Distribution based truncation for variable selection in subspace methods for multivariate regression

Kristian Hovde Liland[1,⋆], Martin Høy[2], Harald Martens[2,3], Solve Sæbø[1]

8th January 2013

1) Norwegian University of Life Sciences, Department of Chemistry, Biotechnology and Food Science
P.O. Box 5003, N-1432 Ås, Norway

2) Nofima, Norwegian Institute of Food, Fisheries and Aquaculture Research
Osloveien 1, N-1430 Ås, Norway,

3) Norwegian University of Life Sciences, Department of Mathematical Sciences and Technology
P.O. Box 5003, N-1432 Ås, Norway

(⋆) Corresponding author: kristian.liland@umb.no, tel: +47 64965830

Distribution based truncation for variable selection in subspace methods for multivariate regression

Abstract

Analysis of data containing a vast number of features, but only a limited number of informative ones, requires methods that can separate true signal from noise variables. One class of methods attempting this are the sparse partial least squares methods for regression (sparse PLS). This paper aims at improving the theoretical foundation, speed and robustness of such methods. A general justification of truncation of PLS loading weights is achieved through distribution theory and the central limit theorem. We also introduce a quick plug-in based truncation procedure based on a novel application of theory intended for analysis of variance for experiments without replicates. The result is a versatile and intuitive method that performs component-wise variable selection very efficiently and in a less ad hoc manner than existing methods. Prediction performance is on par with existing methods, while robustness is ensured through a better theoretical foundation.

1 Introduction

One of the major challenges in recent and coming data analysis is the ever increasing number of variables recorded for each sample. The data matrices become wider and wider. Because of instrumental noise, biological noise and other uncontrollable variations in the recorded signal, variables that should have no signal for a given sample, or be equal across samples, almost never show a zero signal in the final centred data set. And differences between two signals that should be zero are seldom zero in practice. Since predictive multivariate methods like partial least squares regression (PLSR) [1] in their basic forms take into account all variables, the sheer number of non-zero noise variables will often over-shadow the true signal.

Various forms of variable selection approaches have been proposed in the context of regression. Variable selection can also play a role in finding important variables in explorative studies, with the purpose of stabilizing the regression modelling and improving its predictive ability and interpretability. Sometimes the aim is to find which variables influence a certain process causually, or at least convey the most interesting information, e.g. metabolites, genes, wavenumbers, or molecular weights. Depending on the aim of the study different selection strategies may be favourable and the focus on how many variables to retain may be different.

Based on ideas of component-wise variable selection, sparseness and normally distributed noise we propose to use distribution based truncation to identify all unimportant model parameters that are (or appear to be) non-zero due to random errors, and force these towards zero. In the present PLSR context, this means to zero out small, apparently random elements in all the loading weight vectors. The intension is thereby to drastically reduce the problem of non-zero noise contributions. In the following sections we will look at some related methods intended for the same purpose and motivate a simple, intuitive and flexible strategy for truncation of non-informative variables. Applications to real and simulated data and comparison with other methods will also be presented.

2 Background

A basic assumption in statistics is the central limit theorem (CLT). The CLT was first presented by Abraham de Moivre in 1733 and has been formalised and interpreted under varying conditions and degrees of strictness ever since. A simple interpretation is that as the number of observations sampled from a random process increases, the distribution of the mean (and the sum) will approach a normal distribution. More interesting in this context is that many types of random noise are seen as approximately normally distributed, and linear combinations of such will tend even more towards the normal distribution. In this paper we propose to use the CLT to distinguish between variables with expected non-zero loading weights from the noisy variables with loading weights with a zero-expectation. We refer to the new modelling principle as Truncation-PLS in the following, and the resulting methods Truncation-PLSR and Truncation-PLS-DA are described in detail in Section 3.

Many approaches have been invented that attempt to find the interesting information in a cloud of variables – the needle in the haystack. One of the oldest and most varied class of methods for this purpose is variable selection. A large proportion of these methods work univariately, evaluating single variables for inclusion or exclusion. When the number of variables are counted in tens or hundreds of thousands, this strategy will be prone to spurious correlations, hampered by multiple testing problems and vulnerable to low sensitivity or high false discovery rate. Moreover, it can lead to serious misinterpretation: Assume e.g. that the regressor set contains both an "upstream", causally important variable observed with much noise and a "downstream" consequential but unimportant variable observed with little noise, and that the two are strongly intercorrelated. Traditional stepwise variable selection methods will then eliminate the causally important variable to reduce the collinearity.

Subspace-based regression methods such as PCR and PLSR attain an implicit variable selection - not by eliminating individual variables, but by eliminating subspace dimensions- i.e. linear combinations of variables. However, if the number of noisy regressor- or regressand-variables is very high compared to the number of observations, this basic bilinear approach is not good enough: The combined covariation contributions of the noisy variables prevent the bilinear regression methods from finding a useful initial subspace. Therefore, various variable selection strageties have been developed also for PLSR to improve prediction and to simplify interpretation, but without eliminating interesting variables just to reduce collinearity.

One approach is to reduce small parameters towards zero by a general shrinking/expansion of the PLS loading weight elements according to a chosen exponent (Powered PLS[2, 3]). Another approach is to induce sparseness in the data by forcing contributions close to zero to be true zeros. Examples of such methods are the least absolute shrinkage and selection operator (LASSO) [4] and its spin-off the elastic net [5], both inducing constraints on the L_1 norm of the regression vector β. The latter method also applies ridging by penalizing the L_2 norm of β. For PLSR sparseness was introduced by Martens & Næs (1989, p. 160), who suggested the use of rough statistical significance testing of the elements in each individual loading weight vector, followed by a re-orthogonalization. A similar approach was implemented in terms of the soft-threshold-PLS [6] (ST-PLS) and sparse PLS [7] (sPLS). These methods apply a shrinkage towards zero to the PLS loading weights so that many contributions become zero. The amount of shrinkage can be chosen to remove a certain proportion of the variables or it can be chosen by some other criterion. In addition to giving a multivariate approach to variable selection, these methods can also select different variables in each PLS component that is produced. As these two methods, ST-PLS and sPLS, are very similar in the single response case, we choose to compare our method to ST-PLS, as the R-implementation of this method fits models much faster than the sPLS version. We propose to combine the sparseness ideas with the distributional quality of noise in data, e.g. in PLS loading weights, to sort between noise and signal and thereby weighting down or completely truncating what is classified as noise.

In addition to several of the mentioned sparse methods we will include variable selection by the Variable Influence on Projection [8] (VIP) and Selectivity Ratio plot [9] (SR) methods for comparison. These PLS based methods use different criteria for assessing the importance of variables in regression and classification. We will not go into details about how variables are selected by these methods in this paper, but include them as reference standards.

