JOURNAL OF CHEMOMETRICS
J. Chemometrics 2005; **19**: 107–118
Published online 22 September 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cem.915

The prediction error in CLS and PLS: the importance of feature selection prior to multivariate calibration

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

Received 18 September 2004; Revised 25 June 2005; Accepted 25 June 2005

Classical least squares (CLS) and partial least squares (PLS) are two common multivariate regression algorithms in chemometrics. This paper presents an asymptotically exact mathematical analysis of the mean squared error of prediction of CLS and PLS under the linear mixture model commonly assumed in spectroscopy. For CLS regression with a very large calibration set the root mean squared error is approximately equal to the noise per wavelength divided by the length of the net analyte signal vector. It is shown, however, that for a finite training set with n samples in p dimensions there are additional error terms that depend on $\sigma^2 p^2/n^2$, where σ is the noise level per co-ordinate. Therefore in the 'large p—small n' regime, common in spectroscopy, these terms can be quite large and even dominate the overall prediction error. It is demonstrated both theoretically and by simulations that dimensional reduction of the input data via their compact representation with a few features, selected for example by adaptive wavelet compression, can substantially decrease these effects and recover the asymptotic error. This analysis provides a theoretical justification for the need to perform feature selection (dimensional reduction) of the input data prior to application of multivariate regression algorithms. Copyright © 2005 John Wiley & Sons, Ltd.

KEYWORDS: classical least squares; partial least squares; prediction error; dimensional reduction; feature selection

1. INTRODUCTION

Multivariate regression problems arise in the analysis of data in diverse applications. When the number of samples, n, is (much) larger than the number of regressors, p, and the corresponding matrices are well conditioned, standard methods such as ordinary least squares (OLS) can be typically applied. However, in many scientific fields, including chemometrics in general and spectroscopy in particular, the common situation is that the number of samples is much smaller than the number of variables ($n \ll p$), in which case ordinary least squares is indeterminate and thus inapplicable. The remarkable fact that predictions are possible even in this setting stems from the (sometimes hidden) property that, although the data are presented in a high-dimensional space, they actually have a much lower intrinsic dimensionality $d \ll n$. For example, in a spectroscopic measurement of a system with three components at 1000 different wavelengths, although the measured spectrum is represented in a 1000-dimensional space, it is typically assumed to be in a three-dimensional subspace (or at most a five-dimensional subspace if the measuring device adds a random baseline shift and a random slope to the signal).

In this setting, for which standard methods such as OLS fail, classical least squares (CLS) and partial least squares (PLS) are two common and very successful algorithms applied in practice [1–4]. These methods are sometimes viewed as performing dimensional reduction, since in both CLS and PLS the data are projected onto a few data-dependent directions and regression is performed in this lower-dimensional subspace. The two methods differ in the way this subspace is defined and in the regression method employed in it. CLS, also known as the K-matrix method, is a direct method that requires full knowledge of all components in the training samples of the measured system and is thus typically applicable only to very simple systems [3]. Recently, however, modifications of the algorithm to include unmodeled interferences have been suggested [5,6], thus possibly extending its applicability. PLS, on the other hand, is an indirect method that requires only knowledge of the concentration of the substance of interest, is thus more applicable than CLS and is the *de facto* standard calibration method in spectroscopy [4].

An important theoretical and practical question is what is the expected performance of these algorithms on future samples given calibration on a finite and noisy training set, and how does this performance compare with that of competing algorithms such as principal component regression (PCR) and ridge regression (RR)? In the chemometrics

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

literature this problem was tackled mainly by direct application of CLS, PLS and competing algorithms both on real data sets and on simulated data sets that follow a linear mixture model (see e.g. References [7–10]). In their seminal paper, Thomas and Haaland [9] investigated the effects of eight different parameters on the prediction error of CLS, PLS and PCR by extensive Monte Carlo simulation studies. Wentzell and Vega-Montoto [10] also made an extensive numerical comparison of PLS and PCR with simulated data containing many components. The main conclusion of these studies is that most algorithms have a similar performance, with each algorithm having its own regime of superiority so that no one algorithm is everywhere optimal. On the theoretical front, various works have attempted to estimate the prediction error for specific data sets using various approximations for error propagation [11–14], but no explicit formulae for the linear mixture case were given.

In the statistical literature the subject of multivariate calibration has been addressed in many works [15–20]. Much effort was put forth to elucidate the PLS algorithm from a statistical point of view [21–24], although a theory for the performance of PLS under the linear mixture model with a finite and noisy training set was not considered. In terms of theoretical formulae for the expected mean squared error of prediction, most attention has been devoted to the study of other multivariate regression algorithms such as the generalized least squares and best linear predictor algorithms and not of the more common CLS and PLS algorithms. In addition, most works consider only the case of more observations than variables, $n > p$, since these algorithms become indeterminate when $n < p$. To overcome this indeterminacy, minimal length regressors were proposed [18,19]. Theoretical work on the mean squared error of prediction was mainly done on the univariate case (only one component in one dimension), where both asymptotic and exact expressions for the root mean squared error of prediction (RMSEP) as well as confidence regions have been derived for various regressors [16,20,25].

Although both CLS and PLS perform a dimensional reduction, it is known empirically that an initial dimensional reduction of the input data prior to application of these algorithms is often very beneficial in practice. Most work on this type of feature selection prior to application of multivariate algorithms has focused on methods to optimally select a subset of the original variables (wavelength selection). Both Xu and Schechter [26] and Spiegelman *et al.* [27] gave a theoretical justification for wavelength selection based on an approximate analysis of the uncertainty error in the computation of the regression vector under a linear mixture model.

In this paper we extend these results and provide a mathematical analysis of the expected RMSEP for both CLS and PLS under the linear mixture model. For CLS we show that, although the asymptotic error for a very large training set is given by the noise level divided by the length of the net analyte signal vector [11,28], for a finite training set of n noisy samples there are additional correction terms of order $O(1/n)$, $O(1/n^2)$, etc. The interesting property we find is that, although the $1/n$ term is typically multiplied by an $O(1)$ coefficient, the $1/n^2$ term is multiplied by $\sigma^2 p^2$, where σ is the noise per co-ordinate and p is the dimensionality of the input data. Therefore in the 'large p—small n' regime, common in spectroscopy involving many more variables than samples, this correction term may actually dominate the overall error. From a statistical point of view these results are not surprising. In classification problems it is well known that the performance of standard classification algorithms such as Fisher's linear discriminant analysis is greatly degraded in the 'large p—small n' setting, since there appear correction terms of the form $\sigma^2 p/n$ [29,30]. Therefore our results can be viewed as the analogues of these well-known formulae to multivariate calibration problems.

Indeed, many papers in the chemometrics literature show empirically that an initial dimensional reduction prior to application of PLS, typically achieved in practice by wavelength selection, is quite beneficial in decreasing prediction errors. Our error analysis, showing that some error terms are of the form $\sigma^2 p^2/n^2$, provides the theoretical justification for this empirical finding, as also concluded by Spiegelman *et al.* [27] and Xu and Schechter [26]. However, while both these works (as well as many others) suggest wavelength selection as the method of choice to perform this initial dimensional reduction, in this paper we show mathematically that, for complex systems with many interfering components and lack of specificity at any single wavelength, wavelength selection methods have severe limitations and cannot in general achieve optimal prediction errors. In contrast, we propose to use adaptive wavelet feature selection algorithms [31,32] to perform this initial dimensional reduction, and present some simulation results that show their empirical success in achieving near-optimal prediction errors. Thus our analysis provides a justification and a better theoretical understanding of the role of wavelets as a tool for feature selection prior to multivariate calibration. A survey of the recent literature indeed reveals an increasing use of wavelets in the analysis of spectroscopic signals, with empirical reports that this use decreases (sometimes) the prediction errors of multivariate regression algorithms [33–37].

The paper is organized as follows. In Section 2 we define the probabilistic model of the input data and the multivariate calibration problem. The analysis of CLS and PLS under this model is described in Section 3. The issue of feature selection is described in Section 4. Section 5 presents numerical simulations that verify the results of our analysis. We conclude with a discussion and summary in Section 6. Mathematical proofs appear in the Appendix.

2. MULTIVARIATE CALIBRATION UNDER THE LINEAR MIXTURE MODEL

2.1. Notation

We denote vectors by boldface lowercase letters, e.g. $\mathbf{v}$, and matrices by bold capital letters, e.g. $\mathbf{C}$. 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, e.g. u_0 and u_1, while the mean of a random variable u is $E\{u\}$. Noisy estimates of noise-free quantities have a hat on top, e.g. $\hat{\mathbf{v}}$ and $\hat{u}$.

2.2. The linear mixture model

We consider the standard multivariate calibration problem in spectroscopy, namely the determination of analyte concentration from the absorbance spectrum of a complex multicomponent system, under the following probabilistic setting for the input data. We consider a system containing k different components, denoted $u_1, u_2, \ldots, u_k$, where each component u_j is a random variable with mean μ_j and unique spectral response vector $\mathbf{v}_j \in \mathbb{R}^p$. We denote by $\mathbf{C}^p$ the $k \times k$ (population) matrix of second moments of these random variables, with entries $C^p_{i,j} = E\{u_i u_j\}$. If all the averages $\mu_j = 0$, then $\mathbf{C}^p$ is the covariance matrix. Therefore, with some abuse of notation, we sometimes refer to $\mathbf{C}^p$ as the covariance matrix.

