JOURNAL OF CHEMOMETRICS
J. Chemometrics 2005; **19**: 45–54
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cem.906

Partial least squares, Beer's law and the net analyte signal: statistical modeling and analysis

Boaz Nadler* and Ronald R. Coifman
Department of Mathematics, Yale University, New-Haven, CT 06520, USA

Received 15 August 2004; Revised 25 April 2005; Accepted 1 May 2005

Partial least squares (PLS) is one of the most common regression algorithms in chemistry, relating input–output samples (x_i, y_i) by a linear multivariate model. In this paper we analyze the PLS algorithm under a specific probabilistic model for the relation between x and y. Following Beer's law, we assume a linear mixture model in which each data sample (x, y) is a random realization from a joint probability distribution where x is the sum of k components multiplied by their respective characteristic responses, and each of these components is a random variable. We analyze PLS on this model under two idealized settings: one is the ideal case of noise-free samples and the other is the case of an infinite number of noisy training samples. In the noise-free case we prove that, as expected, the regression vector computed by PLS is, up to normalization, the net analyte signal. We prove that PLS computes this vector after at most k iterations, where k is the total number of components. In the case of an infinite training set corrupted by unstructured noise, we show that PLS computes a final regression vector which is not in general purely proportional to the net analyte signal vector, but has the important property of being optimal under a mean squared error of prediction criterion. This result can be viewed as an asymptotic optimality of PLS in the limit of a very large but finite training set. Copyright © 2005 John Wiley & Sons, Ltd.

1. INTRODUCTION

Partial least squares (PLS) is one of the most common regression algorithms in the field of chemometrics in general, and spectroscopy in particular [1–3]. In the typical setting, given a finite training set with n samples ($\mathbf{x}_i$, y_i), PLS builds a linear relationship between $\mathbf{x}$ and y that is then used for prediction of y for new data $\mathbf{x}$.

The main assumption of PLS is that the data $\mathbf{x}$, although possibly residing in a high-dimensional space, depend linearly on only a small number of latent variables. PLS estimates these latent variables as projections of the original input variables of $\mathbf{x}$ and uses them to construct the regression vector relating $\mathbf{x}$ to y [3,4]. Much theoretical work and many simulations have been devoted to explaining what PLS does and why it works so well in practical spectroscopic applications, characterized by high collinearity in the input data and lack of specificity at any individual predictor variable (see e.g. References [1,5,6] and references cited therein). In some simulation studies of PLS the probabilistic model considered for the input and output is that data samples ($\mathbf{x}$, y) follow a joint multivariate Gaussian distribution [7–9]. However, as pointed out by Wold [4], in these models the data matrix reaches full rank as the number of samples tends to infinity, and therefore this modeling approach is inconsistent both with the underlying assumptions of PLS and with the basic physics and chemistry of real systems for which measurements are taken.

Following Wold's observation, in this paper we analyze the PLS algorithm under a probabilistic model for data samples that is motivated by physics and chemistry (Beer's law) and thus more closely resembles typical data measured in actual systems. We therefore assume a linear mixture model where each data sample ($\mathbf{x}_i$, y_i) is a random realization from a generally unknown probability space in which $\mathbf{x}$ is the linear sum of k random components each multiplied by its characteristic (spectral) response vector. While this model has been used in many simulation studies and also as a benchmark for proposed new algorithms [10–13], it seems that the theoretical analysis of PLS on such a model has not been fully explored.

In this paper we consider the PLS algorithm on this model under two different settings. The first is the idealized case of an error-free training set, while the second is the case of a noisy but infinite training set. This second case can be viewed as the asymptotic limit of a finite training set when the number of samples $n \to \infty$. The more complicated (and more interesting) theoretical analysis of PLS predictions based on a finite and noisy training set will be published separately [14].

The main results of our analysis are as follows. First we show that in the ideal noise-free case the regression vector computed by PLS is, as expected and up to a normalization constant, equal to the *net analyte signal* (NAS) vector [15,16], giving a zero prediction error. We prove that, similar to the

*Correspondence to: B. Nadler, Department of Mathematics, Yale University, New-Haven, CT 06520, USA.
E-mail: boaz.nadler@yale.edu

case of principal component regression (PCR), this vector is constructed in at most k iterative steps, where k is the total number of components in the input data. Our analysis also clarifies in statistical terms an issue that has received a lot of attention in the literature, namely the chemical interpretation of the various projections computed by PLS [2,17,18]. In general we find that the projections and loadings computed by PLS are complex linear combinations of the pure spectra of the various components with coefficients that depend both on the physical interference amongst these spectra as well as on their statistical correlations with the substance of interest. Therefore these loadings and projections can differ substantially if a new finite data set is used for calibration, even though the resulting regression vector remains the same. Therefore, much to the regret of the analytical chemist, a chemical interpretation of the loading and projection vectors of PLS is quite difficult if not impossible in complex multi component systems.

In addition we show that the error of prediction in PLS with a sub optimal number of latent variables can in some cases be dominated by interfering components with a large variance, even though these may be totally uncorrelated and unrelated with the substance of interest. This provides a theoretical motivation for various preprocessing algorithms that attempt to remove variability in the data that is uncorrelated with y, such as multiplicative scatter correction (MSC), standard normal variate (SNV) and orthogonal signal correction (OSC) [19–21].

In the setting of an infinite training set of samples corrupted by unstructured noise, we show that PLS is optimal under a mean squared error criterion. Moreover, similar to the recent analysis of Brown [22], we show that the resulting regression vector is not equal to a scaled version of the NAS, but rather depends in general on the level of noise and on all components in the system, their spectral responses and statistical correlations. Therefore, although various preprocessing algorithms such as OSC, hybrid linear analysis (HLA) or others [12,13,21] that attempt to remove variability in the data that is uncorrelated with the substance of interest may give more robust results for small training samples, they are suboptimal for prediction purposes in the asymptotic limit of a large training set.

The paper is organized as follows. In Section 2 we briefly present the PLS algorithm, while in Section 3 we define the specific input model considered in this paper. The main results for the noise-free case are derived in Section 4, while those for the noisy samples appear in Section 5. An application of these results is given in Section 6. We conclude with a summary and discussion in Section 7.

2. THE PLS ALGORITHM

2.1. Notation

We denote vectors by boldface lowercase letters, as in $\mathbf{v}$. The Euclidean norm of a vector $\mathbf{v}$ is denoted $\|\mathbf{v}\|$ and its dot product with a vector $\mathbf{w}$ is denoted $\mathbf{v}\cdot\mathbf{w}$. Random variables are denoted by italic lowercase letters, as in y and u_1. The mean of a random variable y is $\mathbb{E}\{y\}$, its variance is $\mathrm{Var}(y)$ and its covariance with another random variable u is $\mathrm{Cov}(y, u)$.

2.2. A probabilistic version of PLS

Consider a calibration (training) set of n samples $\{\mathbf{x}_i, y_i\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^p$ are the predictor or input variables and $y_i \in \mathbb{R}$ are the response or output variables. PLS computes a linear calibration model based on this finite training set, whose final output is a regression vector $\mathbf{r} \in \mathbb{R}^p$ such that subsequent predictions of y for new data points $\mathbf{x}$ are given by

$$\hat{y} = \bar{y} + \mathbf{r}\cdot(\mathbf{x} - \bar{\mathbf{x}})$$

where $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are the mean values of training set inputs and outputs respectively.