The distribution based truncation approach to variable selection adds to an already long list of methods for variable selection. As described in this article the selection of variables in this approach is motivated from a well established principle in classical statistics. Furthermore, there is only one tuning parameter which needs to be set for variable selection, which makes the method simple and easy to implement. The statistical foundation and the non-complexity of the new method makes it appealing and easy to understand. However, the predictive performance of prediction methods is typically very dependent on the properties of the data, and there is no uniformly best method for prediction and variable selection. Therefore, it is important to expand the statistical toolbox, but at the same time it is important to build an understanding of when the various methods work best. In order to do this we compare the predictive performance of the various methods and attempt to interprete the results in light of the multivariate properties of the data.

3 Methods

Distribution assumptions

In the following the Truncation-PLS is based on loading weights from PLS regression, though the concept is applicable also to regular regression coefficients. Further, the approach could similarly be applied to select Y variables, or to PLS scores in order to eliminate non-informative samples, but these aspects are not covered in this paper. When recording output from some kind of spectroscopic/-metric instrument we expect that the absence of a signal results in white (non-informative) noise, while the presence of a signal will produce a systematic deviation from randomness. The same applies to other types of data, e.g. micro arrays, but the distribution of the noise varies. When creating vectors of loading weights in PLS, we compute the first eigenvector of the matrix product $\mathbf{X}'_{\{a-1\}} \cdot \mathbf{Y}_{\{a-1\}}$ (for component number a). If a given X-variable is uncorrelated with the response variable(s) (for possibly deflated matrices) the loading weight for this variable will be a sum over n equally distributed random variables, and by the CLT it will therefore represent random normal noise, at least approximately. For X-variables correlated to the response variable the theoretical distributions of each loading weight will also be asymptotically normal distributed, but with non-zero mean. However, as the correlation increases the distributions will be increasingly skewed. As the true correlation between an X-variable and the response approaches 1, the limiting distribution of the corresponding loading weight will be a chi-square distribution with non-zero expectation. In Figure 1 (left) the theoretical distributions of three non-normalized loading weights (sample size n=20) are illustrated; a centred normal distribution for an uncorrelated X-variable, and two skewed distributions for two X-variables with correlation -0.6 and 0.6 with the response, respectively. In this figure the distributions have been weighted to reflect a situation where 70% of the variables are distributed according to the central noise distribution and 30% are correlated with the response with either the -0.6 or the 0.6 correlation. In a real data application the loading weights of the informative X-variables will follow different skewed distributions. The sample distribution of the weights will therefore represent a mix of several theoretical distributions and not just three as used in Figure 1 (left). An example of a sample distribution of loading weights is given in Figure 1 (right). The main objective in Truncation-PLS is to find lower and upper cut-offs between which it is assumed that the majority of the loading weights represent noise variables. Hence, the problem boils down to finding an estimate of the central normal distribution of loading weights (or at least selected percentiles) in order to distinguish this from the skewed distributions.

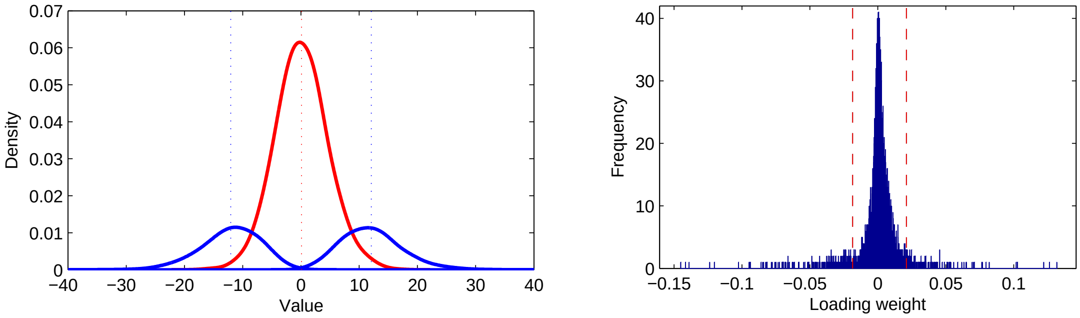


Figure 1: Left: Simulated theoretical distributions of loading weights from X variables with no correlation to the response (red curve, 70 % centred around 0), and correlation of -0.6 and 0.6, respectively (blue curves, 15 % each, centred around -12 and 12, respectively). Right: Histogram of normalized loading weights (milk protein data) illustrates the distributional character of the noninformative loading weights. The red vertical lines indicate the cut-offs between inliers and outliers.

To conform to the classical CLT the observations would need to be independent, but this is not always true in practice. However, CLT theory also exist for observations having weak dependence, and we will only consider the variables where we do not expect any information to be present, supporting independence of these variables.

Algorithm

The idea presented in Section 2 lays the ground for a wide range of possible implementations for classifying data as noise or signal based on their distribution. In principle, the truncation may be applied to several different model parameter types - to object scores in Y or X, to Y-loading weights and to X-loading weights. In this paper we focus on the truncation of the X-loading weights, called **w** in the nomenclature of [10]. The main approach will be to make a confidence interval around the median value of a sorted vector, e.g. PLS loading weights, and truncate or down-weight everything that falls inside the interval, see Algorithm 1. The width of the confidence interval will be estimated using theory from Lenth [11]. A second approach will be to make use of a qq-plot, classifying variables close to the straight line going through a chosen pair of quantiles as inliers. Alternatively one could adapt a normal or Student t distribution to the same vector by direct fitting to the selected distribution, but this can be a time consuming and unstable procedure. The variations have in common that outliers are considered true information, while observations within a certain range of the distribution are classified as noise. In the histogram of loading weights in Figure 1 (right) the estimated cut-offs between inliers and outliers are indicated. The general distribution based truncation algorithm is as follows:

Algorithm 1 General distribution based truncation for a given component

```
• Input candidate loading weight vector $\mathbf{w}$ to be truncated.
• Sort $\mathbf{w} \Rightarrow \mathbf{w}_s$.
• Either
    – compute a confidence interval around the median of $\mathbf{w}_s$, or
    – fit a line through quantiles around the median of $\mathbf{w}_s$.
• Classify outliers as real, informative contributions and inliers as noise.
• Truncate inliers.
```

In practice the distribution based truncation can be plugged into the NIPALS [12] algorithm or kernel based algorithms as a component-wise processing of the candidate PLS loading weights to impose sparseness on the variables, or even truncate the scores to impose sparseness on the objects. In this paper we limit the applications to the single response case, but the procedures are equally relevant in multi-response problems, as well as other multivariate methods like LPLS, PCA, ICA and CCA. Truncation of loading weights will be relevant for most applications as it is more likely that some variables do not contribute to a component than that a set of objects do not contribute. When truncating only loading weights, the following computation of scores ensures that loading weights and scores reflect the same information. If scores are truncated, this will not be reflected in the information of the loading weights, meaning that a re-computation of loading weights and scores may be necessary based on the truncation generated from the scores, or loading weights have to be disregarded when analysing the resulting model. As suggested by Martens & Næs, one could also re-orthogonalize the vectors of loading weights if orthogonality is considered important. Re-orthogonalization may introduce shadowing effect from previous component such that some zero loading weights become non-zero. For the data sets we are using in this paper the changes in regression coefficients are very small with or without re-orthogonalization, and the predictions are equal since the non-orthogonalized and orthogonalized loading weights span the same predictor space.

