Received: 1 December 2019 | Revised: 21 January 2020 | Accepted: 24 January 2020

DOI: 10.1002/cem.3226

RESEARCH ARTICLE

WILEY Journal of CHEMOMETRICS

Comparison of variable selection methods in partial least squares regression

Tahir Mehmood[1,2] | **Solve Sæbø**[2] | **Kristian Hovde Liland**[2,3]

[1]School of Natural Sciences, National University of Sciences and Technology (NUST), Islamabad, Pakistan

[2]Faculty of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences, Ås, Norway

[3]Faculty of Science and Technology, Norwegian University of Life Sciences, Ås, Norway

Correspondence
Tahir Mehmood, School of Natural Sciences, National University of Sciences and Technology (NUST), Islamabad, Pakistan.
Email: tahime@gmail.com

Abstract

Through the remarkable progress in technology, it is getting easier and easier to generate vast amounts of variables from a given sample. The selection of variables is imperative for data reduction and for understanding the modeled relationship. Partial least squares (PLS) regression is among the modeling approaches that address high throughput data. A considerable list of variable selection methods has been introduced in PLS. Most of these methods have been reviewed in a recently conducted study. Motivated by this, we have therefore conducted a comparison of available methods for variable selection within PLS. The main focus of this study was to reveal patterns of dependencies between variable selection method and data properties, which can guide the choice of method in practical data analysis. To this aim, a simulation study was conducted with data sets having diverse properties like the number of variables, the number of samples, model complexity level, and information content. The results indicate that the above factors like the number of variables, number of samples, model complexity level, information content and variant of PLS methods, and their mutual higher-order interactions all significantly define the prediction capabilities of the model and the choice of variable selection strategy.

KEYWORDS

PLS; variable selection

1 | INTRODUCTION

Thanks to the massive use of data generation technologies (spectroscopy, RNAs, satellite images, brain images, etc), a huge amount of data is created in many real-life applications. It enables economic, speedy, and efficient generation of information (variables) of given objects (samples). Within a variety of fields, for instance, bioinformatics, genomics, transcriptomics, proteomics, metabolomics, chemometrics, image processing, geographical information systems, process and analytical technology, we are now witnessing an explosive increase in high dimensional data sets. In order to understand the complexity behind such high-dimensional data sets, multivariate approaches are mandatory to consider. The negative aspect of data generation technologies is the inclusion of irrelevant variables. These irrelevant variables result in a declination of the model performance, in amplification of model complexity, and in the reduction of the understandability of modeled relations. Hence, exclusion of irrelevant variables is important.[1-4]

High-dimensional data sets are often prone to the "large p - small n" problem, ie, many variables and few samples. This problem has been addressed explicitly in partial least squares (PLS) regression (PLSR),[5] and further understanding is

made in previous studies.[6,7] PLS has proven to be a very versatile method for multivariate data analysis. It is a supervised method purposely established to address the problem of making good predictions in multivariate problems, see Martens and Næs.[6] Although we can model high-dimensional data by PLS, very large p and small n can still spoil the PLSR results. For example, in some cases, the PLS estimator with the univariate response is not consistent[1] and a large number of irrelevant variables may cause a large variation in test set prediction.[2] These examples motivate variable selection in PLS to improve the model performance.[8,9] Further, variable selection is important for improved interpretation and understanding of the modeled phenomena. These two extreme end motivations are somehow contradictory, ie, normally increased model performance is achieved with a large number of variables, hence this motivates for a compromise in model performance and a number of selected variables.[10] Thus, variable selection is needed for having an interpretable[3] relationship between explanatory variables and the response, together with acceptable statistical model performance.

PLS in its original form has no direct implementation of variable selection, but a huge range of methods are proposed in PLS that address variable selection. Recently, a large set of variable selection methods in PLS have been reviewed[10] where the variable selection methods are categorized on the bases of their methodological construction into three main categories: filter methods, wrapper methods, and embedded methods. Motivated by this study, we have conducted a comparison of a large selection of available methods for variable selection within PLS.

The main aim of this paper is to unravel patterns of dependencies connecting variable selection methods and data properties. A simulation study is performed where data sets having miscellaneous characteristics like a number of variables, number of samples, model complexity level, and information content has been created. A meta-analysis provides an overview of the performance of PLS methods. This ultimately provides a suggestion for the selection of appropriate methods for the data set in hand.

2 | PLSR ALGORITHM

There is a variety of PLSR algorithms, we start with most pioneering algorithm, called orthogonal scores PLSR,[5] and latte ww will include latest algorithms as well. We assume a set of explanatory variables $\boldsymbol{X}_{(n,p)}$ are linked to a response $\boldsymbol{y}_{(n,1)}$ through the linear relationship $\boldsymbol{y} = \alpha + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with unknown regression parameters α and $\boldsymbol{\beta}$ and error term $\boldsymbol{\epsilon}$. For simplicity, we have considered only the single response case, but the methods can be generalized to multiple responses. Initially, the variables are centered (and optionally scaled) into $\boldsymbol{X}_0 = \boldsymbol{X} - \boldsymbol{1}\bar{\boldsymbol{x}}'$ and $\boldsymbol{y}_0 = \boldsymbol{y} - 1\bar{y}$. Let A be the number of components to be extracted. Then for $a = 1, 2, \ldots, A$, the algorithm goes as follows:

1. Compute the loading weights (LWs) by
$$\boldsymbol{w}_a = \boldsymbol{X}'_{a-1}\boldsymbol{y}_{a-1}.$$
The weights define the direction in the space spanned by $\boldsymbol{X}_{a-1}$ of maximum covariance with $\boldsymbol{y}_{a-1}$. Normalize LWs to have length equal to 1 by
$$\boldsymbol{w}_a \leftarrow \boldsymbol{w}_a / ||\boldsymbol{w}_a||.$$
2. Compute the score vector $\boldsymbol{t}_a$ by
$$\boldsymbol{t}_a = \boldsymbol{X}_{a-1}\boldsymbol{w}_a.$$
3. Compute the X-loadings $\boldsymbol{p}_a$ by regressing the variables in $\boldsymbol{X}_{a-1}$ on the score vector
$$\boldsymbol{p}_a = \boldsymbol{X}'_{a-1}\frac{\boldsymbol{t}_a}{\boldsymbol{t}'_a\boldsymbol{t}_a}.$$
Similarly compute the Y-loading q_a by
$$q_a = \boldsymbol{y}'_{a-1}\frac{\boldsymbol{t}_a}{\boldsymbol{t}'_a\boldsymbol{t}_a}.$$
4. Deflate $\boldsymbol{X}_{a-1}$ and $\boldsymbol{y}_{a-1}$ by subtracting the contribution of $\boldsymbol{t}_a$ as
$$\boldsymbol{X}_a = \boldsymbol{X}_{a-1} - \boldsymbol{t}_a\boldsymbol{p}'_a,$$
$$\boldsymbol{y}_a = \boldsymbol{y}_{a-1} - \boldsymbol{t}_a q_a.$$
5. If $a < A$ return to 1.

