Machine Learning
https://doi.org/10.1007/s10994-024-06520-3

Variable selection for both outcomes and predictors: sparse multivariate principal covariates regression

Soogeun Park[1] · Eva Ceulemans[2] · Katrijn Van Deun[1]

Received: 1 December 2022 / Revised: 9 November 2023 / Accepted: 14 February 2024

Abstract

Datasets comprised of large sets of both predictor and outcome variables are becoming more widely used in research. In addition to the well-known problems of model complexity and predictor variable selection, predictive modelling with such large data also presents a relatively novel and under-studied challenge of outcome variable selection. Certain outcome variables in the data may not be adequately predicted by the given sets of predictors. In this paper, we propose the method of Sparse Multivariate Principal Covariates Regression that addresses these issues altogether by expanding the Principal Covariates Regression model to incorporate sparsity penalties on both of predictor and outcome variables. Our method is one of the first methods that perform variable selection for both predictors and outcomes simultaneously. Moreover, by relying on summary variables that explain the variance in both predictor and outcome variables, the method offers a sparse and succinct model representation of the data. In a simulation study, the method performed better than methods with similar aims such as sparse Partial Least Squares at prediction of the outcome variables and recovery of the population parameters. Lastly, we administered the method on an empirical dataset to illustrate its application in practice.

Keywords Outcome variable selection · Response variable selection · Response selection · Variable selection · Principal covariates regression · Dimension reduction

Editors: Michelangelo Ceci, João Gama, Jose Lozano, André de Carvalho and Paula Brito.

✉ Katrijn Van Deun
k.vandeun@tilburguniversity.edu

Soogeun Park
s.park_1@uvt.nl

[1] Tilburg University, Tilburg, The Netherlands

[2] KU Leuven, Leuven, Belgium

Published online: 17 July 2024

1 Introduction

Following the advancements of technology for data collection, most research disciplines are faced with challenges arising from an abundance of data. In deriving a prediction model, researchers are increasingly encountering a setting where they handle a bulk of data at both ends of predictor and outcome variables. For example, Stein et al. (2010) proposed a model that predicts the volume of each location of the brain (measured by large fMRI data) by numerous predictors from genome-wide association (GWAS) data. Specific genetic polymorphisms that are strongly associated with different parts of the brain were explored and identified therein. Similarly, Mayer et al. (2018) used genome-wide expression data in predicting responses of cell lines to several types of drugs. The study adopted random forests to find subsets of biologically meaningful associations between transcription rates and responses to drugs. Other examples include image-on-image regression in which an image is employed to predict another image (Guo et al., 2022), or multitrait GWAS where multiple correlated phenotypic traits are modelled together by genotypic variables (Kim et al., 2016; Oladzad et al., 2019). Studies that investigate associations between genes (Park & Hastie, 2008), or across protein and DNA (Zamdborg & Ma, 2009) are also along these lines.

Predictive modelling in the presence of such large amounts of data presents two well-known issues. First, a constructed prediction model with many variables is difficult to interpret due to the sheer number of coefficients; studying the predictor-outcome relationship becomes complicated. Second, certain predictor variables may be redundant. In a setting like the fMRI-GWAS study (Stein et al., 2010) where variables are collected without a specific research question, there is a need to screen out non-essential predictors that do not have any predictive power.

One way to deal with the two abovementioned problems pertaining to model interpretability and redundant predictor variables is the method of Principal Covariates Regression (PCovR; De Jong and Kiers, 1992. It is a combination of Principal Component Analysis (PCA) and Ordinary Least Squares (OLS) being applied in fields including chemometrics (Boqué & Smilde, 1999), material science (Helfrecht et al., 2020), health science (Taylor et al., 2019) and clinical psychology (Nelemans et al., 2019). PCovR introduces ‘principal covariates’; a low number of summary variables that condense the information in the large volume of predictor variables, akin to principal components in PCA. The outcome variable is then regressed on the principal covariates, significantly decreasing the number of regression coefficients to be estimated. However, since all of the predictor variables are involved in constructing the principal covariates, a large set of coefficients connecting the predictors with the covariates still has to be estimated. Understanding the nature of the covariates by inspecting these coefficients therefore becomes very cumbersome. To this end, PCovR has been extended to incorporate regularization penalties that induce sparseness in these coefficients (e.g. Van Deun et al., 2018; Park et al., 2020). This not only allows the covariates to be easily interpreted, but also discards the predictors that are redundant.

A related but rarely visited issue is that some of the outcome variables may also be redundant. They may not have a substantial relationship with any of the predictors, meaning that they cannot be predicted by the available sets of predictors. Such outcome variables are expected especially in the context of an exploratory research setup. For example, in a multitrait GWAS setup comparable to the aforementioned studies, not all phenotypes may have strong relationships with the available transcription rates. Researchers may want to identify a subset of phenotypes that are relevant to the available genetic predictors and obtain a more concise and interpretable prediction model.

Removal of unpredictable outcome variables in such cases offers benefits that extend beyond improved model interpretability. It can enhance the quality of predictions because these irrelevant outcomes may obscure relevant predictive relationships pertaining to other outcome variables (Fowlkes & Mallows, 1983; Steinley & Brusco, 2008). A redundant outcome may be mistakenly included in the model, which can impair out-of-sample prediction of relevant outcome variables, as well as the entire set of outcome variables. This concept can be seen analogous to how eliminating redundant predictors in a regression setting can treat overfitting and improve the predictive performance.

Settings that can benefit from the exclusion of irrelevant outcome variables are increasingly common these days with the growing number of investigations that incorporate non-targeted and naturally-occurring sources of data; prior information concerning predictor-outcome relationship is not available. Throughout the paper, we refer to these outcome variables that are not predictable by the given set of predictors as 'inactive outcomes', and otherwise as 'active outcomes'. This terminology is used in other papers that address outcome variable selection (Su et al., 2016; Hu et al., 2022).

While PCovR and its sparse extensions accommodate for issues arising from a large set of predictors, they are primarily designed to address a single outcome variable. Similarly, whereas methods designed to eliminate redundant predictor variables have been extensively studied (Tibshirani, 1996; Zou & Hastie, 2005; Yuan & Lin, 2006), regression problems involving variable selection at the level of outcome variables have not received much attention. There have been many approaches to regress multivariate outcome variables jointly on the predictors instead of modelling the outcomes individually, but most of these works were confined to identifying common sets of predictors that are important for predicting all of the outcome variables (Obozinski et al., 2006; Peng et al., 2010; Luo, 2020). Similarly, while multivariate methods such as Partial Least Squares (PLS) and Reduced Rank Regression (RRR) that have their basis on reducing the dimensionality of the variables have been extended to incorporate sparsity, the majority of these extensions have only targeted predictor variables (Lê Cao et al., 2011; Chung & Keles, 2010; Chen & Huang, 2012). To our knowledge, there has been only a handful of studies that target outcome variable selection; these include regularized regression approaches (An & Zhang, 2017; Hu et al., 2022), a sparse RRR method (Chen et al., 2012) and a method within a framework of envelope modelling (Su et al., 2016).

In this paper, we propose the method of Sparse Multivariate Principal Covariates Regression (SMPCovR), an extension of PCovR methodology that tackles the variable selection problem for both predictor and outcome variables. Starting from the PCovR model, sparseness is promoted in both sides of the model; in constructing the covariates from the predictors and in predicting the outcome variables based on the covariates. The resulting model is not only sparse and easy to interpret, but also eliminates redundant predictor variables and inactive outcome variables, from which improvement with respect to prediction can be expected. It contributes to the under-studied problem of variable selection of outcome variables.

The paper is arranged as follows. The next section provides methodological details of SMPCovR. We begin with a discussion of PCovR since it is the basis of our current method. A simulation study that comparatively evaluates SMPCovR along with other methods devised with similar research aims is presented afterwards. The method is also administered to an empirical dataset for an illustrative purpose, as well as to expand upon the comparison against competitive methods in a practical data setting. The paper concludes with a disussion. The R implementation of SMPCovR can be found on Github: https://github.com/soogs/SMPCovR. The code for generating the results in this paper is also available therein.

2 Methods

2.1 Notation

The following notation is used 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.e., $i \in \{1, 2, \dots, I\}$). Superscripts $^{(X)}$ and $^{(Y)}$ highlight affiliation with predictor and outcome variables, respectively. To denote estimates, a $\hat{}$ over the symbol denoting the population parameter is used. $\mathbf{X}$ refers to a $I \times J$ matrix containing the standardized scores of J predictors obtained from I observation units (that is, each column has mean zero and variance equal to one). $\mathbf{Y}$ denotes a $I \times K$ matrix of K continuous outcome variables that are mean-centered and scaled to variance equal to one, also observed on the same I observation units. The total number of covariates or components is denoted by R.

2.2 PCovR

We begin by discussing the method of PCovR and show how the method extends to the current method of SMPCovR. PCovR (De Jong & Kiers, 1992) is a combination of PCA and OLS. It models the predictor and outcome variables by using principal covariates which can be understood as summary variables. These covariates are linear combination of the predictors which are obtained such that they explain the variance in the predictor and outcome variables simultaneously. PCovR decomposes the predictors $\mathbf{X}$ and the outcome variables $\mathbf{Y}$ as follows:

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\mathbf{W}\mathbf{P}^{(Y)^T} + \mathbf{E}^{(Y)} \\ \mathbf{X} &= \mathbf{X}\mathbf{W}\mathbf{P}^{(X)^T} + \mathbf{E}^{(X)} \end{aligned} \tag{1}$$

where $\mathbf{W}$ denotes the weights matrix of size $J \times R$: the predictor variables are multiplied by the weights to construct principal covariates $\mathbf{T} = \mathbf{XW}$ with w_{jr} the weight corresponding to the jth predictor variable and the rth covariate. It can be seen that both $\mathbf{Y}$ and $\mathbf{X}$ are modelled on the basis of the covariates $\mathbf{XW}$. The first line of equation (1) is the model for the outcome variables: $\mathbf{P}^{(Y)}$ refers to the regression coefficients matrix of size $K \times R$ with $p_{rk}^{(Y)}$ the regression coefficient linking the rth covariate with the kth outcome variable. The residuals pertaining to the outcome variables are denoted by $\mathbf{E}^{(Y)}$. On the other hand, the second line of the equation gives the model for the predictors. $\mathbf{P}^{(X)}$ indicates the loadings matrix of size $J \times R$; $p_{rj}^{(X)}$ is the loading that connects the rth covariate with the jth predictor variable.

The following loss function is minimized when estimating the model parameters:

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

where $0 \leq \alpha \leq 1$ is a user-specified tuning parameter that expresses the balance between focussing on the reconstruction of predictors or the prediction of the outcome variables in deriving the covariates. With α specified as 0, the method boils down to PCA where principal components are found by only considering the predictors. When $\alpha = 1$, the method becomes equivalent to RRR (Izenman, 1975; Kiers & Smilde, 2007). Constraints are needed to identify a unique solution from (2); an orthonormality constraint is usually placed upon the covariates ($\mathbf{T}^T\mathbf{T} = (\mathbf{X}\mathbf{W})^T(\mathbf{X}\mathbf{W}) = \mathbf{I}$).

The principal covariates can be understood as underlying processes that explain the relation of the outcome variables to the predictor variables. Thus, it is often of research interest to interpret the constructed covariates. All of the parameter sets $\mathbf{W}$, $\mathbf{P}^{(X)}$ and $\mathbf{P}^{(Y)}$ can be studied as they offer insights from different angles. The weights matrix $\mathbf{W}$ provides the composition of the covariates as it prescribes how the predictor variables are combined to form the covariates. The loadings matrix $\mathbf{P}^{(X)}$ shows how the covariates recover back the predictors. Additionally, if the covariates are scaled to variance equal to one ($\mathbf{T}^T\mathbf{T} = I\mathbf{I}$), the loadings are equivalent to the correlation between the covariates and the predictors. Lastly, the regression coefficients $\mathbf{P}^{(Y)}$ represent how the covariates are used to predict the outcome variables. Unlike the weights and the loadings matrices, the regression coefficients concern the link between the covariates and the outcome variables.

2.3 SMPCovR

When large sets of predictor variables and outcome variables are present, inspecting the PCovR estimates to understand the nature of the covariates becomes difficult. Also, the dataset may present redundant predictors and inactive outcomes. The novel method of SMPCovR induces sparseness in the weights $\mathbf{W}$ and regression coefficients $\mathbf{P}^{(Y)}$ so that these issues are resolved within the context of PCovR.

2.3.1 Model and objective function

SMPCovR models the predictor and the outcome variables in the same manner as the PCovR model above yet with the additional constraint that only few variables make up the covariates and that not all outcome variables are predictable by (all) covariates. Such a sparse model can be attained by adding penalties to the objective expressed in (2):

$$\begin{aligned} L\left(\mathbf{W}, \mathbf{P}^{(X)}, \mathbf{P}^{(Y)}\right) = &\frac{\alpha}{\|\mathbf{Y}\|_2^2} \left\| \mathbf{Y} - \mathbf{X}\mathbf{W}\mathbf{P}^{(Y)^T} \right\|_2^2 + \frac{1-\alpha}{\|\mathbf{X}\|_2^2} \left\| \mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{P}^{(X)^T} \right\|_2^2 \\ &+ \sum_r^R \lambda_1 \|\mathbf{w}_r\|_1 + \sum_r^R \lambda_2 \|\mathbf{w}_r\|_2^2 \\ &+ \sum_r^R \gamma_1 \left\|\mathbf{p}_r^{(Y)}\right\|_1 + \sum_r^R \gamma_2 \left\|\mathbf{p}_r^{(Y)}\right\|_2^2 \end{aligned} \tag{3}$$

where the loadings associated with the predictors $\mathbf{P}^{(X)}$ are constrained to be column-orthogonal ($\mathbf{P}^{(X)^T}\mathbf{P}^{(X)} = \mathbf{I}$) in order to avoid trivial solutions with very small weights (close to zero) and very large loadings. Just as in the objective criterion for PCovR, the first and the second terms are sum of squares that concern the regression problem and the PCA problem, respectively. The two terms are balanced by specification of the α parameter ($0 \leq \alpha \leq 1$). Note that the constraint on the covariates employed for PCovR is removed for this objective criterion.

The terms with λ_1 and λ_2 respectively refer to the lasso and ridge penalties for the weights, while the terms with γ_1 and γ_2 indicate the lasso and the ridge penalties imposed on the regression coefficients. While the lasso penalty enforces the coefficients to zero and discards variables from the model, the incorporation of the ridge penalty prevents divergence occurring due to covariates being correlated. This combination of the lasso and ridge penalties is also known as the elastic net penalty (Zou & Hastie, 2005). The combination is necessary because the lasso penalty alone is shown to be inconsistent in the high-dimensional case, while the ridge penalty alone does not impose any sparsity. When all of the regression coefficients corresponding to an outcome variable are forced to zero, this outcome variable is modelled by zero and excluded from the model. Likewise, all of the weights corresponding to a predictor being penalized to zero removes the predictor variable from the model.

2.3.2 Algorithm

Estimates of the SMPCovR parameters can be obtained by alternating least squares. In turn, one of the parameter sets among $\mathbf{W}$, $\mathbf{P}^{(X)}$ and $\mathbf{P}^{(Y)}$ is estimated conditionally upon fixed values of the others. The elastic net problems for $\mathbf{W}$ and $\mathbf{P}^{(Y)}$ are convex problems, and they are both tackled via coordinate descent (Friedman et al., 2010). On the other hand, the conditional problem for $\mathbf{P}^{(X)}$ is known as an Orthogonal Procrustes Problem (Schönemann, 1966); it is not convex, but has a closed-form solution (Ten Berge, 1993). Since each of the estimation problems for $\mathbf{W}$, $\mathbf{P}^{(X)}$ and $\mathbf{P}^{(Y)}$ can converge at the global optimum of the conditional (penalized) least squares problem, the resulting alternating least squares procedure is monotonic. However, there is no guarantee of convergence to the global optimum for the combined problem (3), due to its non-convexity. To avoid local minima, we recommend to use multiple random starting values, along with rational starting values based on PCovR. Further details on the algorithm for minimizing the objective function can be found in Appendix 1, including the schematic outline of the algorithm and the derivation of solutions to the conditional updates (Appendices 2, 3, 4).

2.3.3 Model selection

The SMPCovR method entails the following list of tuning parameters that shape the model construction.

- Number of covariates R
- Weighting parameter α
- Lasso parameter concerning weights λ_1
- Ridge parameter concerning weights λ_2

- Lasso parameter concerning regression coefficients γ_1
- Ridge parameter concerning regression coefficients γ_2

We employ k-fold cross-validation (CV) as a standard model selection method for all of the tuning parameters except for the number of covariates R. Although a conventional model selection scheme with CV would consider all possible combinations of different values for all of the tuning parameters involved, such an exhaustive strategy would be computationally intensive, considering that the method is devised to cater for large sets of both predictor and outcome variables. Therefore, a sequential approach is adopted where the parameters are tuned in turn. Such a sequential approach has been shown to be a suitable model selection strategy for the methods that precede SMPCovR: PCovR and sparse PCovR (Vervloet et al., 2016; Park et al., 2020).