While the original PLS algorithm is defined on a finite data set described by two finite input and output matrices $\mathbf{X}$ and $\mathbf{Y}$, in recent years various authors have developed a probabilistic derivation of PLS, thus allowing its interpretation in terms of stochastic variables [6,23,24].

Since in this paper we analyze PLS under a specific stochastic model for inputs and outputs, following References [6,24], we describe this statistically derived version of PLS. Let A be a user-defined maximal number of iterative steps of the algorithm, $\mathbf{x}_0 = \mathbf{x} - \bar{\mathbf{x}}$, and with some abuse of notation we assume that y is already mean centered. For $a = 1 \ldots, A_{\max}$:

1. Find projection $\mathbf{w}_a$ such that $\mathbf{x}_{a-1}\cdot\mathbf{w}$ best correlates with y. Up to a normalization constant the best projection is given by
$$\mathbf{w}_a = \mathbb{E}\{\mathbf{x}_{a-1}y\}$$
2. Compute the coefficient of the ath latent variable:
$$t_a = \mathbf{x}_{a-1}\cdot\mathbf{w}_a$$
3. Compute the ath regression score:
$$q_a = \frac{\mathbb{E}\{yt_a\}}{\mathrm{Var}(t_a)}$$
4. Compute the ath spectral loading:
$$\mathbf{p}_a = \frac{\mathbb{E}\{\mathbf{x}_{a-1}t_a\}}{\mathrm{Var}(t_a)}$$
5. Project the score t_a:
$$\mathbf{x}_a = \mathbf{x}_{a-1} - t_a\mathbf{p}_a$$

The output of PLS after a steps is $\hat{y} = \bar{y} + \sum_{j=1}^{a} q_j t_j$, which can equivalently be written as

$$\hat{y} = \bar{y} + \mathbf{r}_a\cdot(\mathbf{x} - \bar{\mathbf{x}})$$

As derived in References [6,24], this version of PLS is defined in a *population* setting with an infinite amount of data points so that expectations are over the corresponding joint probability densities for $(\mathbf{x}, y)$. However, for a finite number of samples, replacing all expectations over infinite populations by their sample estimates recovers the original sample PLS algorithm.

3. A PROBABILISTIC MODEL OF THE INPUT DATA

While PLS can be applied to multivariate regression problems in many diverse fields, we focus our attention on the typical spectroscopic application, namely the determination

of analyte concentration from the corresponding (typically near-infrared) spectral data. The reason for this is twofold. Firstly, spectroscopy is one of the most important and common applications of the PLS algorithm, and secondly, in this setting we can formulate an approximate *physical* model that relates the input $\mathbf{x}$ to the output y, based on Beer's law.

In this setting the input data $\mathbf{x}$ are usually taken as the logarithm of the absorbance or reflectance spectrum. Beer's law states that in the absence of non-linear effects the logarithm of the spectrum is proportional to the analyte concentration y multiplied by its characteristic response spectrum $\mathbf{v}_0$ [1,2]. In the presence of many substances, under the assumption of additivity and neglecting possible interactions between the substance of interest and other substances, the resulting spectral data $\mathbf{x}$ are then proportional to the sum of all the substances in the material, each multiplied by their characteristic spectrum. In real life measurements, typically there are additional contributions to the resulting spectra due to the measuring device, such as an additive baseline shift, a random or deterministic machine drift or, in general, a more complicated characteristic spectrum of the measuring device. Physically there can also be temperature effects as well as noise from other sources and various other errors introduced by the measurement process [2].

Therefore the actual spectral data that one encounters in real life situations are rather complex. In order to mathematically analyze and understand the results of the PLS algorithm in this setting, some simplifications are obviously needed. In this paper we consider a simplified probabilistic model of the spectral data based on Beer's law and analyze the PLS algorithm under such a model. We thus assume that the noise-free data points $(\mathbf{x}, y)$ are random realizations from an underlying (and generally unknown) probability space with $k+1$ random variables, denoted y and $u_1, \ldots, u_k$, as follows:

$$\mathbf{x} = y\mathbf{v}_0 + \sum_{i=1}^{k} u_i \mathbf{v}_i \tag{1}$$

while the noisy data are given by

$$\tilde{\mathbf{x}} = \mathbf{x} + \sigma\boldsymbol{\xi} \tag{2}$$

In (1), $\mathbf{x}$ is the spectrum, y is the substance of interest in prediction, $\mathbf{v}_0 \in \mathbb{R}^p$ is its characteristic spectral response and the remaining vectors $\mathbf{v}_i$ are the characteristic spectral responses associated with the random variables u_i respectively. In the absence of noise the measured spectrum is $\mathbf{x}$, while in its presence it is $\tilde{\mathbf{x}}$ given by (2), where $\boldsymbol{\xi}$ is a random noise vector in $\mathbb{R}^p$ whose p co-ordinates are independent identically distributed random variables with zero mean and unit variance, and σ is a measure of the level of noise. The random variables u_i, sometimes denoted *components*, can be either other physical substances present in the material or measures of other physical phenomena that contribute to the measured signal. Examples of the latter phenomena include the random amplitude of a deterministic machine signal, particle size-dependent light scattering and random optical path length, to name just a few. Input–output relations of the form (1) have been considered in the literature [3,11,18] and used as a benchmark in various simulation studies and tests of new algorithms [10,12,13,25–27]. In Section 4 we present a theoretical analysis of PLS on error-free inputs of the form (1), while the case of inputs corrupted by noise according to (2) is considered in Section 5.

We note that quite a lot of analysis on the PLS algorithm has been performed on a different, more general and thus abstract model of input–output relations, namely that $\mathbf{x}$ is a multidimensional Gaussian process with correlation matrix $\boldsymbol{\Sigma}$ and $y = \boldsymbol{\beta} \cdot \mathbf{x} + e$, where e is a zero-mean random error (see e.g. References [8,23]). One of the differences in the analysis between this model and Equations (1) and (2) is that in our case the predictions of PLS can be compared with the actual parameters of the problem, and a chemical interpretation of PLS with regard to the spectral responses $\mathbf{v}_i$ can be performed.

3.1. The net analyte signal vector

We start our analysis with the ideal error-free case. The problem at hand is therefore the prediction of the substance y given the error-free signal $\mathbf{x}$, under the assumption that $\mathbf{x}$ and y are related via Equation (1). In principle, if all the response vectors $\mathbf{v}_i$ were known, then any vector $\mathbf{r}$ that is not orthogonal to $\mathbf{v}_0$ but is orthogonal to all $\mathbf{v}_i$ for $i \geq 1$ would give a perfect error-free regression of y, since in that case

$$\mathbf{x} \cdot \mathbf{r} = (\mathbf{v}_0 \cdot \mathbf{r})y + \sum_j (\mathbf{v}_j \cdot \mathbf{r})u_j = \text{Const} \times y$$

which differs from y only by an easily evaluated normalization constant.

