, \sum_{j=1}^{q} TN_j, \sum_{j=1}^{q} FN_j\right) \end{aligned} \quad (5)$$

As Zhang and Zhou (*14*) note, label-based metrics based on macro- and micro-averaging assume equal weights for labels and samples, respectively, but both differ from the example-based metric described above.

2.5. Decision Trees

For the same hierarchy, there are two other ways of assigning class labels. The class labels can be assigned by use of a decision tree, or by a sequence of separate classifications performed at each level. In a decision tree, a sequence of decisions are made, often binary, to reach a set of terminal "leaf" nodes, where a decision is made as to class label, usually by majority vote of the objects

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

assigned to the node. While the decision taken at each of the intermediate nodes could involve assignment of class labels, they are generally not assigned at these intermediate nodes in the hierarchy. Instead, decision trees make use of single-label classification at the terminal "leaf" nodes of the decision tree, but information on the path though the tree and on any ancestor classes are not of interest, no matter how many layers there are in the hierarchy. The node indicated by label A_{12} in Figure 1a is a leaf node, and in a decision tree using the hierarchy of Figure 1a, the decisions made at higher levels in the hierarchy merely identify the path to the terminal node under a specified decision metric, so each of the class nodes in a decision tree can be described by a single label. Note that it is possible, even common, to assign the same class label to *multiple* leaf nodes in a decision tree, unlike class label assignments made in a tree-structured hierarchy. The metric for deciding a partition of the data to increase the class purity is often either the Gini index or the information as measured by the cross-entropy of the data. Neither is optimal for determining rules for separating a set of class labels, and it is usual to prune the classification tree once constructed by using a cross-validation step. That, too, is seldom optimal, in part because of the choice of the single variables used to define the partitions defining the decision paths is somewhat arbitrary, and a criterion must be selected for deciding when to stop the recursive partitioning of any impure space containing objects with more than one class label. The disadvantages of decision tree-based approaches are that the method is essentially univariate; the partitions are all based on single variables, and a "weak" classifier results. Recent practice in many fields where decision trees are used has been to use collections of these cross-validated trees, starting each one from a different (random) partitioning choice – a random forest. Using a random forest helps here because the combined set of classifiers somewhat compensates for weakness in the classifiers. These so-called "bagged" methods can be expected to outperform any single method, as Breiman has demonstrated (*22*). It is also possible to weight classifiers according to their performance of a practice test set, up-weighting ones that perform well and down-weighting ones that perform poorly; this "boosting", when combined with bagging, can often lead to classifiers with very strong predictive performance even though the individual classifiers are not very strong. Advantages of the decision tree include interpretability, ability to use ordinal and other data, and a built-in resistance to missing data; the random forest gives up most of its interpretability for improved performance as compared to a decision tree.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

2.6. Sequential Multi-Label Classification

A hierarchy like that shown in the Figure 1a can also be regarded as describing a *sequence* of binary or higher classifications done on subsets of the data. Like the decision tree, a decision is made at each branch in the hierarchy, but unlike the decision tree classifier, multivariate class rules are developed at each branch node leading to assignment of explicit class labels, and an effort is made to achieve optimal or nearly-optimal partitioning of data as measured by the usual classification metrics. Depending on the specific classifiers used, the data in the hierarchy could be assigned more than two class labels at each branch node

using this approach, something that is more difficult to implement effectively in a decision tree. The end result is a multi-label classification with information on class ancestry, but unlike a flat, multi-label classification, each class label is assigned *independently* of other nodes at the same level of the hierarchy, and each decision is made not on the entire data, but on only a portion of it because, like a decision tree, the sequence of decisions recursively partitions the data into increasingly smaller subsets. The assignment of class labels at each stage results in samples being associated with multiple class labels. In a sequential classification, cross-validation may be used to optimize rules produced at each stage, but no pruning is done.

Where these approaches differ most is in the establishment of class labeling rules. In each case, the class label structure must be established prior to determining the rules for assigning class labels to new data. The classification rules depend both on the methodology used and on the optimization metric. Zhou and Zhang (*14*) review recent research on multi-label classification and point out some of the strengths and weaknesses of various algorithmic approaches to multi-label classification. In this report, only two will be mentioned.

The most common way at present to perform a multi-label classification is to use a conventional decision tree, modified to make (usually) binary decisions and assign levels at each level of the tree, as discussed above. This method, first reported by Clare and King (*16*) has become the *de facto* standard against which other methods are compared, despite its shortcomings as a first-order approach (*14*). In this approach, multiple labels in the leaves are allowed and the informational entropy used to calculate information gain (IG) in the standard C4.5 algorithm of Quinlan (*23*)

$$IG(S,A) = Info(S) - Info_A(S) = Entropy(S) - \sum_a \frac{N_{S_a}}{N_S} Entropy(S_a) \tag{6}$$

is modified slightly, from

$$Entropy = -\sum_{i=1}^{N_C} P_S(c_i) \log P_S(c_i) \tag{7}$$

to

$$Entropy = -\sum_{i=1}^{N_C} P_S(c_i) \log P_S(c_i) + (1 - P_S(c_i)) \log(1 - P_S(c_i))) \tag{8}$$

for a set of N_S samples in training set S, some N_{Sa} of which have value a for attribute A, with each class c_i represented by percentage $P_S(c_i)$ for each class c_i of the N_C classes.

With this modification, the binary splits that result from the application of the C4.5 algorithm consider not only the probability of membership in class i but also the probability of other classes. This approach has the advantage of simplicity and interpretability, but has the disadvantage of retaining C4.5's focus on binary decisions. Gao, et al. have extended the binary, single-feature approach of Clare and King to multi-label classification (*24*).

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

Multi-label C4.5 also retains the classifier's relatively weak classification performance, as it bases classification on a single variable, even with multi-label data (*22*, *24*, *25*). The alternative is to perform the sequential classification discussed above using a sequence of classification steps. Any multi-label, multivariate classifier algorithm can be used for this task, but this paper focuses on a model-based-classifier optimized with the expectation-maximization (EM) algorithm (*26*) using the naïve Bayes classifier (*27*). Model-based classification is closely related to model-based clustering (*28*), which was used to define clusters in this work, so it is convenient to make use of this classification approach. In most comparisons of flat classifiers, the naïve Bayes classifier usually ranks as one of the strongest simple classifiers (*25*, *27*). The naïve Bayes classifier outputs the probability for each sample belonging to each class. The probabilistic estimate of class label permits an estimate of the classification uncertainty for the training data as well as for new samples.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

2.7. Gaining Advantage from Multi-Label Classification

From the discussion above, it should be clear that the training and use of a multi-label classifier is likely to be much more demanding than a training and using a traditional, flat classifier, but studies have shown that for many data sets in text analysis (*13–15*) and even a few in biology and in chemistry, there is an advantage to using the multi-label classifier [e.g., (*29–32*)]. The biggest gains in classification performance are to be expected from data where there are many classes or where the classes show some sort of distance-based or other relationship, as might be found in genomics or in a geospatial set. In both of these situations, the class label ontology is often structured. However, it has been suggested (*13*) that any situation where there are both numerous classes and a large number of objects to be assigned class labels should benefit from a multi-label classification. Increasingly, this is the case in chemometric data, as data become cheaper to collect and interest increases in the fine detail present in label structure.

2.8. Modeling the USGS Surface Water Data

The example shown in this paper concerns the modeling of a dataset composed of water analyses made over a period of about four years on 2685 water samples collected at 328 river sampling sites distributed over the continental USA, and Alaska, the US Virgin Islands and Puerto Rico, and Hawaii. The longitude and latitude of each sample location was recorded, but no class label information was available. A total of 23 trace elements and 2 isotopic ratios were measured at different times during the period 1984-1987. The 23 trace elements were Aluminum (Al), Barium (Ba), Calcium (Ca), Magnesium (Mg), Potassium (K), Chlorine, measured as chloride (Cl), Iron (Fe), Manganese (Mn), Strontium (Sr), Silicon (Si), Zinc (Zn), Sulfur, measured as sulfate (S), Copper (Cu), Nickel (Ni), Fluorine, measured as fluoride (F), Lithium (Li), Beryllium (Be), Cadmium (Cd), Chromium (Cr), Cobalt (Co), Silver (Ag), Vanadium (V) and Selenium (Se). The two isotopic ratios measured are ^{1}H:^{2}H and ^{16}O:^{18}O. Elemental analyses of these water samples were performed using USGS-certified methods using

inductively-coupled plasma atomic emission spectrometry and mass spectrometry. Isotope ratios were obtained by isotope ratio mass spectrometry, as described in (*33*). The distribution of sampling sites is shown in Figure 2.

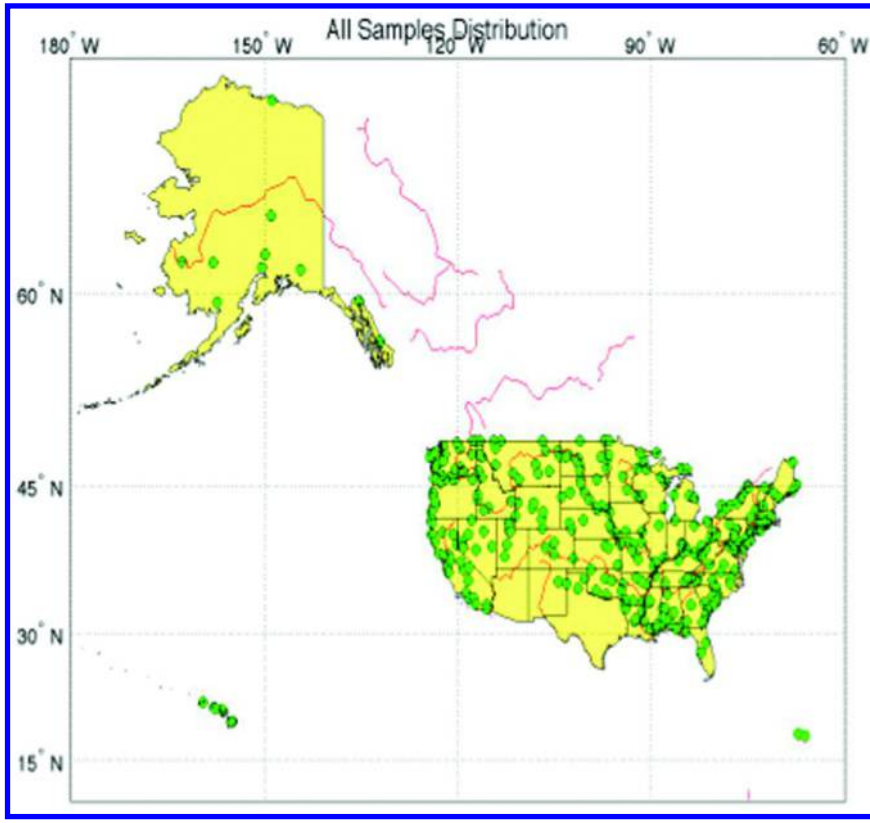

Figure 2. Geographical distribution of sampling sites (in green) for the water study. Major rivers are depicted in red.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

2.8.1. Missing Data Imputation and Preliminary Examination

Environmental studies commonly have missing data, where a particular analysis was not run, and censored data, in which the values observed for a particular analysis were at or below the detection limit. This data set had many missing or censored data, a total of 34.3% of the measurements in the set. As Figure 3 shows, while almost all of the variables had some missing or censored data, several variables were dominated by missing or censored measurements. Each of the last 7 trace elements listed above has at least 80% missing values, and was removed from further consideration. Multiple imputation based on a version of the EM algorithm (*26*) was applied to data that had been log transformed to make variables approximately normally-distributed.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

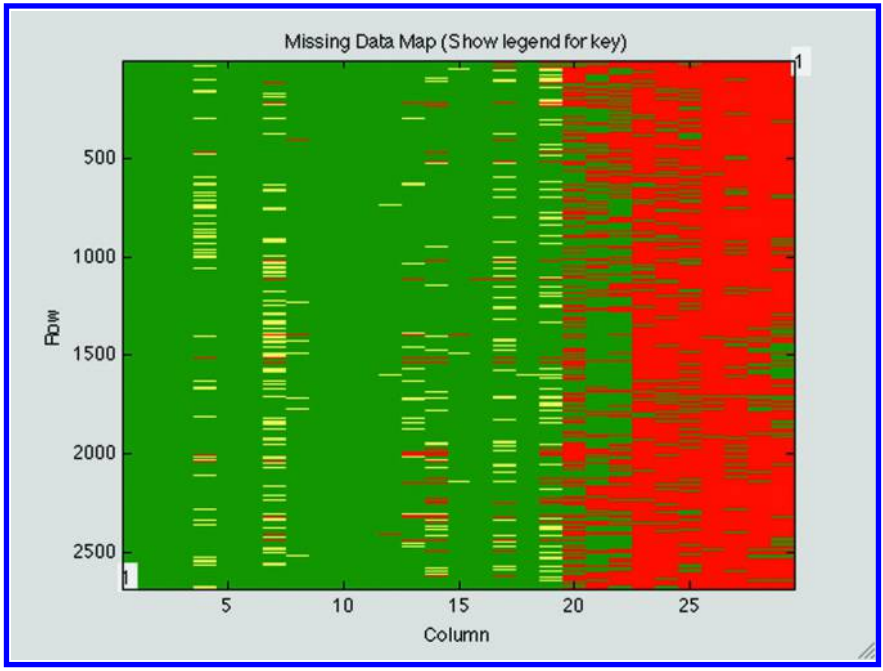

Figure 3. Missing and censored data in the water data. Green indicates no missing entries. Yellow indicates 1 missing entry; red indicates more than 2 missing entries.

The distributions of some of the chemical measurements available for the data set after imputation are summarized as Boxplots in Figure 4.

As seen in Figure 4, apart from measurements of the isotopic ratio, most of the other measurements available for the modeling had a very wide range and a high standard deviation compared with their mean, indicating a large spread in the data. As the Figure shows, there is also right-skewness in much of the data, even after removal of outliers. Considering the extremely broad study area and the time over which the sites were sampled to collect groundwater data, a substantial part of the scatter in the data is likely from seasonal variation and changes in stream flow as well as from the geologic variation that we seek to relate to location. Therefore, given the redundant information and possible undesirable effects brought into the modeling by seasonal variations in the chemical signatures, careful selection of the chemical measurements had to be made to capture as much spatial variation as possible while retaining as little of the seasonal variation as possible so that the spatial variations in chemical measurements chosen at each transitional region overwhelm the seasonal and other temporal effects occurring at each individual sampling site in the same region.

Approximately 15% of the training samples were regarded as outliers and were removed during model construction; the outlier samples were generally ones that were not clustered into the same region when performing majority voting at each sampling site.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

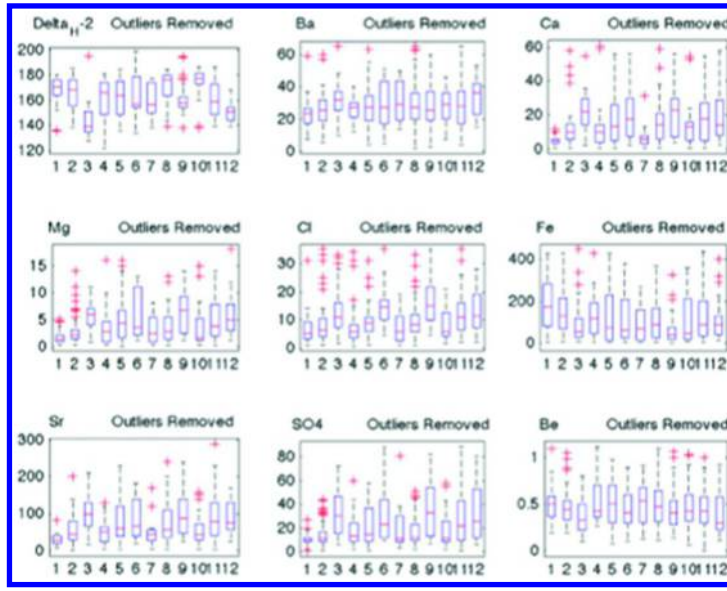

Figure 4. Exploratory data analysis of USGS water data variables for the Southeastern USA.

2.8.2. Dealing with Temporal Effects in the Data

The use of a multi-label classifier requires the ontology of the class labels. Most studies with multi-label data involve an analysis of text, where it is possible to pre-establish a hierarchy of classes from consideration of grammar and usage. Finding the sets of labels for a multi-label set where the ontology has not been pre-established is not routine, and there has not been much systematic research on methods to discover the ontology. For a set of data in which the class ontology is based on distance between class labels, an approach based on distance may be useful, for example the well-known distance-based hierarchical clustering that is common in exploratory analysis or by other clustering methods may give an estimate of the class ontology.

To classify samples on the basis of location, it is important to retain analyte measurements that correlate well to location but to discard those that do not to reduce the dimensionality of the modeling. Water levels at the sampling sites vary with the season, the weather, and the time of day, producing temporal "noise" which has a strong effect on analyte concentrations. While variations in the local and regional geology suggest that certain analyte concentrations should vary with location, the temporal changes in water levels result in analyte concentrations that may or may not have a connection to geology, and hence to the geographical location of the sample. Because samples are taken at fixed locations, but at different times and seasons, some temporal averaging of the data occurs, but because the sampling was neither consistent at all sites nor uniform in

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

timing or frequency, separating the temporal and geological effects on the data is challenging.

As an indication of the extent of the temporal effects, Figure 5 shows the variation in Mg and SO_4^{2-} for just one region (the Ohio valley region) given as a function of month over the three years of this study. Both fluctuate strongly by month and within the month; there is little in the way of an obvious trend in the fluctuations.

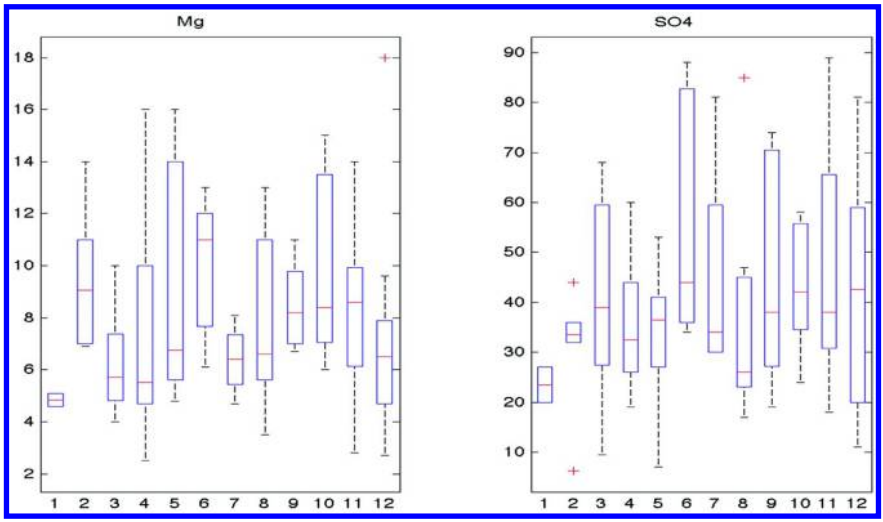

Figure 5. Variation of two of the analytes with season in the Ohio Valley region

In establishing class labels from the data by variable selection and clustering, several guiding principles were assumed:

1. Any classes found in the data should be compact: smaller, regional clusters are preferred over larger, diffusely defined clusters;
2. The class topology is hierarchical, with single parent class structure; and,
3. All variables used in establishing class labels must show a strong relationship with distance over the geographic region defined by the class.

In the work discussed here, model-based clustering on selected variables was used to identify clusters in the trace analysis data (*34*). Because analyte levels may correlate well to geography over some small sub-region of the whole but may correlate much less well over a larger geographic region, variables were selected for each step in the clustering, not globally. To minimize the effects of temporal variation in the data used for clustering, at each level of the class ontology, variograms (*35*)

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \left[z(x_i) - z(x_i + h) \right]^2 \quad (9)$$

were used to identify variables z most strongly related to distance h as measured at $N(h)$ pairs of samples x_i. Finding an increasing trend in γ(h) with distance h in the variogram does not identify the existence of non-random spatial dependence, and it is necessary to compare the experimental variogram to a variogram created for the same variable, but without systematic spatial information, by randomizing the spatial data. A permutation test (*36*) is used to exclude those variables without systematic spatial patterns. The test is performed by randomly permuting the values of the same variable of interest a large number of times, T, and a variogram is made for every random ordering. The permutations scramble any systematic spatial information present in that variable, and generate a randomly varying set of variograms for the permuted versions of z. In Figure 6, the first five permuted variograms for variable z are represented by dashed lines. The experimental variogram for the un-permuted variable z, indicated in bold in Figure 6, shows a clear, increasing trend, while the variograms for that same variable z calculated after random permutation of z to remove systematic spatial contributions are much flatter. These variograms represent the variable z without the systematic spatial dependence in z. The random permutations of the values of variable z define a distribution on $\gamma(h)$, from which critical points (for example, the 2.5th and 97.5th percentiles) of the $\gamma(h)$ values can be calculated for each separation distance h over the range considered for the modeling. The two limiting variograms defined by these critical points define the critical regions for a random distribution defined by the data in variable z; an experimental variogram falling outside the region defined by these limiting variograms thus has non-random spatial behavior over the distance examined at a specified confidence. The two boundary variograms defining the critical limits of the distribution are represented by two solid lines with squares in Figure 6. The size of geographic regions varied with level in the ontology of the classes, as did the variables identified by the variographic analysis.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

2.8.3. Finding Clusters in the Surface Water Data

Clustering of the data into a large number of distinct, compact clusters was not possible in a single step, in part because of variation in sampling density over the regions studied. Regions near the East and West coast, on the Missouri River, the Mississippi River and the Ohio River had high sampling site density, while other regions had low sampling site density or even no sampling sites.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

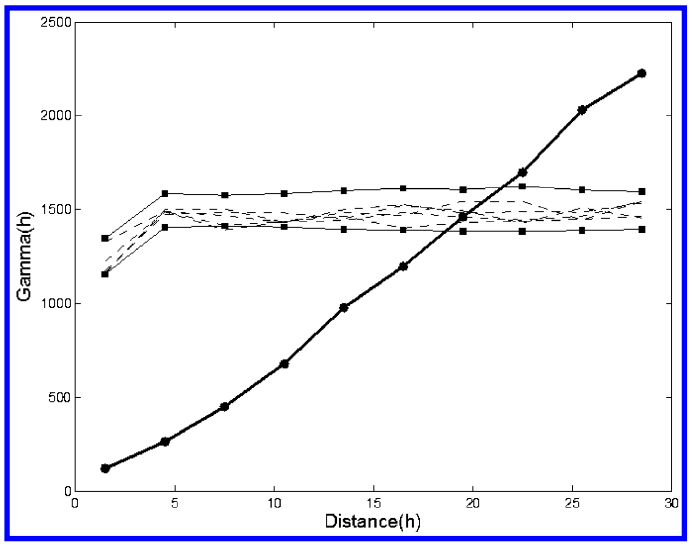

Figure 6. Feature identification of $^{1}H:^{2}H$ from its variogram. The bold, solid line shows the experimental variogram calculated over some region of the spatial data. The two solid lines with squares indicate the variograms for the 2.5th and 97.5th percentiles of the distribution of variograms obtained at each separation distance from the T random perturbations of location in $^{1}H:^{2}H$ data. Only five dashed lines, representing variograms obtained for $^{1}H:^{2}H$ for the first five random perturbations, are shown here for clarity. From reference (32), with permission.

Because the identification of clusters depended on identification of a suitable set of variables sensitive to distance and the set of variables used in the clustering was discovered by choosing samples from a region over which to test the variogram, the two tasks couple. Multiple passes of variable selection and clustering were needed at each level of the hierarchy to find self-consistent sets of compact clusters and distance-related variables. Unfortunately, there is no automated method to perform this simultaneous variable selection and clustering.

To obtain compact clusters, model-based clustering was used (*28*). Hierarchical clustering was used to help establish an initial guess needed to perform the model-based clustering under the Expectation-Maximization (EM) algorithm (*26*). A Gaussian distribution of data within the clusters was assumed, but the cluster size and shapes were permitted to vary as needed. In each clustering experiment, Markov-chain Monte Carlo methods were used with model-based clustering as optimized by the Bayes Information Criterion (BIC) to decide on the number of clusters found and the sample membership of each cluster (*34*). When samples collected at a fixed location but at different times failed to be classified into the same cluster, a majority voting rule was implemented so that all samples from the same geographical sampling location were forced to belong to the same region (cluster). The BIC was maximized in these experiments, identifying both cluster location and shape. Emphasis was placed on finding the best set of stable clusters. Generally, the Monte Carlo simulation was run for sufficient samples

to ensure that all clusters were stable and all cluster memberships were settled. Figure 7 summarizes the process.

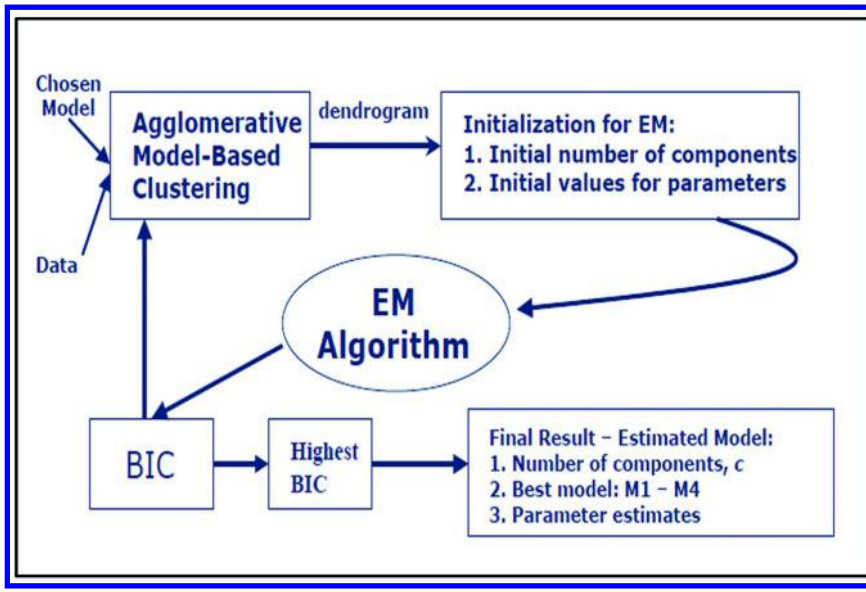

Figure 7. Clustering used to define hierarchical class structure. M1-M4 are different assumed shapes for clusters (34).

2.8.4. Estimating Uncertainty in Clustering

To ensure that clusters discovered in the data were reliable indicators of class, a new method to estimate the sample membership uncertainty during clustering was developed. The MCMC modeling implemented by the Gibbs sampling method can be used to test the uncertainty of membership of water samples from the USGS surface water data. Suppose the clustering results correspond to the watersheds, i.e., different clusters match different watersheds, and then the clustering membership uncertainty can be seen as the stability test of watersheds formed by samples collected at different sites.

To illustrate this process, consider water samples from three adjacent watersheds: Ohio, Mid Atlantic and South Atlantic Gulf. After variable selection using the variogram to find a set of variables that relates strongly to location, model-based clustering is performed on the samples. Three clusters result from the analysis.

These clusters are assigned labels from the mixture parameters that result from the model-based clustering analysis, as shown in Figure 8. However, since these water samples are assigned class labels only once, no assessment of stability or uncertainty is available for each sample. Due to the high variability of the surface water samples, both the location indicated by chemical measurements and the

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

stability of that pattern (watersheds) is desired to permit reliable prediction of new data.

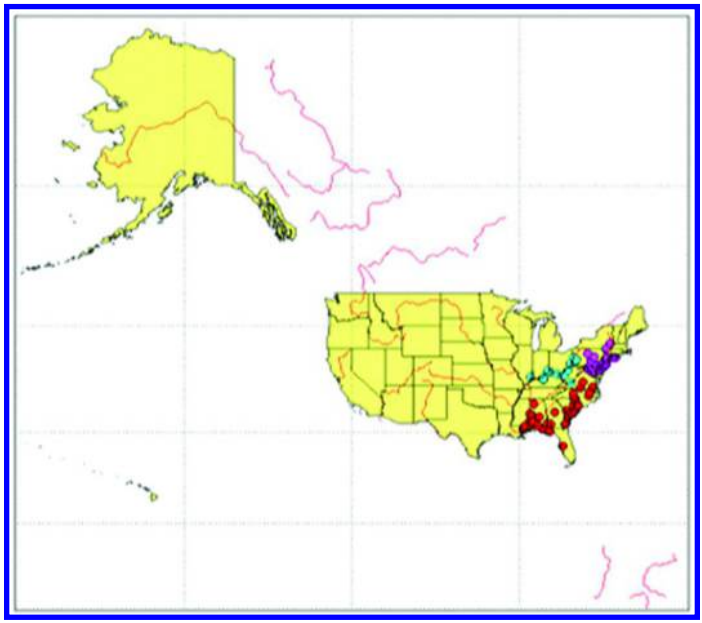

Figure 8. Clustering of three regions in southeastern USA

MCMC modeling implemented via Gibbs sampling can be used to estimate the clustering uncertainty. Through the Gibbs sampling, a joint distribution for every mixture parameter can be obtained instead of a single optimal value, and multiple mixture model can be built to fit the data after sampling from those joint distributions. In this way, the water samples can be clustered multiple times and an estimate of the stability of the clustering can be obtained.

The key in MCMC modeling is to create a Markov process whose stationary distribution is the joint posterior distribution of parameters, and then sample from the conditional posterior distribution of each parameter for long enough (using enough iterations) that the distribution of draws is close enough to a stationary distribution. For each cluster in the Gaussian mixture model, Gibbs sampling is used to estimate the model parameters for each cluster: means μ_k, covariance Σ_k, mixing coefficients τ_k and the classification vector $V = (\upsilon_1, ..., \upsilon_n)$, where υ_k implies that observation x_i is assigned to cluster k. After convergence of the MCMC simulation (which occurs after a sufficient number of burn in iterations), the samples are used as the Bayes estimate of the parameter, and if we continue sampling from the joint posterior distribution of each parameter for more iterations, the samples obtained can be regarded as the true Bayesian estimates of the parameters. The clustering uncertainty can then be calculated by clustering observations multiple times based on those Bayesian estimates (*34*).

In Figure 9 below, the iteration plots of the mixing coefficients τ_k obtained by Gibbs sampling for samples from Ohio, Mid Atlantic and South Atlantic Gulf

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

are shown. They give a good indication of the number of iterations needed for the burn-in period. The initial guess used for the mixing coefficients is 1/3 for every cluster, represented by a "*" sign in the Figures, indicating that each cluster is equally likely for these data. The convergence is almost immediate, and successive draws are independent.

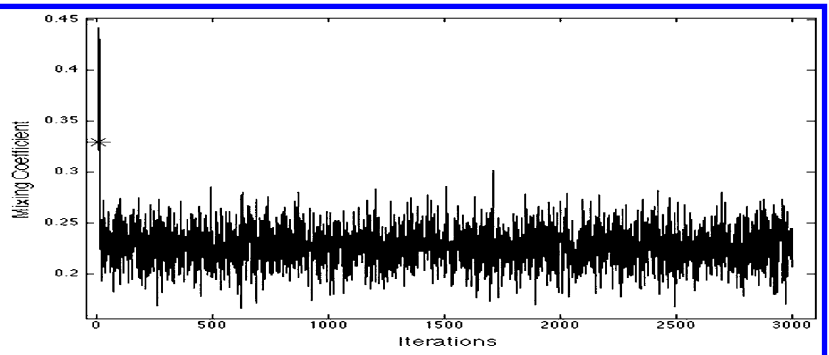

Figure 9a. Mixing coefficient sampling of Ohio samples

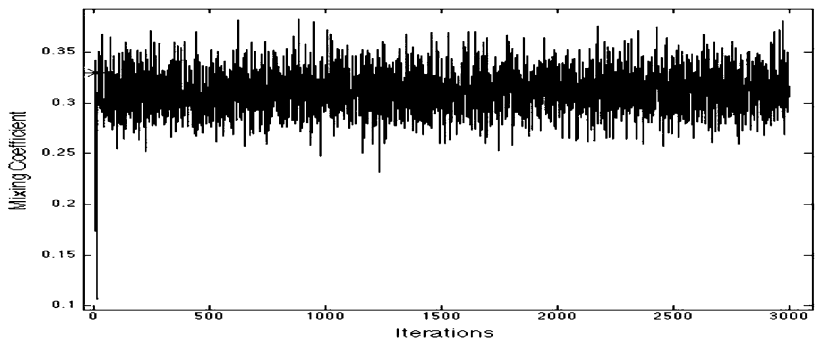

Figure 9b. Mixing coefficient sampling of Mid Atlantic samples

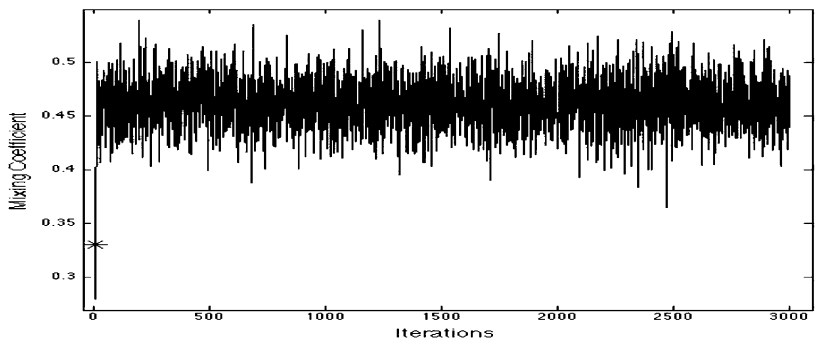

Figure 9c. Mixing coefficient sampling of South Atlantic Gulf samples

Figure 9. MCMC settling of the mixing coefficients in model-based clustering

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

Figure 10 shows the stabilization of the three cluster means μ_k in the MCMC simulation. At the beginning of Gibbs sampling, the starting value for cluster mean is the grand mean of the dataset consisting of the 3 sites, represented by the green dot in the center of the Figure. The cluster means calculated after each iteration are indicated by black "+" signs. In Figure above, only the first 50 iterations of Gibbs sampling for the cluster means are shown. The three colors correspond to the samples from the three different watersheds. The convergence of the means is again immediate, and the estimated mean for each cluster stabilizes in the center of each cluster.

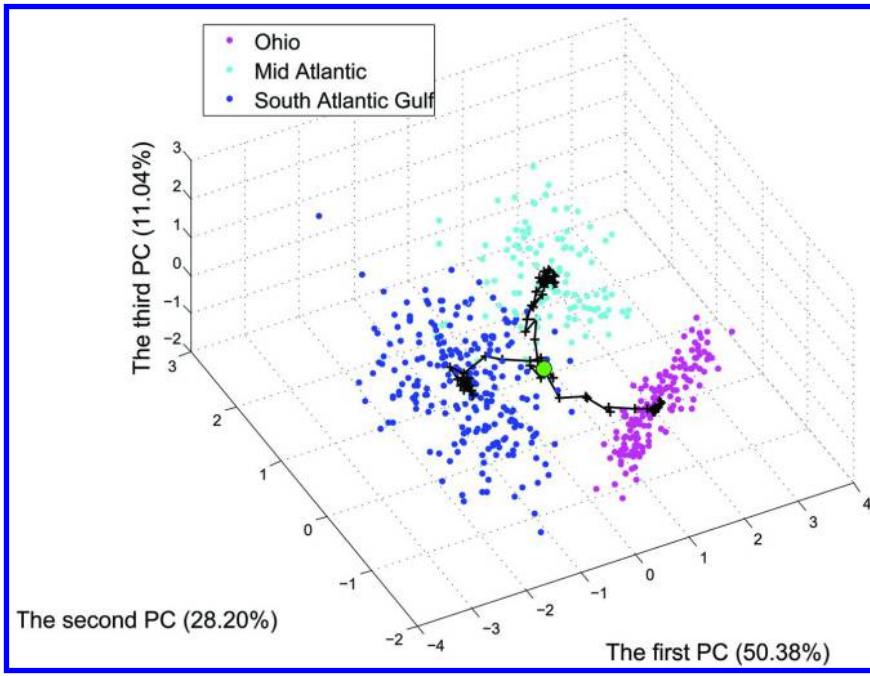

Figure 10. Cluster mean convergence of samples from three watersheds by MCMC via Gibbs sampling

During Gibbs sampling, 3000 iterations are run and the first 1000 iterations are discarded as "burn-in" of the simulation. The mixture parameters samples from the remaining 2000 iterations are regarded as the "true" estimates and used for classifying samples, to access sample clustering uncertainty and for testing class stability. In Figure 11, the stability of the clustering is summarized by samples with different signs. Black dots indicate samples with high stability, while square and circles represent samples with low stability. Most of the samples show a clustering uncertainty of less than 0.1, meaning that at least 90% of the iterations locate them in one, fixed cluster. The low uncertainty of samples produces stable clustering and stable classes. The clustering of the three sites used in this illustration is consistent with that found using an optimal set of mixture parameters from the EM algorithm. Samples from Ohio all have high clustering stability because they

are well separated from the other two clusters. There are only three samples with low cluster stability, all of which lie on the boundary of the Mid-Atlantic and South Atlantic Gulf groups. It is therefore not surprising that these samples are sometimes clustered in Ohio, and sometimes in the adjacent group depending on small changes in the cluster model parameters found. Figures 10 and 11, taken together, demonstrate that the site clustering pattern produced from the systematic identification of variables and clusters using the Gibbs sampling to estimate the clustering parameters is very stable, and most of the samples in the larger set showed a very low clustering uncertainty during 2000 Gibbs sampling runs.

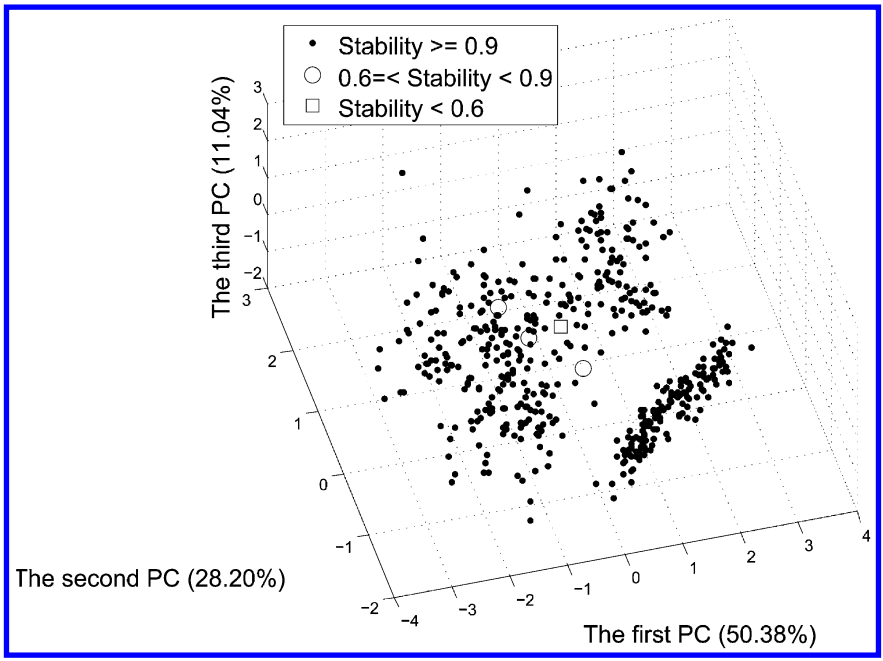

Figure 11. Clustering stability of samples from three watersheds by MCMC via Gibbs sampling

It is interesting that the three regional clusters identified in the Figure 8 above, a subset of the 18 terminal node clusters discovered in all, agreed well with the results from a smaller hydrogeological study on the Ohio, Mid-Atlantic and Gulf regions, where the authors also found three distinct regions of essentially the same shape (*36*).

2.8.5. Results of Clustering the USGS Data

The series of variable selection and clustering steps led to the set of 18 terminal node clusters shown in Figure 12. Because model-based clustering assigns a posterior probability of class membership rather than assigning

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

membership in a binary sense, some overlap in cluster membership was possible at each level in the hierarchy. To determine class labels, however, all samples were identified with the label of the cluster where they showed the largest posterior probability of membership. The compactness of the each of the terminal sub-regions depended largely on the sampling site density.

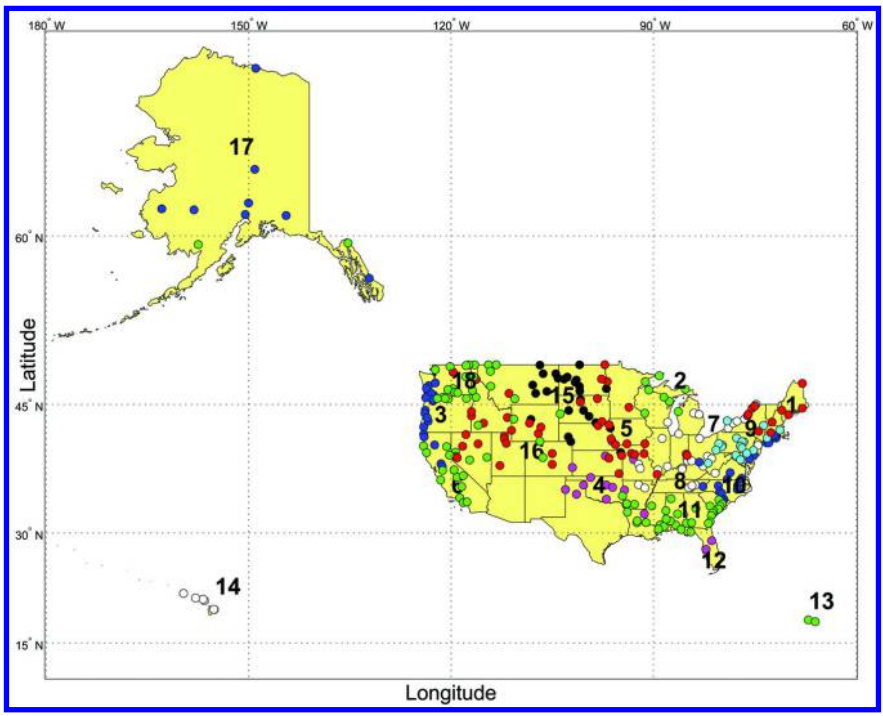

Figure 12. Terminal node clusters produced from clustering of chemical signatures of USGS surface water samples

The set of clusters discovered in this way from chemical signatures was compared with the USGS map of watersheds shown in Figure 13 (*37*). Most of the clusters defining geographic regions were well-separated from each other except for a few sites. Some of the terminal sub-regions discovered by the clustering, in particular those for regions on the West Coast of the USA, Florida and for New England, corresponded well to the published climate zones (*37*) for the United States. Given the sizable differences in the way that the regions were assigned, the similarity is notable.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

Figure 13. Watersheds identified by the US Geological Survey (33).

3. Construction of the Multi-Label, Hierarchical Model

Tree-structured hierarchical, multi-label modeling provides a general route to the estimation of location as well as an estimate of its associated uncertainty for multivariate spatial data, especially with a set of chemical measurements. Building a tree-structured hierarchical model is comprised of three steps.

3.1. Hierarchical Decomposition of the Region

The first step is to decompose some region A containing all of the spatial data into a number of regionally compact sub-regions (A_{11}, A_{221}, …A_{2221}). The decomposition is implemented in an hierarchical way; the spatial data in region A are first decomposed into a set of broad sub-regions (A_1 and A_2), and then each broad sub-region is decomposed in turn into a set of smaller geographic sub-regions, analogous to the decomposition of a data space done in recursive partitioning.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

The geospatially-related grouping in the data is assigned to the separate classes though the cluster analysis performed earlier, and the sequence of cluster analyses determines the ontology of the hierarchical classification model. Thus, spatial identification of a geographically distinct sub-region is achieved through a hierarchy of clustering steps in which samples from different geographic regions form different clusters. Both chemical measurements and geographical features can be used in clustering to decompose regions, though this work focused on geospatial analysis though chemical signatures alone. Unlike a spatial clustering approach that prefers samples from the same cluster to have similar geographical features (e.g., by requiring that they are close to each other spatially), the clustering done here in the formation of hierarchical regions allows samples from the same cluster (a geographic region) to be *separated in space* as long as they form distinct classes at the next level of the hierarchy. For example, samples with class label A_1 might separate into a set of possibly disjoint sub-classes A_{11}, A_{12} , A_{13} and A_{14}.

As shown in Figure 1a, the class model for the water samples can be interpreted as a set of hierarchical regions represented in a tree ontology of class labels, in which every node represents a region as well as a portion of spatial data belonging to that region. In this tree representation, the root node corresponds to the initial, complete geographical region spanning all of the spatial data, and the terminal nodes correspond to compact sub-regions where further spatial subdivision by clustering with a set of spatially relevant features was not successful. Once the clustering is accomplished, each parent node in the hierarchy also describes a transition between two regions by means of a classification based on a set of chemical measurements or geographical features that separate the two (child) sub-regions according to their labels.

3.2. Hierarchical Region Identification

Based on the tree representation already defined by clustering using the sets of spatially significant variables corresponding to each of the sets of hierarchical regions, each of the modeling steps in the tree structure is part of a hierarchical classification sequence, in which classifiers are built at each parent node (sub-region) of the tree by using samples known from the clustering to be of that sub-region as training samples. The aim of this step is to identify a smaller sub-region at every node and to eventually provide a set of distinct, geographically compact sub- regions to enable classification of a test sample with unknown location by means of the top-down sequence of binary classifications. The features used to build the classifiers are limited to the available chemical measurements because geographical features are not available for the unknown samples. At each node, the set of chemical measurements used for regional identification is the same as those used in regional decomposition step if geographical features were not used in the clustering to find sub-regions; otherwise, a variable selection procedure must be performed to find the best subset from among all chemical measurements to best distinguish the particular pair of sub-regions.

Depending on the classification algorithm used for identification of rules for assigning the sub-region labels, the uncertainty in class membership, which is

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

interpreted as one minus the probability of the test sample belonging to a class represented by one of the sub-regions, can be assessed in different ways: by resampling, by bootstrapping, or by computation of the posterior probability of class membership (*38–40*). With the hierarchical structure of the model, a test sample is passed by means of a sequence of classifications along a path from root node to a terminal node (the terminal sub-region). All classifiers along the path are built locally, to best distinguish different portions of data based on different sets of chemical measurements; as a consequence, the posterior probability that the test sample is classified into any parent node is independent from the posterior probability that it is then classified into either of the child nodes. Because of the hierarchical structure of the model, the probability of a test sample belonging to a terminal sub-region is the product of the probabilities computed at each classification node that the sample has passed through. One minus the probability that the test sample belongs to the terminal sub-region gives a measure of the uncertainty that the test sample belongs to the sub-region after a sequence of classification steps.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

3.3. The Hierarchical Classifier

At each node of the hierarchical model, two clusters were chosen based on a dendrogram produced by model-based clustering so that each "parent region" always had at most two "child regions". The dataset was divided into a training set including 2357 samples and a test set including 328 samples; the test set consisted of one sample taken from each of the different sampling sites and was held to test the model later. For each sampling site, the test sample selected for testing was selected as the one that had the smallest multivariate Euclidean distance to the vector consisting of the median values of 18 chemical measurements for all samples collected at the same sampling site, so that the test sample at each location can be expected to be less affected by seasonal variation compared to one chosen by random selection.

The binary, tree-structured hierarchical model produced from the analysis is shown in Figure 14. The terminal nodes indicate the terminal sub-regions, which are represented by circles with the region number inside. The parent nodes indicate the transitional regions, which are represented by rectangles with its child region numbers inside. All the training samples are assigned to only one of the 18 terminal sub-regions to establish the classification rules used in the hierarchical, multi-label model. For each parent node, the chemical measurements used to separate its two child regions in clustering are indicated between the two child nodes. At different levels of the tree, different sets of chemical measurements are used to decompose the regions. Some chemical measurements, especially the $^{1}H:^{2}H$ isotopic ratio, are used more frequently than others p because they were less affected by seasonal effects and tended to show systematic variation with distance; The $^{1}H:^{2}H$ ratio is known to be less affected by stream-flow conditions and therefore by seasonal variations as compared to most other chemical measurements (*41–43*). In addition, the $^{1}H:^{2}H$ ratio was strongly affected by elevation and by climate, both of which should vary spatially. Of the 18 usable chemical measurements, the isotope ratio values had much smaller variances than the rest, which suggests that many of the

extreme values and wide ranges in trace element levels are produced by stream-flow, seasonal and geologic variation. Trace elements with less spread compared with that of other trace elements, for example Ba, Ca and Si, were used more often in this study because for this spatiotemporal USGS surface water dataset, those chemical measurements with lower variance tended to show systematic variations with distance.

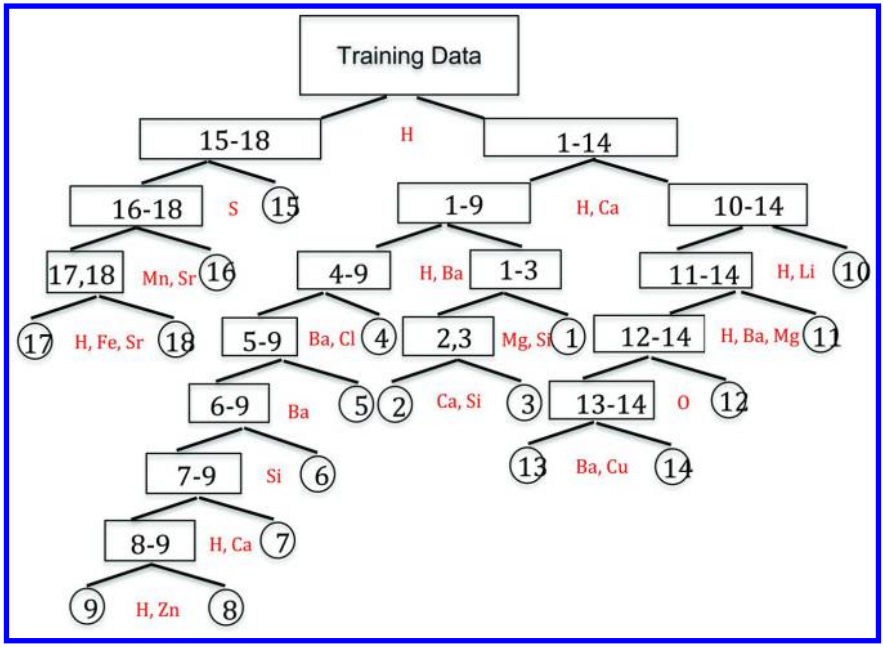

Figure 14. Hierarchical regions produced by the tree-structured hierarchical model of the USGS surface water dataset

3.4. Building the Hierarchical Probabilistic Classifier

At each node of the model, the region label determined from the clustering step was taken as the "true" class label of the samples from that region; a classifier can be built at each parent node to distinguish samples from the two different "child" nodes of that region. To be consistent with the probability-based clustering algorithm used here and to easily assess uncertainty estimates, a Naïve Bayes classifier (*44*, *45*) was built at each non-terminal node to identify the rule for assignment of class labels for its sub-regions. A Naïve Bayes classifier is a simple but effective (*46*) probabilistic classifier based on Bayes' theorem, but requires that the set of selected features were independent of one another within each class. The Bayes classifier is also known to work well when the assumption of feature independence is not strictly valid (*47*). With a naïve Bayes classifier, classes are associated with data in two steps: first, a training step is performed in which parameters of a probability distribution are estimated for features of each

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

of the classes based on the training samples, and a prediction step is performed in which the method computes the posterior probability of the test sample belonging to each class. The test sample is then classified by choosing the class with the larger posterior probability. The same set of chemical measurements used in the clustering step were also used in developing the classifier at each node. Because most of the chemical measurements showed non-Normal distributions, a kernel density function (*44*, *48*) was computed and fit to the samples from each region instead of using the normal distribution for the prior in the Bayes classifier. A Normal kernel, where the kernel function is a standard normal density function, was used here (*18*), and the width of the kernel smoothing window was automatically chosen for each combination of feature and class. The posterior probabilities, obtained by the application of Bayes' rule combining the prior and the likelihood function that the sample belongs to each of the two regions, were computed from the naïve Bayes classifier. The sample was classified into the region with the higher posterior probability, as shown in Figure 15. The prior probabilities for each class were taken as the relative frequencies of each class based on the number of training samples found in each region in the clustering step. The likelihood function depended on the probability distributions of the selected chemical measurements for each class. As discussed in Section *3.2*, the posterior probabilities calculated at each non-terminal node are independent from the posterior probabilities of the other nodes because of the use of independent priors and the differing chemical measurement sets used. The probability of a test sample belonging to a terminal sub-region was the product of the posterior probabilities computed at each classification node that the sample passed through.

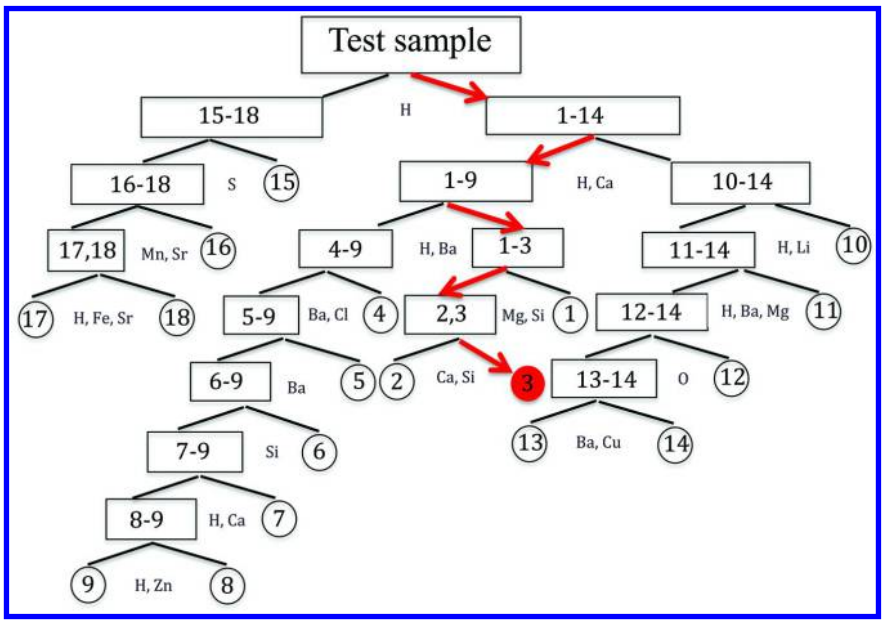

Figure 15. Placing an unknown sample in a geographic region by descending the tree

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Table I. Success in locating test samples

	Number of samples	*Classification accuracy*	*Average probability of classification*
Training set	1931	94%	0.85
Test set 1[a]	328	85%	0.85
Test set 2[b]	500	88%	0.87

[a] Drawn from reserved training data. [b] Kriged unknown data created from training samples.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

3.5. Assessing the Tree-Structured Hierarchical Model

Three approaches are adopted to evaluate the performance of the tree-structured hierarchical modeling algorithm. First, a training set made from the USGS surface water dataset was used to estimate the training error of the model. Second, a test set made from USGS surface water data that were omitted from consideration when building the hierarchical model was used to test the model. A third test was based on a set of test samples generated by spatial interpolation using Kriging, in which the chemical measurements at simulated, unobserved locations are predicted based on the original USGS surface water data. The accuracy of the classification of the data into geographic regions was assessed from application of the model to these three test sets. Results are discussed in turn below.

