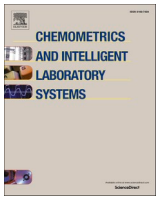
Chemometrics and Intelligent Laboratory Systems 254 (2024) 105223

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics



G-CovSel: Covariance oriented variable clustering

Jean-Michel Roger [a,b,*], Alessandra Biancolillo [c], Bénédicte Favreau [d], Federico Marini [e]

[a] ITAP-INRAE, Institut Agro, University Montpellier, 34196, Montpellier, France
[b] ChemHouse Research Group, 34196, Montpellier, France
[c] Department of Physical and Chemical Sciences, University of L'Aquila, 67100, Coppito, L'Aquila, Italy
[d] CIRAD, UMR AGAP, Montpellier, France
[e] Department of Chemistry, University of Rome "La Sapienza", 00185, Rome, Italy

ARTICLE INFO

Keywords:
Variable selection
Multivariate regression
Multivariate discrimination
Variable clustering
Biomarker identification

ABSTRACT

Dimensionality reduction is an essential step in the processing of analytical chemistry data. When this reduction is carried out by variable selection, it can enable the identification of biochemical pathways. CovSel has been developed to meet this requirement, through a parsimonious selection of non-redundant variables. This article presents the g-CovSel method, which modifies the CovSel algorithm to produce highly complementary groups containing highly correlated variables. This modification requires the theoretical definition of the groups' construction and of the deflation of the data with respect to the selected groups. Two applications, on two extreme case studies, are presented. The first, based on near-infrared spectra related to four chemicals, demonstrates the relevance of the selected groups and the method's ability to handle highly correlated variables. The second, based on genomic data, demonstrates the method's ability to handle very highly multivariate data. Most of the groups formed can be interpreted from a functional point of view, making g-CovSel a tool of choice for biomarker identification in omics. Further work will be carried out to generalize g-CovSel to multi-block and multi-way data.

1. Introduction

Analytical chemistry devices provide large quantities of variables, more or less directly linked to the phenomena of interest. Some measurement methods are highly indirect and provide highly correlated variables, such as optical spectroscopy Ultra Violet (UV), Visible, Near Infrared (NIR), Middle Infrared (MIR) [1] or liquid chromatography (LC), gas chromatography (GC) [2]. Others are more direct and provide less correlated variables, such as mass spectrometry (MS) [3]. What all these techniques have in common is that they produce highly multivariate data. Each sample measured is represented by a vector in a very high-dimensional vector space, which poses specific mathematical problems [4]. Dimensionality reduction is therefore a necessary step in any classification or regression operation [5]. A range of specialized multivariate analysis techniques, known as chemometrics, has been developed over the past few decades [6,7]. One of the strengths of chemometrics is that it provides knowledge about the system under study. In addition to validating the models produced, this feedback makes it possible to identify the factors responsible for the phenomena under study, such as biomarkers [7].

Chemometric analysis of data involves a phase of dimensionality reduction, allowing the practitioner to focus on the useful part of the measured signal, which is modeled in chemometrics as a vector subspace related to the factors of interest. Factorial methods such as Principal Component Analysis (PCA) [8] or Partial Least Squares (PLS) [9] yield loadings/weights which are a basis of the useful subspace. These coefficients are then analyzed to trace the factors behind the variations under study. Even more explicitly, Multivariate Curve Resolution (MCR) can be used to recover the pure spectra of the constituent elements of a mixture, enabling them to be identified directly [10].

Another way of reducing the size of the data is to select variables. Although in chemometrics, variable selection is often used as a preprocessing step to improve model performance [11], it is also very interesting as an analysis technique on its own, with the aim of identifying the physico-chemical phenomena related to the factors studied [12]. Machine learning community offers a huge choice of variable selection methods [13]. Those based on both predictors and responses are of greater interest for chemometrics [14]. Variable Importance in Projection (VIP) [15] has been developed to quantify the importance of each variable in a regression or classification PLS model. Applying a

* Corresponding author. ITAP-INRAE, Institut Agro, University Montpellier, 34196, Montpellier, France.
E-mail address: jean-michel.roger@inrae.fr (J.-M. Roger).

https://doi.org/10.1016/j.chemolab.2024.105223
Received 26 March 2024; Received in revised form 26 August 2024; Accepted 28 August 2024
Available online 29 August 2024

threshold (usually 1) to the calculated VIPs enables the most important variables to be selected. Interval PLS (i-PLS) [16] proposes to subdivide the entire spectral range into intervals, then PLS is applied individually to each segment. This identifies the most informative regions for prediction. Sparse PLS (SPLS) refers to a family of PLS algorithms that produce loadings containing a high proportion of null values. Several algorithms exist, all based on minimizing the L_1 norm (i.e., the sum of the absolute values) of the weights applied to the predictors and responses, as for example in Monteiro et al. [17]. Group Partial Least Squares (G-PLS) regression [18] was developed as a special case of S-PLS, which aims to respect predefined groups of variables. These groups are constructed by examining the correlation maps derived from the data to be analyzed. VIP, i-PLS and SPLS allow selecting the most useful variables to predict the responses of interest. However, as this selection is global, it is difficult to identify the different phenomena causing the link between predictors and responses. CovSel has been developed to answer this requirement, performing parsimonious selection of non-redundant variables in a predictive context [19]. As implied by its name, the relevance of each variable is assessed by estimating the covariance between each predictor and the responses: the variable with the highest covariance is identified and selected first; subsequently, all other predictors and the responses are orthogonalized with respect to this selected variable and the process is repeated until the predetermined number of variables has been selected. This strategy has also been extended to multi-block and multi-way data (SO-CovSel [20] and N-CovSel [21], respectively). Thanks to the deflation performed by CovSel, the selected variables are intended to be linked to different sources of covariance between the measurements and the phenomenon under study. CovSel is therefore a tool of choice to identify the factors responsible for the phenomena under study.

However, in some cases, identifying one variable per factor is not enough. This is the case in MS, for example, where a molecule is represented by a set of peaks linked to fragments or isotopes. In this case, it is useful to have a tool that identifies groups of variables, so that each group is linked to a molecule, or a family of molecules. Clustering of Variables Around Latent Components (CAVALC) [22] is a method for identifying groups of variables, so that each of these clusters is associated to a single latent variable. The latent variable associated to the cluster may be only explicative of the variance in the grouped predictors (i.e., a principal component) or account for the covariance of the predictors with external data or responses. However, since no deflation/orthogonalization steps are involved, there is no guarantee that the identified group of variables are associated to different/unrelated sources of (co-)variance.

This article presents an extension of CovSel, called g-CovSel, designed to perform a selection of groups of variables, so that each group is linked to a phenomenon explaining a set of responses. The first part describes the theoretical aspects of the method. The second part presents the datasets used to illustrate how it works. The third section discusses the properties of g-CovSel, based on the results obtained. The article ends with a conclusion and research prospects.

