





## RESEARCH ARTICLE

# Selection of principal variables through a modified Gram–Schmidt process with and without supervision

Joakim Skogholt<sup>1</sup>  | Kristian H. Liland<sup>1</sup>  | Tormod Næs<sup>2</sup>  |  
Age K. Smilde<sup>3,4</sup>  | Ulf G. Indahl<sup>1</sup> 

<sup>1</sup>Faculty of Science and Technology,  
Norwegian University of Life Sciences, Ås,  
Norway

<sup>2</sup>Nofima AS, Ås, Norway

<sup>3</sup>Swammerdam Institute for Life Sciences,  
University of Amsterdam, Amsterdam,  
The Netherlands

<sup>4</sup>Simula Metropolitan Center for Digital  
Engineering, Oslo, Norway

## Correspondence

Joakim Skogholt, Faculty of Science and  
Technology, Norwegian University of Life  
Sciences, NO-1432 Ås, Norway.  
Email: [joakim.skogholt@nmbu.no](mailto:joakim.skogholt@nmbu.no)

## Abstract

In various situations requiring empirical model building from highly multivariate measurements, modelling based on partial least squares regression (PLSR) may often provide efficient low-dimensional model solutions. In unsupervised situations, the same may be true for principal component analysis (PCA). In both cases, however, it is also of interest to identify subsets of the measured variables useful for obtaining sparser but still comparable models without significant loss of information and performance. In the present paper, we propose a voting approach for sparse overall maximisation of variance analogous to PCA and a similar alternative for deriving sparse regression models influenced closely related to the PLSR method. Both cases yield pivoting strategies for a modified Gram–Schmidt process and its corresponding (partial) QR-factorisation of the underlying data matrix to manage the variable selection process. The proposed methods include score and loading plot possibilities that are acknowledged for providing efficient interpretations of the related PCA and PLS models in chemometric applications.

## KEYWORDS

modified Gram–Schmidt process, partial least squares regression, principal components, principal variables, QR-factorisation, variable selection

## 1 | INTRODUCTION

For highly multivariate datasets with a large number of measured variables, it is often attractive to reduce the data space dimensionality to obtain simpler and more interpretable models. For unsupervised problems, principal component analysis (PCA)<sup>1–4</sup> is probably the best known and most widely used dimension-reducing technique for such purposes. For supervised problems, principal component regression (PCR) and variants of partial least squares regression<sup>5–8</sup> (PLSR) are among the most popular choices.

It follows from the Eckart–Young theorem<sup>9</sup> that retaining the  $A$  principal components (PCs) with the largest associated variance produces the  $A$ -subset of linear combinations (of the  $p$  original variables being measured) which best approximates the full dataset. The latter means that for each fixed dimension  $A < p$ , no other subspace selection alternative is capable of accounting for more information in terms of explained variance than the associated dominant PCA

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2023 The Authors. *Journal of Chemometrics* published by John Wiley & Sons Ltd.

subspace. However, as stated by McCabe<sup>10</sup>: ‘... interpretation of the results and possible subsequent data collection with further analysis still involves all of the variables’. This is because each PC is a linear combination of all the original  $p$  variables. Hence, dimensionality reduction in terms of PCA does not provide a real reduction of complexity with respect to the original variables.

There have been many different approaches to improving the interpretability of PCA components. One approach is sparse PCA<sup>11</sup> that uses the lasso<sup>12</sup> to obtain sparseness in components, but these models are also problematic.<sup>13</sup> With sparse PCA, interpretability is improved by imposing sparseness on the loadings, but one still considers linear combinations of variables. Many other methods using the lasso for sparseness have been developed; see, for example, previous studies<sup>14–16</sup> for a review of some recent developments. Another approach to the problem of interpretability of PCA is due to Krzanowski.<sup>17</sup> Krzanowski criticised the use of only eigenvalues (of the covariance or correlation matrix) to select variables, as such criteria do not necessarily preserve relevant group structure in the data. The solution suggested by Krzanowski was to apply PCA to both the full data matrix and the matrix with only the selected variables, followed by using the norm of the residual after a Procrustes analysis on the two sets of scores (truncated to the estimated dimension of the data set) as a measure of the quality of the approximation. In practice this may require an extensive and computationally expensive search. This approach has been developed further since; see, for example, Guo et al<sup>18</sup> and Wang and Gehan.<sup>19</sup> Other approaches for enhancing interpretability of PCA include a probabilistic approach to variable selection,<sup>20</sup> using entropy optimisation principles,<sup>21</sup> as well as various forms of clustering.<sup>22,23</sup> See also previous studies<sup>24–29</sup> for reviews of various variable selection methods relevant to the present context.

McCabe<sup>10</sup> suggested the concept of principal variables (PVs) as subsets of the original variables that for a given size ( $A$ ) account for as much variance as possible in terms of the original dataset. In a later publication Cadima et al<sup>30</sup> noted that ‘The complete search, among all  $A$ -variable subsets, of a subset which optimises a given criterion, is a task which quickly becomes infeasible even for moderately sized dataset (unless  $A$  when compared with  $p$  is very small or very large)’. Thus, the overwhelming combinatorial nature of an exhaustive variable subset investigation calls for heuristic alternatives that are computationally feasible.

Our approach is based on considering PCA as an iterative voting process to select orthogonal directions maximising the explained variance, where in every step each direction gets a vote. We suggest a selection procedure for identifying PVs by restricting the variance-maximisation (voting) process of PCA to only consider the individual column vectors of the data matrix. In the suggested approach, the variables are selected sequentially according to the variance-maximisation criterion, resulting in an efficient greedy algorithm doing a locally optimal selection in each step. The selected variables will, as illustrated in the examples, often explain a similar amount of variance in the data as is explained by the associated PCA components.

The suggested variable selection procedure is straight forward to expand to regression problems, appearing as a unified approach to variable selection for both unsupervised and supervised problems. In the latter case, the method selects variables according to a combination of variance maximisation and correlations with the response. The predictions of the resulting models are directly comparable to the predictions obtained by PCR and PLSR.

Both alternatives are based on relatively simple and intuitive variable selection criteria and provide appropriate information for generating scores and loading plots of the type much used for model interpretations in chemometric applications of PCA and PLSR. By using the selected variables to model the full variable set, we also obtain a set of regression coefficients useful for describing, understanding, and plotting the relationships between the selected and unselected variables.

The structure of the paper is as follows. In Section 2, we describe the two variable selection algorithms in detail including some computational aspects. We also present and compare the theoretical properties of the suggested methods with Höskuldsson's PVs<sup>31</sup> (unsupervised selection) and the CovSel method of Roger et al.<sup>32</sup> (supervised selection). In Section 3, we apply the methods to two datasets that have been selected to demonstrate the various aspects of information content obtained by the two algorithms. We also discuss the stability in the resulting variable selections of both methods. Finally, in Section 4, we discuss further possible modifications regarding tailoring of the suggested methodology to an extended set of scenarios.

## 2 | THEORY

In the following, we assume that the columns of an  $(n \times p)$  data matrix  $\mathbf{X}$ , representing the measurements of  $p \geq 2$  variables for  $n$  distinct samples, and the associated response vector  $\mathbf{y}$  of dimension  $(n \times 1)$  are all centred (or standardised, if appropriate).

## 2.1 | The pivoted modified Gram–Schmidt (pMGS) process and QR-factorisation

The proposed voting principles for variable selection to be described below are combined with pMGS process.<sup>33,34</sup> The purpose of pMGS is to obtain an orthogonal basis for the column space of a data matrix  $\mathbf{X}$  for some associated problem to be solved.

A completed pMGS process results in a QR-factorisation

$$\mathbf{XW} = \mathbf{QR}, \quad (1)$$

where  $\mathbf{W}$  denotes a  $(p \times p)$ -permutation matrix, the columns of the matrix  $\mathbf{Q}$  span an orthogonal basis for the column space of  $\mathbf{X}$ , and  $\mathbf{R}$  is an upper triangular matrix of corresponding size with diagonal elements of decreasing absolute values. The permutation matrix  $\mathbf{W}$  results from interchanging the rows of the  $(p \times p)$ -identity matrix  $\mathbf{I}$  and represents the ranking (or selection) order of the columns (and corresponding variables) in the data matrix  $\mathbf{X}$ .

Traditionally, the pivoting in QR-factorisation is obtained by comparing the  $L_2$ -norms of the (deflated)  $\mathbf{X}$ -columns with selection of the column (variable) corresponding to the largest (residual) norm. The permutation matrix  $\mathbf{W}$  in (1) is obviously sparse and orthogonal ( $\mathbf{W}\mathbf{W}' = \mathbf{W}'\mathbf{W} = \mathbf{I}$  as the inverse of a permutation matrix is its transpose, where  $'$  denotes matrix transpose). Right multiplication by  $\mathbf{W}'$  in Equation (1) therefore yields the following factorisation of

$$\mathbf{X} = \mathbf{QRW}'. \quad (2)$$

The structural similarity between the factorisation in Equation (2) and the singular value decomposition (SVD)  $\mathbf{X} = \mathbf{USV}'$  presented in Equation (3), with associated comments, should be noted. The major differences between the two are (i) the upper triangular  $\mathbf{R}$  versus the diagonal  $\mathbf{S}$  (i.e., a special case of an upper triangular matrix), (ii) the sparsity of the orthogonal matrix  $\mathbf{W}$  versus the generally non-sparse orthogonal matrix  $\mathbf{V}$ , and (iii) the columns of  $\mathbf{V}$  represent a basis for the row space of  $\mathbf{X}$  whereas the columns of  $\mathbf{W}$  obviously form a basis for  $\mathbb{R}^p$ .

