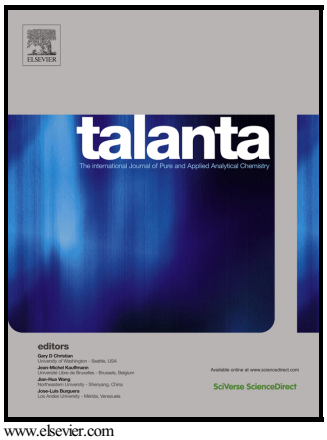
Comparison of the single channel and multichannel(multivariate) concepts of selectivity in analytical chemistry

Zsanett Dorkó, Tatjana Verbić, George Horvai



www.elsevier.com

PII: S0039-9140(15)00112-5
DOI: http://dx.doi.org/10.1016/j.talanta.2015.02.030
Reference: TAL15403

To appear in: *Talanta*

Received date: 21 November 2014
Revised date: 5 February 2015
Accepted date: 17 February 2015

Cite this article as: Zsanett Dorkó, Tatjana Verbić and George Horvai, Comparison of the single channel and multichannel(multivariate) concepts of selectivity in analytical chemistry, *Talanta,* http://dx.doi.org/10.1016/j.talanta.2015.02.030

Comparison of the single channel and multichannel(multivariate) concepts of selectivity in analytical chemistry

Zsanett Dorkó[a], Tatjana Verbić[b], George Horvai[a,c,*]

[a]*Department of Inorganic and Analytical Chemistry, Budapest University of Technology and Economics, Szent Gellert ter4., H-1111 Budapest, Hungary; zsdorko@mail.bme.hu*

[b]*Faculty of Chemistry, University of Belgrade, Studentski Trg 12-16, 11000 Belgrade, Serbia; tatjanad@chem.bg.ac.rs*

[c]*MTA-BME Research Group of Technical Analytical Chemistry, Szent Gellert ter4., H-1111 Budapest, Hungary*
[*]*Phone: +36-1-4631480; e-mail: george.horvai@mail.bme.hu*

Abstract

Different measures of selectivity are in use for single channel and multichannel linear analytical measurements, respectively. It is important to understand that these two measures express related but still distinctly different features of the respective measurements. These relationships are clarified by introducing new arguments. The most widely used selectivity measure of multichannel linear methods (which is based on the net analyte signal, NAS, concept) expresses the *sensitivity to random errors* of a determination where all bias from interferents is computationally eliminated using pure component spectra. The conventional selectivity measure of single channel linear measurements, on the other hand, helps to estimate the *bias caused by an interferent* in a biased measurement. In single channel methods expert knowledge about the samples is used to limit the possible range of interferent concentrations. The same kind of expert knowledge allows improved (lower mean squared error, MSE) analyte determinations also in “classical” multichannel measurements if those are intractable due to perfect collinearity or to high noise inflation. To achieve this goal bias variance tradeoff is employed, hence there remains some *bias* in the results and therefore the concept of single channel selectivity can be extended in a natural way to multichannel measurements. This extended definition and the resulting selectivity measure can also be applied to the so-called inverse multivariate methods like partial least squares regression (PLSR), principal component regression (PCR) and ridge regression (RR).

Keywords: selectivity, error inflation, interference, multivariate, bias variance tradeoff

1. Introduction

Selectivity is a central concept in analytical chemistry[1]. Without selective methods analytical measurement of individual components’ concentrations in mixtures would not be possible. A general definition of selectivity and particularly its quantification are quite difficult [2-3]. However, if a measured signal depends linearly on the concentrations of some components in the sample (e.g., in absorption spectrophotometry), acceptable measures of selectivity can be obtained. There have been, indeed, two main trends for defining the selectivity of linear methods. In measurements on a single channel (e.g., on a single wavelength or with a single sensor) selectivity is commonly defined

[3-5]as the ratio of analyte sensitivity to interferent sensitivity. In multichannel (multivariate) analysis other selectivity measures have been proposed [6-8], and the one introduced by Lorber [7](to be explained later in this paper) appears to be the most accepted.

It will be shown in this paper that this widely accepted multichannel selectivity is not a simple extension of the single channel selectivity concept. The two selectivities reflect two different approaches of analytical chemists to solve the same problem, the determination of an analyte concentration in samples where interferents may be present. It will also be shown that a direct extension of the single channel approach to multichannel measurements is possible. This may result in better analytical results and easier methods and one can also define and measure selectivity in accordance with the single channel methods.

This paper is part of an effort to clarify the concept of analytical selectivity in systems both with linear and nonlinear responses and with one or more measurement channels [2-3, 9].

Scalar quantities will be denoted in this paper by lowercase letters, vectors as lowercase bold face letters, matrices as uppercase letters. Row vectors and column vectors will not be differentiated as this will be clear from the context. Vector multiplication means always the scalar product. Vector norms (Euclidean) are denoted by double vertical lines.

2. Definitions of analytical selectivity in linear methods

2.1. Single channel linear methods

In many analytical methods the measured signal(s) depend linearly on the analyte concentration and also on the concentrations of some interferents potentially present in the investigated samples. Such techniques are, for example, absorption spectrometries, where the Lambert-Beer law has wide validity:

$$A_\lambda = s_A c_A + s_B c_B + s_C c_C + \ldots + \varepsilon \qquad (1)$$

Here A_λ is the measured absorbance signal at wavelength λ, the c-s are concentrations, the s-s are sensitivities (typically all non-negative, and this non-negativity will be assumed throughout this paper) and the lower case indices denote different compounds: A is the analyte, B, C, and possibly others are interferents; ε is the random error of the absorbance measurement (not the molar absorbance coefficient). Let us assume that the sensitivities s_A, s_B, etc., are known accurately and precisely from a preceding calibration and the linear model is also accurate. If the absorbance is measured on a single wavelength (or more generally a single measurement channel is used) then the concentration of the analyte, c_A, cannot be determined from the measured absorbance alone, because the interferent concentrations c_B, c_C, etc., are also unknown and only a single equation is available. But we may have additional information which limits the possible range of c_A. A natural constraint is that all concentrations are non-negative. In many linear methods also the sensitivities are non-negative (see above). These two conditions limit the possible range of c_A between the detection limit and A_λ/s_A (neglecting the random error at this upper limit). This range is still too wide and further information is needed to estimate c_A more sharply, i.e., with less error. Before we show

how, let us calculate) the relative error of c_A in a single channel measurement from Eq.(1). For simplicity we consider only one interferent, B.

$$\frac{\frac{A_\lambda}{s_A} - c_A}{c_A} = \frac{s_B c_B}{s_A c_A} + \frac{\varepsilon}{s_A c_A} \qquad (2)$$

The first term on the right hand side is the interference effect or relative bias. The bias itself is $s_B c_B/s_A$. Both depend on the ratio of the respective sensitivities, but also on the concentration(s). The ratio of sensitivities is then a characteristic quantity. Its reciprocal, s_A/s_B, may be considered the selectivity measure of the method. The higher this selectivity measure is, the less interference (relative bias) will be observed at a given ratio of the concentrations. Since in analytical chemistry the (relative) error of the analyte concentration estimate is very important, this definition of selectivity makes sense. Indeed this is the traditional definition of single channel selectivity [4-5].

The selectivity, s_A/s_B, is the ratio of two sensitivities. As the sensitivities express signal changes per unit concentration change, their ratio shows the necessary change in c_B to bias c_A by one concentration unit. For the same reason *the reciprocal of the selectivity shows the (change of) bias in c_A caused by unit change in c_B*. This formulation will be extended later in this paper to multichannel methods.

Two things need to be noted here. First, the selectivity, s_A/s_B is used in the estimation of the bias, not the random error. This will be very important later in the discussion of multichannel selectivity. Second, s_A/s_B is not sufficient alone to estimate the relative bias. The concentration ratio, c_B/c_A is also needed. Although the individual concentrations c_A and c_B, respectively, are unknown, the analyst may have some information about their ratio. For example the analyst may know from experience with the samples at hand that the ratio $c_B/c_A$is less than 0.01 in all samples, i.e, $c_B/c_A<0.01$. This inequality is a very useful constraint. For example if $s_A/s_B= 2$, then the bias in the determination of the analyte concentration is found from Eq. (2) to be less than 0.5×0.01= 0.005, i.e. 0.5%.

