JOURNAL OF CHEMOMETRICS
J. Chemometrics 2004; **18**: 526–536
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cem.897

A comparative study of linear regression methods in noisy environments

Marco S. Reis* and Pedro M. Saraiva

GEPSI, PSE Group, Department of Chemical Engineering, University of Coimbra, Pólo II, Pinhal de Marrocos, P-3030-290 Coimbra, Portugal

Received 21 May 2004; Revised 25 January 2005; Accepted 22 February 2005

With the development of measurement instrumentation methods and metrology, one is very often able to rigorously specify the uncertainty associated with each measured value (e.g. concentrations, spectra, process sensors). The use of this information, along with the corresponding raw measurements, should, in principle, lead to more sound ways of performing data analysis, since the quality of data can be explicitly taken into account. This should be true, in particular, when noise is heteroscedastic and of a large magnitude. In this paper we focus on alternative multivariate linear regression methods conceived to take into account data uncertainties. We critically investigate their prediction and parameter estimation capabilities and suggest some modifications of well-established approaches. All alternatives are tested under simulation scenarios that cover different noise and data structures. The results thus obtained provide guidelines on which methods to use and when. Interestingly enough, some of the methods that explicitly incorporate uncertainty information in their formulations tend to present not as good performances in the examples studied, whereas others that do not do so present an overall good performance. Copyright © 2005 John Wiley & Sons, Ltd.

KEYWORDS: measurement uncertainty; multivariate least squares; maximum likelihood principal component regression; partial least squares; principal component regression

1. INTRODUCTION

The majority of data analysis tools commonly applied to chemical processes rely on simplified assumptions regarding the nature of the errors to be included in their general statistical model structures. More specifically, the error term is normally considered to arise from several sources, such as modelling mismatch (inadequate model structure), uncontrolled interferences [1] and measurement noise, and their statistical descriptions are based on an assumed homoscedastic behaviour (i.e. with constant variance). This may be a reasonable assumption for the first two error sources (modelling mismatch and uncontrolled interferences), for which we are not usually able to provide additional *a priori* knowledge regarding their behaviour over time. However, with the development of measurement instrumentation methods and metrology, the depth of knowledge regarding measurement quality and features has increased significantly, so that one is very often able to rigorously specify their associated *uncertainty* [2].

Basically, uncertainty is defined as a 'parameter associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand' [2]. The *standard uncertainty*, u (to which we will refer simply as uncertainty), should be expressed in terms of the standard deviation of the values obtained under the same experimental conditions, and it can be obtained either from the analysis of collected data (the so-called type A evaluation) or through an adequate, alternative mean (type B evaluation). All the numerical quantities calculated from uncertain measurements turn out to be also uncertain quantities and therefore should have associated uncertainty values (*combined standard uncertainties*, u_c), calculated through uncertainty propagation formulaes.

That being the case, one should be able to express the uncertainty associated with each single raw data value to be used in any data analysis task. This implies that there should not be only one data table to be explored, but rather two tables: the usual raw data table and another one with the associated uncertainties. Therefore, with this added knowledge at our disposal, we should try to integrate it into our data analysis tasks. In fact, there is already a current trend towards this explicit consideration of measurement uncertainties. Namely, Wentzell *et al.* [3] developed so-called maximum likelihood principal component analysis (MLPCA), which estimates a PCA model in an optimal maximum likelihood sense when data are affected by measurement errors exhibiting a known complex structure, such as cross-correlations along sample or variable dimensions.

*Correspondence to: M. S. Reis, GEPSI, PSE Group, Department of Chemical Engineering, University of Coimbra, Pólo II, Pinhal de Marrocos, P-3030-290 Coimbra, Portugal.
E-mail: marco@eq.uc.pt
Contract/grant sponsor: FCT; Contract/grant number: POCTI/EQU/47638/2002.

The reasoning underlying MLPCA was then applied to multivariate calibration [4], extending the consideration of measurement uncertainties to input/output modelling approaches closely related to PCA, such as principal component regression (PCR) and latent root regression (LRR), giving rise to their maximum likelihood versions MLPCR (maximum likelihood principal component regression) and MLLRR (maximum likelihood latent root regression). Bro *et al.* [5] presented a general framework for integrating data uncertainties into the (maximum likelihood) fitting of models, which includes MLPCA as a special case. The issue of (least squares) model fitting is also referred to by Lira [6], along with the presentation of general expressions for uncertainty propagation in several input/output model structures. Both multivariate least squares (MLS) and its univariate version, bivariate least squares (BLS), were applied in several contexts of linear regression modelling, when all variables are subject to measurement errors [7–9]. All these different techniques have been used in several real world situations, such as multivariate calibration [4], signal processing [5], assessment of accuracy in analytical methods [9] and the presence of bias in method comparison studies [10] when both variables carry measurement errors. On the other hand, Faber and Kowalski [11] explicitly consider the influence of measurement errors in the calculation of confidence intervals for the parameters and predictions made using PCR and PLS, and similar efforts can also be found elsewhere [12–15].

In general, those techniques that are able to integrate data uncertainty into the core of their implementation procedure can lead to new and more flexible data analysis tools, in the sense that they are applicable in more general measurement error contexts, including those whose measurement error structures are not covered by more conventional techniques. Some examples of application contexts where such methods can be quite useful include the analysis of spectra (which often present noise, frequently of a heteroscedastic nature, and in the presence of strong correlations in the predictors), microarray data (where heteroscedasticity is mainly due to different levels of colour definition in the spotted arrays), laboratory data (where measurements of quality variables are often correlated and affected by different levels of uncertainties) and industrial data (as is the case of the example described in Section 5). The main purposes of this paper are thus (i) to gather several techniques with the potential of adequately handling complex noise sources in the context of linear regression modelling, (ii) to propose new developments for some of these approaches that may lead to improved performance and (iii) to conduct a Monte Carlo simulation study to assess the performance of the different alternative techniques under several noisy environments and data structures.

Other comparative studies can be found in the literature. For instance, Höskuldsson [16] compares the performances of several methods, e.g. ridge regression (RR), PLS and PCR, for several data sets using different quality measures, while Frank and Friedman [17] conduct an extended simulation study where the predictive performances of several regression methods are compared (OLS, PCR, PLS, RR and variable subset selection), but for conditions of homoscedastic noise.

Table I. Formulation of the optimization problems underlying OLS and MLS methods

OLS	$\hat{b}_{OLS} = \underset{b=[b_0 \ldots b_p]^T}{\arg\min} \left\{ \sum_{i=1}^{n} (y(i) - \hat{y}(i))^2 \right\}$	(1)
MLS	$\hat{b}_{MLS} = \underset{b=[b_0 \ldots b_p]^T}{\arg\min} \left\{ \sum_{i=1}^{n} \frac{(y(i) - \hat{y}(i))^2}{s_e^2(i)} \right\}$	(2)

In the present study we have used a wider range of noise scenarios that, combined with different data structures, allows one not only to extend their results to new contexts but also to bring into discussion other methods. The techniques considered include not only those available in the literature that are able to integrate uncertainty information (MLS, MLPCR), but also those whose ability to cope well with noisy data (PLS, PCR, RR) is widely recognized. Furthermore, we introduce modifications on some of the above methods in order to explicitly integrate data uncertainty information into their algorithms or fix some potential problems that arise when doing so.

The remaining sections of this paper are organized as follows. Section 2 presents the methods that will be used in our comparative study. Section 3 covers the comparative simulation study, whose results are presented and commented on in detail. Section 4 discusses several relevant issues involving the methods used as well as their relative merit, underlining some counterintuitive results and main features identified. Section 5 presents our conclusions.

