A new IUPAC-consistent approach to the limit of detection in partial least-squares calibration

Journal:	*Analytical Chemistry*
Manuscript ID:	ac-2014-01786u.R1
Manuscript Type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Allegrini, Franco; University of Rosario, Analytical Chemistry Olivieri, Alejandro; University of Rosario, Analytical Chemistry

A new IUPAC-consistent approach to the limit of detection in partial least-squares calibration

Franco Allegrini and Alejandro C. Olivieri*

Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Instituto de Química de Rosario (IQUIR-CONICET), Suipacha 531, Rosario S2002LRK, Argentina

ABSTRACT: There is currently no well-defined procedure for providing the limit of detection (LOD) in multivariate calibration. Defining an estimator for the LOD in this scenario has shown to be more complex than intuitively extending the traditional univariate definition. For these reasons, although many attempts have been made to arrive at a reasonable convention, additional effort is required to achieve full agreement between the univariate and multivariate LOD definitions. In this work, a novel approach is presented to estimate the LOD in partial least-squares (PLS) calibration. Instead of a single LOD value, an interval of LODs is provided, which depends on the variation of the background composition in the calibration space. This is in contrast with previously proposed univariate extensions of the LOD concept. With the present definition, the LOD interval becomes a parameter characterizing the overall PLS calibration model, and not each test sample in particular, as has been proposed in the past. The new approach takes into account IUPAC official recommendations, and also the latest developments in error-in-variables theory for PLS calibration. Both simulated and real analytical systems have been studied for illustrating the properties of the new LOD concept.

Analytical chemistry is the science of chemical measurements, and thus it is of fundamental importance to develop appropriate estimators for the figures of merit which are conventionally used to evaluate the quality of the measurements.[1-3] Among these figures of merit, one of the most controversial ones has been the limit of detection (LOD).[4-7] Its importance lies in the fact that it is a good measure of the quality of a calibration model, because its definition brings together two important analytical concepts: the sensitivity and the precision in the analytical determinations.

Currently, the International Union of Pure and Applied Chemistry (IUPAC) adopts the definition given by the International Standardization Organization (document ISO 11843)[8] for the capability (or limit) of detection as 'the lowest quantity of a substance that can be distinguished from the absence of that substance (a blank value) within a stated confidence limit'.[9-11] This implies that the LOD is the minimum quantity detectable with a pre-set probability of false positives (Type I errors) and false negatives (Type II errors).[9-11]

Regarding LOD estimators, when the analytical signal is univariate and analyte-specific, the recommended detection rule is based on a Neyman-Pearson test that considers false positive and false negative errors for the null hypothesis 'there is no analyte' and the alternative hypothesis 'there is analyte'.[9] The LOD can be directly estimated from the univariate calibration line, as a simple alternative to the original recommendation, in which the LOD is estimated from the average signal level and standard deviations for repeated measurements of a blank sample and for one or more samples at concentrations near the detection limit.[12]

However, when dealing with multivariate calibration, as is the case of partial least-squares (PLS) regression analysis, the application of the above definition is not entirely clear, and some aspects which remain outside the field of application of the ISO norm need to be considered.[13] In fact, there is still no generally accepted LOD estimator for PLS studies. Nevertheless, there is a high interest in the topic,[2] undoubtedly tied to the inclusion of PLS regression in many commercial instruments, particularly those based on near infrared spectral (NIR) measurements,[14] in addition to the continuous emergence of new and more sensitive analytical techniques, and the release of regulations on human or environmental exposure to low levels of chemical health hazards.

The main difficulty in estimating a multivariate limit of detection is that the instrumental signals are not specific for a particular analyte. In response to this, Lorber et al. developed an approach based on the concept of *net analyte signal*.[15] However, the main drawback of this estimator is that it only considers the uncertainty in the signal measurements, making its real application rather limited, because other important sources of uncertainty are the calibration concentrations and signals. Additional strategies, which rely on the standard deviation of the blank based on spectral residuals, suffer from the same drawback.[16]

Rius et al. suggested a multivariate LOD based on the calculation of a response detection which is specific for the analyte of interest, with evaluation of the probabilities of errors of both Types I and II.[7] They presented the interesting idea that an LOD value should be calculated for each test sample, implicitly suggesting the possibility of considering the multivariate LOD as a concentration range rather than as a single concentration value. Nonetheless, the authors exposed the need for further research aimed at the calculation of a non-ambiguous detection response. A similar method, based on a simplified formula for the sample-specific standard error in concentration for PLS regression,[17] has been applied in several literature works.[18,19] However, in all of these approaches, the leverage (a dimensionless parameter measuring the position of the sample in the calibration space) of each sample at zero analyte concentration is only an approximation, and there is no well-established procedure to calculate it.

Finally, Ortiz et al. proposed an LOD estimator which can be directly generated by extending the IUPAC recommendations for univariate methods to multivariate calibration.[13,20] This generalization is based on the mathematical proof that the capability of detection, as defined by ISO and IUPAC for univariate calibration, is invariant for linear transformations of the response. As a consequence, the same capability of detection is obtained using the regression of estimated concentration vs. calibration concentrations. The latter values can be either measured by a reference technique, or nominally assigned when prepared in the laboratory from analyte standards. Although this 'pseudo-univariate' approach sounds valid, it is not in complete agreement with the latest advances in uncertainty propagation in PLS calibration, based on the so-called error-in-variables (EIV) models.[21] In particular, it is not consistent with the idea of a sample-specific LOD value.[2,3]

In this work, a new methodology to estimate the LOD is proposed for PLS multivariate calibration. It is based on several complementary ideas: (1) each test sample has in principle a specifically associated LOD value, (2) the universe of test samples is well-represented by the calibration set of samples, (3) the leverages for the calibration samples can be extrapolated to zero analyte concentration, and (4) a range of LOD values can be easily estimated for the PLS model as a whole. The lower and upper limits of the LOD interval ($\mathrm{LOD}_{\mathrm{min}}$ and $\mathrm{LOD}_{\mathrm{max}}$ respectively) correspond to the calibration samples with the lowest and largest extrapolated leverages to zero analyte concentration. These results allow the mutual relationship between $\mathrm{LOD}_{\mathrm{min}}$, $\mathrm{LOD}_{\mathrm{max}}$ and the pseudo-univariate value ($\mathrm{LOD}_{\mathrm{pu}}$) to be uncovered. Finally, the proposal is tested in several simulated and experimental systems.

THEORY

PLS regression. Partial least-squares has gained popularity in analytical chemistry, as has been extensively described in the literature.[22-24] The PLS model can be interpreted as the result of merging principal component regression (PCR) and multivariate linear regression (MLR). PCR finds factors that capture the greatest amount of variance in the matrix of predictor (**X**) variables (*e.g.,* spectra, matrix size $J \times I$, where J is the number of wavelengths and I the number of samples). MLR seeks to find a single factor that best correlates predictor (**X**) variables with predicted (**y**) variables (*e.g.,* concentrations, of size $I \times 1$). In PLS, on the other hand, the information contained in both **X** and **y** is actively used for the definition of the latent variable space, in such a way that latent factors both capture variance and achieve correlation, maximizing the covariance between the predictor and the variable to be predicted.

