Globally sparse PLS regression

Tzu-Yu Liu, Laura Trinchera, Arthur Tenenhaus, Dennis Wei, Alfred Hero

To appear in 2013 Springer Verlag book series: "New perspectives in Partial Least Squares and Related Methods"

Globally Sparse PLS Regression

Tzu-Yu Liu, Laura Trinchera, Arthur Tenenhaus, Dennis Wei and Alfred O. Hero

Abstract Partial least squares (PLS) regression combines dimensionality reduction and prediction using a latent variable model. It provides better predictive ability than principle component analysis by taking into account both the independent and response variables in the dimension reduction procedure. However, PLS suffers from over-fitting problems for few samples but many variables. We formulate a new criterion for sparse PLS by adding a structured sparsity constraint to the global SIMPLS optimization. The constraint is a sparsity-inducing norm, which is useful for selecting the important variables shared among all the components. The optimization is solved by an augmented Lagrangian method to obtain the PLS components and to perform variable selection simultaneously. We propose a novel greedy algorithm to overcome the computation difficulties. Experiments demonstrate that our approach to PLS regression attains better performance with fewer selected predictors.

Tzu-Yu Liu
Electrical Engineering and Computer Science Department, University of Michigan, USA
e-mail: joyliu@umich.edu

Laura Trinchera
MMIP Departement, AgroParisTech and UMR518 MIA, INRA, France
e-mail: laura.trinchera@agroparistech.fr

Arthur Tenenhaus
Department of Signal Processing and Electronic Systems, Supélec, France
e-mail: Arthur.Tenenhaus@supelec.fr

Dennis Wei
Electrical Engineering and Computer Science Department, University of Michigan, USA
e-mail: dlwei@eecs.umich.edu

Alfred O. Hero
Electrical Engineering and Computer Science Department, University of Michigan, USA
e-mail: hero@eecs.umich.edu

1 Introduction

Partial least squares (PLS) regression combines dimensionality reduction and prediction using a latent variable model. It was first developed for regression analysis in chemometrics [Wold *et al*., 1983], and has been successfully applied to many different areas, including sensory science and more recently genetics. Since PLS-R does not require matrix inversion or diagonalization, it can be applied to problems with large numbers of variables. As predictor dimension increases, variable selection becomes essential to avoid over-fitting, to provide more accurate predictors and to yield more interpretable parameters. For this reason sparse PLS was developed by H. Chun and S. Keles [Chun *et al*., 2010]. The sparse PLS algorithm performs variable selection and dimension reduction simultaneously using an L_1 type variable selection penalty. However, the L_1 penalty used in [Chun *et al*., 2010] penalizes each variable independently and this can result in different sets of variables being selected for each PLS component leading to an excessively large number of variables. In this paper we propose a global variable selection approach that penalizes the total number of variables across all PLS components. Put another way, the proposed global penalty guarantees that the selected variables are shared among the PLS components. This results in improved PLS performance with fewer variables. We formulate PLS with global sparsity as a variational optimization problem with objective function equal to the univariate PLS criterion with added mixed norm sparsity constraint on the weight matrix. The mixed norm sparsity penalty is the L_1 norm of the L_2 norm on the subsets of variables used by each PLS component. A novel augmented Lagrangian method is proposed to solve the optimization problem and soft thresholding for sparsity occurs naturally as part of the iterative solution. Experiment results show that the modified PLS attains better performance (lower mean squared error, MSE) with many fewer selected predictor variables.

2 Partial Least Squares Regression

Partial Least Squares (PLS) methods embrace a suite of data analysis techniques based on algorithms belonging to the PLS family. These algorithms consist of various extensions of the Nonlinear estimation by Iterative PArtial Least Squares (NIPALS) algorithm that was proposed by Herman Wold [Wold, 1966] as an alternative algorithm for implementing a Principal Component Analysis (PCA) [Hotelling, 1933]. The NIPALS approach was slightly modified by Herman Wold son, Svante, and Harald Martens, in order to obtain a regularized component based regression tool, known as PLS Regression (PLS-R) [Wold *et al*., 1983, Wold *et al*., 1984].

