ELSEVIER

Journal of Chromatography A, 839 (1999) 129–139

JOURNAL OF CHROMATOGRAPHY A

Statistical determination of the proper sample size in multicomponent separations

Attila Felinger*, Ervin Vigh[1], András Gelencsér[2]

Department of Analytical Chemistry, University of Veszprém, Veszprém, Egyetem u. 10., H-8200 Hungary

Received 8 September 1998; received in revised form 10 December 1998; accepted 23 December 1998

Abstract

The Fourier analysis of multicomponent chromatograms is applied to chromatograms of diesel fuels. The concentration of the injected sample was varied over several orders of magnitude. The number of components, the average peak width, as well as the retention pattern were determined for all sample concentrations. The lower edge of the practical sample size is limited by the baseline noise, whereas the upper limit is set by column overload. By means of statistical analysis, those two boundaries were isolated and a method is proposed to determine the useful sample size for the analysis of any multicomponent mixture.

Keywords: Diesel fuel; Sample size; Fourier analysis

1. Introduction

The statistical theory of peak overlap has been enormously developed during the past 15 years [1–3]. From one single multicomponent chromatogram, the number of detectable components, the retention pattern, and other chromatographic properties can be estimated.

There are two fundamentally different approaches for the statistical analysis of multicomponent chromatograms. The first theory – developed by Davis and Giddings [1,4] – is based on the analysis of the retention time increments between adjacent single component peaks using the concepts of stochastic point processes.

The other method, the Fourier analysis of multicomponent chromatograms extracts information from either the power spectrum or the autocovariance function of the chromatogram. By means of Fourier analysis, the average peak shape parameters of the chromatogram are also determined, therefore further information on the efficiency of the separation can be obtained.

The concentration of the components present in a multicomponent mixture may span over a huge concentration range. Experimental data show that the distribution of the single component peak areas is very close to the exponential distribution [5,6]. The sample size for a chromatographic analysis must be chosen properly so that the number of detectable components is as large as possible. If the sample size is too small, many components will be buried in the

*Corresponding author.

[1]Present address: Tiszai Vegyi Kombinát Rt., H–3581 Tiszaújváros, P.O. Box 20.

[2]Present address: Department of Earth and Environmental Sciences.

0021-9673/99/$ – see front matter
PII: S0021-9673(99)00068-0

baseline noise. On the contrary, if the sample size is too large, the concentration of some components will be so high that the column will be overloaded. Due to column overload, the peak width will increase, the column efficiency drops, and the probability of peak overlap will increase.

Besides the sample size, the integration method of the multicomponent chromatogram also influences the number of detectable components [7].

The aim of this study is to apply Fourier analysis in order to determine the concentration range, or sample size, which is optimum for the detection of the highest number of components, and accordingly to extract the maximum amount of information.

2. Theory

The fundamental assumption when using statistical estimations relies in the randomness of the retention times. Felinger demonstrated that mixing a few chemical families results in a nearly-Poissonian retention pattern [8]. In a Poissonian chromatogram, the distribution of retention time increments is exponential, and the retention times themselves follow a uniform distribution. This pseudo-random feature of a complex chromatogram makes the statistical estimation possible. When one chemical family is dominant in the sample, and therefore the retention time increments are not disordered, an ordered interval distribution model, such as the normal or the gamma, can be applied [9]. Even in this case however, some randomness is always present in the sample due to experimental uncertainties and contaminants.

Davis and Giddings attributed the randomness to the distribution of the difference of the standard chemical potentials between the stationary and the mobile phases [4]. Herman et. al. [6] and Martin et al. [10] analyzed published retention data, and confirmed the random nature of the retention pattern.

Therefore, we can assume that retention time and peak area (or peak height) are independent random variables in a multicomponent chromatogram. This simple assumption allows the determination of the power spectrum of a multicomponent chromatogram.

The power spectrum of a general Poissonian chromatogram is [11]:

$$P(\omega) = \frac{2a_a^2}{T}\left(\frac{\sigma_a^2}{a_a^2} + 1\right) E\{|g(\omega)|^2\} \tag{1}$$

where a_a is the average area of the single component peaks, σ_a is the standard deviation of the areas, T is the average retention time increment, and $g(\omega)$ is the Fourier transform of the single component peak shape model.

The simplest method is to assume that all single component peak widths are identical, a situation which can be reached by temperature programming in gas chromatography. When the chromatogram is disordered and the peak width and other peak shape parameters of the single component peaks are constant, the following power spectrum is obtained:

$$P(\omega) = \frac{2a_a^2}{T}\left(\frac{\sigma_a^2}{a_a^2} + 1\right) |g(\omega)|^2 \tag{2}$$

When the single component peak shapes are described by the exponentially modified Gaussian (EMG) function – which is identical to the Gaussian peak if $\tau = 0$ – then the power spectrum of a single component peak is [11]:

$$|g(\omega)|^2 = \frac{e^{-\omega^2\sigma^2}}{1 + \omega^2\tau^2} \tag{3}$$

Thus, the power spectrum of a disordered, Poissonian multicomponent chromatogram built up by constant-width EMG peaks is:

$$P(\omega) = \frac{2a_a^2}{T}\left(\frac{\sigma_a^2}{a_a^2} + 1\right) \frac{e^{-\omega^2\sigma^2}}{1 + \omega^2\tau^2} \tag{4}$$