LWs, scores and loadings computed at each iteration of the algorithm can be stored in matrices/vectors $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_A]$, $\boldsymbol{T} = [\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_A]$, $\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_A]$ and $\boldsymbol{q} = [q_1, q_2, \ldots, q_A]$. Then, the PLSR-estimator for the regression coefficients (RCs) for the linear model is found by $\hat{\boldsymbol{\beta}} = \boldsymbol{W}(\boldsymbol{P}'\boldsymbol{W})^{-1}\boldsymbol{q}$ and $\hat{\alpha} = \bar{y} - \bar{\boldsymbol{x}}\hat{\boldsymbol{\beta}}$. In the case of multiple responses, the Y-loadings will be replaced by a loading matrix $\boldsymbol{Q} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_A]$.

3 | VARIABLE SELECTION METHODS IN PLS

Variable selection methods in PLS are classified based on how the variable selection is made in PLS into three categories: filter, wrapper, and embedded methods, for detail, see Mehmood et al.[10] A short description is given below.

3.1 | Filter methods

This is a two-step procedure. At first, a PLS model is built. Then, the output from the PLSR-algorithm is solely used to identify a subset of important variables. Examples are selections based on LWs[11] or RCs,[9] Jack-knife testing (JT),[12] variable importance in projection (VIP),[13] selectivity ratio (SR),[14] and significance multivariate correlation (sMC).[15]

3.1.1 | LWs from PLS

From the fitted PLS model for some number of components which are possibly optimized through cross-validation, the LW ($\boldsymbol{w_a}$) can be used as a measure of importance to select variables.[8,11] For each component, the variables with a LW above a certain threshold in absolute value may be selected; specifically, the threshold is defined as $w_* = median(w)/InterquartileRange(w)$. This is also known as hard thresholding and was suggested by Shao et al.[16]

3.1.2 | RCs from PLS

RCs ($\boldsymbol{\beta}$) are single measures of association between each variable and the response. Again, variables having small $\beta_* = median(\beta)/InterquartileRange(\beta)$ can be eliminated.[9] Also in this case, thresholding may be based on significance considerations from jackknifing or bootstrapping,[12] which has been adopted in a wide range of studies (eg, previous works[17,18]).

3.1.3 | Variable importance in PLS projection

A third filter measure is the variable importance in PLS projections (VIP) introduced by[13] as "variable influence on projection" which is known now as "VIP" termed by Eriksson et al.[19]The v_j weights are measures of the contribution of each variable according to the variance explained by each PLS component where $(w_{aj}/\|\boldsymbol{w}_a\|)^2$ represents the importance of the jth variable. Since the variance explained by each component can be computed by the expression $\boldsymbol{q}_a^2\boldsymbol{t}_a'\boldsymbol{t}_a$,[19] the v_j can be expressed as

$$v_j = \sqrt{p\sum_{a=1}^{A}[(\boldsymbol{q}_a^2\boldsymbol{t}_a'\boldsymbol{t}_a)(w_{aj}/\|\boldsymbol{w}_a\|)^2]/\sum_{a=1}^{A}(\boldsymbol{q}_a^2\boldsymbol{t}_a'\boldsymbol{t}_a)}.$$

Variable j can be eliminated if $v_j < u$ for some user-defined threshold $u \in [0, \infty)$. It is generally accepted that a variable should be selected if $v_j > 1$.[19-21]

3.1.4 | SR from PLS

The SR is based on the target projection (TP) approach.[14] Target projection is based on a postprojection of the predictor variables onto the fitted response vector from the estimated model. This results in a decomposition of the original predictor matrix into a latent (TP)-component and a residual component as follows:

$$\boldsymbol{X} = \hat{\boldsymbol{X}}_{TP} + \boldsymbol{E}_{TP} = \boldsymbol{t}_{TP}\boldsymbol{p}'_{TP} + \boldsymbol{E}_{TP},$$

where $\boldsymbol{t}_{TP} = \boldsymbol{X}\boldsymbol{w}_{TP}$, $\boldsymbol{w}_{TP} = \hat{\boldsymbol{\beta}}_{\boldsymbol{PLS}}/||\hat{\boldsymbol{\beta}}_{\boldsymbol{PLS}}||$ and $\boldsymbol{p}_{TP} = \boldsymbol{X}'\boldsymbol{t}_{TP}/(\boldsymbol{t}'_{TP}\boldsymbol{t}_{TP})$. The loadings from this model can be used as measures of how much each predictor variable contributed to the fitted response from the PLSR-model, and based on this SR, r_j is introduced.[22] For each variable j, the r_j can be computed as

$$r_j = V_{exp,j}/V_{res,j},$$

where $V_{exp,j}$ is the explained variance and $V_{res,j}$ is the residual variance for variable j according to the (TP)-model. The most influential variables will score highest on r_j. In order to set a threshold for r_j, Kvalheim et al presents an F-test under the null hypothesis that "explained and residual variance are equal,"[23] and for each variable the hypothesis can be tested by rejecting the null hypothesis if

$$r_j > F_{\alpha,n-2,n-3}.$$

This provides a probabilistic measure for the selection of variables.[24]

3.1.5 | Significance multivariate correlation with PLS

In sMC,[15] an F-test is also used to assess variables which are statistically significant with respect to their relationship (regression) to $\boldsymbol{y}$, but for an sMC_i test value an F-distribution with 1 numerator and an $n-2$ denominator degrees of freedom is used, $F_{(1-\alpha,1,n-2)}$ where α is the chosen significance level and sMC is defined as

$$SMC_i = \frac{MS_{i,PLS_{regression}}}{MS_{i,PLS_{residuals}}} = \frac{\| \frac{\hat{y}\hat{\beta}_i'}{\|\beta_i'\|^2} \|^2}{\|x_i - \frac{\hat{y}\hat{\beta}_i'}{\|\beta_i'\|^2}\|^2/(n-2)}.$$

3.1.6 | Minimum redundancy maximum relevance in PLS

The minimum redundancy maximum relevance (mRMR) is a variable selection algorithm that tends to select variables having a high correlation with the response variable (relevance) and the least correlation between the selected variables (redundancy).[25] In PLS, the variables with minimum redundancy and maximum relevancy can be found by starting with a variable having maximal mutual information with the response y then it greedily adds variables from PLS LWs with a maximal value of

$$J(W) = I(W, y) - \frac{1}{|S|}\sum_{j \in S} I(W, j),$$

where S is the set of already selected variables and j present a respective variable in S. Variables are ranked based on mRMR, and the top $'m'$ number of variables are marked as influential, where $'m'$ can be determined through validation.

