Sparse common and distinctive covariates regression

Park, Soogeun; Ceulemans, Eva; Van Deun, Katrijn

Published in:
Journal of Chemometrics

DOI:
10.1002/cem.3270

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

Received: 23 December 2019 | Revised: 8 April 2020 | Accepted: 24 May 2020

DOI: 10.1002/cem.3270

SPECIAL ISSUE - RESEARCH ARTICLE

Journal of CHEMOMETRICS WILEY

Sparse common and distinctive covariates regression

Soogeun Park[1] | **Eva Ceulemans**[2] | **Katrijn Van Deun**[1]

[1]Department of Methodology and Statistics, School of Social and Behavioral Sciences, Tilburg University, Tilburg, Netherlands

[2]Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium

Correspondence
Soogeun Park, School of Social and Behavioral Sciences, Tilburg University, Tilburg, Netherlands.
Email: s.park_1@uvt.nl

Funding information
Nederlandse Organisatie voor Wetenschappelijk Onderzoek, Grant/Award Number: VIDI 452.16.012

Abstract

Having large sets of predictors from multiple sources concerning the same observation units and the same criterion is becoming increasingly common in chemometrics. When analyzing such data, chemometricians often have multiple objectives: prediction of the criterion, variable selection, and identification of underlying processes associated to individual predictor sources or to several sources jointly. Existing methods offer solutions regarding the first two aims of uncovering the predictive mechanisms and relevant variables therein for a single block of predictor variables, but the challenge of uncovering joint and distinctive predictive mechanisms and the relevant variables therein in the multisource setting still needs to be addressed. To this end, we present a multiblock extension of principal covariates regression that aims to find the complex mechanisms in which several or single sources may be involved; taken together, these mechanisms predict an outcome of interest. We call this method sparse common and distinctive covariates regression (SCD-CovR). Through a simulation study, we demonstrate that SCD-CovR provides competitive solutions when compared with related methods. The method is also illustrated via an application to a publicly available dataset.

KEYWORDS

common and distinctive processes, data integration, multiblock data, principal covariates regression, variable selection

1 | INTRODUCTION

When predicting an outcome by a number of predictor variables, there often is the additional aim to obtain insight in the mechanisms at play. For example, when modeling vaccine efficacy as a function of mRNA transcription rates soon after vaccination,[1] setting up a prediction tool was not the only aim. The authors also wanted to understand the involved biological processes by finding—in the transcriptomics data—those biological pathways that are associated to the efficacy of the vaccine. To obtain an even deeper understanding of the system under study often, large and heterogeneous collections of data are used, which results in several blocks of predictors pertaining to the same observation units. A prominent example is multi-omics studies. These are used to obtain a better understanding of disease mechanisms by jointly studying several features of the biological system (e.g., genomic, transcriptomic, and proteomic data collected from the same sample of patients and controls).[2] Obtaining insights from such large multiblock data implies revealing (1) the relevant features in the system and (2) the orchestration of the system (which features act jointly and which ones

act individually in shaping the outcome). For example, the emergence of asthma is known to depend on a complex interplay between genetic susceptibility and environmental exposure.[3] A complicating factor in the analysis of the data is that they often consist of large collections of untargeted variables, which implies that it is the data analyst's task to sort out the relevant predictors from the variables that are irrelevant for the process under study. Moreover, such selection of variables is necessary to ease the interpretation of the resulting model and to address model inconsistency in the high-dimensional setting of (many) more variables than cases.[4]

Within chemometrics, partial least squares (PLS) and principal covariates regression (PCovR) are popular methods that target the twofold goals of deriving the components that represent the underlying processes and predicting the criterion variables. Variants of the methods suited for multiblock data have been devised and shown to be useful at extracting insight about the mechanisms while predicting the criterion variable. Examples include incorporating information on physical properties of intermediate granules when modeling the relationship between process variables and crushing strength of finished tablets,[5] predicting sensory attributes of carrot genotypes via finding joint mechanisms concerning dry matter content, non-volatile and volatile compounds,[6] and mapping an interrelated model between consumer preference and sensory information such as odor and flavor pertaining to different flavored water samples.[7] As these multiblock methods are subject to interpretational difficulties due to a large number of predictors, sparse PCovR (SPCovR) and sparse PLS (SPLS) were devised to provide solutions that perform variable selection.[8,9] Furthermore, viewing each block of predictors as representative of a part of the system under study, multiblock data may present two different types of underlying predictive mechanisms; those that pertain only to variables from a single predictor block and the mechanisms that require joint involvement of variables from multiple predictor blocks. We denote the two types of mechanisms by distinctive and (partially) common mechanisms, respectively (with partially indicating mechanisms that pertain to variables from multiple though not all blocks). Identification of these mechanisms has not been fully addressed in the context of criterion prediction by the existing methods.

On the other hand, for purely explorative purposes (this is, only revealing underlying mechanisms without trying to predict a criterion), methods that specifically aim to capture common and distinctive processes have been put forward. Simultaneous component analysis (SCA) with distinctive and common components (DISCO-SCA), joint and individual variation explained (JIVE), and similar other approaches aim to unravel the structure of the underlying processes by separating common and distinctive mechanisms.[10,11] Måge et al[12] provided a comprehensive comparison of the performance of several of these approaches under varying data structures, whereas Smilde et al[13] proposed a general framework for the methods devised to decompose multiblock data into common and distinctive processes. Moreover, to attain more interpretable solutions especially with high dimensional data, sparse methods have been developed that capture the common and distinctive processes by incorporating particular penalty terms or prespecified structures.[14-16]

Along these lines of research, a method is needed that serves the twofold goals of obtaining insightful predictive models in the setting of high dimensional multiblock data. As discussed, such a method should incorporate predictor selection and uncover the common and distinctive predictive mechanisms. The development of such a method could be envisaged both along the PLS and PCovR lines. Yet, in comparison with SPLS, SPCovR has been shown to be more effective in recovering the underlying processes,[8] and it also offers more flexibility concerning the importance assigned to the dual aim of prediction of the criterion variable and the reconstruction of the predictor variables. Therefore, the current paper focuses on PCovR and integrates the sparse PCovR and SCA methods in the new sparse common and distinctive covariates regression (SCD-CovR) method. We evaluate the performance of SCD-CovR by comparing it with other methods that are characterized by similar goals such as sparse generalized canonical correlation analysis (SGCCA) that is based on PLS.[17]

The paper is arranged as follows. First, we describe SCD-CovR in detail, followed by a brief overview of existing related methods. Then, simulation studies that comparatively demonstrate the performance of SCD-CovR and other methods are presented, and their results are discussed. Finally, we conclude the paper by formulating some limitations and directions for future research. The implementation of SCD-CovR was done in R, and it can be found on Github: https://github.com/soogs/SCD-CovR, along with the code used to generate the results reported in this paper.

2 | SPARSE COMMON AND DISTINCTIVE COVARIATES REGRESSION

We will use the following notation throughout the paper: scalars, vectors, and matrices are denoted by italic lowercase, bold lowercase, and bold uppercase letters, respectively. Transposing is indicated by the superscript ^{T}. Lowercase subscripts running from 1 to corresponding uppercase letters denote indexing: $i \in \{1,2, ..., I\}$. Subscript $_C$ indicates

concatenation of multiple data blocks, whereas superscripts $^{(X)}$ and $^{(y)}$ highlight affiliation with predictor and criterion variables, respectively.

2.1 | Model and objective function

SCD-CovR models a criterion in function of multiple blocks of predictors all obtained from the same set of observation units. Let $\mathbf{X}_k$ be a column-centered matrix containing the scores of the I observation units on the J_k predictors in the kth predictor block, with $k \in \{1,2, ..., K\}$. Also, let $\mathbf{y}$ be a centered vector containing the I scores on the criterion.

The SCD-CovR model is based on the well-known principal component analysis (PCA) model that takes the following formulation for $\mathbf{X}_k$:

$$\mathbf{X}_k = \mathbf{X}_k \mathbf{W}_k \left(\mathbf{P}_k^{(X)}\right)^T + \mathbf{E}^{(X)}, \tag{1}$$

where $\mathbf{W}_k$ and $\mathbf{P}_k^{(X)}$ are $J_k \times R$ matrices of component weights and loadings, respectively. To identify the solution, usually, the constraint $\left(\mathbf{P}_k^{(X)}\right)^T \mathbf{P}_k^{(X)} = \mathbf{I}_R$ is added under a principal axes orientation. The weights define how the predictors are combined into the R principal components (viz., $\mathbf{T}_k = \mathbf{X}_k \mathbf{W}_k$, implying $t_{ir} = \sum_{j_k} x_{ij_k} w_{j_k r}$) whereas the loadings express the relationship between them. $\mathbf{E}^{(X)}$ is used to denote the matrix of residuals. This formulation is known as the weight-based model.[14]

PCovR explicitly models the criterion as a function of the components in the PCA model (1):

$$\mathbf{y} = \mathbf{X}_k \mathbf{W}_k \mathbf{p}^{(y)} + \mathbf{e}^{(y)}, \tag{2}$$

with $\mathbf{p}_k^{(y)}$ the vector of R regression coefficients and $\mathbf{e}^{(y)}$ the residuals pertaining to the criterion. The twofold aim of PCovR in reconstructing $\mathbf{X}_k$ and predicting $\mathbf{y}$ is expressed by the objective function to be minimized[18]:

$$L\left(\mathbf{W}_k, \mathbf{P}_k^{(X)}, \mathbf{p}^{(y)}\right) = \alpha \frac{\left\|\mathbf{y} - \mathbf{X}_k \mathbf{W}_k \mathbf{p}^{(y)}\right\|_2^2}{\|\mathbf{y}\|_2^2} + (1-\alpha) \frac{\left\|\mathbf{X}_k - \mathbf{X}_k \mathbf{W}_k \left(\mathbf{P}_k^{(X)}\right)^T\right\|_2^2}{\|\mathbf{X}_k\|_2^2}, \tag{3}$$

with $0 \le \alpha \le 1$ a known constant. The α parameter specifies the balance between modeling the criterion and modeling the block of predictors. With α set at 0, the method is identical to PCA followed by regression, whereas at 1, it becomes equivalent to linear regression (viz., $y_i = \sum_r p_r^{(y)} t_{ir} = \sum_r \left(\sum_{j_k} p_r^{(y)} x_{ij_k} w_{j_k r}\right) = \sum_{j_k} \left(\sum_r p_r^{(y)} w_{j_k r}\right) x_{ij_k}$, with $\sum_r p_r^{(y)} w_{j_k r}$ as a regression coefficient for the j_kth predictor). How to optimally balance α has been explicitly explored by Vervloet et al.[19] Note that to identify the PCovR solution, De Jong and Kiers[18] introduced the constraint $\mathbf{T}^T\mathbf{T} = \mathbf{I}_R$. As pointed out by Vervloet et al,[19] the solution is still subject to rotational freedom.

As PCA and PCovR construct the components by linearly combining all the predictors, the interpretation of the components can be difficult, especially when the number of predictors grows large. The solutions can also be inconsistent in the high-dimensional setup.[20] To overcome these issues, Zou et al[21] devised a sparse PCA method that imposes regularization penalties on the objective function. Note that sparse implies that many of the component weights are penalized to become zero. A sparse variant of PCovR, SPCovR, was also developed in a similar manner.[8] SPCovR finds the solutions by minimizing the following objective function:

$$\begin{aligned} L\left(\mathbf{W}_k, \mathbf{P}_k^{(X)}, \mathbf{p}^{(y)}\right) &= \alpha \frac{\left\|\mathbf{y} - \mathbf{X}_k \mathbf{W}_k \mathbf{p}^{(y)}\right\|_2^2}{\|\mathbf{y}\|_2^2} + (1-\alpha) \frac{\left\|\mathbf{X}_k - \mathbf{X}_k \mathbf{W}_k \left(\mathbf{P}_k^{(X)}\right)^T\right\|_2^2}{\|\mathbf{X}_k\|_2^2} \\ &\quad + \lambda_L |\mathbf{W}_k|_1 + \lambda_R \|\mathbf{W}_k\|_2^2 \end{aligned}, \tag{4}$$

such that $\left(\mathbf{P}_k^{(X)}\right)^T \mathbf{P}_k^{(X)} = \mathbf{I}_R$ and with $\lambda_L \geq 0$, $\lambda_R \geq 0$ and $\alpha \geq 0$. The regularization parameters are the lasso, with $|\mathbf{W}_k|_1 = \sum_{j_k,r} |w_{j_k r}|$, and the ridge $\|\mathbf{W}_k\|_2^2 = \sum_{j_k,r} w_{j_k r}^2$, together forming the elastic net.[22] The former shrinks and forces certain weights to be exactly zero, whereas the latter only shrinks the estimates. Therefore, the lasso penalty is employed to obtain sparse weights, whereas the ridge penalty is required to ensure stable estimates under high-dimensionality. It can also be seen that when both of the tuning parameters, λ_L and λ_R, are 0, the PCovR formulation (3) is retrieved. Note that because of the penalties, the SPCovR model is identified and not subject to rotational freedom. However, the components pertain to permutational freedom and sign invariance.