3.5.1. Testing the Tree-Structured Hierarchical Model Using the Training Set

All of the training samples (with outliers removed) with known terminal sub-region were used as the input to the classification model to estimate their locations. Because the training samples were used to build the tree-structured hierarchical mode, any test of the model using training samples is likely to produce optimistic results. The locations predicted from the model were compared with the true locations of the training samples to determine the training set error. With a hierarchical classification step, the classification training error of the hierarchical model can be easily determined, as shown in Table IV. In Table IV, CE represents the classification error (expressed as a percentage) of the samples that are supposed to be from each terminal sub-region after comparing the terminal sub-region label obtained in model-building and the predicted terminal sub-region label. For any terminal sub-region, if N_o represents the number of training samples in this region based on clustering results, and $\hat{N}_o$ represents the number of samples from N_o training samples that are actually classified into this terminal sub-region, the classification error (CE), defined as $CE = 1 - \frac{\hat{N}_o}{N_o}$,

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

and is within the range [0,1]. The classification accuracy is $\frac{\hat{N}_o}{N_o}$. The quantity TD indicates the depth of the each of the sub-regions in the tree-structured hierarchical model; that is, the number of separate binary classifications needed to reach this region of the tree. For example, in Figure 14, region 15 has a depth of 2 and region 16 has a depth of 3. For all N_o samples in each of the terminal sub-regions, CP, the average of all the N_o posterior probabilities, is an estimate of the average probability that samples of this region are classified into a terminal sub-region. Note that the posterior probability for each of the N_o samples assigned to the region by clustering is the product of the probabilities computed at each classification node that the sample has passed through. The results from Table IV show that most terminal sub-regions had a classification error of less than 5% for the training set. Considering that the data were environmental samples with many imputed values, that there were seasonal effects embedded in the data and that the classification here was not done in a single step, but over multiple steps in a hierarchical structure, a training error 5% is very good. However, there were some terminal sub-regions in the model that showed much higher classification error in the training set. The higher error resulted from limited data for these regions as well as the difficulty of selecting informative chemical measurements and from incomplete removal of interference from the seasonal variations in the data for these regions. By summing classification accuracies from all 18 terminal sub-regions in the data, an overall classification sample accuracy of 94% was calculated for all training samples. Samples from terminal sub-regions with larger model depths (TD) in the hierarchy, for instance regions 7, 8 and 9, were more likely to have lower CP values, so that these sub-regions showed lower probabilities of label assignment and correspondingly higher uncertainties in the class assignment for any sample reaching this sub-region. However, the samples belonging to those terminal sub-regions having higher classification errors were also likely to have higher uncertainty in class membership as established in the clustering step; samples from regions 2, 7 and 10 showed this tendency. Samples classified into an incorrect terminal sub-region (one where the cluster label for the sample differed from that found from the classification) generally showed a higher Bayes posterior membership uncertainty in that assignment, while the samples classified into a correct terminal sub-region typically were associated with a lower Bayes posterior uncertainty. For all the training samples that were correctly classified, the average CP of all the samples was 0.85, while the average CP was 0.55 for all of the misclassified training samples. The classification uncertainty of samples from a specific terminal sub-region were found to be related to the depth of that terminal sub-region in the tree-structured hierarchical model; when the three sub-regions with high classification errors were not considered, the correlation coefficient between TD and CP over 18 terminal sub-regions was -0.7, which suggests a strong negative relationship between the depth of the terminal sub-region and the class posterior membership probability for samples belonging to that terminal sub-region. This result is not surprising, since sub-regions with high depth in the hierarchical scheme require more classification steps, and those additional steps introduce more uncertainty in the assignment of the class label.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

Table II. Training error for the classification step in tree-structured hierarchical modeling of USGS surface water data (modified from (*32*))

Region	*CE(%)*[a]	*TD*	*CP*	*Region*	*CE(%)*	*TD*	*CP*
1	7.1	4	0.82	10	7.6	3	0.67
2	37.8	5	0.70	11	0.6	4	0.84
3	2.3	5	0.84	12	0.0	5	0.78
4	16.8	4	0.82	13	5.6	6	0.81
5	3.4	5	0.85	14	0.0	6	0.89
6	3.6	6	0.85	15	1.1	2	0.92
7	10.2	7	0.77	16	2.5	3	0.92
8	1.0	8	0.77	17	19.0	4	0.85
9	2.5	8	0.79	18	1.3	4	0.95

[a] CE is the classification error. CP is the average posterior probability that all N_o samples belonging to the terminal sub-region predicted from the clustering are correctly classified by the hierarchical classifier. TD indicates the depth of the region in the tree-structured hierarchical model.

Table III. Training error for the classification step in tree-structured hierarchical modeling of reserved samples from the USGS surface water data (modified from (*32*))

Region	*N*[a]	*CE (%)*	*CP*	*Region*	*N*	*CE (%)*	*CP*
1	14	14.3	0.77	10	27	22.2	0.70
2	7	28.6	0.70	11	32	9.4	0.81
3	17	5.9	0.82	12	2	0.0	0.63
4	18	27.8	0.82	13	2	0.0	0.86
5	25	24.0	0.86	14	4	0.0	0.80
6	11	36.4	0.83	15	26	7.7	0.89
7	14	14.3	0.73	16	22	4.5	0.93
8	19	26.3	0.77	17	7	14.3	0.90
9	19	21.0	0.74	18	37	0.0	0.95

[a] N is the number of test samples in each terminal sub-region. CE is the classification error of regional identification. For each terminal sub-region, CP is the average posterior probability that samples are hierarchically classified into the predicted terminal region.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

3.6. Testing the Tree-Structured Hierarchical Model Using an External Test Set

Table III summarizes the prediction results from the tree-structured hierarchical model applied to the test set made from the reserved USGS surface water data. Each test sample was assigned a "true" region number obtained from the results of clustering of similar samples so that the region identification error could be calculated. Since every sampling site of USGS surface water data contained a sample withheld from modeling to permit evaluation of the model, the "true" region number of the test sample was regarded to be the same as the region number of the training samples measured at the same location but at different times.

The overall classification accuracy of terminal sub-region samples was 85%, compared to 94% for training samples. This result is again very good considering that the test set is new data and may have contained seasonal information not incorporated in the model in building the tree-structured hierarchical model. This result demonstrates that the spatial variation contained in the selected chemical measurements at each node dominates any new seasonal information in test samples during the hierarchical classification. As with the training set, the regional identification error found for the test set varied greatly over the different sub-regions. Compared with the training samples, the test set contained far fewer samples for each terminal sub-region. Therefore, any inconsistency in region identification caused by seasonal variation effects contained in an individual sample had a correspondingly larger impact on the sample classification error, CE. As seen in the classification of training data, the probability of class assignment CP was generally lower for those sub-regions with higher classification errors (CE) or greater tree depths (TD) (Table IV). For all test samples that were correctly classified, the average CP was 0.85, while the average CP was 0.66 for all of the test samples that were classified into a terminal sub-region not consistent with the clustering. The probability that test samples were classified into the correct terminal sub-region was usually fairly consistent with that found for the training samples.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

3.7. Testing the Tree-Structured Hierarchical Model Using Kriged Samples

Ordinary Kriging (*49*) was implemented to generate new, simulated USGS surface water samples at locations unobserved in the training set. Ordinary Kriging is not suited to the situation where there are multiple samples measured at the same location. Therefore, the original USGS data with 2685 samples and 25 chemical measurements was trimmed down to a smaller set with 328 samples and 25 chemical measurements by averaging of all samples taken at the same location. This averaging has the effect of modifying the temporal variation at each site. Because the variation may not be random, the averaging may not have reduced the temporal effects in the data.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Kriging amounts to predicting a measurement at a specified location from a weighted linear regression model based on measurements made at nearby locations (*49*). To estimate the chemical measurements at new locations, the following steps were required for each terminal sub-region:

1. The locations of the unobserved data were computed. The new locations were computed by taking the weighted average of the locations of two training samples that were randomly chosen from the entire dataset. The weight was also randomly simulated at every iteration, within the range (0,1). For each terminal sub-region, the number of new Kriged samples generated within the region was made proportional to the number of training samples belonging to the region.
2. The variogram was calculated and fitted for each variable. The Euclidean distance between two samples was calculated from their longitude and latitude coordinates. For each chemical measurement within the maximum Euclidean distance used to define the variogram, a separate experimental variogram was calculated based on the trimmed dataset, which represented the spatial dependence of each of the 25 chemical measurement variables in the original USGS surface water data in the sub-region. A least-squares fit of the theoretical variogram using a spherical model (*50*) was then performed on each experimentally determined variogram.
3. Interpolation. Based on the variogram model found in Step 2, the weights for ordinary Kriging were computed and the expected error was minimized, then the set of 25 chemical measurements expected at the new, unobserved locations found in Step 1 were estimated in turn.

The true regions for Kriged test samples at new locations were taken as the terminal sub-regions from which they were generated. Due to the topology of the sampling sites, 13% of the samples generated by Kriging contained some negative values for chemical concentrations because of the negative Kriging weights combined with high values of location. Samples with any negative concentration values were removed before testing the tree-structured hierarchical model. The newly generated Kriging data consisted of 435 samples that estimated the 25 chemical measurements at locations different from those used for the original USGS surface water data sampling. The results of testing the model by using these data are shown in Table IV.

For the set of 435 useful Kriged samples obtained, the regional sample identification accuracy was 88%, and CP was 0.87. This result is very good considering the fact that new samples at locations unobserved in the training set were hierarchically classified into mostly correct regions. While it should be noted that the Kriging does not make a completely new sample because these are related to existing samples, and the Kriged samples are constrained to a single region to permit successful interpolation, it should also be noted that the Kriged data permit interrogation of the model multiple times and in all regions, something that is not possible with the withheld training data because of limited sampling data.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

Table IV. Error of tree-structured hierarchical modeling predictions estimated using Kriged test data (modified from (*32*))

Region	*N*[a]	*CE(%)*	*CP*	*Region*	*N*	*CE (%)*	*CP*
1	22	0.0	0.83	10	42	19.0	0.71
2	13	30.7	0.79	11	37	2.7	0.83
3	33	3.0	0.86	12	3	0.0	0.70
4	22	9.1	0.94	13	4	0.0	0.84
5	26	19.2	0.88	14	11	0.0	0.90
6	17	11.8	0.85	15	37	8.1	0.94
7	22	4.5	0.83	16	32	9.4	0.91
8	30	26.7	0.73	17	12	25.0	0.79
9	36	22.2	0.81	18	36	11.1	0.91

[a] N is the number of samples generated by ordinary Kriging in each sub-region. CE is the classification error of regional identification. For each terminal sub-region, CP is the average posterior probability that samples were hierarchically classified into the terminal region assigned from clustering.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

For samples at new locations generated by Kriging, the tree-structured hierarchical modeling algorithm can be used to accurately place the sample into the correct terminal sub-region with low uncertainty. The seasonal variation was controlled by averaging chemical measurements at each sampling site which, together with the interpolation of values for chemical variables by ordinary Kriging and slightly more compact test locations as compared to the distribution of original locations, produced location prediction performance for the Kriging samples that was slightly better than that of the test set made up of the original USGS surface water samples.

4. Conclusions

Tree-structured hierarchical modeling is an attractive way to simplify modeling of complex data. The creation of the hierarchy requires careful selection of variables and identification of class ontology, but the combination of a variable selection with model-based clustering offers a systematic way to accomplish the goal. The example presented in this paper demonstrates a systematic approach to prediction of location for USGS surface water samples from their chemical signatures, even in the presence of strong interference from temporal effects. Judging from the results of testing by three different assessment approaches, the tree-structured hierarchical model gives reliable identification of the geographical region from a chemical signature.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

Acknowledgments

This work was supported by the NGA under the NURI academic research program.

References

1. Kowalski, B. R.; Bender, C. F. *J. Am. Chem. Soc.* **1972**, *94*, 5632–5639.
2. Kowalski, B. R.; Bender, C. F. *J. Am. Chem. Soc.* **1973**, *95*, 686–693.
3. Bender, C. F.; Kowalski, B. R. *Anal. Chem.* **1973**, *45*, 590–592.
4. Duewer, D. L.; Kowalski, B. R. *Anal. Chem.* **1975**, *47*, 526–530.
5. Brown, S. D.; Skogerboe, R. K.; Kowalski, B. R. *Chemosphere* **1980**, *9*, 265–276.
6. Wangen, L.; Isenhour, T. L. *Appl. Spectrosc.* **1971**, *25*, 136.
7. Frew, N. M.; Wangen, L.; Isenhour, T. L. *Pattern Recognit.* **1971**, *3*, 281.
8. Vong, R.; Geladi, P.; Wold, S.; Esbensen, K. *J. Chemom.* **1988**, *2*, 281–296.
9. Alsberg, B. K.; Kell, D. B.; Goodacre, R. *Anal. Chem.* **1998**, *70*, 4126–4133.
10. Perestrelo, R.; Silva, C.; Camara, J. S. *J. Sep. Sci.* **2014**, *37*, 1974–1981.
11. Perez, N. F.; Ferre, J.; Boque, R. *Chemom. Intell. Lab. Syst.* **2009**, *95*, 122–128.
12. Brereton, R. G.; Lloyd, G. R. *J. Chemom.* **2014**, *28*, 213–225.
13. Silla, C. N.; Freitas, A. A. *Data Mining Knowl. Disc.* **2010**, *22*, 31–72.
14. Zhang, M.-L.; Zhou, Z.-H. *IEEE Trans. Knowl. Data Eng.* **2014**, *26*, 1819–1837.
15. Gibaja, E.; Ventura, S. *Wiley Interdisc. Rev.: Data Mining Knowl. Disc.* **2014**, *4*, 411–444.
16. Tsoumakas, G.; Katakis, I.; Vlahavas, I. In *Data Mining and Knowledge Discovery Handboo*k, Part 6; Springer: 2010; pp 667−685.
17. McLachlan, G. J. *Discriminant Analysis and Statistical Pattern Recognition*; Wiley-Interscience: New York, NY, 1992.
18. Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*, 2nd ed.; Wiley-Interscience: New York, NY, 2001.
19. Clare, A.; King, R. D. In *Lecture Notes in Computer Science 2168*; De Raedt, L., Siebes, A., Eds.; Springer: Berlin, Germany, 2001; pp 42–53.
20. Elisseeff, A.; Weston, J. In *Advances in Neural Information Processing Systems 14*, Dietterich, T. G., Becker, S., Ghahramani, Z., Eds.; MIT Press: Cambridge, MA, U.S.A., 2002; pp 681–687.
21. Cheng, W.; Hüllermeier, E. *Mach. Learn.* **2009**, *76*, 211–225.
22. Breiman, L. *Mach. Learn.* **2001**, *45*, 5–32.
23. Quinlan, J. R. *C4.5: Programs for machine learning*; Morgan Kaufmann Publishers: San Mateo, CA, 1993.
24. Gao, S.; Wu, W.; Lee, C.-H.; Chua, T.-S. In *Proceedings of the Twenty-first International Conference on Machine Learning* (ICML'04); ACM: 2004; pp 329–336.
25. Wu, X.; Kumar, V.; Quinlan, J. R.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G. J.; Ng, A.; Liu, B.; Yu, P. S.; Zhou, Z.-H.; Steinbach, M.; ·Hand, D. J.; Steinberg, D. *Knowl. Inf. Syst.* **2008**, *14*, 1–37.

26. McLachlan, G. J.; Krishnan, T. *The EM algorithm and extensions*; Wiley: New York, 1997.
27. Domingos, P.; Pazzani, M. *Mach. Learn.* **1997**, *29*, 103–130.
28. Fraley, C.; Raftery, A. E. *J. Am. Stat. Assoc.* **2002**, *97*, 611–631.
29. Otero, F.; Freitas, A.; Johnson, C. *Memetic Comput.* **2010**, *2*, 165–181.
30. Kawai, K.; Takahashi, Y. *Chem-Bio. Inf. J.* **2009**, *4*, 44–51.
31. Ukwatta, E.; Samarabandu, J. In *Canadian Conference on Computer and Robot Vision (CRV '09)*; IEEE: 2009; 132−139.
32. Chen, L.; Brown, S. D. *J. Chemom.* **2014**, *28*, 523–538.
33. United States Geological Survey. National Water Information System. http://waterdata.usgs.gov/nwis/ [accessed 1 February 2014].
34. Chen, L.; Brown, S. D. *J. Chemom.* **2014**, *28*, 358–369.
35. Zhang, H.; Lan, Y.; Lacey, R. E. *Int. Agric. Biol. Eng.* **2009**, *2*, 62–68.
36. Ganio, L. M.; Torgersen, G. E.; Gresswell, R. E. *Front. Ecol. Environ.* **2005**, *3*, 138–144.
37. Fovell, R. G.; Fovell, M. Y. C. *J. Clim.* **1993**, *6*, 2103–2135.
38. Fraley, C.; Raftery, A. E. *Comput. J.* **1999**, *41*, 578–588.
39. Preisner, O.; Lopes, J. A.; Menezes, J. C. *Chemom. Intell. Lab. Syst.* **2008**, *94*, 33–42.
40. Smith, B. M; Gemperline, P. J. *J. Chemom.* **2002**, *16*, 241–246.
41. Bowen, G. J.; Ehleringer, J. R.; Chesson, L. A.; Stange, E.; Cerling, T. E. *Water Resour. Res.* **2007**, *43*, 1–12.
42. Bowen, G. J.; Winter, D. A.; Spero, H. J.; Zierenberg, R. A.; Reeder, M. D.; Cerling, T. E.; Ehleringer, J. R. *Rapid Commun. Mass Spectrom.* **2005**, *19*, 3442–3450.
43. Ehleringer, J. R.; Bowen, G. J.; Chesson, L. A.; West, A. G.; Podlesak, D. W.; Cerling, T. E. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 2788–2793.
44. Murakami, Y.; Mizuguchi, K. *Bioinformatics* **2010**, *26*, 1841–1848.
45. Perez, A.; Larranaga, P.; Inza, I. *Int. J. Approx. Reason.* **2009**, *50*, 341–362.
46. Caruana, R.; Niculescu-Mizil, A. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006; ACM Press: New York, 2006; pp 161–168.
47. Zhang, H. *Int. J. Pattern Recognit.* **2005**, *19*, 183–198.
48. Scott, A. J.; Symons, M. J. *Biometrics* **2013**, *27*, 387–397.
49. Atkinson, P. M.; Lewis, P. *Comput. Geosci.* **2000**, *26*, 361–371.
50. Zhang, H.; Lan, Y.; Lacey, R. E. *Int. J. Agric. Biol. Eng.* **2009**, *2*, 62–68.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch007

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch008

Chapter 8

Improving Investigative Lead Information in the Forensic Examination of Automotive Paints

Barry K. Lavine,* Collin White, Matthew Allen, and Ayuba Fasasi

Department of Chemistry, Oklahoma State University, Stillwater, Oklahoma 74078
***Phone 405-744-5945, e-mail bklab@chem.okstate.edu**

Pattern recognition has been applied to the problem of searching the infrared (IR) spectral libraries of the Paint Data Query (PDQ) automotive paint database to differentiate between similar but nonidentical IR spectra, and to determine the assembly plant, model, and line of an automotive vehicle from a paint chip or smear recovered from a crime scene where damage to a vehicle and/or injury or death to a pedestrian has occurred. Currently, modern automotive paints use thinner undercoat and color coat layers protected by a thicker clear coat layer. All too often, a clear coat is the only layer of the manufacturer's paint left at the crime scene. In these cases, the use of text to encode chemical information about each layer of the manufacturer's paint in PDQ limits the searching of clear coats, as modern clear coats generally have only one of two possible formulations: acrylic melamine styrene or acrylic melamine styrene polyurethane. In these cases, the text based search system of PDQ would return a large (and unusable) number of hits that span multiple manufacturers and models. To assess the evidentiary information content of clear coats, pattern recognition techniques have been applied to the IR spectral libraries of the PDQ database to differentiate between similar but nonidentical automotive paint spectra. A prototype library search system to identify the assembly plant of an automobile from the IR spectrum of a clear coat has been developed. The proposed pattern recognition assisted IR library search system for automotive clear coats described in this chapter consists of two separate but interrelated components: search prefilters

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

to reduce the size of the PDQ library to a specific assembly plant or plants corresponding to the unknown paint sample and a search algorithm that cross correlates the spectrum of the unknown with each IR spectrum in the set identified by the search prefilters to identify the IR spectrum in the truncated PDQ library most similar to the unknown. Even in challenging trials where the clear coat paint samples evaluated were all the same manufacturer (e.g., Chrysler) within a limited production year range (2000-2006), the assembly plant and model of the automobile from which the unknown automotive paint sample was obtained could be correctly identified. The prototype pattern recognition system outperformed commercial library search algorithms.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch008

Introduction

Automotive paint in the form of an intact chip or smear is often recovered from a crime scene where damage to a vehicle and/or injury or death to a pedestrian has occurred. Studies (*1*, *2*) performed by the Royal Canadian Mounted Police (RCMP) approximately 40 years ago demonstrated that vehicles could be differentiated by comparing the color, layer sequence, and chemical composition of each layer in an automotive paint system. To make these comparisons possible, the RCMP developed the Paint Data Query (PDQ) database for forensic automotive paint examinations (*3*, *4*). Today, PDQ contains over 21,000 samples (street samples and factory panels) that correspond to over 84,000 individual paint layers, representing the automotive paint systems used on most domestic and foreign vehicles marketed in North America. Over 53 local, state, and federal forensic laboratories in the United States use PDQ as well as international forensic laboratories including the National Forensic Laboratory Service of the RCMP, the Center of Forensic Sciences in Toronto, Canada, members of the ENFSI network of European forensic science institutes, and the Australian Police Services.

Modern automotive paints typically consist of four layers (*5*, *6*). From top to bottom, they are the clear coat, color coat, surfacer, and primer. PDQ is a database of the physical attributes, chemical composition and the infrared (IR) spectrum of each layer of the manufacturer's original paint system. If the original layers are present in a paint chip or smear, PDQ can assist in identifying the specific manufacturer and model of the automotive vehicle within a limited production year range.

PDQ utilizes a text-based search system that relies on information about the color and the chemical formulation of both the top and undercoat paint layers to identify the vehicle. Although direct searching of IR spectra in PDQ does not exist, a text based search of both the physical and chemical attributes of each layer of automotive paint can serve as a potent pre-screen to a manual IR search (*7*, *8*). Unfortunately, the use of text to encode chemical information about each paint layer limits the searching of clear coats as modern clear coats applied to automotive substrates generally have only one of two possible formulations:

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

acrylic melamine styrene or acrylic melamine styrene polyurethane. Although the presence of urethane in a clear coat paint sample is coded for a sample in PDQ, the amount could be large or small, a feature that can easily be distinguished through a visual inspection of the IR data but cannot be searched using the text-based system of PDQ. Thus, initial PDQ searches for this clear coat sample alone would return a large (and unusable) number of hits that span multiple manufacturers and models.

The vast majority of clear coats are not colored (i.e. tinted), and none contain inorganic fillers or binders with which to further differentiate one clear coat from another. As modern automotive paints use thinner undercoat and color coat layers protected by a thicker clear coat layer, all too often, a clear coat paint smear is the only layer of paint left at the crime scene. In these cases, the text based search system of PDQ cannot identify the automotive vehicle. Although commercial library search algorithms are capable of searching paint layers in combination against the PDQ spectral libraries with some success, clear coat formulations are too similar for these algorithms to generate accurate hit lists by searching clear coat paint spectra alone.

To assess the evidentiary information content of clear coats, pattern recognition techniques have been applied to the IR spectral libraries of the PDQ database to differentiate between similar but nonidentical automotive paint spectra. To tackle the problem of library searching, a prototype library search system to identify the assembly plant of an automobile from the IR spectrum of a clear coat paint smear has been developed. The proposed pattern recognition assisted IR library search system for automotive clear coats described in this chapter consists of two separate but interrelated components: search prefilters to reduce the size of the PDQ library to a specific assembly plant or plants corresponding to the unknown paint sample and a search algorithm that cross correlates the spectrum of the unknown with each IR spectrum in the set identified by the search prefilters to identify the IR spectrum in the truncated PDQ library most similar to the unknown. As the size of the library is culled for a specific match, the search prefilters increase both the selectivity and accuracy of the search. Even in challenging trials where the clear coat paint samples evaluated were all the same manufacturer (e.g., Chrysler) within a limited production year range (2000-2006), the assembly plant and model of the automobile from which the unknown automotive paint sample was obtained could be correctly identified. Furthermore, the prototype pattern recognition system outperformed commercial search algorithms.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch008

Experimental

Method

IR spectra of 1206 clear coats applied to the metal surfaces of automobiles and trucks assembled at 25 General Motors (GM), 12 Chrysler, and 17 Ford assembly plants within a limited production year range (2000 – 2006) were obtained using either a Thermo-Nicolet 6700s, BioRad 40A or BioRad 60 FTIR spectrometer equipped with a DTGS detector. All clear coat IR spectra

were collected in transmission mode using high pressure diamond anvil cells augmented with either a Harrick 4x beam condenser (BioRad 40A and BioRad 60) or Harrick 6x beam condenser (Thermo-Nicolet 6700s instrument).

Data preprocessing of the IR spectra of the clear coats was crucial to ensure the successful development of the search prefilters. In this study, data preprocessing consisted of aligning the optical systems of the Thermo Nicolet and BioRad FTIR spectrometers through normalization of the helium neon frequency, application of the discrete wavelet transform (*9*) to resolve overlapping spectral bands, and variable selection to identify informative wavelet coefficients using a genetic algorithm. Each step is described in detail below.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch008

Spectral Alignment

Sample preparation and placement can affect both wavelength alignment and the measured absorbance. When the sample itself is not the limiting aperture, small wavenumber shifts, nevertheless, can be observed if the sample is not positioned at a fixed angle relative to the IR beam. The beam may be refracted at an angle that displaces the image on the detector as the beam traverses the sample. This can change the pathlength of the beam through the interferometer, altering the relationship of the interferogram to the He-Ne laser, resulting in small wavenumber shifts. For samples that are tilted with respect to the laser beam, the pathlength within the sample can also change altering measured absorbance values. To address these problems, the interferogram of each clear coat paint spectrum was multiplied by the Norton-Beer medium apodization function prior to the application of the Fourier transform to ensure that measured absorbance would be equal to the true absorbance. Each IR spectrum was then normalized to the helium neon laser frequency of 15798.0 cm^{-1}. This laser frequency value corresponds to that measured at the aperture setting to ensure that peak positions are independent of aperture setting. Both operations (apodization and normalization) were performed using OMNIC (Thermo Nicolet). After this preprocessing, each spectrum consisted of 1869 points for the entire mid-IR range of 400 cm^{-1} to 4000 cm^{-1}.

Wavelets

The fingerprint region (667 cm^{-1} to 1640 cm^{-1}) of each spectrum comprised of 506 points was normalized to unit length and subject to wavelet analysis using the MATLAB Wavelet toolbox 3.0.4 (Math Works, Natick, MA). Outside of the fingerprint region, each IR spectrum consisted of only noise (2100 cm^{-1} to 2500 cm^{-1}) due to uncompensated absorption of IR radiation by the diamond anvil cell, and C-H stretching bands which were present in all clear coat spectra. The discrete wavelet transform (*10*, *11*) using the Symlet6 mother wavelet at the eighth level of decomposition was applied to the fingerprint region of each spectrum. Wavelet coefficients generated at all eight levels of decomposition (1150 in total for each spectrum which included the approximation coefficients) were retained for discriminant development.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch008

Genetic Algorithm for Variable Selection

Wavelet coefficients that conveyed information about the manufacturer or the assembly plant of the automotive vehicle were identified by a genetic algorithm (GA) for pattern recognition analysis (*12–17*). The pattern recognition GA identified a set of wavelet coefficients that optimized the separation (i.e., discrimination) of the assembly plants in a plot of the two or three largest principal components of the aligned, preprocessed, and wavelet transformed spectral data. Because principal components maximize variance, the bulk of the information encoded by the selected wavelet coefficients was about the classification problem of interest. The principal component (PC) plot in the fitness function of the pattern recognition GA acted as an embedded information filter. A good PC plot could only be generated by coefficients whose variance or information content is primarily about spectral differences between the assembly plants. The principal component analysis routine incorporated into the fitness function of the pattern recognition GA limited the search to these types of coefficient subsets, thereby significantly reducing the size of the search space. In addition, the pattern recognition GA focused on those assembly plants and/or clear coats that were difficult to classify, as the pattern recognition GA trained by boosting the relative importance (i.e., weights) of those assembly plants and/or paint samples (i.e., vehicles). Clear coats that consistently classified correctly were not as heavily weighted as those samples that were difficult to classify. Over time, the pattern GA was able to learn its optimal parameters in a manner similar to a neural network. The pattern recognition GA integrates aspects of artificial intelligence and evolutionary computations to yield a "smart" one-pass procedure for variable selection and classification. Further details about the operation of the pattern recognition GA can be found elsewhere (*18*, *19*).

Search Prefilters and Infrared Library Searching

Search prefilters (i.e., classifiers) were developed from PDQ library spectra to extract information from an unknown clear coat to yield a response based on the manufacturer and the assembly plant of the vehicle. Most search prefilter identify library spectra dissimilar to the unknown for the purpose of excluding them from the search. This allows for more powerful search algorithms that are also more computationally intensive to be utilized for library matching as the size of the library has been culled for a specific match. In this study, the Chrysler search prefilters limit the search of each validation sample to a specific assembly plant or group of assembly plants whose clear coat spectra are similar to the clear coat spectrum of the validation sample. As the Chrysler clear coat search prefilters only retain those spectra in the library similar to the spectrum of the validation sample, both the accuracy and speed of the search would be expected to be increased.

Spectral features encoded in the wavelet coefficients identified by the pattern recognition GA were used to develop these classifiers. Search prefilters to identify the assembly plant of Chrysler automotive vehicle were developed from IR spectra obtained from 379 clear coats and 12 Chrysler car and truck

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

assembly plants in North America between the years 2000 and 2006. To ensure classification accuracy, search prefilters for assembly plant were developed using a hierarchical classification scheme. This approach simultaneously allowed for removal of irrelevant variation from the data using the pattern recognition GA to identify informative wavelet coefficients and regularization of the classifier for ill posed classification problems using linear models (i.e., principal component plots) of the data. Feature selection and classification was optimized simultaneously at each level. Differentiation between the inter-class and intra-class variation could not be achieved using a horizontal classification structure (where all twelve assembly plants are differentiated simultaneously). For automotive manufacturer (GM versus Chrysler versus Ford), a single classifier was sufficient to identify the make of the vehicle.

Search prefilters developed from the clear coats eliminated dissimilar spectra from the library search providing the forensic scientist with an opportunity to take advantage of more sophisticated but also more time-consuming search algorithms. Commercial infrared library search systems compare IR spectra by summing the squares of the difference between spectra at every wave number. However, these algorithms do not perform well when differentiating between similar spectra as small peak shifts are not handled well and bands of low intensity, which may be highly informative, are largely ignored. For these reasons, library searching was performed using a cross correlation search algorithm to provide the best match between an unknown and the IR spectra in the hit list generated by the search prefilters. The cross correlation function has been shown to correctly identify unknown spectra from similar but nonidentical spectra (*20*). Although cross correlation is slower than conventional search algorithms, it is suitable as a post searching method to rank probable matches that have been selected by a faster algorithm (e.g., search prefilters).

Library matching was performed by cross correlating the unknown with each spectrum in the library subset identified by the search prefilters and then comparing each cross correlated spectrum with the corresponding autocorrelated library spectrum (*21*). Since cross correlation is a measure of the similarity of two time varying functions, cross-correlation can be used to estimate the correlation between two signals by computing the dot product after a suitable time lag has been applied to one of the signals. The cross correlation function C_{ij} for the sampling interval Δt and relative displacement $n\Delta t$ between two signals s_I and s_j is estimated by the following equation

$$C_{ij}(n\Delta t) = \frac{1}{T}\sum_{t=0}^{T} s_i(t)s_j(t), \qquad n = 0,1,2,\ldots\frac{T}{\Delta t} \quad (1)$$

Autocorrelation, which is similar to cross correlation, is the signal cross correlated with itself. Autocorrelation and cross correlation were performed by normalizing all IR spectra to unit length. The cross correlation library searching algorithm identifies the IR spectrum that is most similar to the unknown in the truncated library using three different modes of comparison:

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch008

1. Each autocorrelated library spectrum is compared to each cross-correlated library and unknown spectrum
2. Autocorrelated spectrum of unknown is compared to each cross-correlated unknown and library spectrum.
3. Autocorrelated spectrum of unknown is compared to each autocorrelated library spectrum

Each comparison was made using a range of window sizes centered at the midpoint of the cross-correlated data (which corresponds to the cross correlation between two signals with zero lag) and increased in steps of 10 points or 100 points to include the entire cross correlated spectrum. Because of the symmetry associated with cross correlation, the comparisons were made from only one side of the center burst.

To assess similarity of spectra in the library matching, the Euclidian distance was used to evaluate the similarity index (see Equation 2) where s_{ij} is the similarity of the match, d_{ij} is the Euclidean distance between the cross correlated and autocorrelated spectrum and d_{max} is the largest distance in the set of (cross correlated and autocorrelated) spectra compared. The similarity metric in Equation 2 was used instead of the hit quality index (*22*), as it proved to be more informative for ranking IR spectra.

$$s_{ij} = 1 - \frac{d_{ij}}{d_{max}} \quad (2)$$

Library spectra were arranged in descending order of similarity for each comparison. The five most similar library spectra in each window size were chosen from each comparison, with the sample identities preserved. After every window was analyzed, a histogram was generated depicting the frequency of occurrence for the most similar spectra. The top two library spectra with the highest frequency of occurrence were selected as potential matches.

Library matching was performed using the entire with the exception of the region corresponding to absorption of infrared radiation by the diamond transmission cell. The performance of the prototype pattern recognition driven library searching system (search prefilters and cross-correlation library search algorithm) was compared to OMNIC, a commercial library searching algorithm used in Thermo Nicolet FTIR spectrometers.

Results and Discussion

Development of Search Prefilters for Chrysler

The first step was to differentiate Chrysler clear coats by plant group. To determine the composition of each plant group, Chrysler assembly plants (see Table 1) whose clear coat paint spectra exhibited a doublet for the carbonyl band (acrylic melamine styrene polyurethane) as opposed to a singlet (acrylic melamine styrene**)** were flagged. The two assembly plants (Jefferson North and Newark) whose clear coat paint spectra exhibited a doublet for the carbonyl were placed

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch008

in Plant Group 13, whereas the assembly plants whose clear coat paint spectra exhibited a singlet for the carbonyl band were assigned to other plant groups. Using only clear coat paint spectra, each of the ten remaining assembly plants (see Table 1) was analyzed by principal component analysis (*23*) to assess its class structure. In four of the ten assembly plants (Bramalea/Brampton, Dodge Main, St. Louis, and Toledo), the PC plot of the clear coat paint spectra exhibited two distinct sample clusters (see Figures 1 - 4).

Table 1. Chrysler Assembly Plants and Plant Groups

Plant	*PID# (data label)*	*Divided between groups*	*Group*
Belvidere (BEL)	1000	NO	11
Bloomington (BLO)	1001	NO	12
Bramalea/Brampton (BRA/BRP)	1002	YES	11, 12
Dodge Main (DOD)	1003	YES	11, 12
Jefferson North (JFN)	1004	NO	13
Newark (NEW)	1006	NO	13
Saltillo (SAL)	1007	NO	11
Sterling Heights (STH)	1008	NO	12
Saint Louis (STL)	1009	YES	11, 12
Toledo (TOL)	1010	YES	11, 12
Toluca (TOU)	1011	NO	11
Windsor (WIN)	1012	NO	12

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch008

For the Bramalea/Brampton assembly plant, clustering occurred on the basis of model: Dodge Charger and some Chrysler 300 lines versus Chrysler Concorde, Chrysler LHS, Dodge Intrepid, Dodge Magnum, and other Chrysler 300 lines, whereas for Dodge Main, clustering occurred on the basis of the production year of the vehicle: 2000-2002 versus 2003-2006. For the St. Louis assembly plant, clustering occurred on the basis of both model and line: Dodge Caravan and Chrysler Town and Country versus Dodge Ram, whereas for Toledo, clustering was correlated to a specific vehicle: Jeep Liberty versus the other models and lines assembled at the plant. Because the average clear coat paint compared visually, the four assembly plants were further divided into subplants on the basis of the observed sample clustering.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch008

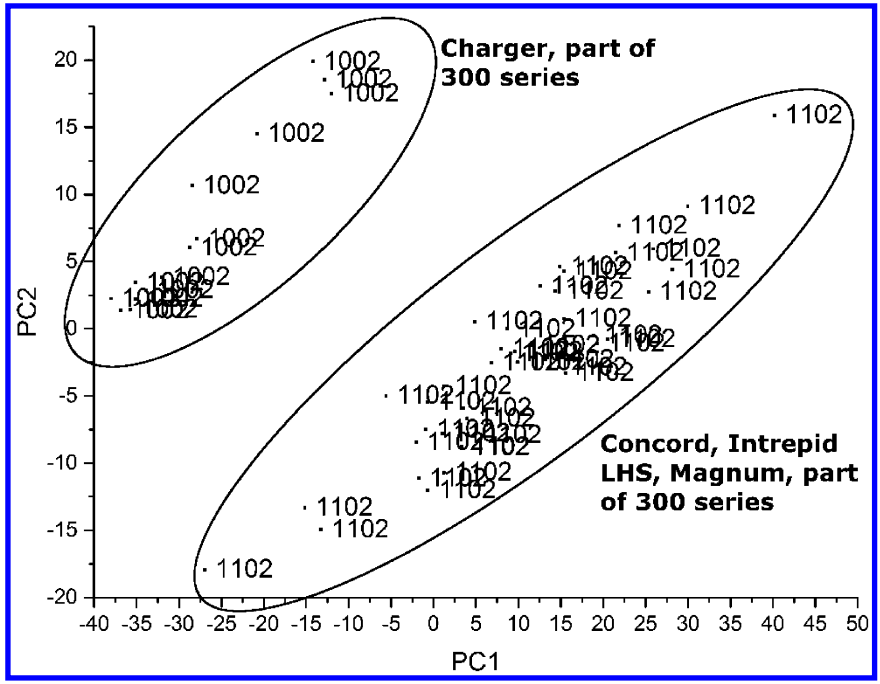

Figure 1. Plot of the two largest principal components of the clear coat paint spectra from the Bramalea/Brampton plant. Two distinct sample clusters are evident in the plot.

To assign the remaining assembly plants and subplants to specific plant groups, the average IR spectrum (clear coat layer) of each assembly plant or subplant was computed. Principal component analysis and hierarchical clustering (*24*) were performed on the average spectra. Figures 5 and 6 summarize the results of the clustering study. Plant Group 11 consists of Belvidere, Bramalea/Brampton (subplant), Dodge Main (subplant), Saltillo, St. Louis (subplant), Toledo (subplant), and Toluca assembly plants, whereas Plant Group 12 is comprised of Bloomington, Bramalea/Brampton (subplant), Dodge Main (subplant), Sterling Heights, St. Louis (subplant), Toledo (subplant), and Windsor assembly plants.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch008

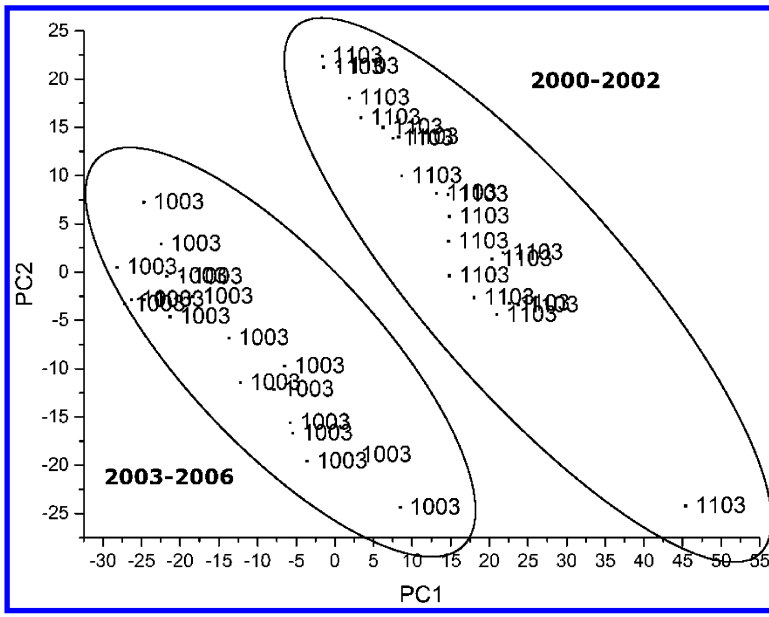

Figure 2. Plot of the two largest principal components of the clear coat paint spectra from the Dodge Main plant. Two distinct sample clusters are evident in the plot.

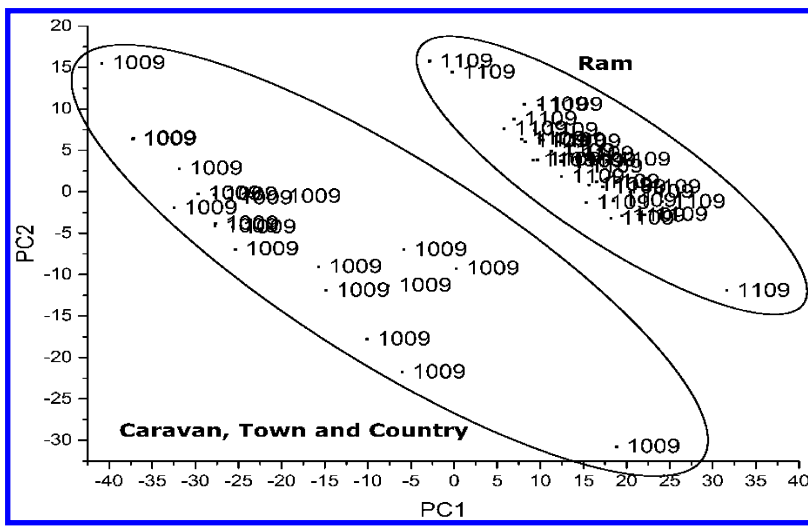

Figure 3. Plot of the two largest principal components of the clear coat paint spectra from the St. Louis plant. Two distinct sample clusters are evident in the plot.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch008

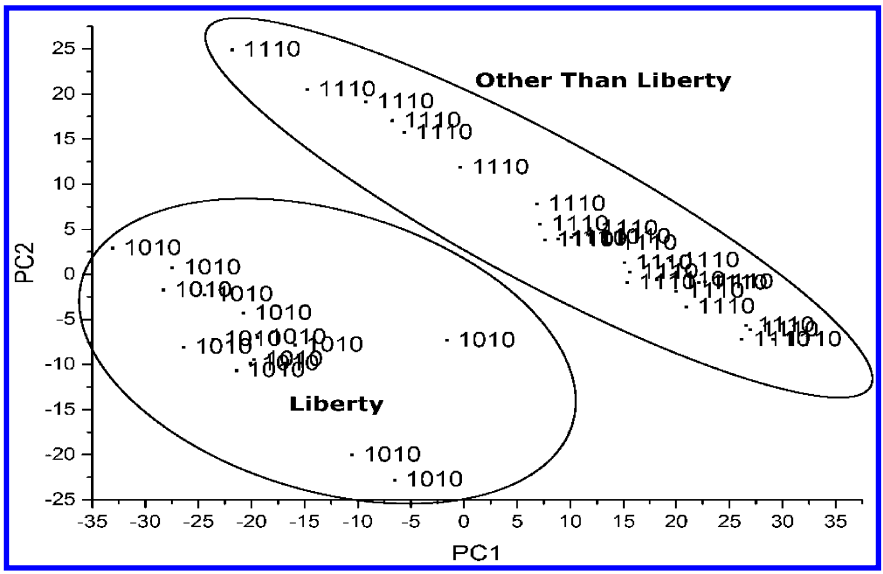

Figure 4. Plot of the two largest principal components of the clear coat paint spectra from the Toledo plant. Two distinct sample clusters are evident in the plot.

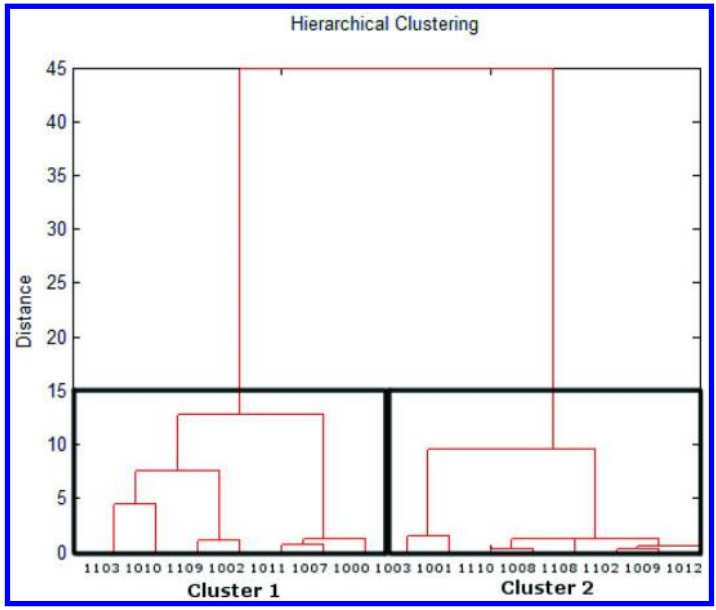

Figure 5. Hierarchical cluster analysis (Wards method) of the average IR spectrum (clear coats) of each assembly plant or subplant. 1000 = Belvidere, 1001 = Bloomington, 1002 = Bramalea/Brampton subplant, 1003 = Dodge Main subplant, 1007 = Saltillo, 1008 = Sterling Heights, 1009 = St. Louis subplant, 1010 = Toledo, 1011 = Toluca, 1012 = Windsor, 1102 = Bramalea/Brampton subplant, 1103 = Dodge Main subplant, 1109 = St. Louis subplant, and 1110 = Toledo subplant.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch008

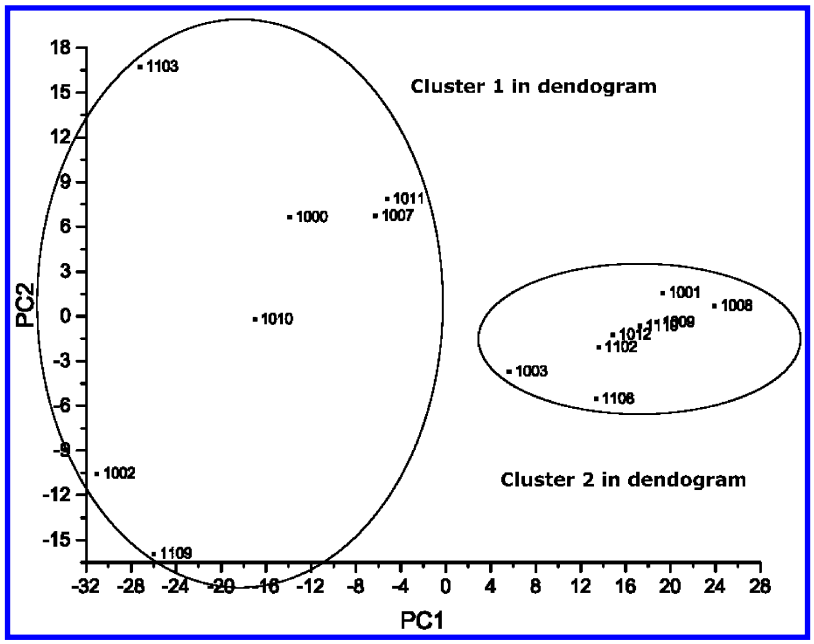

Figure 6. Principal component analysis of the average IR clear coat paint spectrum of each assembly plant or subplant. 1000 = Belvidere, 1001 = Bloomington, 1002 = Bramalea/Brampton subplant, 1003 = Dodge Main subplant, 1007 = Saltillo, 1008 = Sterling Heights, 1009 = St. Louis subplant, 1010 = Toledo, 1011 = Toluca, 1012 = Windsor, 1102 = Bramalea/Brampton subplant, 1103 = Dodge Main subplant, 1109 = St. Louis subplant, and 1110 = Toledo subplant.

Having ascertained the membership of each plant group, the next step was classification. The Chrysler clear coats were divided into a training set of 379 samples and a validation set of 42 samples. The validation set samples, which were selected by random lot, were not part of the training set used to develop the search prefilter for plant group or assembly plant. The training set of 379 clear coats was divided into 3 classes by Plant Group (see Table 1). Figure 7 shows a PC plot of the two largest principal components of the 379 wavelet transformed clear coat IR spectra and the 1150 wavelet coefficients comprising the training set for Plant Group (see Table 1). All wavelet coefficients were autoscaled prior to principal component analysis. Each clear coat is represented as a point in the PC plot of the data. (1 = Plant Group 11, 2 = Plant Group 12, and 3 = Plant Group 13). Although Plant Group 12 is well separated from Plant Groups 11 and 13, the other two plant groups overlap in the PC plot of the wavelet transformed IR spectral data.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch008

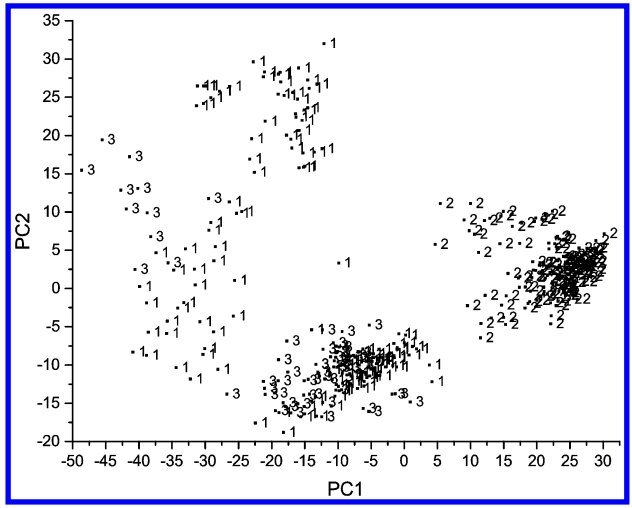

Figure 7. PC plot of the two largest principal components of the 379 wavelet transformed clear coat IR spectra and the 1150 wavelet coefficients comprising the training set data for Plant Group. (1 = Plant Group 11, 2 = Plant Group 12, and 3 = Plant Group 13).

Feature selection was performed to identify wavelet coefficients characteristic of the profile of each plant group. The pattern recognition GA identified informative wavelet coefficients by sampling key feature subsets, scoring their PC plots, and tracking those plant groups/and or IR spectra that were difficult to classify. The boosting routine used this information to steer the population to an optimal solution. After 200 generations, the pattern recognition GA identified 9 wavelet coefficients whose PC plot showed clustering (see Figure 8) of the IR clear coat paint spectra on the basis of plant group.

To assess the predictive ability of the 9 wavelet coefficients identified by the pattern recognition GA, a validation set of 42 IR spectra was used. IR spectra from the validation set were projected directly onto the PC plot developed from the 379 IR spectra and the training set and the 9 wavelet coefficients identified by the pattern recognition GA. Figure 9 shows the projection of the validation set samples onto the PC map of the training set data. All validation set samples are located in a region of the map with clear coats from the same Plant Group.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch008

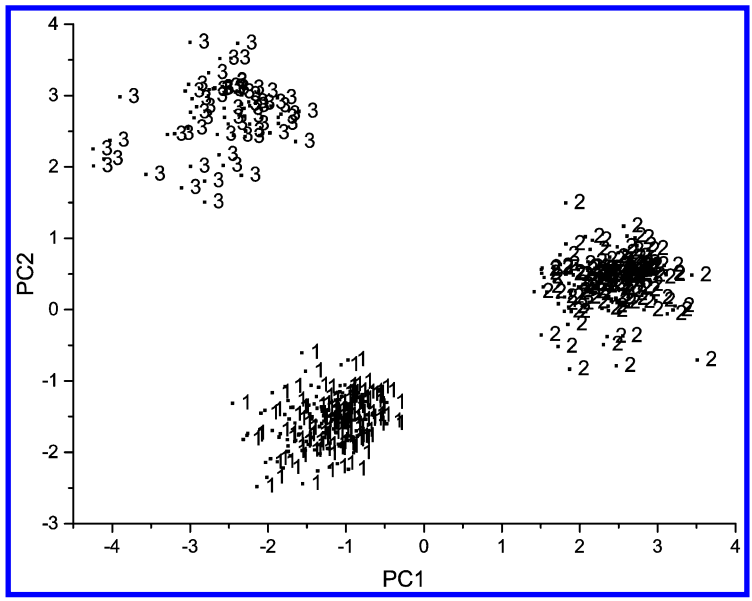

Figure 8. PC plot of the two largest principal components of the 379 training set samples and 9 wavelet coefficients identified by the pattern recognition GA (1 = Plant Group 11, 2 = Plant Group 12, 3 = Plant Group 13).