The PLS calibration stage requires, as a first step, the estimation of the optimum number of latent variables A, which is usually done by a technique known as leave-one-out cross validation.[25] The main result of the calibration is the vector of latent regression coefficients **v** (size $A \times 1$), and two matrices of loading vectors **P** and **W** (both of size $J \times A$). In the subsequent prediction phase, these parameters are employed to estimate the analyte concentration in a test sample ($\hat{y}$, with the 'hat' over the symbol meaning that the parameter is estimated) from its spectrum **x**:

$$\hat{\mathbf{t}} = (\mathbf{W}^{\mathrm{T}}\mathbf{P})^{-1}\,\mathbf{W}^{\mathrm{T}}\,\mathbf{x} \qquad (1)$$

$$\hat{y} = \mathbf{v}^{\mathrm{T}}\,\mathbf{t} + \bar{y}_{\mathrm{cal}} \qquad (2)$$

where $\hat{\mathbf{t}}$ is vector of the so-called scores for the test sample (size $A \times 1$), the superscript 'T' indicates transposition, and $\bar{y}_{\mathrm{cal}}$ is the mean calibration concentration. The latter term appears in equation (2) for mean-centered data, which is the default option in PLS studies.

Equation (2) is defined in the space of the latent variables, although an analogous expression exists in the real variable space, as:

$$\hat{y} = \mathbf{b}^{\mathrm{T}} \mathbf{x} + \bar{y}_{\mathrm{cal}} \qquad (3)$$

where **b** is the vector of regression coefficients in the real space. In the remainder of this work, the hats will be avoided for clarity.

Multivariate LOD. According to the latest IUPAC recommendations, the estimation of the limit of detection should comply with two conditions: (1) it should be based on the theory of hypothesis testing, taking into account the probabilities of false positives and false negative decision, and (2) it should include all the different sources of error, both in calibration and prediction steps which could affect the final result.

Considering the first condition, the multivariate LOD should be based on the same expression as the one used for univariate calibration:[3]

$$\mathrm{LOD} = (t_{\alpha,\nu} + t_{\beta,\nu})\ \mathrm{var}(y_0)^{1/2} \qquad (4)$$

where $\mathrm{var}(y_0)$ is the concentration variance for a blank sample, and $t_{\alpha,\nu}$ and $t_{\beta,\nu}$ are coefficients for a Student's *t* distribution with ν degrees of freedom. The latter two parameters take into account the probability of making Type I errors (assuming that the analyte is present when it is absent) with a probability α, and Type II errors (assuming that the analyte is absent when it is present) with a probability β. Typically, α and β are assigned a value of 0.05 (i.e., a confidence level of 95%), ν is usually large for a multi-sample calibration set, and therefore in practice the factor $(t_{\alpha,\nu} + t_{\beta,\nu})$ in equation (4) takes the approximate value of 3.3.

It is important to notice that in equation (4) the distance from the blank to the LOD is approximated by the sum of two confidence intervals. A more rigorous approach suggests the use of a noncentrality parameter of a noncentral *t* distribution instead of a sum of *t*-coefficients.[26] However, the values provided by these alternative statistical approaches do not significantly differ.[27] In any case, a thorough analysis of the LOD estimators based on prediction intervals has been performed.[28,29]

A key point in regard to equation (4) is the criterion adopted for estimating the variance of the predicted concentration, which concerns the second of the above conditions. In this sense, the basic assumption throughout this work is that the variance in the predicted analyte concentration by a PLS model is given by the well-known expression:[3,16,19,30-32]

$$\mathrm{var}(y) = \mathrm{SEN}^{-2}\,\mathrm{var}(x) + h\,\mathrm{SEN}^{-2}\,\mathrm{var}(x) + h\,\mathrm{var}(y_{\mathrm{cal}}) \qquad (5)$$

where SEN is the sensitivity [given in PLS by the inverse of the length of the regression coefficients, i.e., by 1/|| **b** ||, where **b** is from equation (3) and || || implies the Euclidean norm of a vector],[21,33] var(*x)* is the variance in instrumental signals, *h* is the sample leverage, and var(y_{cal}) the variance in the calibration concentrations. The three terms in the right-hand side of equation (4) account for the propagation of uncertainties derived from: (1) instrumental signals in the test sample data, (2) instrumental signals in the calibration data, and (3) calibration concentrations. The first and probably the most relevant contribution is transmitted directly via the inverse squared sensitivity. The second and third terms arise from calibration uncertainties and are both scaled by the sample leverage. The latter is proportional to the Mahalanobis distance of a sample from the center of the calibration space (for mean-centered data), and can be expressed as a function of concentrations, instrumental variables, or latent variables. In the latter case, an appropriate expression for *h* is:[21]

$$h = \mathbf{t}^{\mathrm{T}}\,(\mathbf{T}^{\mathrm{T}}\,\mathbf{T})^{-1}\,\mathbf{t} \qquad (6)$$

where **T** is the matrix of scores for the calibration samples, which is obtained by projecting the calibration matrix of signals **X** onto the PLS loadings, analogously to equation (1). Appropriate values of var(x) and var(y_{cal}) are usually available from sample replicate analysis or estimated from other sources.[17]

Notice that when both signals and concentrations are mean-centered prior to PLS modeling, two additional terms are required in the right-hand side of equation (5), having the same form as the current last two terms in this equation, with the leverage h replaced by ($1/I$), where I is the number of calibration samples.[21] One simple way of taking this fact into account is to define a new, 'effective' leverage, as ($h + 1/I$) to be used instead of h in equation (5) and in all equations requiring to estimate var(y) for mean-centered data.

To be able to estimate the LOD, equation (4) requires the value of var(y_0), i.e., the concentration variance for a blank sample [the value of var(y) when $y = 0$], which would in principle be available from equation (5). In this regard, the leverage when the analyte concentration is zero (h_0) plays a fundamental role. Surprisingly, though, to the best of our knowledge there are no consistent proposals for estimating this latter parameter. Approximations to h_0 have been suggested, involving the study of samples which are supposed to be near the detection limit.[18,19]

As an extension of the LOD univariate concept, one tends to intuitively think on a single LOD value for the multivariate case, although a deeper analysis indicates that this is not the case. In univariate calibration a single value of h_0 exists, which can be confidently estimated from the calibration parameters.[1] However, in multivariate calibration h_0 assumes different values depending on the sample composition. According to equations (1) and (6), each test sample with zero analyte concentration, but having different levels of other concomitant components, all contributing to the sample spectrum, will generate a specific set of scores, and thus a specific value of the leverage h_0.[2] Therefore, in the framework of PLS calibration it is more reasonable to consider the existence of an LOD interval, whose values depend on the variability of the background composition, rather than a single LOD value.

DATA SETS

Simulated data. Synthetic data sets were created by mimicking a three-component analytical system, with component 1 being the analyte of interest. Each calibration and test spectrum (**x**) was built using the following expression:

$$\mathbf{x} = y_1\,\mathbf{s}_1 + y_2\,\mathbf{s}_2 + y_3\,\mathbf{s}_3 \tag{7}$$

where $\mathbf{s}_1$, $\mathbf{s}_2$ and $\mathbf{s}_3$ are the pure component spectra at unit concentration defined in a range of 100 data points (see Figure 1A), and y_1, y_2 and y_3 are the component concentrations in a specific sample. The pure component signals $\mathbf{s}_1$, $\mathbf{s}_2$ and $\mathbf{s}_3$ are Gaussian shaped functions, centered at sensors 50, 40 and 20 respectively, with full widths at half maximum of 24 sensors in the three cases. All constituents are present in the calibration set, composed of 100 samples with randomly chosen concentrations ranging from 0 to 1. Two types of test samples were created, where: (1) all components have random concentrations in the range from 0 to 1 in 100 different samples, and (2) the analyte of interest (component 1) is absent, and the remaining two components have random concentrations in the range 0-1 in additional 100 different samples.