Instead of applying hard thresholding, where inliers are set to zero and outliers are kept as they are, it could be valuable to shrink according to the probability of being an inlier or outlier. Such a soft shrinkage could be $1 - P(\mathbf{x}_j = \mathit{inlier})$, but estimating this probability would require estimates of the distributions of the outliers. Instead we apply a cumulative distribution function on the observed variables and rescale so that the median is given weight 0 and the largest outlier is given weight 1. As this strategy gives rather poor distinction between inliers and outliers we introduce a parameterized version of these weights to produce weights that are closer to a hard cut-off as illustrated in Figure 2.

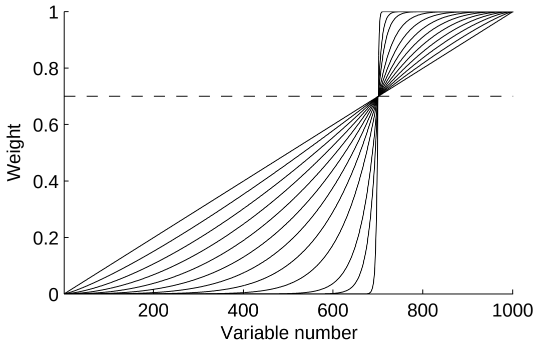


Figure 2: Transformation of scaled weights for gradually steeper transition between inliers and outliers. For this example the weight corresponding to the cut-off between inliers and outliers is set to 0.7.

3.1 Cut-off determination

In order to find cut-offs between inliers and outliers an estimate of the central normal distribution of inliers is needed. Since the distribution is centered in zero the distribution will be fully characterized by an estimate of its variance. In order to distinguish the central noise distribution from the non-central distributions of the informative outliers, a mixture model approach could be adopted. For instance, [13] presented a mixture model approach for sample size determination with false discovery rate control for high-throughput data problems, and a similar approach could be adopted here. However, estimating a set of central and non-central distributions involves iterative procedures (like the EM-algorithm) which would seriously slow down the fitting process of the PLS regression model. Further, only the variance of the central noise distribution is needed, not the properties of the non-central distributions.

A similar problem arises in the analysis of saturated ANOVA models for 2^k-designs without replicates. Then all degrees of freedom are consumed in the estimation of the effects and no conventional error variance estimate can be computed. Still, all effect estimates have the same variance, but a set of non-important effects have zero-expectation. From these a variance estimate for significance testing can be found by the method presented by Lenth [11]. In order to estimate the variance Lenth uses the fact that the standard deviation of a central normal distribution is tightly connected to the median of the absolute value of the random variable. Since the median is rather robust against the influence from outliers, this variance estimate will be only moderately affected by the outliers as long as the majority of the effects (or loading weights in our case) are samples from the central noise distribution. In the setting of this paper the approach of Lenth can be described as follows:

Let $w_1, w_2, ..., w_p$ represent the loading weights computed from the p X-variables at step a of the PLS algorithm. Further, define $s_0 = 1.5 \cdot median\,|w_k|$ for $k = 1, ...p$. It can be shown that s_0 is a fairly good estimate of the standard deviation of the normal distribution of the inliers. In order to make it even more robust and less biased Lenth recommends to make the final estimate, the *pseudo standard error* (PSE), based on a set of inlying values only:

$PSE = 1.5 \cdot \underset{|w_k|<2.5 \cdot s_0}{median} |w_k|.$

Lenth argues that if the w_k are realizations of a $N(0, \tau^2)$ random variable W, the median of $|W|$ is approximately 0.675τ, implying that $1.5 \times median\,|W| \approx 1.01\tau$. And since $Pr(|W| > 2.5\tau) \approx 0.01$, the

PSE is roughly consistent for 1.5 times the $0.495th$ quantile of $|W|$, which is $1.5 \times 0.665\tau \approx \tau$.

The PSE can be combined with a Student t quantile of $d = p/3$ degrees of freedom to give a conservative margin of error (ME) for confidence intervals: $ME = t_{0.975;d} \cdot PSE$ (95% confidence). However, in high-throughput data problems the degrees of freedom will usually be large, and percentiles from the standard normal distribution may be used instead. In the PLS algorithm the cut-offs are thus defined by the limits of a $(1-\alpha)100\%$ confidence interval around the median loading weight with margins of error as described above: $median(\mathbf{w}) \pm ME$, for some chosen confidence level $(1-\alpha)$.

If there is a large asymmetry in the number of positive and negative outliers, the skewness in the distribution of $\mathbf{w}$ may cause ME to be slightly inflated causing a potential loss of informative outliers detected in the lighter tail. This can be avoided by estimating the margin of error separately for positive and negative loading weights. This is accomplished by first finding s_0^- and PSE^- using the absolute values of the negative weights and then computing the marginal error ME^- for the lower tail. Then the same exercise is conducted for the positive loading weights finding s_0^+, PSE^+ and finally ME^+ for the upper tail. Finally, the cut-offs are defined by $ME = min(ME^-, ME^+)$. The increased flexibility can improve the estimation of boundaries between inliers and outliers when there is asymmetry in the distributions. In the rest of this paper we refer to truncation using Lenth's methods as Lenth.

3.2 Outlier detection by qq-plots

An alternative to the above strategy is to use a qq-plot (quantile-quantile plot) as basis, extending an interval around the median value of $\mathbf{w}_s$ minimising the mean squared error (MSE) to the line going through selected quantiles (qq-line), e.g. the 25-th and 75-th percentile of the Student t distribution or normal distribution, see Figure 3. To favour solutions having many inliers the MSE is weighted with the ratio between the total number of points and the number of non-informative inliers $(\frac{n_{tot}}{n_{in}})$. Alternatively one can favour solutions with few informative outliers with MSEs that are not significantly worse than the minimum MSE. Utilizing functions based on golden section search with parabolic interpolation, or similar, the MSE minimization can be solved quickly as a linear search, or a series of such in cases of asymmetry. Visualisation of the sorted $\mathbf{w}$ vector plotted against the final distribution, e.g. Figure 3, can aid in validating and justifying the final truncation.

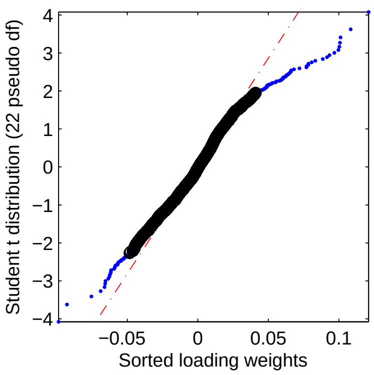


Figure 3: qq-plot of the first vector of loading weights (colon cancer data) against a Student t distribution with 22 pseudo degrees of freedom. Small dots indicate outliers while larger dots indicate inliers. The line going through the 20-th and 80-th percentiles is indicated in dot-dashed form.