SPCovR and the above methods only target data with a single predictor block and hence do not address the questions associated with multiple predictor blocks. These questions can be answered by performing a joint decomposition of the K predictor blocks into components by imposing a SCA model[23]:

$$\mathbf{X}_C = \mathbf{X}_C \mathbf{W}_C \left(\mathbf{P}_C^{(X)}\right)^T + \mathbf{E}^{(X)}, \tag{5}$$

where $\mathbf{X}_C = [\mathbf{X}_1, ..., \mathbf{X}_K]$ (of size $I \times \sum_{k=1}^K J_k^{(X)}$) denotes the supermatrix that concatenates the predictor blocks. Consequently, $\mathbf{W}_C$ and $\mathbf{P}_C^{(X)}$ are weight and loading matrices of size $\sum_{k=1}^K J_k^{(X)} \times R$. Hence, the criterion variable can be modeled using the SCA weights:

$$\mathbf{y} = \mathbf{X}_C \mathbf{W}_C \mathbf{p}^{(y)} + \mathbf{e}^{(y)}, \tag{6}$$

with $\mathbf{p}_C^{(y)}$ a vector of R regression coefficients.

As the interpretation of SCA solutions is even more challenging, sparse SCA methods were devised.[14] Furthermore, a sparse SCA method that explicitly models common and distinctive processes was proposed. This method, sparse common and distinctive SCA (SCaDS), minimizes the following objective function[16]:

$$L\left(\mathbf{W}_C, \mathbf{P}_C^{(X)}\right) = \left\| \mathbf{X}_C - \mathbf{X}_C \mathbf{W}_C \left(\mathbf{P}_C^{(X)}\right)^T \right\|_2^2 + \lambda_L |\mathbf{W}_C|_1 + \lambda_R \|\mathbf{W}_C\|_2^2, \tag{7}$$

such that $\left(\mathbf{P}_C^{(X)}\right)^T \mathbf{P}_C^{(X)} = \mathbf{I}_R$ and subject to *zero block constraints* on $\mathbf{W}_C$ that fix block-specific sets of weights—pertaining to one or several predictor blocks—to zero. This implies that the component is determined only by predictors of those blocks for which the weights have not been fixed to zero. Common components are obtained by not placing such zero block constraints on the component. The lasso penalty is used in addition to the zero block constraints to achieve sparseness within the common and distinctive components. As an alternative to using such a fixed structure, sparse multiblock PCA methods that rely on a group lasso penalty (which has the property to shrink entire groups of coefficients to zero) have also been proposed.[15]

Building upon SCaDS and SPCovR, we propose the SCD-CovR that predicts the criterion, while providing sparse solutions that capture the common and distinctive processes in the predictor blocks. SCD-CovR implies minimizing the following objective function:

$$\begin{aligned} L\left(\mathbf{W}_C, \mathbf{P}_C^{(X)}, \mathbf{p}^{(y)}\right) = {} & \alpha \frac{\left\| \mathbf{y} - \mathbf{X}_C \mathbf{W}_C \mathbf{p}^{(y)} \right\|_2^2}{\|\mathbf{y}\|_2^2} + (1-\alpha) \frac{\left\| \mathbf{X}_C - \mathbf{X}_C \mathbf{W}_C \left(\mathbf{P}_C^{(X)}\right)^T \right\|_2^2}{\|\mathbf{X}_C\|_2^2} \\ & + \lambda_L |\mathbf{W}_C|_1 + \lambda_R \|\mathbf{W}_C\|_2^2 \end{aligned}, \tag{8}$$

such that $\left(\mathbf{P}_C^{(X)}\right)^T \mathbf{P}_C^{(X)} = \mathbf{I}_R$, and subject to zero block constraints on $\mathbf{W}_C$.

As in SCaDS, common and distinctive components can be obtained with SCD-CovR through the zero block constraints on $\mathbf{W}_C$. Similarly, as for SPCovR, the components both account for variation in the criterion *and* predictor

variables with α allowing to flexibly tune prediction and reconstruction. The $\mathbf{W}_C$ weights can be examined to understand which predictors define the derived common and distinct components. It is also easy to see that this method is an adaptation of PCovR. When λ_L and λ_R are equal to zero and with the absence of the zero block constraints, the formulation is identical to PCovR.

2.2 | Algorithm

To solve the optimization problem defined in (8), we use an alternating procedure where the loadings $\mathbf{P}_C^{(X)}$ and the regression coefficients $\mathbf{p}^{(y)}$ are solved for conditional upon fixed values for the weights $\mathbf{W}_C$ and vice versa. A schematic outline of the algorithm is given here below. The optimization procedure that we propose here closely follows those proposed for SCaDS and SPCovR.[8,16] This procedure boils down to solving for all components together (unlike deflation methods that solve for each component in turn) and using a coordinate descent procedure to solve the conditional elastic net problem to estimate the sparse weights. More details on the procedure can be found in Appendix A. The alternating routine ensures that the loss is nonincreasing and the algorithm converges to a stationary point, usually a local minimum. To avoid local minima problems, we recommend to use multiple random and a rational starting value based on PCovR.

Algorithm 1 SCD-CovR

Initialize:
$\mathbf{X}_C$ and $\mathbf{y}$, number of components R, weighting parameter α, regularization parameters λ_L and λ_R, maximum number of iterations T, convergence threshold $\epsilon \geq 0$
Initialize:
$\mathbf{W}_C \leftarrow \mathbf{W}_C{}^{(0)}$, $\mathbf{P}_C^{(X)} \leftarrow \mathbf{P}_C^{(X)(0)}$, $\mathbf{p}^{(y)} \leftarrow \mathbf{p}^{(y)(0)}$, $L_0 \leftarrow$ Initial loss,
Loss difference $d \leftarrow 1$, Iteration counter $t \leftarrow 1$
while $t < T$ **or** $\epsilon < d$ **do**
  Conditional estimation of $\mathbf{P}_C^{(X)(t)}$ and $\mathbf{p}^{(y)(t)}$ given $\mathbf{W}_C{}^{(t-1)}$
  Conditional estimation of $\mathbf{W}_C{}^{(t)}$ given $\mathbf{P}_C^{(X)(t)}$ and $\mathbf{p}^{(y)(t)}$
  $L_u \leftarrow$ updated loss given $\mathbf{W}_C{}^{(t)}$, $\mathbf{P}_C^{(X)(t)}$ and $\mathbf{p}^{(y)(t)}$
  $d \leftarrow L_0 - L_u$
  $t \leftarrow t + 1$
  $L_0 \leftarrow L_u$
end while

2.3 | Model selection

To use our proposed SCD-CovR method, values have to be provided for the number of components R, the weighting parameter α, the number of (partially) common and distinct components, and the ridge and lasso regularization parameters λ_L and λ_R. In order to select a suitable model, these parameters need to be tuned according to some optimality criterion. Several model selection strategies exist targeting different optimality criteria. These include cross-validation that is often recommended within the literature for methods involving regularization parameters. To optimize the optimality criterion, a grid search can be used, which exhaustively compares all possible combinations of the tuning values for the different parameters. A sequential approach where each parameter is tuned in turn can also be considered as it was demonstrated to work well for cross-validation for PCovR.[24] As cross-validation is computationally costly if we consider all combinations of the tuning parameters, we therefore opt to use the sequential approach in the simulation study and the empirical application. The procedures are implemented slightly differently in these two sections because no oracle information is available for the empirical example. However, in general, the procedures first optimize R, λ_R and α simultaneously, followed by tuning the zero block constraints and λ_L. An interesting feature of the sparse PCA or PCovR methods with sparse weights instead of loadings is that the level of sparsity does not closely relate

to the amount of variance explained; models comprised of components with very sparse weights can account for a comparable amount of variance as models that are much less or barely sparse.[16] The weights are used to construct the component scores and these can be approximated very well with few nonzero weights. This even means that distinctive components can still account for a considerable amount of variance in the data block(s) for which the component has all zero weights.

2.4 | Related methods

SCD-CovR is a method with three main objectives. It (a) predicts a criterion, (b) recovers the underlying common and distinctive predictor mechanisms via dimension reduction, and (c) derives sparse and therefore interpretable components. The method offers a solution that achieves all of these objectives in a balanced and a flexible manner. This section lists other component based methods devised to fulfill and balance these multiple objectives. When prediction is the only objective, methods with more emphasis on prediction may outperform SCD-CovR. In a similar vein, Smilde et al[25] commented that PLS usually yields better prediction if the multiple blocks are analyzed as one single "superblock". Accounting for the multiblock structure helps in revealing meaningful insights but may come with lower prediction quality. On the other hand, applying a componentwise approach or explicitly taking into account the multiblock structure regularizes the problem. As such procedures safeguard against overfitting, they may improve the prediction quality especially in unstable settings (e.g., high dimensional data).

A method often used to aim both at prediction and modeling the variation in the block of predictors is principal component regression (PCR). This method first performs PCA and then, in a second and separate step, regresses the criterion on the components. The PCA step can be performed with SCaDS (leading to PCR-SCaDS) to also meet the objectives of finding common and distinctive mechanisms and having sparse component weights. It is closely related to SCD-CovR, as the components found by PCR-SCaDS are equal to the SCD-CovR components that we would obtain if we set the weighting parameter α to zero. Moreover, both methods encourage the recovery of the common and distinctive structure by imposing zero block constraints on the weights matrix. In comparison with SCD-CovR, PCR-SCaDS does not take the regression problem into consideration when deriving the components, implying that the processes that underlie the predictors would be retrieved with higher quality. However, simultaneously, PCR-SCaDS suffers from the weakness that predictor components that explain a lot of variance in the criterion may not be recovered.[24]

SGCCA is another component-based method that addresses the multiple goals of simultaneous prediction and modeling the variation in the predictors. Being an extension of PLS, multiple data blocks are analyzed simultaneously to obtain sparse components while at the same time these components should account for the variation in the criterion.[17] Extracting components that also allow to predict well is similar to SCD-CovR but unlike PCR-SCaDS. However, whereas SCD-CovR provides a flexible framework to weight reconstruction of the predictors and prediction of the criterion, PLS-based methods tend to lean closer to prediction.[8,24] This also means that SGCCA may have more difficulties in recovering the underlying processes. Furthermore, methods based on PLS are often more prone to overfitting than those derived from PCovR, which in turn results in a diminished quality of out-of-sample prediction. Finally, SGCCA does not explicitly facilitate the retrieval of common and distinctive processes.

On top of these two methods, SPCovR can also be considered closely related to SCD-CovR. Their only difference is the zero block constraints on the weights for finding the common and distinctive structure. The two methods are expected to yield similar performance with respect to prediction. However, SCD-CovR can be expected to be better at capturing the common and distinctive underlying processes and thus in giving insight into joint and distinctive mechanisms.