Generalizing what has been said above, Eq. (1) is an underdetermined linear "equation system" consisting of a single equation with two unknown concentrations. To obtain a sufficiently close estimate of the analyte concentration, further information is needed about the concentration variables. Such information may be further equations, which are derived from additional measurements, like absorbance readings at multiple wavelengths. This is the case in multichannel measurements, where the goal may be to completely eliminate the bias due to interferents. This will be discussed later. But the analyst may not want to eliminate the effect of interferents completely, since she *needs only to keep the total uncertainty of c_A below certain, predetermined limit*[10]. Therefore she may be satisfied to know that the first term on the right hand side of Eq.(2) is below a certain limit. For a method with given selectivity this means she needs to make sure that the concentration ratio c_B/c_A is less than a certain limit. Mathematically this is a constraint on the two variables in the form of an inequality:

$$\frac{c_B}{c_A} \leq u_{lim} \qquad (3)$$

where u_{lim} is an upper limit. The analyst may know from experience with earlier samples that this limit is never exceeded; she may even ascertain this in new samples by some semiquantitative tests. Alternatively it may be enough to know or to prove that c_B is less than a certain limit, if one knows simultaneously that c_A is higher than a certain minimum in every sample. Occasionally the analyst may not know these relationships a priori, but she may perform a sample pretreatment operation which leads to the required relation. It is also possible that a method developer includes in the description of the method that some concentrations, or concentration ratios, must not exceed a certain value for the method to be sufficiently accurate. Such considerations are very common in some areas of analytical chemistry, e.g., in ion selective electrode potentiometry. But less explicitly than in potentiometry, they are used in essentially all single channel analytical measurements, because analytical chemists do not bother with interferents which are extremely unlikely to be present in samples in appreciable concentrations compared to the analyte. Interferents with very low sensitivity values can also be mostly disregarded because the bias caused by them is well within the tolerance limits. This sort of neglecting minor interferences does not work in some *multichannel* measurement methods, where the mere assumption that an interferent may be present, can be the cause of very large analytical errors. One goal of this paper is to show how to avoid this situation.

2.2. Selectivity concepts for multichannel linear methods

Some decades ago it became feasible to make quickly and at low cost multichannel analytical measurements, e.g., in the form of full spectra or of sensor array readings. In many instances this had made possible to obtain fully determined or even overdetermined equation systems of the kind of Eq. (1). This means that, if the determinant of the equation system is not zero, and if the pure component spectra are all available, then all concentrations in the equation system can be determined *without bias* caused by other components in the equations. (Other sources of bias, like unmodeled interferents, imprecise calibration, unmodeled nonlinearities, etc. are not being considered here.) Such measurements are therefore totally selective (“specific”) for the analyte (and also for the interferents). It was therefore thought that the necessity for using selectivity, as a measure of bias caused by interferents, became superfluous with multichannel measurements. There were, however, other, new problems discovered, which were attributable to the interferences. Therefore new definitions of selectivity were introduced by several workers to quantitate such new effects.

The relationship between these new, multichannel or multivariate measures of selectivity and the preexisting, above explained single channel selectivity is still not sufficiently clarified. This issue will be investigated here after introducing the probably most widely used multichannel selectivity measure.

When speaking here of underdetermined and well (or fully) determined equation systems, a difference between mathematics and analytical chemistry (and statistics) should be pointed out. The mathematical classification considers the simultaneous determination of all (concentration) variables. For the analytical chemist (and the statistician) the problem may be differently posed when she needs only the analyte concentration and is not interested in the other (interferent) concentrations. This point will be reemphasized later in this paper.

2.3. Multichannel selectivity concept based on the net analyte signal (NAS)

Kaiser [6, 11-12] had extensively discussed the new possibilities arising with the use of multichannel linear methods. Although the usefulness of Kaiser's definitions is unclear (he himself had also expressed his doubts in this respect), they had made a great influence on later researchers.

An important work, which was also considered by its author an improvement of Kaiser's concepts, has been the paper of Lorber [7], where a new selectivity measure was proposed. A IUPAC technical report [13], published twenty years later, in 2006, asserted that this selectivity criterion (which is essentially identical (see[13-15]) to that of Bergmann et al. [8]), is the most suitable among those published. Very recently a review by Olivieri [16] confirms that this is still the prime selectivity measure in multivariate analysis. A tutorial by Kalivas and Lang [17]includes moreover an explanation of the relationship between the selectivity measure of Lorber [7] and the variance inflation factor employed in mathematical statistics. Furthermore, it has been mathematically proved that the widely used partial least squares regression (PLSR) method calculates actually (in a noise free setting) the "net analyte signal" used by Lorber to define selectivity [18]. Therefore, in the rest of this paper the selectivity measure introduced in [7] will be more thoroughly investigated.

In contrast to Kaiser, who was mainly interested in the determination of all unknown concentrations, Lorber [7]was closer to the analytical chemists' approach and developed a selectivity definition which relates to the determination of only a single component of the mixture. One way to appreciate the importance of this idea is to consider a case when the spectra (which can be considered as vectors, see below) of three interferents are coplanar among themselves (but not with the analyte spectrum). This does not prevent or deteriorate the determination of the analyte concentration even though the sensitivity matrix has lower rank than the number of components. In this case it is not possible to calculate the concentration of all components, but the analyte concentration can be determined.

Lorber defined analytical selectivity with reference to the "net analyte signal" and to error inflation. Therefore these quantities need to be discussed before proceeding further.

2.4. The net analyte signal

The idea of the net analyte signal is best understood if absorbance spectra are considered. As a visual help, Fig.1 shows a special case where only two compounds are considered (the analyte A and the interferent B), and absorbance readings are taken only at two wavelength values.(Note that the latter restriction is not necessary because two vectors always define a plane in whatever dimension. Thus if one considers spectra measured at hundreds of wavelength values, Fig.1 would look the same, except for the position of coordinate axes.)

Fig. 1.

Let us consider absorbances $A_{\lambda j}$ measured at wavelengths λ_j ($A_{\lambda 1}$ and $A_{\lambda 2}$ in Fig.1). These absorbance values run from j=1 to j=p and constitute a p-vector. Let us assume that we know the spectrum of

each compound (at unit concentration) exactly (s_A and s_B in Fig.1). These spectra are now *p*-vectors and if organized in rows according to the components, they constitute a $k \times p$ matrix (where *k* is the number of components), which will be called here the sensitivity matrix and its rows the sensitivity vectors. (The same matrix is sometimes denoted in the literature as the K matrix. Note also that in part of the literature the transpose of this matrix is defined as the sensitivity matrix.)One may consider the linear subspace defined by the respective sensitivity vectors of all compounds except for the analyte. In Fig.1 this subspace is the one-dimensional line s_B. One can decompose the sensitivity vector of the analyte into its orthogonal projection s_A' onto this subspace, and into another vector, s_A*, which is orthogonal to the subspace. This decomposition is unique. The component which is orthogonal to the subspace, i.e. s_A* (or just the length of this vector) is called the net analyte signal (NAS) of the analyte at unit concentration. If the spectrum of a mixture, s, is measured, this mixture spectrum-vector can be decomposed in the same way and one obtains the net analyte signal of the mixture, NAS_s. Since

$$\mathbf{s} = c_A\mathbf{s}_A + c_B\mathbf{s}_B = c_A(\mathbf{s}_A^* + \mathbf{s}_A') + c_B\mathbf{s}_B = c_A\mathbf{s}_A^* + (c_A\mathbf{s}_A' + c_B\mathbf{s}_B) \quad (4)$$

one obtains that

$$NAS_S = c_A\|\mathbf{s}_\mathbf{A}^*\| = c_A\|\mathbf{s}_A\|sin\alpha \quad (5)$$

where double vertical lines denote vector length (Euclidean norm) and α is the angle between the two sensitivity vectors.

The analyte concentration can be obtained as:

$$c_A = \frac{NAS_s}{\|\mathbf{s}_A^*\|} \quad (6)$$

Note that Eq.(6) is only valid if α≠0.If α=0, the equation system cannot be solved for c_A. If α≠0 one obtains c_A accurately, i.e., without bias. Assuming that the pure components' spectra had been accurately and precisely measured and the linear model is perfectly valid (these ideal cases can, of course, only be approximated), and any background signal has been subtracted, any uncertainty of c_A will be due to the errors of signal measurement. In the following text it will be assumed that the signal measurement has unbiased random error.

