# INFORMATION CONTENT OF ANALYTICAL SIGNALS OF INSTRUMENTAL METHODS*

Karel ECKSCHLAGER

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

Received August 24th, 1979

Instrumental analytical methods are studied from the point of view of the information content of the signal, whose intensity or both position and intensity are sources of relevant information. The information content is expressed with the aid of the divergence measure in units defined by the unit isoinform. If both the intensity and the position of the signal are sources of information, it is assumed that they are mutually independent or that the measurement of the signal intensity is carried out with accuracy which depends on the signal position.

Instrumental analytical methods can be generally defined as processes of obtaining relevant information about the chemical composition of matter or about the structure of chemical compounds. Such a process proceeds in two steps: in the first one, the information is created and coded in the form of a signal, in the second one it is decoded. The first step (creating information) proceeds with an instrumental analytical method in a real system (apparatus), namely so that a sample enters the system whose output gives a signal bearing information about the composition or structure of the sample. The signal, which is realized by a change of a physical state (usually electric current or voltage), is often two-dimensional, *i.e.*, it has a certain intensity $y_i$ in a certain position $z_i$. In instrumental methods of chemical analysis, the signal intensity $y_i$, is usually the source of relevant information; the signal position $z_i$ is often *a priori* known or the assignment of the signal position to a certain component is done so that the value of $z_i$ is irrelevant as a source of information. In instrumental methods of structural analysis and sometimes also in chemical quantitative analysis both the intensity $y_i$ and the position $z_i$ of the signal can be sources of relevant information. The information content of instrumental analytical methods where only the signal intensity is the source of relevant information has already been studied[1-3].

The present work is devoted to the case where both the intensity $y_i$ and the position $z_i$ of the signal are sources of relevant information and both simultaneously control the information content of the results.

## THEORETICAL

The information content of the results of chemical or structural analyses can be expressed in different ways; the divergence measure proved most suitable for this purpose[4], therefore we shall use it in the present work. The information content (gain)

* Part XV in the series Theory of Information as Applied to Analytical Chemistry; Part XIV: This Journal *45*, 2516 (1980).

expressed by the divergence measure is conditioned only by the *a priori* and *a posteriori* probability distributions. The former characterizes preliminary assumptions about the unknown but fixed value $X_i$ (content of the determined component, measured physical quantity) for the i-th component ($i = 1, 2, \ldots, k$), or about the position $Z_i$ and the intensity $Y_i$ of the signal, and has a probability density $p_0(x)$, $p_0(y)$, and $p_0(z)$ respectively. The result obtained by an analysis or a measurement is represented by a continuous random variable $\xi_i$, which takes on values $x_{i,j}$ ($i = 1, 2, \ldots, k$; $j = 1, 2, \ldots, n$, where $n$ is the number of repetitions), the signal intensity $\eta_i$ takes on values $y_{i,j}$ and the position $\zeta_i$ takes on values $z_{i,j}$. The probability distributions of these random variables are described as *a posteriori* distributions with probability densities $p(x)$, $p(y)$, and $p(z)$ respectively. For each component $i$ we have conditions $p_0(x) > 0$ and $p(x) \geqq 0$ for $x \in \langle x_1, x_2 \rangle$, where $x_1$ is the lowest and $x_2$ the highest value which, according to our presumptions, the true content $X$ can assume. When we consider the intensity and the position of the signal, then $p_0(y) > 0$ and $p(y) \geqq 0$ for $y \in \langle y_{\min}, y_{\max} \rangle$ and $p_0(z) > 0$, $p(z) \geqq 0$ for $z \in \langle z_{\min}, z_{\max} \rangle$, where $y_{\min}$, $z_{\min}$ are the smallest and $y_{\max}$, $z_{\max}$ the highest values of the intensity and the position, which can be measured or recorded by the apparatus used. The relation between $x_1$, $x_2$ and $y_{\min}$, $y_{\max}$ is in substance given by the sensitivity of the apparatus. For the simplest case of a linear dependence $y = bx$, where $b = \mathrm{d}y/\mathrm{d}x = \text{const.}$ in the whole interval $\langle x_1, x_2 \rangle$, this relation is illustrated in Fig. 1.