The term $2a_a^2/T$ can be replaced by chromatographic quantities better suited for our purpose. When m is the number of single components, the total area of the chromatogram is:

$$A_T = ma_a \tag{5}$$

As the mean interval between adjacent peaks is the ratio of the total chromatographic space over the number of single components:

$$T = \frac{X}{m} \tag{6}$$

the power spectrum is expressed as:

$$P(\omega) = \frac{2A_T^2}{mX} \frac{e^{-\omega^2\sigma^2}}{1+\omega^2\tau^2}\left(\frac{\sigma_a^2}{a_a^2}+1\right) \quad (7)$$

Parameters A_T and X are easily accessible from the chromatogram. The number of single components m, the peak widths σ, and in the case of asymmetrical peaks parameter τ can be determined by nonlinear curve fitting. The relative standard deviation of the peak areas can be estimated by the heights of the detected peaks as well.

It has recently been demonstrated quantitatively by Dondi et al. how the saturation of the chromatogram influences the accuracy of the estimation of the peak area dispersion, σ_a^2/a_a^2 [12].

According to the Wiener–Khinchin theorem, the power spectrum and the autocorrelation function form a Fourier pair. The relationship between the power spectrum and autocorrelation function of ergodic stochastic processes is [13]:

$$P(\omega) = 2\int_{-\infty}^{\infty} C(t)e^{-i\omega t}\,dt = 4\int_0^{\infty} C(t)\cos(\omega t)\,dt \quad (8)$$

and

$$C(t) = \frac{1}{4\pi}\int_{-\infty}^{\infty} P(\omega)e^{i\omega t}\,d\omega$$

$$= \frac{1}{2\pi}\int_0^{\infty} P(\omega)\cos(\omega t)\,d\omega \quad (9)$$

On the basis of the above equations, the autocorrelation or autocovariance function and the power spectrum are identical tools to characterize multicomponent chromatograms.

For the sake of convenience, the chromatogram is usually centered around its mean before the calculation of its power spectrum or autocorrelation function. When a mean-centered chromatogram is subjected to the calculations discussed here, the autocovariance function of the chromatogram is obtained instead of the autocorrelation function.

When using the autocovariance function to determine the parameters of multicomponent chromatograms, we can avoid the use of smoothing windows, which is always a demanding step in frequency-domain signal processing [3].

The inverse Fourier transform of the power spectrum given in Eq. 2 can be calculated by the Wiener–Khinchin theorem, and the following autocovariance function is obtained:

$$c(t) = \frac{2A_T^2}{mX}\left(\frac{\sigma_a^2}{a_a^2}+1\right)C_u(t) \quad (10)$$

where $C_u(t)$ is the autocorrelation function of the

Table 1
The contribution of the interval distribution to the power spectrum

Distribution	$f(t)$	$\theta(\omega)$	$2\Re\,\frac{\theta(\omega)}{1-\theta(\omega)}$
Exponential	$\frac{1}{\tau}e^{-t/\tau}$ $(t \geq 0)$	$\frac{1}{1-i\omega T}$	0
Normal	$\frac{e^{-(t-T)^2/2\sigma_T^2}}{\sqrt{2\pi}\sigma_T}$	$e^{-\omega^2\sigma_T^2/2-iT\omega}$	$\frac{2e^{-\omega^2\sigma_T^2}-2\cos(T\omega)e^{-\omega^2\sigma_T^2/2}}{2\cos(T\omega)e^{-\omega^2\sigma_T^2/2}-e^{-\omega^2\sigma_T^2}-1}$
Uniform	$\frac{1}{2T}$ $0 \leq t \leq 2T$	$\frac{\sin(T\omega)}{T\omega}e^{iT\omega}$	$\frac{2-2\cos(2T\omega)-2T\omega\sin(2T\omega)}{\cos(2T\omega)+2T\omega\sin(2T\omega)-2T^2\omega^2-1}$
Gamma	$\frac{\tau^{-p}t^{p-1}e^{-t/\tau}}{\Gamma(p)}$ $t \geq 0$	$\frac{1}{(1-i\omega\tau)^p}$	$\frac{2(\tau^2\omega^2+1)^{-p}-2(\tau^2\omega^2+1)^{-p/2}\cos[p\arctan(\tau\omega)]}{2(\tau^2\omega^2+1)^{-p/2}\cos[p\arctan(\tau\omega)]-(\tau^2\omega^2+1)^{-p}-1}$

unit-area peak shape model of the single component peaks.