One such vector that has received considerable attention in the literature is the *net analyte signal* (NAS) vector [15,16]. This vector is defined as the part of the response $\mathbf{v}_0$ of the substance y that is orthogonal to the response vectors $\mathbf{v}_i$ of all other components. Specifically, if $\{\mathbf{z}_j\}_{j=1}^{m}$ is an orthonormal basis for Span $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$, then

$$\text{NAS}(y) = \mathbf{v}_0 - \sum_{i=1}^{m} (\mathbf{v}_0 \cdot \mathbf{z}_i)\mathbf{z}_i$$

Typically, however, not all response vectors of all variables are known, nor is there even an explicit knowledge of exactly how many components are present in the system. PLS is thus a method to calculate a regression vector $\mathbf{r}$ without such explicit knowledge. In this paper we prove that, as expected and known empirically, in the case of error-free data the output of the PLS algorithm is exactly the NAS up to a normalization constant. Recently, various authors have considered other approaches to compute the NAS directly, either when the response $\mathbf{v}_0$ of the substance of interest y is known, as in the HLA algorithm by Berger *et al.* [12], or even without its explicit knowledge, as in References [13,15,28]. Since in the error-free case all these algorithms are equivalent, e.g. they all compute the appropriately normalized version of the NAS, they should all have similar predictive performance when relatively small noise is added to the signals, as indeed is reported in various simulation studies [25,26]. A detailed theoretical comparison of their performance when there is also error in the spectra will be considered in a separate publication.

4. ANALYSIS OF THE PLS ALGORITHM

4.1. PLS with two components

We start our analysis with one of the simplest possible examples, in which each data point (spectral measurement) $\mathbf{x}$ is the result of only two underlying components y and u_1, with corresponding spectra $\mathbf{v}_0$ and $\mathbf{v}_1$, where y is the substance of interest in prediction. Thus

$$\mathbf{x} = y\mathbf{v}_0 + u_1\mathbf{v}_1 \tag{3}$$

Since prior to application of the PLS algorithm the calibration data are mean centered, then without loss of generality we assume that $\mathbb{E}\{y\} = \mathbb{E}\{u_i\} = 0$. For future use we define the quantities

$$V_y = \mathrm{Var}(y), \qquad c_1 = \mathrm{Cov}(y, u_1), \qquad V_{u_1} = \mathrm{Var}(u_1)$$

We assume that y and u_1 indeed represent two different components, not perfectly correlated and with different spectral signals. These conditions translate mathematically into $c_1^2/V_yV_{u_1} < 1$ and $\mathbf{v}_1 \neq c\mathbf{v}_0$ for any scalar c.

The first step in PLS is to compute the projection $\mathbf{w}_1$ that best correlates with y. Up to normalization it is given by

$$\mathbf{w}_1 = \mathbb{E}\{\mathbf{x}y\} = V_y\mathbf{v}_0 + c_1\mathbf{v}_1 \tag{4}$$

In the case of the more general model (1) with a total of $k+1$ components,

$$\mathbf{w}_1 = V_y\mathbf{v}_0 + \sum_j c_j\mathbf{v}_j$$

where c_j is the covariance between u_j and y. Note that in general the first projection is not proportional to the pure response $\mathbf{v}_0$ of the substance of interest y. Assuming that the response vectors $\{\mathbf{v}_j\}$ are linearly independent, $\mathbf{w}_1$ is proportional to $\mathbf{v}_0$ if and only if all correlations c_j between y and all other components in the system vanish. This finding clarifies in mathematical terms the statement of Haaland and Thomas [11] that 'the quality of $\hat{\mathbf{w}}_1$ is dependent to some degree on relative intensities of spectral bands...'. Equation (4) shows that the important factor determining the first projection $\mathbf{w}_1$ is not the physical interference between $\mathbf{v}_0$ and the spectral responses of other vectors, but rather the *statistical correlation* in the calibration set between y and the other substances.

The second PLS step is the computation of the latent variable t_1 as

$$\begin{aligned} t_1 = \mathbf{x}\cdot\mathbf{w}_1 &= (y\mathbf{v}_0 + u_1\mathbf{v}_1)\cdot(V_y\mathbf{v}_0 + c_1\mathbf{v}_1) \\ &= y(V_y\|\mathbf{v}_0\|^2 + c_1\mathbf{v}_0\cdot\mathbf{v}_1) \\ &\quad + u_1(c_1\|\mathbf{v}_1\|^2 + V_y\mathbf{v}_0\cdot\mathbf{v}_1) \end{aligned} \tag{5}$$

while in the next step we compute the score $q_1 = \mathbb{E}\{t_1y\}/\mathrm{Var}(t_1)$. In PLS with only one latent variable the predicted value for a new data point $\mathbf{x}$ is then given by

$$\hat{\boldsymbol{y}}_1 = q_1t_1 = q_1\mathbf{w}_1\cdot\mathbf{x}$$

To compute q_1 explicitly for this example, we thus need to calculate $\mathbb{E}\{t_1y\}$ and $\mathrm{Var}(t_1)$. For ease of notation we define the two quantities

$$A = V_y\|\mathbf{v}_0\|^2 + c_1\mathbf{v}_0\cdot\mathbf{v}_1, \qquad B = c_1\|\mathbf{v}_1\|^2 + V_y\mathbf{v}_0\cdot\mathbf{v}_1 \tag{6}$$

so that $t_1 = Ay + Bu_1$ and $\mathrm{Var}(t_1) = A^2V_y + B^2Vu + 2ABc_1$. In terms of A and B, q_1 is given by

$$q_1 = \frac{\mathbb{E}\{yt_1\}}{\mathrm{Var}(t_1)} = \frac{AV_y + Bc_1}{A^2V_y + B^2V_{u_1} + 2ABc_1}$$

The most common measure of the performance of PLS is the mean squared error of prediction (MSEP). In the case of only one latent variable we obtain

$$\begin{aligned} \mathrm{MSEP} &= \mathbb{E}\{(y - \hat{\boldsymbol{y}}_1)^2\} \\ &= B^2\frac{V_{u_1}V_y - C_1^2}{A^2V_y + B^2V_{u_1} + 2ABc_1} \end{aligned} \tag{7}$$

Note that if $B = 0$ then the mean squared error of prediction after one latent variable is zero. According to (6), B depends both on the interference between the spectral responses $\mathbf{v}_0$ and $\mathbf{v}_1$ and on the correlation between the random variables y and u_1. Thus B is equal to zero when u_1 is uncorrelated with y and has an orthogonal non-interfering response, or when

$$\frac{\mathrm{Cov}(y, u_1)}{\mathrm{Var}(y)} = -\frac{\mathbf{v}_0\cdot\mathbf{v}_1}{\|\mathbf{v}_1\|^2}$$

This analysis clarifies the importance of the statistical correlation of the substance y with other components in the system. It shows that in general the optimal number of components needed for PLS calibration is not necessarily equal to the rank of the spectral data matrix [11]. In our example, when $B = 0$, PLS with only one latent variable is optimal even though the spectral matrix has a rank of two.

Before proceeding to the next iterative step in the algorithm, we consider the following two examples.

