Journal of Pharmaceutical & Biomedical Analysis
Vol. 9, No. 8, pp. 637–641, 1991
Printed in Great Britain

0731–7085/91 $3.00 + 0.00

Mathematical isolation of component spectra in HPLC/UV–vis and GC–MS. How unique are the resolved spectra?*

ERKKI J. KARJALAINEN

Department of Clinical Chemistry, University of Helsinki, Haartmaninkatu 4, SF-00290 Helsinki, Finland

Abstract: The resolution of overlapping spectra in GC–MS and HPLC/UV–vis is fundamentally limited by the quality of the experimental data. The narrowness of the solution range depends on the degree of overlap between components. If the components are dissimilar, the solutions obtained by all mathematical methods are robust. Small perturbations in the observations do not change the calculated solution very much. Alternating regression (AR) is a useful tool in the analysis of overlapping spectra because AR can be calculated very rapidly. The robustness of the solution can be easily checked with AR. The mathematical analysis is repeated several times after adding different sets of noise. Each time different random spectra are used as a starting point. The range of solutions thus obtained reflects the quality of the data for resolution purposes.

Keywords: *Alternating regression; deconvolution of spectra; spectrum reconstruction; GC–MS; HPLC/UV–vis.*

Introduction

Biological samples are often very complex mixtures. Analysis of these mixtures with hyphenated instruments produces large two-dimensional observation matrices. It is not possible to extract the full information hidden in these data matrices because the conceptual and mathematical tools required are not fully developed.

Ideally, hyphenated instruments like GC–MS and HPLC/UV–vis should produce pure spectra of all sample components. In practice raw spectra are produced that are mixtures of several compounds. In the case of unknown and novel compounds the situation is even more unsatisfactory. Spectrum libraries cannot be used to identify them. For known components the library search routines are very useful tools [1–3].

A "novelty filter" is required, i.e. a process that extracts components that are new. This type of problem is encountered in several fields of analytical chemistry. Environmental analysis and doping analysis in sports are two examples. Identification of new drug metabolites is an important area of pharmaceutical research. Doping analysis provided the initial motivation for developing the mathematical tools described later [4, 5].

Sometimes attempts are made to isolate all component spectra in a GC–MS run. It is hard to reach that goal. As the physical isolation of all components is not always possible, mathematical tools can be used as a complement to physical separations. If resolutions are not successful for some components, attempts are made to separate them by the analysis of data from a hyphenated instrument. In this sense, the technique can be described as 'mathematical chromatography' or 'chromatography by mathematics' [5]. For the analyst the computer is a logical extension of the chromatographic column.

The Problem of Uniqueness

A set of methods has been developed — Alternating Regression (AR) — that is able to decompose a two-dimensional observation matrix into a number of two-dimensional components. The algorithm has been described thoroughly earlier [6]. The method needs an estimate for the number of components from the investigator, and no previous data about the spectra or retention profiles is needed. In

*Presented at the Symposium on "Chemometrics in Pharmaceutical and Biomedical Analysis", November 1990, Stockholm, Sweden.

this sense alternating regression calculates all components from scratch.

A central problem in any method for mathematical resolution is estimating the uniqueness of the solution. The result should lie in a narrow interval and it should be reproducible when starting from different points. The result should not depend on the particular method of analysis, and several different mathematical approaches should produce the same result. If the mathematical tools are applied with proper care the solution is limited only by the quality of the data.

The uniqueness of the solution is a function of the observations, not a function of the algorithm. In most circumstances the observations do not define a single, unique solution. The algorithm can work correctly, but the observations are not distinct enough to produce a unique result. It is necessary to analyse which types of observations produce unique results and which do not.

All solutions to the spectrum reconstruction problem reduce the information content in the solution. When the reduction is sufficient the information matches that in the observations. The output required often has more degrees of freedom than the observations. The limited amount of information in the observations is demonstrated by factor analysis [7, 8]. More numbers in output cannot be produced than we have numbers in input.