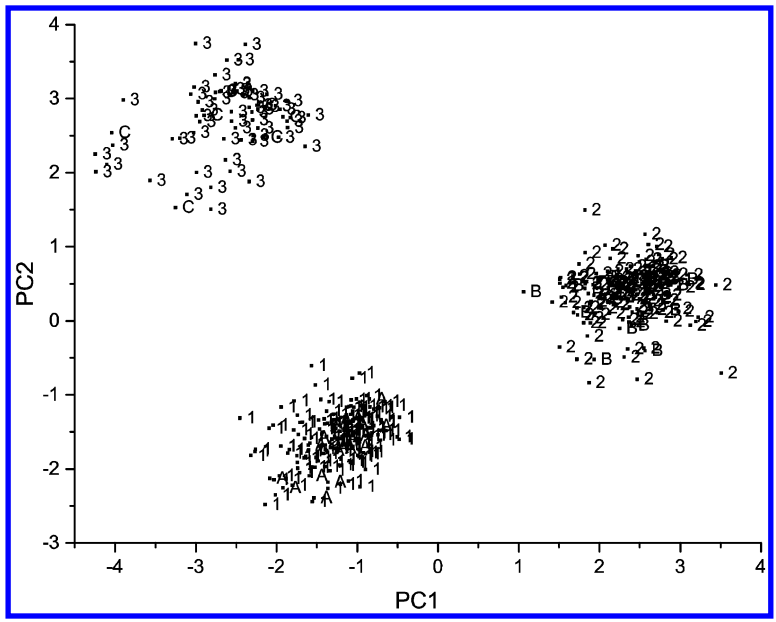

Figure 9. Validation set samples projected onto the PC plot of the data defined by the 379 wavelet transformed clear coat IR spectra of the training set and the 9 wavelet coefficients identified by the pattern recognition GA. (1 = Plant Group 11, 2 = Plant Group 12, and 3 = Plant Group 13, A = Plant Group 11 (validation), B = Plant Group 12 (validation), C = Group 13 (validation)).

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Linear discriminant analysis (*25*) was also used to classify the 379 wavelet transformed IR spectra in the training set. The training set data were divided into 3 classes on the basis of plant group. Linear discriminant analysis was used to develop a classifier to separate the clear coats by plant group. A discriminant developed from the 9 wavelet coefficients identified by the pattern recognition GA achieved a classification success rate of 100% for the training set. To further test the predictive ability of these 9 coefficients and the discriminant associated with them, the validation set of 42 IR spectra of clear coat paint samples was employed. Again, a classification success rate of 100% was achieved for the IR spectra in the validation set. The results obtained from the LDA study are consistent with the results obtained using PCA.

For each plant group, a search prefilter was developed to discriminate the automotive paint samples by assembly plant. Table 2 lists the assembly plants and subplants comprising Plant Group 11, which consists of four subplants (Bramalea/Brampton, Dodge Main, St. Louis, Toledo) and a plant subgroup consisting of three assembly plants (Belvidere, Saltillo and Toluca). Saltillo, Toluca, and Belvidere were combined into a subplant group because the average spectra of the Belvidere, Saltillo and Toluca assembly plants were similar.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch008

Table 2. Assembly and Subplants Comprising Plant Group 11

Plant	*Training*	*Validation*
7110 (Belvidere + Saltillo + Toluca plants)	80	13
1002 (subplant of Bramalea/Brampton)	13	1
1010 (subplant of Toledo)	14	1
1103 (subplant of Dodge Main)	19	2
1109 (subplant of St Louis)	30	2

Figure 10 shows a plot of the two largest principal components of the 156 samples comprising the training set for Plant Group 11 and the 31 wavelet coefficients identified by the pattern recognition GA. Each clear coat is represented as a point in the PC plot. Four subplants (Bramalea Brampton, Dodge Main, St. Louis, and Toledo) form distinct clusters in this PC plot whereas the plant subgroup comprised of Saltillo and Toluca overlap with Belvidere, forming a larger plant subgroup. Projecting the validation set samples assigned to Plant Group 11 onto the PC plot showed that each projected validation set sample is located in a region of the PC plot with samples from the same assembly plant or subplant.

Table 3 lists the assembly plants or subplants comprising Plant Group 12, which consists of one assembly plant (Bloomington), one subplant (Dodge Main) and one plant subgroup consisting of three subplants (Bramalea/Brampton, St. Louis and Toledo), and two assembly plants (Sterling Heights and Windsor).

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Bramalea/Brampton, St. Louis, Toledo, Sterling Heights and Windsor were combined to form a plant subgroup because their average spectra were very similar.

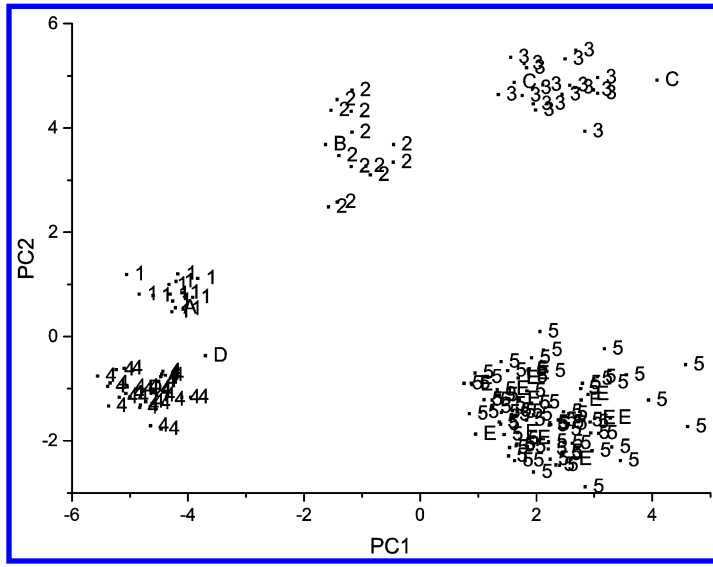

Figure 10. Validation set samples (indicated with P at end of class label) projected onto the PC plot of the data defined by the 156 paint samples comprising Plant Group 11 (training set) and the 31 wavelet coefficients identified by the pattern recognition GA. (1 = subplant of Bramalea/Brampton, 2 = subplant of Toledo, 3 = subplant of Dodge Main, 4 = subplant of St Louis, 5 = plant subgroup containing Belvidere, Saltillo and Toluca, A-E = validation (A corresponding to 1, B corresponding to 2, etc.)).

Table 3. Assembly and Subplants Comprising Plant Group 12

Plant	*Training*	*Validation*
8182 (subplant of Bramalea/Brampton + Sterling Heights + subplant of St. Louis + subplant of Toledo + Windsor)	110	15
1001 (Bloomington)	33	3
1003 (subplant of Dodge Main)	13	1

Figure 11 shows a plot of the two largest principal components of the 157 samples comprising the training set for Plant Group 12 and the 29 wavelet coefficients identified by the pattern recognition GA. Each clear coat is represented as a point in the PC plot. Bloomington and the Dodge Main subplant were well separated from each other and the plant subgroup. Projecting the validation set

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch008

samples assigned to Plant Group 12 onto the PC plot shows that each projected validation set sample is located in a region of the PC map with samples from the same assembly plant, subplant, or plant subgroup.

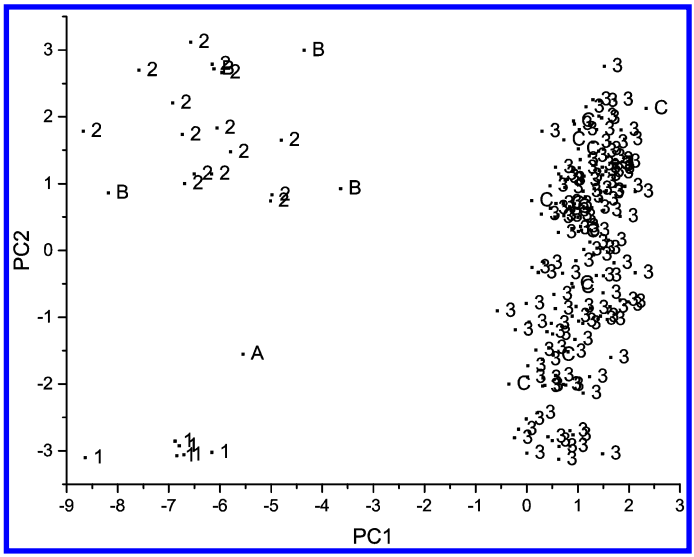

Figure 11. Validation set samples (indicated by P at end of class label) projected onto the PC plot of the data defined by the 157 paint samples comprising Plant Group 12 (training set) and the 29 wavelet coefficients identified by the pattern recognition GA. (1 = Bloomington, 2 = subplant of Dodge Main, 3 = plant subgroup containing subplant of Bramalea/Brampton, Sterling Heights, subplant of St. Louis, and Windsor, A-C = validation (A corresponding to 1, B corresponding to 2, C corresponding to 3).

Table 4 lists the two assembly plants comprising Plant Group 13 for the development of the search prefilter for assembly plant. The pattern recognition GA was not able to identify a set of coefficients from the wavelet transformed IR spectra that could differentiate Jefferson North from the Newark assembly plant. To better understand the reasons for this lack of success, principal component analysis was performed on both the Newark and the Jefferson North assembly plants. Clustering correlated to the production year of the vehicle was observed for the Newark assembly plant (see Figure 12). For this reason, Newark was divided into two subplants. Again, we were not able to identify wavelet coefficients that could solve this three-way classification problem (Jefferson North versus Newark subplant 2000-2002 versus Newark subplant 2002-2006). However, we observed during the course of this GA run that Jefferson North clustered in two distinct groups on the basis of production year. One cluster consisted of Grand Cherokees (2000 and 2006) and Commodores (2006) and the other cluster consisted of Grand Cherokees, Cherokees, and Commodores (2001-2006). Furthermore, one of the Jefferson North clusters merged with a Newark subplant. For this reason,

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch008

the three class study was reconfigured to take into account these subplants (see Table 5).

Figure 13 shows a plot of the two largest principal components of the 64 samples comprising Plant Group 13 and the 3 wavelet coefficients identified by the pattern recognition GA for the training set. Every subplant or plant subgroup is well separated from each other in the plot. Projecting the validation set samples assigned to Plant Group 13 onto the PC plot showed that each projected validation set sample is located in a region of the PC map with samples from the same plant subgroup or subplant.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch008

Table 4. Assembly Plants Comprising Plant Group 13

Plant	*Training*	*Validation*
Jefferson North	34	3
Newark	30	3

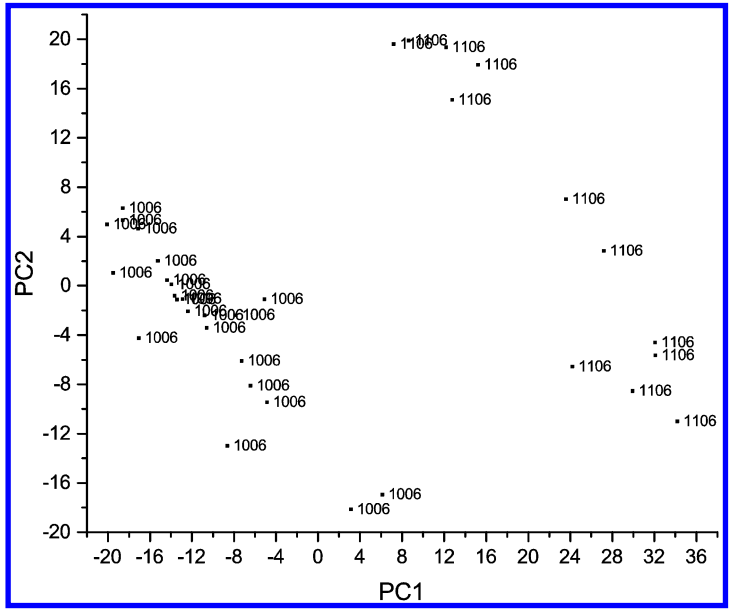

Figure 12. Plot of the two largest principal components of the clear coat paint spectra from the Newark assembly plant. Two distinct sample clusters are evident in the plot.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Table 5. Assembly and Subplants Comprising Plant Group 13

Plant	*Training*	*Validation*
406 (subplant of Jefferson North + subplant of Newark)	42	3
1104 (subplant of Jefferson North)	11	2
1106 (subplant of Newark)	11	1

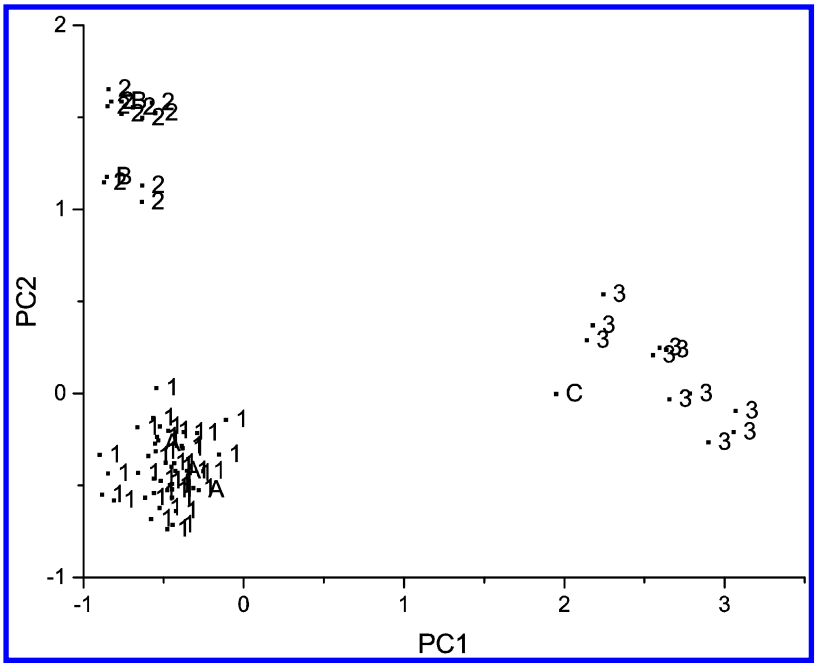

Figure 13. Validation set samples (indicated by P at end of class label) projected onto the PC plot of the data defined by the 157 paint samples comprising Plant Group 13 (training set) and the 3 wavelet coefficients identified by the pattern recognition GA. (1 = subplant of Jefferson North, subplant of Newark. 2 = subplant of Jefferson North, and 3 = subplant of Newark, A-C = validation (A corresponding to 1, B corresponding to 2, C corresponding to 3)).

A summary of the results obtained for the 42 validation set samples is shown in Table 6. All validation set samples were correctly classified by the Chrysler search prefilters at both the Plant Group and assembly plant level.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch008

Table 6. Chrysler Search Prefilters Validation Sample Results

Validation Sample	*Assigned Plant Group*	*Assigned Plant(s)*	*Correct Plant*
1	11	1103	1103
2	11	1109	1109
3	12	1102	1102
4	12	1002, 1008, 1009, 1012, 1108, 1110	1012
5	12	1002, 1008, 1009, 1012, 1108, 1110	1110
6	12	1001	1001
7	12	912	912
8	11	1000, 1007, 1011	1011
9	12	1102	1102
10	11	1000, 1007, 1011	1007
11	11	1000, 1007, 1011	1000
12	13	1104	1104
13	12	1002, 1008, 1009, 1012, 1108, 1110	1108
14	11	1000, 1007, 1011	1007
15	13	1004, 1006	1006
16	13	1004, 1006	1006
17	11	1000, 1007, 1011	1000
18	12	1102	1102
19	11	1000, 1007, 1011	1007
20	11	1109	1109
21	11	1000, 1007, 1011	1011
22	13	1104	1104
23	11	1002	1002
24	12	1002, 1008, 1009, 1012, 1108, 1110	1110
25	11	1103	1103
26	11	1000, 1007, 1011	1007
27	11	1000, 1007, 1011	1011
28	12	1002, 1008, 1009, 1012, 1108, 1110	1008
29	12	1003	1003
30	12	1002, 1008, 1009, 1012, 1108, 1110	1008
31	12	1003	1003

Continued on next page.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch008

Table 6. (Continued). Chrysler Search Prefilters Validation Sample Results

Validation Sample	*Assigned Plant Group*	*Assigned Plant(s)*	*Correct Plant*
32	11	1000, 1007, 1011	1011
33	13	1106	1106
34	13	1004, 1006	1004
35	11	1000, 1007, 1011	1011
36	11	1000, 1007, 1011	1007
37	11	1000, 1007, 1011	1000
38	12	1003	1003
39	12	1002, 1008, 1009, 1012, 1108, 1110	1012
40	12	1002, 1008, 1009, 1012, 1108, 1110	1008
41	12	1003	1003
42	11	1010	1010

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch008

Search Prefilter for Automotive Manufacturer

As the goal of this study is to demonstrate proof of concept, not to develop a field deployable system, the search problem was intentionally made more challenging through selection of samples from the same automotive manufacturer within a limited production year range. This tested the capability of the search prefilters to differentiate between similar but nonidentical IR spectra. To use the Chrysler search prefilters in an automated library search system, it would be necessary to develop a search prefilter to differentiate clear coats by automobile manufacturer. To demonstrate that Chrysler automotive clear coats can be discriminated from those of other manufacturers, a search prefilter was developed to discriminate Chrysler clear coats from those of GM and Ford. For this study, 1206 IR spectra of clear coats (425 GM, 421 Chrysler, and 360 Ford) within a limited production year range (2000-2006) were employed. To develop this search prefilter, the 1206 IR spectra were divided into a training set of 1164 clear coats and a validation set of 42 Chrysler clear coats previously used in the Chrysler search prefilter study. All IR spectra were smoothed using a Savitzky-Golay filter (4^{th} order polynomial, 17 point window) and then wavelet transformed using a Symlet6 mother wavelet at the 8^{th} level of decomposition prior to pattern recognition analysis. Figure 14 shows a PC plot of the 1206 IR spectra of the training set and the 39 wavelet coefficients identified by the pattern recognition GA to differentiate clear coats by automobile manufacturer. Each paint sample is represented as a point in the PC map of the data (1 = GM, 2 = Chrysler, and 3 = Ford). Chrysler, GM, and Ford clear coats can be readily differentiated from each other in the PC plot.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

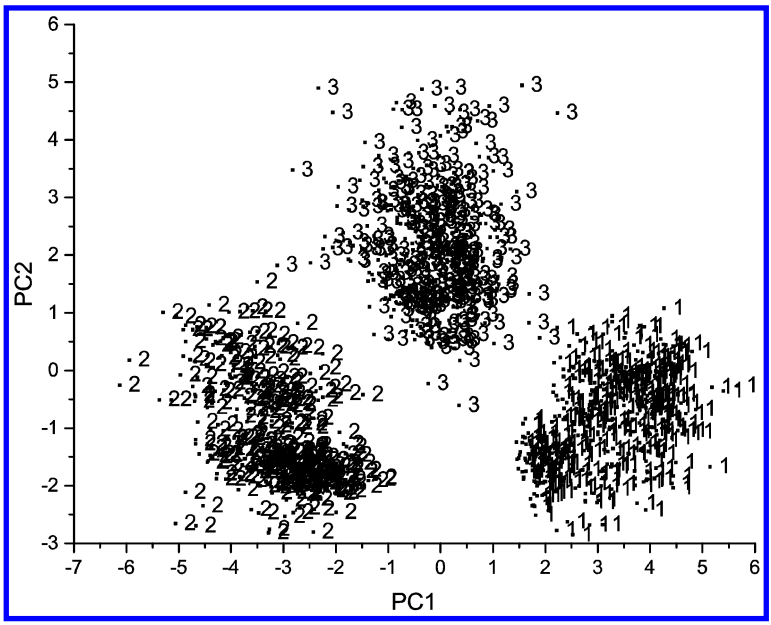

Figure 14. PC plot of the data defined by the 1183 paint samples (training set) and the 53 wavelet coefficients identified by the pattern recognition GA. (1 = GM, 2 = Chrysler (training set), 3 = Ford)

To assess the predictive ability of the 39 wavelet coefficients identified by the pattern recognition GA, we chose to map the 42 Chrysler clear coat IR spectra from the validation set directly onto the PC map developed from the 1164 training set samples and the 39 wavelet coefficients identified by the pattern recognition GA (see Figure 15). Every validation set sample was correctly classified, i.e. each projected Chrysler sample was projected in a region of the map with clear coats from the same manufacturer.

A 3-layer neural network (39-3-3) trained by back propagation (*26*) was also used to classify the 1164 wavelet transformed IR spectra comprising the training set. A discriminant developed from the 39 wavelet coefficients identified by the pattern recognition GA achieved a classification success rate of 100% for the training set. To further test the predictive ability of this network, the validation set of 42 Chrysler clear coats was employed. Again a classification success rate of 100% was achieved for the IR spectra comprising the validation set. The results from the network analysis were consistent with the results obtained from principal component analysis. Evidently, features in the IR spectra of clear coats correlated to automobile manufacturer can be identified by the pattern recognition GA. This suggests that a search prefilter can be developed to discriminate Chrysler clear coats from those of other automotive manufacturers. When used in conjunction with Chrysler search prefilters, it should be possible to identify the assembly plant of a Chrysler vehicle from a paint chip or paint smear recovered from a crime scene.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch008

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch008

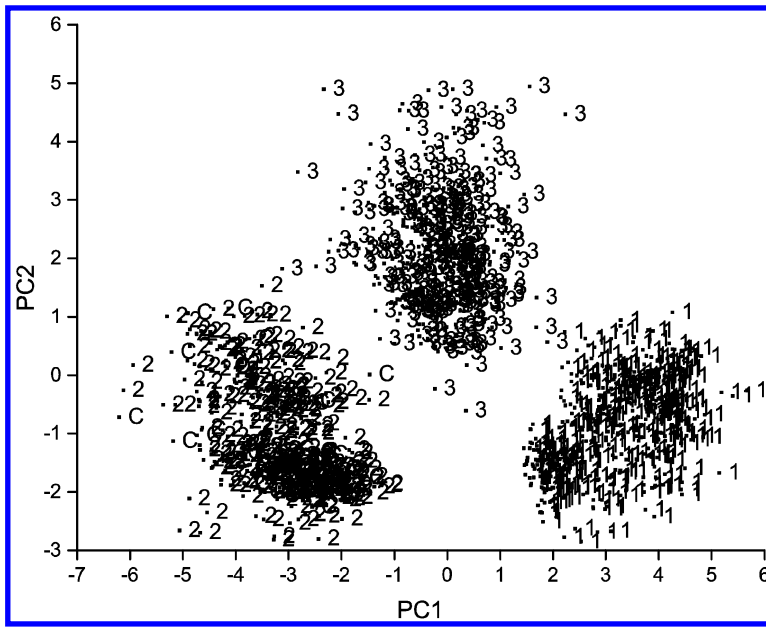

Figure 15. Validation set samples projected onto the PC plot of the data defined by the 1183 paint samples (training set) and the 39 wavelet coefficients identified by the pattern recognition GA. (1 = GM, 2 = Chrysler (training set), 3 = Ford, C = Chrysler (validation set))

Library Searching

To extract information about the vehicle model or line from its IR spectrum, the cross-correlation library search algorithm was used to identify the IR spectrum most similar to the unknown in the set identified by the Chrysler search prefilters. During this phase of the study, each spectrum in the validation set was vector normalized and compared to library spectra using the region from 668 to 1891 cm^{-1} and 2856 to 3675 cm^{-1}. The first one hundred and fifty points (399 to 667 cm^{-1}) and the last one hundred and seventy (3676 to 4000cm^{-1}) were omitted because of noise. The region of the spectrum corresponding to absorption by the diamond transmission cell was also omitted.

The results from the cross-correlation library searching algorithm for the 42 validation set samples are summarized in Table 7. The top five hits from the cross-correlation library searching algorithm for each validation set sample were compared to the top five matches obtained for each validation set sample identified by OMNIC, a commercial IR library search algorithm considered by many workers in the field as the industry standard. For this study, OMNIC was configured using correlation for the search type and Happ-Genzel for apodization as this set of conditions yielded the best results. Library search results for OMNIC are also summarized in Table 7. Clearly, the prototype pattern recognition library system (search prefilters and the cross-correlation library searching algorithm) outperformed OMNIC. For the six validation set samples misclassified by cross

correlation, the PDQ library did not have an IR spectrum of the corresponding model and line for one sample, two of the validation samples were comparatively poor matches to their correct library samples, and the remaining three missed validation samples were similar to their correct library sample, but were more similar to other (incorrect) library samples.

Table 7. Results from Library Search for Validation Samples

Library Searching Method	*Correct Top 5 Matches*	*Total Validation Samples*
Cross-Correlation	36	42
OMNIC 1206 GM, Ford, and Chrysler IR spectra comprising the library	28	42

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch008

Conclusion

The prototype IR library search system for the PDQ database is able to differentiate between similar but nonidentical FTIR spectra. The use of search prefilters increases both the selectivity and accuracy of the search by eliminating spectra from the search that are not from the same assembly plant(s). The prototype system is directly targeted to enhance current approaches to data interpretation in automotive forensic paint examinations and to aid in evidential significance assessment, both at the investigative lead stage and at the courtroom testimony stage.

Information derived from searches using the prototype library search system can serve to quantify the general discrimination power of original automotive paint comparisons encountered in casework, and further efforts to succinctly communicate the significance of the evidence to the courts.

Acknowledgments

BKL wishes to expression his appreciation to Mark Sandercock of the Royal Canadian Mounted Police for his guidance and advice, providing technical information about Chrysler and General Motors automotive paint systems, and supplying IR spectra from the PDQ database used in this study. This research was supported by two grants, 2010-DN-BX-K217 and 2012-DN-BX-K05, awarded by the National Institute of Justice, Office of Justice Programs, United States Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect those of the Department of Justice.

References

1. Rodgers, P. G.; Cameron, R.; Cartwright, N. S.; Clark, W. H.; Deak, J. S.; Norman, E. W. W. *Can. Soc. Forensic Sci. J.* **1976**, *9*, 1–14.
2. Rodgers, P. G.; Cameron, R.; Cartwright, N. S.; Clark, W. H.; Deak, J. S.; Norman, E. W. W. *Can. Soc. Forensic Sci. J.* **1976**, *9*, 49–68.
3. Cartwright, N. S.; Cartwright, L. J.; Norman, E. W. W.; Cameron, R.; Clark, W. H.; MacDougal, D. A. *Can. Soc. Forensic Sci. J.* **1982**, *15*, 105–115.
4. Buckle, J. L.; MacDougal, D. A.; Grant, R. R. *Can. Soc. Forensic Sci. J.* **1997**, *30*, 199–212.
5. Dossel, K. F. Top Coats. In *Automotive Paints and Coatings*; Streitberger, H. J., Dossel, K. F., Eds.; Wiley-VCH: New York, 2008.
6. Ryland, S. G.; Suzuki, E. M. Analysis of Paint Evidence. In *Forensic Chemistry Handbook*; Kobilinsky, L., Ed.; John Wiley & Sons: New York, 2012.
7. Bishea, G.; Buckle, J.; Ryland, S. International Forensic Automotive Paint Database. In *Proceedings of SPIE Conference, Investigation and Forensic Science Technologies*; International Society for Optical Engineering: 1999; Vol. 3576, 73−76.
8. Beveridge A.; Fung T.; MacDougall, D. Use of Infrared Spectroscopy for the Characterization of Paint Fragments. In *Forensic Examination of Glass and Paint: Analysis and Interpretation*; Caddy, B., Ed.; Taylor and Francis: New York, 2001.
9. Walker, J. S. *Primer on Wavelets and Their Scientific Applications*; Chapman & Hall/CRC: Boca Raton, FL, 1999.
10. Hubbard, B. B. *The World According to Wavelets*, 2nd ed.; A. K. Peters: Natick, MA, 1998.
11. Chau, F.; Liang, Y.; Fao, J.; Shao, X. *Chemometrics – From Basics to Wavelet Transform*; John Wiley & Sons: New York, 2004.
12. Lavine, B. K.; Fasasi, A.; Mirjankar, N.; Sandercock, M. *Talanta* **2015**, *120*, 182–190.
13. Lavine, B. K.; Fasasi, A.; Mirjankar, N.; White, C. *Microchem. J.* **2014**, *113*, 30–35.
14. Lavine, B. K.; Fasasi, A.; Mirjankar, N.; Sandercock, M.; Brown, S. D. *J. Chemometr.* **2014**, *28*, 385–394.
15. Lavine, B. K.; Mirjankar, N.; Delwiche, S. *Microchem. J.* **2014**, *117*, 178–182.
16. Lavine, B. K.; Nuguru, K.; Mirjankar, N.; Workman, J. *Appl. Specrosc.* **2012**, *66*, 917–925.
17. Lavine, B. K.; Nuguru, K.; Mirjankar, N.; Workman, J. *Microchem. J.* **2012**, *103*, 21–36.
18. Lavine, B. K.; Nuguru, K.; Mirjankar, N. *J. Chemomet.* **2011**, *25*, 116–129.
19. Lavine, B. K.; Davidson, C. E. Multivariate Approaches to Classification Using Genetic Algorithms. In *Comprehensive Chemometrics*; Brown, S., Tauler, R., Walczak, R., Eds.; Oxford-Elsevier: New York, 2009; Vol. 3.
20. Powell, L. A.; Hieftje, G. M. *Anal. Chim. Acta* **1978**, *100*, 313–327.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch008

21. Fasasi, A.; Mirjankar, N.; Stoian, R.; White, C.; Allen, M.; Sandercock, M.; Lavine, B. K. *Appl. Spectrosc.* **2015**, *69*, 84–94.
22. Smith, B. *Infrared Spectral Interpretation: A Systematic Approach*; CRC Press: Boca Raton, FL, 1999.
23. Jackson, J. E. *A Users Guide to Principal Component Analysis*; John Wiley & Sons: New York, 1991.
24. Massart, D. L.; Kaufman, L. *The Interpretation of Analytical Chemical Data by the use of Cluster Analysis*; John Wiley & Sons: New York, 1983.
25. James, M. *Classification Algorithms*; John Wiley & Sons: New York, 1985.
26. Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*; Wiley-VCH: Weinheim, Germany, 1999.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch008

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch009

Chapter 9

Net Analyte Signal (NAS) for Selection of Multivariate Calibration Models and Development of NAS Sample-Wise Target Calibration Model Attributes

Jonathan Palmer and John H. Kalivas*

Department of Chemistry, Idaho State University, Pocatello, Idaho 83209
***E-mail: kalijohn@isu.edu**

Common approaches to multivariate calibration such as multiple linear regression (MLR), partial least squares (PLS), or ridge regression (RR) require a model selection process (tuning parameter selection). Selection often involves evaluating only the cross validation prediction errors, but assessing multiple criteria is more robust. With multiple model quality indicators, trade-offs between the model indicators can be used to identify acceptable models. Guidelines to assist model selection can be formed with net analyte signal (NAS) assessment measures. By using spatial relationships between calibration model vectors and NAS related components, a geometric NAS construct can be formed. Presented in this paper are NAS attributes derived from the NAS construct. The potential of some of these measures to guide selection of RR and PLS models is studied using near infrared and nuclear magnetic resonance sample sets. Sample-wise NAS target model regression approaches are proposed.

Introduction

In multivariate calibration, various algorithms are employed in seeking a solution to the general equation

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} \tag{1}$$

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

where **y** is an $m \times 1$ vector of known responses (a physical or chemical property of interest, often analyte concentration) for m reference samples, **X** is an $m \times n$ matrix of n measured variables, **b** is the $n \times 1$ regression model vector to be estimated, and **e** is the $n \times 1$ vector accounting for normally distributed random noise with mean zero and covariance $\sigma^2\mathbf{I}$ with **I** being the $n \times n$ identity matrix (*1–3*). The model in Equation 1 is typical for spectroscopic data and in these situations, the variables are typically wavelengths.

The task of regression model formation is generalized as

$$\hat{\mathbf{b}} = \mathbf{X}^{+}\mathbf{y} \tag{2}$$

where the + sign denotes a generalized inverse and the hat symbol over **b** indicates estimated values. Once the model is formed, it can then be used to predict responses for a set of new samples by $\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}}$.

Common algorithms to estimate the regression vector **b** in Equation 1 include partial least squares (PLS), ridge regression (RR), a variant of Tikhonov regularization TR), multiple linear regression (MLR), and principal component regression (PCR), but there are others. These calibration processes vary depending on respective generalized inverses used in Equation 2 and the corresponding tuning parameters involved.

Many processes are available for tuning parameter selection with evaluation of prediction errors from a cross validation (CV) process being commonplace. However, work has recognized potential issues with just prediction error as the sole indicator of model quality as well as development of alternatives (*3–11*).

The concept of net analyte signal (NAS) has been used in the chemistry literature in numerous ways including a preliminary assessment of an NAS figure of merit for model selection (*6*). Evaluated in this paper are additional geometric components of the NAS construct with application towards model tuning parameter selection as well as a calibration approach based on sample-wise NAS target modeling. For the sample-wise target modeling, the NAS for each new unknown sample is regulated as a target in the regression processes in an attempt to better balance the selectivity/sensitivity tradeoff in forming the model for that particular sample. The paper concludes with suggestions towards incorporating the new NAS attributes in an ensemble approach based on the sum of ranking differences (SRD) (*9*) to select models consistent across multiple NAS attributes.

Net Analyte Signal (NAS)

The fundamentals of NAS are well described (*12–18*) and only a brief overview is provided here. The NAS hyperdimensional geometry is the crux of a model selection algorithm as well as the NAS measures evaluated in this paper. For simplicity however, discussion focusses on a distilled three dimensional analogy of the NAS geometry. The underlying foundation of this geometric construct is the set of representative non-analyte spectra (spectra without the analyte also known as the interferent spectra) designated **N** and depicted as the two dimensional plane shown in Figure 1.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch009

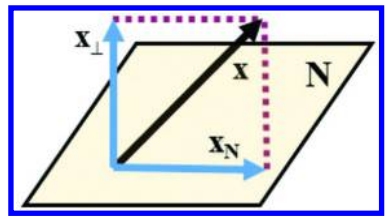

Figure 1. The N space depicted as a two dimensional plane analog with sample vector x and its associated projections into ($\mathbf{x}_\mathbf{N}$) and orthogonal ($\mathbf{x}_\perp$) to N.

Recent work has noted two common definitions of NAS (*14*). The strictest definition designates NAS as that part of a measured spectrum due to the analyte of interest that is orthogonal to the non-analyte components making up the rest of the sample spectrum. The orthogonal NAS vector is computed by projecting the sample spectrum **x** orthogonal to **N** using

$$\mathbf{x}_\perp = \left(\mathbf{I} - \mathbf{V}_k\mathbf{V}_k^{\mathrm{T}}\right)\mathbf{x} \tag{3}$$

where the subscript $\perp$ symbol denotes that $\mathbf{x}_\perp$ is the orthogonal NAS vector, $\mathbf{V}_k$ represents the $n \times k$ matrix of k selected loading vectors obtained from the singular value decomposition (SVD) of the **N** ($\mathbf{N} = \mathbf{USV}^{\mathrm{T}}$) and **I** is the $n \times n$ identity matrix. Under this definition, the reliability of NAS is entirely dependent upon the quality (comprehensiveness) of **N**. Any non-analyte component missing from **N** will manifest into $\mathbf{x}_\perp$ as a false positive NAS characteristic. In the noise free case, it has been shown that the PLS regression vector, up to normalization, is the NAS vector (*17*). It should be noted that for an improved NAS vector, the **x** should first be projected into the relevant calibration space (*15*).

The orthogonal definition of NAS essentially forces the algorithm to exclude certain characteristics of non-analyte spectra that may be surreptitiously useful for prediction, perhaps due in part to significant spectral overlap with the analyte spectrum. In other words, the model vector direction for the selected tuning parameter may not be completely orthogonal to the non-analyte space and not resemble the orthogonal NAS vector (*17–21*) thereby reducing selectivity. As a result, there is a gain in sensitivity (*6*).

A flexible description of NAS can be made in an attempt to counter the strict orthogonality constraint. Redefining NAS as the portion of a sample spectrum useful for prediction (*14*) allows inclusion of spectral variations that may aid model formation and hence, prediction accuracy, regardless of the degree of presupposed NAS association to the analyte or non-analyte. In an effort to achieve this oblique or semi-orthogonal NAS, an additional parameter was tested to control the degree of orthogonality when forming $\mathbf{x}_\perp$ by

$$\mathbf{x}_\rho = \left(\mathbf{I} - \rho\mathbf{V}_k\mathbf{V}_k^{\mathrm{T}}\right)\mathbf{x} \tag{4}$$

where ρ is a scalar in the range of zero to one and the absence of the $\perp$ symbol indicates that $\mathbf{x}_\rho$ may not be orthogonal to the **N** (*21*, *22*). When $\rho = 1$, $\mathbf{x}_\rho = \mathbf{x}_\perp$ and when $\rho = 0$, $\mathbf{x}_\rho = \mathbf{x}$ and no projection is made. Controlling the degree of

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch009

orthogonality set by ρ alters the vector direction of $\mathbf{x}_\rho$ relative to **N**. This becomes more apparent by defining

$$\mathbf{x}_\mathbf{N} = \mathbf{V}_k \mathbf{V}_k^\mathrm{T} \mathbf{x} \tag{5}$$

the net non-analyte signal (NnAS) contained within the sample (the projection of **x** into the plane of **N**). The NnAS vector $\mathbf{x}_\mathbf{N}$ can be substituted into Equation 4 to obtain $\mathbf{x}_\rho = \mathbf{x} - \rho\mathbf{x}_\mathbf{N}$ showing that by reducing ρ from one, $\mathbf{x}_\rho$ can contain some non-analyte information. Shown in Figure 1 is the projection with ρ = 1 forming $\mathbf{x}_\perp$.

The angular freedom brought by ρ allows the NAS vector $\mathbf{x}_\rho$ to extract useful information from **N** conferring value to the prediction process while adjusting the selectivity/sensitivity tradeoff. In practice however, the proper degree of projecting **x** is not known a prior and hence, ρ should ideally be optimized in order to obtain the NAS $\mathbf{x}_\rho$ most useful for prediction on a sample-wise basis, a difficult task. However, it has been shown that the magnitude and direction of $\hat{\mathbf{b}}$ (determined from Equation 2 based on some tuning parameter value) relative to an **N** does reveal information useful for prediction in terms of the selectivity/sensitivity tradeoff (*6*). This relationship is further studied in this paper for tuning parameter selection in the global sense (one model for multiple samples) and in the local sense (a unique model for each sample).

Comparison of the ratio $\|\mathbf{x}_\perp\| / \|\mathbf{x}_\mathbf{N}\|$ based on respective Euclidean 2-norms (L_2 norms symbolized by $\|\square\|$) for the NAS/NnAS ratio, provides a rudimentary estimate of a signal to noise ratio within the sample from an orthogonal NAS perspective.

The method of Equation 4 is the focus of this chapter and net analyte preprocessing (NAP) is not studied. For NAP, two orthogonal projections are used. The first estimates **N** by orthogonalizing **X** to **y** using $\mathbf{N} = \left(\mathbf{I} - \mathbf{y}\left(\mathbf{y}^\mathrm{T}\mathbf{y}\right)^{-1}\mathbf{y}^\mathrm{T}\right)\mathbf{X}$. This projection is followed by orthogonalizing **X** to the estimated **N** by a projection similar to Equation 4. The first orthogonal projection in NAP is comparable to orthogonal signal correction (OSC) (*18*, *23*, *24*).

NAS Measurers for Global Model Selection

Selection of candidate global models and hence tuning parameters, proceeds by comparison of NAS geometric relationships. Global models are used to predict future samples until said models are no longer functional due to changes in measurement conditions. These NAS measures are based on angles and Euclidean distances measured among geometric relationships within the NAS construct (spatial relations between analyte and non-analyte components). The size and orientation of these measures change as tuning parameters are adjusted, revealing information relevant to the tradeoff between selectivity and sensitivity.

Figure 2 shows a possible situation for a model vector from a calibration set using Equation 2 relative to an **N**.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch009

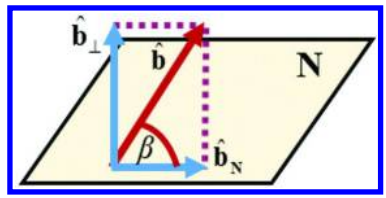

Figure 2. A possible model vector $\hat{\mathbf{b}}$ relative to the N and its associated projections into ($\hat{\mathbf{b}}_N$) and orthogonal ($\hat{\mathbf{b}}_\perp$) to N. The angle between $\hat{\mathbf{b}}$ and $\hat{\mathbf{b}}_N$ is labeled β.

An estimated model vector contains information related to the NAS vector $\mathbf{x}_\perp$ pointing, in principle, in the same direction as $\hat{\mathbf{b}}_\perp$, the model vector orthogonal to the **N** computed by $\hat{\mathbf{b}}_\perp = \left(\mathbf{I} - \mathbf{V}_k\mathbf{V}_k^{\mathrm{T}}\right)\hat{\mathbf{b}}$. Note that there are many directions the model vector can actually point while still maintaining the same angular relationship with **N** defined by the angle β in Figure 2. A similar statement is true for the direction of $\hat{\mathbf{b}}_N$.

Projection of $\hat{\mathbf{b}}$ into the **N** forming $\hat{\mathbf{b}}_N = \mathbf{V}_k\mathbf{V}_k^{\mathrm{T}}\hat{\mathbf{b}}$ represents the non-NAS component of the model vector and characterizes that part of $\hat{\mathbf{b}}$ not in the direction of $\hat{\mathbf{b}}_\perp = \left(\mathbf{I} - \mathbf{V}_k\mathbf{V}_k^{\mathrm{T}}\right)\hat{\mathbf{b}}$, i.e., $\hat{\mathbf{b}} = \hat{\mathbf{b}}_N + \hat{\mathbf{b}}_\perp$. From the NAS geometry in Figure 2, the angular relationship between $\hat{\mathbf{b}}$ and $\hat{\mathbf{b}}_N$ can be expressed as

$$\cos\beta = \frac{\left\|\hat{\mathbf{b}}_N\right\|}{\left\|\hat{\mathbf{b}}\right\|} \tag{6}$$

describing the degree of orthogonality between $\hat{\mathbf{b}}$ and **N** and hence, model selectivity. Because cos β is determined by the tuning parameter value for the calibration modeling method (as well as k), it seems natural to treat zero as the ideal target value and select a model(s) with a small cos β value. Specifically, keeping $\hat{\mathbf{b}}$ largely orthogonal to **N** should cause much of the non-analyte information in a sample spectrum to be zeroed out when multiplied by the regression vector. This interaction stems from writing **x** as (assuming a linear Beer-Lambert law type relationship)

$$\mathbf{x}^{\mathrm{T}} = y_a\mathbf{k}_a^{\mathrm{T}} + \mathbf{1}^{\mathrm{T}}\mathbf{N} + \mathbf{r}^{\mathrm{T}} \tag{7}$$

where y_a and $\mathbf{k}_a$ respectively represent the analyte concentration and pure component analyte spectrum at unit concentration, the **1** is a vector of ones with as many ones as there are spectra in **N**, and **r** denotes the random spectral noise. The non-analyte spectra in **N** represent all components of **x** not due to the analyte such as pure component interferent spectra, spectra characterizing instrumental and/or environmental sources affecting **x** such as scatter, baseline shifts, background, temperature, etc. The spectra in **N** are scaled by the respective quantities in $\mathbf{y}_N$. Prediction of y_a, signified by $\hat{y}_a$, is obtained by multiplying **x** in Equation 7 by a model vector $\hat{\mathbf{b}}$ expressed by

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch009

$$\hat{y}_a = \mathbf{x}^{\mathrm{T}}\hat{\mathbf{b}} = y_a\mathbf{k}_a^{\mathrm{T}}\hat{\mathbf{b}} + \mathbf{1}^{\mathrm{T}}\mathbf{N}\hat{\mathbf{b}} + \mathbf{r}^{\mathrm{T}}\hat{\mathbf{b}} \quad (8)$$

In order to obtain $\hat{y}_a = y_a$ from Equation 8, three conditions must be fulfilled. Specifically, $\mathbf{k}_a^{\mathrm{T}}\hat{\mathbf{b}} = 1$, $\mathbf{N}\hat{\mathbf{b}} = \mathbf{0}$, and $\mathbf{r}^{\mathrm{T}}\hat{\mathbf{b}} = 0$. In most situations, not all three conditions can be simultaneously satisfied. As noted following, the level of success depends on the sample specific situation.

Neglecting the noise component **r**, Equation 8 reveals that predictions are based on the sum of the products $\mathbf{N}\hat{\mathbf{b}}$ and $\mathbf{k}_a^{\mathrm{T}}\hat{\mathbf{b}}$. When the model vector is orthogonal to the **N**, then for each spectrum in **N**

$$\mathbf{n}^{\mathrm{T}}\hat{\mathbf{b}} = \|\mathbf{n}\|\|\hat{\mathbf{b}}\|\cos\theta = 0 \quad (9)$$

where θ is the angle between an **n** and $\hat{\mathbf{b}}$ and hence, there are no contributions from **N** to the predicted analyte value $\hat{y}_a$. The predicted analyte value is now based on how well Equation 10 holds

$$\mathbf{k}_a^{\mathrm{T}}\hat{\mathbf{b}} = \|\mathbf{k}_a\|\|\hat{\mathbf{b}}\|\cos\varphi = 1 \quad (10)$$

where φ is the angle between $\mathbf{k}_a$ and $\hat{\mathbf{b}}$. Any deviation of the model vector from orthogonality to **N** will contribute to the predicted value in Equation 8. In previous work with simple simulated spectral situations, conditions expressed in Equations 9 and 10 were found to be true (*6*). However, with a real data set, the conditions were not generally met and the final predicted values for the analyte did depend on the degree of orthogonality between $\hat{\mathbf{b}}$ and **N**.

Deviations from orthogonality can be characterized by cos β in Equation 6. However, focusing only on cos β to select a tuning parameter was found to not fully assess the selectivity/sensitivity tradeoff to obtain acceptable predictions errors with the selected model (*6*). Instead, better model selection for previously studied data sets was achieved by selecting those models yielding a comparatively small $\|\hat{\mathbf{b}}_{\mathbf{N}}\|$ value in combination with favorable values for other critical measures of model quality such as small values for $\|\hat{\mathbf{b}}\|$ and cos β. Results showing model selection with these measures are expanded upon in this paper with additional development and application to sample-wise target models as described next.

Augmenting Ridge Regression (RR) with Sample-Wise NAS Target Modeling and Localized NAS Diagnostics

Employing RR is a common method of solving the ill-posed mathematical problem that confounds multivariate calibration. With RR, the matrix of **X** typically contains far more columns (measured variables) than rows (distinct samples). The resulting rank deficiency of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ causes issues when attempting to take the corresponding inverse. By adding a controlled amount of diagonalization to an otherwise singular calibration matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, RR facilitates solution of the inverse operation by

$$\hat{\mathbf{b}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda^2\mathbf{I}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \quad (11)$$

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch009

where **I** is the $n \times n$ identity matrix and the scalar λ is an adjustable value that controls the weight applied to the identity matrix to stabilize the inverse operation and simultaneously, penalize the magnitude of **b** as noted in the minimization expression

$$\min\left(\left\|\mathbf{Xb}-\mathbf{y}\right\|^2+\lambda^2\left\|\mathbf{b}\right\|^2\right) \tag{12}$$

with $0 \le \lambda < \infty$. Expression 12 functions as the least squares criterion for

$$\begin{pmatrix}\mathbf{y}\\ \mathbf{0}\end{pmatrix}=\begin{pmatrix}\mathbf{X}\\ \lambda\mathbf{I}\end{pmatrix}\mathbf{b} \tag{13}$$

The two terms in Expression 12 represent measures of bias and variance respectively. Attempting to minimize the two values invokes a method of trade-off analysis.

Explored in this paper are attempts to exploit sample-wise NAS specific information to target the RR algorithm toward the NAS vector of a new unknown sample **x**. The method developed in this paper differs from that previously published (*7*, *25*), albeit the goal is the same, i.e., targeting a prior known vector for the model. For a particular sample spectrum, this sample-wise orthogonal NAS target would be $y\frac{\mathbf{x}_\perp}{\left\|\mathbf{x}_\perp\right\|^2}$. Because y is not known for a new sample, a tuning parameter ζ was used to replace y to control the magnitude of the target NAS. Additionally, since the orthogonality of $\hat{\mathbf{b}}$ relative to **N** may not necessarily be the best for prediction of a particular sample, the solution $\hat{\mathbf{b}}$ is computed by

$$\hat{\mathbf{b}}=\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda^2\mathbf{I}\right)^{-1}\left(\mathbf{X}^{\mathrm{T}}\mathbf{y}+\lambda^2\zeta\frac{\mathbf{x}_\rho}{\left\|\mathbf{x}_\rho\right\|^2}\right) \tag{14}$$

thus targeting the output of RR ($\hat{\mathbf{b}}$) directly towards NAS information by adhering to the minimization expression

$$\min\left(\left\|\mathbf{Xb}-\mathbf{y}\right\|^2+\lambda^2\left\|\mathbf{b}-\zeta\frac{\mathbf{x}_\rho}{\left\|\mathbf{x}_\rho\right\|^2}\right\|^2\right) \tag{15}$$

for the set of equations

$$\begin{pmatrix}\mathbf{y}\\ \lambda\zeta\dfrac{\mathbf{x}_\rho}{\left\|\mathbf{x}_\rho\right\|^2}\end{pmatrix}=\begin{pmatrix}\mathbf{X}\\ \lambda\mathbf{I}\end{pmatrix}\mathbf{b} \tag{16}$$

In using Equation 14 as the solution to Expression 15, there are essentially four tuning parameters. The λ and ζ parameters are obvious, but also needed are ρ and k to calculate $\mathbf{x}_\rho$ using Equation 4. Without prior knowledge or empirical evidence, it is difficult to decide on particular parameter values. In theory, all four

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch009

parameters should be specific to each sample depending on the amount of analyte in the sample relative to the non-analyte and how well **N** spans the non-analyte signal specific to the sample. Described next is the NAS geometry studied to aid in selecting these tuning parameters. The method of PLS can be used by removing the λ term and using LVs instead.

Assuming that $\left\|\hat{\mathbf{b}}_{\mathbf{N}}\right\|$ assesses how much of $\hat{\mathbf{b}}$ lies in the **N** space, a reasonable local sample-wise analog of $\left\|\hat{\mathbf{b}}_{\mathbf{N}}\right\|$ would represent the portion of $\hat{\mathbf{b}}$ that lies within a subspace of **N** specific to each sample. Equation 5 defined $\mathbf{x}_{\mathbf{N}}$ as the NnAS for a sample. The vector $\mathbf{x}_{\mathbf{N}}$ should ideally distinguish the local subspace of **N** for that sample. Projection of $\hat{\mathbf{b}}$ onto this vector can then be considered as that portion of $\hat{\mathbf{b}}$ related to the local non-analyte information identified and computed by

$$\hat{\mathbf{b}}_{\mathbf{x}_{\mathbf{N}}} = \left(\frac{\mathbf{x}_{\mathbf{N}}\mathbf{x}_{\mathbf{N}}^{\mathrm{T}}}{\mathbf{x}_{\mathbf{N}}^{\mathrm{T}}\mathbf{x}_{\mathbf{N}}}\right)\hat{\mathbf{b}} \tag{17}$$

Figure 3 illustrates the projection. The vector L_2 norm $\left\|\hat{\mathbf{b}}_{\mathbf{x}_{\mathbf{N}}}\right\|$ expresses the localized relationship between $\hat{\mathbf{b}}$ and $\mathbf{x}_{\mathbf{N}}$. Similarly, a sample-wise analog of the angle β between $\hat{\mathbf{b}}$ and $\hat{\mathbf{b}}_{\mathbf{x}_{\mathbf{N}}}$ can be obtained.

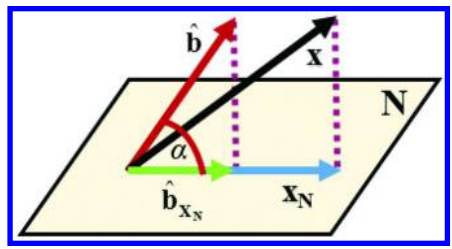

Figure 3. Projection of $\hat{\mathbf{b}}$ onto $\mathbf{x}_{\mathbf{N}}$ forming $\hat{\mathbf{b}}_{\mathbf{x}_{\mathbf{N}}}$ as the portion of the model vector that lies in the local non-analyte space specific to sample x.

This angle will be referred to as α with the intent to guide the process of sample-wise model selection through minimization of cos α computed by

$$\cos\alpha = \frac{\left\|\hat{\mathbf{b}}_{\mathbf{x}_{\mathbf{N}}}\right\|}{\left\|\hat{\mathbf{b}}\right\|} \tag{18}$$

with the goal of orthogonalizing the two vectors. In addition to studying cos α, trends for $\left\|\hat{\mathbf{b}}_{\mathbf{x}_{\mathbf{N}}}\right\|$ are investigated as tuning parameters vary.

From NAS theory and the model selection goals for this paper, it seems reasonable to minimize the inner product between $\mathbf{x}_{\mathbf{N}}$ and $\hat{\mathbf{b}}$ by seeking a small value for $\left|\mathbf{x}_{\mathbf{N}}^{\mathrm{T}}\hat{\mathbf{b}}\right|$. This can be equivalently expressed by $\left|\mathbf{x}_{\mathbf{N}}^{\mathrm{T}}\hat{\mathbf{b}}\right| = \left|\left\|\mathbf{x}_{\mathbf{N}}\right\|\left\|\hat{\mathbf{b}}\right\|\cos\alpha\right|$ showing that there are multiple ways in which the inner product can be small.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch009

Tradeoffs in Model Attributes and L-Shaped Curves

Requiring $\hat{\mathbf{b}}$ to be orthogonal to either $\mathbf{N}$ (for a global model) or $\mathbf{x}_N$ (for a sample-wise local model) might seem ideal, but this forced orthogonality can be an unfair constraint and the model vector is actually regulated by the tuning parameter to have a direction and magnitude that is most useful for prediction. Thus, it is prudent to analyze the tradeoff between orthogonality and other key diagnostic measures opposing orthogonality that maintain prediction error assessment value.