The autocovariance function for a constant-peak-width Poisson chromatogram can be calculated analytically assuming, for instance, Gaussian peak profile [11,14]:

$$c(t) = \frac{A_T^2}{2\sigma\sqrt{\pi}Xm}\left(\frac{\sigma_a^2}{a_a^2} + 1\right) e^{-t^2/4\sigma^2} \qquad (11)$$

The nonlinear parameter estimation – which determines the number of single components and peak shape parameters – can be performed on the basis of the numerically calculated autocovariance function. If this is the case, the calculation of the power spectrum of the chromatogram is not required, all the necessary information can be obtained directly in the time domain. The numerical calculation of the autocovariance function is given as:

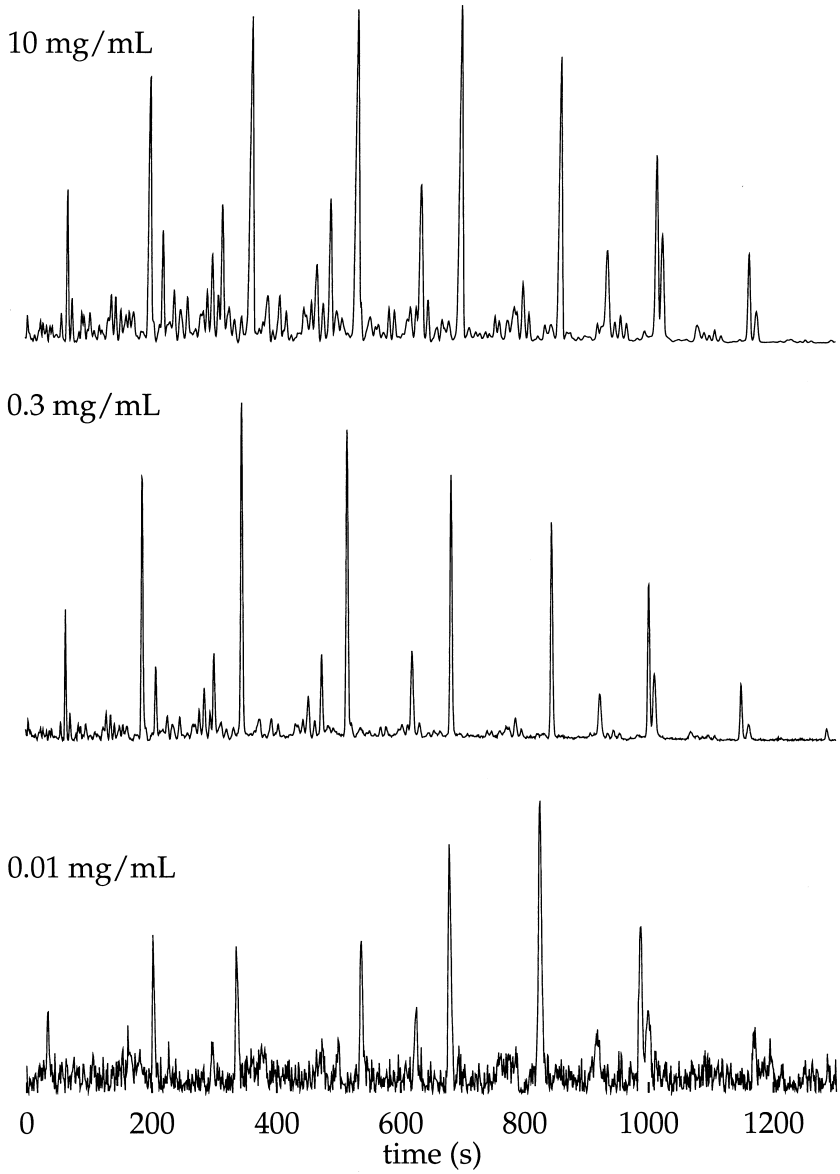

Fig. 1. Chromatograms of the diesel fuel recorded at different sample sizes.

$$c(k) = \frac{1}{M}\sum_{i=1}^{N}\left[Y(i) - \overline{Y}\right]\left[Y(i+k) - \overline{Y}\right] \quad (12)$$

where $N = X/\Delta t$ is the number of digitized points, Δt is the sampling time, $\overline{Y}$ is the mean value of the chromatogram, M is the width for which the autocovariance function is calculated.

When there exists some structural relationship among the components of a multicomponent mixture, the chromatogram is not totally disordered and interval models other than the exponential distribution should be used. The power spectrum of such an uncorrelated multicomponent chromatogram is:

$$P(\omega) = \frac{2A_T^2}{mX}|g(\omega)|^2 \left\{\frac{\sigma_a^2}{a_a^2} + 1 + 2\Re\frac{\theta(\omega)}{1-\theta(\omega)} + \frac{1}{T}\delta(\omega)\right\} \quad (13)$$

where $\theta(\omega)$ is the characteristic function of the retention time increment distribution [9,11,14–16]. When the chromatogram is centered around its mean intensity before the power spectrum is calculated, the Dirac-delta disappears from the power spectrum expression.

The term that depends on the distribution of the retention time increments is given in Table 1 for different distributions.