## 2.2 | PCA and its voting process for PV selection (PVS)

### 2.2.1 | PCA as a voting process

PCA works by compressing the information content of a data matrix  $\mathbf{X}$  with measurements on  $p$  variables into an  $A$ -dimensional subspace ( $A \leq p$ ) spanned by orthogonal linear combinations of the original variable measurements. The coefficients in the linear combinations (the loading vectors) coincide with the eigenvectors of the covariance matrix associated with  $\mathbf{X}$  and with the right singular (column) vectors in  $\mathbf{V}$  resulting from the SVD

$$\mathbf{X} = \mathbf{USV}'. \quad (3)$$

In (3), the columns of both  $\mathbf{U} = [\mathbf{u}_1 \mathbf{u}_2 \dots \mathbf{u}_n]$  and  $\mathbf{V} = [\mathbf{v}_1 \mathbf{v}_2 \dots \mathbf{v}_p]$  are mutually orthogonal, and  $\mathbf{S}$  is  $n \times p$  and diagonal in the sense that it is all zeros outside the diagonal elements  $s_{11} \geq s_{22} \geq \dots \geq s_{rr}$  (the singular values), where  $r = \min(n, p)$  is the rank of  $\mathbf{X}$ . (Below, we will suppress the double indexing and refer to the singular values as  $s_i$ ,  $1 \leq i \leq r$ .) The initial  $A$  column vectors of  $\mathbf{XV}$  are referred to as the first  $A$  PC scores

$$\mathbf{XV}_A = \mathbf{U}_A \mathbf{S}_A, \quad (4)$$

where  $\mathbf{V}_A$  and  $\mathbf{U}_A$  represent the first  $A$  columns of  $\mathbf{V}$  and  $\mathbf{U}$ , respectively, and  $\mathbf{S}_A$  is the initial  $A \times A$  submatrix of  $\mathbf{S}$ .

The score vectors in (4) represent a set of  $n$ -dimensional orthogonal basis vectors for the correspondingly dominant  $A$ -dimensional column subspace of  $\mathbf{X}$  in terms of overall empirical  $\mathbf{X}$ -variance. According to the Eckart–Young theorem,<sup>9</sup>

$$\mathbf{X} \approx \mathbf{U}_A \mathbf{S}_A \mathbf{V}'_A \quad (5)$$

is the best possible rank  $A$  approximation of  $\mathbf{X}$  in both the Frobenius and the spectral norm.

Regarding the column space  $Col(\mathbf{X})$  of  $\mathbf{X} = [\mathbf{x}_1 \mathbf{x}_2 \dots \mathbf{x}_p]$ , a simple algebraic manipulation shows that for any candidate unit vector  $\mathbf{q} \in Col(\mathbf{X}) \subseteq \mathbb{R}^n$ , the amount of empirical  $\mathbf{X}$ -variance projecting onto the direction of  $\mathbf{q}$  is given by the quadratic form

$$v_{\mathbf{X}}(\mathbf{q}) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^p (\mathbf{q}' \mathbf{x}_i)^2 = \frac{1}{n} \mathbf{q}' \mathbf{X} \mathbf{X}' \mathbf{q} = \frac{1}{n} \mathbf{q}' \mathbf{U} \mathbf{S} \mathbf{V}' \mathbf{V}' \mathbf{S}' \mathbf{U}' \mathbf{q} = \frac{1}{n} \sum_{j=1}^r s_j^2 (\mathbf{q}' \mathbf{u}_j)^2. \quad (6)$$

By considering the value  $v_{\mathbf{X}}(\mathbf{q})$  as the ‘ $\mathbf{X}$ -votes’ for the direction of  $\mathbf{q}$ , we can rank any set of candidate vectors from  $Col(\mathbf{X})$  in terms of the resulting votes obtained by plugging their associated unit vectors one-by-one into  $v_{\mathbf{X}}(\cdot)$ .

From inspection of the right-hand side in Equation (6), it follows that the maximum value of  $v_{\mathbf{X}}(\cdot)$  over all possible unit vectors is obtained for the dominant left singular vector  $\mathbf{u}_1$  resulting in the value  $v_{\mathbf{X}}(\mathbf{u}_1) = \frac{s_1^2}{n}$ . By imposing orthogonality constraints w.r.t. the dominant left singular vectors  $\mathbf{u}_1, \dots, \mathbf{u}_{k-1}$ ,  $v_{\mathbf{X}}(\cdot)$  is obviously maximised by  $v_{\mathbf{X}}(\mathbf{u}_k) = \frac{s_k^2}{n}$ , where  $\mathbf{u}_k$  is the  $k$ th column of the left singular vectors in  $\mathbf{U}$ .

## 2.2.2 | Heuristic PVS by variance voting

To obtain a subset of  $PVs$ , one can restrict the PCA-voting function in (6) to consider the problem of maximising the voting function  $v(\cdot)$  over the  $p$  normalised  $\mathbf{X}$ -columns  $\mathbf{z}_i = \mathbf{x}_i / \|\mathbf{x}_i\|$  for  $1 \leq i \leq p$ . The variable associated with the maximum  $v$ -value is selected, and in the case of a tie the tied variable with lowest index is selected.

Evaluation and selection of subsequent variables are constrained by ignoring the voting contributions in the directions of the previously selected variables. Technically, this requirement is dealt with by deflation of the data matrix  $\mathbf{X}$  with respect to the variable/column just selected before repeating the process until an appropriate subset of variables is obtained.

The following algorithm describes the required steps for implementing the PVS votings to select  $A \leq r$   $PVs$  from the (centred) data matrix  $\mathbf{X}$  of size  $n \times p$  and rank  $r$ . A MATLAB implementation of the algorithm is given in Appendix A, and the algorithm is illustrated in Figure 1.

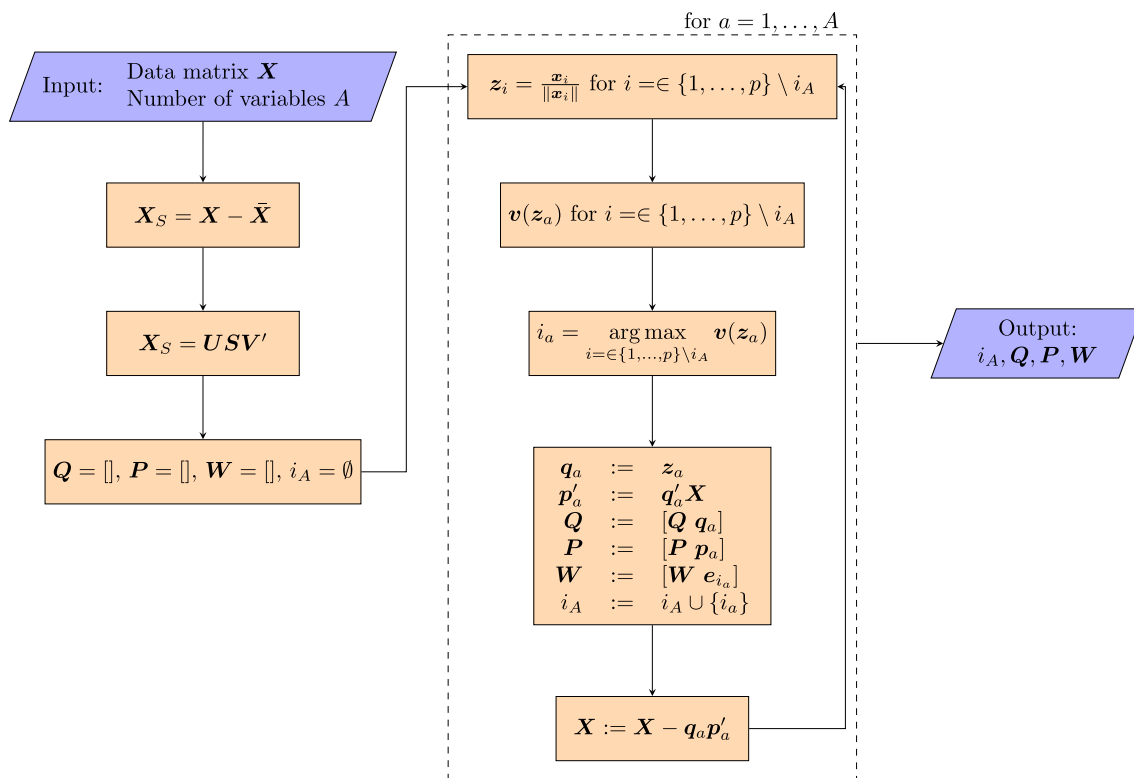


FIGURE 1 Flow chart for the PVS algorithm.