2. MULTIVARIATE LINEAR REGRESSION METHODS

In this section we briefly review the methods that will be used later on in our Monte Carlo simulation comparative study. For the sake of clarity we organize them under four main groups according to their mutual affinities.

2.1. OLS group

Ordinary least squares (OLS) and multivariate least squares (MLS) [8,18] provide as estimates for the linear regression model parameters, those that derive from the solution of the optimization problems presented in Table I.

OLS tacitly assumes a homoscedastic behaviour (i.e. with constant variance) for the noise error term in the standard linear regression model. On the other hand, MLS is built upon an errors-in-variables (EIV) functional relationship among the true values for both the input and output variables, which are then affected by zero-mean random errors with a given covariance structure (assumedly known). In the denominator of Equation (2) we can find a term, $s_e^2(i)$, that results from the summation of the uncertainties associated with the response and the ones arising from the propagation of uncertainties of the predictors to the response (according to a formula derived from error propagation theory [6,18]):

$$s_e^2(i) = u_y(i)^2 + \sum_{j=1}^{p} \hat{b}_j^2 u_X(i,j)^2 - 2\sum_{j=1}^{p} \hat{b}_j \operatorname{cov}\left(\Delta\zeta(i), \Delta\xi_j(i)\right) + 2\sum_{j=1}^{p}\sum_{k=j+1}^{p} \hat{b}_j \hat{b}_k \operatorname{cov}\left(\Delta\xi_j(i), \Delta\xi_k(i)\right) \quad (3)$$

Table II. Formulation of the optimization problems underlying RR and rMLS methods

RR	$\hat{b}_{RR} = \underset{b=[b_0 \ldots b_p]^T}{\arg\min} \left\{ \sum_{i=1}^{n} (y(i) - \hat{y}(i))^2 + \lambda \sum_{j=1}^{p} b(j)^2 \right\}$	(4)
rMLS	$\hat{b}_{rMLS} = \underset{b=[b_0 \ldots b_p]^T}{\arg\min} \left\{ \sum_{i=1}^{n} \frac{(y(i) - \hat{y}(i))^2}{s_e^2(i)} + \lambda \sum_{j=1}^{p} b(j)^2 \right\}$	(5)

where $uX(i,j)$ and $uy(i)$ are the uncertainties associated with the ith observation of the jth input and output variables respectively, $\Delta\xi_j(i)$ and $\Delta\zeta(i)$ are the random errors affecting the ith measurement of predictor j and response, respectively, and $\hat{b}_j$ represents the coefficient of the linear regression model associated with variable j.

2.2. RR group

A well-known characteristic of the OLS method is the fact that the variance of its parameter estimates increases when the input variables get more correlated. Computational simulations showed us that the same applies to MLS. One possible way to address this issue consists of enforcing an effective shrinkage of the coefficients under estimation. This can be done by adopting the ridge regression (RR) regularization approach. It basically consists of adding an extra term to the objective function, which penalizes the occurrence of large solutions (in a square norm sense). The optimization formulation underlying RR estimates [19,20] and the one proposed for its counterpart based on MLS, rMLS (standing for 'ridge MLS'), are presented in Table II.

2.3. PCR group

PCR [1,21] is another methodology that handles collinearity among predictor variables. It uses those uncorrelated linear combinations of the input variables that most explain the input space variability, provided by principal component analysis (PCA), as the new set of predictors onto which the response is to be regressed. These predictors are orthogonal and therefore the collinearity problem is solved if we disregard the linear combinations with the smallest variability explanation power [22].

After developing MLPCA, which estimates the PCA subspace in an optimal maximum likelihood sense when data are affected by measurement errors with a known uncertainty structure [3], Wentzell *et al.* [4] applied it in the context of developing a PCR methodology that incorporates this additional knowledge regarding measurement uncertainties (MLPCR). As in PCR, the MLPCR methodology consists of first estimating a PCA model, now accomplished through MLPCA, in order to calculate the scores, using non-orthogonal (maximum likelihood) projections onto the estimated MLPCA subspace (instead of the PCA orthogonal projections), and then applying OLS to develop the final predictive model. This technique makes use of the available uncertainty information in the former phases (estimation of an MLPCA model and calculation of scores) but not during the stage where OLS is applied. Therefore Martínez *et al.* [18] propose a modification to the regression phase, in order to make it consistent with the efforts of integrating uncertainty information carried out in the initial phase, by replacing OLS with MLS (we will call this modification MLPCR1). In order to implement MLS in the second phase, estimated scores uncertainties for the ith observation are given by the diagonal elements of the following matrix [18]:

$$Z_i = \left\{ P^T [diag(uX(i,:))]^{-1} P \right\}^{-1} \quad (6)$$

where *diag* is an operator that converts a vector into a diagonal matrix and P is the matrix of maximum likelihood loads.

2.4. PLS group

PLS [1,16,21,23–27] is widely used by the chemometrics community in several contexts (such as multivariate calibration, QSAR and experimental design). It also adequately handles noisy data with correlated predictors in the estimation of a linear multivariate model. As in PCR, PLS finds a set of uncorrelated linear combinations of the predictors, belonging to some lower-dimensional subspace in the X-variables space, onto which y is to be regressed. In PLS this subspace is the one that, while still covering well the X-variability, provides a good description of the variability exhibited by the Y-variable(s). The algorithmic nature of PLS [16,23] can be translated into the solutions of a succession of optimization subproblems [1,21,24], as presented in the first column of Table III for one of its common versions, relative to the case of a single response variable (PLS1). However, if besides having available raw data, $[X|y]$, we also know their respective uncertainties, $[uX|uy]$, then one way to incorporate this additional information into a PLS algorithm would be through an adequate reformulation of the optimization subtasks appearing in its algorithmic structure. Therefore we propose a modification of the objective functions underlying each optimization subproblem in order to incorporate measurement uncertainties, but still preserving the successful algorithmic structure of PLS. Such a sequence of optimization subproblems is presented in the second column of Table III.

More details about implementation issues related to this modification of PLS1 (here called uncPLS1), which explicitly incorporates uncertainty information, are presented in Appendix 1.

3. A MONTE CARLO SIMULATION COMPARATIVE STUDY

In this section we describe the main results obtained through the application of all the different linear regression methods presented in Section 2 (PLS, uncPLS1, RR, rMLS, PCR, MLPCR, MLPCR1, OLS, MLS) to different data structure and noise conditions. The complete set of conditions employed is organized under a total of six case studies representing different noise patterns. Furthermore, for each case study, several simulation scenarios are covered, by varying some data structure and noise parameters, in order to enable a finer comparison between the different methods. The case studies explored cover the following situations: (1) complete heteroscedastic noise; (2) complete heteroscedastic noise plus bias; (3) all variables have similar levels of noise from the standpoint of their range of variation (*structured row-wise noise*); (4) noise as described in situation (3) plus bias; (5) proportional noise; (6) two levels of noise (very high

Table III. PLS1 as a succession of optimization subproblems (first column) and its counterpart that makes use of data uncertainties, uncPLS1 (second column)

