International Journal of Pharmaceutics 417 (2011) 280–290

Contents lists available at ScienceDirect

International Journal of Pharmaceutics

journal homepage: www.elsevier.com/locate/ijpharm

Review

Multivariate data analysis in pharmaceutics: A tutorial review

Tarja Rajalahti*, Olav M. Kvalheim

Department of Chemistry, University of Bergen, Allégaten 41, N-5007 Bergen, Norway

ARTICLE INFO

Article history:
Received 18 January 2011
Accepted 10 February 2011
Available online 16 February 2011

Keywords:
Multivariate analysis
Latent variables
Pretreatment
Variable selection
PAT
Pharmaceutics

ABSTRACT

We provide an overview of latent variable methods used in pharmaceutics and integrated with advanced characterization techniques such as vibrational spectroscopy. The basics of the most common latent variable methods, principal component analysis (PCA), principal component regression (PCR) and partial least-squares (PLS) regression, are presented. Multiple linear regression (MLR) and methods for improved interpretation, variable selection, classification and validation are also briefly discussed. Extensive use of the methods is demonstrated by compilation of the recent literature.

Contents

1. Introduction

Measured data are not the same as information. Therefore an important issue in all empirical sciences, including pharmaceutical sciences, is how to reveal the relevant information in the data. Chemometrics can be defined as "information aspects of chemistry" (Wold and Sjostrom, 1998) where statistical and mathematical methods are used (i) to produce "good data", and, (ii) to extract relevant information from measured data. The first aim can be achieved by using design of experiments (DoE) to provide a small number of information-rich experiments. Multivariate data analysis can be employed for the second purpose. In addition visu-

* Corresponding author. Tel.: +47 55583366; fax: +47 55589490.
E-mail address: Tarja.Rajalahti@kj.uib.no (T. Rajalahti).

0378-5173/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijpharm.2011.02.019

alization of the data represents an important issue. The methods used in chemometrics are fully applicable in pharmaceutical sciences. Multivariate projection methods can be used to simplify complex pharmaceutical data and thus make the visualization easier. Furthermore they make for example classification of samples and prediction of outcome possible.

Instrumentation developed in the field of process analytical chemistry (PAC) supply data about the state of a process (Callis et al., 1987). Off-line instrumentation requires manual sampling and transport to a laboratory with the analytical instrument. At-line instrumentation includes also manual sampling but the analyzer is located close to the process line. On-line instrumentation consists of automated sampling system in combination with an automated analyzer. In-line instrumentation performs the analysis *in situ* using a probe located in the process stream. In noninvasive instrumentation the probe does not have a physical contact with the sample. This represents the most desired situation since sampling problems are greatly reduced. Vibrational spectroscopy techniques such as infrared (IR), near infrared (NIR), and Raman, and imaging techniques are characterization methods that have been applied in pharmaceutical industry to monitor physical and chemical phenomena occurring during the processes. These techniques produce data with high dimensionality, since each sample is described with hundreds or even thousands of variables. Combination of PAC instrumentation and multivariate analysis provides tools for effective process monitoring and control enabling detection of multivariate relationships between different variables such as raw materials, process conditions, and end products. Thus multivariate methods can play a critical role in process understanding, multivariate statistical process control (MSPC) (MacGregor and Kourti, 1995), fault detection and diagnosis, process control and process scale-up.

Process analytical technology (PAT) has its roots in PAC (Kourti, 2006). The aim of the PAT initiative is to increase process understanding and control and at the same time reduce the uncertainty and variation in the quality of the end product (United States Food and Drug Administration (FDA), 2004). The objective is to assure and build in quality throughout manufacturing process, also referred as quality by design (QbD), and enable prompt problem solving if necessary (Yu, 2008). Chemometric techniques, both multivariate data analysis and DoE, have a central role in PAT initiative.

This tutorial review covers the area of multivariate data analysis and theoretical background to the methods is provided. Several pharmaceutical applications employing advanced characterization techniques in combination with multivariate data analysis are reported. Some of the recent ones and their corresponding references are presented in Table 1. The present paper does not aim at detailed discussion of the applications, but to show the variability of pharmaceutical applications and to give an overview of the possibilities that multivariate data analysis method can provide. Multivariate data analysis has proven to be a powerful tool when combined with advanced characterization techniques. Theory of DoE is not included here and literature with references covering the field of experimental design and optimization can be found elsewhere (Box et al., 1978; Gabrielsson et al., 2002; Lundstedt et al., 1998; Mandenius and Brundin, 2008).

2. Theory

2.1. Background

Models can be seen as tools to describe reality. Empirical models based on the experimental data can be estimated and used for interpretation and prediction. All models are more or less erroneous, since there are always noise and other irrelevant features in the data. Experimental error is produced by both known and unknown disturbing factors that may confound important effects wholly or partially. This can be reduced and sometimes almost eliminated by using DoE and statistical analysis. Confusion of correlation with causation is a common problem in all empirical researches. Correlation between two variables often occurs because they are both associated with a third factor meaning that correlation does not automatically imply that the two variables have a causal relationship. One famous example of this is the positive correlation between the number of inhabitants and the number of storks observed in the German city Oldenburg in the 1930s (Box et al., 1978). Correlations are necessary for prediction purposes but should never be interpreted as direct causality. Validation, interpretation and reduction of multivariate regression models estimated from non-designed collinear data represent other challenges.

2.2. Notation

Generally, bold uppercase characters (e.g. $\mathbf{X}$) represent matrices, bold lowercase characters (e.g. $\mathbf{x}$) represent vectors, and italic characters (e.g. N) represent scalars. The transpose is indicated by a superscript T (e.g. $\mathbf{X}^{\mathrm{T}}$). The transpose of a column vector is a row vector and vice versa. Vectors are by default column vectors – a transposed vector is therefore a row vector. Similarly, the transpose of a matrix means that the matrix is rearranged by switching rows and columns. The inverse of a matrix is indicated by a superscript -1 (e.g. $\mathbf{X}^{-1}$).

2.3. Multiple linear regression (MLR)

DoE represents a special case of predictive modelling. The objective of predictive modelling is to determine the relationship between several x-variables (often called independent or explanatory variables) and one or more y-variables (dependent or response variables). This objective can be achieved by means of a model, where the observed result, i.e. response (y), is described as a function of the x-variables, usually called factors ($x_1, x_2, \ldots, x_N$) in DoE. The noise is left in the residual (e_y).

$$y = f(x_1, x_2, \ldots, x_N) + e_y \tag{1}$$

For practical purposes, the function f can usually be approximated by using polynomial functions. For instance, a model for N non-interacting x-variables linearly correlated to y can be written as:

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_Nx_N + e_y \tag{2}$$

where b_i ($i = 0, 1, 2, \ldots, N$) are regression coefficients describing the effect of each calculated term. Eq. (2) can be written in matrix form:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}_y \tag{3}$$

The parameters $\mathbf{b}$ can be estimated by a least squares fit minimizing the sum of squared residuals. Multiple linear regression (MLR) is used for estimating the regression vector $\mathbf{b}$. From Eq. (3) we obtain:

$$\mathbf{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{4}$$

If all the x-variables can be controlled, we can select discrete levels for each x-variable so as to enforce orthogonality between them and their derived interactions and squared terms. The matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ then becomes a diagonal matrix and $\mathbf{b}$ is easily calculated.

When the x-variables are not controlled or the number of x-variables is exceeding the number of experiments, co-linearity arises between x-variables. In the latter case, the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ has no longer full rank implying that the usual inverse of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ no longer exists. The same may happen in the former case even if the number of experiments is larger than the number of variables. Regression

Table 1
Pharmaceutical applications where advanced characterization techniques are used in combination with multivariate data analysis methods.