2. Theory

Let $\mathbf{X}$ be a matrix (N, P) of N individuals described by P descriptors (predictor variables); let $\mathbf{Y}$ be a matrix (N, Q) of the same N individuals described by Q responses (dependent variables). The aim of CovSel [19] is to select the variables in $\mathbf{X}$ that best explain the variables in $\mathbf{Y}$. CovSel uses the following iteration:

1. Define the number L of variables to be selected.
2. Select the variable of $\mathbf{X}$ with the highest squared covariance with $\mathbf{Y}$.
3. Deflate $\mathbf{X}$ and $\mathbf{Y}$ of the information present in the selected variable.
4. Continue from step 2 until the value defined in step 1 is reached.

g-CovSel is based on the same four steps, but at step 2, a group of variables is selected, in place of one variable. Step 3 is also changed, in order to deflate the data of the group of variables. These two steps are detailed in the following.

2.1. Building the groups

The building of a group of variables G begins with the selection of a single variable I, based on the same principle as CovSel, i.e., as the one which satisfies $I = \underset{i}{\operatorname{argmax}}\left(\mathbf{x}_i'\mathbf{Y}\mathbf{Y}'\mathbf{x}_i\right)$, where $\mathbf{x}_i$ is the ith column of $\mathbf{X}$. This variable will serve as the seed for the identification of G and will thus be called the *seed*. This condition ensures that g-CovSel will act as an extension of CovSel and so that CovSel is a particular case of g-CovSel. The variables members of G should explain $\mathbf{Y}$ and should be related to the seed.

Let define two functions $CR(.)$ and $CV(.)$- where $(.)$ is used as a placeholder for the function argument - by:

$$CR(\mathbf{x}_k) = R^2(\mathbf{x}_I, \mathbf{x}_k) \tag{1}$$

$$CV(\mathbf{x}_k) = \frac{\mathbf{x}_k'\mathbf{Y}\mathbf{Y}'\mathbf{x}_k}{\mathbf{x}_I'\mathbf{Y}\mathbf{Y}'\mathbf{x}_I} \tag{2}$$

Where R^2 is the squared correlation coefficient. We have: $CR(\mathbf{x}_k) \in [0,1]$, $CV(\mathbf{x}_k) \in [0,1]$, $CR(\mathbf{x}_I) = 1$ and $CV(\mathbf{x}_I) = 1$. Using $CR(\mathbf{x}_k)$ as abscissa and $CV(\mathbf{x}_k)$ as ordinate, all the variables of $\mathbf{X}$ can be represented as points in the square $[0,1] \times [0,1]$. The seed is located at the upper right corner of the square; variables in the upper part of the square are the most linked to $\mathbf{Y}$ responses; variables falling in the rightmost region of the square are the most correlated with the seed.

Thus, the group sought should reasonably contain variables with high values of both the abscissa and the ordinate, i.e., predictors falling close to the upper right corner. Based on this principle, two group formation algorithms are proposed hereafter:

- Double threshold-based grouping (DT):

Two separate thresholds τ_R and τ_V are identified for $CR(\boldsymbol{x}_k)$ and $CV(\boldsymbol{x}_k)$, respectively, so that

$$k \in G \text{ if } CR(\boldsymbol{x}_k) > \tau_R \text{ and } CV(\boldsymbol{x}_k) > \tau_V \; ; \; \tau_R \in [0,1] \; ; \; \tau_V \in [0,1]$$

- RV based grouping (RV):

All the variables k of $\mathbf{X}$ are sorted according to their increasing squared distance to the seed, in the space defined by $CR(\boldsymbol{x}_k)$ and $CV(\boldsymbol{x}_k)$, i.e. according to the increasing value of $(1 - CR(\boldsymbol{x}_k))^2 + (1 - CV(\boldsymbol{x}_k))^2$. Then, variables are progressively included in G, following the order given by the sorting, and the RV coefficient between $\mathbf{X}_G = \left[\mathbf{x}_I, \mathbf{x}_{G,1}, \mathbf{x}_{G,2}, \cdots, \mathbf{x}_{G,k}\right]$ and $\mathbf{Y}$ is calculated:

$$RV(\mathbf{X}_G, \mathbf{Y}) = \frac{trace\left(\mathbf{X}_G\mathbf{X}_G'\mathbf{Y}\mathbf{Y}'\right)}{\sqrt{trace\left(\mathbf{X}_G\mathbf{X}_G'\mathbf{X}_G\mathbf{X}_G'\right)trace\left(\mathbf{Y}\mathbf{Y}'\mathbf{Y}\mathbf{Y}'\right)}} \tag{3}$$

Variables are added to G as long as $RV(\mathbf{X}_G, \mathbf{Y})$ increases.

The DT grouping is very simple but requires that the user have a good idea of the correlations between the $\mathbf{X}$ variables. So for near-infrared spectral data, τ_R can certainly take values very close to 1, such as 0.95 or more, but for less correlated data, as mass spectrometry abundances, the value of τ_R is likely to be much lower. The RV grouping is more complicated to calculate, but it adapts more easily to complex situations, where groups of variables need to be defined according to their overall relevance to the responses. Moreover, RV-based selection method can be run automatically, i.e. without the need of any parameter tuning.

2.2. Deflating the data

The role of deflation done before each new selection is to remove from **X** and **Y** the variance carried by the selected variables. This offers a certain guarantee that the selected groups will be as complementary as possible. It also enables to decompose the total variance of **X** and **Y**, so that it is possible to examine the evolution of the variance explained as a function of the number of groups selected. In factorial methods such as PCA and PLS, and also in CovSel, deflation of **X** (and/or **Y**) consists in projecting **X** (and/or **Y**) orthogonally to the vector of scores calculated in the current step. Because the score vector is one-dimensional, this deflation consumes one dimension in $Col(\mathbf{X})$. If we were to proceed in the same way for g-CovSel, i.e. project orthogonally to all the variables in the selected group, there would be the risk of drastically reducing the rank of $Col(\mathbf{X})$. In chemometrics applications, the total rank of $Col(\mathbf{X})$ can be limited, because there often is less individuals than variables. The risk would be to rapidly erode the entire variance of **X**. It is therefore necessary to perform a rank 1 deflation. To do this, a score vector **t** is defined as follows:

- Let G_K be the set of indexes of the group, and I_K be the index of the seed at step K.
- Let $w_i = sign\left(\mathbf{x}_i^T \mathbf{x}_{I_K}\right)\sqrt{CV_i \times CR_i}$ if $i \in G_K$ and $w_i = 0$ otherwise
  - $\mathbf{t} = \mathrm{Xw}$.
  - $\mathbf{t} = \frac{\mathbf{t}}{\|\mathbf{t}\|}$.

Then, the deflation is done as usually by:

- $\mathbf{X} = \mathbf{X} - \mathbf{t}\,\mathbf{t}^T\,\mathbf{X}$.
- $\mathbf{Y} = \mathbf{Y} - \mathbf{t}^T\mathbf{t}\,\mathbf{Y}$.

3. Material and methods

A first test of g-CovSel was carried out on the well-known corn dataset that usually serves as a reference for calibration transfer, and which was initially provided by Mike Blackburn of Cargill. This dataset can be accessed via Eigenvector Research at https://eigenvector.com/resources/data-sets/. It includes reference values for protein, starch, moisture and oil, which were evaluated on $N_C = 80$ separate corn samples using three NIR instruments (1100–2498 nm, $P_C = 700$ wavelengths). Spectra from the M5 instrument were retained for testing g-CovSel. To enhance the underlying peaks, second derivative of the spectra was calculated using a Savitsky and Golay filter [23] (width = 51, polynomial degree = 3, order of derivation = 2). The corn dataset therefore consisted of $\mathbf{X}_C(80, 700)$, $\mathbf{Y}_C(80, 4)$. Matrix $\mathbf{X}_C$ was column mean centred and matrix $\mathbf{Y}_C$ was column autoscaled.