PLS1	uncPLS1
Step 1. Pre-treatment Centre X and y; Scale X and y **Begin For Cycle** $a = 1$:# latent variables	Step 1. Pre-treatment Center X and y; Scale X and y. Scale X and y uncertainties **Begin For Cycle** $a = 1$:# latent variables
Step 2. Calculate ath X-weights vector (w) $w = \arg\min_w \sum_{i=1}^{n}\sum_{j=1}^{m}(X(i,j) - u(i)\times w(j))^2$ $w_{\text{new}} \leftarrow w_{\text{old}}/\|w_{\text{old}}\|$ Note: for $a = 1$ the Y-scores, u, are equal to y	Step 2. Calculate ath X-weights vector (w) $w = \arg\min_w \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(X(i,j) - u(i)\times w(j))^2}{uX(i,j)^2 + w(j)^2 \times uy(i)^2}$ $w_{\text{new}} \leftarrow w_{\text{old}}/\|w_{\text{old}}\|$
Step 3. Calculate ath X-scores vector (t) $t = \arg\min_t \sum_{i=1}^{n}\sum_{j=1}^{m}(X(i,j) - t(i)\times w(j))^2$	Step 3. Calculate ath X-scores vector (t) $t = \arg\min_t \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(X(i,j) - t(i)\times w(j))^2}{uX(i,j)^2}$
Step 4. Calculate ath X-loadings vector (p) $p = \arg\min_p \sum_{i=1}^{n}\sum_{j=1}^{m}(X(i,j) - t(i)\times p(j))^2$	Step 4. Calculate ath X-loadings vector (p) $p = \arg\min_p \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(X(i,j) - t(i)\times p(j))^2}{uX(i,j)^2 + p(j)^2 \times ut(i)^2}$
Step 5. Rescale X-scores and X-weights $p_{\text{new}} \leftarrow p_{\text{old}}/\|p_{\text{old}}\|$ $t_{\text{new}} \leftarrow t_{\text{old}} \times \|p_{\text{old}}\|$ $w_{\text{new}} \leftarrow w_{\text{old}} \times \|p_{\text{old}}\|$	Step 5. Rescale X-scores and X-weights $p_{\text{new}} \leftarrow p_{\text{old}}/\|p_{\text{old}}\|$ $t_{\text{new}} \leftarrow t_{\text{old}} \times \|p_{\text{old}}\|$ $w_{\text{new}} \leftarrow w_{\text{old}} \times \|p_{\text{old}}\|$ Step 5.1. Update $ut(i), i = 1:n$
Step 6. Regression of u on t (b) $b = \arg\min_b \sum_{i=1}^{n}(u(i) - t(i)\times b)^2$	Step 6. Regression of u on t (b) $b = \arg\min_b \sum_{i=1}^{n}\frac{(u(i) - b\times t(i))^2}{uu(i)^2 + b^2 \times ut(i)^2}$
Step 7. Calculation of X- and Y-residuals $E_a = E_{a-1} - t_a p_a^T \ (X = E_0)$ $F_a = F_{a-1} - b_a t_a \ (y = F_0)$ Note: continue the calculations with E_a playing the role of X and F_a the role of y (u)	Step 7. Calculation of X- and Y-residuals $E_a = E_{a-1} - t_a p_a^T \ (X = E_0)$ $F_a = F_{a-1} - b_a t_a \ (y = F_0)$ Step 7.1. Update $\{uE(i,j),\ uF(i)\}_{i=1:n,j=1:m}$
End For Cycle	End For Cycle

and low), where the occurrence of a particular level follows a binomial distribution.

Each simulation begins with the generation of noiseless signals using a linear model of the type $y = b_0 + b_1X_1 + \cdots + b_pX_p$ ($b_i = 1,\ i = 0:p,\ p = 10$). The X-data are generated from a multivariate normal distribution, and the variables can present two levels of correlation between themselves: 0.9 and 0.1. We will refer to these levels through the variable COST (standing for correlation structure): COST = 1 means a mutual correlation of 0.9 and COST = 2 a mutual correlation of 0.1. Zero-mean Gaussian noise is then added to the noiseless data (predictors and responses) according to the case study being considered and the associated noise parameter settings. Two noise parameters are accounted for: noise level (NOISEL) and heterogeneity level (HLEV). NOISEL represents the level of the average standard uncertainty affecting each variable (i.e. the level of the mean standard deviation for the additive noise that affects each variable) and is given by the multiplication of $K_1 = 0.01$ (if NOISEL = 1; low noise level) or $K_1 = 0.5$ (if NOISEL = 2; high noise level) by the theoretical standard deviation of the noiseless variables, i.e. $\bar{u}(X_i) = K_1(\text{NOISEL}) \times \sigma_{X_i}$. The other noise parameter, HLEV, represents the degree to which uncertainties vary along the observation index for a given variable (i.e. the degree of noise heterogeneity or heteroscedasticity for a given variable). HLEV = 1 means low variation of noise uncertainty or standard deviation from observation to observation, while HLEV = 2 represents highly heteroscedastic noise uncertainty behaviour. For variable X_i, uncertainties along the observation index are randomly generated from a uniform distribution centred at $\bar{u}(X_i)$, whose range is given by $R(\text{HLEV}) = K_2(\text{HLEV}) \times \bar{u}(X_i)$, where $K_2 = 0.01$ (if HLEV = 1; low heterogeneity level) or $K_2 = 1$ (if HLEV = 2; high heterogeneity level):

$$u(X_i(k)) \sim U\left[\bar{u}(X_i) - \frac{R(\text{HLEV})}{2}, \bar{u}(X_i) + \frac{R(\text{HLEV})}{2}\right]$$

The simulations conducted for all the scenarios covered share the same common structure, as follows.

1. First, the most adequate tuning parameter for each method is set (number of latent dimensions for PLS and PCR methods, ridge parameter for RR methods) regarding a given simulation scenario (each scenario is defined by a particular combination of levels for COST, NOISEL and HLEV). This is done by using (fivefold) cross-validation for PLS and PCR or by using (also fivefold) cross-validation plus the generation of a logarithmic grid in the range of plausible values for the ridge parameter (the parameter used in cross-validation is RMSEPW, defined below). This procedure is repeated 10 times and the median of the best values is chosen as the tuning

parameter to be used in our simulations (the median was used instead of the mean in order to provide some outlier protection for unusual parameter values that might be obtained during the 10 preliminary trials). Variables are 'autoscaled' in all methods except OLS and MLS. In the maximum likelihood versions of PCR (MLPCR and MLPCR1) the maximum number of latent variables was set equal to nine (total number of variables minus one) in order to avoid convergence problems that MLPCA runs into when the number of latent variables is the same as the number of predictor variables.

2. For a given case study and simulation scenario, two noiseless data sets are generated according to the linear model presented: a training or reference data set and a test data set, both with 100 multivariate observations. Furthermore, a random sequence of uncertainties (noise standard deviations) for all the observations belonging to each variable is also generated according to the NOISEL and HLEV parameters associated with them.
3. Then, zero-mean Gaussian noise with standard deviation given by the uncertainties generated (as explained in step 2) is added to the training and test data sets, based upon which a model is estimated by each method using the training data set and its prediction performance evaluated using the test data set. This process is repeated 100 times and the performance metrics are averaged over these trials.

