



Evaluation of the Maximum Achievable Information Content in Quantitative Spectrochemical Analysis

J. Dubrovkin

Computer Department, Western Galilee College
2421 Acre, Israel

Abstract: The existence of the local maximum of information obtained in univariate analysis was discovered for the first time. The information content was shown to be maximal at the point where the predicted concentration is close to the average of the calibration set concentrations. This result was obtained using a new expression for the systematic error of the measured concentration caused by systematic errors in the preparation of standard samples. The information content of the analysis of multicomponent mixtures was calculated numerically using simulated Gaussian spectra and the spectrum of Excedrin tablet components. It was shown that the informational approach allows balancing the ratio between random and systematic errors. The use of information measures for calculating the optimal number of components for the gasoline octane number determination by the Partial Least Squares method was found to be not effective.

Keywords: quantitative spectrochemical analysis, information theory, regularization, Partial Least Squares, errors

I. Introduction

The application of the information theory to the methods of analytical chemistry [1] had a great impact on the development of the major concepts of this science. In particular, the information theory was employed in assessing the information obtained by means of derivative spectroscopy [2]. Unfortunately, analytical chemists have been paying little attention to the information theory for more than a decade. An interesting exception was the recent article [3], in which the information criteria were employed for selecting the calibration curves. In our opinion, the potentialities of using information measure as the performance criterion of spectrochemical analysis are far from being fully exploited and require further studies.

It is known that the amount of information acquired in the quantitative analysis is determined by the random and systematic errors [1], which are important metrological characteristics of analysis [4]. For example, the validation of analytical methods is performed by comparing the errors of a particular analysis with the requirements of the relevant regulatory document. In many cases, mathematical methods of chemometrics make it possible to significantly reduce these errors down to certain minimum attainable values that can be evaluated by computer modelling [5]. However, it may well happen that the algorithmic means of computer-aided quantitative analysis give the errors in the concentration measurements close to the minimum achievable level. In this case, it is hardly reasonable to make great efforts to improve the employed mathematical procedures in an attempt to obtain better metrological characteristics of the analysis.

In our previous work [5], optimization of the quantitative analysis of multicomponent mixtures was achieved by minimizing the total standard deviation in the concentration determination. The ratio between the random and systematic components was not taken into account because these components appeared in the theoretical expression for the total error with equal weights [6]. In contrast to this, the calculation of the amount of information acquired in quantitative analysis performed in [1] did not have this drawback. In view of the above, the goal of the present research was using information measures for the optimization of quantitative analytical procedures.

Standard algebraic notations are used throughout the paper. Bold upper-case and lower-case letters denote matrices and column vectors, respectively. Upper-case and lower-case italicized letters represent scalars. All calculations were performed and the plots were built using the MATLAB program.

II. Theory

The information content of univariate quantitative analysis is represented [1] as

$$I = \ln \frac{R_c}{\sqrt{2\pi e^k} \sigma_{pr}} - 0.5 \left(\frac{\delta_{pr}}{\sigma_{pr}} \right)^2, \quad (1)$$

where R_c is the *a priori* defined range of the uniformly distributed predicted analyte concentration (c_{pr}), $k = (\sigma_{et}/\sigma_{pr})^2$, σ_{et} and σ_{pr} are the standard deviations of the standard (calibration etalon) sample concentration and of c_{pr} , respectively, and δ_{pr} is the systematic error of the c_{pr} determination. It is supposed that the errors of the standard sample and the analyte concentrations are normally distributed.

Eq. 1 can be simplified under the assumption of $R_c = 1$ (i. e., $c_{pr} = 0 \div 1$) as follows:

$$I = -0.5 \ln(2\pi) - \ln \sigma_{pr} - 0.5 \left(\frac{\delta_{eff}}{\sigma_{pr}} \right)^2, \quad (2)$$

where $\delta_{eff} = \sqrt{\sigma_{et}^2 + \delta_{pr}^2}$.

According to the assumption of $R_c = 1$, we set the range of standard deviations σ_{pr} to be $0.001 \div 0.01$. Then, the corresponding minimum relative random errors of concentration measurements are in the range of $0.1 \div 1\%$. Thus, the first two members in Eq. 2, which equal $-0.5 \ln 2\pi - \ln \sigma_{pr} \approx 3.7 \div 6.0$ (nit), define the *global* information maximum achieved in the ideal case of $\delta_{eff}=0$ (prediction error δ_{pr} and the random error of the standard sample concentration, σ_{et} , being negligible as compared to σ_{pr}).