The number of parameters in the solution can be reduced by several means. When the number of parameters in the solution is small enough a stable solution to the reconstruction problem is obtained. Different methods use different ways to reduce the degrees of freedom in the solution.

Factor analysis is currently the most popular way to reduce the dimensionality of the solution [9, 10]. The solution found by eigenanalysis is transformed by some transformations into physical spectra [11]. Alternating regression does not use factor analysis. It operates directly in the space of spectra and concentration profiles. The solution is stabilized by forcing the concentration profiles to a unimodal shape.

Analysing the Problem of Uniqueness

The degree of difficulty in solving a given spectrum decomposition problem has been discussed previously [6]. If the overlap between components is very extensive, the results are not uniquely defined. Instead of one solution we get results in a certain range of variation.

In statistical terms the covariances between the component matrices are a measure of the inherent difficulty of mathematical analysis. If it is required to reduce the estimate to a single number then a condition number can be used for the component matrix that contains one column for each compound. The column corresponding to a given compound contains the product of the spectrum and concentration vector for that compound. The original rectangular matrix containing the outer product is "stretched" into a vector to form a column in component matrix. The condition number is calculated as the ratio between the largest and smallest eigenvalue of the component matrix.

Some expressions of the component differences are quite familiar to practical analysts. In fact, they are used in daily work.

In the simplest case where one mass number in the observations is specific for one compound very little mathematical analysis is needed. In practice it is difficult to prove that there are no other sources for the specific ion. If there are mass numbers specific for one molecule, the solution found by all decomposition methods is very robust and tolerates noise in the observations.

Another favourable situation is the absence of a given ion for a compound. If this mass number is present in other compounds but missing in one, the contrast makes it possible to find a unique solution. It is not possible to set up a simple calibration curve between the missing intensity and concentration of the compound. Still, the specific absence of a mass number makes it possible to get a unique solution for this component. GC–MS spectra often have useful gaps that make them dissimilar.

Usually the situation is not black or white, it is grey. There are no compound-specific ions or compound-specific gaps. The information is not sufficient to define a unique solution. The only way to force a unique solution is to bring in additional information in the form of *a priori* constraints. It can be assumed for example, that the chromatographic peaks have a certain shape [12]. The shape functions reduce the dimensionality of the solution. Assumptions can also be made about how many spectrum lines are present at most. Some rules can be

incorporated that define dependencies between fragments in a mass spectrum.

Introducing constraints has its dangers. The solution can be too subjective. By introducing constraints it is possible to obtain a definite solution. Still, the result may be worthless because it incorporates too many subjective elements.

The best way to get a more unique solution is to do more experimental work. Additional information should be added until an answer is obtained with a narrow confidence interval. This often means combining a number of different spectroscopies. If one spectroscopy is used, the analyst manipulates experimental variables to obtain a different data set than the original [13].

If successful, a new expanded data set is obtained where the components are more orthogonal than in the first one. Overall, better separation is obtained. With luck, some of the new observations have more specific peaks or more specific missing peaks. Finally, the robust nature of the solution should be verified by adding noise and solving the spectrum reconstruction problem several times.

Some Numerical Experience

HPLC/UV–vis data are quite different from GC–MS data. UV spectra are continuous, MS spectra are line spectra. The information content is much higher in mass spectra. AR works well with HPLC/UV–vis data and the resulting spectra are smooth. A demonstration of the AR with synthetic data is shown in Fig. 1 and Fig. 2A–D. The speed of convergence is typically faster with GC–MS than with HPLC/UV–vis. One possibility is to use the smoothness of optical spectra as a further constraint.

The quantitative fitting of a complete GC–MS run brings out some deficiencies in the data. Invariably, a small proportion of outliers is found in the data. These outliers are not outliers on the intensity scale. They are outliers on the mass axis. The frequency of outliers is about one point in one thousand.

The reason for the outliers is the algorithm in the computer of the GC–MS that converts the raw readings into a line spectrum. Because the algorithm is forced to work in real time, there is a loss of accuracy in some assigned mass numbers. The next generation of mass spectrometers should store the complete primary data on disk. This would permit even better utilization of all instrumental data, as true two-dimensional measurements could be analysed.

