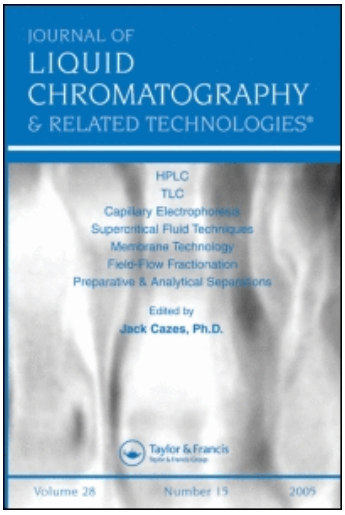

Journal of Liquid Chromatography & Related Technologies
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597273

SPEAK, A PROCEDURE FOR NOISE REDUCTION IN MULTICHANNEL CHROMATOGRAPHIC DATA

J. L. Beltrán[a]; J. Guiteras[a]; R. Ferrer[a]
[a] Departament de Química Analítica, Universitat de Barcelona, Barcelona, Spain

Online publication date: 17 November 1999

J. LIQ. CHROM. & REL. TECHNOL., 22(20), 3177–3187 (1999)

SPEAK, A PROCEDURE FOR NOISE REDUCTION IN MULTICHANNEL CHROMATOGRAPHIC DATA

J. L. Beltrán, J. Guiteras, R. Ferrer

Departament de Química Analítica
Universitat de Barcelona
Avgda. Diagonal, 647
E-08028 Barcelona, Spain

ABSTRACT

A new procedure to reduce the background noise of multichannel chromatographic data, named SPEAK, has been developed. It is based on the assumption of gaussian elution profiles, but it can also deal with non-gaussian peaks.

The method has been tested with simulated data matrices, containing high noise levels, and real data matrices obtained by a fast-scanning fluorescence detector coupled to a HPLC system. Results indicate that SPEAK is able to lower the background noise by a factor of about 4 - 5, without any significant losses of the relevant information.

INTRODUCTION

There are several mathematical procedures that have been applied to the resolution or smoothing of chromatographic data. For single channel chromatograms, the peaks can be adjusted to different statistical functions, as the Gauss, Log-normal, Gamma, Weibull, modified Gauss or exponentially modified Gauss.[1-3] These fit functions allow the resolution of overlapped peaks.

In some cases, the chromatograms required some kind of smoothing, to reduce the signal-noise ratio in order to achieve better detection limits. In these cases, several different filtering algorithms can be used, as the Fast-Fourier Transform or the Savitzky-Golay filter.[4-6] For multichannel chromatograms, such as those obtained by HPLC-DAD (absorption spectrum), several chemometric procedures based on factor analysis can be used for the smoothing of the data matrix,[7] or for the deconvolution of chromatographic peaks, such as the self-modelling curve resolution (SMCR), principal components regression (PCR) or rank-annihilation by evolving factor analysis (RAEFA).[8-10] In principle, methods based on factor analysis do not need previous assumptions about the data matrix (except unimodality constraints and non-negative detector responses and chromatographic profiles), and therefore, they differ from the methods used in single channel detection, because these require the use of a predefined fitting function. Other procedures, such as the Kalman filter or the generalized rank annihilation (GRAM), have also been used successfully for multichannel chromatography and capillary electrophoresis.[11-13]

The procedure SPEAK, presented here, is intended to find the corresponding SPEctra and PEAKs of the different compounds determined by multichannel chromatography, in order to reduce the background noise significantly, and it is based on the reconstruction of the mean profile of a chromatogram by the successive addition of gaussian peaks.

To test its ability to reduce the instrumental noise, the SPEAK procedure was first used for the reconstruction of a simulated multichannel chromatogram, over which several levels of random noise, ranging in amplitude from 10% to 50% of the maximum value of the signal, had been superimposed. The procedure was then tested with a set of chromatograms recorded with a fast-scanning fluorescence detector coupled to a HPLC system (HPLC-FSFS).[14] This detection system allows one to obtain a fluorescence spectrum of the effluent of the chromatographic column each 2.7 seconds, at a scanning speed of 50 nm s^{-1}. In these conditions, the integration time is very low (about 0.04 s) and some signals can be severely affected by the background noise, particularly in the case of compounds whose concentration levels are close to their detection limits. The results indicate that the instrumental noise can be reduced about 4 - 5 times, with few losses in data reproduction.