3.2 | Wrapper methods

The variables identified by the filter methods are used to refit the PLS model. This refitted PLS model is then used again with filter methods. This procedure is repeated a certain number of times. The wrapper methods are mainly distinguished by the choice of underlying filter-method and how the "wrapping" is executed. Wrapper methods can further be categorized[3] based on the search algorithm: deterministic or randomized. Randomized search algorithms utilize some kind of randomness in the selection of subset while deterministic ones do not. An example from randomized wrapper methods is Monte-Carlo variable elimination with PLS (MVE),[26] while examples from deterministic wrapper methods are sub-window permutation analysis with PLS (SPA),[27] backward variable elimination in PLS (BVE-PSL),[8] and regularized elimination procedure in PLS (REP).[3]

3.2.1 | Genetic algorithm combined with PLS regression

Hasegawa et al combine genetic algorithm (GA) with PLS in the GA method.[28] Genetic algorithms involve the following steps:

1. Building an initial population of variable sets by setting bits for each variable randomly, where bit "1" represents a selection of the corresponding variable while "0" presents nonselection. The approximate size of the variable sets must be set in advance.
2. Fitting a PLSR-model to each variable set and computing the performance by, for instance, a leave-one-out cross-validation procedure.
3. A collection of variable sets with higher performance are selected to survive until the next "generation."
4. Crossover and mutation: New variable sets are formed (a) by the crossover of selected variables between the surviving variable sets and (b) by changing (mutating) the bit value for each variable by a small ratio R_0. Indicating the variable selection is defined by this mutating ratio R_0.
5. The surviving and modified variable sets from the population serve as input to point 2.

The steps 2 to 5 are repeated a preset number of times. Upon completion of the GA-algorithm, the best variable set (or a combination of a collection of the best sets) in terms of performance is selected.

3.2.2 | Uninformative variable elimination in PLS

Centner et al[29] introduced uninformative variable elimination in PLS (UVE), where artificial noise variables are added to the predictor set before the PLSR model is fitted. All the original variables having lower "importance" than the artificial noise variables are eliminated before the procedure is repeated until a stop criterion is reached. The steps involved in UVE are as follos:

1. Generate a noise matrix $\boldsymbol{N}$, having the same dimension as $\boldsymbol{X}$, where entries are randomly drawn from uniform distribution in the interval 0.0 to 1.0.
2. Combine the $\boldsymbol{N}$ and $\boldsymbol{X}$ matrices into a new matrix of variables $\boldsymbol{Z}=[\boldsymbol{X},\boldsymbol{N}]$.
3. Fit the PLSR model to the combined matrix $\boldsymbol{Z}$ and validate by means of leave-one-out cross-validation.
4. Cross-validation results are used to compute a test statistic for each variable as $c_j = mean(\hat{\beta}_j)/sd(\hat{\beta}_j)$, for $j = 1, 2, \ldots, 2p$.
5. Set the threshold *cmax* as the maximum of absolute value c among the noise variables. Original variables with an absolute value of c smaller than *cmax* are assumed to be noise variables and are eliminated.

Steps 2 to 6 are repeated unless the performance of the models start decreasing. Artificially added random variables could influence the model if random variables are not properly selected.[30]

3.2.3 | Sub-window permutation analysis coupled with PLS

Sub-window permutation analysis coupled with PLS (SPA)[27] provides the influence of each variable without considering the influence of the rest of the variables. The use of subsets of variables makes SPA more efficient and fast for huge datasets. Steps involved in SPA are as follows:

1. Sub-dataset sampling in both sample and variable space into N test and N training data sets.
2. For each randomly sampled training set, a PLS model is built and a normal prediction error, NPE, is measured on the corresponding test set. This NPE will be associated with all the predictors in the given training data set.
3. Prediction performance is also measured for each sampled training set by iteratively permuting each predictor variable in the training set to obtain a permuted prediction error, PPE, associated with each predictor.
4. The variable importance of each variable j is assessed upon completion of the sub-dataset sampling by comparing the distributions of normal prediction errors (*NPE*s) and *PPE* of any given variable. Hence, for variable j, the statistical assessment of importance is computed as $D_j = mean(PPE_j) - mean(NPE_j)$.
5. All variables for which $D_j > 0$ are considered informative, in the sense that they will with large probability improve the prediction performance of a model if they are included.

3.2.4 | Iterative predictor weighting PLS

Forina et al[31] introduced an iterative procedure for variable elimination, called Iterative predictor weighting PLS (IPW). This is an iterative elimination procedure where a measure of predictor importance is computed after fitting a PLSR model (with complexity chosen based on predictive performance). The importance measure β is used both to rescale the original X-variables and to eliminate the least important variables before subsequent model refitting.

3.2.5 | Backward variable elimination PLS

Backward variable elimination procedures are also developed for the elimination of non-informative variables, first introduced by Frank et al[8] and then Fernandez et al. In general, variables are first sorted for some important measure, and usually, one of the filter measures described above is used. Second, a threshold is used to eliminate a subset of the least informative variables. Then, a model has fitted again to the remaining variables and performance is measured. The procedure is repeated until maximum model performance is achieved.

3.2.6 | Regularized elimination procedure in PLS

Mehmood et al[3] introduced a regularized variable elimination procedure (REP) for parsimonious variable selection, where also a stepwise elimination is carried out. A stability-based variable selection procedure is adopted, where the samples

have been split randomly into a predefined number of training and test sets. For each split, g, the following stepwise procedure is adopted to select the variables: Let $\boldsymbol{Z}_0 = \boldsymbol{X}$ and s_j (eg, w_j, β_j, or r_j) be one of the filter criteria for variable j.

1. For iteration g, run $\boldsymbol{Y}$ and $\boldsymbol{Z}_g$ through cross validated PLS. The matrix $\boldsymbol{Z}_g$ has p_g columns, and we get the same number of criterion values, sorted in ascending order as $s_{(1)}, \ldots, s_{(p_g)}$.
2. Assume there are M criterion values below some predefined cutoff u. If $M = 0$, terminate the algorithm here.
3. Else, let $N = \lceil fM \rceil$ for some fraction $f \in \langle 0, 1]$. Eliminate the variables corresponding to the N most unfavorable criterion values.
4. If there is still more than one variable left, let $\boldsymbol{Z}_{g+1}$ contain these variables, and return to (1).

The fraction f determines the "steplength" of the elimination algorithm, where an f close to 0 will only eliminate a few variables in every iteration. The fraction f and the cutoff u can be determined through cross validation.

3.2.7 | Hotelling T^2 based variable selection in PLS (T^2)

PLS LW matrix $\boldsymbol{W}$ against the validated optimal model can be used to classify the informative variables through Hotelling T^2.[32] The procedure follows these steps:

1. Extract PLS LWs matrix W'.
2. From LWs matrix W' compute
$$T^2 = p(\bar{W}_i - \bar{\bar{W}})' S_W^{-1} (\bar{W}_i - \bar{\bar{W}})$$
and
$$Upperlimit = C(p, A^*) F_{(A^*, p-A^*, \alpha_{T^2})}.$$
3. Remove the inlier as non-informative from the model.