The main strength of AR is the speed of calculation. Data can be processed in a time frame that corresponds to the time needed for measurements. In theory non-linear optimization (NLP) can be used to solve the same problem as AR [14]. In practice AR is much faster than NLP. This makes it usable even on small computers. Speed can be used to perform robustness analysis by solving the problem several times after perturbations by added noise. In future all published spectra should have an indication about confidence intervals.

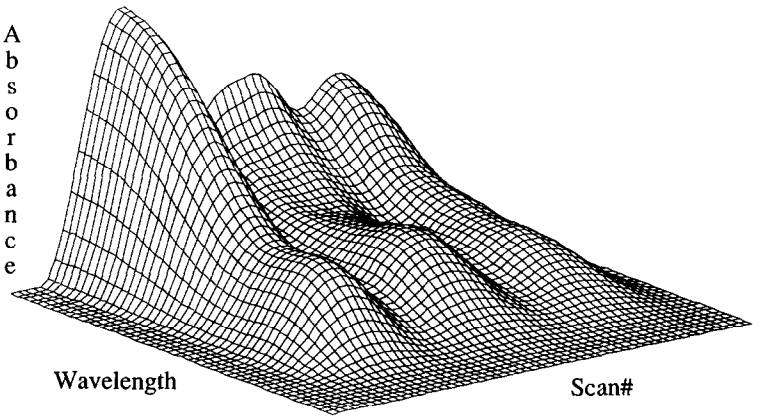

Figure 1
A synthetic observation matrix for HPLC/UV–vis. A mixture of five components was formed from the original spectra shown in Fig. 2(A). Most of the raw spectra contain three components.