1. Centre  $\mathbf{X}$  by subtracting its  $p$  column means.
2. Compute the ‘economy version’ SVD of  $\mathbf{X} = \mathbf{U}_r \mathbf{S}_r \mathbf{V}'_r$ , required for the voting function (6) to be used in Step 4(b).
3. Initialise the ‘bookkeeping’ set of selections  $\mathbf{i}_A = \emptyset$  and let  $a = 1$ .
4. For  $a \leq A$ , find the  $a$ th PV by:
  - a. Calculate direction vectors  $\mathbf{z}_i = \mathbf{x}_i / \|\mathbf{x}_i\|$  for  $i \in \{1, \dots, p\} \setminus \mathbf{i}_A$ .
  - b. Calculate the votes  $v(\mathbf{z}_i)$  for each vector from the step above.
  - c. Identify the index  $i_a$  associated with the largest ‘voting’ value in 4(b).
  - d. Define  $\mathbf{q}_a = \mathbf{z}_{i_a}$ . Calculate the  $\mathbf{q}_a$ -coordinates  $\mathbf{p}_a \in \mathbb{R}^p$  of each  $\mathbf{X}$ -column:  $\mathbf{p}'_a = \mathbf{q}'_a \mathbf{X}$  and set  $\mathbf{w}_a = \mathbf{e}_{i_a} \in \mathbb{R}^p$ , where  $\mathbf{e}_{i_a}$  is the  $i_a$ th column of the  $p \times p$  identity matrix (the  $i_a$ th standard basis vector). Update the set of selections  $\mathbf{i}_A = \mathbf{i}_A \cup \{i_a\}$ .
  - e. Deflate  $\mathbf{X}$  with respect to the direction of the chosen variable:  
 $\mathbf{X} := \mathbf{X} - \mathbf{q}_a \mathbf{p}'_a$  (the modified Gram–Schmidt step).
  - f. Set  $a = a + 1$  and repeat 4.
5. Set  $\mathbf{Q}_A = [\mathbf{q}_1 \mathbf{q}_2 \dots \mathbf{q}_A]$ ,  $\mathbf{P}_A = [\mathbf{p}_1 \mathbf{p}_2 \dots \mathbf{p}_A]$ ,  $\mathbf{W}_A = [\tilde{\mathbf{W}}_A \mathbf{J}_A]$  where  $\tilde{\mathbf{W}}_A = [\mathbf{w}_1 \mathbf{w}_2 \dots \mathbf{w}_A]$  and  $\mathbf{J}_A$  is the reduced matrix obtained by eliminating from the  $p \times p$  identity matrix all columns corresponding to the completed selection  $\mathbf{i}_A = \{i_1, i_2, \dots, i_A\}$ .

Note that  $\mathbf{X}'\mathbf{Q}_A = \mathbf{P}_A$  is the matrix of  $\mathbf{X}$ -loadings associated with the orthogonal columns of  $\mathbf{Q}_A$  that spans the same subspace as the selected variables  $\mathbf{X}\tilde{\mathbf{W}}_A$ . The  $A \times p$ -matrix  $\mathbf{R}_A = \mathbf{P}'_A \mathbf{W}_A$  is rectangular and in upper triangular form (everything below the diagonal is zero), where the permutation matrix  $\mathbf{W}_A$  acts on the columns of  $\mathbf{P}'_A$ , and the  $a$ th selected variable is identified by the non-zero element in the  $a$ th column of  $\mathbf{W}_A$ . The associated PVS of  $A < r$  variables results in a partial and pivoted QR-factorisation of the column permuted data matrix:

$$\mathbf{X}\mathbf{W}_A \approx \mathbf{Q}_A \mathbf{Q}'_A \mathbf{X}\mathbf{W}_A = \mathbf{Q}_A \mathbf{P}'_A \mathbf{W}_A = \mathbf{Q}_A \mathbf{R}_A. \quad (7)$$

Right multiplication with  $\mathbf{W}'_A$  in Equation (7) yields a rank  $A$  approximation of the original  $\mathbf{X}$  analogous to (5), that is,

$$\mathbf{X} \approx \mathbf{Q}_A \mathbf{P}'_A = \mathbf{Q}_A \mathbf{R}_A \mathbf{W}'_A. \quad (8)$$

Note that the latter approximation includes an exact reproduction of the selected  $\mathbf{X}$ -columns specified by the indices in  $\mathbf{i}_A$ , that is, the first  $A$  columns of (7) yield the exact pivoted QR-factorisation of the submatrix

$$\mathbf{X}\tilde{\mathbf{W}}_A = \mathbf{Q}_A \tilde{\mathbf{R}}_A, \quad (9)$$

associated with the selected PVs. The square upper triangular  $\tilde{\mathbf{R}}_A$  denotes the first  $A$  columns of  $\mathbf{R}_A$ . As the left-hand side of Equation (9) consists of the  $A$  selected variables in  $\mathbf{X}$  according to the selection order, the right-hand side simply gives the QR-factorisation of this matrix. Right multiplication with  $\tilde{\mathbf{R}}_A^{-1}$  results in an expression for the orthogonal basis of the column space of the  $\mathbf{X}$ -columns selected by PVS:

$$\mathbf{Q}_A = \mathbf{X}\tilde{\mathbf{W}}_A \tilde{\mathbf{R}}_A^{-1}. \quad (10)$$

The approximation given in Equation (8) is in exact arithmetic equivalent to the approximation obtained by regressing each variable in the data set on the variables selected by the PVS algorithm. The associated regression coefficients  $\mathbf{C}_A$  to model the full  $\mathbf{X}$  matrix from the selected PVs  $\mathbf{X}\tilde{\mathbf{W}}_A$  are given by the least squares solution with respect to  $\mathbf{C}_A$  of  $\mathbf{X} \approx (\mathbf{X}\tilde{\mathbf{W}}_A)\mathbf{C}_A = (\mathbf{Q}_A \tilde{\mathbf{R}}_A)\mathbf{C}_A$ , that is, the minimiser of

$$\|(\mathbf{X}\tilde{\mathbf{W}}_A)\mathbf{C}_A - \mathbf{X}\|^2 = \|(\mathbf{Q}_A \tilde{\mathbf{R}}_A)\mathbf{C}_A - \mathbf{X}\|^2, \quad (11)$$

with respect to  $\mathbf{C}_A$ :

$$\hat{\mathbf{C}}_A = \tilde{\mathbf{R}}_A^{-1} \mathbf{Q}'_A \mathbf{X} = \tilde{\mathbf{R}}_A^{-1} \mathbf{P}'_A. \quad (12)$$

### 2.2.3 | Plotting tools for model interpretations

The PVS method(s) yields plotting possibilities similar to those of PCA and PLS models. The *score plots* of a PVS model are obtained as scatter plots based on selected pairs (or triples) of the first few columns in the  $\mathbf{Q}_A$  matrix. The *loading plots* obtained by plotting the columns of the  $\mathbf{P}_A$ -matrix provide an effective visualisation of the relationships between the selected variable directions in  $\mathbf{Q}_A$  and the original  $\mathbf{X}$ -data (as  $\mathbf{P}_A = \mathbf{X}'\mathbf{Q}_A$ ). To explore the relationships between the  $A$  selected (principal) variables  $\mathbf{X}\tilde{\mathbf{W}}_A$  and the complete set of variables in  $\mathbf{X}$ , one can consider visualisation of the  $\hat{\mathbf{C}}_A$  regression coefficients in a fashion similar to the  $\mathbf{P}_A$ -loadings. Note that while the scores are orthonormal, this is not the case for the loadings. To our knowledge, this is the case for all variable selection methods where there is correlation between the selected variables. The loadings are non-orthogonal, but are in the row space of  $\mathbf{X}$ . The non-orthogonality of the loadings may in some cases make interpretation difficult.<sup>35</sup>

### 2.2.4 | Equivalence of PVS and Cadima and Jolliffe's RM-criterion

The PVS variable selection strategy presented here is mathematically equivalent to Cadima and Jolliffe's RM-criterion,<sup>30,36</sup> which is also equivalent to the second of McCabe's four criteria<sup>10</sup> when the variable selection is done greedily. Their criterion is the maximisation of the correlation between the (centred) data matrix and its projection onto the column space of a subset of its columns. By letting  $\mathbf{H}_K$  denote the projection matrix onto a subset of  $K$  columns from the data matrix, Cadima and Jolliffe,<sup>36</sup> p. 68 states:

The maximisation of  $\text{corr}(\mathbf{X}, \mathbf{H}_K \mathbf{X})$  therefore selects the  $k$ -variable subset that maximises the same criterion (variance) as PCA, though here we are restricted to subsets of the observed variables rather than the subsets of all linear combination of those variables.

Hence, the PVS variable selection strategy is equivalent to the RM-criterion when the variable selection is done greedily. The advantage of the above PVS algorithm is that it is a computationally efficient method for implementing the RM-criterion and, as demonstrated in the examples, the selected variables are directly comparable to the PCs scores in the amount of explained variance. The variable subsets selected by PVS are not necessarily optimal, and different variable subsets may be obtained using methods such as simulated annealing and genetic algorithms, as emphasised in Cadima et al.<sup>30</sup> The RM-criterion only applies to unsupervised problems, and below, we will later discuss how the ideas of PVS can be adjusted to obtain useful selections also for supervised problems. The PVS criterion is also similar to Wei and Billings<sup>37</sup> and Whitley et al.,<sup>38</sup> but they consider explained correlation rather than variance. Puggini and McLoone<sup>39</sup> consider the same criterion as PVS, but the algorithm used to obtain the variables is different.

### 2.2.5 | Höskuldsson's PVs—Discussion and comparison with PVS

In the examples, we have included a comparison between PVS and a variable selection method by Höskuldsson.<sup>31</sup> The latter method is briefly reviewed here. Höskuldsson<sup>31</sup> suggests a slightly different variable selection criterion where variables are selected according to

$$\arg \max_{i=1, \dots, p} \left[ (\mathbf{x}'_1 \mathbf{x}_i)^2 + (\mathbf{x}'_2 \mathbf{x}_i)^2 + \dots + (\mathbf{x}'_p \mathbf{x}_i)^2 \right]. \quad (13)$$

After selecting a variable, the matrix  $\mathbf{X}$  is deflated w.r.t. the chosen variable before the next variable is selected.

This variable selection procedure is similar but not equivalent to the PVS algorithm suggested in the present work. As there is no normalisation of the selected direction in Höskuldsson's criterion, the selected variables depend not only on the direction (which PVS does) but also on the size of the numbers in the matrix. Hence, PVS is better at picking out hidden correlations in the data, as a column with relatively small norm that is correlated with other columns can potentially explain more of the variance in the data than the column with the largest norm. Furthermore, as we will see in the examples, it appears that Höskuldsson's criterion may be more affected by random noise.