Analysis of tradeoffs between model attributes has been shown to aid in model selection (*3*, *5–11*) and the NAS-derived fitness measures can be used to enhance this process. For example, the tradeoff between RMSEC and the model vector L_2 norm is well characterized (*7–10*). Recall from the previous discussion that RMSEC directly quantifies model bias while the model vector L_2 norm assesses the potential prediction variance. Because the two diagnostic measures are in direct competition with each other (the bias/variance tradeoff), a plot of one against the other is expected to form an L-shaped curve. Model selection can be performed by choosing the model closest to the origin at the point of maximum curvature. This area resides in the corner region of the L-shaped curve. However, model selection from the standard L-curve can be somewhat subjective. Despite this subjectivity, the standard L-curve by itself has been used to obtain models with sufficient prediction accuracy. This paper evaluates the NAS measures used as selection criteria in conjunction with standard L-curves. As shown in Results and Discussion section, the NAS measures also tend to form L shaped curves when plotted similarly.

Experimental

Calibration

The NAS model measures are obtained from leave multiple out cross validation (CV) using 100 random splits with 70% for calibration and 30% for validation. Calibration sets are mean centered and the corresponding validation set is mean centered to the calibration set mean. The $\mathbf{N}$ matrix is mean centered to the mean of $\mathbf{N}$. Mean values of the NAS and model indicators from the CV are reported. The singular value decomposition (SVD) is used for obtaining the $\mathbf{V}$ eigenvectors from respective data set specific $\mathbf{N}$ matrices. Assessing the NAS measures requires determination of an appropriate value for k (number of eigenvectors to retain for projections relative to $\mathbf{N}$). Data set specific values for k are such that over 98% of the variation in $\mathbf{N}$ is retained and values used for k are reported in respective data set descriptions. Eighty λ tuning parameters are used for RR with specific values given in respective data set descriptions. Similarly, the maximum number of PLS latent variables (LVs) evaluated are provided in respective data set descriptions. The sample-wise NAS target method is tested with RR and PLS using only the temperature data set. Fifty ζ and ρ values are used with actual values described in the temperature data set section. In this case, no CV is used as described in the temperature data set section.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch009

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch009

Temperature Data

A three component temperature data set comprising 22 samples of varying mole fractions of ethanol, water, and isopropanol measured over 200 wavelengths from 850 to 1049 nm was analyzed using ethanol as the analyte (*26*, *27*). The data set is available at five temperatures and 30 ºC was used. Three of the 22 samples are pure component spectra as shown in Figure 4a. The two pure component interferent spectra and one blank sample (containing only water and isopropanol) are used for **N**. Two eigenvectors from the SVD of **N** are used for respective projections relative to **N**. These three non-analyte samples are not part of the calibration set. The pure component analyte spectrum was also not used for calibration. The remaining spectra are used two ways. One consists of a non-CV situation based on the literature with 13 and 6 samples in the calibration and validation sets, respectively. This data split was used for the sample-wise NAS target method. The 50 ζ and ρ values used for this process ranged from 0 to 1 in increments of 0.0204. The second way the data set was used involves the CV process previously described in the Calibration section. In both cases, the 80 λ values increased exponentially from 1.0×10^{-4} to 20.

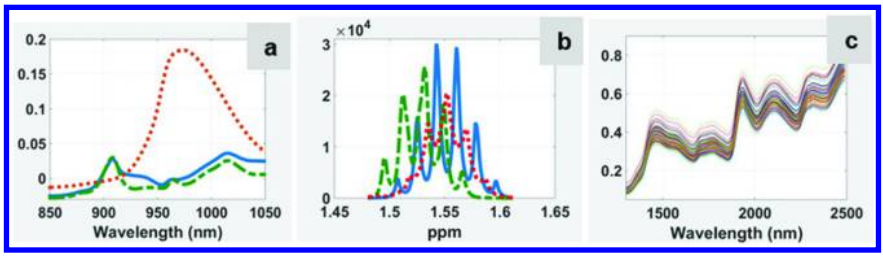

Figure 4. Pure component spectra for (a) temperature data with ethanol (solid), isopropanol (dot-dash), and water (dot); (b) NMR with propanol (solid), butane (dot-dash), and pentanol (dot); and (c) the corn data set.

Nuclear Magnetic Resonance Data

The nuclear magnetic resonance (NMR) data set comprises 231 samples of a three component mixture of propanol, butanol, and pentanol measured from 0.6425 ppm to 3.8431 ppm at 2.3×10^{-4} ppm increments for 14,000 response variables (*27*, *28*). The spectral region was reduced from 1.4814 ppm to 1.6099 ppm using every fifth response for a total of 113 responses. This region was chosen because of the significant overlap among the signals for all three components shown in Figure 4b. Propanol is presented as the analyte. Sample concentrations for each component ranged from 0 to 100 %. In addition to pure component spectra, each component has 21 blank samples with respect to the other two components. For each component tested as the analyte, the respective two pure component interferent spectra and five blank samples are removed from the 231 sample set. The five blank and two pure component interferent spectra are used to construct respective analyte specific **N** matrices and five eigenvectors are used for respective projections relative to **N**. The seven samples for each **N** ranged

from 0, 5, 25, 50, 75, 95, and 100 % of each non-analyte component. The 80 λ values increased exponentially from 100 to 8.0×10^5 for all three components.

Corn Data

The near infrared (NIR) spectral data set consists of 80 corn samples measured across 700 wavelengths from 1100 to 2498nm at 2nm intervals (*27*, *29*). The data set is reduced to 100 wavelengths from 1302 nm to 2490 nm at 12 nm increments as plotted in Figure 4c. The set includes samples measured from three different instruments and the mp5 instrument is used. Concentration information as percent composition is provided for moisture, oil, protein, and starch. Because there are no pure component spectra or blank samples, spectral differences between two spectra with the same analyte amount within a concentration tolerance are used. For moisture, there are two pairs of spectra that exactly match in moisture content to three decimal places. The remaining samples used for **N** consist of three sample pairs that differ in moisture content by 1×10^{-3}, one pair differing by 3×10^{-3}, and one pair by 5×10^{-3}, for a total of seven difference spectra. The maximum tolerances used to form **N** matrices for oil, protein, and starch are 3.0×10^{-3}, 3.0×10^{-3}, and 9.0×10^{-3}, respectively. Each analyte had an **N** with seven difference spectra. In each case, five eigenvectors from the SVD of **N** are used for projections. The 80 $\boldsymbol{\lambda}$ tuning parameters increased exponentially from 1.0×10^{-8} to 37 for moisture, oil, and starch and from 1.0×10^{-8} to 74 for protein.

Results and Discussion

Global Model Tuning Parameter Selection

NMR Data

Previous work had evaluated the NAS attributes using RR for the temperature data (*6*) and as with that data set, the NMR data has all pure component spectra and non-analyte spectra to select from and form a data set represented by **N**. Plotted in Figure 5a are the inner product between **N** and PLS model vectors with the same for RR in Figure 5b. From these two figures, it is observed that Equation 9 is not absolute and inner products are not always equal to the target zero value for most models. Displayed in Figure 5c are the inner products between the analyte pure component propanol spectrum and the RR and PLS model vectors revealing that this inner product approaches the target value of 1 for Equations 10 and 8. Figure 5d presents the $\left\|\hat{\mathbf{b}}_N\right\|$ plot for RR and PLS. Local minima of these plots identify RR model 15 with tuning parameter value 491.68 and the PLS 12 LV model as potential models to be selected. From the corresponding RMSECV plots in Figure 5e, the selected models in Figure 5d have acceptable bias/variance tradeoffs relative to the underlying selectivity/sensitivity tradeoffs characterized by the $\cos\beta$ plots for RR and PLS in Figure 5e. Thus, as with the temperature data in previous work (*6*), the local minimum in a plot of $\left\|\hat{\mathbf{b}}_N\right\|$ is able select acceptable tuning parameter values as a single criterion. This observation was true for the

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch009

other NMR analytes. In the previous temperature data study with ethanol as the analyte, models tended to be selected by $\|\hat{\mathbf{b}}_N\|$ with greater selectivity at a sacrifice to sensitivity. With propanol, the opposite seems to be occurring with selectivity being sacrificed for sensitivity.

A trend observed across the analytes for this data set is that the minimum of $\|\hat{\mathbf{b}}_N\|$ occurs at the point on the RMSECV curve where no appreciable improvement in RMSECV values can be obtained by further increasing $\|\hat{\mathbf{b}}\|$. This trend is also observed for the temperature data, the corn data set (next), and other data sets.

Corn Data

The corn data does not have any pure component or non-analyte spectra to use for **N**. In this case, spectra with constant or nearly constant analyte are differenced thereby removing the analyte contribution leaving non-analyte related spectra. Using these spectra for **N** as described in the Experimental section, the plots in Figure 6 characterize the NAS measures for moisture as the analyte. For RR and PLS, the respective models at the larger model vector L_2 norms are fairly orthogonal to each of the spectra in **N**, but not exactly. The plot of cos β in Figure 6e shows that relative to the complete space of **N** (as defined by the five eigenvectors used to span **N** from the SVD of **N**), these model vectors with greater $\|\hat{\mathbf{b}}\|$ values are quite orthogonal to **N**. This angular relationship indicates that sensitivity is being sacrificed for improved selectivity. The greater selectivity at the larger model vector L2 norms is confirmed by the plot of $\|\hat{\mathbf{b}}_N\|$ in Figure 6c where the local minima are used to select tuning parameter values for RR and PLS (ridge value and LV, respectively). The local minima of these plots identify RR model 40 with tuning parameter value 4.63×10^{-4} and the PLS 15 LV model as potential models for selection. From the RMSECV plots in Figure 6d, these models produce acceptable prediction errors balancing the bias/variance tradeoff. The global RMSECV minima in Figure 6d occur at RR model 37 with tuning parameter value 1.98×10^{-4} and 21 LVs for PLS.

Sample-Wise NAS Target Modeling

The temperature, NMR, and corn data sets were all studied for sample-wise NAS target modeling using Equations 14 – 16. Results were similar for all three data sets and presented are the temperature data results.

Using NAS indicators for global modeling requires selecting two tuning parameters (ridge value for RR or LVs for PLS and the number of eigenvectors from **N** to form the NAS measures). In that work, the number of eigenvectors was set for at least 98% of the variation in **N** being captured and the NAS attribute $\|\hat{\mathbf{b}}_N\|$ was used to select the model tuning parameter. For sample-wise NAS target modeling, these same two tuning parameters need to be determined.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch009

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch009

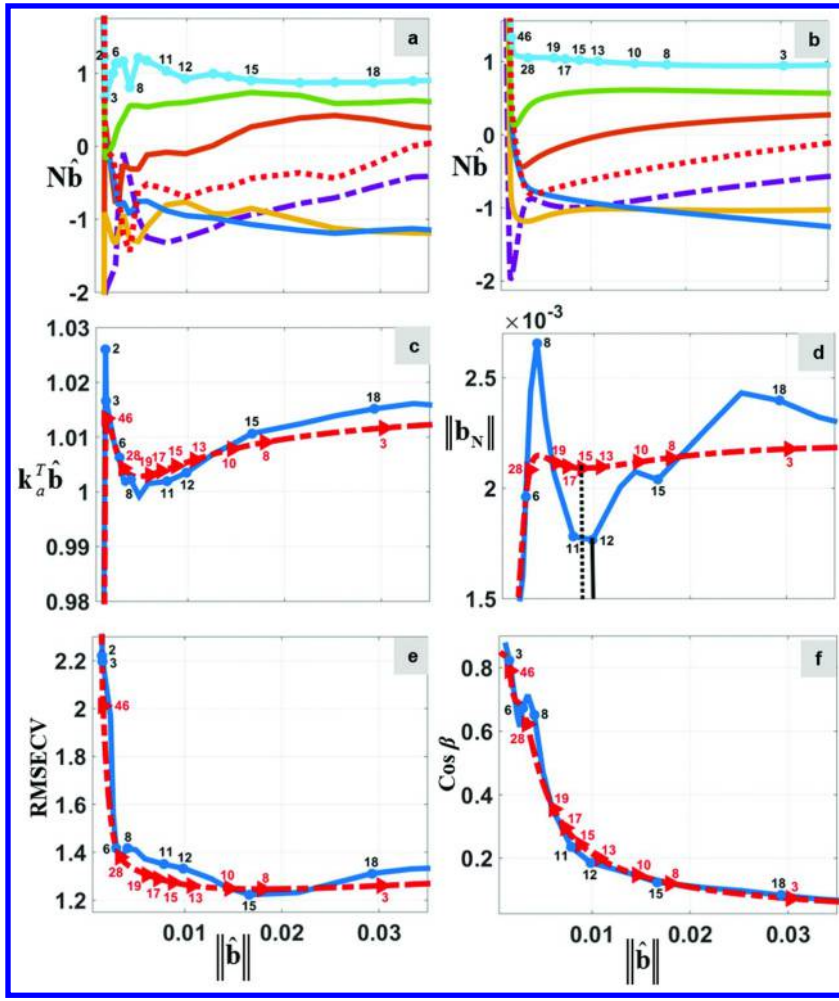

Figure 5. NMR data RR and PLS plots of NAS attributes and RMSECV against model regression vector L_2 norms. Numbers on plots are either the RR model number or the number of PLS LVs. The RR λ values increase in order with RR model numbers. (a) and (b) contain inner products between **N** *and RR and PLS model vectors, respectively, with (dot-dash) and (dot) being the pure component interferent spectra for butanol and pentanol, respectively. For (c) – (f), plot symbols for RR and PLS are (dot-dash) and (solid) respectively. (c) inner products between the analyte propanol pure component spectrum and model vectors. (d)* $\|\hat{\mathbf{b}}_N\|$ *plots with vertical lines marking selected RR (dash) and PLS (solid) models at local minima. (e) RMSECV. (f) cos β.*

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch009

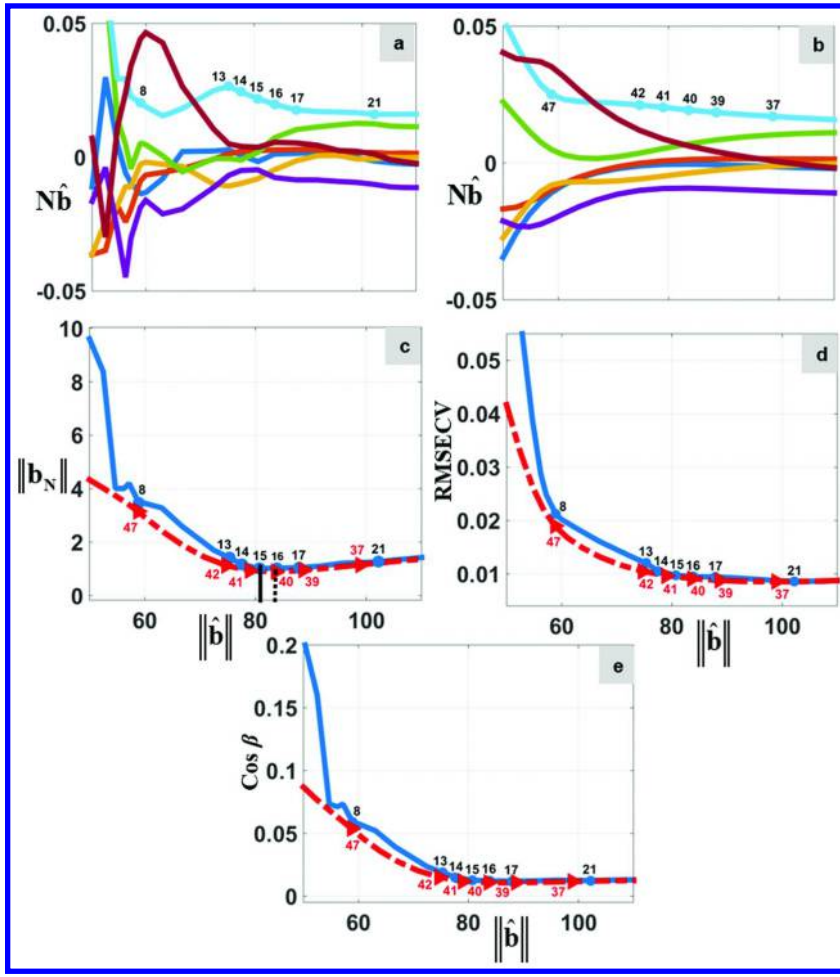

Figure 6. Corn data RR and PLS plots of NAS attributes and RMSECV against model regression vector L_2 norms. Numbers on plots are either the RR model number or the number of PLS LVs. The RR λ values increase in order with RR model numbers. (a) and (b) are, respectively, inner products between N and PLS and RR model vectors. For (c) – (e), plot symbols for RR and PLS are (dot-dash) and (solid) respectively. (c) $\|\hat{\mathbf{b}}_N\|$ plots with vertical lines marking selected RR (dash) and PLS (solid) models at local minima. (d) RMSECV. (e) cos β.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

In addition, values for the tuning parameters ρ (for the degree of orthogonality in the projection relative to **N**) and ζ (for the magnitude of the target vector) need to be selected. Using the same number of eigenvectors (two) as in the global model selection study, the effects of varying ρ and ζ on the NAS measures and predication errors were studied.

Shown in Figures 7 a and b are images of RMSEV values at two ρ values while ζ and the RR tuning parameter λ vary. At ρ = 1, the RMSEV image in Figure 7b shows localized regions with small RMSEV values while the ζ and λ values vary. This complexity suggests the difficulty expected in selecting values for λ, ρ, and ζ at a fixed number of eigenvectors. However, as ρ decreases from 1 to 0.003, the RMSEV values become essentially independent of ζ, i.e., the variation of λ is the same across the ζ values in Figure 7a. This independence of ζ is further confirmed from graphical analysis of the images for the minimization term $\left\| \mathbf{b} - \zeta \frac{\mathbf{x}_\rho}{\|\mathbf{x}_\rho\|^2} \right\|$ in Expression 15 shown in Figures 7c and d. The images and plots in Figures 7a-d deviate very little as ρ goes to zero. At these small values for ρ, there is effectively no projection and the NAS target is the validation sample itself. The algorithm is apparently focusing on the NAS most useful for prediction and not necessarily the orthogonal NAS. Local NAS measures $\|\hat{\mathbf{b}}_{\mathbf{x}_N}\|$ and cos α, (the local analogies to $\|\hat{\mathbf{b}}_N\|$ and cos β), have the same characteristics as Figure 7a-d. Analogous results were obtained for PLS by replacing the RR tuning parameter with LVs. At the small ρ values, consistent L-curves at each ζ value can be obtained. The goal was for these L-curves to correlate with prediction error in the corner region as with L-curves for global models allowing selection of λ independent of ρ and ζ. However, these L shaped curves do not always correlate with prediction errors of each validation sample for selection of λ or LV values (true for the sample in Figure 7). Thus, using the sample-wise NAS target approach was not feasible in the current format due to the inconsistencies of the results.

When ζ was fixed to specific values and λ and ρ were allowed to vary, results similar to those in Figure 7 were obtained for RR (as well as for PLS). Additionally, as ζ converges to zero, the sample-wise NAS target modeling process reduces to RR or PLS as the case may be. In this situation, the local minima in $\|\hat{\mathbf{b}}_N\|$ can be used to select the global model as in the previous section.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch009

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch009

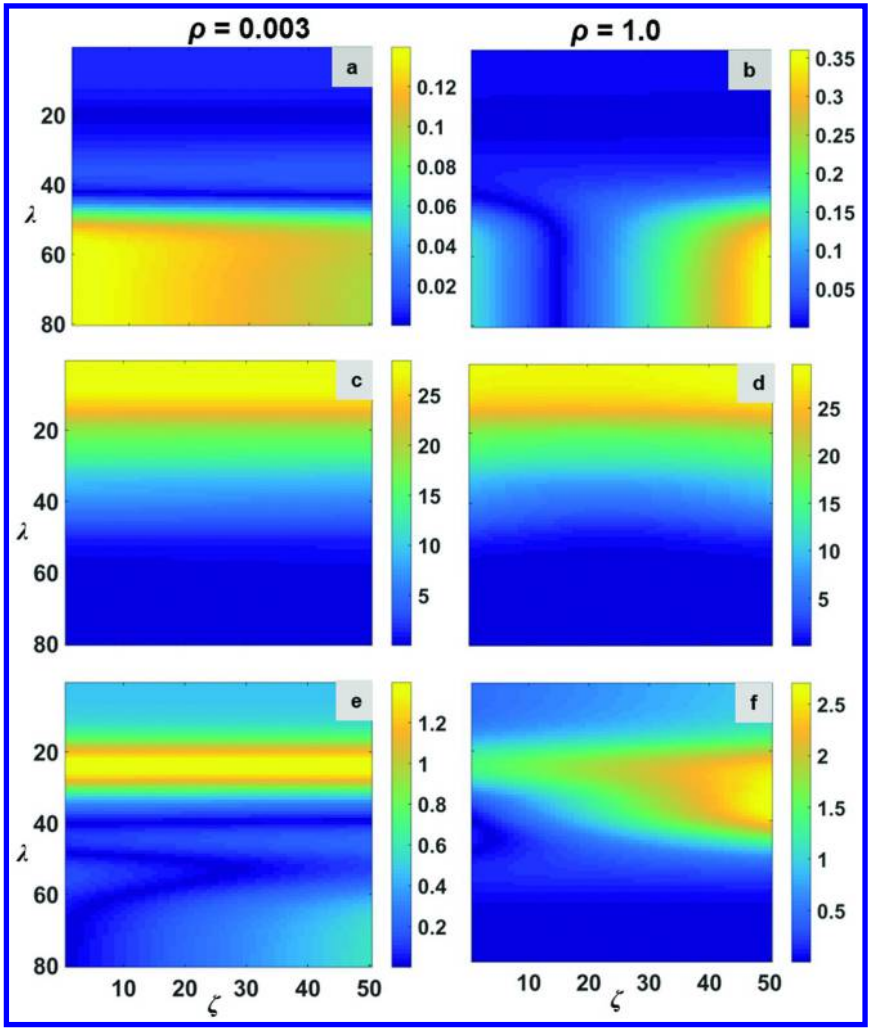

Figure 7. Temperature data RR images of model and NAS measures for ρ = 0.003 on the left and ρ = 1 on the right. The λ and ζ values increase in order of respective axis indices (tuning parameter numbers). (a) and (b) are RMSEV. (c) and (d) are $\left\| \mathbf{b} - \zeta \frac{\mathbf{x}_\rho}{\|\mathbf{x}_\rho\|^2} \right\|$. *(e) and (f) are* $\left\| \hat{\mathbf{b}}_{\mathbf{x}_N} \right\|$.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch009

The difficulty in using a sample-wise NAS target approach probably resides in the need for a local non-analyte set of sample, i.e., the discrete relationships of each new sample relative to the global **N** must be reconciled. Perhaps it is no surprise that NAS model selection methods should begin to fail when attempting to seek localized answers from an algorithm that is based on global non-analyte information. An analogy is local calibration based on selecting calibration samples best for a particular new sample. Increased model prediction accuracy can be obtained compared to a global model by matching the new sample to a subset of calibration samples closely matrix matched to the new sample. Similarly, for sample-wise NAS target modeling to function in a local capacity for a certain sample, the space of **N** should be selected to closely match the new sample as well as any environmental and instrumental conditions present when the new sample is measured. In the current study, **N** represents an aggregate of known types of non-analyte information anticipated in all new samples. However, not every sample can be expected to contain all non-analyte information spanned by **N**. Therefore, the projection of **x** into the full **N** space represents an unrealistic interpretation of the sample-wise non-analyte signal $\mathbf{x}_N$. Consequently, the angle α between **b** and $\mathbf{x}_N$ is actually the angle between the model vector and a pseudo-global non-analyte space that probably does not represent the actual non-analyte space of any one sample. This conclusion is based on observations that some samples seem to perform extremely well under the sample-wise NAS target approach while others are quite difficult to predict accurately, such as in Figure 7.

Advancing NAS Modeling

It was hoped to be possible to select sample-wise models by reducing the number of tuning parameters through setting the ρ tuning parameter to essentially zero in order to obtain L shaped curves correlated to sample-wise prediction errors as a function of λ. Unfortunately, this worked for some samples, assumed to be appropriately characterized by **N**, but not for other samples. Hence, the conclusion that in order to work best, one needs a localized **N** particular to each new sample. However, it may still be possible to select sample-wise models using a global based **N** by using a process that allows multiple tuning parameters to be optimized simultaneously. In recent work, the method of sum of ranking differences (SRD) was used to select respective RR and PLS tuning parameters based on an ensemble of multiple model fitness criteria (*9*). Recent work has advanced SRD to select two tuning parameters (*30*). Such an approach is being investigated for selecting values for λ, ρ, and ζ as the number of eigenvectors vary, but even this may prove to be to much.

Because the best $\mathbf{x}_\rho$ is not known, a study was undertaken to estimate it in a post global model analysis by projecting **x** onto each respective tuning parameter based RR and PLS $\hat{\mathbf{b}}$ using $\hat{\mathbf{x}}_\rho = \hat{\mathbf{b}}\hat{\mathbf{b}}^{+}\mathbf{x} = \dfrac{\hat{\mathbf{b}}\hat{\mathbf{b}}^{\mathrm{T}}}{\left\|\hat{\mathbf{b}}\right\|^2}\mathbf{x}$. If the model vector is completely orthogonal to **N**, then ρ =1 and $\hat{\mathbf{x}}_\rho = \mathbf{x}_\perp$, the orthogonal NAS of **x**. Substituting $\hat{\mathbf{x}}_\rho$ for $\mathbf{x}_\rho$ in Equation 4 allows solution for the corresponding value

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

of ρ. Specifically, $\hat{\mathbf{x}}_\rho = \mathbf{x} - \rho\mathbf{V}\mathbf{V}^{\mathrm{T}}\mathbf{x} = \mathbf{x} - \rho\mathbf{x}_{\mathrm{N}}$ and defining $\mathbf{d} = \mathbf{x} - \hat{\mathbf{x}}_\rho$, the solution for ρ becomes $\rho = \dfrac{\mathbf{x}_{\mathrm{N}}^{\mathrm{T}}\mathbf{d}}{\|\mathbf{x}_{\mathrm{N}}\|^2}$. Plotted in Figure 8 are the RR and PLS mean values of ρ across all calibration samples against the respective $\|\hat{\mathbf{b}}\|$ for the NMR data. At the models selected by the local minima of $\|\hat{\mathbf{b}}_{\mathrm{N}}\|$, the values are $\rho_{\mathrm{RR}} = 0.986$ and $\rho_{\mathrm{PLS}} = 0.990$ indicating that the selected models are nearly orthogonal to **N**. This agrees with the plot in Figure 5f. The plots in Figure 8 also show that as the sensitivity degrades with increasing values of $\|\hat{\mathbf{b}}\|$, the selectivity (orthogonality) of $\mathbf{x}_\rho$ increases. Similar results to Figure 8 were obtained for the NMR validation data as well as for the corn and temperature data.

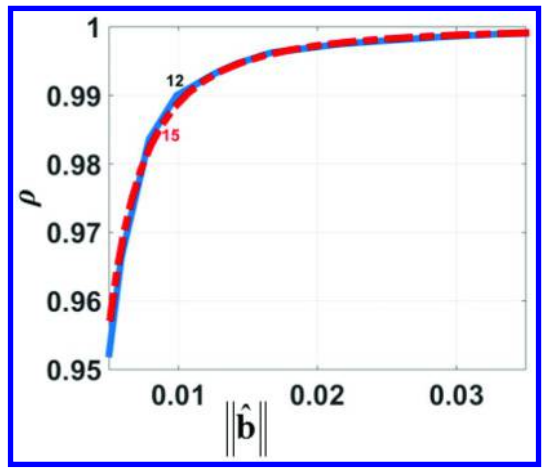

Figure 8. NMR calibration data for RR (dot-dash) and PLS (solid) plots of ρ against model regression vector L_2 norms. Numbers on plots are either the RR model number or the number of PLS LVs. The RR λ values increase in order with RR model numbers.

Current work also involves adjusting the sample-wise target concept to an NAS global target model using SRD for model selection. Specifically, rather than using the sample-wise target $\zeta \dfrac{\mathbf{x}_\rho}{\|\mathbf{x}_\rho\|^2}$ as in Expression 15, the target $\dfrac{\bar{\mathbf{s}}_\rho}{\|\bar{\mathbf{s}}_\rho\|^2}$ can be used where $\bar{\mathbf{s}}_\rho$ denotes the mean NAS vector at unit concentration where the mean is computed from the projection operation $\mathbf{s}_\rho = \left[\left(\mathbf{I} - \rho\mathbf{V}_k\mathbf{V}_k^{\mathrm{T}}\right)\mathbf{x}\right] / y$ for each calibration sample with $\mathbf{V}_k$ defined as before. In this way the ζ tuning parameter is removed and a calibration model is sought in the mean NAS direction. With SRD, it will also be possible to use the individual NAS vectors rather than the mean vectors. An uncertainty is if the **N** matrix should only be used to generate the NAS measures clarifying the degree of orthogonality (model direction) for model

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch009

selection or whether it should also be included in calculation of the regression vector. For example, the minimization expression

$$\min\left(\left\|\mathbf{Xb}-\mathbf{y}\right\|^2+\eta^2\left\|\mathbf{Nb}\right\|^2+\lambda^2\left\|\mathbf{b}-\frac{\overline{\mathbf{s}}_\rho}{\left\|\overline{\mathbf{s}}_\rho\right\|^2}\right\|^2\right) \quad (19)$$

could be solved. Expression 19 without the third penalty target term has been used in previous work (*21*, *31*). There are many vectors that satisfy the second penalty term (orthogonality to **N** for model direction). Including the third term for a specific target vector direction may assist the algorithm to compute a more acceptable solution with improved tradeoffs. Such work is ongoing in our laboratory.

In this paper, the non-analyte matrix **N** was always measured at the same measurement conditions as the calibration samples for forming a global model. Recent work has shown that **N** can also span a set of sample conditions (secondary conditions) differing from those for the calibration samples in **X** (primary conditions) (*21*, *31*). This approach is also under investigation with the new NAS measures and NAS target regression methods.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch009

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. CHE-1111053 (co-funded by MPS Chemistry and the OCI Venture Fund) and is gratefully acknowledged by the authors.

References

1. Næs, T.; Isaksson, T.; Fern, T.; Davies, T. *A User Friendly Guide to Multivariate Calibration and Classification*; NIR Publications: Chichester, U.K., 2002.
2. Hastie, T. J.; Tibshirani, R. J.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed; Springer-Verlag: New York, NY, 2009.
3. Kalivas, J. H. In *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*; Brown, S. D., Tauler, R., Walczak, B., Eds.; Elsevier: Amsterdam, 2009; Vol. 3, pp 1−32.
4. Höskuldsson, A. *Chemom. Intell. Lab. Syst.* **1996**, *32*, 37–55.
5. Green, R. L.; Kalivas, J. H. *Chemom. Intell. Lab. Syst.* **2002**, *60*, 173–188.
6. Kalivas, J. H.; Palmer, J. *J. Chemom.* **2015**, *28*, 347–357.
7. Hansen, P. C. *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*; SIAM: Philadelphia, PA, 1998.
8. Gowen, A. A.; Downey, G.; Esquerre, C.; O'Donnell, C. P. *J. Chemom.* **2011**, *25*, 375–381.
9. Kalivas, J. H.; Héberger, K.; Andries, E. *Anal. Chim. Acta* **2015**, *869*, 21–33.
10. Forrester, J. B.; Kalivas, J. H. *J. Chemom.* **2004**, *18*, 372–384.

11. Pinto, L. A.; Galvão, R. K. H.; Araújo, M. C. U. *Anal. Chim. Acta* **2010**, *682*, 37–47.
12. Booksh, K.; Kowalski, B. R. *Anal. Chem.* **1994**, *66*, 782A–791A.
13. Lorber, A.; Faber, K.; Kowalski, B. R. *Anal. Chem.* **1997**, *69*, 1620–1626.
14. Ferré, J.; Faber, N. M. *Chemom. Intell. Lab. Syst.* **2003**, *69*, 123–136.
15. Ferré, J.; Brown, S. D.; Rius, F. X. *J. Chemom.* **2001**, *15*, 537–553.
16. Bro, R.; Anderson, C. M. *J. Chemom.* **2003**, *17*, 646–652.
17. Nadler, B.; Coifman, R. R. *J. Chemom.* **2005**, *19*, 45–54.
18. Goicoechea, H. C.; Olivieri, A. C. *Chemom. Intell. Lab. Syst.* **2001**, *56*, 73–81.
19. Brown, C. D.; Green, R. L. *Trends Anal. Chem.* **2009**, *28*, 506–514.
20. Brown, C. D. *Anal. Chem.* **2004**, *76*, 4364–4373.
21. Andries, E.; Kalivas, J. H. *J. Chemom.* **2013**, *27*, 126–140.
22. Shi, Z.; Cogdill, R. P.; Martens, H.; Anderson, C. A. *J. Chemom.* **2010**, *24*, 288–299.
23. Wold, S.; Antti, H.; Lingren, F.; Feld, M. S. *Anal. Chem.* **2007**, *79*, 234–239.
24. Ni, W.; Brown, S. D.; Man, R. *Chemom. Intell. Lab. Syst.* **2009**, *98*, 97–107.
25. Shih, W. C.; Bechtel, K. L.; Feld, M. S. *Anal. Chem.* **2007**, *79*, 234–239.
26. Wülfert, F.; Kok, W. T.; Smilde, A. K. *Anal. Chem.* **1998**, *70*, 1761–1767.
27. http://www.models.life.ku.dk/datasets, accessed March 15, 2015.
28. Winning, H.; Larsen, F. H.; Bro, R.; Engelsen, S. B. *J. Magn. Resonance* **2008**, *190*, 26–32.
29. Eigenvector Research Incorporated, Manson, Washington. http://www.eigenvector.com/data/Corn/index.html, accessed March 15, 2015.
30. Tencate, A.; White, A.; Kalivas, J. H. In preparation.
31. Ottaway, J.; Farrell, J.; Kalivas, J. H. *Anal. Chem.* **2013**, *85*, 1509–1516.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch009

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch010

Chapter 10

Adaptive Regression via Subspace Elimination

Joshua Ottaway, Joseph P. Smith, and Karl S. Booksh*

Department of Chemistry and Biochemistry, University of Delaware, Newark, Delaware 19716
***E-mail: kbooksh@udel.edu**

The two primary goals in the creation of any multivariate calibration model are effectiveness and longevity. A model must predict accurately and be able to do so over an extended period of time. The primary reason models fail when applied to future samples is the presence of uncalibrated interferents. Uncalibrated interferents represent any change in the future samples not described by the original calibration set. Most often uncalibrated interferents result from the addition of chemical constituents, changes to analytical instrumentation, and changes to environmental conditions. Presented within this chapter is a novel algorithm, Adaptive Regression via Subspace Elimination, for handling uncalibrated chemical interferents via an adaptive variable selection approach. Results are presented for synthetic Near Infrared (NIR) and Infrared (IR) data.

Introduction

Multivariate calibration models are a widely accepted means of quantitatively determining analyte properties, such as pH, oxidation state, and most commonly, concentration (*1*, *2*). In spectroscopy, calibration models consist of a set of observed spectra with known reference values. These observed spectra describe the calibration space as well as the set of all instrumental, environmental, and chemical effects captured by the spectra. These calibration spectra and associated reference values often require extensive laboratory time and expense to collect. Due to this time and expense, many methods for updating an existing model to predict future spectra with uncalibrated interferents have been developed (*3–11*).

Calibration maintenance studies have been the subject of review articles (*12*). These methods typically fall into two broad categories: robust model

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

building and model updating (*13–23*). The goal of any robust modeling approach is to construct the calibration set such that it spans all possible future chemical, environment, and instrumental interferents that may appear in future samples. This proves impractical in the majority of situations for two reasons. First, it is extremely difficult to predict all the possible interferents that may appear in future samples. Second, even if all possible future interferents were known, the calibration set would have to be constructed containing all said interferents in a sufficiently robust experimental design. Thus, the size of such a calibration set could quickly expand well past a practical size.

The other broad category of calibration maintenance is to update the calibration set. This can be accomplished by collecting or creating a large number of new samples containing the uncalibrated interferent and subsequently augmenting the original calibration set. Alternatively, updating the calibration set can be done by collecting a few samples and weighting those appropriately when augmenting the calibration set. These approaches have been well reviewed in literature but suffer the drawback of requiring reference values for those new samples (*12*).

Additionally, both strategies suffer from an expanding 'interferent space'. Including more interferents in the experimental design decreases the net analyte signal (NAS). From the standpoint of NAS, an optimal experimental design would include only the interferents present in a future sample. Decreasing the NAS degrades the noise handling properties of the multivariate model. When the NAS becomes sufficiently small, the ability to reliably estimate properties of future samples is lost.

The goal of this new calibration maintenance process, Adaptive Regression via Subspace Elimination (ARSE), is not to update a calibration set but rather eliminate the contribution of the uncalibrated interferent. By determining the set of variables in the test spectrum that have a contribution from the uncalibrated interferent, those variables can be eliminated. This updated calibration set, which eliminated the contaminated variables, can be reanalyzed to construct a new calibration model. Effectively, ARSE is trading bias for an increase in variance. The variables most biased by the uncalibrated interferent are identified and eliminated. The remaining subset of variables has consequently diminished the capacity to average the effects of random errors.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch010

Mathematics and the Approach

The ARSE algorithm is based on the assumption that any uncalibrated interferent (UI) will contaminate a subset of the variables in a future sample. For instance, if $\mathbf{S}_c$ represents the multivariate space described by the variables in the calibration set, then a future contaminated sample, $\mathbf{x}_f$, can be described as the set of variables that lie within $\mathbf{S}_c$, $\mathbf{x}_{fa}$, and the set of variables that would lie within $\mathbf{S}_c$ if not for the contribution of the UI, $\mathbf{x}_{fb}$, as seen in equation 1.

$$\boldsymbol{x_f = x_{fa} \subseteq S_c + (x_{fb} \subseteq S_c + UI)} \quad \textbf{Eq. 1}$$

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

This then becomes a combinatorics problem. If, for example, the calibration set is described by 40 variables and a new sample has an uncalibrated interferent contaminating 30 of those variables, one 10-variable subspace that is not contaminated by the interferent exists out of 8 * 10^8 possible 10-variable subspaces. Not only is this computationally impractical for an exhaustive search, but it is also dependent on knowledge concerning exact amounts of contaminated variables.

However, in many spectroscopic applications, observed interferent spectra have contributions at many spectroscopically informative wavelengths in a calibration model. Furthermore, the probability of finding a range of wavelengths that are uncontaminated is low. For this reason, methods like secured principal component regression (s-PCR) (*24*) have not seen much practical utility.

To overcome the problems of a large search space with too few uncontaminated variables, ARSE was preformed following transformation of all spectra into the wavelet domain. As described in Goswami et al. this is a frequency and phase transformation that preserves all the information contained within the spectra (*25*). This transform significantly increase the number of variables unique to the analyte of interest with respect to the uncalibrated interferent (Figure 1). ARSE was then preformed on a finite number of random subsets of variables. This allowed the computational issues to be avoided while also evaluating the usefulness of each variable in combination with many different variables.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch010

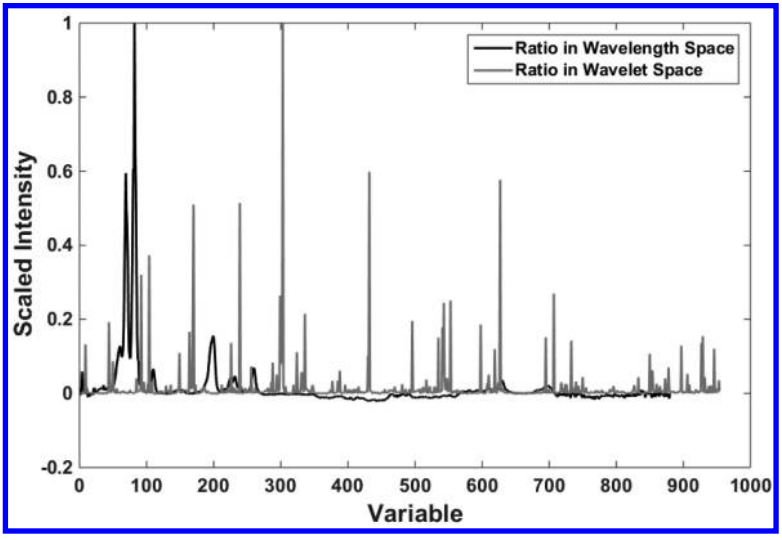

Figure 1. Ratio of analyte of interest pure component to uncalibrated interferent pure component for data set 1 in both the wavelength and absolute-valued wavelet space

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

The goal of ARSE is to determine the variables described by the calibration set with high predictive ability while having minimal contribution from the UIs. The presences of a UI can be identified through the projection error,

$$ProjE = \boldsymbol{x} * (\boldsymbol{I} - \boldsymbol{V} * \boldsymbol{V}')^2 \qquad \textbf{Eq. 2}$$

In equation 2, **x** represents a future, possibly contaminated sample with *m* variables, **I** represents an *m* x *m* identity matrix, and **V** represents the eigenvectors of the calibration space. The projection error for any k-variable subspace of the calibration space can be equivalently determined by selecting any *k* of the *m* variables of **x**. The identity matrix becomes *k* x *k*, and the eigenvector matrix, **V**, must be recalculated from just the *k* variables in the calibration set. This distance outside of the calibration space is then a measure of the contamination within each set of *k* variables.

The net predictive ability of a k-variable subspace is determined by calculating the prediction error for a set of k-variables,

$$PredE = \sqrt{\frac{\sum(y-\hat{y})^2}{n}} \quad \textbf{Eq. 3}$$

Where *y* represent the calibration reference values, $\hat{y}$ represent the partial least squares (PLS) calculated estimate of the reference values, and *n* represents the number of samples in the calibration set. A separate multivariate model is generated for each unique subspace analogous to moving window partial least squares regression (*26*), except the set of variables employed is not contiguous and is in the wavelet domain.

After many k-variable subspaces are analyzed, the average projection and prediction errors can be calculated for each variable based on the observed projection and prediction errors of subspaces when each variable is employed. This, therefore, gives a measure of how contaminated each variable is by an uncalibrated interferent and how informative each variable is with respect to determining the analyte of interest in the calibration space.

After a large set of potential subspaces have been analyzed, the average projection and prediction errors are used to establish the best subset of variables to use to predict the sample with UI. This is accomplished by rank ordering these two errors and then examining the union of the two ordered sets. The union of the two sets can be expanded until the desired number of samples are present: for example both errors may need to be expanded until the best performing 30 variables, by each metric, are present before the union of the two sets contains five variables. Once the desired number of variables are present in the union, those variables can be selected to build a model and predict the possibly contaminated test sample. The algorithm in its entirety can be seen below in Figure 2.

This approach can then be repeated for each sample in the set of samples with UI. By analyzing each uncalibrated sample independently, the calibration model can be rebuilt to fit the needs of each specific sample. This allows the algorithm to compensate for a set of samples that may have different interferents in different samples.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch010

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch010

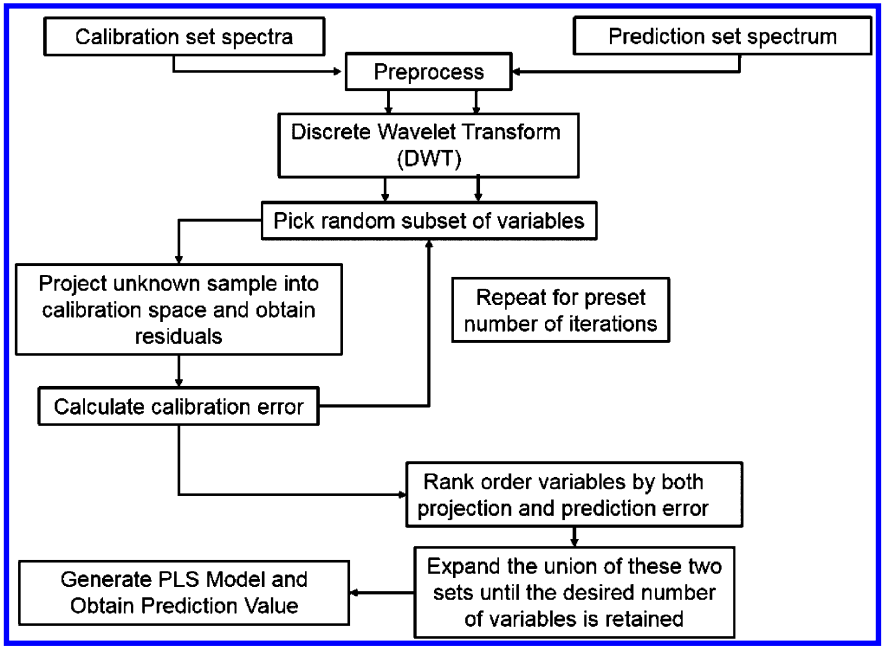

Figure 2. Diagram of the ARSE algorithm

Experimental

Software

Programs for new methods were written in Matlab 8.4 (The Mathworks, Natick, MA). PLS programs were used from the PLS toolbox version 7.95 (Eigenvector Research, Inc., Manson, WA).

Data Sets

Data Set 1

Synthetic mixture spectra were made from pure component spectra obtained from the EPA Vapor-Phase IR Library. The pure component spectra were measured at 4 cm^{-1} resolution from 450 to 4000 cm^{-1}. The pure component spectra (Figure 3) were then used to create 40 calibration samples containing three species and 25 test samples containing the original three calibration species plus an uncalibrated interferent. Concentrations were randomly determined values between 0 and 1, and the concentration of the uncalibrated interferent in the test samples was entirely independent of the concentration of the analyte of interest.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

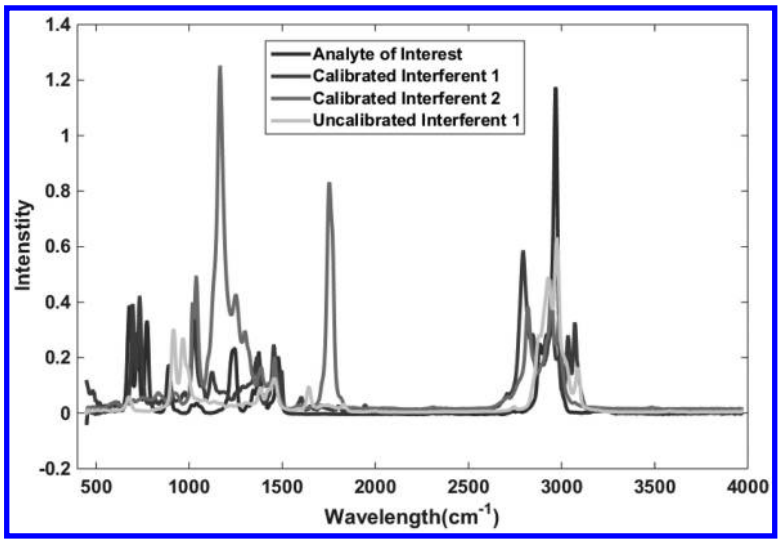

Figure 3. Pure component spectra for data set 1

Data Set 2

Synthetic mixture spectra using pure component UV-VIS spectra were measured in-house on a HP 8452a UV-VIS with 2 nm resolution from 190 to 820 nm. The pure components (Figure 4) consist of three dyes, Eosin Y, green food coloring (Blue 1 and Yellow 5), and Rhodamine B were used for the calibration set, where Eosin Y was treated as the analyte of interest. The test set consisted of two dyes, Methyl Red and Quinaldine Red, to act as different uncalibrated interferents. These were then used to create a calibration set containing 60 samples and two different 40 sample test sets with each test set containing one of the uncalibrated interferents. As with the previous set, concentrations were randomly distributed values from 0 to 1, U(0,1) and the concentration of the uncalibrated interferents were independent of those of the analyte of interest and also followed a U(0,1) distribution.

Data Preprocessing and Algorithm Parameters

All spectra were wavelet transformed using a symlet 8 wavelet, keeping both the detail and approximation coefficents of the wavelet transform to potentially increase the number of variables that are unique to the calibration set. All spectra and corresponding reference values were mean centered before any PLS model was built. The algorithm was allowed to run for 1 million iterations while selecting $k = 6$ random variables for each iteration. The ProjE (eq. 2) and PredE (eq. 3) were calculated for the 6-variable subspace. Preliminary results show that these parameters work well, but may not be optimized for general use nor optimal for any particular data set. Further optimization and validation of the algorithm parameters is an ongoing project.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch010

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch010

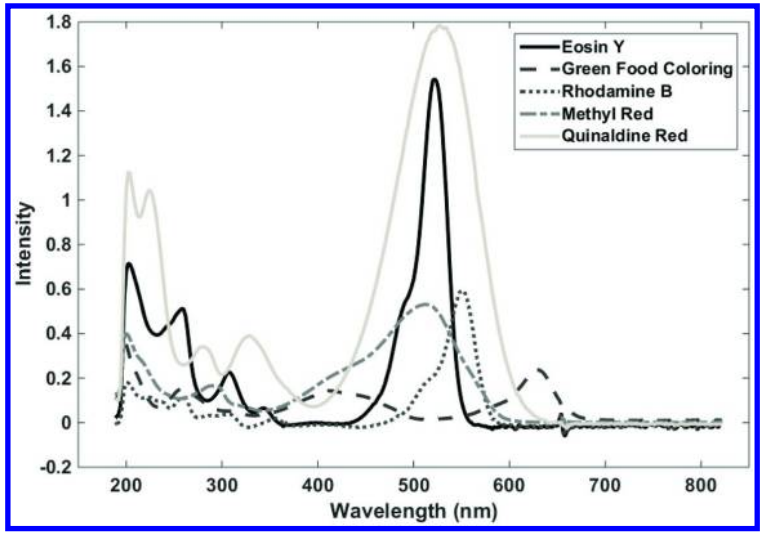

Figure 4. Pure component spectra for data set 2

Results and Discussion

Data Set 1

The synthetic IR data are perfectly estimated by a three factor PLS model when no noise or uncalibrated interferent are present for the calibration model construction in either the wavelength or wavelet space. When estimating concentrations in the presence of uncalibrated interferents, analysis in the wavelet and wavelength spaces perform equivalently. In the wavelength space, estimating the analyte concentration from the 25 spectra containing an uncalibrated interferent presents a root mean squared error of prediction (RMSEP) of 0.2738 for samples with a mean nominal concentration of 0.5 (Table 1). These 25 samples have a mean error of -0.2406 with a standard deviation of 0.1227 based solely on the distribution of added uncalibrated interferent. In the wavelet space, estimating the analyte concentration from the 25 spectra containing an uncalibrated interferent presents a RMSEP of 0.2786 for samples with a mean nominal concentration of 0.5 (Table 2). These 25 samples have a mean error of -0.2503 with a standard deviation of 0.1248 based solely on the distribution of added uncalibrated interferent. The distribution of prediction errors is shown in Figure 5.

The application of ARSE performs significantly better on the wavelet transformed data than on the wavelength-space data. This is due to a lack of interferent-free variables in the wavelength space. Applying ARSE in the wavelength space does not improve the RMSEP with either a 6-variable or 12-variable model (Table 1). The RMSEP converges to the RMSEP without ARSE as more variables are included.

Table 1. Results for data set 1 with no noise in wavelength space

Method	*Mean Error*	*STD*	*RMSEP*
PLS	-0.2460	0.1227	0.2738
ARSE plus PLS 6 variables	-1.3785	0.6875	1.5342
ARSE plus PLS 12 variables	-0.7455	0.3718	0.8298

Table 2. Results for data set 1 with no noise in wavelet space

Method	*Mean Error*	*STD*	*RMSEP*
PLS	-0.2503	0.1248	0.2786
ARSE plus PLS 6 variables	0.0470	0.0234	0.0523
ARSE plus PLS 12 variables	0.0470	0.0234	0.0523

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch010

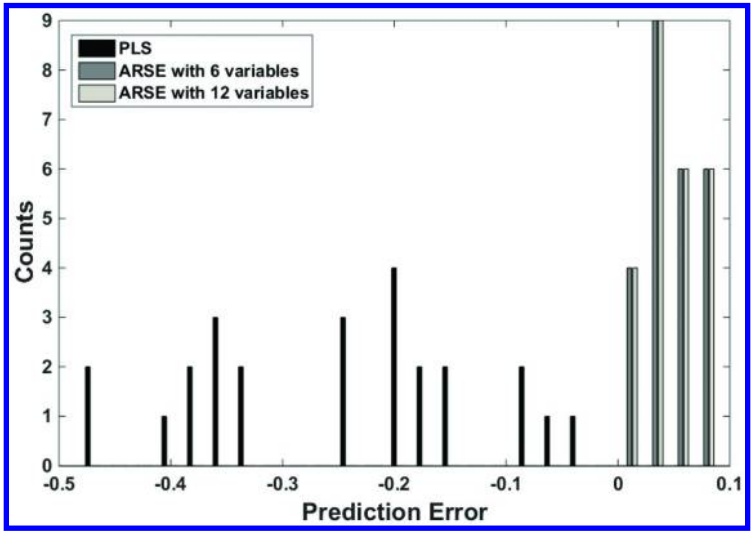

Figure 5. Histogram of errors for noiseless data set

Applying ARSE to the noiseless data reduces the RMSEP by a factor of 5.3 for both the best 6-variable and best 12-variable ARSE models (compare rows 2 and 3 to row 1 in Table 2). The absolute mean error and the standard deviation of observed errors are also reduced by a factor of 5.3. That is to say the PLS model is 530% more biased without ARSE than following the application of ARSE. The distribution of errors following ARSE are virtually identical for the 6-variable and

12-variable ARSE treatments (Table 2). The bias shift from negative to positive is a consequence of which subspace is kept. This should not be interpreted as an 'overcorrection' by ARSE. Upon selection of variables with minimal uncalibrated interferent contribution, there was a greater net overlap with positive weighted variables in the regression vector than that of negatively weighted variables in the regression vector when the PLS model was rebuilt within the retained subspace.