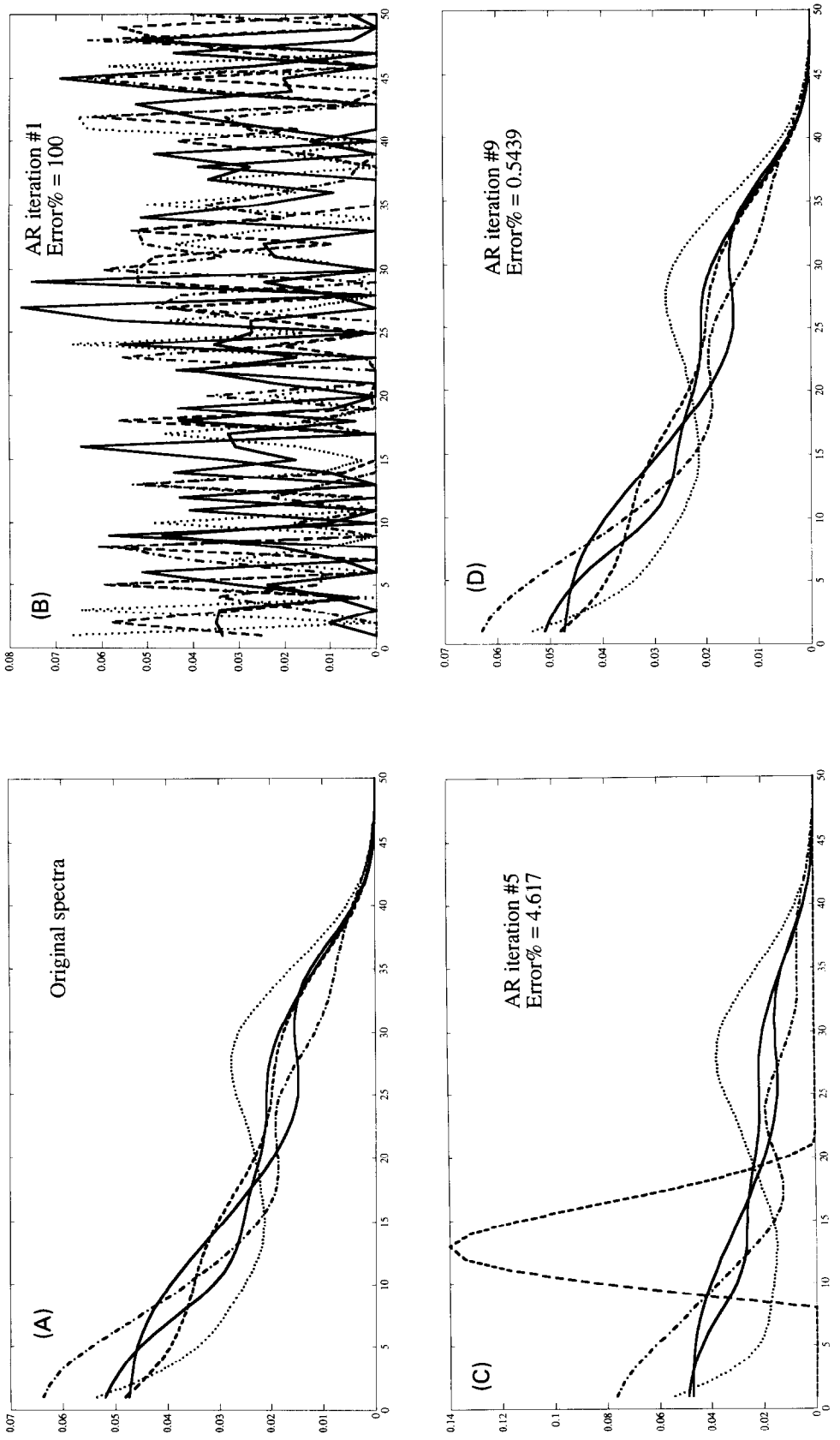

Figure 2
(A) The original spectra that were used to synthesize the overlapping observations. (B) These spectra are the starting values for AR. They are random numbers. (C) After five iterations the current spectra show some resemblance to the original spectra. (D) After nine iterations the fit is good and the resulting spectra resemble the original spectra that were used to form the observations.

References

[1] B.A. Knock, I.C. Smith, D.E. Wright and R.G. Ridley, *Anal. Chem.* **42**, 1516–1520 (1970).
[2] H.S. Hertz, R.A. Hites and K. Biemann, *Anal. Chem.* **43**, 681–691 (1971).
[3] D.B. Stauffer and F.W. McLafferty, *Anal. Chem.* **57**, 1056–1060 (1985).
[4] K. Kuoppasalmi and U. Karjalainen, in *Clinical Chemistry Research Foundation Library*, Vol. 1, pp. 1–40. United Laboratories Ltd, Helsinki (1984).
[5] E.J. Karjalainen and U.P. Karjalainen, in *Clinical Chemistry Research Foundation Library*, Vol. 2, pp. 1–48. United Laboratories Ltd, Helsinki (1987).
[6] E.J. Karjalainen, *Chemometrics and Intelligent Laboratory Systems* **7**, 31–38 (1989).
[7] J.M. Halket, *J. Chromatogr.* **175**, 229–241 (1979).
[8] M.A. Sharaf and B.R. Kowalski, *Anal. Chem.* **53**, 518–522 (1981).
[9] W. Lindberg, J. Öhman and S. Wold, *Anal. Chem.* **58**, 299–303 (1986).
[10] B. Vandeginste, R. Essers, T. Bosman, J. Reijnen and G. Kateman, *Anal. Chem.* **57**, 971–985 (1985).
[11] J.K. Strasters, H.A.H. Billiet, L. de Galan, B.G.M. Vandeginste and G. Kateman, *Anal. Chem.* **60**, 2745–2751 (1988).
[12] J.P. Foley, *Anal. Chem.* **59**, 1984–1987 (1987).
[13] L.S. Ramos, J.E. Burger and B.R. Kowalski, *Anal. Chem.* **57**, 2620–2625 (1985).
[14] M.D. King and G.S. King, *Anal. Chem.* **57**, 1049–1056 (1985).

[*Received for review 26 November 1990*]