Fig.1 also shows, for comparison and later use, the vector s_1, which is the spectrum of a solution containing compound A in concentration c_A, (just like in solution s) but not containing any B.

2.5. Error inflation

If repeated measurements are made on the same sample, the resulting spectra will include different experimental errors. The endpoints of the spectrum vectors will be scattered in some pattern due to the noise. Lines of constant probability density may be plotted, e.g., like the two circles in Fig.2. The circle around point $s_1$represents the error distribution for a pure A solution, the circle around s shows

the error distribution for a mixture containing both A and B. An assumption in many papers on multichannel selectivity has been that the noise is independent and identically distributed (e.g. with Gaussian distribution) and therefore the constant probability lines are circles and the radius (*R*) of these circles (which is here the standard deviation of the noise, often denoted by σ_ε) is independent from the sample composition and thus also from the actual absorbance values. (Note, however, that this assumption is not naturally valid for all analytical measurements [15, 19].Definition of multichannel selectivity becomes, however, somewhat more complicated in other cases, and is not considered here for simplicity.)

Fig. 2.

It has been recognized that, under the assumptions of the noise features made above, the determination of c_A can only be done with error inflation. By this one means that the relative standard deviation of the measured c_A is higher when measured in a mixture than when measured in pure A solution. This is quite simply shown in Fig.2. In pure A solution the distance between the two triangles (2*R*) gives the error bar of c_A (two times its standard deviation). In the case of the mixture s the distance between the two squares (2*R*) gives the error bar of NAS_s. The ratio of the two relative errors has been called the error inflation factor. By denoting the radius of the noise circle by *R*, the error inflation factor (eif_A) is given by the ratio of the respective noise/signal values and turns out to be equal to the reciprocal of sinα:

$$eif_A = \frac{\frac{R}{c_A \left\| \mathbf{s}_\mathbf{A}^* \right\|}}{\frac{R}{c_A \left\| \mathbf{s}_\mathbf{A} \right\|}} = \frac{1}{\sin\alpha} \qquad (7)$$

Obviously, if α is a small angle, the error inflation is very high.

A remarkable property of the error inflation factor is that it does not depend on the actual concentrations since it is completely defined by sinα, and thus by the directions of the vectors s_A and s_B. This concentration independence means that the error inflation factor can be very high even if the interferent concentration is zero, i.e., if there is no interferent in the sample. The error inflation factor arises by including the interferent B into the measurement model, not because B is present in the actual sample. In other words the error inflation arises because one cannot exclude the presence of the interferent in the samples.

2.6. Selectivity based on the net analyte signal

Selectivity has been defined by Lorber as a measure of the overlap of the analyte spectrum s_A with the interferent spectrum or spectra. He had recognized that several choices to measure the degree of overlap may be used. He chose the following definition:

$$Sel_A = \frac{\left\|\mathbf{s}_A^*\right\|}{\left\|\mathbf{s}_A\right\|} = sin\alpha \qquad (8)$$

with the argument that the selectivity defined by Eq. (8) is the reciprocal of the error inflation factor of Eq. (7). Note that with this definition, if the selectivity is small, error inflation is large. If the selectivity (and thus α) tends to zero, the error inflation factor (which is the reciprocal of Sel_A) goes to infinity.

In the following sections arguments will be made to show that this definition of selectivity contradicts the traditional single channel selectivity definition explained earlier in this paper. It will also be shown that it is not a sufficiently general measure of error inflation. Moreover it will be shown that it is possible to define multichannel selectivity in a manner which is a natural extension of the traditional selectivity concept.

3. Comparison of the single channel and multichannel selectivity definitions

Both the single channel and the NAS based multichannel selectivity defined above are useful measures of the respective methods' tolerance against interference related errors. At the same time the two types of measure are also very different. The single channel selectivity expresses tolerance against bias (systematic error) caused by an interferent. The multichannel selectivity shows tolerance against random error inflation when the bias is eliminated completely.

Single channel selectivity and the NAS based multichannel selectivity represent two different approaches to the problem of measuring an analyte's concentration in the presence of interferents. If one has multiple channels available for measurement and none of these is specific (fully selective) for the analyte then one is looking for a scalar (and monotonous and hopefully linear) function of the ensemble of measured responses (i.e., of the sample "spectrum") which is totally selective for the analyte. In a linear system (Eq. (4), where the spectra of all pure components are known, i.e., Eq.(4) is extended to several interferents and the equation system can be solved for c_A) the scalar product of the sample spectrum with the NAS vector is just such a function. This was an ingenious discovery but since this function behaves just like a totally selective single channel response (i.e., it depends only on the analyte concentration, independently from the concentrations of the interferents), it is not meaningful to speak about the selectivity of such multichannel analyses in the same sense as in single channel analyses. A problem with this function is that it may deliver analyte concentrations which are very sensitive to the random measurement errors of the multichannel (spectral) measurement. This is likely to happen if the length of the NAS vector (at unit analyte concentration) is small compared to the noise. This is why Sel_A is a useful quantity: it tells something about the noise sensitivity of the totally selective response. Yet it is also clear that the selectivity of a totally selective response is infinitely high (in the single channel sense) and therefore the two terminologies (single channel and multichannel) become contradicting.

The contradiction between the two selectivities (single channel vs. multichannel) may be appreciated by the following numerical example.

Mixtures consisting of three components are considered. Table 1 shows signals of the three compounds on three channels (sensors) at unit concentrations of each compound. One may see immediately that the third line of the table is equal to the second line plus 0.001 times the first line. Therefore the determinant of this 3×3 matrix is zero.

Table 1. Signals of three compounds on three channels at unit concentrations – the multichannel selectivity is zero for the analyte

Components	Channel 1	Channel 2	Channel 3
A	100	20	100
B	0.2	0.1	0.1
C	0.3	0.12	0.2

Each of the three sensors in this example is individually very selective for compound A, yet they constitute together a sensor array with zero multichannel selectivity for A. This peculiar situation can be regarded a simple consequence of the different definitions of single channel and multichannel selectivity, or as a sign of contradiction in terminology.

An obvious, but not very important further difference between single channel and multichannel selectivity, is that their numerical values span different ranges: 0 to infinity for single channel and 0 to 1 for multichannel.

The most recent IUPAC recommendation on the terminology for selectivity [20] discussed both single channel and multichannel selectivity but did not mention the differences between these concepts. Therefore the present work may complement that recommendation in this respect.

4. Practical consequences: reducing error inflation at the cost of bias

While discovering a conflict of terminology is certainly interesting, it is not the only conclusion from the above discussion. To go a step further, one should recognize that in multichannel measurements it may be impossible to find a scalar function of the multiple responses which is totally selective for the analyte concentration. For example, in linear systems described by Eq.(4) (extended to several interferents) no such response is found (at least as a linear combination of the measured signals) if the NAS vector is zero, i.e., when the analyte spectrum is a linear combination of some interferents' spectra. And even if the NAS vector is not zero, but its length at unit analyte concentration is small compared to noise, it is not very useful to calculate the analyte concentration from the NAS component of the sample spectrum.

This apparently hopeless situation may be resolved in the same way as it is done in the single channel case. One should recognize that in the multichannel case one was looking for a scalar function of the

measured signals at all channels so that this function depends exclusively (i.e., totally selectively) on the analyte concentration *at any mathematically possible values of the interferent concentrations*. On the other hand, if one has only a single channel for measurement and this channel is not fully selective to the analyte, then the analyst asks: do I really need to expect *any mathematically possible sample compositions*? Very often the answer is no. One may know from experience with a given sample type that the composition of the samples will be in some limited range. The question is then if in this limited range one may estimate c_A as if the single signal depended only on the analyte concentration. If the analyte is determined under this assumption one will have some bias in the result (Eq. (2)). So the question to ask is if the maximum bias which may occur in the limited concentration range is tolerable. It was shown earlier in this paper as a sequel to Eq.(2) that in single channel measurements the bias may be tolerable. There is no *a priori* reason why this procedure could not be extended to multichannel measurements *with known spectra for all compounds*. How this may work is best shown by a numerical example. Later this example will be generalized and compared to the so-called inverse methods where this tradeoff between bias and variance is routinely used, but requires a set of samples which is representative for all future samples to be available.