Example 1

Consider the case in which $\mathbf{v}_0\cdot\mathbf{v}_1 = 0$ and $c_1 = 0$. Then, according to (6), $B = 0$ and thus $t_1 = V_y\|\mathbf{v}_0\|^2y$ is a constant multiple of y. The normalization factor $q_1 = 1/V_y\|\mathbf{v}_0\|^2$, so that indeed $\hat{\boldsymbol{y}}_1 = q_1t_1 = y$, yielding a zero prediction error after only one latent variable, in accordance with (7). As already noted by Kvalheim and Karstang [17], this result also extends to the more general model (1) with an arbitrary number of components, as long as they are all uncorrelated with y and all their responses are orthogonal to $\mathbf{v}_0$. This, however, is a very idealized case seldom if ever encountered in practice.

Example 2

Consider now the more common case in which u_1 is not correlated with y, i.e. $c_1 = 0$ but $\mathbf{v}_0\cdot\mathbf{v}_1 \neq 0$. Now $B \neq 0$ and the first latent variable t_1 contains a contribution from u_1:

$$t_1 = V_y\|\mathbf{v}_0\|^2y + V_y(\mathbf{v}_0\cdot\mathbf{v}_1)u_1$$

In this case

$$\begin{aligned} q_1 = \frac{\mathbb{E}\{t_1y\}}{V(t_1)} &= \frac{V_y^2\|\mathbf{v}_0\|^2}{V_y^3\|\mathbf{v}_0\|^4 + V_y^2(\mathbf{v}_0\cdot\mathbf{v}_1)^2V_{u_1}} \\ &= \frac{\|\mathbf{v}_0\|^2}{V_y\|\mathbf{v}_0\|^4 + (\mathbf{v}_0\cdot\mathbf{v}_1)^2V_{u_1}} \end{aligned}$$

and obviously $\hat{y}_1 \neq y$. The mean squared error of prediction with one latent variable is

$$\begin{aligned} \mathrm{MSEP} &= \mathbb{E}\{(y - q_1 t_1)^2\} \\ &= V_y \frac{(\mathbf{v}_0 \cdot \mathbf{v}_1)^2 V_{u_1}}{V_y \|\mathbf{v}_0\|^4 + (\mathbf{v}_0 \cdot \mathbf{v}_1)^2 V_{u_1}} \end{aligned} \tag{8}$$

The mean squared error is obviously less than the normalization factor V_y, which is the mean squared error obtained by the trivial (and uninformative) prediction $\hat{y} = 0$, which corresponds to predicting $\hat{y} = \bar{y}$ before mean centering. As seen from (8), the prediction error of PLS with only one variable depends both on the variance of u_1 and on the physical interference between the characteristic spectrum of u_1 and that of y. For example, if u_1 has a much larger variance than y and a non-negligible interference of its response with that of y, then prediction with only one latent variable will be extremely poor. This phenomenon of an unrelated component with a much larger spectral variation than that of the substance of interest occurs rather frequently in spectroscopy. Three physical examples are a random baseline shift, scattering effects and particle size effects. Equation (8) shows the importance of detection and removal of such large effects in the spectral data prior to the application of the PLS algorithm, thus providing a theoretical motivation for various preprocessing algorithms with this goal in mind, such as MSC and SNV [1] or the more recent orthogonal signal correction [21]. From our analysis it is also clear that removal of such components should lead to a decrease in the optimal number of latent variables needed for calibration, as indeed is observed in simulations and in real data sets.

4.2. PLS with two substances and two latent variables

We now proceed with the computation of the loading vector $\mathbf{p}_1$ and the second latent variable t_2. For simplicity we compute this loading vector explicitly only for the second example in which $c_1 = 0$ but $\mathbf{v}_0 \cdot \mathbf{v}_1 \neq 0$.

Since the projection $\mathbf{w}_1$ is defined up to a normalization constant, we choose $\mathbf{w}_1 = \mathbf{v}_0 / \|\mathbf{v}_0\|^2$ and denote $\eta = \mathbf{v}_0 \cdot \mathbf{v}_1 / \|\mathbf{v}_0\|^2$. With these definitions the first latent variable and its variance simplify to

$$t_1 = y + \eta u_1, \qquad \mathrm{Var}(t_1) = V_y + \eta^2 V_{u_1}$$

Therefore

$$\begin{aligned} \mathbf{p}_1 &= \frac{\mathbb{E}\{\mathbf{x} t_1\}}{\mathrm{Var}(t_1)} = \frac{\mathbb{E}\{(y\mathbf{v}_0 + u_1 \mathbf{v}_1)(y + \eta u_1)\}}{1 + \eta^2} \\ &= \frac{1}{V_y + \eta^2 V_{u_1}} (V_y \mathbf{v}_0 + \eta V_{\mathbf{v}_1} \mathbf{v}_1) \end{aligned} \tag{9}$$

The data $\mathbf{x}_1$ are computed by subtraction of the product $t_1 \mathbf{p}_1$ from the original (mean-centered) data $\mathbf{x}$:

$$\begin{aligned} \mathbf{x}_1 &= \mathbf{x} - t_1 \mathbf{p}_1 \\ &= \frac{1}{V_y + \eta^2 V_{u_1}} (V_y V u_1 - \eta V_{u_1} y)(\mathbf{v}_1 - \eta \mathbf{v}_0) \end{aligned}$$

Under our assumptions, $\mathbf{v}_1 \neq \eta \mathbf{v}_0$, so that $\mathbf{x}_1 \neq 0$. The second projection computed by PLS is

$$\mathbf{w}_2 = \mathbb{E}\{y \mathbf{x}_1\} = \frac{V_{u_1} V_y}{V_y + \eta^2 V_{u_1}} (-\eta)(\mathbf{v}_1 - \eta \mathbf{v}_0)$$

Since this projection is defined up to a normalization constant, we choose to take

$$\mathbf{w}_2 = \frac{\mathbf{v}_1 - \eta \mathbf{v}_0}{\|\mathbf{v}_1\|^2 - \eta^2 \|\mathbf{v}_0\|^2} (V_y + \eta^2 V_{u_1}) \tag{10}$$

Thus the second latent variable is $t_2 = \mathbf{x}_1 \cdot \mathbf{w}_2 = V_y u_1 - \eta V_{u_1} y$ and its corresponding score q_2 is given by

$$q_2 = \frac{\mathbb{E}\{y t_2\}}{\mathrm{Var}(t_2)} = -\frac{\eta}{\eta^2 V_{u_1} + V_y}$$

Indeed, $\hat{y} = q_1 t_1 + q_2 t_2 = y$, so that, as expected, PLS with two latent variables yields a zero prediction error, since the model (3) contains two components and no noise.