Gaussian independent and identically distributed noise was added in three different manners: (1) only in calibration concentrations, (2) only in calibration and test sample signals, (3) in all concentrations and signals. Figure 1B shows some typical calibration signals including signal noise. For each of these noise addition modes, the PLS calibration/prediction process was repeated 1,000 times (both signal and concentration data were mean-centered) and a pseudo-univariate calibration line was obtained by regressing predicted analyte concentration values against nominal concentrations for the calibration set. The statistical parameters of the calibration lines were employed to estimate $\mathrm{LOD_{pu}}$ in each Monte Carlo cycle, as proposed by Ortiz et al. for estimating the LOD (see below).[13] The mean $\mathrm{LOD_{pu}}$ value was then compared with the extremes of the presently proposed LOD, estimated from equations (12) and (13) using in both cases the 'effective' leverages ($h_{0\mathrm{min}} + 1/I$) and ($h_{0\mathrm{max}} + 1/I$).

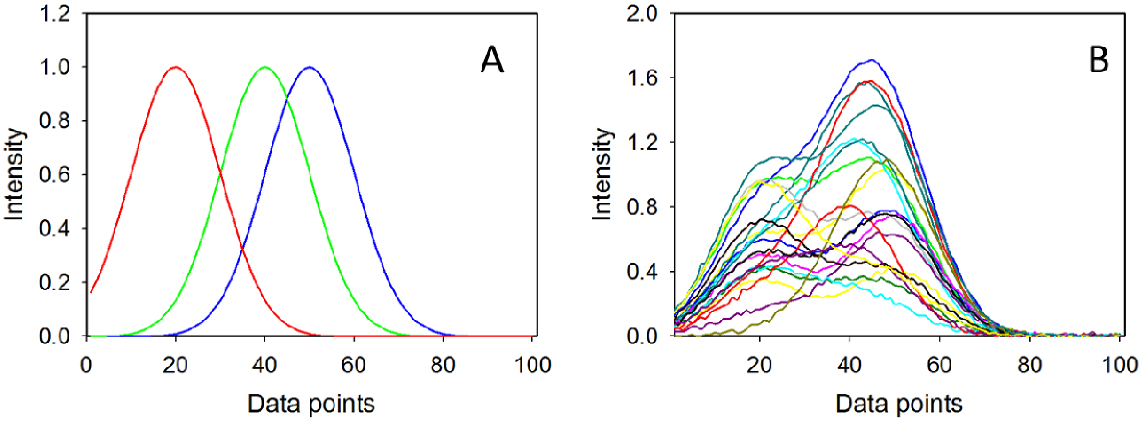


Figure 1. A) Pure component spectra employed to build the synthetic data sets: blue line, analyte of interest, green and red lines, additional sample components. B) Representative calibration spectra created from the noiseless profiles shown in A), including random instrumental noise.

Experimental data. Several experimental data sets, previously analyzed using PLS regression, were employed to assess the detection limit with the newly proposed approach, and also with univariate extensions of the LOD. They comprise the following analytes of interest and sample types: (1) fluoride ion in natural waters containing sulphate as potential interferent,[34] (2) 2-sec-butyl-4,6-dinitrophenol (DINOSEB) in a complex reacting mixture containing aromatic hydrocarbons,[35] (3) bromhexine in anti-coughing syrups,[36] (4) the antibiotic tetracycline in human sera,[37] (5) biodiesel in mixtures with diesel oil[38] and (6) humidity in corn seeds.[39] The spectral data measured for these systems were as follows: (1), (2) and (3), UV-visible spectra, (4), synchronous fluorescence spectra, and (5) and (6), NIR spectra. Experimental details on the preparation of calibration standards and test samples, measurement of instrumental signals and PLS modeling can be found in refs. 34-37. Data set No. (6) is available on the internet at http://www.eigenvector.com/data/Corn/. In all cases, both signal and concentration data were mean-centered prior to PLS modeling. All these data sets have been included as Supporting Information.

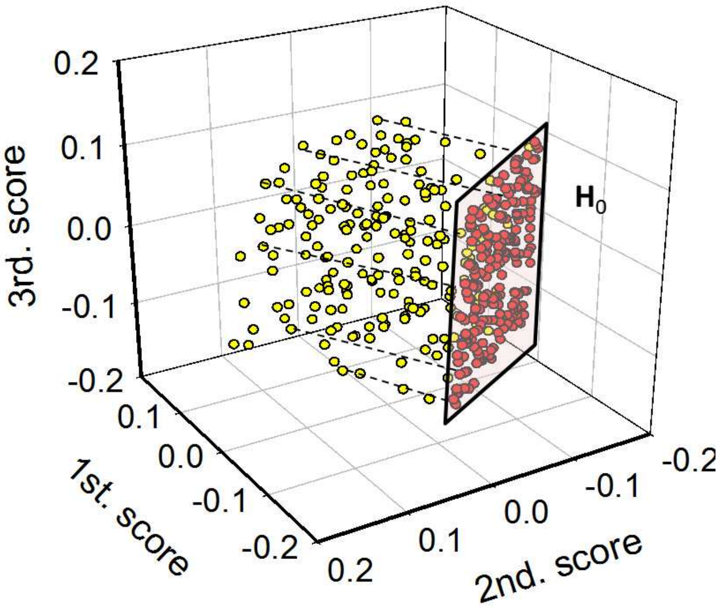


Figure 2. Location of samples in PLS score space for the ternary synthetic data set: yellow circles, samples having random concentrations of the three components (in the range from 0 to 1), red circles, samples having zero analyte concentration and random concentrations of the two additional components (in the same range of values).

