Behavior Research Methods
https://doi.org/10.3758/s13428-022-02011-2

Logistic regression with sparse common and distinctive covariates

S. Park[1] · E. Ceulemans[2] · K. Van Deun[1]

Accepted: 10 October 2022

Abstract

Having large sets of predictor variables from multiple sources concerning the same individuals is becoming increasingly common in behavioral research. On top of the variable selection problem, predicting a categorical outcome using such data gives rise to an additional challenge of identifying the processes at play underneath the predictors. These processes are of particular interest in the setting of multi-source data because they can either be associated individually with a single data source or jointly with multiple sources. Although many methods have addressed the classification problem in high dimensionality, the additional challenge of distinguishing such underlying predictor processes from multi-source data has not received sufficient attention. To this end, we propose the method of Sparse Common and Distinctive Covariates Logistic Regression (SCD-Cov-logR). The method is a multi-source extension of principal covariates regression that combines with generalized linear modeling framework to allow classification of a categorical outcome. In a simulation study, SCD-Cov-logR resulted in outperformance compared to related methods commonly used in behavioral sciences. We also demonstrate the practical usage of the method under an empirical dataset.

Keywords Multiblock data · Principal covariates regression · Common and distinctive processes · Data integration · Classification · Logistic regression

Introduction

In behavioral research, it is often of interest to classify subjects, e.g., by constructing a logistic regression model. For example, in mental health research, scores on various tests are used to classify subjects into having versus not having a disorder such as alcoholism (Babor, Higgins-Biddle, Saunders, & Monteiro, 2001), dementia (Mioshi, Dawson, Mitchell, Arnold, & Hodges, 2006), and eating disorders (Hill, Reid, Morgan, & Lacey, 2010; Botella, Huang, & Suero, 2015). By constructing a classification model, the factors predicting class membership can be investigated. For example, Barnes et al. (2009) studied the importance of various measures such as genotype, fMRI, and cognitive tests in predicting dementia among older adults through logistic regression. As a result, a risk index that stratifies older adults into different risk groups depending on their scores on certain risk factors was put forward.

Many studies in behavioral sciences of today involve datasets comprised of multiple blocks of predictor variables obtained for the same individuals, with each block of variables originating from different measurement instruments. Examples of such blocks include demographic data, social media, genetic profiling, and questionnaires. These joint datasets are referred to as multiblock data (more details on the conceptual framework are given in Van Mechelen and Smilde, 2010). A unique feature of multiblock data is that they can reveal two different kinds of sources of interindividual variation; those that concern single individual data blocks and those that jointly encompass multiple blocks. These sources of variation are referred to as distinctive and common, respectively, and they are used to reveal the processes underlying the emergence of particular conditions. To explain more concretely, let us consider a block of genotype data and another block of self-reported health behavior data collected from two groups of children; ADHD-diagnosed and healthy. Studying the onset of ADHD by adopting this multiblock dataset, processes that only underlies the genotype data may be found. For example, a dopaminergic pathway involving dopamine transporter gene (DAT1) and a serotonergic pathway incorporating serotonin transporter gene (5HTTT) have been reported to play a role in ADHD (Gizer, Ficks, & Waldman, 2009). These biological

✉ S. Park
s.park_1@tilburguniversity.edu

[1] Tilburg University, Tilburg, Netherlands

[2] KU Leuven, Leuven, Belgium

Published online: 13 February 2023

pathways would be considered as distinctive processes as they entail only the genotype data block. On the other hand, the multiblock data could also reveal a process that involves both blocks of genotype and health behavior. Kahn, Khoury, Nichols, and Lanphear (2003) found the combination of maternal prenatal smoking with a DAT1 genotype leading to ADHD, while in another study, maternal stress during pregnancy together with dopamine receptor 4 gene (DRD4) were associated with severity of ADHD symptoms (Grizenko et al., 2012). Such cases of gene–environment interplay are examples of common processes as they involve multiple data blocks.