Suppose that the data consists of n samples of independent variables $X \in R^{n\times p}$ and dependent variables (responses) $Y \in R^{n\times q}$. In standard PLS Regression the aim is to define orthogonal latent components in R^p, and then use such latent components as predictors for Y in an ordinary least squares framework.The X weights used to compute the latent components can be specified by using iterative algorithms be-

longing to the NIPALS family or by a sequence of eigen-decompositions. Moreover, in the univariate response case, it does not make sense to calculate components in the unidimensional response space. For the k-th component, the X weights can be directly computed as a function of Y. In particular, for the first component the X weights are defined such that the covariance between the predictors and the univariate response is maximized. In both the univariate and multivariate cases, the general underlying model behind the PLS Regression is $X = TP^T + E$ and $Y = TQ^T + F$, where T is the latent component matrix, P and Q are the loading matrices, E and F are the residual terms.

2.1 Univariate response

We assume, without loss of generality, that all the variables have been centered in a pre-processing step. For univariate Y, i.e $q = 1$, PLS Regression, also often denoted as PLS1, successively finds X weights $R = [\,\mathbf{r}_1\ \mathbf{r}_2\ ...\ \mathbf{r}_K\,]$ as the solution to the constrained optimization

$$\mathbf{r}_k = \arg\max_{\mathbf{r}} \{\mathbf{r}^T X_{(k-1)}^T Y_{k-1} Y_{k-1}^T X_{(k-1)} \mathbf{r}\} s.t.\ \mathbf{r}^T \mathbf{r} = 1 \tag{1}$$

where $X_{(k-1)}$ is the matrix of the residuals (i.e. the deflated matrix) from the regression of the X-variables on the first $k-1$ latent components, and $X_0 = X$. Due to the deflation on data after each iteration for finding the weight vector $\mathbf{r}_k$, the orthogonality constraint is satisfied by construction. These weights are then used to find the orthogonal latent components $T = X_{(k-1)}R$. Such components can be also expressed in terms of original variables (instead of deflated variables), i.e. as $T = XW$, where W is the matrix containing the weights to be applied to the original variables in order to exactly obtain the latent components [Tenenhaus, 1998].

For a fixed number of components, the response variable Y is predicted in an ordinary least squares regression model where the latent components play the role of the exogenous variables

$$\arg\min_{Q} \{||Y - TQ^T||_2\} = (T^T T)^{-1} T^T Y \tag{2}$$

This provides the regression coefficients $\hat{\beta}^{PLS} = W\hat{Q}^T$ for the model $Y = X\beta^{PLS} + F$.

Depending on the number of selected latent components the length $\|\hat{\beta}^{PLS}\|_2$ of the vector of the PLS coefficient estimators changes. In particular, de Jong [de Jong, 1995] has shown that the sequence of these coefficient vectors have lengths that are strictly increasing as the number of component increases. This sequence converges to the ordinary least squares coefficient vector and the maximum number of latent components obtainable equals the rank of the X matrix. Thus, by using a number of latent components $K < rank(X)$, PLS-R performs a dimension reduction by shrinking the β vector. Hence, PLS-R is a suitable tool for problems with data containing many more variables p than observations n.

The objective function in (1) can be interpreted as maximizing the squared covariance between Y and the latent component: $corr^2(Y, X_{k-1}\mathbf{r}_k)\text{var}(X_{k-1}\mathbf{r}_k)$. Because the response Y has been taken into account to formulate the latent matrix, PLS has better performance in prediction problems than principle component analysis (PCA) does [De Jong, 2005]. This is one of the main difference between PLS and principle component analysis (PCA) [Boulesteix *et al*., 2007].

2.2 Multivariate response

Similarly to univariate response PLS-R, multivariate response PLS-R selects latent components in R^p and R^q , i.e. $\mathbf{t}_k$ and $\mathbf{v}_k$, such that the covariance between $\mathbf{t}_k$ and $\mathbf{v}_k$ is maximized. For a specific component, the sets of weights $\mathbf{r}_k \in R^p$ and $\mathbf{c}_k \in R^q$ are obtained by solving

$$\max\{\mathbf{t}^T\mathbf{v}\} = \max\{\mathbf{r}^T X_{k-1}^T Y_{k-1}\mathbf{c}\} s.t.\ \mathbf{r}^T\mathbf{r} = \mathbf{c}^T\mathbf{c} = 1 \tag{3}$$

where $\mathbf{t}_k = X_{(k-1)}\mathbf{r}_k$, $\mathbf{v}_k = Y_{(k-1)}\mathbf{c}_k$, and $X_{(k-1)}$ and $Y_{(k-1)}$ are the deflated matrices associated to X and Y. Notice that the optimal solution $\mathbf{c}_k$ should be proportional to $Y_{k-1}^T X_{k-1}\mathbf{r}_k$. Therefore, the optimization in (3) is equivalent to

$$\max_{\mathbf{r}}\{\mathbf{r}^T X_{k-1}^T Y_{k-1} Y_{k-1}^T X_{k-1}\mathbf{r}\} s.t.\ \mathbf{r}^T\mathbf{r} = 1 \tag{4}$$