Here, A^* presents the used number of PLS components and $C(p, A^*) = \frac{A^*(p-1)}{p-A^*}$. Here, a second PLS is fitting on selected variables to select the α_{T^2} values. The α_{T^2} determines the variable selection in PLS and is mainly dependent on the upper limits $C(p, A^*) F_{(A^*, p-A^*, \alpha_{T^2})}$.

3.3 | Embedded methods

The variable selection is an assimilated part of a modified PLS-algorithm. Examples include soft-threshold PLS (ST),[33] the sparse-PLS (SPLS),[34] and distribution based truncation for variable selection in PLS (TR).[35]

3.3.1 | ST PLS

Spbø et al[33] introduced a soft-thresholding step in the PLS algorithm (ST) based on ideas from the nearest shrunken centroid method.[36] The ST approach is more or less identical to the Sparse-PLS presented independently by Lê Cao et al.[34] At each step of the sequential ST algorithm, the LWs are modified as follows:

1. Scaling:
$\boldsymbol{w}_k \leftarrow \boldsymbol{w}_k / \max_j |w_{k,j}|$, for $j = 1, \ldots, p$.
2. Soft-thresholding:
$w_{k,j} \leftarrow \text{sign}(w_{k,j})(|w_{k,j}| - \delta)_+$, for $j = 1, \ldots, p$ and some $\delta \in [0, 1\rangle$. Here $(\ldots)_+$ means $\max(0, \ldots)$.
3. Normalizing:
$\boldsymbol{w}_k \leftarrow \boldsymbol{w}_k / \|\boldsymbol{w}_k\|$.

The shrinkage $\delta \in [0, 1)$ sets the degree of thresholding, ie, a larger δ gives a smaller selected set of variables. Cross validation is used to define this threshold.[33]

3.3.2 | Distribution-based truncation for variable selection in PLS (TRUNC)

For the selection of variables, the magnitude of PLS LWs (the columns of W) is an established criterion. Liland et al[35] suggested a distribution-based truncation for the variable selection. At each step of the sequential PLS algorithm, the weights were modified as follows:

1. Sort LWs $\boldsymbol{w}$ as $\boldsymbol{w}_{\boldsymbol{s}}$.
2. Compute a confidence interval around the median of $\boldsymbol{w}_{\boldsymbol{s}}$, which is based on a threshold pr.
4. Classify outliers as real, informative contributions and inliers as noise.
5. Truncate inliers.

3.3.3 | Weighted variable contribution in PLS

Weighted variable contribution with PLS (PLS-WVC) is based on to the first singular value of the covariance matrix for each PLS component.[37] In each PLS iteration, the singular value decomposition of covariance of $X^t y$ can be written as $s_a = w_a^t X_a^t y_a q_a = (w_{a,1}^2 + w_{a,2}^2 + \ldots + w_{a,p}^2)$. This means the contribution of the ith certain variable to the first singular value of the correlation matrix between X_a and y_a is $w_{a,i}^2 s_a$. The weighted contribution of the ith variable can be defined as

$$WVC(i) = \sqrt{\frac{m \sum_{a}^{A} \alpha_a w_{a,i}^2 s_a}{\sum_{a}^{A} \alpha_a s_a}},$$

where

$$\alpha_a = \frac{w_a^t X_a^t y_a q_a}{w_a^t X_a^t y_a w_a}.$$

This expression seems to be a hybrid between two of the four possible criteria in the original article and will now work due to the dimensions not fitting each other in the denominator. In each iteration, $w_a \to 0$ if $WVC(i) < \eta$ followed by the LW normalization, ie, $\boldsymbol{w}_k \leftarrow \boldsymbol{w}_k / \|\boldsymbol{w}_k\|$. The optimal η can be determined through validation.

4 | COMPUTATIONAL STRUCTURE

4.1 | Simulation for the comparison

We are interested in understanding the strengths and weaknesses of different variable selection methods in PLS. This goal can be achieved by constructing an experimental design for running simulation experiments, with different data properties that may influence the methods' performances. The chosen factors are the number of variables (p), the number of training samples (n), the number of relevant predictors (q), information content (R^2), a parameter ($\gamma(> 0)$) defining the decline of the eigenvalues of the predictor covariance matrix (Σ_x), the number of latent relevant components (m), and their position pos in the set of indices of declining eigenvalues of (Σ_x). As an example, if $pos = (1, 5)$, it means that the eigenvectors associated with the largest and the fifth largest eigenvalues of (Σ_x) define the relevant components for the prediction of the response y. The eigenvalues of (Σ_x) are assumed to decline exponentially for $j = 1, \ldots, p$ according to $\exp(-\gamma(j-1))$. That is, a large γ implies rapid decline and high multi-collinearity.

These factors represent the data properties and define the factors to be varied in our simulation study. We have considered different levels of these factors, which are listed in Table 1. The simulations were conducted using the simrel-package[38] and the plsVarSel-package[39] in R,[40] and further details regarding the simulation model may be found therein.

4.2 | Assessment of performance

We have simulated independent test data for the assessment of the prediction performance of the selected variables from the different methods. In particular, the root mean square error of prediction ($RMSEP$) and the root mean square error of prediction relative to minimum achievable error ($RMSEP_{minA} = RMSEP/\sqrt{1-R^2}$) were measured. The latter can be

TABLE 1 The input factors together with their description and levels considered in the simulation study are presented

Input Factors	Data Properties	Levels	No. of levels
p	The number of predictors	100 and 2000	2
n	The number of training samples	50 and 200	2
q	The number or relevant predictors	0.05p and 0.2p	2
m	The number of relevant components	set by length of each *pos* variable	2
pos	The position of relevant components	(1,3),(1, 3, 5, 7, 9)	2
γ	The decline in eigenvalues of Σ_x	0.1, 0.95	2
R^2	The information content	0.6 and 0.9	2

computed since the lower bound of prediction error is known from the simulation design. Further, accuracy (*Accuracy*) of selected variables was computed for each method.

4.3 | Optimization and parameter tuning

In the process of model fitting, the complexity parameters of the models had to be set, and for many methods, there are given recommended values of these parameters, whereas for others, the tuning of the parameters must be done according to some criteria. We have applied the most commonly used criterion in practical modeling and variable selection with PLS, namely to minimize cross-validated prediction error measured as the root mean squared error of prediction (RMSEP).

All PLS-methods have at least one common complexity parameter, the number of components. In addition, there are some method-specific parameters for variable selection that need to be set for each method. To set the values of these parameters, cross-validation is again a frequently adopted procedure. Hence, double cross-validation is recommended where the training data is further divided into test and training set for model tuning. Double cross-validation is recommended for filter and embedded selection methods, while for wrapper methods, the tuning of both the number of PLS components and the other additional parameters at the same time is not recommended. In this situation, triple cross-validation (TCV)[3,41] provides a solution. Here, the training data are divided into test and training data for tuning the variable selection parameters and then a further split of training and test data is used for selecting the number of components.