RESULTS AND DISCUSSION

LOD interval for PLS calibration. For the simulated ternary system consisting of one analyte to be quantitated, in the presence of two additional components, the number of calibration latent variables for constructing a PLS model is three. This means that each sample has an associated score vector **t** of size 3×1, and can thus be plotted as a point in three-dimensional score space. Figure 2 shows the location of a number of test samples, where it can be seen that: (1) the samples with zero analyte concentration (red circles) lie in a definite region $\mathbf{H}_0$ of the π_0 plane, and (2) the projections of the positions of the remaining test samples (yellow circles), perpendicular to π_0, do also lie within $\mathbf{H}_0$. This suggests that the latter region embraces all possible blank samples (from the point of view of component 1 as the analyte of interest) which are represented by the chosen calibration set. The overall idea of the present work is to find the limits of $\mathbf{H}_0$ in score space, *even if blank samples were not included in the calibration set.*

In general, a hyperplane π_0 exists for every calibration set, representing the scores of the samples for which the analyte of interest is absent, i.e., the specific background for each sample. Resorting to equation (2), the hyperplane in *A*-dimensional score space can be defined by the following equation (signal and concentration mean-centering is assumed):

$$\pi_0\text{: } \mathbf{v}^{\mathrm{T}}\,\mathbf{t} + \bar{y}_{\mathrm{cal}} = 0 \qquad (8)$$

Since the LOD is a function of the variance in the predicted analyte concentration for a blank sample, which is in turn a function of h_0, estimating the LOD interval consists on finding the minimum ($h_{0\mathrm{min}}$) and the maximum ($h_{0\mathrm{max}}$) value of this parameter for a certain calibration set. From a geometrical point of view, $h_{0\mathrm{min}}$ is the minimum distance between π_0 and the center of a normalized calibration score space (see Appendix), i.e., the perpendicular distance from π_0 to the center. Interestingly, the Appendix shows that $h_{0\mathrm{min}}$ is simply given by:

$$h_{0\,\mathrm{min}} = \frac{\bar{y}_{\mathrm{cal}}^{2}}{\sum_{i=1}^{I} y_i^2} \qquad (9)$$

where y_i is the centered concentration for the *i*th calibration sample. The leverage in equation (9) corresponds to the value obtained in univariate calibration with a given calibration set, provided other sample components are absent.[1] On the other hand, the upper limit $h_{0\max}$ can be estimated by first computing the leverages for the projections ($h_{0\text{cal}}$) of all calibration samples onto π_0 (see Appendix):

$$h_{0\text{cal}} = h_{\text{cal}} + h_{0\min}\left[1-\left(\frac{y_{\text{cal}}}{\bar{y}_{\text{cal}}}\right)^2\right] \qquad (10)$$

where h_{cal} and y_{cal} are the leverage and (centered) analyte concentration of a generic calibration sample. Then the maximum of all possible $h_{0\text{cal}}$ values is found:

$$h_{0\max} = \max(h_{0\text{cal}}) \qquad (11)$$

The values of $h_{0\min}$ and $h_{0\max}$ [or the 'effective' leverages ($h_{0\min}$ +1/I) and ($h_{0\max}$ +1/I) for mean-centered data] can subsequently be inserted in equations (4) and (5) to obtain the lower and upper limits of the LOD interval:

$$\text{LOD}_{\min} = 3.3\ [\text{SEN}^{-2}\ \text{var}(x) + h_{0\min}\ \text{SEN}^{-2}\ \text{var}(x) + h_{0\min}\ \text{var}(y_{\text{cal}})]^{1/2} \qquad (12)$$

$$\text{LOD}_{\max} = 3.3\ [\text{SEN}^{-2}\ \text{var}(x) + h_{0\max}\ \text{SEN}^{-2}\ \text{var}(x) + h_{0\max}\ \text{var}(y_{\text{cal}})]^{1/2} \qquad (13)$$

These limits can be reported for a PLS calibration based on a given set of samples, and characterize the overall model and not a specific test sample.

It should be noticed that $\text{LOD}_{\min}$ and $\text{LOD}_{\max}$ depend on the leverage, which is a function of the calibration score matrix **T**. Since this matrix depends on the calibration design, i.e., the set of samples selected for calibration and the number of calibration latent variables, the limits of the LOD interval will also be depend on these two factors. The importance of methodologies to determine a number of factors that avoid overfitting, and to choose a set of samples with spectral features which span most of the expected variability of future samples in spectral space, has been treated in detail in the literature.[25, 40] This implies that the assumption throughout this work is that the correct design of the calibrations leads to an unbiased prediction.

Decision rules for detection. Once the limits of the LOD interval are set, the analyst may declare that the analyte is not detected in a given test sample if its predicted concentration is below $\mathrm{LOD_{min}}$, or that it is present if its predicted concentration is above $\mathrm{LOD_{max}}$. In principle, the question remains unsolved for samples whose predicted analyte concentrations lie within both LOD interval limits. Figure 3 provides a schematic representation of the three possible situations that can be found in practice.

In the concentration range $\mathrm{LOD_{min}} < y < \mathrm{LOD_{max}}$, the question can be solved by estimating a specific LOD value for the test sample, approximating its real leverage h to the leverage h_0 which would correspond to its background components, i.e., in the absence of analyte. This is equivalent to taking the sample as if it were a blank, which is conceivable since its analyte concentration is most probably very low. The obtained LOD value can then be employed to check whether the predicted concentration is below (analyte absent) or above (analyte present) the sample-specific LOD.

Pseudo-univariate LOD. In this approach, the analyte concentrations estimated for the calibration set of samples by the PLS model are plotted against their nominal or measured concentrations.[13] The result is a pseudo-univariate calibration graph in which the vertical scale is the estimated analyte concentration instead of either instrumental or latent variables. The graph is processed as in univariate calibration, assuming that the detection limit is insensitive to any linear transformation applied to the signal.[13] This leads to an $\mathrm{LOD_{pu}}$ value, estimated from the classical univariate equation:[4]

$$\mathrm{LOD_{pu}} = 3.3\ s_{\mathrm{pu}}^{-1}\ [(1 + h_{0\mathrm{min}} + 1/I)\ \mathrm{var_{pu}}]^{1/2} \qquad (14)$$

where s_{pu} is the slope of the pseudo-univariate line and $\mathrm{var_{pu}}$ is the variance of the regression residuals. Equation (14) does not include a term accounting for calibration concentration uncertainties, as is customary in univariate calibration.

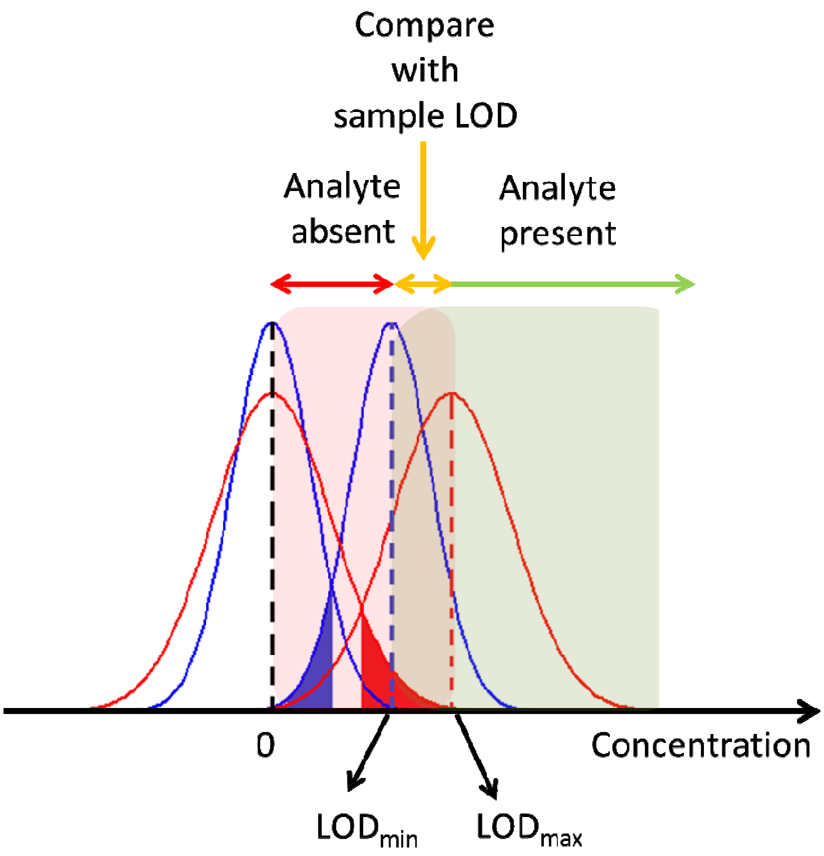


Figure 3. Schematic representation of the minimum and maximum LOD values proposed in the present report, and the decisions concerning the presence or absence of the analyte in different concentration ranges. The blue shaded region corresponds to Type I errors for the minimum LOD, while the red shaded region to Type II errors for the maximum LOD.

