Anal. Chem. **1997,** *69,* 1620−1626

Net Analyte Signal Calculation in Multivariate Calibration

Avraham Lorber*

Nuclear Research Centre-Negev, P. O. Box 9001, Beer-Sheva 84190, Israel

Klaas Faber[†] and Bruce R. Kowalski

Laboratory for Chemometrics, Department of Chemistry, Box 351700, University of Washington, Seattle, Washington 98195

Net analyte signal plays an important role in the calculation of figures of merit for characterizing a calibration model. Until now, its computation has only been feasible for the direct calibration model, which requires knowledge of pure spectra or concentrations of all contributing species in the calibration samples. An increasingly important calibration model is the inverse calibration model, which also allows for quantitation if the knowledge about the interferents is incomplete. This paper shows that net analyte signal computation is possible for the inverse calibration case. Application to the determination of protein content in wheat samples by near-infrared spectrometry is presented. Net analyte signal calculation was used to estimate selectivities (ratio of signal available for quantitation to the total measured signal). The selectivities were found to range between 0 and 2% of the measured reflectance signal. A new measure for outlier diagnosis based on the correlation of the net analyte signal to the regression coefficients vector is introduced and tested on the same data.

Calibration relates measured variables (response measurements) to objects (samples) with known properties (concentration values) via a mathematical model. A regression model is constructed using these data. The model is then used to predict the same properties from the measured variables of unknown objects. Calibration in the context of its application to chemical analysis may be viewed as a subset of the more general regression analysis, in which the emphasis is to have the best predictive model.

The analyst is not only interested in the final output of the calibration; rather, figures of merit characterizing the whole calibration process as well as for each forthcoming sample determined by the model are required in order for the output to make sense. The figures of merit needed to characterize the prediction were summarized by Kaiser:[1] (a) signal-to-noise ratio, (b) precision or relative precision expressed as percent relative standard deviation, (c) accuracy, which is the sum of precision and bias from the true model, and (d) net analyte signal. In addition, three other figures of merit are useful to characterize the calibration model: (a) expected prediction error, (b) sensitivity, which for zeroth-order data is the slope of the calibration curve, and (c) selectivity, which accounts for the degree of interference.

Those figures of merit were introduced to characterize a univariate model, i.e., modeling with a single detector. Lorber[2] presented a generalization that allows estimation of figures of merit for multivariate data. Further considerations to generalize for more complicated data structure types were presented by Wang et al.[3] and by Messick et al.[4] The subject has been reviewed recently by Booksh and Kowalski[5] and by Kalivas and Lang.[6]

Until now, the usefulness of the figures of merit in multivariate calibration has been limited. The reason is that estimation of those figures of merit was restricted to the classical calibration model (when the pure spectra, or concentrations of all components in the calibration set, are known). Unfortunately, most of the applications in chemical analysis are of the second type of calibration, i.e., inverse calibration, which is much less demanding and only requires the knowledge of the concentrations of the analyte of interest in the calibration set.

This paper shows that the calculation of the net analyte signal is not necessarily limited by having knowledge of the pure spectra of all components present in the calibration set. It is shown that with a limited amount of information, typically found in complex situations, where only the measured spectra of a set of calibration samples and the concentrations of the analyte of interest are available, net analyte signal computation is still feasible.

THEORY

Computation of Net Analyte Signal in Classical Univariate Calibration. For pedagogical purposes we will start with a discussion of the simplest form of calibration, zeroth-order calibration (univariate calibration). The linear univariate calibration model is formulated as

$$r = cs + d + \epsilon \tag{1}$$

where r is the response of the analyzer, c is the concentration of analyte of interest, s is the sensitivity of the analyzer to the presence of the analyte, d is a constant contribution to the measured response (background or constant interfering constitu-

[†] Current address: Netherlands Forensic Science Institute, Volmerlaan 17, 2288 GD Rijswijk, The Netherlands.

[1] Kaiser, H. *Spectochim. Acta, Part B* **1978**, *33*, 551.

[2] Lorber, A. *Anal. Chem.* **1986**, *58*, 1167.

[3] Wang, Y.; Borgen, O. S.; Kowalski, B. R.; Gu, M.; Turecek, F. *J. Chemom.* **1993**, *7*, 117.

[4] Messick, N. J.; Kalivas, J. H.; Lang, P. M. *Anal. Chem.* **1996**, *68*, 1572.

[5] Booksh, K. S.; Kowalski, B. R. *Anal. Chem.* **1994**, *66*, 782A.

[6] Kalivas, J. H.; Lang, P. M. *Chemom. Intell. Lab. Syst.* **1996**, *32*, 135.

S0003-2700(96)00862-1 CCC: $14.00

ents), and ϵ is the measurement error. For a set of I measurements ($i = 1, ..., I$), the solution of the regression equation aims to estimate s, which in the case of a linear model is equal to the slope of the calibration curve. Because of the simple structure of the data, it is quite easy to generalize the model to account for nonlinearities and for various error distribution models, a feat that is difficult if not impossible in multivariate regression.

For the purpose of characterizing the calibration, it is not sufficient to consider the total response r, rather, the net analyte signal r^{net}, calculated by $r^{\text{net}} = r - d$ is a more relevant quantity, since it considers only the part of the signal usable for quantitation. The net signal is the basis for further calculations such as sensitivity, signal-to-noise ratio, and limit of detection. A related quantity, selectivity, which is defined as the ratio r^{net}/r, describes which part of the measured response is available for quantitation.