For each component, the solution to this criterion can be obtained by using a so called PLS2 algorithm. A detailed description of the iterative algorithm as presented by Höskuldsson is in Algorithm 1 [Höskuldsson, 1988].

In 1993 de Jong proposed a variant of the PLS2 algorithm, called Straightforward Implementation of a statistically inspired Modification of PLS (SIMPLS), which calculates the PLS latent components directly as linear combinations of the original variables [de Jong, 1993]. The SIMPLS was first developed as an optimality problem and solve the optimization

$$\begin{aligned} \mathbf{w}_k &= \arg\max_{\mathbf{w}}(\mathbf{w}^T X^T Y Y^T X\mathbf{w}) \\ s.t.\ \mathbf{w}^T\mathbf{w} &= 1,\ \ \mathbf{w}^T X^T X\mathbf{w}_j = 0\ for\ j = 1,...,k-1. \end{aligned} \tag{5}$$

Ter Braak and de Jong [ter Braak *et al*., 1998] provided a detailed comparison between the objective functions for PLS2 in (4) and SIMPLS in (5) and shown that the successive weight vectors $\mathbf{w}_k$ can be derived either from the deflated data matrices or original variables in PLS2 and SIMPLS respectively. Let W^+ be the Moore-Penrose inverse of $W = [\mathbf{w}_1\ \mathbf{w}_2\ ...\ \mathbf{w}_{k-1}]$. The PLS2 algorithm (Algorithm 1) is equivalent to solving the optimization

$$\mathbf{w}_k = \arg\max_{\mathbf{w}}(\mathbf{w}^T X^T Y Y^T X\mathbf{w})$$

for *k=1:K* **do**
  initialize $\mathbf{r}$;
  $X = X_{new}$;
  $Y = Y_{new}$;
  while *solution has not converged* **do**
    $\mathbf{t} = X\mathbf{r}$;
    $\mathbf{c} = Y^T\mathbf{t}$;
    Scale $\mathbf{c}$ to length 1;
    $\mathbf{v} = Y\mathbf{c}$;
    $\mathbf{r} = X^T\mathbf{v}$;
    Scale $\mathbf{r}$ to length 1;
  end
  loading vector $\mathbf{p} = X^T\mathbf{t}/(\mathbf{t}^T\mathbf{t})$;
  deflate $X_{new} = X - \mathbf{t}\mathbf{p}^T$;
  regression $\mathbf{b} = Y^T\mathbf{t}/(\mathbf{t}^T\mathbf{t})$;
  deflate $Y_{new} = Y - \mathbf{t}\mathbf{b}^T$;
  $\mathbf{r_k} = \mathbf{r}$;
end

Algorithm 1: PLS2 algorithm

$$s.t. \mathbf{w}^T(I - WW^+)\mathbf{w} = 1, \mathbf{w}^T X^T X \mathbf{w}_i = 0 \; for \; i = 1, ..., k-1. \tag{6}$$

Both NIPALS and SIMPLS have the same objective function but each are maximized under different constraints. NIPALS and SIMPLS are equivalent when Y is univariate, but provide slightly different weight vectors in multivariate scenarios. The performance depends on the nature of the data, but SIMPLS appears easier to interpret since it does not involve deflation of the data sets [de Jong, 1993]. However NIPALS can manage missing data when SIMPLS needs complete data. We develop our globally sparse PLS based on the SIMPLS optimization formulation.