The number of covariates R is tuned as the first step of the sequential approach. PCA is performed on the concatenated data matrix $[\mathbf{Y} \quad \mathbf{X}]$ to find a suitable number of principal components. This number of components would be adopted as the number of covariates R for SMPCovR. A typical approach is the use of scree plot in which an 'elbow' is searched for from a plot that illustrates the amount of variance each principal component explains. However, since this location of the elbow can involve a subjective opinion, the acceleration factor technique (Raîche et al., 2013) is employed instead. It is an objective method that finds at which principal component the amount of explained variance changes most abruptly. The method retains the components that precede the component where the abrupt change in variance takes place (Cattell, 1966). It is along the same line as other strategies devised to objectively search for the elbow, such as the Convex Hull method (Wilderjans et al., 2013). We make use of the implementation in the R package "nFactors" (Raîche & Magis, 2020).

The subsequent step is to determine the values of α, γ_1 and γ_2 simultaneously via CV. In doing so, the number of covariates found in the previous step is used. Also, the parameters pertaining to the weights λ_1 and λ_2 are fixed at a small value of 10^{-7}. For the CV, we employ the R^2 measure computed from the CV test set to evaluate the out-of-sample prediction quality of the model parameters:

$$R^2_{\text{cv}} = 1 - \frac{\left\| \mathbf{Y}^{\text{test}} - \mathbf{X}^{\text{test}} \hat{\mathbf{W}} \hat{\mathbf{P}}^{(Y)^T} \right\|_2^2}{\left\| \mathbf{Y}^{\text{test}} \right\|_2^2}, \tag{4}$$

where $\mathbf{Y}^{\text{test}}$ and $\mathbf{X}^{\text{test}}$ refer to the outcome and predictor variables in the CV test set. We rely on the one standard error rule (1SE rule; Hastie et al., 2009) to select the final model after CV. Among the model configurations that fall within the 1 SE region from the maximum R^2_{cv}, the 1SE rule would favour the model with the lowest model complexity. Therefore, within the 1SE region, the models with the smallest α, the largest γ_1 and the smallest γ_2 values are selected. While smaller α values are linked with lower model complexity because they are more robust to overfitting than larger α, we found in our experiments that the combination of larger values of γ_1 and smaller values of γ_2 promote greater sparsity in the coefficients.

By using the selected values of α, γ_1 and γ_2, the parameters for the weights, λ_1 and λ_2, are tuned in the final stage of the procedure. This is because the impact of these parameters towards the model fit is relatively small, compared to the other parameters. In an investigation into different model selection procedures for sparse PCA, de Schipper and Van Deun (2021) reported that even a model with very sparse weights can result in good recovery of the true underlying component scores. Furthermore, methods that precede SMPCovR have also adopted this procedure to select the sparsity parameters for the weights at the final stage, resulting in good retrieval of the true model parameters (Park et al., 2020, 2023). Employing the 1SE rule again, models with the largest λ_1 and the lowest λ_2 values were selected, as they lead to more sparsity in the coefficients. A concrete demonstration of this model selection procedure for SMPCovR is provided in Sect. 4.1.2, where details of administering each of the steps of this procedure on an empirical dataset are presented.

2.3.4 Related methods

Our proposed method of SMPCovR accommodates three goals: (a) it is a prediction method for multiple continuous outcome variables, (b) it represents underlying predictive processes by covariates, and (c) it provides sparse coefficients and performs variable selection at both sides of predictor and outcome variables. This section compares SMPCovR to other methods that are devised with a similar set of aims.

Sparse PCovR Sparse PCovR (SPCovR; Van Deun et al., 2018) is an immediate predecessor of SMPCovR. The method finds sparse weights, but sparseness is not imposed to the regression coefficients $\mathbf{P}^{(Y)}$. In fact, SMPCovR without the lasso penalty on the regression coefficients boils down to SPCovR. Although the covariates can be found considering the multiple outcomes, an entire set of regression coefficients $\mathbf{P}^{(Y)}$ is estimated which is burdening for model interpretation. This also implies that inactive outcome variables are not filtered out from the model, and they may hinder the prediction quality.

Sparse Partial Least Squares (sPLS; Lê Cao et al., 2008; Chung and Keles, 2010) is a sparse extension to PLS, which is a well-known method in the same spirit as PCovR; it models predictor and outcome variables simultaneously by introducing summary variables (Wold, 1982; Wold et al., 1984). Just like in PCovR, these summary variables account for variance in both predictor and outcome variables. However, PLS does not incorporate the balancing parameter α. Although sPLS can model multiple outcome variables and performs variable selection for the predictors, it has not been extended to also enforce sparseness on coefficients that connect the summary variables with the outcome variables. However, outcome variable selection has been addressed within the framework of envelope modelling (Su et al., 2016). Envelope modelling[1] has been shown to be connected with PLS; the two methods target the same population parameters, but they differ in the method of estimation (Cook et al., 2013). Yet, the method in Su et al. (2016) is only designated for variable selection for the outcomes, and not for the predictor variables; the authors suggest a prior subset selection of predictor variables in the case of high dimensionality. Therefore, similarly to sPLS, the method does not address the complete set of goals of SMPCovR.

[1] Envelope modelling (Cook et al., 2010) is a recent branch of methods that identifies ‘material’ and ‘immaterial’ parts of predictor and outcome variables. A linear model is constructed only on the basis of the useful ‘material’ parts, which allows efficient estimation and overcomes problems such as collinearity.

3 Simulation study

We have conducted a simulation study in which we examine the performance of SMPCovR, SPCovR and sPLS with respect to the retrieval of underlying processes and the prediction of the multiple outcome variables. These underlying processes are specified by covariates that underlie the simulated data. The covariates were defined to only explain the variance for the subsets of predictor and outcome variables; other predictors and outcomes were defined to be redundant and inactive, respectively. We excluded the envelope method (Su et al., 2016) as it does not have a publicly available software implementation.

Owing to the sparsity penalty imposed upon the regression coefficients, we expect SMPCovR to outperform the other two methods in prediction when some outcome variables are inactive. By filtering out the inactive variables that are not related with the underlying covariates, overfitting of these inactive variables would be avoided. As a consequence, this would result in better prediction quality of the outcome variables overall, compared to SPCovR and sPLS.

Since the defined covariates underlie both predictor and outcome variables, the quality of retrieval of the underlying processes can be studied from two angles: (1) covariate-predictor relationships and (2) covariate-outcome relationships. With respect to the covariate-predictor relationships, it is anticipated that SMPCovR and SPCovR would show comparable performance because they are equipped with the same set of sparsity penalties on the weights. In contrast, sPLS is hypothesized to underperform as PLS-based methods have shown to be less effective in recovering the weights that prescribe the relationships between covariates and predictors (Park et al., 2020). On the other hand, it is natural that SMPCovR would provide better recovery of regression coefficients that represent the covariate-outcome relationships than the other two methods. Owing to the sparsity penalty imposed on the regression coefficients, SMPCovR would be able to discern between the important and unimportant covariate-outcome associations, while the other methods would only provide non-zero coefficients.

3.1 Design and procedure

Fixing the number of observations I to 100, the predictor variables were generated from an underlying model comprised of three covariates. While varying the number of outcome variables $\mathbf{Y}$ to be at either $K = 5$ or $K = 20$, we generated $J = 200$ predictor variables for the high-dimensional setting and $J = 30$ for the low-dimensional setting. The following setup was used.

$$
\begin{aligned}
&\mathbf{T} \sim \mathcal{MVN}(\mathbf{0}, \boldsymbol{\Sigma} = 50^2\mathbf{I}) \\
&\mathbf{E}^{(X)} \sim \mathcal{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_{E^{(X)}} = \sigma^2\mathbf{I}) \\
&\mathbf{E}^{(Y)} \sim \mathcal{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_{E^{(Y)}} = \sigma^2\mathbf{I}) \\
&\mathbf{X} \leftarrow \mathbf{T}\mathbf{W}^T + \mathbf{E}^{(X)} \\
&\mathbf{Y} \leftarrow \mathbf{T}\mathbf{P}^{(Y)^T} + \mathbf{E}^{(Y)}
\end{aligned}
\tag{5}
$$

$\mathbf{T}$ (size 100×3) is the covariate scores matrix which is generated from a multivariate normal distribution characterized by the mean vector $\boldsymbol{\mu} = \mathbf{0}$ and the diagonal covariance matrix $\boldsymbol{\Sigma}$ with diagonal elements fixed at 50^2. Therefore, the three covariates are the same

Table 1 Weights defined for the low-dimensional setup

1	2	3
0.5	0	0
0.5	0	0
0.5	0	0
0.5	0	0
0	0.354	0
0	0.354	0
0	0.354	0
0	0.354	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0.354	0
0	0.354	0
0	0.354	0
0	0.354	0
0	0	0.5
0	0	0.5
0	0	0.5
0	0	0.5
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0

size in variance and are uncorrelated. The weights matrix $\mathbf{W}$ (size $J \times 3$) is defined with 82% and 85% level of sparsity for low and high-dimensional setups, respectively. Furthermore, it is ensured that the columns of the weights matrix are orthogonal to each other ($\mathbf{W}^T\mathbf{W} = \mathbf{I}$; this constraint is not included in our objective function; it is used specifically for the data generation here). Since the covariates are defined to be uncorrelated, the model we use here can be seen as a PCA decomposition where the weights are equal to loadings. This is how $\mathbf{X}$ is defined by multiplying $\mathbf{T}$ and $\mathbf{W}$. The weights matrix defined for a low-dimensional setup can be seen in Table 1.

It can be seen that out of the 30 predictors in the low-dimensional setting, 14 predictors are redundant; they are not related with any covariates. In the high-dimensional setting, 110 predictors out of the 200 are defined as being redundant. Similarly, in specifying the regression coefficients $\mathbf{P}^{(Y)}$ (size $K \times 3$), 40% of the outcome variables are always defined as inactive; more details regarding the regression coefficients follow below.

$\mathbf{E}^{(X)}$ (size $100 \times J$) and $\mathbf{E}^{(Y)}$ (size $100 \times K$) denote the residual matrices corresponding to the predictor and outcome variables, respectively. They are drawn from multivariate normal distributions with zero mean vector and diagonal covariance matrices $\mathbf{\Sigma}_{E^{(X)}}$ and $\mathbf{\Sigma}_{E^{(Y)}}$, respectively. The two residual matrices are generated such that they are uncorrelated with each other, and also with the covariate scores. The variance of the residual matrices are governed by the design factors of the simulation study (given below): proportion of variance in $\mathbf{X}$ and $\mathbf{Y}$ explained by the underlying covariates. Four data characteristics were manipulated, based on the data generating model provided above. The different levels of the manipulated factors are given by square brackets.

Study setup

1. Number of predictors J: [200], [30]
2. Number of outcome variables: [5], [20]
3. Proportion of variance in $\mathbf{X}$ and $\mathbf{Y}$ explained by the covariates (Variance Accounted For; VAF): [0.9], [0.5]

The first and the second design factors concern the dimensionality of the predictor and outcome variables, respectively. The $\mathbf{P}^{(Y)}$ matrices created by the third design factor are shown below. We show the matrices corresponding to 5 outcome variables; the coefficients were defined in a similar manner for the case with 20 outcome variables, (provided in Appendix 5).

$$\begin{array}{c} \begin{matrix} 1 & 2 & 3 \end{matrix} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \end{array}$$

The columns indicate the regression coefficients corresponding to each covariate. As aforementioned, 40% of the outcome variables (2 out of 5) are not linked with any of the covariates. Fully crossing the design factors and generating 20 datasets per condition, $2 \times 2 \times 2 \times 50 = 400$ datases were produced. Three different analyses were administered to each of these datasets: SMPCovR, SPCovR and sPLS.

This data generating model is comprised of weights that have 'simple structure', with each important predictor linked only to a single covariate. However, in practice, the underlying processes may not be as straightforward, as multiple covariates may be associated with the predictors. To account for this, we conducted an additional simulation study where the weights are defined to be not in the simple structure. The findings are provided in Appendix 6, and they are in agreement with the results obtained from this simulation study.

3.2 Model selection

The model selection procedure for SMPCovR in the simulation study follows the procedure detailed in Sect. 2.3.3, except for the number of covariates which is fixed at three, by following the true covariate structure. For the first round of 5-fold CV, the weighting parameter α and the regularization parameters for regression coefficients γ_1 and γ_2 were selected simultaneously, while the other parameters λ_1 and λ_2 were fixed at a small value of 10^{-7}. The following ranges were considered:

1. α: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]
2. γ_1: equally distanced sequence of size 7 from 10^{-5} to 0.5 on the natural log scale
3. γ_2: equally distanced sequence of size 7 from 10^{-5} to 0.5 on the natural log scale

Crossing these ranges, $9 \times 7 \times 7 = 441$ model configurations were assessed with CV. The 1SE rule was used as described in Sect. 2.3.3 to select the values. With these parameters fixed, the parameters for the weights λ_1 and λ_2 were tuned by 5-fold CV. The following ranges were considered:

1. λ_1: equally distanced sequence of size 7 from 10^{-5} to 0.5 on the natural log scale
2. λ_2: equally distanced sequence of size 7 from 10^{-5} to 0.5 on the natural log scale

For this second round of CV, $7 \times 7 = 49$ models were considered. The 1SE rule was employed again to select the values for these parameters.

With regards to SPCovR, the number of covariates was fixed at three following the true number of covariates. Then, the following ranges of the parameters were considered simultaneously with 5-fold CV:

1. α: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]
2. λ_1: equally distanced sequence of size 10 from 10^{-5} to 0.5 on the natural log scale
3. λ_2: equally distanced sequence of size 7 from 10^{-5} to 0.5 on the natural log scale

In total, $9 \times 10 \times 7 = 630$ models were evaluated. The 1SE rule was employed in such a way that the model with the smallest α, the largest lasso, and the smallest ridge parameters was chosen, as they encourage a more sparse model to be found.

Lastly, the number of covariates for sPLS was fixed at three again. The number of non-zero coefficients (linking the predictor variables with the covariates) to be included in the sPLS model was chosen through 5-fold CV. The range of [4, 8, 12, 16, 20, 28] non-zero coefficients per covariate was considered for the low-dimensional setup ($6^3 = 216$ models in total), while the range of [25, 40, 50, 75, 80, 100, 120, 125, 150, 160, 175, 200] was employed for the high-dimensional setup ($12^3 = 1728$ models in total). We used the 1SE rule to pick out the model with the least number of non-zero coefficients.

3.3 Evaluation criteria

The following four measures were employed to study the performance of the methods:

1. R^2_{out}: proportion of explained variance in the entire set of the outcome variables in the out-of-sample test dataset.
2. $R^2_{\text{out}_{active}}$: proportion of explained variance in the subset of active outcome variables in the out-of-sample test dataset.
3. Correct weights classification rate: proportion of the elements in $\mathbf{W}$ correctly classified as zero and non-zero elements relative to the total number of coefficients.
4. Correct regression coefficients classification rate: proportion of the elements in $\mathbf{P}^{(Y)}$ correctly classified as zero and non-zero elements relative to the total number of coefficients (computed only for SMPCovR)

An independent test set (of 100 observation units) needed for computing the out-of-sample R^2 measures was generated following the same data generating procedures as the data used for model-fitting. The out-of-sample R^2 measures are defined as in the following equations:

$$R^2_{\text{out}} = 1 - \frac{\left\| \mathbf{Y}^{\text{out}} - \mathbf{X}^{\text{out}} \hat{\mathbf{W}} \hat{\mathbf{P}}^{(Y)^T} \right\|_2^2}{\left\| \mathbf{Y}^{\text{out}} \right\|_2^2,} \tag{6}$$

$$R^2_{\text{out}_{active}} = 1 - \frac{\left\| \mathbf{Y}^{\text{out}}_{K^\star} - \mathbf{X}^{\text{out}} \hat{\mathbf{W}} \hat{\mathbf{P}}^{(Y)^T}_{K^\star} \right\|_2^2}{\left\| \mathbf{Y}^{\text{out}}_{K^\star} \right\|_2^2} \tag{7}$$

where $\mathbf{Y}^{\text{out}}$ and $\mathbf{X}^{\text{out}}$ indicate the outcome and predictor variables, respectively, from the out-of-sample test data. The subscript $_{K^\star}$ denotes a subset within the sequence of indices for outcome variables $K^\star \subseteq \{1, 2, \dots, K\}$. It comprises of indices corresponding to the outcomes defined as being active. When SMPCovR excludes outcome variables from the model, it provides zero-predictions for these outcomes. In computing the R^2 measures, these zero-predictions are compared against $\mathbf{Y}^{\text{out}}$.

The correct classification rates concerning the weights and the regression coefficients represent the method's ability in retrieving the underlying processes. As SPCovR and sPLS only provide non-zero regression coefficients, they were excluded for the criterion of correct regression coefficients classification.