A second g-CovSel test was carried out on genomic data measured on *Eucalyptus grandis* trees. These data contained 24 individuals, divided into 4 replicates of 6 modalities. These modalities resulted from crossing 3 nutrition modalities C (control), Na (sodium addition), K (potassium addition) with 2 levels of water stress FR (full rainfall) and RR (restricted rainfall). The experimental design was balanced, so that each of the 6 classes K + FR, Na + FR, C + FR, K + RR, Na + RR, C + RR contained 4 individuals. Each individual was described by the expression of $P_E = 4890$ genes. A complete description of this dataset can be found in Ref. [24]. The two classes K + FR and K + RR were used to test g-CovSel. The eucalyptus dataset therefore consisted of $\mathbf{X}_E(8, 4890)$, $\mathbf{Y}_E(8, 2)$, where $\mathbf{Y}_E$ contained the membership degrees of the individuals in the 2 classes. A logarithmic transformation was applied to $\mathbf{X}_E$. Then both matrices $\mathbf{X}_E$ and $\mathbf{Y}_E$ were column mean centred.

The two datasets were then processed according to the following workflows. First, a CovSel selection was performed to get an initial idea of the number of factors underlying the datasets, as well as a selection of unique variables. The g-CovSel method was then applied, using the DT method on the corn dataset and both DT and RV methods on the eucalyptus dataset. The program used was developed in Matlab R2015b (The Mathworks, Natick, MA, USA). Matlab source code is available for download at https://forgemia.inra.fr/chemhouse/octave/g_covsel.

4. Results and discussion

The functioning and properties of g-CovSel are presented and discussed in this section, based on the application of the two datasets presented above.

4.1. Results of CovSel on the corn and eucalyptus datasets

CovSel was first applied to both corn and eucalyptus datasets, requesting the selection of 10 variables. Fig. 1a and b show the selected variables on the x-axis, and the evolution of explained variances on the y-axis. It can be seen that the two datasets behave very differently. For corn data (Fig. 1-a), the proportion of variance explained for the **X** block is significantly higher than for the **Y** block. In contrast, for the eucalyptus data (Fig. 1-b), the proportion of variance of **Y** is greater than that of **X**. It's well known that the variance of NIR spectra is dominated by multiplicative and additive effects, so that the first component of a PCA generally accounts for over 80 % of the variance. This is the phenomenon we find here, confirmed by the fact that the first variable selected corresponds to the extreme left of the spectra. In fact, Fig. 2 clearly shows that this part of the spectra has a high variance. On the contrary, as the $\mathbf{X}_E$ variables have already been pre-sorted, there is little variance unrelated to $\mathbf{Y}_E$ in the selected variables. In view of these graphs, it was decided to set the number of groups to be searched at 8 for corn dataset and at 3 for eucalyptus dataset.

Fig. 2 shows the spectra of the corn dataset, with the first eight selected wavelengths indicated. We can see that all the selected wavelengths are related to physico-chemical factors, such as baseline changes, moisture, fat, starch and protein contents.

The first three genes selected by CovSel from the eucalyptus dataset are Eucgr.G03028, Eucgr.K03413 and Eucgr.K00207 (see the x-axis labels of the graph in Fig. 1 - b). The first gene corresponds to a protein of the Laccase/Diphenol oxidase family. These proteins are involved in the lignin biosynthesis pathway, which is affected by the level of water restriction on potassium-fertilized trees [24]. The second gene selected by CovSel corresponds to a heat stable protein with antimicrobial and antifungal activities. This gene is involved in the Eucalyptus response to sodium fertilization [24]. The third gene selected by CovSel is a MADS-BOX gene. MADS-BOX are key transcription factors involved in abiotic stress-related regulatory networks regulating development and growth [26].

5. Results of g-CovSel on the corn dataset

Fig. 3 shows the (CR, CV) maps built by g-CovSel during the eight first steps of the algorithm on the corn dataset. At each step of the algorithm, few variables are close to the seed, situated at the upper right-hand corner $(1,1)$, making it easy to identify groups. These variables are aligned on a quasi-linear and quasi-vertical structure, showing that they are highly correlated with the group seed, but that the covariance with **Y** decreases very rapidly when distancing from the seed. This is typical of near-infrared spectral data, for which a high CR threshold and a low CV threshold should be used. Fig. 4 shows the CV values as a function of the wavelengths during the eight first steps of the algorithm. Note that each curve has peaks and that the seed corresponds to the maximum of a peak. Note again that, thanks to deflation, peaks selected at a given step disappear from the CV curve at the next step, as do correlated peaks. This is the case, for example, between the sixth and seventh steps. The peak selected at the sixth step, centred at 1904 nm, disappears at the seventh step, as does the peak centred at 1400 nm, as both these peaks are in a zone linked to the OH bonds of water. Fig. 4 clearly shows the benefits of building groups on a dual criterion of correlation and

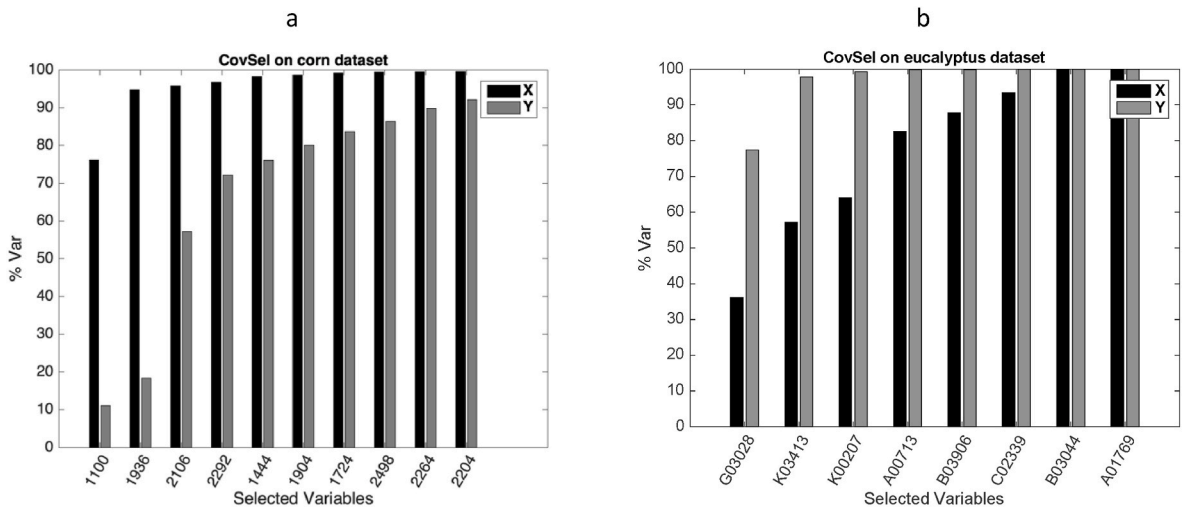


Fig. 1. Results of CovSel on corn dataset (a) and eucalyptus dataset (b). Abscissa indicates the selected variables; ordinate reports the percentage of variance explained. For eucalyptus data (b), the "Eucgr." prefix for each gene has been omitted for clarity.