No.	Characterization method	Application	MVA method	Reference
1	NIR	API and excipient content in pharmaceutical formulations (powders and tablets)	PCA, PLS	Sarraguça and Lopes (2009)
2	NIR	API content in intravenous injections	PLS	Lopez-Arellano et al. (2009)
3	NIR	API content in pellets	PLS	Mantanus et al. (2010a)
4	NIR	API content in pellets	PCA, PLS	Mantanus et al. (2010b)
5	NIR	API content in syrups	PLS	Ziemons et al. (2010)
6	NIR	API content in tablets	PLS, PCR	Chalus et al. (2005)
7	NIR	Calibration transfer for solid formulations	PLS	Bergman et al. (2006)
8	NIR	Characterization of polymorphs	PCA, PLS	Blanco et al. (2004)
9	NIR	Characterization of polymorphs	MCR	Blanco et al. (2006)
10	NIR	Characterization of powder blending	DoE, PLS	Shi et al. (2008)
11	NIR	Granulation process analysis	DoE, PCA, PLS	Rantanen et al. (2005)
12	NIR	Monitoring of API fermentation and downstream purification	PCA, PLS, SOM[a]	Lopes et al. (2004)
13	NIR	Monitoring of fluidized drying	PLS	Peinado et al. (2011)
14	NIR	Monitoring of moving solids	PCA, PLS, SIMCA	Andersson et al. (2005)
15	NIR	Monitoring of powder mixing	PLS	Vanarase et al. (2010)
16	NIR	Monitoring of powder quality	PLS	Märk et al. (2010)
17	NIR	Quantification of API and excipients in powder blends	DoE, MLR, PCR, PLS	Wu et al. (2009)
18	NIR	Quantification of film thickness on pellets	PLS	Lee et al. (2011)
19	NIR	PAT application in powder blending	DoE, PCA	El-Hagrasy et al. (2006a)
20	NIR	PAT application in powder blending	PCA, SIMCA	El-Hagrasy et al. (2006b)
21	NIR	PAT application in powder blending	MLR, PCR, PLS	El-Hagrasy and Drennen (2006)
22	NIR	PAT application in tabletting process	PLS	Moes et al. (2008)
23	NIR	Pharmaceutical batch process analysis	N-way methods	Stordrange et al. (2004)
24	NIR	Powder flow characterization of pharmaceutical formulations	PLS	Benedetti et al. (2007)
25	NIR	Water content in freeze drying	PCA, PLS	Grohganz et al. (2009)
26	NIR	Water content in pellets	PLS	Mantanus et al. (2009)
27	NIR imaging	Characterization of counterfeit tablets	MCR	Lopes et al. (2010)
28	NIR imaging	Characterization of powder blends	PCA, PLS-DA	Ma and Anderson (2008)
29	NIR imaging	Estimation of differences in textures of pharmaceutical tablets	PCA, PLS	Svensson et al. (2006)
30	NIR imaging	Quantification and distribution of components in pharmaceutical tablets	CLS[b], MCR	Amigo and Ravn (2009)
31	Raman	API content in capsules	PLS	Kim et al. (2007a)
32	Raman	API content in liquids	PLS	Kim et al. (2007b)
33	Raman	API content in suspensions	PLS	Park et al. (2007)
34	Raman	API content in tablets	PLS, MCR	Fransson et al. (2010)
35	Raman	API content in tablets	PLS	Johansson et al. (2005)
36	Raman	Content uniformity of tablets	PCA, PLS	Wikström et al. (2006)
37	Raman	Detection of counterfeit tablets	PCA, HCA[c]	de Veij et al. (2007)
38	Raman	Identification of tablets	SVM[d], PLS	Roggo et al. (2010)
39	Raman	Monitoring of freeze drying	DoE, PCA, MCR	De Beer et al. (2007)
40	Raman	PAT application in active coating	PLS, MCR	Müller et al. (2010)
41	Raman	Quantitative analysis of tablets and capsules	PLS	Johansson et al. (2007)
42	Raman	Quantification of polymorphs in powder mixtures	PLS, ANN[e]	Braun et al. (2010)
43	Raman	Tablet coating thickness	TFA[f], PCR, PCA	Kauffman et al. (2007)
44	Raman	Tablet coating thickness and characterization	PLS	Romero-Torres et al. (2006)
45	Raman	Tablet coating variability	PLS	Romero-Torres et al. (2005)
46	Raman imaging	Characterization of tablets	PCA, MCR	Zhang et al. (2005)
47	IR	Monitoring of crystallization process	MSPC, PCA, PLS	Pöllänen et al. (2005)
48	IR	Permeation of model drugs through membrane and human skin	TFA	Russeau et al. (2009)
49	IR imaging	API and excipient content in pharmaceutical formulations	CLS, PLS	Gendrin et al. (2007)
50	IR imaging	Qualitative analysis of solid forms	PCA	Roggo et al. (2005)
51	NIR + Raman	API content in wafers	PLS	Haag et al. (2009)
52	NIR + Raman	Monitoring of dehydration behaviour	PLS-DA	Kogermann et al. (2007)
53	NIR + Raman	Monitoring of freeze drying	PCA, MCR	De Beer et al. (2009)
54	NIR + Raman + NIR imaging	Prediction of physical properties of matrix tablets	PLS	Shah et al. (2007)
55	NIR + IR	Four examples from pharmaceutical industry	DoE, PCA, PLS, SIMCA	Lundstedt-Enkel et al. (2006)
56	NIR + IR	New tablet formulation	DoE, PCA, PLS	Gabrielsson et al. (2006a)
57	NIR + IR	Robustness testing in new tablet formulation	DoE, PCA, PLS	Gabrielsson et al. (2006b)
58	IR + dissolution curves	Differentiation of crystalline polymorphs of API	PCA	Maggio et al. (2009)
59	Raman + X-ray diffractometry	Solid state analysis	PCA	Jorgensen et al. (2006)
60	HPLC	Characterization of herbal medicine	TP	Chau et al. (2009)
61	HPLC	Chromatographic purity analysis	PCA, MSPC	Laursen et al. (2010)
62	LC/MS	Characterization of impurities in pharmaceuticals	MCR	Zomer et al. (2005)
63	Laser diffractometry	API batch to batch variation	PCA	Hagsten et al. (2008)
64	Laser diffractometry	Dry powder inhaled (DPI) formulations	DoE, PCA	Guenette et al. (2009)
65	Sieve analysis	Characterization of compaction and tablet properties	PCA, PLS	Haware et al. (2009)

Table 1 (*Continued*)

No.	Characterization method	Application	MVA method	Reference
66	Spatial filtering (SFT)	Fluid bed granulator	DoE, PLS	Närvänen et al. (2009)
67	Acoustic emission	Determination of end-product granule size distribution	N-way methods	Matero et al. (2010)
68	Mass distribution profiles	Characterization of the performance of nebulizers	PCA, O-PLS	Shi et al. (2009)
69	Digital images + SEM[g]	Prediction of packing efficiency and different metrics of flowability	PLS, PCA	Sandler and Wilson (2010)
70	Digital images	Coating uniformity in immediate release tablets	PCA, PLS	García-Muñoz and Gierer (2010)
71	Digital images	Visual characterization of pharmaceutical solids	DoE, PCA, PLS	García-Muñoz and Carmody (2010)
72	Digital images	Visual characterization of pharmaceutical solids	PCA, PLS	Laitinen et al. (2004)

[a] SOM: self organized maps.
[b] CLS: classical least squares.
[c] HCA: hierarchical cluster analysis.
[d] SVM: support vector machines.
[e] ANN: artificial neural networks.
[f] TFA: target factor analysis.
[g] SEM: scanning electron microscope.

coefficients can still be calculated by introducing the so-called generalized inverse $\mathbf{X}^+$:

$$\mathbf{X}^+ = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \quad (5)$$

By introducing the generalized inverse into Eq. (4), we obtain the expression for calculation of the regression vector as:

$$\mathbf{b} = \mathbf{X}^+\mathbf{y} \quad (6)$$

Except for cases where *x*-variables are controlled in designed experimentation, measured data in pharmaceutical applications are typically multivariate and collinear and MLR cannot be used. This is a main reason why latent variable regression (LVR) methods such as partial least squares (PLS) have become popular. Instead of using the original variables in the regression, we calculate a new set of orthogonal (latent) variables leading to reduced dimensionality and perform the least-square estimation based on these latent variables.

2.4. Latent variable methods

Characterization of pharmaceutical systems using common instrumental measurement methods produces multivariate, collinear data. Measured variables, which describe partially or fully the same property of a system, provide similar information content. Collinear variables can be combined and described by fewer, so-called factors or latent variables (LVs), which describe the underlying structure in the data. In modelling, the prime aim is to separate information from noise and find the crucial patterns in the data. The concept of factors or latent variables was first applied in psychology and provided the mathematical foundation for psychometrics (Horst, 1965, 1992; Thurstone, 1947). Following the introduction of computers and computerized measurement techniques, LV methodology has penetrated nearly all areas where complex systems are measured and modelled, and it is especially powerful when huge amounts of data are produced and systematic approaches are needed to reveal the information in the data.

2.4.1. Geometric presentation of a data matrix and latent variable projections

Data, for example acquired spectra, are arranged into a table (matrix) in such a way that each row represents one sample and each column one measured variable (e.g., a wavelength). Any matrix can be presented in two co-existing spaces, variable space and object space, which together contain all available information in a data matrix (Kvalheim, 1988). This is illustrated in Fig. 1. Each object (sample) *i* is described by the same *N* measured variables thus forming an object (row) vector, $\mathbf{x}_i^T$. Similarly, each variable *j* is described by its values for all the *M* objects, making up a variable (column) vector, $\mathbf{x}_j$. To visualize the data structure, object vectors can be plotted in variable space, where the number of axes is equal to the number of variables *N*. In this way all the information in $\mathbf{X}$ regarding the relationships (similarities or differences) between objects can be displayed. Similarly, variable vectors can be plotted in object space, where the number of axes is equal to the number of objects *M*. In this way the relationships (correlations or co-variances, depending on pretreatment) between variables can be quantitatively displayed. Since the object space shows common variation in a set of variables, it also displays the underlying factors or LVs. When the number of variables increases, the challenge is to find low-dimensional, information-rich projections of both variable and object space since the full spaces cannot be displayed and comprehended in a simple manner. This task can be achieved by projecting onto LVs. Different projections can be calculated using a generalization of the NIPALS algorithm (Box 1) (Kvalheim, 1987).

The score vector $\mathbf{t}_a$ and the loading vector $\mathbf{p}_a$ represent different presentations of the same LV, carrying information about samples in variable space and variables in object space, respectively (Fig. 1). The weight vector $\mathbf{w}_a$ defines the LV uniquely and any LV method can be derived from the definition of $\mathbf{w}_a$ (Box 2). Several criteria are available and used for decomposition of matrices, that is, to determine the axes for projections (Box 3). We shall discuss some of these below.

Box 1: Successive orthogonal projections.

(i) Select $\mathbf{w}_a$
(ii) Project objects on $\mathbf{w}_a$:

$$\mathbf{t}_a = \mathbf{X}_a\mathbf{w}_a$$

(iii) Project variable vectors on $\mathbf{t}_a$:

$$\mathbf{p}_a^T = \frac{\mathbf{t}_a^T\mathbf{X}_a}{\mathbf{t}_a^T\mathbf{t}_a}$$

(iv) Remove the latent-variable *a* from $\mathbf{X}_a$, i.e. substitute $\mathbf{X}_a$ with $\mathbf{X}_a - \mathbf{t}_a\mathbf{p}_a^T$.