3.4 Results

3.4.1 Out-of-sample $\text{R}^2{}_{\text{out}}$ and $R^2_{\text{out}_{active}}$

Figure 1 clearly shows the outperformance of SMPCovR over the other methods. None of the study design factors led to results pointing in another direction. When the VAF was lower, the performance of SMPCovR stood out more prominently. The outperformance of SMPCovR comes from the fact that the method screens out the inactive outcome variables, while the other methods include these outcome variables. This can be understood as a case of overfitting, since the other methods are modelling inactive outcomes which are only comprised of error variance.

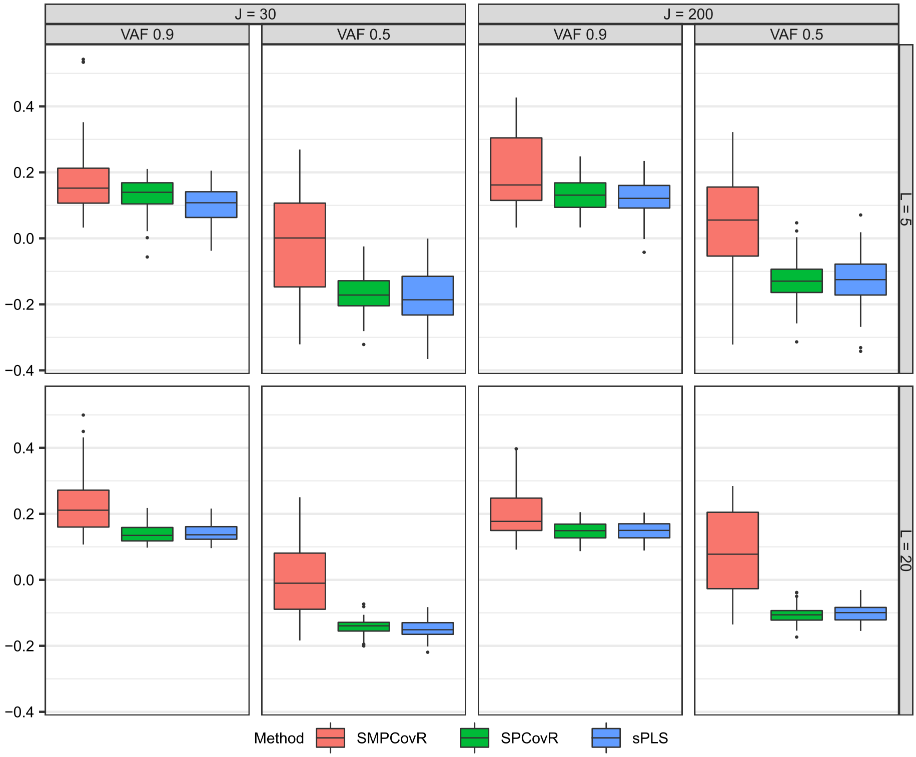

Fig. 1 Boxplots of the out of sample R^2_{out}. Each panel corresponds to one of the 8 conditions

Figure 2 reports the $R^2_{\text{out}_{active}}$ values computed only on the basis of active outcome variables; it can be seen that the three methods result in similar quality of prediction for the active outcomes. Hence, the strength of SMPCovR originates from correct identification of active and inactive outcome variables. It appears that when the underlying model is a covariate model, prediction of active outcomes appears to be straightforward for these methods. This is sensible, because once the covariates are accurately retrieved (in our simulation study, the results from the correct weights classification rate imply good recovery of covariates from all three methods), the task of predicting active outcomes becomes a low-dimensional regression problem. However, a major challenge in setups where inactive outcomes are expected may be to discern between active and inactive outcomes. It not only singles out relevant outcomes, but also contributes to the overall quality of prediction. By not distinguishing between the two types of outcomes, SPCovR and sPLS modelled the inactive outcomes and hence resulted in diminished prediction quality concerning the entire set of outcome variables. On the other hand, SMPCovR excluded the inactive outcomes and gained in overall prediction performance.

3.4.2 Correct weights classification rate

Figure 3 portrays that the most impactful design factor in the comparative performance with respect to correct identification of the zero versus non-zero weights is the dimensionality of the predictors. In the low-dimensional setting, SMPCovR and SPCovR resulted in

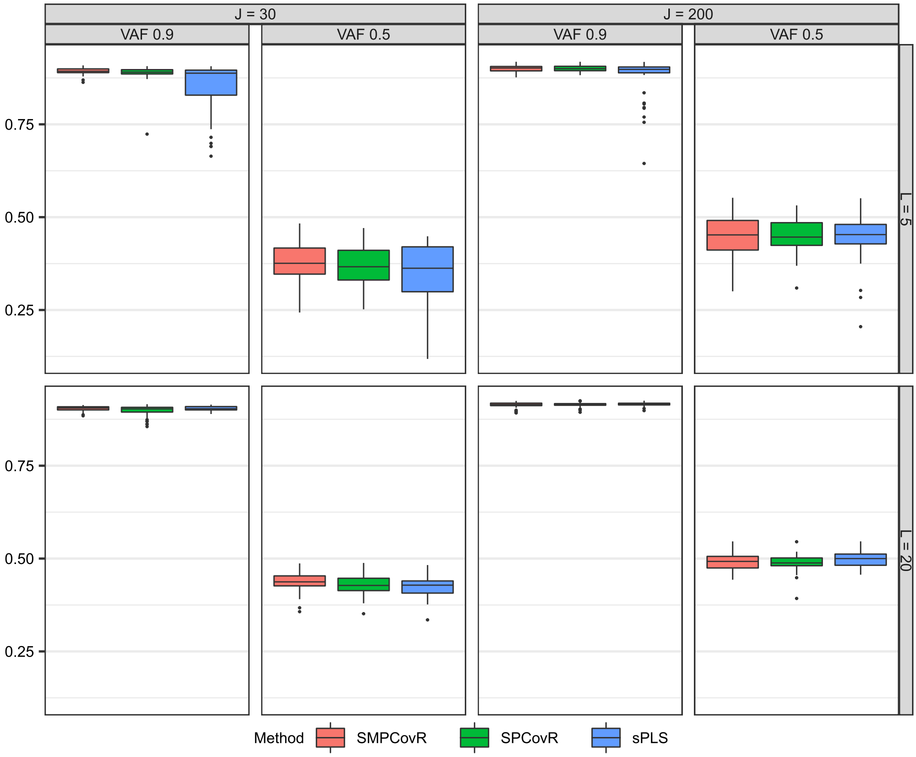

Fig. 2 $R^2_{\mathrm{out}_{active}}$ computed only on the basis of active outcomes. Each panel corresponds to one of the 8 conditions

comparable levels of correct classification rate which are higher than that of sPLS. In contrast, when the number of predictor variables exceeds the number of observations, the three methods have resulted in similar classification rates. Nevertheless, across most of the data conditions, it can be seen that similar levels of classification rates were obtained between the three methods.

3.4.3 Correct classification rate for regression coefficients

It appears that the true structure of the regression coefficients is recovered well in most conditions by SMPCovR. In addition, Appendix 7 shows the rate of correctly classified outcome variables. It can be seen that the method is able to provide a fairly good classification between active and inactive outcomes in most of the replicate datasets. Furthermore, to evaluate the performance of SPCovR and sPLS in handling the true zero regression coefficients, we inspected the coefficients that the two methods provided for the true zero regression coefficients. The mean absolute discrepancy of the estimated coefficients from zero are reported in Appendix 8. It can be observed that the coefficients from the two methods are quite far away from zero under low dimensionality. For high-dimensional data, while the mean discrepancy of SPCovR becomes near-zero, sPLS shows high discrepancy. This finding supports the use of a sparsity-inducing penalty on the regression coefficients, because without it, the methods struggle to derive near-zero values (Fig. 4).

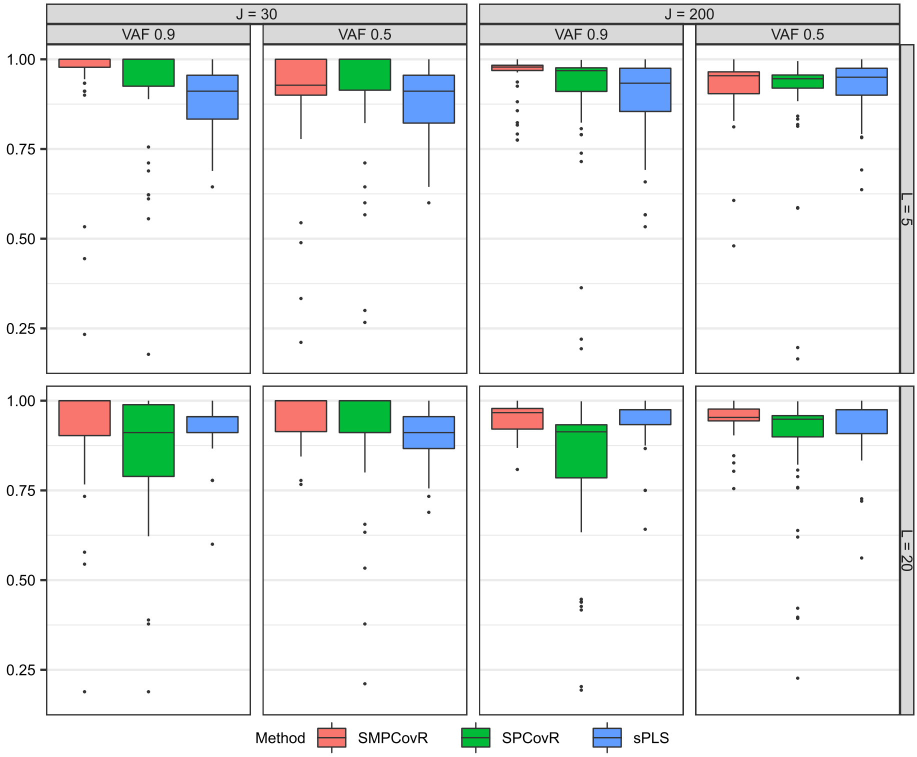

Fig. 3 Boxplots of the correct classification rate for the **W**. Each panel corresponds to one of the 8 conditions

4 Empirical illustration

We illustrate the use of SMPCovR by administering the method to an empirical dataset. We also apply SPCovR and sPLS on the same dataset to evaluate the effectiveness of our proposed method in a pratical setting.

4.1 Pittsburgh cold study

4.1.1 Dataset and pre-processing

We adopted the dataset from the third wave of the Pittsburgh Cold Study (PCS) which took place from 2007 to 2011.[2] Healthy participants were invited and administered nasal drops of rhinovirus that causes symptoms of common cold. Severity of 16 types of symptoms related to cold and flu were self-reported each day up to five days after the virus exposure. Out of the 16, there were 8 symptoms that were known to comprise the common cold:

[2] The data were collected by the Laboratory for the Study of Stress, Immunity, and Disease at Carnegie Mellon University under the directorship of Sheldon Cohen, PhD; and were accessed via the Common Cold Project website (www.commoncoldproject.com; grant number NCCIH AT006694).

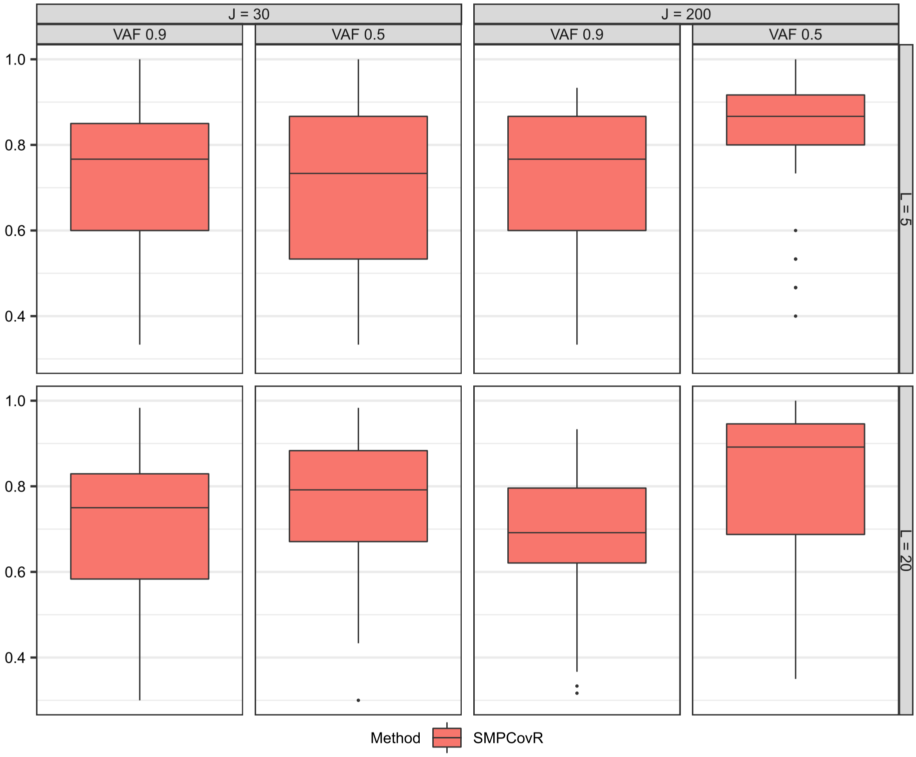

Fig. 4 Boxplots of the correct classification rate for the $\mathbf{P}^{(Y)}$. Each panel corresponds to one of the 8 conditions

headache, sneezing, chills, sore throat, runny nose, nasal congestion, cough and malaise (Jackson et al., 1958). Among the other symptoms, fever, muscle ache, joint ache and poor appetite have been identified as symptoms of flu (Monto et al., 2000), while there were 4 other related symptoms such as chest congestion, sinus pain, earache and sweating. Hence, it can be expected that the participants are more likely to develop the 8 cold symptoms than the other symptoms, as they were exposed to rhinovirus. Furthermore, 187 variables regarding the participants were also collected under various themes including blood chemistry, health practices and psychosocial states.

The participants are categorized into two groups according to the diagnosis of cold infection. This diagnosis was conducted by combining the serological testing of blood and illness criteria, and most of the participants were not diagnosed of cold infection. Therefore, we selected a subset of 46 participants by excluding the observations with missing values in the variables and to obtain a balance between the size of two diagnosis groups. Using the symptom variables as the outcome and the other variables as the predictors, we conduct SMPCovR to target the regression problem of symptom severity while constructing a model that describes the underlying predictive processes characterized by subsets of important predictor and outcome variables.

Table 2 The configurations of the models that fall within the 1 SE region from the maximum R^2_{cv}, from the first round of CV. SE denotes the standard error of R^2_{cv}, while 'Outcome included' refers to the number of outcome variables included. Note that the total numbers of weights and regression coefficients are $187 \times 2 = 374$ and $16 \times 2 = 32$, respectively

Model	α	γ_1	γ_2	R^2_{cv}	SE	Nonzero weights	Nonzero reg	Outcome included
1	0.6	0.0824	0.0452	0.0587	0.0145	374	14	13
2	0.6	0.0452	0.5000	0.0575	0.0411	374	21	16
3	0.8	0.0136	0.0824	0.0655	0.0291	374	27	15
4	0.8	0.0136	0.5000	0.0586	0.0109	374	29	15
5	0.9	0.0136	0.1503	0.0673	0.0277	374	29	16
6	0.5	0.0074	0.1503	0.0604	0.0196	374	30	16
7	0.7	0.0074	0.5000	0.0845	0.0373	374	32	16
8	0.9	0.0074	0.1503	0.0808	0.0515	374	32	16
9	0.9	0.0041	0.5000	0.0615	0.0669	374	32	16
10	0.3	0.0022	0.0001	0.0633	0.0496	374	32	16

4.1.2 Model selection

SMPCovR Prior to the model selection and estimation, both predictor and outcome variables were centered and standardized such that the variance of each variable was equal to 1. We followed the model selection strategy outlined in Sect. 2.3.3. First, with the acceleration factor technique, the number of covariates was determined to be two. Appendix 9 shows the proportion of variance explained with increasing number of components. The first round of 5-fold CV was administered to select the values for α, γ_1 and γ_2, by employing the following ranges for the parameters.

1. α: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]
2. γ_1: 0 and equally distanced sequence of size 19 from 10^{-5} to 0.5 on the natural log scale
3. γ_2: 0 and equally distanced sequence of size 19 from 10^{-5} to 0.5 on the natural log scale

Crossing these ranges, $9 \times 19 \times 19 = 3249$ model configurations were assessed with CV. There were 34 models that fall within the 1 SE region from the maximum R^2_{cv}. Table 2 presents 10 of these, arranged in an descending order of γ_1, followed by ascending orders of α and γ_2.