Computation of Net Analyte Signal in Classical Multivariate Calibration. As with zeroth-order calibration, derivation of figures of merit is easy after a formula for calculating the *net analyte signal* is derived. Because in zeroth-order calibration the contribution of a single analyte plus a constant term is modeled, the net analyte signal is calculated as the measured response minus the constant term. In first-order calibration, computation of net analyte signal is more complicated because one or more interfering species with varying concentration may contribute to the measured response. Lorber[2] observed that the net analyte signal of the kth analyte may be computed as the part of its spectrum orthogonal to the contribution of other coexisting constituents. The reasoning behind this postulation is not arbitrary and stems from the properties of solving a set of linear equations. The part of the spectrum that is not orthogonal to the contribution of the others is a linear combination of their pure spectra. Only the orthogonal part is unique to the sought-for analyte. Therefore, only this part is useful for quantitation or, in other words, the net analyte signal.

The part of a vector $\mathbf{u}$, orthogonal to the column space of a matrix $\mathbf{X}$, is computed by

$$\mathbf{v} = (\mathbf{I} - \mathbf{X}\mathbf{X}^{+})\mathbf{u} \tag{2}$$

where $\mathbf{v}$ is the orthogonal part, the superscript "+" symbolizes the Moore−Penrose pseudoinverse, and $\mathbf{I}$ is the identity matrix. The matrix $\mathbf{I} - \mathbf{X}\mathbf{X}^{+}$ is a projection matrix that projects $\mathbf{u}$ onto the null space of the rows of $\mathbf{X}$, which is the orthogonal complement of the column space of $\mathbf{X}$. A scalar measure for the degree of overlap is given by the ratio $\alpha = ||\mathbf{v}||/||\mathbf{u}||$, which is the sine of the angle between the two vectors. In the multivariate scenario it is an expression for selectivity, as shown below.

Having a matrix of sensitivities $\mathbf{S}$ (responses divided by the concentration of the analyte in the sample), constructed from the pure spectra $\mathbf{s}_k = (k = 1, ..., K)$ measured at J wavelengths, the part of the kth analyte sensitivities vector $\mathbf{s}_k$ that is orthogonal to the other constituents, may be computed as in eq 2 as

$$\hat{\mathbf{s}}_k^{\text{net}} = (\mathbf{I} - \mathbf{S}_{-k}\mathbf{S}_{-k}^{+})\mathbf{s}_k \tag{3}$$

where $\hat{\mathbf{s}}_k^{\text{net}}$ is the estimated net part of the kth constituent sensitivity vector and $\mathbf{S}_{-k}$ is the matrix of sensitivities of all except the kth analyte. The part of the unknown sample response vector that is unique is given by

$$\hat{\mathbf{r}}_{k,\text{un}}^{\text{net}} = (\mathbf{I} - \mathbf{S}_{-k}\mathbf{S}_{-k}^{+})\mathbf{r}_{\text{un}} \tag{4}$$

Both left-hand sides of eqs 3 and 4 contain the contribution of only the kth analyte, which in turn means that $\hat{\mathbf{r}}_{k,\text{un}}^{\text{net}}$ must be proportional to $\hat{\mathbf{s}}_k^{\text{net}}$. This proportionality factor is actually the sought-for analyte concentration, or

$$\hat{\mathbf{r}}_{k,\text{un}}^{\text{net}} = \hat{c}_{k,\text{un}}\hat{\mathbf{s}}_k^{\text{net}} \tag{5}$$

which may be rearranged as

$$\hat{c}_{k,\text{un}} = (\hat{\mathbf{r}}_{k,\text{un}}^{\text{net}})^{\text{T}}\hat{\mathbf{s}}_k^{\text{net}}/||\hat{\mathbf{s}}_k^{\text{net}}||^2 \tag{6}$$

Using the idempotent property of the projection matrix $\mathbf{I} - \mathbf{S}_{-k}\mathbf{S}_{-k}^{+}$, eq 6 may be further simplified to

$$\hat{c}_{k,\text{un}} = \mathbf{r}_{\text{un}}^{\text{T}}\hat{\mathbf{s}}_k^{\text{net}}/||\hat{\mathbf{s}}_k^{\text{net}}||^2 \tag{7}$$

Equation 7 has the same form as prediction with the regression vector $\hat{\beta}_k$, obtained by inverse multivariate calibration where the prediction equation is

$$\hat{c}_{k,\text{un}} = \mathbf{r}_{\text{un}}^{\text{T}}\hat{\beta}_k \tag{8}$$

Comparing the two equations reveals the interesting identity,

$$\hat{\beta}_k = \hat{\mathbf{s}}_k^{\text{net}}/||\hat{\mathbf{s}}_k^{\text{net}}||^2 \tag{9}$$

This equation gives insight into the meaning of the regression coefficients. It says that the regression coefficients are actually proportional to the part of the pure component spectrum that is orthogonal to the spectra of the other components in the model. This finding is also an outcome of tensor theory[7,8] where the regression coefficients vector is identified as the contravariant vector. Through the relationship in eq 9, it is possible to estimate the sensitivity, defined as $||\hat{\mathbf{s}}_k^{\text{net}}||$. The importance of the sensitivity as a contributor to prediction error and its relation to other diagnostic measures such as the variance factor were discussed by Kalivas and Lang[6] and Faber et al.[9]