Summarizing, the four methods can be expected to perform differently in terms of prediction and recovering the underlying components when administered to the same data. Concerning prediction, PCR-SCaDS is expected to underperform because it would be unable to capture an underlying process that is strongly associated to the criterion but accounts only for a minor portion of the variation in the predictor variables. We anticipate SGCCA to be more prone to overfitting than the other methods. Regarding correct recovery of the component weights, SGCCA would be relatively worse than the other methods due to its stronger focus on the prediction. Lastly, SCD-CovR and PCR-SCaDS are expected to recover the underlying common and distinctive processes more effectively than the other methods as they specifically target these processes through the zero block constraints.

3 | SIMULATION STUDY

Although adaptations of PLS, PCR, and PCovR have been compared in previous research,[8,24] they have not been put to test in settings where underlying common and distinctive processes are expected. Also, the effectiveness of the methods may depend on certain data characteristics. Therefore, we have conducted a simulation study in which we examine the performance of the methods with respect to sparse retrieval of the underlying processes, identification of common and distinctive components, and the prediction of the criterion.

3.1 | Design and procedure

Fixing the number of observations I to 100, two blocks of predictor variables were generated to represent three components with a common and distinctive structure. Two components represented processes distinctive to predictor block 1 and 2, respectively. The remaining component reflects a common process involving both of the blocks. We defined the three components such that one of them explains 50% of the true structural variance in the predictors, another one 40%, and the remaining one 10%. Adopting the terminology from Vervloet et al,[24] we refer to the first two components as "strong" components and call the third one a "weak" component. On the other hand, the three components also differ in "relevance" for predicting the criterion, in that one of them explains 66.7% of the true criterion variance and the other two 16.67% each. Finally, 70% of the weights and the loadings were made sparse.

We manipulated five data characteristics that are listed in the overview below. Each level within the manipulated factors is provided in square brackets. For the second and third factor which concern the strength and the relevance of the components, the proportion of variance explained is provided in the following order: [component distinctive to block 1, component distinctive to block 2, and common component].

Study setup

1. Number of predictors J_k in each block: [100], [10]
2. Strength of the three components: [50%, 40%, 10%], [10%, 40%, 50%]; in the first case, the common component is weak. In the second case, the first distinctive component is weak.
3. Relevance of the three components: [16.67%, 16.67%, 66.67%], [66.67%, 16.67%, 16.67%]; in the first case, the common component is the most relevant. In the second case, the first distinctive is.
4. Proportion of error in $\mathbf{X}_C$: [10%], [50%]
5. Proportion of error in $\mathbf{y}$: [10%], [50%]

To obtain two predictor blocks that correspond to the settings described above, the following procedure was followed. The true predictor matrix $\mathbf{X}_C^*$ is defined by the model $\mathbf{X}_C^* = \mathbf{X}_C^*\mathbf{W}_C\left(\mathbf{P}_C^{(X)}\right)^T$ where the weights and the loadings are equal and column-orthogonal: $\mathbf{W}_C = \mathbf{P}_C^{(X)}$. First, a random column-centered matrix $\mathbf{T}^*$ of size $I \times R$ was generated from a multivariate normal distribution with the identity matrix as covariance matrix. Subsequently, $\mathbf{T}^*$ was centered and column-orthogonalized to yield $\mathbf{T}$. Second, to obtain a sparse and orthogonal weights matrix, we started by generating a random weights matrix of $\mathbf{W}_C^*$ of size $\sum_k J_k \times R$ from a uniform distribution over the interval of [0, 1]. To create one distinctive component for each of the two predictor blocks, the weights of the predictors on this component were set to zero in the other block. In the remaining nonzero parts, randomly chosen elements were replaced by zeros to attain a sparsity level of 70% when computed across the full matrix. The resulting matrix was orthogonalized using a Gram–Schmidt procedure in a manner that the sparse elements are retained to yield the true weights matrix $\mathbf{W}_C$. Furthermore, a diagonal matrix $\mathbf{D}$ was created with the diagonal values representing the relative proportion of variance accounted for by the components (i.e., reflecting their strength). Because $\mathbf{W}_C = \mathbf{P}_C^{(X)}$, the true predictor matrix $\mathbf{X}_C^*$ was then obtained as $\mathbf{X}_C{}^* = \mathbf{TD}\left(\mathbf{P}_C^{(X)}\right)^T = \mathbf{X}_C^*\mathbf{W}_C\left(\mathbf{P}_C^{(X)}\right)^T$. Finally, residuals were added generated from a standard normal distribution and scaled such that the predictor blocks contain the desired level of error to yield $\mathbf{X}_C$. The proportion of error is defined as the proportion of total variance in the observed $\mathbf{X}_C$ or $\mathbf{y}$ that is due to error. The scores on the criterion variable were obtained in a similar fashion with the equation $\mathbf{y} = \mathbf{TDp}^{(y)} + \mathbf{e}^{(y)} = \mathbf{X}_C^*\mathbf{W}_C\mathbf{p}^{(y)} + \mathbf{e}^{(y)}$. To specify

the regression coefficients $\mathbf{p}^{(y)}$, we first fixed the coefficient pertaining to the second component to −0.3. This second component is constantly irrelevant across the conditions. The other two coefficients were specified according to the different levels of strength and relevance.

Fully crossing the conditions and generating 50 replicate datasets per condition, $2 \times 2 \times 2 \times 2 \times 2 \times 50 = 1600$ datasets were produced. Each of the 1600 datasets was subjected to eight different analyses. The different analysis methods resulted from crossing the following four methods with two different numbers of extracted components.

Analysis methods

1. Method: [SCD-CovR], [SPCovR], [PCR-SCaDS], [SGCCA]
2. Number of components extracted: [2], [3]

Although a three-component model was used for data generation, we varied the extracted number of components because we aim to understand the behavior and the performance of different methods at identifying the components. When methods extract two components from data generated using a three-component model, methods can focus on different aspects and thus yield different subsets of components. As the relevance and the strength of the three components are manipulated across the conditions, we can observe how both aspects determine which two components are extracted. For example, as mentioned in Section 2.4, we expect PCR-SCaDS to recover the strong components rather than the relevant components.

3.2 | Model selection

The number of components R extracted for all four methods is fixed by the study design. A few other tuning parameters were fixed such as to correspond to the true model structure. Suitable values for the tuning parameters that were not fixed were found sequentially, for each data set and each analysis method.

For SCD-CovR, using the given R, we first simultaneously tuned the weighting parameter α and the ridge penalty λ_R via 10-fold cross-validation, keeping the lasso penalty λ_L at 0 (which therefore does not induce any sparsity) and the zero block constraints such that no distinctive components are imposed. We adopted the one standard error (SE) rule to select a parameter that yields the most general model among the set of parameters with errors within one SE from the minimal cross validation error. Usually, generality of models indicates that the model is unsaturated and thus easy to interpret and unlikely to overfit. Because higher α values place more emphasis on criterion prediction and therefore lead to a model more prone to overfitting, we chose the lowest α value via the one SE rule. Second, a suitable common and distinctive component structure was determined. When extracting two components, the zero block constraints on $\mathbf{W}_C$ that provide the structure of the common and distinctive components were chosen through 10-fold cross-validation. We selected the common and distinctive structure of the two components which led to the smallest cross validation error. The one SE rule was not used because it is difficult to define what a general model is with regards to the common and distinctive structure. On the other hand, for retrieving three components, the defined true structure was provided. The lasso parameter was tuned by selecting the value that results in the correct number of zero component weights.

For SPCovR, the set of tuning parameters is the same as SCD-CovR except for the zero block constraints. As α and λ_R were selected without any zero block constraints for SCD-CovR, these values were adopted for SPCovR (note that when the zero block constraints do not impose distinctive components, SCD-CovR is equivalent to SPCovR). Also here, λ_L was tuned to return the correct number of zero coefficients.

For PCR-SCaDS, the number of common and distinctive components as well as λ_R and λ_L needs to be determined. We started the sequential approach by performing 10-fold cross-validation with the one SE rule for determining λ_R. Next, the zero block constraints and λ_L were found as previously discussed for SCD-CovR.

Finally, for SGCCA, the λ_L tuning parameter was fixed to yield the same number of zero-coefficients as in the generated data. The ridge penalty in SGCCA was tuned using the default setting the package provides.

3.3 | Evaluation criteria

The four considered methods serve multiple aims: predicting a criterion, capturing possible common and distinctive underlying processes, and providing sparse solutions for better interpretation. To assess the effectiveness of the methods at meeting these aims, we employed two evaluation criteria.

1. Out-of-sample R^2: equivalent to the R^2 measure for ordinary least squares (OLS) but applied for an independent out-of-sample test set.
2. Correct classification rate: proportion of $\mathbf{W}_C$ coefficients correctly classified as zero and nonzero elements relative to the total number of coefficients.

The independent test set (of 100 observation units) needed for computing the out-of-sample R^2 was generated following the same underlying model and the procedures as the data used for estimation. The out-of-sample R^2 measure is computed by the following equation:

$$R^2_{out\text{-}of\text{-}sample} = 1 - \frac{\|\mathbf{y}_{test} - \hat{\mathbf{y}}_{test}\|_2^2}{\|\mathbf{y}_{test}\|_2^2}, \tag{9}$$

where $\mathbf{y}_{test}$ refers to the $\mathbf{y}$ scores from an out-of-sample test set and $\hat{\mathbf{y}}_{test}$ indicates the predicted score that corresponds to $\mathbf{y}_{test}$. Therefore, $\frac{\|\mathbf{y}_{test} - \hat{\mathbf{y}}_{test}\|_2^2}{\|\mathbf{y}_{test}\|_2^2}$ refers to the scaled sum of squared prediction error. Because this scaled sum of prediction error can be larger than one, it is possible for the out-of-sample R^2 to take a negative value. The correct classification rate is computed by comparing the true and the estimated $\mathbf{W}_C$ weights matrices. To handle the permutational freedom and the sign invariance of the estimated components, we calculated Tucker congruence between the columns of the true $\mathbf{W}_C$ matrix and those of the estimated $\mathbf{W}_C$ matrix. After pairing the true and estimated $\mathbf{W}_C$ columns that resulted in the highest Tucker congruence, the correct classification rate is calculated from the matching pairs of true and estimated $\mathbf{W}_C$ columns. This strategy was also used when only two components were extracted: they were matched to those two components of the three true ones that yield the highest Tucker congruence.

3.4 | Results

3.4.1 | Out-of-sample R^2

First, we consider the performance of the four methods in terms of how well they predict new data. The results are summarized in Figures 1-3, and 4. The first two of these refer to the results obtained when extracting two components only, whereas the latter two refer to the analyses with three extracted components.

The aggregated results over all conditions, for the analyses with two extracted components, can be found in Figure 1. It can be observed that on average, PCR-SCaDS has smaller out-of-sample R^2 than the other three methods. The latter show similar performance among each other. To examine whether there is an effect of the design factors and of the used method on out-of-sample R^2, we studied how the out-of-sample R^2 changes according to each of the conditions in the design by observing the boxplots.