As parameters concerning the weights were fixed at 10^{-7}, it can be observed that all of the model configurations in the 1 SE region provided only non-zero values for the weights. As outlined in Sect. 2.3.3, we selected the model comprised with the largest γ_1, the smallest α and the smallest γ_2 values. Model 1 characterized by the parameters $\alpha = 0.6$, $\gamma_1 = 0.082$ and $\gamma_2 = 0.045$ was therefore chosen. With these parameters fixed, the parameters for the weights λ_1 and λ_2 were tuned by 5-fold CV. The following ranges were considered:

1. λ_1: 0 and equally distanced sequence of size 19 from 10^{-5} to 0.5 on the natural log scale
2. λ_2: 0 and equally distanced sequence of size 19 from 10^{-5} to 0.5 on the natural log scale

Table 3 The configurations of the models that fall within the 1 SE region from the maximum R^2_{cv}, from the second round of CV. SE denotes the standard error of R^2_{cv}, while 'Outcome included' refers to the number of outcome variables included. Note that the total numbers of weights and regression coefficients are $187 \times 2 = 374$ and $16 \times 2 = 32$, respectively

Model	λ_1	λ_2	R^2_{cv}	SE	Nonzero weights	Nonzero reg	Outcome included
1	0.0041	1e−04	0.0767	0.0267	33	11	11
2	0.0012	0e+00	0.0895	0.0261	69	12	12
3	0.0012	1e−04	0.0927	0.0164	71	12	12
4	0.0007	0e+00	0.0787	0.0101	102	13	13
5	0.0007	2e−04	0.0805	0.0208	112	12	12
6	0.0007	4e−04	0.0791	0.0164	125	12	12
7	0.0004	0e+00	0.0790	0.0167	132	13	13
8	0.0004	1e−04	0.0797	0.0222	136	13	13
9	0.0004	4e−04	0.0848	0.0244	159	13	13

For this second round of CV, $19 \times 19 = 361$ models were considered. Within the 1 SE region, 9 models were found, which are provided in Table 3, arranged in a descending order of λ_1 combined with an ascending order of λ_2.

Following the rationale of selecting the largest λ_1 and the smallest λ_2 values, Model 1 was chosen with $\lambda_1 = 0.0041$ and $\lambda_2 = 10^{-4}$. This model 'provided the least number of non-zero weights and regression coefficients, while including the least number of outcomes. Table 4 displays the weights and regression coefficients of this model.

4.1.3 Results

Table 4 presents the weights and regression coefficients found by the chosen model. It first shows that only the first covariate is able to predict the cold symptoms; the model has excluded the second covariate in predicting the outcome variables. Out of the 187 predictor variables, 21 predictor variables compose the first covariate. IL-6, IL-8, IL-10 and TNF alpha are concentrations of nasal cytokine. These concentrations were measured each day for five days after the viral exposure and summed. Among the total 7 variables present in the data concerning cytokine, these 4 were picked out by the model. The model also selected weight and concentration measures of corpuscular hemoglobin, Non-fasting glucose and Urea nitrogen among the 29 blood chemistry variables measured before the viral exposure. Whereas lower levels of hemoglobin appears to result in more cold-related symptoms, glucose and nitrogen levels seem to have the opposite effect. # weekdays alcohol refers to the amount of alcohol usually consumed during weekdays. The alcohol consumption appears to be positively associated with the cold-related symptoms. This was the only variable chosen among 17 variables regarding health practices such as smoking, sleeping and physical activity. The next 5 variables concern measures from various psychosocial assessment scales measured before the viral exposure. Sadness and fatigue were found to be related with cold symptoms from the 13 PANAS (Positive and Negative Affect Schedule; Watson et al., 1988) measures that target mood and affect. Similarly, the ECR (Experiences in Close Relationships; Fraley et al., 2000) scale concerns adult attachment types. Along the same line, social participation and loneliness were also results from

Table 4 Weights and regression coefficients derived by SMPCovR from the PCS dataset. The weights are only provided for the predictors chosen by the model out of the total 187. $\hat{\mathbf{w}}_1$ and $\hat{\mathbf{w}}_2$ indicate the weights corresponding to the first and second covariates respectively. The regression coefficients corresponding to all of the outcome variables in the dataset are provided

$\hat{w}_2$		
	1	
IL-6	0.339	
IL-8	0.536	
IL-10	0.461	
TNF alpha	1.860	
Corpuscular Hgb (weight)	−0.048	
Corpuscular Hgb (conc)	−0.533	
Non-fasting glucose	0.002	
Urea nitrogen	0.210	
# weekdays alcohol	0.616	
PANAS: sadness	0.938	
PANAS: fatigue	0.476	
ECR: anxiety	0.123	
Social participation	−0.154	
Loneliness	0.156	
Daily: loneliness subscale	0.065	
Daily: loneliness	0.090	
Daily: negative subscale	0.457	
Daily: fatigue subscale	0.546	
Daily: fatigue	0.672	
Daily: tiredness	0.069	
Daily: anger subscale	0.401	
$\hat{\mathbf{w}}_2$		
	2	
PANAS: joviality	0.483	
PANAS: positive	0.638	
IPIP: emo. Stable	0.147	
IPIP: agreeable	0.275	
Opener: total	0.124	
ECR: avoid	−0.136	
GS-ISEL: total	0.121	
PWB: self-acceptance	0.214	
PWB: env. mastery	0.180	
PWB: positive rel	0.350	
PWB: Psych well-being	1.151	
# days: hugs	0.078	
$\hat{\mathbf{P}}^{(Y)}$		
	1	2
Sneezing	0.084	0
Runny nose	0.062	0
Nasal congestion	0.060	0
Cough	0.099	0
Sore throat	0.058	0
Headache	0.002	0
Chills	0.000	0

Table 4 (continued)

$\hat{w}_2$		
	1	
Malaise	0.035	0
Chest congestion	0.062	0
Sinus pain	0.070	0
Earache	0.000	0
Muscle ache	0.000	0
Joint ache	0.000	0
Sweating	0.000	0
Fever	0.042	0
Poor appetite	0.074	0

self-reported scales before the viral exposure. Lastly, the daily loneliness, negative affect, fatigue, tiredness and anger variables come from daily interviews conducted prior to the viral exposure. Altogether, the first covariate represents the combined effect of these physiological and behavioural elements in leading to the various cold symptoms.

Eleven symptoms out of the total 16 were indicated to be in relation with the first covariate. Seven out of 8 symptoms characterizing the common cold according to Jackson et al. (1958)[3] were included in the model; it excluded chills. It is also interesting to see that symptoms typically associated with flu such as fever and poor appetite are also included (Monto et al., 2000), while the participants were not exposed to an influenza virus known to cause flu.

The second covariate which is not relevant in predicting the symptoms is constructed with 12 predictor variables, most of which originate from the psychosocial assessment scales. In addition to the PANAS and ECR scales featured for the first covariate, variables from IPIP

(International Personality Item Pool; Goldberg et al., 1999), a well-known scale for the big five personality, the Opener scale (Miller et al., 1983) which assesses the tendency to "open up" to others, GS-ISEL (Giving Support - Interpersonal Support Evaluation List; Cohen et al., 1985) that measures the perceived extent of providing social support to others, and Ryff scales of Psychological Well-Being (Ryff, 1989) were found to compose the second covariate. Additionally, the number of days experiencing hugs from the daily interview was also included. Together, the second covariate can represent a process that is a mixture of social openness, psychological well-being and positive affect.

Although not in relation with the cold symptoms, we found that the second covariate explains much more variance in the predictor variables than the first covariate comprised of 21 variables. While the two covariates together explained 14.3% of variance in the predictors, the first covariate took account of 5.3% while the second covariate explained the remainnig 9%.

To evaluate the quality of this model in outcome variable prediction, the R^2 measures were computed. We have calculated six different types of R^2 measures: $R^2_{\text{fit}_{all}}$, $R^2_{\text{fit}_{sub}}$, $R^2_{\text{fit}_{active}}$ $R^2_{\text{loocv}_{all}}$, $R^2_{\text{loocv}_{sub}}$ and $R^2_{\text{loocv}_{active}}$. The first three measures were computed on the basis of in-

[3] headache, sneezing, chills, sore throat, runny nose, nasal congestion, cough and malaise.

Table 5 R^2 measures attained from the three methods from the PCS data

	SMPCovR	SPCovR	sPLS
fit_{all}	0.151	0.266	0.206
fit_{sub}	0.220	0.353	0.261
fit_{active}	0.190	0.316	0.197
loocv_{all}	0.113	0.112	0.052
loocv_{sub}	0.131	0.111	0.023
loocv_{active}	0.128	0.105	−0.009

sample data while the next three measures were results from leave-one-out CV. The measures with the subscript 'all' were computed with respect to all of the outcome variables in the dataset, while the ones with the subscript 'sub' were derived on the basis of the subset of 11 outcome variables selected by the SMPCovR model. Lastly, the measures with the subscript 'active' were computed from the 8 symptoms known to characterize the common cold. These 8 outcomes are considered as active outcomes, because the participants were exposed to virus causing common cold. Appendix 10 provides the formulae for these measures. To obtain a comparative insight about the quality of the SMPCovR method under the PCS dataset, we also computed the R^2 values using SPCovR and sPLS that were employed in the simulation study. We extracted two covariates for both methods in order to match the SMPCovR model. As done in the simulation study, 5-fold CV and the 1SE rule were employed to select the parameters for SPCovR and the number of non-zero coefficients for sPLS. Appendix 11 provides the ranges of tuning parameters adopted to generate the models for the two methods. Table 5 reports the six different types of R^2 measures computed for the three methods.

It can be seen that SMPCovR resulted in the highest R^2_{loocv} measures which represent the quality of out-of-sample prediction. While SPCovR showed comparable results with SMPCovR, sPLS fell short by a big margin. Whereas both SPCovR and sPLS performed well for in-sample prediction with high R^2_{fit} values, the large discrepancy in the values compared to the R^2_{loocv} measures signal possible occurrence of overfitting. The models constructed by SPCovR and sPLS can be found in Appendix 11. It can be seen that although SPCovR led to similar out-of-sample prediction quality as SMPCovR, the method found considerably more non-zero weights (100 and 55 for the two covariates, respectively.) On the other hand, the sPLS model found was very sparse, only comprised of 6 and 1 non-zero coefficients.

Lastly, we inspected the SMPCovR model by plotting the covariate scores with the additional grouping information of diagnosis of cold infection (diagnosed using serological testing and illness criteria). Although this grouping information was not provided as a predictor, the two groups of cold and no cold can be fairly distinguished. As portrayed by the regression coefficients shown in Table 4, it appears that the first covariate is much more related with cold diagnosis than the second covariate. To conclude, the SMPCovR method was able to meet its goals when analyzing the PCS dataset. It derived a predictive model where some of the inactive outcome variables are filtered out while summarizing the predictor processes into interpretable covariates comprised of a small subset of predictor variables (Fig. 5).

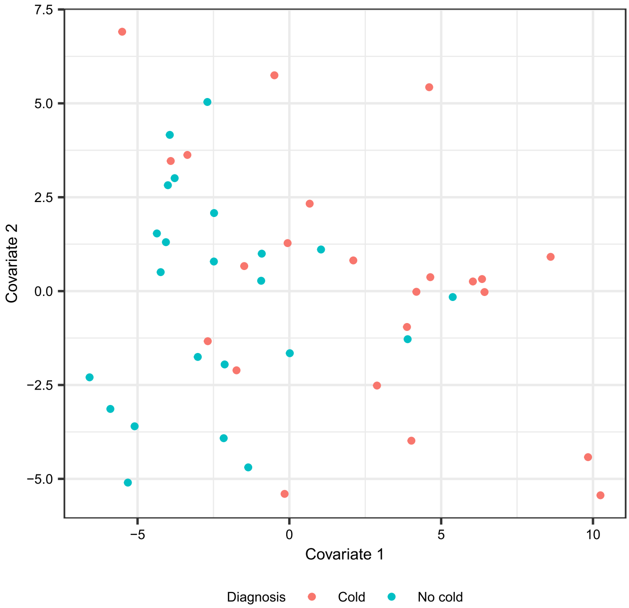

Fig. 5 Scatterplot of the two covariates found by SMPCovR. The colours represent the cold diagnosis

5 Discussion

Predictive modelling in the presence of large numbers of predictor and outcome variables presents multiple challenges. Constructed models feature a huge number of estimated coefficients, rendering the interpretation infeasible. Moreover, there may be subsets of both predictor and outcome variables that are not important. Certain predictor variables may be redundant in predicting any of the outcome variables, while some outcome variables may not at all be adequately predicted by the available predictors.

In this paper, we proposed the method of SMPCovR that accommodates for these issues by relying on PCovR methodology and incorporating sparsity penalties at both sides of predictors and outcomes. Through a simulation study, it was shown that our method performs well at retrieving the coefficients that represent how processes underneath data underlie the predictor and outcome variables. With regards to prediction of outcome, SMPCovR showed outperformance in the prediction of the entire set of outcome variables, owing to the correct exclusion of inactive outcomes. However, concerning exclusively the active outcomes, SMPCovR did not exhibit a notable improvement in predictive quality compared to the other methods. This may be attributed to the fact that the datasets were generated from covariates, as discussed in Section 3.4. In other settings where the data generating model is not characterized by covariate structures, different results may be expected. For example, in the case of the PCS data, where the underlying model is unknown, SMPCovR exhibited better prediction of the active outcomes than SPCovR and sPLS. To further investigate into this scenario where the underlying model is not based on covariates, we conducted an additional simulation study where a linear model was used instead (Appendix 12). In line with the results from the PCS data, SMPCovR provided better predictions for the active outcomes than the other methods.

The PCovR methodology provides an advantageous position in the settings with large numbers of predictor and outcome variables. The predictors and outcomes are linked with the reduced dimensions of the covariates, instead of being directly connected with each other. This reduces the number of estimated coefficients by far. In total, $(J + K) \times R$ coefficients need to be found by SMPCovR, while $J \times K$ coefficients need estimation in a regularized regression setup with predictors and outcomes directly connected. Using the example of the PCS dataset in Section 4, SMPCovR model would comprise of $(187 + 16) \times 2 = 406$ coefficients at maximum, while a regression model can consist of $187 \times 16 = 2992$ coefficients. By imposing further sparsity penalties on the coefficients, SMPCovR can derive an even more sparse and concise model representation. Furthermore, the reduction of the number of coefficients also implies that less number of coefficients need to be forced to zero to exclude a variable (both predictor and outcome) altogether from the model. As a consequence, SMPCovR is a prediction method with multivariate outcomes that conducts variable selection in an effective manner. These strengths also apply generally to other regression methods based on dimension reduction such as PLS.

There are limitations to our proposed method. Being characterized with 6 different tuning parameters, model selection is a natural complication. To reduce the compuational burden of CV, we employed a sequential model selection approach where sets of model parameters were separately tuned in turn, instead of simultaneous selection. The sequential approach has been shown suitable for PCovR and SPCovR (Vervloet et al., 2016; Park et al., 2020). In this study, the strategy resulted in good results in both the simulation and empirical studies. However, we did not conduct an extensive investigation focused on the model selection approaches due to the scope of our paper.

Besides the weakness of the sequential approach that the selection the parameters is detached from each other, there are additional concerns regarding the acceleration factor technique for determining the number of covariates. While various studies compared the scree test along with many other methods for selecting the number of components, the scree test has not been selected as the optimal choice (Jackson, 1993; Ferré, 1995; Henry et al., 1999). In fact, there has not been a clear consensus on the best method. Although employing CV also to choose the number of covariates could have been a choice that is well-aligned with the rest of the model selection procedure, it increases the computational burden. Furthermore, it has been reported that CV tends to include too many covariates for PCovR (Vervloet et al., 2016). Consequently, the decision to opt for the scree plot approach with the acceleration factor technique was guided by its intuitive nature and computational efficiency. Further research is needed to gain a deeper understanding of the most effective covariate selection strategies for PCovR and its sparse extensions, such as our current method of SMPCovR.

In a similar vein, it is worth noting that estimating the SMPCovR model involves a notable time cost. To provide an indication, we ran the SMPCovR model presented in Sect. 4.1.3 one hundred times on a laptop equipped with a four-core Intel i5-10210U processor with a base clock speed of 2.11 GHz and 8GB of RAM. On average, each run took approximately 0.729 s to complete. In contrast, the sPLS model discussed in the same section had an average runtime of just 0.031 s. Considering that this difference in time would be amplified with datasets that are larger than the PCS data, combined with the extensive list of parameters to be tuned, it appears that SMPCovR may not be the most efficient choice in practice and that there is room to improve the implementation of the method.

Lastly, our method targets a non-convex problem by alternating least squares, which can lead to convergence to local minima (please refer to Sect. 2.3.2). While we recommend employing multiple random starting values as well as rational starting values based on PCovR, strategies for avoiding local minima, such as simulated annealing (Kirkpatrick et al., 1983; e.g. Ceulemans et al., 2007) and (De Jong, 1975) can be considered for future research.

Our proposed method is one of the first regression methods that conducts variable selection in both predictor and outcome variables. With growing availability of large datasets and increasing use of data collected without specific research aims, we believe such methods are becoming more relevant. The literature also seems to be steering towards this direction, with Hu et al. (2022) hinting at an adaptation to the objective criterion to allow predictor variable selection on top of the outcome variable selection offered in Hu et al. (2022). We expect that PCovR and other multivariate methods that leverage from dimension reduction to bear great potential in taking the lead in this under-studied research problem.

Appendix 1. SMPCovR algorithm

The SMPCovR loss (3) can be minimized by an alternating least squares procedure. A schematic outline of the algorithm is provided in what follows. It is similar to the procedures proposed to solve SCaDS (de Schipper & Van Deun, 2018), SPCovR (Van Deun et al., 2018) and SSCovR (Park et al., 2020). The algorithm involves solving for all covariates together (unlike the deflation approach in which one covariate is solved in turn). The routine continues until the algorithm converges into a stationary point, usually a local minium. To avoid local minima problems, we recommend to use multiple random and a rational starting value based on PCovR.