EXPERIMENTAL

Materials

Simulated Data

A simulated data matrix (containing 50 readings for every one of 50 different retention times) was calculated from three hypothetical compounds having

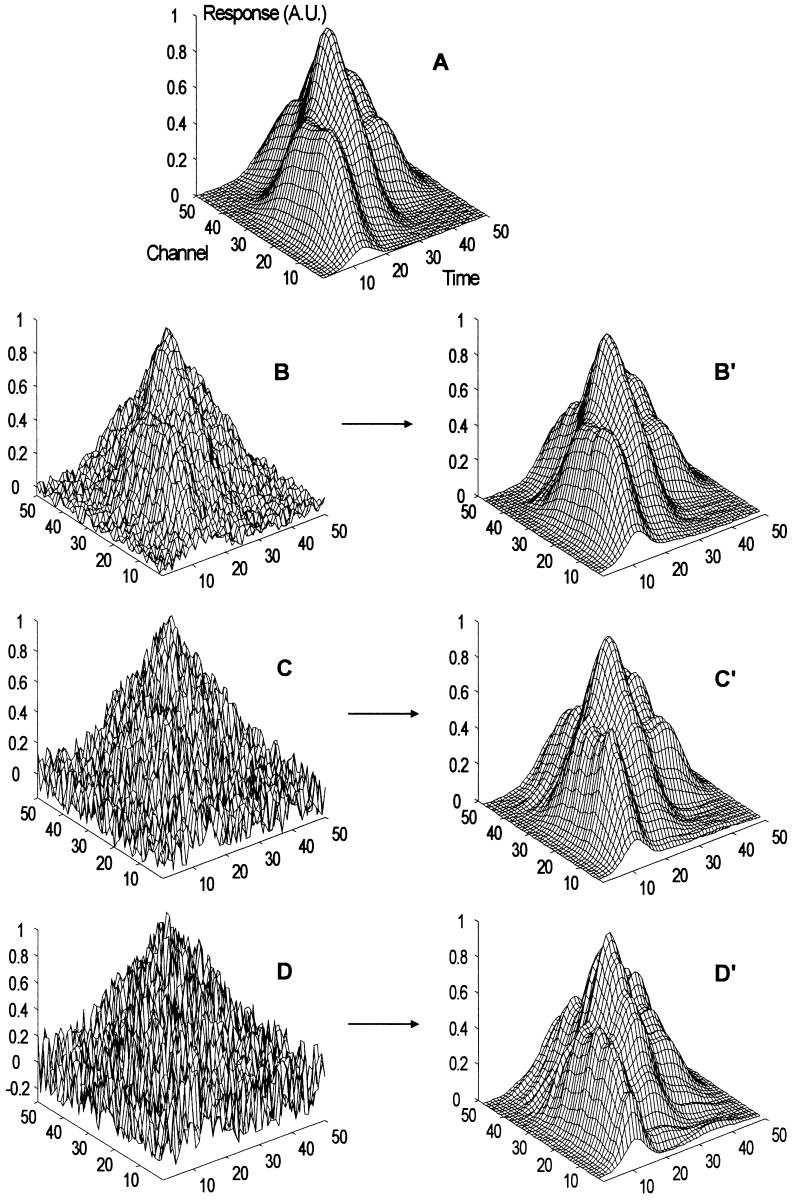

Figure 1. A: Original synthetic data matrix. B, C, and D: matrix A with added noise levels equivalent to 10%, 30 % and 50% of the maximum signal, respectively. B', C' and D': B, C, and D matrices reconstructed by the SPEAK procedure.

overlapped gaussian peaks and different spectra. The data matrix is represented in Figure 1A. Several levels of random noise, with amplitudes equivalent to 10%, 20%, 30%, 40% and 50% of the maximum value of the signal, were added to the matrix.

Figures 1B, 1C and 1D show the data corresponding to 10%, 30% and 50% of noise added, respectively.

Real Chromatographic Data

HPLC-FSFS chromatograms were obtained for solutions of several polycyclic aromatic hydrocarbons (PAHs) as described previously.[14] The first data set corresponds to a fraction of a chromatogram of a synthetic sample, prepared from standards of several PAHs.

The HPLC-FSFS system was used to detect benzo[b]fluoranthene, benzo[k]fluoranthene, benzo[a]pyrene and dibenz[a]-anthracene by setting the excitation wavelength at 300 nm, and recording the emission fluorescence in the range from 335 to 435 nm.