When using a Student-t distribution the number of degrees of freedom needs to be specified. Calculating exactly how many degrees of freedom that are consumed by a PLS component is not trivial, but a rough estimate is the following leverage-based estimate (pseudo degrees of freedom): $\sum_i \frac{\mathbf{t}_a^2}{\underset{i}{max}(\mathbf{t}_a^2)}$, where $\mathbf{t}_a$ is the a-th PLS-score vector and i is the sample number. As the truncation is robust to changes in number of degrees of freedom, we do not need the exact degrees of freedom. Note that the number of degrees of freedom consumed will change after truncation. In the rest of this paper we refer to truncation using qq-plots as qq-line.

Note that for both the Lenth and the qq-line method the number of variables selected as informative may vary from one component to another. Furthermore, the same variable may be selected in several components. Hence, the total number of selected variables may not be set exactly, but can be to some extent controlled by the number of PLS-components and the chosen width of the interval around the median weight.

3.3 Reference methods

The truncation procedures are compared to ST-PLS, Elastic net, variable selection by VIP and SR, and PLS without any modifications. This is a small subset of representative methods. For more PLS based variable selection methods we recommend the papers of Mehmood et al. [14] and Roger et al [15]. To make comparisons fair we optimize each method separately with regard to classification/prediction. The performance of each method is evaluated on test set data or by cross-validation in terms of classification errors for the classification problems and root mean square error of prediction (RMSEP) for the prediction problems. With the Elastic net the optimization is performed over a reasonable grid of ridging values (0.1 to 1, where the value 1 gives the Lasso) and L_1 shrinkages (automatically chosen [16]). The shrinkage of ST-PLS is varied over a relevant range (0.05 to 0.95), and the cut-off for VIP is varied from 0.8 to 1.2 [17]. For SR we optimize the cut-off between 0.05 and 0.5, as the cut-off suggested by the authors (0.5) selects too few variables to obtain good predictions on the data sets tested in this paper. Because there are so many models, not all parameter combinations will be reported.

There are several sparse PLS regression methods to chose between, but we found that their resulting variable selections were quite similar, especially when optimizing the sparseness parameter with regard to prediction. We have selected ST-PLS [6] as a common representative, though any of [7, 18, 19] would have been a good alternative.

In addition to the results associated with parameters giving the lowest prediction errors we will present models that have slightly higher prediction errors but give more sparse loading weights and regression coefficients (simplified models). For the data sets where repeated cross-validation is used, the simplified models should have no more than one standard error higher prediction error, while for the data sets where test set prediction is used common additions to the error of 0.001 and 0.01 are used (see the Results section).

4 Examples

4.1 Data sets

The distribution based truncation method for variable selection is compared to the reference methods on both a set of real data sets and to simulated data. These data sets represent a wide range of high-dimensional data types with different properties, and the results will be discussed in light of these. In order to summarize the data properties we use the approach of Helland and Almøy [20] and Sæbø et al. [6] who study the eigenvalue structure of the sample covariance matrix of the predictors and the covariance between the principal components and the response. In the following we refer to the latter property as the relevance of a latent component, following the notation of Næs and Helland [21]. We summarize the data structures in eigenvalue-covariance plots. Helland and Almøy [20] conclude in their study that prediction, using PLSR methods at least, is most difficult in cases where there are irrelevant components having large eigenvalues, or contrary, if there are relevant components having small eigenvalues. In these cases we therefore expect that variable selection methods based on latent components will be less favourable.

4.1.1 Simulated data

These are simulated data containing two correlating, informative features and a variable number of uninformative variables as described in [22, 23]. The total number of variables range from 100 to 20 000, and the number of observations in each of two classes are 100 and 50 for the calibration and validation data, respectively. The simulation study is replicated exactly to be comparable to the papers it has appeared in previously.

4.1.2 Colon cancer data

These are expression levels of 2000 genes on 62 patients as presented by Alon et al. [24]. Among the patients 20 were healthy while 42 had colon cancer. As can be seen from Figure 4 there are several large eigenvalues which indicate several directions in the predictor space of large variance. At the same time these directions appear to be relevant for prediction by having large covariances with the response. Hence, prediction using PLS based methods should be relatively easy, but might require a few components.

4.1.3 Prostate cancer data

These are expression levels of 12 600 genes on 102 patients as presented by Singh et al. [25]. Among the samples 52 were tumor specimens and 50 were normal. From Figure 4 we observe a rapid drop in eigenvalues implying strong dependence between the predictor variables. However, some directions of small variability (small eigenvalues) have some of the largest covariances with the response. This is an example of a data set where there are relevant components with small eigenvalues which according to Helland and Almøy [20] is not favourable for PLS prediction. We therefore expect that the PLS-based variable selection methods will not perform well for this data set.

4.1.4 Fish oil data

These are Raman spectra from 45 oil samples extracted from farmed salmon (Salmo salar) [26]. Raman spectroscopy with a UV laser has been conducted. As a fat indicator the iodine value has been chosen as the response for regression. The spectra are pre-processed by asymmetric least squares [27] ($\lambda = 7$, $p = 0.11$ [28]) wrapped in a customized baseline correction [29] to reduce baseline flexibility under a broad cluster of peaks. The spectra have been cut down to 2263 wavlengths to remove artifacts at the ends of the spectra. These data have a structure resembling the colon data with several directions in the predictor space with high variability and high relevance. Prediction should be relatively easy using a few components in the PLS model.

4.1.5 Milk protein data

These are matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) spectra from 45 milk mixtures (x 4 spot replicates) of cow, goat and ewe milk [3]. Another set of 45 mixtures from a technical replicate is used as validation data. Spectral values from 5000 m/z to 20 000 m/z (6179 variables) are used for predicting the percentage of cow milk in the mixtures, i.e. the degree of adulteration. If the truncation procedure is plugged into canonical PLS (CPLS) [30], the percentage of goat and ewe milk can be used as additional responses to obtain more parsimonious solutions. The eigenvalues for these data imply strong variable dependence with one or two relevant components. Prediction should be quite easy with few components using PLS regression.