Since $I \geq 0$, the maximal values of the $\delta_{eff} / \sigma_{pr}$ ratio range from 2.7 to 3.5.

Eq. 2 has a *local* maximum at $\delta_{eff} = \sigma_{pr}$ (see Appendix A):

$$I_{loc} = -\ln(\sqrt{2\pi}e^k \sigma_{pr}) - 0.5 \approx 3.2 \div 5.5 \text{ (nit)}. \quad (3)$$

The existence of the global and local maxima of Eq. 2 is demonstrated in Fig. 1.

In conclusion, the low information content of the analysis shows a significant dominance of the systematic errors over the random ones, i.e., the disbalance of these two types of errors.

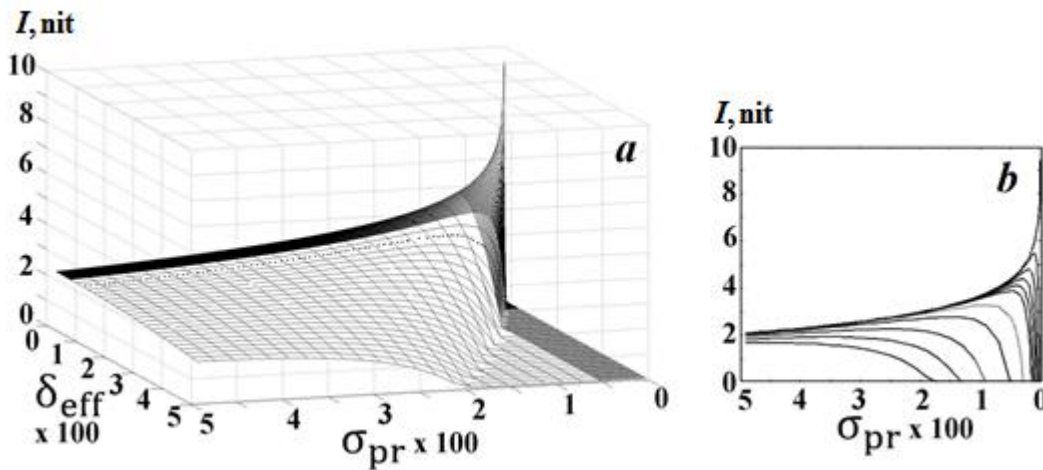


Fig. 1. (a) Plot of function $I(\delta_{eff}, \sigma_{pr})$. (b) Family of $I(\sigma_{pr})$ curves. The δ_{eff} values of (45, 35, 25, 19, 9, 5, 4, 3, 2, 1) $\times 10^{-2}$ correspond to the curves in the "bottom-up" order. Identical dependences are denoted by dotted line in (a) and (b).

III. Univariate Analysis

In Appendix B (Eq. B6), it is shown that the systematic error of the quantitative analysis of one-component solutions which is performed using classical univariate calibration is

$$\delta_{pr} \leq \delta_{et}^{max} \left[1 + 3 \frac{a_{pr} - \bar{a}}{a_{max} - a_{min}} \right], \quad (4)$$

where δ_{et}^{max} is the maximal systematic error of the standard solution concentration, a_{pr} is the absorbance of the test solution, a_{max} , $\bar{a}$, and a_{min} are the maximal, the mean, and the minimal absorbance of the standard solutions, respectively, at a given wavelength. If $a_{pr} = \bar{a}$, then Eq. 4 reaches its minimum:

$$0 \leq \delta_{pr}^{min} \leq \delta_{et}^{max} \quad (5)$$

The maximum error, δ_{et}^{max} , can be made negligible as compared to the random error, σ_{et} , by careful preparation of standard solutions. Besides, the standard deviation of the standard concentrations does not exceed the corresponding values for the test concentrations:

$$0 \leq \delta_{pr}^{min} \leq \delta_{et}^{max} \ll \sigma_{et} \leq \sigma_{pr} \quad (6)$$

Then $\sigma_{et} \leq \delta_{eff} \leq \sigma_{pr}$ and thus the information content of the analysis appears to be close to the local maximum given by Eq. 3.