Methods based on PCA have been actively proposed to disentangle the common and distinctive processes from multiblock data, but without considering the prediction problem of an outcome variable (e.g., simultaneous component analysis with distinctive and common components, DISCO-SCA; Schouteden et al., 2013). As multiblock datasets are often characterized by a large number of variables, these PCA based methods have been further extended. The presence of many variables complicates the interpretation of the components derived by SCA as they are associated with a large set of variables. The introduction of sparseness penalties—limiting the number of variables associated with a component—yields interpretable components that represent common and distinctive processes (e.g., sparse common and distinctive SCA (SCaDS); de Schipper & Van Deun, 2018).

Recently, a method that identifies common and distinctive processes from a multiblock dataset in the context of a regression problem for a continuous outcome has been proposed (Sparse Common and Distinctive Covariates Regression (SCD-CovR); Park et al., 2020). The method is an extension of principal covariates regression (PCovR) which finds summary variables that explain variance in both predictors and outcome by combining PCA and linear regression (De Jong & Kiers, 1992). SCD-CovR incorporates SCaDS into the PCovR framework to obtain sparse common and distinctive predictor processes. In order to address the classification problem, the current paper extends the SCD-CovR method to logistic regression; this means that here we develop sparse common and distinctive covariates logistic regression method (SCD-Cov-logR). SCD-Cov-logR reveals the common and distinctive predictor processes that play a role in classification of the outcome and does so in an interpretable/insightful way by relying on sparse representations.

The paper is arranged as follows. First, we provide the methodological background and mathematical details of SCD-Cov-logR. Then, the results from simulation studies that comparatively demonstrate the performance of SCD-Cov-logR against an existing method with a similar set of objectives are presented. After further illustration of the current method on an empirical multiblock dataset, the paper is concluded by formulating some limitations and directions for future research. The implementation of SCD-Cov-logR was done in R and Rcpp, which can be found on GitHub: https://github.com/soogs/SCD-Cov-logR, along with the code used to generate the results reported in the paper.

Methods

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: $i \in \{1, 2, \ldots, I\}$. Subscript $_C$ indicates concatenation of multiple data blocks, while superscripts $^{(X)}$, $^{(y)}$ and $^{(g)}$ highlight affiliation with predictor, continuous outcome and binary outcome variables, respectively. To denote estimates, a $\hat{}$ over the symbol denoting the population parameter is used (i.e., $\hat{\mathbf{b}}$ is the estimated logistic regression coefficients). $\mathbf{X}$ refers to a matrix containing the standardized scores of J predictors corresponding to I observation units (that is, each column has mean zero and variance equal to one). In the context of multiple predictor blocks, $\mathbf{X}_k$ (with size $I \times J_k$) indicates a kth predictor block matrix with its predictors column-scaled and standardized; with $k \in \{1, 2, \ldots, K\}$. $\mathbf{X}_C = [\mathbf{X}_1, \ldots, \mathbf{X}_K]$ (of size $I \times \sum_{k=1}^{K} J_k$) denotes the supermatrix that concatenates the predictor blocks. $\mathbf{g}$ indicates a dummy vector of size I containing the scores on the binary outcome variable, while $\mathbf{y}$ is a vector of size I of a continuous outcome. In the context of an outcome variable with multiple categories, $\mathbf{G}$ (with size $I \times M$) refers to a dummy matrix for the categorical outcome with M total categories. For the ith observation unit, $g_{im} = 1$ if the response is in the mth category and $g_{im} = 0$ otherwise. Lastly, $\mathbf{I}_a$ denotes a $a \times a$ identity matrix where the subscript a indicates the size of the matrix.

Model and objective function

SCD-Cov-logR is a classification method for a categorical outcome. The method is particularly suitable when multiple large blocks of predictor variables are available as it allows to take the block structure into account and to limit the number of variables contributing to the predictive processes. SCD-Cov-logR constructs two types of summary covariates: distinctive covariates based on a linear combination of the predictor variables of one single data block and common covariates that combine variables of multiple data blocks. Identification of different types of predictor processes helps understanding of processes that play important roles in the classification of the outcome. To further facilitate the interpretation of these processes, SCD-Cov-logR introduces regularization penalties to select a subset of the predictor variables in constructing the common and distinctive covariates. Taken together, an effective classification method results where common and distinctive predictor processes are identified in a sparse and therefore interpretable manner; the method is also flexible in the sense that it includes several other methods as a special case such as logistic regression and PCovR for categorical outcomes. We start with a brief description of the building blocks, namely logistic regression and PCovR, before moving onto SCD-Cov-logR. While the current method allows classification of both binary and multiclass outcome variables via logistic regression, we focus on binary logistic regression in the following subsections in describing our method. The multiclass classification via multinomial logistic regression will be discussed thereafter, as it is a straightforward extension of the binary problem.