3.1. Comparison metrics

The performance metric used for parameter estimation is the mean value of the relative (absolute) deviation or error (MRAE) of the estimated variable coefficients from the true ones. The estimate of the intercept was kept out of our calculations, as it often happens that this term dominates the overall error without being particularly relevant for the prediction results. Thus for each of the 100 simulations the following value is calculated:

$$\mathrm{MRAE}(i) = \frac{1}{p}\sum_{i=1}^{p}\frac{|b_i - \hat{b}_i|}{b_i}, \quad i = 1:100$$

where p is the number of predictor variables. For prediction assessment the square root of the weighted prediction mean square error in the test set (RMSEPW) is calculated, where the weights are the result of combining the predictor and response uncertainties, as is also suggested by the MLS criterion:

$$\mathrm{RMSEPW}(i) = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\frac{(y(k) - \hat{y}(k))^2}{uy(k)^2 + \left(uX(k,:)^{*2}\right)^T B^{*2}}}, \quad i = 1:100$$

where n represents the number of observations in the test set and B^{*2} is the Hadamard product of the coefficient vector (without intercept) with itself. The more familiar RMSEP is also calculated:

$$\mathrm{RMSEP}(i) = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(y(k) - \hat{y}(k))^2}, \quad i = 1:100$$

At the end of the 100 simulations the values of each of the above metrics are averaged and their standard deviations calculated and saved for further analysis.

The means of the quality metrics thus obtained allow us to make a rough comparison of the expected performance of the methods involved, while their standard deviations bring their variability into discussion. However, these quantities are still not enough by themselves to draw more in-depth conclusions about the relative performance of the methods because of the strong correlations between the values of the quality metrics for the various methods calculated during the 100 simulations. Therefore, for each case study and simulation scenario, paired t-tests were also conducted to determine whether method A is better than method B (a 'Win' for method A), performs worse (a 'loss') or there is no significant difference between methods A and B (a 'tie') for a given significance level (we used $\alpha = 0.01$). For the sake of simplicity we will only present here the number of 'wins', 'losses' and 'ties' that each method obtained for each simulation scenario. Alternatively, multiple comparison methods [28,29] could also have been adopted, especially if we want to have tight control over the overall significance level of the test performed. However, these methods are usually quite conservative, getting less sensitive to differences as the number of methods under comparison increases. For instance, a study where six methods were involved and significant differences apparently did exist resulted in no difference being detected between any of the methods at a reasonable level of significance using a Tukey's test-based multiple comparison approach [30]. Since we are comparing nine methods altogether, the sensitivity of such a test would be even more affected and therefore our choice went towards the adoption of an alternative, more sensitive approach. This comes at the cost of incurring in higher overall type I errors rates than the significance level used for each method, but as long as we keep this limitation in mind, our results still provide a sound basis for establishing the kind of general guidelines we are interested in identifying.

3.2. Case study 1

In this case study the various methodologies are tested under zero-mean heteroscedastic Gaussian noise with standard deviation (or uncertainty) randomly extracted from a uniform distribution. Eight simulation scenarios are explored in this case study, whose results are presented graphically to facilitate the extraction of general trends regarding their relative performance. Figure 1 reports the methods' performance in prediction (complete results are shown in Appendix II).

Regarding prediction results (Figure 1 and Table VIII in Appendix II), we can see that, for a low noise level (NOISEL = 1), methods MLPCR and uncPLS1 tend to perform comparatively worse, except for the scenario where the X-variables are highly correlated (COST = 1) and the noise is almost homoscedastic (HLEV = 1). For a high noise level (NOISEL = 2), MLPCR and uncPLS1 present very interesting results, whereas MLPCR1 and MLS present worse results for COST = 1. Curiously, OLS performs better than MLS for NOISEL = 2 (noisy conditions), especially when inputs are collinear. This is somewhat counterintuitive given the fact that MLS was supposed to take advantage of the knowledge of data uncertainties. Regarding parameter estimation (Table IX), for a low noise level (NOISEL = 1), method

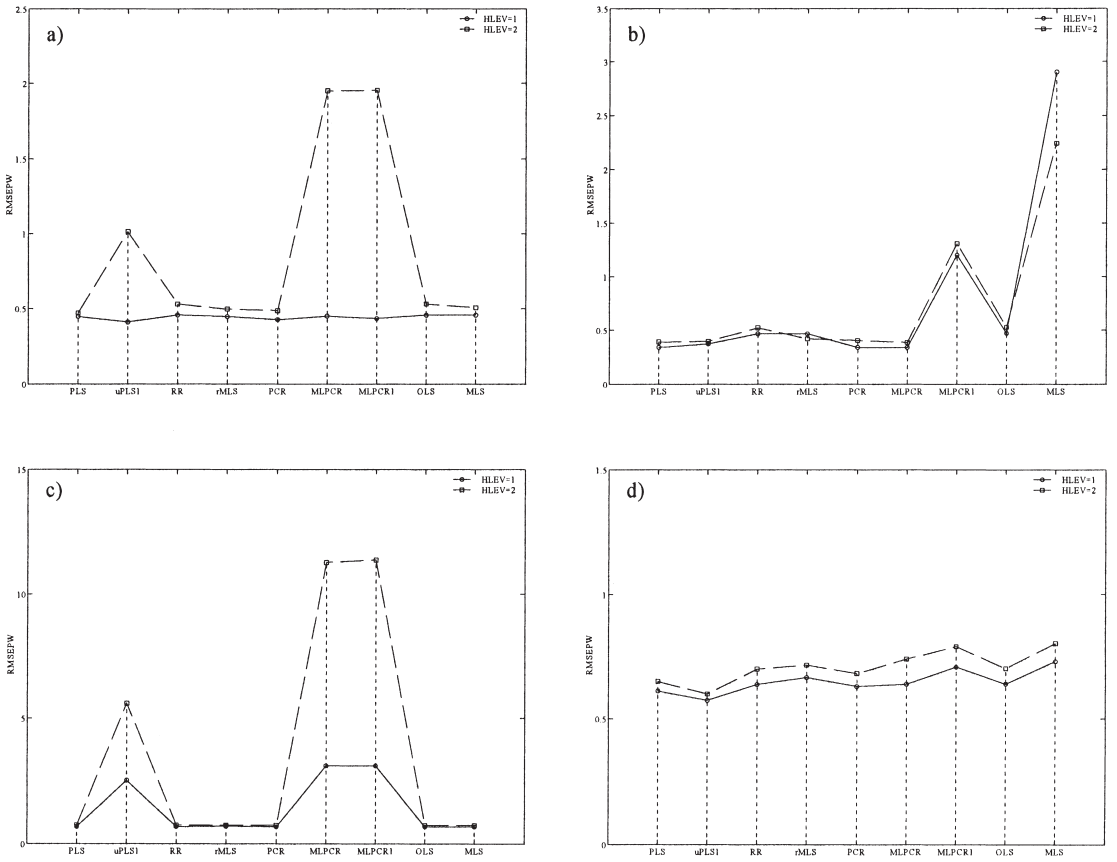

Figure 1. Prediction results (using RMSEPW) for case study 1: (a) NOISEL = 1 and COST = 1; (b) NOISEL = 2 and COST = 1; (c) NOISEL = 1 and COST = 2; (d) NOISEL = 2 and COST = 2.

uncPLS1 tends to perform better than all the others for COST = 1, but its results for COST = 2 are quite bad, in line with its poor prediction performance. At a high noise level the bad estimation performance of MLS for collinear inputs becomes quite evident; for COST = 2, uncPLS1 shows the best performance among all the methods. PLS and PCR present quite consistent performances, never failing completely and sometimes performing quite well. No relevant difference in the trends of prediction results obtained using either RMSEPW or RMSEP was identified.