Figure 2 presents these boxplots of out-of-sample R^2 arranged for each condition, conveying that the proportion of error variance in $\mathbf{y}$ plays an influential role in the performance of the methods. In the conditions where the error

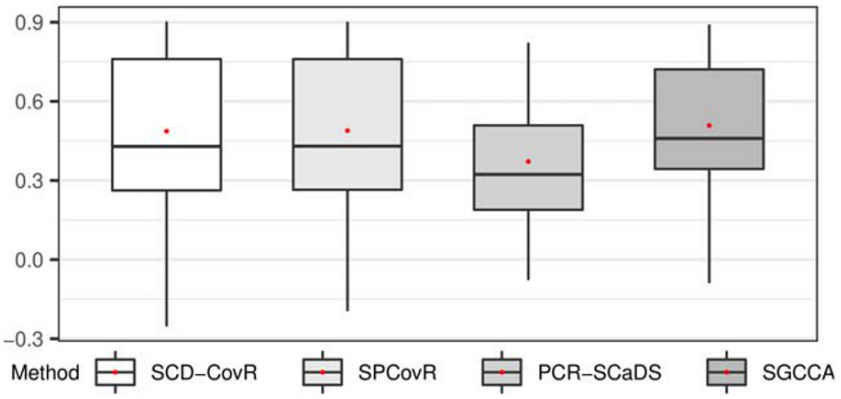

FIGURE 1 Box plots of the out-of-sample R^2 when two components are extracted: aggregated results. The red dot indicates the mean. PCR-SCaDS, principal component regression followed by sparse common and distinctive simultaneous component analysis, SCD-CovR, sparse common and distinctive covariates regression; SGCCA, sparse generalized canonical correlation analysis; SPCovR, sparse principal covariates regression

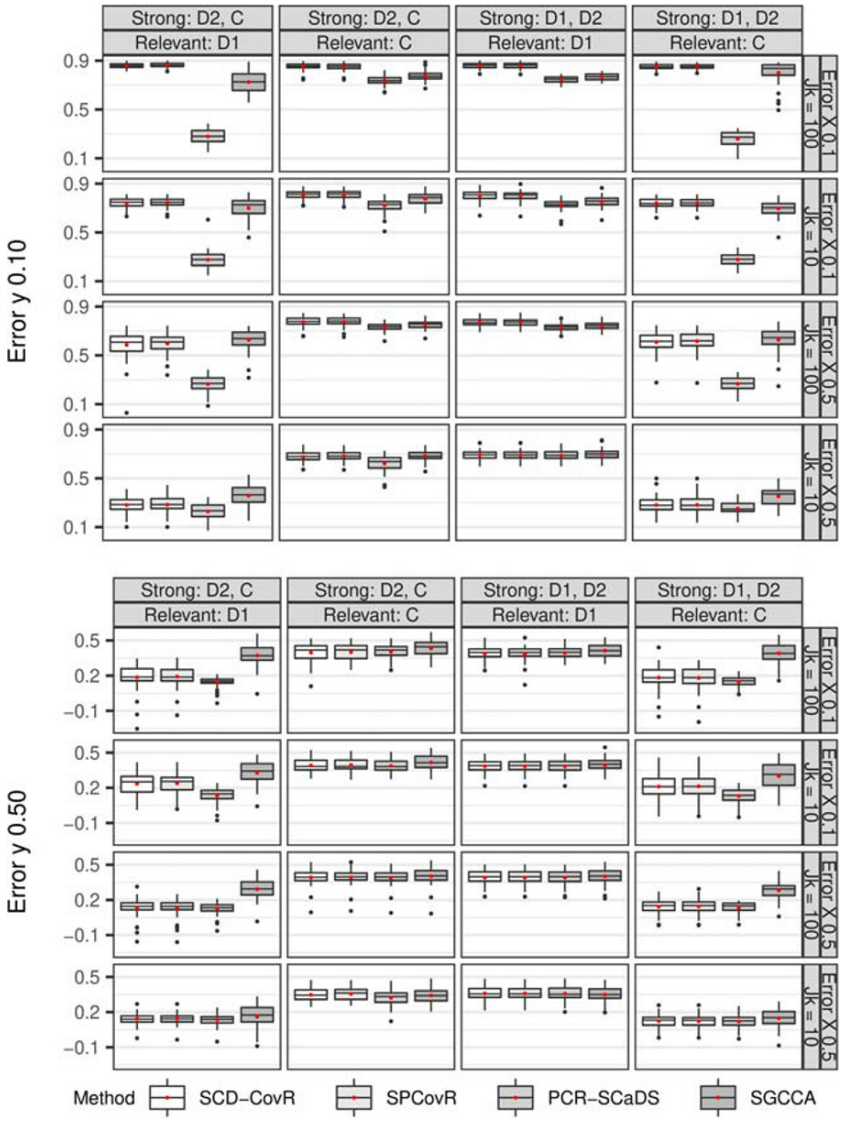

FIGURE 2 Box plots of the out-of-sample R^2 when two components are extracted; each panel corresponds to one of the 16 conditions. The column panels indicate the manipulated strength and relevance of the three components; D1 and D2 denote the components distinctive to block 1 and 2, respectively, whereas C refers to the common component. The row panels indicate the number of variables Jk in each predictor block. PCR-SCaDS, principal component regression followed by sparse common and distinctive simultaneous component analysis, SCD-CovR, sparse common and distinctive covariates regression; SGCCA, sparse generalized canonical correlation analysis; SPCovR, sparse principal covariates regression

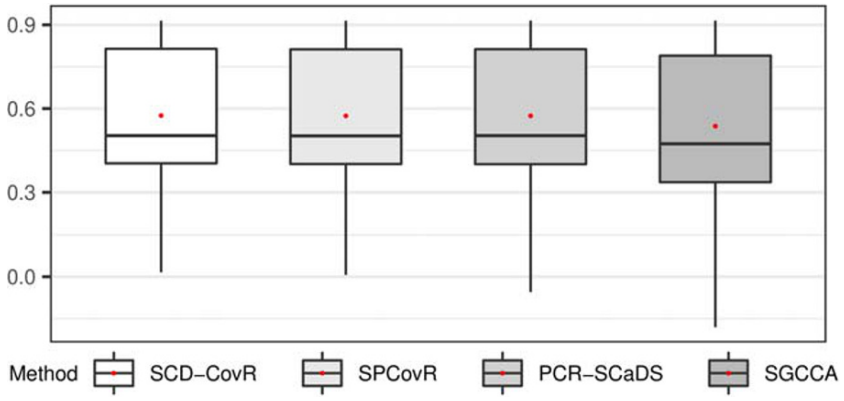

FIGURE 3 Box plots of the out-of-sample R^2 when three components are extracted: aggregated results. The red dot indicates the mean. PCR-SCaDS, principal component regression followed by sparse common and distinctive simultaneous component analysis, SCD-CovR, sparse common and distinctive covariates regression; SGCCA, sparse generalized canonical correlation analysis; SPCovR, sparse principal covariates regression

variance in **y** equals 10%, the four methods have comparable levels of prediction performance in those situations where the strong component is relevant for prediction (the two columns in the middle). In contrast, when the component relevant for prediction is a weak one, the out-of-sample R^2 of PCR-SCaDS decreases considerably. On the other hand, although this trend of underperformance of PCR-SCaDS can also be found in the 50% error on **y** conditions, it is not as pronounced.

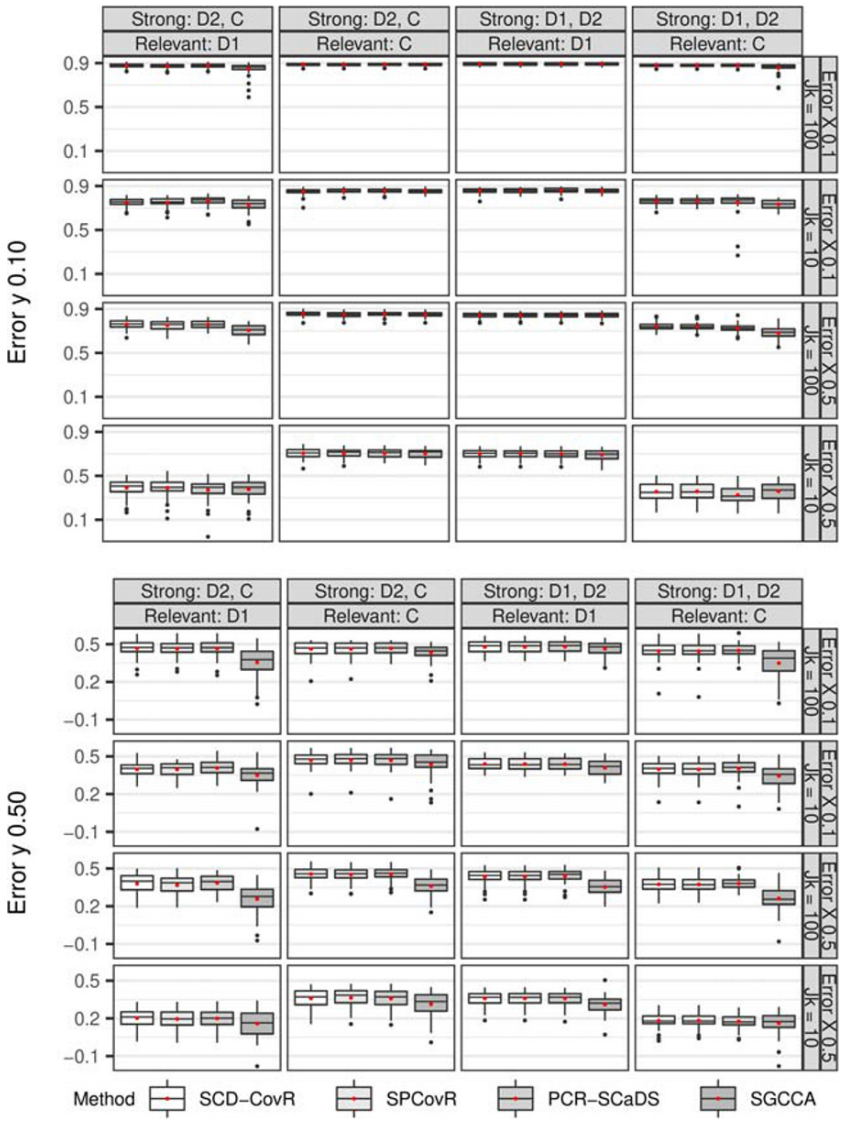

FIGURE 4 Box plots of the out-of-sample R^2 when three components are extracted; each panel corresponds to one of the 16 conditions. The column panels indicate the manipulated strength and relevance of the three components; D1 and D2 denote the components distinctive to block 1 and 2, respectively, whereas C refers to the common component. The row panels indicate the number of variables Jk in each predictor block. PCR-SCaDS, principal component regression followed by sparse common and distinctive simultaneous component analysis, SCD-CovR, sparse common and distinctive covariates regression; SGCCA, sparse generalized canonical correlation analysis; SPCovR, sparse principal covariates regression

SGCCA is more sensitive to whether the relevant component is strong or weak; when a strong component is relevant, the method has comparable out-of-sample R^2 with the other three methods. However, for datasets, where the weak component is relevant, SGCCA outperforms the other methods. SCD-CovR and SPCovR outperform PCR-SCaDS with respect to prediction in all conditions; they perform similar to or a bit better than SGCCA in terms of prediction when the strong component is also the relevant one, but SGCCA has better predictive performance when the relevant component is a weak component. The underperformance of PCR-SCaDS is not a surprising outcome because it only considers the predictor variables in constructing the components. Therefore, the variance explained in **y** by a weak but relevant component is not effectively captured by the method, because it extracts the two strong though irrelevant components.

Figure 3 summarizes the out-of-sample R^2 obtained when each of the methods extracted three components. SGCCA appears to stand out with a slightly lower out-of-sample R^2 on average, whereas the other three methods show very similar performance. Figure 4 shows the results laid out in function of the factors.

In most of the conditions in Figure 4, we can observe the trend conveyed in Figure 3: SGCCA shows a lower level of out-of-sample R^2, whereas the other three methods perform comparably. The underperformance of SGCCA is clearer in the conditions in which the proportion of error in **y** is 50%. This result can be attributed to overfitting: for these conditions where SGCCA showed low levels of R^2, the residuals (in-sample errors) were considerably smaller than the prediction error computed with the out-of-sample observation of **y**. On the other hand, the two different types of errors were comparable for the three other methods. In contrast to Figure 2 with two-component models, the prediction

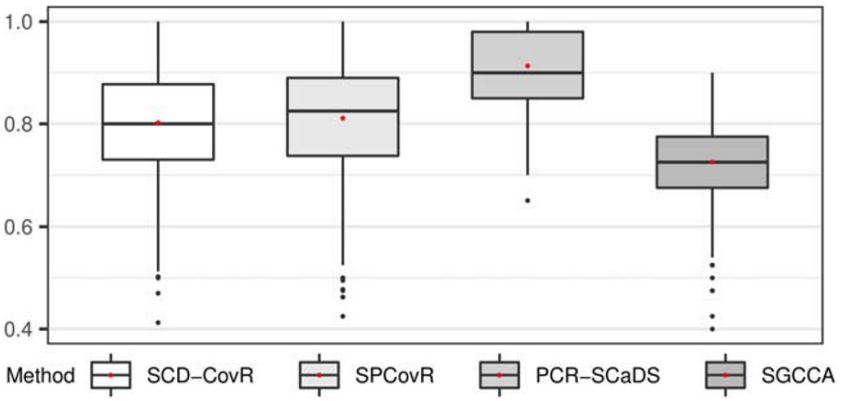

FIGURE 5 Box plots of the correct classification rate when two components are extracted: aggregated results. The red dot indicates the mean. PCR-SCaDS, principal component regression followed by sparse common and distinctive simultaneous component analysis, SCD-CovR, sparse common and distinctive covariates regression; SGCCA, sparse generalized canonical correlation analysis; SPCovR, sparse principal covariates regression

quality of PCR-SCaDS is similar with the one shown by SCD-CovR and SPCovR. This is reasonable, as in this setup, where all three underlying components are extracted, PCR-SCaDS is able to extract the relevant but weak component.