3 Globally Sparse PLS Regression

One approach to sparse PLS is to add the L_1 norm of the weight vector, a sparsity inducing penalty, to (5). The solution for the first component would be obtained by solving

$$\mathbf{w}_1 = \arg\max_{\mathbf{w}} (\mathbf{w}^T X^T Y Y^T X \mathbf{w}) \; s.t. \; \mathbf{w}^T\mathbf{w} = 1, \; ||\mathbf{w}||_1 \leq \lambda. \tag{7}$$

The addition of the L_1 norm is similar to SCOTLASS (simplified component lasso technique), the sparse PCA proposed by Jolliffe [Jolliffe *et al*., 2003]. However, the solution of SCOTLASS is not sufficiently sparse, and the same issue remains in (7). Chun and Keles [Chun *et al*., 2010] reformulated the problem, promoting the exact zero property by imposing the L_1 penalty on a surrogate of the weight vector

instead of the original weight vector [Chun *et al.*, 2010], as shown in (8). For the first component, they solve the following optimization by alternating between updating $\mathbf{w}$ and $\mathbf{z}$ (block coordinate descent). The L_2 norm addresses the potential singularity problem when solving for $\mathbf{z}$.

$$\mathbf{w}_1, \mathbf{z}_1 = \arg\min_{\mathbf{w},\mathbf{z}} \{-\kappa \mathbf{w}^T X^T Y Y^T X \mathbf{w} + (1-\kappa)(\mathbf{z}-\mathbf{w})^T X^T Y Y^T X (\mathbf{z}-\mathbf{w}) + \lambda_1 ||\mathbf{z}||_1 + \lambda_2 ||\mathbf{z}||_2^2\}$$
$$s.t.\ \mathbf{w}^T \mathbf{w} = 1 \tag{8}$$

As mentioned in the Introduction, this formulation penalizes the variables in each PLS component independently. This paper proposes an alternative in which variables are penalized simultaneously over all directions. First, we define the global weight matrix, consisting of the K weight vectors, as

$$W = \begin{bmatrix} | & | & & | \\ \mathbf{w}_1 & \mathbf{w}_2 & \cdots & \mathbf{w}_K \\ | & | & & | \end{bmatrix} = \begin{bmatrix} - & \mathbf{w}_{(1)}^T & - \\ - & \mathbf{w}_{(2)}^T & - \\ & \vdots & \\ - & \mathbf{w}_{(p)}^T & - \end{bmatrix}$$

Notice that the elements in a particular row of W, i.e. $\mathbf{w}_{(j)}^T$, are all associated with the same predictor variable $\mathbf{x}_j$. Therefore, rows of zeros correspond to variables that are not selected. To illustrate the drawbacks of penalizing each variable independently, as in [Chun *et al.*, 2010], suppose that each entry in W is selected independently with probability p_1. The probability that the $(j)_{th}$ variable is not selected becomes $(1-p_1)^K$, and the probability that all the variables are selected for at least one weight vector is $[1-(1-p_1)^K]^p$, which increases as the number of weight vectors K increases. This suggests that for large K the local variable selection approach of [Chun *et al.*, 2010] may not lead to an overall sparse and parsimonious PLS model. In such cases a group sparsity constraint is necessary to limit the number of selected variables. The globally sparse PLS variable selection problem is to find the top K weight vectors that best relate X to Y, while using limited number of variables.

$$W = \arg\min_W -\frac{1}{n^2} \sum_{k=1}^{K} \mathbf{w}_k^T X^T Y Y^T X \mathbf{w}_k + \lambda \sum_{j=1}^{p} ||\mathbf{w}_{(j)}||_2 \tag{9}$$
$$s.t.\ \mathbf{w}_k^T \mathbf{w}_k = 1\ \forall\ k \ and\ \mathbf{w}_k^T X^T X \mathbf{w}_i = 0\ \forall\ i \neq k$$

The objective function (9) is the summation of the first K terms in the SIMPLS objective. Instead of the sequential greedy solution in PLS2 algorithm, the proposed globally sparse PLS must solve for the K weight vectors simultaneously. The L_2 norm of each row of W promotes grouping entries in W that relate to the same

predictor variable, whereas the L_1 norm promotes a small number of groups, as in (7).