Logistic regression

Logistic regression assumes that the log-odds (logit) of the binary outcome are linearly dependent on the predictor variables. Let $\mathbf{x}_i$ be the vector of predictor scores for subject i and g_i the score on the outcome (either 0 or 1). The log-odds for subject i is modeled by:

$$\log\left(\frac{p(g_i = 1)}{1 - p(g_i = 1)}\right) = \mathbf{x}_i^T \mathbf{b} + b_0 \tag{1}$$

where $p(g_i = 1)$ denotes the probability that the ith subject would fall under the category represented by a 1. The vector $\mathbf{b}$ indicates the logistic regression weights and the scalar b_0 the intercept. From this model, it follows that

$$\begin{aligned} p(g_i = 1) &= \frac{1}{1 + e^{-(\mathbf{x}_i^T \mathbf{b} + b_0)}} \\ p(g_i = 0) &= 1 - p(g_i = 1), \end{aligned} \tag{2}$$

which can be used to set up the likelihood equation. The estimates of the logistic regression parameters can then be obtained by maximizing the log-likelihood or minimizing the negative log-likelihood; here, the latter will be used for integration with the PCovR objective. The following negative log-likelihood is minimized:

$$L(\mathbf{b}, b_0) = -\sum_{i}^{I} (g_i(b_0 + \mathbf{x}_i^T \mathbf{b}) - \log(1 + e^{(b_0 + \mathbf{x}_i^T \mathbf{b})})). \tag{3}$$

Typically, the minimum of this function is found via a numerical procedure as it has no closed form. A popular approach is the Newton–Raphson method for finding the root of the first derivative which amounts to iteratively reweighted least squares. It boils down to formulating local quadratic approximations of the negative log-likelihood in an iterative scheme that, after initialization, uses the minimum of the quadratic approximation for updating in the next iteration.

PCovR

In a setting with a large set of predictor variables, the ordinary (least-squares) approach to linear regression involves several drawbacks. It is difficult to interpret the large set of regression coefficients corresponding to each of the predictors. Also, in the case of multicollinearity (highly correlated predictors), the estimates are instable. When the number of predictors exceeds the number of observations (high-dimensionality), the method has no unique solution. In order to alleviate these difficulties, Principal Covariates Regression (PCovR; De Jong & Kiers 1992) was put forward by combining PCA with linear regression. PCovR introduces summary variables, the so-called ‘principal covariates’, in modeling the predictor and outcome variables. The covariates summarize the predictors by a linear combination of the original variables that is obtained in such a way that they account for variation in both predictor and outcome variables. Regression coefficients are found for these limited number of covariates instead of for each of the original predictor variables, resolving the challenges of finding a unique and stable regression model in the setting of a large number of predictors. Since the covariates summarize the predictors, they can be understood to represent the predictor processes behind the outcome. Let R be the pre-specified number of covariates to be derived. PCovR then assumes the following models for the predictor and outcome variables:

$$\begin{aligned} \mathbf{y} &= \mathbf{XWp}^{(y)} + \mathbf{e}^{(y)} \\ \mathbf{X} &= \mathbf{XW}(\mathbf{P}^{(X)})^T + \mathbf{E}^{(X)}. \end{aligned} \tag{4}$$

Both the models for the outcome $\mathbf{y}$ and for the predictor variables $\mathbf{X}$ rely on the same summary predictor scores $\mathbf{XW}$ with $\mathbf{W}$ referring to the weights matrix of size $J \times R$. The weights prescribe the linear combination of the predictors to compose the principal covariates (namely, $\mathbf{T} = \mathbf{XW}$). The first line of Eq. 4 shows the model underlying the outcome; in that model $\mathbf{p}^{(y)}$ indicates a vector of R regression coefficients while $\mathbf{e}^{(y)}$ denotes the residuals pertaining to the outcome. The second line of Eq. 4 gives the model for the predictors. $\mathbf{P}^{(X)}$ indicates the loadings matrix of size $J \times R$. Similar to the regression coefficients $\mathbf{p}^{(y)}$ for the outcome variable in the first line, the loadings matrix linearly combine the covariates to reconstruct back the predictors. It can be seen as regression coefficients obtained from regressing the predictor variables on the principal covariates. Note that this model formulation also underlies the methods of principal components regression (PCR; see

Jolliffe, 1982) and partial least squares (PLS; Wold, 1982; Wold et al., 1983).