We assume that $\mathbf{C}^p$ is of full rank and that the vectors $\{\mathbf{v}_j\}_{j=1}^k$ are linearly independent in $\mathbb{R}^p$, as otherwise a reduced model with fewer random components can be formulated. Based on Beer's law, we further assume that the noise-free logarithm of the spectrum, denoted $\mathbf{x}$, is linearly related to the components via

$$\mathbf{x} = \sum_{j=1}^{k} u_j \mathbf{v}_j \tag{1}$$

whereas the measured spectrum is noisy and given by

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

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. We assume that u_1 is the substance of interest and, without loss of generality, scale all the other interfering components $u_2, \ldots, u_k$ so that their corresponding spectral responses have unit norm ($\|\mathbf{v}_j\| = 1$ for $j \geq 2$). This scaling has no effect on the final prediction of u_1.

The basic multivariate calibration problem can be cast as follows. Given a finite training set of n noisy samples, $\{\tilde{\mathbf{x}}_i, \mathbf{u}_i\}_{i=1}^n$, related via Equations (1) and (2), with $\mathbf{u}_i = (u_{i,1}, u_{i,2}, \ldots, u_{i,k})$ the vector of components for the ith sample, construct a regression fzunction $f\colon \mathbb{R}^p \to \mathbb{R}$ to accurately predict u_1 from future samples $\tilde{\mathbf{x}}$. Since we assume a linear relation between components and spectra, in this paper we focus on linear regressors of the form

$$\hat{u}_1 = f(\tilde{\mathbf{x}}) = \mathbf{r} \cdot \tilde{\mathbf{x}}$$

where $\mathbf{r}$ is the constructed regression vector. Note that in this paper we consider models without an intercept and therefore we do not mean center the data. As described below, mean centering, which is a preprocessing step typically employed in practice, does not qualitatively change our results.

Although this paper is written with a focus on chemometric applications, referring to $\tilde{\mathbf{x}}$ as the spectrum and u_i as the analyte concentrations, our analysis is general and thus applicable to any other data modeled by Equations (1) and (2). In the statistics literature the linear mixture model (1) is also known as the standard multivariate linear regression model [38], while problems in which the predictor variables are noisy as in Equation (2) are generally termed 'error-in-variables' (EIV) problems.

The model (1)–(2) has been used extensively as a benchmark in many simulation studies and in tests of new algorithms [9,10,26,39]. In this paper we present an asymptotic theory for the prediction error of both CLS and PLS on this model. For simple systems with a single component we obtain explicit formulae, asymptotically exact in the limit of small noise, for the expected mean squared error of prediction as a function of the number of training samples, n, the noise level σ and the dimension p of the signals. Although for complex multicomponent systems the explicit computation of the different constants is essentially algebraically intractable, the prediction error has similar qualitative features as in the case of a single-component system, where an explicit formula is available.

3. THE EXPECTED PREDICTION ERROR

3.1. Classical least squares

For the paper to be reasonably self-contained, we first briefly describe the steps in the classical least squares algorithm. Given a finite training set $\{\tilde{\mathbf{x}}_i, \mathbf{u}_i\}_{i=1}^n$, in CLS we first compute estimates $\{\hat{\mathbf{v}}_j\}$ for the (unknown) spectral responses $\{\mathbf{v}_j\}$ by least squares minimization:

$$\min_{\{\mathbf{v}_j\}} \sum_{i=1}^{n} \left\| \tilde{\mathbf{x}}_i - \sum_{j=1}^{k} u_{i,j} \mathbf{v}_j \right\|^2$$

The solution is

$$\begin{pmatrix} \hat{\mathbf{v}}_1 \\ \hat{\mathbf{v}}_2 \\ \vdots \\ \hat{\mathbf{v}}_k \end{pmatrix} = \mathbf{C}^{-1} \begin{pmatrix} E\{\tilde{\mathbf{x}} u_1\} \\ E\{\tilde{\mathbf{x}} u_2\} \\ \vdots \\ E\{\tilde{\mathbf{x}} u_k\} \end{pmatrix} \tag{3}$$

where $\mathbf{C}$ is the $k \times k$ matrix of second moments of the k components $u_1, \ldots, u_k$ in the training set, assumed to be of full rank.

We denote by $\hat{\mathbf{V}}$ the $k \times k$ matrix of spectral interferences, with entries $\hat{V}_{i,j} = \hat{\mathbf{v}}_i \cdot \hat{\mathbf{v}}_j$. Then the regression vectors computed by CLS for the k different components are given by

$$\begin{pmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \vdots \\ \mathbf{r}_k \end{pmatrix} = \hat{\mathbf{V}}^{-1} \begin{pmatrix} \hat{\mathbf{v}}_1 \\ \hat{\mathbf{v}}_2 \\ \vdots \\ \hat{\mathbf{v}}_k \end{pmatrix} \tag{4}$$

Finally, prediction of u_1 for new spectra $\tilde{\mathbf{x}}$ is given by

$$\hat{u}_1 = \tilde{\mathbf{x}} \cdot \mathbf{r}_1$$

The question considered in this paper is how well $\hat{u}_1$ approximates the unknown value u_1, and specifically what can be said about the mean squared error of prediction $E\{(\hat{u}_1 - u_1)^2\}$, when the regression vector $\mathbf{r}_1$ is constructed from a finite and noisy training set. Before considering the case of finite n, we first state the following well-known result about CLS regression as the number of training samples approaches infinity.

Theorem 1

As $n \to \infty$, the regression vector computed by CLS for the jth component is given by

$$\mathbf{r}_j = \frac{\mathbf{v}_j^{\perp}}{\|\mathbf{v}_j^{\perp}\|^2}$$

where $\mathbf{v}_j^{\perp}$ is the net analyte signal vector of the jth component [11]. The corresponding root mean squared error of prediction is given by

$$\mathrm{RMSEP}(\mathrm{CLS}, n=\infty) = \frac{\sigma}{\|\mathbf{v}_j^{\perp}\|} \tag{5}$$

A proof of this theorem appears in the Appendix. It shows that, as $n \to \infty$, CLS computes all spectral responses $\{\mathbf{v}_j\}$ without error, and by Equation (4) also computes an error-free net analyte signal vector. The prediction error is therefore due only to the noise in the new unseen spectral data, and for an unbiased estimator CLS yields the optimal prediction possible under a mean squared error criterion.

When the regression vector is computed from a finite set of noisy samples, the prediction errors may be significantly larger than in Equation (5), since various estimates in the CLS algorithm become noisy. Intuitively, multivariate calibration is more difficult either when different components are highly correlated in the training set or when there are non-negligible interferences amongst the different spectral responses $\mathbf{v}_j$. In order to quantify these effects, we define

$$s_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n} u_{i,j}^2} \tag{6}$$

and denote by λ_0 the minimal eigenvalue of the covariance matrix $\mathbf{C}$ of the training set. In addition, we define $\mathbf{V}$ to be the $k \times k$ matrix of interferences of the noise-free spectral responses $\mathbf{v}_1, \ldots, \mathbf{v}_k$, with entries $V_{i,j} = \mathbf{v}_i \cdot \mathbf{v}_j$, and denote by μ_0 its smallest eigenvalue.

The following theorem and its corollary, both proven in the Appendix, quantify the prediction error in CLS with a finite training set.

Theorem 2

Let $\mathbf{r}_1$ denote the estimated regression vector computed by CLS with a finite number of training samples. Then

$$\mathbf{r}_1 = \frac{\mathbf{v}_1^{\perp}}{\|\mathbf{v}_1^{\perp}\|^2} + \frac{\sigma}{\sqrt{n}}\frac{vs}{\lambda_0\mu_0}\boldsymbol{\zeta}_1 + \frac{\sigma^2}{n}\frac{s^2}{\lambda_0^2\mu_0}\boldsymbol{\zeta}_2 + O(\sigma^3) \tag{7}$$

where $v = \max_j \|\mathbf{v}_j\|$, $s = \max_j s_j$, $\boldsymbol{\zeta}_1$ is a random noise vector in $\mathbb{R}^p$ whose p co-ordinates all have zero mean and $O(1)$ variance and $\boldsymbol{\zeta}_2$ is a vector whose p co-ordinates are all $O(p)$. The p co-ordinates of $\boldsymbol{\zeta}_1$ are all linear combinations of the noises $\boldsymbol{\xi}_i$ in the training set, with the exact coefficients being complex functions of the covariance matrix $\mathbf{C}$ and the spectral responses $\mathbf{v}_j$. The vector $\boldsymbol{\zeta}_2$, on the other hand, is a complex quadratic function of the original noises in the training set, such that all its p co-ordinates are $O(p)$.

Equation (7) shows that for the case of a finite training set there can be substantial differences between the noise-free net analyte signal and the one estimated by CLS. The following corollary quantifies the expected mean squared error of prediction given the form (7) for the estimated regression vector. The expected squared prediction error is defined as the squared prediction error averaged over all possible noises in the n spectra of the training set, while keeping the concentration values fixed.