The situation described above, i.e., the presence of knowledge of the pure spectra of all analytes (or equivalently, their concentrations in the training samples) is considered suitable for *classical calibration*. As seen, calculation of the net signal requires knowledge of the space spanned by the other constituents, which is generally considered to be available under classical calibration. In *inverse calibration*, only knowledge of the concentrations of the analyte of interest is enough to calculate the regression coefficients, hence, to obtain a predictive model. Only through the connection to regression coefficients made in eq 9 was the calculation of a figure of merit, i.e., sensitivity, possible. Calculation of other figures of merit still requires calculation of the net analyte signal. It is emphasized that eqs 3−9 were derived for the vector of sensitivities, not the actual measured responses of the pure component. In the review by Booksh and Kowalski,[5]

[7] Sanchez, E.; Kowalski, B. R. *J. Chemom.* **1988**, *2*, 247.
[8] Sanchez, E.; Kowalski, B. R. *J. Chemom.* **1988**, *2*, 265.
[9] Faber, N. M.; Buydens, L. M. C.; Kateman G. *J. Chemom.* **1994**, *8*, 181.

this distinction is lacking. Therefore, it was assumed that net analyte signal computation is possible even for inverse calibration from its relationship with the regression vector.

Computation of Net Analyte Signal in Inverse Multivariate Calibration. In an inverse calibration situation, the spectra at J ($j = 1, ..., J$) wavelengths (variables) are taken on I ($i = 1, ..., I$) calibration samples, giving rise to a calibration response matrix $\mathbf{R}$ of dimension $I \times J$. Additionally, it is required that the concentrations of the kth analyte, $\mathbf{c}_k$, are known for all I samples. These data are sufficient to build a calibration model.[10,11]

Having the same data, $\mathbf{R}$ and $\mathbf{c}_k$, it is also possible to eliminate the contribution of the kth constituent to the spectra in $\mathbf{R}$. In other words, it is possible to calculate the space of the spectra spanned by all constituents except the kth analyte. This is done by solving a rank annihilation problem. First, $\mathbf{R}$ is rebuilt using only A significant components, yielding the matrix $\hat{\mathbf{R}}$. This step is often necessary in the inverse calibration setting in order to avoid the inversion of the singular matrix $\mathbf{R}^T\mathbf{R}$ in the calculation of the regression coefficients. It is the common situation in, for example, near-infrared applications where the number of wavelengths J usually exceeds the number of calibration samples I. In practice, principal component regression (PCR) or partial least squares (PLS) are popular methods for the calculation of the significant components. Subsequently, the rank annihilation step in the A-dimensional space is given by

$$\hat{\mathbf{R}}_{-k} = \hat{\mathbf{R}} - \alpha\hat{\mathbf{c}}_k\hat{\mathbf{r}}^T \tag{10}$$

where $\hat{\mathbf{c}}_k$ is the concentration vector $\mathbf{c}_k$ projected down onto the A-dimensional space, i.e., $\hat{\mathbf{c}}_k = \hat{\mathbf{R}}\hat{\mathbf{R}}^+\mathbf{c}_k$, and $\hat{r}$ is a linear combination of the rows of $\hat{\mathbf{R}}$, which should include a contribution from the spectrum of the kth constituent. Although in the regular setting of rank annihilation the vector $\hat{\mathbf{r}}$ is considered to be the pure spectrum of analyte k, it is easily proven that the choice of different linear combinations will only affect the value of α. Rank annihilation will make $\hat{\mathbf{R}}_{-k}$ free from the contribution of the kth component. It is noted that the matrix $\hat{\mathbf{R}}_{-k}$ has the same size as the original response matrix $\mathbf{R}$ (and $\hat{\mathbf{R}}$) whereas $\mathbf{S}_{-k}$ has one column less than $\mathbf{S}$. However, the rank of $\hat{\mathbf{R}}_{-k}$ is A-1. The scalar α is computable as

$$\alpha = 1/\hat{\mathbf{r}}^T\hat{\mathbf{R}}^+\hat{\mathbf{c}}_k \tag{11}$$

The vector $\hat{\mathbf{R}}^+\hat{\mathbf{c}}_k$ in the dominator is the estimated regression vector $\hat{\beta}_k$ of the inverse calibration model. A derivation of eq 10 is given in the Appendix.

The net analyte signal in the most general scenario is then calculated by

$$\hat{\mathbf{r}}^{\mathrm{net}}_{k,\mathrm{un}} = (\mathbf{I} - \hat{\mathbf{R}}^T_{-k}(\hat{\mathbf{R}}^T_{-k})^+)\mathbf{r}_{\mathrm{un}} \tag{12}$$

Since $\hat{\mathbf{r}}^{\mathrm{net}}_{k,\mathrm{un}}$ is free from interference, it is possible to replace it by a scalar representation without loss of information, e.g., its Euclidean norm. With this choice of scalar representation one obtains

$$\mathrm{N\hat{A}S}_{k,\mathrm{un}} = ||\hat{\mathbf{r}}^{\mathrm{net}}_{k,\mathrm{un}}|| \tag{13}$$

It is important to note that the term net analyte signal (NAS) is used in the literature to indicate the vector as well as its size (norm). In this paper, the meaning should be clear from the context.