The aim of PCovR to find covariates that effectively reconstruct $\mathbf{X}$ and simultaneously predict $\mathbf{y}$ is expressed by the following joint loss function (De Jong & Kiers, 1992):

$$L(\mathbf{W}, \mathbf{P}^{(X)}, \mathbf{p}^{(y)}) = \alpha \frac{\|\mathbf{y} - \mathbf{X}\mathbf{W}\mathbf{p}^{(y)}\|_2^2}{\|\mathbf{y}\|_2^2} + (1-\alpha)\frac{\|\mathbf{X} - \mathbf{X}\mathbf{W}(\mathbf{P}^{(X)})^T\|_2^2}{\|\mathbf{X}\|_2^2}, \quad (5)$$

with $0 \leq \alpha \leq 1$, a known constant which specifies the balance between fitting the outcome and the predictors. With α set at 0, the method is the same as PCR where the outcome variable is regressed on the principal components found by PCA. On the other hand, with $\alpha = 1$, the method is equivalent to linear regression.[1] The solution of Eq. 5 is not identifiable without imposing constraints. Therefore, the covariates are often constrained to be orthonormal ($\mathbf{T}^T\mathbf{T} = \mathbf{I}_R$) to identify the solution (De Jong & Kiers, 1992).

The principal covariates in the PCovR model are used to represent the processes that underlie both the predictor and outcome variables. Therefore, it is important to interpret the derived covariates to understand the nature of these processes. There are two ways of interpreting the covariates. Firstly, the loadings matrix $\mathbf{P}^{(X)}$ can be studied. When the principal covariates are scaled to variance equal to one ($\mathbf{T}^T\mathbf{T} = I\mathbf{I}_R$) and the predictor variables have been centered and scaled to variance equal to one, the loadings are equal to the correlation between the principal covariates and the predictor variables. Therefore, $\mathbf{P}^{(X)}$ can be conveniently interpreted in two ways; regression coefficients that reconstruct the predictors (namely, $\mathbf{T}(\mathbf{P}^{(X)})^T = (\mathbf{X}\mathbf{W})(\mathbf{P}^{(X)})^T = \mathbf{X}$) and covariate-predictor correlations. The loadings derived within PCA are also commonly studied in the same manner. On the other hand, the second way to understand the covariates is by observing the weights matrix $\mathbf{W}$. The weights are used in combining the predictors to construct the covariates, and therefore they describe the composition of the covariates. They also play an important role in applying the model to new data, in the context of prediction for new observations, as they are used to transform the new predictor variables to covariate scores. Studying the loadings or the weights are both valid ways to understand the nature of the covariates and the two estimates can both be inspected in a complementary manner. However, if one of the estimates should be chosen for inspection, the choice should depend on the research aim of interest; loadings reflect the strength of association of the predictor variables with the principal covariates while weights prescribe how the covariates are constructed. We refer to Guerra-Urzola, Van Deun, Vera, and Sijtsma (2021) for a thorough discussion of the issue of loadings versus weights in the context of sparse PCA.

SCD-Cov-logR

Here, we propose a method for binary classification that is suitable for multiblock data where several blocks of predictor variables are available: besides the fact that the method can handle many predictors or even high-dimensional data, it yields particular insight in the data by revealing common and distinctive predictor processes in a sparse and therefore interpretable manner.