Repeat (i)–(iv) for $a = 1, 2, \ldots, A$, where *A* is the dimension of the model. $\mathbf{X}_1 = \mathbf{X}$.

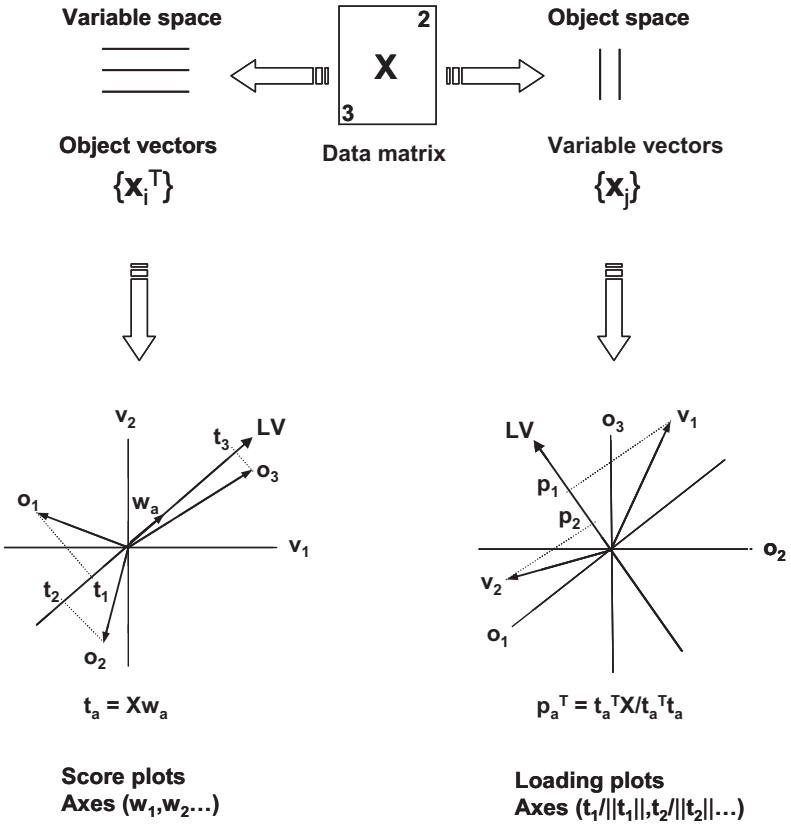

Fig. 1. The two alternative ways to look at a data matrix **X** and the principle of latent variable (LV) projections. Three vectors, $\mathbf{w}_a$, $\mathbf{t}_a$, and $\mathbf{p}_a$, are needed to define the LV in the two spaces (see algorithm in Box 1). Axes or vectors related to objects and variables are labelled with 'o' and 'v', respectively. In order to have a simple illustration, only three objects characterized by two variables are used.

Box 2: Method overview

PCA/SVD	$\mathbf{w}_a = \mathbf{p}_a/\|\|\mathbf{p}_a\|\|$
PLS	$\mathbf{w_a} = \mathbf{y_a}^T\mathbf{x_a}/\|\|\mathbf{y_a}^T\mathbf{x_a}\|\|$
TP	$\mathbf{w}_a = b/\|\|b\|\|$ (only one component)

PCA, principal component analysis; SVD, singular value decomposition; PLS, partial least squares; TP, target projection.

Box 3: Decomposition criteria

PCA ⇒ Maximum variance
PLS ⇒ Relevant components
TP ⇒ "Real" factors

2.4.2. Principal component analysis (PCA)

The oldest and most common latent variable projection method is principal component analysis (PCA) (Jackson, 1991; Wold et al., 1987). The data matrix **X** is decomposed into a number of principal components (PCs) that maximize explained variance in the data on each successive component under the constraint of being orthogonal to the previous PCs. The result is a bilinear model, a product of scores **T** and loadings **P** matrices:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} = \mathbf{t}_1\mathbf{p}_1^{\mathrm{T}} + \mathbf{t}_2\mathbf{p}_2^{\mathrm{T}} + \cdots + \mathbf{t}_A\mathbf{p}_A^{\mathrm{T}} + \mathbf{E} \quad (7)$$

X is an $M \times N$ matrix, consisting of M samples (rows) with N measured variables (columns). **T** is an $M \times A$ matrix and $\mathbf{P}^{\mathrm{T}}$ is an $A \times N$ matrix, where A is the number of calculated PCs. **T** and **P** consist of orthogonal and orthonormal vectors, respectively. **E** is an $M \times N$ matrix containing the residuals, that is, variance not explained by the PCs. Eq. (7) also shows the latent variable decomposition of **X** as a sum of products of score $\mathbf{t}_a$ and loading $\mathbf{p}_a$ vectors; $a = 1, 2, \ldots, A$. PCA is uniquely defined from the algorithm in Box 1 by using the constraint that the weights $\mathbf{w}_a$ are equal to the loadings $\mathbf{p}_a$. This is obtained by iterating steps (i)–(iii) until convergence, reducing the procedure to the traditional NIPALS algorithm (Horst, 1965; Wold et al., 1987).

PCA is a data visualization technique. Since each object gets a score value on each PC, objects can be presented in score plots. Score plots can reveal patterns, such as clusters, trends and outliers, in the data. In the same manner variables can be presented in loading plots, since each variable gets a loading value on each PC. Loading plots reveal covariances among variables and can be used to interpret patterns observed in the score plot. Together scores and loadings map the co-variance structure in the data. The maximum number of PCs is equal to min[M, N], but only the PCs that map the dominant variation patterns in the data are usually extracted. Noise is left in the residuals.

2.4.3. Principal component regression (PCR) and partial least squares (PLS) regression

One of the most common tasks in data analysis is to calculate a model which shows how one or several response variables, can be explained by means of a set of predictor variables. If the number of the x-variables is rather low and the x-variables are almost linearly independent (as in the case of DoE) and contain little noise compared to the noise in responses, MLR works well. In most pharmaceutical applications, however, the x-variables are correlated. This is always the case when working with spectral profiles.

A straightforward solution to the problem of collinear x-variables is to perform the regression using the PC scores, that is, principal component regression (PCR). This provides orthogonal

predictor variables which make the calculation of the inverse and thus the regression vector trivial, i.e. $\mathbf{b} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{y}$ where $(\mathbf{T}^T\mathbf{T})$ is a diagonal matrix of dimension A. Furthermore, by leaving out minor components, PCA provides noise reduction in $\mathbf{X}$ and thus the regression assumption of almost error-free predictor variables is obeyed.

A criticism against PCR is that the major principal components may model variation in the x-variables of little or no relevance to the y-variables. PLS regression was suggested as a modelling technique to overcome this problem (Geladi and Kowalski, 1986; Wold et al., 1984, 2001). Similar to PCA, PLS calculates a set of LVs leading to reduced dimensionality, but uses another criterion than maximum variance for the decomposition step. A normalized weight vector for PLS is calculated as the covariance between the response $\mathbf{y}$ and the data matrix $\mathbf{X}$:

$$\mathbf{w}_{\mathrm{PLS},1}^T = \frac{\mathbf{y}^T\mathbf{X}}{\left\|\mathbf{y}^T\mathbf{X}\right\|} \tag{8}$$

Scores and loadings for the PLS components are calculated successively by projecting the spectral variables $\mathbf{X}$ on $\mathbf{w}_{\mathrm{PLS},1}$ and by projecting $\mathbf{X}$ on the resulting score vectors as shown in Box 1. Each component is checked for predictive power by using some kind of cross validation (Bro et al., 2008; Filmoser et al., 2009; Smit et al., 2007; Stone, 1974; Wold, 1978). The part of $\mathbf{X}$ explained by a pair of PLS score and loading vectors in each step is removed before the next pair is calculated. The PLS decomposition can be written formally the same way as PCA (Eq. (7)) with a minor modification: the last PLS loading vector $\mathbf{p}_A$ has to be substituted by the corresponding PLS weight vector $\mathbf{w}_A$ (Pell et al., 2007). For PLS, the score vectors are orthogonal, while the loading vectors are neither orthogonal nor of unit length. However, since the PLS score vectors are orthogonal, PLS also leads to simple calculations in the inversion step and the regression vector.

For both PCA and PLS, the decomposition of $\mathbf{X}$ can be expressed as a product of three matrices:

$$\mathbf{X} = \mathbf{URW}^T \tag{9}$$

For PCA, $\mathbf{R}$ is a diagonal matrix with elements $||\mathbf{t}_a||$, $a = 1, 2, \ldots, A$, $\mathbf{U}$ is the matrix of normalized scores and $\mathbf{W}$ is the matrix of PCA weights which are identical to the loadings $\mathbf{P}$. This formulation of PCA is often referred to as the singular value decomposition (SVD). For PLS, $\mathbf{R}$ is bidiagonal matrix (Manne, 1987). From the formulation in Eq. (9), the generalized inverse for PLS can be expressed in terms of the original x-variables:

$$\mathbf{X}^+ = \mathbf{WR}^{-1}\mathbf{U}^T \tag{10}$$

Since both $\mathbf{W}$ and $\mathbf{U}$ are orthonormal and the bidiagonal matrix $\mathbf{R}$ is trivial to invert, see e.g. Kvalheim (1990), the general inverse and thus the regression coefficients for PLS expressed by the original x-variables can easily be calculated.

PLS regression can also be used as a supervised classification method. The response variable is then a binary vector of zeros and ones, describing the class membership for each sample in the investigated groups. The method is called PLS-discriminant analysis (PLS-DA) (Sjöström et al., 1986). Classification using latent variable methods is discussed in Section 2.4.6.