In conclusion, the results for the out-of-sample R^2 show that SCD-CovR yields a relatively high quality of prediction. When two components are extracted, it outperforms PCR-SCaDS, whereas with three extracted components, the method results greater R^2 than SGCCA. Additionally, the performance of SPCovR is comparable with that of SCD-CovR. It should be noted, however, that when not all components are extracted and there is a weak component that is relevant for prediction, then SGCCA is the preferred method in terms of prediction.

3.4.2 | Correct classification rate

Figures 5 and 6 present the results of the correct classification rate. In Figure 5, which pertains to the analyses with two extracted components, PCR-SCaDS yields the highest rate of weights correctly classified as zero or nonzero, closely followed by SCD-CovR and SPCovR. SGCCA has a considerably lower correct classification rate. SCD-CovR, SPCovR, and PCR-SCaDS again show comparable and high correct classification rates also when three components were extracted (Figure 6), where SGCCA underperforms again. This general trend seen in Figures 5 and 6 is largely consistent across conditions.

The outperformance of PCR-SCaDS and SCD-CovR is sensible. On top of the lasso penalty that induces sparsity, these methods also constrain the weights such that an entire set of weights belonging to a predictor block are made sparse. When three components are extracted, the oracle information of the common and distinctive component structure is provided which further eases the correct classification. In contrast, SPCovR and SGCCA do not explicitly cater for capturing common and distinctive processes and thus are expected to show a diminished rate of correct classification. However, SPCovR resulted in a very similar level of performance as SCD-CovR, and this can be attributed to the usage of rational starting values based on PCovR. Because the predictor variables were generated with an underlying true unrotated structure of PCA, initializing the convergence with PCovR solutions helps SPCovR in correctly retrieving the weights.

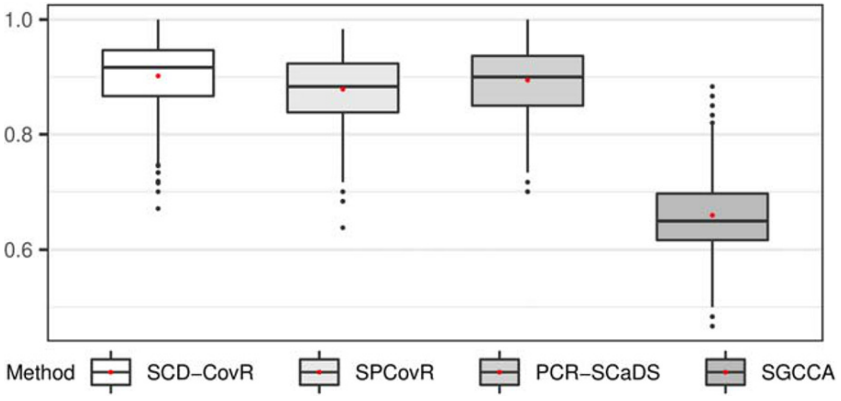

FIGURE 6 Box plots of the correct classification rate when three components are extracted: aggregated results. The red dot indicates the mean. PCR-SCaDS, principal component regression followed by sparse common and distinctive simultaneous component analysis, SCD-CovR, sparse common and distinctive covariates regression; SGCCA, sparse generalized canonical correlation analysis; SPCovR, sparse principal covariates regression

To conclude, the results from the correct classification rate suggest that SCD-CovR and SPCovR return weights that are of similar quality as those obtained with PCR-SCaDS which emphasizes the recovery of the weights more.

3.4.3 | Capturing common and distinctive components

On top of the prediction quality and the correct retrieval of sparse weights, SCD-CovR also targets another objective, namely to capture common and distinctive predictive processes. For each of the 1600 simulated datasets that the methods were administered to, we counted the number of common and distinctive components found by the methods. Regardless of the presence of zero block constraints, a column of the estimated $\mathbf{W}_C$ matrix that contains only zeroes for a predictor block is considered a distinctive component. Otherwise, when nonzero weights are found for both blocks, the component is a common component. For instances where an entire column is zero, the corresponding component is identified as neither common nor distinctive. Table 1 provides the numbers of these components (note that we generated all of the replicate datasets by a three-component model with two components distinctive to each predictor block and one common component).

Concerning analyses with two components where the zero block constraints are selected via cross-validation for SCD-CovR and PCR-SCaDS, it can be seen that almost all of the components found by PCR-SCaDS were distinctive. SCD-CovR identified about 41% of the estimated components as distinctive components. SPCovR and SGCCA, which do not impose an explicit constraint for the distinctive components, mostly identified common components, naturally. With respect to the three-component models, SCD-CovR and PCR-SCaDS with the oracle information on the common and distinctive structure returned the components reflecting the structure effectively. However, it can be seen that SCD-CovR provided a few more distinctive components than defined. These are instances where the lasso penalty sparsifies the weights corresponding to an entire predictor block, while the respective component is a common component. Although SPCovR and SGCCA do not provide sufficient numbers of distinctive components, SPCovR derived a lot more of those than in the two-component setting. Interestingly, the number of components did not appear to influence the effectiveness of SGCCA in capturing the common and distinctive components. Also, a component distinctive to the first predictor block was found much more frequently than the other distinctive component by SGCCA.

These numbers of retrieved common and distinctive components suggest that SCD-CovR is as effective as PCR-SCaDS with heavy emphasis on reconstructing the predictors when the correct common and distinctive structure is given. SPCovR that has similar performance with SCD-CovR at correct classification of the weights falls short at providing enough distinctive components, when the correct number of three components is used. This implies in practice that far more components extracted by SPCovR would be interpreted as a common process rather than a distinctive one, than the components derived using SCD-CovR. Evaluating the performance of the methods under

TABLE 1 Number of common and distinctive components considering the weights matrix

	SCD-CovR	SPCovR	PCR-SCaDS	SGCCA
Two-component model				
D1	666	101	1596	197
D2	641	138	1599	9
C	1893	2961	0	2994
Three-component model				
D1	1636	643	1601	200
D2	1601	840	1601	8
C	1563	3317	1595	4592

Note. D1 and D2 indicate components distinctive to block 1 and 2, respectively, and C refers to a common component. There were 1600 replicate datasets, and thus, the total numbers of estimated components for the analyses with two and three components were 3200 and 4800, respectively.

Abbreviations: PCR-SCaDS, principal component regression followed by sparse common and distinctive simultaneous component analysis, SCD-CovR, sparse common and distinctive covariates regression; SGCCA, sparse generalized canonical correlation analysis; SPCovR, sparse principal covariates regression.

two-component model is less straightforward than three-component model, because now, the methods have to summarize the structural variation governed by three true components by estimating only two. Methods can choose certain favorable components or may create composite components that combine multiple true components. In such cases, simply deriving more distinctive components perhaps does not directly link to outperformance. Although 50% of the replicate datasets were characterized by the common component being a strong component, PCR-SCaDS extracted only distinctive components. This indicates the method's strong inclination towards finding distinctive components. At the same time, although the other 50% of the datasets did not feature the common component being strong, a vast majority of the components retrieved by SPCovR and SGCCA were common components. This implies that these two methods favor common components. In contrast, 59% of the components retrieved by SCD-CovR were common components, and this appears to address the true component structure better than the other methods. To conclude, our results from two-component models suggest that SCD-CovR is more capable than the other methods in finding an adequate balance between common and distinctive components in reflecting the underlying component structure.

4 | ILLUSTRATIVE APPLICATION

In this section, we illustrate SCD-CovR by applying it to an empirical dataset. We also compare with results that are obtained with the related methods to examine the practical effectiveness of SCD-CovR.

4.1 | Dataset and preprocessing

We analyzed a dataset originally from Thybo et al[26] regarding texture measurements of potatoes. The dataset consists of 20 potato samples that were analyzed using three measurement platforms: chemical analysis, uniaxial compression, and sensory analysis. The chemical analysis block contains 14 variables regarding chemical aspects of the potatoes, such as the chemical composition. The uniaxial compression block with 36 variables provides measurements obtained from administering uniaxial compression at six deformation rates on cooked potato samples. The sensory analysis block is composed of nine sensory variables reported by trained experts. Here, we conduct SCD-CovR with the aim to predict the sensory experience, while also exploring the underlying common and distinctive predictive processes in the chemical and uniaxial compression blocks.

To this end, we constructed a univariate criterion from the sensory analysis data block by extracting the first principal component. All variables were first centered and scaled to unit sum of squares. Next, in order to account for the differing size of the two predictor data blocks, we scaled these blocks so that the sum of squares of each data block is equal. We administered SCD-CovR along with the three related methods employed in the simulation study to assess the performance of the methods when being applied to an empirical dataset.

4.2 | Model selection

The model selection strategy for this empirical dataset was largely in line with the strategy used in the simulation study, applying the same tuning sequence. However, the true number of components as well as their status (common, distinctive for block one or two) and the level of sparseness were unknown in this setting. We found the number of components through a residual test where we observe the change of sum of squared residuals $\|\mathbf{y}-\widehat{\mathbf{y}}\|_2^2$ (where $\mathbf{y}$ and $\widehat{\mathbf{y}}$ indicate the observed criterion and the fitted values, respectively) while increasing the number of components. For the test, we fixed the ridge and lasso penalties λ_R and λ_L to 0.01 (to account for high dimensionality) and 0, respectively. As the common and distinctive structures of the model may interact with the number of components needed, we included all the possible combinations of the common and distinctive components in the residual test. Concerning the weighting parameter α, we used the maximum likelihood approach discussed in Vervloet et al.[19] The following formula was used:

$$\alpha_{ML} = 1 - \frac{\|\mathbf{X}_C\|_2^2}{\|\mathbf{X}_C\|_2^2 + \|\mathbf{y}\|_2^2 \frac{\sigma^2_{\mathbf{E}^{(X)}}}{\sigma^2_{\mathbf{e}^{(y)}}}}, \tag{10}$$

where $\sigma^2_{\mathbf{E}^{(X)}}$ and $\sigma^2_{\mathbf{e}^{(y)}}$ refer to the error variances to be estimated (see Vervloet et al.[27] for details). The results from the residual test are shown in Figure B1. Within each number of components, models comprised mostly of distinctive components resulted in larger sums of residuals. However, when observing the overall trend, the sum of squared residuals decreases sharply at three components independently of the common and distinctive structures. The sum of residuals then stabilizes with subsequent numbers of components. The residual test using the aforementioned tuning parameters therefore resulted in the choice of three components. In order to make the method comparison fair, we also used three components when applying the other methods.

Given this number of components, we used the same model selection procedure as in the simulation study. This procedure consists of conducting cross-validation for α and λ_R simultaneously, followed by cross-validation for the zero block constraints. Both procedures employed 10-folds. The one SE rule was adopted for α and λ_R but not for the zero block constraints. Out of the three different configurations of zero block constraints, which resulted in similar levels of cross validation error, (D1,D2,C), (C,C,C), and (D2,C,C), the configuration with the smallest error, (D1, D2, C) was selected (Figure B2). We acknowledge that it is hard to tell which of these three structures is the true underlying common and distinctive structure, however. Because the oracle level of sparsity is unavailable for this empirical example, λ_L was determined through 10-fold cross-validation with the one SE rule as well (Figure B3). The plots that depict the cross-validation errors and the corresponding SEs can be found in Appendix B.