Corollary 1

The expected mean squared error of prediction for u_1 admits the form

$$\begin{aligned} E\{\mathrm{MSEP}(\mathrm{CLS}, n)\} &= E\{(\hat{u}_1 - u_1)^2\} \\ &= \frac{\sigma^2}{\|\mathbf{v}_1^{\perp}\|^2}\left(1 + \frac{c_1}{n} + \frac{\sigma^2 p^2}{n^2}(c_2 + o(1))\right) \end{aligned} \tag{8}$$

where the constants c_1 and c_2 are complicated functions of the covariance matrices $\mathbf{C}$ and $\mathbf{C}^p$ and the spectral responses $\mathbf{v}_j$ but are independent of σ, n and p.

Example 1

In the case of a system with a single component u_1 ($k=1$) the coefficients in (7) and (8) can be evaluated explicitly. Specifically, given a training set of n samples of the form $\tilde{\mathbf{x}}_i = u_i\mathbf{v} + \sigma\boldsymbol{\xi}_i$, where for simplicity the subscript notation is dropped from $\mathbf{v}_1$ and u_1, the estimate $\hat{\mathbf{v}}$ can be written as

$$\hat{\mathbf{v}} = \mathbf{v} + \frac{\sigma}{\sqrt{n}s}\hat{\boldsymbol{\xi}} \tag{9}$$

where s is given by (6) and

$$\hat{\boldsymbol{\xi}} = \frac{1}{\sqrt{n}s}\sum_{i=1}^{n} u_i\boldsymbol{\xi}_i$$

is a normal random variable in $\mathbb{R}^p$ whose p co-ordinates all have zero mean and unit variance. The regression vector is $\mathbf{r} = \hat{\mathbf{v}}/\|\hat{\mathbf{v}}\|^2$, leading to the following predicted value $\hat{u}$ for a new noisy sample $\tilde{\mathbf{x}}$:

$$\hat{u} = u\left(\frac{\|\mathbf{v}\|^2}{\|\hat{\mathbf{v}}\|^2} + \frac{\sigma}{\sqrt{n}s}\frac{\hat{\boldsymbol{\xi}}\cdot\mathbf{v}}{\|\hat{\mathbf{v}}\|^2}\right) + \sigma\frac{\boldsymbol{\xi}\cdot\hat{\mathbf{v}}}{\|\hat{\mathbf{v}}\|^2}$$

The corresponding expected mean squared error of prediction is

$$\begin{aligned} E\{\mathrm{MSEP}\} = \frac{\sigma^2}{\|\mathbf{v}\|^2}\Bigg[1 + \frac{1}{n}\frac{E\{u^2\}}{s^2} \\ + \frac{\sigma^2}{\|\mathbf{v}\|^2 s^2}\left(\frac{E\{u^2\}}{s^2}\frac{p^2 - 8p - 24}{n^2} - \frac{p-4}{n}\right) + O(\sigma^4)\Bigg] \end{aligned} \tag{10}$$

A detailed derivation of this formula appears in the Appendix. In principle it is obtained by an expansion of various quantities as power series in σ. Note that when $p=1$ we recover the well-known asymptotic formula for the expected MSEP in univariate calibration [40], up to an additional $1/n$ term due to mean centering (obviously absent in our formula, as we did not mean center the data).

We note that, for the case of a single component, Nishii and Krishnaiah [41] derived an exact formula for the expected MSEP in terms of moments of a Poisson random variable. Therefore Equation (10) can be derived by computing the asymptotic expansion of the moments of the Poisson random variable in their formula. This approach, however, is not applicable to more complex systems where exact expressions for the MSEP are unknown.

3.2. Partial least squares

While CLS is a direct calibration procedure requiring knowledge of all components in the system, PLS is an inverse calibration method that requires knowledge of only the analyte of interest. Owing to the difference between direct and inverse calibration, in the presence of noisy data the regression vectors of PLS and CLS differ even in the asymptotic limit $n \to \infty$. The following theorem, proven in Reference [24], characterizes the limiting behavior of PLS as $n \to \infty$ on inputs of the form (1) and (2). For the analysis we assume that k latent variables are needed in the noise-free case to reach a zero prediction error.

Theorem 3

Assume that training samples $\{\tilde{\mathbf{x}}_i, \mathbf{u}_i\}_{i=1}^n$ are random realizations from a population model with covariance matrix $\mathbf{C}^p$. Then $\mathbf{C} \to \mathbf{C}^p$ as $n \to \infty$ and the regression vector computed by PLS with k latent variables converges (with probability one) to the optimal one under a mean squared error criterion.

In other words, as $n \to \infty$, the regression vector computed by PLS, denoted $\mathbf{r}_{\text{PLS}}$, is equal to the optimal vector that minimizes the (population) mean squared prediction error

$$\min_{\mathbf{r} \in \mathbb{R}^p} E\{(u_1 - \tilde{\mathbf{x}} \cdot \mathbf{r})^2\}$$

where averaging is with respect to all possible values for the k concentrations $u_1, \ldots, u_k$ and over all possible noise vectors $\boldsymbol{\xi}$ in $\tilde{\mathbf{x}}$, all weighted by their corresponding probabilities. In general, owing to the presence of noise, $\mathbf{r}_{\text{PLS}}$ is not directly proportional to the net analyte signal vector [24]. However, it can be shown that

$$\mathbf{r}_{\text{PLS}} = \frac{\mathbf{v}_1^{\perp}}{\|\mathbf{v}_1^{\perp}\|^2} + \sigma \sum_j \beta_j \mathbf{v}_j + O(\sigma^2)$$

where the coefficients β_j are complex functions of the covariance matrix $\mathbf{C}^p$ and the spectral responses $\mathbf{v}_j$. The following theorem shows that, nonetheless, the effect of a finite and noisy training set on PLS is similar to its effect on CLS.

Theorem 4

Let $\mathbf{r}_{\text{PLS}}(n)$ denote the regression vector computed by PLS with k latent variables based on a finite number, n, of training samples. Then

$$\mathbf{r}_{\text{PLS}}(n) = \mathbf{r}_{\text{PLS}} + \alpha_1 \frac{\sigma}{\sqrt{n}} \boldsymbol{\zeta}_1 + \alpha_2 \frac{\sigma^2}{n} \boldsymbol{\zeta}_2 + O(\sigma^3) \tag{11}$$

where $\boldsymbol{\zeta}_1$ is a random noise vector in $\mathbb{R}^p$ whose p co-ordinates all have zero mean and unit variance, while $\boldsymbol{\zeta}_2$ is a random noise vector whose p co-ordinates are all $O(p)$. The p co-ordinates of $\boldsymbol{\zeta}_1$ are all linear combinations of the noises $\boldsymbol{\xi}_i$ in the training set, with the exact coefficients being complex functions of $\mathbf{C}$ and $\mathbf{v}_j$. Similarly, all p co-ordinates of $\boldsymbol{\zeta}_2$ are quadratic in the noises $\boldsymbol{\xi}_i$, and the coefficients α_1 and α_2 depend only on $\mathbf{C}$ and $\mathbf{v}_j$ but not on σ, n and p.

In contrast to CLS, where explicit bounds on α_1 and α_2 can be derived, similar formulae for PLS are much more difficult to compute, since PLS is an iterative algorithm. In the Appendix we follow the first few steps in the PLS algorithm, sketching the proof that the regression vector has the form (11). A regression vector of this form, in turn, leads to the following estimate for the expected prediction errors.

Corollary 2

The expected mean squared error of prediction of PLS admits the asymptotic form

$$E\{\text{MSEP(PLS)}\} = \text{MSEP}(\text{PLS}, n = \infty) \left(1 + \frac{c_1}{n} + c_2 \frac{\sigma^2 p^2}{n^2}(1 + o(1))\right) \tag{12}$$

where c_1 and c_2 are complex functions of $\mathbf{C}, \mathbf{C}^p$ and the spectral responses $\mathbf{v}_j$ but are independent of σ, n and p.

The proof of this corollary is essentially the same as that for the case of CLS and is thus omitted.

Example 2

Consider PLS on a single-component system. In this case, exact calculations of c_1 and c_2 are possible. First we consider the limit $n \to \infty$, where the optimal regression vector is given by

$$\mathbf{r}_{\text{PLS}} = \frac{V_1}{V_1 \|\mathbf{v}\|^2 + \sigma^2} \mathbf{v}$$

with $V_1 = E\{u_1^2\}$ and the subscript notation dropped from $\mathbf{v}_1$. The corresponding optimal root mean squared prediction error is

$$\text{RMSEP}(\text{PLS}, n = \infty) = \frac{\sigma}{\|\mathbf{v}\|} \frac{1}{\sqrt{1 + \sigma^2/\|\mathbf{v}\|^2 V_1}} \tag{13}$$

However, in the case of a finite training set and up to $O(\sigma^2)$ the expected MSEP is given by

$$E\{\text{MSEP(PLS)}\} \approx \frac{\sigma^2}{\|\mathbf{v}\|^2 B^2} \left[1 + \frac{1}{n} \frac{V_1}{s^2} + \frac{\sigma^2}{\|\mathbf{v}\|^2 s^2} \left(\frac{V_1}{s^2} \frac{(p+n)^2 + 4(p+n)}{n^2} + O\left(\frac{p}{n}\right) \right) \right] \tag{14}$$

where

$$B^2 = \left(1 + \frac{\sigma^2}{\|\mathbf{v}\|^2 s^2}\right)^2 + \frac{\sigma^2}{\|\mathbf{v}\|^2 s^2 n}(4p + 2n) + O(\sigma^4)$$

The derivation of (14) is similar to that of (10) for CLS with a single component and is therefore not described in detail. Comparison of (14) and (10) reveals that both PLS and CLS with one component have a similar performance and a similar behavior in the 'large p—small n' regime as long as $\sigma/s\|\mathbf{v}\| \ll 1$. Only when $\sigma/s\|\mathbf{v}\|$ is significantly larger than zero is B significantly larger than one, so that the shrinkage of PLS and its associated MSE superiority over CLS are evident.