2.4.4. Methods for improved interpretation

PCR or PLS models can be used to predict the responses from x-variables such as spectral profiles. Unfortunately, whether PCR or PLS is used for modelling, numerous components are usually needed to describe the variation in $\mathbf{X}$. This makes interpretation of PCR and PLS models difficult since the information about the response is scattered between the components. Target projection (TP) and orthogonal PLS (O-PLS) are methods developed to circumvent this problem. During decomposition of $\mathbf{X}$, O-PLS first models the information in the x-variables orthogonal to the response, i.e. the so-called orthogonal components, and then calculate a predictive PLS component as the last step. TP projects the systematic information in the x-variables described by a PLS model onto the response variable to obtain a single latent variable (the target-projected component). The target-projected component represents the direction in the multivariate predictive space with strongest relation to the response for any given latent-variable decomposition (Kvalheim, 1990; Kvalheim and Karstang, 1989). Thus, TP represents the optimal way of relating a latent variable decomposition to a known target vector (response variable).

The regression vector $\mathbf{b}$, obtained from the PCR or PLS models, defines the direction in variable space with strongest relation to the response. Target-projected scores $\mathbf{t}_{\mathrm{TP}}$, proportional to the predicted response $\hat{y}$, and target-projected loadings $\mathbf{p}_{\mathrm{TP}}$ are obtained by using the normalized regression vector as weight vector, i.e. $\mathbf{w}_{\mathrm{TP}} = \mathbf{b}/||\mathbf{b}||$ in the algorithm in Box 1.

We can insert the score vector $\mathbf{t}_{\mathrm{TP}}$ in object space in Fig. 1. Since this vector is proportional to the vector of predicted responses $\hat{\mathbf{y}}$ and the TP loadings are the projections of the x-variable vectors onto this vector, TP loadings represent the features in the x-variables explaining and predicting the response variable. It follows that the TP loadings should be optimal for interpretations of the y-related predictive variation in $\mathbf{X}$, while many researchers wrongly use regression coefficients for interpretations. However, as discussed below, the TP loadings may not be the optimal choice for variable selection since they may be dominated by x-variables with high variance, but comparatively small correlation to the response.

After target projection a PCR or PLS model is reduced to a single-component TP model:

$$\mathbf{X} = \hat{\mathbf{X}}_{\mathrm{TP}} + \mathbf{E}_{\mathrm{TP}} = \mathbf{t}_{\mathrm{TP}}\mathbf{p}_{\mathrm{TP}}^T + \mathbf{E}_{\mathrm{TP}} \tag{11}$$

Ergon (2005) has developed an approach which in addition to the predictive y-related component obtained by TP also incorporates components describing the x-related or y-orthogonal variation in a PLS model.

The so-called O-PLS method (Trygg and Wold, 2002) represents another approach to obtain a single predictive latent variable. In this approach, the weights $\mathbf{w}_1$ for the first O-PLS component are selected as the difference vector $(\mathbf{p}_0 - \mathbf{w}_0)/||\mathbf{p}_0 - \mathbf{w}_0||$ where $\mathbf{w}_0$ and $\mathbf{p}_0$ are the weights and loadings respectively for the first component in a "standard" PLS regression. The y-orthogonal variation is then extracted as the first $(A-1)$ O-PLS component by updating the weight vector as $(\mathbf{p}_a - \mathbf{w}_0)/||\mathbf{p}_a - \mathbf{w}_0||$ and using the algorithm in Box 1. The predictive component A is calculated when either $\mathbf{X}$ is exhausted for orthogonal variation, meaning that $(\mathbf{p}_a - \mathbf{w}_0) \to 0$, or by cross validation in $\mathbf{X}$ to obtain the orthogonal components in $\mathbf{X}$ describing systematic variation.

With the same number of PLS components, the TP component is identical to the predictive component obtained from O-PLS. Thus, TP and O-PLS represent different algorithms to achieve the same goal (Kvalheim et al., 2009). A detailed analysis of interpretation of PLS/TP regression models and thus also the predictive O-PLS component can be found in a recent paper by Kvalheim (2010).

2.4.5. Variable selection in regression

A typical feature for data obtained from instrumental techniques (e.g., full spectral profiling) is that the number of objects is often very small compared to the number of variables (i.e., tables are "short and fat"). However, many of the variables are actually irrelevant as they represent variation not related to the investigated response. Therefore the number of variables can often be drastically reduced with minor loss of information. The challenge is to find the most significant variables. Variable selection methods aim at selecting a smaller panel of variables that are related to the

response variable and thus needed for a good predictive model. When a large number of variables are measured it is impossible to test all the variable combinations in question; for instance, there are 2.46×10^{20} (500!/(490!10!)) possible combinations to pick 10 variables out of 500. Variable selection strategies are therefore needed to search for appropriate combinations.

Univariate variable selection methods treat each variable (e.g., peaks in a spectral profile) independently. Statistical values are calculated for each variable after testing differences between profiles from two sample groups. t-statistics and analysis of variance (ANOVA) are methods often used for this purpose. However, these methods do not take into account collinearity in the data and they cannot handle properly the situation with few samples compared to the number of measured variables. It is relatively easy to get high correlation by pure chance and an irrelevant model will be created. In addition many of the traditional statistical tests assume that the data obey normal distribution, which is not always the case in real-life applications.

Several multivariate variable selection methods are available based on, for example, the co-variances between the response and each variable, i.e. the PLS weights (Hoskuldsson, 2001), size of regression coefficients (Centner et al., 1996), variable importance on projection (VIP) (Eriksson et al., 2001), interval PLS (Norgaard et al., 2000), and genetic algorithm (Lavine et al., 2004). All these methods have their weaknesses. Co-variance between the response and *x*-variables may be high because of high variance in *x*-variables. If *x*-variables are standardized to unit variance before regression, this pitfall is avoided, but this pretreatment procedure may enhance noise from minor *x*-variables in the model (Kvalheim, 1985). A selection based purely on size of regression coefficients may remove unimportant *x*-variables, but are also impacted by *x*-variables with high variance, but low correlation with the response. In addition, multicollinearity between *x*-variables and variation not related to the response (interference, orthogonal variation) introduces problems. VIP is also strongly influenced by orthogonal variation and therefore not useful for variable selection in regression situations. Wiklund et al. (2008) invented the so-called S-plot to cope with the situation. S-plot is a scatter plot showing covariance and correlation between the scores for the predictive O-PLS component and the spectral variables; the most important variables should have both high covariance and high correlation to the score on the predictive component. With many variables, the plot becomes crowded. Recently, we developed a visualization method called selectivity ratio (SR) for searching for important variables (Rajalahti et al., 2009a). The ratio between explained and residual (unexplained) variance for each variable in the TP model (or the predictive O-PLS component) defines an SR for the variable in question. The statistical significance of the SR method can be determined using e.g. a non-parametric test called the discriminating variable (DIVA) test (Rajalahti et al., 2009b). A nice feature of the SR plot is that it looks like a spectrum or chromatogram, but highlights the *x*-variables with strongest predictive ability and correlation to the response.

A recent tutorial by Andersen and Bro (2010) provides a practical guide to variable selection in regression-based calibration models.

2.4.6. Classification using latent variables

In unsupervised classification no *a priori* information about class membership for samples is used in the model building, that is, the modelling is based on *x*-variables only. PCA can be used as an unsupervised classification method. Soft independent modelling of class analogies (SIMCA) is a supervised classification technique based on PCA since it uses *a priori* information to split a data set into groups or classes of similar objects (Wold, 1976). Models for classes are calculated using the appropriate number of PCs determined from cross validation (with confidence intervals) and new samples are then projected onto the class models. Samples fitting inside the boundaries of a certain class can be assigned to that class. Samples outside confidence intervals are classified as outliers to that class.

PLS can be used as a supervised classification method, PLS discriminant analysis (PLS-DA). For binary classification a response vector can be created with values 1 or 0 according to the class membership of the samples and a PLS-DA model can be calculated. When new samples are measured and predicted using a PLS-DA model response values close to 1 or 0 should be obtained. For the binary case with a balanced number of samples in each group, the threshold 0.5 can be used to decide the class membership for the tested samples. The threshold can of course be varied from case to case since the optimal choice is problem and sample dependent. Balancing false positives against false negatives is often used as criterion for deciding the threshold. In multiclass problems two strategies are possible: either a single model, including all groups, or several binary models, modelling the groups pairwise.

2.5. Data pretreatment

There are many experimental and instrumental effects that are not related to compositional differences between samples and thus make comparison of profiles from different samples difficult. Examples of sources of variation are, for example, sample collection, sample preparation and instrumental artefacts. In order to remove these disturbing factors and ensure that collected spectra can be analyzed jointly, proper data pretreatment is necessary prior to data analysis. Pretreatment has a significant effect on the final results and should therefore be carefully considered. A good pretreatment procedure enhances the chemical/compositional information content in the data while a wrong pretreatment procedure destroys it by affecting the compositional correlation structure. Crucial factors affecting the data analysis depend on the analytical technique used and there is no single recipe that can be used for all data. Some relevant references are mentioned here but a thorough discussion of all possibilities needs a paper on its own.

Stordrange et al. (2002) compared different recipes for preprocessing NIR data using standard methods like normalization, differentiation and multiplicative scatter correction (MSC) (Geladi et al., 1985). In addition orthogonal signal correction (OSC) (Wold et al., 1998) and optimized scaling (OS) (Karstang and Manne, 1992) were tested. Chalus et al. (2005) compared the influence of standard normal variate (SNV) (Barnes et al., 1989), MSC, second derivative, and OSC (separately and combined) on NIR data. Luypaert et al. (2004) applied SNV, detrend correction, offset correction, and first and second derivation on the removal of spectral variations in NIR spectroscopy. Artursson et al. (2000) applied various preprocessing methods on data generated by X-ray powder diffraction. Several wavelet transforms, Fourier transform (FT), Savitzky–Golay (Savitzky and Golay, 1964), OSC, and combinations of wavelet transform and OSC, and FT and OSC were studied to enhance the predictive ability of PLS models.