Algorithm 1 SMPCovR

```
Inputs:
    $\mathbf{X}$ and $\mathbf{Y}$, number of covariates $R$, weighting parameter $\alpha$, regularization parameters for $\mathbf{W}$ $\lambda_{Lr}$ and $\lambda_{Rr}$, regularization parameters for $\mathbf{P}^{(Y)}$ $\gamma_{L_r}$, and $\gamma_{R_r}$, maximum number of iterations $T$, convergence threshold $\epsilon \geq 0$
Initialize:
    $\mathbf{W} \leftarrow \mathbf{W}^{(0)}$ $L_0 \leftarrow$ Initial loss,
    Loss difference $d \leftarrow 1$, Iteration counter $t \leftarrow 1$
while $t < T$ or $\epsilon < d$ do
    Conditional estimation of ${\mathbf{P}^{(X)}}^{(t)}$, ${\mathbf{P}^{(Y)}}^{(t)}$ given $\mathbf{W}^{(t)}$
    Conditional estimation of $\mathbf{W}^{(t+1)}$ given ${\mathbf{P}^{(X)}}^{(t+1)}$ and ${\mathbf{P}^{(Y)}}^{(t+1)}$
    $L_u \leftarrow$ updated loss given $\mathbf{W}^{(t+1)}$, ${\mathbf{P}^{(X)}}^{(t+1)}$ and ${\mathbf{P}^{(Y)}}^{(t+1)}$
    $d \leftarrow L_0 - L_u$
    $t \leftarrow t + 1$
    $L_0 \leftarrow L_u$
end while
```

Appendix 2. Estimation of W

Conditional estimation of $\mathbf{W}$ given the other parameters $\mathbf{P}^{(X)}, \mathbf{P}^{(Y)}$ pertains to an elastic net regression problem. The SMPCovR objective function (3) is first arranged with respect to the hth element of the weights corresponding to the covariate component r^*: w_{hr^*}.

$$\begin{aligned} L\left(w_{hr^*}\right) = {} & \frac{\alpha}{\|\mathbf{Y}\|_2^2} \sum_i^N \left\| \mathbf{y}_i - \sum_r^R \sum_{j \neq h}^J x_{ij} w_{jr} \mathbf{p}_r^{(Y)} - \sum_{r \neq r^*}^R x_{ih} w_{hr} \mathbf{p}_r^{(Y)} - x_{ih} w_{hr^*} \mathbf{p}_{r*}^{(Y)} \right\|_2^2 \\ & + \frac{1-\alpha}{\|\mathbf{X}\|_2^2} \sum_i^N \left\| \mathbf{x}_i - \sum_r^R \sum_{j \neq h}^J x_{ij} w_{jr} \mathbf{p}_r^{(X)} - \sum_{r \neq r^*}^R x_{ih} w_{hr} \mathbf{p}_r^{(X)} - x_{ih} w_{hr^*} \mathbf{p}_{r^*}^{(X)} \right\|_2^2 \\ & + \lambda_1 \left| w_{hr^*} \right| + \lambda_2 \ w_{hr^*}^2 + \gamma_1 \left| \mathbf{p}_{r^*}^{(Y)} \right|_1 + \gamma_2 \left\| \mathbf{p}_{r^*}^{(Y)} \right\|_2^2. \end{aligned} \tag{8}$$

Taking the derivative with respect to w_{hr^*} we get:

$$\begin{aligned} & \frac{-2\alpha}{\|\mathbf{Y}\|_2^2} \sum_i^N \mathbf{p}_{r^*}^{(Y)^T} \left(\mathbf{r}_{ih} - x_{ih} w_{hr^*} \mathbf{p}_{r^*}^{(Y)} \right) x_{ih} - \frac{2(1-\alpha)}{\|\mathbf{X}\|_2^2} \sum_i^N \left(s_{ih} - x_{ih} w_{hr^*} \right) x_{ih} \\ & \quad + \lambda_1 \partial \left| w_{hr^*} \right| + 2\lambda_2 \ w_{hr^*} \end{aligned} \tag{9}$$

where

$$\begin{aligned} \mathbf{r}_{ih} &= \mathbf{y}_i \sum_r^R \sum_{j \neq h}^J x_{ij} w_{jr} \mathbf{p}_r^{(Y)} - \sum_{r \neq r^*}^R x_{ih} w_{hr} \mathbf{p}_r^{(Y)} \\ s_{ih} &= \mathbf{p}_{r^*}^{(X)^T} \mathbf{x}_i - \sum_{j \neq h}^J x_{ij} w_{jr^*}. \end{aligned} \tag{10}$$

We can equate the derivative to zero to satisfy the optimality conditions for $\hat{w}_{hr^*}$, which can be summarized by the following:

$$\hat{w}_{hr^*} = \frac{S\left(\sum_i^N \left[\frac{2\alpha}{\|\mathbf{Y}\|_2^2} \left(\mathbf{p}_{r^*}^{(Y)^T} \mathbf{r}_{ih} + \frac{2(1-\alpha)}{\|\mathbf{X}_C\|_2^2} s_{ih} \right) x_{ih} \right], \lambda_1 \right)}{\sum_i^N \left(\frac{2\alpha}{\|\mathbf{Y}\|_2^2} \left\| \mathbf{p}_{r^*}^{(Y)} \right\|_2^2 + \frac{2(1-\alpha)}{\|\mathbf{X}_C\|_2^2} \right) {x_{ih}}^2 + 2\lambda_2} \tag{11}$$

where S(.) is a element-wise soft-thresholding operator. With these conditions, we can set up the following coordinate descent algorithm.

Algorithm 2 Coordinate descent for the weights

```
for $r^*$ in $1:R$ do
    for $h$ in $1:J$ do
        $\hat{w}_{hr^*} \leftarrow \frac{S\left( \sum_i^N \left[ \frac{2\alpha}{\|\mathbf{Y}\|_2^2} \left( \mathbf{p}_{r^*}^{(Y)^T} \mathbf{r}_{ih} + \frac{2(1-\alpha)}{\|\mathbf{X}_C\|_2^2} s_{ih} \right) x_{ih} \right], \lambda_1 \right)}{\sum_i^N \left( \frac{2\alpha}{\|\mathbf{Y}\|_2^2} \left\| \mathbf{p}_{r^*}^{(Y)} \right\|_2^2 + \frac{2(1-\alpha)}{\|\mathbf{X}_C\|_2^2} \right) {x_{ih}}^2 + 2\,\lambda_2}$
```

Appendix 3. Estimation of $\mathbf{P}^{(Y)}$

Conditional estimation of $\mathbf{P}^{(Y)}$ given the other parameters $\mathbf{W}_C, \mathbf{P}_C^{(X)}$ is an elastic net regression problem. The SMPCovR objective function (3) is first arranged with respect to the regression coefficients corresponding to hth outcome variable and r^*th covariate:

$$L\left(p_{hr^*}^{(Y)}\right) = \frac{\alpha}{\|\mathbf{Y}\|_2^2} \sum_i^N \left(y_{ih} - \sum_{r \neq r^*}^R \mathbf{x}_i{}^T \mathbf{w}_r p_{hr}{}^{(Y)} - \mathbf{x}_i{}^T \mathbf{w}_{r^*} p_{hr^*}^{(Y)} \right)^2 + \gamma_1 \left| p_{hr^*}^{(Y)} \right| + \gamma_2 \; p_{hr^*}^{(Y)\,2}. \tag{12}$$

Taking the derivative with respect to $p_{hr^*}{}^{(Y)}$:

$$\frac{-2\alpha}{\|\mathbf{Y}\|_2^2} \sum_i^N \mathbf{x}_i{}^T \mathbf{w}_{r^*} \left(t_{ih} - \mathbf{x}_i{}^T \mathbf{w}_{r^*} p_{hr^*}^{(Y)} \right) + \gamma_1 \partial \left| p_{hr^*}{}^{(Y)} \right| + 2\gamma_2 \; p_{hr^*}^{(Y)} \tag{13}$$

where

$$t_{ih} = y_{ih} - \sum_{r \neq r^*}^R \mathbf{x}_i{}^T \mathbf{w}_{r^*} p_{hr}^{(Y)}. \tag{14}$$

We can equate the derivative to zero to satisfy the optimality conditions for $\hat{p}_{hr^*}^{(Y)}$, which can be summarized by the following:

$$\hat{p}_{hr^*}^{(Y)} = \frac{S\left(\sum_i^N \left(\mathbf{x}_i{}^T \mathbf{w}_{r^*} \right) t_{ih}^{(r^*)}, \frac{\|\mathbf{Y}\|_2^2 \gamma_1}{2\alpha} \right)}{\sum_i^N \left(\mathbf{x}_i{}^T \mathbf{w}_{r^*} \right)^2 + \left(\|\mathbf{Y}\|_2^2 / \alpha \right) \gamma_2}. \tag{15}$$

With these conditions, we can set up the following coordinate descent algorithm.

Algorithm 3 Coordinate descent for the regression coefficients $\mathbf{P}^{(Y)}$

```
for $r^*$ in $1 : R$ do
    for $h$ in $1 : K$ do
        $\hat{p}_{hr^*}^{(Y)} \leftarrow \frac{S\left( \sum_i^N \left( \mathbf{x}_i{}^T \mathbf{w}_{r^*} \right) t_{ih}^{(r^*)}, \frac{\|\mathbf{Y}\|_2^2 \gamma_1}{2\alpha} \right)}{\sum_i^N \left( \mathbf{x}_i{}^T \mathbf{w}_{r^*} \right)^2 + \left( \|\mathbf{Y}\|_2^2 / \alpha \right) \gamma_2}$
```

Appendix 4. Estimation of $\mathbf{P}^{(X)}$

The loadings $\mathbf{P}^{(X)}$ such that $\mathbf{P}^{(X)^T} \mathbf{P}^{(X)} = \mathbf{I}_R$ are obtained via a closed-form solution; $\mathbf{P}^{(X)} = \mathbf{U}\mathbf{V}^T$ where $\mathbf{U}$ and $\mathbf{V}$ are found through singular value decomposition of $\mathbf{X}^T\mathbf{X}\mathbf{W} = \mathbf{U}\mathbf{D}\mathbf{V}^T$.

Appendix 5. Regression coefficients $\mathbf{P}^{(\boldsymbol{Y})}$ defined for the case with 20 outcome variables

$$\begin{array}{c}
\begin{matrix} 1 & 2 & 3 \end{matrix} \\
\begin{pmatrix}
1 & 0 & 0 \\
1 & 0 & 0 \\
1 & 0 & 0 \\
1 & 1 & 1 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
1 & 1 & 1 \\
0 & 0 & 1 \\
0 & 0 & 1 \\
0 & 0 & 1 \\
1 & 1 & 1 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{pmatrix}
\end{array}$$

Appendix 6. Simulation study with weights not in simple structure

In this section, we conducted an additional simulation study where the true weights are defined to be not in the simple structure. In the original simulation study featured in the main text, each important predictor was defined to be linked only to one covariate. In the current simulation study in this section, some of the important predictors were defined to be associated with multiple covariates. All of the settings employed in the original simulation study were directly adopted in this study, except for the weights matrix used for the data generating model, and the number of replicate datasets generated. The following weights matrix was used for the low-dimensional setup (Table 6).

In comparison with the weights matrix presented in Table 1 employed for the main simulation study, it can be seen that 4 additional predictors were defined to each be linked with two covariates. For the high-dimensional setting comprised of 200 total predictors, 20 additional predictors were defined in this way. Using this new weights

Table 6 Weights defined for the low-dimensional setup for the simulation study with weights not having simple structure

W		
1	2	3
0.474	0	0
0.474	0	0
0.474	0	0
0.474	0	0
0	0.335	0
0	0.335	0
0	0.335	0
0	0.335	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0.335	0
0	0.335	0
0	0.335	0
0	0.335	0
0	0	0.474
0	0	0.474
0	0	0.474
0	0	0.474
0.200	−0.141	0
0.1	−0.283	0
0.2	0	−0.141
0.1	0	0.283
0	0	0
0	0	0
0	0	0

matrix, the setup provided in (5) was employed to generate each dataset. In line with the original simulation study, the number of observations was fixed at 100.

The design factors in the original simulation study (number of predictors, number of outcomes and VAF) were directly adopted (please refer to Section 3.1). Fully crossing the factors and generating 20 datasets per condition, $2 \times 2 \times 2 \times 20 = 160$ datasets were produced. The three methods (SMPCovR, SPCovR and sPLS) were administered to each of these datasets, with the model selection procedures kept the same as the original simulation study. The four evaluation criteria in the original study were also used here to assess the performance of the methods. The next section provides the results.

Results

Out-of-sample R^2_{out} and $R^2_{\text{out}_{active}}$

See Figs. 6, 7.

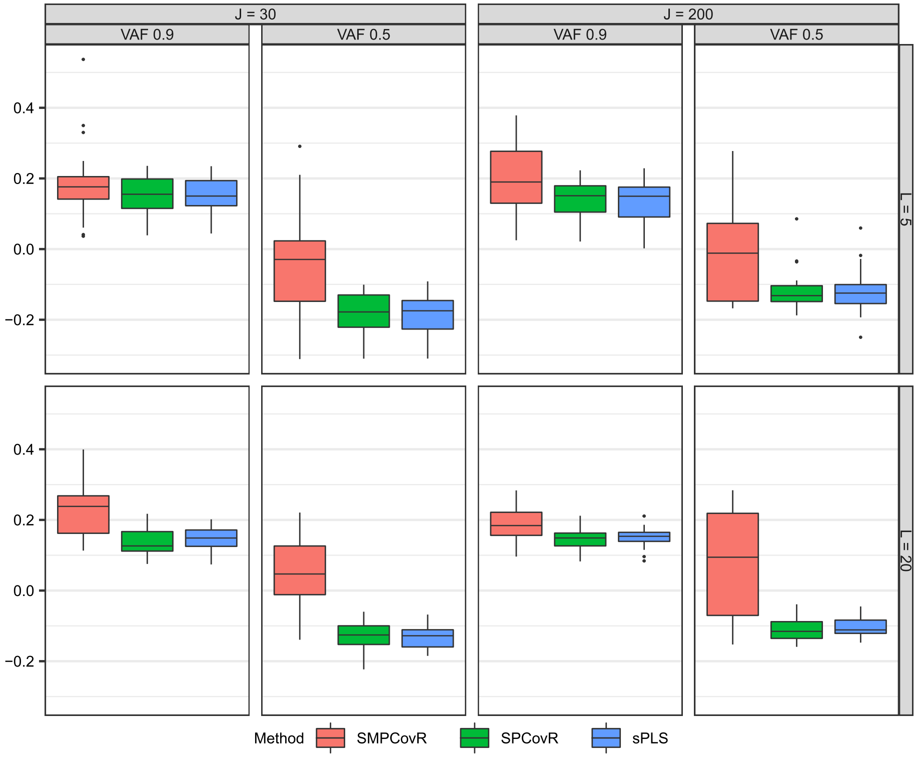

Fig. 6 Boxplots of the out of sample R^2_{out}. Each panel corresponds to one of the 8 conditions. From the simulation study with weights not in simple structure

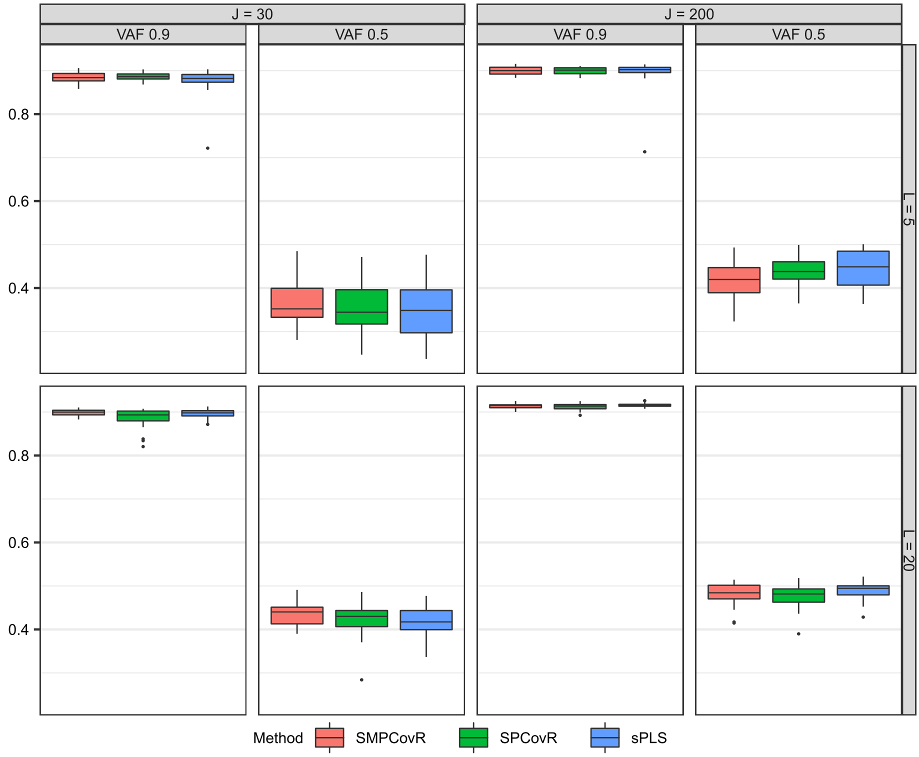

Fig. 7 Boxplots of the out of sample $R^2_{\text{out}_{active}}$. Each panel corresponds to one of the 8 conditions. From the simulation study with weights not in simple structure

Correct weights classification rate

Correct classification rate for regression coefficients

Altogether, the results concerning all of the four evaluation criteria were very much in line with those from the original simulation study. The only small difference pertains to the correct classification of weights displayed in Fig. 8; The performance of SPCovR is worse for the current simulation study where the weights are not in the simple structure than in the original study (Fig. 9).