3.3. Case studies 2–4

As the results for case studies 2 (heteroscedastic noise plus bias), 3 (row-wise structured noise) and 4 (row-wise structured noise plus bias) have similar performance trends for the various methods, we group them all under this single subsection. Case 2 addresses the situation where data bias is also present. Only scenarios where the noise parameters have high values were considered (NOISEL = 2 and HLEV = 2 for COST = {1,2}). For each of these two scenarios, after generation of the noise standard deviations (following the same procedure as in case study 1) a positive value (bias) was added to each datum, which amounts to 10% of the respective noise standard deviation, along with the randomly generated noise component. In case 3, instead of allowing the noise characteristics for each value to vary randomly according to the noise parameters, we forced a similar variation pattern in all the values belonging to the same row (simulating what happens if the whole line of collected values experiences similar oscillations of measurement quality). To do so, we simulated a single univariate random pattern along the observation index, $u(k)$, which will be used to establish the noise standard deviations for all the values in each row. As in case study 2, we consider only the scenarios defined by NOISEL = 2 and HLEV = 2 for COST = {1,2}. In case 4 the same procedure for generating noise standard deviations as for case 3 was adopted, but a bias term was added as for case 2 (10% of the generated noise standard deviation). The prediction (RMSEPW) and estimation (MRAE) results obtained for case study 2 are presented in Table IV.

Regarding prediction performance, the results obtained show that, for the situation where the X-variables are highly collinear (COST = 1), MLPCR presents the best metrics, whereas MLS maintains its sensitivity to this type of data structure. MLPCR1 also faces problems in this scenario. For low collinearity (COST = 2) the performance of the various methods is more uniform, but uncPLS1 and PLS tend to perform better. Parameter estimation results reveal that methods MLS, RR and OLS tend also to perform poorly for COST = 1, followed by rMLS. PLS and uncPLS1 present good estimation performances.

3.4. Case study 5

In this case study, zero-mean heteroscedastic Gaussian noise, whose standard deviation is proportional to the noiseless

Table IV. Summary of results regarding comparison metrics RMSEPW and MRAE for the simulation scenarios covered in case study 2. The mean, standard deviation (SD) and number of 'losses' (L), 'ties' (T) and 'wins' (W) are indicated for each method and simulation scenario. Simulation scenario settings are identified through the code 'case study/COST NOISEL HLEV'

	RMSEPW		MRAE	
Case/ Scenario Method	2/122 Mean/SD L T W	2/222 Mean/SD L T W	2/122 Mean/SD L T W	2/222 Mean/SD L T W
PLS	0.37/0.04 1 1 6	0.64/0.06 1 0 7	6.09/2.26 1 0 5	19.13/3.91 1 0 5
uncPLS1	0.38/0.05 3 0 5	0.60/0.05 0 0 8	13.37/12.09 2 0 4	11.90/2.81 0 0 6
RR	0.50/0.08 5 0 3	0.67/0.06 2 1 5	69.04/18.47 4 0 2	22.78/5.23 2 3 1
rMLS	0.41/0.04 4 0 4	0.70/0.07 5 1 2	38.19/11.00 3 0 3	23.78/5.58 2 3 1
PCR	0.37/0.04 1 1 6	0.66/0.06 2 2 4	5.78/2.35 0 0 6	22.50/4.70 2 3 1
MLPCR	0.36/0.04 0 0 8	0.71/0.06 5 1 2		
MLPCR1	1.29/1.03 7 0 1	0.76/0.09 7 1 0		
OLS	0.51/0.08 6 0 2	0.67/0.06 3 1 4	70.62/18.95 5 0 1	22.79/5.24 2 3 1
MLS	2.68/3.20 8 0 0	0.76/0.09 7 1 0	551.23/703.33 6 0 0	27.18/6.52 6 0 0

Table V. Summary of results regarding comparison metrics RMSEPW and MRAE for the simulation scenarios covered in case study 5. The mean, standard deviation (SD) and number of 'losses' (L), 'ties' (T) and 'wins' (W) are indicated for each method and simulation scenario. Simulation scenario settings are identified through the code 'case study/COST NOISEL HLEV'

	RMSEPW		MRAE	
Case/ scenario Method	5/122 Mean/SD L T W	5/222 Mean/SD L T W	5/122 Mean/SD L T W	5/222 Mean/SD L T W
PLS	0.35/0.05 1 2 5	0.68/0.07 0 4 4	6.45/2.22 1 1 4	21.55/3.55 0 1 5
uncPLS1	0.34/0.04 0 0 8	0.76/0.06 8 0 0	5.94/1.68 0 2 4	33.97/3.35 6 0 0
RR	0.50/0.07 5 0 3	0.68/0.09 0 3 5	71.85/20.88 4 0 2	22.75/5.04 2 1 3
rMLS	0.41/0.05 4 0 4	0.68/0.08 0 4 4	46.29/13.97 3 0 3	21.80/5.22 0 1 5
PCR	0.35/0.05 1 2 5	0.72/0.08 5 1 2	5.99/2.48 0 1 5	25.52/4.62 4 1 1
MLPCR	0.35/0.05 1 2 5	0.68/0.07 0 4 4		
MLPCR1	1.28/1.29 7 0 1	0.74/0.10 5 2 1		
OLS	0.51/0.07 6 0 2	0.68/0.09 1 3 4	73.45/21.38 5 0 1	22.76/5.05 2 1 3
MLS	2.42/2.43 8 0 0	0.75/0.10 6 1 1	481.75/472.82 6 0 0	25.56/7.10 4 1 1

signal level, is generated and added to the noiseless signals for each variable. Such uncertainties increase from a minimum value, $\min u(X_i)$, obtained for the minimum value of X_i, $\min X_i$, proportionally to the relative deviation of the level of the variable from this minimum value:

$$u(X_i(k)) = \min u(X_i) + \frac{X_i(k) - \min X_i}{\max X_i - \min X_i} \max u(X_i) \qquad (7)$$

The minimum value of the uncertainty was set to $\min u(X_i) = 0.001\sigma_{X_i}$ (σ_{X_i} is the theoretical standard deviation of the noiseless variable X_i), whereas $\max u(X_i) = \sigma_{X_i}$.

When predictors are weakly correlated (COST = 2), the performance of the methods is quite similar (Table V). However, when collinearity is present, MLS faces problems as well as MLPCR1. The uncPLS1 alternative behaves quite well, but its performance worsens when collinearity in the X-variables is removed.

3.5. Case study 6 (binomial heteroscedastic noise)

In this final case study we tested the various methods under a quite extreme scenario of heteroscedasticity, where uncertainties associated with the observed values of each variable are only allowed to vary between two possible levels: a quite high level ($u(X_i) = 3\sigma_{X_i}$) and a reasonably low level ($u(X_i) = 0.1\sigma_{X_i}$). The attribution of these two levels of uncertainty to observations is conducted by extracting them randomly from a binomial distribution, where the probability associated with the high level is equal to 0.9.

When prediction performance is evaluated in terms of RMSEPW, we registered improvements according to the following sequences: PLS → uncPLS1; RR → rMLS; PCR → MLPCR → MLPCR1 (Table VI). However, trends regarding RMSEP are quite different: PCR and MLPCR present the best performances, but MLPCR1 does not perform comparatively well. In terms of the mean values, RMSEPW results are more heterogeneous, in the sense that we can spot more clearly differences between the various methods, while they tend to present more similar prediction performances when compared in terms of RMSEP.

