INFORMATION PROPERTIES OF AN ANALYTICAL SYSTEM*

Karel ECKSCHLAGER

Institute of Inorganic Chemistry,
Czechoslovak Academy of Sciences, 160 00 Prague 6

Received May 7th, 1981

The connection of information measures such as the uncertainty after analysis and the information gain with inputs and outputs or with input-output relations in the system in which an analysis runs through as a process of obtaining information is shown on examples of instrumental or chromatographic qualitative analyses and of quantitative and trace analyses.

In preceding papers of this series[1-3] it has been shown that the information gain or the aposteriori uncertainty of quantitative analysis results depend on their precision and unbiasness; as far as results of instrumental quantitative analyses have been considered also the importance of the selectivity of the analytical procedure for the amount of information obtained by the analysis has been mentioned. For trace analyses[4] the dependence of the information gain on the determination limit has been studied. Here we will point out the close connection of the information aposteriori uncertainty of results of chromatographic or instrumental qualitative analyses with the selectivity of the procedure.

Precision, unbiasness, the determination limit, and selectivity are given by the input-output relation of the analytical system in which the analysis runs through as a process of obtaining information about the chemical composition of the analyzed sample. Various types of analyses can be described by various models for inputs, outputs, and input-output relations. Adequately chosen measures are mathematical means enabling the description of information properties of individual models. Therefore we will show on some examples how formulae for the amount of information are linked with the model of the input-output relation of the analytical system.

THEORETICAL

Analysis as a process of obtaining information always runs through in a system with an input-output relation[5]. In the case of instrumental (*e.g.*, emission spectrographic or IR spectrometric) or chromatographic (PC or TLC) qualitative analyses[6,7] the

* Part XVII in the series Theory of Information as Applied to Analytical Chemistry; Part XVI: This Journal *47*, 1195 (1982).

input is represented by a set of n possible components x_i ($i = 1, 2, \ldots, n$), the output by a set of signals in positions z_j ($j = 1, 2, \ldots, m$; $m \geqq n$) and the input-output relation can be described by a matrix of conditional probabilities $\|a_{ij}\|$[6] of the dimension $n \times m$. The matrix elements $a_{ij} = P(i \mid j)$ are probabilities that the component x_i is present in the input if we get a signal in the position z_j in the output. For a perfectly selective procedure of a qualitative or identification analysis the matrix $\|a_{ij}\|$ is square $n \times n$ and diagonal with elements $a_{ii} = 1$, $a_{ij} = 0$, $i \neq j$. H. Kaiser[8] introduced the following quantity for the quantification of the selectivity of a multicomponent quantitative analysis procedure

$$\Xi = \min_{i = 1, \ldots, n} \left(\frac{a_{ii}}{\sum_{j=1}^{m} a_{ij} - a_{ii}} - 1 \right), \tag{1}$$

where $a_{ij} = \gamma_{ij}$ is the partial sensitivity of the determination of the ith component by means of a signal in position z_j. This quantity can be analogously used to express the selectivity of a procedure of qualitative or identification analysis if we substitute $a_{ij} = P(i \mid j)$. It is true in either case that the selectivity of an analysis is the more perfect the greater is Ξ; indeed, if $a_{ij} = P(i \mid j)$, it does not assume such high values even if the selectivity is good as it does for $a_{ij} = \gamma_{ij}$ where γ_{ij} can be large enough ($\approx 10^3$). The aposteriori uncertainty of a qualitative or identification analysis, *i.e.*, the uncertainty after analysis when a signal in the position z_j in the output was measured is given by Shannon's entropy

$$H[P(i \mid j)] = -\sum_{j=1}^{m} P(i \mid j) \log P(i \mid j) = \sum_{j=1}^{m} a_{ij} \log \frac{1}{a_{ij}}. \tag{2}$$

The value of $H[P(i \mid j)]$ is obviously small if the selectivity is high and the entropy can take on different values for the same value of the quantity Ξ according to the spread of the a_{ij}'s around a_{ii}. An example of the influence of conditional probabilities upon the uncertainty has been illustrated by Liteanu and Rica[6]. Instead of entropy given by (2) it is sometimes more advantageous, for assessing the selectivity of a quantitative analysis procedure, to use its "relative" value

$$H_r = -\frac{\sum_{i=1}^{m} a_{ij} \log a_{ij}}{\log m} \tag{3}$$

with $0 \leqq H_r \leqq 1$ because the maximum value of $H[P(i \mid j)]$ is $\log m$; it takes on this value for the discrete uniform distribution when $P(i \mid j) = 1/m$, $j = 1, 2, \ldots, m$.