A pretreatment strategy for mass spectral data that account for baseline effects, shifts in m/z values (alignment/synchronization problem), structured noise (heteroscedasticity), and differences in signal intensities caused by analytical workup and the instrumental technique (normalization problem) has been developed (Arneberg et al., 2007). Heteroscedasticity may seriously influence the correlation structure when samples have to be normalized and should be minimized before the normalization step (Kvalheim et al., 1994). Other important pretreatment steps to be considered are smoothing, such as methods for moving average and Savitsky–Golay, use of 1st and 2nd derivatives to remove background and data reduction using, for example, binning. Scaling of variables to unit variance not only enhances small signals at the expense of larger signals, but

also increases noise (Kvalheim, 1985). A better approach for mass spectral and chromatographic data is to reduce heteroscedasticity and influence of major signals simultaneously by the *n*th-root transform (Arneberg et al., 2007).

2.6. Validation of models

For the case of many more variables than objects, overfitting of models represents a serious pitfall (Brereton, 2006). Therefore, it is mandatory to check models for predictive performance. There are several options for validating models for predictive ability (Anderssen et al., 2006; Faber and Rajko, 2007; Westerhuis et al., 2008). Cross validation is the preferred method. Different procedures have been developed, but in all algorithms, the data are somehow partitioned into a training set and a validation set. For regression models, the validation is performed on the response (with the exception of O-PLS where validation of the orthogonal components has to be performed on **X**). An overview of some common methods for cross validation can be found in the paper by Bro et al. (2008). Another way of validating regression models is to randomly permute the response values and create a distribution of cross validated prediction estimates. The predictive performance of the "true" model should stand out compared with the null distribution from models with permuted responses.

3. Applications

Table 1 compiles some recent pharmaceutical applications where advanced characterization techniques are used in combination with multivariate data analysis methods.

3.1. Vibrational spectroscopy

IR, NIR and Raman spectroscopy have been used for many applications, such as qualitative and quantitative analysis of different pharmaceutical formulations, and monitoring of pharmaceutical processes. Quantification of active pharmaceutical ingredients (API) and excipients is a typical example.

NIR and Raman enable rapid and non-destructive measurements that can be performed remotely through optical fibres and no sampling is thus needed. Due to these factors these spectral techniques are particularly useful for process analysis and can be implemented in PAT. In addition, spectroscopic analysis of solids may offer probing of solid state properties such as crystallinity and sample density, parameters that are entirely lost by chromatography and other wet-chemistry methods (Johansson et al., 2002).

Spectroscopic techniques are usually employed in combination with multivariate data analysis methods. Usually a PLS calibration model is first built relating measured spectra to a reference technique. Validated model can then be used for on-line monitoring of the process and predicting for example API concentration in real-time. Identification of raw materials or intermediate products is also of interest in the pharmaceutical industry. Spectra from new compounds are compared with spectra of already approved compounds in e.g. NIR libraries and are then classified similar or dissimilar. Either unsupervised (PCA) or supervised (SIMCA, PLS-DA) classification methods can be utilized for this purpose.

A recent example of the successful use of NIR in PAT is the development, validation and transfer of a NIR method to determine the end point of a fluidized drying process by Peinado et al. (2011). Samples were taken from batches that were produced at full commercial scale and moisture content was measured with in-line NIR probe throughout the drying process. PLS regression (calibration model and validation) was employed as the multivariate method and robustness assessment was performed using PCA. The developed NIR method is currently implemented as a primary in-line method for controlling the drying end point in real-time for commercial production of solid oral-dose medicine. This has resulted in approx. 10% savings in energy efficiency and operational time for this particular process.

Extensive reviews on combined NIR spectroscopy and multivariate data analysis in pharmaceutical technology have been published (Luypaert et al., 2007; Reich, 2005; Roggo et al., 2007). Aaltonen et al. (2008) discussed topics related to spectroscopic analysis of pharmaceutical solids. One of the highlighted areas in this review was the importance of multivariate methods when using spectroscopic techniques. The use of Raman spectroscopy for quantitative analysis of pharmaceutical solids is reviewed in Strachan et al. (2007).

3.2. Imaging

Imaging techniques utilized in pharmaceutical applications vary from digital images (monochromatic or colour) to chemical imaging using spectroscopic techniques. Spectroscopic (hyperspectral) imaging techniques, in particular IR, NIR and Raman imaging, have become an attractive alternative because of instrumental development. A recent review on hyperspectral imaging of solid dosage forms was published by Amigo (2010).

Several multivariate exploratory and resolution methods can be applied to image analysis techniques to provide information about pure compounds in a sample (de Juan et al., 2004). Multivariate curve resolution (MCR) comprises a family of chemometric methods intended for the analysis of complex multicomponent systems and data produced with e.g. hyphenated instruments like GC/MS and LC/MS. These methods are not discussed in this paper and a good overview of the progress of MCR methods can be found elsewhere (de Juan and Tauler, 2006). Multiway analysis (N-way methods) is another family of multivariate methods often applied in image analysis and hyphenated instruments (Bro, 2006; Smilde et al., 2004). Different spectroscopic imaging techniques in pharmaceutical applications and the data analysis methods suitable for image analysis are reviewed by Gendrin et al. (2008).

3.3. Other characterization techniques

Other characterization techniques applied in pharmaceutics are for example, gas and liquid chromatography (GC and LC), mass spectrometry (MS), laser and X-ray diffractometry, and acoustic emission. In addition to laser diffractometry, also sieve analysis, spatial filtering technique and imaging are employed for particle size distribution measurement.

Hagsten et al. (2008) investigated 131 API batches to identify sources of batch to batch variation in the full scale processability by extrusion. Combination of low-pressure compression with particle size measurement provided a suitable tool for powder characterization. Particle size distributions were measured by laser diffractometry.

If the measured variables were evaluated separately none of them explained the batch to batch variation. Multivariate analysis by PCA revealed grouping of the batches according to their quality and the variables mainly contributing to this clustering could be detected. The amount of added granulation liquid reflected the investigated variation, and the batch quality was found out to be influenced by particle size, specific surface areas and packing behaviour.

A new strategy for revealing and ranking the bioactive components in natural products from chromatographic profiling was presented by Chau et al. (2009). The approach is based on PLS/TP analysis of the chromatographic profiles and utilizes selectivity ratios (SR) for the detection and ranking of the bioactive com-

ponents. This study represents a new way to analyze complex multicomponent samples from herbal formulations.

4. Conclusions

This tutorial review has given an introduction to multivariate data analysis methods commonly used in pharmaceutics in combination with advanced characterization techniques. As shown by several applications published in this area these methods are nowadays widely used in the pharmaceutical sciences and have a central role in PAT initiatives in the industry.

Since there is usually no trivial answer to a given data-analytical problem, the analyst should be able to recognise what is relevant and suitable for the given purpose. A recent paper by Kjeldahl and Bro (2010) discusses a number of common misunderstandings and pitfalls in practical multivariate data analysis that one should be aware of. Among these are for example, selection of relevant samples and variables, diagnosis and interpretation of the models, and the use of software packages. Pitfalls when using PLS regression in NIR applications are also discussed by Xiang et al. (2009).

Acknowledgement

University of Bergen is thanked for the grant (open researcher initiated projects) to Olav M. Kvalheim.