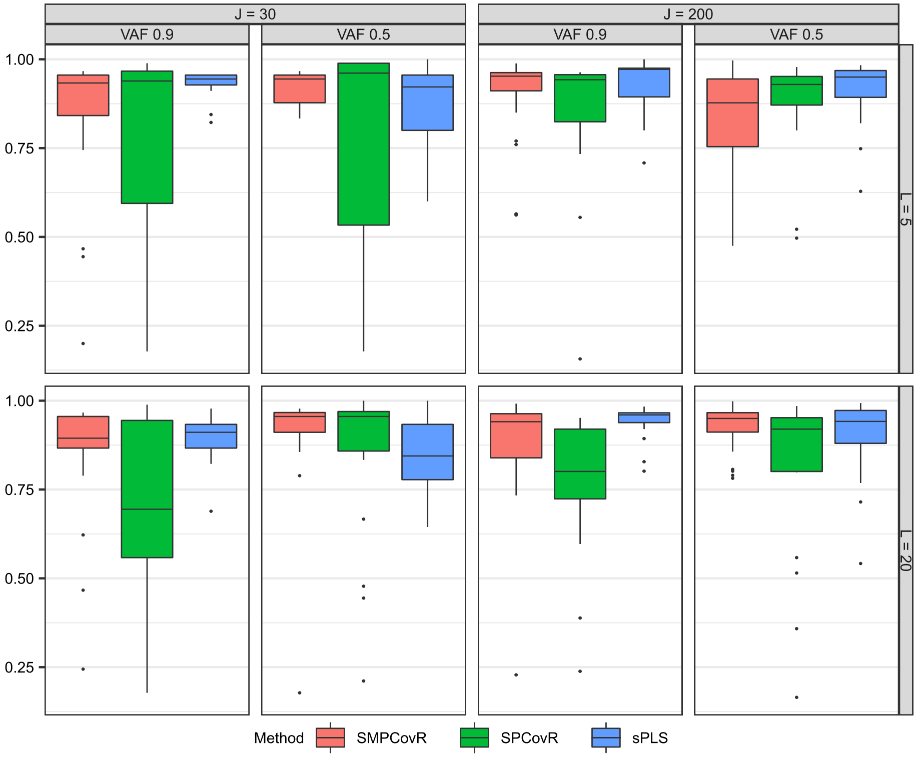

Fig. 8 Boxplots of the correct classification rate for the **W**. Each panel corresponds to one of the 8 conditions. From the simulation study with weights not in simple structure

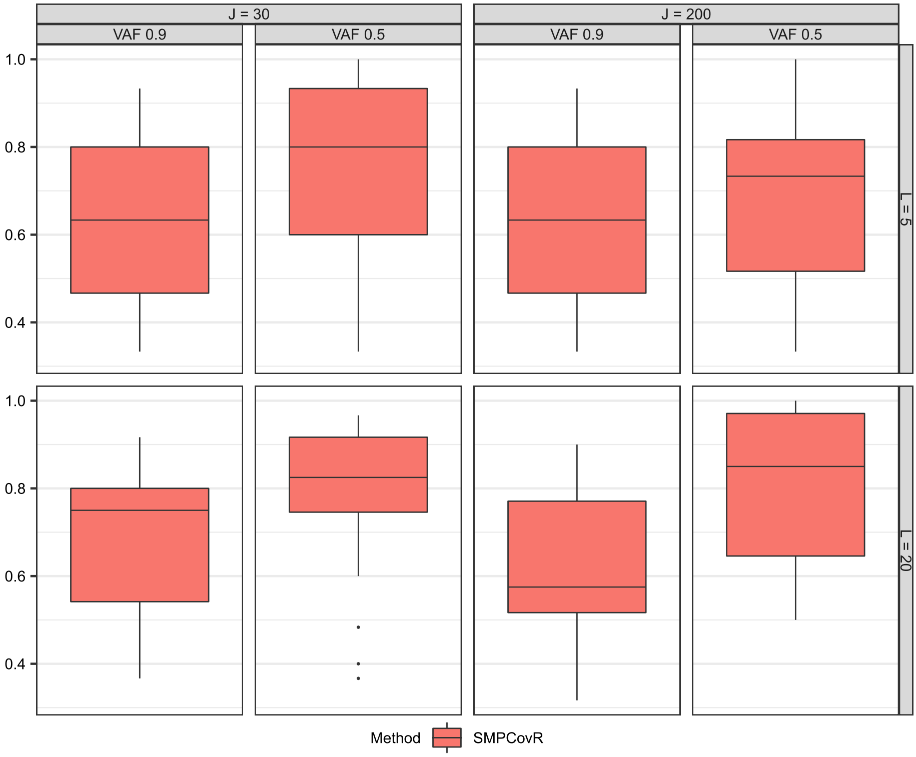

Fig. 9 Boxplots of the correct classification rate for the $\mathbf{P}^{(Y)}$. Each panel corresponds to one of the 8 conditions. From the simulation study with weights not in simple structure

Appendix 7. Simulation study: proportion of outcomes correctly identified as active and inactive by SMPCovR

See Fig. 10.

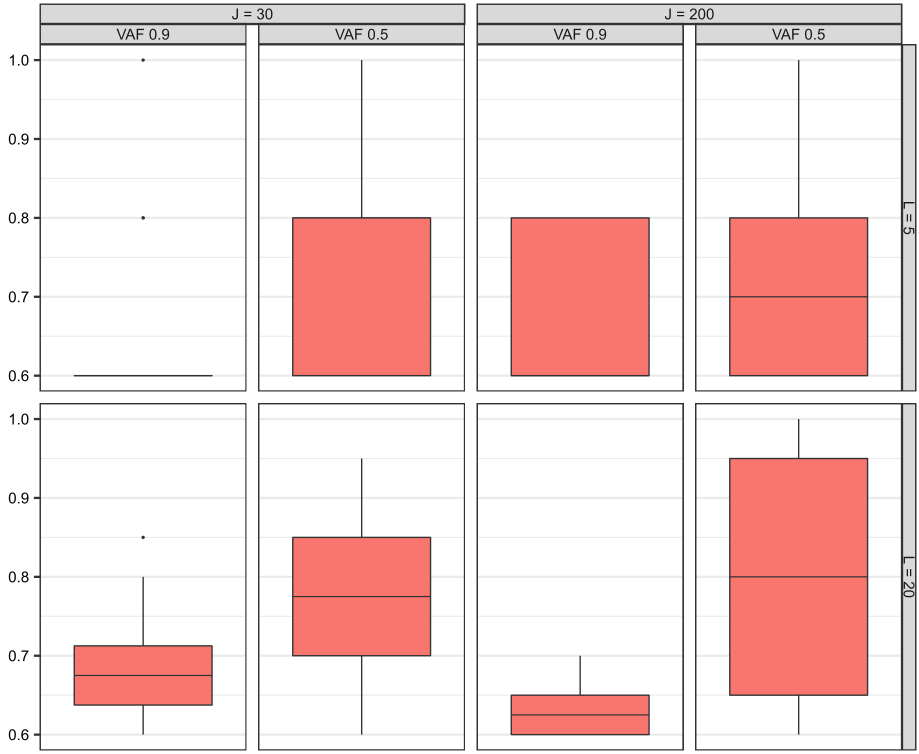

Fig. 10 Proportion of outcomes correctly identified as active and inactive. Each panel corresponds to one of the 8 conditions

Appendix 8. Simulation study: discrepancy from zero regression coefficients from SPCovR and sPLS

See Fig. 11.

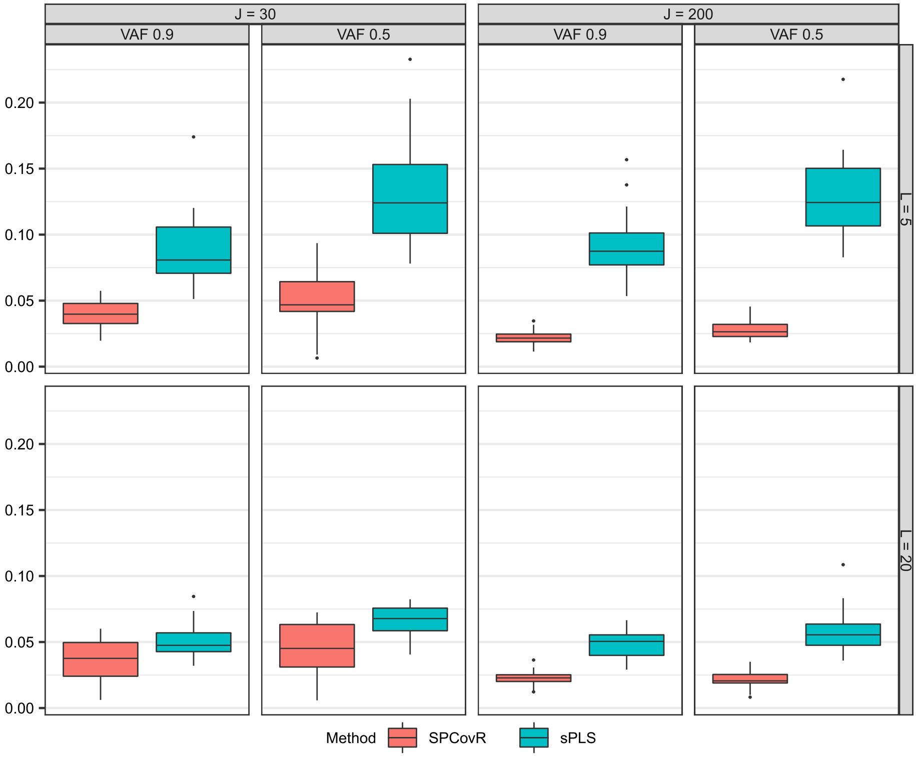

Fig. 11 Mean absolute discrepancy of the zero regression coefficients. Each panel corresponds to one of the 8 conditions

Appendix 9. The scree test with acceleration factor conducted to determine the number of covariates for the PCS dataset

See Fig. 12.

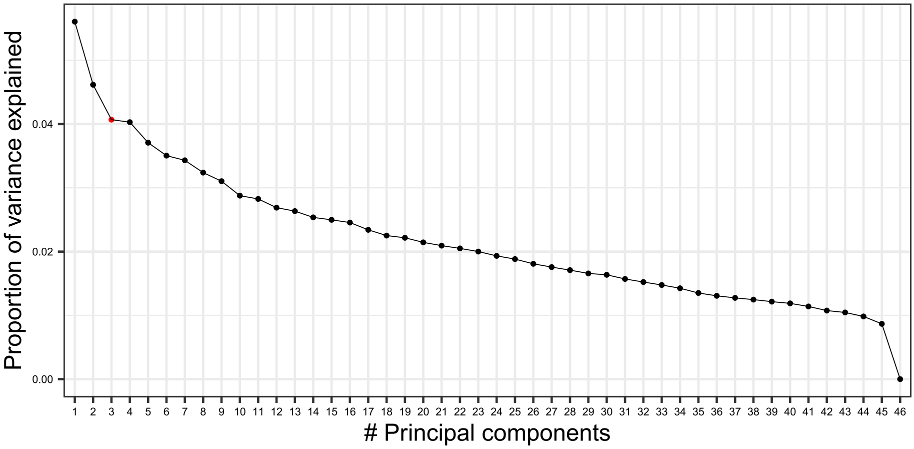

Fig. 12 It can be seen that the sharpest change of slopes occurs at the third principal component. Therefore, the number of SMPCovR covariate is determined as two

Appendix 10. The R^2 measures computed on the PCS dataset

$R^2_{\text{fit}_{all}}$, $R^2_{\text{fit}_{sub}}$, $R^2_{\text{fit}_{active}}$, $R^2_{\text{loocv}_{all}}$, $R^2_{\text{loocv}_{sub}}$ and $R^2_{\text{loocv}_{active}}$ employed to evaluate the models fitted on the PCS dataset were calculated by the following equations.

$R^2_{\text{fit}_{all}}$ is the R^2 measure computed on the in-sample data on the basis of all of the outcome variables. This can be considered as the conventional R^2 measure:

$$R^2_{\text{fit}_{all}} = 1 - \frac{\left\| \mathbf{Y} - \mathbf{X}\hat{\mathbf{W}}\hat{\mathbf{P}}^{(Y)^T} \right\|_2^2}{\|\mathbf{Y}\|_2^2}. \tag{16}$$

The $R^2_{\text{fit}_{sub}}$ measure is computed on the in-sample data, however on the basis of outcome variables selected as being active by the SMPCovR model:

$$R^2_{\text{fit}_{sub}} = 1 - \frac{\left\| \mathbf{Y}_{K^*} - \mathbf{X}\hat{\mathbf{W}}\hat{\mathbf{P}}_{K^*}^{(Y)^T} \right\|_2^2}{\|\mathbf{Y}_{K^*}\|_2^2} \tag{17}$$

with the subscript $_{K^*}$ indicating a subset within the sequence of indices for outcome variables $K^* \subseteq \{1, 2, \dots, 16\}$. It comprises of indices corresponding to the active outcomes selected by SMPCovR. As reported in Table 4, $\mathbf{Y}_{K^*}$ would comprise of the 11 following outcomes: sneezing, runny nose, nasal congestion, cough, sore throat, headache, malaise, chest congestion, sinus pain, fever and poor appetite. Since an outcome variable is removed from the SMPCovR model if its corresponding row in the estimated regression coefficients matrix $\hat{\mathbf{P}}^{(Y)}$ is a zero-vector, the indices of non-zero rows of $\hat{\mathbf{P}}^{(Y)}$ make up the set K^*. $\hat{\mathbf{P}}^{(Y)}_{K^*}$ denotes the submatrix of $\hat{\mathbf{P}}^{(Y)}$ with non-zero rows.

Similarly, the $R^2_{\mathrm{fit}_{active}}$ measure is computed on the in-sample data, on the basis of outcome variables deemed active according to the theory. These are the 8 symptoms known to comprise the common cold: headache, sneezing, chills, sore throat, runny nose, nasal congestion, cough and malaise (Jackson et al., 1958).

$$R^2_{\mathrm{fit}_{active}} = 1 - \frac{\left\| \mathbf{Y}_{K^\star} - \mathbf{X}\hat{\mathbf{W}}\hat{\mathbf{P}}^{(Y)^T}_{K^\star} \right\|_2^2}{\left\| \mathbf{Y}_{K^\star} \right\|_2^2} \tag{18}$$

with the subscript $_{K^\star}$ indicating a subset within the sequence of indices for outcome variables $K^\star \subseteq \{1, 2, \dots, 16\}$. It comprises of indices corresponding to the 8 symptoms.

$R^2_{\mathrm{loocv}_{all}}$ is calculated via leave-one-out CV. All of the outcome variables in the PCS dataset are incorporated:

$$R^2_{\mathrm{loocv}_{all}} = 1 - \frac{\left\| \mathbf{y}^{\mathrm{test}} - \mathbf{x}^{\mathrm{test}^T}\hat{\mathbf{W}}\hat{\mathbf{P}}^{(Y)^T} \right\|_2^2}{\left\| \mathbf{y}^{\mathrm{test}} \right\|_2^2} \tag{19}$$

where $\mathbf{y}^{test}$ and $\mathbf{x}^{test}$ refer to the outcome and predictor variables in the CV test set (it is a vector, since leave-one-out CV uses one observation unit for each test set).

$R^2_{\mathrm{loocv}_{sub}}$ is also calculated via leave-one-out CV, but on the basis of active outcome variables selected by the SMPCovR model:

$$R^2_{\mathrm{loocv}_{sub}} = 1 - \frac{\left\| \mathbf{y}^{\mathrm{test}}_{K^*} - \mathbf{x}^{\mathrm{test}^T}\hat{\mathbf{W}}\hat{\mathbf{P}}^{(Y)^T}_{K^*} \right\|_2^2}{\left\| \mathbf{y}^{\mathrm{test}}_{K^*} \right\|_2^2}. \tag{20}$$

As for the formula for $R^2_{\mathrm{loocv}_{all}}$, $\mathbf{y}^{test}$ and $\mathbf{x}^{test}$ refer to the outcome and predictor variables in the CV test set. As for the formula for $R^2_{\mathrm{fit}_{sub}}$, the subscript $_{K^*}$ denotes a subset within the sequence of indices for outcome variables $K^* \subseteq \{1, 2, \dots, 16\}$. It comprises of indices corresponding to the active outcomes selected by SMPCovR. As reported in Table 4, $\mathbf{Y}_{K^*}$ would comprise of the 11 following outcomes: sneezing, runny nose, nasal congestion, cough, sore throat, headache, malaise, chest congestion, sinus pain, fever and poor appetite.