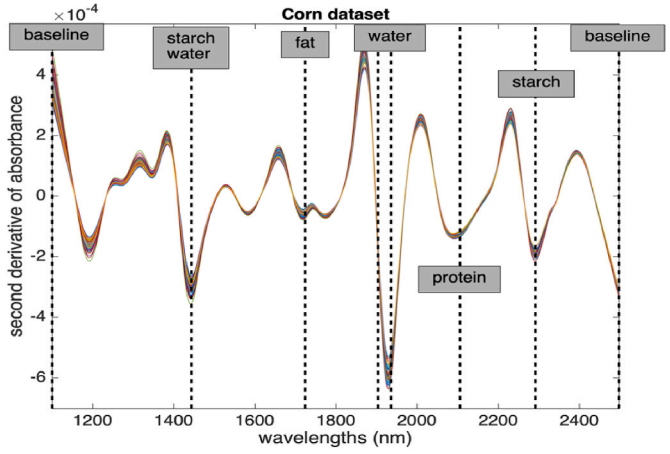


Fig. 2. Corn spectra. Vertical dashed lines show the CovSel selection. Text-boxes indicate physico-chemical factors linked to the variables, based on [25].

covariance. For example, in the fourth step, three peaks show covariances well above the 0.5 threshold: at 1904, 2264 and 2292 nm. Taking correlation into account enables g-CovSel to place the two peaks at 2264 and 2292 nm in the same group, which corresponds to a wavelength zone related to starch, proteins and fat [27], and to leave aside the peak at 1904 nm, corresponding to water [25] and which is selected in the sixth step.

While most of the selected groups contain a single peak, the second and fourth contain two. Let's take a look at the second group; it contains two peaks centred at 1876 and 1936 nm. These two *CV* peaks correspond to the positive and negative peaks shown in Fig. 2. As the spectra treated here are second derivatives of absorption spectra, the positive peak corresponds to the left foot of the absorption peak for OH bonds in water at 1876 nm, and the negative peak to the top of this peak at 1936 nm [25]. This group is therefore linked to the height of the water OH bond absorption peak, and therefore to moisture. The peak selected in the seventh group is centred on 1904 nm, which corresponds to a zone where the second derivative cancels out (see Fig. 2), i.e., the left inflection point of the water OH bond absorption peak. This group is therefore certainly linked to the width or position of this peak, and therefore to the state of water. This rapid analysis shows that the selected groups correspond to different underlying physico-chemical phenomena.

In the second step of Fig. 4, we can see that the wavelength 2250 nm has been selected to be part of the group made up of peaks at 1876 and 1936 nm. This is an artefact due to the threshold on *CR* or *CV* being too low. Indeed, in Fig. 3, second step, we can see that the selection is a little too wide and includes a point belonging to a structure other than the one in the top right-hand corner of the map. However, this error did not prevent the peaks at 2240 nm and 2292 nm from being correctly selected in the fourth step.

Finally, we note that the groups formed are completely disjoint. This shows that deflation has worked. In fact, the aim of deflation is to erase the vector subspace generated by the selected variables. This means that the variance carried by these variables becomes very low, and consequently the covariance with **Y**. It is therefore highly unlikely that variables will belong to more than one group. However, it is conceivable that, if the groups are unclear, the user might be obliged to use very low thresholds, which would make the deflation process very "soft", and open the door to this possibility. Further study will be carried out on this aspect, in order to explore the link with the concept of interactions between factors.

In [19], it is shown that CovSel is a special case of PLS. Similarly, since g-CovSel performs variable group selection by optimizing a covariance criterion between **X** and **Y**, it may be relevant to compare it with SPLS, considering that the non-null values of each sparse loading define a variable group. To this end, corn data were processed by SPLS, using the algorithm described in Ref. [17]. In this algorithm, sparsity is set with a penalty term, λ varying between 1 (maximum sparsity) and $\sqrt{P}$ (minimum sparsity), where P is the number of variables in **X**. A first run, with $\lambda = 1$, selected the 1100 nm wavelength for the first loading, which is entirely consistent with CovSel, since this variable had the maximum covariance with **Y**. Successive runs were then made, with an increasing value of λ, until the same group was obtained as the first one selected by g-CovSel, Fig. 4, step 1. The value of λ corresponding to this limit was $\lambda = 0.253$. Fig. 5-top shows the two groups formed by the first two SPLS latent variables, with $\lambda = 0.253$. The first block is consistent with that of g-CovSel, by construction. However, the second group is made up of the peak at 2104 nm, which corresponds to the third group selected by g-CovSel (see Fig. 4). The group consisting of the two peaks at 1876 and 1936 nm, selected in second position by g-CovSel (Fig. 4, step 2), has disappeared from the covariance curve of the second SPLS step (Fig. 5-top, step 2). This indicates that the deflation performed by SPLS at the first step was too abrupt and led to the disappearance of the information carried by the two peaks at 1876 and 1936 nm. This shows the value of the deflation proposed by g-CovSel, which allows the information carried by one group to be removed without altering the other

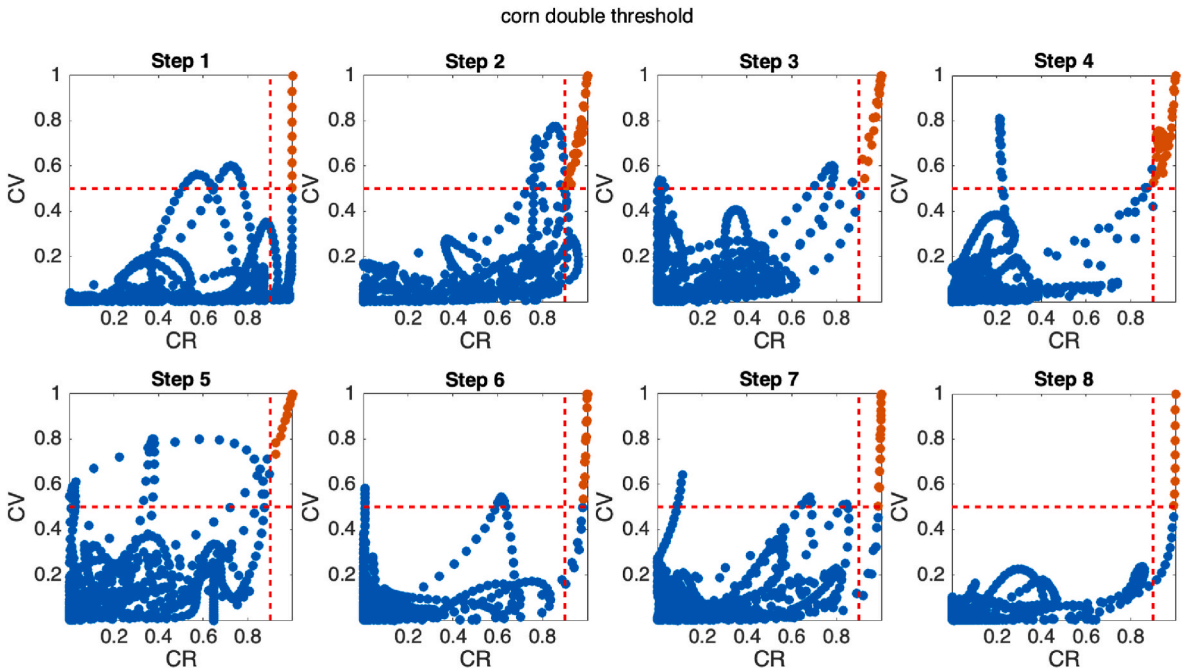


Fig. 3. (CR,CV) maps produced par g-CovSel on the corn dataset, along the 8 first steps of the algorithm. Dotted red lines indicate the thresholds used: 0.5 on CV and 0.9 on CR. Variables selected in the groups are shown in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