The second data set corresponds to a water sample from a lagoon near Barcelona airport. In this case, the PAHs were extracted from the water sample by a cloud-point extraction procedure, using Triton X-114 as non-ionic surfactant.[15] A fraction of the chromatogram of this sample, which contained the same compounds as the synthetic sample, was selected to compare results.

Methods

SPEAK Procedure

The SPEAK procedure, developed in MATLAB™ 4.2 environment, is intended to reduce the background noise in multichannel chromatograms. SPEAK is based on two hypotheses: a) the peaks of the different analytes follow a gaussian distribution, and b) the background noise has a random distribution.

The procedure consists of several steps, which are described below:

1) The data matrix **D,** obtained with a multichannel chromatographic detector and which contains **c** channels measured at **t** different retention times, is first reduced to a single vector, **d,** by calculating the mean value of the signal for all the channels at the different retention times:

$$D(cxt) \rightarrow d(t) \qquad (1)$$

This data reduction has two properties: all the channels are represented and vector **d** has a lower noise than the original data matrix, **D**.

2) A gaussian peak ($\mathbf{p_1}$) is fitted to vector **d**. The solution is found by the iterative Gauss-Newton non-linear regression algorithm[16], in which the parameters $\boldsymbol{\sigma_1}$, $\mathbf{t_{o1}}$ and $\mathbf{A_1}$ of the gaussian distribution are optimized in order to find a new vector, $\mathbf{p_1}$, which is as close as possible to **d**:

$$p_1(i) = \frac{A_1}{\sigma_1\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{t(i) - t_{o1}}{\sigma_1}\right)^2\right\} \qquad (2)$$

An error vector, $\mathbf{e_1}$ is obtained as the difference between **d** and $\mathbf{p_1}$. In this step, the errors are important, because only one compound is used to model the whole chromatogram of a complex mixture.

3) The error vector, $\mathbf{e_1}$, is then fitted to a second gaussian distribution, a second set of parameters, $\boldsymbol{\sigma_2}$, $\mathbf{t_{o2}}$ and $\mathbf{A_2}$ is calculated and a new vector, $\mathbf{p_2}$, which corresponds to the estimated elution profile of another component of the mixture, is obtained.

A new error vector, $\mathbf{e_2}$, is calculated as:

$$e_2 = d - p_1 - p_2 \qquad (3)$$

The error is minimised by successive refinements of the parameters that describe peaks $\mathbf{p_1}$ and $\mathbf{p_2}$ until the relative difference of the errors of two consecutive refinements steps is less than 0.1%.

The process described in step 3 is repeated until either 99.5% of the area described by the original chromatogram, represented by vector **d**, is reached, or the relative difference between two consecutive additions of peaks is less than 0.1%. Let **k** be the number of peaks required to achieve the reconstruction.

4) As the result of step 3, a set of independent peaks, $\mathbf{p_1}$, $\mathbf{p_2}$, ...,$\mathbf{p_k}$, which describe a matrix of chromatographic profiles, **P** (k x t), is obtained. These profiles have all the gaussian properties, such as unimodality and positive values, and can be related to the original data matrix **D** in order to obtain a set of pure spectra **S** (c x k) of the analytes:

$$D = S \bullet P \qquad (4)$$

In fact, direct matrix operation would give the pure spectra of the **k** compounds, but the use of a non-negative-least-squares routine is preferable because it gives spectra that are positive in all cases.

5) A Savitzky-Golay[6] cubic filter smoothes the spectra of the different compounds calculated by the above-described procedure, and a $\mathbf{S_R}$ matrix of pure spectra is obtained. This smoothing implies the loss of the first two

Table 1

Error in the Reproduction of the Synthetic Data

Percentage of Error Added	Raw Data Matrix Standard Deviation	Raw Data Matrix RRSMF (%)	Recovered Data Matrix Standard Deviation	Recovered Data Matrix RRSMF (%)
10	0.0295	8.697	0.0074	2.190
20	0.0583	17.221	0.0095	2.814
30	0.0885	26.118	0.0211	6.227
40	0.1138	33.599	0.0260	7.679
50	0.1416	41.782	0.0281	8.295

points and the last two points of each spectrum, but it reduces the presence of unexpected maxima or minima. Finally, a reconstructed data matrix ($\mathbf{D_R}$) is obtained as the product of the matrix of pure spectra and the matrix of chromatographic profiles.

$$D_R = S_R \bullet P \tag{5}$$

The reconstructed data matrix has four channels less than the original data matrix, as a result of the smoothing of the pure spectra.