3. Experimental

Analyses were carried out with a GC–MS system (Fisons Instruments, model GC 8000 gas chromatograph; MS Trio 1000, model EI & CI 4521 mass spectrometer). Separations were carried out on a 30 m×0.32 mm open tubular column containing SPB-1 polydimethylsiloxan as stationary phase with a 0.25-

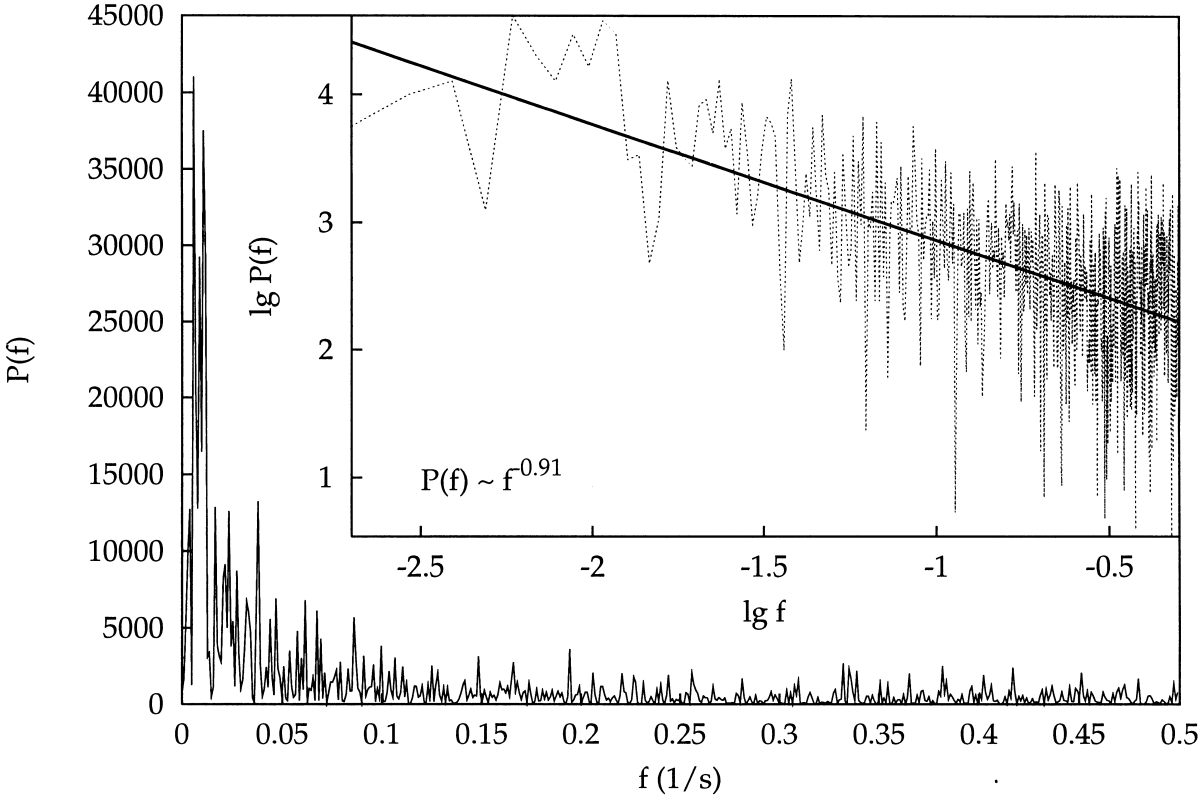

Fig. 2. Plot of the power spectrum of the baseline noise. The insert shows the logarithmic plot of the power spectrum.

μm film thickness. Helium was applied as carrier gas. The sample was introduced with splitless injection. A temperature program was applied to assure nearly constant peak width along the chromatogram.

A 100 mg/ml solution of the diesel fuel in *n*-hexane was prepared, and that stock solution was diluted to the desired sample concentration.

In this study, the results obtained from the total ion current chromatograms are summarized.

4. Results and discussion

A GC–MS analysis of a diesel fuel was carried out in a wide concentration range of the sample (0.01–30 mg/ml). Fig. 1 clearly shows that the sample size has a complex effect on the chromatogram. When the concentration of the sample is 0.01 mg/ml, only some major components can be detected. Most of the components are lost in the excessive baseline noise. When the sample concentration is increased to 0.3 mg/ml, the noise level is very low and the determination of several minor components is possible. A further significant increase of the sample concentration – to 10 mg/ml – results in the appearance of many further peaks.

To understand the effect of baseline noise on the results obtained by Fourier analysis, we have to calculate the power spectrum and the autocovariance function of the noise. The effect of baseline noise can be neglected provided that white noise perturbs our measurements as all disturbing effects of the white noise accumulate at the origin of the autocovariance function [14]. Unfortunately, the most common noise type of chromatographic detectors is the more structured flicker noise [3] whose power spectrum varies as $1/f$. The effect of flicker baseline noise on the accuracy of Fourier analysis cannot be neglected.