We propose to solve the optimization (9) by augmented Lagrangian methods, which allows one to solve (9) by variable splitting iterations. Augmented Lagrangian methods introduce a new variable M, constrained such that $M = W$, such that the row vectors $\mathbf{m}_{(j)}$ of M obey the same structural pattern as the rows of W:

$$\min_{W,M} -\frac{1}{n^2}\sum_{k=1}^{K}\mathbf{w}_k^T X^T Y Y^T X \mathbf{w}_k + \lambda \sum_{j=1}^{p} ||\mathbf{m}_{(j)}||_2 \qquad (10)$$
$$s.t.\ \mathbf{w}_k^T \mathbf{w}_k = 1\ \forall\ k\ ,\ \mathbf{w}_k^T X^T X \mathbf{w}_i = 0\ \forall\ i \neq k,\ and\ M = W$$

The optimization (10) can be solved by replacing the constrained problem by an unconstrained one with an additional penalty on the Frobenius norm of the difference $M - W$. This penalized optimization can be iteratively solved by a block coordinate descent method that alternates between optimizing over W and over M (See algorithm 2). We initialize the algorithm 2 with $M(0)$ equals to the solution of standard PLS, and $D(0)$ equals to the zero matrix. Once the algorithm converges, the final PLS regression coefficients are obtained by applying the standard PLS regression on the selected variables keeping the same number of components K. The optimization over W can be further simplified to a secular equation problem, whereas the optimization over M can be shown to reduce to solving a soft thresholding operation. As described later in the experimental comparisons section, the parameters λ and K are decided by cross validation.

set $\tau = 0$, choose $\mu > 0, M(0), W(0), D(0)$;
while *stopping criterion is not satisfied* **do**
  $W(\tau+1) = \arg\min_{W} -\frac{1}{n^2}\sum_{k=1}^{K}\mathbf{w}_k^T X^T Y Y^T X \mathbf{w}_k + \frac{\mu}{2}||W - M(\tau) - D(\tau)||_F^2$
  $s.t.\ \mathbf{w}_k^T \mathbf{w}_k = 1\ \forall k,\ \mathbf{w}_k^T X^T X \mathbf{w}_i = 0\ \forall\ i \neq k;$
  $M(\tau+1) = \arg\min_{M} \lambda \sum_{j=1}^{p} ||\mathbf{m}_{(j)}||_2 + \frac{\mu}{2}||W(\tau+1) - M - D(\tau)||_F^2;$
  $D(\tau+1) = D(\tau) - W(\tau+1) + M(\tau+1);$
end

Algorithm 2: Exact solution of the global PLS variable selection problem using the augmented Lagrangian method

4 Experimental Comparisons

In this section we show experimental results obtained by comparing standard PLS-R, L_1 penalized PLS-R [Chun *et al.*, 2010], our proposed globally sparse PLS-R,

and Correlated Component Regression [Magidson, 2010]. All the methods have been applied on the Octane data set (see [Tenenhaus, 1998]). The Octane data is a real data set consisting of 39 gasoline samples for which the digitized Octane spectra have been recorded at 225 wavelengths (in nm). The aim is to predict the Octane number, a key measurement of the physical properties of gasoline, using the spectra as predictors. This is of major interest in real applications, because the conventional procedure to calculate the Octane number is time consuming and involves expensive and maintenance-intensive equipment as well as skilled labor.

The experiments are composed of 150 trials. In each trial we randomly split the 39 samples into 26 training samples and 13 test samples. The regularization parameter λ and number of components K are selected by 2-fold cross validation on the training set, while μ is fixed to 2000. The averaged results over the 150 trials are shown in Table 1. All the methods but CCR perform reasonably in terms of MSE on the test set. We further show the variable selection frequencies for the first three PLS methods over the 150 trials superimposed on the octane data in Fig. 1. In chemometrics, the rule of thumb is to look for variables that have large amplitudes in first derivatives with respect to wavelength. Notice that both L_1 penalized PLS-R and globally sparse PLS have selected variables around 1200 and 1350 nm, and the selected region in the latter case is more confined. Box and Whisker plots for comparing the MSE, number of selected variables, and number of components of these three PLS formulations are shown in Fig. 2. Comparing our proposed globally sparse PLS with standard PLS and L_1 penalized PLS [Chun *et al*., 2010], we see that PLS with global variable selection attains better performance in terms of MSE, the number of predictors, and the number of components.

Table 1 Performance of the PLS with global variable selection compared with standard PLS and L_1 penalized PLS

methods	MSE	number of var.	number of comp.
PLS-R	0.0564	225	5.5
L_1 penalized PLS-R	0.0509	87.3	4.5
globally sparse PLS-R	0.0481	38.5	3.8
CCR	0.8284	19.1	6

5 Conclusion

The formulation of the SIMPLS objective function with an added group sparsity penalty greatly reduces the number of variables used to predict the response. This suggests that when multiple components are desired, the variable selection tech-

nique should take into account the sparsity structure for the same variables among all the components. Our proposed globally sparse PLS algorithm is able to achieve as good or better performance with fewer predictor variables and fewer components as compared to competing methods. It is useful for performing dimension reduction and variable selection simultaneously in applications with large dimensional data but comparatively few samples ($n < p$). In future work, we will apply globally sparse PLS algorithms to multivariate response datasets.