With the use of the divergence measure, when we consider the true content $X_i$ and the information about its value $\xi_i$ obtained by an instrumental method, the information content is given as

$$I(a, p_0) = \int_{x_1}^{x_2} p(x) \ln \frac{p(x)}{p_0(x)} \, \mathrm{d}x \,. \qquad (1)$$

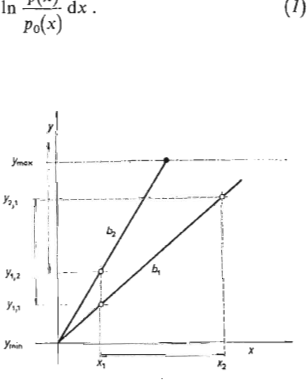


FIG. 1

Linear Depedence $y = b(x)$, where $b = \mathrm{d}y/\mathrm{d}x = \text{const.}$ in the whole Interval $\langle x_1, x_2 \rangle$.

The *a priori* probability distribution will depend on our preliminary knowledge of the value of $X_i$. If we do not know anything more than an interval, in which the value of $X_i$ can lie anywhere with the same probability, we use the continuous rectangular distribution:

$$p_0(x) = \begin{cases} 1/(x_2 - x_1) & \text{for } x \in \langle x_1, x_2 \rangle \\ 0 & \text{otherwise} \end{cases} \qquad (2a)$$

If we assume that the fixed unknown quantity $X_i$ is equal to the values of $\mu_0^{(x)}$ which lies in the interval $\langle x_1, x_2 \rangle$, namely $\mu_0^{(x)} = 0{\cdot}5(x_1 + x_2)$, we can consider the *a priori* distribution to be normal with a variance $(\sigma_0^{(x)})^2$, where $\sigma_0^{(x)} = (x_2 - x_1)/K$, hence

$$p_0(x) = \frac{K}{(x_2 - x_1)\sqrt{(2\pi)}} \exp\left[-\frac{1}{2}\left(\frac{K(x - \mu_0^{(x)})}{x_2 - x_1}\right)^2\right]. \qquad (2b)$$

For $K \geqq 6$ we have $\int_{x_1}^{x_2} p_0(x)\,dx \approx 1$, $\int_{x_1}^{x_2} x p_0(x)\,dx \approx \mu_0^{(x)}$ and $\int_{x_1}^{x_2} (x - \mu_0^{(x)})^2\, p_0(x)\,.\,dx \approx (\sigma_0^{(x)})^2$. Since the results of measurements in instrumental chemical and structural analysis are usually normally distributed, the *a posteriori* probability distribution is given as

$$p(x) = \frac{1}{\sigma^{(x)}\sqrt{(2\pi)}} \exp\left[-\frac{1}{2}\left(\frac{x - \mu^{(x)}}{\sigma(x)}\right)\right], \qquad (3)$$

where necessarily $x_1 + 3\sigma^{(x)} \leqq \mu^{(x)} \leqq x_2 - 3\sigma^{(x)}$. Then the information content takes on, in the case of a rectangular *a priori* distribution the value

$$I(p, p_0) = \ln \frac{(x_2 - x_1)\sqrt{n}}{\sigma^{(x)}\sqrt{(2\pi e)}}, \qquad (4)$$

where $\sqrt{(2\pi e)} = 4{\cdot}13273$. In the case of a normal *a priori* distribution according to Eq. (*2b*), the information content is given as

$$I(p, p_0) = \ln \frac{(x_2 - x_1)\sqrt{n}}{\sigma^{(x)}\sqrt{(K^2 e)}} + \frac{K^2}{2}\left[\left(\frac{\mu^{(x)} - \mu_0^{(x)}}{x_2 - x_1}\right)^2 + \left(\frac{\sigma^{(x)}}{x_2 - x_1}\right)^2\right]. \qquad (5a)$$

For $\sigma^{(x)} \ll x_2 - x_1$ and $\mu_0^{(x)} \approx \mu^{(x)}$, which is common in analytical practice, it takes on the form

$$I(p, p_0) = \ln \frac{(x_2 - x_1)\sqrt{n}}{\sigma^{(x)}\sqrt{(K^2 e)}}, \qquad (5b)$$

which is analogous to Eq. (*4*) with the only difference that for $K = 6$ we have $\sqrt{(K^2 e)} = 9{\cdot}89233$.

