SEPARATION OF ANALYTES AS A SOURCE OF INFORMATION*

Karel ECKSCHLAGER

Institute of Inorganic Chemistry, Czechoslovak Academy of Sciences, 250 68 Řež

Received October 17, 1988
Accepted November 24, 1988

Dedicated to Jaroslav Janák, Corresponding Member of the Czechoslovak Academy of Sciences, on the occasion of his 65th birthday.

It is shown how the amount of information gained by separation and detection or determination of several components depends on the selectivity of the analytical procedure in question. The selectivity is characterized in terms of the information entropy. Rules useful in the optimization of multicomponent analysis are derived from the results of information field mapping.

Most important methods of multicomponent molecular analysis, separation procedures often provide a considerable amount of information. This amount will apparently be the higher the more perfect is the selectivity of determination and the higher number of components can be determined simultaneously. Selectivity of quantitative multicomponent analysis was defined by H. Kaiser[1] as early as 1972; later his definition was modified by Bergmann and coworkers[2,3]. Otto and Wegescheider[4] define selectivity in terms of the so-called condition numbers. An alternative approach to the characterization of the selectivity of qualitative and/or quantitative multicomponent analysis is in terms of the information entropy[5,6].

With respect to the selectivity of simultaneous determination of several ($n \geqq 2$) components, a continuous series of cases can occur, from the optimum situation where the analytical signal is constituted by n isolated peaks, each component being associated with a single peak, to the case of a continuous signal where, although peaks corresponding to the individual components cannot be resolved, the components can be determined with a higher or lower degree of uncertainty, such as in some field flow fractionation methods[7].

Previously[8] the information field was mapped for quantitative single-component analysis, and some dependences of utility in the practice were obtained. The present paper gives results derived from a similar mapping for multicomponent analysis. Only the information gain associated with the component separation is considered, neglecting whether the signal is employed for a qualitative, identification or quantitative multicomponent analysis.

* Part XXII in the series Theory of Information as Applied in Analytical Chemistry; Part XXI: Collect. Czech. Chem. Commun. *53*, 3021 (1988).

THEORETICAL

Instrumental methods of multicomponent analysis provide information on the sample composition so that the a priori information (sample contains a maximum of n different components) reduces to an a posteriori uncertainty, given by the fact that the identities of the components A_i and their amounts in sample X_i ($i = 1,2, \ldots, n$) are only determined with an uncertainty. Identities of analytes A_i and their concentrations X_i cannot be determined directly but only through measurement of a quantity that is related with the kind and amounts of analytes or only with the amounts of the mutually separated analytes. Thus, the a posteriori uncertainty of multicomponent instrumental analysis is given by

a) the uncertainty of the component separation by the separating operation;

b) the uncertainty of measurement of the positions and intensities of the more or less overlapping peaks;

c) the uncertainty associated with the transformation of information about the peak positions and intensities to information about the analyte identities A_i and concentrations X_i ($i = 1, 2, \ldots, n$).

The two last contributions will be omitted from the ensuing treatment (in fact, the uncertainty of determination of the analytical signal position and intensity has been studied previously[9]). Hence, we are going to deal with information gained by the separation of n components solely, bearing in mind that the actual information gained from the identification, detection or determination of the components is (occasionally very markedly) lower than as obtained in this treatment; this is the case particularly if the uncertainty of the results is further contributed to by the numerical resolution of overlapping signals.

The a posteriori uncertainty after separation of n components can be expressed by Shannon's entropy

$$H(a_{ij}) = -\sum_{i=1}^{n} a_{ij} \operatorname{ld} a_{ij} \quad (j = \text{const}), \tag{1}$$

where ld denotes the binary logarithm; thus the uncertainty and information gain will be expressed in bits. When using Eq. (*1*), the condition

$$\sum_{i=1}^{n} a_{ij} = 1 \tag{2}$$

($i = 1, 2, \ldots, n$) must be fulfilled. We put $(-a_{ij} \operatorname{ld} a_{ij}) = 0$ for $a_{ij} = 0$. The way of determining the a_{ij} values is irrelevant in the treatment (precision is not included in the a posteriori uncertainty); let us only mention that for qualitative and identification analysis, a_{ij} is equal to $P(A_i \mid Z_j)$, the conditional probability that component A_i is present if a signal occurs in position Z_j, and for quantitative analysis a_{ij} is the relative partial sensitivity of determination of component A_i by signal measure-

ment in position Z_j. Details can be found in monograph[6], p. 29, 30 and 53. Basically, $a_{ij} \in \langle 0, 1 \rangle$ is the contribution of component A_i to the total signal in position Z_j $(j = 1, 2, \ldots, k, k \geqq n)$.