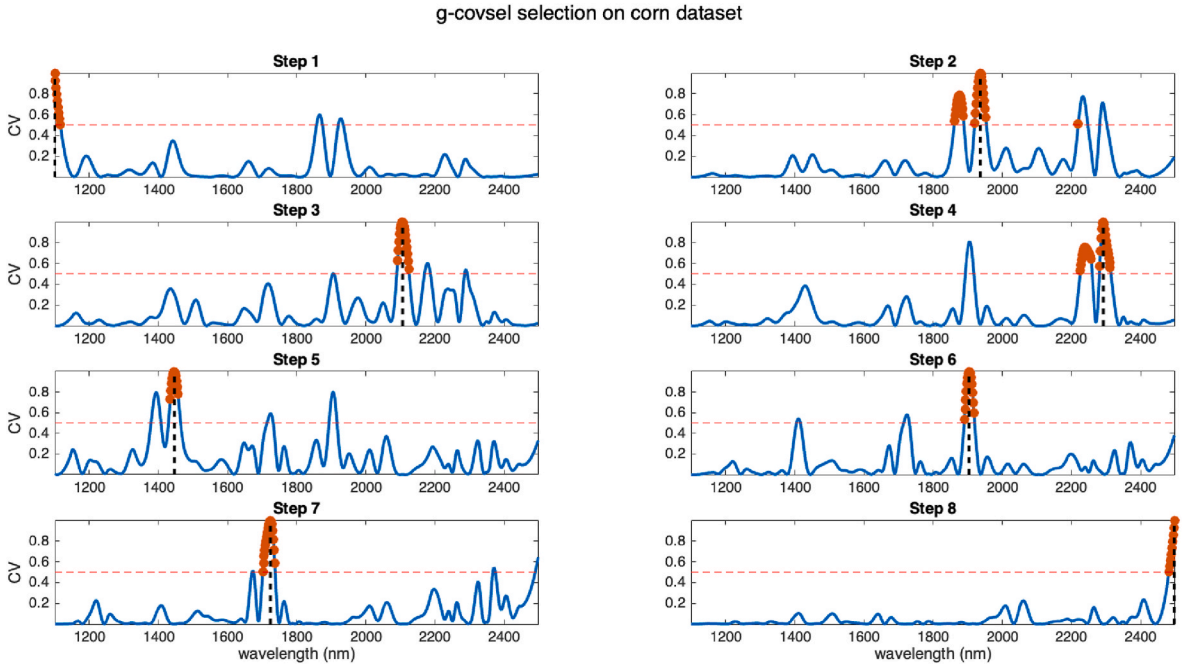


Fig. 4. CV spectra produced par g-CovSel on the corn dataset, along the 8 first steps of the algorithm. Dotted horizontal red lines indicate the threshold used on CV (0.5). Variables selected in the groups are shown in red. Dotted vertical black lines indicate the group seeds. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

groups, thanks to the joint consideration of covariance and correlation. Fig. 5-bottom shows the two groups formed by the first two SPLS latent variables, with $\lambda = 0.3$. The first group now contains also the two peaks at 1876 and 1936 nm. Referring to Fig. 3, step 1, we can see that these two peaks, although having a fairly high normalized covariance with ($CV > 0.5$), are only weakly correlated with the group seed ($CR < 0.75$). The SPLS selects them because it only considers covariance. Here we see the value of the g-CovSel algorithm, which simultaneously takes into account the covariance with **Y** and the correlation with the group seed.

In order to validate the relevance of the groups selected by g-CovSel, eight PLS2 models were built using variables from G_1,$\{G_1,G_2\}$,…,$\{G_1,\ldots,G_8\}$, to predict the four responses: moisture, fat, proteins and starch. In addition, a PLS2 model was built, using all the variables in **X**. A calibration set of 60 samples and a test set of 20 samples were constructed by performing a PCA on the four responses, then sorting the first vector of scores, and taking one sample out of four for the test, and the rest for

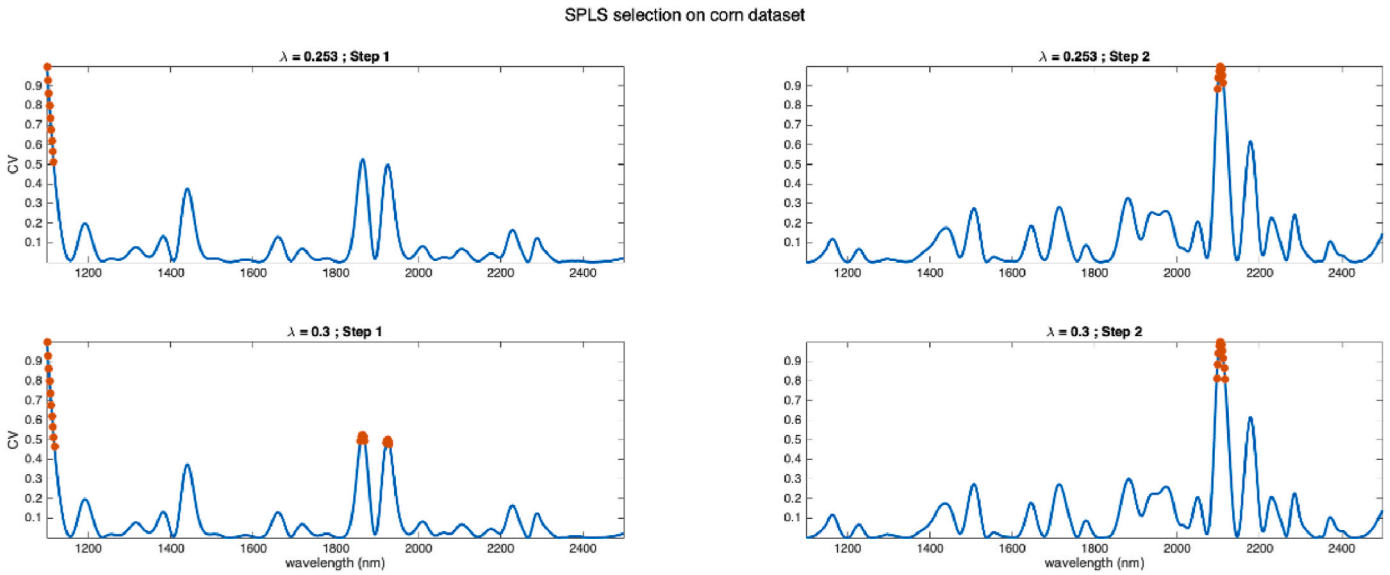


Fig. 5. CV spectra produced par SPLS on the corn dataset, along the 2 first latent variables, with two sparsity levels. Top: sparsity index at 0.253, giving the same first block as g-CovSel. Bottom: sparsity index at 0.3, giving a larger block.

calibration. All the models were submitted to a two-fold cross-validation repeated 30 times, using the same splits for all models. Fig. 6 reports the RMSECV curves of these models, for each response in **Y**. Firstly, it can be noted that, for all predicted responses, the RMSECV curve of the model using the eight groups is very close to that of the model using all **X** variables. This validates the overall relevance of the eight selected groups. Fig. 6-a shows that the performance of the PLS2 model in predicting moisture improves markedly when the G_2 and G_6 groups are added, in perfect agreement with the variables of these groups, which are linked to water (see Fig. 4). Fig. 6-b shows that fat prediction performance improves with the introduction of groups G_4 and G_7. Group G_4 is linked to the 2220–2320 nm spectral zone, which contains several absorption peaks related to amino acids, starch and fat [27]. Its variables are therefore particularly useful for building a PLS2 model predicting these compounds. Group G_7 is linked to the fat absorption peak at 1725 nm [27]. The same comments on G_4 can be drawn from Fig. 6-c, regarding the protein content prediction. As far as starch prediction is concerned (Fig. 6-d), it can be noted that almost all groups are required to achieve a good model.