Calibration and Prediction by Net Analyte Signal. The possibility of computing a scalar value free from interferences from a vector containing contributions of unknowns (eq 12) enables the formulation of a new way to perform multivariate calibration, Net analyte signal (NAS) calibration. The regression step involves I calculations of NAS for all calibration samples, such as in eq 12,

$$\hat{\mathbf{r}}^{\mathrm{net}}_{k,i} = (\mathbf{I} - \hat{\mathbf{R}}^T_{-k})(\hat{\mathbf{R}}^T_{-k})^+)r_i \tag{14}$$

where $\mathbf{r}_i$ is the spectrum of the ith calibration sample.

Once all the $\mathrm{NAS}_{k,i}$ are computed as the Euclidean norm of the vectors, it is straightforward to establish a bivariate regression model between $c_{k,i}$ and $\mathrm{NAS}_{k,i}$ as

$$c_{k,i} = f_k(\mathrm{NAS}_{k,i}) + \epsilon_i \tag{15}$$

where the determination function f_k to be estimated may take any nonlinear form and ϵ_i is a residual. The advantage of NAS calibration is this extra flexibility to handle nonlinearities. Using eq 15, prediction for an unknown sample may be written as

$$\hat{c}_{k,\mathrm{un}} = \hat{f}_k(\mathrm{N\hat{A}S}_{k,\mathrm{un}}) \tag{16}$$

Equation 16 specializes in the case of the linear model to

$$\hat{c}_{k,\mathrm{un}} = \mathrm{N\hat{A}S}_{k,\mathrm{un}}/||\hat{\mathbf{s}}^{\mathrm{net}}_k|| \tag{17}$$

Since the paper concentrates on figures of merit, the differences from traditional calibration are not discussed here.

Sensitivity. Sensitivity is a figure that characterizes the calibration model and tells to what extent the response due to the particular analyte varies as a function of its concentration. Sensitivity in the context of univariate calibration is defined as the slope of the calibration curve. It is essentially a differential of response with regard to concentration. As shown in eq 9, there is a direct relationship between the regression coefficients and the sensitivities. Therefore, even in the inverse calibration case, it was still possible to estimate the sensitivities through this relationship.

With the new way to calculate the NAS presented here, it is possible to calculate the vector of sensitivities for each calibration sample i as

$$\hat{\mathbf{s}}_{k,i} = \hat{\mathbf{r}}^{\mathrm{net}}_{k,i}/c_{k,i} \tag{18}$$

In an errorless situation (both in concentration and responses), all calibration samples should produce the same vector of sensitivities. This is no longer true in real applications, and the individual estimates from eq 18 may then be combined to produce an estimate that is representative of the entire calibration set.

[10] Lorber, A. *Anal. Chim. Acta* **1984**, *164*, 293.
[11] Martens, H.; Næs, T. *Multivariate Calibration*; Wiley: New York, 1989.

Depending on the specific procedure used, the resulting sensitivity vector will be different from the one calculated from eq 9.

Selectivity. Selectivity is a measure of degree of overlap aimed to indicate what part of the total signal is lost due to overlap. Dividing $\hat{\mathrm{NAS}}_{k,\mathrm{un}}$ by the length of the original spectrum of the unknown sample gives the desired value:

$$\hat{\mathrm{SEL}}_{k,\mathrm{un}} = \hat{\mathrm{NAS}}_{k,\mathrm{un}}/||\mathbf{r}_{\mathrm{un}}|| \qquad (19)$$

It is important to note that the selectivity defined here is sample dependent and will be determined by the amount of the analyte of interest relative to the interferents. In the previous definition, the pure spectra were used for the definition and the result was characteristic for the calibration procedure. The selectivity values obtained for the individual prediction samples will be lower than the one defined for the pure spectra, and therefore, the previous definition may be regarded as the maximum obtainable selectivity. The new definition is introduced in order to overcome the impossibility of calculating the maximum obtainable selectivity in the indirect calibration scenario, since the total contribution of the analyte to the unknown sample spectrum cannot be calculated without having its pure spectrum. However, the new definition might actually be more relevant for the analyst, since it gives information on the particular situation and not on the ideal case.

Signal-to-Noise Ratio. Being able to calculate the NAS allows us to relate the useful part of the signal to the measurement noise. The signal-to-noise (S/N) ratio may be directly calculated as

$$\mathrm{S}\hat{/}\mathrm{N}_{k,\mathrm{un}} = \hat{\mathrm{NAS}}_{k,\mathrm{un}}/\delta \mathrm{r} \qquad (20)$$

where δr is an estimate for the standard deviation of the measurement errors in the response values.

In eq 20 it is assumed that the measurement errors are uncorrelated and have a constant variance. In another paper we show how to derive a more general expression for S/N that also allows for correlated and heteroscedastic errors.[12] In addition, it is assumed that the projection matrix needed to calculate $\hat{\mathbf{r}}^{\mathrm{net}}_{k,\mathrm{un}}$ in eq 12 is “relatively” precise. The latter assumption is often reasonable in practical applications and can be justified as follows. Assume that the projection matrix is calculated by means of PCA, which corresponds to the use of PCR for the estimation of the regression vector. The noise-averaging properties of PCA are well-known. They are due to the fact that the principal components (PCs) are linear combinations of an often much larger number of variables (columns of the data matrix subject to PCA), hence the uncertainty in the PCs will be (much) lower than the uncertainty in the original variables and the projection matrix constructed from those PCs will be relatively immune to noise. (The validity of this conjecture has been confirmed by calculations at our laboratory.) A similar reasoning is valid if PLS is used to estimate the regression vector.