## 2.3 | PLSR and an associated voting process for selecting principal regression variables (PRVs)

### 2.3.1 | PLSR

PLSR<sup>5–7</sup> represents an alternative data compression strategy resembling PCA that also takes into consideration the measured responses in the dimension reduction process. The goal of PLSR is to achieve good linear modelling of the responses from low-dimensional linear combinations of the predictors. PLS regression is therefore much used to handle situations where the number of measured predictors is considerably larger than the number of measured samples ( $p \gg n$ ).

Similar to PCA, PLS can be considered as a ‘voting’ process working by repeated maximisation of the overall covariance between the (deflated) predictors ( $\mathbf{X}$ -columns) and the associated (deflated)  $n$ -dimensional response vector  $\mathbf{y}$ . The extraction of  $A$  PLS components results in the matrix product identity

$$\mathbf{X}\mathbf{W}_{A,PLS} = \mathbf{T}_{A,PLS}\mathbf{B}_{A,PLS}, \quad (14)$$

where the normalised  $n$ -dimensional column vectors of  $\mathbf{T}_{A,PLS}$  are orthogonal, the normalised  $p$ -dimensional column vectors of  $\mathbf{W}_{A,PLS}$  (the non-sparse loading weights) are also orthogonal, and  $\mathbf{B}_{A,PLS}$  is  $A \times A$  upper bidiagonal (yet another special case of an upper triangular matrix).

From  $\mathbf{P}_{A,PLS} = \mathbf{X}'\mathbf{T}_{A,PLS}$  (known as the *PLS loadings*), the matrix  $\mathbf{B}_{A,PLS}$  is given by  $\mathbf{B}_{A,PLS} = \mathbf{P}'_{A,PLS}\mathbf{W}_{A,PLS}$ . Unless there is some subspace of the column space of  $\mathbf{X}$  that is exactly orthogonal to the response vector  $\mathbf{y}$ , the maximum possible number of PLS components is equal to the rank ( $r$ ) of  $\mathbf{X}$ .

In PLS, the first column ( $\mathbf{w}_{PLS}$ ) of  $\mathbf{W}_{A,PLS}$  is the normalised version of the vector  $\mathbf{X}'\mathbf{y}$  and the corresponding first column  $\mathbf{t}_{PLS}$  of  $\mathbf{T}_{A,PLS}$  is the normalised version of the associated vector  $\mathbf{X}\mathbf{w}_{PLS}$  that solves the covariance-maximisation problem. After computing the residual  $\mathbf{X}$  and  $\mathbf{y}$  by the deflation steps  $\mathbf{X} := \mathbf{X} - \mathbf{t}_{PLS}(\mathbf{t}'_{PLS}\mathbf{X})$  and  $\mathbf{y} := \mathbf{y} - \mathbf{t}_{PLS}(\mathbf{t}'_{PLS}\mathbf{y})$ , respectively, the covariance-maximisation process is repeated to extract the subsequent  $A - 1$  components to be accounted for by the columns of  $\mathbf{W}_{A,PLS}$  and  $\mathbf{T}_{A,PLS}$ . The implementation of this process is known as the NIPALS algorithm<sup>5</sup> for PLS regression. Alternative and faster PLS regression algorithms<sup>7</sup> giving the factorisations in Equation (14) are also available, but the variable selection process described below will include deflation steps similar to those just described.

### 2.3.2 | Selection of PRVs by variance multiplication voting

The fundamental idea of PLS is to favour directions (linear combinations)  $\mathbf{q}$  in the  $\mathbf{X}$ -variable columns that simultaneously captures both dominant  $\mathbf{X}$ - and  $\mathbf{y}$ -variance. A voting function candidate for this purpose can be obtained by multiplying together the empirical  $\mathbf{X}$ - and  $\mathbf{y}$ -variances that are simultaneously accounted for in the  $\mathbf{q}$ -direction, that is,

$$v_{y,reg}(\mathbf{q}) = \frac{1}{n^2} \mathbf{q}'\mathbf{X}\mathbf{X}'\mathbf{q} \cdot \mathbf{q}'\mathbf{y}\mathbf{y}'\mathbf{q} = v_{\mathbf{X}}(\mathbf{q}) \cdot v_{\mathbf{y}}(\mathbf{q}), \quad (15)$$

where  $v_{\mathbf{y}}(\cdot)$  is analogous to the PVS-voting function defined in Equation (6). The corresponding pseudo-algorithm for implementing the *PRV* voting principle is obtained by making only three minor modifications of the proposed PVS algorithm in Section 2.2.2:

- Step 1: Centre the response  $\mathbf{y}$  as well as the data matrix  $\mathbf{X}$ .
- Step 4(b): Replace  $v_{\mathbf{X}}(\mathbf{z}_i)$  by the voting function  $v_{y,reg}(\mathbf{z}_i)$ .
- Step 4(e): Extend the step to include the deflation of  $\mathbf{y} := \mathbf{y} - \mathbf{q}_a(\mathbf{q}'_a\mathbf{y})$ .

The regression coefficients are obtained by regressing the response onto the selected variables. A MATLAB implementation of the PRV algorithm is given in Appendix A. This modified algorithm also allows for an approximation to

the data matrix  $\mathbf{X}$  from  $\mathbf{XW}_A$  as described for PVS. Loading and score plots are obtained in the same way as with PVS—the only difference is in the selected variables described by  $\mathbf{W}_A$ . The suggested criterion is similar to the one considered by Billings and Wei,<sup>40</sup> but their criterion selects variables that maximise correlation with the response without considering the explained  $\mathbf{X}$ -variance.

### 2.3.3 | CovSel—Description and comparison with PRV

In the examples below, we compare PRV with the variable selection method known as CovSel.<sup>32</sup> CovSel also works with multiple responses, but in this paper, we only consider univariate problems. As both CovSel and PRV are methods inspired by PLS, CovSel is a natural benchmark for PRV. Further, the differences between the two methods are useful for illustrating the properties of PRV.

At each iteration, CovSel selects the variable maximising the covariance with the response and more precisely maximises

$$\arg \max_{i=1, \dots, p} (\mathbf{x}'_i \mathbf{y})^2, \quad (16)$$

followed by a deflation of the predictor matrix with respect to the selected variable.

Although CovSel and PRV are quite similar, the differences in the variable selection criterion may have a large impact on some datasets. To understand the differences between the two variable selection methods, we rewrite both criteria slightly to make the comparison easier. The CovSel criterion can be written as

$$\arg \max_{i=1, \dots, p} (\|\mathbf{y}\| \cdot \cos(\theta_i))^2 \cdot \|\mathbf{x}_i\|^2 = \arg \max_{i=1, \dots, p} \cos^2(\theta_i) \cdot (\mathbf{x}'_i \mathbf{x}_i)^2, \quad (17)$$

where  $\theta_i$  is the angle between  $\mathbf{y}$  and  $\mathbf{x}_i$ . The PRV criterion can be written as

$$\arg \max_{i=1, \dots, p} (\|\mathbf{y}\| \cdot \cos(\theta_i))^2 \cdot \sum_{j=1}^r (s_j \cdot \cos(\phi_{ij}))^2 = \arg \max_{i=1, \dots, p} \cos^2(\theta_i) \cdot \sum_{j=1}^r (s_j \cdot \cos(\phi_{ij}))^2, \quad (18)$$

where  $\theta_i$  again is the angle between  $\mathbf{y}$  and  $\mathbf{x}_i$  and  $\phi_{ij}$  is the angle between  $\mathbf{x}_i$  and the  $j$ th PC. By comparing (17) and (18), we see that CovSel favours variables that are both highly correlated with the response and have high norm, whereas PRV prefers variables that are both highly correlated with the response and highly correlated with PCs with larger (relative to the other) singular values.

The deflation in CovSel causes a large reduction in the norm of any remaining variables that are highly correlated to the selected variable. Hence, CovSel will typically not choose multiple highly correlated variables. There is a similar effect in PRV, but since PRV maximises a weighted sum of squared correlations with the PCs, it is not as sensitive to the reductions in norm caused by deflations as CovSel is.

An alternative view to the difference between the PRV and CovSel is that, subject to correlation with the response, CovSel identifies variables explaining local variance (as CovSel considers each variable in isolation without looking at the total explained variance in the data), whereas PRV identifies variables capturing global variance (as PRV considers the total explained variance in the data by each variable). A consequence of this difference is that CovSel is typically more stable in the variable selections than PRV. When the collected data are from a single distribution, the variances in isolated variables are less likely to be affected by small disturbances in the training set, and hence, CovSel will generally select variables more consistently. For dataset with highly correlated variables, there may be numerous correlated variables explaining approximately the same fraction of the total variance in the data, and hence, PRV will be more sensitive to such disturbances in the training data. As demonstrated in the examples below, PRV and CovSel are both useful alternatives, and the choice of method depends on properties of the data set subject to analysis.



## 3 | EXAMPLES

### 3.1 | Datasets

#### 3.1.1 | Beer data

The data set consists of 60 NIR spectra of beer<sup>41</sup> and is shown in Figure 2A. Wavelengths were measured in 2 nm intervals from 400 to 2250 nm giving measurements for a total of 926 wavelengths. The response variable is the amount of real extract concentration, which is useful for determining the ability of yeast to ferment alcohol. From the plot of the spectra we see that the region from about 1400 to 2250 nm is particularly noisy, and this is partly because of water.<sup>41</sup> See also Andersen and Bro<sup>42</sup> for an analysis of the data set in a variable selection context.