In terms of the original input $\mathbf{x}$, the PLS regression is

$$\begin{aligned} \hat{y} &= \frac{1}{\|\mathbf{v}_0\|^2 - \frac{(\mathbf{v}_0 \cdot \mathbf{v}_1)}{\|\mathbf{v}_1\|^2}} \left(\mathbf{v}_0 - \frac{\mathbf{v}_0 \cdot \mathbf{v}_1}{\|\mathbf{v}_1\|^2} \mathbf{v}_1 \right) \cdot \mathbf{x} \\ &= \mathbf{r} \cdot \mathbf{x} \end{aligned} \tag{11}$$

where the regression vector $\mathbf{r}$ is, up to normalization, equal to the NAS $\mathbf{v}_0 - \mathbf{v}_0 \cdot \mathbf{v}_1 / \|\mathbf{v}_1\|^2 \mathbf{v}_1$. Thus, in the absence of noise and when u_1 is not correlated with y, the PLS algorithm computes the net analyte signal as its regression vector without explicit knowledge of the response vectors $\mathbf{v}_0$ and $\mathbf{v}_1$. This result is not specific for the case of two substances or the case of no correlation between them. In the next subsection we prove that in a general noise-free setting with an arbitrary number of components the regression vector computed by PLS is equal to the net analyte signal up to a normalization constant. Obviously, in this ideal setting this is also the regression vector computed by PCR and other multivariate algorithms [28].

4.3. The number of latent variables

In the PLS literature it is often stated that the total number of latent variables should be at least equal to the total number of expected factors in the measured spectrum [1,3]. In this subsection we formalize this statement in the context of the probabilistic input model (1). We thus assume that each spectral data point depends on the substance of interest in prediction, y, and on an additional set of k random variables, $u_1, \ldots, u_k$, through relation (1). We assume that $\mathbf{v}_0 \notin \mathrm{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ and that the $(k+1) \times (k+1)$ covariance matrix C of y and $u_1, \ldots, u_k$ is of full rank. Otherwise, as in the case of closures, at least one random variable is linearly dependent on the others and an equivalent model with a smaller number of random variables can be formulated. Note that the last assumption implies that we have a calibration set of at least $k+1$ data points. Similarly we assume that all vectors $\mathbf{v}_i$ are linearly independent. Otherwise an equivalent reduced model with fewer random variables can be defined. We can now prove the following theorem.

Theorem 1

Under the above conditions, after at most $k+1$ latent variables, the error produced by the PLS algorithm is zero.

Proof

According to Reference [24], at each step of the PLS algorithm a new score t is computed such that it is uncorrelated with all previous scores and such that it has maximal correlation with y. Under the assumptions of our probabilistic model, each such score is a linear combination of all the substances, i.e.

$$t_i = \alpha_0^{(i)} y + \sum_{j=1}^{k} \alpha_j^{(i)} u_j$$

where the coefficients $\alpha_j^{(i)}$ depend on the correlations between the various random variables as well as on the dot products of their characteristic spectral responses. The condition of no correlation between t_i and all other latent variables t_j means that the vectors $\boldsymbol{\alpha}^{(i)} = (\alpha_0^{(i)}, \ldots, \alpha_k^{(i)})$ are orthogonal under the quadratic form C, since

$$\mathbb{E}\{t_i t_j\} = \boldsymbol{\alpha}^{(i)} C (\boldsymbol{\alpha}^{(j)})' = 0$$

where $\boldsymbol{\alpha}'$ is the transpose of the row vector $\boldsymbol{\alpha}$. Since we assumed that C is a full rank matrix, it follows that the vectors $\boldsymbol{\alpha}^{(i)}$ are independent in $\mathbb{R}^{k+1}$. Therefore, with at most $k+1$ such vectors, it is possible to find a linear combination of them that gives the first variable y. Once this is achieved, the error in the prediction of y is zero and the PLS algorithm stops. □

The proof of this theorem, as well as the analysis of PLS with two latent variables presented in the previous subsection, also clarifies some points concerning the interpretation of the latent variables and of the projection vectors $\mathbf{w}_i$. According to this statistical analysis, the latent variables t_i are all linear combinations of the original components, and the projection vectors $\mathbf{w}_j$ are linear combinations of the spectral responses of these components. The exact coefficients for all these combinations are in general complex functions of the *statistical correlations* between the different components and of the *physical interferences* between their corresponding spectral responses. Therefore, as is well known empirically, for complex systems with many possibly correlated components, chemical interpretation of these vectors and latent variables is usually not feasible. Moreover, as proven in the next subsection, although different finite data sets with different correlations between the various components in them would yield different intermediate projections, their end result remains the same, an appropriately scaled version of the NAS.

4.4. PLS = NAS

Theorem 2

Under the above conditions, in the absence of noise, the regression vector computed by PLS corresponding to a zero prediction error is a constant times the net analyte signal.

Proof

According to the previous theorem, after at most $k+1$ steps, PLS achieves a zero prediction error. Assume that PLS achieves this zero prediction error after l steps. The corresponding predictor is thus given by

$$\hat{y} = \sum_{i=1}^{l} q_i t_i$$

The coefficients t_i are all given by projections $\tilde{\mathbf{w}}_i$ of the original data. The projections $\tilde{\mathbf{w}}_i$ are in general different from the original projections $\tilde{\mathbf{w}}_i$, because at each iteration of the PLS algorithm we subtract the quantity $t_a \mathbf{p}_a$ from the current data $\mathbf{x}_{a-1}$. Thus

$$\hat{y} = \mathbf{x} \cdot \left(\sum_{i=1}^{l} q_i \tilde{\mathbf{w}}_i \right)$$

and the regression vector $\mathbf{r}$ is given by

$$\mathbf{r} = \sum_{i=1}^{l} q_i \tilde{\mathbf{w}}_i$$

By construction, each of these projections $\tilde{\mathbf{w}}_i$ is a linear combination of the original vectors $\mathbf{v}_i$. Therefore there exist coefficients $\beta_0, \ldots, \beta_k$, such that

$$\mathbf{r} = \beta_0 \mathbf{v}_0 + \sum_{i=1}^{k} \beta_i \mathbf{v}_i$$

Since $\hat{y} = y$ (perfect prediction with zero error), $\beta_0 \neq 0$ and the vector $\mathbf{r}$ must be orthogonal to all the other vectors $\{\mathbf{v}_i\}_{i \geq 1}$. Therefore, by definition, the vector $\mathbf{r}$ is equal to the net analyte signal up to a normalization constant. □

5. PLS IN THE PRESENCE OF NOISE

In this section we relax the idealized assumption of error-free data and consider the effect of noise in the spectral data on the PLS algorithm. We restrict our analysis to the population setting with an infinite training set. These results can therefore be viewed as the asymptotic limit of PLS on noisy data as the size of the training set $n \to \infty$. The analysis of PLS with a finite and noisy calibration set is considered in Reference [14].

We first consider a system with a single component, for which we assume input data of the form

$$\tilde{\mathbf{x}} = y \mathbf{v}_0 + \sigma \boldsymbol{\xi} \tag{12}$$

where $\boldsymbol{\xi}$ is a random noise vector in $\mathbb{R}^p$ whose p co-ordinates are all independent identically distributed random variables, uncorrelated with the substance y and normally distributed with zero mean and unit variance, and σ is a measure of noise strength.