References

Aaltonen, J., Gordon, K.C., Strachan, C.J., Rades, T., 2008. Perspectives in the use of spectroscopy to characterise pharmaceutical solids. Int. J. Pharm. 364, 159–169.
Amigo, J.M., 2010. Practical issues of hyperspectral imaging analysis of solid dosage forms. Anal. Bioanal. Chem. 398, 93–109.
Amigo, J.M., Ravn, C., 2009. Direct quantification and distribution assessment of major and minor components in pharmaceutical tablets by NIR-chemical imaging. Eur. J. Pharm. Sci. 37, 76–82.
Andersen, C.M., Bro, R., 2010. Variable selection in regression—a tutorial. J. Chemometr. 24, 728–737.
Anderssen, E., Dyrstad, K., Westad, F., Martens, H., 2006. Reducing over-optimism in variable selection by cross-model validation. Chemometr. Intell. Lab. Syst. 84, 69–74.
Andersson, M., Svensson, O., Folestad, S., Josefson, M., Wahlund, K.G., 2005. NIR spectroscopy on moving solids using a scanning grating spectrometer—impact on multivariate process analysis. Chemometr. Intell. Lab. Syst. 75, 1–11.
Arneberg, R., Rajalahti, T., Flikka, K., Berven, F.S., Kroksveen, A.C., Berle, M., Myhr, K.M., Vedeler, C.A., Ulvik, R.J., Kvalheim, O.M., 2007. Pretreatment of mass spectral profiles: application to proteomic data. Anal. Chem. 79, 7014–7026.
Artursson, T., Hagman, A., Bjork, S., Trygg, J., Wold, S., Jacobsson, S.P., 2000. Study of preprocessing methods for the determination of crystalline phases in binary mixtures of drug substances by X-ray powder diffraction and multivariate calibration. Appl. Spectrosc. 54, 1222–1230.
Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Appl. Spectrosc. 43, 772–777.
Benedetti, C., Abatzoglou, N., Simard, J.S., McDermott, L., Leonarda, G., Cartilier, L., 2007. Cohesive, multicomponent, dense powder flow characterization by NIR. Int. J. Pharm. 336, 292–301.
Bergman, E.L., Brage, H., Josefson, M., Svensson, O., Sparen, A., 2006. Transfer of NIR calibrations for pharmaceutical formulations between different instruments. J. Pharm. Biomed. Anal. 41, 89–98.
Blanco, M., Alcala, M., Gonzalez, J.M., Torras, E., 2006. Near infrared spectroscopy in the study of polymorphic transformations. Anal. Chim. Acta 567, 262–268.
Blanco, M., Valdes, D., Bayod, M.S., Fernandez-Mari, F., Llorente, I., 2004. Characterization and analysis of polymorphs by near-infrared spectrometry. Anal. Chim. Acta 502, 221–227.
Box, G.E.P., Hunter, W.G., Hunter, J.S., 1978. Statistics for Experimenters. John Wiley & Sons, New York.
Braun, D.E., Maas, S.G., Zencirci, N., Langes, C., Urbanetz, N.A., Griesser, U.J., 2010. Simultaneous quantitative analysis of ternary mixtures of D-mannitol polymorphs by FT-Raman spectroscopy and multivariate calibration models. Int. J. Pharm. 385, 29–36.
Brereton, R.G., 2006. Consequences of sample size, variable selection, and model validation and optimisation, for predicting classification ability from analytical data. TrAC: Trends Anal. Chem. 25, 1103–1111.
Bro, R., 2006. Review on multiway analysis in chemistry—2000–2005. Crit. Rev. Anal. Chem. 36, 279–293.
Bro, R., Kjeldahl, K., Smilde, A.K., Kiers, H.A.L., 2008. Cross-validation of component models: a critical look at current methods. Anal. Bioanal. Chem. 390, 1241–1251.
Callis, J.B., Illman, D.L., Kowalski, B.R., 1987. Process analytical-chemistry. Anal. Chem. 59, A624.
Centner, V., Massart, D.L., deNoord, O.E., deJong, S., Vandeginste, B.M., Sterna, C., 1996. Elimination of uninformative variables for multivariate calibration. Anal. Chem. 68, 3851–3858.
Chalus, P., Roggo, Y., Walter, S., Ulmschneider, M., 2005. Near-infrared determination of active substance content in intact low-dosage tablets. Talanta 66, 1294–1302.
Chau, F.T., Chan, H.Y., Cheung, C.Y., Xu, C.J., Liang, Y., Kvalheim, O.M., 2009. Recipe for uncovering the bioactive components in herbal medicine. Anal. Chem. 81, 7217–7225.
De Beer, T.R.M., Alleso, M., Goethals, F., Coppens, A., Heyden, Y.V., De Diego, H.L., Rantanen, J., Verpoort, F., Vervaet, C., Remon, J.P., Baeyens, W.R.G., 2007. Implementation of a process analytical technology system in a freeze-drying process using Raman spectroscopy for in-line process monitoring. Anal. Chem. 79, 7992–8003.
De Beer, T.R.M., Vercruysse, P., Burggraeve, A., Quinten, T., Ouyang, J., Zhang, X., Vervaet, C., Remon, J.P., Baeyens, W.R.G., 2009. In-line and real-time process monitoring of a freeze drying process using Raman and NIR spectroscopy as complementary process analytical technology (PAT) tools. J. Pharm. Sci. 98, 3430–3446.
de Juan, A., Tauler, R., 2006. Multivariate curve resolution (MCR) from 2000: progress in concepts and applications. Crit. Rev. Anal. Chem. 36, 163–176.
de Juan, A., Tauler, R., Dyson, R., Marcolli, C., Rault, M., Maeder, M., 2004. Spectroscopic imaging and chemometrics: a powerful combination for global and local sample analysis. TrAC: Trends Anal. Chem. 23, 70–79.
de Veij, M., Vandenabeele, P., Hall, K.A., Fernandez, F.M., Green, M.D., White, N.J., Dondorp, A.M., Newton, P.N., Moens, L., 2007. Fast detection and identification of counterfeit antimalarial tablets by Raman spectroscopy. J. Raman Spectrosc. 38, 181–187.
El-Hagrasy, A.S., D'Amico, F., Drennen, J.K., 2006a. A process analytical technology approach to near-infrared process control of pharmaceutical powder blending. Part I. D-optimal design for characterization of powder mixing and preliminary spectral data evaluation. J. Pharm. Sci. 95, 392–406.
El-Hagrasy, A.S., Delgado-Lopez, M., Drennen, J.K., 2006b. A process analytical technology approach to near-infrared process control of pharmaceutical powder blending. Part II. Qualitative near-infrared models for prediction of blend homogeneity. J. Pharm. Sci. 95, 407–421.
El-Hagrasy, A.S., Drennen, J.K., 2006. A process analytical technology approach to near-infrared process control of pharmaceutical powder blending. Part III. Quantitative near-infrared calibration for prediction of blend homogeneity and characterization of powder mixing kinetics. J. Pharm. Sci. 95, 422–434.
Ergon, R., 2005. PLS post-processing by similarity transformation (PLS plus ST): a simple alternative to OPLS. J. Chemometr. 19, 1–4.
Eriksson, L., Johansson, E., Kettaneh-Wold, N., Wold, S., 2001. Multi- and Megavariate Data Analysis: Principles and Applications. Umetrics Academy, Umeå, Sweden.
Faber, N.M., Rajko, R., 2007. How to avoid over-fitting in multivariate calibration—the conventional validation approach and an alternative. Anal. Chim. Acta 595, 98–106.
Filmoser, P., Liebmann, B., Varmuza, K., 2009. Repeated double cross validation. J. Chemometr. 23, 160–171.
Fransson, M., Johansson, J., Sparén, A., Svensson, O., 2010. Comparison of multivariate methods for quantitative determination with transmission Raman spectroscopy in pharmaceutical formulations. J. Chemometr. 24, 674–680.
Gabrielsson, J., Lindberg, N.O., Lundstedt, T., 2002. Multivariate methods in pharmaceutical applications. J. Chemometr. 16, 141–160.
Gabrielsson, J., Sjostrom, M., Lindberg, N.O., Pihl, A.C., Lundstedt, T., 2006a. Multivariate methods in the development of a new tablet formulation: excipient mixtures and principal properties. Drug Dev. Ind. Pharm. 32, 7–20.
Gabrielsson, J., Sjostrom, M., Lindberg, N.O., Pihl, A.C., Lundstedt, T., 2006b. Robustness testing of a tablet formulation using multivariate design. Drug Dev. Ind. Pharm. 32, 297–307.
García-Muñoz, S., Carmody, A., 2010. Multivariate wavelet texture analysis for pharmaceutical solid product characterization. Int. J. Pharm. 398, 97–106.
García-Muñoz, S., Gierer, D.S., 2010. Coating uniformity assessment for colored immediate release tablets using multivariate image analysis. Int. J. Pharm. 395, 104–113.
Geladi, P., Kowalski, B.R., 1986. Partial least-squares regression—a tutorial. Anal. Chim. Acta 185, 1–17.
Geladi, P., Macdougall, D., Martens, H., 1985. Linearization and scatter-correction for near-infrared reflectance spectra of meat. Appl. Spectrosc. 39, 491–500.
Gendrin, C., Roggo, Y., Collet, C., 2007. Content uniformity of pharmaceutical solid dosage forms by near infrared hyperspectral imaging: a feasibility study. Talanta 73, 733–741.
Gendrin, C., Roggo, Y., Collet, C., 2008. Pharmaceutical applications of vibrational chemical imaging and chemometrics: a review. J. Pharm. Biomed. Anal. 48, 533–553.
Grohganz, H., Fonteyne, M., Skibsted, E., Falck, T., Palmqvist, B., Rantanen, J., 2009. Role of excipients in the quantification of water in lyophilised mixtures using NIR spectroscopy. J. Pharm. Biomed. Anal. 49, 901–907.
Guenette, E., Barrett, A., Kraus, D., Brody, R., Harding, L., Magee, G., 2009. Understanding the effect of lactose particle size on the properties of DPI formulations using experimental design. Int. J. Pharm. 380, 80–88.