Model We make use of a model formulation that integrates the logistic regression and PCovR models in Eqs. 2 and 4. More specifically, the model for the outcome variable is adapted. Let the vector $\mathbf{x}_{Ci}$ denote the ith row of the supermatix $\mathbf{X}_C$ resulting from the concatenation of the predictor blocks and let $\mathbf{W}_C$ of size $\sum_{k=1}^K J_k \times R$ denote the corresponding weights matrix, then the log-odds of the binary outcome can be modeled by the principal covariates as follows:

$$\begin{aligned} \log\left(\frac{p(g_i = 1)}{1 - p(g_i = 1)}\right) &= \mathbf{x}_{C_i}^T \mathbf{W}_C \mathbf{p}^{(g)} + p_0^{(g)} \\ \mathbf{x}_{Ci} &= \left[\mathbf{x}_{C_i}^T \mathbf{W}_C (\mathbf{P}_C^{(X)})^T\right]^T + \mathbf{e}_i^{(X)}, \end{aligned} \quad (6)$$

where $\mathbf{p}^{(g)}$ in the first line of the equation denotes the vector of R regression coefficients and $p_0^{(g)}$ the intercept. As in the PCovR model (Eq. 4), the weights matrix dictates the composition of the covariates ($\mathbf{T}_C = \mathbf{X}_C\mathbf{W}_C$). In the second line, $\mathbf{P}_C^{(X)}$ indicates the loadings matrix of size $\sum_{k=1}^K J_k \times R$. They recover the predictor variables from the covariates, as done in the PCovR model. Therefore, the covariates in this model explain both the variance of predictor variables and the log-odds of the binary outcome variable.

The model in Eq. 6 includes all predictor variables in constructing the principal covariates while often it is of interest to find the subset of variables that are relevant for the predictor processes represented by the principal covariates. Hence, our proposed model is subject to a sparsity inducing penalty that limits the number of predictor variables contributing to the covariates. SCD-Cov-logR therefore imposes the sparsity on the weights, as we are interested in finding a subset of predictors that together make up the predictor processes. In this way, understanding the covariates becomes much easier as they are based on a smaller subset of predictors.

To understand the composition of the covariates not only at the level of the individual variables but also at the level

[1] $\hat{y}_i = \sum_r p_r^{(y)} t_{ir} = \sum_r (\sum_j p_r^{(y)} x_{ij} w_{jr}) = \sum_j (\sum_r p_r^{(y)} w_{jr}) x_{ij}$, with $\sum_r p_r^{(y)} w_{jr}$ as a regression coefficient for the jth predictor, where r is an index for each covariate.

of the blocks, sparsity is imposed in two ways: On the one hand at the level of the blocks (blockwise sparsity) and, on the other hand, at the level of the individual variables (elementwise sparsity). Blockwise sparsity refers to forcing the weights corresponding to an entire set of predictors in a data block to zero. By doing so, distinctive covariates which are only comprised of predictors from a single data block can be obtained. If more than one predictor blocks but not all make up a covariate, that would be referred to as a locally common covariate, as opposed to a globally common covariate where all of the predictor blocks are involved in deriving the covariate (Måge, Smilde, & Van der Kloet, 2019). Elementwise sparsity indicates dropping individual predictors out of the model. Combining these two types of sparsity encouraged at different levels, only a subset of predictors within the blocks that are chosen by blockwise sparsity would be left in the model to make up a covariate. Common and distinctive covariates that are comprised of a small interpretable subset of predictors can therefore be found to represent the underlying predictor processes.

Objective function In setting up the objective function of SCD-Cov-logR, the objectives for logistic regression and PCovR are combined. As discussed, for a binary outcome the log-odds are regressed on the covariates. Hence, the squared error pertaining to the outcome (the left term in Eq. 5) is replaced by a negative log-likelihood function based on the PCovR logistic regression model (first line in Eq. 6). Furthermore, the two types of sparsity on the weights $\mathbf{W}_C$ are accomplished by imposing two different penalties. We employ the group lasso penalty (Yuan & Lin, 2006) which shrinks and sparsifies the weights at the block level, and the lasso penalty (Tibshirani, 1996) that does the same but for individual weights. This combination of penalties is also known as the sparse group lasso (Friedman, Hastie, & Tibshirani, 2010a; Simon, Friedman, Hastie, & Tibshirani, 2013). The objective of SCD-Cov-logR is to minimize the following loss function,