We follow the steps of the PLS algorithm. In the infinite population setting, the best projection $\mathbf{w}_1$ is unaffected by noise and in this case given by $\mathbf{w}_1 = \mathbb{E}\{y\tilde{\mathbf{x}}\} = V_y \mathbf{v}_0$. The corresponding latent variable t_1, however, does contain a noise contribution:

$$t_1 = \mathbf{w}_1 \cdot \tilde{\mathbf{x}} = V_y (\|\mathbf{v}_0\|^2 y + \sigma \boldsymbol{\xi} \cdot \mathbf{v}_0)$$

A simple computation shows that

$$q_1 = \frac{\mathbb{E}\{y t_1\}}{\mathrm{Var}(t_1)} = \frac{1}{V_y \|\mathbf{v}_0\|^2 + \sigma^2}$$

Therefore the PLS predictor is

$$\hat{y} = q_1 t_1 = \frac{V_y}{V_y \|v_0\|^2 + \sigma^2} (\|v_0\|^2 y + \sigma \boldsymbol{\xi} \cdot v_0)$$

which corresponds to a shrunk version of the noise free regression vector

$$\mathbf{r} = \frac{V_y \|\mathbf{v}_0\|^2}{V_y \|\mathbf{v}_0\|^2 + \sigma^2} \frac{\mathbf{v}_0}{\|\mathbf{v}_0\|^2} \tag{13}$$

This *shrinkage* of the regression vector leads to a prediction bias towards the mean

$$|\mathbb{E}\{\hat{y} \,|\, y\}| = \frac{V_y}{V_y + \sigma^2} |y| < |y|$$

with a corresponding mean squared error of prediction

$$\text{MSEP} = \frac{\sigma^2}{\|\mathbf{v}_0\|^2} \frac{1}{1 + \dfrac{\sigma^2}{V_y \|\mathbf{v}\|^2}}$$

This shrinkage, due to the fact that PLS is an inverse calibration procedure, has been studied extensively by various authors [1,7,29,30]. An interesting property in the context of the assumed model (12) for the spectra is that with an infinite training set the regression vector (13) is *optimal* under a mean squared error of prediction criterion. As we now show, this optimality is true in the more general case of (2) with an arbitrary number of components.

Theorem 3

For an infinite training dataset sampled according to (1) and (2), PLS regression is optimal under a mean squared error of prediction criterion.

Proof

For simplicity we prove the theorem for the case in which for error-free samples PLS requires $k+1$ latent variables for a zero prediction error. This implies in particular that the vectors $\{\mathbf{v}_j\}_{j=0}^{k}$ are linearly independent. Let $\mathbf{r}_{\text{opt}}$ be the optimal regression vector that minimizes the mean squared error of prediction

$$\min_{\mathbf{r} \in \mathbb{R}^p} \mathbb{E}\{(\hat{y} - y)^2\} \tag{14}$$

where $\hat{y} = \mathbf{x} \cdot \mathbf{r}$, and let $\mathbf{r}_{\text{PLS}}$ be the regression vector computed by PLS.

To find the optimal regression vector $\mathbf{r}_{\text{opt}}$, we decompose it as

$$\mathbf{r}_{\text{opt}} = \sum_{j=0}^{k} \alpha_j \mathbf{v}_j + \sum_{j=1}^{p-k-1} \beta_j \mathbf{v}_j^{\perp} \tag{15}$$

where the set $\{\mathbf{v}_j^{\perp}\}_{j=1}^{p-k-1}$ is an orthogonal completion of $\{\mathbf{v}_j\}_{j=1}^{k}$ to a basis of $\mathbb{R}^p$. Inserting (15) into (14) and taking partial derivatives with respect to α_j and β_j, we obtain a linear system whose solution gives $\beta_j = 0$ for all j. Therefore $\mathbf{r}_{\text{opt}} \in \text{Span}\{\mathbf{v}_j\}_{j=0}^{k}$. We now consider the regression vector computed by PLS. By construction,

$$\hat{y}_{\text{PLS}} = \sum_{i=0}^{k} q_i t_i$$

where $t_i = \mathbf{x} \cdot \tilde{\mathbf{w}}_i$, and thus

$$\mathbf{r}_{\text{PLS}} = \sum_i q_i \tilde{\mathbf{w}}_i$$

A careful examination of the steps in PLS shows that, in the case of unstructured noise, both $\{\mathbf{w}_i\}$ and $\{\mathbf{p}_i\}$ belong to $\text{Span}\{\mathbf{v}_j\}_{j=0}^{k}$. In addition, each $\tilde{\mathbf{w}}_i$ can be written as

$$\tilde{\mathbf{w}}_j = \mathbf{w}_i - \sum_{j<i} a_{i,j} \mathbf{w}_j$$

for some coefficients $a_{i,j}$. Since $\{\mathbf{w}_i\}$ are orthogonal [5], it follows that $\{\tilde{\mathbf{w}}_i\}$ are linearly independent and therefore form a basis for $\text{Span}\{\mathbf{v}_j\}_{j=0}^{k}$. The regression vector computed by PLS is simply the result of an ordinary least squares regression on these projections. This is equivalent to minimizing (14) under the restriction that $\mathbf{r}_{\text{PLS}} \in \text{Span}\{\tilde{\mathbf{w}}_j\}_{j=0}^{k}$. However, since the result is independent of the choice of a basis and depends only on the underlying vector space, and since $\text{Span}\{\tilde{\mathbf{w}}_j\}_{j=0}^{k} = \text{Span}\{\mathbf{v}_j\}_{j=0}^{k}$, it follows that $r\{\text{PLS}\} = r\{\text{opt}\}$.

We remark that PLS is not the only procedure that is asymptotically optimal. A similar proof shows that PCR is also asymptotically optimal in the presence of noise under the same conditions. Finally, note that in the presence of noise the optimal regression vector is not purely proportional to the NAS. Rather, it contains small perturbations of the order of σ^2 in the directions of all the other components in the system, with coefficients that depend on their statistical correlations and physical interferences. Therefore preprocessing algorithms that attempt to remove variability in the spectrum that is uncorrelated with y are in general asymptotically suboptimal for prediction purposes. They may, however, yield better and more robust predictions in the case of small finite training sets, since these methods require in general a smaller number of latent variables.

6. EXAMPLES

6.1. Three components with closure

In the paper by Haaland and Thomas [11] the following example of a system with three components is presented. The noise-free spectral signal is given by

$$\mathbf{x} = y\mathbf{v}_0 + u_1\mathbf{v}_1 + u_2\mathbf{v}_2 \tag{16}$$

where the variables y, u_1 and u_2 satisfy the closure constraint

$$y + u_1 + u_2 = 1 \tag{17}$$

From a calibration set of 16 noise-free samples the first projection vector $\mathbf{w}_1$ as well as the resulting regression vector found by PLS are calculated numerically and a discussion of the interpretation of these two vectors is given.