4.4 | Meta analysis using mixed-effects ANOVA

The design factors with their chosen levels, as listed in Table 1, lead to $2^6 = 64$ design levels. For each design level, we simulated five data sets, which were analyzed by all 18 variable selection methods. From the parameters defining the simulation models, the theoretical minimum prediction error was known for each simulation design. In order to compare the prediction performance of the various models and for different designs, it is convenient to compare observed prediction error relative to the minimum achievable. We have referred to this variable above as ($RMSEP_{minA}$). A mixed-effects model was fitted to this response with the design factors *p*, *n*, *pos*, *gamma*, and *method* as main effects. In addition all 2. and 3. order interactions between these were included. Further, a random effect of data set, with 64 * 5 = 320, levels were included in the model to account for the inherent random "prediction difficulty" associated with each simulated training data set. The design factor R^2 was not included in the analysis since the minimum achievable error has a 1:1 relation to this factor, and hence, the effect of R^2 has been removed when we study the relative prediction error. After an initial model fit in R and subsequent backward elimination of nonsignificant effects (5% test-level), a reduced model was found and the results are given below.

5 | RESULTS AND DISCUSSIONS

This study provides the comparison of 17 variable selection methods in PLS. The optimized values of the tuning parameters for the various methods, as described in the methods section, and as set by the optimization criterion (ie RMSEP, see Section 4.3) are presented in Table 2.

For real-life data, the prediction error is the only possibility to assess the performance of the method, while in a simulation study where we know which variables are actually important, we have other possibilities. The core of the study is to focus on the differences between methods, in prediction error and variable selection accuracy. In Figure 1, we have displayed the average accuracy across all simulation sets for each method versus the average prediction error. The methods which on average perform best with regard to selection accuracy are the SR and the sMC which reach an accuracy of 0.9, but the SR has lower prediction error. A large group of methods performs more or less equally with regard to both the accuracy of variable selection and prediction error: Trunc, JT, RC, BVE, ST, and REP. The lowest accuracy is observed for SPA and JT. In reference to prediction capability, JT and mRMR both perform worst while remaining methods have similar lower prediction capability. Since prediction capability is the main focus of the article, we have therefore discarded the mRMR and JT for further analysis.

Since variable selection methods are classified based on their construction into three groups, called filter, wrapper, and embedded. In reference to computational time, groups of methods can be ranked from faster to slower as a filter, embedded, and wrapper. Since wrapper and embedded are more conscious about prediction error, hence a group of methods can be ranked from high performance to low performance as embedded, wrapper, and filter. This is also supported by

TABLE 2 The values of the tuning parameters of the various methods set by the optimization criteria RMSEP

Variable Selection Methods	Parameters	Value
Filter methods		
Loading weights (LW)	w_*	0.05
Regression coefficients (RC)	β_*	0.17
Jackknife testing (JT)	α	0.15
Variable importance in projection (VIP)	v	0.90
Selectivity ratio (SR)	α	0.32
Significance multiple correlation (sMC)	α	25.2
Min. redundancy max. relevance (mRMR)	m	0.75
Wrapper methods		
Genetic algorithm with PLS (GA)	R_0	8.39
Monte-Carlo elimination with PLS (MCUVE)	c	2.01
Sub-window permutation in PLS (SPA)	α	0.33
Iterative predictor weighting PLS (IPW)	β	0.01
Backward variable elimination in PLS (BVE)	v	1.12
Regularized elimination in PLS (REP)	v	1.15
Hotelling T^2 based selection in PLS (T^2)	α	0.10
Embedded methods		
Soft-thresholding PLS (ST)	δ	0.54
Truncation PLS (TRUNC)	pr	0.90
Weighted variable contribution in PLS (WVC)	η	0.17

Note. In addition to these tuned parameters, some additional parameters were kept fixed, for instance, we used IPW with 10 iterations, GA with linear kernels, and population of size $1/p$, SPA with 10 variables to be sampled in each run, BVE with $\alpha = 0.05$, and RE with $\alpha = 0.05$ and $f = 1$.

the results because in reference to better accuracy and minimum prediction error methods can be ranked as T2, BVE, and SR. Among the embedded method, T2 considers the multivariate structure of LWs which results in better performance. Similarly among wrapper methods, BVE selects the optimal model from steps wise elimination in contrast to REP which compromises the performance for a smaller number of variables. Moreover, among filter methods, SR associates the statistical significance based on target projections for variable selection.

```
Anova Table (Type II tests)
Response: RMSEP_minA
                    Sum Sq   Df   F value   Pr(>F)
p                    0.002    1    0.0302   0.8620110
n                   86.987    1 1465.8391   < 2.2e-16 ***
q                    2.934    1   49.4375   2.340e-12 ***
pos                  3.224    1   54.3282   1.992e-13 ***
gamma              144.743    1 2439.1073   < 2.2e-16 ***
PLSMethod           18.785   14   22.6107   < 2.2e-16 ***
p:pos                0.588    1    9.9117   0.0016525 **
p:gamma              0.652    1   10.9786   0.0009286 ***
n:q                  0.691    1   11.6409   0.0006506 ***
n:pos                1.193    1   20.0977   7.532e-06 ***
n:gamma             46.926    1  790.7716   < 2.2e-16 ***
q:gamma              5.851    1   98.6028   < 2.2e-16 ***
p:PLSMethod          4.835   14    5.8199   1.937e-11 ***
gamma:PLSMethod      8.136   14    9.7926   < 2.2e-16 ***
p:gamma:PLSMethod    4.419   14    5.3195   3.629e-10 ***
Residuals          280.809 4732
---
Signif. codes: 0  *** 0.001 ** 0.01 * 0.05 . 0.1  1
```