Hagsten, A., Larsen, C.C., Sonnergaard, J.M., Rantanen, J., Hovgaard, L., 2008. Identifying sources of batch to batch variation in processability. Powder Technol. 183, 213–219.
Haware, R.V., Tho, I., Bauer-Brandl, A., 2009. Application of multivariate methods to compression behavior evaluation of directly compressible materials. Eur. J. Pharm. Biopharm. 72, 148–155.
Horst, P., 1965. Factor Analysis of Data Matrices. Holt, Rinehart and Winston, Inc, New York.
Horst, P., 1992. Sixty years with latent variables and still more to come. Chemometr. Intell. Lab. Syst. 14, 5–21.
Hoskuldsson, A., 2001. Variable and subset selection in PLS regression. Chemometr. Intell. Lab. Syst. 55, 23–38.
Haag, M., Bruning, M., Molt, K., 2009. Quantitative analysis of diphenhydramine hydrochloride in pharmaceutical wafers using near infrared and Raman spectroscopy. Anal. Bioanal. Chem. 395, 1777–1785.
Jackson, J.E., 1991. A Users' Guide to Principal Components. Wiley, New York.
Johansson, J., Folestad, S., Josefson, M., Sparen, A., Abrahamsson, C., Andersson-Engels, S., Svanberg, S., 2002. Time-resolved NIR/Vis spectroscopy for analysis of solids: pharmaceutical tablets. Appl. Spectrosc. 56, 725–731.
Johansson, J., Pettersson, S., Folestad, S., 2005. Characterization of different laser irradiation methods for quantitative Raman tablet assessment. J. Pharm. Biomed. Anal. 39, 510–516.
Johansson, J., Sparen, A., Svensson, O., Folestad, S., Claybourn, M., 2007. Quantitative transmission Raman spectroscopy of pharmaceutical tablets and capsules. Appl. Spectrosc. 61, 1211–1218.
Jorgensen, A.C., Miroshnyk, I., Karjalainen, M., Jouppila, K., Siiria, S., Antikainen, O., Rantanen, J., 2006. Multivariate data analysis as a fast tool in evaluation of solid state phenomena. J. Pharm. Sci. 95, 906–916.
Karstang, T.V., Manne, R., 1992. Optimized scaling—a novel-approach to linear calibration with closed data sets. Chemometr. Intell. Lab. Syst. 14, 165–173.
Kauffman, J.F., Dellibovi, M., Cunningham, C.R., 2007. Raman spectroscopy of coated pharmaceutical tablets and physical models for multivariate calibration to tablet coating thickness. J. Pharm. Biomed. Anal. 43, 39–48.
Kim, J., Noh, J., Chung, H., Woo, Y.A., Kemper, M.S., Lee, Y., 2007a. Direct, non-destructive quantitative measurement of an active pharmaceutical ingredient in an intact capsule formulation using Raman spectroscopy. Anal. Chim. Acta 598, 280–285.
Kim, M., Chung, H., Woo, Y., Kemper, M.S., 2007b. A new non-invasive, quantitative Raman technique for the determination of an active ingredient in pharmaceutical liquids by direct measurement through a plastic bottle. Anal. Chim. Acta 587, 200–207.
Kjeldahl, K., Bro, R., 2010. Some common misunderstandings in chemometrics. J. Chemometr. 24, 558–564.
Kogermann, K., Aaltonen, J., Strachan, C.J., Pollanen, K., Veski, P., Heinamaki, J., Yliruusi, J., Rantanen, J., 2007. Qualitative in situ analysis of multiple solid-state forms using spectroscopy and partial least squares discriminant modeling. J. Pharm. Sci. 96, 1802–1820.
Kourti, T., 2006. Process analytical technology beyond real-time analyzers: the role of multivariate analysis. Crit. Rev. Anal. Chem. 36, 257–278.
Kvalheim, O.M., 1985. Scaling of analytical data. Anal. Chim. Acta 177, 71–79.
Kvalheim, O.M., 1987. Latent-structure decompositions (projections) of multivariate data. Chemometr. Intell. Lab. Syst. 2, 283–290.
Kvalheim, O.M., 1988. Interpretation of direct latent-variable projection methods and their aims and use in the analysis of multicomponent spectroscopic and chromatographic data. Chemometr. Intell. Lab. Syst. 4, 11–25.
Kvalheim, O.M., 1990. Latent-variable regression-models with higher-order terms—an extension of response modeling by orthogonal design and multiple linear-regression. Chemometr. Intell. Lab. Syst. 8, 59–67.
Kvalheim, O.M., 2010. Interpretation of partial least squares regression models by means of target projection and selectivity ratio plots. J. Chemometr. 24, 496–504.
Kvalheim, O.M., Brakstad, F., Liang, Y.Z., 1994. Preprocessing of analytical profiles in the presence of homoscedastic or heteroscedastic noise. Anal. Chem. 66, 43–51.
Kvalheim, O.M., Karstang, T.V., 1989. Interpretation of latent-variable regression models. Chemometr. Intell. Lab. Syst. 7, 39–51.
Kvalheim, O.M., Rajalahti, T., Arneberg, R., 2009. X-tended target projection (XTP)-comparison with orthogonal partial least squares (OPLS) and PLS post-processing by similarity transformation (PLS plus ST). J. Chemometr. 23, 49–55.
Laitinen, N., Antikainen, O., Rantanen, J., Yliruusi, J., 2004. New perspectives for visual characterization of pharmaceutical solids. J. Pharm. Sci. 93, 165–176.
Laursen, K., Frederiksen, S.S., Leuenhagen, C., Bro, R., 2010. Chemometric quality control of chromatographic purity. J. Chromatogr. A 1217, 6503–6510.
Lavine, B.K., Davidson, C.E., Rayens, W.S., 2004. Machine learning based pattern recognition applied to microarray data. Comb. Chem. High Throughput Screen. 7, 115–131.
Lee, M.-J., Seo, D.-Y., Lee, H.-E., Wang, I.-C., Kim, W.-S., Jeong, M.-Y., Choi, G.J., 2011. In line NIR quantification of film thickness on pharmaceutical pellets during a fluid bed coating process. Int. J. Pharm. 403, 66–72.
Lopes, J.A., Costa, P.F., Alves, T.P., Menezes, J.C., 2004. Chemometrics in bioprocess engineering: process analytical technology (PAT) applications. Chemometr. Intell. Lab. Syst. 74, 269–275.
Lopes, M.B., Wolff, J.C., Bioucas-Dias, J.M., Figueiredo, M.A.T., 2010. Near-infrared hyperspectral unmixing based on a minimum volume criterion for fast and accurate chemometric characterization of counterfeit tablets. Anal. Chem. 82, 1462–1469.
Lopez-Arellano, R., Santander-Garcia, E.A., Andrade-Garda, J.M., Alvarez-Avila, G., Garduno-Rosas, J.A., Morales-Hipolito, E.A., 2009. Quantification of lysine clonixinate in intravenous injections by NIR spectroscopy. Vib. Spectrosc. 51, 255–262.
Lundstedt-Enkel, K., Gabrielsson, J., Olsman, H., Seifert, E., Pettersen, J., Lek, P.M., Boman, A., Lundstedt, T., 2006. Different multivariate approaches to material discovery, process development, PAT and environmental process monitoring. Chemometr. Intell. Lab. Syst. 84, 201–207.
Lundstedt, T., Seifert, E., Abramo, L., Thelin, B., Nystrom, A., Pettersen, J., Bergman, R., 1998. Experimental design and optimization. Chemometr. Intell. Lab. Syst. 42, 3–40.
Luypaert, J., Heuerding, S., Vander Heyden, Y., Massart, D.L., 2004. The effect of preprocessing methods in reducing interfering variability from near-infrared measurements of creams. J. Pharm. Biomed. Anal. 36, 495–503.
Luypaert, J., Massart, D.L., Heyden, Y.V., 2007. Near-infrared spectroscopy applications in pharmaceutical analysis. Talanta 72, 865–883.
Ma, H., Anderson, C.A., 2008. Characterization of pharmaceutical powder blends by NIR chemical imaging. J. Pharm. Sci. 97, 3305–3320.
MacGregor, J.F., Kourti, T., 1995. Statistical process control of multivariate processes. Control Eng. Pract. 3, 403–414.
Maggio, R.M., Castellano, P.M., Kaufman, T.S., 2009. PCA-CR analysis of dissolution profiles. A chemometric approach to probe the polymorphic form of the active pharmaceutical ingredient in a drug product. Int. J. Pharm. 378, 187–193.
Mandenius, C.F., Brundin, A., 2008. Bioprocess optimization using design-of-experiments methodology. Biotechnol. Prog. 24, 1191–1203.
Manne, R., 1987. Analysis of 2 partial-least-squares algorithms for multivariate calibration. Chemometr. Intell. Lab. Syst. 2, 187–197.
Mantanus, J., Ziemons, E., Lebrun, P., Rozet, E., Klinkenberg, R., Streel, B., Evrard, B., Hubert, P., 2009. Moisture content determination of pharmaceutical pellets by near infrared spectroscopy: method development and validation. Anal. Chim. Acta 642, 186–192.
Mantanus, J., Ziémons, E., Lebrun, P., Rozet, E., Klinkenberg, R., Streel, B., Evrard, B., Hubert, P., 2010a. Active content determination of non-coated pharmaceutical pellets by near infrared spectroscopy: method development, validation and reliability evaluation. Talanta 80, 1750–1757.
Mantanus, J., Ziémons, E., Rozet, E., Streel, B., Klinkenberg, R., Evrard, B., Rantanen, J., Hubert, P., 2010b. Building the quality into pellet manufacturing environment—feasibility study and validation of an in-line quantitative near infrared (NIR) method. Talanta 83, 305–311.
Matero, S., Poutiainen, S., Leskinen, J., Järvinen, K., Ketolainen, J., Poso, A., Reinikainen, S.P., 2010. Estimation of granule size distribution for batch fluidized bed granulation process using acoustic emission and N-way PLS. J. Chemometr. 24, 464–471.
Moes, J.J., Ruijken, M.M., Gout, E., Frijlink, H.W., Ugwoke, M.I., 2008. Application of process analytical technology in tablet process development using NIR spectroscopy: blend uniformity, content uniformity and coating thickness measurements. Int. J. Pharm. 357, 108–118.
Müller, J., Knop, K., Thies, J., Uerpmann, C., Kleinebudde, P., 2010. Feasibility of Raman spectroscopy as PAT tool in active coating. Drug Dev. Ind. Pharm. 36, 234–243.
Märk, J., Andre, M., Karner, M., Huck, C.W., 2010. Prospects for multivariate classification of a pharmaceutical intermediate with near-infrared spectroscopy as a process analytical technology (PAT) production control supplement. Eur. J. Pharm. Biopharm. 76, 320–327.
Norgaard, L., Saudland, A., Wagner, J., Nielsen, J.P., Munck, L., Engelsen, S.B., 2000. Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy. Appl. Spectrosc. 54, 413–419.
Närvänen, T., Antikainen, O., Yliruusi, J., 2009. Predicting particle size during fluid bed granulation using process measurement data. AAPS PharmSciTech 10, 1268–1275.
Park, S.C., Kim, M., Noh, J., Chung, H., Woo, Y., Lee, J., Kemper, M.S., 2007. Reliable and fast quantitative analysis of active ingredient in pharmaceutical suspension using Raman spectroscopy. Anal. Chim. Acta 593, 46–53.
Peinado, A., Hammond, J., Scott, A., 2011. Development, validation and transfer of a near infrared method to determine in-line the end point of a fluidised drying process for commercial production batches of an approved oral solid dose pharmaceutical product. J. Pharm. Biomed. Anal. 54, 13–20.
Pell, R.J., Ramos, L.S., Manne, R., 2007. The model space in partial least squares regression. J. Chemometr. 21, 165–172.
Pöllänen, K., Hakkinen, A., Reinikainen, S.P., Rantanen, J., Karjalainen, M., Louhi-Kultanen, M., Nystrom, L., 2005. IR spectroscopy together with multivariate data analysis as a process analytical tool for in-line monitoring of crystallization process and solid-state analysis of crystalline product. J. Pharm. Biomed. Anal. 38, 275–284.
Rajalahti, T., Arneberg, R., Berven, F.S., Myhr, K.M., Ulvik, R.J., Kvalheim, O.M., 2009a. Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. Chemometr. Intell. Lab. Syst. 95, 35–48.
Rajalahti, T., Arneberg, R., Kroksveen, A.C., Berle, M., Myhr, K.M., Kvalheim, O.M., 2009b. Discriminating variable test and selectivity ratio plot: quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles. Anal. Chem. 81, 2581–2590.
Rantanen, J., Wikstrom, H., Turner, R., Taylor, L.S., 2005. Use of in-line near-infrared spectroscopy in combination with chemometrics for improved understanding of pharmaceutical processes. Anal. Chem. 77, 556–563.
Reich, G., 2005. Near-infrared spectroscopy and imaging: basic principles and pharmaceutical applications. Adv. Drug Deliv. Rev. 57, 1109–1143.