In the case when the results of instrumental chemical or structural analysis do not refer to a value but rather to an intensity $Y_i$ and, contingently to a position $Z_i$ of the signal, we can use, as an *a priori* distribution, the rectangular one, for example for the signal intensity

$$p_0(y) = \begin{cases} 1/(y_{max} - y_{min}) & \text{for} \quad y \in \langle y_{min}, y_{max} \rangle \\ 0 & \text{otherwise} \end{cases} \tag{2c}$$

and analogously for its position, or we can use the normal distributions for arguments $y$ and $z$ with functions analogous to (*2b*). If only the intensity $\eta_i$ is the source of relevant information, then the information content is for a rectangular *a priori* distribution, of the size

$$I(p, p_0) = \ln \frac{(y_{max} - y_{min})\sqrt{n}}{\sigma^{(y)}\sqrt{(2\pi e)}} \tag{6a}$$

and for a normal *a priori* distribution

$$I(p, p_0) = \ln \frac{(y_{max} - y_{min})\sqrt{n}}{\sigma^{(y)}\sqrt{(K^2 e)}}, \tag{6b}$$

where $K \geqq 6$. These expressions are analogous to (*4*) and (*5b*), but they have identical meaning only in a special and in the analytical practice rather exceptional case. While the quantities $x_1$ and $x_2$ in Eqs. (*4*) and (*5b*) characterize our *a priori* knowledge about the value of $X$, $y_{min}$ and $y_{max}$ in (*6a*, *b*) are given by technical parameters of the device used in the analysis.

If both the intensity $\eta_i$ and the position $\zeta_i$ of the signal are sources of relevant information and they are mutually independent or the parameters of their *a priori* distributions are functionally interrelated, we define the information content with the use of the divergence measure as

$$I(p, p_0) = \int_{y_{min}}^{y_{max}} p(y) \ln \frac{p(y)}{p_0(y)} \, dy + \int_{z_{min}}^{z_{max}} p(z) \ln \frac{p(z)}{p_0(z)} \, dz . \tag{7}$$

In the case of *a priori* rectangular distributions $p_0(y)$ and $p_0(z)$ according to (*2c*) and normal *a posteriori* distributions $p(y)$ and $p(z)$ with probability densities analogous to Eq. (*3*) and with the assumption that $\eta_i$ and $\zeta_i$ are mutually independent, the information content equals

$$I(p, p_0) = \ln \frac{(y_{max} - y_{min})(z_{max} - z_{min})\, n}{2\pi e \sigma^{(y)} \sigma^{(z)}}, \tag{7a}$$

where $2\pi e = 17{\cdot}07947$. For normal *a priori* distributions $p_0(y)$ and $p_0(z)$ with probability densities analogous to (*2b*) and for independent $\eta_i$ and $\xi_i$, the information content is

$$I(p, p_0) = \ln \frac{(y_{max} - y_{min})(z_{max} - z_{min})\, n}{K^2\, e\sigma^{(y)}\sigma^{(z)}}, \qquad (7b)$$

where $K \geqq 6$. If the parameters of the *a posteriori* distributions $p(y)$ and $p(z)$ are functionally interrelated (this, of course, does not mean a stochastic dependence of the signal position and intensity), the information content can be determined analogously. A dependence of the type $\sigma^{(y)} = f(\mu^{(z)})$ is rather frequent in the analytical practice; then the information content for a rectangular *a priori* distribution becomes

$$I(p, p_0) = \ln \frac{(y_{max} - y_{min})(z_{max} - z_{min})\, n}{2\pi e\sigma^{(z)} f(\mu^{(z)})} \qquad (8a)$$

and for a normal *a priori* distribution,

$$I(p, p_0) = \ln \frac{(y_{max} - y_{min})(z_{max} - z_{min})\, n}{K^2\, e\sigma^{(z)} f(\mu^{(z)})}. \qquad (8b)$$