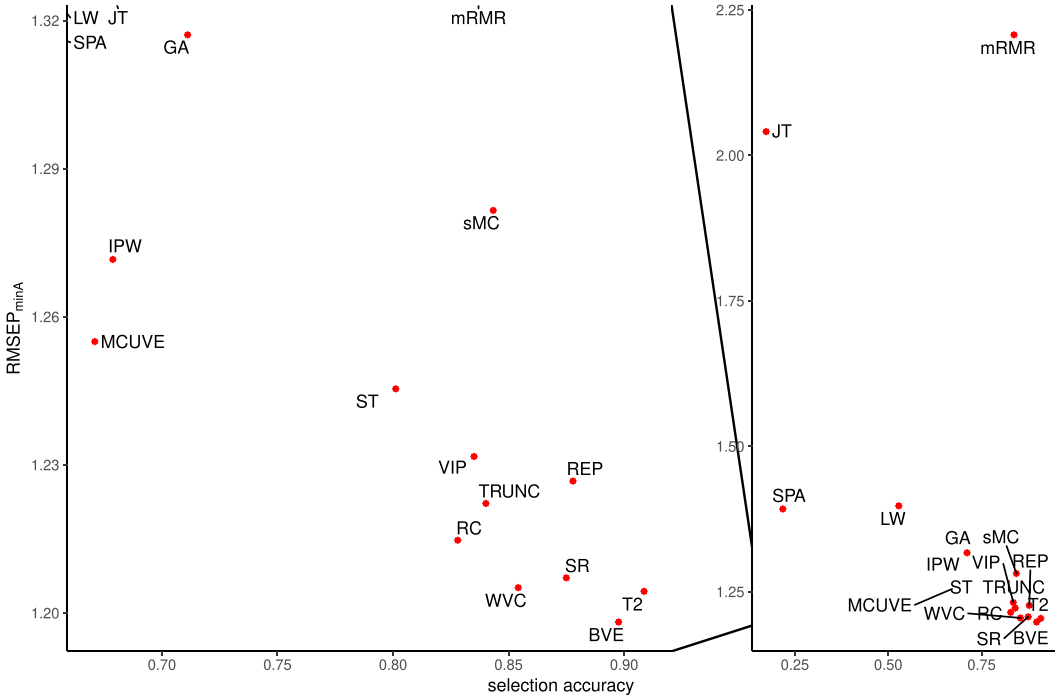

FIGURE 1 Average relative prediction error across all simulations versus the average selection accuracy. Left-hand panel is a magnified view of the lower right part of the right-hand panel

As can be seen from the ANOVA table, there are several second-order interactions between design factors and the variable selection methods that appear important. Further, all main effects are significant and the effects of these were more or less as expected. The distribution of prediction error relative to the minimum achievable ($RMSEP_{minA}$) over the levels of significant second-order interaction effects that appeared in the ANOVA are presented in Figure 2. To explore the conditions under which methods differ and on which basis, we start with the interaction plots for the significant second-order interactions.

The prediction error is in general decreasing when the number of observations increases and when the relevant information for the response is located in components with large variances (eigenvalues). This latter observation confirms the findings of Helland and Almøy (1994).[42] The results also indicate an increase in prediction error when the number of variables increases and when the relevant information for the response is located in components with large variances (eigenvalues), but with larger components, a reversed trend in prediction error is observed with the increase of prediction error. The prediction error decreases as the fraction of relevant important variables get decreased. Moreover, there is a decrease in prediction error with increasing number of variables, p, but this is likely an artifact of the simulation design where the number of samples n was set as a function of p, and the largest values of p were thus associated with the largest values of n. A closer inspection of prediction errors for values of p with equal numbers of n showed that prediction error was quite constant with increasing p, indicating that PLS-regression is capable of ignoring irrelevant predictors even without strict variable selection. This is the essence of the PLS algorithm which is achieved because of dimension reduction through the use of latent variables in the presence of response. We find that the prediction error decreases when the number of observations increases and when the degree of decline, *gamma*, of the eigenvalues of (Σ_x) increases. The interaction plot of a number of variables and degree of decline *gamma* indicates that with a low degree of decline $gamma = 0.01$, the prediction error increases with the increase of a number of variables. However, it shows a reversed trend when the degree of decline increases with $gamma = 0.95$. Moreover, the prediction error increases when the number of relevant important variables increases and when the degree of decline *gamma*, of the eigenvalues of (Σ_x) decreases.

With regard to revealing any patterns between selection method and data properties, these are the most interesting results when it comes to the prediction performances of the methods. In Figure 2, the interaction between the degree of decline, *gamma*, of the eigenvalues of (Σ_x) and PLS variable selection method is displayed. On the vertical axis, a value of one means a prediction performance equal to the minimum achievable. As we move from left to right in the figure, the relative information in the predictor matrix becomes more hidden in principal directions with steadily smaller variances

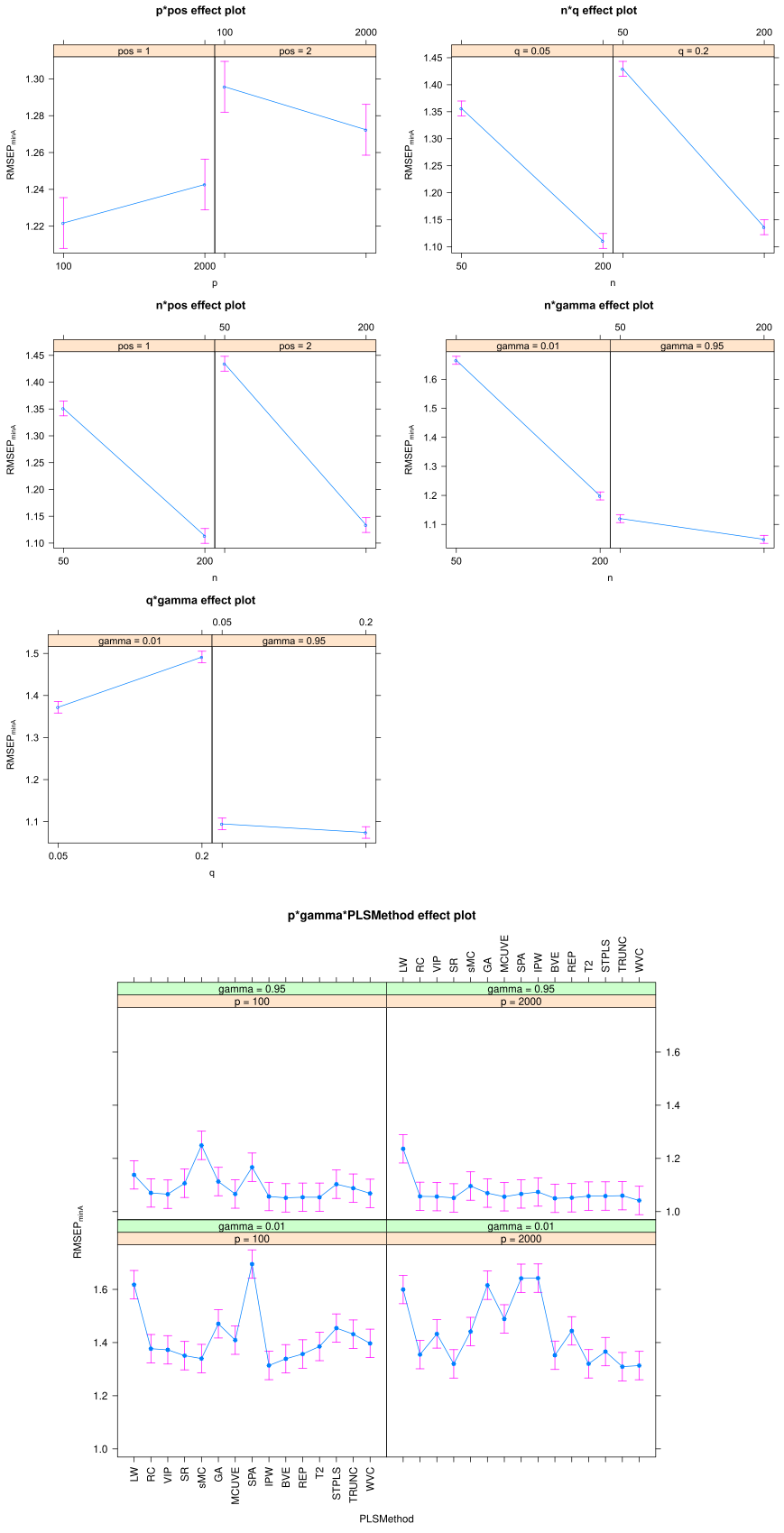

FIGURE 2 The distribution of prediction error relative to the minimum achievable ($RMSEP_{minA}$) over the levels of significant second-order interaction effects that appeared in the ANOVA is presented