To further demonstrate the efficacy of this new algorithm, two levels of normally distributed noise, N (0, 1), were added to the calibration and test data sets. The noise levels were scaled to be 1% and 5%of the net spectral intensity of each variable in the wavelength space. The spectra were then converted from wavelength space to wavelet space prior to ARSE application.

The addition of 1% and 5% noise does not significantly impact the performance of the PLS model prior to treatment by ARSE. The RMSEP, mean bias, and standard deviation of observed biases are all within 0.2% of the values obtained by PLS analysis of the noiseless data. (Compare Table 3, first row to Table 2, first row). The errors of analysis are dominated by the bias derived from the uncalibrated interferent not the added random spectral noise.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch010

Table 3. Results for data set 1 with noise added in wavelet space

Method	*1 Percent Noise*			*5 Percent Noise*		
	Mean Error	*STD*	*RMSEP*	*Mean Error*	*STD*	*RMSEP*
PLS	-0.2497	0.1247	0.2780	-0.2458	0.1268	0.2754
ARSE plus PLS 6 variable	0.0448	0.0733	0.0846	0.0001	0.2028	0.1987
ARSE plus PLS 12 variable	0.0332	0.0604	0.0679	0.0111	0.1518	0.1491

A Monty Carlo noise sensitivity analysis highlights the impact of random errors on the ability of ARSE to determine robust calibration models. With 1% normally distributed noise, both the 6 variable and 12 variable ARSE models show improvement on both accuracy (~3x-4x) and precision (~2x) (Table 3, Figure 6). With 5% normally distributed noise, the tradeoff between accuracy and precision using ARSE becomes evident (Figure 7). While a greater than 10x improvement in accuracy is realized, the precision is degraded substantially. In this example, the 12 variable model yields a better precision without loss of accuracy than the 6 variable model due to signal averaging, in which least 12 uncontaminated variables exist in the wavelet space. In addition to signal averaging across more variables, where appropriate, loss in precision can be recovered by averaging replicate samples.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch010

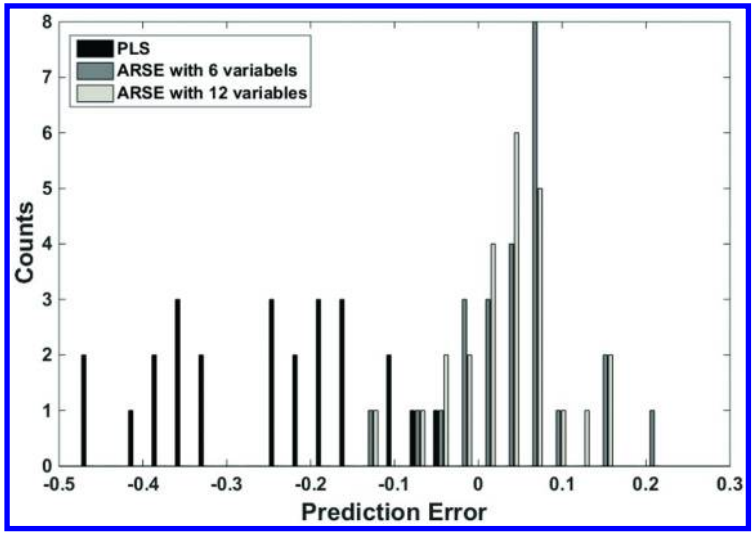

Figure 6. Histogram of errors for 1% noise data set

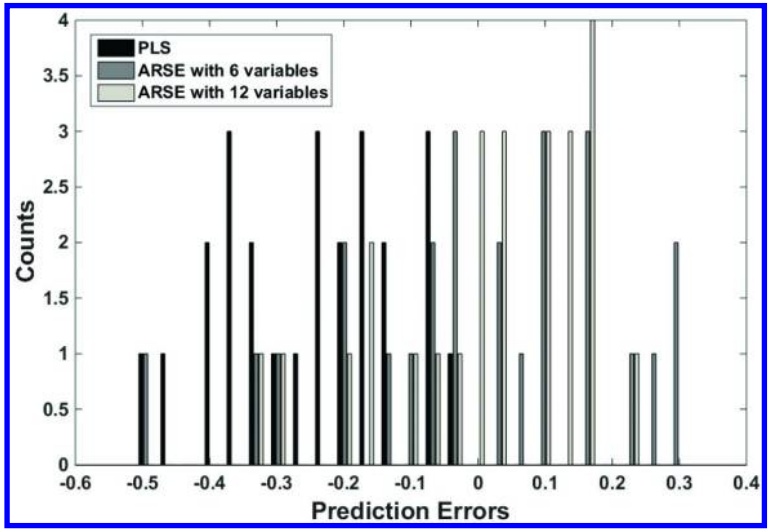

Figure 7. Histogram of errors for 5% noise data set

Ultimately, the accuracy and precision of ARSE is based on which variables are selected. Of interest is the effect on variable selection by the noise present. Each test set sample was replicated 20 times with different realizations of 1% and 5% noise. Due to similarities in results, only the 1% noise will be discussed. Interestingly, although each sample was analyzed independently of all other samples, the ARSE algorithm under-corrected samples with low analyte concentration and over-corrected samples with high analyte concentration

(Figure 8). Observing the frequency usage of each variable (Figure 9) in the 12-variable model shows that 5 variables were employed at least 90%of the time and 11 variables where employed at least 50% of the time**.** When the 12 most commonly employed variables for the 5 lowest concentrations are used to form a model applied to all samples, a similar under-correction is observed; when the 12 most commonly employed variables for the 5 highest concentration are used to form a model applied to all samples, a similar over-correction is observed. Consequently, it is concluded that the under-/over-correction problem is a function of the variables chosen, and is not intrinsic to the PLS model or the constructed data. Disaggregating the data shows that the major difference among the variables selected for the high concentration and low concentration samples is a simple shift in the variables being chosen, i.e. for high concentration variable 937 is chosen and for the low concentration variable 940 is chosen.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch010

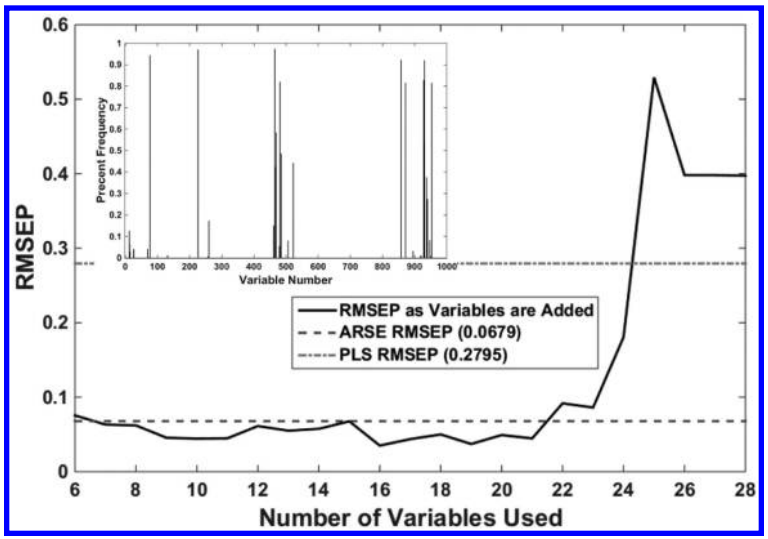

Figure 8. RMSEP as a function of number of variables used. Inset histogram of the percentage a variable is chosen

A more accurate and precise model can be realized with a different selection of variables. From the histogram of employed variables (Figure 8), models were constructed using the *k* most frequently selected variables and applied to all the samples, each with 20 different realizations of noise (Figure 9). The RMSEP was determined for the 20 replicates of all 25 samples from Figure 9 as a function of number of variables used (Figure 8). Clearly, more accurate and precise models can be obtained by choosing variables in a manner that is more robust to random noise. Based on the observed RMSEP for different calibration subspaces in Figure 8, judicious selection of variables through an improved ARSE algorithm could lead to a further 50% improvement in robustness against uncalibrated interferents.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch010

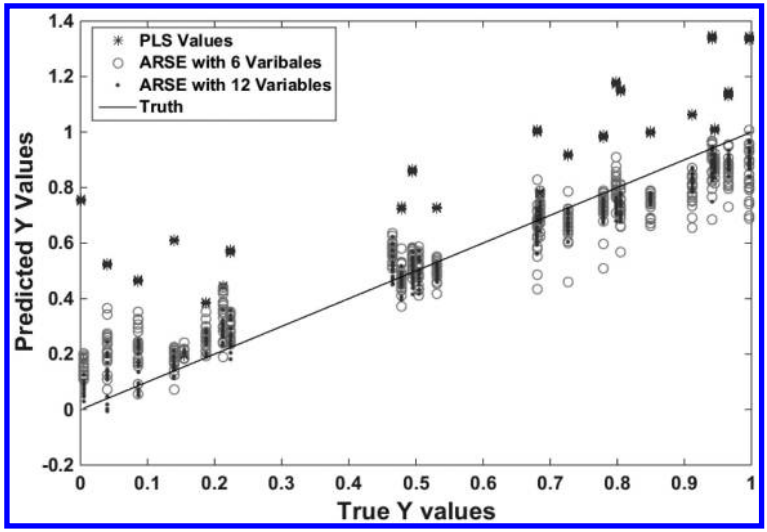

Figure 9. Predicted vs True Y values for repeated 1% noise samples

Data Set 2

The UV-Vis absorption data set represents a more challenging scenario for ARSE. Not only are there no analytically useful uncontaminated variables in the wavelength space, there are few analytically useful uncontaminated variables in the wavelet space. When building PLS models across all variables in the presences of Quinaline Red, RMSEPs of 0.4872 and 0.4953 in the wavelet and wavelength space were obtained, respectively; when in the presence of Methyl Red, RMSEPs of 0.1841 and 0.1887 in the wavelet and wavelength space were obtained, respectively.

Employing ARSE to the data with Quinaline Red as the uncalibrated interferent, a 4.2x improvement is realized for the 6-variable model and 3.8x improvement is realized for the 12-variable model (Table 4, Figure 10). The ARSE models also presents a 1.9x improvement in model accuracy and 2.4x improvement in model precision for the 6-variable model and 3.8x improvement in model accuracy and 4.0x improvement in model precision for the 12-variable model. When examining the data with Methyl Red as the uncalibrated interferent, a slight increase in RMSEP (1.14x) is observed for the 6-variable model; however, a 1.8x improvement is observed for the 12-variable model (Table 6, Figure 11). The accuracy improves for both the 6-variable and 12-variable model, whereas the precision only improves for the 12-variable model. This degradation of the precision for the 6-variable model results in the higher overall prediction error when compared to the PLS model.

As with the IR data, two levels of normally distributed noise, N (0, 1), were added to the calibration and both test data sets. The noise levels were scaled to be 1 percent and 5 percent of the net spectral intensity of each variable in the wavelength space. The spectra were then converted from wavelength space to wavelet space prior to ARSE application.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch010

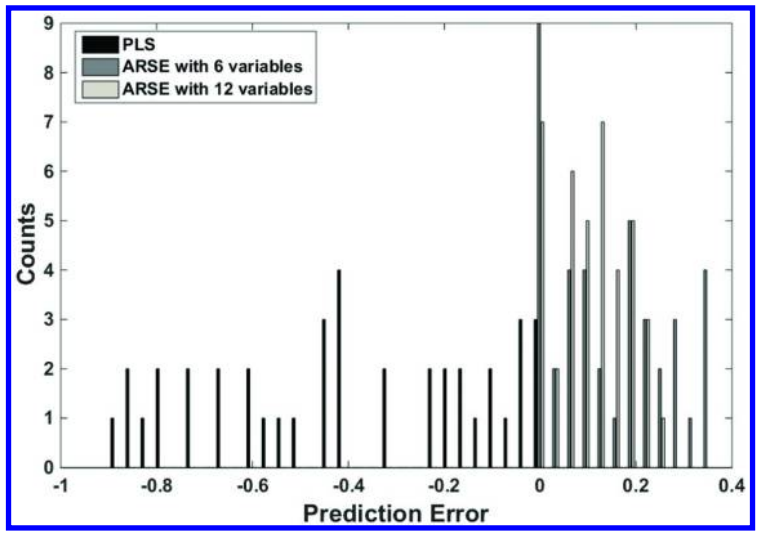

Figure 10. Histogram of prediction errors with Quinaldine Red as uncalibrated interferent and no noise

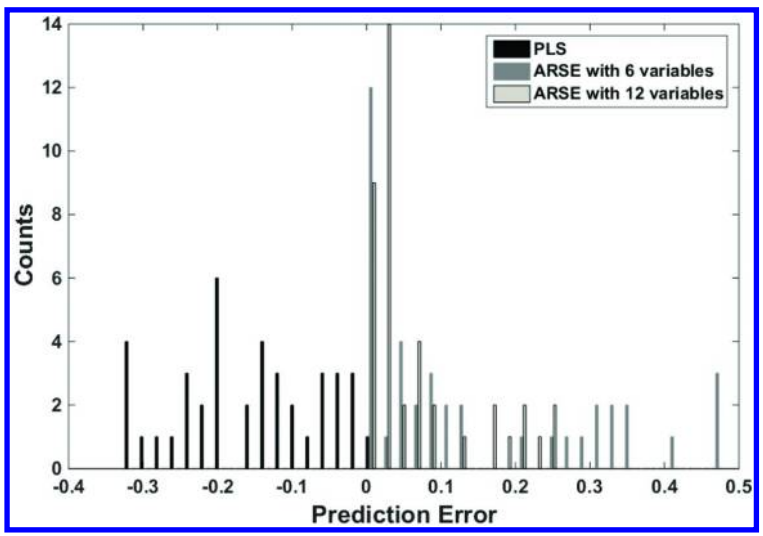

Figure 11. Histogram of prediction error with Methyl Red as uncalibrated interferent and no noise

Similarly, the addition of noise had no significant effect on the PLS model. The RMSEP is within 0.1% of the noiseless value for the 1% noise data set and is exactly the same to 4 decimal places for the 5% noise data set (Compare row 1 Tables 4 and 5, and row 1 Tables 6 and 7). The model precision and accuracy also vary by, at most, 0.1% when comparing the noiseless data to the data sets with additional noise. This again demonstrates that the error in the model is inherent

to the presence of the uncalibrated interferent rather than any additional spectral noise.

Table 4. Data set 2 with Quinaldine Red as uncalibrated interferent in wavelet space

Method	*Mean*	*STD*	*RMSEP*
PLS	-0.3976	0.2852	0.4872
ARSE plus PLS 6 variable	0.2139	0.1184	0.1184
ARSE plus PLS 12 variable	0.1039	0.0712	0.1254

Table 5. Data set 2 with Quinaldine Red as uncalibrated interferent in wavelet space with added noise

Method	*1 Percent Noise*			*5 Percent Noise*		
	Mean Error	*STD*	*RMSEP*	*Mean Error*	*STD*	*RMSEP*
PLS	-0.3969	0.2885	0.4869	-0.3988	0.2834	0.4872
ARSE plus PLS 6 variable	0.1303	0.2481	0.2774	1.1586	1.0178	1.5337
ARSE plus PLS 12 variable	0.9202	0.7289	1.1683	1.2792	1.1098	1.6844

When Quinaldine Red is the uncalibrated interferent, ARSE is able to improve the RMSEP by a factor of 1.75x for a 6-variable model on the 1% noise data (Table 5). However, for both the 12-variable model on the 1% noise data and both models on the 5% noise data, there is a significant increase in all three factors of merit. This increase is due to the inherent difficulty of this data set. The difficulty of the small number of analytically useful uncontaminated variables in this data set is further complicated by the introduction of any noise.

Similarly to with Quinalidine Red, when Methyl Red is the uncalibrated interferent ARSE is able to improve the RMSEP by a factor of 1.4x for the 6-variable model in the 1% noise data set (Table 7). Again, however, for the 12-variable model for the 1% noise data set and both models for the 5% noise data set, there is a significant degradation in the prediction error, though less than in the case of Quinalidine Red.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch010

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Table 6. Data set 2 with Methyl Red as uncalibrated interferent in wavelet space

Method	*Mean*	*STD*	*RMSEP*
PLS	-0.1570	0.0973	0.1841
ARSE plus PLS 6 variable	0.1460	0.1583	0.2139
ARSE plus PLS 12 variable	0.0685	0.0761	0.1016

Table 7. Data set 2 with Methyl Red as uncalibrated interferent in wavelet space with added noise

Method	*1 Percent Noise*			*5 Percent Noise*		
	Mean Error	*STD*	*RMSEP*	*Mean Error*	*STD*	*RMSEP*
PLS	-0.1568	0.0971	0.1838	-0.1596	0.0992	0.1873
ARSE plus PLS 6 variable	0.0848	0.1055	0.1344	0.2465	0.2648	0.3593
ARSE plus PLS 12 variable	0.1917	0.1455	0.2396	0.2573	0.2703	0.3707

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch010

Conclusions

The results in this chapter show that a solution to the problem of uncalibrated interferents in future samples exists in the form of determining uncontaminated variables and subsequently re-building a model with just those variables. This approach is not without obstacles that must be overcome. First is the creation of variables that are both analytically relevant to the analyte of interest and uncontaminated by interferents. Within this chapter, that was accomplished via a symlet based wavelet transform. Future work will focus on analyzing other possible wavelet families as well as wavelet preprocessing to eliminate irrelevant variables prior to the application of ARSE.

Once appropriate variables are created, the obstacle becomes selecting those variables. This work has shown that with an ARSE-like algorithm it is possible to select variables and build a model that improves model accuracy with minimal increase in model precision. However, the selection process must be further modified to be more robust to the effect of sample to sample variation.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

References

1. Beebe, K.; Kowalski, B. *Anal. Chem.* **1987**, *59*, A1007.
2. Thomas, E.; Haaland, D *Anal. Chem.* **1990**, *62*, 1091–1099.
3. Blanco, M.; Coello, J.; Iturriaga, H.; Maspoch, S.; Rovira, E. Wavelength. *Appl. Spectrosc.* **1995**, *49*, 593–597.
4. Bouveresse, E.; Massart, D. L.; Dardenne, P. *Anal. Chem.* **1995**, *67*, 1381–1389.
5. Candolfi, A.; Massart, D. L. *Appl. Spectrosc.* **2000**, *54*, 48–53.
6. Capron, X.; Walczak, B.; de Noord, O. E.; Massart, D. L. *Chemom. Intell. Lab. Syst.* **2005**, *76*, 205–214.
7. Chen, Z.; Morris, J.; Martin, E. *Anal. Chem.* **2006**, *78*, 7674–7681.
8. Greensill, C. V.; Wolfs, P. J.; Spiegelman, C. H.; Walsh, K. B. *Appl. Spectrosc.* **2001**, *55*, 647–653.
9. Anderson, C. E.; Kalivas, J. H. *Appl. Spectrosc.* **1999**, *53*, 1268–1276.
10. Igne, B.; Roger, J.; Roussel, S.; Bellon-Maurel, V.; Hurburgh, C. R. *Chemom. Intell. Lab. Syst.* **2009**, *99*, 57–65.
11. Isaksson, T.; Kowalski, B. Piece-Wise. *Appl. Spectrosc.* **1993**, *47*, 702–709.
12. Feundale, R.; Woody, N.; Tan, H.; Myles, A.; Brown, S.; Ferre, J. *Chemom. Intell. Lab. Syst* **2002**, *64*, 181–192.
13. Adhihetty, I. S.; Mcguire, J. A.; Wangmaneerat, B.; Niemczyk, T. M.; Haaland, D. M. *Anal. Chem.* **1991**, *63*, 2329–2338.
14. Andrews, D. T.; Wentzell, P. D. *Anal. Chim. Acta* **1997**, *350*, 341–352.
15. Barring, H. K.; Boelens, H. F. M.; de Noord, O. E.; Smilde, A. K. *Appl. Spectrosc.* **2001**, *55*, 458–466.
16. Blanco, M.; Coello, J.; Iturriaga, H.; Maspoch, S.; Rovira, E. *Appl. Spectrosc.* **1995**, *49*, 593–597.
17. Bouveresse, E.; Massart, D. L. *Chemom. Intell. Lab. Syst.* **1996**, *32*, 201–213.
18. Bouveresse, E.; Massart, D. L.; Dardenne, P. *Anal. Chim. Acta* **1994**, *297*, 405–416.
19. Chu, X. L.; Yuan, H. F.; Lu, W. Z. *Spectrosc. Spectral Anal.* **2001**, *21*, 881–885.
20. Denoord, O. E. *Chemom. Intell. Lab. Syst.* **1994**, *25*, 85–97.
21. Duponchel, L.; Ruckebusch, C.; Huvenne, J. P.; Legrand, P. *J. Near Infrared Spectrosc.* **1999**, *7*, 155–166.
22. Fearn, T. *J. Near Infrared Spectrosc.* **2001**, *9*, 229–244.
23. Geladi, P.; Barring, H.; Dabakk, E.; Trygg, J.; Antti, H.; Wold, S.; Karlberg, B. *J. Near Infrared Spectrosc.* **1999**, *7*, 251–264.
24. Vogt, F.; Mizaikoff, B. *J. Chemom.* **2003**, *17*, 225–236.
25. Goswami, J. C.; Chan, A. K. *Fundamentals of wavelets : theory, algorithms, and applications*; Wiley: New York, 1999.
26. Jiang, J. H.; Berry, R. J.; Siesler, H. W.; Ozaki, Y. *Anal. Chem.* **2002**, *74*, 3555–3565.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch010

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch011

Chapter 11

The Essential Aspects of Multivariate Calibration Transfer

Jerome J. Workman, Jr.*,[1]

Department of Health and Human Services, National University, 9388 Lightwave Avenue, San Diego, California 92123, United States
***E-mail: jworkman04@gsb.columbia.edu**
[1]Current Address: Unity Scientific, 117 Old State Road, Brookfield, Connecticut 06804, United States

The technical issues associated with multivariate calibration transfer (or calibration transfer) for spectroscopic instruments using absorption spectroscopy are addressed in this chapter. Calibration transfer refers to a series of analytical approaches or chemometric techniques used to attempt to apply a single spectral database, and the calibration model developed using that database, to two or more instruments, with retained accuracy and precision. One may paraphrase this definition of calibration transfer as, "calibration transfer means the ability for a multivariate calibration to provide the same analytical result for the same sample measured on a second (child) instrument as it does on the instrument on which the calibration model was created (parent instrument)", as described in H. Mark, and J. Workman, *Spectroscopy* 2013, 128(2), 1-9. There are many technical aspects involved in successful calibration transfer, related to the measuring instrument reproducibility and repeatability, the reference chemical values used for the calibration, the multivariate mathematics used for calibration, and so forth. Ideally a multivariate model developed on a single instrument would provide a statistically equivalent analysis when used on other instruments following transfer.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Instrumentation Issues

Calibration transfer involves measuring a set of reference samples, developing a multivariate calibration, and transferring that calibration to other instruments (*1*). The basic spectra are initially measured on at least one instrument (i.e., the parent, primary, or Master instrument) and combined with the corresponding reference chemical information (i.e., Actual or Reference values) for the initial development of a multivariate calibration model. These models are applied and maintained on the original instrument over time and are transferred to other instruments (i.e., child, secondary, or transfer instruments) to enable analysis using the child instruments with minimal intervention and recalibration. (Note that the issue of calibration transfer disappears if the instruments are precisely alike.) If instruments are the "same" then any one sample placed on any spectrophotometer will predict or report precisely the "same" analytical result. Since instruments are not precisely alike, and in fact are different from the moment of manufacture, and they drift over time, the use of calibration transfer techniques is often applied to produce the best attempt at calibration model or data transfer with minimal variation across instruments. The discussion of calibration transfer, from a purely chemometric and instrumentation standpoint, does not address differences in reference laboratories used for the Y-block or reference data. This is a separate problem not addressed within this chapter. However, from a theoretical perspective, if instrumentation is identical and constant then the same physical sample will yield the same predicted result when using the same multivariate calibration.

In practice, the current state-of-the-art for multivariate calibration transfer is to apply one or more software algorithms, and to measure physical standards on multiple instruments. These standard spectra are used to align the spectra from different instruments and to move (or transfer) calibrations from one instrument to another. All the techniques used to date involve measuring samples on the calibration instrument (i.e., parent), and the transfer instrument (i.e., child) and then applying a variety of approaches to complete the transfer procedure.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch011

Types of Spectrophotometers

Absorption-based spectrophotometers exist in several design types. There are instruments based on the grating monochromator with mechanical drive, grating monochromator with encoder drive, the Michelson interferometer in various forms, dispersive gratings with array detectors, interference and linear variable filter instruments with single detectors, linear variable filters with array detection, MEMS based Fabry-Perot interferometers, Digital Transform actuator technologies, Acousto-optic tunable filters (AOTF), Hadamard transform spectrometers, laser diodes, tunable lasers, various interference filter types, multivariate optical computing devices, and others. Calibration transfer from one design type spectrometer to the same type and manufacturer is challenging, but transfer of sophisticated multi-factor PLS models across instruments of different design types is not currently well understood. The requirement of precise spectral shapes across instruments requires more sophisticated spectral transforms than

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

simple first-order X- (wavelength or frequency),Y- axes (photometric value), and smoothing (lineshape simulation) corrections. Note that using a smoothing function to simulate lineshape or resolution differences is not adequate, since lineshape varies with wavelength, and smoothing functions may more closely mimic spectral shapes in one wavelength region, yet cause greater differences in other regions of the spectrum. Multivariate calibrations are developed for many data dimensions and require delicate adjustments across multi-dimensional spectral space to fit measured data precisely. The center wavelength, data spacing, photometric response, photometric linearity, resolution, instrument line shape and symmetry, and other parameters must be nearly identical for multivariate calibrations to yield equivalent prediction results across different spectrometers and different design types of spectrometers.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch011

Common Calibration Transfer Practices

The most ubiquitous approach to calibration transfer involves applying an existing PLS model to a transfer or child instrument using a bias or slope correction for predicted results. In this procedure a set of 10 to 40 test or transfer samples is measured on both the parent and child instruments, and the resulting analytical results are "adjusted" on the child instrument using a bias and slope procedure (i.e., linear regression) to best fit the child instrument results to those of the parent. This process has been demonstrated to be ineffective at creating a parity between results reported from both parent and child instruments over time. Some of the mathematics of this approach are discussed in this chapter.

The second and third most used approaches to multivariate calibration transfer involves the application of Direct Standardization (DS), and Piecewise Direct Standardization (PDS) (*2–5*). These approaches are described in detail with application descriptions, examples, and equations in reference (*5*). These approaches are also often combined with small adjustments in bias or slope of predicted values to compensate for small differences not accounted for by using standardization algorithms. Note that the frequency with which standardization approaches must be applied to child instruments is dependent upon the frequency of calibration updates required and the spectral shape drift of the child instruments with respect to the parent (or calibration instrument).

For the DS method, the test sample set is measured on the Parent and Child instruments as typically Absorbance (A) with respect to wavelength (k). The spectral data has *k* specific wavelengths. A transformation matrix (T) is used to match the child instrument data (A_C) to the Parent instrument data (A_P). And so Equation 1 demonstrates the matrix notation. Note that for DS a linear relationship is assumed between the parent and child measurement values.

$$A_P = A_C \hat{T} + E \tag{1}$$

where A_P is the parent data for the test sample set as an *n x k* matrix (*n* samples and *k* wavelengths), A_C is the child instrument data for the test samples as an *n x k* matrix, $\hat{T}$ is the k *x k* transformation matrix, and *E* is the unmodeled residual error matrix.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

The transformation matrix ($\hat{T}$) is computed as Equation 2.

$$\hat{T} = A_C^{+} A_P \tag{2}$$

where A_C^{+} is the pseudoinverse approximated using singular value decomposition (SVD) of the *n x k* spectral data matrix for a set of transfer or standardization samples measured on the child instrument, A_P is the *n x k* spectral data matrix for the same set of transfer or standardization samples measured on the parent instrument. The transform matrix is used to convert a single spectrum measured on the child instrument to be converted to "look" like a parent instrument spectrum.

For the PDS method, the DS method is used piecewise or with a windowing method to more closely match the spectral nuances and varying resolution and lineshape of spectra across the full spectral region, and there is no assumption of linearity between the parent and child prediction results. The transformation matrix is formed in an iterative manner across multiple windows of the spectral data in a piecewise fashion. Many other approaches have been published and compared, but for many users these are not practicable and have not been adopted for various reasons; these methods are reviewed in reference (*6*). Note for calibration transfer the alignment of the wavelength axis between instruments is most critical since PLS assumes data channel (i.e., wavelength) integrity for all spectral data.

If the basic methods for calibration transfer do not produce satisfactory results, the user begins to measure more samples on the transfer (i.e., child) instrument until the model is basically updated based on the child instrument characteristics. One might note that to date using the same sample set to develop an entirely new calibration on a child instrument yields the optimum results for calibration accuracy; however this is not considered calibration transfer *per se*; *recalibration* would be a more precise term for this process. Imagine the scenario where a user has multiple products and constituents and must check each constituent for the efficacy of calibration transfer. This is accomplished by measuring 10-20 product samples for each product and each constituent in order to compare the average laboratory reference value to the average predicted value for each constituent, and then adjusting each constituent model with a new bias value. This exercise obviously results in a ponderous procedure; however this is less exasperating than recalibrating each instrument using hundreds of samples for each product and constituent.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch011

Calibration Modeling Approaches to Transfer

Various methods have been proposed to produce a universal model, or a calibration that is mostly robust against standard instrument changes as are common to modern commercial instruments. These have been referred to as robust models, or global models. In this case various experimental designs are constructed to better represent the product, reference values, and instrument calibration space to include typical changes and interferents that should be

included within the model space for predicted values to be broadly applicable. Using this approach one might design a detailed factorial experiment for the composition of the learning or calibration set to include multiple variations typically encountered during routine analysis. A list of some of these variations may consist of: differences in pathlength, sample temperature, moisture content, flow rate, particle size, interferent content, instrument type, constituent ratios, sampling parameters, process or manufacturing conditions, and the like (*7*). These approaches will work for a period until the instrument performance characteristics drift or the product or constituent chemistry changes significantly. These types of changes are expected and thus routine recalibration (i.e., model updating) would normally be required as a standard procedure if any of the changes are considered significant.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch011

a. Standardization Methods

The most common method for calibration transfer using near infrared spectroscopy involves measuring spectra and developing calibration models on a parent or primary instrument, sometimes referred to as a "Master" instrument and transferring a calibration using a set of transfer samples measured on a child or secondary instrument (*8*, *9*). It is commonly understood that the results of calibration transfer often require a bias on the child instrument or a significant number of samples measured on the child instrument to develop a suitable working calibration. This practice most often increases the error of analysis on the second instrument. There are multiple papers and standards describing the statistical techniques used for analytical method comparisons on two or more instruments, some of which are described in references (*10–13*).

b. Instrument Comparison and Evaluation Methods

One of the essential aspects for determining the efficacy and quality of calibration transfer is to make certain the spectrometer instrumentation is essentially identical, or as similar as possible, prior to the calibration transfer experiment. There are many standard tests that are used to determine alikeness between spectrophotometers. Eight basic tests are quite useful for characterizing instrument measurement performance. These tests include: wavelength accuracy, wavelength repeatability, absorbance/response accuracy, absorbance/response repeatability, photometric linearity, photometric noise, signal averaging (noise) tests, and the instrument line shape (ILS) test. If carefully conducted, these experiments provide specific information for diagnosing mechanical, optical, and electronic variations associated with basic design limitations, tolerance problems, or implementation errors. The results of these tests provide objective data for correcting and refining instrument repeatability and reproducibility. These tests have been described in some detail within previous publications (*14–17*).

The Mathematical Aspects of Calibration Transfer

The Basic Premises

A review of the application of chemometrics to spectroscopic methods and practices have been described (*18–20*). Calibration for both quantitative and qualitative methods have been discussed and reviewed and multi-way methods have also been described in detail. To summarize, the process of multivariate calibration transfer involves the following steps:

(1) a comprehensive set of teaching (or calibration) spectra are measured on at least one instrument (i.e., the parent, primary, or Master instrument) and combined with the corresponding reference chemical information (i.e., Actual or Reference values) for the initial development of calibration models. These models are maintained on the original instrument over time and are used to make the initial calibration.
(2) The initial calibration is transferred to one or more additional instruments (i.e., child, secondary, or transfer instruments) to enable analysis using the child instruments with minimal correction, biasing, or other intervention.
(3) A set of transfer samples, representing a subset of the full teaching or calibration set, is measured on the child instrument.
(4) The process of applying a standardization algorithm, such as DS or PDS, is applied.
(5) Residual mean differences are biased and or slope corrected by regressing the child to predict the parent instrument reported analytical values on the transfer sample set.

For emphasis note that the issue of calibration transfer disappears if the instruments are precisely alike. If the parent and child instruments are the "same" over time and temperature conditions then one sample placed on any of these instruments will produce precisely the "same" result using the same multivariate calibration model. Since instruments are not alike from the time of manufacture, and also change significantly over time, the use of calibration transfer mathematics is often applied to produce the best attempt at model or spectral data transfer.

There is another critical factor that has sometimes been overlooked when using absorption spectroscopy for analytical chemistry, that is, that *spectroscopy measures the weight per volume (or moles per volume) fractions of the various components of a mixture.* The reference values may be based on one of several physical or chemical properties that are only vaguely related to the measured volume fraction. These would include: the weight fraction of materials, the volume percent of composition with unequal densities, the physical or chemical residue after some processing or separation technique, the weight fraction of an element found in a larger molecule (such as total nitrogen versus protein), and other measured or inferred properties. The non-linearity caused by differences in the volume fraction measured by spectroscopy and the reported reference values must be compensated for by using specific mathematics for non linear fitting during calibration modeling. This non-linear compensation during calibration often involves additional factors when using partial least squares (PLS), or

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch011

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

additional wavelengths when using multiple linear wavelength regression (MLR). If the analyst is not using mass per volume fractions as the units for reference values the nuances of instrumental differences will be amplified since the spectroscopy is not necessarily directly or linearly measuring these other types of concentration units.

In the process of transferring calibrations from a parent to a child instrument, one may take four different fundamental strategies for matching the predicted values across instruments. Each of these strategies varies in complexity and efficacy. One may (1) adjust the calibration model (i.e., the regression or b-vector), (2) the instrument as it measures spectra (i.e., the X and Y axes), (3) the spectra (using various spectral transformations, such as matching X and Y axes and apparent lineshapes via smoothing), or (4) adjusting the final predicted results (via bias or slope adjustments). All of these methods have been applied individually or in combination in an attempt to match the reported predicted results derived from parent and child instruments. Ideally one would adjust all spectra to look alike across instruments, such that calibration equations all give the same results irrespective of the specific instrument used. This is the challenge for instrument designers and manufacturers.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch011

a. Instrument Correction

A basic instrument correction must be applied to align the wavelength axis and photometric response for each instrument to make their measurement spectra somewhat alike. This process will create spectra and measurement characteristics that are more similar and repeatable. The correction (or internal calibration) procedure requires photometric and wavelength measurement of reference materials that are stable over time and can be relied upon to have accurate and repeatable characteristics. It is of paramount importance that the reference standards used to align the wavelength and photometric axes do not change appreciably over time. Stable reference standards of known stability and low uncertainty for measurand values may be used at any time to bring the instrument alignment back to its original and accurate state for wavelength and photometric registration. One may note that all spectrophotometers will change over time due to lamp color temperature drift, mechanical wear of moving parts, vibration, electronic component and detector aging; and variations associated with the instrument operating environment, such as temperature, vibration, dust, mechanical and optical fouling, and humidity.

b. Virtual Instrument Standardization

The concept of virtual instrument standardization has been reported to be successful and was demonstrated commercially using high and low finesse etalons and a laser crystal for instrument alignment (*21*, *22*). This technology was limited to a single instrument diode array design and used a proprietary set of special materials and algorithms. These methods and material definitions were never published in sufficient detail to replicate and thus the exact elements remain a trade secret, except for information disclosed within an issued U.S. Patent (*23*).

Comparing Results from Test Sets of Transfer Samples

For spectroscopic-based measurement using multivariate calibration, one may compare the standard or reference concentrations for a set of samples to the spectroscopic-based predicted values. One may also compare the response of the parent Instrument to that of the child instrument. In making these comparisons one may perform statistical tests for bias, correlation, and slope. A statistically significant difference in bias should result in a change of the bias. A statistically significant result in correlation or slope should result in a basic multivariate recalibration, unless one can demonstrate that the differences between the compared values have some real slope variation between them due to fundamental scientific principles. A test for differences between parent and child instrument predictions will indicate the similarity between instruments and the spectra measured between parent and child systems.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch011

a. Bias or Slope Adjustments of Predicted Results across Parent and Child Instruments

A significant bias between parent and child predicted values on the same test sample set is mainly caused by instrument measurement differences. Other sources of significant bias changes between reference values and spectroscopy-based predicted values are due to chemical or spectral interferences. These cause significant bias in the measured analyte concentration due to the effect of another component, property of the sample, or analytical measurement anomaly (*24*).

b. Bias (means) One-Sample t-Test between Parent and Child Instruments

The concept of mean differences (or bias) requires a test of significance to determine whether a bias between mean values is meaningful. An appropriate statistical test will reveal if the variation in the mean values (bias) between sample sets of predicted values is within the expected random fluctuation for a normally distributed population of measurements. The larger the sample set used to test calibration transfer bias the more accurate is the estimate of the true bias value. Thus, using 20 test samples would give a more accurate estimate of the bias than 10 samples. The smaller the standard error of the mean, the greater the confidence will be of the true bias value. The standard error of the mean is given as the SEM (Equation 4) with an example demonstrating the use of 20 rather than 10 samples for bias testing. In this case, a more powerful estimate of the true bias is demonstrated by a factor of $\sqrt{20}/\sqrt{10} = 1.41$. This is true only if there is no significant slope difference. Note: it is acceptable to make the statistically significant bias correction even if there is a slope difference.

The statistical test used to determine bias significance is a simple parametric one-sample t-test. For this test the average predicted mean value for a set of samples on the parent instrument is computed, and compared to the set of predicted values for the same sample set on the child instrument. In the case of NIR data, we designate the mean NIR value for a set of reference samples measured on the parent (i.e., calibration) instrument as $\overline{X}_{Parent}$ (For this test only the mean value is

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

required, not the individual analysis values). The mean NIR value for the same set of reference samples as measured on the child (i.e., transfer) instrument is computed as $\overline{X}_{Child}$. Then for this experiment the test hypotheses is as

$$H_o: \overline{X}_{Parent} = \overline{X}_{Child}$$
$$H_A: \overline{X}_{Parent} \neq \overline{X}_{Child}$$

The Standard Deviation for the child instrument (s_{Child}) is computed using the NIR predicted data on the measured set of samples as given in 3. Where $\overline{y_i}$ is the mean predicted value for the sample set for the child instrument; and y_i are the individual predicted values for the set of test samples for the child instrument.

$$s_{Child} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2}{n-1}} \quad (3)$$

Then the Standard Error of the Mean (SEM) is computed for the child instrument NIR data as Equation 4.

$$SEM = \frac{s_{Child}}{\sqrt{n}} \quad (4)$$

This experiment tests whether the predicted value mean is statistically the same for the parent and child instruments. There are many variations of testing mean differences, but this is a basic test for comparing means when the sample size of two groups is identical. This test determines whether the average predicted values are statistically the same for the test set used. Note that the reference values are not assumed to be known for the test set. For this t test statistic one is able to compute the t-test for mean bias significance as Equation 5.

$$t = \frac{\overline{X}_{Child} - \overline{X}_{Parent}}{SEM} \quad (5)$$

If this resulting t-test is greater than the t critical value for n-1 degrees of freedom, the bias is significant and it should be changed. If the t value computed is less than the critical value of t it is not significant and should not be changed. If the bias is significant, the difference between the slope of each of the two sets of prediction results are compared (i.e., parent versus child); or the correlation coefficients are compared. Note that n in the case of statistical comparisons for bias and slope is the number of samples in the test set, for this example, 20 samples.

c. Bias (means) Two-Sample t-Test between Parent and Child Instruments

The predicted values for the set of test samples is determined for both the parent and child instruments in order to apply a parametric two-sample t-test. Note we are using an identical number of samples for each bias test, so the sample sizes for each test of the parent and child instruments are identical. There is no

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch011

assumption that the reference values for the set of test samples are known. For this test the mean NIR value for the set of reference samples measured on the parent (i.e., calibration) instrument as $\overline{X}_{Parent}$ is computed. Then the mean NIR value for the same set of test samples as measured on the child (i.e., transfer) instrument as $\overline{X}_{Child}$ is computed. For this test the null and test hypotheses are the same as in the one-sample t-test as

$$H_o: \overline{X}_{Parent} = \overline{X}_{Child}$$
$$H_A: \overline{X}_{Parent} \neq \overline{X}_{Child}$$

The Standard Deviations are computed for both the parent (s_{Parent}) and child (s_{Child}) NIR predicted data on the measured set of test samples as follows. Where $\overline{x_i}$ and $\overline{y_i}$ are the mean predicted values for the test set for each instrument; and x_i and y_i are the individual predicted values for the set of test samples for each the parent and child instruments, respectively. The standard deviation calculation for each instrument are given in Equations 6 and 7.

$$s_{Parent} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x_i})^2}{n-1}} \quad (6)$$

$$s_{Child} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \overline{y_i})^2}{n-1}} \quad (7)$$

Next the t-test for mean bias significance is computed as Equation 8.

$$t = \frac{\overline{X}_{Child} - \overline{X}_{Parent}}{\sqrt{\frac{s_{Child}}{n_c} + \frac{s_{Parent}}{n_p}}} \quad (8)$$

If this t-test value is greater than the t critical value for $n_c + n_p - 2$ degrees of freedom, we know the bias to be significant (accept H_A), thus the bias should be changed. If the t value computed is less than the critical value of t it should not be changed and is not significant (accept H_0). If the bias is significant the difference between the slope of the two lines should be tested, or a comparison of the correlation coefficients should be made.

d. Comparing the Correlation Coefficients between Parent and Child Instruments Using the (r-to-z transform) Significance Test

The slope should not be changed to adjust the predicted values following calibration transfer. However the slope significance should be computed as an

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch011

indication of a need to recalibrate the instrument using a new multivariate model on the child instrument. For this test the correlation coefficients are compared for the parent instrument and the child instrument for the same test set of samples; the correlation here refers to Pearson's r statistic.

For this test one computes the r-to-z transformation (aka Fisher's z′ transform) from the Pearson's r correlation. The r-to-z transformation statistic is used to compare two correlation coefficients to see if they are significantly different, such as another correlation possibly obtained in the same or a similar experiment. As an example, if the null hypothesis test is r = 1, it will indicate if the measured correlation is significantly different from 1.

For comparing two correlation coefficients using parametric statistics, the r-to-z transformation is used. This refers to the "r-to-z transformation" for comparing two correlation coefficients, particularly when at least one r-value does not equal zero. For this example case the null (H_0) and test (H_A) hypotheses, respectively, are as H_0: r=1, and H_A: r<1 (or could be stated as r is not equal to 1).

For this question, use Equation 9 for Z_{obs} in conjunction with Table 2 listing the r to z transformation and the critical values for the t-distribution table (Table 1 gives the critical t values for any degree of freedom for different significant levels). For a hypothetical example: given a correlation coefficient of r_{Parent}=0.999 (as representing the correlation coefficient obtained by regressing the parent instrument predictions for a test sample set for time 1 (X) versus time 2 (Y), for example, one week apart. (Compute the r-value and report it as $r_{Parent.}$) Next the correlation is computed for the child instrument (r_{Child}) using the same multivariate calibration and the same test samples on the Child instrument. (Compute the r-value and report it as $r_{Child.}$) The z observed statistic is then computed for this correlation comparison to determine whether it is significantly different from 0.999 (the Parent instrument reference r value). To continue, if r_{Child} is 0.992, the test will be to compare 0.992 to 0.999 to see if they are statistically the same. For this example, we note the test sample set is 20 samples with the predictions from these samples used to compute the correlation values; therefore n_{Parent} = 20 and n_{Child} = 20.

$$Z_{obs} = \frac{\left|Z_{Parent} - Z_{Child}\right|}{\sqrt{\frac{1}{n_{Parent} - 3} + \frac{1}{n_{Child} - 3}}} \tag{9}$$

For this example, calculate the r to z′ transformation for both Parent and Child r-values as: 0.999 and 0.992, respectively. (An r to z′ table (Table 2) or Equation 10 may be used to compute the r to z′ transformation.) For this example, the z_{Parent} = 3.800 and the z_{Child} = 2.759. One may now compute the Z_{obs} = 3.04. (Compare this value to the Critical Values for the t Distribution as Table 1, Significance level of 0.025 for a one-tailed test and the critical value is 2.13.)

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch011

Table 1. Critical Values for the t Distribution (20 test samples)

Confidence Level (%)	*Significance Level (α), One-Tailed Test*	*Critical Value of t*
90	0.05	1.75
95	0.025	2.13
98	0.01	2.60
99	0.005	2.95
99.9	0.0005	4.07

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch011

Table 2. For r to z′ transformation

r (correlation computed)	*z transform (z′)*
0.9999999999	11.859
0.9999999	8.406
0.999999	7.254
0.99999	6.103
0.9999	4.952
0.999	3.800
0.998	3.453
0.997	3.250
0.996	3.106
0.995	2.994
0.994	2.903
0.993	2.826
0.992	2.759
0.991	2.700
0.990	2.647
0.985	2.443
0.980	2.298
0.975	2.185
0.970	2.092
0.965	2.014
0.960	1.946
0.955	1.886
0.950	1.832

Continued on next page.

Table 2. (Continued). For r to z′ transformation

r (correlation computed)	*z transform (z')*
0.945	1.783
0.940	1.738
0.935	1.697
0.930	1.658
0.925	1.623
0.920	1.589
0.915	1.557
0.910	1.528
0.905	1.499
0.900	1.472

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch011

Conclusion for This Example

Since 3.04>2.13 one rejects the null hypothesis and accepts the alternate hypothesis, thus r_{Parent} is not the same as r_{Child} and r<1 for the Child predicted values is the conclusion at 95% confidence. The correlation values are not the same for this test and the predicted values for parent and child have different correlation values. This indicates the predicted values from the Child are not the same as the Parent.

Equation for Computing r to z Transformation

The sampling distribution of Pearson's r is not normally distributed. Fisher developed a transformation now called "Fisher's z′ transformation" which converts the Pearson's r to the normally distributed variable z′. The r to z′ transformation equation is as Equation 10. This statistic is normally distributed with a standard error of $1/\{n-3\}^{1/2}$ as published in the references (*25*).

$$z' = 0.5[\ln(1+r) - \ln(1-r)] \quad (10)$$

e. Slope Significance Limit Test between Parent and Child Instruments

If one expects the child instrument to perform identically to the parent instrument for predictions, one should compute the confidence limits for comparison between the parent and child predicted values using criteria computed only from the parent instrument. So for a slope significance test one needs to look at the slope change between the parent instrument predicted values for a set of test samples over time and retain these results for this test. To accomplish this

one would compute the predicted values for a set of 20 test samples on the parent instrument one-week apart. Thus one would have two sets of predicted values designated by different measurement times as x_i for time 0 (the reference values), and for the same set of samples measured 1 week later designated as (y_i). (Note these test samples must be chemically stable over the 1 week period. This time duration depends upon the stability of the instruments and the recommended time between instrument alignment calibrations.)

For the computation of the acceptable slope Confidence Interval (C.I.) for the child instrument to be considered alike to the parent instrument, compute three basic sets of numbers. These are: (1) the set of parent predicted values at time 0 (x_i), (2) the set of parent predicted values at Week 1 (y_i), and (3) the set of predicted values from the simple linear regression between these x_i and y_i values, designated as ($\hat{y}_i$). Note that n is the number of x_i, y_i pairs (i.e., 20 for this example). From these three sets of values one may compute the standard deviation of the residuals for the predicted values, from the regression ($\hat{y}_i = b + mx_i$) as Equation 11.

$$s_{y/x} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} \tag{11}$$

Then compute the standard deviation for the desired slope as Equation 12.

$$s_m = \frac{s_{y/x}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}} \tag{12}$$

And now the confidence limit for the slope is given as Equation 13.

$$C..I._m = m \pm t \cdot s_m \tag{13}$$

Where t is the critical value of the t distribution, two-tailed test, at alpha = 0.05 and with degrees of freedom as n - 2 = 18. This value is 2.10 and can be found in any table of the Critical Values of the t distribution.

Using this test criteria we may test the child instrument slope following calibration transfer and compare the slope obtained to the confidence limits of the test measured on the same parent instrument over the experimental time interval. To complete this parent and child comparison one would designate the parent instrument results as x_i and the child measured results as y_i. Then the slope would be computed for the regression and must fall within the computed confidence limits for the time 0 and week 1 parent tests. This is a test of equivalence in slope between parent and child predicted values.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch011

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch011

Developing Global or Robust Models Including Variation between Instruments

Various methods have been proposed to produce the universal model, or a calibration that is mostly robust against standard instrument differences or changes with time as are common to commercial instruments of today. These have been referred to as robust models, or global models. For computing a robust model, various experimental designs have been constructed to better represent the product, reference values, and instrument calibration space and to include typical changes and interferents that should be included within the model for it to be broadly applicable. Using this approach one might design a factorial experiment for the composition of the calibration set to include multiple variations typically encountered during routine analysis. A list of some of these variations may consist of differences in sample pathlength, sample holder type, sample temperature, sample moisture content, flow rate, particle size, interferent content, instrument type, constituent ratios, sampling parameters, and others (*7*). These approaches will work for a period of time until the instrument conditions drift or the product or constituent chemistry changes. These types of changes are expected and thus routine recalibration (i.e., multivariate model updating) would be required as a standard procedure if the chemistry or instrument measurement changes are considered significant. A method for selection of specific robust wavelengths in MLR models that are more forgiving toward wavelength differences in interference filter based instruments has been demonstrated to be effective (*26*). A similar approach might be applied for computing robust PLS or PCR scores and loadings.

A method to reduce the effect of interference on NIR measurements has been demonstrated. This is a pre-processing method applying orthogonal signal correction (OSC). The goal of OSC is to remove variation from the spectral data, X, that is orthogonal to Y (*27*). This orthogonal variation is modeled by additional components for X and results in the decomposition, $X = t \cdot p' + t_o p_o' + e$, where t_o and p_o represent the scores and loadings for the orthogonal component and e represents the residual. By removing the Y-orthogonal variation from the data via $X - t_o p_o'$, OSC maximizes correlation and covariance between the X and Y scores to achieve more accurate NIR prediction (*28*).

a. Augmenting Models over Time

If instrument differences are significant, the predicted values between parent and child instruments will be unacceptably large. In these cases, one may begin to collect more samples on the child instrument and then re-compute the multivariate model using the majority of high leverage sample data from the child instrument. This in effect uses the calibration transfer as a temporary solution or bridge to building an accurate model on the child instrument. Such a practice is often used when the instrument differences are too significant for direct and accurate calibration transfer.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch011

b. Sample Selection To Improve Spectral Data

Sample selection methods have been used and perfected since the beginning of chemometric methods and spectroscopy. There are many methods and a variety of nomenclatures for these techniques. The purpose of such methods are to remove the redundancy in spectral data such that the most repetitive samples do not have excessive influence or leverage on the multivariate regression model. This provides a basis such that the regression line is more appropriately fitted to the extreme samples, including those with high and low analyte concentrations. Such methods of sample selection include: random subset selection, manual subset selection, spectral subtraction methods for "uniqueness" tests, stratified sample selection, discriminant based selection techniques using spectral distances, correlation matching techniques, PCA methods, and others. These methods are described in more detail in reference (*18*). One of the first successful approaches for sample selection was a process that used spectral subtraction to remove the unusual spectra from all of the other spectra described in reference (*29*). Even early use of these methods significantly improved the Standard Error of Prediction (SEP) for multiple constituents in forage analysis (*30*).

c. Spectral Data Transformation

This process consists in altering spectral data from the child instrument to be more like that measured on the parent instrument. Direct Standardization (DS), and Piecewise Direct Standardization (PDS) have been used most often for this procedure (*2*–*5*).

d. Local Methods

Locally weighted regression or local regression methods use spectral data and corresponding reference data to build a "local" calibration using only those samples near the unknown or test sample spectrum. For example, the unknown spectrum is measured and the sample spectra most like the unknown are selected from a resident database. The multivariate calibration model is then computed using only the local samples. The samples can be down weighted for use in the regression model based on distance from the unknown sample. This allows quite accurate prediction analysis when a variety of samples and instrument type data is incorporated into a spectral database. The first description of the use of this method for spectroscopy is referenced based on original work from the statistics community (*31*, *32*). A disadvantage is the requirement for large resident databases with reference chemistry values, and increased computational load requirements in order to provide real-time analytical results.

e. Use of Indicator Variables

A method has been used previously that simultaneously optimizes the calibration for multiple instruments and provides parametric t-tests for the differences between them. The method creates the calibration by running the

samples on several instruments (the more the better). All the data is added into the calibration, and indicator variables are used between the specific instruments. With this approach one obtains the optimum calibration corrected for all instruments; the coefficients of the indicator variables are the computed biases between the instrument results, and the t-tests reported from this method are valid for the corresponding bias values (*33*).