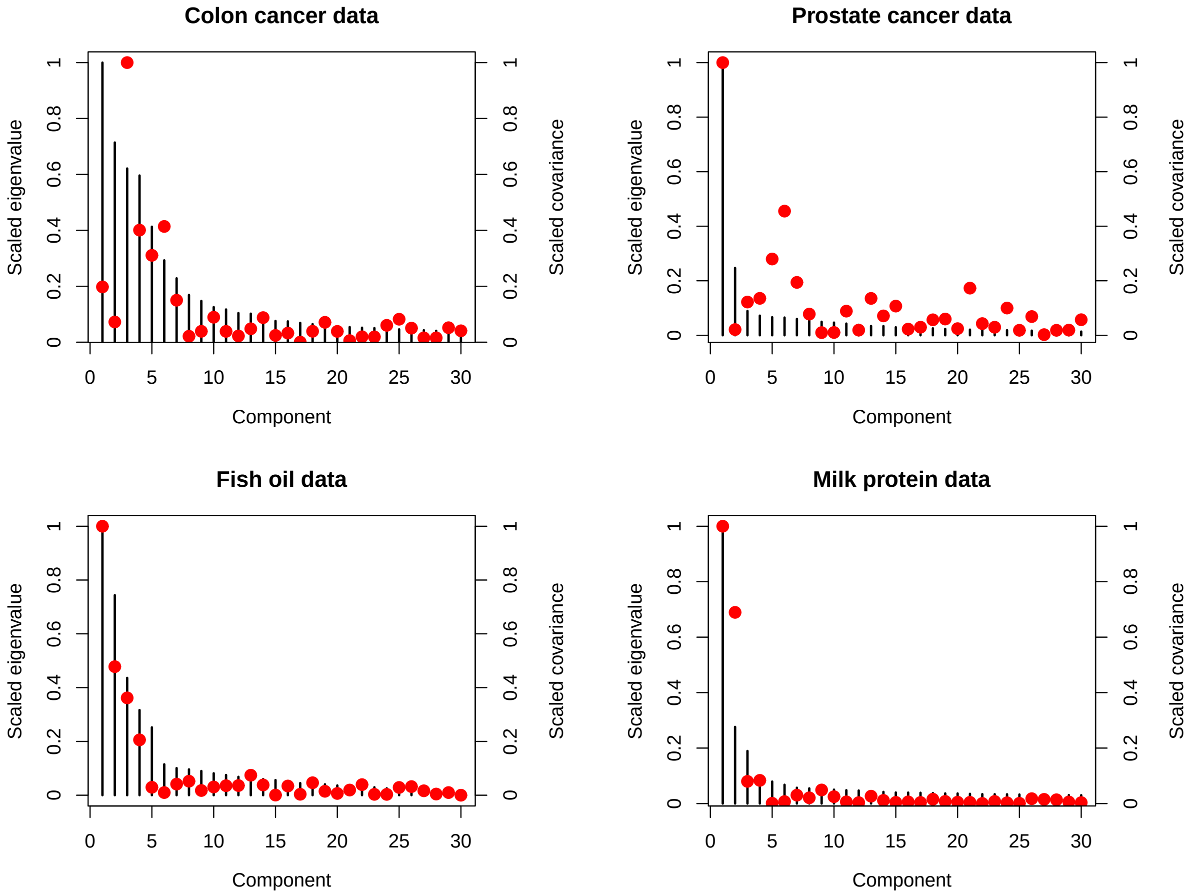


Figure 4: Summaries of data properties for the real data sets. Eigenvalues of the sample covariance matrix (scaled by the largest) are marked by the height of bars. Covariances (scaled by the largest) between principal components and the response are marked by red dots.

4.2 Results

4.2.1 Simulated data

Following the proposed simulation scheme of [22] as was done with PLS and sPLS in [23], we obtain the results shown in Figure 5. Choosing two different widths of the confidence intervals of Lenth's method we find classification errors almost identical to what was shown using sPLS and greatly improved compared to the conventional PLS regression. However, the widest Lenth confidence interval (99.9 %) gives almost perfect classification regardless of number of uninformative variables. These optimistic results are caused by a simulation procedure that highly favours sparse modelling methods, and so should not be over-interpreted.

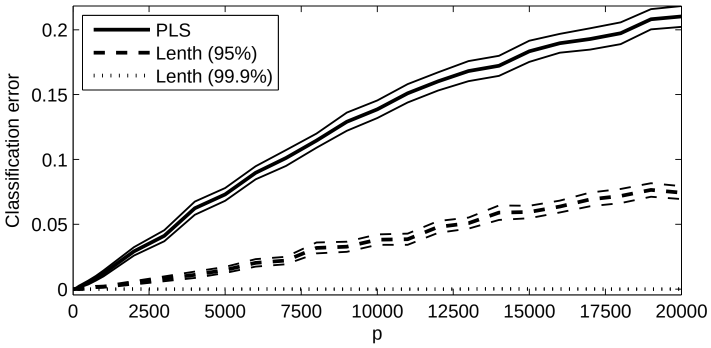


Figure 5: Classification error of two class simulated data. Two regressor variables are informative for the regressand variable, while the total number of regressor variables are indicated on the first axis as p.

4.2.2 Colon cancer data

Figure 6a shows the average classification error of patients from 200 random 10-fold cross-validations [31]. Linear discriminant analysis with empirical priors is used for the classification. It is evident that one component is not enough to obtain good classification regardless of the PLS method used. Elastic net performs approximately at the same level as the one-component PLS variants. The ST-PLS and qq-line Truncation-PLS have the best combinations of few non-zero variables and low classification error (bottom left corner of the figure). The VIP ans SR methods with two and three PLS components have a slightly worse combination of sparseness and error, together with Lenth and Weighted Lenth.