Table 2. Signals of three compounds on three channels at unit concentrations – the error inflation is very high

Components	A (λ_1)	A (λ_2)	A (λ_3)
A	101	20	100
B	20	101	100
C	120	120	200

Table 2 shows absorbances (e.g. in mAU units) of three compounds at unit concentrations of each at three wavelength values. One may immediately see that the third line of the table is very close to the sum of the first two lines. Therefore the determinant of this 3×3 matrix is close to zero and the ratio of the highest to the lowest singular value of the matrix is high. (The singular values are 321.6, 81.0 and 0.622). The value of Sel_A=sin α is 0.0075, i.e., very small. Therefore error inflation will be very high. For example if the composition of an analytical sample is the following:

c_A=1; c_B=1; c_C=0.01

then an error free measurement of the three absorbances would yield 122.2, 122.2 and 202 (e.g. in mAU). From these values the correct concentrations can be obtained by using the matrix inverse. However for slightly erroneous absorbances, like 121.4, 122.2 and 202.65 the analyte concentration would be calculated as c_A=0.205 instead of c_A=1.00, i.e., with -79.5% error. (Note that this is a worst case example, but shows how enormous the effect can be. Using Eq. (7) one can see that 0.5 unit (mAU) standard deviation of the absorbance causes ca. 46% standard deviation in the analyte

concentration.) However if the analyst happens to know that the concentration of compound C is at most 0.01 in all samples while the analyte concentration, c_A , is at least 1.00, she may disregard the presence of C in the samples. In other words she considers the samples to contain only A and B and therefore she uses only the first two lines in Table 2. If she measures now again the same sample as above and obtains the same error loaded values as above, i.e., 121.4, 122.2 and 202.65, she will obtain for the analyte concentration (by using the pseudoinverse of the first two lines) c_A=1.01, meaning only a 1% error. This error is partly the bias due to disregarding the presence of C, while the random error inflation is negligible in agreement with the singular values of the first two lines, 186.1 and 81.0(ratio 2.3) and the corresponding sin α being 0.73.

In this example it was assumed that the concentration of C was less than 0.01 in all samples. If this is not the case it may be possible to reduce selectively the concentration of C to this limit by some kind of sample pretreatment.

It is also easy to see that the previous problem (observed with c_C=0.01) would have existed and the same solution would have worked if the concentration of C would have been, e.g., 100 times higher but the absorbance of C at unit concentration 100 times lower.

In the above example it would have been possible to drop the measurements at one wavelength when C was no more included in the model. The remaining two concentrations, c_A and c_B, can be computed from measurements on merely two channels. Taking for example only the two first λ-s and using the error loaded measurement values 121.4 and 122.2, the result for c_A would be 1.002. This reduction of the number of channels is usually not important in spectroscopy. With sensors, however, this can be a significant advantage. It is possible that one has two sensors, both being sensitive to all three sample components A, B and C, respectively, but a third sensor would be difficult to get. In such a case the concentration of the analyte A cannot be generally determined. But again if the concentrations are limited, for example just as in the example above, then two sensors may be enough as the numerical example has just shown. This reduction of the number of channels may appear similar to variable selection in inverse methods (which see later). Yet the possibility of reducing the number of channels(i.e., the number of measured variables) arose here out of selecting a compound to be deleted from the model. Such compound selection cannot be done in inverse methods. Put it in another way: one would not gain anything in this example from dropping the third channel (i.e., measurements on the third wavelength) without deleting compound C from the model.

These examples show that in multichannel measurements with high or even infinite error inflation it can be useful to resort to the fundamental idea of single channel selectivity: one should leave out from the linear model a compound (or some compounds) which have low effect on the measured signals but contribute to the high error inflation factor. One should remind again that the error inflation factor is independent from the interferent concentrations, it is merely due to including the interferents in the model. The proposed procedure can only be used if there is indeed at least one interferent which has low effect on the measured signals but contributes to the high error inflation factor.

The effect of an interferent on the estimate of the analyte concentration can be easily estimated. To this end one need to calculate the apparent analyte concentration caused by unit concentration of the interferent, using the model without the interferent. For example in the case of Table 2 the signal caused by unit interferent concentration is just the third line of the table. The apparent analyte

concentration can be calculated by multiplying this vector with the Moore-Penrose inverse of the first two lines. The result is 0.997 (and the apparent concentration of B is also obtained as 0.997). This value has the same meaning as the reciprocal selectivity s_C/s_A from the single channel equation Eq. (1), because from Eq. (1) at c_C=1 the signal is $s_C c_C$ and the apparent analyte concentration is $s_C c_C/s_A$. In the example the bias due to C is 0.997c_C. For instance if c_C is 0.01, then the bias is 0.997x0.01= 0.00997≈0.01, close to the value obtained above in the numerical example below Table 2.

The example shows that the single channel selectivity concept can be extended to multichannel measurements in a natural way. Selectivity can also be calculated in an analogous way to single channel selectivity and with the same meaning: the reciprocal of the apparent concentration change (bias) due to unit concentration (change) of the interferent. This extension is different from the NAS based multichannel selectivity.

5. Generalization: Selectivity in classical and inverse multivariate linear methods

5.1. The analytical task and some potential difficulties

In the rest of this paper the results of the previous sections are generalized and compared with other methods of multivariate analysis.

As seen in the previous sections, multivariate linear methods allow the determination of the composition of multicomponent mixtures. Their goal may be the determination of all, spectrally active components' concentration in the sample, or only the determination of a single analyte. In the latter case the other components are considered interferents. This paper deals only with this case, i.e. with the determination of a single analyte in the presence of interferents. The terminology of absorption spectroscopy is often used here for all linear multichannel measurements to avoid complicated wording. Second and higher order methods are not considered.

For each channel one can write a linear equation to express the measured signal. For the j-th channel:

$$s_j = c_A s_{AJ} + c_B s_{BJ} + \ldots + \varepsilon_J \qquad (9)$$

and the spectrum of a mixture may be written in a vector equation as:

$$\mathbf{s} = c_A \mathbf{s}_A + c_B \mathbf{s}_B + \ldots + \boldsymbol{\varepsilon} \qquad (10)$$

where ε_i and ε, respectively, denote the random measurement noise and other notations are as before. From a *mathematical point of view*, at least *k* channels are needed for the measurement of *k* compounds. If more than *k* channels are available then the data can be redundant and regression methods are used. But even *k* or more than *k* channels are not always sufficient to determine all concentrations. The equation system must also be solvable for the concentrations to be determined from it. If the analyte is the compound A, then common sense dictates that the equation system must be solvable for c_A.

But in *analytical chemistry* all this is not always true. This is because one need not determine c_A with 100% accuracy. A certain level of bias is tolerable. This is clearly seen in the single channel case discussed earlier in this paper. There the presence of interferent B was acceptable, although a single sensor and consequently *a single equation was available for two concentrations*. It was only important that the signal due to B should be a small fraction of the signal due to A. (Note that the discussion of single channel measurements rarely includes the case when c_A and c_B are correlated. If this correlation is sufficiently strong, the calibration should be made with calibrating solutions reflecting the same correlation. This would then eliminate the bias.)

In *multichannel* measurements a situation similar to the single channel case is found if *the number of channels is less than k, the number of components*. This does not necessarily exclude the possibility of measuring c_A, but some bias must be accepted. This case was exemplified in this paper by the numerical example where two sensors were sufficient to measure c_A in the presence of two interferents, B and C.

Even if *the number of available channels is equal to k, the number of components, or even higher than k*, there may exist two further types of problem. One is when the spectrum vector of the analyte is the linear combination of some interferent spectra. Formally this excludes the possibility of determining c_A. But again, since perfect accuracy is not the goal, c_A may be estimated unless the bias caused by the interferent(s) becomes too high. A simple example for this is when there is only one interferent, B, and s_A and s_B are parallel vectors. Mathematically, c_A cannot be determined in this case. But it is obvious that in this case one has the same situation as when a single channel is used. One has to replace only s_A by $||s_A||$ and s_B by $||s_B||$.