In the case of quantitative analyses the input into the system in which analytical information arises is given by a fixed but unknown value of the content X_i of the ith component and the output is a signal in position z_j with intensity η_{ij} behaving as a continuous random variable. The input-output relation can be represented by a calibration function $f_{ij}^{(K)}$ or by an analytical function $f_{ij}^{(A)}$. The final result ξ_i is taken as a continuous random variable with a probability density $p(x)$. The uncertainty after a quantitative analysis depends on whether the input-output relation, *i.e.*, $f_{ij}^{(K)}$ and $f_{ij}^{(A)}$, or its realization in processing the analytical signal (the stoichiometric "factor", the calibration curve or straight line, the standard addition, the inner standard, *etc.*) excludes or cannot exclude the possibility of the rise of a systematic error. If always $E[\xi_i] = X_i$, *i.e.*, the results are unbiased, the uncertainty can be expressed by Shannon's enetropy

$$H(p) = -\int_{x_1}^{x_2} p(x) \log_b p(x)\, dx\,, \tag{4}$$

where $\int_{x_1}^{x_2} p(x)\, dx = 1$. For normally distributed results of a quantitative analysis we get the uncertainty after analysis (using natural logarithms)

$$H(p) = \tfrac{1}{2} \ln 2\pi e\sigma_i^2 \tag{5}$$

i.e., it depends only on the variance σ_i^2 characterizing the precision of the results. If the analytical system cannot exlude the rise of a systematic error $\delta_i = |X_i - \mu_i|$ where $\mu_i = E[\xi_i]$ we have to evaluate the inaccuracy of the aposteriori distribution by the means of the Kerridge–Bongard measure of inaccuracy

$$H(q \mid p) = -\int_{x_1}^{x_2} q(x) \log_b p(x)\, dx\,, \tag{6}$$

where $q(x)$ is the true distribution, $\int_{x_1}^{x_2} q(x)\, dx = 1$, and $p(x) > 0$ for $x \in \langle x_1, x_2 \rangle$ is the distribution found by the analysis, $1 - \varepsilon \leqq \int_{x_1}^{x_2} p(x)\, dx \leqq 1$ ($\varepsilon \ll 1$). Now it depends on how we substitute the true ditsribution $q(x)$: as far as we take it for an arbitrary but very narrow continuous distribution with the expectation X_i we obtain for normally distributed results

$$H(X \mid p) = \frac{1}{2}\left[\ln 2\pi\sigma_i^2 + \left(\frac{\delta_i}{\sigma_i}\right)^2\right]. \tag{7}$$

The negative value of this quantity was chosen[2] to characterize quantitative analytical methods or as an objective function[1,2] to optimize quantitative analytical procedures. However, $q(x)$ can be substituted by a distribution of the same type and variance as

$p(x)$ but with the expectation X_i; then we obtain for normally distributed results

$$H(q \mid p) = \frac{1}{2}\left[\ln 2\pi e\sigma_i^2 + \left(\frac{\delta_i}{\sigma_i}\right)^2\right] \tag{8}$$

so that $H(q \mid p) = H(X \mid p) + \frac{1}{2}$.

The information gain of a quantitative analysis provided that the input-output relation can exclude the rise of a systematic error is given by the divergence measure[6]

$$I(p, p_0) = H(p \mid p_0) - H(p) = \int_{x_1}^{x_2} p(x) \ln \frac{p(x)}{p_0(x)}\, \mathrm{d}x\,. \tag{9}$$

Its properties were described earlier[9,10]; here we wish to recall only that it depends on the precision of the results and on the extent to what the results confirmed the apriori assumptions, *i.e.*, on "the moment of surprise" at the results. If the input--output relation cannot eliminate the systematic error we evaluate the inaccuracies before and after analysis by the means of the Kerridge–Bongard measure (6). Then the information gain can be decomposed as

$$I(r; p, p_0) = H(r \mid p_0) - H(r \mid p) = \int_{x_1}^{x_2} r(x) \ln \frac{p(x)}{p_0(x)}\, \mathrm{d}x\,, \tag{10}$$

where $r(x)$, $p_0(x)$, and $p(x)$ are the true, the apriori and the aposteriori probability distributions, respectively. In that specific case when we choose the uniform $p_0(x)$, the normal $p(x)$ with an expectation $\mu_i \neq X_i$, and the normal $r(x)$ with the expected value X_i and with the same variance as has $p(x)$ the formula (*10*) yields

$$I(r; p, p_0) = \ln \frac{x_2 - x_1}{\sigma_i \sqrt{(2\pi e)}} - \frac{1}{2}\left(\frac{\delta_i}{\sigma_i}\right)^2 \tag{11}$$

which was derived for the information gain of results subject to a systematic error $\delta_i = |X_i - \mu_i|$ in[11] and in Section 6.4 of the monograph[5]. If all the three distributions are normal and namely the true one with the mean value X_i and with the variance σ_r^2, the apriori one with μ_0 and σ_0^2, and the aposteriori one with parameters μ and σ^2 then