#### 3.1.2 | Raman spectra of pork adipose tissue

The data set consists of 77 Raman spectra of pork adipose tissue.<sup>43,44</sup> Raman intensity was measured in intervals of  $0.6 \text{ cm}^{-1}$ , and after truncation, the included wavenumbers are in the range  $819.6\text{--}3099 \text{ cm}^{-1}$ . The data set contains responses for saturated fatty acids (SFA), monounsaturated fatty acids (MUFA), polyunsaturated fatty acids (PUFA), and the iodine value. Here, we only include the results for the PRV modelling of iodine. Prior to analysis, the data were pre-processed using EMSC pre-processing<sup>45</sup> with degree 6 polynomial correction and interferent correction, as done in Liland et al.<sup>44</sup> For the reference spectrum in the EMSC pre-processing the first right singular vector of the data set (or training set when appropriate) was used. The pre-processed spectra are shown in Figure 3A. For inspection of the raw data, see Liland et al.<sup>44</sup>

## 3.2 | Unsupervised

### 3.2.1 | Methodology (unsupervised)

For unsupervised problems, we compare PVS with PCA and Höskuldsson's PV. We use the full datasets, and for the spectroscopic data, we also compare the variables selected with the regression variants of the methods to illustrate how the inclusion of a response affects the variable selection.

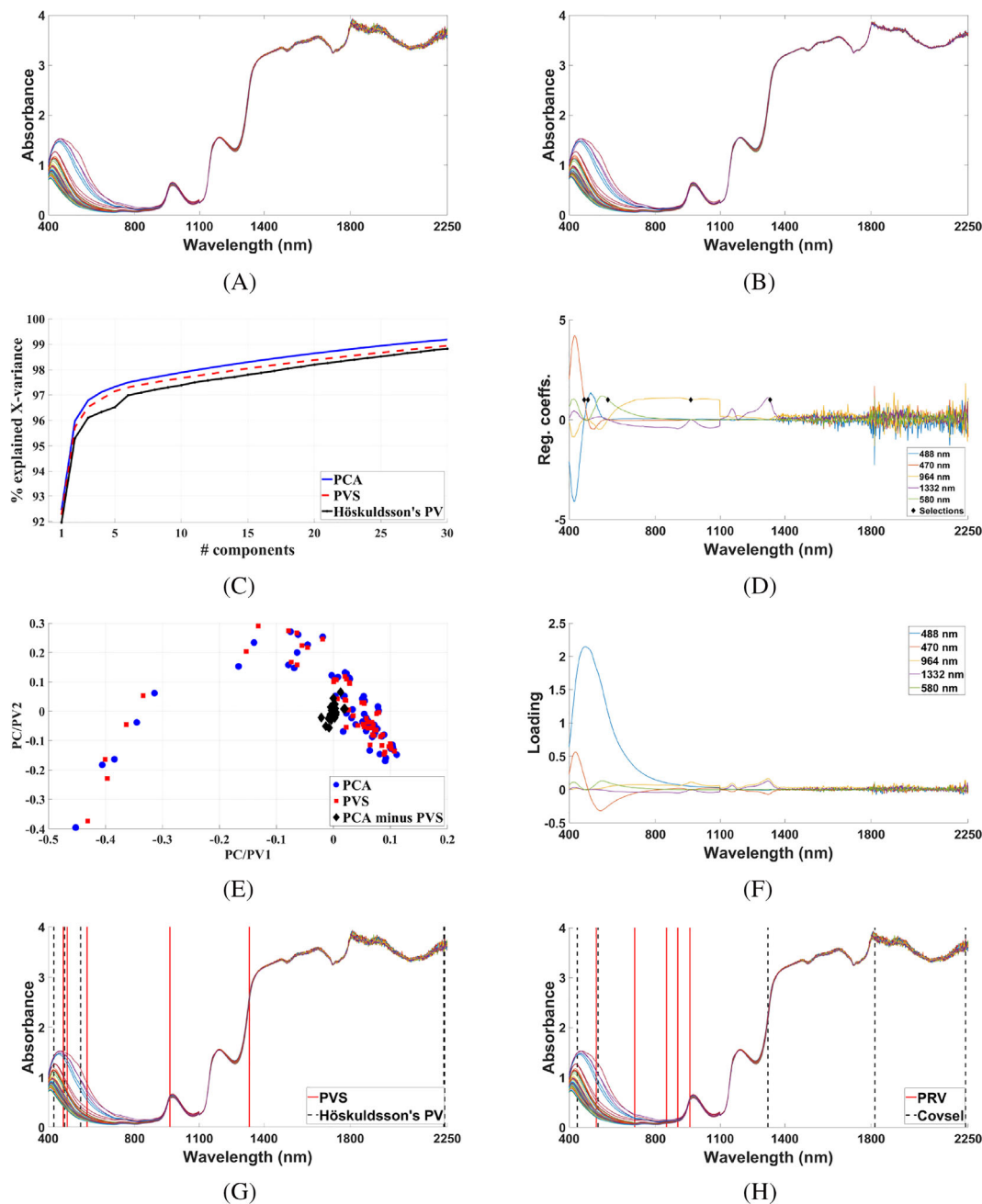
#### 3.2.2 | Beer data

Figure 2C shows explained variance as a function of the number of components (PCA) and variables (PVS and Höskuldsson's PV). We see that the variance explained by PVS closely matches the variance explained by PCA, and hence, it is possible to obtain a lower dimensional model comparable in explained variance to PCA by selecting a handful of variables instead of using linear combinations of all the variables.

Höskuldsson's PV does not explain as much of the variance in the data as PCA and PVS, which is a consequence of the difference between the two variable selection criteria. This can also be seen in Figure 2G where the first five variables selected by PVS and Höskuldsson's PV are shown. Two of the first five variables selected by Höskuldsson's criterion are in the noisy part of the data (notably, the first three variables selected are not from the noisy part of the data), whereas none of the first five variables selected by PVS are from that part of the spectra.

Of the first 10 variables selected, Höskuldsson's criterion selects six from the noisy part of the spectra, and PVS selects three variables from the noisy part of the spectra.

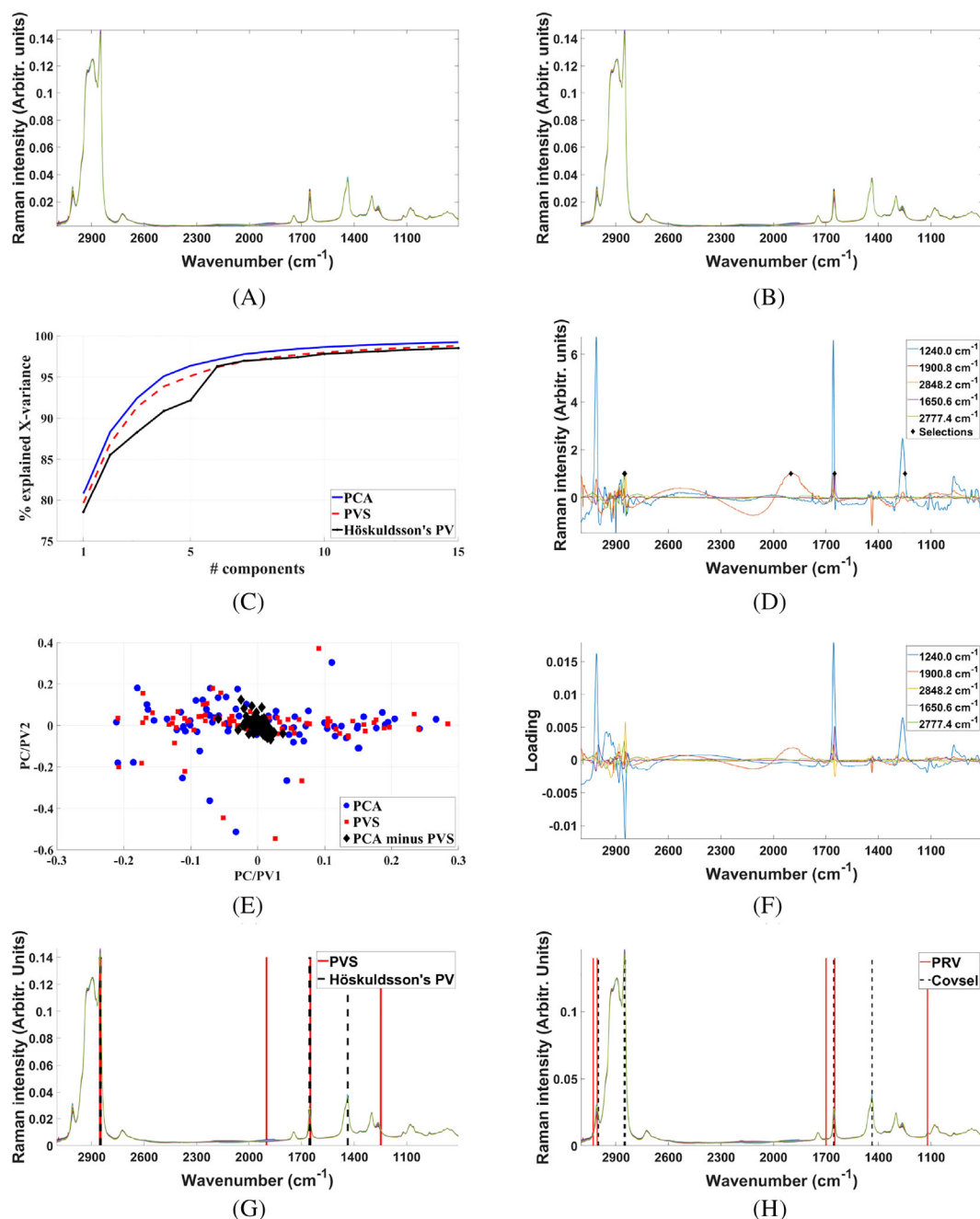
For exploring some of the differences in the variables selected by the unsupervised (PVS and Höskuldsson's method) and supervised (PRV and CovSel) approaches, the first five variables selected by the different methods applied to the whole data set are shown in Figure 2G,H.