$$
\begin{aligned}
&L(\mathbf{W}_C, \mathbf{P}_C^{(X)}, \mathbf{p}^{(g)}, p_0^{(g)}) \\
&= \frac{\alpha}{l_0}\left[-\sum_i^I (g_i(p_0^{(g)} + \mathbf{x}_{C_i}^T \mathbf{W}_C \mathbf{p}^{(g)}) - \log(1 + e^{(p_0^{(g)} + \mathbf{x}_{C_i}^T \mathbf{W}_C \mathbf{p}^{(g)})}))\right] \\
&+ \frac{1-\alpha}{\|\mathbf{X}_C\|_2^2} \sum_i^I \left\| \mathbf{x}_{C_i}^T - \mathbf{x}_{C_i}^T \mathbf{W}_C (\mathbf{P}_C^{(X)})^T \right\|_2^2 \\
&+ \sum_r^R \lambda_{Lr} \, |\mathbf{w}_{Cr}|_1 + \sum_r^R \sum_k^K \lambda_{Gr} \sqrt{J_k} \left\| \mathbf{w}_r^{(k)} \right\|_2 + \lambda_R \left\| \mathbf{p}^{(g)} \right\|_2^2 \qquad (7)
\end{aligned}
$$

where the loadings associated with the predictors $\mathbf{P}_C^{(X)}$ are constrained to be column-orthogonal ($(\mathbf{P}_C^{(X)})^T \mathbf{P}_C^{(X)} = \mathbf{I}_R$) in order to identify the solution (and to avoid an ill-posed problem resulting in ever-decreasing weights compensated by ever-increasing loadings). l_0 refers to the negative log-likelihood of the null model fitted without any predictors $l_0 = -\sum_i^I (g_i \log(\bar{p}) + (1 - g_i) \log(1 - \bar{p}))$, where $\bar{p} = \frac{1}{I}\sum_i^I g_i$ is the proportion of observations in the first category. The terms with λ_{Gr} and λ_{Lr} refer to the group lasso and the lasso penalties corresponding to the rth covariate. $\mathbf{w}_r^{(k)}$ indicates the weights corresponding to the covariate r and the predictor block k. The last term denotes the ridge penalty imposed on the regression coefficients $\mathbf{p}^{(g)}$ to prevent divergence occurring due to covariates being correlated.

The first term of the loss function represents the negative log-likelihood function based on Eq. 6. It is in the same format as the negative log-likelihood function commonly used for logistic regression, except that it has been adapted according to the multiblock PCovR model structure. This term is divided by the log-likelihood of the null model[2] l_0, while the second term of sum of squared predictor errors is divided by the total sum of squared predictor scores. The two types of losses are therefore placed within a comparable scale between 0 and 1. With respect to the penalties on the weights, it can be seen that the group lasso penalty $\|\cdot\|_2$ concerns a group of weights connecting the predictors in the kth predictor block with the rth covariate, while the lasso penalty $|\cdot|_1$ is imposed on all of the $\sum_{k=1}^K J_k$ individual weights corresponding to rth covariate. The two penalties together make up the sparse group lasso.

It is possible to re-express the objective function by scaling the α parameter such that it already takes account of the negative log-likelihood of the null model l_0 and the sum of squared predictor scores $\|\mathbf{X}_C\|_2^2$. The scaled weighting parameter β is defined by:

$$
\beta = \frac{\alpha \, \|\mathbf{X}_C\|_2^2}{\alpha \, \|\mathbf{X}_C\|_2^2 + (1 - \alpha) l_0} \qquad (8)
$$

β can then replace $\frac{\alpha}{l_0}$ in the objective function (7) while $(1 - \beta)$ replaces $\frac{(1-\alpha)}{\|\mathbf{X}_C\|_2^2}$, leading to a different expression of the same objective. Such rescaling of the weighting parameter has been shown in Vervloet, Van Deun, Van den Noortgate, and Ceulemans (2013).

Relation to existing methods Several existing methods rely on objective functions that are similar to the objective introduced here in Eq. 7. A method called Sparse Principal Component Regression (SPCR; Kawano et al., 2018) has been proposed and combined with generalized linear modeling. SPCR and SCD-Cov-logR are characterized by similar objective functions; our method can be viewed as an extension of SPCR for the setting of multiple predictor blocks.

[2] This ratio of negative log-likelihoods is used in computation of McFadden's pseudo R^2 (McFadden & et al. 1973) that provides insight on explained variance in the context of logistic regression.