For the case where the maximum number of components present is known to be n, the a priori uncertainty is determined as the entropy in Eq. (*1*), for $a_{ij} = 1/n$, i.e. as $H(a_{ij}) = \mathrm{ld}\, n$.

The information content of a signal in position Z_j is

$$I_{sep}^{(j)} = \mathrm{ld}\, n + \sum_{i=1}^{n} a_{ij} \,\mathrm{ld}\, a_{ij} \quad (j = \mathrm{const}) \tag{3}$$

We have $0 \leqq I_{sep}^{(j)} \leqq \mathrm{ld}\, n$, the value being zero if no separation takes place and maximum, $I_{sep}^{(j)} = \mathrm{ld}\, n$, if the separation is complete. A measure of completeness of separation in multicomponent analysis where n components are determined by measuring the signal in positions Z_j $(j = 1, 2, \ldots, k, k \geqq n)$ is the information gain

$$I_{sep} = k \,\mathrm{ld}\, n + \sum_{i=1}^{n} \sum_{j=1}^{k} a_{ij} \,\mathrm{ld}\, a_{ij} \,. \tag{4}$$

In practice the positions Z_j correspond to the maxima of peaks of the individual components or, for a continuous signal, to certain components of the sample. For $k = n$ the matrix $\|a_{ij}\|$ is a square one, and for a perfectly selective procedure this is a diagonal unit matrix ($a_{ii} = 1$, $a_{ij} = 0$ for $j \neq i$); the information gain I_{sep} in Eq. (*4*) then attains its maximum, $I_{sep} = n \,\mathrm{ld}\, n$.

Information gained by component separation can be characterized by the mean value with respect to one component when measuring $k \geqq n$ peaks (or in $k \geqq n$ positions of a continuous signal); this is

$$I = (1/n)\, I_{sep} = (k/n) \,\mathrm{ld}\, n + (1/n) \sum_{i=1}^{n} \sum_{j=1}^{k} a_{ij} \,\mathrm{ld}\, a_{ij} \tag{5}$$

and lies in the region of $0 \leqq I \leqq (k/n) \,\mathrm{ld}\, n$. The value is zero if no separation takes place, and $I = \mathrm{ld}\, n$ if the selectivity is complete and $k = n$. The theoretical maximum, $I = (k/n) \,\mathrm{ld}\, n$ with $k > n$, will be hardly encountered in practice because measurements in $k > n$ positions, leading to an overdetermined system of equations in the calculation treatment, are accomplished only in the case of a few selective procedures where the second right-hand term in Eq. (*5*) has a rather high negative value.

It should be mentioned that multicomponent qualitative analysis procedures have also been characterized in terms of equivocation[6,10], which is the entropy expected for the conditional probability $P(A_i \mid Z_j)$. However, since this quantity can attain different values for the same matrix $\|a_{ij}\|$, hence, for identically selective procedures, it is not used here for characterization of the information gain associated with component separation.

RESULTS AND DISCUSSION

Information field mapping for the case of determination of n components by measurement in $k \geqq n$ signal positions was performed by evaluating the I values (Eq. (5)) for Gaussian and Lorentzian peaks which overlapped to various extent; a Hewlett–-Packard 9 825 computer was employed.

The peak profile for a unit half width ($b = 1$) is

$$y_G = y_{max} \exp(-2{\cdot}7726x^2) \tag{6}$$

for the Gaussian shape and

$$y_L = y_{max}[1 + (2x)^2]^{-1} \tag{7}$$

for the Lorentzian shape; $x = (z - z_{max})$, where z_{max} is the position where the peak attains its maximum value y_{max}. For a comparison of the two profiles and other details see ref.[11]. A continuous signal is the result of summation of a larger number of signals with small differences $\Delta = |z_{1,max} - z_{2,max}|$ where $z_{1,max}$ and $z_{2,max}$ are the positions of maxima of two neighbouring peaks 1, 2.