Limit of Detection. The limit of detection (LOD) is a useful figure of merit for methods such as atomic emission spectrometry and mass spectrometry where calibration curves may be extended to the background level of the instrument. In some other spectroscopic techniques, such as near-infrared modeling where the calibration model is applied in a narrow range only, this figure is not useful.

The International Union of Pure and Applied Chemistry (IUPAC) recommends checking two hypotheses in order to arrive at the decision of whether the analyte is present at a detectable level.[13] The null hypothesis states that no analyte is present whereas under the alternative hypothesis the analyte is present at the unknown limit of detection. Checking both hypotheses allows for simultaneously controlling the rates of false positives (type I error) and false negatives (type II error). According to the IUPAC recommendation, the practical implementation may proceed in signal space or in concentration space. The principle of testing both hypotheses has very recently been applied by Boqué and Rius for the classical multivariate calibration model.[14] They formulated LOD in terms of concentrations and obtained excellent results in terms of the prediction of the rates of type I and type II errors. The following equations summarize the reformulation of the LOD of Boqué and Rius in terms of the estimated NAS calculated in the inverse multivariate calibration model:

$$\text{null hypothesis:}\quad \mathrm{H_o}:\ \hat{\mathrm{NAS}}_{k,\mathrm{un}} = r_{\mathrm{o}}$$

$$\text{alternative hypothesis}\quad \mathrm{H_A}:\ \hat{\mathrm{NAS}}_{k,\mathrm{un}} > r_{\mathrm{o}}$$

$$\text{probability of type I error:}\quad \alpha = \mathrm{pr}\{t > t_\alpha\}$$

$$\text{probability of type II error:}\quad \beta = 1 - \mathrm{pr}\{t(\Delta) > t_\alpha\}$$

Here r_{o} is the NAS at zero concentration value, $t = (\hat{\mathrm{NAS}}_{k,\mathrm{un}} - r_{\mathrm{o}})/\delta\mathrm{r}$ and Δ is the noncentrality parameter of a noncentral *t*-distribution. Under the assumption $r_{\mathrm{o}} = \mathrm{o}$, *t* is actually the signal-to-noise ratio.

The LOD may be estimated as

$$\hat{\mathrm{LOD}}_{k,\mathrm{un}} = \Delta(\alpha,\beta)\delta\mathrm{r} \qquad (21)$$

where the factor $\Delta(\alpha,\beta)$ can be obtained either numerically[14] or from statistical tables.[15] The assumptions underlying eq 20 are discussed in the preceding section. For more details, see Boqué and Rius.[14]

EXPERIMENTAL SECTION

The NAS calculation was applied to the data published by Fearn.[16] He represented the linear regression of the percentage protein in ground wheat samples against the logarithm of near-infrared reflectance intensities at six wavelengths. The reference values for protein content were determined by the Kjeldahl nitrogen method. The data consist of measurements taken on 50 samples. The data set was divided according to the original division made by Fearn into 24 calibration samples and 26 prediction samples. Only mean centering was applied as preliminary data treatment.

The equations were implemented in the Matlab language and run on an IBM compatible personal computer.

RESULTS AND DISCUSSION

The data of Fearn are used here to demonstrate some of the benefits obtainable by the possibility of calculating NAS for inverse

[12] Faber, N. M.; Lorber, A.; Kowalski, B. R. *J. Chemom.*, submitted.

[13] Currie, L. A. *Pure Appl. Chem.* **1995**, *67*, 1699.

[14] Boqué, R.; Rius, F. X. *Trends Anal. Chem.*, in preparation.

[15] Owen, D. B. *J. Am. Stat. Assoc.* **1965**, *60*, 320.

[16] Fearn, T. *Appl. Stat.* **1983**, *32*, 73.

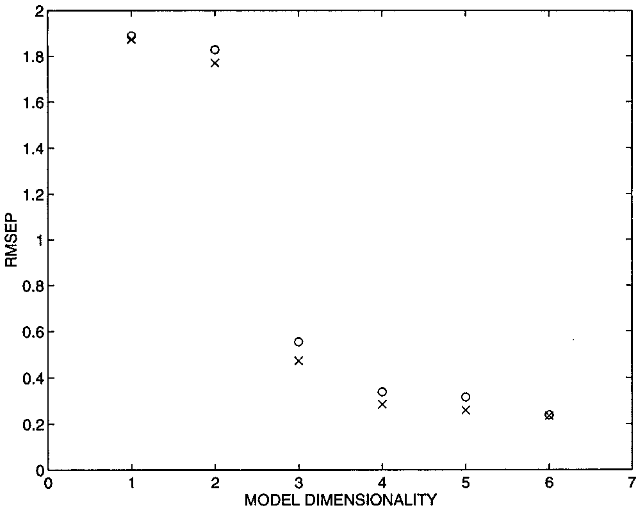

Figure 1. RMSEP versus PCR model dimensionality calculated from reference values of samples 1−26 (○) and samples 2−26 (×).