According to the Beer-Lambert-Bouguer law, the absorbance changes linearly with the analyte concentration; therefore, the analysis should be optimal at the solution concentration close to the average concentration of the standard solutions. In this case the leverage (h), which quantifies the position of the prediction sample relative to the calibration set [7]:

$$h = (c_{pr} - \bar{c})^2 / \sum_i^p (c_i - \bar{c})^2 = 0, \quad (7)$$

where $\bar{c}$ is the average analyte concentration for the calibration standards (c_i).

This result can be generalized to the case of multicomponent quantitative analysis, where the information content is the sum of the values for each component obtained separately from Eq. 2. It is assumed that each test component is of equal information importance for the mixture analysis [1]. For example, in the case of

three-component mixture analysis [8], it was found empirically that the errors are minimal when the point corresponding to the concentration vector of the test sample is located close to the center of the concentration diagram for the calibration set. According to our findings, in this case the information content of the analysis will be maximal. We calculated the leverage for the three-component mixtures [7]:

$$h = \mathbf{c}_{pr}^T (\mathbf{C}\mathbf{C}^T)^{-1} \mathbf{c}_{pr}, \quad (8)$$

where $\mathbf{c}_{pr}$ is the concentration vector of the test sample, $\mathbf{C}$ is the concentration matrix of the calibration set. It was found that in the point of the maximal information content the leverage is very small; therefore the impact of the calibration errors on the uncertainty of analysis is minimal [7].

IV. Multivariate Analysis

To evaluate the minimal uncertainty of the quantitative analysis of mixtures with known qualitative composition ("white mixtures") [5] and to select the number of components in the Partial Least Squares (PLS) regression [8], we will use the information measure.

Consider the ideal linear model of the absorbance spectrum of an n -component mixture:

$$\mathbf{y} = \mathbf{A}\mathbf{c} + \boldsymbol{\varepsilon}, \quad (9)$$

where $\mathbf{A} = \mathbf{KW}$, $\mathbf{K}$ is the $m \times n$ matrix, whose columns are the standard spectra of the pure-component solutions of given concentrations ($\mathbf{c}_{et}$) (the path length is fixed); m is the wavelength number; $\mathbf{c}$ is the vector of the relative component concentrations (fractions); $\mathbf{W}$ is a diagonal weight matrix with unit trace; $\boldsymbol{\varepsilon}$ is the $m \times 1$ vector of normal noise with zero mean and variance σ_y^2 .

The minimum uncertainty (the relative mean squared error, *RMSE*) of vector $\mathbf{c}$ is obtained for the optimal matrix $\mathbf{W}_{opt}$ [5], which changes the ratios of the optical densities of the mixture components. Matrix $\mathbf{A} \approx \mathbf{KW}_{opt}$, which is close to the optimal one, can be obtained by selecting the analytical wavelengths and also by varying the physicochemical factors affecting absorbance (e.g., solvent pH).

We will solve Eq. 9 using the Tikhonov regularization method [9] since it is very effective in increasing the stability of calculations that are extremely sensitive to the noise of the measurements in the case of strongly overlapping pure-component spectra of similar forms.

Regularized predictor $\hat{\mathbf{c}}$ can be represented as [8]:

$$\hat{\mathbf{c}} = (\mathbf{A}^T \mathbf{A} + \alpha \mathbf{I})^{-1} \mathbf{A}^T \mathbf{y}, \quad (10)$$

where α is the regularization parameter and $\mathbf{I}$ is the identity matrix. With the increase of α , the random error of $\hat{\mathbf{c}}$ decreases, while its systematic errors (bias) increase. The choice of the optimal value (α_{opt}) is made by finding a compromise between these two types of errors. For α_{opt} , the value of *RMSE* is minimal [6].

The PLS algorithm is described in [8].

V. Numerical Experiments

The numerical experiments were performed by computer modeling. The calibration data for determining the octane number of gasoline (NIR-spectra of 60 samples and their octane numbers) were taken from the MATLAB program package [10]. The errors were calculated using cross-validation [8]: the octane number was calculated for each spectrum, while the remaining 59 spectra were regarded as the calibration set. To evaluate the minimum uncertainty of the multicomponent analysis of "white" mixtures, the columns of matrix $\mathbf{K}$ were composed of the Gaussian bands and the UV-absorption spectra of Excedrin tablet components [11] (Fig. 2).