The baseline noise of the GC–MS system was

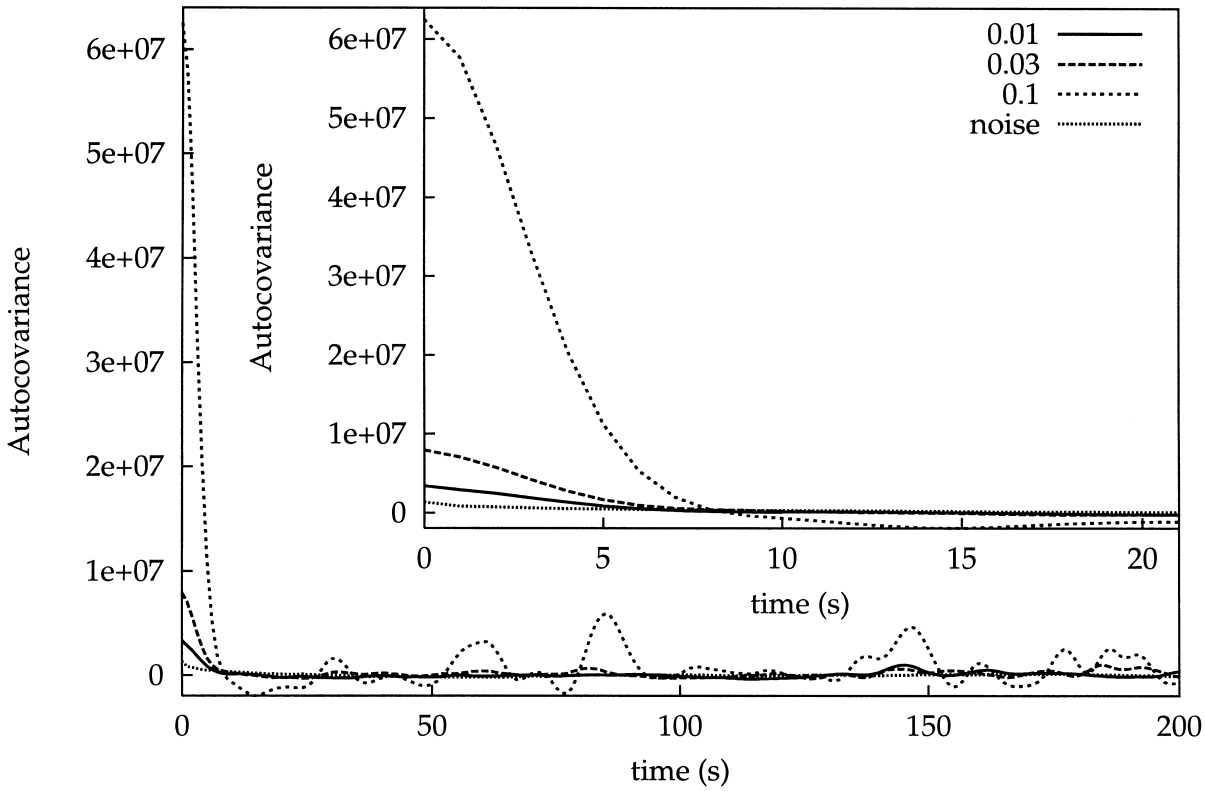

Fig. 3. Autocovariance function of the baseline noise and that of the chromatograms of small sample sizes.

recorded, and its power spectrum is plotted in Fig. 2. The power spectrum is plotted on a log–log scale in the insert. A straight line was fitted to the power spectrum plotted on a logarithmic scale. The slope of the line is −0.91. That value is very close to the theoretical value −1 of the pure flicker noise. This result confirms that the baseline noise of the GC–MS system is flicker noise, and noise will disturb the

Table 2
Numerical results of fitting the power spectrum to chromatograms of varying sample sizes[a]

c (mg/ml)	σ_h^2/a_h^2	p	Model	m	σ (s)	σ_T, p	rss[b]
0.01	0.114	8	E	23	2.160	–	0.129
			U	21	2.291	–	0.066
			N	20	2.363	47.86	0.078
			G	19	2.393	1.622	0.109
0.03	0.427	13	E	27	1.988	–	0.043
			U	27	1.978	–	0.107
			N	27	1.966	71.00	0.038
			G	29	1.932	0.822	0.041
0.1	0.908	19	E	33	1.891	–	0.030
			U	30	1.953	–	0.047
			N	32	1.903	53.64	0.028
			G	30	1.964	1.441	0.027
0.3	2.321	62	E	78	1.968	–	0.021
			U	53	2.151	–	0.048
			N	74	1.992	10.80	0.023
			G	83	1.951	0.903	0.020
1	2.811	66	E	99	2.003	–	0.055
			U	92	2.091	–	0.059
			N	90	2.108	13.09	0.054
			G	89	2.090	1.980	0.052
3	2.363	81	E	151	2.320	–	0.028
			U	128	2.561	–	0.044
			N	145	2.367	12.29	0.030
			G	149	2.331	1.025	0.029
10	2.107	94	E	151	3.000	–	0.041
			U	120	3.453	–	0.063
			N	146	3.050	12.96	0.045
			G	148	3.006	1.059	0.042
30	1.499	95	E	135	4.671	–	0.080
			U	101	5.386	–	0.112
			N	127	4.730	14.24	0.087
			G	145	4.599	0.822	0.079
100	0.792	69	E	126	9.162	–	0.099
			U	78	10.460	–	0.124
			N	136	8.503	15.23	0.101
			G	111	9.259	1.268	0.100

[a] σ_T and p are the fitting parameters of the normal and the gamma distribution, respectively (see Table 1).
[b] rss stands for the residual sum of squares.

power spectrum and the autocovariance function of the chromatograms at small sample sizes.

The autocovariance functions of three chromatograms of small sample sizes and that of the noise are plotted in Fig. 3. From the visual inspection of Fig. 3 we expect that noise has a disturbing effect only on the smallest sample sizes.

The total ion current chromatograms recorded at many sample sizes were subjected to Fourier analysis. Eq. 13 was fitted to the power spectra of the experimental chromatograms for all four interval models given in Table 1. The numerical technique is detailed elsewhere [15].