Table 1. Projection Matrix for Net Analyte Signal Calculation Based on a Four-Dimensional PCR Model

0.5164	−0.1585	−0.2493	−0.3333	0.0256	0.2253
	0.7540	−0.2036	−0.1736	0.0364	−0.2958
		0.7935	−0.2153	−0.0081	−0.1179
			0.7426	−0.0526	0.0289
				0.0049	−0.0069
					0.1886

Table 2. Prediction Results and Diagnostics for Test Samples

sample	content (% protein)	deviation (% protein)	selectivity (% total signal)	correlation	leverage
1	8.66	−0.93	1.14	−0.8739	6.85
2	7.90	−0.12	0.98	−0.9887	0.57
3	9.27	0.58	0.11	−0.4958	0.19
4	11.77	0.19	0.98	0.9973	0.48
5	9.70	0.48	0.14	0.6923	0.01
6	10.46	0.52	0.51	0.9861	0.18
7	10.17	0.31	0.28	0.8465	0.11
8	11.10	−0.02	0.47	0.9942	0.06
9	12.03	−0.48	0.71	0.9885	0.33
10	9.43	−0.01	0.49	−0.7145	0.75
11	8.66	−0.41	0.84	−0.9848	0.52
12	14.44	−0.29	2.13	0.9997	0.92
13	8.50	−0.02	0.69	−0.9764	0.36
14	10.41	−0.17	0.17	0.9087	0.50
15	9.72	−0.22	0.32	−0.6293	0.30
16	11.69	−0.30	0.74	0.9956	0.50
17	12.19	−0.12	1.03	0.9961	0.63
18	11.59	−0.04	0.81	0.9993	0.41
19	8.76	−0.27	0.65	−0.9805	0.20
20	8.60	−0.13	0.79	−0.9910	0.39
21	8.54	0.17	0.57	−0.9925	0.30
22	9.34	−0.19	0.43	−0.9483	0.26
23	10.09	0.10	0.17	0.6362	0.06
24	8.72	−0.16	0.58	−0.9703	0.24
25	10.87	−0.06	0.40	0.9631	0.19
26	10.89	0.52	0.70	0.9475	0.83

multivariate calibration. Equations 10−13 were evaluated using the four-dimensional PCR model. The results obtained for PLS are very similar and will not be presented here. The choice for four-dimensional PCR model is based on the root-mean-squared error of prediction (RMSEP), which is obtained in the usual way as

$$\mathrm{RMSEP} = \sqrt{\frac{1}{I_{\mathrm{p}}}\sum_{i=1}^{I_{\mathrm{p}}}(\hat{c}_i - c_i)^2} \qquad (22)$$

where I_{p} denotes the number of prediction samples, $\hat{c}_i$ is the model estimate for the protein content, and c_i is the reference value obtained from the Kjeldahl method. Figure 1 shows the dependency of the RMSEP estimate on the PCR model dimensionality. RMSEP is estimated using the full prediction set (26 samples) and using all prediction samples except number one. This sample is known to be an extreme outlier (see below), and it makes sense to also estimate the RMSEP without including it in the test set. Since the RMSEP estimate from eq 22 is based on squared deviations, it is sensitive to samples that may have an exceptionally high prediction error. Excluding such samples will give a more robust estimate. It is seen from this plot that only a marginal decrease in estimated RMSEP is obtained by increasing the model dimensionality beyond four. The reduction of the RMSEP estimate resulting from excluding the outlying sample is appreciable for the four-dimensional model.

Table 1 gives the projection matrix that is used for the net analyte signal calculation in eq 12. Since an orthogonal projection matrix is symmetric, only the upper right corner is shown. It is easily verified that this matrix has rank three, which is the dimension of the PCR model minus one. For a projection matrix, the rank should be equal to the sum of the diagonal elements (the trace). This property can be used to test the adequacy of the calculations leading to the projection matrix.

Table 2 summarizes the prediction results and diagnostics for the entire test set. The second column lists the reference values that are used to validate the model in terms of the RMSEP. The third column gives the differences between the model estimates and the reference values, which are the individual contributions to eq 22. The next column gives the selectivities calculated according to eq 19. It is seen that the selectivities are indeed sample dependent. The fifth column lists the correlations between the net analyte signal vector and the regression coefficients. A value close to −1 or +1 is indicative of a NAS that is not degraded by undesired influences. The last column gives the leverage, which is closely related to the Mahalanobis distance. The Mahalanobis distance is a measure for the position of a sample in calibration space. It is independent of the number of calibration samples. The leverage, however, decreases with increasing number of calibration samples. The leverage converges to zero if the number of calibration samples approaches infinity and the samples spread out well enough. Outliers are identified by comparing their leverage to the average value obtained for the calibration set. If PCR is used to estimate the model parameters, this value is calculated as the ratio of the model dimensionality and the number of calibration samples.[11] The resulting average for the calibration set is 4/24 ≈ 0.17. It is clear from comparing this number to the numbers in the last column of Table 2 that the first prediction sample is an extreme outlier. As expected, the prediction error (0.93%) is much higher than the RMSEP estimated from the reduced set (0.29%). In the next sections, the quantitative information collected in Table 2 will be displayed graphically and discussed in more detail.

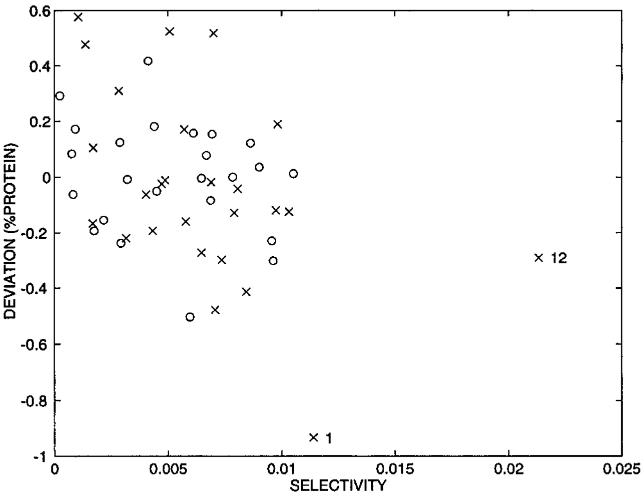

Figure 2. Prediction error versus selectivity for calibration (○) and prediction samples (×).