Lastly, the $R^2_{\mathrm{loocv}_{active}}$ measure is computed by leave-one-out CV, on the basis of the 8 outcome variables deemed active according to the theory on common cold.

$$R^2_{\mathrm{loocv}_{sub}} = 1 - \frac{\left\| \mathbf{y}^{\mathrm{test}}_{K^\star} - \mathbf{x}^{\mathrm{test}^T}\hat{\mathbf{W}}\hat{\mathbf{P}}^{(Y)^T}_{K^\star} \right\|_2^2}{\left\| \mathbf{y}^{\mathrm{test}}_{K^\star} \right\|_2^2} \tag{21}$$

with the subscript $_{K^\star}$ indicating a subset within the sequence of indices for outcome variables $K^\star \subseteq \{1, 2, \dots, 16\}$. It comprises of indices corresponding to the 8 symptoms.

Appendix 11. Model selection for SPCovR and sPLS for the PCS dataset

For SPCovR, the α parameter, lasso parameter and the ridge parameter were tuned by 5-fold CV. We adopted the sequence [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9] for α. For both the lasso and the ridge parameters, 0 and equally distanced sequence of size 9 from 10^{-5} to 0.5 on the natural log scale was employed as the range. Crossing the three ranges, $9 \times 10 \times 10 = 900$ different models were evaluated by CV. With regards to sPLS, the range considered for the number of non-zero coefficients per component was the multiples of 6 running from 6 to 180 along with 1 and 187 (minimal and maximal number of non-zero coefficients). With the number of components fixed at two, the 5-fold CV was performed for $32^2 = 1024$ total models. After the CV, the 1SE rule was used to select the final model for both methods.

Tables 7 and 8 present the weights and regression coefficients from the SPCovR model. The model is comprised of 100 and 55 non-zero weights for each covariate. The method retrieved much more non-zero weights than SMPCovR across the multiple predictor themes: blood chemistry, health practices, psychosocial assessment scales, childhood experiences and daily interviews.

The weights and regression coefficients found by sPLS are provided in Table 9. The model was comprised of much smaller number of non-zero weights than SMPCovR and SPCovR. The first covariate only consisted of TNF alpha, one of the 7 variables regarding nasal cytokine. The second covariate was associated with 6 variables from daily interviews. With respect to the regression coefficients, most of them were far away from zero. This was also in line with our finding in the simulation study, where sPLS did not provide near-zero coefficients as estimates for the true zero regression coefficients (see Appendix 8).

Table 7 Weights derived by SPCovR from the PCS dataset. $\hat{\mathbf{w}}_1$ and $\hat{\mathbf{w}}_2$ indicate the weights corresponding to the first and second covariates respectively

$\hat{w}_1$	
	1
IL-6	0.025
Red blood cells	0.089
Neutrophil proportion	0.007
Corpuscular volume	−0.108
Corpuscular Hgb (weight)	−0.104
Chloride	−0.032
Calcium	0.054
Non-fasting glucose	0.013
Protein	0.011
Albumin	0.011
# weekdays alcohol	−0.006
# drinks on weekdays	−0.068
# weekly physical activity	−0.002
PSQI: too hot	0.014
FES: conflict	0.005
FES: total	−0.011
PBI: care	−0.057
PBI: total	−0.038
RF: total	0.051
Stress at 10	0.017
Stress at 15	0.017
Tot. stress upbringing	0.022
PANAS: fear	0.020
PANAS: hostility	0.040
PANAS: guilt	0.110
PANAS: sadness	0.115
PANAS: joviality	−0.177
PANAS: self−assurance	−0.088
PANAS: attentive	−0.055
PANAS: fatigue	0.110
PANAS: serenity	−0.143
PANAS: surprise	−0.077
PANAS: negative	0.075
PANAS: positive	−0.126
IPIP: conscientiousness	−0.095
IPIP: emotional stability	−0.164
IPIP: extraversion	−0.054
IPIP: agreeableness	−0.116
LOT-R: optimism	−0.095
Opener: total	−0.170
COMM: total	−0.012
CM-Ho: total	0.015
Shyness	0.042
ECR: avoid	0.178

Table 7 (continued)

$\hat{w}_1$	
	1
ECR: anxiety	0.059
TSC: network size	−0.037
ISEL: total	−0.073
GS-ISEL: total	−0.124
SNI: total roles	−0.022
PCOM: total	−0.141
Loneliness	0.139
PSS (10-item): total	0.170
PSS (4-item): total	0.139
PWB: self-acceptance	−0.203
PWB: env. mastery	−0.184
PWB: positive rel	−0.176
PWB: purpose	−0.061
PWB: total	−0.218
ERQ: suppression	0.076
# days: hugs	−0.138
# days: tension	0.019
# days: romance	−0.107
Daily: # domains interacted	−0.011
Daily: # social partners	−0.006
Daily: # social interaction	−0.068
# days: social interaction	−0.009
Daily: # other activities	−0.013
Daily: # errands	−0.021
Daily: # leisure activities	0.037
Daily: nap duration	−0.020
# days: rested	−0.026
Daily: sleep duration	0.009
Daily: minutes in bed	0.013
Daily: loneliness subscale	0.164
Daily: isolated	0.125
Daily: loneliness	0.139
Daily: negative subscale	0.143
Daily: negative affect	0.141
Daily: fatigue subscale	0.058
Daily: depressed subscale	0.139
Daily: anxiety subscale	0.086
Daily: happy	−0.146
Daily: tired	0.040
Daily: calm	−0.102
Daily: sad	0.078
Daily: energetic	−0.106
Daily: hostile	0.087
Daily: on edge	0.060

Table 7 (continued)

$\hat{w}_1$	
	1
Daily: fatigue	0.055
Daily: lively	−0.131
Daily: angry	0.085
Daily: cheerful	−0.131
Daily: tense	0.099
Daily: at ease	−0.112
Daily: unhappy	0.161
Daily: well−being subscale	−0.148
Daily: vigor subscale	−0.133
Daily: calm subscale	−0.115
Daily: positive affect	−0.179
Daily: anger subscale	0.101
$\hat{\mathbf{w}}_2$	
	2
IL-6	0.066
IL-10	0.074
TNF alpha	0.709
Red blood cells	0.053
Corpuscular Hgb (weight)	−0.073
Corpuscular Hgb (conc)	−0.089
Potassium	0.020
Bilirubin	0.061
Non-fasting glucose	0.026
Urea nitrogen	0.069
Creatine	0.004
# weekdays alcohol	0.004
# weekend days alcohol	0.036
# drinks on weekdays	0.124
# drinks on weekend days	0.008
PSQI: use bathroom	0.141
PSQI: too hot	−0.189
PSQI: bad dreams	−0.008
FES: expressiveness	0.069
FES: conflict	0.009
PSP: total	−0.058
Stress at 10	−0.056
PLI: physical environment	0.097
PLI: social environment	−0.054
PANAS: guilt	−0.032
PANAS: fatigue	0.087
IPIP: extraversion	0.058
COMM: total	0.198
TSC: network size	−0.079
TSC: indirect	0.078

Table 7 (continued)

$\hat{w}_1$	
	1
GS-ISEL: total	0.102
NAR: total	−0.110
SNI: # network members	0.013
Social participation	−0.237
Perceived community score	0.231
Loneliness	0.019
ERQ: reappraisal	−0.023
ERQ: suppression	−0.014
# days: sharing	0.116
Daily: # social partners	0.008
Daily: # errands	0.028
Daily: # leisure activities	−0.157
# days: at home	−0.140
Daily: loneliness subscale	−0.040
Daily: isolated	−0.003
Daily: fatigue subscale	0.029
Daily: happy	0.009
Daily: tired	0.188
Daily: calm	0.082
Daily: hostile	0.022
Daily: on edge	−0.029
Daily: fatigue	0.243
Daily: lively	−0.013
Daily: calm subscale	0.039
Daily: anger subscale	0.057

PSQI = Pittsburgh Sleep Quality Index; FES = Family Environment Scale; PBI = Parental Bonding Instrument; RF = Risky Families Questionnaire; PANAS = Positive and Negative Affect Schedule; IPIP = International Personality Item Pool; LOT-R = Revised Life Orientation Test; COMM = Communal Orientation; CM-Ho = Cook-Medley Hostility Scale; ECR = Experiences in Close Relationships Scale; TSC = Tucker Social Control; ISEL = Interpersonal Support Evaluation List; GS-ISEL = Giving Support - Interpersonal Support Evaluation List; SNI = Social Network Index; PCOM = Perceived Community Scale; PSS = Perceived Stress Scale; PWB = Ryff's Psychological Well-being Scale; ERQ = Emotion Regulation Questionnaire; PSP = Parental Social Participation; PLI: Childhood Places You've Lived Interview; NAR = Negative Aspects of Relationships; more information about the predictors can be found on: https://www.cmu.edu/common-cold-project/measures-by-study/index.html

Table 8 Regression coefficients derived by SPCovR from the PCS dataset

$\hat{P}^{(Y)}$		
	1	2
Sneezing	0.056	0.513
Runny nose	0.059	0.454
Nasal congestion	0.057	0.496
Cough	0.085	0.548
Sore throat	0.025	0.445
Headache	0.028	0.288
Chills	0.019	0.139
Malaise	0.069	0.319
Chest congestion	0.051	0.426
Sinus pain	0.039	0.500
Earache	0.045	0.034
Muscle ache	−0.018	0.318
Joint ache	−0.012	0.044
Sweating	0.061	0.193
Fever	0.043	0.427
Poor appetite	0.084	0.405

Table 9 Weights and regression coefficients derived by sPLS from the PCS dataset. $\hat{\mathbf{w}}_1$ and $\hat{\mathbf{w}}_2$ indicate the weights corresponding to the first and second covariates respectively

$\hat{w}_1$		
	1	
TNF alpha	1	
$\hat{w}_2$		
	2	
Daily loneliness	0.033	
Daily negative affect & Fatigue	0.411	
Daily negative affect	0.019	
Daily fatigue subscale	0.608	
Daily tiredness	0.436	
Daily fatigue	0.518	
$\hat{P}^{(Y)}$		
	1	2
Sneezing	−0.200	−0.259
Runny nose	−0.245	−0.247
Nasal congestion	−0.240	−0.285
Cough	−0.341	−0.398
Sore throat	−0.316	−0.264
Headache	−0.088	−0.224
Chills	−0.047	−0.222
Malaise	−0.090	−0.308
Chest congestion	−0.328	−0.259
Sinus pain	−0.404	−0.237
Earache	−0.022	−0.394

Table 9 (continued)

$\hat{P}^{(Y)}$		
	1	2
Muscle ache	−0.198	−0.104
Joint ache	0.126	−0.086
Sweating	−0.125	0.062
Fever	−0.354	0.059
Poor appetite	−0.376	−0.253

Appendix 12. Simulation study linear model

A simulation study where the data is generated from a linear model instead of a covariate model is conducted in this section. The aim is to examine the performance of SMPCovR, SPCovR and sPLS under datasets not generated from a covariate model.

Design and procedure

The datasets were generated from a linear model, with 50 predictor variables and 10 outcome variables. The true predictor variables $\mathbf{X}^{\mathrm{true}}$ were drawn from the multivariate normal distribution with a zero vector for the mean and an identity matrix for the covariance. The $\mathbf{B}$ (size 50×10) matrix consisting of true regression parameters was defined such that only the first seven outcome variables are linked with the first seven predictor variables. The remaining three outcomes were inactive, while the remaining 43 predictors were redundant. The (7×7) upper left submatrix of the regression parameters is as the following:

$$\begin{array}{cccccccc}
 & \mathrm{y}_1^{\mathrm{true}} & \mathrm{y}_2^{\mathrm{true}} & \mathrm{y}_3^{\mathrm{true}} & \mathrm{x}_4^{\mathrm{true}} & \mathrm{y}_5^{\mathrm{true}} & \mathrm{y}_6^{\mathrm{true}} & \mathrm{y}_7^{\mathrm{true}} \\
\mathrm{x}_1^{\mathrm{true}} & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
\mathrm{x}_2^{\mathrm{true}} & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
\mathrm{x}_3^{\mathrm{true}} & 0 & 0 & 0.4 & 0.4 & 0.4 & 0.4 & 0.4 \\
\mathrm{x}_4^{\mathrm{true}} & 0 & 0 & 0.4 & 0.4 & 0.4 & 0.4 & 0.4 \\
\mathrm{x}_5^{\mathrm{true}} & 0 & 0 & 0.4 & 0.4 & 0.4 & 0.4 & 0.4 \\
\mathrm{x}_6^{\mathrm{true}} & 0 & 0 & 0.4 & 0.4 & 0.4 & 0.4 & 0.4 \\
\mathrm{x}_7^{\mathrm{true}} & 0 & 0 & 0.4 & 0.4 & 0.4 & 0.4 & 0.4
\end{array}$$

The remaining elements from the $\mathbf{B}$ matrix were defined as zero. The true outcome variables were generated by the matrix product: $\mathbf{Y}^{\mathrm{true}} = \mathbf{X}^{\mathrm{true}}\mathbf{B}$.

As the final step of data generation, the residuals for the predictors and outcomes, $\mathbf{E}^{(X)}$ and $\mathbf{E}^{(Y)}$, drawn from multivariate normal distribution with a zero vector for the mean and an identity matrix for the covariance, were added to obtain the observed data: $\mathbf{X} = \mathbf{X}^{\mathrm{true}} + \mathbf{E}^{(X)}, \mathbf{Y} = \mathbf{Y}^{\mathrm{true}} + \mathbf{E}^{(Y)}$. The residual matrices were scaled to control the level of VAF by the true matrices, according to the design factors given below.

Study setup

1. Number of observations each in the train and test set: [100], [25]
2. Proportion of variance in observed data $\mathbf{X}$ and $\mathbf{Y}$ explained by the true matrices $\mathbf{X}^{\mathrm{true}}$ and $\mathbf{Y}^{\mathrm{true}}$ (VAF): [90%], [50%]

As the numbers of predictors and outcomes are fixed at $J = 50$ and $K = 10$, when the number of observations $I = 25$, the dataset is high-dimensional. Crossing the design factors and generating 20 datasets per condition, $2 \times 2 \times 20 = 80$ datasets were produced. Three different analyses were administered to each of these datasets: SMPCovR, SPCovR and sPLS.

Model selection

The procedure detailed in Sect. 2.3.3 is followed for SMPCovR. The number of covariates selected was two, which is aligned with the model used for data generation. From the regression parameters **B**, it can be observed that the there are two groups of correlated active outcomes. The following ranges of parameters were used to conduct the two rounds of 5-fold CV with the 1SE rule.

SMPCovR tuning parameter ranges for the first round of CV

1. α: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
2. γ_1: 0 and equally distanced sequence of size 14 from 10^{-5} to 0.5 on the natural log scale
3. γ_2: 0 and equally distanced sequence of size 14 from 10^{-5} to 0.5 on the natural log scale

SMPCovR tuning parameter ranges for the second round of CV

1. λ_1: 0 and equally distanced sequence of size 14 from 10^{-5} to 0.5 on the natural log scale
2. λ_2: 0 and equally distanced sequence of size 14 from 10^{-5} to 0.5 on the natural log scale

With regards to SPCovR, with the number of covariates fixed at two, the following ranges of parameters were used for 5-fold CV with the 1SE rule.

SPCovR tuning parameter ranges

- α: 0.1, 0.3, 0.5, 0.7, 0.9
- λ_1: 0 and equally distanced sequence of size 9 from 10^{-5} to 0.5 on the natural log scale
- λ_1: 0 and equally distanced sequence of size 9 from 10^{-5} to 0.5 on the natural log scale

Finally, with the number of covariates for sPLS fixed at three, the number of weights that link the predictors to covariates was chosen via 5-fold CV with the 1SE rule. The range of [1, 2, 3, 5, 10, 20, 30, 40, 50] non-zero coefficients per covariate was considered.

After the model selection procedures, the two different types of out-of-sample R^2 measures (6, 7) employed in the original simulation study in the main text were used again to assess the quality of prediction.