The global RMSECVs of the PLS2 model using the variables of the eight groups and of the PLS2 model using all the variables were calculated as the quadratic mean of the four RMSECV. For both models, 13 latent variables was chosen as the optimal dimension. Table 1 summarizes the prediction performances of the two models on the test set samples. It can be noticed that the performances of the g-CovSel-PLS2 model are very close to the ones of the PLS2 model using all the variables, except for starch. In fact, the prediction of this response by NIRS is problematic, and relies not only on isolated peaks, but also on more global features of the spectrum, such as slopes or curvatures [28]. So it's not surprising that g-CovSel doesn't produce a very powerful model for this response.

5.1. Results of g-CovSel on the eucalyptus dataset

In this case study, g-CovSel was applied to the discrimination of the two classes K + FR and K + RR. The eucalyptus dataset therefore consisted of $\mathbf{X}_E(8,4890), \mathbf{Y}_E(8,2)$. Given the very limited number of samples, in this case the focus was on showing that, even with very ill-conditioned data sets, the method could anyway be able to produce reasonable and interpretable results, comparable with those produced in the original publication [24]. The first three groups of genes in relation to the two classes, i.e. in relation to the difference in water treatment in the presence of potassium-rich nutrition, were sought. Both DT ($t_{cov} = t_{cor} = 0.4$) and RV-based grouping algorithms were used. Table 2 lists the genes making up the groups. It can be seen that the two algorithms returned different numbers of genes in each group, but that for all groups, the smallest selection is always included in the largest selection. This shows good agreement on group construction between the two algorithms, but that they diverge on group size.

Group 1 is dominated by genes from the Purple network (19 out of 22). This network was identified in Ref. [24] as significant for the K + RR response. Note that the LightCyan network, which had also been identified as significant for this response, is completely absent from the g-CovSel selection. It is likely that this second network would have been selected with lower t_{cov} and t_{cor} thresholds, but that the deflation of the information carried by group 1 eliminated it. Only one Purple gene is found in another group. There is therefore a good match between group 1 and the Purple network, showing that g-CovSel is fairly consistent with the Weighted Gene Co-expression Network Analysis method (WGCNA) used in Ref. [24]. Gene D01509 was selected in third position in group 1, meaning that it is close to the group seed, and therefore very important for K + FR vs K + RR discrimination. On the other hand, it was assigned by WGCNA in a different network from that predominantly associated with the other genes in group 1. We note that it had a lower number of links to other genes (66), as detected by WGCNA, compared with the other genes in this group (over 200). All this makes this variable suspect. A detailed examination of the gene count shows that one sample among the eight was an outlier, with 193 counts compared to about 15 counts for the other individuals. If we replace the suspect value by the median of the values presented by the other individuals in the group, g-CovSel no longer selects this gene. The other selected genes remain unchanged, illustrating the robustness of g-CovSel. The Eucgr.K01002 gene shows also a disagreement between WGCNA and g-CovSel. It was selected by g-CovSel in sixth position by both DT and RV algorithms, whereas it was not placed in any network by WGCNA. A detailed examination of the data for this gene shows that it was over-expressed under water stress, and was significant for many of the factors studied in Ref. [24]. It is highly likely that WGCNA did not select it because it only uses correlation between genes, unlike g-CovSel, which also uses covariance with the response. The Eucgr.A02802 gene was placed in group 1 by g-CovSel, while WGCNA placed it in the green network, with a low number of counts (74). An examination of this variable for the eight individuals