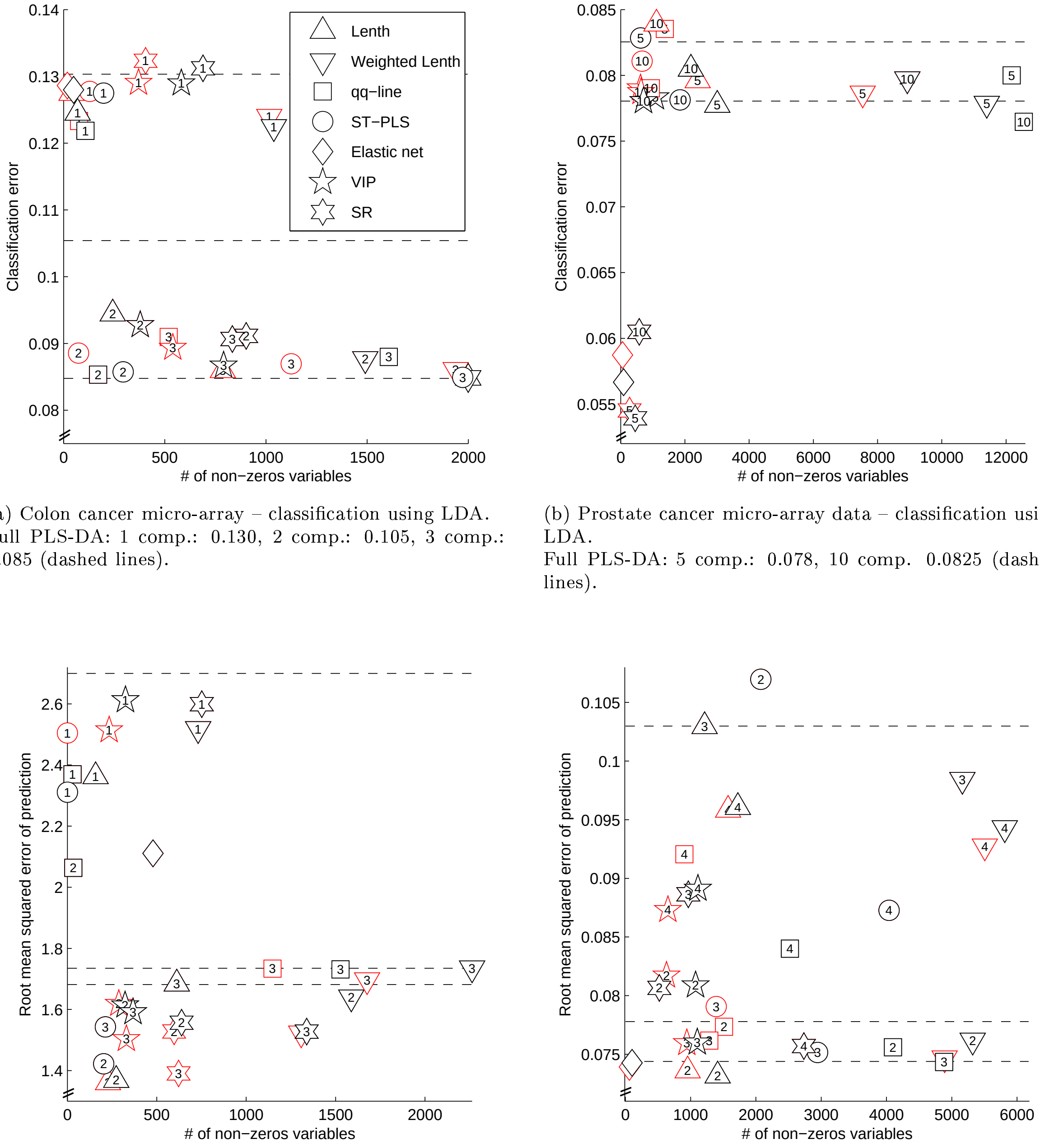
We also observe that choosing a model with slightly higher error than the best model can greatly reduce the number of non-zero variables, especially for Lenth's method. Depending on the aim of the analysis, e.g. variable selection or stable predictions, the choice of truncation type and parameter settings may differ, especially since all the presented models using two and three components lie within a 1 % error margin.

The most sparse two component models (average number of non-zero variables in parentheses) are ST-PLS (74, simplified model), qq-line (171), Lenth (243) and ST-PLS (294). All of these models have a higher average precision compared to the ordinary two component PLS solution, and are very close to the precision of the three component PLS solution.

4.2.3 Prostate cancer data

Figure 6b shows the average classification error of patients from 100 random 10-fold cross-validations. We observe that the best predictions are found when using 5 component PLS models with variable selection by SR. Following closely is the Elastic net. Both of these methods give very sparse solutions. There is almost a 2 % gap down to the rest of the methods. Here variable selection by VIP, qq-line (simplified model), ST-PLS and Lenth give the most sparse solutions while Weighted Lenth gives marginally better classification.

For this data set it seems that the small variation in the discriminating information favours Elastic net and SR while the sparse PLS methods and VIP obtain proportions correctly classified similar to only using PLS with all variables.



(a) Colon cancer micro-array – classification using LDA. Full PLS-DA: 1 comp.: 0.130, 2 comp.: 0.105, 3 comp.: 0.085 (dashed lines).

(b) Prostate cancer micro-array data – classification using LDA.
Full PLS-DA: 5 comp.: 0.078, 10 comp. 0.0825 (dashed lines).

(c) Fish oil Raman data – prediction of iodine.
Full PLSR: 1 comp.: 2.70, 2 comp.: 1.68, 3 comp. 1.74 (dashed lines).

(d) Milk protein MALDI-TOF data – prediction of adulteration.
Full PLSR: 1 comp.: 0.103, 2 comp.: 0.074, 3 comp.: 0.078 (dashed lines).

Figure 6: Repeated random 10-fold cross-validated classification (subfigures a and b) and test set predictions (subfigures c and d) using varying numbers of PLS components. The symbols indicate different variable selection strategies and their numbers of components. Black symbols are associated with the parameters giving the highest precision, while red symbols indicate models using fewer variables while retaining most of their precision.

4.2.4 Fish oil data

In Figure 6c we see the results of test set predictions using the same methods as above. Parameters have been chosen by cross-validation. The best combination of prediction and sparseness is observed for Lenth and ST-PLS. Precisions of these predictions are much better than only using PLS. The RMSEP values from Elastic net are somewhere between the one component PLS models and the two/three component models. As the parameters and simplifications are chosen on the cross-validation results, we observe both reductions and increases in RMSEP when using simplified models.

4.2.5 Milk protein data

In addition to comparison with the reference methods this data set is included both to show how one can obtain parsimonious models by plugging the truncation algorithm into a different NIPALS algorithm, the canonical PLS, and to show how interpretation of spectral data can be made easier by imposing sparseness. The CPLS algorithm differs from the regular PLS in the way that additional sample information (like design variables) may be included as extra response variables to stabilize the extraction of the latent components. This has the typical effect that the number of components is reduced compared to PLS regression. As mentioned in the description of the data the percentage of goat and ewe milk was included as additional responses in the analysis of the cow milk data. In Figure 6d we see the results of test set predictions using the same methods as above. Parameters have been chosen by cross-validation. Here Elastic net is the winner considering the combination of prediction and sparseness. However, prediction-wise the other methods are very close behind. Among the PLS based methods, Lenth has the best combination of prediction and sparseness, having marginally better prediction than Elastic net using less than 1/6 of the variables with the simplified model.

Figure 7 shows the prediction error of PLS and CPLS regression used separately and combined with a pre-chosen truncation (99.9 % confidence interval (Lenth's method) with sharp cut-off). We observe that for models using few components truncation has no effect on prediction with PLS, but gives a minor improvement when combined with CPLS. Also, CPLS has much lower prediction error for one and two component models. Looking only at prediction, the best balance between prediction error and complexity is a two component CPLS model with truncation.

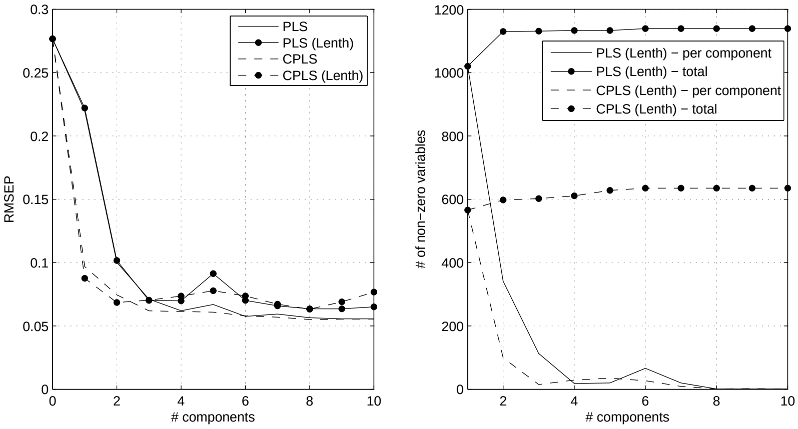


Figure 7: Prediction of cow milk proportions in milk mixtures from MALDI-TOF spectra (left) and the number of non-zero variables per component/in total using truncation (right). The total number of variables was 6179.