Results

Out-of-sample R^2_{out} and $R^2_{\text{out}_{active}}$

Unlike the findings in the original simulation study using the covariate model, the prediction performances of the methods evaluated with R^2_{out} and $R^2_{\text{out}_{active}}$ are in agreement. When the true matrices accounted for 90% of the variance in the observed data under low dimensionality, all three methods exhibited similar quality in prediction. However, when the VAF reduced to 50% or the dimensionality was high, SMPCovR demonstrated superior performance. It appears that when the data is generated from a linear model,

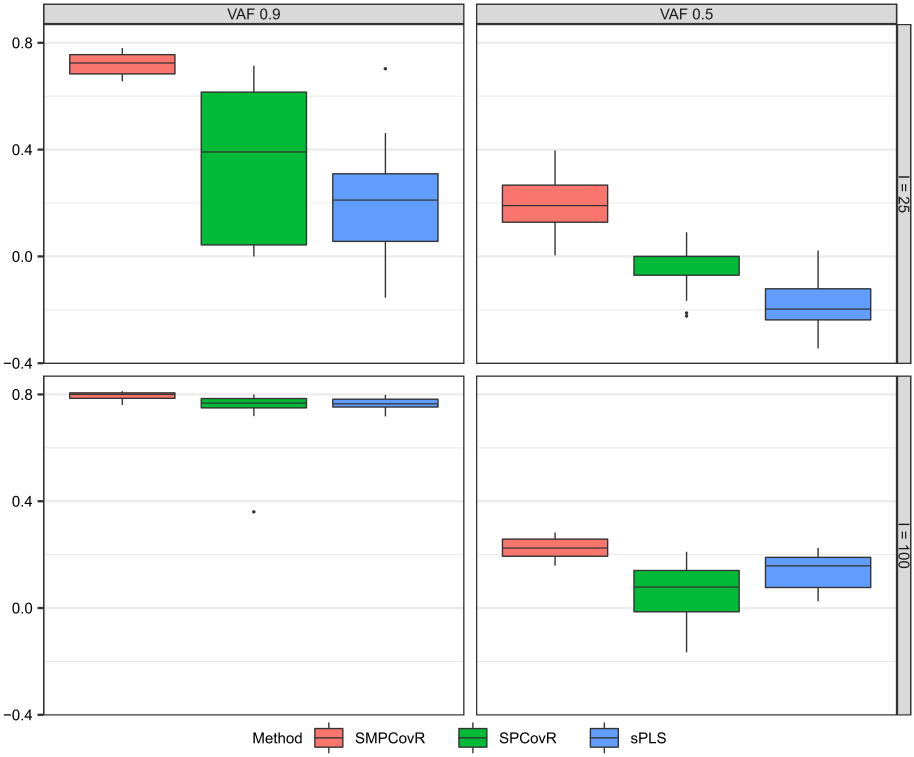

Fig. 13 Boxplots of the out of sample R^2_{out}. Each panel corresponds to one of the 8 conditions

SMPCovR with outcome selection is better suited for prediction for both the active and the entire set of outcome variables, than SPCovR and sPLS. As the effect of overfitting is expected to be more pronounced under these settings of lower VAF and higher dimensionality, this outperformance of SMPCovR could be attributed to its robustness against overfitting due to outcome selection, compared to the two other methods (Figs. 13, 14).

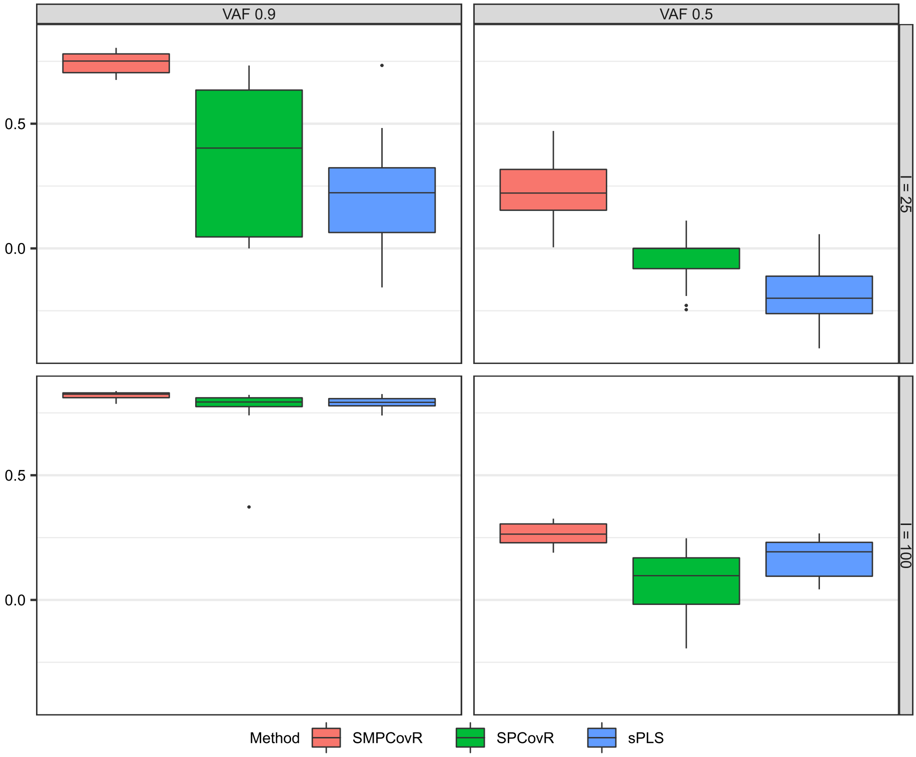

Fig. 14 Boxplots of the out of sample $R^2_{\mathrm{out}_{active}}$. Each panel corresponds to one of the 8 conditions

Author Contributions SP, KVD and EC: method conception, study design, interpretation of results, discussion. SP: algorithm implementation, manuscript drafting. All authors critically read and approved of the manuscript.

Funding 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.

Availability of data and materials Code used to generate the results reported in the manuscript is available on Github: https://github.com/soogs/SMPCovR. Pittsburgh Cold Study data were collected by the Laboratory for the Study of Stress, Immunity, and Disease at Carnegie Mellon University under the directorship of Sheldon Cohen, PhD. The data were accessed via the Common Cold Project website (www.commoncoldproject.com; grant number NCCIH AT006694).

Declaration

Conflict of interest The authors have no Conflict of interest to declare that are relevant to the content of this article.

Ethics approval Ethics approval is not applicable to the current research and manuscript.

Consent to participate Consent to participate is not applicable to the current research and manuscript.

Consent for publication Consent for publication is not applicalbe to the current research and manuscript.

Code availability The implementation of SMPCovR was done in R and Rcpp, which can be found on Github: https://github.com/soogs/SMPCovR.

References

An, B., & Zhang, B. (2017). Simultaneous selection of predictors and responses for high dimensional multivariate linear regression. *Statistics & Probability Letters, 127*, 173–177.

Boqué, R., & Smilde, A. K. (1999). Monitoring and diagnosing batch processes with multiway covariates regression models. *AIChE Journal, 45*(7), 1504–1520.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245–276.

Ceulemans, E., Van Mechelen, I., & Leenen, I. (2007). The local minima problem in hierarchical classes analysis: An evaluation of a simulated annealing algorithm and various multistart procedures. *Psychometrika, 72*(3), 377–391.

Chen, K., Chan, K.-S., & Stenseth, N. C. (2012). Reduced rank stochastic regression with a sparse singular value decomposition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 74*(2), 203–221.

Chen, L., & Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association, 107*(500), 1533–1545.

Chung, D., & Keles, S. (2010). Sparse partial least squares classification for high dimensional data. *Statistical Applications in Genetics and Molecular Biology*. https://doi.org/10.2202/1544-6115.1492

Cohen, S., Kamarck, T., Mermelstein, R., et al. (1994). Perceived stress scale. *Measuring Stress: A Guide for Health and Social Scientists, 10*(2), 1–2.

Cohen, S., Mermelstein, R., Kamarck, T. & Hoberman, H. M. (1985). Measuring the functional components of social support. In *Social support: Theory, research and applications*, pp. 73–94. Springer.

Cook, R. D., Helland, I., & Su, Z. (2013). Envelopes and partial least squares regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75*(5), 851–877.

Cook, R. D., Li, B., & Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica, 20*, 927–960.

De Jong, K. A. (1975). *An analysis of the behavior of a class of genetic adaptive systems*. University of Michigan.

De Jong, S., & Kiers, H. A. (1992). Principal covariates regression: part i. theory. *Chemometrics and Intelligent Laboratory Systems, 14*(1–3), 155–164.

de Schipper, N. C., & Van Deun, K. (2018). Revealing the joint mechanisms in traditional data linked with big data. *Zeitschrift für Psychologie, 226*(4), 212–231.

de Schipper, N. C., & Van Deun, K. (2021). Model selection techniques for sparse weight-based principal component analysis. *Journal of Chemometrics, 35*(2), e3289.

Ferré, L. (1995). Selection of components in principal component analysis: A comparison of methods. *Computational Statistics & Data Analysis, 19*(6), 669–682.

Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association, 78*(383), 553–569.

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology, 78*(2), 350.

Friedman, J., Hastie, T. & Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. *arXiv preprint*arXiv:1001.0736.

Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N. & Qian, J. (2021). Package 'glmnet'. *CRAN R Repositary 595*.

Goldberg, L. R., et al. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe, 7*(1), 7–28.

Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology, 85*(2), 348.

Guo, C., Kang, J., & Johnson, T. D. (2022). A spatial bayesian latent factor model for image-on-image regression. *Biometrics, 78*(1), 72–84.

Gvaladze, S., Vervloet, M., Van Deun, K., Kiers, H. A., & Ceulemans, E. (2021). Pcovr2: A flexible principal covariates regression approach to parsimoniously handle multiple criterion variables. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01508-y

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction,* (Vol. 2). Springer.

Helfrecht, B. A., Cersonsky, R. K., Fraux, G., & Ceriotti, M. (2020). Structure-property maps with kernel principal covariates regression. *Machine Learning: Science and Technology, 1*(4), 045021.

Henry, R. C., Park, E. S., & Spiegelman, C. H. (1999). Comparing a new algorithm with the classic methods for estimating the number of factors. *Chemometrics and Intelligent Laboratory Systems, 48*(1), 91–97.

Hu, J., Huang, J., Liu, X., & Liu, X. (2022). Response best-subset selector for multivariate regression with high-dimensional response variables. *Biometrika, 110*(1), 205–223.

Hu, J., Liu, X., Liu, X., & Xia, N. (2022). Some aspects of response variable selection and estimation in multivariate linear regression. *Journal of Multivariate Analysis, 188*, 104821.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis, 5*(2), 248–264.

Jackson, D. A. (1993). Stopping rules in principal components analysis: A comparison of heuristical and statistical approaches. *Ecology, 74*(8), 2204–2214.

Jackson, G. G., Dowling, H. F., Spiesman, I. G., & Boand, A. V. (1958). Transmission of the common cold to volunteers under controlled conditions: I: The common cold as a clinical entity. *AMA Archives of Internal Medicine, 101*(2), 267–278.

Kawano, S. (2021). Sparse principal component regression via singular value decomposition approach. *Advances in Data Analysis and Classification, 15*, 795–823.

Kawano, S., Fujisawa, H., Takada, T., & Shiroishi, T. (2015). Sparse principal component regression with adaptive loading. *Computational Statistics & Data Analysis, 89*, 192–203.

Kiers, H. A., & Smilde, A. K. (2007). A comparison of various methods for multivariate regression with highly collinear variables. *Statistical Methods and Applications, 16*, 193–228.

Kim, J., Zhang, Y., & Pan, W. (2016). Powerful and adaptive testing for multi-trait and multi-snp associations with gwas and sequencing data. *Genetics, 203*(2), 715–731.

Kirkpatrick, S., Gelatt, C. D., Jr., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science, 220*(4598), 671–680.

Lê Cao, K.-A., Boitard, S., & Besse, P. (2011). Sparse pls discriminant analysis: Biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics, 12*(1), 253.

Lê Cao, K.-A., Rossouw, D., Robert-Granié, C., & Besse, P. (2008). A sparse pls for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*. https://doi.org/10.2202/1544-6115.1390

Luo, S. (2020). Variable selection in high-dimensional sparse multiresponse linear regression models. *Statistical Papers, 61*(3), 1245–1267.

Mayer, J., Rahman, R., Ghosh, S., & Pal, R. (2018). Sequential feature selection and inference using multivariate random forests. *Bioinformatics, 34*(8), 1336–1344.

Miller, L. C., Berg, J. H., & Archer, R. L. (1983). Openers: Individuals who elicit intimate self-disclosure. *Journal of Personality and Social Psychology, 44*(6), 1234.

Monto, A. S., Gravenstein, S., Elliott, M., Colopy, M., & Schweinle, J. (2000). Clinical signs and symptoms predicting influenza infection. *Archives of Internal Medicine, 160*(21), 3243–3247.

Moos, R. H. (1990). Conceptual and empirical approaches to developing family-based assessment procedures: Resolving the case of the family environment scale. *Family Process, 29*(2), 199–208.

Nelemans, S. A., Van Assche, E., Bijttebier, P., Colpin, H., Van Leeuwen, K., Verschueren, K., Claes, S., Van Den Noortgate, W., & Goossens, L. (2019). Parenting interacts with oxytocin polymorphisms to predict adolescent social anxiety symptom development: A novel polygenic approach. *Journal of Abnormal Child Psychology, 47*(7), 1107–1120.

Obozinski, G., Taskar, B., & Jordan, M. (2006). Multi-task feature selection. *Statistics Department, , UC Berkeley, Tech. Rep, 2*(2.2), 2.

Oladzad, A., Porch, T., Rosas, J. C., Moghaddam, S. M., Beaver, J., Beebe, S. E., Burridge, J., Jochua, C. N., Miguel, M. A., Miklas, P. N., et al. (2019). Single and multi-trait gwas identify genetic factors associated with production traits in common bean under abiotic stress environments. *G3: Genes, Genomes, Genetics, 9*(6), 1881–1892.

Park, M. Y., & Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics, 9*(1), 30–50.

Park, S., Ceulemans, E., & Van Deun, K. (2020). Sparse common and distinctive covariates regression. *Journal of Chemometrics, 35*, e3270.

Park, S., Ceulemans, E., & Van Deun, K. (2023). Logistic regression with sparse common and distinctive covariates. *Behavior Research Methods, 55*, 4143.

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J. R., & Wang, P. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The Annals of Applied Statistics, 4*(1), 53.

Raîche, G. & Magis, D. (2020). Package 'nfactors'. *Repository CRAN*, 1–58.

Raîche, G., Walls, T. A., Magis, D., Riopel, M., & Blais, J.-G. (2013). Non-graphical solutions for cattell's scree test. *Methodology*. https://doi.org/10.1027/1614-2241/a000051

Ryff, C. D. (1989). Happiness is everything, or is it? Explorations on the meaning of psychological well-being. *Journal of Personality and Social Psychology, 57*(6), 1069.

Schönemann, P. H. (1966). A generalized solution of the orthogonal procrustes problem. *Psychometrika, 31*(1), 1–10.

Stein, J. L., Hua, X., Lee, S., Ho, A. J., Leow, A. D., Toga, A. W., Saykin, A. J., Shen, L., Foroud, T., Pankratz, N., et al. (2010). Voxelwise genome-wide association study (vgwas). *Neuroimage, 53*(3), 1160–1174.

Steinley, D., & Brusco, M. J. (2008). Selection of variables in cluster analysis: An empirical comparison of eight procedures. *Psychometrika, 73*(1), 125–144.

Su, Z., Zhu, G., Chen, X., & Yang, Y. (2016). Sparse envelope model: Efficient estimation and response variable selection in multivariate linear regression. *Biometrika, 103*(3), 579–593.

Taylor, M. K., Sullivan, D. K., Ellerbeck, E. F., Gajewski, B. J., & Gibbs, H. D. (2019). Nutrition literacy predicts adherence to healthy/unhealthy diet patterns in adults with a nutrition-related chronic condition. *Public Health Nutrition, 22*(12), 2157–2169.

Ten Berge, J. M. (1993). *Least squares optimization in multivariate analysis*. Leiden University Leiden: DSWO Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288.

Tucker, J. S. (2002). Health-related social control within older adults' relationships. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 57*(5), 387–395.

Van Deun, K., Crompvoets, E. A., & Ceulemans, E. (2018). Obtaining insights from high-dimensional data: Sparse principal covariates regression. *BMC Bioinformatics, 19*(1), 104.

Vervloet, M., Van Deun, K., Van den Noortgate, W., & Ceulemans, E. (2016). Model selection in principal covariates regression. *Chemometrics and Intelligent Laboratory Systems, 151*, 26–33.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The panas scales. *Journal of Personality and Social Psychology, 54*(6), 1063.

Wilderjans, T. F., Ceulemans, E., & Meers, K. (2013). Chull: A generic convex-hull-based model selection method. *Behavior Research Methods, 45*(1), 1–15.

Wold, H. O. A. (1982). Soft modeling: The basic design and some extensions. In K. G. Jöreskog & H. O. A. Wold (Eds.), *Systems Under Indirect Observation* (Vol. 2, pp. 1–53). Amsterdam: North-Holland.

Wold, S., Ruhe, A., Wold, H., & Dunn, W. Iii. (1984). The collinearity problem in linear regression: The partial least squares (pls) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing, 5*(3), 735–743.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68*(1), 49–67.

Zamdborg, L., & Ma, P. (2009). Discovery of protein-dna interactions by penalized multivariate regression. *Nucleic Acids Research, 37*(16), 5246–5254.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301–320.