$$I(r; p, p_0) = \ln \frac{\sigma_0}{\sigma} + \frac{1}{2}\left[\left(\frac{\delta_0}{\sigma_0}\right)^2 - \left(\frac{\delta}{\sigma}\right)^2 + \left(\frac{\sigma_r}{\sigma}\right)^2 + \left(\frac{\sigma_r}{\sigma_0}\right)^2\right], \tag{12}$$

where $\delta_0 = |X_i - \mu_0|$ and $\delta = |X_i - \mu|$. For $\mu = X_i$ and $\sigma_r = \sigma$ the result in (*12*)

changes into

$$I(r;\, p,\, p_0) = \ln\frac{\sigma_0}{\sigma} + \frac{1}{2}\left[\left(\frac{\mu - \mu_0}{\sigma_0}\right)^2 - \frac{\sigma^2 - \sigma_0^2}{\sigma_0^2}\right]$$

shown in Section 6.6 of the monograph[5] and valid for the divergence measure of the information gain when $p(x)$ and $p_0(x)$ are normal.

Thus the aposteriori uncertainty of the results of a quantitative analysis depends either only on the precision (*4*) unless the results are subject to a systematic error or on the precision and the unbiasness (*6*) provided that a systematic error is present. The information gain always depends on "the moment of surprise" and either only on the precision as in case (*9*) or also on the unbiasness as in (*10*) according to how we express the apriori and aposteriori uncertainties.

In trace analyses the input into an analytical system is given by a small but unknown value X_i and the output is a signal in position z_j. The intensity of this signal η_{ij} can be so small that it cannot be distinguished from the background noise and the analysis results in the finding that the content of the sought-for component $\xi_i \leqq x_0$, *i.e.*, that it does not surpass the determination limit x_0, or that it is distinguishable from the noise of the background and enables the determination of the *i*th component, *i.e.*, $\xi_i > x_0$. In (ref.[4]) and in Section 6.7 of[5] it has been shown that the information gain of trace analyses depends on x_0 in such a way that it increases with decreasing values of x_0. The following cases have been considered:

1) $X_i \leqq x_0$; then $I(p, p_0) = \ln\dfrac{x_1}{x_0}$ (x_1 is the highest anticipated content) (*13*)

2) $X_i > x_0$ and the distribution of the results is normal with $x_0 < \mu_i \leqq x_0 + 3\sigma_i$; then

$$I(p, p_0) = \ln\frac{x_1}{\sigma_i\sqrt{2\pi e}} + \frac{1}{2}\frac{z_0\,\varphi(z_0)}{1 - \phi(z_0)} - \ln\left[1 - \phi(z_0)\right], \qquad (14)$$

where $z_0 = (\mu_i - x_0)/\sigma_i$, $\varphi(z_0)$ and $\phi(z_0)$ are the frequency and the distribution function of the standardized normal variable, respectively. If $\mu_i > x_0 + 3\sigma_i$ the formula (*14*) changes into

$$I(p, p_0) = \ln\frac{x_1}{\sigma_i\sqrt{(2\pi e)}}. \qquad (15)$$

3) $X_i > x_0$ and the distribution of the results is log-normal; then we get

$$I(p, p_0) = \ln\frac{x_1}{k x_0 \sigma_i\sqrt{(2\pi e)}}, \qquad (16)$$

where k is the coefficient of asymmetry of the aposteriori distribution defined in[4].

In last two cases ($X_i > x_0$) the information gain depends on the distance $\mu_i - x_0$ and, with the increase of this distance, its value tends to the value given in (*15*) which is valid for the determination of higher than trace contents of the *i*th component.

RESULTS AND DISCUSSION

Information quantities such as the aposteriori uncertainty, the information gain or the equivocation are either explicitly dependent upon or tightly linked with parameters of the analytical system as, *e.g.*, selectivity, precision, accuracy or the determination limit. These parameters follow from the input-output relation in the analytical system.