(eigenvalues). In the upper panel where $gamma = 0.95$ and the decline in eigenvalues are large, these components have very small variances and the information is hard to find. For $gamma = 0.01$ (lower panels), the variances of the compo-

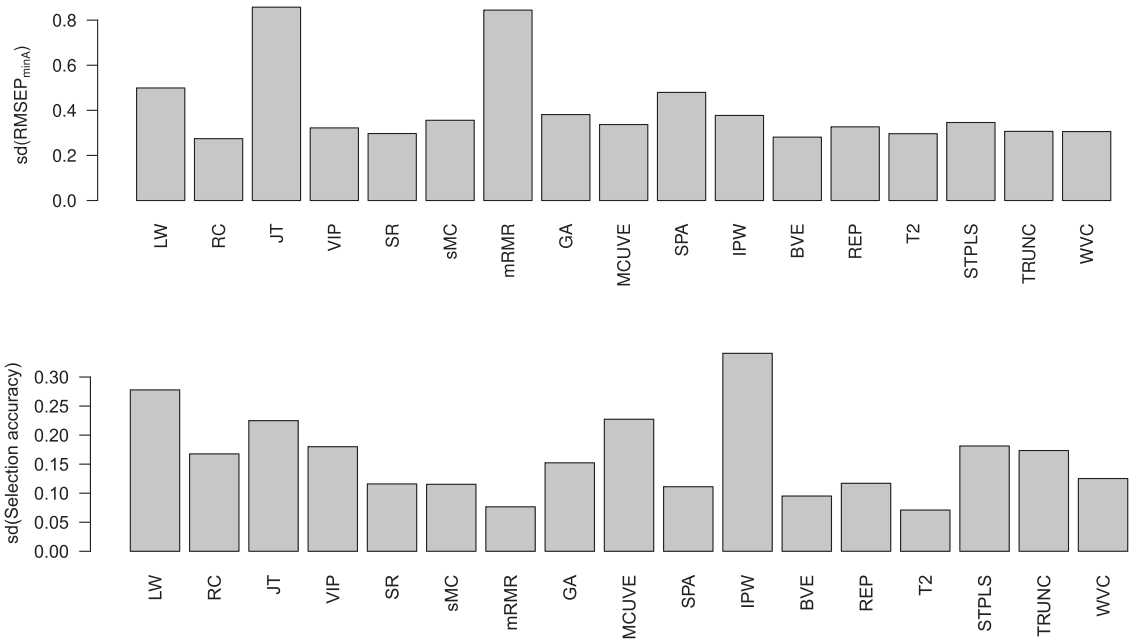

FIGURE 3 The standard deviations of all $RMSEP_{minA}$ values and all selection accuracies for all methods

nents are more similar with a less steep decline in eigenvalues and information is easier to find and the prediction error is larger.

With regard to the performance of the methods we can make the following observations:

- The LW which is purely based on the selection on LWs performs badly for all data types.[6]
- Some methods struggle more than others when $gamma = 0.01$, that is, when there is a relatively low degree of collinearity in the data. This applies especially to RC, SR, IPW, BVE, T2, TRUNC, and WVC. These methods rely less on borrowing information between variables than the others. These methods perform better when collinearity is larger and it is easier to locate the informative directions in the predictor space.[43]
- Several methods are quite similar in their performance across all data designs, but among these, the SR appears to be better.[10]
- The commonly used RC performs well in all cases.[44]

When choosing a method for variable selection, the stability of the method is also an issue to take into consideration. In Figure 3, the standard deviations of all prediction errors are displayed for each method. The variation is largest for JT and mRMR, followed by LW. All other methods have similar and lower variability in the prediction errors for the test data. Also included is the standard deviations of the selection accuracies. Except for LW, the pattern is a bit different from IPW and MCUVE as the most unstable, together with LW. T2 seems to be the best compromise with regard to overall stability.

6 | CONCLUSIONS

In this work, we have compared a large range of PLS variable selection methods. Some of these are tailored specially for PLS, while others have more general application. The simulation framework applied allows for the creation of data that mimics several important aspects of real data, particularly with regard to the eigenvalue structure of the predictive variables.

From the results and discussion, it is evident that variable selection is no guarantee for improving predictions, though many methods improve on the pure PLS model in situations with low variable correlation. Some methods can, in general, be left aside due to low predictive performance (JT, mRMR) or bad selection accuracy (SPA, LW). Among the remaining methods, the structure of the data and the emphasis on either good selection, eg, for biomarkers, or good prediction will be important for the final choice of method. Among the best overall performers, we would like to point out BVE, T2, SR, and WVC, with BVE as a marginal winner on average $RMSEP_{minA}$ and T2 as the most accurate and stable selector among the best performers.

ORCID

Tahir Mehmood https://orcid.org/0000-0001-9775-8093
Kristian Hovde Liland https://orcid.org/0000-0001-6468-9423