The parameter LOD_{pu} has the advantage of being a single figure of merit characterizing the overall PLS calibration model. However, the underlying idea is not consistent with the LOD interval described above, and it is not clear which is the relationship among LOD_{pu} and the lower and upper interval values LOD_{min} and LOD_{max}. One of the purposes of the present work was to uncover such a relationship, which will be discussed in the next sections.

Simulated data. The simulated data set was employed to calculate and compare the pseudo-univariate PLS detection limit defined by Ortiz et al. (LOD_{pu}),[13] with the

LOD interval proposed in this work (from $\mathrm{LOD_{min}}$ to $\mathrm{LOD_{max}}$). Monte Carlo simulations allowed to study the behavior of both estimators under the effect of different noise sources. The simulations were performed in the following way: after creating a data set with a predefined sensitivity given by the relative position of the analyte peak with respect to the interfering agents, noise was added in the three different manners described in the relevant section. Mean-centered (both in signal and concentration) PLS models were built using three calibration latent variables, and analyte concentrations were predicted in the calibration and in the test samples. The calibration/prediction process was repeated 1000 times using different random seeds for the signal and/or concentration uncertainties, depending on the manner in which noise was added to the synthetic data. In each of these cycles, predicted analyte concentrations in the calibration samples were regressed against their nominal concentrations, estimating the $\mathrm{LOD_{pu}}$ value with equation (14) as described by Ortiz et al., considering the latter regression as a true univariate calibration.[13]

Table 1. Comparison of LOD values in the simulated system.[a]

Uncertainty in instrumental signals	Uncertainty in calibration concentrations	Mean $\mathrm{LOD_{pu}}$	$\mathrm{LOD_{min}}/\mathrm{LOD_{max}}$
0.005	0	0.0067	0.0067/0.0069
0	0.005	0.017	0.0033/0.0052
0.005	0.005	0.018	0.0075/0.0086
0.01	0	0.013	0.013/0.014
0	0.01	0.033	0.0047/0.0073
0.01	0.01	0.036	0.014/0.016
0.008	0.001	0.0111	0.0106/0.0108

[a] All values are given in arbitrary signal and concentration units.

While LOD_{min} and LOD_{max} did not significantly change from run to run, the Monte Carlo LOD_{pu} values follow a Gaussian behavior, as shown in Figure 4 in two typical cases. The means of the LOD_{pu} distributions are compared in Table 1 with the lower and upper limits of the LOD interval (LOD_{min} and LOD_{max}) in several different cases. It is interesting to note that the LOD_{pu} distribution is centered at the lower limit LOD_{min} of the presently proposed LOD interval, provided the noise in calibration concentrations is negligible compared to the level of noise in instrumental signals (Table 1 and Figure 4A). This result can be explained on the following facts regarding the estimation of LOD_{pu}: (1) the variance of the pseudo-univariate regression residuals var_{pu} approaches [SEN^{-2} var(x)],[41] and (2) the regression slope s_{pu} is expected to be close to 1. Introduction of these parameters in equation (14) leads to an LOD_{pu} identical to LOD_{min} [equation (12) with var(y_{cal}) ≈ 0 and and 'effective' leverage ($h_{0min} + 1/I$)].

In contrast, when concentration uncertainties compete with the instrumental noise in relative size, the mutual relationship among LOD_{pu}, LOD_{min} and LOD_{max} is less clear. As shown in Table 1 and also illustrated in Figure 4B, the LOD_{pu} value can be even larger than the upper limit LOD_{max}. This can be explained on the basis of how the errors in calibration concentrations var(y_{cal}) are incorporated into the LOD definitions. In the estimation of both LOD_{min} and LOD_{max}, the latter contribution is scaled by the leverage, but in LOD_{pu} it is directly incorporated into the first, test sample-dependent term of the LOD expression. In the latter case, the 'signal' is replaced by the estimated concentrations, and therefore concentrations errors are directly propagated to the standard error in predicted concentration. In any case, the conceptual approach to LOD_{pu} is radically different than the presently proposed range of LOD values, which should in principle lead to a better insight into the PLS detection capabilities.

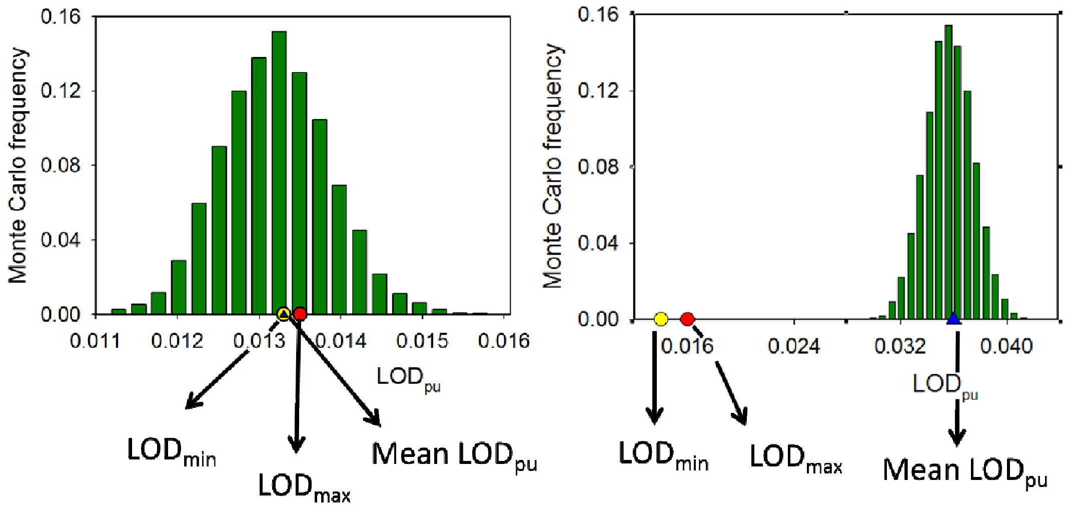


Figure 4. Distribution histograms of LOD$_{pu}$ values after repeated Monte Carlo calculations in a typical simulated data set, for negligible (A) and finite (B) uncertainties in calibration concentrations. The mutual relationship among the mean LOD$_{pu}$ value, LOD$_{min}$ and LOD$_{max}$ are shown. Specific uncertainties employed in (A) and (B) are: concentration, 0 and 0.01, signal, 0.01 and 0.01 units respectively.