**FIGURE 2** (A) Plot of the raw NIR spectra. (B) Rank 5 approximation to the spectra using PVS variables. (C) Plot of the cumulative explained  $X$ -variance as a function of the number of components (PCA) and variables (PVS and Höskuldsson's PV). (D) Plot of the regression coefficients obtained when regressing the full data set on the selected variables. (E) Score plot of the data with scores from PCA and PVS. The pairwise differences between the PCA and PVS scores are also shown to illustrate the similarities between the two sets of scores. (F) Loading plot for the first five variables selected by PVS. (G) The first five variables selected by PVS and Höskuldsson's PV. (H) The first five variables selected by PRV and CovSel for the whole data set.

Figure 2B shows the approximation obtained by regressing the data set on the first five variables selected by PVS. Note that a regularising effect is exhibited in the noisy part of the spectra in the approximation. Figure 2C shows the associated regression coefficients for each of the selected variables.

The PVS scores and loadings, that is, the first columns in the  $Q_A$ -matrix and  $P_A$ -matrix, respectively, are given in Figure 2E,F, respectively. Figure 2E also includes PCA scores for comparison, and we see that the PVS scores are very similar to the PCA scores. This is a consequence of PVS capturing approximately the same variance as PCA with two variables as shown in Figure 2C.



**FIGURE 3** (A) Plot of the pre-processed Raman data. (B) Rank 5 approximation to the spectra using PVS variables. (C) Plot of the cumulative explained X-variance as a function of the number of components (PCA) and variables (PVS and Höskuldsson's PV). (D) Plot of the regression coefficients obtained when regressing the full data set on the selected variables. (E) Score plot of the data with scores from PCA and PVS. The pairwise differences between the PCA and PVS scores are also shown to illustrate the similarities between the two sets of scores. (F) Loading plot for the first five variables selected by PVS. (G) The first five variables selected by PVS and Höskuldsson's PV. (H) The first five variables selected by PRV and CovSel for the whole data set.

### 3.2.3 | Raman data

Figure 3C shows the explained variance as a function of the number of components for PCA, PVS, and Höskuldsson's PV. For the first five components, PVS explains more of the total variance than Höskuldsson's PV, but notably for exactly six and seven components, it is the other way around. Although PVS typically will explain more variance than Höskuldsson's PV (with the same number of variables), PVS (like Höskuldsson's PV) is a greedy algorithm and will not

necessarily find a global optimum for the explained variance. The first five variables selected for the data set for Höskuldsson's PLS and PLS are shown in Figure 3G, and the first five variables selected with CovSel and PRV are shown in Figure 3H.

Note that in both the unsupervised and the supervised case, PLS/PRV select at least one variable from a flat part of the spectra that do not appear to be particularly relevant for prediction. These variables are favoured by PLS/PRV because they explain a significant amount of total variance in the data, even though the variance of each variable is small. Moreover, it is worthwhile to note that the approximation of the data set using the PLS variables, shown in Figure 3B, clearly reproduces some useless (in terms of prediction) variation in the spectra (approximately between 1800 and 2600  $\text{cm}^{-1}$ ). Figure 3E,F shows the PLS scores and loading plots, respectively. As with the beer data, we see that the score plot obtained by PLS is very similar to the score plot by PCA.

### 3.3 | Supervised

#### 3.3.1 | Methodology (supervised)

In the following examples, we compare PRV with results obtained by PCR, PLS, and CovSel. To compare the different variable selection methods, the following methodology was applied: The data set was randomly divided into 5000 train/validation/test splits. Fifty per cent of the data was used for training, 25% for validation, and the remaining 25% for testing. For each data split, the models were trained on the training set, and the number of components of the model used for the test set was the model with the lowest RMSE on the validation set. For PRV and CovSel, the selected variables were used as predictor variables in ordinary least squares regression.

For the beer data, models were built both on the full data set and after truncating the spectra to wavelengths in the range 800–1400 nm. The latter truncation contains the interval that is relevant for prediction<sup>41,42</sup> and avoids the irrelevant for prediction<sup>41,42</sup> systematic variation in the 400–800 nm range and the noisy region of the spectra.

For the Raman data, the interferent was estimated with the difference of two spectra with similar value for the response, as done in Liland et al.<sup>44</sup> This difference spectrum was added to the EMSC model, and the projection of the spectra onto this subspace was removed from the remaining spectra. To avoid data leakage, the two spectra used to estimate the interferent were removed from the data set, and the pre-processing model was calculated only from the training set for each data split.

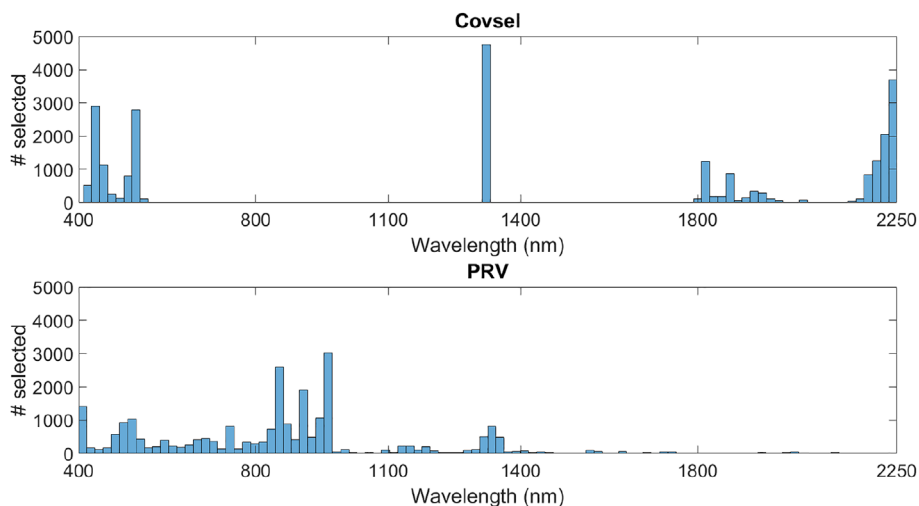
#### 3.3.2 | Results—Beer data

From Table 1, we see that the average RMSEP on the whole spectra is by far lowest for PRV. PRV also has the lowest standard deviation of RMSEP indicating little sensitivity to the train/validation/test set split, whereas the other methods appear to be more sensitive to the data split.

PRV mostly avoids the noisy region, and this appears to be the reason why the PRV models are better. This is supported by the results for the truncated spectra also given in Table 1. By truncating the spectra to the relevant region, all methods give better models, and the standard deviation of RMSEP is much lower indicating that all methods have some problems with the uninteresting parts of the spectra. The histogram in Figure 4, showing the first five variables

**TABLE 1** Beer: Summary of results across all training test set splits for full spectra and spectra truncated to the region 800–1400 nm.

|   | Cov  | PRV  | PLS  | PCR  |
|---|------|------|------|------|
| Average number of variables/components (full spectra)   | 3.6  | 8.2  | 7.4  | 24.4 |
| Average number of variables/components (trunc. spectra) | 7.1  | 7.4  | 6.4  | 14.4 |
| Average RMSEP (full spectra)                            | 0.55 | 0.35 | 0.88 | 0.89 |
| Average RMSEP (trunc. spectra)                          | 0.33 | 0.20 | 0.21 | 0.22 |
| Standard deviation of RMSEP (full spectra)              | 0.55 | 0.20 | 0.39 | 0.40 |
| Standard deviation of RMSEP (trunc. spectra)            | 0.09 | 0.06 | 0.06 | 0.05 |



**FIGURE 4** Beer data—Histogram indicating the number of times a variable is selected among the first five variables for CovSel (top) and PRV (bottom). The part of the spectrum relevant for prediction of the response is 800–1400 nm.

**TABLE 2** Raman data: Summary of results across all training test set splits.

|   | Cov  | PRV  | PLS  | PCR  |
|---|------|------|------|------|
| Average number of variables/components (full spectra)                     | 5.9  | 5.0  | 4.9  | 12.9 |
| Average number of variables/components (trunc. to intervals around peaks) | 6.6  | 5.4  | 5.0  | 11.8 |
| Average number of variables/components (trunc. to eight variables)        | 4.2  | 3.0  | 4.0  | 4.9  |
| Average RMSEP (full spectra)  | 1.12 | 1.22 | 1.03 | 1.03 |
| Average RMSEP (trunc. to intervals around peaks)                          | 1.11 | 1.26 | 1.05 | 1.04 |
| Average RMSEP (trunc. to eight variables)                                 | 1.12 | 1.03 | 1.08 | 1.08 |
| Standard deviation of RMSEP (full spectra)                                | 0.37 | 0.41 | 0.38 | 0.38 |
| Standard deviation of RMSEP (trunc. to intervals around peaks)            | 0.38 | 0.43 | 0.39 | 0.38 |
| Standard deviation of RMSEP (trunc. to eight variables)                   | 0.39 | 0.40 | 0.40 | 0.40 |

*Note:* In the interval truncation, the spectra are truncated to intervals around peaks, and for the truncation to eight variables, the functional group frequencies from Olsen et al<sup>43</sup> is used.

selected by CovSel and PRV across all datasets, also gives a clear indication of why PRV works better than CovSel on the beer data set. The most commonly selected variable by CovSel is (for almost all data splits) from the part of the spectra relevant for prediction (according to Nørgaard et al<sup>41</sup> and Andersen and Bro<sup>42</sup>), but the other four variables are typically from either the noisy part of the spectrum or the region containing the systematic variation that is not relevant for prediction of the response. This is a consequence of the CovSel criterion as discussed in Section 2.3.3.