The other possible problem is that the analyte spectrum is *approximately* but not exactly equal to the linear combination of some interferent spectra. In this case one can resolve the equation system accurately for c_A, but when repeated measurements are made on the same sample, the reproducibility of the c_A values is very bad, even if the measurement noise is quite normal (i.e., it would be tolerable in a single channel measurement or a single component multichannel measurement).

5.2. Classical and inverse methods for solving the analytical task

Independently from all these problems, the analytical task of measuring c_A may take (at least) two different forms. In one of these, pure substance spectra can be and are determined for all components of the mixture and the analyst knows from experience, or from other sources, those concentration combinations of analyte and interferents which she may expect to encounter in future samples. In the other typical situation pure component spectra are not available and the expectable concentration combinations are not known either, but one has access to a number of samples which are representative of all future samples, and one can also measure by an independent and accurate analytical method the concentration of the analyte, c_A, in all these representative samples. Analytical methods devised for the first of these two cases are called *classical methods* while those devised for the second are called *inverse methods*. Both types of methods may be plagued by error inflation and by bias in the estimated c_A. Mixed cases also exist. i.e., when the inverse method is used but some or all pure component spectra are also available.

5.3. The possibility of bias variance tradeoff

Both in the classical and the inverse methods c_A is estimated in an unknown sample as a scalar function of the sample spectrum in the following form:

$$c_A = \mathbf{sb} \tag{11}$$

Here s is the spectrum vector of the sample and b is a vector chosen in such a way that within the range of all expected future samples c_A may be determined with sufficiently low error. It is very important to specify the kind of error meant here. One possibility is to require zero bias, and in this case the error is the standard deviation or its square, the variance of c_A. Another possibility is to allow some bias and consider the so called mean squared error (MSE), which is the square root of the sum of the variance (arising due to instrumental noise) and the squared bias (arising due to interferent effects, since other biasing effects like imprecise pure component spectra in the classical method are not considered here):

$$MSE = \sqrt{Var + bias^2} \tag{12}$$

In the classical method one may try to decompose any sample spectrum vector, s, into the pure component spectrum vectors. This may not always be possible (see above: when s_A is the linear combination of some interferent spectra) but if it is, then the resulting coefficient of the analyte spectrum s_A will be an unbiased estimate of c_A. It is shown in statistics texts[21]that this estimate has the lowest possible variance among all possible unbiased estimates of c_A. However, it may not have the lowest possible MSE.

The analyst is usually more interested in lowering the total error, MSE, than in obtaining by all means an unbiased estimate. So the question is how to find a suitable b vector. From Fig.2 it is clear that by choosing b to be the NAS vector one executes a simple vector decomposition and, as expected, the bias of c_A is zero. As long as α is not small, this is quite satisfactory, because the instrumental noise is only slightly inflated. But if α is small, the noise is highly inflated and it may exceed the tolerance limits of the analysis, so it can be worth looking for a solution with some bias but with lower MSE.

The search for lower MSE occurs by varying the b vector. One must be aware, however, that it may not be possible to find any b vector with much lower MSE than in the unbiased case. On the other hand, if one can lower the MSE, one need not find the b with the *lowest* MSE. One needs only to bring the MSE within the predefined tolerance limits of the analysis. Therefore many different methods may exist for finding a suitable b vector and the obtained b vectors can be different.

The search for a good b, i.e., the bias variance tradeoff, is not a trivial task because the MSE should be low enough for essentially *all* future samples. Thus one needs to know and use in this search the expected concentration combinations in all future samples. In the inverse method this is achieved by using for the calibration a set of (hopefully) representative samples with measured c_A concentrations. In the classical method the analyst's educated guess about future sample compositions should be used. Both methods may introduce hidden errors. For finding a good b one needs to know or model the instrumental noise, too. This may be iid (independent and identically distributed) as in previous sections of this paper but may also have other properties. Zero bias of the instrumental noise is

assumed here to avoid confusing it with bias due to interferents. Heteroscedasticity and nonzero error covariances can usually be handled by standard mathematical methods.

5.4. Bias and selectivity

Whatever method is used to arrive at some satisfactory b vector, the bias of the method caused by an interferent, say B, in any particular sample can be expressed as:

$$Bias_{AB} = c_B \mathbf{s}_B \mathbf{b} \tag{13}$$

simply because

$$c_A = \mathbf{sb} = (c_A \mathbf{s}_A + c_B \mathbf{s}_B + \ldots + \boldsymbol{\varepsilon})\mathbf{b} = c_A \mathbf{s}_A \mathbf{b} + c_B \mathbf{s}_B \mathbf{b} + \ldots + \boldsymbol{\varepsilon}\mathbf{b} \tag{14}$$

and because it makes sense (although it is not absolutely necessary) to choose b such that in the absence of any interferent and noise the estimate of c_A should equal the true value of c_A:

$$\hat{c}_A = c_A \mathbf{s}_A \mathbf{b} \tag{15}$$

Therefore

$$\mathbf{s}_A \mathbf{b} = 1 \tag{16}$$

and hence

$$\hat{c}_A = c_A + c_B \mathbf{s}_B \mathbf{b} + \ldots + \boldsymbol{\varepsilon}\mathbf{b} \tag{17}$$

showing that the bias due to B is indeed $c_B s_B b$.

On the other hand the equality $s_A b=1$means that

$$\|\mathbf{b}\| = 1 / \|\mathbf{s}_A\| \cos\gamma \tag{18}$$

where γ is the angle between s_A and b. Since a unit concentration change in c_B changes the estimated c_A by $s_B b$, the selectivity is

$$Sel_{AB} = 1 / \mathbf{s}_B \mathbf{b} = \|\mathbf{s_A}\| \cos\gamma \|\mathbf{s_B}\| \cos\beta \tag{19}$$

where β is the angle between s_B and b. This selectivity value is the reciprocal of the analyte concentration bias caused by a unit change in interferent concentration, just like the single channel selectivity. Note that in single channel measurements $b=s_A/||s_A||$is the only reasonable choice (to avoid unnecessary error inflation), so that γ=0 and β=0. Thus the present formula returns the conventional value for selectivity, s_A/s_B. On the other hand if in multichannel measurements b is selected to point into the NAS vector direction then γ=π/2-α and β=π/2. The selectivity is therefore infinitely large, in agreement with the fact that the bias from the interferents is completely eliminated. Therefore the selectivity defined here is the logical generalization of the single channel selectivity to multivariate methods and might well replace the NAS based selectivity, sinα.

The selectivity depends on the expected sample compositions' distribution because the direction of b depends on these. Thus even though one can define and calculate the selectivity of the method after having fixed b, this selectivity value can only be used within the range of the expected or estimated future sample compositions. This is, of course, not a very bad limitation.

Although the selectivity formula derived above applies also to inverse methods, its application is hindered when the individual component spectra are not available. This can be sometimes helped with by carrying out a few extra measurements. Arnold et al [22] presented an experimental method to test the selectivity of a PLS inverse method. They had built the PLS model for a compound (denoted here for simplicity as A) with calibration mixtures and then submitted to this model the spectra of each individual compound (the analyte A and two interferents, B and C) in the same matrix, each compound at several concentrations. They established the *apparent* A concentrations for every spectrum and plotted calibration lines for A, B and C. They considered as selectivity the ratio of the slope of the analyte calibration line to the slope of an interferent's calibration line. This definition and procedure is in agreement with the here proposed selectivity definition. Although Arnold et al did not derive the equations presented here, their work shows that practical analytical chemists would intuitively understand selectivity in the same way as presented here.

In conclusion one can extend the single channel selectivity concept to multichannel (multivariate) analyses, both in classical and inverse methods. The obtained selectivity measure is different from the NAS selectivity, which is sin α. Although the NAS selectivity is advantageously independent of the future sample compositions, it shows only that there is a problem with error inflation, and does not quantitate any particular interferent's effect on the measurement bias. Moreover if in the classical method the b vector is not parallel to the NAS vector, then 1/sinα is no more the measure of noise inflation, either. As one may see from the above derivation (using Eqs. (17) and (18)) the noise inflation factor is then 1/cosγ.

6. A simple way for bias variance tradeoff in the classical method

In this paper numerical examples have been shown for the classical linear multivariate analytical method. In these examples the expected composition of future samples was approximately specified like c_A=1; c_B=1; c_C=0.01 (where $c_C \leq 0.01$ might have been written because the bias is less if the interferent concentration is less). The following algorithm is a generalization of the examples.