There are many conventional and unconventional approaches to calibration transfer. However the fact remains that significant differences in the instrument response between parent and child instruments causes the greatest variation in predicted results following calibration transfer. If instrument spectral profiles can be made statistically alike between instruments the transfer issue disappears. The additional challenges of relating specific reference laboratory results to results predicted using spectroscopy is another ongoing area of discovery and represents yet another problem set still to be resolved. Noting that absorption spectroscopy directly measures weight or moles per unit volume and not other arbitrary chemical or physical characteristics of samples is helpful.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch011

Formal Statistical Methods of Uncertainty

There are prescribed statistical methods for measuring the agreement between instruments following calibration transfer. These statistical methods used for evaluating the agreement between two or more instruments (or methods) for reported analytical results are formalized for commerce or medical devices. The emphasis is on acceptable analytical accuracy and confidence levels using two standard approaches: Standard Uncertainty/Relative Standard Uncertainty, and Bland-Altman "Limits of Agreement".

How To Tell if Two Instrument Predictions, or Method Results, Are Statistically Alike?

The main question when comparing parent to child instrument predictions, or a reference laboratory method to an instrument prediction, or results from two completely different reference methods, is how to know if the differences are meaningful or significant and when they are not. There is always some difference expected, since an imperfect world allows for a certain amount of "natural" variation. However when are those differences considered statistically significant differences, or when are the differences too great to be acceptable? There are a number of reference papers and guides to tell us how to compute differences, diagnose their significance, and describe the types of errors involved between methods, instruments, and analytical techniques of many types. Whether the analytical method is based on spectroscopy and multivariate calibration methods, other instrumental methods, or even gravimetric methods. One classic reference of importance is noted for comparing methods (*34*). This reference describes details regarding collaborative laboratory tests, ranking of laboratories for accuracy, outlier determination, ruggedness tests for methods, and diagnosing the various types of errors in analytical results.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch011

a. Standard Uncertainty and Relative Standard Uncertainty

The definitions of uncertainty are described by the U.S. National Institute of Standards and Technology (NIST), a National Metrological Institute (NMI), which is a non-regulatory agency of the United States Department of Commerce. NIST's purpose is to advance measurement science, measurement standards, and measurement technology. Their charter is to define measurements from first principles that can be verified world-wide and used as standards for making measurements of any kind related to commerce or technology. The NIST definition for Uncertainty is quite specific and is as follows (*35*, *36*).

Uncertainty Defined

NIST Definitions are given for uncertainty concepts.

"The **standard uncertainty** $u(y)$ of a measurement result y is the estimated standard deviation of y."

"The **relative standard uncertainty** $u_r(y)$ of a measurement result y is defined by $u_r(y) = u(y)/|y|$, where y is not equal to 0."

Meaning of Uncertainty

If the probability distribution characterized by the measurement result y and its standard uncertainty $u(y)$ is approximately normal (Gaussian), and $u(y)$ is a reliable estimate of the standard deviation of y, then the interval $y - u(y)$ to $y + u(y)$ is expected to encompass approximately 68 % of the distribution of values that could reasonably be attributed to the value of the quantity Y of which y is an estimate. This implies that it is believed with an approximate level of confidence of 68 % that Y is greater than or equal to $y - u(y)$, and is less than or equal to $y + u(y)$, which is commonly written as $Y = y \pm u(y)$.

Use of concise notation If, for example, $y = 1\,234.567\,89$ U and $u(y) = 0.000\,11$ U, where U is the unit of y, then $Y = (1\,234.567\,89 \pm 0.000\,11)$ U. A more concise form of this expression, and one that is in common use, is $Y = 1\,234.567\,89(11)$ U, where it understood that the number in parentheses is the numerical value of the standard uncertainty referred to the corresponding last digits of the quoted result.

The following use of the NIST nomenclature will demonstrate their definitions for uncertainty. In Equation 14 it is noted that

$$y = f(X_1, X_2, ..., X_N) \tag{14}$$

where y (is the estimated analytical value for any sample as a function of a series of measurement quantities such as X_1, X_2, ..., X_N; and where each X_i is an independent observation (or measurement). When using this nomenclature each set of measurements for every test sample is denoted as X_i measurements. We note the value (y_i) for each sample measurement is estimated as the sample mean from N independent measurements and is denoted as $X_{i,k}$, giving the relationship as follows in Equation 15.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

$$y_i = \overline{X}_i = \frac{1}{N}\sum_{k=1}^{N} X_{i,k} \tag{15}$$

So the estimated analytical value (y_i) is the mean for a number of measurements of a sample set ($\overline{X_i}$) using the analytical method prescribed. And it follows that the standard uncertainty $u(X_i)$ with reference to the measured values (X_i) is equal to the estimated standard deviation of the mean as Equation 16.

$$u(X_i) = s(\overline{X}_i) = \left(\frac{1}{n(n-1)} \sum_{k=1}^{N} (X_{i,k} - \overline{X}_i)^2 \right)^{\frac{1}{2}} \tag{16}$$

So to apply this to a set of comparative data for measurement set X_1 and X_2, the following relationship applies as Equation 17.

$$u(y_i) = u(X_i) = s(\overline{X}_i) = \left(\frac{1}{n(n-1)} \sum_{k=1}^{N} (X_{i,k} - \overline{X}_i)^2 \right)^{\frac{1}{2}} \tag{17}$$

where $u(y_i)$ is the estimated standard uncertainty for a series of measurements on multiple samples where the mean value of the measurements for each sample is used for comparison. The Equations 14-17 above are often used for multiple measurements of a single physical constant. For the calibration transfer application the date is as X_1 (parent) and X_2 (child) measurements for each test sample. The variance is computed for each of the 20 test samples and the pooled results are tabulated to yield an estimate of standard uncertainty $u(y_i)$ as Equation 18.

$$u(y_i) = u(X_i) = s(\overline{X}_i) = \left(\frac{1}{n(n-1)} \sum_{k=1}^{N} (\sigma_{i,k})^2 \right)^{\frac{1}{2}} \tag{18}$$

Relative Standard Uncertainty

This is denoted as $u_r(y_i) = u(y_i)/|y_i|$ and so the *relative standard uncertainty* is reported as Equation 19.

$$u_r(y_i) = u(y_i / |y_i|) \tag{19}$$

Confidence Levels are reported as follows for expressions 21 and 22.

$$For\_\_68\%: y_i \pm u(y_i) \tag{20}$$

$$For\_\_95\%: y_i \pm 2 \cdot u(y_i) \tag{21}$$

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch011

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch011

b. Using Bland-Altman "Limits of Agreement"

The Bland-Altman plot is used broadly in medical or clinical analysis. It represents a standard nomenclature for clinical data in an industry with stringent requirements for analytical accuracy. There is an entire series of publications related to this method. Table 3 data illustrates a virtual comparison of four instruments or four methods used for the same sample set. The results shown are for analyte concentration for each sample (each row is the same sample).

Table 3. Analytical data used for illustration. (Different instruments or methods are designated as A, B, C, and D)

Sample No.	*A*	*B*	*C*	*D*
1	12.4	12.1	14.9	16.1
2	12.9	12.5	15.5	16.8
3	14.0	13.9	16.8	18.2
4	16.0	15.9	19.2	20.8
5	13.2	12.9	15.8	17.2
6	12.8	12.7	15.4	16.6
7	14.5	14.9	17.4	18.9
8	13.0	13.4	15.6	16.9
9	13.6	13.5	16.3	17.7
10	12.7	12.6	15.2	16.5
11	14.2	14.6	17.0	18.5
12	16.3	16.4	19.6	21.2
13	17.8	17.9	21.4	23.1
14	18.0	18.5	21.6	23.4
15	14.5	13.9	17.4	18.9
16	17.2	17.5	20.6	22.4
17	14.4	14.6	17.3	18.7
18	15.2	15.7	18.2	19.8
19	16.6	16.5	19.9	21.6
20	13.5	13.1	16.2	17.6

The Bland-Altman paper describes the errors often made when comparing analytical methods (*10*). The authors summarize the contents of this paper as follows, "In clinical measurement comparison of a new measurement technique with an established one is often needed to see whether they agree sufficiently for

the new to replace the old. Such investigations are often analyzed inappropriately, notably by using correlation coefficients. The use of correlation is misleading. An alternative approach, based on graphical techniques and simple calculations, is described....” So what can be learned by using the techniques described in this paper to compare results from analytical methods? When methods are compared following calibration, one attempts to assess the degree of agreement between them. Bland and Altman discount completely the use of correlation as a useful parameter to assess analytical agreement. Their arguments are given in this discussion.

For this illustration, an initial comparison of different instruments or analytical methods are made as measurements A through D for each sample (Table 3). A line of equality plot is then made to compare the results. The various X,Y data points are plotted against a perfectly straight line of equality. The authors make the point that correlation (r) measures the strength of the relationship between two variables, but it does not measure the agreement between them (Figure 1). Perfect agreement is indicated by the data points lying directly on the line of equality. A perfect correlation is indicated if the points lie along any straight line. The authors emphasize that: (1) correlation indicates the strength of a relationship between variables - not that the analytical results agree; (2) a change in scale does not affect correlation, but drastically affects agreement; (3) correlation depends upon the range of the true quantity (analyte) in the sample; (4) tests of significance are mostly irrelevant between two similar analytical methods, and (5) data in poor agreement analytically can be highly correlated (Figure 2). Figure 2 shows three analytical sets with perfect correlation but poor agreement.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch011

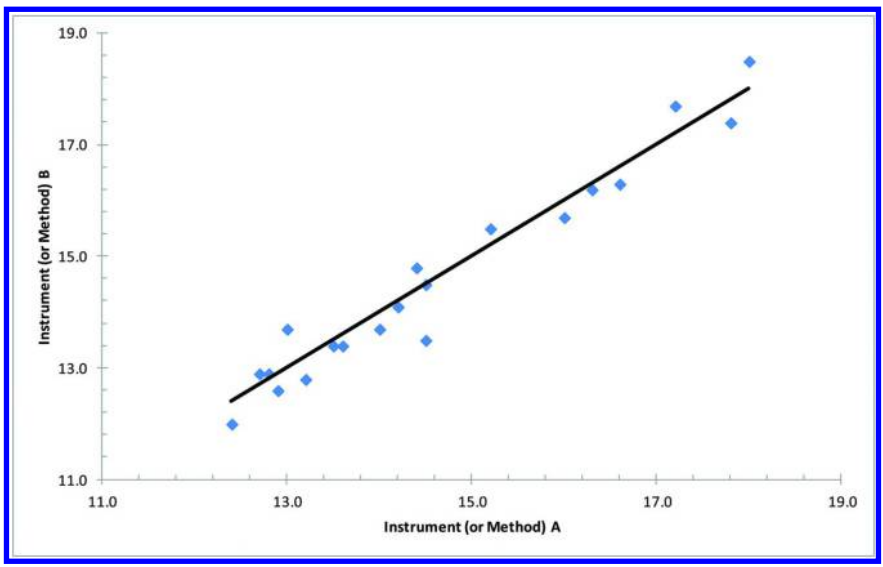

Figure 1. Instrument or Method A (x-axis) as compared to Instrument or Method B (y-axis), with data points compared to a perfect line of equality.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch011

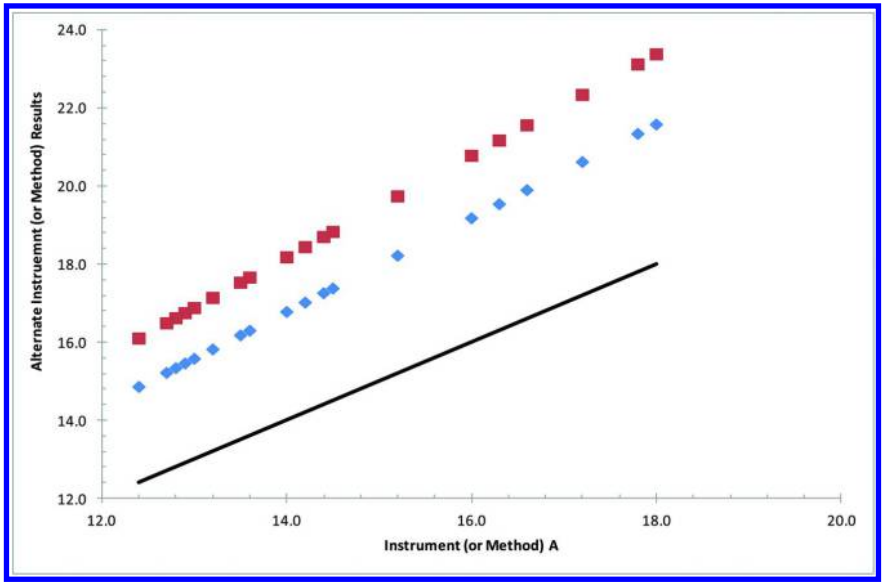

Figure 2. Instrument or Method A (x-axis) as compared to Instruments (or Methods) C and D (middle and top scatter plots), showing line of equality for perfect agreement with A (solid line). Note that C and D have a perfect correlation to A (r = 1.00), but are not in analytical agreement. This indicates correlation as an imperfect representation of agreement between methods.

A Bland-Altman plot (Figure 3), extremely familiar to clinical analysts, demonstrates a good visual comparison technique to evaluate the agreement between two methods or instruments, using the data comparison found in Table 3. The data indicates a comparison of different instruments or different methods for identical samples (as rows). The x-axis (abscissa) for each sample is represented by the average value for each sample obtained from the comparison results (using two methods or two instruments). The y-axis (ordinate) for each sample is represented by the difference between one method and the second method (or measurements from instruments A and B used for this example) for each test sample. Such a plot uses the mean and plus or minus two standard deviations as the upper and lower comparison thresholds.

To assess if the data are in close enough agreement for analytical purposes between A1 and B1 the bias or mean difference ($\overline{d}$) , the standard deviation of the differences (s or SD), and the expected "limits of agreement" are all computed. These are expressed as $\bar{d} \pm 2 \cdot s$ for a 95% confidence level.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch011

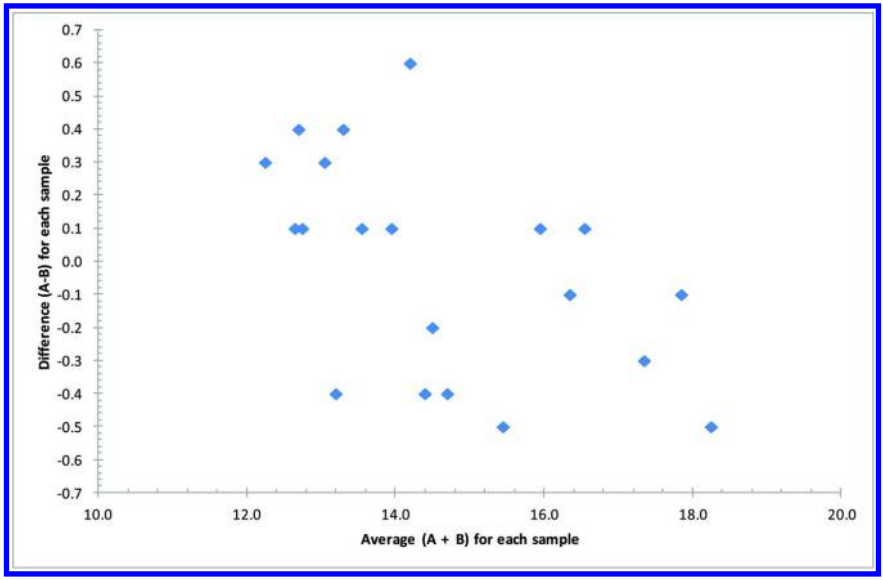

Figure 3. The Bland-Altman Plot indicating the difference plotted against the mean for each sample for Instrument (or Methods) for A and B. The average for A + B for each sample is plotted as abscissa(x-axis), versus difference of A-B plotted on y-axis. Perfect agreement demonstrates a horizontal line along the 0.0 y-axis.

The mean difference is computed as the average of all the differences between the comparative instruments, for each sample as Equation 22.

$$\bar{d}_i = \frac{\sum_{i=1}^{n}(A_i - B_i)}{n} \tag{22}$$

The standard deviation for this comparison for a set of sample measurements A and B is computed as Equation 23.

$$s_i = \sqrt{\frac{\sum_{i=1}^{n}(A_i - B_i)^2}{2n}} = \sqrt{\frac{\sum_{i=1}^{n} D_i^2}{2n}} \tag{23}$$

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

If the Bias, Standard deviation for all samples, and a comparison with the 95% confidence limits ($\overline{d}\pm 2s$) indicates the measured differences are considered too large for a 95% confidence of the result agreement between methods, the methods are not considered equivalent. On the other hand if these limits of agreement are acceptable for the application where they are used the results are equivalent. In a clinical situation a physician determines the level of accuracy or agreement required for critical intervention decision making; this would be analogous to a process control supervisor or analytical scientist assessing the acceptable level of agreement between comparative methods in order to use the alternate method or child instrument following calibration transfer as an acceptable and approved analytical substitute.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch011

Conclusions

There are multiple methods that have developed for use in transferring calibrations and for testing the efficacy of calibration transfer. Continued advancements in the design and manufacturing of instrumentation, as well as the application of chemometric and statistical methods is providing an increasingly scientific and metrical basis for the methods and validation of multivariate calibration transfer. Classic and advanced methods have been applied to analytical results in commerce and clinical analysis. Included below are references for the reader's further study of this subject of comparing two or more analytical methods or instruments following the transfer of calibrations. Other references discussing the mathematical details of comparing analytical methods are given in references (*37–42*).

References

1. Mark, H.; Workman, J. *Spectroscopy* **2013**, *128*, 1–9.
2. Wang, Y.; Veltkamp, D. J.; Kowalski, B. R. *Anal. Chem.* **1991**, *63*, 2750–2756.
3. Wang, Z.; Dean, T.; Kowalski, B. R. *Anal. Chem.* **1995**, *67*, 2379–2385.
4. Zhang, L.; Small, G. W.; Arnold, M. A. *Anal. Chem.* **2003**, *75*, 5905–5915.
5. Kowalski, B. R.; Veltkamp, D. J.; Wang, Y. U.S. Patent No. 5,459,677, Oct. 17, 1995.
6. Feudale, R. N.; Woody, N. A.; Tan, H.; Myles, A. J.; Brown, S. D.; Ferre, J. *Chemom. Intell. Lab. Syst.* **2002**, *64*, 181–192.
7. Abookasis, D. A.; Workman, J. J. *J. Biomed. Opt.* **2011**, *16*, 027001–027001-9.
8. Shenk, J. S.; Westerhaus, M. O. *Crop Sci.* **1991**, *31*, 1694–1696.
9. Shenk, J. S.; Westerhaus, M. O. U.S. Patent No. 4,866,644, September 12, 1989.
10. Bland, J. M.; Altman, D. G. *Lancet* **1986**, *1*, 307–310.
11. Mark, H.; Workman, J. *Chemometrics in Spectroscopy*; Elsevier, Academic Press: Amsterdam, 2007.

12. *ASTM E1655 - 05 Standard Practices for Infrared Multivariate Quantitative Analysis*; ASTM International: West Conshohocken, PA: 2012.
13. *ASTM E1944 - 98 Standard Practice for Describing and Measuring Performance of Laboratory Fourier Transform Near-Infrared (FT-NIR) Spectrometers: Level Zero and Level One Tests*; ASTM International, West Conshohocken, PA: 2007.
14. Workman, J.; Mark, H. *Spectroscopy* **2013**, *28*, 12–25.
15. Workman, J.; Mark, H. *Spectroscopy* **2013**, *28*, 28–35.
16. Workman, J.; Mark, H. *Spectroscopy* **2014**, *29*, 18–27.
17. Workman, J.; Mark, H. *Spectroscopy* **2014**, *29*, 14–21.
18. Workman, J.; Mobley, P.; Kowalski, B.; Bro, R. *Appl. Spectrosc. Rev.* **1996**, *31*, 73–124.
19. Mobley, P.; Kowalski, B.; Workman, J.; Bro, R. *Appl. Spectrosc. Rev.* **1996**, *31*, 347–368.
20. Bro, R.; Workman, J.; Mobley, P.; Kowalski, B. *Appl. Spectrosc. Rev.* **1997**, *32*, 237–261.
21. Workman, J.; McDermott, L. *JPAC* **1996**, *2*, 444–450.
22. Workman, J. In *Spectrophotometry, Luminescence and Colour; Science and Compliance*; Burgess, C., Jones, D. G., Eds.; Elsevier: Amsterdam, 1995; pp 369−384.
23. Tracy, D.; Hoult, R.; Ganz, A. U.S. Patent No. 5,303,165, April 12, 1994.
24. *Interference Testing in Clinical Chemistry; Approved Guideline*, 2nd ed.; CLSI document EP07-A2; Clinical and Laboratory Standards Institute: Wayne, PA, USA, 2005.
25. Fisher, R. A. *Biometrika* **1915**, *10*, 507–521.
26. Mark, H.; Workman, J. *Spectroscopy* **1988**, *3*, 28–36.
27. S. Wold, S.; Antti, H.; Lindgren, F.; Ohman, J. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 175–185.
28. Goicoechea, H. C.; Oliveri, A. C. *Chemom. Intell. Lab. Syst.* **2001**, *56*, 73–81.
29. Honigs, D. E.; Hieftje, G. M.; Mark, H. L.; Hirschfeld, T. B. *Appl. Spectrosc.* **1985**, *57*, 2299–2303.
30. Workman, J. *Proceedings of the 1986 Forage and Grasslands Conference*, Athens, Georgia, April 17−18, 1986; American Forage and Grass Council: Lexington, KY, 1986.
31. Naes, T.; Isaksson, T.; Kowalski, B. *Anal. Chem.* **1990**, *62*, 664–673.
32. Cleveland, W. S.; Deviin, S. J. *J. Am. Stat. Assoc.* **1988**, *83*, 596–610.
33. Burns D. A.; Ciurczak, E. W. Indicator Variables. *Handbook of Near-Infrared Analysis*, 2nd ed.; Marcel Dekker: New York, 2001; pp 351−362.
34. Youden, W. J.; Steiner, E. H. *Statistical Manual of the AOAC*; Arlington, VA, 1984.
35. Reference: See NIST definitions of uncertainty at URL: http://physics.nist.gov/cgi-bin/cuu/Info/Constants/definitions.html.
36. Taylor, B. N.; Kuyatt, C. E. *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results*; NIST Technical Note 1297; September 1994 Edition, 24 pages.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch011

37. Horwitz, W. *Anal. Chem.* **1982**, *54*, 67A–76A.
38. Hall, P.; Selinger, B. *Anal. Chem.* **1989**, *61*, 1465–1466.
39. Rocke, D.; Lorenzato, S. *Technometrics* **1995**, *37*, 176–184.
40. Miller, J. C; Miller, J. N. *Statistics for Analytical Chemistry*, 2nd ed.; Ellis Horwood: Upper Saddle River, NJ, 1992; pp 63–64.
41. Dixon, W. J.; Massey, F. J., Jr. In *Introduction to Statistical Analysis*, 4th ed.; Dixon, W. J., Ed.; McGraw-Hill: New York, 1983; pp 377, 548.
42. Rohrabacher, D. B. *Anal. Chem.* **1991**, *63*, 139–146.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch011

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch012

Chapter 12

Approaching the Chemometric Modeling of Realistically Diverse Biochemical Data

Jeffrey A. Cramer*

Naval Research Laboratory, 4555 Overlook Avenue SW, Washington, DC 20375, United States
***E-mail: jeffrey.cramer@nrl.navy.mil**

Every biomolecule can potentially interact with every other biomolecule with which it comes into contact. This large and diverse array of potential biomolecules and potential interactions becomes problematic when attempting to produce predictive biochemical models intended to reliably derive actionable information under the maximum possible number of realistic circumstances. The following chapter presents an overview of how biochemical modeling challenges have historically been addressed in –omics fields using the tools found in chemometrics and related statistical analysis and data modeling fields, as well as how the challenge of realistic and robust biochemical analysis and modeling might be addressed in the future.

Introduction

Realistic attempts to comprehensively model or otherwise characterize biochemical data must contend with the high levels of chemical complexity found in metabolically-active biochemical systems (*1*) to achieve high levels of reliability. Grappling with this complexity is especially important within the various fields of biologically-oriented "-omics" research (such as metabolomics, proteomics, and genomics) that routinely utilize high-throughput techniques, such as hyphenated mass spectrometry (MS) (*2*) paired with database searches and other specialized software applications (*3*), to produce large amounts of biochemically meaningful data (*4*), usually with some biochemically meaningful endpoint as an analytical target (*5*). Because biological processes are typically the results of

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

interactions between many biomolecules (*6*, *7*), and because these biomolecules have a diverse array of chemical and physical properties (*8*), up to and including their three-dimensional structures (*9*) and how these three-dimensional structures interact with one another (*10*) and the potential therapeutics to be introduced as a consequence of analysis (*11*), comprehensive biochemical modeling quickly becomes a non-trivial task. This non-triviality is further compounded by the realistic scope of comprehensive biomolecule classification (*12*), the dynamic biomolecular changes brought on by factors as random as mutations (*13*), the fact that any given biomolecule might not fundamentally interact with the employed analytical method in an anticipated fashion (*14*, *15*), and all potential interferents to be found in the extended environments within which targeted biomolecules may exist (*16*), up to and including the presence of distinct microbiomes within larger individuals (*17*).

Perhaps understandably, then, the primary method by which to address the challenges presented by complex biochemical interactions is to not address them at all, instead finding biochemical actors and relationships that might be limited in scope, but are believed (correctly or not (*18*)) to be robust enough to be found in a majority of circumstances, regardless of the competing interactions of individual biomolecules and their environment. While this limited approach can sometimes be a very successful one, more in-depth and far-reaching analyses can potentially encompass a greater degree of realistic biodiversity, thus providing more reliable and more useful information more often to the end-user.

This chapter delivers an overview of how the modeling and analysis tools found in the field of chemometrics, as well as related statistical analysis and data modeling fields, have historically been deployed to address complex biochemical challenges in –omics fields. Three overarching trends are apparent in this history, two chemically-specific trends and one functionality-specific trend. The two chemically-specific modeling trends make for a bifurcated "forest and trees" approach to chemically-specific biochemical modeling: in the "trees" branch of the approach, discrete biomarkers are identified and characterized, whereas, in the "forest" branch of the approach, the interactions of multiple biochemical factors are identified and characterized. In contrast, the functionality-specific trend foregoes chemical specificity to, as directly as possible, answer the question that the end-user would actually like answered by the data. An extrapolation will also be made regarding how these three trends might be combined to address biochemical challenges in a robust and comprehensive fashion in the future.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch012

Timeline Note

Although this book is intended to cover four decades' worth of the history of chemometrics, the –omics fields as a whole have not existed for forty years (the term "proteomics" only dates to 1995, for instance (*19*)), and the interface between chemometrics and –omics is a similarly recent conceptual space. This is the reason that the vast majority of references in this chapter have appeared in the literature within the past ten to fifteen years.

Biomarkers: Modeling the Trees

When modeling discrete biological phenomena, it is sometimes possible to find similarly discrete, biochemically meaningful indicators (i.e. biomarkers) that directly infer the actionable information desired (*20*). If a biomarker, such as a protein or a peptide or even an abnormal quantity of a single atom or ion, is only produced in response to the specific biological phenomenon being interrogated, then the biomarker is completely selective to that phenomenon, and even biomarkers that are not completely selective can still be useful analytically. The use of biomarkers is also aided by the fact that living systems, generally speaking, at least attempt to maintain homeostasis in the presence of external stressors (*21*), potentially allowing biomarker-based diagnostics to remain relevant despite said stressors.

Finding these biomarkers in the raw data in the first place, of course, can be a challenge in and of itself. Although direct comparisons between data sets collected from samples with and without biomarkers are sometimes possible (*22*) (and can sometimes be further aided by feature detection algorithms (*23*) and biochemically relevant statistical techniques (*24*)), chemometrics tools and related data analysis tools have also been effectively deployed in the recent past to extract biomarker information from large, complex biochemical data sets.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch012

Hyphenated MS Data Preprocessing

One of the most potent tools in identifiying chemically-specific biomarkers is the chemically-specific analysis techniques of hyphenated MS, and one of the most notable phenomena to discuss in the context of hyphenated MS data is that of misaligned peak data. Misaligned peaks cause nonlinear data variations that render linear modeling procedures and feature-based database searches problematic. Many hyphenated MS peak alignment approaches have already been collated in previous -omics reviews (*25*, *26*), and a few biomarker-specific examples of peak alignment strategies have even appeared in the literature. Some examples:

- *2004*: An alignment strategy specifically geared towards biomarker discovery in liquid chromatography MS (LC-MS) data, based on the maximization of spectral similarities, was developed (*27*).
- *2005*: A software tool called PepMatch was developed to align peptide features, extracted from LC-MS data, across multiple samples (*28*).
- *2006*: The peptide- and protein-specific alignment technique of ordered bijective interpolated warping (OBI-Warp) was developed and applied to electrospray ionization liquid chromatography MS (ESI-LC-MS) data (*29*).
- *2006*: Nonlinear robust ridge regression was applied in the alignment of the LC-MS data collected from biochemically complex plant cell culture samples (these samples differed in the manner in which they were illuminated prior to sampling, resulting in detectable biomarker changes) (*30*).

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

- *2010*: Metabolite-oriented peak alignment was extended to two-dimensional gas chromatography - time-of-flight-MS (GC x GC-TOF-MS) using landmark peak identifications (acquired from discrete biomarkers) and the local partial linear fitting of these peaks to compensate for unwanted time shifts (*31*).

It should be noted here that, even after data alignment is satisfactorialy accomplished, the aligned data may still require preprocessing to smooth out unwanted data features to facilitate biomarker discovery. This smoothing can be accomplished through the use of a hierarchical, non-global multivariate curve resolution (MCR) strategy, as was seen in 2005, in work that also included validation in the form of unsupervised and supervised biomarker-based pattern recognitions (*32*).

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch012

Database Searches

During the course of identifying discrete biomarkers, it is routinely necessary to compare data collected in-house to archived databases in order to objectively establish biomarker identities. Feature detection and selection are critical in the context of effective database searches, and biomarker-specific algorithmic search tools have been developed to facilitate these operations. Some examples:

- *2006*: A quality control metric called mass deviance, itself based on comparisons between the masses of identified features and theoretical peptide masses as they would have appeared in the original LC-MS data set, was developed (*33*).
- *2010*: A type of proteomics-oriented tandem MS spectral deconvolution, in which the collections of peaks arising from related fragment ions possessing the same chemical formulae and charge states, but arising from different isotopic distributions, was produced to improve database comparisons (*34*).
- *2013*: The *t* test, Mann-Whitney-Wilcoxon (mww) test, nearest shrunken centroid (NSC), linear support vector machine-recursive features elimination (SVM-RFE), principal component discriminant analysis (PCDA), and partial least squares discriminant analysis (PLSDA) were all compared as potential methods by which to select biomarker-relevant features in LC-MS data sets (*35*).

Once a database search is performed and biomarker identities have been assigned, however, a validation procedure may still be necessary to ensure that the assigned identity is, in fact, accurate. Such a validation procedure was produced in 2008 by constructing a Bayesian classfier, using an expectation-maximization (EM) algorithm, that incorporates decoy (i.e. false) peptide matches to allow for validation in those circumstances in which more common validation software would not perform as expected (*36*).

It should also be noted that, even after biomarkers have been identified and confirmed to exist in a given data set *qualitatively*, the actual data used to produce

these identities might not provide accurate *quantitative* information, as could be the case with hyphenated MS techniques. To compensate for this phenomenon, in 2008, parallel factor analysis (PARAFAC) was applied to GC x GC-TOF-MS data to more accurately quantify peak volumes and, hence, more accurately quantify the metabolic activities of 44 distinct biomarkers (*37*).

Pattern Recognition

In those cases of exploratory data analysis in which biomarker identities are not necessarily known *a priori*, chemometrics-based pattern recognition and machine learning tools can and have been deployed. Some examples:

- *2006*: Fisher ratios were used to find variances between sets of GC x GC-TOF-MS data, a technique used in the referenced work to generate a list of biomarkers from a set of urine metabolite samples (*38*).
- *2008*: PLS and PCA were applied to GC-MS data to obtain regression coeffecients that were subsequently used to select discrete biomarkers corresponding to acute liver failure in rats (*39*)
- *2008*: PCA and classification modeling were applied to trace element data collected from plant samples to determine which specific elements could be used to indirectly characterize aspects of plant physiology (*40*).
- *2009*: Electrospray ionization MS (ESI-MS) data collected from two sets of mouse liver extracts, each with different cholesterol levels, were subjected to PCA, not only to determine how the data clustered in the subsequent PC space, but also to obtain loadings indicating which biomarkers contributed to the observed clustering (*41*).
- *2010*: Decision tree-based classifiers were applied to an established metabolomics database to mine hyphenated MS data features for the presence of unidentified biomarkers (*42*).

Data Degeneracy

Obviously, the search for discrete biomarkers has certain limitations. In the context of proteomics, for instance, different proteins can share extensive portions of their respective peptide sequences, a phenomenon known as peptide degeneracy. This degeneracy sometimes renders attempts to infer the presence of a given protein using peptide-based analytical data problematic, especially when accommodating realistically complex biological samples (*43*). Though it should be noted that proteins sharing a great deal of their peptide sequences might be similar enough for the purposes of deriving actionable information from said peptides (*44*), this will not always be the case, and realistically comprehensive data analysis and modeling strategies must take these degeneracies into account, as was done in the following examples:

- *2008*: A hierarchical statistical modeling methodology was developed to simultaneously account for the uncertainties in both tandem MS-based

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch012

peptide identifications and protein identifications based on these same peptide identifications (*45*).

- *2010*: A normalized spectral abundance factor (itself a quantification based on the counted number of hyphenated MS spectra associated with a given protein identification) was modified in such a manner that *shared* spectral counts corresponding to shared peptide identifications were distributed based on *unique* spectral counts corresponding to unique peptide identifications (*46*).
- *2010*: A Bayesian method was employed to reduce the relevance of degenerate peptide information during the course of MS-based protein biomarker identification (*47*).

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch012

Interactions: Modeling the Forest

As indicated previously, real-world biochemistry does not occur in a metaphorical vacuum. The discovery and modeling of discrete biomarkers, and knowing where said biomarkers are generally located at a macroscopic level (thus allowing proper sampling strategies to be developed), will not necessarily take all possible (or even all realistic) biochemical interactions and interferences into account. This potential limitation to robust biochemical analysis and modeling has, of course, not escaped the notice of -omics practitioners (*48*, *49*).

Pattern Recognition

Several methods have been developed using chemometrcs and related analysis and modeling techniques to more thoroughly evaluate these complex interactions in a biologically realistic fashion. Some examples:

- *2005*: Both consensus PCA (CPCA) and multi-block PLS (MBPLS) were used to analyze GC-MS and LC-MS data simultaneously in a fused fashion. This was done specifically to correlate this data directly to a fermentation reaction in a manner unavailable when making use of either data set separately, due to the different metabolites detected using each technique and the complex metabolic pathways being utilized by all of the detected metabolites during the course of the fermentation (*50*).
- *2006*: PCA was applied to three sets of data (collected using nuclear magnetic resonance, or NMR, spectroscopy; ultra-performance LC-MS, or UPLC-MS; and GC-MS) obtained from plasma collected from two different rat populations, and the resulting score clusters varied in a manner that could be used to discriminate between these two sample populations based on multiple interrelated metabolic differences (*51*).
- *2006*: A hybrid approach (combining genetic algorithms, or GA, and artificial neural networks, or ANN) was developed to find groupings of interrelated biomarkers within the data collected from gene expression microarrays. In the approach, GA was used to select subsets of the

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

expressed genes, and an ANN was employed to inform the production of the GA fitness function (*52*).

- *2008*: Six different dimensionality reduction strategies (PCA; linear discriminant analysis, or LDA; classical multidimensional scaling; isometric mapping; locally linear embedding, or LLE; and Laplacian eigenmaps) were applied to gene and protein expression data. These strategies were evaluated in terms of how well the subsequently-generated, dimension-reduced data sets maintained their abilities to represent known classes, and find novel subclasses, within the parent biochemical data (*53*).
- *2008*: PLS, uninformative variable elimination PLS (UVE-PLS), continuum power regression (CPR), a hybrid technique combining UVE-PLS and CPR, classification and regression trees (CART), stepwise multiple linear regression (MLR), and genetic algorithm MLR (GA-MLR) were all employed to reduce microarray data dimensionalities (and, hence, increase data interpretability) through feature selection in an attempt to uncover interrelated gene expressions (*54*).
- *2012*: A novel sparse MBPLS (sMBPLS) modeling procedure was applied to a multi-dimensional data set consisting of DNA methylation (DM) data, copy number variation data, microRNA expression (ME) data, and gene expression (GE) data to more comprehensively elucidate the complex and interacting mechanisms behind gene regulation (*55*).
- *2012*: A joint non-negative matrix factorization was simultaneously applied to sets of DM, ME, and GE data to elucidate complex cellular activities in order to classify sample sub-groups (*56*).
- *2012*: Goeman's global test (a technique used in genomics for finding differences in how groups of genes express themselves in RNA microarray data) was applied to metabolomics-based hyphenated MS data. The use of this technique is intended to establish differences between metabolic conditions at the metabolic pathway level (*57*).
- *2012*: A technique dubbed individual differences scaling (INDSCAL, itself related to PARAFAC) was used to directly focus upon and correlate the differences found between multiple metabolites as reported in metabolite profiles (*58*).
- *2012*: Simultaneous component analysis (SCA) was combined with INDSCAL-like constraints to more comprehensively interrogate the overall metabolic differences between the specific compound measurements of individual plant samples (*59*).
- *2013*: The technique of weighted correction network analysis (WGCNA) was used to derive eigenvalues from the hyphenated MS data collected from maize kernels, resulting in the non-targeted collection of biomarkers. These eigenvalues and their associated biomarkers, in turn, elucidated metabolic pathways (*60*).
- *2013*: Both the primary and secondary metabolism of plant samples was modeled using a combination of GC-MS and LC-MS with the explicit goal of understanding plant-environment interactions. This work combined the GC-MS and LC-MS data sets in the context of a

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch012

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

single modeling challenge and employed Granger causality metrics (derived from the field of econometrics) to find metabolically relevant relationships in the data (*61*).

- *2013*: MBPLS and multi-block PCA (MBPCA) were used to combine two separate GC-MS data sets to facilitate the discovery of not only the discrete biomarkers associated with meat spoilage, but also (by means of a Bayesian network analysis) how these biomarkers are interrelated to one another (*62*).

Larger-Scale Interactions

The biochemical interactions that were interrogated in the previous examples were primarily focused on the level of individual biomolecular interactions. However, there are other, larger-scale interactions that might be explored using chemometric tools. For example:

Multi-Protein Complexes

The three-dimensional structures of individual and disparate proteins, due to both chemical and geometric affinities, often spatially associate with one another in very specific protein complexes (*63*). These three-dimensional associations are, unfortunately, difficult to perceive using analytical techniques, such as hyphenated MS, unless modifications are made to the complexes themselves to render them more obvious. Protein complexes, therefore, are typically interrogated experimentally using various forms of chemical cross-linking, a technique that produces additional chemical bonds between proteins in a given complex. Upon digestion, these additional chemical bonds hold portions of the original proteins together in a manner commensurate with their original configurations within the complex, information that can then be subsequently extracted using hyphenated MS (*64*), matrix-assisted laser desorption ionization MS (MALDI-MS) (*65*), or combinations of these and other experimental techniques and statistical methods (*66*) combined with commensurate software applications (*67*).

It should be noted here that alternatives that measure more generalized protein-protein affinities than those described above are available if less precise three-dimensional information is necessary (as might be the case if extensive databases of pre-existing information are to be employed) (*68*). Also, mathematical models (up to and including Quantitative Structure-Activity Relationships, or QSARs (*69*)) can be used to extrapolate at least some aspects of how individual proteins are likely to interact with one another in three-dimensional spaces (*70*, *71*).

Imaging

Knowing where individual biomarkers, and the interactions that can be found between them, are located macroscopically in living and/or environmental samples

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch012

might also be used to obtain a more complete picture of overarching biological phenomena. Biologically meaningful results have already been extracted from imaging data using chemometric modeling and related data analysis strategies as applied to morphological phenomena such as size and shape (*72–75*).

Metaproteomics

The full characterization of the biomolecular complement of mixed communities of organisms (*76*) has its utility in the commensurately full characterization of the inter-organism interactions associated with these communities and the large-scale effects that result from these interactions (*77*). For example, marine biofilms are collections of many disparate organisms, and metaproteomic work was performed in 2012 to determine how best to obtain insights into the overall compositions of sampled biofilms from LC-MS/MS data sets (*78*).

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch012

Functionality: Answering the End-User's Question

Characterizing individual biomolecules and how they interact in overall biochemical systems is a necessary step in comprehensively understanding what is happening in a given biochemical system. Obviously, however, there are many instances in which a comprehensive understanding of a sample's biochemistry is not necessary to address basic macromolecular, macroscopic questions, such as in the case of bioprocess quality control (*79*). In these circumstances, data as compositionally informative as hyphenated MS data might not even be necessary, and other types of robust, lower-cost instrumentation (typically still possessing the second-order advantage (*80*) to compensate for uncalibrated biochemical and environmental interferents as well as some irregularities in instrumentation) can be employed (*81*) provided that one has taken into account the challenges associated with employing chemically non-specific techniques when analyzing chemically and morphologically diverse samples (*82*). Of course, the use of data that are less chemically specific than hyphenated MS data renders chemometrics and other statistical methods no less useful and, in some cases, necessary to derive actionable information.

It should be noted that the following examples of this avenue of research tend slightly towards the applications of variable selection techniques, which makes sense in the present context because such techniques can potentially be quite useful with respect to the modeling of discrete biological phenomena in the presence of realistic biochemical interferences:

- *2001*: Two-dimensional scanning fluorimetry data collected from a biofilm were correlated to the biological degradation of chlorinated organic compounds by said biofilm using ANNs (*83*).
- *2003*: The musts of white grapes were classified according to their corresponding grape variety by fusing data from aroma sensors,

Fourier transform infrared (FT-IR) spectrometry, and ultraviolet (UV) spectrometry using a Bayesian approach (*84*).

- *2005*: Microbial growth, in terms of both biomass and the presence of specific biogenic components, was modeled using PLS, radial basis function (RBF) networks, and variable selection via self-organizing maps (SOMs), as applied to both dielectric spectroscopy and two-dimensional fluorescence spectroscopy (*85*).
- *2006*: Consensus PLS (cPLS), in which multiple predictions obtained from multiple individual models are combined into a single prediction, was applied to near-infrared (NIR) data collected from corn samples to predict moisture, oil, protein, and starch contents (*86*).
- *2006*: The behaviors of multiple fluorescent compounds, and how these behaviors correlate to the overall growth of yeast and the glycerol contents of individual samples, were modeled using a combination of PARAFAC and PLS as applied to multiwavelength, two-dimensional fluorescence measurements (*87*).
- *2007*: Several versions of PLS, each making use of a different variable selection strategy, were applied to NIR data collected from apples to predict soluble solid contents (*88*).
- *2009*: Dynamic time warping was used to align the peaks of a GC x GC data set (with the utilized MS detector's signal summed as a single total ion current) in order to more effectively allow both PCA and independent component analysis (ICA) to differentiate between the three types of tobacco used to produce the data (*89*).
- *2010*: Nicotine content was correlated quantitatively to the NIR data collected from tobacco samples by applying boosting PLS (*90*).
- *2010*: An antibiotic production process was monitored with multiwavelength fluorescence spectroscopy (the PARAFAC loadings of which were correlated to known fluorophores) and gas analyzer data; PLS, multilinear PLS, and locally weighted regression (LWR), in concert with multiple variable selection strategies, were used to correlate the data to biomass and amino acid concentration (*91*).

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch012

The Future of Functionality

Based on the three general trends presented in this chapter, the manner in which to achieve robust and realistic biochemically-oriented chemometrics modeling might seem obvious: combine the use of hyphenated MS data (which allows for the chemically-specific discovery of discrete biomarkers and the extrapolation of chemically-specific biomarker interactions) with practical, actionable functionality modeling. This would, essentially, allow for functionality-level modeling that would have direct, thorough, and robust biochemical information "baked in" to the raw data, thus allowing modeling and analysis operations to address the goals of end-users as directly, thoroughly, and robustly as possible. Imaging versions of MS already exist (*92*) to collect chemically-specific large-scale morphological data, if as much were required to

address biological heterogeneities (though the use of chromatographic separations would need to be re-evaluated in this context). One could even envision hyphenated MS data sets sufficiently comprehensive enough to address the challenges of metaproteomics and data analysis strategies sufficiently (artificially) intelligent enough to predict and accommodate the existence of multi-protein complexes.

Generally speaking, however, models are only as good as the data upon which they are based. Unless one intends to include every biomolecule that has ever and could theoretically ever exist on this planet with inordinate amounts of real and theoretical training data, determinations will likely have to be made regarding exactly how robust and realistic modeling solutions arising from this course of action can actually be. Even if the ideal of infinitely robust and realistic modeling were to be abandoned, though, attempting to model data sets representing large numbers of potential biomolecules and relatively few numbers of sampled individuals is already a recipe for overfitting (*93*), and this becomes all the more problematic if the challenge becomes one of accomodating the maximum possible amount of chemically-specific biodiversity for realistically comprehensive functionality-specific hyphenated MS data modeling.

The direct modeling of functionality using biomolecular data in a manner even approaching realistic comprehensiveness, therefore, must be accompanied by a methodology to accommodate unknown and unknowable biomolecules and biomolecular interactions. The beginnings of such a methodology might be found in the general idea of situational awareness (*94*), which combines the high-level or information-level data fusion of multiple types of data with probabilistic machine learning algorithms (such as Bayesian belief networks (*95*)) to identify meaningful events under a maximally diverse array of conditions. Combined with adaptive learning methodologies such as particle swarm optimization (*96*), it would likely be possible to construct robust and realistic (though not necessarily *perfectly* robust and realistic) biochemical functionality models that can adjust themselves dynamically to new, untrained conditions.

Other Considerations

Secondary Metabolites

Metabolically active biomolecules that are not strictly necessary for maintaining essential metabolic functions can be modeled both in and of themselves and as a means of indirectly inferring more critical pieces of actionable information. For example, using secondary metabolites, the presence and identity of individual microbes can be established (*97*) and biochemically meaningful comparisons can be made between them (*98*). The use of secondary metabolites becomes especially interesting as a methodology by which to characterize entire microbial colonies because such multi-microbe collectives would be expected to produce collections of secondary metabolites indicative of overall biochemical conditions throughout the colony. The relevant data from these multi-microbe colonies could also be collected in a minimally-invasive, minimally-destructive fashion, as was accomplished in 2012 using nanospray desorption ESI-MS (*99*).

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.
Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch012

Experimental Design

It should finally be noted here that the analysis of large data sets for the purposes of biomarker discovery, and perhaps even functionality modeling, could potentially benefit from the application of deliberate experimental design principles to minimize duplicated effort (*100*). This is especially true in those cases in which a data feature corresponding to a biomarker might not be as biochemically important in the context of a single experiment as in the context of multiple experiments (*101*). Such a circumstance would present the challenge of minimizing the number of samples necessary to ensure that such a biochemical importance would remain apparent.

Conclusions

The preceding chapter has presented an overview of ongoing research progress along the interface of chemometrics and the various –omics fields. This progress was reported upon in the context of three overall trends and how these trends might be coordinated into a single overarching methodology to address realistically diverse and complex biochemical challenges. This potential coordination notwithstanding, efforts to model and otherwise characterize biomarkers, biochemical interactions, and biochemical functionality will most likely experience future increases in accuracy, precision, and scope for as long as the analysis of –omics biochemical data is deemed a productive avenue of research.