The *RMSE* values were calculated in two ways.

1. At the global minimum of the mean value of the two-variable function:

$$RMSE(\mathbf{W}, \alpha)_{\min} = \min_{\mathbf{W}, \alpha} \frac{\sum_{k=1}^n ((\hat{\mathbf{c}} - \mathbf{c}_{pr})_k)^2}{\sum_{k=1}^n (\mathbf{c}_{pr})_k^2}, \quad (11)$$

where $(\mathbf{c}_{pr})_k$ is the concentration vector of the k^{th} test mixture (Table 1).

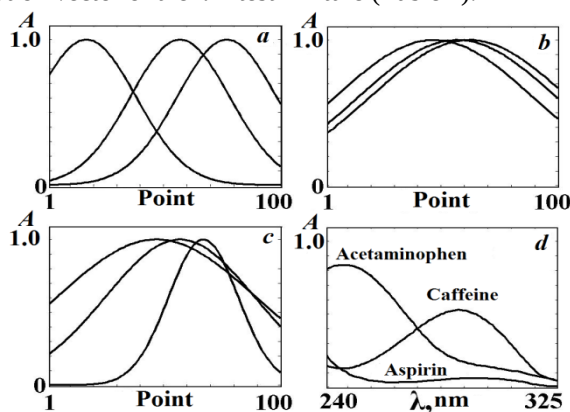


Fig. 2. Component spectra of data sets 1, 2, 3 (curves *a*, *b*, *c*, respectively) and of Excedrin (curves *d*). The component concentrations were 20 mg mL⁻¹ in water:methanol:glacial acetic acid (69:28:3) solution. Reproduced with permission from www.dissolutiontech.com.

Table 1. Fractions (%) of the test mixture components

Mixture Component	1	2	3	4	5	6	7	8	9	10
1	50	25	25	33	80	10	10	90	5	5
2	25	50	25	33	10	80	10	5	90	5
3	25	25	50	33	10	10	80	5	5	90

2. At the minimum of the mean value of the single-variable function (11), $RMSE(\mathbf{W}, \alpha_{opt}(\mathbf{W}))_{\min}$. For each value of matrix $\mathbf{W}$, the regularization parameter, α_{opt} , was calculated at the point of the maximum of the mean information content of the analysis:

$$I_{\max(\alpha_{opt})} = \max_{\alpha} (\overline{\sum_{k=1}^n I_k}). \quad (12)$$

The random (σ_{pr}) and the systematic (δ_{pr}) errors in the determination of each component concentration were evaluated separately. It was assumed that $\delta_{eff} = \delta_{pr}$ since, in this case, $\delta_{pr} \gg \sigma_{et}$.

The number of the components (nc) in the PLS-based determination of the octane number was evaluated based on two values as criteria: the minimum $RMSE$ value: $\min_{nc}(RMSE)$ and the maximum information gain: $\max_{nc}(I)$.

In all cases the results of 100 independent calculations were averaged. In each calculation, a newly generated normal noise was added to the mixture spectrum.

VI. Results and Discussion

The results obtained in this study are represented in Figs. 3 and 4. In Fig. 3, the dependences of the information content of the analysis on $\sqrt{RMSE}$ ($RRMSE$) are shown for three data sets (Fig. 2 a-c). For each data set, 10 mixtures were studied using criteria (11) and (12). The straight lines represent liner approximations of the discrete data. The $I(RRMSE)$ dependences for 10 mixtures of the Excedrin tablet components are presented in Fig. 4a.

To interpret the obtained results, we need first to consider some theoretical issues.

It is well-known that for non-regularized solutions, the random errors are determined by the condition number of matrix $\mathbf{K}$ ($\text{cond}(\mathbf{K})$) [6]. In the case of strongly overlapping pure-component spectra having similar forms, the value of $\text{cond}(\mathbf{K})$ can be very large. For regularized solutions, the $RMSE$ values are minimised due to the reduction of the random errors. On the other hand, the increase of the regularization parameter results in increased systematic errors [6]. The contribution of the last term of Eq. 2, $(\delta_{eff}/\sigma_{pr})^2$, increases and the information content of the analysis decreases.