Experimental data. In all the experimental systems, the PLS models were built as already reported in the literature,[34-37] using a number of calibration samples and latent variables as summarized in Table 2. The values of LOD$_{pu}$ were estimated as described above, from the pseudo-univariate plot of estimated vs. nominal (or measured, depending on the system) analyte concentrations in the calibration set of samples. For the estimation of the LOD interval proposed in the present work, equations (12) and (13) were employed, inserting appropriate values of the following parameters: (1) sensitivity, as the inverse of the length of the vector of regression coefficients computed with the PLS model, (2) the minimum and maximum 'effective' leverage values ($h_{0\min}$ + $1/I$) and ($h_{0\max}$ + $1/I$), because mean-centering was employed. The variance in spectral signals was estimated from the consideration of the average spectral residuals when modeling the test set of samples (Table 2). Regarding the variance in concentrations, when the calibration samples are prepared starting from analyte standards, the uncertainties are usually known by the analyst from uncertainty propagation analysis. This occurs in the first five examples of Table 2. In the last entry of this table, on the other hand, humidity values were measured by a reference technique, and hence the uncertainty can in principle be estimated from replicate analysis. In the absence of this information, we have employed the average uncertainty when predicting the calibration concentrations by the PLS model. This discussion highlights the need of estimating the calibration concentration uncertainties in a reliable manner (either from replicate reference measurements or from error propagation considerations), because they constitute a key aspect in the present LOD calculations.

As can be appreciated in the first five cases of Table 2, the LOD_{pu} values are larger than the maximum values LOD_{max} of the presently proposed LOD range. This is probably due to the fact that in these cases the calibration concentration errors are relevant, as in most analytical systems, and agrees with the conclusions reached during the simulation study. In the case of the calibration for humidity in seeds (last entry in Table 2), the reference values were measured by a very precise gravimetric method. Under very small concentration uncertainties, the LOD_{pu} approaches LOD_{min}, in agreement with the simulation results.

The example where tetracycline was detected in human sera (Table 2) deserves a special attention. In ref. 37, a rather cumbersome experimental procedure was employed to approximate the detection limit, preparing a large set of experimental samples having various analyte concentration levels near the expected LOD value. A detailed statistical analysis was then undertaken to detect the analyte concentration for which the predicted concentration was statistically different than zero. The reported LOD value was of ca. 0.30 mg L^{-1},[37] which can now be favorably compared with the limits of the LOD interval quoted in Table 2. This implies that the LOD for this PLS model could have been adequately estimated from the calibration set, without the need of preparing an additional set of low analyte concentration samples.

Table 2. Comparison of LOD values in experimental systems.[a]

System	Fluoride in natural waters	DINOSEB in a reacting mixture	Bromhexine in syrups	Tetracycline in sera	Biodiesel in diesel oil	Humidity in corn
Spectra	UV-visible	UV-visible	UV-visible	Synchronous fluorescence	NIR	NIR
Concentration range	0-1.4 mg L^{-1}	0-261 mg L^{-1}	1.55-2.66×10^{-4} mol L^{-1}	0-4 mg L^{-1}	0-20 %	9.4-10.9 %
I	36	10	12	50	48	50
A	4	2	3	4	11	13
$[\mathrm{var}(x)]^{1/2}$	0.001	0.001	0.006	3	0.001	0.001
$[\mathrm{var}(y_{\mathrm{cal}})]^{1/2}$	0.01	0.3	1×10^{-6}	0.15	0.01	0.005
LOD$_{\mathrm{pu}}$	0.18	1.7	0.065	0.30	2.8	0.080
LOD$_{\mathrm{min}}$	0.028	0.47	0.053	0.16	0.74	0.080
LOD$_{\mathrm{max}}$	0.040	0.77	0.057	0.27	1.1	0.081

[a] I = number of calibration samples. A = number of PLS latent variables. All LOD values are given in the same units as the corresponding concentration range. Signal uncertainties $[\mathrm{var}(x)]^{1/2}$ are given in absorbance units, except for tetracycline in sera, which are in arbitrary fluorescence intensity units. Concentration uncertainties $[\mathrm{var}(y_{\mathrm{cal}})]^{1/2}$ are given in the same units as the corresponding concentration ranges.

CONCLUSIONS

A new way of calculating the limit of detection in partial least-squares regression was investigated, together with the corresponding results towards both simulated and experimental data sets. The method is based on a geometrical analysis of the multivariate leverage definition in the latent space, and combines mathematical and analytical criteria, leading to a new LOD estimator which adopts the form of a detection interval. This proposal represents an adequate trade-off between the two main current trends regarding the multivariate LOD definition: one aiming to calculate a sample-dependent LOD based on the EIV model, and the other one extending the ISO/IUPAC univariate definition to ascribe a unique LOD value to a given calibration model. The presently proposed estimator can be easily extended to other inverse multivariate models, although further studies should be made to apply it to more complex multiway data.

Acknowledgements

Universidad Nacional de Rosario, CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas), ANPCyT (Agencia Nacional de Promoción Científica y Tecnológica, Project No. PICT-2010-0084) are gratefully acknowledged for financial support. F. A. thanks CONICET for a doctoral fellowship and Fundación Josefina Prats for financial support.

Supporting Information

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

References

1 Danzer, K.; Currie, L. A. *Pure Appl. Chem*. **1998**, *70*, 993-1014.

2 Olivieri, A. C. *Chem. Rev*. **2014**, *114*, 5358-5378.

3 Olivieri, A. C.; Faber, N. M.; Ferré, J.; Boqué, R; Kalivas, J. H.; Mark, H. *Pure Appl. Chem*. **2006**, *78*, 633-661.

4 Currie, L. A. *Anal. Chim. Acta* **1999**, *391*, 127-134.

5 Loock, H. P.; Wentzell, P. D. *Sensor. Actuat. B-Chem* **2012**, *173*, 157-163.

6 Boqué, R.; Rius, F. X. *Chemom. Intell. Lab. Syst.* **1996**, *32*, 11-23.

7 Boqué, R.; Larrechi, M. S.; Rius, F. X. *Chemom. Intell. Lab. Syst.* **1999**, *45*, 397-408.

8 ISO 11843-1, *Capability of detection*, Genève, Switzerland, 1997.

9 ISO 11843-2, *Capability of detection*, Genève, Switzerland, 2000.

10 McNaught, A. D.; Wilkinson, A. IUPAC, *Compendium of Chemical Terminology* 2nd ed., Blackwell, Oxford, 1997.

11 Inczédy, J.; Lengyel, T.; Ure, A.M.; Gelencsér, A.; Hulanicki, A. IUPAC Analytical Chemistry Division, *Compendium of Analytical Nomenclature*, 3rd. ed., Blackwell, Oxford, 1998.

12 MacDougall, D.; Crummett, W. B. *Anal. Chem.* **1980**, *52*, 2242-2249.

13 Ortiz, M. C.; Sarabia, L. A.; Herrero, A.; Sánchez, M. S.; Sanz, M. B.; Rueda, M. E.; Giménez, D.; Meléndez, M. E. *Chemom. Intell. Lab. Syst*. **2003**, *69*, 21-33.

14 Burns, D. A.; Ciurczak, E. W. *Handbook of near-infrared analysis*, 3rd ed., *Practical Spectroscopy Series*, CRC Press, Boca Raton, USA, Vol. 35, 2008.