Obtain the spectra of all pure compounds(approximate spectra may suffice for low level interferents).

Determine or estimate the noise of the signal (absorbance) measurements (may be wavelength/channel dependent). Proceed only if this noise level would be acceptable for measuring the pure analyte in its expected concentration range. (Otherwise the method will never be satisfactory because this noise can only be inflated, not reduced.)

Calculate the error inflation factor of the classical (vector decomposition) method. If the inflated error (noise) is still acceptable, use the classical method. This will eliminate all bias, independently from the interferent concentrations.

If the inflated error is unacceptable, estimate the maximum expected concentrations of the interferents and the expected range of analyte concentrations. Check if deleting the spectrum vectors of those interferents for which $c_J s_J$ (where the index J denotes generally an interferent)is small compared to $c_A s_A$, reduces significantly the noise inflation factor (as calculated from the NAS vector relative to the remaining interferents pectra). If the reduced noise inflation is acceptable, calculate the maximum bias caused by the interferents whose spectra were deleted. If this (and the resulting MSE) is also acceptable the method development is finished. As a result, some of the interferences will be completely eliminated while those for which the spectrum was deleted will remain, but cause only a tolerable bias. The procedure described in this paragraph is the one suggested by the numerical examples in this work. Simulations (not shown here) indicate that this method works well in a wide parameter range.

If the resulting bias is too high even after the procedure of the preceding paragraph, one may go one step back and reintegrate the spectrum of one interferent. If this reduces the bias substantially, without inflating the measurement error too much, one may continue this procedure until arriving at the limit of acceptability. If this method does not help, one has to consider projection of the pure component spectra on a vector intermediate between the last NAS direction and s_A. Selection may be done by some optimization algorithm (to obtain acceptable MSE) or by simulating sample spectra in the expected concentration range and employing inverse regression to them. This is a somewhat underdefined task, just like all inverse regression methods. The procedure proposed in the previous paragraph helps to avoid the need for such methods.

Note, however, that it may not be possible to arrive at an acceptable analytical method by any classical or inverse method with the measuring system at hand and the prevailing concentration distributions. In this case one has to improve the measuring system, e.g., by reducing instrumental noise, by decreasing the concentration of some interferents by sample pretreatment or by using other sensors.

The method proposed in this paper is slightly suboptimal because it uses always the projection of the sample and pure component vectors onto a NAS vector (that defined by the analyte and the not discarded interferent spectra), without considering other possible b vectors. On the other hand, however, it is transparent and relatively simple, particularly if the number of components to be considered is small. Notably it does not require the development of a special search algorithm for finding the b vector. Besides it can also handle the case of multicollinearity, just like the traditional single channel method does and unlike the vector decomposition (ordinary least squares, OLS) method.

To the authors' knowledge there have been no other methods suggested yet to apply the classical method to the specific problem of treating low-level bias in the high error inflation situation. This is surprising because in single channel measurements this is perhaps the most prevalent problem and it has been solved intuitively for such a long time, that the solution is considered a commonplace in

analytical chemistry. Closest to the present work are papers by Brown and coworkers [23-24]who had very elegantly presented the statistical foundations of both the classical and inverse methods. They did not derive, however, the simple formula for selectivity shown in the present work and they also did not discuss the particularly relevant case of low level bias under high error inflation.

Most other papers on multivariate linear methods use the NAS based selectivity definition. Some of these specifically mention the relationship with single channel selectivity. Faber et al.[25] expressed the view that it is actually the old single channel definition of selectivity which should be abandoned and replaced by a new one derived from the NAS selectivity concept. The single channel selectivity definition proposed in that paper is based on applying the inverse calibration method also in the single channel case. This means calibration with representative samples of known and varying analyte concentration, c_A and unknown and varying interferent concentration, c_B values. The linear calibration plot (signal vs. c_A) has a positive intercept, due essentially to the average level of bias in the representative samples. The measured signals in each sample are corrected with the intercept. Selectivity is defined as the ratio of the corrected signal to the uncorrected signal. Due to this shift the average bias disappears but the signal (the instrumental noise was assumed to be zero) becomes noisy. In that paper this is called the bias variance tradeoff. The latter definition goes against the meaning of bias variance tradeoff in statistics [21], which is a tradeoff of bias caused by interferents against the (inflated) instrumental noise. It is also unclear how this single channel selectivity could be used in practice.

In contrast to that paper, two more recent IUPAC documents [5, 13]retain the old selectivity definition of single channel measurements, i.e., s_A/s_B. For multivariate methods the second IUPAC document [13] states, that since interferences can be adequately modeled using multivariate data, the numerical assessment of multivariate selectivity has always been approached differently from the single channel case. The authors consider the NAS selectivity definition of Lorber [7]and Bergman et al. [8] the most suitable one. This is based in this IUPAC document on the statement that a prediction sample's spectrum can always be decomposed "in two orthogonal parts: a part that can be uniquely assigned to the analyte of interest (the NAS), and the remaining part that contains the (possibly varying) contribution from other components". This statement is, however, only true for the simple vector decomposition (ordinary least squares). For biased regression methods, i.e., the majority of practically used methods like PCR, RR, PLS, and even for the classical method shown in this paper this is not true. This IUPAC document shortly mentions alternative works of Brown and Arnold [22-24],but only because the latter authors define pairwise selectivity relative to individual interferents, unlike the NAS selectivity. The paper concludes on these alternatives: "The radically different standpoint taken in that work may lead to a critical reexamination of multivariate selectivity assessment." This reexamination may not have taken place yet since in a very recent paper by Olivieri [16] the statements of the IUPAC paper[13] are essentially repeated. It should be noted that in the Arnold paper the NAS selectivity is called spectral selectivity. This might be a reasonable differentiating name for the NAS selectivity.

7. The meaning and the limitations of the NAS based selectivity

As noted earlier, an interesting property of the single channel selectivity is that it relates to an equation which is underdefined for the determination of the analyte concentration, since a single

equation includes two unknowns: the concentrations of the analyte and the interferent, respectively. In contrast to this, nearly all papers on multivariate selectivity [e.g., [13-16]) refer to equation systems which can be solved (have a unique solution) for the analyte concentration. This means also that they consider only one version of the classical method (ordinary least squares) where all bias is completely compensated. However, it should be impossible to define selectivity for this method, at least if one wishes to use selectivity to estimate measurement bias due to an interferent. On the other hand in inverse methods and in some versions of the classical method (which see above)there is bias variance tradeoff and consequently there is nearly always some bias in the analyte concentration estimate. Therefore in these multivariate methods selectivity can be defined (as has been done in this manuscript) in accordance with the single channel selectivity.

The background of the contradiction between the single channel and multichannel (NAS based) selectivity definitions can be clearly understood by reference to Faber et al.[14]. In that paper the effect of instrumental noise on analyte concentration estimates is discussed for the classical, least squares multivariate method. It is shown in the paper that the NAS selectivity "entirely accounts for the effect of interferences". This remark makes it obvious that the NAS based multichannel selectivity is used to express the noise inflation which arises due to the complete elimination of bias by interferents. In this sense the NAS based selectivity is indeed related to interferents, albeit only indirectly. But in the other multivariate methods, which use bias variance tradeoff, bias is not completely eliminated. The arising optimal b vector has generally a direction different from the NAS vector and the error inflation factor cannot be given any more by $1/\sin \alpha$. This means that the NAS selectivity is not applicable even for the limited task, for which it had been designed, i.e., to quantitate the error inflation due to bias reduction in multivariate linear methods.

Ultimately the NAS based selectivity concept is a consequence of narrowing the range of linear multivariate analysis methods to the classical method with unbiased estimation.

For an easy comparison Table 3 shows the relationship between the properties of different multivariate linear methods.