In this subsection we show how these two quantities can be computed analytically by our analysis and discuss some of the implications of these results for interpretation purposes. From the graphs presented in Reference [11], we estimate the unspecified response vectors by

$$\mathbf{v}_0 = 0.7\exp\left[-\left(\frac{t-15}{6}\right)^2\right] + 2\exp\left[-\left(\frac{t-55}{4}\right)^2\right]$$

$$\mathbf{v}_1 = 2.35\exp\left[-\left(\frac{t-45}{5}\right)^2\right] + 1.4\exp\left[-\left(\frac{t-82}{6}\right)^2\right]$$

$$\mathbf{v}_2 = 1.4\exp\left[-\left(\frac{t-51}{5}\right)^2\right] + \exp\left[-\left(\frac{t-78}{5}\right)^2\right]$$

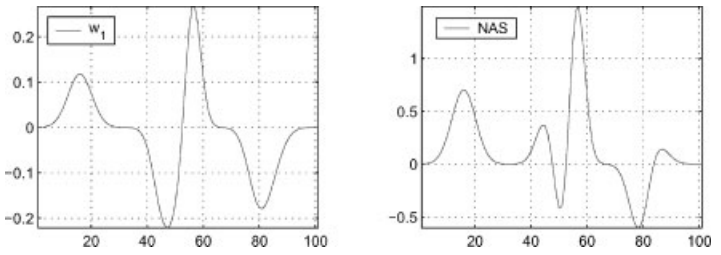

Figure 1. The first projection vector $\mathbf{w}_1$ (left) and the net analyte signal (right) corresponding to the example of Reference [11].

Combining Equations (16) and (17), one can equivalently write the signal as

$$\mathbf{x} = \mathbf{v}_2 + y(\mathbf{v}_0 - \mathbf{v}_2) + u_1(\mathbf{v}_1 - \mathbf{v}_2)$$

From the 16 samples given in the simulated example of Reference [11], we have that $\mathbb{E}\{y\} = \mathbb{E}\{u_1\} = 1/3$. Therefore, after mean centering, we obtain a signal with only two components, i.e.

$$\mathbf{x} = (y - 1/3)(\mathbf{v}_0 - \mathbf{v}_2) + (u_1 - 1/3)(\mathbf{v}_1 - \mathbf{v}_2)$$

In addition, $\mathrm{Var}(y) = 14/900 = 0.0155$ and $\mathrm{Cov}((y - 1/3), (u_1 - 1/3)) = 0.1038$. Therefore, according to (4), the first projection is

$$\begin{aligned}\mathbf{w}_1 &= \mathrm{Var}(y)(\mathbf{v}_0 - \mathbf{v}_2) \\ &\quad + \mathrm{Cov}((y - 1/3), (u_1 - 1/3))(\mathbf{v}_1 - \mathbf{v}_2) \\ &= 0.0155\mathbf{v}_0 + 0.1038\mathbf{v}_1 - 0.119\mathbf{v}_2\end{aligned}$$

Note that this vector is not equal to $\mathbf{v}_0$ owing to the correlations between the different components in the system. This vector is plotted in Figure 1 (left) and indeed resembles the one computed numerically in Reference [11] by the PLS algorithm (their Figure 3-B). According to our analysis, the regression coefficient found by PLS should be equal (up to normalization) to the net analyte signal. In Figure 1 (right) we have plotted the net analyte signal corresponding to $\mathbf{v}_0$. This vector indeed resembles in shape the one found computationally in Reference [11] (their Figure 3-C). The differences between the two vectors are due to our approximation of the unspecified response vectors.

Another few remarks about the simulation data of Reference [11] are relevant. First of all, as an outcome of our analysis, we see that PLS with two latent variables will give a zero prediction error even though there are three components in the spectral data. As noted by Haaland and Thomas [11], this is due to the constraint (17), which mathematically yields a new model with only two random variables. Addition of a random baseline and a random drift means mathematically addition of two more components (random variables). Therefore, according to our analysis, since these four response vectors are linearly independent, the noise-free data would require four latent variables to achieve a zero prediction error, as is indeed observed in their simulations.

Another important point concerns calibration design, i.e. how should calibration samples be chosen for an optimal prediction. Our analysis shows that in the noise-free case the exact population of the calibration set (with possible different correlations between the random variables) is unimportant as long as all possible different spectral components are present in the calibration set, because, by the end of the day, the regression vector computed by the PLS algorithm is the net analyte signal times a multiplicative constant. In the ideal noise-free case this result is *independent* of the training set size and the exact correlations amongst the different random variables present in the training set. An answer to this question in the case of noisy measurements is the subject of future research.

6.2. A three-component system

As a second example we consider a system with three independent components y, u_1 and u_2 and corresponding spectra $\mathbf{v}_0$, $\mathbf{v}_1$ and $\mathbf{v}_2$, shown in Figure 2 (left). All three spectra are sums of Gaussians with different centers, widths and amplitudes. In this example, only the substance u_1 has a spectrum interfering with that of y. However, PLS requires three latent variables to achieve a zero prediction error, with the resulting regression vector shown in Figure 2 (right). The intuitive explanation for this result is as follows. To compute the value of y from the signal $\mathbf{x}$, we need to take out the contribution of the only interfering substance u_1, for example by estimating the value of u_1 and its spectral response $\mathbf{v}_1$. However, substance u_2 is interfering with u_1, and this interference (with u_1!) also needs to be taken into

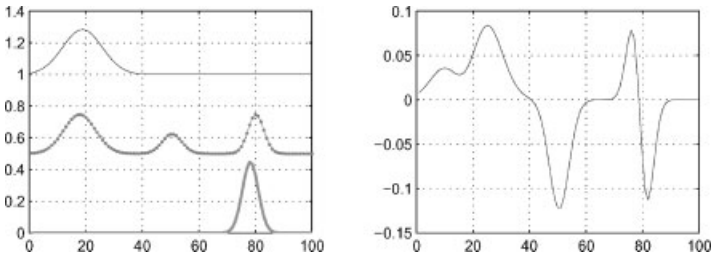

Figure 2. The three spectra of the three components (left), from top to bottom $\mathbf{v}_0$, $\mathbf{v}_1$ and $\mathbf{v}_2$ respectively, and the corresponding net analyte signal (right).

account. Therefore, even though u_2 does not directly interfere with y, PLS requires three latent variables for a zero prediction error.

Another point shown in this figure is that the final regression vector can have a region of non-vanishing coefficients where the original substance does not have any signal response at all [31]. Therefore wavelength selection schemes that retain regions where the regression vector is large are not necessarily optimal, as pointed out also in Reference [32]. In this specific example, retaining only the first 60 wavelengths gives a model with only two latent variables instead of three. In the case of finite and noisy calibration data, depending on the noise level and on the number of calibration samples, taking only these wavelengths could possibly yield improved overall performance of the final regression model.

6.3. A noisy two-component system

Consider an infinite training set with noisy data according to

$$\tilde{\mathbf{x}} = y\mathbf{v}_0 + u_1\mathbf{v}_1 + \sigma\boldsymbol{\xi}$$

For simplicity we analytically analyze the optimal regression in the case where $\|\mathbf{v}_0\| = \|\mathbf{v}_1\| = V_y = V_{u_1} = 1$, with $c_1 = 0$, so there is no correlation between y and u_1. In this case the optimal regression vector (also computed by PLS) is

$$\mathbf{r}_{\text{opt}} = \frac{1}{(1+\sigma^2)^2 - \eta^2}[(\mathbf{v}_0 - \eta\mathbf{v}_1) + \sigma^2\mathbf{v}_0]$$

where $\eta = \mathbf{v}_0 \cdot \mathbf{v}_1$, so that $\mathbf{v}_0 - \eta\mathbf{v}_1$ is the net analyte signal vector and the corresponding optimal mean square error of prediction is

$$\begin{aligned}\text{MSEP}_{\text{opt}} &= \sigma^2 \frac{1+\sigma^2}{(1+\sigma^2)^2 - \eta^2} \\ &= \frac{\sigma^2}{1-\eta^2} \frac{1+\sigma^2}{1 + \dfrac{2\sigma^2 + \sigma^4}{1-\eta^2}}\end{aligned} \tag{18}$$

Note that in the presence of noise the optimal regression vector is *not* equal to the net analyte signal. Rather, it is a shrunk version of it added by a small amount of the pure spectral response $\mathbf{v}_0$. The larger the noise, the more the regression vector is tilted in the direction of the pure spectrum $\mathbf{v}_0$, as also seen in the simulations of Brown [22].