The number of peaks counted in the chromatograms, as well as the number of single components and the average peak width – determined by a nonlinear curve fitting procedure, as described above – are summarized in Table 2. On the basis of the best fitting model for retention pattern, the number of components and the average peak width were determined. The number of peaks counted in the chromatogram, as well as the number of single components and the average peak width determined from the best fitting model, are plotted in Fig. 4.

As it is expected, both the number of peaks and the number of single components rise when the sample concentration is increased, although in the 0.01–0.1 mg/ml range there is no real change. In that range, only about a dozen peaks can be counted. In the 0.1–3 mg/ml range, the number of counted peaks suddenly increases, and so does the number of estimated single components. The higher the sample concentration, the more detectable peaks emerge out of the baseline noise. The average peak width is nearly constant in that region.

When the sample concentration is further increased, the average peak width suddenly rises, and the number of single components start to drop. At

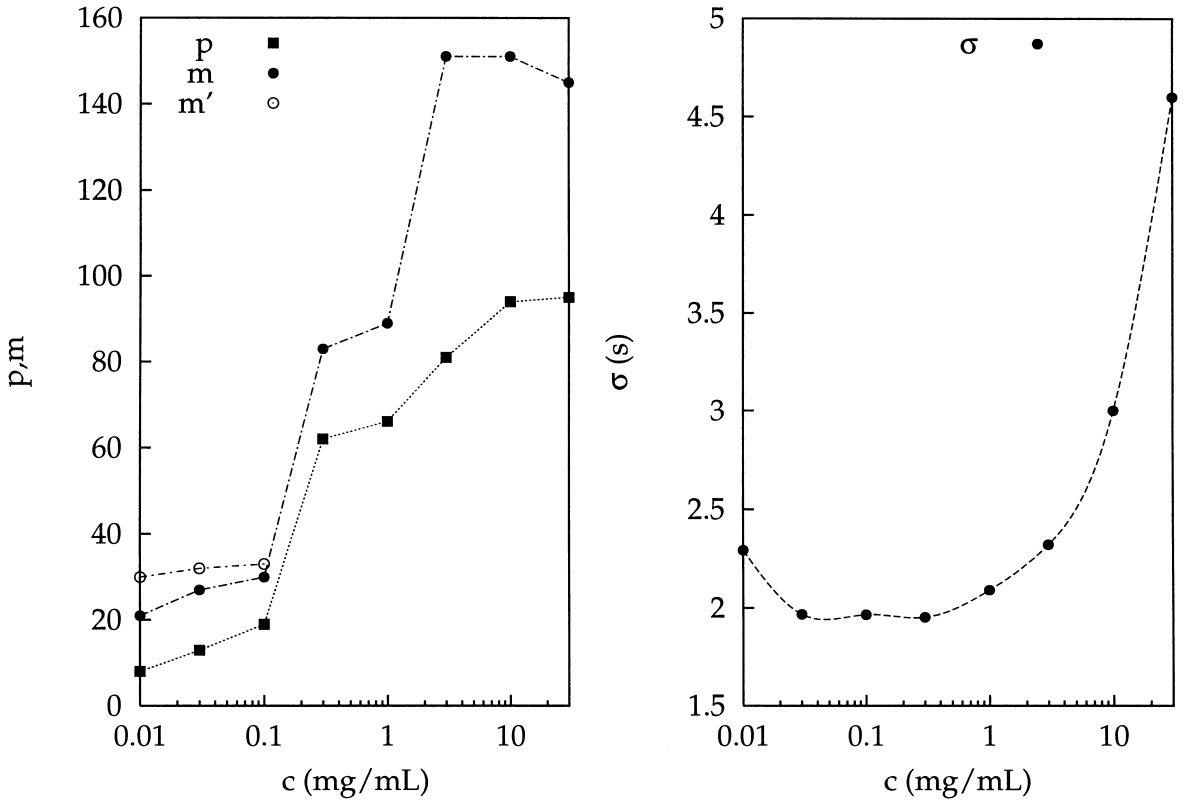

Fig. 4. Plot of the number of counted peaks, number of single components, and corrected number of single components against the sample size (left); plot of the average peak width against the sample size (right).

that point, for some sample components the column is overloaded. One particular peak of a major component is plotted in Fig. 5. The overload effect is clearly seen on the peak shape. As the relative concentration of this component is high, its peak shape is distorted already at smaller sample concentrations than one would conclude from Fig. 4. The plot of the number of single components and average peak width in Fig. 4 shows at what sample size the 'global overload' of the column is reached.

Since the concentration of the sample components spans over several orders of magnitude in a multicomponent mixture, major components reach the nonlinear section of their isotherms first, therefore their band distorts first. For minor components, the effect of concentration overload can only be observed at higher sample sizes. The effect of 'global overload' is determined from the normalized autocovariance function. Global overload is reached when due to the sample concentration, the shape of the autocovariance function, and accordingly that of an average peak alters. This effect of large sample size on the shape of the autocovariance function was already described by Pietrogrande et al. [17].