4. DISCUSSION

The results presented in the previous section provide a basis for establishing a detailed comparison among the methods studied, in the restricted context of the various scenarios chosen for our simulation study. By checking the results for the situations that most resemble a given practical application scenario, we can rank the methods that present greater potential to accomplish certain objectives regarding prediction or parameter estimation goals. However, we can go one step further in the analysis of results and try to extract some more global performance trends over the various methods. For instance, we can look at such an overall performance in terms of the number of 'losses', 'wins' and 'ties' obtained by each method for all the situations covered. However, we will restrict our analysis here to those situations where NOISEL = 2 (the majority), so that only truly noisy data structures support this comparison. Figure 2 presents such results for prediction (using RMSEPW), showing that PLS and uncPLS1 receive the best scores, followed by PCR and MLPCR. This is especially interesting given the fact that PLS

Table VI. Summary of results regarding comparison metrics RMSEPW, RMSEP and MRAE for the simulation scenarios covered in case study 6. The mean, standard deviation (SD) and number of 'losses' (L), 'ties' (T) and 'wins' (W) are indicated for each method and simulation scenario. Simulation scenario settings are identified through the code 'case study/COST NOISEL HLEV'

	RMSEPW		RMSEP		MRAE	
Case/ scenario Method	6/122 Mean/SD L T W	6/222 Mean/SD L T W	6/122 Mean/SD L T W	6/222 Mean/SD L T W	6/122 Mean/SD L T W	6/222 Mean/SD L T W
PLS	1.83/0.81 3 0 5	2.29/0.62 3 2 3	68.37/12.30 2 0 6	39.76/5.77 2 0 6	65.40/19.10 1 2 3	81.01/13.5 2 2 2
uncPLS1	0.88/1.02 1 0 7	2.22/2.26 1 1 6	101.21/145.69 332	71.90/69.39 6 1 1	90.26/165.20 0 3 3	76.37/72.7 1 0 5
RR	2.23/0.81 5 1 2	2.31/0.64 3 2 3	84.41/16.16 4 1 3	42.13/6.04 3 0 5	96.33/22.50 4 0 2	81.19/13.9 3 1 2
rMLS	0.59/0.12 0 0 8	1.48/0.41 0 0 8	79.02/11.99 3 1 4	56.18/6.21 5 0 3	67.12/19.10 1 2 3	52.52/13.0 0 0 6
PCR	2.38/0.77 7 0 1	2.43/0.54 6 1 1	56.03/7.93 0 1 7	34.55/4.27 1 0 7	57.64/16.70 0 1 5	90.39/11.0 5 0 1
MLPCR	2.21/0.84 4 2 2	2.42/0.54 6 1 1	56.95/10.87 0 1 7	32.44/3.30 0 0 8		
MLPCR1	1.56/2.32 2 0 6	2.41/3.38 1 1 6	191.40/271.39 7 0 1	83.72/88.04 6 1 1		
OLS	2.23/0.81 4 1 3	2.31/0.64 3 2 3	85.14/16.48 5 1 2	42.40/6.13 4 0 4	97.28/22.90 5 0 1	81.11/14.02 2 1 3
MLS	5.97/7.49 8 0 0	11.22/17.30 8 0 0	702.28/877.03 8 0 0	320.72/430.7 8 0 0	906.53/1121 6 0 0	375.36/516.5 6 0 0

does not incorporate *a priori* information regarding uncertainties in its algorithm. We can also see that the proposed modification to MLS, rMLS, leads to improved results.

As stated initially, this analysis is a general one and as such does not reflect 'local' problems that some methods face in practice. Therefore, keeping in mind their limitations, by looking into our detailed simulations results and Table VII, which tries to summarize them, one gets important insights for a sound selection process of the methods to be used in future applications, depending upon their goals and features.

The good results obtained with PLS are coherent with its well known ability to handle noisy situations [27], but still quite surprising is the fact that it sometimes outperforms other methods that do make explicit use of data uncertainties. Frank and Friedman [17] provide some general comments regarding PLS performance, focusing on its shrinkage properties along the eigendirections, which are smoother than the ones obtained using PCR, as more components are considered. Helland [25,26] further elaborates on this issue by stating some desirable and undesirable shrinkage properties of PLS, also referring to its usually higher parsimony in terms of the number of latent variables needed to achieve optimal predictive performance. However, we should exercise some care when extending their comments and results to the present situation, owing to the complex noise scenarios considered, which may interfere with the previous explanations, mostly based on models with homoscedastic error structures. In general terms, our interpretation of the good results obtained by PLS is closely linked to the effective way it provides for estimating the lower-dimensional predictive space (i.e. the one spanned by the set of weight vectors) onto which the regressors are projected prior to being used for predicting the response. In our opinion, this projection operation acts as a quite effective filter that removes or

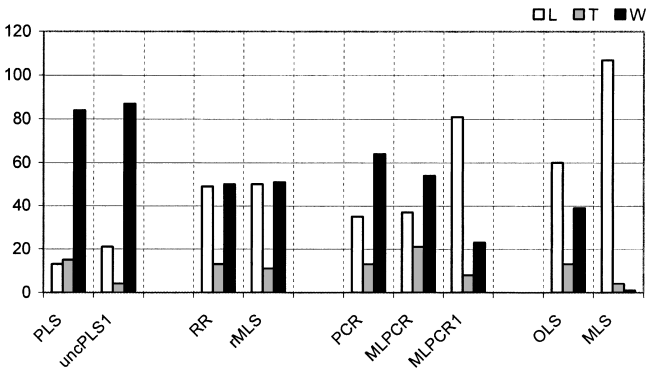

Figure 2. Results for number of 'losses' (L), 'ties' (T) and 'wins' (W) in the global assessment of all the methods studied regarding their prediction ability (using RMSEPW).

Table VII. Situations where our detailed simulation results advise *against* the use of certain methods

Do not use method	Under the following conditions		
	COST	NOISEL	HLEV
uncPLS1	2	1	{1,2}
	1	1	2
RR	1	2	2
MLPCR1	1	1	2
	1	2	{1,2}
	2	1	{1,2}
MLS	1	2	{1,2}

minimizes interferences due to noise (in fact, many filters can be formulated in terms of projection operations onto suitable bases, such as wavelets, sinusoids or the Karhunen–Loève transform). This 'estimation effectiveness' is related to what Höskuldsson [31] calls 'the stability of predictors derived from PLS methods', later interpreted through a fit/variance trade-off under the framework of the H-principle [16]. Furthermore, the use of orthogonal projections in this context also seems to play in favour of PLS for highly heteroscedastic data, when compared with uncertainty-based non-orthogonal projections or maximum likelihood projections used by the uncPLS1 and MLPCR methods. Simulation results show some evidence towards a lower variance of the orthogonal projection scores relative to the one exhibited by maximum likelihood projection scores, something that does not occur under homoscedastic situations. It seems to be the case that, for highly heteroscedastic scenarios, oscillations in the non-orthogonal projection line bring some added variability to the scores, other than the one strictly arising from the variability due to noise sources. This increased dispersion in the reduced space of the scores, usually the one relevant for prediction purposes, can increase prediction uncertainty due to poorly estimated models. Since heteroscedastic conditions prevail in the scenarios studied in this work, methods that use such non-orthogonal projections may be affected in their performance because of this feature.

In spite of the fact that method uncPLS1 represents an effort towards the explicit integration of uncertainty information into the algorithmic structure of PLS, some simplifications were introduced into it. Namely, the uncertainty of loading vectors and weights was neglected. Future developments should consider these issues, with the same concerns applying also to MLPCR methods, where uncertainty in the loads is also neglected when the propagation of uncertainties to the scores is carried out. On the other hand, the extensive solution of small optimization problems makes uncPLS1 more prone to numerical convergence problems than the original PLS method. However, this type of numerical problem does not represent a serious drawback in the solution of bivariate estimation problems in uncPLS1 (unless the X- and Y-scores are highly uncorrelated). The assumed independence of uncertainties in the scores for the regression step in MLPCR1 may also deserve more attention in future studies.