In the information field mapping, the elements of the matrix $\|a_{ij}\|$ were calculated as the conditional probabilities $P(A_i \mid Z_j)$ or $P(X_i \mid Z_j)$, i.e. probabilities that analyte A_i is present or that it is present in a concentration $X_i \geqq x_{DL}$ (x_{DL} is the detection

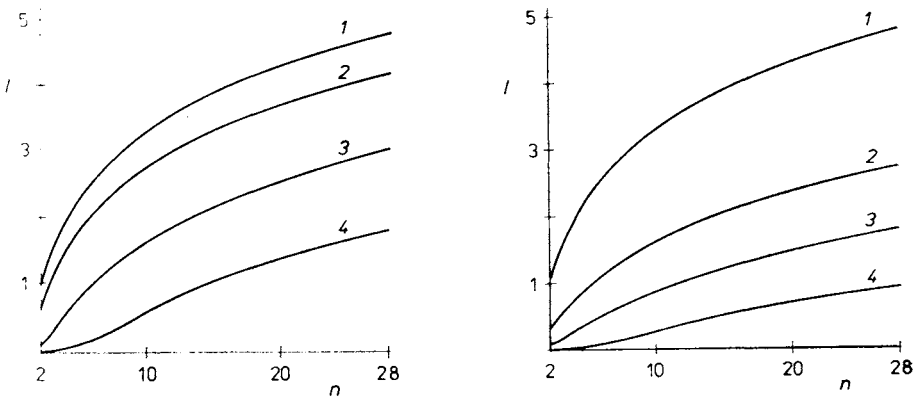

FIG. 1

Dependence of I on the number of separated components n for equidistant peak distribution and Gaussian (*a*) and Lorentzian (*b*) peak shapes. Curves: *1* perfectly selective procedure, *2* $\Delta = 1b$, *3* $\Delta = 0{\cdot}5b$, *4* $\Delta = 0{\cdot}2b$