In the noisy part of the spectrum, there will be some spurious correlations with the response, and the noise also causes the variables in this part of the spectrum to have fairly large norm/variance. As CovSel maximises the product of the squared correlation with the response and the norm of a variable, it seems clear that CovSel will select variables from the noisy part of the spectrum. For the beer data set, PRV is able to avoid this area of the spectrum as the PCs covering the variation in the noisy part of the spectra correspond to smaller singular values and hence are weighted down by the PRV criterion.

### 3.3.3 | Results—Raman data

From Table 2, we see that for the full spectra, on average, PLS and PCR have the lowest RMSEP, followed by CovSel and then PRV. Although the differences in the RMSEP averages are smaller than 1 SE error of RMSEP, it does seem to

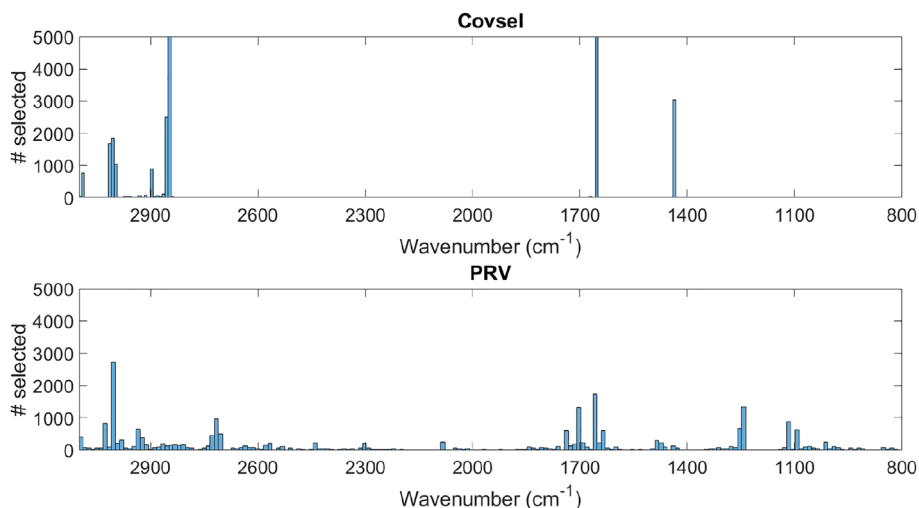


FIGURE 5 Raman data—Histogram indicating the number of times a variable is selected among the first five variables for CovSel (top) and PRV (bottom).

be the case that the CovSel models are better than the PRV models as CovSel generally selects more chemically relevant variables than PRV. The histograms in Figure 5 show the first five variables selected by PRV and CovSel across all dataset splits. As with the beer data, CovSel is much more stable in the variable selection than PRV, which selects a much wider variety of variables in the spectra. This is also the case for both the beer data and the Raman data with PVS (more variety in the selected variables) and Höskuldsson's PVs (more stable in the selected variables relative to PVS). The four largest peaks in the top histogram in Figure 5 are close to peaks in the data that are relevant for the prediction of iodine values, but this is only the case for the two largest peaks in the histogram for PRV. Rerunning the analysis allowing PRV to use a maximum of two variables gives almost the same average RMSEP (1.23), supporting the fact that PRV mostly selects two variables that are relevant for prediction. Inspection of the variables selected shows that PRV selects several variables outside of the peaks in the data. The reason for this is that although the variance in each such variable is small, the total  $\mathbf{X}$ -variance in all these variables combined is quite large, and as part of the PRV criterion is to maximise the total variance explained in the data it is more susceptible to selecting such variables. CovSel does not have this issue as it looks at each variable in isolation rather than the total explained variance. Hence, the advantage PRV has over CovSel for the beer data set is a disadvantage for PRV for this data set. From Table 2, we see that the best results for PRV are obtained by truncating the spectra to the eight functional group frequencies shown in fig. 1 in Olsen et al.<sup>43</sup>

Investigations with other data set indicate that CovSel in general has better prediction results on Raman data than PRV. It should also be noted that the results are highly dependent on the pre-processing of the data, making any general conclusions difficult.

## 4 | DISCUSSION AND CONCLUSION

In applications where sparseness and interpretations are important, it may be useful to consider PVS and PRV as alternatives to PCA and PLSR. Similar to PCA/PLSR, PVS/PRV also provides directions spanning large amounts of  $\mathbf{X}$ - and/or  $\mathbf{y}$ -variance, and associated score plots, with the obvious advantage of addressing pairs of selected variables rather than pairs of more complex linear combinations including all the available variables. Both PVS and PRV are based on intuitive choices of optimisation criteria, and by using a greedy approach to the variable selection, both methods are efficient to implement for practical applications. These methods should also be considered as useful for data exploration independent of the variables included in some final resulting model. The sparser models may also make the requirements for future data collection simpler/cheaper. As long as the required selections of variables in number are similar to the number of components selected by PCR or PLS, this can also contribute to making issues regarding model interpretations correspondingly simpler.



For regression problems, we consider both CovSel and PRV as interesting alternatives, where (prior) domain knowledge may be helpful in deciding the choice of method that is more likely to provide the better variable selection. For spectroscopic data of the types described in Section 3, the choice of method may also depend on the choice of pre-processing and truncation (if any) of the measured spectra. For data with local regions containing 'high' noise, such as the beer data, applications of PRV seem to work well without any truncation of the spectral region, in contrast to CovSel where removal of the noisy parts of the spectra seems to be a requirement. For data sets similar to the Raman data examples where the local presence of noise seems to be low, but adds up to a significant amount of total variance, the choice of method is more likely to be the opposite. In the case of spectroscopic data with a presence of wavelength shift issues, it may be safer to avoid variable selection, and in this case, non-sparse methods such as PCR and PLS may give more robust models. Further, for non-sparse methods using linear combinations of all variables, there is typically some noise cancelling of normally distributed noise. This is lost with variable selection methods. At the same time, as PVS and PRV select variables maximising the total explained variance in the data, these methods should have some robustness to this type of noise in the data.

It should be noted that the PRV criterion in Equation (15) gives the same weight to the  $\mathbf{X}$ - and  $\mathbf{y}$ -variance. In some situations, it may be useful to introduce some alternative weighing of the two variance terms, in a fashion similar to the Powered PLS.<sup>46</sup> This can be done by considering powers of the two voting terms and replacing the criterion of (15) by a criterion of the form  $v_{\mathbf{X}}(\mathbf{q})^{\frac{1-\alpha}{\alpha}} \cdot v_{\mathbf{y}}(\mathbf{q})^{\frac{\alpha}{1-\alpha}}$ . Within this formulation, the choice  $\alpha = \frac{1}{2}$  corresponds to the PRV criterion, and for other choices of  $\alpha$ , we may obtain more emphasis on one of the two variance terms in the associated voting function alternative.

Although classification problems have not been considered in this paper, both PVS and PRV can also be applied to such problems in the same fashion as described for the CovSel<sup>32</sup> method by first using PVS or PRV for the variable selection and then using the selected variables as inputs for the chosen classification method.

Several modifications and extensions of PRV are obviously possible. Although we have suggested using all PCs/singular values in the voting functions, it is straightforward to eliminate certain (undesired) PCs from a voting function if it is clear from domain knowledge that the variances of some PCs for some reason are irrelevant or misleading.

## PEER REVIEW

The peer review history for this article is available at <https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/cem.3510>.

## DATA AVAILABILITY STATEMENT

All data used in the manuscript are from previously published articles and available from the corresponding author upon request.

## ORCID

Joakim Skogholt  <https://orcid.org/0000-0001-8511-993X>

Kristian H. Liland  <https://orcid.org/0000-0001-6468-9423>

Tormod Næs  <https://orcid.org/0000-0001-5610-3955>

Age K. Smilde  <https://orcid.org/0000-0002-3052-4644>

Ulf G. Indahl  <https://orcid.org/0000-0002-3236-463X>

## REFERENCES

1. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. *London, Edinburgh, Dublin Philosoph Mag J Sci.* 1901; 2(11):559-572.
2. Hotelling H. Analysis of a complex of statistical variables into principal components. *J Educ Psychol.* 1933;24(6):417.
3. Jolliffe IT. *Principal Component Analysis.* 2nd ed. Springer-Verlag; 2002.
4. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. *Philosoph Trans R Soc A: Math Phys Eng Sci.* 2016;374(2065):20150202.
5. Wold S, Ruhe A, Wold H, Dunn III WJ. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM J Scientif Stat Comput.* 1984;5:735-743.
6. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemometr Intell Lab Syst.* 2001;58:109-130.
7. Björck Å, Indahl UG. Fast and stable partial least squares modelling: A benchmark study with theoretical comments. *J Chemometr.* 2017;31:e2898.