With regards to SPCovR, we adopted the same number of components, α and λ_R as used for SCD-CovR. The lasso penalty λ_L was chosen through 10-fold cross-validation with the one SE rule (Figure B4). For PCR-SCaDS, the procedures from the simulation study were taken (Figure B5, B6). λ_L was determined through 10-fold cross-validation with the one SE rule (Figure B7). Lastly, SGCCA only needs tuning of the lasso penalty governing the level of sparsity, this penalty was tuned via 10-fold cross-validation with the one SE rule as well (Figure B8). The plots in Appendix B can be consulted for the cross-validation results.

4.3 | Results

The four methods were administered with the tuning parameters in Table 2. The table also provides the R^2 values of each method, calculated by $R^2 = 1 - \left(\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2 / \|\mathbf{y}\|_2^2 \right)$, where $\mathbf{y}$ and $\hat{\mathbf{y}}$ indicate the observed criterion and the fitted values, respectively.

The R^2 values are very high, except for PCR-SCaDS. In order to also test for out-of-sample prediction quality, we conducted 10-fold cross-validation. The results can be seen in Figure 7 and are in agreement with those found in our simulation study. SCD-CovR and SPCovR produced less cross-validation errors than PCR-SCaDS and SGCCA; cross-validation error is comparable with prediction error. Inspecting the weights matrix produced by the two outperforming methods, we found that SPCovR produced two common components and one component distinctive to the chemical block, whereas SCD-CovR found one common component and one distinctive component for each predictor block. It is difficult to determine which of the both solutions is more interpretable, but this finding indicates that SCD-CovR is capable of capturing more distinctive components than SPCovR while providing competitive quality in prediction.

For interpretation of the final SCD-CovR model, we can first study the retrieved sparse weights matrix (Table C1 in Appendix C). It displays that the resulting weights matrix is very sparse; there are only 7, 5, and 4 nonzero weights that

TABLE 2 Tuning parameters and R^2 per method

	R	α	Ridge	Lasso	Block	R^2
SCD-CovR	3	0.7	0.005	3.579	C, D1, D2	0.981
SPCovR	3	0.7	0.005	5.477	NA	0.933
PCR-SCaDS	3	NA	0.001	0.011	D1, D2, D2	0.663
SGCCA	3	NA	NA	0.277	NA	0.954

Note. "Block" refers to the zero block constraints.

Abbreviations: PCR-SCaDS, principal component regression followed by sparse common and distinctive simultaneous component analysis, SCD-CovR, sparse common and distinctive covariates regression; SGCCA, sparse generalized canonical correlation analysis; SPCovR, sparse principal covariates regression.

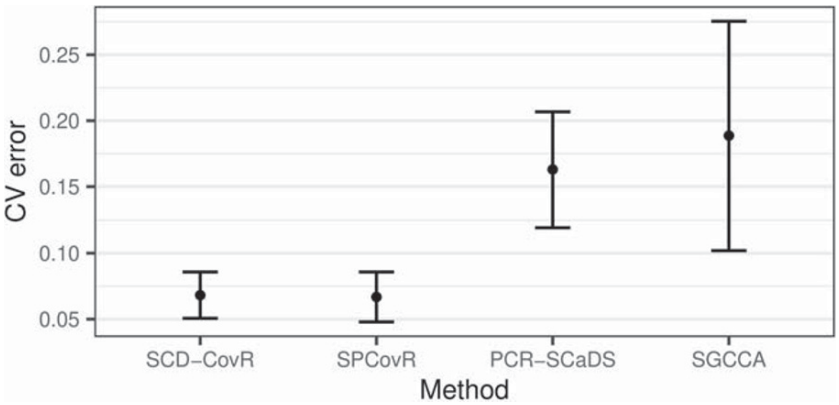

FIGURE 7 Cross-validation error and the corresponding standard error of the four methods. PCR-SCaDS, principal component regression followed by sparse common and distinctive simultaneous component analysis, SCD-CovR, sparse common and distinctive covariates regression; SGCCA, sparse generalized canonical correlation analysis; SPCovR, sparse principal covariates regression. CV, cross-validation

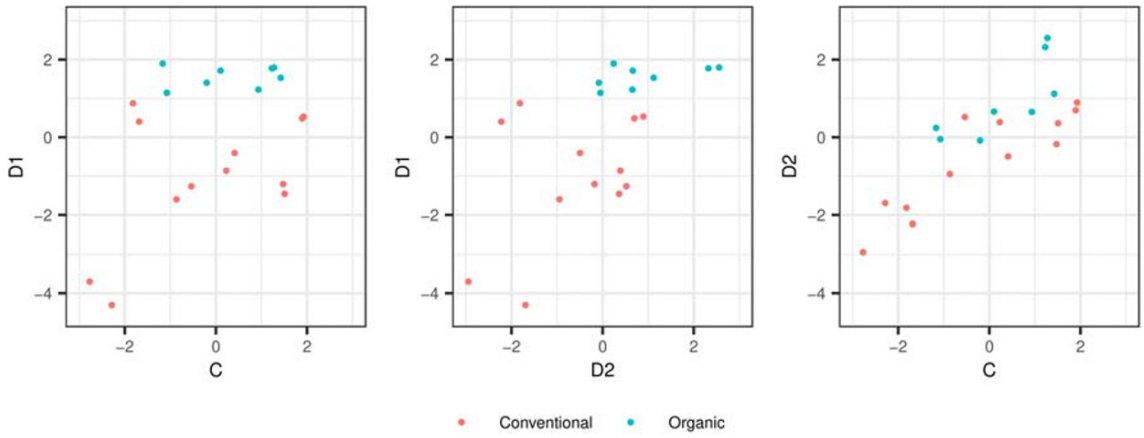

FIGURE 8 Component scores of the potato samples. The two types of potato samples are displayed in different colors. C, D1, and D2 indicate the type of the components (i.e., D1 refers to the component distinctive to the first predictor block that is the chemical analysis)

correspond to the three components respectively. As dictated by the tuned zero-block constraints, the weights matrix contains nonzero coefficients from both predictor blocks only in the column that corresponds to the common component.

We further investigated the model by inspecting Figure 8. This figure plots the component scores of the potato samples. Out of the 20 potato samples, 12 were grown conventionally, and 8 were grown organically. Although this information was not incorporated when fitting the model, the two types can be clearly distinguished using the two distinctive components. Therefore, these components found by SCD-CovR not only are capable of predicting the response variable but also reveal existing structural variation. In summary, the exploration of the final model shows that the method is able to fulfill its aims. It retrieves common and distinctive components that are sparse and thus more interpretable. The components also adequately explain the variance in both response and the predictors.

5 | DISCUSSION

Data originating from multiple sources can be analyzed with several objectives: prediction of a criterion, selection of relevant variables, and uncovering the common and distinctive underlying mechanisms. We proposed SCD-CovR to address these three aims simultaneously.

Through a simulation study incorporating multiple evaluation criteria that reflect these aims, we demonstrated that SCD-CovR outperforms three related methods that serve a subset of these goals; SPCovR, PCR-SCaDS, and SGCCA. Our method resulted in better prediction than PCR-SCADS and was also more effective than SGCCA for prediction under certain conditions. The coefficients retrieved by SCD-CovR better reflected the true underlying coefficients than those found by SGCCA. Lastly, with respect to finding common and distinctive processes, the method outperformed

SPCovR and SGCCA in capturing the block structure of common and distinctive components. We further illustrated this comparative advantage of SCD-CovR by reanalyzing a publicly available empirical dataset. The SCD-CovR cross-validation error was lower than that of PCR-SCaDS and SGCCA. At the same time, SCD-CovR retrieved more distinctive components than SPCovR.

These results provide further insight into the strengths of our proposed method. The outperformance in prediction compared with PCR-SCaDS reiterates previous comparisons of PCovR and PCR.[24,28] Deriving components while taking the criterion into account is more effective for prediction, than adopting a two-step approach of first constructing the components and then subsequently using them for prediction. Similarly, PLS methods have been found to be more prone to overfitting than PCovR methods,[8] and our outcome of the simulation study shows the same, with SCD-CovR yielding better out-of-sample prediction under several conditions. Moreover, SPCovR and SCD-CovR being more effective than SGCCA exhibit the benefits of the weighting parameter α. It enables a good balance between focusing on the predictors or the criterion, whereas SGCCA emphasizes the criterion more strongly. Our results are all based on α values established through cross-validation and thus indicate the effectiveness of the weighting parameter even within a data-driven approach. Lastly, concerning the identification of common and distinctive components, the simulation results from the three-component models illustrate the outperformance of SCD-CovR when the zero block constraints are correctly specified. This implies that the method can be especially effective when supported by an adequate model selection strategy.

Our proposed method also comes with some weaknesses. Model selection is an obvious challenge. As the method is devised to serve multiple aims, it involves many parameters to be tuned. The weighting parameter α, the number of components, the common and distinctive component structure, and the penalization parameters are all influential, and the retrieved model heavily depends on the choice of these parameters. Furthermore, identifying and discerning common and distinctive processes for data fusion methods is a very complicated task as it often interacts with other aspects such as the number of components.[12] In the same vein, the weighting parameter α involved with PCovR is also difficult to tune.[24] However, as the current paper focuses more on the proposal and the illustration of the new SCD-CovR method, this intricate problem of model selection has not been extensively addressed.

The examples presented in the current study only concern a scenario with two data blocks, but it is possible to extend our method to a situation with more blocks. In that case, a component that is constructed by predictors from a single data block would be defined as a distinctive component. Components pertaining to predictors from multiple but not all blocks would be called partially or locally common, as opposed to globally common components that involve predictors from all of the data blocks. These terminologies are in line with the previous literature such as Måge et al.[12] In such data circumstances, the challenge of model selection would involve heavy computational burden because our method caters for capturing of common and distinctive underlying processes by means of the prespecified zero block constraints. Given K data blocks and R components, no less than $\binom{(2^K-1)+R-1}{R}$ different zero block constraints should be evaluated. Considering that the method also involves several other parameters for retrieving the sparse solutions, the model selection procedure becomes a particularly intensive task.

As it holds for many other methods that rely on the lasso and elastic net penalties to attain sparsity, SCD-CovR is not free from the shortcoming that nonzero coefficients may be overly shrunken towards zero. Alternatives have been proposed, including the adaptive lasso[29] and the SCADS penalty,[30] which apply different degrees of shrinkage depending on the value of the coefficients. Stability selection[31] is another effective method for variable selection that does not shrink the nonzero coefficients. However, some degree of shrinkage of the nonzero coefficients may be beneficial in terms of bias-variance tradeoff as it helps to stabilize the OLS estimates.[32]

There are several future directions that the method can extend towards. Handier solutions to retrieve the distinctive components such as the Group lasso penalty can be adopted to greatly relieve the computational demand of the zero block constraints. Gu and Van Deun[33] have implemented the Group lasso to find distinctive components within the multiblock sparse PCA setting, and this could be one of the possible future directions in extending the SCD-CovR method. Another natural extension is to allow multiple criterion variables, as the current method only entails the univariate regression problem. Furthermore, the method can be adapted to incorporate more diverse structures of underlying processes. The current simulation study assumes that the data generating model follows the properties of PCA where the weights and the loadings are equal. However, true structures where this equality does not hold may exist. It would be interesting to examine the applicability of the method within such circumstances, as both weights and loadings would need to be considered for interpretation. Similarly, our proposed method only enforces sparsity in the weights, but the true structure may also include sparse loadings. Looking further into these other possible models

where loadings or both weights and loadings are sparse can also be a plausible direction in devising a predictive method that is more interpretable, in a modern multiblock setting.

ACKNOWLEDGEMENTS

This research was funded by a personal grant from the Netherlands Organisation for Scientific Research [NWO-VIDI 452.16.012] awarded to Katrijn Van Deun. The funder did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. We thank the anonymous reviewers for providing their valuable comments and suggestions on improving the manuscript.

