

INFORMATION CONTENT OF INSTRUMENTAL ANALYSIS RESULTS*

K. ECKSCHLAGER

*Institute of Inorganic Chemistry,
Czechoslovak Academy of Sciences, 250 68 Prague - Řež*

Received July 3rd, 1975

The information content is calculated for the results of instrumental analysis, where signals in the form of a peak indicate by their position the presence and by their intensity (maximum height or area) the concentration of individual components of the analysed sample.

In our preceding work the information content of a qualitative proof¹ or of various cases of a quantitative determination^{1,3} was defined. We assumed always that the concentration range in which the component to be determined is present is known beforehand^{1,3} or that its content is even determined by a preliminary analysis². In the analytical practice, however, it is often the case that a sample of an entirely unknown qualitative composition is to be analysed by an instrumental method in which a peak-shaped signal in a position z_i of an intensity $y_i \geq y_{\min}$ (y_{\min} is the least signal distinguishable from the zero noise) indicates the presence of a component i ($i = 1, 2, \dots, k$), its intensity $y_i \in \langle y_{\min}, y_{\max} \rangle$ being proportional to the actual concentration, ξ_i , of this component. If no signal distinguishable from the zero noise is present in the position z_j corresponding to the component j ($j = 1, 2, \dots, l$), it follows that the j -th component is present in a lower concentration than would correspond to the signal y_{\min} , *i.e.*, in a lower concentration than the determination limit of the method used, $x_j < x_{\min}$. The dependence of the signal intensity, which is a random variable with the mean value $\mu_{y,i}$, on the actual content of the determined component is expressed as $\mu_{y,i} = f_1(\xi_i)$. The result of the analysis, $x_i = f_2(y_i)$, where the dependence f_2 is found, as a rule, empirically (by calibration), is then a randomly distributed quantity with a probability density $p(x) = (1/\sigma_i \sqrt{2\pi}) \exp[-(x_i - \mu_i)^2/2\sigma_i^2]$, where μ_i is the most probable result. It is, as a rule, estimated from n parallel determinations as the average: $\hat{\mu}_i \equiv \bar{x}_i = \sum_{q=1}^n x_{iq}/n$.

We shall express the information content of instrumental analysis, in which either qualitative or both qualitative and quantitative analyses were performed of an entirely unknown sample or of a sample with an assumed highest possible content of the components.

THEORETICAL

The information content of a qualitative proof was defined earlier with the use of the Brillouine measure^{1,4} as $I = \log_b(n_0/n)$, where the number of possible but unidenti-

* Part VIII in the series Theory of Information as Applied to Analytical Chemistry; Part VII: This Journal 40, 3627 (1975).

fied components of the analysed sample prior to analysis is n_0 and after the analysis n . The base, b , of the logarithm modifies the units in which the information content is expressed¹. In the case of instrumental analysis, it is more appropriate to express the highest obtainable, *i.e.*, potential information content of a qualitative proof with the use of the divergence measure^{2,4} as

$$I(p, p_0) = \log_2 \frac{z_{\max} - z_{\min}}{\Delta_1 z} \equiv \log_2 N_1 \quad (1)$$

for $z \in \langle z_{\min}, z_{\max} \rangle$, where $\Delta_1 z$ denotes the smallest distance of the maximums of the signals necessary for their distinguishing, \log_2 binary logarithm and N_1 maximum number of possible distinguishings. Analogously, with the use of the Shannon relation⁴, Dupuis and Dijkstra⁵ define the information content of a qualitative proof carried out by gas chromatography. The number of distinguishings N_1 in Eq. (1) is for the given interval $\langle z_{\min}, z_{\max} \rangle$ given only by the value of $\Delta_1 z$, which, of course, depends on the form of the signal. In analytical practice, we normally consider⁶ symmetrical signals of the Lorentz type, $y = h/[1 + (2z/a)^2]$, where a is the half-width of the signal, or of the Gauss type, $y = h \exp[-\alpha(z/\sigma)^2]$, where the half-width is given as $a = 2\sigma(\ln 2/\alpha)^{1/2}$, and h is the maximum peak height in both cases. The perfectness of the resolution of two peaks can be for the purpose of a qualitative proof characterized by the relative depth of the minimum between them, $M = (y_2 - y_1)/y_2$ for $y_2 \leq y_1$ (Fig. 1). Doerffel⁷ assumes that it is sufficient to distinguish both signals on the level $3\sigma_N$ from the noise, where σ characterizes the "white noise", and considers sufficient for $\sigma_N/h = 0.05$ such a resolution, where $M \geq 0.2$. With symmetrical peaks of the same maximum height ($h_1 = h_2$) and of the

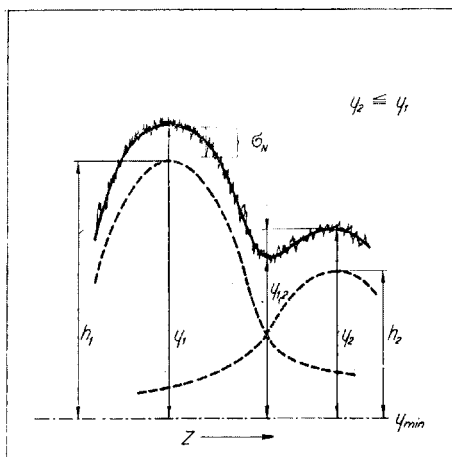


Fig. 1
Distinguishing of Two Partially Overlapping Signals
Dashed curves: Real form of the signals.
Solid curve: Resulting form from the superposition of both signals.

same half-width ($a = a_1 = a_2$) we have for $M = 0.2$ $\Delta_1 z = 1.06a$ for the Lorentz form and $\Delta_1 z = 1.16a \approx 2.73\sigma$ for the Gauss form with $\alpha = 1/2$; the value of $\Delta_1 z$ is larger when the maximum heights of both peaks are not the same. However, under the most unfavourable conditions is $\Delta_1 z \geq 2a$ for a symmetrical peak of the Lorentz or Gauss type. It is obvious that for larger $\Delta_1 z$ values the information content of the instrumental analysis will be smaller.