RESULTS AND DISCUSSION

Simulated Data

The SPEAK procedure was first used with simulated data matrices, which had several different levels of noise superimposed, in order to evaluate its abilities for the reconstruction of the original data. The error in the reproduction of the original data matrix was evaluated from the standard deviation between the original and the reconstructed data, and also from the Relative Root Mean Squared Error or Fit (RRMSEF):

$$\mathrm{RRMSEF}(\%) = 100 \sqrt{\frac{\sum_{i=3}^{c-2}\sum_{j=1}^{t}\left(D_{R,i,j} - D_{i,j}\right)^2}{\sum_{i=3}^{c-2}\sum_{j=1}^{t}\left(D_{i,j}\right)^2}} \tag{6}$$

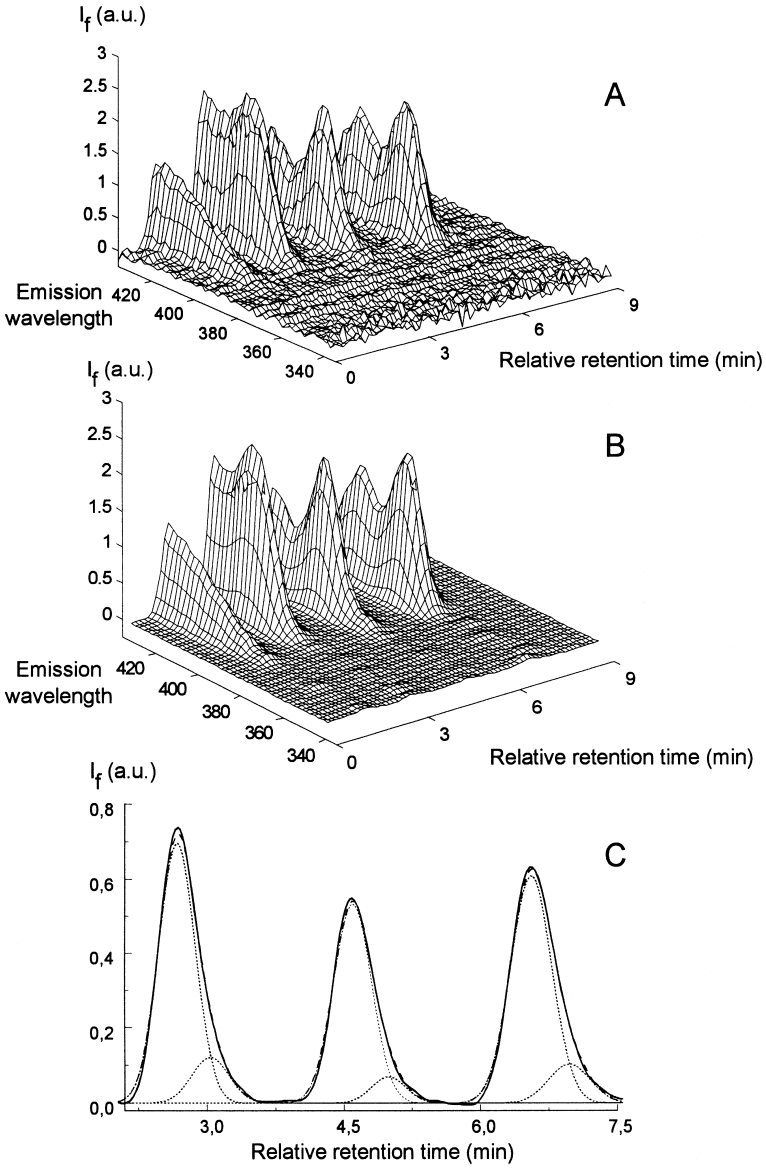

Figure 2. A: Real chromatogram of a mixture of benzo[b]fluoranthene, benzo[k]fluoranthene, benzo[a]pyrene and dibenz[a]anthracene. Injection volume: 25 μL; concentrations were 29.7, 13.3, 14.4 and 36.6 μg l^{-1}, respectively. B: Reconstructed chromatogram. C: Solid line: mean elution profile of A. Dotted lines: individual profiles of six gaussian peaks. Dashed line: elution profile reconstructed by the addition of the calculated gaussian peaks.