REFERENCES

1. Keleş S, Chun H. Comments on: augmenting the bootstrap to analyze high dimensional genomic data. *TEST*. 2008;17(1):36-39.
2. Höskuldsson A. Variable and subset selection in PLS regression. *Chemometrics Intel Lab Syst*. 2001;55(1-2):23-38.
3. Mehmood T, Martens H, Sæbø S, Warringer J, Snipen L. A partial least squares based algorithm for parsimonious variable selection. *Alg Mol Biol*. 2011;6:27.
4. Mehmood T, Martens H, Sæbø S, Warringer J, Snipen L. Mining for genotype-phenotype relations in saccharomyces using partial least squares. *BMC Bioinformatics*. 2011;12:318.
5. Wold S, Martens H, Wold H. The multivariate calibration problem in chemistry solved by the PLS method. In: Conference Proceeding Matrix pencils Springer; 1983; Berlin, Heidelberg:286-293.
6. Martens H, Næs T.. *Multivariate calibration*. Chichester, UK: Wiley; 1989.
7. Helland IS. Some theoretical aspects of partial least squares regression. *Chemom Intell Lab Syst*. 2001;58(2):97-107.
8. Frank IE. Intermediate least squares regression method. *Chemometr Intell Lab Syst*. 1987;1(3):233-242.
9. Frenich AG, Jouan-Rimbaud D, Massart D, Kuttatharmmakul S, Galera MM, Vidal JLM. Wavelength selection method for multicomponent spectrophotometric determinations using partial least squares. *Anal*. 1995;120(12):2787-2792.
10. Mehmood T, Liland KH, Snipen L, Sæbø S. A review of variable selection methods in partial least squares regression. *Chemometr Intell Lab Syst*. 2012;118:62-69.
11. Martens M. Sensory and chemical quality criteria for white cabbage studied by multivariate data analysis. *Lebensmittel-Wissenschaft Technologie*. 1985;18(2):100-104.
12. Efron B, Tibshirani R. *An introduction to the bootstrap*, Vol. 57, New York, USA: Chapman & Hall/CRC; 1993.
13. Wold S, Johansson E, Cocchi M. PLS: partial least squares projections to latent structures. *3D QSAR Drug Design*. 1993;1:523-550.
14. Kvalheim OM, Karstang TV. Interpretation of latent-variable regression models. *Chemom Intell Lab Syst*. 1989;7(1-2):39-51.
15. Tran TN, Afanador NL, Buydens LM, Blanchet L. Interpretation of variable importance in partial least squares with significance multivariate correlation (SMC). *Chemom Intell Lab Syst*. 2014;138:153-160.
16. Shao R, Jia F, Martin E, Morris A. Wavelets and non-linear principal components analysis for process monitoring. *Control Eng Pract*. 1999;7(7):865-879.
17. Ferreira AP, Alves TP, Menezes JC. Monitoring complex media fermentations with near-infrared spectroscopy: comparison of different variable selection methods. *Biotech. Bioeng.* 2005;91(4):474-481.
18. Xu H, Liu Z, Cai W, Shao X. A wavelength selection method based on randomization test for near-infrared spectral analysis. *Chemom. Intell. Lab. Syst.* 2009;97(2):189-193.
19. Eriksson L, Johansson E, Kettaneh-Wold N, Wold S. *Multi-and megavariate data analysis*, Umeå, Sweden: Umetrics Umeå; 2001.
20. Chong G, Jun CH. Performance of some variable selection methods when multicollinearity is present. *Chemom Intell Lab Syst*. 2005;78:103-112.
21. Gosselin R, Rodrigue D, Duchesne C. A bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications. *Chemom Intell Lab Syst*. 2010;100(1):12-21.
22. Rajalahti T, Arneberg R, Kroksveen AC, Berle M, Myhr KM, Kvalheim OM. Discriminating variable test and selectivity ratio plot: quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles. *Anal. Chem.* 2009;81(7):2581-2590.
23. Kvalheim OM, Rajalahti T, Arneberg R. X-tended target projection (XTP)ï£¡comparison with orthogonal partial least squares (OPLS) and PLS post-processing by similarity transformation (PLS+ ST). *J Chemometrics*. 2009;23(1):49-55.
24. Kvalheim OM. Interpretation of partial least squares regression models by means of target projection and selectivity ratio plots. *J Chemom*. 2010;24:496-504.
25. Talukdar U, Hazarika SM, Gan JQ. A kernel partial least square based feature selection method. *Pattern Recognit*. 2018;83:91-106.
26. Cai W, Li Y, Shao X. A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. *Chemom Intell Lab Syst*. 2008;90(2):188-194.
27. Li HD, Zeng MM, Tan BB, Liang YZ, Xu QS, Cao DS. Recipe for revealing informative metabolites based on model population analysis. *Metabolomics*. 2010;6(3):1-9.
28. Hasegawa K, Miyashita Y, Funatsu K. Ga strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists? *J Chem Inf Comput Sci*. 1997;37(2):306-310.
29. Centner V, Massart DL, de Noord OE, de Jong S, Vandeginste BM, Sterna C. Elimination of uninformative variables for multivariate calibration. *Anal Chem*. 1996;68(21):3851-3858.
30. Faber NM, Meinders MJ, Geladi P, Sjostrom M, Buydens LMC, Kateman G. Random error bias in principal component analysis. Part I. Derivation of theoretical predictions. *Analytica chimica acta*. 1995;304(3):257-271.
31. Forina M, Casolino C, Pizarro Millan C. Iterative predictor weighting (IPW) PLS: a technique for the elimination of useless predictors in regression problems. *J Chemom*. 1999;13(2):165-184.

32. Mehmood T. Hotelling t2 based variable selection in partial least squares regression. *Chemom Intell Lab Syst*. 2016;154:23-28.
33. Sæbø S, Almøy T, Aarøe J, Aastveit AH. ST-PLS: a multi-dimensional nearest shrunken centroid type classifier via PLS. *J Chemometrics*. 2007;20:54-62.
34. Lê Cao KA, Rossouw D, Robert-Granié C, Besse P. A sparse PLS for variable selection when integrating omics data. *Stat Appl Genet Mol Biol*. 2008;7(1):35.
35. Liland KH, Høy M., Martens H, Sæbø S. Distribution based truncation for variable selection in subspace methods for multivariate regression. *Chemom Intell Lab Syst*. 2013;122:103-111.
36. Tibshirani R, Hastie T, Narasimhan B, Chu G. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science*. 2003;18(1):104-117.
37. Lin W, Hang H, Zhuang Y, Zhang S. Variable selection in partial least squares with the weighted variable contribution to the first singular value of the covariance matrix. *Chemom Intell Lab Syst*. 2018;183:113-121.
38. Sæbø S, Almøy T, Helland IS. simrel—Aaversatile tool for linear model data simulation based on the concept of a relevant subspace and relevant predictors. *Chemom Intell Lab Syst*. 2015;146:128-135.
39. Liland KH, Mehmood T, Sæbø S. plsvarsel: Variable selection in partial least squares. https://CRAN.R-project.org/package=plsVarSel,_Rpackageversion0.8; 2016.
40. Ihaka R, Gentleman R. R: a language for data analysis and graphics. *J Comput Graph Stat*. 1996;5(3):299-314.
41. Anderssen E, Dyrstad K, Westad F, Martens H. Reducing over-optimism in variable selection by cross-model validation. *Chemom Intell Lab Syst*. 2006;84(1-2):69-74.
42. Helland IS, Almøy T. Comparison of prediction methods when only a few components are relevant. *J Am Stat Assoc*. 1994;89(426):583-591.
43. Boulesteix AL. PLS dimension reduction for classification with microarray data. *Stat Appl Genet Mol Biol*. 2004;3(1):1075.
44. Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS). *J Chemometrics*. 2002;16(3):119-128.

How to cite this article: Mehmood T, Sæbø S, Liland K. Comparison of variable selection methods in Partial Least Squares Regression. *Journal of Chemometrics*. 2020;34:e3226. https://doi.org/10.1002/cem.3226