8. Indahl UG, Liland KH, Næs T. Canonical partial least squares—a unified PLS approach to classification and regression problems. *J Chemometr: J Chemometr Soc.* 2009;23(9):495-504.
9. Eckart C, Young G. The approximation of one matrix by another of lower rank. *Psychometrika.* 1936;1(3):211-218.
10. McCabe GP. Principal variables. *Technometrics.* 1984;26(2):137-144.
11. Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. *J Comput Graph Stat.* 2006;15(2):265-286.
12. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc: Ser B (Methodol).* 1996;58(1):267-288.
13. Camacho J, Smilde AK, Saccenti E, Westerhuis JA. All sparse pca models are wrong, but some are useful. Part I: computation of scores, residuals and explained variance. *Chemometr Intell Lab Syst.* 2020;196:103907.
14. Trendafilov NT. From simple structure to sparse components: a review. *Comput Stat.* 2014;29(3):431-454.
15. Zou H, Xue L. A selective overview of sparse principal component analysis. *Proc IEEE.* 2018;106(8):1311-1320.
16. Van Deun K, Thorrez L, Coccia M, et al. Weighted sparse principal component analysis. *Chemometr Intell Lab Syst.* 2019;195:103875.
17. Krzanowski WJ. Selection of variables to preserve multivariate data structure, using principal components. *J R Stat Soc: Ser C (Appl Stat).* 1987;36(1):22-33.
18. Guo Q, Wu W, Massart DL, Boucon C, De Jong S. Feature selection in principal component analysis of analytical data. *Chemometr Intell Lab Syst.* 2002;61(1-2):123-132.
19. Wang A, Gehan EA. Gene selection for microarray data analysis using principal component analysis. *Stat Med.* 2005;24(13):2069-2087.
20. Boutsidis C, Mahoney MW, Drineas P. An improved approximation algorithm for the column subset selection problem. In: Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms. SIAM; 2009:968-977.
21. Hooda BK, Mishra K, Singh KP. A procedure for identification of principal variables by least generalized dependence. *Commun Stat-Simul Comput.* 2007;37(1):167-177.
22. Enki DG, Trendafilov NT, Jolliffe IT. A clustering approach to interpretable principal components. *J Appl Stat.* 2013;40(3):583-599.
23. Lu Y, Cohen I, Zhou XS, Tian Q. Feature selection using principal feature analysis. In: Proceedings of the 15th ACM International Conference on Multimedia. Association for Computing Machinery; 2007:301-304.
24. Al-Kandari NM, Jolliffe IT. Variable selection and interpretation of covariance principal components. *Commun Stat-Simul Comput.* 2001;30(2):339-354.
25. Cumming JA, Wooff DA. Dimension reduction via principal variables. *Comput Stat Data Anal.* 2007;52(1):550-565.
26. Anzanello MJ, Fogliatto FS. A review of recent variable selection methods in industrial and chemometrics applications. *Europ J Industr Eng.* 2014;8(5):619-645.
27. Øvergaard SI, Pierna JAF, Baeten V, Dardenne P, Isaksson T. Prediction error improvements using variable selection on small calibration sets—a comparison of some recent methods. *J Near Infrared Spectroscopy.* 2012;20(3):329-337.
28. 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.
29. McLeod G, Clelland K, Tapp H, et al. A comparison of variate pre-selection methods for use in partial least squares regression: a case study on nir spectroscopy applied to monitoring beer fermentation. *J Food Eng.* 2009;90(2):300-307.
30. Cadima J, Cerdeira JO, Minhoto M. Computational aspects of algorithms for variable selection in the context of principal components. *Comput Stat Data Anal.* 2004;47(2):225-236.
31. Höskuldsson A. The H-principle: new ideas, algorithms and methods in applied mathematics and statistics. *Chemometr Intell Lab Syst.* 1994;23(1):1-28.
32. Roger J-M, Palagos B, Bertrand D, Fernandez-Ahumada E. CovSel: Variable selection for highly multivariate and multi-response calibration: application to IR spectroscopy. *Chemometr Intell Lab Syst.* 2011;106(2):216-223.
33. Trefethen LN, Bau D. *Numerical Linear Algebra.* SIAM; 1997.
34. Björck Å. *Numerical Methods in Matrix Computations.* Springer; 2015.
35. Camacho J, Smilde AK, Saccenti E, Westerhuis JA, Bro R. All sparse PCA models are wrong, but some are useful. Part II: limitations and problems of deflation. *Chemometr Intell Lab Syst.* 2021;208:104212.
36. Cadima J, Jolliffe IT. Variable selection and the interpretation of principal subspaces. *J Agricult Biolog Environm Stat.* 2001;6(1):62-79.
37. Wei H-L, Billings SA. Feature subset selection and ranking for data dimensionality reduction. *IEEE Trans Pattern Anal Machine Intell.* 2006;29(1):162-166.
38. Whitley DC, Ford MG, Livingstone DJ. Unsupervised forward selection: a method for eliminating redundant variables. *J Chem Inform Comput Sci.* 2000;40(5):1160-1168.
39. Puggini L, McLoone S. Forward selection component analysis: algorithms and applications. *IEEE Trans Pattern Anal Machine Intell.* 2017;39(12):2395-2408.
40. Billings SA, Wei H-L. A Multiple Sequential Orthogonal Least Squares Algorithm for Feature Ranking and Subset Selection. ACSE Research Report no. 908, Automatic Control and Systems Engineering, University of Sheffield; 2005.
41. Nørgaard L, Saudland A, Wagner J, Nielsen JP, Munck L, Engelsen SB. Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy. *Appl Spectroscopy.* 2000;54(3):413-419.
42. Andersen CM, Bro R. Variable selection in regression—a tutorial. *J Chemometr.* 2010;24(11-12):728-737.
43. Olsen EF, Rukke E-O, Flåtten A, Isaksson T. Quantitative determination of saturated-, monounsaturated- and polyunsaturated fatty acids in pork adipose tissue with non-destructive Raman spectroscopy. *Meat Sci.* 2007;76(4):628-634.

44. Liland KH, Kohler A, Afseth NK. Model-based pre-processing in Raman spectroscopy of biological samples. *J Raman Spectrosc.* 2016; 47(6):643-650.
45. Martens H, Stark E. Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy. *J Pharmaceut Biomed Anal.* 1991;9(8):625-635.
46. Indahl UG. A twist to partial least squares regression. *J Chemometr: J Chemometr Soc.* 2005;19(1):32-44.

**How to cite this article:** Skogholt J, Liland KH, Næs T, Smilde AK, Indahl UG. Selection of principal variables through a modified Gram–Schmidt process with and without supervision. *Journal of Chemometrics.* 2023;e3510. doi:[10.1002/cem.3510](https://doi.org/10.1002/cem.3510)

## APPENDIX A: MATLAB IMPLEMENTATIONS

```

1 function [inds, ssTX, ssEX, Q, R, P, C] = PVS(X, A)
2
3 [n,p] = size(X);
4 X = bsxfun(@minus, X, mean(X,1));
5 [U,s, ] = svd(X, 'econ');
6 s = diag(s);
7 inds = zeros(A,1);
8 candInds = 1:p;
9 ssTX = sum(s.^2);
10 ssEX = zeros(A,1);
11 T = U.*s;
12 P = zeros(p,A);
13 Q = zeros(n,A);
14
15 for i=1:A
16     votes = zeros(p,1);
17
18     for j=candInds
19         qj = X(:,j) / norm(X(:,j));
20         if abs(norm(qj)) > 1e-14
21             votes(j) = sum((qj * T).^2);
22         end
23     end
24
25     [ , inds(i)] = max(votes);
26     candInds = setdiff(candInds,inds(i));
27     qi = X(:,inds(i)) / norm(X(:,inds(i)));
28     Q(:,i) = qi;
29     P(:,i) = X * qi;
30     ssEX(i) = 100 * sum((qi * T).^2) / ssTX;
31     X = X - qi*(qi * X);
32 end
33
34 W = eye(p);
35 W(:,1:A) = W(:,inds);
36 R = P * W;
37 R = R(:,1:A);
38 C = R \ P;

```

```

1 function [betas, ids, Q, R, vperm, U, s, ssEX, ssEY, ni, P, C] = PRV(X, Y, A, B)
2
3 mX = mean(X); X = bsxfun(@minus,X,mX);
4 mY = mean(Y); Y = bsxfun(@minus,Y,mY);
5 [U, S, ] = svd(X, 'econ');
6 s = diag(S);
7 Y0 = Y;
8 if nargin == 2
9     A = 2;
10 else A = min(A, length(s));
11 end
12 eps = 1e-10;
13 [n, p] = size(X);
14 ids = NaN(A,1);
15 Q = zeros(n,A);
16 P = zeros(p,A);
17 ssEX = zeros(A,1);
18 ssEY = zeros(A,1);
19 ni = zeros(A,1);
20 vAll = zeros(1,p);
21 if nargin == 4
22     B = max(B,2);
23 elseif nargin < 4
24     B = length(s);
25 end
26
27 T = U(:,1:B).*s(1:B);
28 c = zeros(1,A);
29 ssTX = sum(s.^2);
30 ssTy = sum(Y(:).^2);
31 idn = 1:p;
32
33 for i=1:A
34     Xidn = X(:,idn);
35     sX2 = sum(Xidn.^2);
36     t = sum((T * Xidn).^2) ./ sX2;
37     vAll(idn) = prod((Y * Xidn).^2) ./ sX2,1) .* t;
38     [ , si] = max(vAll);
39     ni(i) = norm(X(:,si));
40     if ni(i) > eps
41         ids(i) = si;
42         qa = X(:,si) ./ ni(i);
43         P(idn,i) = Xidn * qa;
44         ssEX(i) = 100*sum((qa * T).^2) / ssTX;
45         ssEY(i) = 100*sum((qa * Y).^2) / ssTy;
46         X(:,idn) = Xidn - qa * (qa * Xidn);
47         c(i) = qa * Y;
48         Y = Y - qa * c(i);
49         Q(:,i) = qa;
50     end
51     idn = setdiff(idn,si);
52     vAll(si)=0;
53 end
54

```

```
55 W          = eye(p);
56 W(:,1:A) = W(:,ids);
57 R          = P * W;
58 R          = R(:,1:A);
59 C          = R \ P;
60
61 vperm = [ids; setdiff((1:p) , ids)];
62 betas = cumsum((eye(A)/(P(ids,:)) ) .* c, 2);
63 betas = [mY - mX(ids)*betas; betas];
```