The aposteriori uncertainty can be evaluated for various sorts of analyses. Specificity of the input and the output or of the input-output relation in the system, for instance for qualitative, quantitative or trace analyses, enables to obtain its particular values. The difference of apriori and aposteriori uncertainties or the equivocation as well as the information gain depend on probability distributions characterizing the input (the apriori distribution) and the output (the aposteriori distribution) of the analytical system, including conditional distributions. Thus information quantities are due to the model with which we describe the input, the output, and the input--output relation of a given analytical system. Several such models have been introduced above: Thus, *e.g.*, the relation between a set of possible components in the input of a system for instrumental or chromatographic qualitative or identification analysis and a set of signals in the output is described by the matrix of conditional probabilities[6] and we can enumerate, from any its row, the uncertainty after analysis $H[P(i \mid j)]$ given by (*2*) or its relative value H_r by (*3*) or Kaiser's parameter of selectivity Ξ in (*1*). In the case of quantitative analyses we can take into account, for specific apriori and aposteriori distributions, two different input-output relations: for one in which the calibration function $f_{ij}^{(K)}$ or the analytical function $f_{ij}^{(A)}$ yield accurate results we evaluate the aposteriori uncertainty by Shannon's entropy in (*4*) and the information gain by the divergence measure (*9*); for the other one when the rise of a systematic error must be admitted measures in (*6*) and (*10*) have to be employed. It means that, for specific distributions, we get varied formulae for aposteriori uncertainties or inaccuracies (*5*), (*7*), and (*8*) or for information gains (*9*) and (*10*) according to the input-output relations. Another case is encountered in trace analyses: here we distinguish, for the same input and equal input-output relations, three different aposteriori probability distributions in the output and information gains are evaluated by three different formulae in (*13*), (*14*) and (*16*).

The preceding findings documented by examples from the fields of instrumental or chromatographic qualitative or identification analyses and of quantitative and trace analyses have practical importance for the application of information theory in judging and optimizing analytical devices as well as systems in which processes of

creating analytical information run through: All above mentioned features of analytical systems, *i.e.*, selectivity, precision, accuracy, and the determination limit, restricting achievements of analyses result, in instrumental methods, from both the technical parameters of the devices and from the procedure, *e.g.*, calibration, the way of processing the analytical signals, the substracting a blank experiment, *etc.* Therefore the choice of an instrumental method or of a type of the device can be implemented by the use of information profitability[11], which characterizes rather inaffectable features given by technical parameters of the devices, and the analytical procedure can be optimized by the use of aposteriori uncertainty or of the information gain as an objective function.

Thus it is expedient to judge or to optimize different analytical procedures by those parameters that affect the information quantity (the uncertainty or the amount of information) most. So, for instance, we will judge an instrumental or chromatographic qualitative or identification analysis according to the selectivity of the procedure or we will use an appropriate measure of selectivity as in (*1*) or of uncertainty (*2*) or (*3*) or another objective function, connected with selectivity[12], as a response-function in the optimization. The selectivity in instrumental analyses is indeed given by parameters of the devices, first of all by the discrimination capability, and/or by the way of processing the analytical signal ("the separation" of bands in the IR spectrometry by the means of a computer, the Fourier transformation, *etc.*) while it is given mainly by the procedure in chromatographic qualitative or identification analyses. Instrumental methods of a quantitative analysis will be rated according to precision – it results from both the parameters of the device and from the procedure – and accuracy (unbiasness) of the results. The unbiasness is indeed mainly a matter of suitable calibration or of the elimination of the matrix effect and thus it is affected rather by the analytical procedure than by the parameters of the device. In the optimization the formulae (*7*), (*8*), (*11*) or (*12*) can be adopted as objective functions. Yet optimization has to be carried out for different ways of calibrating or we have to find out the most suitable way for the given case beforehand and to optimize the entire procedure for this way. In trace analyses methods the determination limit appears as the parameter having the greatest effect upon the aposteriori uncertainty or upon the information gain; precision and unbiasness are of less use. Moreover a low value of the determination limit cuts down the frequency of cases when $X_i \leqq x_0$ and when the result gives a smaller information gain than that in (*14*) or (*16*) as it has already been shown earlier[4].

REFERENCES

1. Eckschlager K.: This Journal *44*, 2373 (1979).
2. Eckschlager K., Štěpánek V.: This Journal *45*, 2516 (1980).
3. Eckschlager K.: This Journal *41*, 1875 (1976); *46*, 478 (1981).
4. Eckschlager K., Štěpánek V.: Mikrochim. Acta *1978* I., 107; *1981*, 2 143.

5. Eckschlager K., Štěpánek V.: *Information Theory as Applied to Chemical Analysis*. Wiley-Interscience, New York 1979.
6. Liteanu C., Rica I.: Anal. Chem. *51*, 1986 (1979).
7. Cleij P., Dijkstra A.: Fresenius Z. Anal. Chem. *298*, 97 (1979).
8. Kaiser H.: Fresenius Z. Anal. Chem. *260*, 252 (1972).
9. Eckschlager K., Štěpánek V.: This Journal *4,7* 1195 (1982).
10. Vajda I., Eckschlager K.: Kybernetika *16*, 120 (1980).
11. Danzer K., Eckschlager K.: Talanta *25*, 725 (1978).
12. Eckschlager K.: Chem. Listy *76*, 21 (1982).

Translation by V. Štěpánek.