Selectivity Determination. Selectivity measures the degree of overlap of the compound of interest. The results obtained for Fearn's data are displayed in Figure 2. As seen in the plot, the selectivity values are lower than 0.025, the large majority having selectivities even smaller than 0.01. The upper bound for the selectivity of 0.025 means that the analyst may expect an error propagation greater than 40 for this data set.[2] Although this appears to be a bad situation, it must be borne in mind that the determination of protein content in wheat samples is considered to be an easy application for near-infrared spectrometry. Building a global calibration model is feasible, and manufacturers of instruments can bring down the between-instrument variation to a value much lower than 0.025. Applications such as the determination of octane number of gasoline are considered to be more difficult and prone to variation within and between instruments.

It is clear from Figure 2 that the selectivity itself has no correlation to the deviation of determined protein content from the actually measured one. For example, prediction sample 1 has a "reasonable" selectivity but a large prediction error and prediction sample 12 has a "high" selectivity but this is not reflected in a proportionally low prediction error. Some afterthought shows that the observed behavior is to be expected, since low selectivity values may arise from (at least) two entirely different sources. First, there are samples having a concentration of the analyte close to the mean. These samples will have a low NAS because the data are mean-centered. However, these samples occupy a favorable position in calibration space, and consequently, the prediction error should be low, since the uncertainty in the estimated model contributes only little to prediction error. The situation is entirely different for samples that are relatively far away from the center. An eccentric position may lead to a relatively low NAS because the estimated model and, hence, the projection matrix is not reliable for such a sample. From a plot like Figure 2, those two cases are not distinguishable. This is simply the consequence of using a mean-centered model. Independent of these effects, there is another reason that precludes a straightforward interpretation of Figure 2. The reference values from which the prediction errors are estimated are contaminated by a measurement error with a standard deviation of ~0.2%.[11] This means that the true deviations are poorly estimated for this application. Whereas the former complications have an impact on the abscissa values, this complication has an effect on the ordinate values.

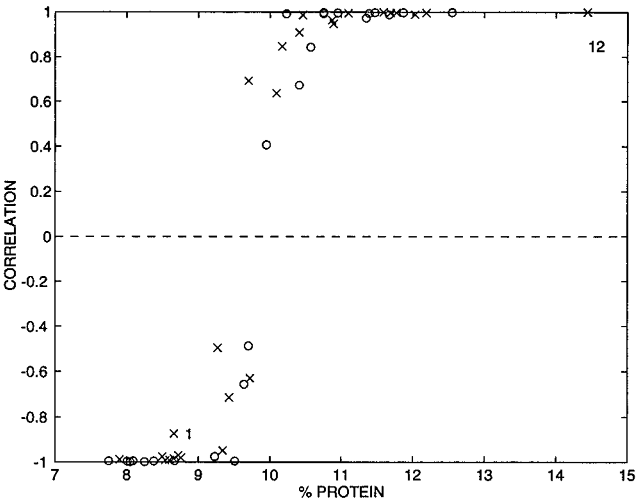

Figure 3. Correlation of NAS with regression coefficients vector versus percentage protein for calibration (○) and prediction samples (×).

Outlier Detection in Prediction Samples. The Mahalanobis distance (similar information as prediction leverage) and spectral residuals are the tools currently used for diagnostic purposes. The Mahalanobis distance provides information about to what degree the particular prediction sample is far away from the design points used in the calibration set. The spectral residuals are also a useful diagnostic tool, but they may be obtained only for overdetermined cases as in PCR or PLS.

The possibility to calculate the NAS presented here allows the introduction of an additional tool. Consider the following correlation,

$$\mathrm{corr}(\hat{\mathbf{r}}^{\mathrm{net}}_{k,\mathrm{un}},\hat{\beta}_k) \qquad (23)$$

as a measure of *degree of purity* of the estimated net analyte signal. The regression vector $\hat{\beta}_k$ is obtained from the calibration set; thus it may be considered as the best estimator of the NAS in the calibration set (up to a scalar multiplication). The estimated NAS of a prediction sample contains a contribution from two legitimate sources, i.e., the actual net signal and random noise. In addition, it also grabs contributions from new sources of variability that were not accounted for in the calibration phase. Thus, the correlation presented in eq 23 is a measure of purity of the estimated NAS.