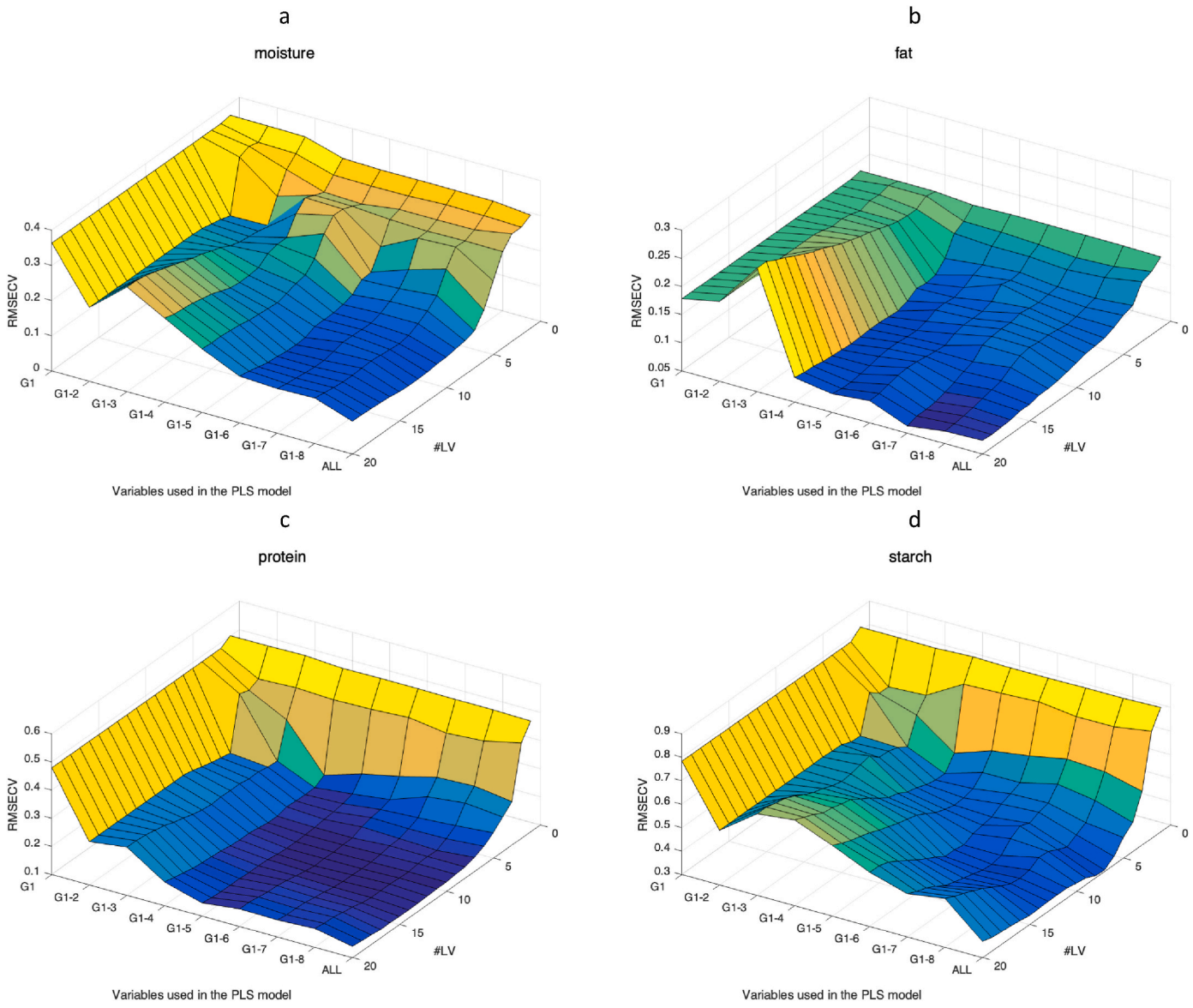


Fig. 6. Evolution of PLSR cross-validation error curves, as variable groups G1 through G8 are introduced into the model. The last curve (ALL) is for a model using all X variables.

Table 1
Prediction results of a PLS2 using all the X variables and of PLS2 using the variables of the eight groups defined by g-CovSel, for each response of the corn dataset. SEP is the standard deviation of the prediction error, Bias is the mean of the prediction error.

		moisture	fat	protein	starch
SEP	PLS2 with all variables	0.062	0.071	0.064	0.213
	PLS2 with the 8 groups	0.066	0.080	0.067	0.265
Bias	PLS2 with all variables	0.016	0.013	0.005	0.020
	PLS2 with the 8 groups	0.013	0.014	0.018	0.023

studied shows no outliers, but very low values (5 counts on average for K + FR and 24 for K + RR), compared with the average value for group 1 (around 71). This leads to greater uncertainty in the calculation of correlations by WGCNA and covariances by g-CovSel, which may explain the divergence in ranking. The function of this gene is unknown, but a research in Phyto Mine shows that it is linked to genes encoding heat shock proteins, involved in the plant's stress response.

Group 2 contains fewer genes than group 1. Six of the group's seven genes were selected by the DT algorithm. In other words, the RV algorithm selected only one gene (Eucgr.K00207, the group seed) and therefore failed to create a group around the seed. On the other hand, the genes selected in group 2 do not belong to any particular network. All genes of group 2 are linked to fertilization, whatever the stress. It is therefore normal that they belong to different WGCNA networks. Group 2 appears to express phenomena orthogonal to those linked to group 1 (fertilization vs. stress). Further study is required to determine whether this group has biological significance or whether the identified seed is isolated, as suggested by the RV algorithm.

Group 3 also contains seven genes. The DT algorithm selected only one gene (Eucgr.K00207, the group seed). This means that at this stage of the g-CovSel run, all the genes had very low values (CV, CR) (<0.4). This may be an indication of the limitations of the DT algorithm, which nevertheless performed satisfactorily for groups 1 and 2. The threshold values $t_{cov} = 0.4$ and $t_{cor} = 0.4$ chosen globally do not seem to be suitable for all steps of the g-CovSel algorithm. Note that the first four genes selected belong to the brown and blue networks, both linked to potassium.

Finally, we note that the majority of the genes selected have a very

Table 2
Groups selected by g-CovSel, either by DT or by RV algorithm. The Factor column indicates the significance for fertilization (F), water restriction (R), or the interaction of the two (F*R). The Network column indicates the gene network identified by the WGCNA method. The Degree column indicates the number of gene connections in this network.

N°	Selected variables		Gene ID	Factor	Network	Degree
1	Eucgr.G03028	(DT + RV)	Laccase/Diphenol oxidase family protein	F	Purple	234
	Eucgr.H03155	(DT + RV)	TPX2 (targeting protein for Xklp2) protein family	F*R	Purple	234
	Eucgr.D01509	(DT + RV)	HXXXD-type acyl-transferase family protein	F	Tan	66
	Eucgr.H03662	(DT + RV)	Major facilitator superfamily protein	F	Purple	212
	Eucgr.L00251	(DT + RV)	galactinol synthase 2	F*R	Purple	247
	Eucgr.K01002	(DT + RV)	Ankyrin repeat family protein	F,R	NA	–
	Eucgr.H01247	(DT + RV)	UDP-glucosyl transferase 72E1	F*R	Purple	241
	Eucgr.K02541	(RV)	COBRA-like glycosyl-phosphatidyl inositol-anchored protein family	F*R	Purple	246
	Eucgr.H03662	(RV)	Major facilitator superfamily protein	F	Purple	212
	Eucgr.D00406	(RV)	senescence regulator	F	Purple	243
	Eucgr.D02334	(RV)	BCL-2-associated athanogene 6	F,R	Purple	231
	Eucgr.B02486	(RV)	FASCICLIN-like arabinogalactan-protein 12	F*R	Purple	246
	Eucgr.C00773	(RV)	Riboflavin synthase-like superfamily protein	F*R	Purple	232
	Eucgr.L03244	(RV)	galactinol synthase 2	F*R	Purple	246
	Eucgr.F01203	(RV)	IQ-domain 10	F	Purple	190
	Eucgr.A02802	(RV)	NA	F	Green	228
	Eucgr.H03616	(RV)	ARM repeat superfamily protein	F*R	Purple	237
	Eucgr.E00460	(RV)	Haloacid dehalogenase-like hydrolase (HAD) superfamily protein	F*R	Purple	246
	Eucgr.E04087	(RV)	S-locus lectin protein kinase family protein	F	Purple	244
	Eucgr.H05151	(RV)	NA	F	Purple	223
	Eucgr.L03739	(RV)	NA	F	Purple	246
	Eucgr.J00951	(RV)	2-oxoglutarate and Fe(II)-dependent oxygenase superfamily protein	F,R	Purple	245
2	Eucgr.K03413	(DT + RV)	heat stable protein 1	F	Greenyellow	213
	Eucgr.A01768	(DT)	Cyclin D6; 1	F	Brown	191
	Eucgr.H01090	(DT)	NA	F,R	Green	74
	Eucgr.H03090	(DT)	Integrase-type DNA-binding superfamily protein	F	Tan	102
	Eucgr.L01034	(DT)	ATP-dependent caseinolytic protease/crotonase family protein	F	Greenyellow	183
	Eucgr.B03126	(DT)	sulfotransferase 12	F	Greenyellow	196
	Eucgr.C02911	(DT)	Phosphorylase superfamily protein	F	Purple	231
3	Eucgr.K00207	(DT + RV)	AGAMOUS-like 20	F	Brown	167
	Eucgr.K00204	(RV)	AGAMOUS-like 20	F	Brown	213
	Eucgr.K00203	(RV)	AGAMOUS-like 20	F	Brown	182
	Eucgr.H04691	(RV)	NA	F	Blue	261
	Eucgr.G00601	(RV)	vesicle-associated membrane protein 725	R	Green	183
	Eucgr.C02508	(RV)	FAD/NAD(P)-binding oxidoreductase family protein	R	NA	–
	Eucgr.H04975	(RV)	CCAAT-binding factor	R	NA	–

high number of links with other genes in their network (degree>200), showing that g-CovSel has prioritized the selection of genes that are potentially considered to be hub genes. This clearly demonstrates the relevance of the (CV, CR) coordinate system used.