15 Lorber, A.; Faber, K.; Kowalski, B. R. *Anal. Chem.* **1997**, *69*, 1620-1626.

16 Ostra, M.; Ubide, C.; Vidal, M.; Zuriarrain, J. *Analyst* **2008**, *133*, 532-539.

17 Faber, N. M.; Bro, R. *Chemom. Intell. Lab. Syst.* **2002**, *61*, 133-149.

18 Blanco, M.; Castillo, M.; Peinado, A.; Beneyto, R. *Anal. Chim. Acta* **2007**, *581*, 318-323.

19 Wu, Z.; Sui, C.; Xu, B.; Ai, L.; Ma, Q.; Shi, X.; Qiao, Y. *J. Pharm. Biomed. Anal.* **2013**, *77*, 16-20.

20 Ortiz, M. C.; Sarabia, L. A.; Sánchez, M. S. *Anal. Chim. Acta*, **2010**, *674*, 123-142.

21 Faber, K.; Kowalski, B. R. *J. Chemometr.* **1997**, *11*, 181-238.

22 Martens, H.; Næs, T. *Multivariate Calibration*, John Wiley, Chichester, 1989**.**

23 Wold, S.; Sjöström, M.; Eriksson, L. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109-130.

24 Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; De Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of chemometrics and qualimetrics*, Elsevier, Amsterdam, 1997.

25 Haaland, D. M.; Thomas, E. V. *Anal. Chem.* **1988**, *60*, 1193-1202.

26 Clayton, C. A.; Hines, J. W.; Elkins, P. D. *Anal. Chem.* **1987**, *59*, 2506-2514.

27 Del Río Bocio, F. J.; Riu, J.; Boqué, R.; Rius, F. X. *J. Chemometr.* **2003**, *17*, 413-421.

28 Voigtman E. *Spectrochim. Acta B* **2008**, *63*, 129-141.

29 Voigtman, E. *Spectrochim. Acta B* **2008**, *63*, 115-128.

30 Fernández Pierna, J. A.; Jin, L.; Wahl, F.; Faber, N. M.; Massart D. L. *Chemom. Intell. Lab. Syst.* **2003**, *65*, 281-291.

31 Faber, N. M.; Song, X. H.; Hopke, P. K. *Trends Anal. Chem.* **2003**, *22*, 330-334.

32 Bro, R.; Rinnan, Å.; Faber, N. M. *Chemom. Intell. Lab. Syst.* **2005**, *75*, 69-76.

33 Faber, K.; Lorber, A.; Kowalski, B. R. *J. Chemometr.* **1997**, *11*, 419-461.

34 Arancibia, J. A.; Rullo, A.; Olivieri, A. C.; Di Nezio, S.; Pistonesi, M.; Lista, A.; Fernández Band, B. S. *Anal. Chim. Acta* **2004**, *512*, 157-163.

35 Arancibia, J. A.; Martínez Delfa, G.; Boschetti, C. E.; Escandar, G. M.; Olivieri, A. C. *Anal. Chim. Acta* **2005**, *553*, 141-147.

36 Goicoechea, H. C.; Olivieri, A. C. *Talanta* **1999**, *49*, 793-800.

37 Goicoechea, H. C.; Olivieri, A. C. *Anal. Chem*. **1999**, *19*, 4361-4368.

38 Pedrido, M. L.; Bortolato, S.; González Sierra, M.; Olivieri, A. C.; Boschetti, C. E. *LabCiencia* **2008**, *3*, 14-18.

39 Allegrini, F.; Olivieri, A. C. *Anal. Chim. Acta* **2011**, *699*, 18-25.

40 Kennard, R. W.; Stone, L. A. *Technometrics* **1969,** *11*, 137-148.

41 Olivieri, A. C. *Anal. Chem*. **2005**, *77*, 4936-4946.

APPENDIX

In this Appendix some relevant results concerning the presently proposed LOD interval for PLS calibration are derived. It is first important to recognize that the leverages are squared distances in score space, once the latter ones are properly normalized [cf. equation (6)], i.e., each score element t_a is multiplied by the factor f_a, which is the ath diagonal element of the $A \times A$ square matrix $(\mathbf{T}^{\mathrm{T}}\ \mathbf{T})^{-1/2}$ ($\mathbf{T}$ is the calibration score matrix). In what follows, we will call the normalized score vectors as: $\mathbf{t}_{\mathrm{N}}$ for a generic sample, $\mathbf{t}_{\mathrm{Ncal}}$ for a calibration sample, and $\mathbf{t}_{\mathrm{N0cal}}$ for the projection of a calibration sample perpendicular to the π_0 hyperplane defined by zero analyte concentration. Specific ath elements of these vectors will be called $t_{a\mathrm{N}}$, $t_{a\mathrm{Ncal}}$ and $t_{a\mathrm{N0cal}}$ respectively.

The expression defining π_0 in score space is (mean-centered signal and concentration data are assumed):

$$\pi_0\text{: } \mathbf{v}^{\mathrm{T}}\ \mathbf{t} + \bar{y}_{\mathrm{cal}} = 0 \tag{A-1}$$

which can be written in terms of normalized scores as follows:

$$\sum_{a=1}^{A} \frac{v_a t_{a\mathrm{N}}}{f_a \bar{y}_{\mathrm{cal}}} = -1 \tag{A-2}$$

A calibration sample located at $\mathbf{t}_{\mathrm{Ncal}}$ can be projected perpendicular to π_0 along the parametric straight line:

$$t_{a\mathrm{N}} = -\frac{v_a}{f_a \bar{y}_{\mathrm{cal}}} k + t_{a\mathrm{Ncal}} \tag{A-3}$$

where k is a variable parameter. The intersection of the latter line with π_0 occurs at the following point:

$$\sum_{a=1}^{A} k \left(\frac{v_a}{f_a \bar{y}_{\mathrm{cal}}} \right)^2 - \frac{v_a t_{a\mathrm{Ncal}}}{f_a \bar{y}_{\mathrm{cal}}} = 1 \tag{A-4}$$

from which k can be calculated as:

$$k = \frac{\bar{y}_{\mathrm{cal}}^2 + \bar{y}_{\mathrm{cal}} \sum_{a=1}^{A} \frac{v_a t_{a\mathrm{Ncal}}}{f_a}}{\sum_{a=1}^{A} \left(\frac{v_a}{f_a} \right)^2} \qquad \text{(A-5)}$$

Thus a generic coordinate of the intersecting point is:

$$t_{a\mathrm{N0cal}} = -\left(\frac{v_a}{f_a} \right) \frac{\bar{y}_{\mathrm{cal}} + \sum_{a=1}^{A} \frac{v_a t_{a\mathrm{Ncal}}}{f_a}}{\sum_{a=1}^{A} \left(\frac{v_a}{f_a} \right)^2} + t_{a\mathrm{Ncal}} \qquad \text{(A-6)}$$

Since the value of $\left(\sum_{a=1}^{A} \frac{v_a t_{a\mathrm{Ncal}}}{f_a} \right)$ is equal to the centered concentration of a given calibration sample (y_{cal}), equation (A-6) can be rearranged to:

$$t_{a\mathrm{N0cal}} = -\frac{v_a(\bar{y}_{\mathrm{cal}} + y_{\mathrm{cal}})}{f_a \sum_{a=1}^{A} \left(\frac{v_a}{f_a} \right)^2} + t_{a\mathrm{Ncal}} \qquad \text{(A-7)}$$