References

1. Jansen, J. J.; Westerhuis, J. A. *Metabolomics* **2012**, *8*, S1–S2.
2. Lokhov, P. G.; Archakov, A. I. *Biochemistry (Mosc) Suppl. Ser. B: Biomed. Chem.* **2009**, *3*, 1–9.
3. Haga, S. W.; Wu, H.-F. *J. Mass Spectrom.* **2014**, *49*, 959–969.
4. Madsen, R.; Lundstedt, T.; Trygg, J. *Anal. Chim. Acta* **2010**, *659*, 23–33.
5. McShane, L. M.; Cavenagh, M. M.; Lively, T. G.; Eberhard, D. A.; Bigbee, W. L.; Williams, P. M.; Mesirov, J. P.; Polley, M.-Y. C.; Kim, K. Y.; Tricoli, J. V.; Taylor, J. M. G.; Shuman, D. J.; Simon, R. M.; Doroshow, J. H.; Conley, B. A. *Nature* **2013**, *502*, 317–320.
6. Kitano, H. *Nature* **2002**, *420*, 206–209.
7. Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M.-C.; Yates, J. R., III. *Chem. Rev.* **2013**, *113*, 2343–2394.
8. Sugimoto, M.; Kawakami, M.; Robert, M.; Soga, T.; Tomita, M. *Curr. Bioinfor.* **2012**, *7*, 96–108.
9. Arnaud, C. H. *Chem. Eng. News* **2013**, *91* (20), 11–17.
10. Wilkins, M. *Expert Rev. Proteom.* **2009**, *6*, 599–603.
11. Moellering, R. E.; Cravatt, B. F. *Chem. Biol.* **2012**, *19*, 11–22.
12. Meyer, B.; Papasotiriou, D. G.; Karas, M. *Amino Acids* **2011**, *41*, 291–310.
13. Li, X.-B.; Qiao, B.; Yuan, Y.-J. *Biotechnol. Appl. Biochem.* **2006**, *45*, 107–118.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch012

14. Leary, D. H.; Hervey, W. J., IV; Deschamps, J. R.; Kusterbeck, A. W.; Vora, G. J. *Mol. Cell. Probes* **2014**, *27*, 193–199.
15. Pierson, N. A.; Chen, L.; Valentine, S. J.; Russell, D. H.; Clemmer, D. E. *J. Am. Chem. Soc.* **2011**, *133*, 13810–13813.
16. Zhu, P.; Bowden, P.; Zhang, D.; Marshall, J. G. *Mass Spectrom. Rev.* **2011**, *30*, 685–732.
17. Deatherage Kaiser, B. L.; Li, J.; Sanford, J. A.; Kim, Y.-M.; Kronewitter, S. R.; Jones, M. B.; Peterson, C. T.; Peterson, S. N.; Frank, B. C.; Purvine, S. O.; Brown, J. N.; Metz, T. O.; Smith, R. D.; Heffron, F.; Adkins, J. N. *PLoS ONE* **2013**, *8*, e67155 (1−13).
18. Ioannadis, J. P. A. *PLoS Med.* **2005**, *2*, e124 (0696−0701).
19. Yadav, S. P. *J. Biomol. Technique* **2007**, *18*, 277.
20. Baxter, I.; Hosmani, P. S.; Rus, A.; Lahner, B.; Borevitz, J. O.; Muthukumar, B.; Mickelbart, M. V.; Schreiber, L.; Franke, R. B.; Salt, D. E. *PLoS Genet.* **2009**, *5*, e1000492 (1−12).
21. Stelling, J.; Sauer, U.; Szallasi, Z.; Doyle, F. J., III; Doyle, J. *Cell* **2004**, *118*, 675–685.
22. Wang, W.; Zhou, H.; Lin, H.; Roy, S.; Shaler, T. A.; Hill, L. R.; Norton, S.; Kumar, P.; Anderle, M.; Becker, C. H. *Anal. Chem.* **2003**, *75*, 4818–4826.
23. Prakash, A.; Mallick, P.; Whiteaker, J.; Zhang, H.; Paulovich, A.; Flory, M.; Lee, H.; Aebersold, R.; Schwikowski, B. *Mol. Cell. Proteom.* **2006**, *5*, 423–432.
24. Guo, W.; Yang, M.; Xing, C.; Peddada, S. D. *BMC Bioinf.* **2012**, *13*, 177 (1−8).
25. Katajamaa, M.; Orešič, M. *J. Chromatogr. A* **2007**, *1158*, 318–328.
26. America, A. H. P.; Cordewener, J. H. G. *Proteomics* **2008**, *8*, 731–749.
27. Radulovic, D.; Jelveh, S.; Ryu, S.; Hamilton, T. G.; Foss, E.; Mao, Y.; Emili, A. *Mol. Cell. Proteom.* **2004**, *3*, 984–997.
28. Li, X.; Yi, E. C.; Kemp, C. J.; Zhang, H.; Aebersold, R. *Mol. Cell. Proteom.* **2005**, *4*, 1328–1340.
29. Prince, J. T.; Marcotte, E. M. *Anal. Chem.* **2006**, *78*, 6140–6152.
30. Fischer, B.; Grossmann, J.; Roth, V.; Gruissem, W.; Baginsky, S.; Buhmann, J. M. *Bioinformatics* **2006**, *22*, e132–e140.
31. Wang, B.; Fang, A.; Heim, J.; Bogdanov, B.; Pugh, S.; Libardoni, M.; Zhang, X. *Anal. Chem.* **2010**, *82*, 5069–5081.
32. Jonsson, P.; Johansson, A. I.; Gullberg, J.; Trygg, J.; A, J.; Grung, B.; Marklund, S.; Sjöström, M.; Antti, H.; Moritz, T. *Anal. Chem.* **2005**, *77*, 5635–5642.
33. Piening, B. D.; Wang, P.; Bangur, C. S.; Whiteaker, J.; Zhang, H.; Feng, L.-C.; Keane, J. F.; Eng, J. K.; Tang, H.; Prakash, A.; McIntosh, M. W.; Paulovich, A. *J. Proteom. Res.* **2006**, *5*, 1527–1534.
34. Liu, X.; Inbar, Y.; Dorrestein, P. C.; Wynne, C.; Edwards, N.; Souda, P.; Whitelegge, J. P.; Bafna, V.; Pevzner, P. A. *Mol. Cell. Proteom.* **2010**, *9*, 2772–2782.
35. Christin, C.; Hoefsloot, C. J.; Smilde, A. K.; Hoekman, B.; Suits, F.; Bischoff, R.; Horvatovich, P. *Mol. Cell. Proteom.* **2013**, *12*, 263–276.
36. Choi, H.; Nesvizhskii, A. I. *J. Proteom. Res.* **2008**, *7*, 254–265.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch012

37. Mohler, R. E.; Tu, B. P.; Dombek, K. M.; Hoggard, J. C.; Young, E. T.; Synovec, R. E. *J. Chromatogr. A* **2008**, *1186*, 401–411.
38. Pierce, K. M.; Hoggard, J. C.; Hope, J. L.; Rainey, P. M.; Hoofnagle, A. N.; Jack, R. M.; Wright, B. W.; Synovec, R. E. *Anal. Chem.* **2006**, *78*, 5068–5075.
39. Huang, X.; Shao, L.; Gong, Y.; Mao, Y.; Liu, C.; Qu, H.; Cheng, Y. *J. Chromatogr. B* **2008**, *870*, 178–185.
40. Baxter, I. R.; Vitek, O.; Lahner, B.; Muthukumar, B.; Borghi, M.; Morrissey, J.; Guerinot, M. L.; Salt, D. E. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 12081–12086.
41. Yang, L.; Bennett, R.; Strum, J.; Ellsworth, B. B.; Hamilton, D.; Tomlinson, M.; Wolf, R. W.; Housley, M.; Roberts, B. A.; Welsh, J.; Jackson, B. J.; Wood, S. G.; Banka, C. L.; Thulin, C. D.; Linford, M. R. *Anal. Bioanal. Chem.* **2009**, *393*, 643–654.
42. Hummel, J.; Strehmel, N.; Selbig, J.; Walther, D.; Kopka, J. *Metabolomics* **2010**, *6*, 322–333.
43. Li, Y. F.; Radivojac, P. *BMC Bioinf.* **2012**, *13* (Suppl. 16), S4 (1−19).
44. Jin, S.; Daly, D. S.; Springer, D. L.; Miller, J. H. *J. Proteom. Res.* **2008**, *7*, 164–169.
45. Shen, C.; Wang, Z.; Shanker, G.; Zhang, X.; Li, L. *Bioinformatics* **2008**, *24*, 202–208.
46. Zhang, Y.; Wen, Z.; Washburn, M. P.; Florens, L. *Anal. Chem.* **2010**, *82*, 2272–2281.
47. Serang, O.; MacCoss, M. J.; Noble, W. S. *J. Proteom. Res.* **2010**, *9*, 5346–5357.
48. Godovac-Zimmermann, J.; Brown, L. R. *Mass Spectrom. Rev.* **2001**, *20*, 1–57.
49. Joyce, A. R.; Palsson, B. Ø. *Nat. Rev. Mol. Cell Biol.* **2006**, *7*, 198–210.
50. Smilde, A. K.; van der Werf, M. J.; Bijlsma, S.; van der Werff-van der Vat, B. J. C.; Jellema, R. H. *Anal. Chem.* **2005**, *77*, 6729–6736.
51. Williams, R.; Lenz, E. M.; Wilson, A. J.; Granger, J.; Wilson, I. D.; Major, H.; Stumpf, C.; Plumb, R. *Mol. BioSyst.* **2006**, *2*, 174–183.
52. Bevilacqua, V.; Mastronardi, G.; Menolascina, F. In *Computational Intelligence and Bioinformatics*; Huang, D.-S., Li, K., Irwin, G. W., Eds.; Lecture Notes in Computer Science; Springer: Berlin, Heidelberg, 2006; Vol. 4115, pp 475−484.
53. Lee, G.; Rodriguez, C.; Madabhushi, A. *IEEE/ACM Trans. Comp. Biol. Bioinf.* **2008**, *5*, 368–384.
54. Czekaj, T.; Wu, W.; Walczak, B. *Talanta* **2008**, *76*, 564–574.
55. Li, W.; Zhang, S.; Liu, C.-C.; Zhou, X. J. *Bioinformatics* **2012**, *28*, 2458–2466.
56. Zhang, S.; Liu, C.-C.; Li, W.; Shen, H.; Laird, P. W.; Zhou, X. J. *Nucl. Acids Res.* **2012**, *40*, 9379–9391.
57. Hendrickx, D. M.; Hoefsloot, H. C. J.; Hendriks, M. M. W. B.; Canelas, A. B.; Smilde, A. K. *Anal. Chim. Acta* **2012**, *719*, 8–15.
58. Jansen, J. J.; Szymańska, E.; Hoefsloot, H. C. J.; Jacobs, D. M.; Strassburg, K.; Smilde, A. K. *Metabolomics* **2012**, *8*, 422–432.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch012

59. Jansen, J. J.; Szymańska, E.; Hoefsloot, H. C. J.; Smilde, A. K. *Metabolomics* **2012**, *8*, S94–S104.
60. Shen, M.; Broeckling, C. D.; Chu, E. Y.; Ziegler, G.; Baxter, I. R.; Prenni, J. E.; Hoekenga, O. A. *PLoS ONE* **2013**, *8*, e57667 (1−8).
61. Doerfler, H.; Lyon, D.; Nägele, T.; Sun, X.; Fragner, L.; Hadacek, F.; Egelhofer, V.; Weckwerth, W. *Metabolomics* **2013**, *9*, 564–574.
62. Xu, Y.; Correa, E.; Goodacre, R. *Anal. Bioanal. Chem.* **2013**, *405*, 5063–5074.
63. Sharon, M. *J. Am. Soc. Mass Spectrom.* **2010**, *21*, 487–500.
64. Vasilescu, J.; Figeys, D. *Curr. Opin. Biotechnol.* **2006**, *17*, 394–399.
65. Toews, J.; Rogalski, J. C.; Clark, T. J.; Kast, J. *Anal. Chim. Acta* **2008**, *618*, 168–183.
66. Wilson, I. A.; Ward, A. B.; Sali, A. *Science* **2013**, *339*, 913–915.
67. Singh, P.; Panchaud, A.; Goodlett, D. R. *Anal. Chem.* **2010**, *82*, 2636–2642.
68. Cho, S.; Park, S. G.; Lee, D. H.; Park, B. C. *J. Biochem. Mol. Biol.* **2004**, *37*, 45–52.
69. Goyal, S.; Grover, S.; Dhanjal, J. K.; Tyagi, C.; Goyal, M.; Grover, A. *J. Mol. Graph. Model.* **2014**, *51*, 64–72.
70. Massanet, R.; Caminal, P.; Perera, A. *Proceedings of the 8th IEEE International Conference on Bioinformatics and BioEngineering, Athens, Greece*; IEEE: 2008; pp 1−5.
71. Liu, W.; Srivastava, A.; Zhang, J. *PLoS Comput. Biol.* **2011**, *7*, e1001075 (1−10).
72. Hernández-Cisneros, R. R.; Terashima-Marín, H. *Proceedings of the 2006 IEEE Congress on Evolutionary Computation, Vancouber, BC, Canada*; IEEE: 2006; pp 2459−2466.
73. Broersen, A.; van Liere, R.; Altelaar, A. F. M.; Heeren, R. M. A.; McDonnell, L. A. *J. Am. Soc. Mass Spectrom.* **2008**, *19*, 823–832.
74. Mavandadi, S.; Dimitrov, S.; Feng, S.; Yu, F.; Yu, R.; Sikora, U.; Ozcan, A. *Lab Chip* **2012**, *12*, 4102–4106.
75. McConico, M. B.; Horton, R. B.; Witt, K. K.; Vogt, F. *J. Chemom.* **2012**, *26*, 585–597.
76. Wilmes, P.; Bond, P. *Trends Microbiol.* **2006**, *14*, 92–97.
77. Ward, J. P.; King, J. R.; Koerber, A. J.; Croft, J. M.; Sockett, R. E.; Williams, P. *J. Math. Biol.* **2003**, *47*, 23–55.
78. Leary, D. H.; Hervey, W. J., IV; Li, R. W.; Deschamps, J. R.; Kusterbeck, A. W.; Vora, G. J. *Anal. Chem.* **2012**, *84*, 4006–4013.
79. Schubert, J.; Simutis, R.; Dors, M.; Havlik, I.; Lübbert, A. *J. Biotechnol.* **1994**, *35*, 51–68.
80. Gómez, V.; Callao, M. P. *Anal. Chim. Acta* **2008**, *627*, 169–183.
81. Ho, C.; Kelly, M. B.; Stubbs, C. D. *Biochim. Biophys. Acta* **1994**, *1193*, 307–315.
82. Vishwanath, K.; Mycek, M.-A. *Opt. Lett.* **2004**, *29*, 1512–1514.
83. Wolf, G.; Almeida, J. S.; Pinheiro, C.; Correia, V.; Rodrigues, C.; Reis, M. A. M.; Crespo, J. G. *Biotechnol. Bioeng.* **2001**, *72*, 297–306.
84. Roussel, S.; Bellon-Maurel, V.; Roger, J.-M.; Grenier, P. *Chemom. Intell. Lab. Syst.* **2003**, *65*, 209–219.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch012

85. Franz, C.; Jürgen, K.; Florentina, P.; Karl, B. *J. Biotechnol.* **2005**, *120*, 183–196.
86. Su, Z.; Tong, W.; Shi, L.; Shao, X.; Cai, W. *Anal. Lett.* **2006**, *39*, 2073–2083.
87. Surribas, A.; Amigo, J. M.; Coello, J.; Montesinos, J. L.; Valero, F.; Maspoch, S. *Anal. Bioanal. Chem.* **2006**, *385*, 1281–1288.
88. Xiaobo, Z.; Jiewen, Z.; Xingyi, H.; Yanxiao, L. *Chemom. Intell. Lab. Syst.* **2007**, *43*–51.
89. Vial, J.; Noçairi, H.; Sassiat, P.; Mallipatu, S.; Cognon, G.; Thiébaut, D.; Teillet, B.; Rutledge, D. N. *J. Chromatogr. A* **2009**, 2866–2872.
90. Tan, C.; Wang, J.; Wu, T.; Qin, X.; Li, M. *Vib. Spectrosc.* **2010**, *54*, 35–41.
91. Ödman, P.; Johansen, C. L.; Olsson, L.; Gernaey, K. V.; Lantz, A. E. *Appl. Microbiol. Biotechnol.* **2010**, *86*, 1745–1759.
92. McDonnell, L. A.; Heeren, R. M. A. *Mass Spectrom. Rev.* **2007**, *26*, 606–643.
93. Westerhuis, J. A.; Hoefsloot, H. C. J.; Smit, S.; Vis, D. J.; Smilde, A. K.; van Velzen, E. J. J.; van Duijnhoven, J. P. M.; van Dorsten, F. A. *Metabolomics* **2008**, *4*, 81–89.
94. Kokar, M. M.; Ng, G. W. *Info. Fusion* **2009**, *10*, 2–5.
95. Abdo, A.; Leclère, V.; Jacques, P.; Salim, N.; Pupin, M. *J. Chem. Inf. Model.* **2014**, *54*, 30–36.
96. Del Valle, Y.; Venayagamoorthy, G. K.; Mohagheghi, S.; Hernandez, J.-C.; Harley, R. G. *IEEE Trans. Evolut. Computat.* **2008**, *12*, 171–195.
97. Kim, J.; Choi, J. N.; Kim, P.; Sok, D.-E.; Nam, S.-W.; Lee, C. H. *J. Microbiol. Biotechnol.* **2009**, *19*, 51–54.
98. Krug, D.; Zurek, G.; Revermann, O.; Vos, M.; Velicer, G. J.; Müller, R. *Appl. Environ. Microbiol.* **2008**, *74*, 3058–3068.
99. Watrous, J.; Roach, P.; Alexandrov, T.; Heath, B. S.; Yang, J. Y.; Kersten, R. D.; van der Voort, M.; Pogliano, K.; Gross, H.; Raaijmakers, J. M.; Moore, B. S.; Laskin, J.; Bandeira, N.; Dorrestein, P. C. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, E1743–E1752.
100. Noble, W. S.; MacCoss, M. J. *PLoS Comp. Biol.* **2012**, *8*, e1002296 (1−6).
101. Bowen, B. P.; Northen, T. R. *J. Am. Soc. Mass Spectrom.* **2010**, *21*, 1471–1476.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch012

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch013

Chapter 13

Fusing Spectral Data To Improve Protein Secondary Structure Analysis: Data Fusion

Olayinka O. Oshokoya and Renee D. JiJi*

Department of Chemistry, University of Missouri-Columbia, 601 S. College Avenue, Columbia, Missouri 65211, United States
***E-mail: jijir@missouri.edu**

The determination of protein secondary structure has become an area of great significance as this knowledge is important for understanding relationships between protein structure and, more importantly, how the changes in structure affect function. Previous studies suggest that a complementary use of spectroscopic data from optical methods such as circular dichroism (CD), infrared (IR) and ultraviolet resonance Raman (UVRR) coupled with multivariate calibration techniques like multivariate curve resolution-alternating least squares (MCR-ALS) is the preferred route for real-time and accurate evaluation of protein secondary structure. This study presents a new strategy for the improvement of secondary structure determination of proteins by fusing CD and UVRR spectroscopic data. Also, a new method for determining the structural composition of each protein is employed, which is based on the relative abundance of the (φ,ψ) dihedral angles of the peptide backbone as they correspond to each type of secondary structure. Comparison of the predicted protein secondary structures from MCR-ALS analysis of CD, UVRR and fused data with definitions obtained from dihedral angles of the peptide backbone, yields lower overall root mean squared errors of calibration for helical, β-sheet, poly-proline II type and total unfolded secondary structures with fused data.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch013

Introduction

Protein secondary structure quantification has become an area of intense biochemical and biophysical research due to the effects of secondary structure on tertiary and quaternary protein structure. There are four levels of protein structure and a change at any level can result in changes in protein function. The primary structure of a protein is the amino acid sequence while the secondary structure refers to the structural motifs within the protein that are defined by the phi (φ) and psi (ψ) dihedral angles of the amide backbone (Figure 1).

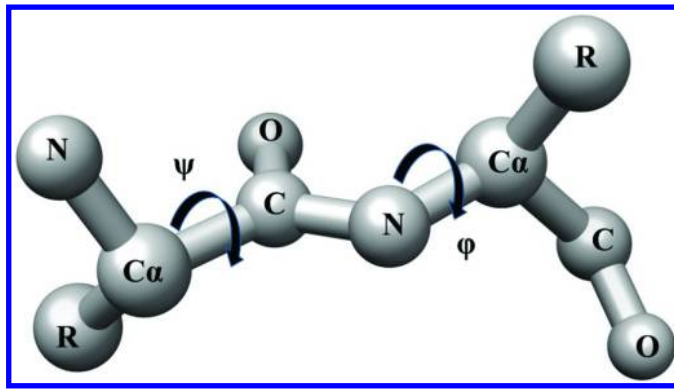

Figure 1. Peptide backbone showing phi (φ) and psi (ψ) dihedral angles.

The three dimensional arrangement of these secondary structure motifs is the tertiary structure and finally the arrangement of protein tertiary subunits to each other in larger complexes that function as a single unit is the quaternary structure (*1–6*). Since secondary structure plays a significant role in protein function and select diseases, it is therefore of substantial interest to rapidly and accurately quantify protein secondary structure especially in an environment that mimics physiological conditions. Traditional methods of protein secondary structure quantification such as x-ray crystallography (XRC) (*7*), nuclear magnetic resonance (NMR) (*8*) and circular dichroism (CD) (*9–11*) are now complimented by a host of vibrational methods, in particular, ultraviolet resonance Raman (UVRR) spectroscopy which has proven useful due to its structural sensitivity to the amide backbone. CD is the current standard in secondary structure analysis of proteins and UVRR is an up-and-coming technique (*12*).

Previous studies show that multivariate analysis of CD and UVRR data results in relatively accurate prediction of helical (α- (-57°, -47°), + 3_{10}- (-49°, -26°)) and β-sheet (anti parallel (-139°, 135°), + parallel (-119°, 113°)) secondary structures in proteins, respectively, but relatively poor prediction of the other secondary structures (*13*). This is because α-helical secondary structure has the largest relative signal intensity in CD spectra of proteins whilst β-sheet secondary structure has the highest relative signal intensity in UVRR spectra of proteins. Combining the predicted amounts of helical and β-sheet contents from CD and

UVRR enables a more accurate estimation of the disordered content, thus, more accurate predictions of secondary structure content (*13*).

In this study, we describe a new addition to the toolbox for protein secondary structure determination by taking advantage of the partial selectivity's of both CD and UVRR spectroscopies. The best estimates by MCR-ALS analysis are achieved with fused data from both spectroscopic techniques. Data fusion refers to methods that combine multiple data types into a single data array, with the expectation that the resulting fused data will be more informative than the individual input sources (*14–17*). Generally, performing data fusion offers advantages which include improved detection, confidence and reliability (*18–24*). Data fusion can be executed in one of three fashions; data level fusion, where the raw data generated by multiple sources are combined directly, or after appropriate normalization has been carried out so that the data are commensurate; feature-level fusion, where feature extraction methods are used to generate representations of the raw data which are then combined; and decision level fusion which involves combining decisions that have been arrived at independently by the available sources (*17*). In this study, we utilize data level fusion, fusing the raw or preprocessed UVRR and CD data before any other analysis is carried out.

We have compared different preprocessing methods for the fused data to determine which method improves protein secondary structure prediction. We have also defined the structural classifications of secondary structure based on the relative distribution of (φ,ψ) dihedral angles of the amide backbone in each protein. We show that by redefining secondary structure based on dihedral angles and application of data fusion to CD and UVRR spectroscopic data, we can improve the determination of not only the helical or β-sheet contents of proteins but also other secondary structures most notably the poly-proline II (PPII) type structure.

PPII-type structure was first identified by Tiffany and Krimm (*25–27*) in poly-L-lysine and poly-L-glutamic acid and has since been shown to be the predominant structure in unfolded or disordered protein regions. PPII-type structure has (-79°, 150°) dihedral angles and is stabilized by water hydrogen bonding with the peptide backbone. Unfortunately, this structure is not defined in the protein data bank and thus difficult to quantify and distinguish from other unfolded or less prevalent structures. Less prevalent structures include left handed α-helices (57°, 47°) and turns, which typically make up less than 5% of the protein's secondary structure. Turns are more complicated as the (φ, ψ) dihedral angles are not repetitive and differ depending on the type of turn. Thus, for quantitative purposes, it makes more sense to define each protein's structural composition based on the abundance of (φ, ψ) dihedral angles.

Materials and Methods

Sample Preparation

Nine globular proteins with varying secondary structure content (Figure 2), amino acids L-phenylalanine and L-tyrosine were obtained from Sigma Aldrich (St Louis, MO) and used without further purification. The proteins and amino

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch013

acids were dissolved in 10 mM phosphate buffer solution (pH 7.2). Protein and aromatic amino acid concentrations were determined by UV-Visible absorption using a Hewlett Packard 8453 spectrometer (Palo Alto, CA), and were 0.5 mg ml^{-1} for protein solutions and 200 μM for amino acid solutions for UVRR analysis and 0.2 mg $ml^{-1}$for CD measurements. Protein coordinate files for the nine proteins were downloaded from the protein data bank, PDB (www.rscb.org) (*28*) and a dihedral angle calculator readily available online (http://cib.cf.ocha.ac.jp/bitool/DIHED2/) (*29*) was used to determine the relative abundance of the (φ, ψ) dihedral angles in each protein for secondary structure content distribution as displayed in Figure 2. The selected proteins are readily soluble in aqueous solution, have a well-distributed combination of the major secondary structures and are relatively inexpensive, making them an ideal set of calibration proteins.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch013

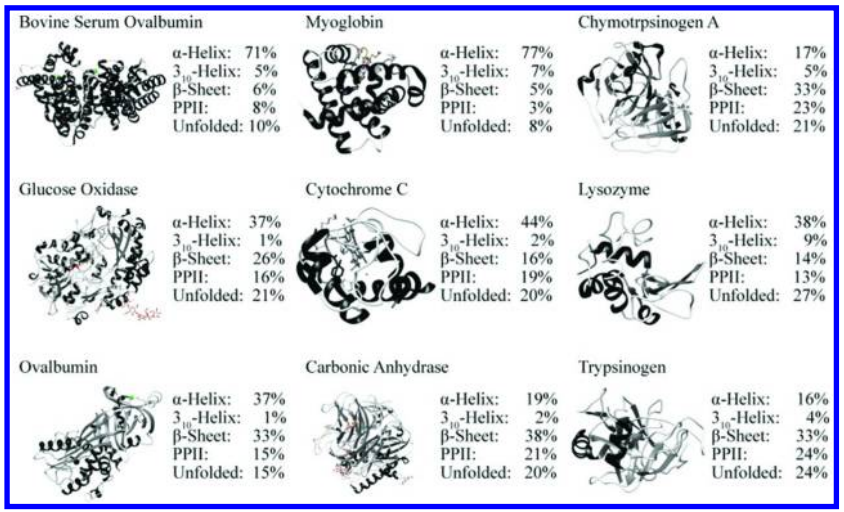

Figure 2. Secondary structure content (%) of proteins used calculated from (φ, ψ) dihedral angles as found on the Research Collaboratory for Structural Bioinformatics (RSCB) Protein Data Bank.

UVRR Spectra Acquisition

The UVRR instrument used to collect protein spectra has been previously described (*30*). Briefly, the fourth harmonic of a tunable Ti:Sapphire laser (Coherent Inc., Santa Clara, CA) was employed to generate an excitation wavelength of 197 nm. The sample was circulated by a Minipuls2 peristaltic pump (Gilson Inc., Middleton, WI) through two nitinol wires (Small Parts Inc., Miramar, FL) to create a thin film under a nitrogen purge to remove ambient oxygen. The temperature of the sample was held at 4°C in a water-jacketed reservoir (Mid Rivers Glassblowing, Saint Charles, MO) using a bath recirculator (Isotemp 3016D, Fisher Scientific, Pittsburgh, PA). Raman scattering was collected in the 135° backscattering geometry and directed into a 1.2 m spectrometer (Horiba Jobin Yvon Inc., Edison, NJ) equipped with a Symphony CCD detector, which

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

was controlled by Synerjy software (Horiba Jobin Yvon Inc., Edison, NJ). Each spectrum was the sum of 3 hours of signal collection. A small aliquot of a 1 M sodium perchlorate solution was added to each sample, for a final concentration of 200 mM, as an internal intensity standard. All spectra were collected in triplicate and calibrated using a standard cyclohexane spectrum (*31*, *32*).

CD Spectra Acquisition

All samples used for UVRR analysis were additionally measured for their corresponding CD spectra. An AVIV 62DS circular dichroism (Aviv Biomedical Inc., Lakewood Township, NJ) spectrometer and a quartz cell with a 1 mm optical path length (Hellman USA, Plainview, NY) were used to collect CD spectra. All spectra were collected between 190 and 250 nm with a resolution of 0.1 nm at room temperature. Every sample was measured five times with a scan speed of 1 nm/5 s and averaged. Each experiment was repeated in triplicate. Corresponding background spectra were collected in the same manner and subtracted from sample spectra.

Data Processing

All data analyses were carried out in MATLAB (version 7.11, Mathworks, Natick MA). Cosmic rays in the UVRR spectra were removed using an in-house program (*33*), and the spectra were base-lined using the MATLAB curve-fitting toolbox. Contributions to spectra from aromatic side chains were subtracted using the phenyalanine band at 1003 cm^{-1} (F12) and tyrosine band at 853 cm^{-1} (Y1) as previously described (*13*). Contributions from tryptophan were disregarded due to its negligible intensity in deep-UVRR spectra (λ_{ex} <210 nm). Areas that appeared to be negative in the spectrum after subtraction of aromatic contribution were set to zero and each resulting spectrum truncated to the 1266–1759 cm^{-1} spectral range so that only the amide regions were used for modeling. For CD data, the mean residue ellipticiity (Θ_{MRE}) was calculated as previously described (*30*).

A MCR-ALS algorithm was used based on that outlined by Bro and Sidiropoulous (*34*). MCR-ALS was selected because on average it performed better in previous studies (*13*) compared to classical least squares and partial least squares for spectral resolution and secondary structure prediction.

For both UVRR and CD, the triplicate spectra were compiled to obtain 27 individual spectra (Figure 3). The UVRR and CD data were then fused according to the model in Figure 4 to give a single data matrix. To evaluate the potential predictive ability of the MCR-ALS models, the root mean squared error of calibration (RMSEC) was used (Equation 1).

$$RMSEC = \left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right]^{1/2} \quad (1)$$

In Equation 1, n is the number of samples, y_i is the abundance of each secondary structure element obtained from the (φ,ψ) dihedral angles as displayed in Figure 2

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch013

and $\hat{y}_i$ is the estimated value obtained from least squares regression of the resolved composition profiles from the MCR-ALS algorithm.

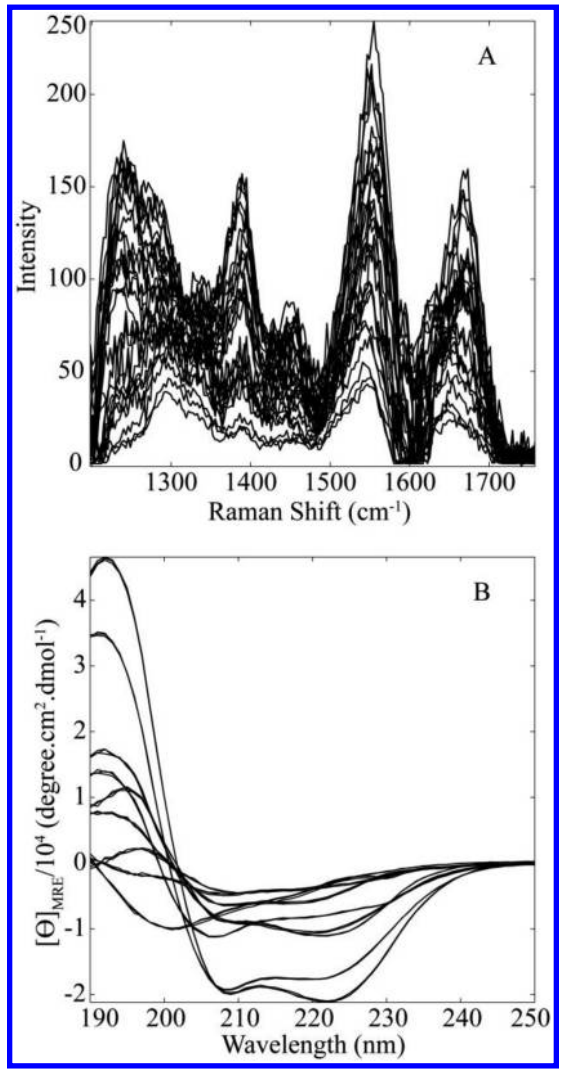

Figure 3. UVRR (A) and CD (B) spectra used for multivariate analysis.

Results and Discussion

Protein Secondary Structure and UVRR and CD Spectra

Ideally, proteins with similar secondary structure contents should have similar CD and UVRR spectra. However, while protein UVRR and CD spectra are highly reproducible, proteins with similar secondary structural content can have very different spectra. For instance, while carbonic anhydrase and chymotrypsinogen

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch013

A have similar secondary structure distributions with high β-sheet and relatively low helical contents (Figure 4), there is a clear difference in their CD spectra but their UVRR spectra are overlapped. Bovine serum albumin and myoglobin also have similar secondary structure distributions but with high helical contents and no β-sheet structure; the CD spectra for both proteins are quite similar but in this case their UVRR spectra, while similar in shape are clearly differing in overall intensity.

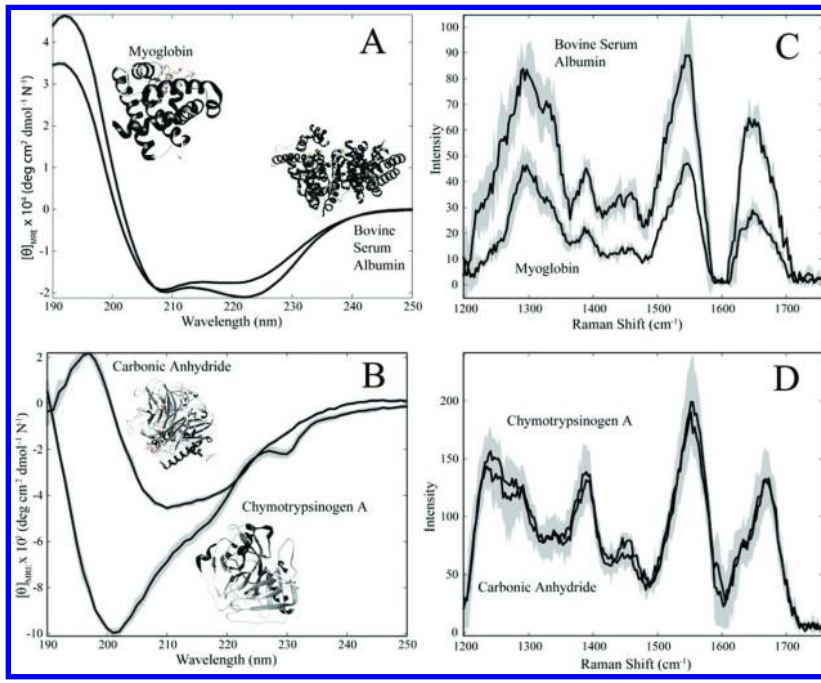

Figure 4. CD (A and B) and UVRR (C and D) spectra of proteins with similar secondary structure compositions. The shaded gray area about the lines represent the standard deviation of three measurements at each variable.

It can be concluded that greater differences in the measured spectra of proteins with similar structural compositions will be observed when the dominant secondary structure type has a low relative signal intensity as compared to the other types of secondary structure. Therefore, poorer prediction of these structures is almost certain if only one technique is employed. In order to take advantage of the predictive capabilities of each technique (CD and UVRR), a data fusion approach was employed.

Effect of Preprocessing on Estimation of Composition Profiles

The UVRR and CD data were fused according to the model in Figure 5 to yield a single data matrix. MCR-ALS was employed to resolve the underlying

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch013

compositional and spectral profiles prior to preprocessing, and after normalization, auto scaling and variance scaling (Figure 6).

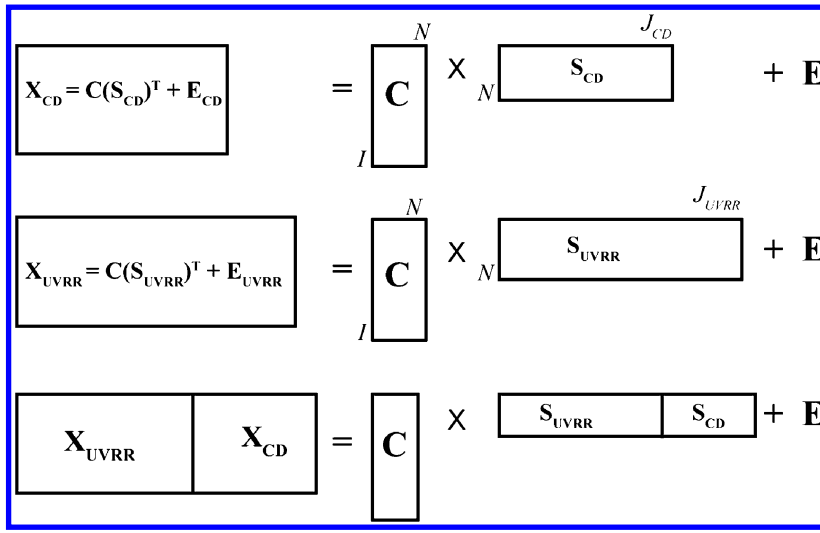

Figure 5. Data fusion model for multivariate analysis for protein secondary structure determination.

A 4-component model was employed because a 5-component model resulted in poorer predictions of the three most prominent structures; helical, β-sheet and PP-II and did not enable resolution of α- and 3_{10}-helical structure or parallel and antiparallel structures. The composition profiles from MCR-ALS analysis assigned to (helical (α- + 3_{10}-helices), β-sheet/strand, PP-II and unfolded (everything else)) were regressed using the secondary structure compositions obtained from the relative abundance of the (φ,ψ) dihedral angles of the peptide backbone for each protein (Figure 2). The resultant regression model was then used to repredict secondary structure of the test protein samples.

Prediction accuracies of about 5% can be achieved for either helical content using CD or β-sheet content using UVRR (Figure 7). A comparison of the RMSEC values versus each preprocessing method is summarized in Table 1. When no preprocessing was employed, the results were similar to the use of CD data alone (Table 1 and Figure 7). RMSEC of β-sheet is high when no preprocessing is employed because the greater intensity of the CD spectra has a greater influence on the model (Figure 6), and the UVRR information is lost. Auto scaling of the fused data resulted in significantly higher RMSEC's for all secondary structure types. Normalization to unit variance did not improve prediction of β-sheet structure (RMSEC = 40%) and appeared to worsen prediction of helical structure, essentially doubling the RMSEC. Ultimately variance scaling improved prediction of β-sheet structure (RMSEC = 5.4%) without much penalty to the prediction of the other structures (Table 1 and Figure 7).

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch013

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch013

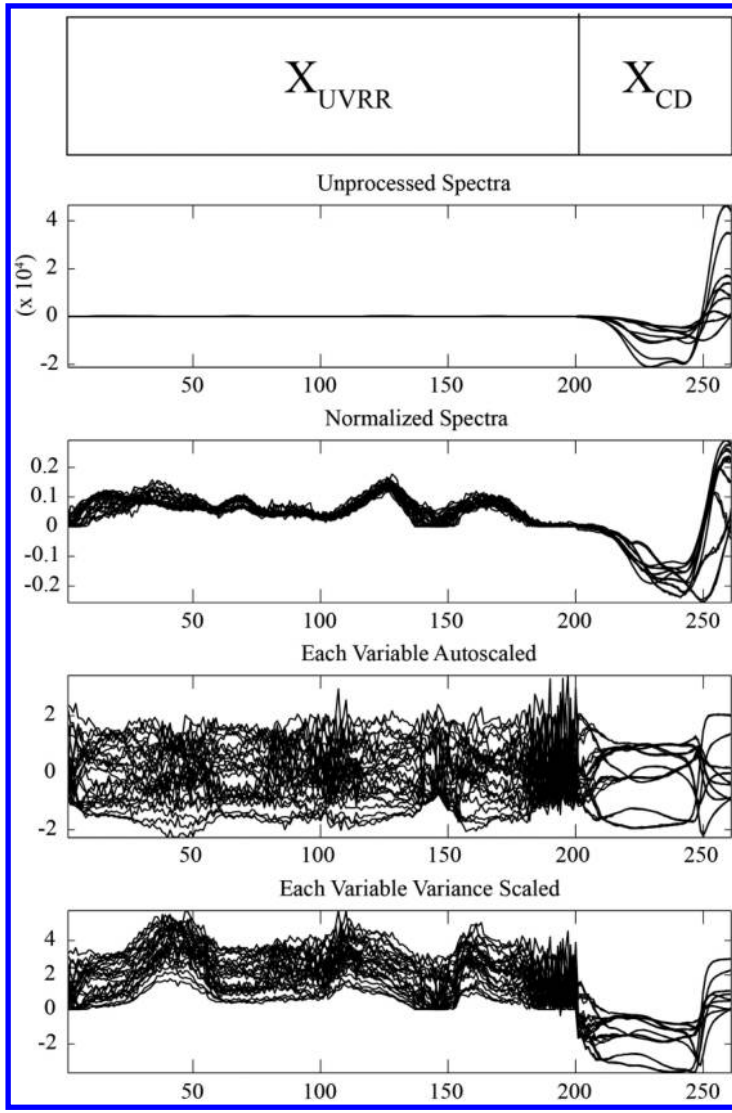

Figure 6. Fused CD and UVRR spectra after application of each preprocessing method.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch013

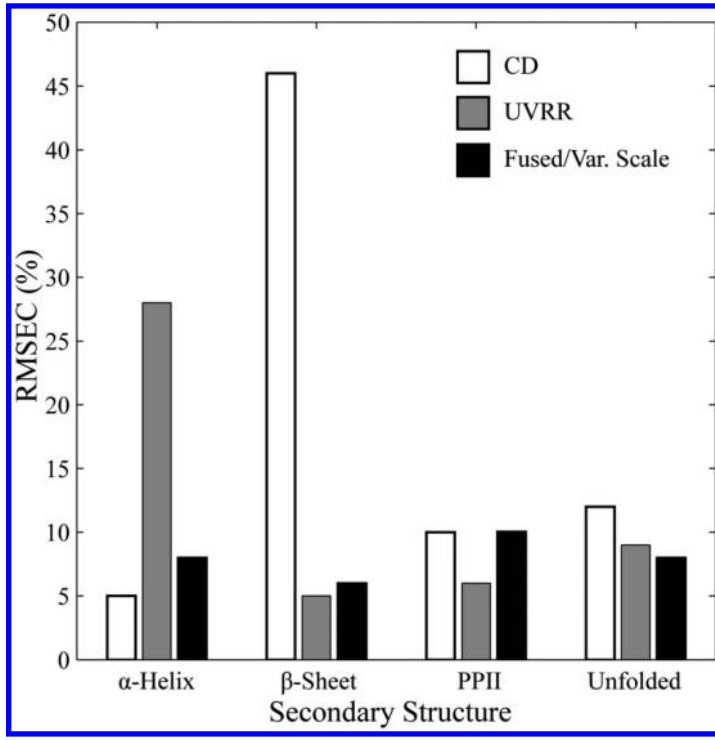

Figure 7. Root mean square error of calibration (RMSEC) for prediction of protein secondary structure using CD, UVRR and fused CD-UVRR spectroscopic data.

Table 1. Root mean square error of calibration (RMSEC) of MCR-ALS model employing different preprocessing methods

Pre-processing method	*Helix*	*Sheet*	*PPII*	*Unfolded*
Unprocessed	5.7%	54.2%	8.7%	11.0%
Normalized	12.9%	40.2%	3.7%	8.0%
Auto-scaled	296.4%	57.0%	82.3%	27.1%
Variance-scaled	6.6%	5.4%	10.7%	8.6%

Conclusions

In this work, we have developed a new approach to protein secondary structure determination by applying multivariate analysis to fused spectroscopic data. The advantage to this approach where CD and UVRR data are fused over individual analysis of both spectroscopic methods is that we can exploit the selective predictive capabilities of each technique (helical structure for CD and β-sheet structure for UVRR) and further improve predictions of other secondary

structures including the PPII-type structure. We have also demonstrated that the most appropriate preprocessing method prior to multivariate analysis is the variance scaling method.

While helical structure prediction is improved using multivariate analysis of the fused data, the limitation of separating α- and 3_{10}-helical structures still looms. This is because both structures have very similar spectra both in CD and UVRR hence making them statistically indistinguishable. Also, less prevalent structures like turns and α-L, which occur in very small quantities, are not yet quantifiable as their CD and UVRR spectra are not distinct enough. Expansion to include other structurally sensitive techniques such as Raman optical activity or vibrational circular dichroism may increase the number of quantifiable secondary structures.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch013

Acknowledgments

The authors thank Dr. Michael Henzl and Dr. Anmin Tan for help with the CD measurements. The authors also thank the NSF (Grant # CHE-1151533), University of Missouri Research Council, University of Missouri Research Board and University of Missouri Department of Chemistry for funding.

References

1. Blake, C. C.; Geisow, M. J.; Oatley, S. J.; Rerat, B.; Rerat, C. *J. Mol. Biol.* **1978**, *121*, 339–356.
2. Herczenik, E.; Gebbink, M. F. B. G. *FASEB J.* **2008**, *22*, 2115–2133.
3. Moglich, A.; Yang, X.; Ayers, R. A.; Moffat, K. *Annu. Rev. Plant Biol.* **2010**, *61*, 21–47.
4. Prusiner, S. B. *Proc. Natl. Acad. Sci. U.S.A* **1998**, *95*, 13363–13383.
5. Weissmann, C. *Nat. Rev. Microbiol.* **2004**, *2*, 861–871.
6. Voet, D.; Voet, J. G. *Biochemistry*, 3rd ed.; John Wiley & Sons, Inc.: Hoboken, NJ, 2004.
7. Higashiura, A.; Ohta, K.; Masaki, M.; Sato, M.; Inaka, K.; Tanaka, H.; Nakagawa, A. *J. Synchrotron Radiat.* **2013**, *20*, 989–993.
8. Castellani, F.; van Rossum, B.; Diehl, A.; Schubert, M.; Rehbein, K.; Oschkinat, H. *Nature* **2002**, *420*, 98–102.
9. Greenfield, N. J. *Anal. Biochem.* **1996**, *235*, 1–10.
10. Greenfield, N. J. *Nat. Protoc.* **2006**, *1*, 2876–2890.
11. Greenfield, N. J.; Fasman, G. D. *Biochemistry* **1969**, *8*, 4108–4116.
12. Roach, C. A.; Simpson, J. V.; Jiji, R. D. *Analyst* **2012**, *137*, 555–562.
13. Oshokoya, O. O.; Roach, C. A.; Jiji, R. D. *Anal. Methods* **2014**, *6*, 1691–1699.
14. Hall, D. L.; McMullen, S. A. H. *Mathematical Techniques in Multisensor Data Fusion*; Artech House: Norwood, MA, 2004.
15. Liggins, M., II; Hall, D.; Llinas, J. *Handbook of Multisensor Data Fusion: Theory and Practice*; CRC Press: Boca Raton, FL, 2008.
16. Mitchell, H. B. *Multi-Sensor Data Fusion: An Introduction*; Springer: Berlin, 2007.

17. Klein, L. A. *Sensor and Data Fusion Concepts and Applications*, 2nd ed.; SPIE Optical Engineering Press: Bellingham, 1999.
18. Ardeshir Goshtasby, A.; Nikolov, S. *Inf. Fusion* **2007**, *8*, 114–118.
19. Bloch, I. *IEEE Trans. Syst. Man. Cybern., Part A Syst. Humans* **1996**, *26*, 52–67.
20. Corona, I.; Giacinto, G.; Mazzariello, C.; Roli, F.; Sansone, C. *Inf. Fusion* **2009**, *10*, 274–284.
21. Hall, D. L.; Llinas, J. *Proc. IEEE* **1997**, *85*, 6–23.
22. Rogova, G. L.; Nimier, V. *Proceedings of the Seventh International Conference on Information Fusion, FUSION 2004*, Stockholm, 2004; International Society of Information Fusion: 2004.
23. Smith, D.; Singh, S. *IEEE. Trans. Knowl. Data Eng.* **2006**, *18*, 1696–1710.
24. Yao, J.; Raghavan, V. V.; Wu, Z. *Inf. Fusion* **2008**, *9*, 446–449.
25. Tiffany, M. L.; Krimm, S. *Biopolymers* **1968**, *6*, 1379–1382.
26. Tiffany, M. L.; Krimm, S. *Biopolymers* **1968**, *6*, 1767–1770.
27. Tiffany, M. L.; Krimm, S. *Biopolymers* **1969**, *8*, 347–359.
28. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235–242.
29. Liljas, A.; Liljas, L.; Piskur, J.; Lindblom, G.; Nissen P.; Kjeldgaard, M. *Textbook of Structural Biology*; World Scientific: 2009.
30. Wang, M.; Jiji, R. D. *Biophys. Chem.* **2011**, *158*, 96–103.
31. Ferraro, J. R.; Nakamoto, K. *Introductory Raman Spectroscopy*; Academic Press, Inc.: San Diego, CA, 1994.
32. Austin, J. C.; Jordan, T.; Spiro, T. G. *Adv. Spectrosc. (Chichester, U. K.)* **1993**, *20*, 55–127.
33. Simpson, J. V.; Oshokoya, O.; Wagner, N.; Liu, J.; Jiji, R. D. *Analyst* **2011**, *136*, 1239–1247.
34. Bro, R.; Sidiropoulos, N. D. *J. Chemometr.* **1998**, *12*, 223–247.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch013

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch014

Chapter 14

Chemometric Modeling of Environmental Impacts on the Chemical Composition and Growth Dynamics of Microalgae Cultures

Frank Vogt*

Department of Chemistry, University of Tennessee, Knoxville, Tennessee 37996, United States
***E-mail: fvogt@utk.edu**

Via photosynthesis, ubiquitous marine microalgae cells sequester large quantities of inorganic compounds (nutrients) into biomass. Such a large-scale compound transformation contributes for instance to counter-balancing anthropogenic releases of the greenhouse gas CO_2 and hence has considerable environmental relevance. Since phytoplankton is a chemically active system which adapts to its ambient conditions, the latter determine the cells' compound transformation performance and the biomass production. For assessing phytoplankton's ecological impacts, interactions between cells and their chemical and biological ambient conditions have been studied by means of FTIR spectroscopy, imaging, and chemometric modeling. Of particular interest is to investigate how the nutrient availability and competing species' presence determine the chemical signature of the cells as well as the quantities of resulting biomass. In order to gain a comprehensive understanding of chemical interactions between cells and their environments, innovations in chemometrics have been required. It is demonstrated that chemometric 'hard-modeling' is a viable route to derive interpretable models.

Introduction

With an increase of industrialization, the production of anthropogenic CO_2 is rising (*1*) and the fate of this high-impact greenhouse gas has become a serious

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

concern (*2*, *3*). Research regarding CO_2 sequestering is being directed towards microorganisms because it has been estimated that about half of the global primary carbon production is due to algal photosynthesis (*4–10*). Gaining a more detailed understanding of these processes is especially crucial as there are first indications that the carbon storage capacity of the oceans has started to diminish (*11*). On the other hand, $CO_{2(aq)}$ produces carbonic acid which dissociates into bicarbonate, a crucial algae nutrient, and H^+ which lowers the ocean's pH (*12*, *13*). While microalgae-based CO_2 assimilation is beneficial, lowered pH levels have detrimental consequences on calcifying organisms such as corals (*12*, *13*). Furthermore, phytoplankton's sequestration of nitrogen compounds which originate from over-fertilization in agricultural areas can cause harmful algae blooms (*14*, *15*). Thus, microalgal transformation of inorganic nutrients (C, N, P, Fe, S (*16–20*)) into bioorganic materials has many ecological and economical consequences.

Despite phytoplankton being an important player in large scale ecosystems, a critical gap in knowledge exists due to the chemically complex relationship between inorganic nutrients, microorganisms, and the consequences of these compound transformations. *It is anticipated that through chemometric modeling more detailed knowledge about microalgae's counterbalancing of anthropogenic CO_2, algae-mediated pH modifications of the ocean, and origins of harmful algal blooming can be gained.* These novel modeling methodologies will also gain new insights in chemical and biological shifts of ecosystems and will open new research directions in environmental chemistry, ecology, and marine biology.

Microalgae cells interact with their growing environment through nutrient uptake and competition with each other for these nutrients. This study reports that the chemical and biological parameters in an ecosystem determine algae's chemical composition and the quantity of phytoplankton. In return, phytoplankton, being at the bottom of the foodweb, determine the biodiversity of higher organisms based on their nutritional value and availability. Therefore, microalgae and their interactions with marine ecosystems are linking environmental chemistry and ecology. In this context, the following four aspects are relevant and have been studied in this project; for these investigations, chemometric innovations have been developed focusing on hard-modeling to ensure model interpretability:

- How does the nutrient availability determine the chemical composition of microalgal biomass? (see ref. (*18*))
- How does the presence of nutrient competitors determine the chemical composition of microalgal biomass? (see ref. (*21*))
- How does nutrient availability determine the growth dynamics and the quantity of produced microalgal biomass? (see ref. (*19*, *20*))
- How does the presence of nutrient competitors determine the growth dynamics and the quantity of microalgal biomass? (see ref. (*22*, *23*))

Experimental Preliminaries

Starting cultures for three sea water species, i.e. *Dunaliella salina, Dunaliella parva*, and *Nannochloropsis oculata*, were obtained from The Culture Collection

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch014

of Algae at the University of Texas, Austin. Cultures were inoculated in "enriched seawater, artificial water" (ESAW) medium which contains the cells' nutrients (*24*, *25*). Such cultures are then exposed for multiple days to continuous illumination while being maintained at 20°C (Figure 1, left). Increasing or decreasing nutrient concentrations in the ESAW medium was the chosen pathway to simulate changing ambient conditions. In these studies, cultures were grown under multiple series of inorganic carbon and nitrogen concentrations, the two most important algae nutrients. Inorganic carbon concentrations were adjusted via dissolving different amounts of sodium bicarbonate ($NaHCO_3$) into the growth medium; varying amounts of either sodium nitrate ($NaNO_3$) or ammonium chloride (NH_4Cl) dissolved in ESAW served as nitrogen sources. In order to generate starving, normal, and excess situations, the following carbon concentrations were realized for these studies: 160µM, 1110µM, 2071µM (normal condition), 3630µM, 5180µM, 6720µM, 8260µM; as nitrogen concentrations, 160µM, 350µM, 549 µM (normal condition), 873µM, 1280µM, 1470µM, and 1650µM were chosen. In order to incorporate unavoidable replicate-to-replicate fluctuations into calibration models, five independent replicate cultures were grown for each condition. Since the ambient conditions in these experiments were known, the chemical signatures of the resulting biomass and its quantity could be related to the ambient conditions.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch014

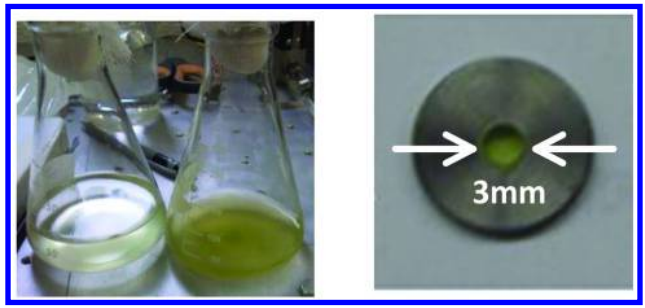

*Figure 1. (**left picture**, left Erlenmeyer) Culture of the sea water microalgae species Dunaliella parva after inoculation when the cell concentration is low; (**left picture**, right Erlenmeyer) the same culture after eight days of growth – (**right picture**) potassium bromide plus dried algal biomass pressed into a 'pellet' from which FTIR transmission spectra were acquired. (F. Vogt, unpublished)*

For harvesting microalgae cells, 1µL of Lugol's solution (Sigma-Aldrich) was added per 1mL of algal suspension in order to fix the cells in their current state. Cells were then extracted from their solutions through centrifugation (4400 rpm) followed by washing the extracts twice with an isoosmotic solution (0.1 M) of ammonium formate (Alfa Aesar) to minimize medium carryover. After gently drying (4-5 days at 60˚C), the algae material was mixed with IR-transparent KBr powder (0.6 weight % of algae) and pressed into a pellet (Figure 1, right picture (*26*)). Due to its sensitivity to a large number of biologically relevant analytes (*27*), FTIR spectroscopy has been found particularly useful for chemical

analyses of microalgal biomass (*18*, *28–32*). FTIR transmission spectra were recorded (3500 – 950 cm^{-1}, 4cm^{-1} resolution, 128 scans). The region 2700 – 1850 cm^{-1} was excluded as it comprised spectroscopic artifacts induced e.g. by fluctuating atmospheric pCO_2 within the spectrometer. As different species have different chemical compositions, FTIR also enabled species classification (*33*, *34*). Measuring time and environment dependent production of biomass has been based on time series of microscope images from which cell counts and cell size distributions had been determined. For confirmation purposes, hemocytometer-assisted cell counts were conducted as well.

Chemometrics Preliminaries

The research topics listed in the bullet list above could have been addressed by conducting a series of experiments presented in the remainder of this manuscript. However, probing a few isolated scenarios does not generate comprehensive insights into the relation between environmental parameters and phytoplankton. This limitation is underscored by the fact that many environmental parameters are coupled and have a nonlinear impact on the biomass (*18*). In order to gain a fundamental understanding of the key parameters and the driving forces behind environment ↔ phytoplankton interactions, chemometric models are required. For this purpose, chemometric 'hard-modeling' is a more promising approach than the ubiquitous 'soft-modeling' which empirically explains structures in data sets (*35*). Hard-modeling is based on theoretical considerations which then lead to model equations that *explicitly* describe the chemical, physical, or biological functionalities of a system. Computing scenario(environment)-specific values for such chemically interpretable parameters (e.g. via least-squares) then deduces information regarding a system's chemical, physical, or biological state. However, hard-modeling faces more challenges than soft-modeling: (i) Many relevant systems feature a considerable level of complexity which must be properly incorporated into hard-models. (ii) Often, the derived model equation requires applying nonlinear least-squares which imposes practical challenges (*36*). On the other hand, even when a nonlinear hard-model equation has to be approximated as a (multivariate) Taylor expansion, one has to keep in mind that the polynomial parameters still carry explicit chemical or physical information (*37*). In addition to gaining new, *fundamental* knowledge about a chemical/physical/biological system, hard-modeling can directly take advantage of additional chemical information by expressing such insights as regression constraints (*38*). Typical examples of extra information are non-negativity of concentrations, concentration ratios, realistic concentration ranges, sum parameters, etc. which are readily implementable in hard-models whereas their translation into latent variables and/or scores is less straightforward.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch014

Impacts of Nutrient Availability on the Chemical Composition of Microalgae

It has been observed that microalgae are actively adapting to their ambience, namely their nutrient situation which in turn causes their spectroscopic signature to change. From an ecological perspective, investigating this will provide knowledge regarding the cells' nutrient utilization and thus their compound transformation capabilities. From an analytical chemistry perspective, this is of interest because a reproducible relation between environment and microalgae's chemical signature enables innovations in environmental monitoring based on microalgae acting as *in-situ* sensors.

Ref. (*18*) has focused on this idea and has demonstrated that the cells' chemical responses to environmental shifts are reproducible, albeit governed by nonlinear responses to multiple, cross-linked environmental parameters. Therefore, conventional, linear multivariate regression (*40*) (MLR) is not applicable for quantitative analyses of such systems. More robust techniques such as Principal Component Regression (*40*, *41*) (PCR) and Partial Least-Squares (*42*, *43*) (PLS) also have limitations; they can approximate nonlinear behavior to a certain extent (*44*) but cannot -due to their empirical nature- shed light on which ambient parameters are linked and how. The multivariate Response Surface (*45*, *46*) methodology on the other hand introduces higher-order predictor variables and derives insights in their coupling. However, Response Surfaces are based on Inverse Least Squares (*42*) (ILS) which generally involves a rather large number (W) of predictor variables to explain a small number of response variables (N). While a large number of adjustable model parameters ensures a high robustness towards unknown signal features, ILS requires $K \geq W \geq N$ calibration samples which for large N can render this approach unfeasible. This is particularly true for measurement techniques such as optical spectroscopy which produce collinear data sets (N large).