In Figure 8 we see the first two vectors of loading weights from PLS and CPLS regression with and without truncation. The contrast is high with a high level of noise in the upper spectra and only a few remaining peaks in the lower spectra. Here the truncated spectra seem to have an advantage when used for interpretation and protein assignment.

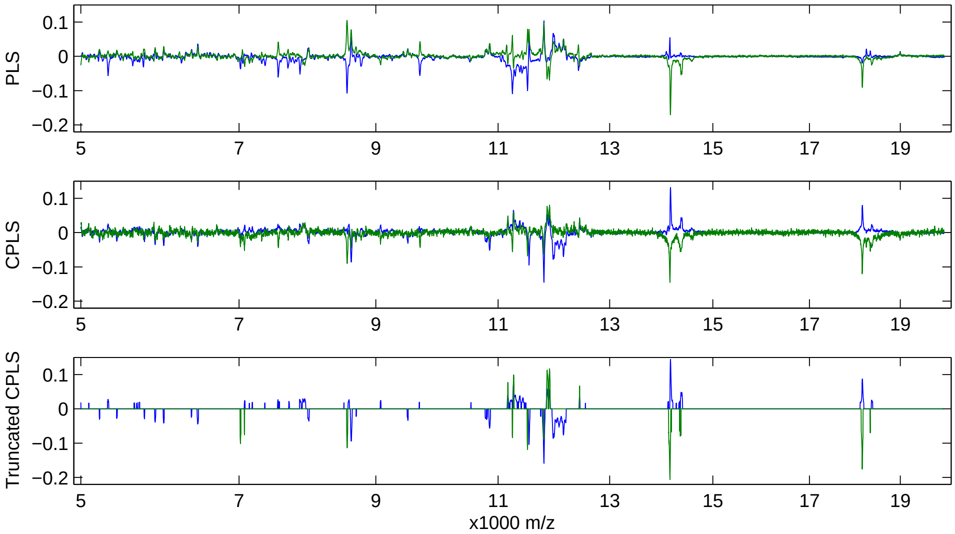


Figure 8: Loading weight vectors from MALDI-TOF spectra of milk (two first components). The top spectra come from ordinary PLS, the middle spectra from CPLS, while the bottom spectra come from truncated CPLS with truncation parameters selected to reflect a typical choice applicable for many types of data.

5 Discussion

Through this paper we have formalised some aspects of the family of sparse PLS methods. Firstly we have have justified truncation of loading weights through the central limit theorem and the distributions of loading weights with no correlation to the response. Secondly we have proposed a new truncation founded on classical statistical asymptotic principles. This is introduced through a novel application of Lenth's theory for creating confidence intervals in saturated ANOVA models for 2^k-designs without replicates. The effect is that the user only has to choose a significance level for the confidence interval, resulting in a less ad hoc approach.

Truncation in this paper is achieved using a general and flexible plug-in which can easily be adjusted and implemented also in other projection based methods like PCA [32], ICA [33], PCR, CPLS and PPLS. PLS regression is an iterative algorithm and component wise truncation will inevitably slow down the algorithm, but Lenth's method is extremely quick, i.e. there is a minimal lag compared to just running regular PLSR. The alternative approach based on the qq-line is also quite quick, and appears to give slightly better results in some situations.

With regard to prediction performance the truncation PLS is mostly on par with ST-PLS, sometimes a little better, sometimes a little worse. As with all statistical methods, this is highly data dependent. However, there are few parameters to tune and they have statistical interpretations. For the data sets included in this paper we see that Elastic net sometimes performs significantly better than the sparse PLS methods, while it trails behind when used on other data sets. This is also the case for the variable selection by Selectivity Ratio plots and to some extent the Variable Influence on Prediction method. The Lasso was also tested with the included data sets, but being a special case of the Elastic net it never performed better in practice. But prediction is not the only goal for a statistical method. The truncation methods have also shown consistent good results, are based on intuitive theory, are quite robust to the choice of parameters and are extremely quick.

The performance of the various methods may to some extent be explained by the structure of the data. The PLS-based methods perform relatively better when there are many directions in the predictor space with both a high variance (high eigenvalue) and a high relevance. This was the case for both the colon cancer data and the fish oil data, and here also the PLS-based variable selection methods performed well, with the new truncation method and ST-PLS slightly ahead of the others. For the prostate data these methods performed worse, and this result confirms the expectations based on the data properties that PLS methods have trouble making good predictions for this kind of data where there are directions in the predictor space of low variance, but with high relevance. However, an exception is the SR method based on the 5 component PLS model. This can be explained by the fact that the SR method is adjusted to be more favourable than ordinary PLS when there are variables with low variances, but with high correlations with the response [9]. This is exactly what is the case here according to Figure 4. Apparently the elastic net has a similar behaviour, which can be explained by the fact that this method, like the ordinary least squares, gives higher weight to variables with high correlations to the response, as opposed to the more covariance-focused PLS. The results indicate that in cases where there is a strong correlation structure in the data (prostate cancer data and milk protein data) the elastic net is a good choice of method for variable selection. When choosing a method for analysis and variable selection it may therefore be worthwhile to study the data properties in terms of eigenvalues and component-response covariances.

One side-effect of applying truncation to vectors of loading weights is that they are no longer orthogonal. In most applications, small deviations from orthogonality can be disregarded. However, when orthogonal vectors of loading weights is important, a re-orthogonalization step can be included after the truncation, forcing the current vector of loading weights to be orthogonal to the previous vectors extracted. The down-side to this is that shadow effects from previous loading weights may appear in the re-orthogonalized loading weights, causing zero weights of regressors already used in previous components to become non-zero. For the data sets we have used in this paper, the shadow effect was so small that they were invisible in plots, and only appeared a few times in measurable sizes. The total number of non-zero regression coefficients should not be affected.