Instead of PLS, we now consider a regression model built by HLA [12], one of many methods that first attempt to remove the effects of u_1 from the noisy spectrum. In this case the resulting regression vector constructed from an infinite training set is simply the net analyte signal vector, with a mean squared error of prediction given by

$$\text{MSEP}_{\text{HLA}} = \frac{\sigma^2}{1-\eta^2} \tag{19}$$

Comparing (19) and (18), we see that if σ and η are small then both methods have a similar performance. The advantage of PLS comes about when either the noise σ is large or the interference between the spectral responses $\mathbf{v}_0$ and $\mathbf{v}_1$ is large, leading to an η close to unity.

7. SUMMARY

In this paper we have presented a mathematical and statistical analysis of PLS under a specific probabilistic model of the input data based on Beer's law. Our analysis provides a theoretical verification of some empirically well-known and observed properties of PLS. It shows that in the error-free case the output of PLS is a scaled version of the net analyte signal vector. In the case of an infinite but noisy training set, while PLS is optimal under a mean squared error criterion, the resulting regression vector is not purely proportional to the net analyte signal, but rather depends in general on all the components and spectral responses in the system.

Our analysis is limited to the cases of either a finite error-free setting or a noisy but infinite population setting. While many simulations have studied the effects of various parameters on PLS and other competing algorithms in the presence of a finite and noisy training set, a theoretical statistical analysis in this case is still an open research problem [14].

Acknowledgements

The authors would like to thank Frederick Warner for many interesting discussions.

REFERENCES

1. Martens H, Naes T. *Multivariate Calibration*. Wiley: Chichester, 1989.
2. Naes T, Isaksson T, Fearn T, Davies T. *User-friendly Guide to Multivariate Calibration and Classification*. NIR Publications, Chichester, 2002.
3. Wold S, Sjostrom M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemometrics Intell. Lab. Syst.* 2001; **58**: 109–130.
4. Wold S. Discussion. *Technometrics* 1993; **35**: 149–156.
5. Hoskuldsson A. PLS regression methods. *J. Chemometrics* 1988; **2**: 211–228.
6. Helland IS. Some theoretical aspects of partial least squares regression. *Chemometrics Intell. Lab. Syst.* 2001; **58**: 97–107.
7. Frank I, Friedman JH. A statistical view of some chemometrics regression tools. *Technometrics* 1993; **35**: 109–148.
8. Helland IS, Almoy T. Comparison of prediction methods when only a few components are relevant. *J. Am. Statist. Assoc.* 1994; **89**: 583–591.
9. Garthwaite PH. An Interpretation of partial least squares. *J. Am. Statist. Assoc.* 1994; **89**: 122–127.
10. Thomas E, Haaland D. Comparison of multivariate calibration methods for quantitative spectral analysis. *Anal. Chem.* 1990; **62**: 1091–1099.
11. Haaland D, Thomas E. Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Anal. Chem.* 1988; **60**: 1193–1202.
12. Berger A, Koo TW, Itzkan I, Feld MS. An enhanced algorithm for linear multivariate calibration. *Anal. Chem.* 1998; **70**: 623–627.
13. Xu L, Schechter I. A calibration method free of optimum factor number selection for automated multivariate analysis. Experimental and theoretical study. *Anal. Chem.* 1997; **69**: 3722–3730.
14. Nadler B, Coifman RR. An exact asymptotic formula for the error in CLS and in PLS: the importance of feature selection in multivariate calibration. *J. Chemometrics* (submitted).
15. Lorber A, Faber K, Kowalski BR. Net analyte signal calculation in multivariate calibration. *Anal. Chem.* 1997; **69**: 1620–1626.

16. Lorber A. Error propagation and figures of merit for quantification by solving matrix equations. *Anal. Chem.* 1986; **58**: 1167–1172.
17. Kvalheim OM, Karstang TV. Interpretation of latent-variable regression models. *Chemometrics Intell. Lab. Syst.* 1989; **7**: 39–51.
18. Seasholtz MB, Kowalski BR. Qualitative information from multivariate calibration models. *Appl. Spectrosc.* 1990; **44**: 1337–1348.
19. Geladi P, MacDougall D, Martens H. Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Appl. Spectrosc.* 1985; **39**: 491–500.
20. Barnes RJ, Dhanoa MS, Lister SJ. Standard normal variate transformation and de-trending of near-infrared difiuse reflectance spectra. *Appl. Spectrosc.* 1989; **43**: 772–777.
21. Wold S, Antti H, Lindgren F, Ohman J. Orthogonal signal correction of near-infrared spectra. *Chemometrics Intell. Lab. Syst.* 1998; **44**: 175–185.
22. Brown CD. Discordance between net analyte signal theory and practical multivariate calibration. *Anal. Chem.* 2004; **76**: 4364–4373.
23. Helland IS. Partial least squares regression and statistical models. *Scand. J. Statist.* 1990; **17**: 97–114.
24. Gustafsson M. A probabilistic derivation of the partial least-squares algorithm. *J. Chem. Info. Comput. Sci.* 2001; **41**: 288–294.
25. Goicoechea H, Olivieri A. Enhanced synchronous spectrofluorometric determination of tetracycline in blood serum by chemometric analysis. Comparison of partial least-squares and hybrid linear analysis calibrations. *Anal. Chem.* 1999; **71**: 4361–4368.
26. Goicoechea H, Olivieri A. Wavelength selection by net analyte signals calculated with multivariate factor-based hybrid linear analysis (HLA). A theoretical and experimental comparison with partial least-squares (PLS). *Analyst* 1999; **124**: 725–731.
27. Coelho CJ, Harrop Galvo RK, Araujo M, Pimentel MF, Cirino da Silva E. A solution to the wavelet transform optimization problem in multicomponent analysis. *Chemometrics Intell. Lab. Syst.* 2003; **66**: 205–217.
28. Bro R, Andersen C. Theory of net analyte signal vectors in inverse regression. *J. Chemometrics* 2003; **17**: 646–652.
29. Butler NA, Denham MC. The peculiar shrinkage properties of partial least squares regression. *J. R. Statist. Soc. B* 2000; **62**: 585–593.
30. De Jong S. PLS shrinks. *J. Chemometrics* 1995; **9**: 323–326.
31. Lorber A, Kowalski BR. The effect of interferences and calibration design on accuracy: implications for sensor and sample selection. *J. Chemometrics* 1988; **2**: 67–79.
32. Brenchley JM, Horchner U, Kalivas JH. Wavelength selection characterization from NIR spectra. *Appl. Spectrosc.* 1997; **51**: 689–699.