Roggo, Y., Chalus, P., Maurer, L., Lema-Martinez, C., Edmond, A., Jent, N., 2007. A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. J. Pharm. Biomed. Anal. 44, 683–700.
Roggo, Y., Degardin, K., Margot, P., 2010. Identification of pharmaceutical tablets by Raman spectroscopy and chemometrics. Talanta 81, 988–995.
Roggo, Y., Edmond, A., Chalus, P., Ulmschneider, M., 2005. Infrared hyperspectral imaging for qualitative analysis of pharmaceutical solid forms. Anal. Chim. Acta 535, 79–87.
Romero-Torres, S., Perez-Ramos, J.D., Morris, K.R., Grant, E.R., 2005. Raman spectroscopic measurement of tablet-to-tablet coating variability. J. Pharm. Biomed. Anal. 38, 270–274.
Romero-Torres, S., Perez-Ramos, J.D., Morris, K.R., Grant, E.R., 2006. Raman spectroscopy for tablet coating thickness quantification and coating characterization in the presence of strong fluorescent interference. J. Pharm. Biomed. Anal. 41, 811–819.
Russeau, W., Mitchell, J., Tetteh, J., Lane, M.E., Hadgraft, J., 2009. Investigation of the permeation of model formulations and a commercial ibuprofen formulation in Carbosil® and human skin using ATR-FTIR and multivariate spectral analysis. Int. J. Pharm. 374, 17–25.
Sandler, N., Wilson, D., 2010. Prediction of granule packing and flow behavior based on particle size and shape analysis. J. Pharm. Sci. 99, 958–968.
Sarraguça, M.C., Lopes, J.A., 2009. Quality control of pharmaceuticals with NIR: from lab to process line. Vib. Spectrosc. 49, 204–210.
Savitzky, A., Golay, M.J.E., 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36, 1627–1639.
Shah, R.B., Tawakkul, M.A., Khan, M.A., 2007. Process analytical technology: chemometric analysis of Raman and near infra-red spectroscopic data for predicting physical properties of extended release matrix tablets. J. Pharm. Sci. 96, 1356–1365.
Shi, S., Ashley, E.S., Alexander, B.D., Hickey, A.J., 2009. Initial characterization of micafungin pulmonary delivery via two different nebulizers and multivariate data analysis of aerosol mass distribution profiles. AAPS PharmSciTech 10, 129–137.
Shi, Z.Q., Cogdill, R.P., Short, S.M., Anderson, C.A., 2008. Process characterization of powder blending by near-infrared spectroscopy: blend end-points and beyond. J. Pharm. Biomed. Anal. 47, 738–745.
Sjöström, M., Wold, S., Söderström, B., 1986. PLS discriminant plots. In: Pattern Recognition in Practice II. Elsevier Science Publ. B.V., Holland, pp. 461–470.
Smilde, A.K., Bro, R., Geladi, P., 2004. Multi-way Analysis with Applications in the Chemical Sciences. Wiley, Chichester.
Smit, S., van Breemen, M.J., Hoefsloot, H.C.J., Smilde, A.K., Aerts, J., de Koster, C.G., 2007. Assessing the statistical validity of proteomics based biomarkers. Anal. Chim. Acta 592, 210–217.
Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc. Ser. B: Methodol. 36, 111–147.
Stordrange, L., Libnau, F.L., Malthe-Sorenssen, D., Kvalheim, O.M., 2002. Feasibility study of NIR for surveillance of a pharmaceutical process, including a study of different preprocessing techniques. J. Chemometr. 16, 529–541.
Stordrange, L., Rajalahti, T., Libnau, F.O., 2004. Multiway methods to explore and model NIR data from a batch process. Chemometr. Intell. Lab. Syst. 70, 137–145.
Strachan, C.J., Rades, T., Gordon, K.C., Rantanen, J., 2007. Raman spectroscopy for quantitative analysis of pharmaceutical solids. J. Pharm. Pharmacol. 59, 179–192.
Svensson, O., Abrahamsson, K., Engelbrektsson, J., Nicholas, M., Wikstrom, H., Josefson, M., 2006. An evaluation of 2D-wavelet filters for estimation of differences in textures of pharmaceutical tablets. Chemometr. Intell. Lab. Syst. 84, 3–8.
Thurstone, L.L., 1947. Multiple Factor Analysis. University of Chicago Press, Chicago, IL.
Trygg, J., Wold, S., 2002. Orthogonal projections to latent structures (O-PLS). J. Chemometr. 16, 119–128.
United States Food and Drug Administration (FDA), 2004. Guidance for Industry: PAT—A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance.
Vanarase, A.U., Alcala, M., Rozo, J.I.J., Muzzio, F.J., Romanach, R.J., 2010. Real-time monitoring of drug concentration in a continuous powder mixing process using NIR spectroscopy. Chem. Eng. Sci. 65, 5728–5733.
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., 2008. Assessment of PLSDA cross validation. Metabolomics 4, 81–89.
Wiklund, S., Johansson, E., Sjostrom, L., Mellerowicz, E.J., Edlund, U., Shockcor, J.P., Gottfries, J., Moritz, T., Trygg, J., 2008. Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models. Anal. Chem. 80, 115–122.
Wikström, H., Romero-Torres, S., Wongweragiat, S., Williams, J.A.S., Grant, E.R., Taylor, L.S., 2006. On-line content uniformity determination of tablets using low-resolution Raman spectroscopy. Appl. Spectrosc. 60, 672–681.
Wold, S., 1976. Pattern-recognition by means of disjoint principal components models. Pattern Recognit. 8, 127–139.
Wold, S., 1978. Cross-validatory estimation of number of components in factor and principal components models. Technometrics 20, 397–405.
Wold, S., Antti, H., Lindgren, F., Ohman, J., 1998. Orthogonal signal correction of near-infrared spectra. Chemometr. Intell. Lab. Syst. 44, 175–185.
Wold, S., Esbensen, K., Geladi, P., 1987. Principal component analysis. Chemometr. Intell. Lab. Syst. 2, 37–52.
Wold, S., Ruhe, A., Wold, H., Dunn, W.J., 1984. The collinearity problem in linear-regression—the partial least-squares (PLS) approach to generalized inverses. SIAM J. Sci. Stat. Comput. 5, 735–743.
Wold, S., Sjostrom, M., 1998. Chemometrics, present and future success. Chemometr. Intell. Lab. Syst. 44, 3–14.
Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemometr. Intell. Lab. Syst. 58, 109–130.
Wu, H., Tawakkul, M., White, M., Khan, M.A., 2009. Quality-by-Design (QbD): an integrated multivariate approach for the component quantification in powder blends. Int. J. Pharm. 372, 39–48.
Xiang, D., Berry, J., Buntz, S., Gargiulo, P., Cheney, J., Joshi, Y., Wabuyele, B., Wu, H.Q., Hamed, M., Hussain, A.S., Khan, M.A., 2009. Robust calibration design in the pharmaceutical quantitative measurements with near-infrared (NIR) spectroscopy: avoiding the chemometric pitfalls. J. Pharm. Sci. 98, 1155–1166.
Yu, L., 2008. Pharmaceutical quality by design: product and process development, understanding, and control. Pharm. Res. 25, 781–791.
Zhang, L., Henson, M.J., Sekulic, S.S., 2005. Multivariate data analysis for Raman imaging of a model pharmaceutical tablet. Anal. Chim. Acta 545, 262–278.
Ziemons, E., Mantanus, J., Lebrun, P., Rozet, E., Evrard, B., Hubert, P., 2010. Acetaminophen determination in low-dose pharmaceutical syrup by NIR spectroscopy. J. Pharm. Biomed. Anal. 53, 510–516.
Zomer, S., Brereton, R.G., Wolff, J.C., Airiau, C.Y., Smallwood, C., 2005. Component detection weighted index of analogy: similarity recognition on liquid chromatographic mass spectral data for the characterization of route/process specific impurities in pharmaceutical tablets. Anal. Chem. 77, 1607–1621.