A note should be made on the different roles of the X loading weights, $\mathbf{w}_a$, and the X loadings, $\mathbf{p}_a$. It is important to remember that the loading weights contain the covariance information between $X_{\{a-1\}}$ and $Y_{\{a-1\}}$ (the first eigenvector of the covariance matrix if Y is multi response) and give us the weights that each explanatory variable has when creating scores and loadings. The scores, $\mathbf{t}_a$, are just linear combinations of the explanatory variables weighted by the loading weights. The loadings, however, are found by projecting each explanatory variable of $X_{\{a-1\}}$ on the scores, $\mathbf{t}_a$. Loading weights and loadings can look quite similar when no truncation has been applied, especially for spectroscopic data. With truncation, however, the loading weights obtain a lot of "zero holes", while the loadings retain a more continuous shape (at least for spectroscopic data). The upshot is that fully truncated variables are not completely lost, and their role in the system may be interpreted graphically since their loadings are intact. Depending on the application, either loading weights or loadings can be interpreted, having roles similar to the regression coefficients with and without "zero holes".

In some applications it may be interesting to apply truncation without ending up with zeros in the resulting regression coefficients, analogous to focusing on loadings instead of loading weights. This can be justified by the need to remove noise in the computation of PLS components and at the same time producing continuous regression coefficients. From the early days of PLSR we find approximate estimates of regression coefficients that produce the desired effect. Two alternatives have been proposed. Firstly the approximated regression coefficients can simply be estimated by the product of the X and y loadings: $\hat{\beta}^{\dagger} = Pq'$. A more elaborate strategy is to produce new approximated X scores, y loadings and regression coefficients by full projection on the X loadings: $T^{\star} = XP(P'P)^{-1}$, $q^{\star} = y'T^{\star}(T^{\star\prime}T^{\star})^{-1}$, and finally: $\hat{\beta}^{\star} = Pq^{\star\prime}$. Both strategies will produce regression vectors without "zero holes".

References

[1] Wold, S., Martens, H. & Wold, H. The multivariate calibration problem in chemistry solved by the PLS methods. *Lecture notes in mathematics* **973**, 286 – 293 (1983).

[2] Indahl, U. A twist to partial least squares regression. *Journal of Chemometrics* **19**, 32 – 44 (2005).

[3] Liland, K. H., Mevik, B.-H., Rukke, E.-O., Almøy, T. & Isaksson, T. Quantitative whole spectrum analysis with MALDI-TOF MS, Part II: Determining the concentration of milk in mixtures. *Chemometrics and Intelligent Laboratory Systems* **99**, 39 – 48 (2009).

[4] Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B* **58**, 267 – 288 (1996).

[5] Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B* **67**, 301 – 320 (2005).

[6] Sæbø, S., Almøy, T., Aarøe, J. & Aastveit, A. H. ST-PLS: a multi-directional nearest shrunken centroid type classifier via pls. *Journal of Chemometrics* **22**, 54 – 62 (2008).

[7] Lê Cao, K., Rossouw, D., Robert-Granié, C. & Besse, P. A sparse pls for variable selection when integrating omics data. *Statistical applications in genetics and molecular biology* **7** (2008).

[8] Wold, S., Johansson, E. & Cocchi, M. *3D QSAR in drug design: theory, methods and applications* (ESCOM Science Publishers B.V., Leiden, The Netherlands, 1993).

[9] Rajalahti, T. *et al.* Discriminating variable test and selectivity ratio plot: Quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles. *Analytical Chemistry* **81**, 2581 – 2590 (2009).

[10] Martens, H. & Næs, T. *Multivariate calibration* (John Wiley and Sons, Chichester, UK, 1989).

[11] Lenth, R. V. Quick and easy analysis of unreplicated factorials. *Technometrics* **31**, 469 – 473 (1989).

[12] Wold, H. *Estimation of principal component and related models by iterative least squares*, vol. Multivariate analysis (Academic Press, New York, USA, 1966).

[13] Jørstad, T., Midelfart, H. & Bones, A. A mixture model approach to sample size estimation in two-sample comparative microarray experiments. *BMC Bioinformatics* **9** (2008).

[14] Mehmood, T., Liland, K. H., Snipen, L. & Sæbø, S. A review of variable selection methods in partial least squares regression. *Chemometrics and Intelligent Laboratory Systems* **118**, 62 – 69 (2012).

[15] Roger, J., Palagos, B., Bertrand, D. & Fernandez-Ahumada, E. Covsel: Variable selection for highly multivariate and multi-response calibration: Application to IR spectroscopy. *Chemometrics and Intelligent Laboratory Systems* **106**, 216 – 223 (2011). URL http://www.sciencedirect.com/science/article/pii/S0169743910001978.

[16] Friedman, J. & Hastie, T. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33** (2010).

[17] Chong, I. & Jun, C. Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems* **78**, 103 – 112 (2005).

[18] Chun, H. & Keleş. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**, 3–25 (2010).

[19] Lee, D., Lee, W., Lee, Y. & Pawitan, Y. Sparse partial least-squares regression and its applications to high-throughput data analysis. *Chemometrics and Intelligent Laboratory Systems* **109**, 1 – 8 (2011).

[20] Helland, I. S. & Almøy, T. Comparison of prediction methods when only a few components are relevant. *Journal of the American Statistical Association* **89**, 583 – 591 (1994).

[21] Næs, T. & Helland, I. S. Relevant components in regression. *Scandinavian Journal of Statistics* **20**, 239 – 250 (1993).

[22] Qiao, Z., Zhou, L. & Huang, J. Z. Sparse linear discriminant analysis with applications to high dimensional low sample size data. *International Journal of Applied Mathematics* **39**, 48 – 60 (2009).

[23] Filzmosera, P., Gschwandtnera, M. & Todorov, V. Review of sparse methods in regression and classification with application to chemometrics. *Journal* **26**, 42 – 51 (2012).

[24] Alon, U. *et al.* Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *P. Natl. Acad. Sci.* **96**, 6745 – 6750 (1996).

[25] Singh, D. *et al.* Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1**, 203 – 209 (2002).

[26] Afseth, N. K., Segtnan, V. H. & Wold, J. P. Raman spectra of biological samples: A study of preprocessing methods. *Applied Spectroscopy* **60**, 1358–1367 (2006).

[27] Eilers, P. H. Parametric time warping. *Analytical Chemistry* **76**, 404–411 (2004).

[28] Liland, K. H., Almøy, T. & Mevik, B.-H. Optimal choice of baseline correction for multivariate calibration of spectra. *Applied Spectroscopy* **64**, 1007 – 1016 (2010).

[29] Liland, K. H., Rukke, E.-O., Olsen, E. F. & Isaksson, T. Customized baseline correction. *Chemometrics and Intelligent Laboratory Systems* **109**, 51 – 56 (2011).

[30] Indahl, U. G., Liland, K. H. & Næs, T. Canonical partial least squares – a unified pls approach to classification and regression problems. *Journal of Chemometrics* **23**, 495 – 504 (2009).

[31] Stone, M. Cross-validatory choice and assesment of statistical predictions. *Journal of the Royal Statistical Society, Series B—Methodological* **36**, 111 – 147 (1974).

[32] Pearson, K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2**, 559 – 572 (1901).

[33] Comon, P. Independent component analysis, A new concept? *Signal processing* **36**, 287 – 314 (1994).