ORCID

Soogeun Park https://orcid.org/0000-0001-8302-9690

REFERENCES

1. Nakaya HI, Wrammert J, Lee EK, et al. Systems biology of vaccination for seasonal influenza in humans. *Nat Immunol*. 2011;12(8):786.
2. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biol*. 2017;18(1):83.
3. Gallagher J, Hudgens E, Williams A, et al. Mechanistic indicators of childhood asthma (mica) study: piloting an integrative design for evaluating environmental health. *BMC Public Health*. 2011;11(1):344.
4. Van Deun K, Thorrez L, Coccia M, et al. Weighted sparse principal component analysis. *Chemomet Intell Lab Syst*. 2019;195:103875.
5. Westerhuis JA, Coenegracht PierreMJ. Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares. *J Chemomet J Chemomet Soc*. 1997;11(5):379-392.
6. Kreutzmann S, Svensson VT, Thybo AK, Bro R, Petersen MA. Prediction of sensory quality in raw carrots (Daucus carota l.) using multi-block ls-parpls. *Food Qual Prefer*. 2008;19(7):609-617.
7. Måge I, Menichelli E, Næs T. Preference mapping by PO-PLS: separating common and unique information in several data blocks. *Food Qual and Prefer*. 2012;24(1):8-16.
8. Van Deun K, Crompvoets EAV, Ceulemans E. Obtaining insights from high-dimensional data: sparse principal covariates regression. *BMC Bioinform*. 2018;19(1):104.
9. LêCao K-A, Rossouw D, Robert-Granié C, Besse P. A sparse PLS for variable selection when integrating omics data. *Stat Appl Gen Molecul Biol*. 2008;7(1):1-29.
10. Schouteden M, Van Deun K, Pattyn S, Van Mechelen I. SCA with rotation to distinguish common and distinctive information in linked data. *Behav Res Methods*. 2013;45(3):822-833.
11. Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and individual variation explained (jive) for integrated analysis of multiple data types. *Ann Appl Stat*. 2013;7(1):523.
12. Måge I, Smilde AK, van der Kloet FM. Performance of methods that separate common and distinct variation in multiple data blocks. *J Chemometrics*. 2019;33(1):e3085.
13. Smilde AK, Måge I, Naes T, et al. Common and distinct components in data fusion. *J Chemometrics*. 2017;31(7):e2900.
14. Van Deun K, Wilderjans TF, Van den Berg RA, Antoniadis A, Van Mechelen I. A flexible framework for sparse simultaneous component based data integration. *BMC Bioinformatics*. 2011;12(1):448.
15. Gu Z, Van Deun K. A variable selection method for simultaneous component based data integration. *Chemometrics Intel Lab Syst*. 2016;158:187-199.
16. deSchipper N, VanDeun K. Revealing the joint mechanisms in traditional data linked with big data. *Zeitschrift für Psychologie*. 2018;226(4):212–231.
17. Tenenhaus A, Philippe C, Guillemot V, Le Cao K-A, Grill J, Frouin V. Variable selection for generalized canonical correlation analysis. *Biostatistics*. 2014;15(3):569-583.
18. De Jong S, Kiers HAL. Principal covariates regression: part I. theory. *Chemometrics Intel Lab Syst*. 1992;14(1-3):155-164.
19. Vervloet M, Van Deun K, Van den Noortgate W, Ceulemans E. On the selection of the weighting parameter value in principal covariates regression. *Chemometrics Intel Lab Syst*. 2013;123:36-43.
20. Johnstone IM, Lu AY. On consistency and sparsity for principal components analysis in high dimensions. *J Am Stat Assoc*. 2009;104(486):682-693.
21. Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. *J Comput Graph Stat*. 2006;15(2):265-286.
22. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B (Stat Methodol)*. 2005;67(2):301-320.
23. Kiers HAL, ten Berge JMF. Alternating least squares algorithms for simultaneous components analysis with equal component weight matrices in two or more populations. *Psychometrika*. 1989;54(3):467-473.
24. Vervloet M, Van Deun K, Van den Noortgate W, Ceulemans E. Model selection in principal covariates regression. *Chemometrics Intel Lab Syst*. 2016;151:26-33.
25. Smilde AK, Westerhuis JA, Boque R. Multiway multiblock component and covariates regression models. *J Chemometrics: J Chemometrics Soc*. 2000;14(3):301-331.
26. Thybo AK, Bechmann IE, Martens M, Engelsen SB. Prediction of sensory texture of cooked potatoes using uniaxial compression, near infrared spectroscopy and low field1h NMR spectroscopy. *LWT-Food Sci Technol*. 2000;33(2):103-111.

27. Vervloet M, Kiers HAL, Van den Noortgate W, Ceulemans E. Pcovr: An R package for principal covariates regression. *J Stat Softw*. 2015; 65(8):1-14.
28. Heij C, Groenen PJF, van Dijk D. Forecast comparison of principal component regression and principal covariate regression. *Comput Stat Data Anal*. 2007;51(7):3612-3625.
29. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc*. 2006;101(476):1418-1429.
30. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc*. 2001;96(456):1348-1360.
31. Meinshausen N, Bühlmann P. Stability selection. *J R Stat Soc Ser B (Stat Methodol)*. 2010;72(4):417-473.
32. Breiman L. Better subset regression using the nonnegative garrote. *Technometrics*. 1995;37(4):373-384.
33. Gu Z, Van Deun K. Regularized SCA: regularized simultaneous component analysis of multiblock data in R. *Behav Res Methods*. 2019; 51(5):2268-2289.

How to cite this article: Park S, Ceulemans E, Van Deun K. Sparse common and distinctive covariates regression. *Journal of Chemometrics*. 2020;e3270. https://doi.org/10.1002/cem.3270

APPENDIX A: ALTERNATING LEAST SQUARES FOR SCD-COVR

As given in Section 2, the objective function to be minimized is

$$L\left(\mathbf{W}_C, \mathbf{P}_C^{(X)}, \mathbf{p}_C^{(y)}\right) = \alpha \frac{\left\|\mathbf{y} - \mathbf{X}_C \mathbf{W}_C \mathbf{p}_C^{(y)^T}\right\|_2^2}{\|\mathbf{y}\|_2^2} + (1-\alpha) \frac{\left\|\mathbf{X}_C - \mathbf{X}_C \mathbf{W}_C \mathbf{P}_C^{(X)^T}\right\|_2^2}{\|\mathbf{X}_C\|_2^2} + \lambda_L |\mathbf{W}_C|_1 + \lambda_R \|\mathbf{W}_C\|_2^2, \tag{A1}$$

such that $\left(\mathbf{P}_C^{(X)}\right)^T \mathbf{P}_C^{(X)} = \mathbf{I}_R$, $\lambda_L, \lambda_R \geq 0$, $\alpha \geq 0$ and zero block constraint on $\mathbf{W}_C$.

The solutions are found through an alternating procedure where the objective is minimized with regards to $\mathbf{P}_C^{(X)}$ and $\mathbf{p}^{(y)}$ conditional on a fixed value of $\mathbf{W}_C$ and vice versa. The procedure iterates until a convergence criterion is met. Many methods that attain sparse solutions from PCA through regularization penalty have adopted this approach to find the solutions.[8,16,21] The procedure for SCD-CovR is similar to these methods, but the minimization with respect to $\mathbf{P}_C^{(X)}$ and $\mathbf{p}^{(y)}$ given $\mathbf{W}_C$ is slightly different. The loadings $\mathbf{P}_C^{(X)}$ are obtained via an analytical solution; $\mathbf{P}_C^{(X)} = \mathbf{U}\mathbf{V}^T$ where $\mathbf{U}$ and $\mathbf{V}$ are found through singular value decomposition of $\mathbf{X}_C^T \mathbf{X}_C \mathbf{W}_C = \mathbf{U}\mathbf{D}\mathbf{V}^T$. The regression coefficients $\mathbf{y}$ are given by the ridge regression estimates; $\mathbf{p}^{(y)} = \left(\mathbf{X}_C^T \mathbf{X}_C + \lambda_R \mathbf{I}\right)^{-1} \mathbf{X}_C^T \mathbf{y}$, where $\mathbf{I}$ is a $\left(\sum_k^K J_k\right) \times \left(\sum_k^K J_k\right)$ identity matrix and λ_R is a ridge penalty. Conditional on these values, the weights $\mathbf{W}$ are found through the coordinate descent algorithm. The zero block constraint specifies the elements that will be put to zero to encourage the common and distinctive processes. The details on the conditional estimation of $\mathbf{W}$ given $\mathbf{P}_C^{(X)}$ and $\mathbf{p}^{(y)}$ can be found in de Schipper and Van Deun.[16]

APPENDIX B: MODEL SELECTION FOR THE ILLUSTRATIVE APPLICATION

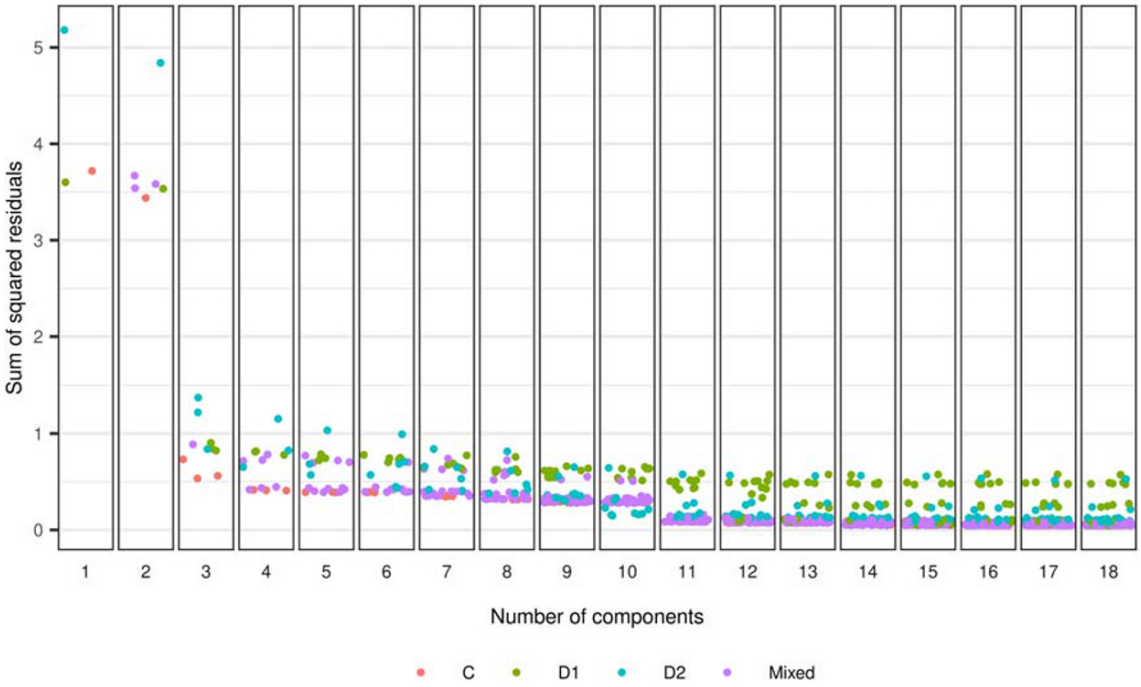

FIGURE B1 Sparse common and distinctive covariates regression (SCD-CovR): residual plot for determining the number of components. Each dot represents one model with a certain common and distinctive component structure. The colors indicate the type of component that occupies more than 65% of the total number of components in a model (e.g., when common components make up more than 65% of the total set of components, the model is colored red). When one particular type of component does not dominate the model, it is indicated by purple. D1 and D2 denote models dominated by components distinctive to block 1 and 2, respectively, whereas C refers to the common component