Results are shown in Table 1, where it can be observed that, for matrices having a random noise added, the RRMSEF values varied between 8.7% and 41.8%, for superimposed noise levels between 10% and 50%, respectively. The reconstruction of the data by the SPEAK procedure led to important reductions in the noise, as the RRMSEF was lowered to levels between 2.2% and 8.3%. The effectiveness of SPEAK is clearly appreciated in Figures 1.B', 1.C' and 1.D', depicting the reconstructed data matrices from data which had noise levels of 10%, 30% and 50% superimposed, and are compared with Figure 1.A, which represents the original data matrix. It should be noticed that these reconstructed matrices had, in all cases, RRMSF values that were about four or five times lower than the corresponding original matrices.

Real Data

The same procedure was applied to the real HPLC-FSFS chromatographic data aforementioned. The synthetic sample (see Figure 2.A) contained four PAHs which were well resolved and had sufficient sensitivity. However, only one compound could be modelled by a single gaussian peak, because the remaining peaks had non-gaussian shapes provoked by the fact that the chromatographic conditions used included an increase in the flow of the mobile phase from 1 mL min^{-1} to 4 mL min^{-1} in this zone. As a consequence, two gaussian peaks were needed in these cases to fit the individual profiles of these compounds. This can be seen in Figure 2.C, where the mean profile and the calculated individual profiles are shown.

In Figure 2.B, the final reconstructed data matrix is represented. The difference between the original and the reconstructed matrices had a mean value of -0.0045 and a standard deviation of 0.0461, and no particular trend of the residuals could be observed.

The last example is a fraction of a chromatogram of a real water sample, which contains the same compounds as in the previous case. This sample was first analysed for PAHs content using a preconcentration step by a cloud-point extraction procedure,[15] and determined by HPLC with wavelength programmed fluorescence detection.[17] The results indicated that the four compounds were present in the water sample in the part-per-trillion range (benzo[b]fluoranthene: 162 ng l^{-1}; benzo[k]fluoranthene: 102 ng l^{-1}; benzo[a]pyrene: 139 ng l^{-1}, and dibenz[a]anthracene: 6 ng l^{-1}). A fraction of the HPLC-FSFS chromatogram of the extract is indicated in Figure 3.A.

Despite the preconcentration factor (25 times), the concentration of PAHs in the extract is very low, so only one component, benzo[k]fluoranthene, having the most intense fluorescence of the four compounds, is clearly identified, whereas the other compounds are masked by the background noise.