Acknowledgements We would like to give special thanks to Douglas Rutledge, professor in AgroParisTech, for his expert knowledge in chemometrics to interpret the selected variables in octane data.

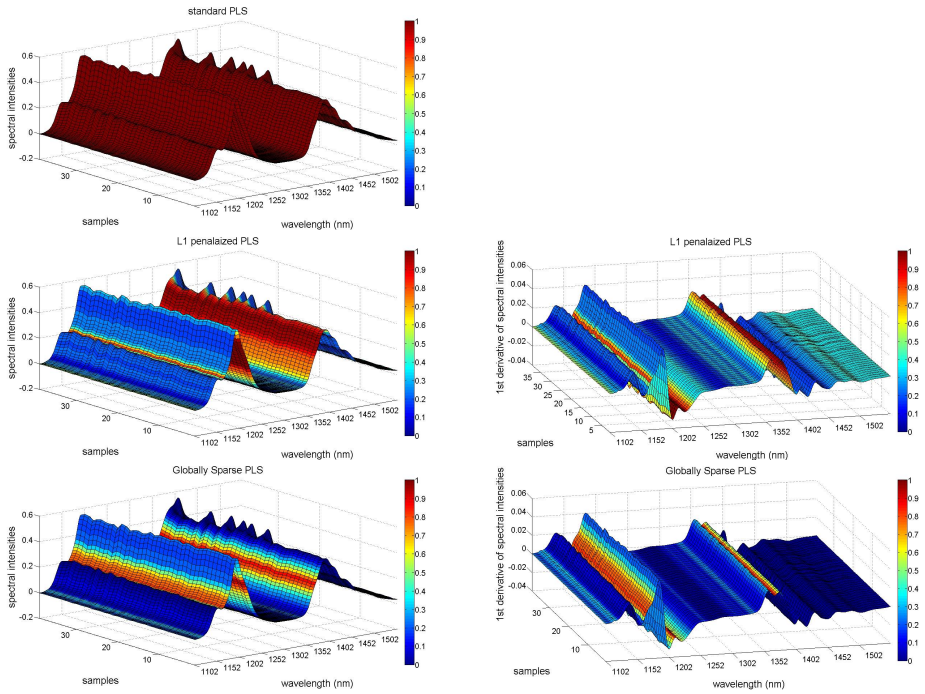

Fig. 1 Variable selection frequency superimposed on the octane data: The hight of the surfaces represents the exact value of the data over 225 variables for the 39 samples. The color of the surface shows the selection frequency of the variables as depicted on the colorbar.