3.3. Implications and applications

3.3.1. 'The good, the bad and the ugly' in multivariate calibration

Equations (7) and (8) for CLS and their analogues for PLS show that there are three main factors influencing the expected mean squared prediction error. The first factor (the good) is the sensitivity $\sigma/\|\mathbf{v}_1^{\perp}\|$, i.e. the noise strength divided by the length of the net analyte signal vector. The norm of the net analyte signal measures how unique the spectrum of the analyte of interest is in comparison with the spectra of the other interfering components, and the larger it is, the smaller are the prediction errors. The second factor (the bad) is the degree of statistical correlation between all components in the training set and the amount of interference of their spectral components, as measured by the eigenvalues λ_0 and μ_0 respectively. The larger the correlations or spectral interferences, leading to smaller values of λ_0 or μ_0, the worse is the expected prediction error. The last factor (the ugly) is the effect of the dimensionality of the signals. With all other parameters kept fixed, the higher the dimension, the more noisy are the various estimates, possibly leading to quite large prediction errors.

3.3.2. *Errors in the 'large* p*—small* n*' regime*

Many papers on univariate calibration consider the asymptotic expansion of the error only up to the $O(1/n)$ terms and do not explicitly consider the higher-order $O(1/n^2)$ term. While in univariate calibration the $O(1/n^2)$ term is indeed typically negligible with respect to the lower-order terms in n, this is not necessarily so in multivariate calibration, and more so in the 'large p–small n' regime typical of spectroscopic applications. This is because, as seen from (8) and (12), these terms are multiplied by p^2, which can be very large. Therefore in the 'large p—small n' regime, common in spectroscopy, it is this third term that can be the dominant one.

3.3.3. *Calibration design*

Equation (7) for the form of the regression vector in CLS provides a hint toward a good calibration design. A training set with large correlations between the different components, leading to a small value of λ_0, yields a relatively large error in the regression vector and thus a larger prediction error. Therefore in a controlled calibration setting the different components should have as large a variance as possible (if data are mean centered) and be as uncorrelated with each other as possible.

3.3.4. *The effect of mean centering*

Mean centering of the spectral signals $\mathbf{x}$ and the concentrations u_j according to their mean values in the training set is a common preprocessing algorithm, typically used to remove a baseline shift (not present in our model). One interesting question is the effect of mean centering on the prediction performance of CLS or PLS. With the aid of (10) and (14) we note that mean centering the concentrations leads to a decrease in the value of s_1, now being equal to the standard deviation of the training set concentrations. This in turn increases the value of $\sigma^2/s_1^2\|\mathbf{v}\|^2$ and thus increases the root mean squared error of prediction. This analysis provides a mathematical explanation for the numerical study of Seasholtz and Kowalski [42] that reported an increase in the RMSEP upon mean centering of their simulated data.

4. DIMENSIONAL REDUCTION, FEATURE SELECTION AND WAVELET COMPRESSION

Historically, CLS and PLS as well as PCR and other multivariate calibration algorithms were regarded as full spectrum methods, which eliminate the need for wavelength selection [26]. Part of this is due to the fact that, as proven by Lorber and Kowalski [12], the length of the net analyte signal vector is an increasing function of the number of co-ordinates, so asymptotically as $n \to \infty$ there is no need for wavelength selection under the linear mixture model. However, in the chemometrics literature there are many papers showing that, although PLS performs a dimensional reduction through its computed projections, an initial dimensional reduction *prior* to application of PLS is often beneficial (if not critical) in decreasing prediction errors. This initial step, in which the dimensionality of the signals is reduced, is often referred to as feature or variable selection, with the most common methods being algorithms that choose a subset of the wavelengths (see e.g. References [26,27,43]). In this section we formalize this empirical finding in the context of our mathematical analysis. However, we also show mathematically that, for complex systems with many interfering components and lack of specificity at any single wavelength, wavelength selection techniques cannot in general achieve optimal prediction errors. Based on the vast statistical literature on the properties and asymptotic optimality of wavelets for signal compression (dimensional reduction) [44,45], we propose to use adaptive wavelet feature selection for this purpose [31,32].

Let $T\colon \mathbb{R}^p \to \mathbb{R}^k$ denote a feature selection or dimensional reduction transformation that takes the original p-dimensional signals and outputs a k-dimensional representation with $k \ll p$. Specifically, we consider the family of possible feature selection transformations defined via the projection of the original signals into k orthonormal vectors $\mathbf{w}_1, \ldots, \mathbf{w}_k$:

$$T(\mathbf{x}) = (\mathbf{w}_1 \cdot \mathbf{x}, \mathbf{w}_2 \cdot \mathbf{x}, \ldots, \mathbf{w}_k \cdot \mathbf{x}) \tag{15}$$

In particular, this family includes all wavelength selection methods, where each projection $\mathbf{w}_j$ chooses a single wavelength.

Consider the expected prediction error when applying CLS for example on the reduced signals. Since T is a linear operator, for a signal $\mathbf{x}$ corrupted by a noise vector $\boldsymbol{\xi}_p \in \mathbb{R}^p$, we have that

$$T(\tilde{\mathbf{x}}) = T(\mathbf{x} + \sigma\boldsymbol{\xi}_p) = T(\mathbf{x}) + \sigma T(\boldsymbol{\xi}_p)$$

Since T is composed of k orthonormal projections, under the Gaussian noise model it follows that, if $\boldsymbol{\xi}_p$ is a standard multivariate normal random variable in p dimensions, then $T(\boldsymbol{\xi}_p)$ is a standard multivariate normal in k dimensions. Therefore the reduced signals $T(\mathbf{x})$ follow the same linear mixture model (1)–(2), only in a k dimensional space with the spectral responses $\mathbf{v}_j$ replaced by $T(\mathbf{v}_j)$. Therefore the same formulae for the expected MSEP, Equations (8) and (12), apply, only with the net analyte signal vector $\mathbf{v}_1^{\perp}$ replaced by $T(\mathbf{v}_1^{\perp})$ and the number of co-ordinates p, replaced by k. Specifically, for CLS we have that

$$E\{\mathrm{MSEP}(\mathrm{CLS}(T\mathbf{x}))\} = \frac{\sigma^2}{\left\|T(\mathbf{v}_1^{\perp})\right\|^2}\left(1 + \frac{c_1}{n} + c_2\frac{\sigma^2 k^2}{n^2}(1 + o(1))\right) \tag{16}$$

This formula reveals the requirements from the dimensional reduction operator T. It should be constructed such that the length of the net analyte signal vector is almost preserved ($\|T(\mathbf{v}_1^{\perp})\| \approx \|\mathbf{v}_1^{\perp}\|$) and yet the signals are represented by as few features as possible ($k \ll p$). If the net analyte signal vector was known, then a single projection with $\mathbf{w}_1 = \mathbf{v}_1^{\perp}$ would suffice. Since it is typically unknown *a priori*, a different criterion for the construction of T is needed. Wavelength selection schemes, and specifically those that choose specific wavelengths based on their individual predictive ability, are obviously suboptimal in complex systems, even though they may still achieve better prediction errors as compared with full spectrum methods. The reason is that it is possible to have spectral regions which are totally uncorrelated with the substance of interest and yet carry important information for calibration, and thus have non-negligible values of the net analyte signal in them (see e.g. Reference [24] and the numerical example in the next section).

Since the net analyte signal is typically unknown *a priori*, a possible different criterion is to simultaneously compress all signals in the training set as best as possible by representing each of them with only k features. If this compression is almost perfect, then $\|T(\mathbf{v}_1^{\perp})\| \approx \|\mathbf{v}_1^{\perp}\|$. For a single smooth signal corrupted by noise and sampled at p points (wavelengths), retaining only the set of wavelet coefficients above a certain threshold is asymptotically almost optimal [44]. The problem with this approach is that each signal will have its own specific set of k wavelet features, whereas for multivariate calibration we require a set of k features that can simultaneously describe all the spectra in the training set. While there are many possible methods to solve this problem, in this paper we focus on its solution via the joint best basis (JBB) algorithm [31,32]. In the JBB algorithm an orthonormal basis of wavelet features is constructed that best describes the data under a given (additive) cost functional, for example minimization of the overall entropy of the signals as measured in this basis. A dimensional reduction transformation T can then be defined by choosing the k most significant coefficients (with the highest entropy) in this construction. In the next section we present numerical results that show this approach to be almost optimal for prediction purposes.