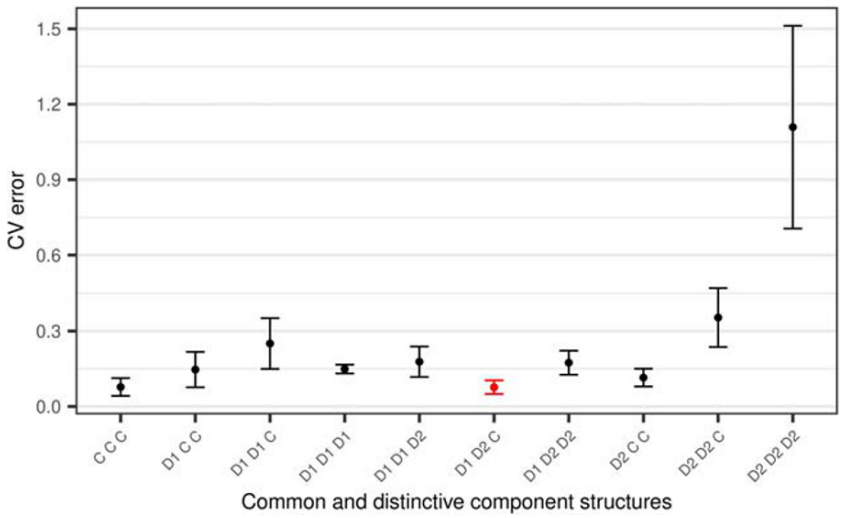

FIGURE B2 Sparse common and distinctive covariates regression (SCD-CovR): cross-validation error and corresponding standard error for zero block constraint for common and distinctive structure. D1 and D2 indicate components distinctive to block 1 and 2, whereas C denotes the common component. The selected zero block constraint with the smallest cross-validation error is displayed in red. CV, cross-validation

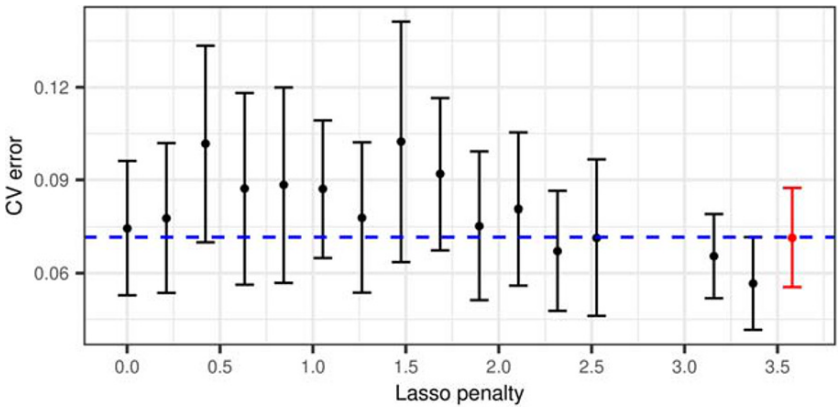

FIGURE B3 Sparse common and distinctive covariates regression (SCD-CovR): cross-validation error and corresponding standard error for the lasso penalty. The blue dashed line indicates the bound used for the one standard error (SE) rule, and the selected lasso value is shown in red. CV, cross-validation

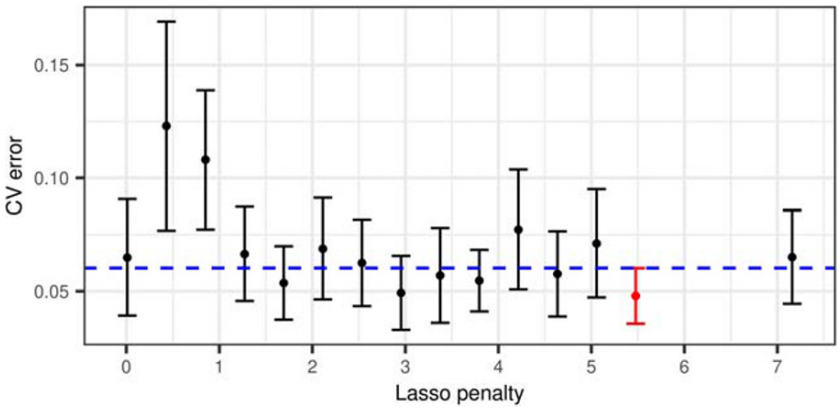

FIGURE B4 Sparse principal covariates regression (SPCovR): cross-validation error and corresponding standard error for the lasso penalty. The blue dashed line indicates the bound used for the one standard error (SE) rule and the selected lasso value is shown in red. CV, cross-validation

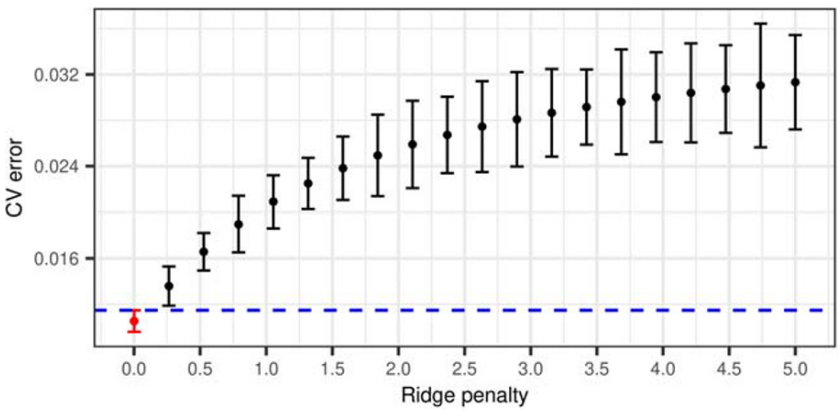

FIGURE B5 Principal component regression followed by sparse common and distinctive simultaneous component analysis (PCR-SCaDS): cross-validation error and corresponding standard error for the ridge penalty. The blue dashed line indicates the bound used for the one standard error (SE) rule and the selected ridge value is shown in red. CV, cross-validation

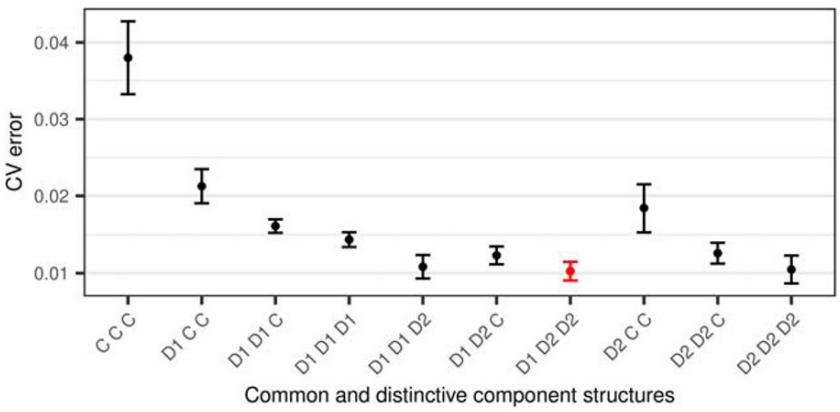

FIGURE B6 Principal component regression followed by sparse common and distinctive simultaneous component analysis (PCR-SCaDS): cross-validation error and corresponding standard error for zero block constraint for the common and distinctive structure. D1 and D2 indicate components distinctive to block 1 and 2, respectively, whereas C denotes the common component. The selected zero block constraint with the smallest cross-validation error is displayed in red. CV, cross-validation

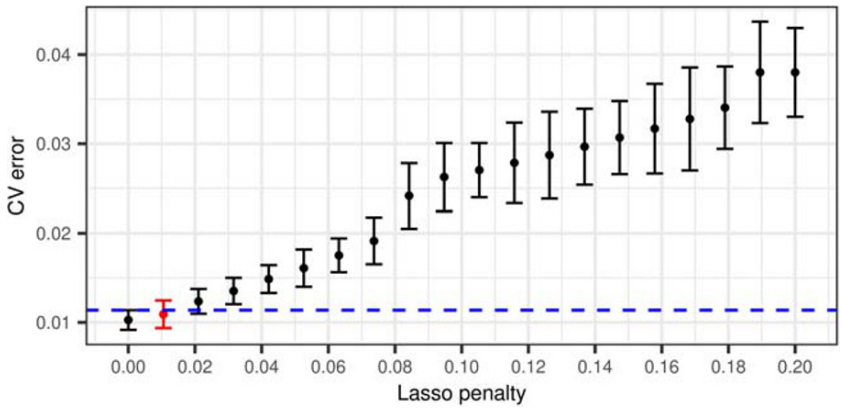

FIGURE B7 Principal component regression followed by sparse common and distinctive simultaneous component analysis (PCR-SCaDS): cross-validation error and corresponding standard error for the lasso penalty. The blue dashed line indicates the bound used for the one standard error (SE) rule, and the selected lasso value is shown in red. CV, cross-validation

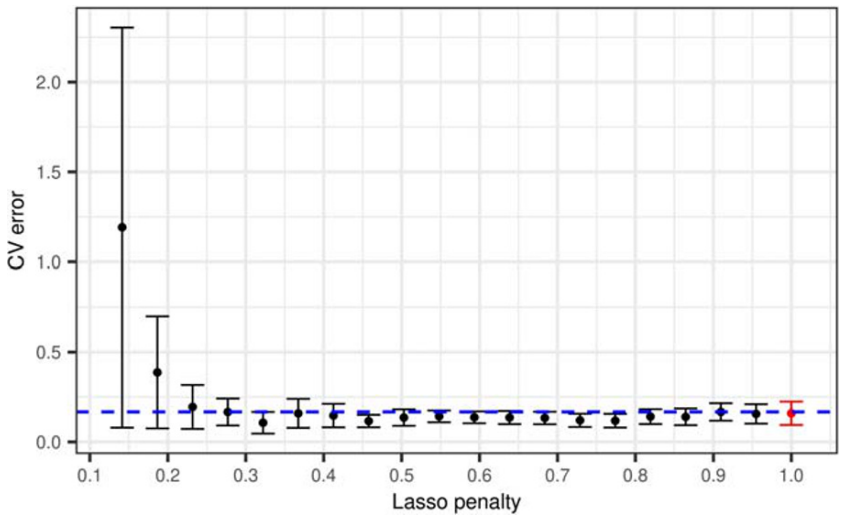

FIGURE B8 Sparse generalized canonical correlation analysis (SGCCA): cross-validation error and corresponding standard error for the lasso penalty. The blue dashed line indicates the bound used for the one standard error (SE) rule and the selected lasso value is shown in red. CV, cross-validation

APPENDIX C: ILLUSTRATIVE APPLICATION RETRIEVED WEIGHTS

TABLE C1 Weights retrieved by the final SCD-CovR model from the illustrative application

	C	D1	D2
Chemical analysis			
PEU	0	0	0
starch	0	-6.004	0
TotalN	0	0	0
phytic	0	-0.024	0
Ca	0	-0.494	0
Mg	0	0	0
Na	0	-0.003	0
K	-2.519	0	0
his1	0	0	0
his2	0.138	0	0
his3	0.306	0	0

(Continues)

TABLE C1 (Continued)

	C	D1	D2
his4	0	1.130	0
his5	0.480	0	0
his6	0	0	0
Uniaxial compression			
FractureWork20	0	0	0
BreakWork20	0	0	4.947
stressT20	0	0	0
strainH20	0	0	0
modulus20	0	0	0
slope20	0	0	0
FractureWor100	0	0	0
BreakWork100	0	0	0
stressT100	0	0	0
strainH100	0.133	0	0
modulus100	0	0	0
slope100	0	0	0
FractureWor250	0	0	0
BreakWork250	0	0	0
stressT250	0	0	0
strainH250	0	0	0
modulus250	0	0	1.766
slope250	0	0	0
FractureWor500	0	0	0
BreakWork500	0	0	0
stressT500	0	0	0
strainH500	0	0	0
modulus500	0	0	0
slope500	0	0	0
FractureWor750	0	0	0
BreakWork750	0	0	0
stressT750	0	0	0
strainH750	0	0	0
modulus750	0	0	1.663
slope750	0	0	0
FractureWor1000	0	0	0
BreakWork1000	5.758	0	0
stressT1000	0	0	0
strainH1000	0	0	0
modulus1000	0	0	1.104
slope1000	0.155	0	0

Note. The table above presents weights corresponding to the chemical analysis block, the one below corresponding to the uniaxial compression block.

Abbreviation: SCD-CovR, sparse common and distinctive covariates regression.