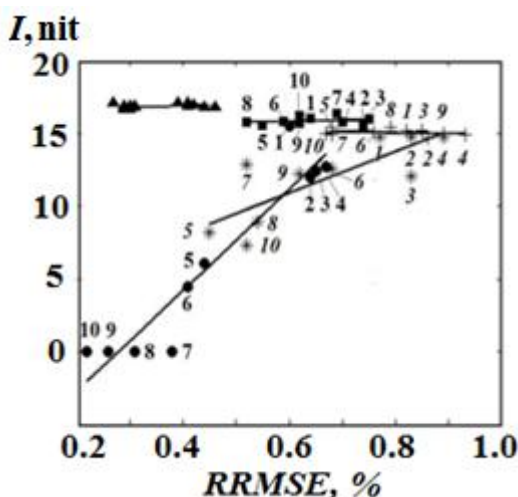


Fig. 3. The $I(RRMSE)$ dependencies obtained for the Tikhonov regularization-based analysis of the mixtures, whose spectra are a sum of Gaussians. Data set and optimization criteria (11) and (12): 1, $\blacktriangle$, $\blacktriangle$; 2, $\bullet$, $\blacksquare$; 3, $+$, $+$, respectively. $\sigma_y = 0.002$. The mixture numbers are given next to symbols (for data set 2, in italic).

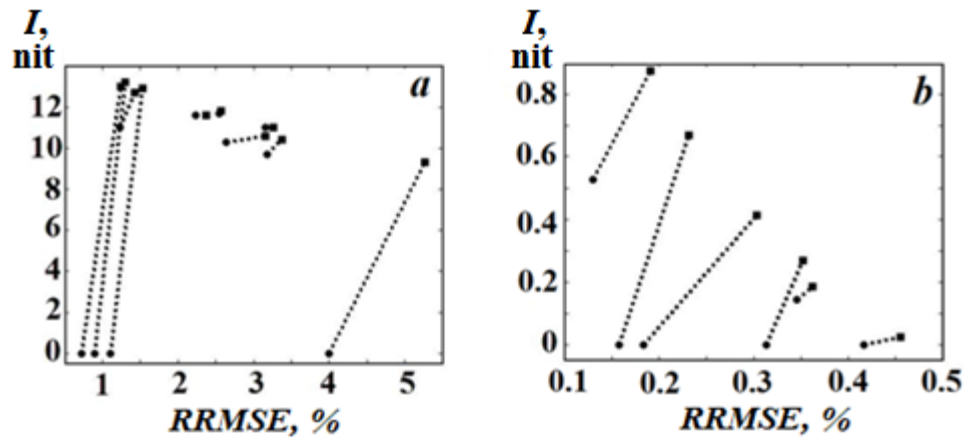


Fig. 4. The $I(RRMSE)$ dependencies: (a) obtained for the Tikhonov regularization-based analysis of the mixtures of the Excedrin tablet components; $\sigma_y = 0.002$. (b) for the PLS-based octane number determination; $\sigma_y = 5 \times 10^{-4}$. Optimization criteria: (11) (●) and (12) (■). Points corresponding to each mixture are connected by dotted lines.

The steepness of the $I(RRMSE)$ dependences obtained using criterion (11) (Fig. 3) determines the information gain which is achieved due to the increase of the $RRMSE$ values. The $I(RRMSE)$ plot obtained for data set 2 is much steeper than the one obtained for data set 3 (see Fig. 3). This fact can be readily understood taking into account the dependence of steepness on the corresponding $cond(\mathbf{K})$ value and stronger overlapping of the component spectra for data set 2 (Fig. 2b) than for data set 3 (Fig. 2c). For data set 1, whose component spectra are well-resolved (Fig. 2a), the obtained information is almost independent on the minimum $RRMSE$ values. In this case, using criteria (11) or (12) gives very similar values of information content (the corresponding points are marked by triangles in the left-top corner of Fig. 3). It can be seen that the information content is approximately constant if the analysis is based on criterion (12).