We note that performing this initial dimensional reduction with the joint best basis algorithm has some additional advantages. The first is that this algorithm is fast, with complexity of the order of $O(np\log p)$ operations. The second is that wavelet operations are required only at the model-building stage. Once the regression coefficient is found in terms of these wavelet features, a regression vector in terms of the original variables can be constructed and used subsequently for new predictions without requiring any wavelet transformations. Finally, for this dimensional reduction step it is possible to use additional spectral samples for which knowledge of their chemical concentrations is unknown, as is the case when a large test set is already available. This leads to much better estimates of the best spectral features and to smaller reconstruction and prediction errors. Note that this use of additional test samples for the initial dimensional reduction is not possible with many wavelength selection algorithms.

Although under the linear mixture model wavelength selection schemes are in general inferior to wavelet compression, for practical data sets wavelength selection may be beneficial in at least two cases not covered in our simple model: (i) removal of spectral regions where the spectral intensities are not linearly related to the concentrations; (ii) removal of spectral regions with much higher noise than others. Therefore a combination of initial wavelength selection followed by dimensional reduction via adaptive wavelet features may prove to yield better prediction errors on real data sets.

5. NUMERICAL RESULTS

In this section we present the results of Monte Carlo simulations on data with one and two components that follow Equations (1) and (2). In the case of a single component we also compare the results with the theoretical Equations (10) and (14). The vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ used in the simulations are the digitized versions of

$$\begin{aligned} \mathbf{v}_1(t) &= \exp\left[-\left(\frac{t-0.15}{0.1}\right)^2\right] + 2\exp\left[-\left(\frac{t-0.7}{0.1}\right)^2\right] \\ \mathbf{v}_2(t) &= \exp\left[-\left(\frac{t-0.725}{0.1}\right)^2\right] + 0.5\exp\left[-\left(\frac{t-0.6}{0.05}\right)^2\right] \end{aligned} \tag{17}$$

sampled at p equidistant points in the unit interval $t \in [0, 1]$ and normalized to have unit L_2 norm. The functions $\mathbf{v}_1(t)$ and $\mathbf{v}_2(t)$ are shown in Figure 1 (left). As obvious from our analysis, the exact shape of the vectors is unimportant, as it is only the length of the net analyte signal that affects the overall error. We note that, in our simulations, increasing the dimension p while keeping the noise level per

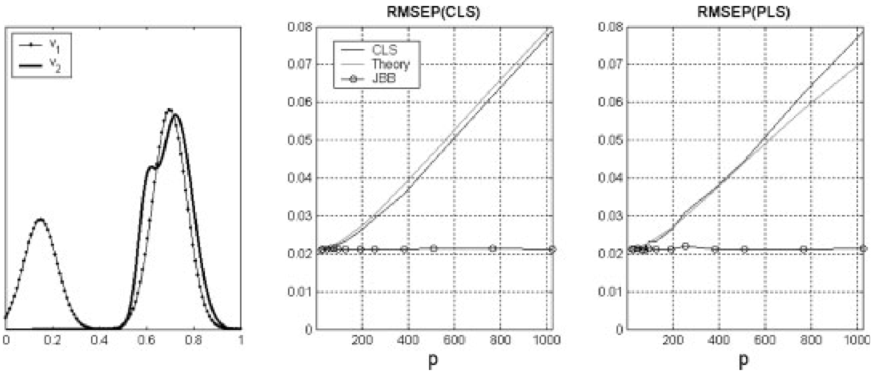

Figure 1. The two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ used in the simulations (left). The root mean squared error of prediction as a function of dimension for CLS regression with one component (middle) and for PLS (right).

co-ordinate and the norms of the vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ fixed yields a harder calibration problem. This is in contrast to the simulations of Thomas and Haaland [9] in which an increase in the dimension yields an easier calibration problem owing to the increase in the norm of the spectral response vectors.

In Figure 1 (middle) the root mean squared error of prediction for CLS is compared with the theoretical Equation (10), while in Figure 1 (right) the performance of PLS is compared with Equation (14). For these graphs we took u_1 to be uniformly distributed in the region $[0, 1]$, $\sigma = 0.02$, $n = 15$ training samples and 8000 test samples. Keeping the values of u_1 in the training and test samples fixed, spectral signals in different dimensions p between 24 and 1024 were randomly generated by adding Gaussian noise to the signals, afterwards building a regression vector from the training set and testing its performance on the test samples. This procedure was repeated 80 times (with the u_1s fixed, generating new noises each time) for statistical accuracy.

The choices of these parameters lead to values of $\sigma^2 p^2/n^2 s_1^2 \|\mathbf{v}_1\|^2$ between 0.01 and about 22.4. As seen from the graphs, in this range of values there is excellent agreement between theory and simulations. For small p the RMSEP is only slightly larger than the asymptotic value of $\sigma/\|\mathbf{v}_1\| = 0.02$. However, for large values of p e.g. $p = 800$, the RMSEP is more than three times as large. Another point shown in the graphs is that indeed, as predicted by the theoretical formulae, CLS and PLS have a similar performance, since in these simulations $\sigma/s_1\|\mathbf{v}_1\| \ll 1$.

In the middle and right figures we also present the results of first applying the joint best basis algorithm, retaining only the best 10 features, and then applying CLS or PLS respectively. For the computations we used Coiflets of order two as the underlying wavelets, although the specific choice of the wavelet does not much affect the results. As seen from the graphs, an initial dimensional reduction by this data-driven adaptive wavelet compression yields almost the optimal asymptotic error. The choice of $k = 10$ features was somewhat arbitrary. In principle the number of features k, can be viewed as a meta-parameter with the optimal value chosen by cross-validation.

For the case of two components the length of the net analyte signal vector is $\|\mathbf{v}_1^{\perp}\| \approx 0.4933$. Therefore the optimal error of CLS is $\sigma/\|\mathbf{v}_1^{\perp}\| \approx 0.0405$, with PLS having a similar asymptotic error since $\sigma/s\|\mathbf{v}_1^{\perp}\| \ll 1$. In Figure 2 we show the RMSEP as a function of dimension for both PLS on the full spectrum and PLS on the best 10 features as computed by the joint best basis algorithm. These runs were done with $n = 20$ training samples, with the second component u_2 also uniformly distributed on [0,1] but independent of u_1. Once again the full spectrum method suffers from a sharp increase in the prediction error as a function of dimension, while applying an initial feature selection yields much smaller prediction errors, only slightly larger than the optimal prediction error, regardless of the initial dimension.

We now compare the performance of the common wavelength selection algorithm of Reference [43] with feature selection by the JBB algorithm. We present results for $p = 128$, where PLS with two latent variables on the full spectrum leads to a root mean squared error of prediction of 0.077, while PLS-JBB gives an error of 0.047. In the wavelength selection scheme [43], each wavelength $1 \leqslant j \leqslant p$ is regressed on the concentration u_1 and a regression coefficient β_j is computed as well as an estimate for the variance of the residual error, σ_j^2. Wavelengths are ordered according to $|\beta_j|/\sigma_j^2$ and the top k are used to construct a model. In this scheme, k is a meta-parameter whose value is chosen such that it minimizes the prediction error on an independent test set. By looking at Figure 2 (left), it is evident that the net analyte signal vector has a non-negligible contribution from the right half of the spectrum, where, owing to interferences with u_2, none of these wavelengths are highly correlated with u_1. Therefore most of these wavelengths are not chosen, which leads to suboptimal prediction errors. The prediction error as a function of k for this wavelength selection scheme is plotted in Figure 2 (right), showing that wavelength selection is beneficial compared with the full spectrum method, as it is able to decrease the RMSEP from 0.077 for the full spectrum method to about 0.063 with about $k = 35$ wavelengths. However, its performance is still far from the asymptotic one and much worse than that of PLS on the first 10 wavelet features of the JBB algorithm, shown in the same figure.

6. SUMMARY AND DISCUSSION

In this paper we have presented a mathematical analysis of the expected prediction errors of CLS and PLS under the

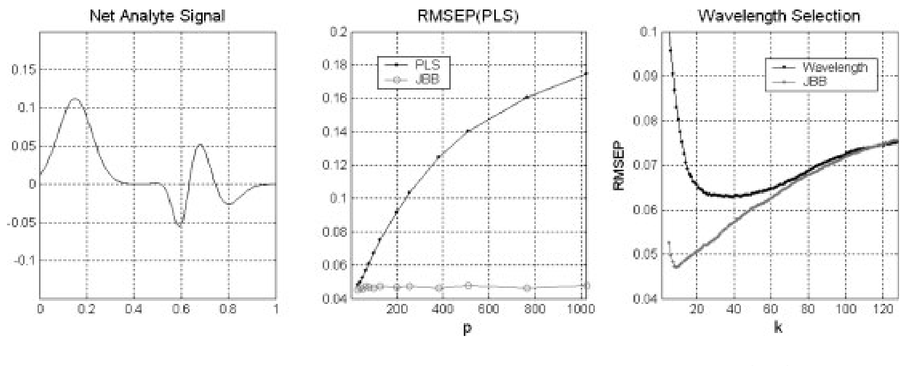

Figure 2. The net analyte signal vector for the two-component system (left). The root mean squared error of prediction as a function of dimension for full spectrum PLS and for PLS on the first 10 JBB co-ordinates (middle). The RMSEP as a function of the number of wavelengths chosen, k out of $p = 128$, in comparison with the prediction error of JBB as a function of the number of wavelet features, k (right).

linear mixture model, showing large errors of the order of $\sigma^2 p^2/n^2$ when there are many more variables than observations. The same analysis applies to many other full spectrum methods, such as PCR, HLA, OSC and ridge regression, and stresses the importance of feature selection *prior* to multivariate calibration.