The function $f(\mu^{(z)})$ can be usually expressed by a polynomial of the form $\sigma^{(y)} = \sum_{j=0}^{r} \sigma_j^{(z)} (\mu^{(z)})^j$ or $\sigma^{(y)} = \sum_{j=0}^{r} \sigma_j^{(z)} (\mu^{(z)})^{-j}$ for $r \leqq 4$.

Instrumental analytical methods lead usually to the determination of several components simultaneously, so that we shall consider rather the amount of information obtained from all signals simultaneously. The case where solely the signal intensity is the source of relevant information was studied by Danzer[2]. If both the signal position and the intensity are sources of relevant information, then the amount of information is

$$M(p, p_0) = \sum_{i=1}^{k} I(p, p_0)_i, \qquad (9)$$

where $i = 1, 2, \ldots k$, ($k$ is the number of determined components) and $I(p, p_0)_i$ is the information content of the determination of the i-th component according to (*7a, b*) or (*8a, b*).

In this way it is possible to determine the amount of information only in the case of a perfectly selective analytical method, where the analytical signals do not overlap. If they do overlap, so that we must separate them by calculation, the precision of the determination of the signal intensity decreases. In the relatively favourable case when two neighbouring signals partially overlap and their separation causes an increase

of the standard deviation of each of them, $\sigma_s^{(y)} = \sqrt{[(\sigma_1^{(y)})^2 + (\sigma_2^{(y)})^2]}$, and assuming $\sigma_1^{(y)} = \sigma_2^{(y)} = \sigma^{(y)}$ we have $\sigma_s^{(y)} = \sigma^{(y)}\sqrt{2}$. The information content for the cases expressed by Eqs (*6a*, *b*) or (*7a*, *b*) is then smaller by $\ln(1/\sqrt{2}) = 0{\cdot}34657$ natural units. In practice, of course, the increase in $\sigma^{(y)}$ caused by separation of the signals can be much larger and the amount of information smaller than in the indicated case. However, if we do not separate the signals, then the value of $k$ in Eq. (*9*) decreases by 2 for every couple of mutually overlapping signals and the amount of information decreases much more than by increasing the value of $\sigma^{(y)}$ to $\sigma_s^{(y)}$.

The discussion on the properties of Eqs (*4*), (*5b*), (*6a*, *b*) through (*9*) would be of no special practical value and we shall return to this point in our later work dealing with analytical problems. Here, in conclusion, we shall mention the reliability of analytical results for the case where both the intensity and the position of the signals are sources of relevant information. (The case when only the intensity is the source of relevant information was already discussed[5] with the use of the quantity $A(p/q)$.) Here, assuming that no stochastic dependence exists between the position and the intensity, we shall introduce the quantity

$$A[p_0(y)/p(y),\, p_0(z)/p(z)] = -\ln\left(2\pi\sigma^{(y)}\sigma^{(z)}\right) - \frac{1}{2}\left[\left(\frac{\delta^{(y)}}{\sigma^{(y)}}\right)^2 + \left(\frac{\delta^{(z)}}{\sigma^{(z)}}\right)^2\right]. \qquad (10)$$

Similarly as in the quoted paper, we can also here introduce the mean accuracy $\bar{A}$ and the total accuracy $\tilde{A}$, and use them especially in optimization of the conditions under which the instrumental or structural analyses are carried out.

REFERENCES

1. Eckschlager K.: This Journal *41*, 1875 (1976).
2. Danzer K.: Z. Chem. *15*, 158 (1975).
3. Eckschlager K.: Z. Chem. *16*, 111 (1976).
4. Eckschlager K., Štěpánek V.: *Information Theory as Applied to Chemical Analysis*. Wiley, New York 1979.
5. Eckschlager K., Štěpánek V.: This Journal *45*, 2516 (1980).

Translated by K. Micka.