The comparison of the I values and the $RRMSE$ values obtained using criteria (11) and (12) for each mixture (Fig. 3) also shows that the cost of the information gain is the increase of the $RRMSE$ values.

It is interesting that for some mixtures of the Excedrin tablet components, the loss of the analysis accuracy (the increase of the $RRMSE$ -values) that occurs when criterion (12) is used instead of criterion (11) is not significant as compared to the information gain (Fig. 4a). This gain results from the sharp decrease of the $|\delta_{eff}/\sigma_{pr}|$ ratio (2) for some mixture components observed when criterion (12) is used (instead of (11) - bold figures in Table 2).

Thus, the use of information criterion (12) for the optimization of the Tikhonov regularization-based analysis eliminates the significant prevalence of systematic errors over random ones.

Our studies of the PLS-based determination of the gasoline octane number showed that the application of criterion (12) for evaluating the optimal number of PLS components is very limited. The information gain was observed only for 4 out of 60 samples (Fig. 4b). For the remaining 56 samples, the number of the components determined according to criteria (11) and (12) coincided. Moreover, it was found that the calculated values were highly sensitive to changes in the standard deviation of the noise in the mixture spectrum.

Table 2. The ratio of systematic to random errors for the analysis of Excedrin tablet components.

Mixture	Criterion	$ \delta_{eff}/\sigma_{pr} ^*$		
		Acetaminophen	Aspirin	Caffeine
7	(11)	0.0043	0.73	6.22
	(12)	0.11	0.33	0.90
8	(11)	5.6	0.0022	0.9
	(12)	0.78	0.22	0.19
9	(11)	0.95	8.1	7.4
	(12)	0.051	0.16	0.15
10	(11)	0.36	1.12	10.1
	(12)	0.70	0.0071	0.82

* Abnormally large values are marked in bold.

This result can be explained in the framework of the spectral coordinate transformation method [12], where, for the purpose of improving the analysis, a spectrum or a set of spectra is replaced by its linear orthogonal transform (as in the case of, e.g., the Principal Component Analysis [6]). In this method, various regression algorithms can be described as filters of the principal components [13]. In the PLS method, the principal components having small singular values are discarded (abruptly truncated), while in the Tikhonov regularization method, their values are reduced smoothly by choosing the regularization parameter. Therefore, in contrast to finding the optimal number of principal components, choosing the optimal regularization parameter based on information measure is a "smooth" and, hence, a stable process.

References

- [1] K. Eckschlager, V. Štěpánek and K. Danzer, "A review of information theory in analytical chemometrics", J. Chemometrics, vol. 4, 1989, pp. 195-216.
- [2] J. M. Dubrovkin, "Change in the amount of information contained in a spectrum on differentiation", J. Appl. Spectr., vol. 35, 1981, pp. 1149-1153.
- [3] E. Rozet, E. Ziemons, R.D. Marini and Ph. Hubert, "Usefulness of information criteria for the selection of calibration curves", Anal. Chem., vol. 85, 2013, pp. 6327-6335.
- [4] K. Danzer, Analytical Chemistry: Theoretical and Metrological Fundamentals. Berlin: Springer-Verlag, 2007.
- [5] J. M. Dubrovkin, "Estimation of the minimum uncertainty of quantitative spectroscopic analysis", Bulletin of Slovak Spectroscopic Society, vol. 19, 2012, pp. 4-7.
- [6] G. A. F. Sever and A. J. Lee, Linear Regression Analysis, 2nd ed., New Jersey: John Wiley and Sons, 2003.
- [7] N. M. Faber, J. Ferré, R. Boqué and J. H. Kalivas, "Quantifying selectivity in spectrophotometric multicomponent analysis", Trends in Anal. Chem., vol. 22, 2003, pp. 352-361.
- [8] H. Martens and T. Næs, Multivariate Calibration. New York: Wiley, 1992.
- [9] A. N. Tikhonov and V. A. Arsenin, Solution of Ill-posed Problems. Washington: Winston & Sons, 1977.
- [10] MATLAB software for Windows from The MathWorks, Inc., 24 Prime Park Way, Natick, Mass. 01760-1500.
- [11] M. J. Simonian, URL: [http://www.dissolutiontech.com/DTresour/1995 Articles/DT199505_A02.pdf](http://www.dissolutiontech.com/DTresour/1995%20Articles/DT199505_A02.pdf)
- [12] J. M. Dubrovkin, "Effectiveness of spectral coordinate transformation method in evaluation of the unknown spectral parameters", J. Appl. Spectr., vol. 38, 1983, pp. 191-194.
- [13] E. Andries and J. H. Kalivas, "Multivariate calibration leverages and spectral, *F*-ratios via the filter factor representation", J. Chemometrics, vol. 24, 2010, pp. 249-260.