This finding is contrary to the typical description of these methods in the literature as performing dimensional reduction and thus eliminating the need for feature selection. Therefore, the use of PLS as a dimensional reduction tool in spectroscopy, as well as its recent use in other areas with a large number of variables and a small number of samples, such as functional MRI [46] and gene expression data [47], should be carefully re-examined.

For the case of continuous signals, as in spectroscopy, we demonstrated the usefulness of data-driven wavelet methods in performing this initial dimensional reduction step, and their advantage over standard wavelength selection schemes. For the case of non-continuous signals, as in gene arrays, wavelet analysis is not applicable and other feature selection methods need to be derived.

Acknowledgements

B.N. would like to thank David Donoho and Sam Oman for interesting discussions and for pointing out References [29] and [41] respectively. The authors would also like to thank the anonymous referee for many helpful suggestions.

APPENDIX

A.1. Proofs of Theorems 1 and 2

A.1.1. Proof of Theorem 1

The estimates $\hat{\mathbf{v}}_j$ are given by (3), where $\mathbf{C}$ is the covariance matrix of the components u_j in the training set. As $n \to \infty$, according to the law of large numbers, $\mathbf{C} \to \mathbf{C}^p$ and $E\{\tilde{\mathbf{x}}u_j\} \to E\{\mathbf{x}u_j\}$ both with probability one. Therefore in the limit $n = \infty$ the estimates computed by CLS are independent of the noises and coincide with those for noise-free data. In particular, $\hat{\mathbf{v}}_j = \mathbf{v}_j$ and thus

$$\begin{pmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \vdots \\ \mathbf{r}_k \end{pmatrix} = \mathbf{V}^{-1} \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_k \end{pmatrix} = \begin{pmatrix} \mathbf{v}_1^{\perp}/\left\|\mathbf{v}_1^{\perp}\right\|^2 \\ \mathbf{v}_2^{\perp}/\left\|\mathbf{v}_2^{\perp}\right\|^2 \\ \vdots \\ \mathbf{v}_k^{\perp}/\left\|\mathbf{v}_k^{\perp}\right\|^2 \end{pmatrix}$$

A.1.2. Proof of Theorem 2

According to (3), the estimates $\hat{\mathbf{v}}_j$ depend on the averages $E\{\tilde{\mathbf{x}}u_j\}$, which can be written as

$$E\{\tilde{\mathbf{x}}u_j\} = E\{\mathbf{x}u_j\} + \frac{\sigma}{n}\sum_{i=1}^{n} u_{j,i}\boldsymbol{\xi}_i = E\{\mathbf{x}u_j\} + \frac{\sigma}{\sqrt{n}} s_j \boldsymbol{\eta}_j \tag{18}$$

where each $\boldsymbol{\eta}_j$ is a linear combination of the original n noises $\boldsymbol{\xi}_i$, appropriately scaled such that its p co-ordinates have zero mean and unit variance. Therefore combining (18), (3) and Theorem 1 gives

$$\begin{pmatrix} \hat{\mathbf{v}}_1 \\ \hat{\mathbf{v}}_2 \\ \vdots \\ \hat{\mathbf{v}}_k \end{pmatrix} = \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_k \end{pmatrix} + \frac{\sigma}{\sqrt{n}}\mathbf{C}^{-1} \begin{pmatrix} s_1\boldsymbol{\eta}_1 \\ s_2\boldsymbol{\eta}_2 \\ \vdots \\ s_k\boldsymbol{\eta}_k \end{pmatrix}$$

Since the matrix $\mathbf{C}$ is symmetric and positive definite, it has a set of k eigenvalues $0 < \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{k-1}$. The eigenvalues of $\mathbf{C}^{-1}$ are $1/\lambda_0 > 1/\lambda_1 > \cdots > 1/\lambda_{k-1}$. Therefore the noises $\boldsymbol{\eta}_j$ are expanded in length by at most $1/\lambda_0$. Specifically, each of the estimated spectral response vectors can be written as

$$\hat{\mathbf{v}}_j = \mathbf{v}_j + \frac{\sigma}{\sqrt{n}}\frac{s}{\lambda_0}\sum_{i=0}^{k} \alpha_{ji}\boldsymbol{\eta}_i$$

for some coefficients α_{ji} which depend on the covariance matrix $\mathbf{C}$ but are all $O(1)$ and are independent of the noises $\boldsymbol{\eta}_i$, and where $s = \max s_j$. Therefore we can write

$$\hat{\mathbf{v}}_j = \mathbf{v}_j + \varepsilon\hat{\boldsymbol{\xi}}_j \tag{19}$$

where $\varepsilon = \sigma s/\sqrt{n}\lambda_0$ and each $\hat{\boldsymbol{\xi}}_j$ is a random noise vector, linearly dependent on the original training noises $\boldsymbol{\xi}_i$, whose p co-ordinates all have zero mean and $O(1)$ variance.

Recall that $\hat{\mathbf{V}}$ is the $k \times k$ matrix of spectral interferences computed from the noisy estimates $\hat{\mathbf{v}}$ and that $\mathbf{V}$ is the corresponding noise-free matrix. Then, using (19),

$$\hat{\mathbf{V}}_{i,j} = \hat{\mathbf{v}}_i \cdot \hat{\mathbf{v}}_j = \mathbf{V}_{i,j} + \varepsilon\left(\hat{\boldsymbol{\xi}}_i \cdot \mathbf{v}_j + \hat{\boldsymbol{\xi}}_j \cdot \mathbf{v}_i\right) + \varepsilon^2\hat{\boldsymbol{\xi}}_i \cdot \hat{\boldsymbol{\xi}}_j \tag{20}$$

or, in matrix notation, $\hat{\mathbf{V}} = \mathbf{V} + \varepsilon\mathbf{R}_1 + \varepsilon^2\mathbf{R}_2$, where

$$(\mathbf{R}_1)_{i,j} = \hat{\boldsymbol{\xi}}_i \cdot \mathbf{v}_j + \hat{\boldsymbol{\xi}}_j \cdot \mathbf{v}_i, \qquad (\mathbf{R}_2)_{i,j} = \hat{\boldsymbol{\xi}}_i \cdot \hat{\boldsymbol{\xi}}_j \tag{21}$$

Note that the entries of $\mathbf{R}_1$ are all linear in the noises and are all $O(v)$, where $v = \max\|\mathbf{v}_j\|$. In addition, the entries of $\mathbf{R}_2$ are all quadratic in the noises. Moreover, the diagonal entries of $\mathbf{R}_2$, equal to $\|\boldsymbol{\xi}_j\|^2$, are all $O(p)$. The reason is that all p co-ordinates of the vector $\boldsymbol{\xi}_j$ have zero mean and $O(1)$ variance and therefore $\|\hat{\boldsymbol{\xi}}_j\|^2 = \sum_{i=1}^{p}\hat{\boldsymbol{\xi}}_{j,i}^2 = O(p)$.

We assume the noises are small enough so that the perturbed matrix $\hat{\mathbf{V}}$ is invertible (see also the discussion in Section A.2). In that case, to leading order in ε the inverse of $\hat{\mathbf{V}}$ is given by

$$\hat{\mathbf{V}}^{-1} = \left[\mathbf{I} - \varepsilon\mathbf{V}^{-1}\mathbf{R}_1 - \varepsilon^2\mathbf{V}^{-1}\mathbf{R}_2 + \varepsilon^2(\mathbf{V}^{-1}\mathbf{R}_1)^2 + O(\varepsilon^3)\right]\mathbf{V}^{-1} \tag{22}$$

Inserting (22) and (19) into (4) gives the following expansion for the estimated regression vectors:

$$\begin{aligned} \hat{\mathbf{r}} = {} & \mathbf{V}^{-1}\mathbf{v} + \varepsilon\mathbf{V}^{-1}\left(\hat{\boldsymbol{\xi}} - \mathbf{R}_1\mathbf{r}\right) \\ & + \varepsilon^2\left[(\mathbf{V}^{-1}\mathbf{R}_1)^2\mathbf{r} - \mathbf{V}^{-1}\mathbf{R}_2\mathbf{r} - \mathbf{V}^{-1}(\mathbf{R}_1\mathbf{V}^{-1} - \mathbf{R}_2)\hat{\boldsymbol{\xi}}\right] + O(\varepsilon^3) \end{aligned} \tag{23}$$

where $\hat{\mathbf{r}} = (\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2, \ldots, \hat{\mathbf{r}}_k)$ and $\hat{\boldsymbol{\xi}} = (\hat{\boldsymbol{\xi}}_1, \hat{\boldsymbol{\xi}}_2, \ldots, \hat{\boldsymbol{\xi}}_k)$, with similar definitions for $\mathbf{v}$ and $\mathbf{r}$.