In order to investigate nonlinear chemical systems with coupled predictors while requiring a reasonable number of calibration samples, the novel data modeling technique 'Predictor Surfaces' has been developed (*18*). Predictor Surfaces have been based on multivariate Taylor expansions up to order P and therefore enable an interpretation and assessment of coupled and higher-order predictor variables (*37*). While Response Surfaces map measured data (e.g. spectra) onto ambient parameters and hence introduce a problematically high number of model parameters, Predictor Surfaces explain measured data (spectra) in terms of a few (Q) predictor variables (ambient parameters). Predictor Surfaces utilize model equations $\mathbf{Y}_{\beta}(\mathbf{x})$ (1) which contain (here) spectroscopic information and predictor variables $\mathbf{x}$ (2) which are (here) concentrations of various nutrients. As will be shown below, coupling of multiple predictors reflect that microalgae cells require a certain *nutrient mix* to thrive.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch014

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

$$Y_{\boldsymbol{\beta}_n}(\mathbf{x}) = \left(\beta_{n,0}\cdot 1\right)_{p=0} + \left(\sum_{q_1=1}^{Q} \beta_{n,q_1}\cdot x_{q_1}\right)_{p=1} + \left(\sum_{q_1=1}^{Q}\sum_{q_2=q_1}^{Q} \beta_{n,q_1\,q_2}\cdot x_{q_1}\cdot x_{q_2}\right)_{p=2} + \left(\sum_{q_1=1}^{Q}\sum_{q_2=q_1}^{Q}\sum_{q_3=q_2}^{Q} \beta_{n,q_1\,q_2\,q_3}\cdot x_{q_1}\cdot x_{q_2}\cdot x_{q_3}\right)_{p=3} + (\cdots)_{p=4} + \cdots = \boldsymbol{\beta}_{\boldsymbol{n}(1\times W)}\cdot \mathbf{x}_{(W\times 1)} \quad (1)$$

with $\mathbf{Y}_{\boldsymbol{\beta}}(\mathbf{x}) = (Y_{\boldsymbol{\beta}_{n=1}}(\mathbf{x}) \quad \cdots \quad Y_{\boldsymbol{\beta}_{n=N}}(\mathbf{x}))^{\mathrm{T}}$ and $\mathbf{x}_{(W\times 1)} =$

$$\left(\underbrace{x_0 = 1}_{p=0} \quad \underbrace{x_1 \quad \cdots \quad x_N}_{p=1} \quad \underbrace{x_1x_1 \quad x_1x_2 \quad x_1x_3 \quad \cdots \quad x_Nx_N}_{p=2} \quad \underbrace{x_1x_1x_1 \quad x_1x_1x_2 \quad \cdots \quad x_Nx_Nx_N}_{p=3} \quad \cdots\right)^{\mathrm{T}} \quad (2)$$

Thus, a measured spectrum $\mathbf{y}_{(N\times 1)} = \mathbf{Y}_{\boldsymbol{\beta}}(\mathbf{x}) + \boldsymbol{\delta}_{(N\times 1)}$ covering N wavenumber positions equals the model $\boldsymbol{\beta}_{(N\times W)}\cdot\mathbf{x}_{(W\times 1)}$ plus a vector $\boldsymbol{\delta}$ containing measurement errors (or more accurately: any signal features not explained by the model). For determining which of the numerous combinations of x and which P are relevant, Analysis of Variance (*47*, *48*) (ANOVA) has been applied. The 'extra-sum-of-squares principle' (*39*) has been found to be a particularly straightforward implementation of ANOVA.

Prior to utilizing the model (1) to predict chemical parameters in the cells' growing environment, a calibration has to be performed during which $\boldsymbol{\beta}_{(N\times W)}$ is determined. In the given application, numerous algae cultures were grown under series of nutrient concentrations **x** followd by measuring FTIR spectra **y** of dried algae material. Since any nonlinearity between response and predictor variables has been built into **x** (2), equation (1) is linear in the model parameters $\beta_{n,w}$ and hence the calibration of Predictor Surfaces can be done by means of a multivariate least-squares regression (*39*):

$$\widehat{\boldsymbol{\beta}}_{(N\times W)} = \left[y^{\mathrm{cal}}_{n,k}\right]\cdot\left[x^{\mathrm{cal}}_{w,k}\right]^{\mathrm{T}}\cdot\left(\left[x^{\mathrm{cal}}_{w,k}\right]\cdot\left[x^{\mathrm{cal}}_{w,k}\right]^{\mathrm{T}}\right)^{-1} \quad (3)$$

The only difference between a conventional MLR calibration and the calibration of Predictor Surfaces (3) is that for the latter the Q predictor variables $x^{\mathrm{cal}}_{1,k}, \ldots, x^{\mathrm{cal}}_{Q,k}$ of the k^{th} calibration sample need to be compiled into a vector $\mathbf{x}^{\mathrm{cal}}_{k\ (W\times 1)}$ (2) prior to using it as the k^{th} column in $\left[x^{\mathrm{cal}}_{w,k}\right]$ (3).

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch014

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch014

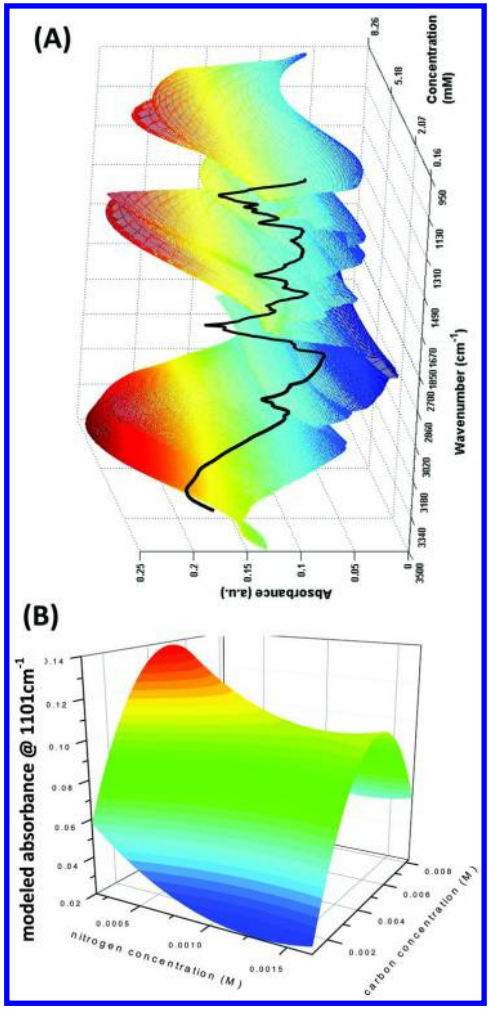

Figure 2. ***(A)*** *This Prediction Surface* $\mathbf{Y}_{\hat{\boldsymbol{\beta}}}(\mathbf{x})$ *(1) with P = 3 describes how the mid-IR signature of the microalgae species Dunaliella parva changes as a function of the bicarbonate concentration in the growing environment. The black spectrum* ***y****(N×1) was obtained from a Dunaliella parva sample grown under an unknown bicarbonate concentration; the latter has been determined by fitting the spectrum onto the Predictor Surface. (Derived work with permission from Elsevier).* ***(B)*** *this 'slice' of a Predictor Hypersurface (Q = 2) describes how the biomass' absorbance at 1101cm-1 changes with shifts in the concentrations of two nutrients, i.e.* $x_1 = c_{\mathrm{HCO}_3^-}$ *and* $x_2 = c_{\mathrm{NO}_3^-}$

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch014

Figure 2(A) depicts a Predictor Surface $\mathbf{Y}_{\hat{\boldsymbol{\beta}}}(\mathbf{x})$ obtained for $P = 3$ which describes how the bicarbonate concentration $\mathbf{x}_{(Q=1,1)} = c_{\mathrm{HCO_3^-}}$ as a sole predictor variable determines the wavenumber dependent mid-IR absorbances of *Dunaliella parva* biomass. For $Q > 1$, a Predictor Surface becomes a Predictor *Hyper*surface which, in the $Q = 2$ case, can be plotted one wavenumber at a time. As shown in Figure 2(B), the given portion of the Predictor Hypersurface, the '1101cm^{-1} slice', describes a nonlinear relation between two nutrient concentrations and the cells' IR-absorbance at 1101cm^{-1}. Obviously the cells produce (*49*) IR-absorbing material dependent on the combination of $c_{\mathrm{HCO_3^-}}$ and $c_{\mathrm{NO_3^-}}$. Thus, Predictor Hypersurfaces are capable of modeling highly nonlinear chemical systems and enable insights into the biological consequences of the cells' chemical environment.

In order to predict $\mathbf{x}^{\mathrm{unk}}$, an unknown response vector $\mathbf{y}^{\mathrm{unk}}$ (black spectrum in Figure 2(A)) is fitted to the calibration model $\mathbf{Y}_{\hat{\boldsymbol{\beta}}}(\mathbf{x})$ (1). Technically, a solution for $\mathbf{x}^{\mathrm{unk}}$ could be estimated via another linear multivariate least-squares regression (MLR) utilizing (3), i.e.: $\hat{\mathbf{x}}^{\mathrm{unk}} = \left(\hat{\boldsymbol{\beta}}^{\mathbf{T}} \cdot \hat{\boldsymbol{\beta}}\right)^{-1} \cdot \hat{\boldsymbol{\beta}}^{\mathbf{T}} \cdot \mathbf{y}^{\mathrm{unk}}$. However, this should *not* be utilized because it would consider each elements of $\hat{\mathbf{x}}^{\mathrm{unk}}_{(W\times 1)}$ (2) as an independent predictor variable. This is clearly not the case since $\mathbf{x}^{\mathrm{unk}}$ contains W elements which are various products of fewer predictor variables $x_{q=1,\ldots Q<W}$. Consequently, searching a solution in a W-dimensional space leads most likely to a chemically meaningless result $\hat{\mathbf{x}}^{\mathrm{unk}}$. To avoid this, $\hat{\mathbf{x}}^{\mathrm{unk}}$ is derived via nonlinear least-squares regression via minimizing the sum of squared errors:

$$\mathrm{SSE}\left(\hat{x}_1^{\mathrm{unk}}, \ldots, \hat{x}_Q^{\mathrm{unk}}\right) = \|\boldsymbol{\delta}\|_2^2 = \left\|\mathbf{y}^{\mathrm{unk}} - \mathbf{Y}_{\hat{\boldsymbol{\beta}}}\left(\hat{\mathbf{x}}^{\mathrm{unk}}\right)\right\|_2^2 = \sum_{n=1}^{N}\left[y_n^{\mathrm{unk}} - Y_{\hat{\boldsymbol{\beta}}_\mathbf{n}}\left(\hat{\mathbf{x}}^{\mathrm{unk}}\right)\right]^2$$

The minimum of the SSE is characterized by (4) in which $\nabla Y_{\hat{\boldsymbol{\beta}}_\mathbf{n}}$ introduces nonlinearities. Thus, nonlinear least-squares has been employed to derive $\hat{x}^{\mathrm{unk}}_{1,\ldots,Q}$:

$$\nabla \mathrm{SSE}\left(\hat{x}_1^{\mathrm{unk}}, \ldots, \hat{x}_Q^{\mathrm{unk}}\right) = \mathbf{0} = -2 \cdot \sum_{n=1}^{N}\left[y_n^{\mathrm{unk}} - Y_{\hat{\boldsymbol{\beta}}_\mathbf{n}}\left(\hat{\mathbf{x}}^{\mathrm{unk}}\right)\right] \cdot \nabla Y_{\hat{\boldsymbol{\beta}}_\mathbf{n}}\left(\hat{\mathbf{x}}^{\mathrm{unk}}\right) \qquad (4)$$

It is important to remember that spectra were not acquired from concentration containing samples (i.e. ESAW media) but from algae which *reflect impacts* of environmental conditions. Thus, this application is based on an indirect measurement and the cells can be interpreted as 'measurement mediators'. This approach together with novel chemometric methodologies enables innovations in environmental monitoring.

For quantitative experiments, two sets of samples have been prepared – one for calibration, i.e. for deriving a Predictor Surfaces (3), and one for assessing its prediction power. Both data sets comprised of completely independent samples, i.e. different nutrient situations were included in the calibration and the test set. When preparing calibration samples for the first test, the concentration of one nutrient ($Q = 1$) was varied whereas all other nutrients were kept at the standard ESAW concentrations (*24*, *25*). For deriving the Predictor Surface shown in

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Figure 2(A), bicarbonate at concentrations of 160µM, 2071µM, 5180µM, and 8260µM were used; bicarbonate concentrations of 1110µM, 3630µM, and 6720µM were included into the test set. With four different values for the sole predictor variable $x_1 = c_{HCO_3^-}$ contained in the calibration sets, $P = 3$ is the highest polynomial order (1) that should be realized. Therefore, three Predictor Surfaces were built for $P = 1$, 2, and 3 along with a Principal Component Regression (PCR) for comparison purposes. Figure 3 compares predicted HCO_3^- concentrations in the growing medium as calculated with the three different Predictor Surfaces and the PCR to the 'true' values. The dashed lines in both graphs indicate where predicted equal true concentrations. Obviously, the $P = 1$ Predictor Surface, i.e. a standard linear MLR, is unable to reliably or even reproducibly predict the bicarbonate level in the growing medium. This MLR failure is due to the inappropriateness of linear models for nonlinear data. PCR performed somewhat better – apparently, an elevated number of incorporated principal components ($6 > Q = 1$)) helped to empirically describe nonlinear relations between predictor and response variables. When expanding the Predictor Surface models to quadratic ($P = 2$) and to cubic models ($P = 3$), the prediction quality considerably improves, with $P = 3$ being most precise and reproducible.

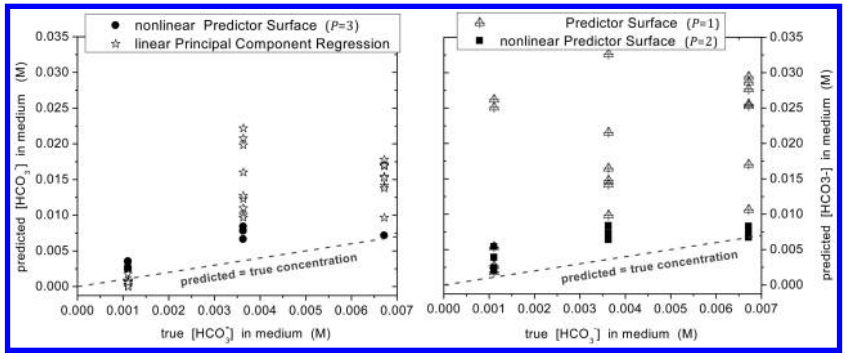

Figure 3. Comparing the prediction power of the novel nonlinear Predictor Surfaces (P = 1 (i.e. MLR), 2, and 3) versus a conventional Principal Component Regression (PCR). Derived work with permission from Elsevier

Impacts of Nutrient Competitors on the Chemical Composition of Microalgae

In the previous section, it has been demonstrated that microalgal biomass chemically adapts to its chemical environment. It has also been hypothesized that the microalgal species composition in a culture ('biological environment') plays a role in the biomass' chemical composition (*21*). The reasoning behind this hypothesis is that different species mutually change their chemical environment through nutrient uptake. The following experiments were conducted to investigate

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch014

impacts due to nutrient competition: Two microalgae species, *Dunaliella parva* and *Nannochlorpsis oculata*, were cultured individually as well as in mixture. From all three culture types (2x single, 1x mix), multiple samples were prepared as outlined above followed by recording of FTIR spectra. By keeping all other parameters in particular the nutrient concentrations the same, any significant spectroscopic differences between single-species cultures and mixed-species cultures is attributed to competition effects. In order to detect any significant spectroscopic modifications among natural replicate-to-replicate fluctuations, a *t*-testing procedure has been performed (95% confidence): From replicate spectra of a certain culture type (single- or mixed-species), mean spectra and their wavenumber dependent standard deviations were computed (Figure 4, top). Said means and their standard deviations were then *t*-tested at every single wavenumber position. However, *t*-testing of a single-species versus a mixed-species culture's spectrum cannot discriminate features originating from competition-induced chemical changes and features of the mixture's other species. To overcome this, spectroscopic features of a mixed culture were compared to *both* single-species cultures contained in the mix. Every single *t*-test determines one of three possible outcomes: (i) there is *no* significant difference between the single and the mixed culture, (ii) the single species has a lower absorbance than the mixture, or (iii) the single species has a higher absorbance than the mixture. Encoded in these two *t*-tests -or more specifically within their $3^2 = 9$ possible outcomes- are two chemically interesting cases: (i) Bands in the mixture spectrum are observed which have not been present in any single-species spectrum. (ii) Bands which are contained in the single-species spectra are missing in the mixture's spectrum. The remaining seven 'situations' are listed in Table 1 and depicted in Figure 4 (top, dark yellow curve). This procedure had been performed for seven bicarbonate concentrations and composed into 2D 'situation planes' (Figure 4, bottom). In such situation planes, the x-axis covers the concentration c of a specific nutrient and the y-axis runs along the wavenumber $\tilde{\upsilon}$. Black rectangles in these $c - \tilde{\upsilon}$ planes indicate under what nutrient concentration c a significant spectroscopic change has been found at which wavenumber $\tilde{\upsilon}$. White areas indicate absence of significant spectroscopic changes for that particular 'situation #'. Further investigations are required to interpret the chemical signatures found within binary-species cultures that cannot be explained by a combination of the two individual species. For orientation, situation #8 in the top graph has been pointed out in the corresponding situation panel below. It was also found that situation #0 (= no competition-induced biomass modification) is the most common one. Please see ref. (*21*) for a presentation of more extensive results.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch014

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch014

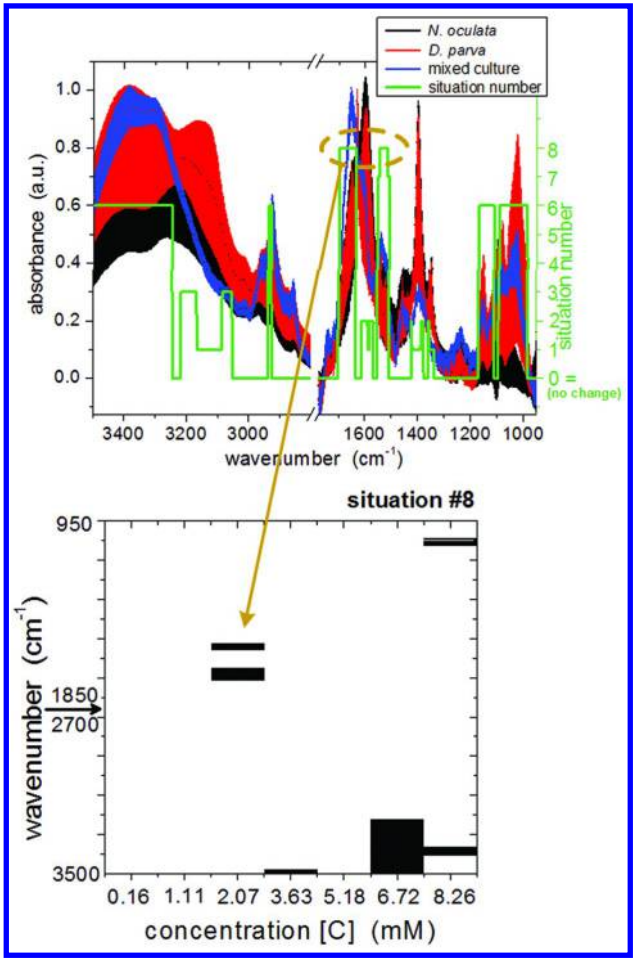

Figure 4. ***(top)*** *FTIR spectra acquired from algae cultures (shaded area = errorbars from replicates) of Nannochloropsis oculata and Dunaliella parva when cultured separately (black and red) and in mixture (blue), respectively (standard conditions: 2071µM [HCO_3^-] and 549µM [NO_3^-]); statistically significant differences between spectra from mixed cultures versus a combination of singly-grown samples were found and assigned a 'situation number' (see Table 1);* ***(bottom)*** *concentration dependency of 'situation 8' occurrences (black areas)*

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch014

Table 1. Nine possible outcomes (situations) for comparing spectra of two single species versus their binary species mixtures (Figure 4, top); these situations are determined at every wavenumber position separately

situation	absorb. of species #1 is:	absorb. of species #2 is:	interpretation
0	same as mix	same as mix	no discernible impact
1	higher than mix	higher than mix	reduction of both species' features in mix
2	higher than mix	same as mix	reduction of feature in species #1 in mix
3	same as mix	higher than mix	reduction of feature in species #2 in mix
4	higher than mix	lower than mix	feature's origin shifts from #1 to #2 in mix
5	lower than mix	higher than mix	feature's origin shifts from #2 to #1 in mix
6	lower than mix	same as mix	feature in mix is due to #2 in mix
7	same as mix	lower than mix	feature in mix is due to #1 in mix
8	lower than mix	lower than mix	generation of new feature in mix

Impacts of Nutrient Availability and Nutrient Competitors on the Growth Dynamics of Microalgae

In the preceding section, it has been presented that microalgae develop biomass of different chemical composition depending on nutrient availability and the presence of nutrient competitors. Based on these findings, it has been hypothesized that the *amount* of produced phytoplankton and the *production dynamics* are also influenced by nutrient availability and competition. Measuring biomass amounts has been based on analyses of images acquired from cell cultures. From microscope images, cell counts and cell size distributions were determined. Recording cell counts over the course of several days gained insights into growth dynamics. Performing such experiments under different nutrient conditions as well as in presence or absence of competitors enabled deducing impacts of these ambient parameters on the biomass production.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch014

Monitoring Cell Culture Growth by Image Analyses

Measuring cell numbers and their size distributions can either be done by means of flow cytometry or based on image analyses recently developed (*19*). While both experimental techniques facilitate *in-situ*, contact free analyses of large numbers of cells, imaging has been chosen as it introduces less sample handling/ disturbance plus it requires less equipment and is thus more widely applicable.

A digital image can be depicted as a 3D plot in which x and y represent the spatial dimensions and z the light intensity at a given pixel (Figure 5 (A) vs. (B)). In transmission microscopy, a homogenous illumination of the sample has been described by a constant background light level a_0 onto which the cells' shadows are superimposed. Each cell's shadow has been modeled as a down-pointing 2D Gaussian Figure 5 (C)) which not only represents the data well but also enables a straightforward measure (*50*) of the cell's cross section via the 2D-Gaussians widths σ_x and σ_y. Assuming an elliptic cell cross section, cell sizes were calculated as $\pi \cdot \sigma_x \cdot \sigma_y$. However, since randomly oriented, non-spherical, 3-dimensional cells were analyzed by a 2D imaging technique, a size distribution rather than one common cell size was found (Figure 5 (D)). As an initial proof-of-concept (*19*), all three microalgae species were cultured under various concentrations of bicarbonate, ammonium, and nitrate. As expected, the two *Dunaliellas* are larger than *Nannocloropsis oculata* and *D. salina* was found to be slightly larger than *D. parva*.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch014

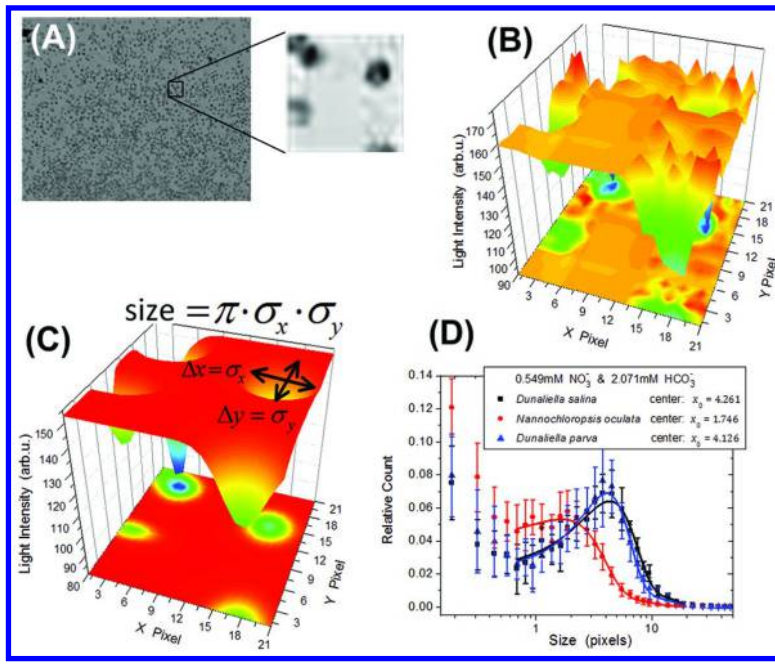

Figure 5. ***(A)*** *transmission microscopy image of Dunaliella salina cells;* ***(B)*** *close-up 3D representation of the cells' shadows (z-axis = light intensity);* ***(C)*** *nonlinear least-squares fits (= 2D Gaussians) of the shadows from which the cells' dimensions in x and y direction are determined via the 2D-Gaussians' 'standard deviations';* ***(D)*** *cell species can be discriminated based on their size distribution. Derived work with permission from John Wiley and Sons*

It was also of interest to assess whether these size distributions reflect in a quantifiable way to determine the nutrient situation of the corresponding cell culture. To build such quantitative calibration models, linear PLS and nonlinear Predictor Surfaces (*18*) have been employed utilizing size distributions in an equivalent way to FTIR spectra in Figure 2. It was found (*19*) that the size distribution of *N. oculata* reflects the concentrations of all three nutrients (HCO_3^-, NO_3^-, and NH_4^+). The size distributions of the two *Dunaliellas* could only predict the two nitrogen containing nutrients but at a higher precision than *N. oculata*. In general, nonlinear Predictor Surfaces achieved a somewhat higher precision than MLR and PCR (see Figure 3) presumably because of fairly strong nonlinearities between size distribution and nutrient concentration.

In a subsequent study (*20*), these chemometric methodologies were augmented in order to mathematically rather than empirically describe the impact of multiple nutrients. This established a novel chemometric tool for investigating

nonlinear and interrelated impacts of multiple ambient chemical parameters on the cells' physical parameters. Advancing chemometrics along these lines will open new research opportunities in biology, ecology, and medicine to study cells' responses to shifts in their ambient conditions.

Modeling Amounts of Produced Biomass

The previous section discussed the impact of ambient chemical parameters (nutrients) on phytoplankton's size distribution. It also has been presented that nutrient competitors impact microalgae's chemical composition (Figure 4). Based on these findings, it has been hypothesized that nutrient competitors may also influence the cell size distributions. Such a competition scenario is given in natural ecosystems and –if found relevant- would need to be considered when assessing phytoplankton's sequestration of inorganic compounds into biomass.

For assessing competition impacts on biomass production namely cell concentrations and cell size distributions, the following approach had been chosen (*22*, *23*): Cultures of *Nannochloropsis oculata* and *Dunaliella parva* were grown individually as well as in binary species mixtures. Providing the same nutrient conditions to both culture types (single vs. mixed) followed by comparing the cell numbers and size distributions, light was shed onto consequences of nutrient competition. For these investigations, six different carbon concentrations (1110µM, 2071µM, 3630µM, 5180µM, 6720µM, and 8260µM) were supplied to cultures via sodium bicarbonate dissolved into the culture media; five nitrogen concentrations (350µM, 549µM, 873µM, 1280µM, and 1470µM) were provided either via ammonium chloride or sodium nitrate. Two different nitrogen sources were incorporated to test for impacts of the nitrogen source. To take the naturally occurring variability in biological materials into account, five replicate cultures were grown for each culture per nutrient condition.

The image analysis method presented before (*19*) (Figure 5 (A)-(C)) had been employed to acquire cell size distributions, *S*, from single-species and binary-species cultures. After composing the measured cell sizes found in a given culture into the discrete bins of a histogram, Poisson-shaped distributions have been observed (Figure 6 (A)) (*22*). These distributions peak at the species-dependent, average cell size S_λ which is covered by bin number λ. While a Poisson distribution $p(k)$ describes the probability for a certain cell to fall into the k^{th} bin, λ and thus S_λ are unknown in this application. However, determining λ is of central interest here as competition induced changes in the cell sizes are reflected in λ. Thus, λ is derived by fitting a 'Poisson-shaped' function $S(k) = A \cdot \frac{\lambda^k}{k!} \cdot \exp\{-\lambda\}$ to cell size histograms. For regression purposes, λ is handled as a continuous fit parameter rather than the integer bin number. Furthermore, an additional fit parameter, A, has been introduced above to describe the maximum height of a histogram which is not normalized like a Poisson distribution.

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch014

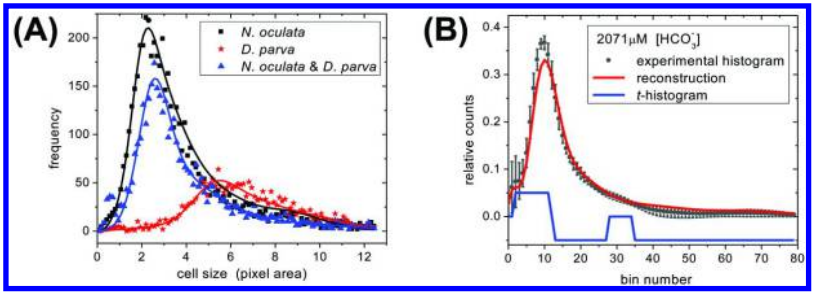

Figure 6. ***(A)*** *size distribution of two mono-species cultures (black, red) in comparison to a mixed-species cultures (blue);* ***(B)*** *statistically significant competition impacts on cell size distributions. (see text; reprinted with permission from John Wiley and Sons (51))*

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch014

If M species are present in a culture, this equation needs to be modified to reflect that a cell size distribution is a superposition of M Poisson-shaped functions: $S(k) = \sum_{m=1}^{M} A_m \cdot \frac{\lambda_m^k}{k!} \cdot \exp\{-\lambda_m\}$. Furthermore, from a practical perspective, M = 20 had been chosen (i.e. >#species) to model out-of-focus cells and randomly oriented, non-spherical cells. Figure 6 (A) shows the resulting fit curves as solid lines.

Under identical nutrient conditions, competition impacts are manifested in differences between a histogram $S_{\text{N.o.\&D.p.}}(k)$ obtained from a binary-species culture and a linear combination of the two single-species histograms $S_{\text{N.o.}}(k)$ and $S_{\text{D.p.}}(k)$. Hence, testing whether the rank of a three-column matrix containing $S_{\text{N.o.}}$, $S_{\text{D.p.}}$, and $S_{\text{N.o.\&D.p.}}$ has rank two or three would theoretically test for absence or presence of competition impacts on the cell size distribution. However, this approach faces two challenges, the difficulty to reliably discriminate between rank two and three and to determine how the size distribution has shifted. Furthermore, noise in the histograms mandates a statistical assessment of size distributions. To overcome the aforementioned limitations, a multivariate least-squares fit solving

$$\begin{pmatrix} S_{\text{N.o.}}(1) & S_{\text{D.p.}}(1) \\ \vdots & \vdots \\ S_{\text{N.o.}}(K) & S_{\text{D.p.}}(K) \end{pmatrix} \cdot \begin{pmatrix} w_{\text{N.o.}} \\ w_{\text{D.p.}} \end{pmatrix} = \begin{pmatrix} S_{\text{N.o.\& D.p.}}(1) \\ \vdots \\ S_{\text{N.o.\& D.p.}}(K) \end{pmatrix} + \boldsymbol{\delta}_{(K\times 1)} \qquad (5)$$

for the two weight factors, $w_{\text{N.o.}}$ and $w_{\text{D.p.}}$, has been chosen. Equation (5) expresses the mixed-species size distribution as a combination of the individual species size distributions. The reconstructed mix-species size distribution (Figure 6 (B)), i.e. $\begin{pmatrix} S_{\text{N.o.}}(1) & S_{\text{D.p.}}(1) \\ \vdots & \vdots \\ S_{\text{N.o.}}(K) & S_{\text{D.p.}}(K) \end{pmatrix} \cdot \begin{pmatrix} \widehat{w}_{\text{N.o.}} \\ \widehat{w}_{\text{D.p.}} \end{pmatrix}$ then describes how much of the

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

measured mix-species size distribution $\begin{pmatrix} S_{\text{N.o.\& D.p.}}(1) \\ \vdots \\ S_{\text{N.o.\& D.p.}}(K) \end{pmatrix}$ can be explained in terms of the single-species distributions. Any differences between measured mix-species and its reconstruction using single-species distributions are then assigned to competition impacts.

As the measured histograms were determined from replicate cultures, errorbars were available for each histogram bin (see Figure 6 (B)). Performing bin-wise *t*-tests (95% confidence) of the mean measured histograms versus their reconstructed counterparts reveals whether these two histograms' bins contain significantly different counts. Based on *t*-test results, the nature of competition-induced shifts in cell size distributions can be determined. In Figure 6 (B), a so-called *t*-histogram is shown that reflects the outcomes of said *t*-tests: a value of 0 in this *t*-histogram indicates no significant difference between measured and reconstructed histogram. A positive value in the *t*-histogram indicates that the measured histogram contains significantly more counts in a given bin than its reconstruction. A negative value means that the reconstruction has more counts in that bin than experimentally determined. The example shown in Figure 6 (B) therefore indicates that, under 2071μM bicarbonate, the mixed-culture's size distribution (black) is shifted to smaller cell sizes than expected due to the reconstruction (red) made from single-species size distributions. Competition impacts in different nutrient situations are presented in ref. (*22*). These results clearly indicate that the cell size distribution is impacted by nutrient competition and suggest further studies to determine the physiological origin of these cell size modifications.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch014

Modeling a Culture's Growth Dynamics

Investigating the extent to which the dynamics of biomass production is related to nutrient availability and/or nutrient competitors is of additional interest as this will determine the microalgal biodiversity in an ecosystem. Since all species have somewhat different nutrient allocation and utilization characteristics, this will also impact an ecosystem's overall transformation of inorganic compounds into biomass. Investigating these topics has been based on a culture's time-dependent cell concentration. Such growth curves (*22*, *23*) describe the cell concentration (*52*) $y(t)$ (6) in a culture which had been started at $t = t_0$ by inoculating y_0 cells per mL into the culturing medium. A culture's growth dynamics is described by its growth rate s which is inversely proportional to the time span τ within which the cell concentration doubles. Eventually, the growth slows down and asymptotically reaches a maximum cell concentration $y_{\max}$ (see appendix in ref. (*22*)). Figure 7 (A) and (B) depict two microalgal growth curves $y(t)$ with s and $y_{\max}$ being clearly species dependent.

$$y(t) = y_{\max} \cdot \frac{1}{1 + \exp\left\{-\frac{y_{\max}}{y_{\max} - y_0} \cdot s \cdot (t - t_0)\right\} \cdot \left(\frac{y_{\max}}{y_0} - 1\right)} \qquad (6)$$

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch014

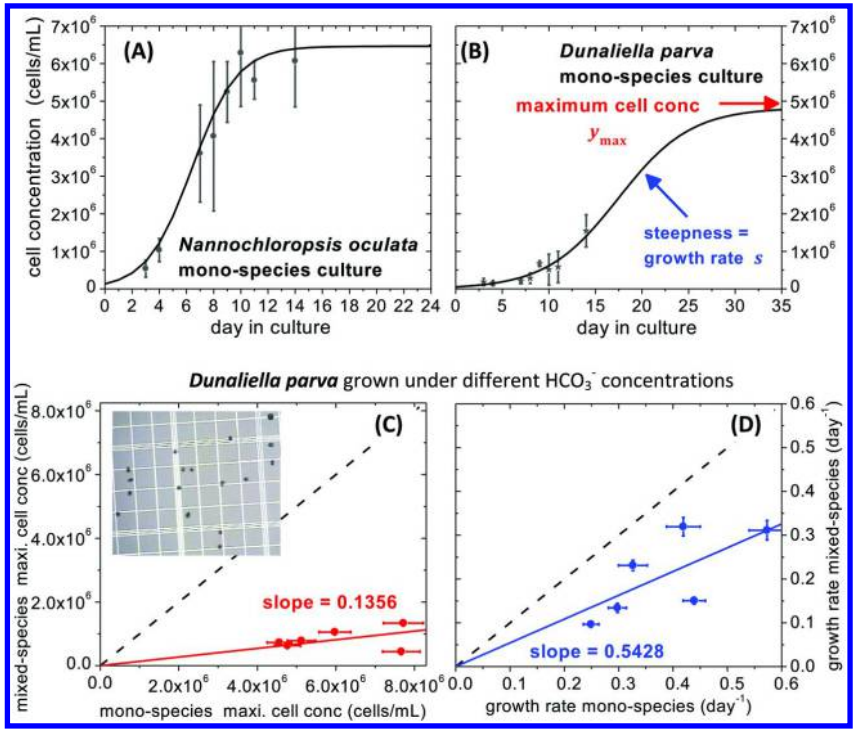

Figure 7. ***(A)*** *&* ***(B)*** *a given growing environment can produce a species-specific maximum cell concentration* y_{max} *at a species-specific growth rate s (6);* ***(C)*** *maximum cell concentration* y_{max} *and* ***(D)*** *and growth rate s are clearly impacted by introducing a nutrient competitor; inset in (C): accurate cell concentrations were obtained with a hemocytometer under a 10x microscope. (reprinted with permission from John Wiley and Sons)*

In order to measure s along with $y_{\max}$ and $_y0$, cell counts $y(t)$ have been obtained on multiple days t via hemocytometer counting (Figure 7 (C) inset) to which equation (6) has been fitted. While $y_{\max}$ is not related to growth dynamics, it represents valuable information for assessing the amount of biomass produced by a culture (see previous section 'Modeling Amounts of Produced Biomass'). From single-species cultures of *Nannochloropsis oculata* (N.o.) and *Dunaliella parva* (D.p.), $s_{\text{N.o.(single)}}$ and $s_{\text{D.p.(single)}}$ have been determined for all nutrient situations listed above and for all replicate cultures (*22*). These growth rates were then compared to their counterparts derived from binary-species cultures (*53*), i.e. $s_{\text{N.o.(mix)}}$, and $s_{\text{N.o.(single)}}$. Comparing $s_{\text{N.o.(single)}}$ to $s_{\text{N.o.(mix)}}$ as well as $s_{\text{D.p.(single)}}$ to $s_{\text{D.p.(mix)}}$ as obtained from otherwise identical culturing conditions reveals impacts of nutrient competition onto phytoplankton growth dynamics. Figure 7 (D) depicts one selected example of growth rates as derived from replicated mixed-species cultures versus replicated single-species cultures' values. A 45°-line (gray dashed) has been included to indicate where $s_{\text{(single)}} = s_{\text{(mix)}}$. If a data point falls below this 45°-line, $s_{\text{(single)}} > s_{\text{(mix)}}$ indicating the growth rate is

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

reduced in species mixtures compared to the single-species culture containing the same nutrient situation. If a data point is above the 45°-line, the growth rate is enhanced in these mixtures. The blue line is a linear regression line of $s_{(mix)}$ versus $s_{(single)}$ under the constraint that the regression function passes through the origin. Such fit lines' slopes then measure how the growth rates of the mixture and the single-cell cultures deviate. In the given example, the impact of a competing species on the growth rate is considerable.

From the same data and the same fit (6), a comparison of the maximum cell concentration y_{max} has been performed in an equivalent manner. Figure 7 (C) shows one representative example of $y_{max(mix)}$ versus $y_{max(single)}$ together with gray, dashed 45°-lines and a red regression line constrained to pass through the graph's origin. In the example shown here, the species' maximum concentration has been massively reduced due to the presence of a nutrient competitor. On the other hand, an equivalent analysis for said competitor revealed (*22*) (not shown here) that that species' y_{max} has only been altered minutely. Therefore, one species outcompetes the other and drives the biodiversity to a different equilibrium than one would expect based on single species' growth curves.

For these investigations, microalgae cells on a cell culture level had been considered. In order to better understand what physiological origins these competition impacts have, ref. (*23*) describes this scenario from a single-cell level based on a species' nutrient uptake characteristics. These studies also demonstrate how hard-modeling chemometrics enables in-depth investigations of complex chemical/biological/ecological systems (*35*).

Conclusions

This study reports on innovations in chemical sensing with the overarching goal to investigate chemical interactions between microalgae cells and their chemical and biological environment. Such investigations aim at increasing the understanding of the key parameters that drive phytoplankton based compound transformation from inorganics dissolved in marine environments into biomass. Due to the interrelation among environmental parameters and the impacts they leave on algal biomass, novel chemometrics modeling strategies are mandatory and have been developed in the course of this research. In particular, hard-modeling was deemed to be highly advantageous compared to conventional soft-modeling because only the former enables a chemical interpretation of the models. This interpretability then increases the understanding of chemical mechanisms at the interface between chemistry in the ambience and biological samples. Furthermore, modifications of the biomass' chemical signature could be related to nutrient concentrations in the cells' culturing environment and thus enable novel approaches for embedded chemical sensors (=cells). Regarding the chemical composition of microalgae, it was found that the nutrient mix rather than the concentrations of individual nutrients is of fundamental importance. It was also demonstrated that the presence of multiple microalgae species considerably changes the picture. One interpretation of this effect is that different species through nutrient uptake mutually modify their chemical environment.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch014

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

The production of biomass has been investigated as a function of microalgae's chemical and biological environment. It was found that the chemical environment alters the cell size distribution so strongly that such distributions reflect ambient conditions in a quantifiable way. More striking results were achieved by demonstrating the impact of nutrient competitions on biomass production dynamics and efficiency. This will enable studies of biodiversity from an environmental chemistry perspective and thus can bridge the gap between chemistry and ecology.

Overall, results of this study demonstrates that chemical analyses of life biological samples cannot be performed without considering the samples' chemical and biological environment. In the author's opinion, nonlinear (hard)modeling of interactions between samples and their environment is an emerging field in chemometrics.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch014

Acknowledgments

This work was supported by the National Science Foundation under CHE-1058695 and CHE-1112269.

References

1. Peters, G.; Marland, G.; Le Quéré, C.; Boden, T.; Canadell, J.; Raupach, M. *Nat. Clim. Change* **2012**, *2*, 2–4.
2. Eby, M.; Zickfeld, K.; Montenero, A.; Archer, D.; Meissner, K.; Weaver, A. *J. Clim.* **2009**, *22*, 2501–2511.
3. Blunden, J., Arndt, D., Baringer, M., Eds. *Bull. Am. Meteorol. Soc.* **2011**, *92*, S1−S266.
4. Hays, G. ; Richardson, A.; Robinson, C. *Trends Ecol. Evol.* **2005**, *20*, 337–344.
5. Bardgett, R.; Freeman, C.; Ostle, N. *ISME J.* **2008**, *2*, 805–814.
6. Zhou, J.; Xue, K.; Xie, J.; Deng, Y.; Wu, L.; Cheng, X.; Fei, S.; Deng, S.; He, Z.; Van Nostrand, J.; Luo, Y. *Nat. Clim. Change* **2012**, *2*, 106–110.
7. Field, C.; Behrenfeld, M.; Randerson, J.; Falkowski, P. *Science* **1998**, *281*, 237–240.
8. Behrenfeld, M.; O'Malley, R.; Siegel, D.; McClain, C.; Sarmiento, J.; Feldman, G.; Milligan, A.; Falkowski, P.; Letelier, R.; Boss, E. *Nature* **2006**, *444*, 752–755.
9. Martinez, E.; Antoine, D.; D'Ortenzio, F.; Gentili, B. *Science* **2009**, *326*, 1253–1256.
10. Raven, J.; Giordano, M.; Beardall, J.; Maberly, S. *Photosynth. Res.* **2011**, *109*, 281–296.
11. Quéré, C.; Rödenbeck, C.; Buitnehuis, E.; Conway, T.; Langefelds, R.; Gomez, A. *Science* **2007**, *316*, 1735–1738.
12. Milligan, A. *Nat. Clim. Change* **2012**, *2*, 489–490.
13. Flynn, K.; Blackford, J.; Baird, M.; Raven, J.; Clark, D.; Beardall, J.; Brownlee, C.; Fabian, H.; Wheeler, G. *Nat. Clim. Change* **2012**, *2*, 510–513.

14. Walsh, J.; Jolliff, J.; Darrow, B.; Lenes, J.; Milroy, S.; Remsen, A.; Dieterle, D.; Carder, K.; Chen, F.; Vargo, G.; Weisberg, R.; Fanning, K.; Muller-Karger, F.; Shinn, E.; Steidinger, K.; Heil, C.; Tomas, C.; Prospero, J.; Lee, T.; Kirkpatrick, G.; Whitledge, T.; Stockwell, D.; Villareal, T.; Jochens, A.; Bontempi, P. *J. Geophys. Res.* **2006**, *111*, C11003.
15. Boss, E.; Behrenfeld, M. *Geophys. Res. Lett.* **2012**, *37*, L18603.
16. Behrenfeld, M; Halsey, K.; Milligan, A. *Philos. Trans. R. Soc. B* **2008**, *363*, 2687–2703.
17. Bilanovic, D.; Andargatchew, A.; Kroeger, T.; Shelef, G. *Energy Convers. Manage.* **2009**, *50*, 262–267.
18. Horton, R.; McConico, M.; Landry, C.; Tran, T.; Vogt, F. *Anal. Chim. Acta* **2012**, *746*, 1–14.
19. McConico, M.; Horton, R.; Witt, K.; Vogt, F. *J. Chemom.* **2012**, *26*, 585–597.
20. McConico, M.; Vogt, F. *J. Chemom.* **2013**, *27*, 217–219.
21. McConico, M.; Vogt, F. *Anal. Lett.* **2013**, *46*, 2752–2766.
22. White, L.; Martin, D.; Witt, K.; Vogt, F. *J. Chemom.* **2014**, *28*, 448–461.
23. Fleming, S.; Vogt, F. *J. Chemom.* **2015**, *29*, 139–141.
24. Andersen R. *Algae Culturing Techniques*; Elsevier Academic Press: Amsterdam, 2005.
25. Berges, J.; Franklin, D. *J. Phycol.* **2001**, *37*, 1138–1145.
26. *Note*: Accuracy and precision of subsequent, quantitative analyses have been improved by dividing each spectrum by the mass of IR absorbing algae material contained in a certain KBr-algae pellet. This step normalized for the mass contained in a sample and thus reduced spectroscopic fluctuation among replicates induced by imperfections in manual sample preparation (*27*).
27. Horton, R.; Duranty, E.; McConico, M.; Vogt, F. *Appl. Spectrosc.* **2011**, *65*, 442–453.
28. Murdock, J.; Wetzel, D. *Appl. Spectrosc. Rev.* **2009**, *44*, 335–361.
29. Kansiz, M.; Heraud, P.; Wood, B.; Burden, F.; Beardall, J.; MacNaughton, D. *Phytochemistry* **1999**, *52*, 407–417.
30. Giordano, M.; Kansiz, M.; Heraud, P.; Beardall, J.; Wood, B.; McNaughton, D. *J. Phycol.* **2001**, *37*, 271–279.
31. Stehfest, K.; Soepel, J.; Wilhelm, C. *Plant Physiol. Biochem.* **2005**, *43*, 717–726.
32. Domenighini, A.; Giordano, M. *J. Phycol.* **2009**, *45*, 522–531.
33. Giordano, M.; Ratti, S.; Domighini, A.; Vogt, F. *Plant Ecol. Divers.* **2009**, *2*, 155–164.
34. *Note*: For some applications such as live cell monitoring, drying the cells is highly unwanted but on the other hand, applying conventional mid-IR spectroscopy is hampered by live cells' water contents. In order to avoid water blocking-out wavenumber regions, utilizing a synchrotron as a powerful IR light source has been proposed (*54−57*). Alternatively, Raman spectroscopy has been utilized (*58−60*) taking advantage of water molecules not being Raman active. Yet, Raman spectroscopy is not as widely used as FTIR possibly because only a limited number of biologically relevant molecules feature a sufficient signal-to-noise.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch014

35. Vogt, F. *J. Chemom.* **2014**, *28*, 785–788.
36. Vogt, F. *J. Chemom.* **2015**, *29*, 71–79.
37. *Note*: For example, if k is a time constant in a theoretically derived model function $\exp\{-k \cdot t\} \approx 1 - k \cdot t = a_0 + a_1 \cdot t$, calculating numerical values for $k = -a_1$ via fitting data to the stated linear approximation conserves the chemical/physical interpretation. If a higher order approximation is required, fitting experimental data to e.g. a second order polynomial $\exp\{-k \cdot t\} \approx 1 - k \cdot t + (1/2) \cdot k \cdot t^2 = a_0 + a_1 \cdot t + a_2 \cdot t^2$ still derives chemical/physical insights via $k = -a_1$. Furthermore, the regression can be made more robust by utilizing theoretical knowledge in form of an equality constraint $-(1/2) \cdot a_1 = a_2$. This very simple example outlines how even approximated theoretical considerations supplemented by calculus can advance hard-modeling.
38. Vogt, F. *Anal. Chim. Acta* **2013**, *797*, 20–29.
39. Draper, N.; Smith H. *Applied Regression Analysis*, 3rd ed.; John Wiley & Sons: New York, NY, 1998.
40. Jolliffe, I. *Principal Component Analysis*, 2nd ed.; Springer: New York, NY, 2002.
41. Gilbert, M.; Luttrell, R.; Stout, D.; Vogt, F. *J. Chem. Educ.* **2008**, *85*, 135–137.
42. Haaland, D.; Thomas, E. *Anal. Chem.* **1988**, *60*, 1193–1202.
43. Martens, H.; Naes T. *Multivariate Calibration*, 2nd ed.; John Wiley & Sons: New York, NY, 1991.
44. Vogt, F.; Klocke, U.; Rebstock, K.; Schmidtke, G.; Wander, V.; Tacke, M. *Appl. Spectrosc.* **1999**, *53*, 1352–1360.
45. Box, G.; Draper, N. *Empirical Model-Building and Response Surfaces*; John Wiley & Sons: New York, 1987.
46. Vogt, F.; Gritti, F.; Giuochon, G. *J. Chemom.* **2011**, *25*, 575–585.
47. Turner, J.; Thayer, J. *Introduction to Analysis of Variance*; Sage Publications: Thousand Oaks, CA, 2001.
48. Crow, E.; Davis, F.; Maxfield, M. *Statistics Manual*; Dover Publishing: Mineola, NY, 2011.
49. *Note*: If the samples would obey a linear model, the (x_1, x_2)-surfaces would be a plane linearly increasing in both x_1 and x_2 directions. Instead, the modeled absorbance at 1101cm^{-1} is a strongly curved surface.
50. *Note*: As a model equation for nonlinear least-squares, a homogeneous illumination a_0 of the samples is superimposed by a cell's shadow of 2D Gaussian shape ($A < 0$): $Z(x,y) = a_0 + A \cdot \exp\left\{-\frac{1}{2} \cdot \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix}^{\mathbf{T}} \cdot \mathbf{\Sigma^{-1}} \cdot \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix}\right\}$. The matrix $\mathbf{\Sigma} = \begin{pmatrix} \sigma_x^2 & \rho_{xy} \cdot \sigma_x \cdot \sigma_y \\ \rho_{xy} \cdot \sigma_y \cdot \sigma_x & \sigma_y^2 \end{pmatrix}$ whose *inverse* is contained in Z is a 'variance-covariance matrix' with σ_x and σ_y describing the widths of a 2D Gaussian along its two principal axes. A 2D Gaussian whose principal axes are not aligned along the x and y-axes introduces a correlation ($\rho_{xy} \neq 0$) between x and y and hence nonzero off-diagonal elements in $\mathbf{\Sigma}$. Ref. (*19*)

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch014

presents details of this fitting procedure including expanding the model equation to multiple cells.

51. *Note*: Theoretically, the originally measured histograms could have been used instead of histograms decomposed into Poisson-shaped functions (5). However, the considerable noise level in the histograms would have obliterated clear trends among size changes (see supplemental material in ref. (*22*)).
52. *Note*: The time dependent number of microalgae cells in a culture, $y(t)$ in (6), is not to be confused with the spectrum $\mathbf{y}_{(N\times 1)}$ (black curve in Figure 2) acquired from a cell culture.
53. *Note*: Since the chosen species have clearly different sizes (Figure 5 (D), Figure 6 (A)), a visual discrimination while hemocytometer-based cell counting of both species contained in a mixture was feasible.
54. Heraud, P.; Wood, B.; Tobin, M.; Beardall, J.; McNaughton, D. *FEMS Microbiol. Lett.* **2005**, *249*, 219–225.
55. Hirschmugl, C.; Zuheir-El Bayarri, B.; Bunta, M.; Holt, J.; Giordano, M. *Infrared Phys. Technol.* **2006**, *49*, 57–63.
56. Nasse, M.; Walsh, M.; Mattson, E.; Reininger, R.; Kajdacsy-Balla, A.; Macias, V.; Bhargava, R.; Hirschmugl, C. *Nat. Methods* **2011**, *8*, 413–416.
57. Hirschmugl, C.; Gough, K. *Appl. Spectrosc.* **2012**, *66*, 475–491.
58. Diem, M.; Romeo, M.; Boydston-White, S.; Miljkovic, M.; Matthaus, C. *Analyst* **2004**, *129*, 880–885.
59. Heraud, P.; Beardall, J.; McNaughton, D.; Bayden, R. *FEMS Microbiol. Lett.* **2007**, *275*, 24–30.
60. Huang, Y.; Beal, C.; Cai, W.; Ruoff, R.; Terentjev, E. *Biotechnol. Bioeng.* **2010**, *105*, 889–898.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ch014

Subject Index

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ix002

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.; ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

D

K

M

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.; ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ix002

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ix002

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

N

P

R

S

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ix002

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.

W

Publication Date (Web): October 7, 2015 | doi: 10.1021/bk-2015-1199.ix002

In 40 Years of Chemometrics – From Bruce Kowalski to the Future; Lavine, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2015.