The new correlation measure was applied to Fearn's data, and the results are presented in Figures 3 and 4. Figure 3 is a scatterplot of the calculated correlation against the reference values of protein content. It is seen that usually the correlation tends to approach the anticipated ±1 values. As expected (due to mean-centering the data), correlation values around the mean concentration value for the calibration set (9.97%) tend to be lower. In this data set, there is one sample with 8.66% protein content whose correlation value is −0.8739. It has a behavior different from the samples in its close proximity. This sample (number 1) has been identified above as an outlier by its extremely large prediction leverage. Furthermore, it is interesting to note that the model has been used for extrapolation in the case of sample

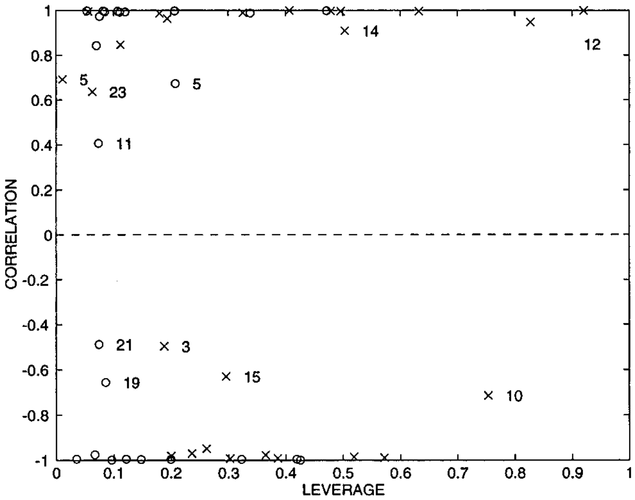

Figure 4. Correlation of NAS with regression coefficients vector versus leverage for calibration (○) and prediction samples (×).

12. This sample has the highest correlation value (0.9997) of the entire test set. This excellent correlation lends credibility to the obtained concentration estimate and strongly indicates the usefulness of the correlation criteria. Figure 4 is a scatterplot of the calculated correlation against the leverage. This figure further illustrates the tendency of the correlation to be small for a sample that is close to the model center in multivariate spectral space. This holds for the calibration samples (e.g., 5, 11, 19, and 21) as well as the prediction samples (e.g., 3, 5, 15, and 23). In addition, there are high leverage points with low correlation (e.g., 10) and extremely high correlation (e.g., 12). The prediction result for sample 10 should be considered with care. The prediction error based on comparison with the reference value is −0.01%. This value appears to be very good. However, it is much smaller than the measurement error in the reference method (0.20%) and therefore cannot be a realistic estimate for the true prediction error. The extreme outlier (sample 1) is not displayed in this plot for ease of presentation. It has already been discussed above.

CONCLUSIONS

It has been known for 15 years that quantitation is possible for inverse multivariate calibration. However, until now it has not been noticed that the same mathematical fact that allows for quantitation also allows computation of the space required to calculate the NAS. The usefulness of this finding to calculate figures of merit is an obvious outcome. As demonstrated, the possibility to calculate selectivity in near-infrared applications can help the analyst to better appreciate the limitation/advantages of the spectroscopic method. The proposed diagnostic tool of correlating the NAS to the regression coefficients vector has the potential to add to the arsenal needed to diagnose when the model is taken to its limits.

We are currently exploring different cases where NAS calculation can enhance the interpretation of multivariate calibration results.

ACKNOWLEDGMENT

This work was supported by the Center for Process Analytical Chemistry (CPAC), a National Science Foundation Industry/University Cooperative Research Center at the University of Washington. Constructive criticism by the reviewers is appreciated by the authors.

APPENDIX: DERIVATION OF EQ 10

In the following derivation, the "hats" are dropped from the symbols to simplify the notation. The calibration response matrix **R** can be expressed in terms of pure component contributions as

$$\mathbf{R} = \sum_{l=1}^{K} \mathbf{c}_l \mathbf{s}_l^{\mathrm{T}} \qquad (24)$$

Consequently, the linear combination **r** that is used in the rank annihilation step is given by

$$\mathbf{r}^{\mathrm{T}} = \mathbf{w}^{\mathrm{T}}\mathbf{R} = \mathbf{w}^{\mathrm{T}}\sum_{l=1}^{K} \mathbf{c}_l \mathbf{s}_l^{\mathrm{T}} \qquad (25)$$

where the weight vector **w** should be selected in such a way that **r** includes a contribution of $\mathbf{s}_k$. Using eq 25 eq 10 is worked out as

$$\mathbf{R}_{-k} = \mathbf{R} - \alpha\mathbf{c}_k\mathbf{w}^{\mathrm{T}}\mathbf{c}_k\mathbf{s}_k^{\mathrm{T}} - \alpha\mathbf{c}_k\mathbf{w}^{\mathrm{T}}\sum_{l\neq k}^{K} \mathbf{c}_l \mathbf{s}_l^{\mathrm{T}} \qquad (26)$$

It is seen that only the second term on the right-hand side of eq 26 can annihilate the contribution of the *k*th component to **R**. This term can be rewritten as

$$\alpha\mathbf{c}_{\mathrm{k}}\mathbf{w}^{\mathrm{T}}\mathbf{c}_k\mathbf{s}_k^{\mathrm{T}} = \alpha\cdot\mathbf{w}^{\mathrm{T}}\mathbf{c}_k\cdot\mathbf{c}_k\mathbf{s}_{\mathrm{k}}^{\mathrm{T}} \qquad (27)$$

and it is evident that the rank annihilation step is successful if

$$\alpha = 1/\mathbf{w}^{\mathrm{T}}\mathbf{c}_k \qquad (28)$$

Finally, from eq 25 it follows that the least-squares estimator for the weight vector **w** is given by

$$\mathbf{w}^{\mathrm{T}} = \mathbf{r}^{\mathrm{T}}\mathbf{R}^{+} \qquad (29)$$

and inserting eq 29 in eq 28 gives eq 11.

Received for review August 23, 1996. Accepted January 7, 1997.[⊗]

AC960862B

[⊗] Abstract published in *Advance ACS Abstracts,* February 15, 1997.