Since $\hat{\boldsymbol{\xi}}$ and the entries of $\mathbf{R}_1$ are all linear in the original noises, we obtain that the $O(\varepsilon)$ correction to $\mathbf{r}_j$ is linear in the original noise vectors. In terms of magnitude, these noise vectors are expanded at most by $1/\mu_0$, where μ_0 is the lowest eigenvalue of the noise-free matrix $\mathbf{V}$. Similarly, the $O(\varepsilon^2)$

term is quadratic in the noises. The leading order term multiplying ε^2 is the second one involving $\mathbf{V}^{-1}\mathbf{R}_2\mathbf{r}$, since all the diagonal entries in $\mathbf{R}_2$ are $O(p)$. This term also scales as ε^2/μ_0. Therefore Equation (7) follows.

A.1.3. *Proof of Corollary 1*

Given the form (7) for the regression vector, the predicted value $\hat{u}_1$ for a new noisy sample $\tilde{\mathbf{x}}$ can be written as

$$\hat{u}_1 = \tilde{\mathbf{x}} \cdot \hat{\mathbf{r}}_1 = (\mathbf{x} + \sigma\boldsymbol{\xi}) \cdot \left(\mathbf{r}_1 + \alpha_1 \frac{\sigma}{\sqrt{n}} \boldsymbol{\zeta}_1 + \alpha_2 \frac{\sigma^2}{n} \boldsymbol{\zeta}_2 + O(\sigma^3)\right)$$

with suitably defined constants α_1 and α_2. Since the noise $\boldsymbol{\xi}$ is independent of $\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2$ and the concentrations in $\mathbf{x}$, we have that

$$\begin{aligned} E\{(\hat{u}_1 - u_1)^2\} &= \sigma^2 E\{(\boldsymbol{\xi} \cdot \mathbf{r}_1)^2\} + \alpha_1^2 \frac{\sigma^2}{n} E\{(\mathbf{x} \cdot \boldsymbol{\zeta}_1)^2\} \\ &\quad + \alpha_2^2 \frac{\sigma^4}{n^2} E\{(\mathbf{x} \cdot \boldsymbol{\zeta}_2)^2\} + O(\sigma^6) \end{aligned} \tag{24}$$

As $n \to \infty$, all terms other than the first vanish and we recover Equation (5), since

$$E\{(\boldsymbol{\xi} \cdot \mathbf{r}_1)^2\} = \|\mathbf{r}_1\|^2 = \frac{1}{\|\mathbf{v}_j^{\perp}\|^2}$$

The expected value of the second term in (24) gives a constant $c_1 = O(E\{\|\mathbf{x}\|^2\})$ which is independent of p, since all p co-ordinates of $\boldsymbol{\zeta}_1$ have $O(1)$ variance. The expected value of the third term, however, is of the order of $p^2 E\{\|\mathbf{x}\|^2\}$, since all p co-ordinates of $\boldsymbol{\zeta}_2$ are $O(p)$. Hence Equation (8) follows.

A.2. Derivation of Equation (10)

We rewrite Equation (9) for $\hat{\mathbf{v}}$ as

$$\hat{\mathbf{v}} = \|\mathbf{v}\| \left(\mathbf{v}_0 + \varepsilon\hat{\boldsymbol{\xi}}\right)$$

where $\mathbf{v}_0 = \mathbf{v}/\|\mathbf{v}\|$ is a vector of unit norm and $\varepsilon = \sigma/(s\sqrt{n}\|\mathbf{v}\|)$. Then

$$\|\hat{\mathbf{v}}\|^2 = \|\mathbf{v}\|^2 (1 + 2\varepsilon\mathbf{v}_0 \cdot \hat{\boldsymbol{\xi}} + \varepsilon^2 \|\hat{\boldsymbol{\xi}}\|^2)$$

and

$$\hat{u} - u = -u\varepsilon \frac{\mathbf{v}_0 \cdot \hat{\boldsymbol{\xi}} + \varepsilon\|\hat{\boldsymbol{\xi}}\|^2}{1 + 2\varepsilon\mathbf{v}_0 \cdot \hat{\boldsymbol{\xi}} + \varepsilon^2\|\hat{\boldsymbol{\xi}}\|^2} + \sigma \frac{\boldsymbol{\xi} \cdot \hat{\mathbf{v}}}{\|\hat{\mathbf{v}}\|^2}$$

Since the noise $\boldsymbol{\xi}$ in a new sample is independent of the noise $\hat{\boldsymbol{\xi}}$ derived from the noises in the training set, we can compute the averages of the first and second terms separately. We use the property that if $\boldsymbol{\xi}$ is a standard multivariate Gaussian random variable in $\mathbb{R}^p$ then $E\{(\boldsymbol{\xi} \cdot \mathbf{v})^2\} = \|\mathbf{v}\|^2$ to obtain that

$$E\{(\hat{u} - u)^2\} = \varepsilon^2 E\{u^2\} \frac{(\hat{\boldsymbol{\xi}} \cdot \mathbf{v}_0 + \varepsilon\|\hat{\boldsymbol{\xi}}\|^2)^2}{(1 + 2\varepsilon\mathbf{v}_0 \cdot \hat{\boldsymbol{\xi}} + \varepsilon^2\|\hat{\boldsymbol{\xi}}\|^2)^2} + \frac{\sigma^2}{\|\mathbf{v}\|^2} \tag{25}$$

Note that, as $n \to \infty$, $\varepsilon \to 0$ and we recover the single-component version of (5). To compute the expected MSEP for finite n, we need to average the right-hand side of (25) over all possible noise vectors $\hat{\boldsymbol{\xi}}$ due to all possible noises in the training set. Under the assumption that the noises follow a Gaussian distribution, the vector $\hat{\boldsymbol{\xi}}$ can in principle obtain any value. Therefore there is an exponentially small probability that $\hat{\mathbf{v}} = \mathbf{0}$, leading to an infinite expected MSEP. The same phenomenon occurs in the analysis of the expected MSEP in the univariate case. This technical problem is overcome by assuming that the noise has compact support with a cut-off at s standard deviations, with $1 \ll s < \|\mathbf{v}\|/\varepsilon$, so that $\mathbf{v}$ is never zero.

In this case we define the random variables

$$\begin{aligned} A &= (\hat{\boldsymbol{\xi}} \cdot \mathbf{v}_0)^2 + 2\varepsilon(\hat{\boldsymbol{\xi}} \cdot \mathbf{v}_0)\|\hat{\boldsymbol{\xi}}\|^2 + \varepsilon^2\|\hat{\boldsymbol{\xi}}\|^4 \\ B &= \|\mathbf{v}_0 + \varepsilon\hat{\boldsymbol{\xi}}\|^4 \\ C &= 1 + 2\varepsilon\hat{\boldsymbol{\xi}} \cdot \mathbf{v}_0 + \varepsilon^2\|\hat{\boldsymbol{\xi}}\|^2 \end{aligned} \tag{26}$$

and denote by μ_A, μ_B and μ_C their respective means. In terms of these random variables the expected MSEP of the CLS predictor for a training set with given fixed values $\{u_i\}_{i=1}^n$ is

$$E\{\text{MSEP}(\text{CLS}(n))\} = \varepsilon^2 E\{u^2\} E\left\{\frac{A}{B}\right\} + \frac{\sigma^2}{\|\mathbf{v}\|^2} E\left\{\frac{1}{C}\right\} \tag{27}$$

where expectancies are over the truncated noise vector $\hat{\boldsymbol{\xi}}$ in the finite training set. In the limit of $\varepsilon^2 p \ll 1$ and with the truncation of the noise $\hat{\boldsymbol{\xi}}$ at s standard deviations, $|C - \mu_C|/\mu_C < 1$ with probability one and it is thus possible to approximate

$$\begin{aligned} E\left\{\frac{1}{C}\right\} &= E\left\{\frac{1}{\mu_c(1 + \frac{C - \mu_C}{\mu_C})}\right\} \approx \frac{1}{\mu_C}\left[1 - E\left\{\frac{C - \mu_C}{\mu_C}\right\}\right. \\ &\quad \left. + E\left\{\left(\frac{C - \mu_C}{\mu_C}\right)^2\right\}\right] = \frac{1}{1 + \varepsilon^2 p}(1 + 4\varepsilon^2 + O(\varepsilon^4)) \end{aligned} \tag{28}$$

and, similarly,

$$E\left\{\frac{A}{B}\right\} \approx \frac{1}{\mu_B}\left(2\mu_A - \frac{E\{AB\}}{\mu_B} + \frac{E\{A(B - \mu_B)^2\}}{\mu_B^2}\right) \tag{29}$$

where

$$\begin{aligned} \mu_A &= E\{A\} = 1 + \varepsilon^2(p^2 + 2p) \\ \mu_B &= E\{B\} = 1 + \varepsilon^2(2p + 4) + \varepsilon^4(p^2 + 2p) \end{aligned} \tag{30}$$

and

$$\begin{aligned} E\{AB\} &= 1 + \varepsilon^2(p^2 + 12p + 72) \\ &\quad + \varepsilon^4(2p^3 + 21p^2 + 70p + 72) + O(\varepsilon^6) \\ E\{A(B - \mu_B)^2\} &= 48\varepsilon^2 + O(\varepsilon^4) \end{aligned} \tag{31}$$

Combining (27)–(31) and the definition of ε yields Equation (10).