Table 3.Properties of multivariate linear methods

	Classical methods		Inverse methods
	OLS	non OLS	
All pure component spectra needed	Yes	Yes	No
MSE vs. Var	MSE=Var	MSE≤Var	MSE≤Var
Bias	0	≥0	≥0
b can be only parallel with the NAS vector	Yes	No	No
Selectivity value defined by bias	∞	≤∞	≤∞
Error inflation factor	$1/\sin \alpha$	$1/\sin \gamma$	$1/\sin \gamma$

8. Conclusion

Selectivity is, in its conventional single channel meaning, a property of such analytical measurements (be they linear or nonlinear) where the estimated analyte concentration is biased by the presence of some interfering compound(s). The corresponding measure of selectivity is the reciprocal of the number which shows the change of the estimated analyte concentration due to a unit (in nonlinear cases a small, essentially differential) change in the interferent's concentration.

There is no guarantee that the selectivity is independent from the actual concentrations of the analyte, the interferent and other components in the sample. Thus the selectivity is not necessarily a property of the analytical method, and is not always quantifiable independently from the sample compositions. If, however, the selectivity is a constant in a certain concentration range (which may be the full possible range of *expected* sample concentrations), then it is a useful quantity to predict the bias caused by different values of the interferent concentration.

In some multivariate linear methods of chemical analysis interferences can be completely eliminated, so that no bias remains. For such methods a new definition of selectivity was adopted some time ago, based on the net analyte signal (NAS) concept. This selectivity was intended to measure (through its reciprocal, the error inflation factor, eif) the inflation of instrumental measurement noise. This noise inflation arises as the consequence of the mathematical operations which eliminate the bias. Thus the noise inflation is indirectly a consequence of interference. This is the reason why the reciprocal of eif was named selectivity.

The multilinear method for which the NAS based selectivity was defined is a so-called classical method. In contrast to the so-called inverse methods, in the classical methods one needs to know the pure component spectra of all spectrally active components of the analytical samples. This information allows eliminating of all interference related bias from the estimated analyte concentration. The mathematical method to do this is vector decomposition or in the statistical sense the method of ordinary least squares (OLS). The NAS selectivity was developed for this particular method. The OLS method has, however, the disadvantage that if the pure component spectra are nearly multicollinear (there is strong "spectral overlap") then the noise inflation is very large. This problem can be eased by alternative computational methods, which reduce noise inflation at the cost of some bias, so that the total error, called the mean squared error (MSE) is lowered. In contrast to an apparently widely held belief, such bias variance tradeoff is not the unique property of inverse methods, but can also be employed in classical methods, i.e., using the pure component spectra. However, it is not easily realized, because bias variance tradeoff requires information about the future analytical samples' composition.In this work it has been recognized that in the typical selectivity problem, i.e., when the interferent effects on the measured signal are relatively small, the bias variance tradeoff is possible in a very general form.

In this paper, beyond clarifying the relationships explained above, general formulas have been derived for calculating the bias, the bias related selectivity and the variance inflation factor for all multilinear methods, be they classical or inverse, OLS or non OLS. To do this, the conventional single channel selectivity concept was used, i.e., that the selectivity is the reciprocal of the bias caused by unit concentration of an interferent. It was shown that the error inflation factor of non OLS methods, be they classical or inverse, is generally different from the eif calculated from the NAS. Further on a simple method has been proposed for carrying out bias variance tradeoff in the framework of the classical method. The proposed method consists of deleting some minor interferences from the

model. This is an extension of the conventional treatment of interferences used in single channel methods.

Some interesting related topics, where this discussion might be extended later are: variable selection, extension of the calibration and higher order methods.

Acknowledgments

The financial support of OTKA, Hungary (Grant No. K104724) and the Ministry of Education, Science, and Technological Development of Serbia (Grant No. 172008) and discussions with Dr P. Horvai are gratefully acknowledged.

References

[1] M. Valcarcel,A. Gomez-Hens,S. Rubio, Selectivity in analytical chemistry revisited, Trends Anal. Chem.20 (2001) 386-393.

[2] T. Verbic,Z. Dorko,G. Horvai, Selectivity in analytical chemistry, Rev. Roum. Chim.58 (2013) 569-575.

[3] Z. Dorko,T. Verbic,G. Horvai, Selectivity in analytical chemistry: two interpretations for univariate methods, Talanta132 (2015) 680-684.

[4] D. Harvey, Modern Analytical Chemistry, The McGraw-Hill Companies, Inc., USA, 2000. pp. 40.

[5] M. Thompson,S. L. R. Ellison,R. Wood, Harmonized guidelines for single-laboratory validation of methods of analysis (IUPAC technical report), Pure Appl. Chem.74 (2002) 835-855.

[6] H. Kaiser, Guiding concepts relating to trace analysis, Pure Appl. Chem.34 (1973) 35-61.

[7] A. Lorber, Error propagation and figures of merit for quantification by solving matrix equations, Anal. Chem.58 (1986) 1167-1172.

[8] G. Bergmann,B. Von Oepen,P. Zinn, Improvement in the definitions of sensitivity and selectivity, Anal. Chem.59 (1987) 2522-2526.

[9] Z. Dorko,T. Verbic,G. Horvai, Selectivity of nonlinear analytical methods, in preparation.

[10] JCGM (Joint Committee for Guides in Metrology), International Vocabulary of Metrology-Basic and General Concepts and Associated terms (VIM), third ed., 2012. pp. 41.

[11] H. Kaiser, Zur Definition von Selektivität, Spezifität und Empfindlichkeit von Analysenverfahren, Z Anal. Chem. 260 (1972) 252-260.

[12] H. Kaiser, Foundations for the critical discussion of analytical methods, Spectrochim. Acta, Part B33 (1978) 551-576.

[13] A. C. Olivieri,N. K. M. Faber,J. Ferre,R. Boque,J. H. Kalivas,H. Mark, Uncertainty estimation and figures of merit for multivariate calibration, Pure Appl. Chem.78 (2006) 633-661.

[14] N. M. Faber,J. Ferre,R. Boque,J. H. Kalivas, Quantifying selectivity in spectrophotometric multicomponent analysis, Trends Anal. Chem.22 (2003) 352-361.

[15] G. Bauer,W. Wegscheider,H. M. Ortner, Selectivity and error-estimates in multivariate calibration - application to sequential ICP-OES, Spectrochim. Acta, Part B46 (1991) 1185-1196.

[16] A. C. Olivieri, Analytical figures of merit: from univariate to multiway calibration, Chem. Rev.114 (2014) 5358-5378.

[17] J. H. Kalivas,P. M. Lang, Interrelationships between sensitivity and selectivity measures for spectroscopic analysis, Chemom. Intell. Lab. Syst.32 (1996) 135-149.

[18] B. Nadler,R. R. Coifman, Partial least squares, Beer's law and the net analyte signal: statistical modeling and analysis, J. Chemom.19 (2005) 45-54.

[19] G. Bauer,W. Wegscheider,H. M. Ortner, Selectivity and limits of detection in inductively coupled plasma optical-emission spectrometry using multivariate calibration, Spectrochim. Acta, Part B47 (1992) 179-188.

[20] J. Vessman,R. I. Stefan,J. F. Van Staden,K. Danzer,W. Lindner,D. T. Burns,A. Fajgelj,H. Muller, Selectivity in analytical chemistry (IUPAC Recommendations 2001), Pure Appl. Chem.73 (2001) 1381-1386.

[21] D. C. Montgomery, E. A. Peck, G. G. Vining, Introduction to Linear Regression Analysis, John Wiley & Sons, fifth ed., 2012.

[22] M. A. Arnold,G. W. Small,D. Xiang,J. Qui,D. W. Murhammer, Pure component selectivity analysis of multivariate calibration models from near-infrared spectra, Anal. Chem.76 (2004) 2583-2590.

[23] C. D. Brown,T. D. Ridder, Framework multivariate selectivity analysis, part I: Theoretical and practical merits, Appl. Spectrosc.59 (2005) 787-803.

[24] T. D. Ridder,C. D. Brown,B. J. V. Steeg, Framework for multivariate selectivity analysis, part II: Experimental applications, Appl. Spectrosc.59 (2005) 804-815.

[25] K. Faber,A. Lorber,B. R. Kowalski, Analytical figures of merit for tensorial calibration, J. Chemom.11 (1997) 419-461.