The effect of global column overload on the normalized autocovariance functions is illustrated in Fig. 6. The plot of the normalized autocovariance functions demonstrates that from a statistical point of view there is no significant difference between the chromatograms as long as the sample size is not higher than 1 mg/ml. At large sample sizes, the shape of the autocovariance function changes when the region of global column overload is reached. The increase of the average peak width broadens the autocovariance function. Therefore, the autocovariance function of a chromatogram is a very simple and useful means to detect global column overload.

The effect of baseline noise on the autocovariance function can simply be eliminated. As noise and

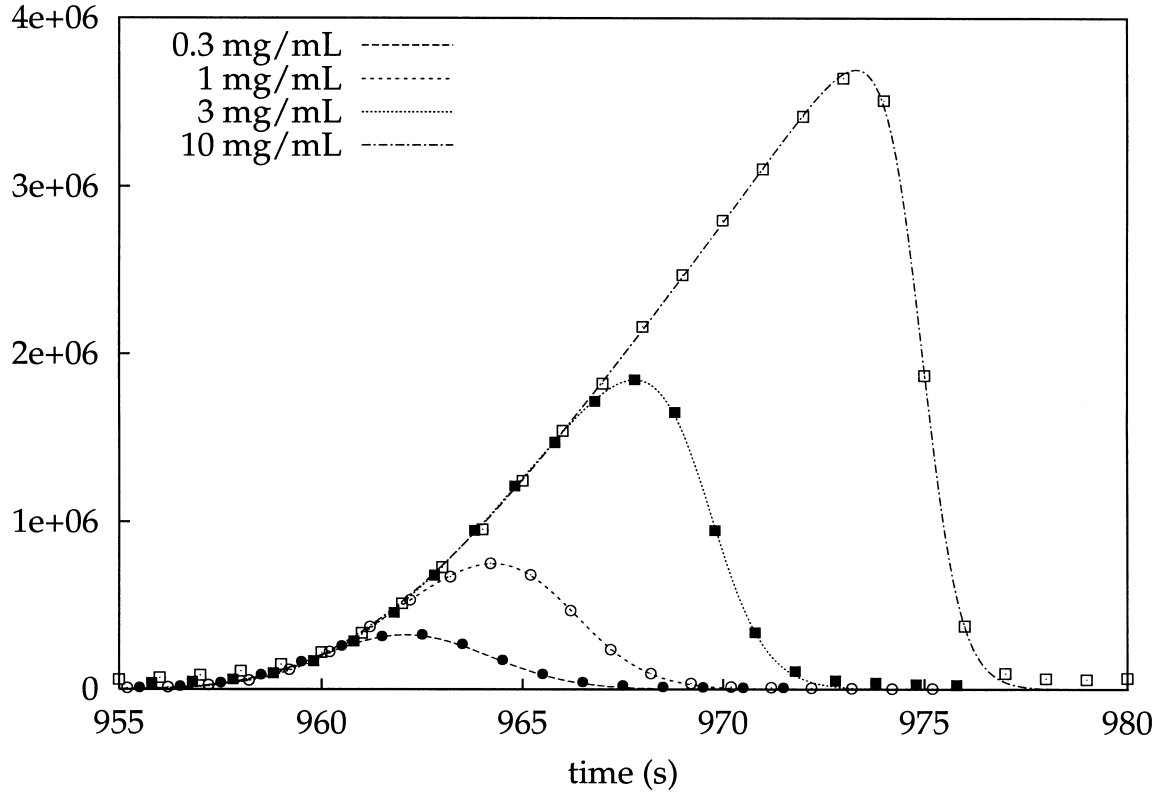

Fig. 5. Effect of sample concentration on one particular peak in the chromatogram of the diesel fuel.