6. Conclusion

This article proposes a new feature selection method, g-CovSel, which extends the CovSel approach to select groups of variables. Whereas CovSel was designed to provide a parsimonious selection of non-redundant variables, g-CovSel proposes to enrich this selection by creating groups around the variables that would have been selected by CovSel. The CovSel algorithm has been modified to take into account both the covariance of the selected variables with the response to be predicted and the inter-group correlation. It's the combination of these two features that makes the g-CovSel method so novel. Two algorithms for creating groups are proposed and illustrated on two very different examples: infrared spectra and genomic data. The two case studies demonstrate the relevance of the proposed theoretical framework. The CovSel variables are found at the center of the groups formed, making g-CovSel a generalization of CovSel. Most of the groups formed can be interpreted from a functional point of view, making g-CovSel a tool of choice for biomarker identification in omics. The basic g-CovSel idea of projecting variables into a reduced space (CR, CV) has been validated, but automatic group generation remains problematic. Indeed, the aim of the RV algorithm is to start from the corner (1,1) of the (CR, CV) square and to wind up, little by little, the variables close to this corner. In cases where there are many variables close to the corner (e.g. with NIR), the first variables will be close to the corner, and the group formation is not problematic. In cases where few variables populate the top-left corner, the RV algorithm will quickly jump to variables that are either in the center of the square, or near the bottom or left borders. Further studies will be carried out on such complex data, in order to test the limits of the proposed clustering algorithms and propose new strategies. A user interface will also be developed to enable domain expertise to be used to define group sizes. Further work will also be carried out to generalize g-CovSel to multi-block and multi-way data, e.g. following the scheme proposed by SO-CovSel [20] and N-CovSel [21].

CRediT authorship contribution statement

Jean-Michel Roger: Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Alessandra Biancolillo:** Writing – review & editing, Validation, Methodology, Formal analysis. **Bénédicte Favreau:** Writing – review & editing, Validation, Data curation. **Federico Marini:** Writing – review & editing, Validation, Methodology, Investigation, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This study was carried out within the Agritech National Research Center and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4 – D.D. 1032 17/06/2022, CN00000022). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

References

[1] W.R. Brode, Chemical Spectroscopy, John Wiley & Sons, 1939.
[2] R.P. Scott, Techniques and Practice of Chromatography, vol. 70, CRC Press, 1995.
[3] H.E. Duckworth, R.C. Barber, V.S. Venkatasubramanian, Mass Spectroscopy, 1986.
[4] F.Y. Kuo, I.H. Sloan, Lifting the curse of dimensionality, Notices of the AMS 52 (11) (2005) 1320–1328.
[5] M.F. Kabir, T. Chen, S.A. Ludwig, A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction, Healthcare Analytics 3 (2023) 100125.
[6] R.G. Brereton, J. Jansen, J. Lopes, F. Marini, A. Pomerantsev, O. Rodionova, R. Tauler, Chemometrics in analytical chemistry—part I: history, experimental design and data analysis tools, Anal. Bioanal. Chem. 409 (2017) 5891–5899.
[7] R.G. Brereton, J. Jansen, J. Lopes, F. Marini, A. Pomerantsev, O. Rodionova, R. Tauler, Chemometrics in analytical chemistry—part II: modeling, validation, and applications, Anal. Bioanal. Chem. 410 (2018) 6691–6704.
[8] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods 6 (9) (2014) 2812–2831.
[9] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta 185 (1986) 1–17.
[10] R. Tauler, Multivariate curve resolution applied to second order data, Chemometr. Intell. Lab. Syst. 30 (1) (1995) 133–146.
[11] T. Mehmood, K.H. Liland, L. Snipen, S. Sæbø, A review of variable selection methods in partial least squares regression, Chemometr. Intell. Lab. Syst. 118 (2012) 62–69.
[12] H. Liu, H. Motoda, Feature Selection for Knowledge Discovery and Data Mining, vol. 454, Springer Science & Business Media, 2012.
[13] P. Dhal, C. Azad, A comprehensive survey on feature selection in the various fields of machine learning, Appl. Intell. 52 (4) (2022) 4543–4581.
[14] M.J. Anzanello, F.S. Fogliatto, A review of recent variable selection methods in industrial and chemometrics applications, Eur. J. Ind. Eng. 8 (5) (2014) 619–645.
[15] S. Wold, E. Johansson, M. Cocchi, PLS: partial least squares projections to latent structures, in: 3D QSAR in Drug Design: Theory, Methods and Applications, Kluwer ESCOM Science Publisher, 1993, pp. 523–550.
[16] L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Interval partial least-squares regression (i PLS): a comparative chemometric study with an example from near-infrared spectroscopy, Appl. Spectrosc. 54 (3) (2000) 413–419.
[17] J.M. Monteiro, A. Rao, J. Shawe-Taylor, J. Mourao-Miranda, Alzheimer's Disease Initiative, A multiple hold-out framework for sparse partial least squares, J. Neurosci. Methods 271 (2016) 182–194.
[18] J. Camacho, E. Saccenti, Group-wise partial least square regression, J. Chemometr. 32 (3) (2018) e2964.
[19] J.M. Roger, et al., CovSel: variable selection for highly multivariate and multi-response calibration: application to IR spectroscopy, Chemometr. Intell. Lab. Syst. 106 (2) (2011) 216–223.
[20] Alessandra Biancolillo, Federico Marini, Jean-Michel Roger, SO-CovSel: a novel method for variable selection in a multiblock framework, J. Chemometr. 34 (2) (2020) e3120.
[21] Alessandra Biancolillo, Jean-Michel Roger, Federico Marini, N-CovSel, a new strategy for feature selection in N-way data, Anal. Chim. Acta 1231 (2022) 340433.
[22] E. Vigneau, E.M. Qannari, Clustering of variables around latent components, Commun. Stat. Simulat. Comput. 32 (4) (2003) 1131–1150.
[23] A. Savitzky, M.J. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (8) (1964) 1627–1639.
[24] B. Favreau, M. Denis, R. Ployet, F. Mounet, H. Peireira da Silva, L. Franceschini, H. Carrer, Distinct leaf transcriptomic response of water deficient *Eucalyptus grandis* submitted to potassium and sodium fertilization, PLoS One 14 (6) (2019) e0218528.
[25] P. Williams, K. Norris, Near-infrared Technology in the Agricultural and Food Industries, American Association of Cereal Chemists, Inc, 1987.
[26] N. Castelán-Muñoz, J. Herrera, W. Cajero-Sánchez, M. Arrizubieta, C. Trejo, B. García-Ponce, A. Garay-Arroyo, MADS-box genes are key components of genetic regulatory networks involved in abiotic stress and plastic developmental responses in plants, Front. Plant Sci. 10 (2019) 853.
[27] B.G. Osborne, T. Fearn, P.H. Hindle, Practical Near Infrared Spectroscopy with Applications in Food and Beverage Analysis, Logman Scientific & Technical, Essex, England, 1993.
[28] H. Jiang, J. Lu, Using an optimal CC-PLSR-RBFNN model and NIR spectroscopy for the starch content determination in corn, Spectrochim. Acta Mol. Biomol. Spectrosc. 196 (2018) 131–140.