The poor performance of MLS when predictors are highly correlated may indicate that the inversion operation undertaken at each iteration is interfering with its performance. In fact, the matrix to be inverted in this method becomes quite ill-conditioned under collinear situations of the predictors. That being the case, the results obtained for the ridge regularization of MLS (rMLS) show that an effective stabilization of this inversion operation was achieved and the collinearity problem therefore minimized.

5. CONCLUSIONS

In this paper we present the results of a comparative study that involved the assessment of the prediction and parameter estimation performance of various methods under different noise and data structure scenarios. PLS methods (PLS, uncPLS1), as well as MLPCR and PCR, show good overall performances. Several real world applications are associated with contexts where uncertainty-based methods can be used with potential benefits. They can also be applied with added value to the analysis of industrial data sets, where sparsity is often a problem, due the presence of variables with different acquisition rates, along with randomly missing data. Under such circumstances an option consists of performing the data analysis on a coarser time scale (days or weeks) than the one suggested by raw data acquisition (minutes or hours). However, when taking appropriate averages, some variables summarize much more information than others (sampled less frequently) and therefore should be weighted differently in the analysis task. The integration of data uncertainty information regarding these averages on a coarser scale in our data analysis, through the types of methods addressed in this paper, provides a sound way to achieve this goal.

Our study covers a variety of data structures and noisy scenarios, but other remaining ones are interesting enough to deserve being addressed in future works, as is the case of data structures arising from latent variable frameworks [32] and of correlated noise, especially relevant in spectroscopic applications [33].

Acknowledgements

The authors would like to acknowledge FCT for financial support through research project POCTI/EQU/47638/2002. We also acknowledge all the quite relevant inputs received from the referees.

APPENDIX I: uncPLS1 IMPLEMENTATION DETAILS

AI.1. Computation of X-scores vector (t)

The calculation of the X-scores vector for each dimension involves solving the optimization problem formulated in step 3 of Table III. Its analytical solution can be derived using multivariate calculus [34], but provides the same numerical results as the maximum likelihood projection formula for calculation of X-scores in MLPCA presented in Reference [4]. Another issue in the calculation of the X-scores is related to the computation of the associated uncertainties. In our uncPLS1 procedure we calculate uncertainties propagated to the scores, assuming that uncertainties in the weights or loadings are negligible (a more complete treatment can be built upon the results of Goodman and Haberman [35]). As the scores can be given as maximum likelihood projections onto the subspace spanned by the weight vector, we can use an expression similar to Equation (6) in order to calculate uncertainty propagated to the ath X-scores. Furthermore, we assumed errors affecting variables to be independent.

AI.2. Computation of X-weights (w) and X-loadings (p) vectors

In the calculation of the X-weights vector the optimization problem can be seen as a succession of univariate regression problems of the Y-score, u, onto $X(:, j)$ (the jth column of X), with zero intercept. However, as both u and $X(:, j)$ have

associated uncertainties, the most adequate way of estimating the $w(j)$ coefficient, in the sense of the optimization subtask formulated in step 2, is by using BLS (without intercept). The same applies to the calculation of the X-loadings, where BLS is now applied to the regression of t onto $X(:,j)$, with the score uncertainties calculated as referred to above and the X-uncertainties provided as inputs or calculated for the residual matrices, obtained after deflation, as shown below.

AI.3. Computation of uncertainties for the X and y residual matrices

After deflation, in order to carry on with uncPLS1, we need to update the uncertainties associated with residual matrices E_a and F_a, which play, for $a > 1$, the same role that X and y have played during the calculations for $a = 1$. This can be done by applying error propagation theory (once again, we have assumed that only the scores carry significant uncertainties).

APPENDIX II: DETAILED RESULTS FOR CASE STUDY 1

Table VIII. Summary of results regarding comparison metric RMSEPW for the simulation scenarios covered in case study 1. The mean, standard deviation (SD) and number of 'losses' (L), 'ties' (T) and 'wins' (W) are indicated for each method and simulation scenario. Simulation scenario settings are identified through the code 'case study/COST NOISEL HLEV'

Case/ scenario Method	1/111 Mean/SD L T W	1/112 Mean/SD L T W	1/121 Mean/SD L T W	1/122 Mean/SD L T W	1/211 Mean/SD L T W	1/212 Mean/SD L T W	1/221 Mean/SD L T W	1/222 Mean/SD L T W
PLS	0.45/0.06 3 2 3	0.47/0.04 0 0 8	0.34/0.03 0 2 6	0.39/0.04 0 2 6	0.68/0.07 0 5 3	0.73/0.07 2 3 3	0.61/0.05 1 0 7	0.65/0.06 1 0 7
uncPLS1	0.41/0.04 0 0 8	1.01/0.09 6 0 2	0.37/0.05 3 0 5	0.40/0.04 1 1 6	0.68/0.44 6 0 2	5.60/0.17 6 0 2	0.57/0.05 0 0 8	0.60/0.05 0 0 8
RR	0.46/0.07 5 2 1	0.53/0.07 4 1 3	0.46/0.05 4 1 3	0.52/0.07 5 0 3	0.68/0.07 0 5 3	0.73/0.07 2 3 3	0.64/0.06 2 2 4	0.70/0.07 3 0 5
rMLS	0.45/0.06 3 2 3	0.50/0.06 1 1 6	0.47/0.05 4 2 2	0.42/0.05 4 0 4	0.68/0.07 0 5 3	0.73/0.06 0 1 7	0.67/0.07 6 0 2	0.72/0.08 5 0 3
PCR	0.43/0.04 1 1 6	0.49/0.06 1 1 6	0.34/0.03 0 2 6	0.40/0.05 3 0 5	0.68/0.07 0 5 3	0.73/0.07 2 3 3	0.63/0.06 2 2 4	0.68/0.07 2 0 6
MLPCR	0.45/0.06 3 5 0	1.95/0.23 7 1 0	0.34/0.03 0 2 6	0.38/0.05 0 1 7	0.68/0.14 7 0 1	11.26/0.09 7 0 1	0.64/0.06 3 2 3	0.74/0.08 6 0 2
MLPCR1	0.43/0.05 1 1 6	1.95/0.23 7 1 0	1.20/0.76 7 0 1	1.31/1.89 7 0 1	0.68/0.14 8 0 0	11.36/0.09 8 0 0	0.71/0.08 7 0 1	0.79/0.09 7 1 0
OLS	0.46/0.07 5 2 1	0.53/0.07 4 1 3	0.47/0.05 5 1 2	0.53/0.08 6 0 2	0.68/0.07 0 5 3	0.73/0.07 2 3 3	0.64/0.06 3 2 3	0.70/0.07 4 0 4
MLS	0.46/0.07 7 1 0	0.51/0.07 3 0 5	2.90/3.59 8 0 0	2.24/3.45 8 0 0	0.68/0.07 0 5 3	0.73/0.06 0 1 7	0.73/0.09 8 0 0	0.80/0.11 7 1 0

Table IX. Summary of results regarding comparison metric MRAE for the simulation scenarios covered in case study 1. The mean, standard deviation (SD) and number of 'losses' (L), 'ties' (T) and 'wins' (W) are indicated for each method and simulation scenario. Simulation scenario settings are identified through the code 'case study/COST NOISEL HLEV'