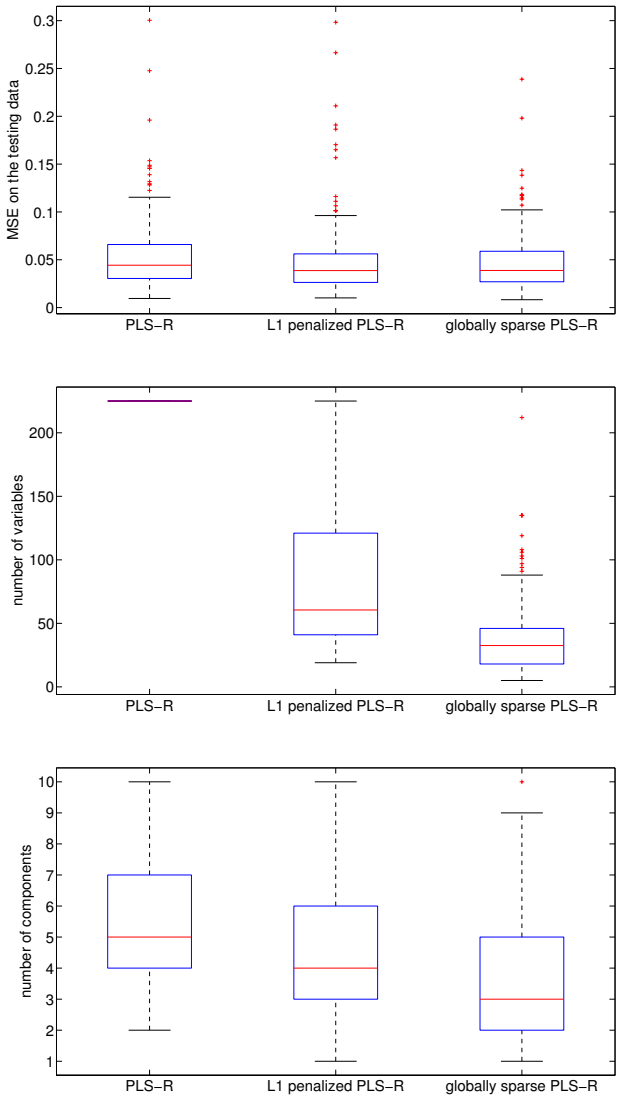

Fig. 2 The Box and Whisker plot for comparing MSE, and number of selected variables, and number of components on the test samples.

References

[de Jong, 1995] de Jong, S. (1995). PLS shrinks. *Journal of Chemometrics*, **9**(4), 323-326.

[Wold, 1966] Wold, H. (1966). Nonlinear estimation by iterative least squares procedures. *Research papers in statistics*, 411-444.

[Hotelling, 1933] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, **24**(6), 417.

[Wold *et al*., 1984] Wold, S., Ruhe, A., Wold, H., and Dunn III, W. J. (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, **5**(3), 735-743.

[Martens *et al*., 1989] Martens, H., and Naes, T. (1989). *Multivariate calibration*. Wiley.

[Rossouw *et al*., 2008] Rossouw, D., Robert-Granié, C., and Besse, P. (2008). A sparse PLS for variable selection when integrating omics data. *Genetics and Molecular Biology*, **7**(1), 35.

[Höskuldsson, 1988] Höskuldsson, A. (1988). PLS regression methods. *Journal of Chemometrics*, **2**(3), 211-228.

[Wold, 1975] Wold, H. (1975). Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach. *Perspectives in Probability and Statistics, In Honor of MS Bartlett*, 117-144.

[Wold *et al*., 1983] Wold, S., Martens, H., and Wold, H. (1983). The multivariate calibration problem in chemistry solved by the PLS method. *Proceedings of the Conference on Matrix Pencils. Lectures Notes in Mathematics*, 286-293.

[Chun *et al*., 2010] Chun, H., and Kele, S. (2010). 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.

[Bach, 2008] Bach, F. R. (2008). Consistency of the group Lasso and multiple kernel learning. *The Journal of Machine Learning Research*, **9**, 1179-1225.

[de Jong, 1993] de Jong, S. (1993). SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, **18**(3), 251-263.

[Tenenhaus, 1998] Tenenhaus, M. (1998). *La Régression PLS: théorie et pratique*. Editions Technip.

[Boulesteix *et al*., 2007] Boulesteix, A. L., and Strimmer, K. (2007). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in bioinformatics*, **8**(1), 32-44.

[ter Braak *et al*., 1998] ter Braak, C. J., and de Jong, S. (1998). The objective function of partial least squares regression. *Journal of chemometrics*, **12**(1), 41-54.

[Jolliffe *et al*., 2003] Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. (2003). A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics*, **12**(3), 531-547.

[Gander *et al*., 1989] Gander, W., Golub, G. H., and von Matt, U. (1989). A constrained eigenvalue problem. *Linear Algebra and its applications*, **114**, 815-839.

[Beck *et al*., 2006] Beck, A., Ben-Tal, A., and Teboulle, M. (2006). Finding a global optimal solution for a quadratically constrained fractional quadratic problem with applications to the regularized total least squares. *SIAM Journal on Matrix Analysis and Applications*, **28**(2), 425-445.

[Jolliffe, 2002] Jolliffe, I. (2002). *Principal component analysis*. John Wiley & Sons, Ltd.

[De Jong, 2005] de Jong, S. (2005). PLS fits closer than PCR. *Journal of chemometrics*, **7**(6), 551-557.

[ter Braak *et al*., 1998] ter Braak, C. J., and de Jong, S. (1998). The objective function of partial least squares regression. *Journal of chemometrics*, **12**(1), 41-54.

[Magidson, 2010] Magidson, J. (2010). Correlated Component Regression: A Prediction/Classification Methodology for Possibly Many Features. *Proceedings of the American Statistical Association*.