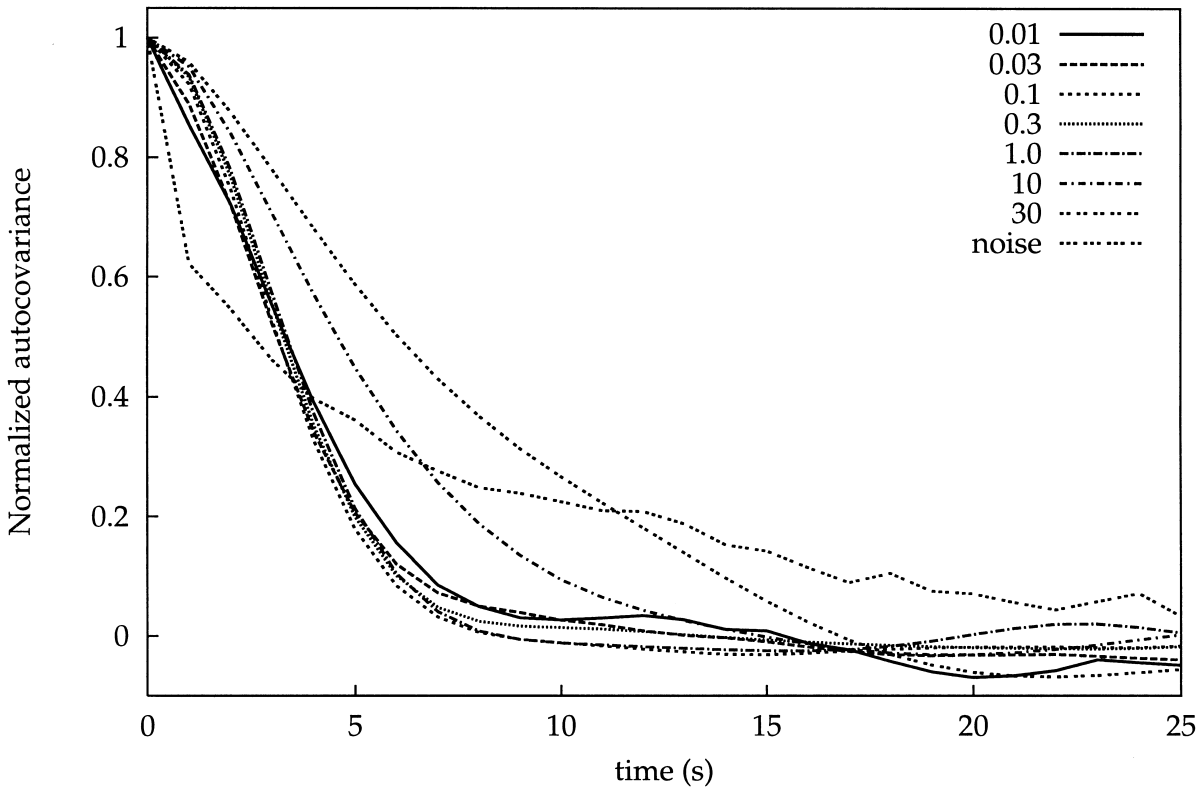

Fig. 6. Normalized autocovariance functions at different sample sizes.

signal are independent of each other, the power spectra – and therefore the autocovariance functions – of noise and signal are additive. For this reason, we can subtract the autocovariance function of the noise from that of the noisy chromatogram, and use that corrected autocovariance function for more accurate parameter estimation. The results of the more accurate estimation are also plotted in Fig. 4 (m'). As it is expected, the effect of noise lessens as sample size increases, therefore the difference between the estimated values of m' and m diminishes as sample concentration increases. For this reason m' has not been determined for sample concentrations higher than 0.1 mg/ml.

The peak height dispersion (σ_h^2/a_h^2) is also reported in Table 2, for each sample size. At the smallest sample size the peaks of only a few major components can be distinguished from baseline noise. As the major components do not span over a wide concentration range, the peak height dispersion is very small in this instance. The concentration range of the detectable components increases with increasing sample size, and so does the peak height dispersion. When the sample size is in the domain where the on-set of the global overload can be observed, the peak height dispersion starts to decrease. This is, probably, due to the peak-width increase of the major components, which suppresses several detectable minor components.

The fitting parameters of the normal and the gamma distribution are also listed in Table 2. For both distributions the extra parameters take values that are close to the characteristic values of the disordered Poisson chromatogram ($\sigma_T = T$ and $p = 1$). Accordingly, we have no reason to assume any deviation from the Poisson model.

For this particular example investigated here, we can conclude that the optimum sample concentration is found at around 3 mg/ml. Smaller samples would lose several components in the baseline noise, larger

samples would destroy the column efficiency due to column overload.

Acknowledgements

This work was supported in part by grant T 025458 of the Hungarian National Research Found (OTKA) and grant FKFP 609/1997 of the Hungarian Ministry of Education.

References

[1] J.M. Davis, Adv. Chromatogr. 34 (1994) 109.
[2] A. Felinger, Adv. Chromatogr. 39 (1998) 201.
[3] A. Felinger, Data Analysis and Signal Processing in Chromatography, Elsevier, Amsterdam, 1998.
[4] J.M. Davis, J.C. Giddings, Anal. Chem. 55 (1983) 418.
[5] L.J. Nagels, W.L. Creten, P.M. Vanpeperstraete, Anal. Chem. 55 (1983) 216.
[6] D.P. Herman, M.-F. Gonnord, G. Guiochon, Anal. Chem. 56 (1984) 995.
[7] D. Bowlin, C. Hott, J.M. Davis, J. Chromatogr. A 677 (1994) 307.
[8] A. Felinger, Anal. Chem. 67 (1995) 2078.
[9] M.C. Pietrogrande, F. Dondi, A. Felinger, J.M. Davis, Chemometr. Intell. Lab. Syst. 28 (1995) 239.
[10] M. Martin, D.P. Herman, G. Guiochon, Anal. Chem. 58 (1986) 2200.
[11] A. Felinger, L. Pasti, F. Dondi, Anal. Chem. 62 (1990) 1846.
[12] F. Dondi, A. Bassi, A. Cavazzini, M.C. Pietrogrande, Anal. Chem. 70 (1998) 766.
[13] R.N. Bracewell, The Fourier Transform and Its Applications, 2nd ed, McGraw-Hill, New York, 1986.
[14] A. Felinger, L. Pasti, P. Reschiglian, F. Dondi, Anal. Chem. 62 (1990) 1854.
[15] A. Felinger, L. Pasti, F. Dondi, Anal. Chem. 64 (1990) 2164.
[16] F. Dondi, A. Betti, L. Pasti, M.C. Pietrogrande, A. Felinger, Anal. Chem. 65 (1993) 2209.
[17] M.C. Pietrogrande, F. Dondi, A. Felinger, J. High Resol. Chromatogr. 19 (1996) 327.