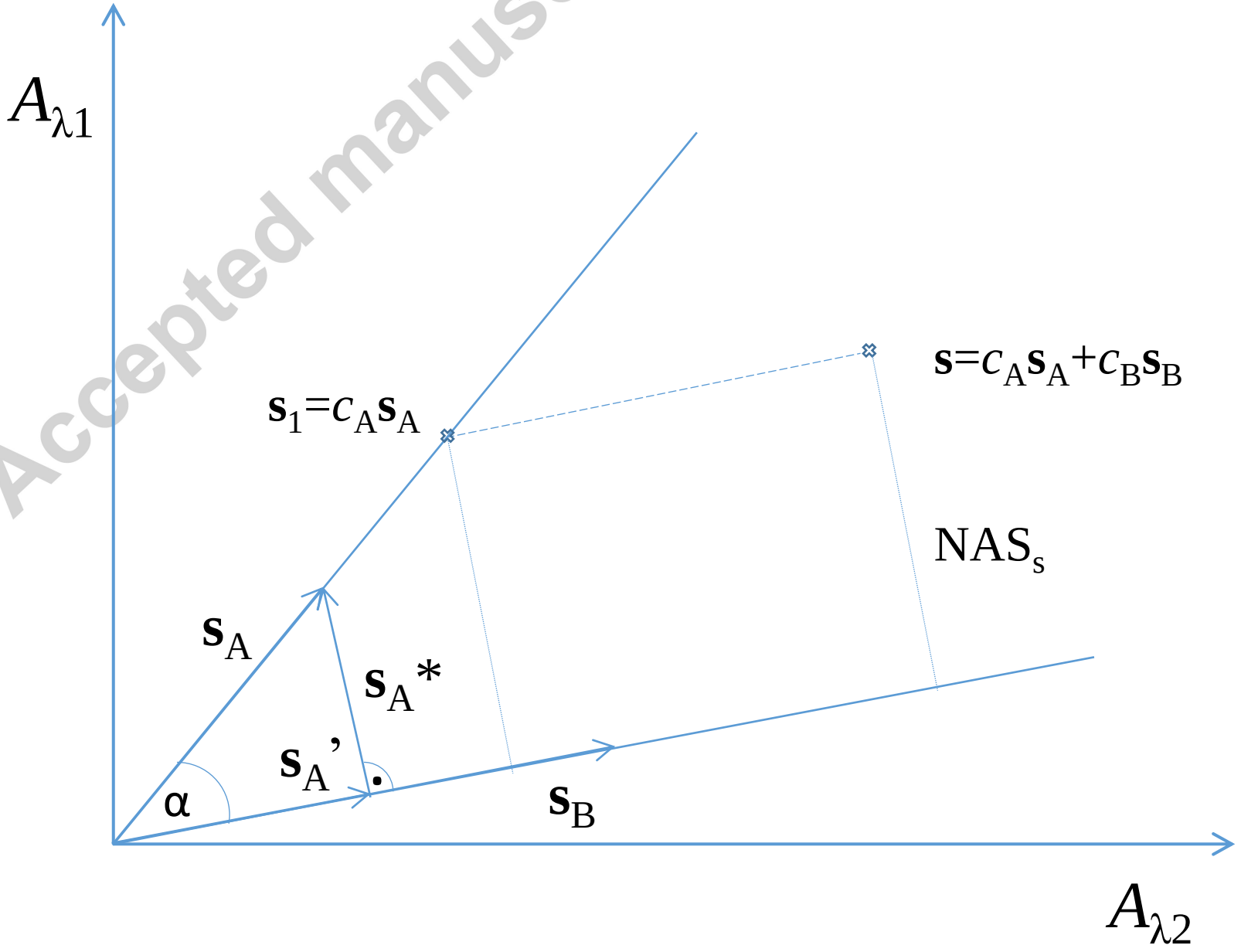
Fig.1: A simple example for a multichannel linear analytical method: absorbances of two components measured at two wavelength values.

A is the analyte, B is the interferent. The axes show the measured absorbances $A_{\lambda 2}$ and $A_{\lambda 1}$, respectively; **s** is the spectrum of a solution containing both A and B. In the sample $\mathbf{s}_1$ there is only A but no B; $\mathbf{s}_A$ and $\mathbf{s}_B$ are the spectra of A and B, respectively, at unit concentrations; $\mathbf{s}_A^*$ is the net analyte signal vector at unit concentration. NAS_s is the length of the net analytical signal vector of sample **s**.

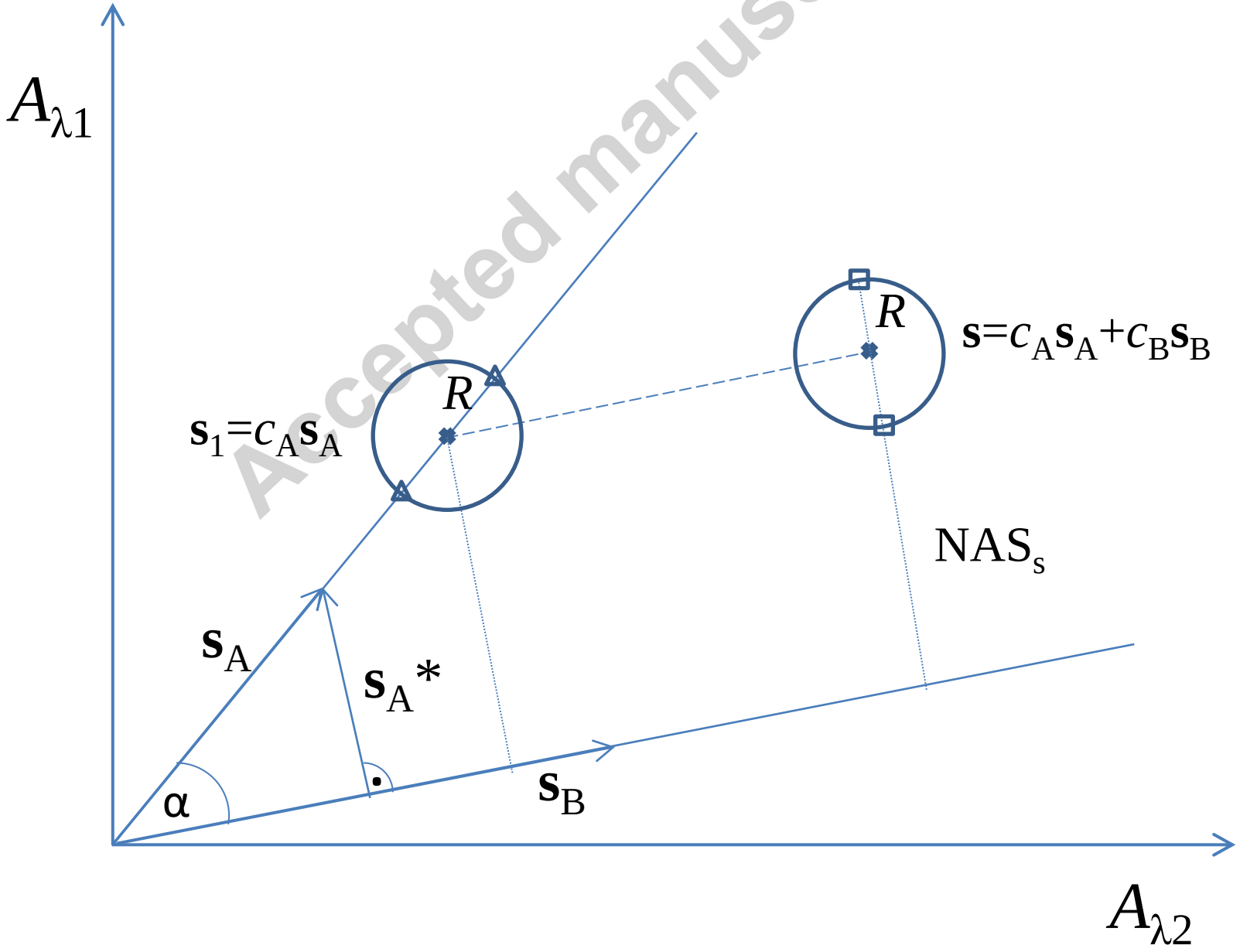
Fig.2: Determination of the error bars in a mixed solution of compounds A and B and in a solution of A alone.

R represents the radius of a constant probability density circle of the assumed normal distribution and corresponds to one standard deviation. Other notations are as in Fig.1.

Figure



Figure



*Graphical Abstract (for review)