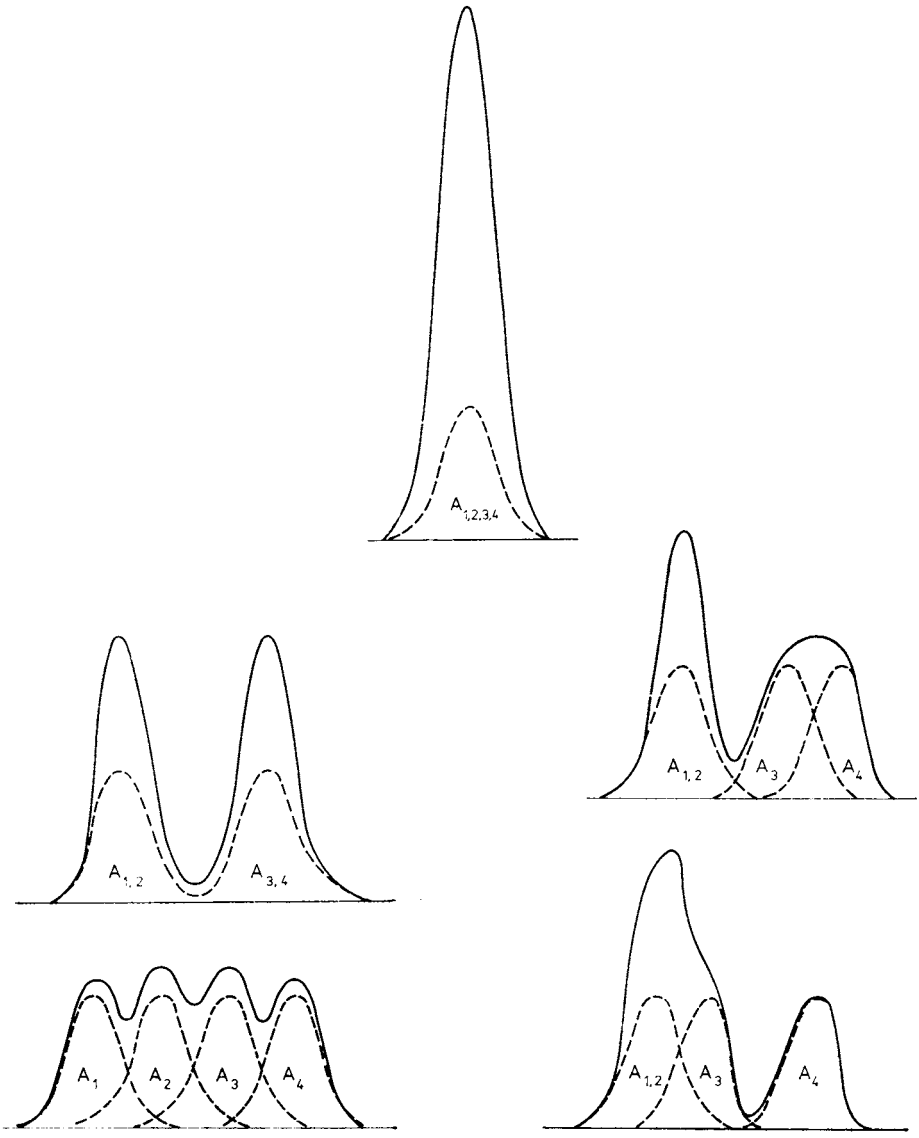

FIG. 2

Dependence of I on the separation of four Gaussian peaks. *a* no separation, $I = 0{\cdot}000$; *b* two and two components cannot be resolved, $I = 1{\cdot}000$ bit; *c* two components do not separate, two separate incompletely, $I = 1{\cdot}087$ bit; *d* two components do not separate, the third is incompletely separated from the two previous components, the fourth is separated completely, $I = 1{\cdot}193$ bit; *e* the four components are mutually incompletely separated, $I = 1{\cdot}264$ bit. For a perfect separation of the four components, $I = \mathrm{ld}\ 4 = 2{\cdot}000$ bit

limit) if a signal occurs in position Z_j. These probabilities can be determined readily by means of the Bayes relation; details can be found, e.g., in ref.[6]

The results of the computational information field mapping for qualitative or quantitative multicomponent analysis can be summarized as follows:

1) Information gained by separation of n components increases with increasing n and is the higher the higher is (Δ/b), where b is the half peak width. For the same n and (Δ/b), this information is higher for Gaussian peaks than for Lorentzian peaks, due to the fact that the latter affect their neighbours to a greater extent (see ref.[11]). The dependence of I in Eq. (5) on n for various Δ values and for equidistant peak spacings is shown in Fig. 1.

2) For a sufficiently high Δ the peaks are virtually separate and the information gained by separation is maximum ($I = \mathrm{ld}\, n$); even for a continuous signal the information gain is appreciable, particularly in the separation of a large number of components. Selectivity should be regarded as a continuous property, although in analytical signal treatment it is convenient to distinguish between complete and incomplete selectivity, where signal overlap can or cannot be neglected (refs[11,12]).

3) For a given number of components the amount of information obtained by their separation is also determined by the signal overlap and by the peak distribution over the region examined, and also by the positions the peaks are sought or measured, their number and choice. Measurement in the positions of the expected z_{max} may not be optimum for the I value according to Eq. (5); however, measurement on steep peak sides is unfavourable as far as accuracy is concerned[11]. Some dependences are apparent from Figs 2 and 3.

In addition to these conditions relevant to the information gain associated with the separation of several components, other factors affecting the accuracy of signal position and intensity determination also contribute in practice[11].

The author wishes to thank Dr Jiří Fusek for setting up the requisite computer programs and performing the computer calculations.

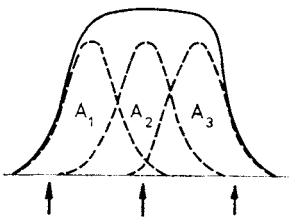

FIG. 3

Effect of position Z_j, where the signal intensity is measured after the separation of three components. *a* Measurement in positions z_{max}, $I = 0{\cdot}912$ bit; *b* for A_1 and A_3, measurement in positions shown by arrows, where the signal of component A_2 does not interfere appreciably, $I = 1{\cdot}21$ bit. For a perfect separation of the three components, $I = 1{\cdot}585$ bit

REFERENCES

1. Kaiser H., Fresenius Z. Anal. Chem. *260*, 252 (1972).
2. Junker A., Bergmann G.: Fresenius Z. Anal. Chem. *272*, 267 (1974).
3. Bergmann G., Oepen B., Zinn P.: Anal. Chem. *59*, 2522 (1987).
4. Otto M., Wegscheider W.: Anal. Chim. Acta *180*, 445 (1986).
5. Eckschlager K.: Collect. Czech. Chem. Commun. *47*, 1580 (1982).
6. Eckschlager K., Štěpánek V.: *Analytical Measurement and Information.* Research Studies Press, Letchworth 1985.
7. Janča J.: *Field-Flow Fractionation.* Dekker, New York 1987.
8. Eckschlager K., Fusek J.: Collect. Czech. Chem. Commun. *53*, 3021 (1988).
9. Eckschlager K.: Collect. Czech. Chem. Commun. *46*, 478 (1981).
10. Cleij P., Dijkstra A.: Fresenius Z. Anal. Chem. *289*, 97 (1979).
11. Doerffel K., Eckschlager K.: *Optimale Strategien in der Analytik.* VEB Deutscher Verlag für Grunstoffindustrie, Leipzig 1981.
12. Doerffel K.: Fresenius Z. Anal. Chem. *330*, 24 (1988).

Translated by P. Adámek.