In equation (A-7), $\sum_{a=1}^{A} \left(\frac{v_a}{f_a} \right)^2$ can be converted to calibration concentrations by noting that the $\mathbf{t}_a$ columns of the $\mathbf{T}$ matrix are orthogonal, i.e., $\mathbf{t}_a^{\mathrm{T}} \mathbf{t}_{a'} = \sum_{i=1}^{I} t_{ia} t_{ia'} = 0$ if $a \neq a'$, which implies the following result:

$$\sum_{a=1}^{A} \left(\frac{v_a}{f_a} \right)^2 = \sum_{a=1}^{A} v_a^2 \mathbf{t}_a^{\mathrm{T}} \mathbf{t}_a = \sum_{a=1}^{A} (v_a \sum_{i=1}^{I} t_{ia}^2) = \sum_{i=1}^{I} \left(\sum_{a=1}^{A} v_a t_{ia} \right)^2 \approx \sum_{i=1}^{I} y_i^2 \qquad \text{(A-8)}$$

where y_i is the centered concentration for the ith calibration sample, estimated from the product of regression coefficients v_a and sample scores t_{ia}.

We now define the minimum projected leverage $h_{0\min}$ as the known expression for the pseudo-univariate leverage for a blank sample:

$$h_{0\min} \approx \frac{\bar{y}_{\mathrm{cal}}^2}{\sum_{i=1}^{I} y_i^2} \qquad \text{(A-9)}$$

From these results, it is possible to transform equation (A-7) in the following simple expression:

$$t_{a\mathrm{N0cal}} = -\frac{v_a(\bar{y}_{\mathrm{cal}} + y_{\mathrm{cal}})}{f_a \bar{y}_{\mathrm{cal}}^2} h_{0\mathrm{min}} + t_{a\mathrm{Ncal}} \qquad \text{(A-10)}$$

The squared length of the vector $\mathbf{t}_{\mathrm{N0cal}}$ [with coordinates given in equation (A-10)] is the leverage ($h_{0\mathrm{cal}}$) of a sample of zero analyte concentration, hypothetically projected perpendicular to π_0. From the above expressions it can be shown that:

$$h_{0\mathrm{cal}} = h_{\mathrm{cal}} + h_{0\mathrm{min}}\left[1 - \left(\frac{y_{\mathrm{cal}}}{\bar{y}_{\mathrm{cal}}}\right)^2\right] \qquad \text{(A-11)}$$

where h_{cal} is the leverage for the calibration sample and y_{cal} is centered. It can easily be seen that at the calibration center, where both h_{cal} and y_{cal} are zero, the minimum projection to π_0 is obtained, i.e., $h_{0\mathrm{cal}} = h_{0\mathrm{min}}$, hence the name $h_{0\mathrm{min}}$ in equation (A-9).

Interestingly, equation (A-11) can be derived from simple trigonometric arguments:

$$h_{0\mathrm{cal}} = h_{0\mathrm{min}} + Q^2 = h_{0\mathrm{min}} + (h_{\mathrm{cal}} - M^2) \qquad \text{(A-12)}$$

where the segments M and Q are defined in Figure 5. From this figure, if the leverages are interpreted as squared distances proportional to concentration, then $M^2 = h_{0\mathrm{min}}\left(\frac{y_{\mathrm{cal}} - \bar{y}_{\mathrm{cal}}}{\bar{y}_{\mathrm{cal}}}\right)^2$, and equation (A-11) immediately follows from equation (A-12).

The conclusion is that at zero analyte level, a range of sample leverages occur, which depend on the variability of the background composition, with two extreme values: the minimum ($h_{0\mathrm{min}}$) given by equation (A-9), and the maximum of all $h_{0\mathrm{cal}}$ values which are provided by equation (A-11).

It should be noticed that all the leverage expressions discussed above correspond to mean-centered data (both signals and concentrations). Before inserting any of these leverages, particularly the minimum and maximum $h_{0\mathrm{min}}$ and $h_{0\mathrm{max}}$ values, in the corresponding expression for the concentration uncertainty, they have to be converted into 'effective' leverages, i.e., ($h_{0\mathrm{min}} + 1/I$) and $h_{0\mathrm{max}} + 1/I$).

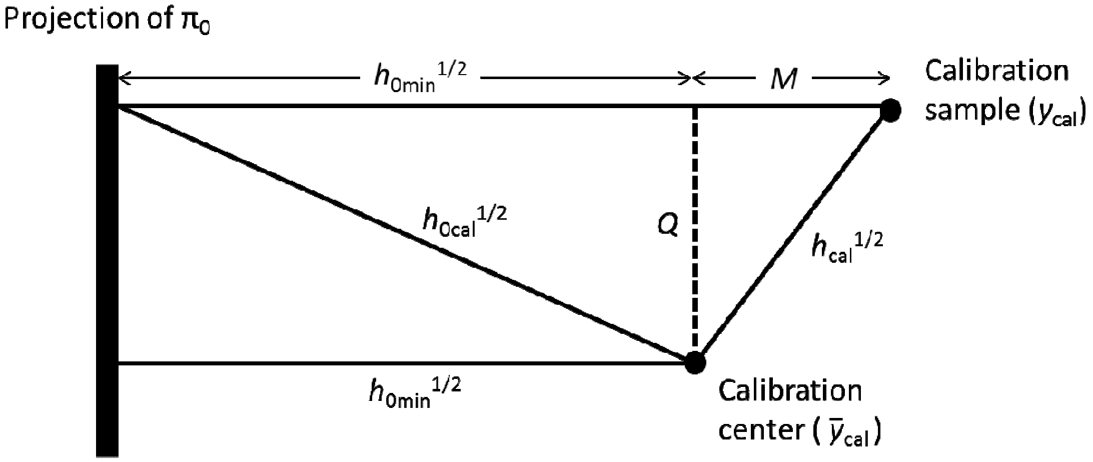
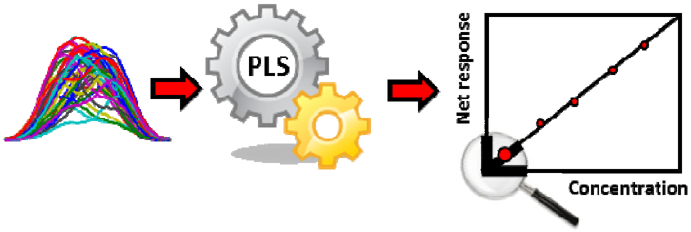

Projection of π_0
$h_{0\mathrm{min}}^{1/2}$
M
Calibration sample (y_{cal})
$h_{0\mathrm{cal}}^{1/2}$
Q
$h_{\mathrm{cal}}^{1/2}$
$h_{0\mathrm{min}}^{1/2}$
Calibration center ($\bar{y}_{\mathrm{cal}}$)

Figure 5. Schematic representation of the leverage parameters relevant to the present work. The thick black line implies the projection of the π_0 plane, the black circles indicate the location of the calibration center (analyte concentration = $\bar{y}_{\mathrm{cal}}$), and a given calibration sample (analyte concentration = y_{cal}). Additional 'distances' in score space (square roots of leverage values) are noted.

For TOC only


PLS
Net response
Concentration