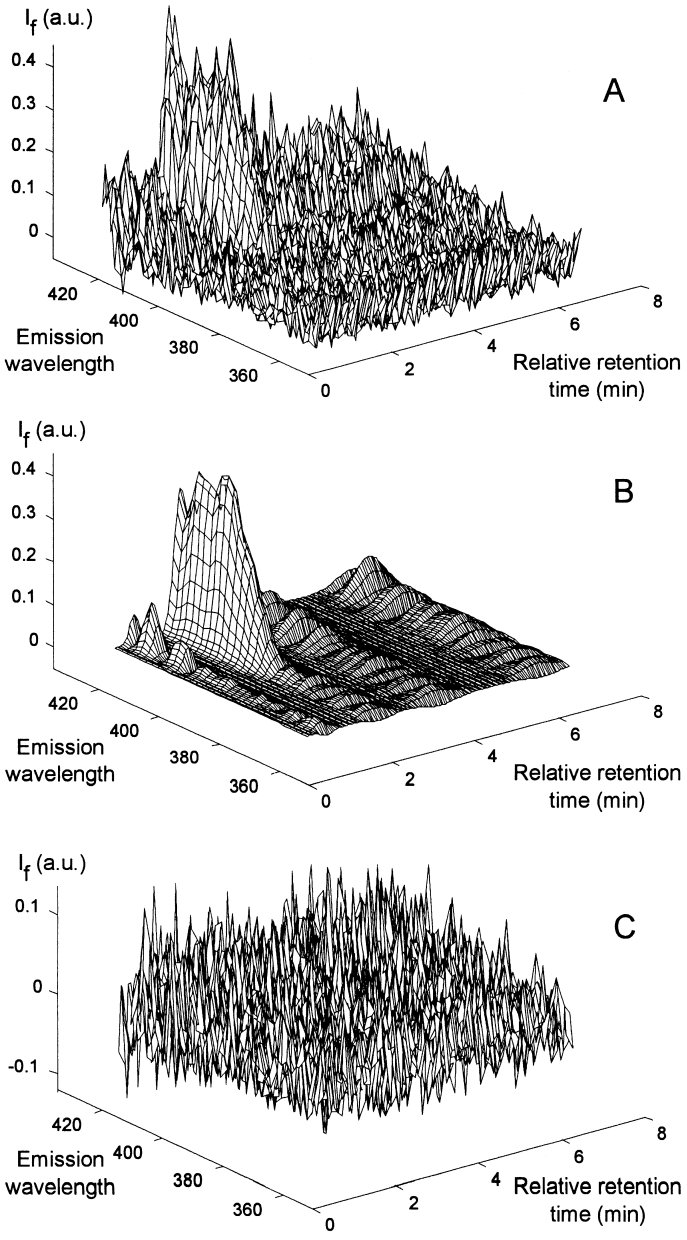

Figure 3. A: Real chromatogram of a lagoon water extract (see text for details). The identified compounds are the same as in Figure 2. Injection volume: 25 μL. B: Reconstructed chromatogram. C: Eliminated noise, calculated as the difference between A and B.

The chromatogram reconstructed by SPEAK is indicated in Figure 3.B, showing the peaks corresponding to the four PAHs. In Figure 3.C, the differences between the raw and the reconstructed data matrices are represented. In this case, the mean value of the residuals was -0.0065, with a standard deviation of 0.0346.

CONCLUSIONS

A new smoothing procedure, SPEAK, for the pre-treatment of multichannel chromatographic data, has been developed. It is based on the assumption that the chromatographic profiles can be fitted to gaussian shapes. In the case of non-gaussian profiles, the fit is obtained by the addition of several gaussian peaks.

The procedure has been tested with simulated data matrices with several levels of random noise superimposed, and with real data matrices. Results indicate that instrumental noise can be reduced to a significant extent, without loss of relevant information. In all cases, regardless of the noise level, SPEAK was able to detect and model all the compounds present in the mixtures.

As noted before, SPEAK works in MATLAB 4.2 environment, and it has been tested under MS Windows 3.1 and MS Windows 95 operating systems on PC-compatible computers, and under IBM AIX 4.1 operating system on an IBM RS6000 computer. The source code (m-files) is available upon request to the authors.

Although SPEAK has been tested with data obtained by HPLC-FSFS, its characteristics make it suitable for use with any spectrometric multichannel detection system used in chromatography.

ACKNOWLEDGMENT

The authors thank the Comisión Interministerial de Ciencia y Tecnología (AMB98-0327) for supporting this study.

REFERENCES

1. J. Grimal, H. Iturriaga, X. Tomas, Anal. Chim. Acta, **139**, 155 (1982).

2. J. P. Foley, J. G. Dorsey, Anal. Chem., **55**, 730 (1983).

3. M. S. Jeansonne, J. P. Foley, J. Chromatogr. Sci., **29**, 258 (1991).

4. G. Horlick, Anal. Chem., **44**, 943 (1972).

5. S. Lamb, T. S. Isenhour, Anal. Chem., **53**, 1179 (1981).

6. A. Savitzky, M. J. E. Golay, Anal. Chem., **36**, 1627 (1964).

7. E. R. Malinowski, **Factor Analysis in Chemistry**, 2nd. Ed., John Wiley, New York 1991.

8. J. G. Hamilton, P. J. Gemperline, J. Chemometrics, **4**, 1 (1990).

9. R. G. Bereton, Analyst, **120**, 2313 (1995).

10. H. Gampp, M. Maeder, C. J. Meyer, A. D. Zuberbuehler, Anal. Chim. Acta, **193**, 287 (1987).

11. S. C. Rutan, Anal. Chem., **63**, 1103A (1991).

12. R. B. Poe, S. C. Rutan, Anal. Chim. Acta, **283**, 645 (1993).

13. S. Xiong, J. Li, J. Cheng, J. Liq. Chrom. & Rel. Technol., **19**, 2129 (1996).

14. R. Ferrer, J. Guiteras, J. L. Beltrán, J. Chromatography A, **779**, 123 (1997).

15. R. Ferrer, J. L. Beltrán, J. Guiteras, Anal. Chim. Acta, **330**, 477 (1996).

16. E. Durand, **Solutions Numeriques des Equations Algébriques, Tome II: Systèmes de Plusieurs Equations**, Masson, Paris, 1972.

17. R. Ferrer, J. Guiteras, J. L. Beltrán, Anal. Letters, **29**, 2201 (1996).

Received December 17, 1998
Accepted April 4, 1999
Manuscript 4951

Request Permission or Order Reprints Instantly!