If qualitative and quantitative instrumental analyses of an entirely unknown sample are carried out simultaneously, it is necessary to express the information content of the quantitative determination with the aid of the Kullback divergence measure^{2,4} $I(p, p_0) = \int_{-\infty}^{\infty} p(x) \log_2 [p(x)/p_0(x)] dx$ with the use of the probability density of the result prior to the analysis, $p_0(x)$, or after the analysis, $p(x)$, and not analogously to Eq. (1), where we used the position of the signal with the probability density of the distribution of the signal intensity y_i . In this case is the value of x_i influenced by the function f_1 (hence, in substance by the sensitivity of the determination) and by the dependence f_2 determined by calibration. The standard deviation σ_i , which characterizes the accuracy of the x_i value, is also influenced by the functions f_1 and f_2 . The value of μ_i is besides dependent on the number of parallel determinations, n . The probability density of the distribution of the results, which we assume before the analysis, $p_0(x)$, corresponds to a rectangular distribution and is given as $p_0(x) = 1/(x_1 - x_0)$ for $x \in \langle x_0, x_1 \rangle$, where x_0 is the lowest and x_1 the highest assumed content of the determined component. The values of x_0 and x_1 are, of course, not related to y_{\min} and y_{\max} , which are given by the properties of the analytical method or apparatus; the only condition is that the signal intensity y_1 corresponding to the value of x_1 must not exceed y_{\max} , i.e., x_1 must be determinable by the analytical method used. On the other hand, x_0 can be smaller than the determination limit x_{\min} , for example, it can be equal to zero. In the definition of the probability density after the analysis, $p(x)$, two possible cases must be distinguished:

a) If $y_1 \geq y_{\min}$, i.e., if we can determine the i -th component quantitatively, then $p(x)$ is the probability density of normally distributed results and the information content of the determination of the i -th component is, e.g., according to ref.^{1,4}

$$I(p, p_0)_i = \log_2 \frac{(x_{i,1} - x_{i,0}) \sqrt{n}}{\sigma_i \sqrt{2\pi e}} \quad (2)$$

b) If the signal intensity y_j is smaller than y_{\min} for such a position of the signal z_j which would correspond to the component j ($j = 1, 2, \dots, 1$), the content x_j is smaller than the determination limit $x_{j,\min}$. Then $p(x) = 1/x_{j,\min}$ for $x_j \in \langle 0, x_{j,\min} \rangle$ and the information content is^{1,3,4}

$$I(p, p_0)_j = \log_2 [(x_{j,1} - x_{j,0})/x_{j,\min}] \quad (3)$$

or for the most frequent case, where $x_{j,0} = 0$ and $x_{j,1} \geq x_{j,\min}$

$$I(p, p_0)_j = \log_2 (x_{j,1}/x_{j,\min}). \quad (4)$$

The total, actual information content of the qualitative as well as quantitative instrumental analysis is then given as

$$I(p, p_0) = \sum_{i=1}^k \log_2 \frac{(x_{i,1} - x_{i,0}) \sqrt{n}}{\sigma_i \sqrt{2\pi e}} + \sum_{j=1}^l \log_2 (x_{j,i}/x_{j,\min}). \quad (5)$$

When performing a quantitative analysis, we must consider the maximum number of possible distinguishings, $N_2 = (z_{\max} - z_{\min})/\Delta_2 z$, where $\Delta_2 z$ is the distance corresponding to the isolated signal for the case of measuring the maximum peak height h or its area in the sense in which Doerffel⁶ defines an isolated signal for both cases. It is obvious that $k + l \leq N_2$, since every distinguishable position of the signal z need not correspond to a real component of the analysed sample. It should be noted that the value of $\Delta_2 z$ is always much larger than $\Delta_1 z$; *e.g.*, if we measure the area of a Lorentz type peak, then for $h_1 = h_2$ is $\Delta_2 z \approx 8a$ and can be even much larger in an unfavourable case⁶.

Eq. (5) applies quite generally and does not involve any restrictive assumptions. The maximum possible content, *i.e.*, the largest amount of information obtainable by performing both qualitative and quantitative analyses simultaneously by emission spectrography was given by Danzer⁸. The different meaning of Eq. (5) and the equation for the maximum (potential) information content after Danzer⁸ was pointed out by Eckschlager⁹.

REFERENCES

1. Eckschlager K.: This Journal 36, 3016 (1971).
2. Eckschlager K., Vajda I.: This Journal 39, 3076 (1974).
3. Eckschlager K.: This Journal 40, 3627 (1975).
4. Eckschlager K.: Chem. Listy 69, 810 (1975).
5. Dupuis F., Dijkstra A.: Anal. Chem. 47, 379 (1975).
6. Doerffel K.: Chem. Anal. (Warsaw) 17, 615 (1972).
7. Doerffel K.: Chem.-Ing.-Tech. 25, 9 (1973).
8. Danzer K.: Z. Chem. 15, 158 (1975).
9. Eckschlager K.: Z. Chem. 16, 111 (1976).

Translated by K. Micka.