Case/ scenario Method	1/111 Mean/SD L T W	1/112 Mean/SD L T W	1/121 Mean/SD L T W	1/122 Mean/SD L T W	1/211 Mean/SD L T W	1/212 Mean/SD L T W	1/221 Mean/SD L T W	1/222 Mean/SD L T W
PLS	2.40/0.60 2 1 3	1.97/0.04 1 1 4	5.81/2.23 1 0 5	5.94/2.53 0 0 6	0.47/0.11 0 5 1	0.49/0.13 2 3 1	18.80/3.92 1 0 5	19.47/4.04 1 0 5
uncPLS1	1.94/0.03 0 1 5	1.68/0.05 0 0 6	28.77/18.6 2 0 4	13.6/12.1 1 0 5	0.68/0.74 6 0 0	7.24/0.22 6 0 0	9.48/3.92 0 0 6	15.87/4.42 0 0 6
RR	2.55/0.67 4 1 1	2.65/0.62 5 1 0	69.3/17.7 3 1 2	68.5/19.4 4 0 2	0.68/0.11 0 5 1	0.49/0.13 2 3 1	22.17/5.02 2 2 2	24.68/5.54 3 1 2
rMLS	2.41/0.62 2 1 3	2.31/0.53 3 0 3	70.1/16.9 3 2 1	36.3/9.73 3 0 3	0.68/0.11 0 5 1	0.46/0.12 0 1 5	23.43/5.53 5 0 1	24.38/6.59 2 3 1
PCR	1.98/0.44 0 1 5	2.04/0.46 1 1 4	5.47/2.33 0 0 6	18.2/10.9 2 0 4	0.68/0.11 0 5 1	0.49/0.13 2 3 1	21.69/4.51 2 2 2	23.33/5.08 2 1 3
OLS	2.55/0.67 4 1 1	2.65/0.62 5 1 0	70.7/18.2 4 1 1	70.1/19.9 5 0 1	0.68/0.11 0 5 1	0.49/0.13 2 3 1	22.17/5.03 2 2 2	24.70/5.56 4 1 1
MLS	2.55/0.67 6 0 0	2.43/0.57 4 0 2	630/783 6 0 0	442/647 6 0 0	0.68/0.11 0 5 1	0.46/0.12 0 1 5	27.38/6.88 6 0 0	29.56/9.13 6 0 0

REFERENCES

1. Martens H, Naes T. *Multivariate Calibration*. Wiley: Chichester, 1989.
2. ISO. *Guide to the Expression of Uncertainty*. ISO: Geneva, 1993.
3. Wentzell PD, Andrews DT, Hamilton DC, Faber K, Kowalski BR. Maximum likelihood principal component analysis. *J. Chemometrics* 1997; **11**: 339–366.
4. Wentzell PD, Andrews DT, Kowalski BR. Maximum likelihood multivariate calibration. *Anal. Chem.* 1997; **69**: 2299–2311.
5. Bro R, Sidiropoulos ND, Smilde AK. Maximum likelihood fitting using ordinary least squares algorithms. *J. Chemometrics* 2002; **16**: 387–400.
6. Lira I. *Evaluating the Measurement Uncertainty*. Institute of Physics Publishing: Bristol, 2002.
7. Martínez À, Riu J, Rius FX. Lack of fit in linear regression considering errors in both axes. *Chemometrics Intell. Lab. Syst.* 2000; **54**: 61–73.
8. Río FJ, Riu J, Rius FX. Prediction intervals in linear regression taking into account errors in both axes. *J. Chemometrics* 2001; **15**: 773–788.
9. Riu J, Rius FX. Assessing the accuracy of analytical methods using linear regression with errors in both axes. *Anal. Chem.* 1996; **68**: 1851–1857.
10. Martínez À, Riu J, Rius FX. Evaluating bias in method comparison studies using linear regression with errors in both axes. *J. Chemometrics* 2002; **16**: 41–53.
11. Faber K, Kowalski BR. Propagation of measurement errors for the validation of predictions obtained by principal component regression and partial least squares. *J. Chemometrics* 1997; **11**: 181–238.
12. Faber K. Comparison of two recently proposed expressions for partial least squares regression prediction error. *Chemometrics Intell. Lab. Syst.* 2000; **52**: 123–134.
13. Faber K, Bro R. Standard error of prediction for multiway PLS: 1. Background and a simulation study. *Chemometrics Intell. Lab. Syst.* 2002; **61**: 133–149.
14. Phatak A, Reilly PM, Penlidis A. An approach to interval estimation in partial least squares regression. *Anal. Chim. Acta* 1993; **277**: 495–501.
15. Pierna JAF, Jin L, Wahl F, Faber NM, Massart DL. Estimation of partial least squares regression prediction uncertainty when the reference values carry a sizeable measurement error. *Chemometrics Intell. Lab. Syst.* 2003; **65**: 281–291.
16. Höskuldsson A. *Prediction Methods in Science and Technology*. Thor Publishing, 1996.
17. Frank IE, Friedman JH. A statistical view of some chemometrics regression tools. *Technometrics* 1993; **35**: 109–135.
18. Martínez À, Riu J, Rius FX. Application of the multivariate least squares regression method to PCR and maximum likelihood PCR techniques. *J. Chemometrics* 2002; **16**: 189–197.
19. Draper NR, Smith H. *Applied Regression Analysis* (3rd edn). Wiley: New York, 1998.
20. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. Springer: New York, 2001.
21. Jackson JE. *A User's Guide to Principal Components*. Wiley: New York, 1991.
22. Martens H, Mevik B-H. Understanding the collinearity problem in regression and discriminant analysis. *J. Chemometrics* 2001; **15**: 413–426.
23. Geladi P, Kowalski BR. Partial least-squares regression: a tutorial. *Anal. Chim. Acta* 1986; **185**: 1–17.
24. Haaland DM, Thomas EV. Partial least-squares methods for spectral analysis. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Anal. Chem.* 1988; **60**: 1193–1202.
25. Helland IS. Some theoretical aspects of partial least squares regression. *Chemometrics Intell. Lab. Syst.* 2001; **58**: 97–107.
26. Helland IS. On the structure of partial least squares regression. *Commun. Statist.-Simul.* 1988; **17**: 581–607.
27. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemometrics Intell. Lab. Syst.* 2001; **58**: 109–130.
28. Kendall M, Stuart A, Ord JK. *The Advanced Theory of Statistics* (4th edn), vol. 3. Charles Griffin: London, 1983.
29. Scheffé H. *The Analysis of Variance*. Wiley: New York, 1959.
30. Indahl UG, Naes T. Evaluation of alternative spectral feature extraction methods of textural images for multivariate modeling. *J. Chemometrics* 1998; **12**: 261–278.
31. Höskuldsson A. PLS regression methods. *J. Chemometrics* 1988; **2**: 211–228.
32. Burnham AJ, Macgregor JF, Viveros R. Latent variable multivariate regression modeling. *Chemometrics Intell. Lab. Syst.* 1999; **48**: 167–180.
33. Wentzell PD, Lohnes MT. Maximum likelihood principal component analysis with correlated measurement errors: theoretical and practical considerations. *Chemometrics Intell. Lab. Syst.* 1999; **45**: 65–85.
34. Magnus JR, Neudecker H. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley: Chichester, 1988.
35. Goodman LA, Haberman SJ. The analysis of nonadditivity in two-way analysis of variance. *J. Am. Statist. Assoc.* 1990; **85**: 109–135.