Appendix A

Determination of the local maximum of the information content of single-component quantitative analysis

Eq. 2 reaches its maximum at $\delta_{eff} = \sigma_{pr}$ since at this point the first derivative of the function with respect to σ_{pr} is zero:

$$\frac{dl}{d\sigma_{pr}} = -\frac{1}{\sigma_{pr}} + \frac{\delta_{eff}^2}{\sigma_{pr}^3} = 0 \quad (A1)$$

and the second derivative is negative:

$$\frac{d^2l}{d\sigma_{pr}^2} = \frac{1}{\sigma_{pr}^2} - \frac{3\delta_{eff}^2}{\sigma_{pr}^4} < 0 \quad (A2)$$

Appendix B

One-component solution: Classical calibration. Evaluation of the bias of the concentration estimator

In the case of one-component solution, the concentration estimator, which uses a single point in the absorption spectrum, has the form [4, 6]:

$$c_{pr} = \bar{c} + \left(\sum_{i=1}^p (a_i - \bar{a})(c_i - \bar{c}) \right) (a_{pr} - \bar{a}) / S_{aa}, \quad (B1)$$

where p is the number of standard solutions (the size of the calibration set), a_i and c_i are the absorbance and concentration of solution i in the calibration set, respectively; a_{pr} is the absorbance of the test solution, and $S_{aa} = \sum_{i=1}^p (a_i - \bar{a})^2$.

Suppose that the systematic concentration error in preparing calibration set solution i is δc_i . The corresponding systematic error for the analyte concentration can be found by differentiating Eq. B1 with respect to c_i :

$$\delta c_{pr}^i \approx \left[1/p + (a_i - \bar{a})(a_{pr} - \bar{a}) / S_{aa} \right] \delta c_i \quad (B2)$$

Since systematic error δc_i may be positive or negative, consider a hypothetical worst case of

$$\delta c_i = \text{sign}(a_i - \bar{a}) \delta_{et}^{max}, \quad (B3)$$

where sign is used to extract the sign of a number, $\delta_{et}^{max} > 0$ is the maximum systematic concentration error for the standard solution. For the sake of simplicity, suppose that the absorbance of the standard solution spectrum, a_i , is uniformly distributed over its mean value with constant step h :

$$a_i - \bar{a} = i * h, \quad (B4)$$

where $i = -q, -q+1, \dots, 0, 1, 2, \dots, q, p = 2q+1$.

Then, in accordance with Eq. B3,

$$\delta_{pr} \leq \sum_{i=1}^p \delta c_{pr}^i = \delta_{et}^{max} + (a_{pr} - \bar{a}) S_a / S_{aa}, \quad (B5)$$

where $S_a = \sum_{i=1}^p (a_i - \bar{a}) \delta c_i = 2h \delta_{et}^{max} \sum_{i=1}^q i = hq(q+1) \delta_{et}^{max}$, $S_{aa} = \sum_{i=1}^p (a_i - \bar{a})^2 = 2h^2 \sum_{i=1}^q i^2 = h^2 q(q+1)(2q+1)/3$.

It follows from Eq. B5 that

$$\delta_{pr} \leq \delta_{et}^{max} \left[1 + 3 \frac{a_{pr} - \bar{a}}{ph} \right] \leq \delta_{pr} \leq \delta_{et}^{max} \left[1 + 3 \frac{a_{pr} - \bar{a}}{a_{max} - a_{min}} \right], \quad (B6)$$

where a_{max} and a_{min} are the maximum and the minimum absorbance of the standard solution spectra at a given wavelength.

In conclusion, it should be noted that, in the general case, the above calculations are valid for linear relationship between the spectral response of the device and the analyte concentration if the contributions of the interfering components in the mixture are constant.