A.3. Proof of Theorem 4

We sketch the proof that PLS has similar correction terms to its computed regression vector as in CLS regression by following the first few steps of the algorithm. In the first step an estimate of the first projection is computed as

$$\hat{\mathbf{w}}_1 = \frac{1}{n}\sum_{j=1}^{n} \tilde{\mathbf{x}}_j u_{1,j} = \mathbf{w}_1 + \frac{\sigma}{\sqrt{n}} s_1 \hat{\boldsymbol{\xi}}_1$$

where

$$\hat{\boldsymbol{\xi}}_1 = \frac{1}{\sqrt{n}s_1}\sum_{j=1}^{n} u_{1,j}\boldsymbol{\xi}_j$$

is a random variable with zero mean and unit variance in all of its p co-ordinates. The next step in PLS is the computation of the scores in the training set:

$$\hat{t}_j = \tilde{\mathbf{x}}_j \cdot \hat{\mathbf{w}}_j = t_j + \sigma\left(\boldsymbol{\xi}_j \cdot \mathbf{w}_1 + \frac{s}{\sqrt{n}}\hat{\boldsymbol{\xi}}_1 \cdot x_j\right) + \sigma^2 \frac{s_1}{\sqrt{n}}\boldsymbol{\xi}_j \cdot \hat{\boldsymbol{\xi}}_1$$

Notice that, since $\hat{\boldsymbol{\xi}}_1$ is a linear combination of the training set noises, one of the terms in $\boldsymbol{\xi}_j \cdot \hat{\boldsymbol{\xi}}_1$ involves$\|\boldsymbol{\xi}_j\|^2 = O(p)$. When taking all the multiplying factors into account, each term $\hat{t}_j$ is corrupted, amongst other terms, by a quantity of the order of $u_{1,j}\sigma^2 O(p)/n$. An analysis of the next iterative steps in PLS shows that the $O(\sigma)$ corrections to both the first score and the first spectral loading are linear in the noises, while the $O(\sigma^2)$ corrections contain terms which are $O(p)$. The same analysis applies to all the subsequent iterative steps in PLS, even though the exact computation of the coefficients is mathematically intractable.

REFERENCES

1. Haaland DM, Easterling RG, Vopicka DA. *Appl. Spectrosc.* 1985; **39**: 73–83.
2. Haaland DM. Multivariate calibration methods applied to quantitative FT-IR analyses. In *Practical Fourier Transform Infrared Spectroscopy*, Ferraro JR, Krishnan K (eds). Academic Press: New York, NY, 1989; 395–468.
3. Martens H, Naes T. *Multivariate Calibration*. Wiley: Chichester, 1989.
4. Wold S, Sjostrom M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemometrics Intell. Lab. Syst.* 2001; **58**: 109–130.
5. Haaland DM, Melgaard DK. New prediction-augmented classical least-squares (PACLS) methods: application to unmodeled interferents. *Appl. Spectrosc.* 2000; **54**: 1303–1312.
6. Melgaard DK, Haaland DM, Wehlburg CM. Concentration residual augmented classical least squares (CRACLS): a multivariate calibration method with advantages over partial least squares. *Appl. Spectrosc.* 2002; **56**: 615–624.
7. Estienne F, Pasti L, Centner V, Walczak B, Despagne F, Jouan-Rimbaud D, De Noord O, Massart D. A comparison of multivariate calibration techniques applied to experimental NIR data sets: Part II. Predictive ability under extrapolation conditions. *Chemometrics Intell. Lab. Syst.* 2001; **58**: 195–211.
8. Centner V, Verdu-Andres J, Walczak B, Jouan-Rimbaud D, Despagne F, Pasti L, Poppi R, Massart D and De Noord O. Comparison of multivariate calibration techniques applied to experimental NIR data sets. *Appl. Spectrosc.* 2000; **54**: 608–623.
9. Thomas EV, Haaland DM. Comparison of multivariate calibration methods for quantitative spectral analysis. *Anal. Chem.* 1990; **62**: 1091–1099.
10. Wentzell PD, Vega-Montoto L. Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures. *Chemometrics Intell. Lab. Syst.* 2003; **65**: 257–279.
11. Lorber A. Error propagation and figures of merit for quantification by solving matrix equations. *Anal. Chem.* 1986; **58**: 1167–1172.
12. 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.
13. Lorber A, Kowalski BR. Estimation of prediction error for multivariate calibration. *J. Chemometrics* 1988; **2**: 93–109.
14. De Vries S, Ter Braak CJF. Prediction error in partial least squares regression: a critique on the deviation used in the Unscrambler. *Chemometrics Intell. Lab. Syst.* 1995; **30**: 239–245.
15. Brown PJ. *Measurement, Regression and Calibration*. Oxford University Press: Oxford, 1993.
16. Brown PJ. Multivariate calibration. *J. R. Statist. Soc. B* 1982; **44**: 287–321.
17. Sundberg R. When is the inverse regression estimator MSE-superior to the standard regression estimator in multivariate controlled calibration situations. *Statist. Prob. Lett.* 1985; **3**: 75–79.
18. Sundberg R, Brown PJ. Multivariate calibration with more variables than observations. *Tehcnometrics* 1989; **31**: 365–371.
19. Denham MC, Brown PJ. Calibration with many variables. *Appl. Statist.* 1993; **42**: 515–528.
20. Oman SD, Srivastava MS. Exact mean squared error comparisons of the inverse and classical estimators in multi-univariate linear calibration. *Scand. J. Statist.* 1996; **23**: 473–488.
21. Hoskuldsson A. PLS regression methods. *J. Chemometrics* 1988; **2**: 211–228.
22. Garthwaite PH. An interpretation of partial least squares. *J. Am. Statist. Assoc.* 1994; **89**: 122–127.
23. Helland IS. Some theoretical aspects of partial least squares regression. *Chemometrics Intell. Lab. Syst.* 2001; **58**: 97–107.
24. Nadler B, Coifman RR. Partial least squares, Beer's law and the net analyte signal: statistical modeling and analysis. *J. Chemometrics* 2005; **19**: 45–54.
25. Oman SD. An exact formula for the mean squared error of the inverse estimator in the linear calibration problem. *J. Statist. Plan. Infer.* 1985; **11**: 189–196.
26. Xu L, Schechter I. Wavelength selection for simultaneous spectroscopic analysis. Experimental and theoretical study. *Anal. Chem.* 1996; **68**: 2392–2400.
27. Spiegelman CH, McShane MJ, Goetz MJ, Motamedi M, Yue QL, Cote GL. Theoretical justification of wavelength selection in PLS calibration: development of a new algorithm. *Anal. Chem.* 1998; **70**: 35–44.
28. Booksh K, Kowalski BR. Theory of analytical chemistry. *Anal. Chem.* 1994; **66**: 782–791.
29. Buckheit J, Donoho DL. Improved linear discrimination using time frequency dictionaries. *Proc. SPIE* 1995; **2569**: 540–551.
30. Raudys S, Young DM. Results in statistical discriminant analysis: a review of the former Soviet Union literature. *J. Mult. Anal.* 2004; **89**: 1–35.
31. Coifman RR, Wickerhauser MV. Entropy-based algorithms for best basis selection. *IEEE Trans. Info. Theory* 1992; **32**: 712–718.
32. Saito N, Coifman RR. On local orthonormal bases for classification and regression. *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing* 1995; 1529–1532.
33. Trygg J, Wold S. PLS regression on wavelet compressed NIR spectra. *Chemometrics Intell. Lab. Syst.* 1998; **42**: 209–22.
34. Teppola P, Minkkinen P. Wavelet–PLS regression models for both exploratory data analysis and process monitoring. *J. Chemometrics* 2000; **14**: 383–399.
35. Amato U, Antoniadis A, De Feis I. Dimension reduction in functional regression with applications. *Comput. Statist. Data Anal.* 2005; in press.
36. Lavine B, Workman JJ. Chemometrics. *Anal. Chem.* 2004; **76**: 3365–3372.
37. Brown PJ, Fearn T, Vannucci M. Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *J. Am. Statist. Assoc.* 2001; **96**: 398–408.
38. Sundberg R. Multivariate calibration—direct and indirect regression methodology. *Scand. J. Statist.* 1999; **26**: 161–207.

39. Berger A, Koo TW, Itzkan I, Feld MS. An enhanced algorithm for linear multivariate calibration. *Anal. Chem.* 1998; **70**: 623–627.
40. Tellinghuisen J. Inverse vs. classical calibration for small data sets. *Fresenius. J. Anal. Chem.* 2000; **368**: 585–588.
41. Nishii R, Krishnaiah PR. On the moments of classical estimates of explanatory variables under a multivariate calibration model. *Sankhya—Indian J. Statist. A* 1988; **50**: 137–148.
42. Seasholtz MB, Kowalski BR. The effect of mean centering on prediction in multivariate calibration. *J. Chemometrics* 1992; **6**: 103–111.
43. Brown PJ, Spiegelman CH, Denham MC. Chemometrics and spectral frequency selection. *Philos. Trans. R. Soc. Lond. A* 1991; **337**: 311–322.
44. Donoho DL, Johnstone IM. Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 1994; **81**: 425–455.
45. Vidakovic B. *Statistical Modeling by Wavelets*. Wiley: New York, NY, 1999.
46. McIntosh AR, Lobaugh NJ. Partial least squares analysis of neuroimaging data: applications and advances. *Neuroimage* 2004; **23**: 250–263.
47. Nguyen DV, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* 2002; **18**: 39–50.