Likewise, several other methods can be seen as a special case of the objective function in Eq. 7. First, if the balancing parameter α is fixed at zero, common and distinctive sparse covariates would be found only optimizing the fit to the predictor variables. This solution would be equivalent to that of SCaDS (de Schipper & Van Deun, 2018), which finds common and distinctive sparse components from multiblock data. For this reason, and also because the algorithm for SCD-Cov-logR is infeasible when α is equal to exactly zero, we rely on SCaDS to find the solutions when $\alpha = 0$. Second, if the negative log-likelihood term is replaced by squared error pertaining to a continuous outcome ($\left\|\mathbf{y} - \mathbf{X}_C\mathbf{W}_C\mathbf{p}^{(y)}\right\|_2^2 / \|\mathbf{y}\|_2^2$), the objective function becomes that of SCD-CovR (Park, Ceulemans, & Van Deun, 2020), which shares the same aims as SCD-Cov-logR except it targets a continuous outcome. Third, starting from the SCD-CovR formulation, if the group lasso parameter is fixed at zero and only a single block of predictors are employed, the problem boils down to SPCovR (Van Deun, Crompvoets, & Ceulemans, 2018) which finds sparse covariates. As these methods serve as the basis for the current SCD-Cov-logR, further details of these directly related methods are provided in Appendix A. Finally, fixing the lasso and group lasso parameters at zero such that weights are found without sparsity, the problem can be seen as an extension to PCovR to account for a binary classification problem.

Algorithm The minimizing solution of Eq. 7 can be found by an alternating procedure where the loadings $\mathbf{P}_C^{(X)}$ and the regression coefficients $\mathbf{p}^{(g)}$ and $p_0^{(g)}$ are solved for conditional upon fixed values for the weights $\mathbf{W}_C$ and vice versa. Such an alternating approach has been effective for SCaDS, SCD-CovR and SPCovR. To treat the minimization of Eq. 7 which is complicated by the negative log-likelihood term, we make use of a local quadratic approximation, similar to the iteratively reweighted least squares approach that is usually taken to solve the logistic regression problem (Friedman, Hastie, & Tibshirani, 2010b). The alternating routine continues until the algorithm converges to a stationary point, usually a local minimum. Since the iteratively reweighted least squares procedure is known to sometimes lead to divergence, we also employ the maximum number of iteration of 5000 as another form of stopping criterion. As the objective function in Eq. 7 is not a convex problem, it is subject to local minima. We recommend using multiple random starting values, along with rational starting values based on PCovR (administered by treating the binary outcome as a continuous variable). Furthermore, employing multiple starting values is particularly important because the estimation of $\mathbf{W}_C$ is often a high dimensional regression problem prone to instable estimates (Jia & Yu, 2010; Guerra-Urzola, Van Deun, Vera, & Sijtsma, 2021), meaning that different starting values may result in different estimates. The sparse group lasso problem for $\mathbf{W}_C$ is treated via coordinate descent (Friedman et al., 2010a), while closed-form solutions exist for the conditional updates of $\mathbf{P}_C^{(X)}$, $\mathbf{p}^{(g)}$ and $p_0^{(g)}$. Further details on the algorithm for minimizing the objective function can be found in the Appendix B, including the schematic outline of the algorithm and the derivation of solutions to the conditional updates (Appendices C and D).

Multiclass classification

Our method can be slightly adapted to address a classification problem in the presence of more than two categories. The method is posed in the same manner as the binary problem, except it relies on multinomial logistic regression. The logit model in Eq. 6 is generalized to a 'baseline-category logit model' (Agresti, 2003) which is a common approach to extend logistic regression to a multiclass problem. Let $p(g_{im} = 1)$ and $p(g_{iM} = 1)$ denote the probability that subject i would fall under the category m and the last category M, respectively. Treating the last category as the baseline, the log-odds of the ith observation being in category m as opposed to being in the baseline category is modeled:

$$
\begin{aligned}
\log\left(\frac{p(g_{im}=1)}{p(g_{iM}=1)}\right) &= {\mathbf{x}_C}_i^T\mathbf{W}_C\mathbf{p}_m^{(g)} + p_{0m}^{(g)}, \\
&\quad \text{for } m = 1, \ldots, M-1 \\
{\mathbf{x}_C}_i &= \left[{\mathbf{x}_C}_i^T\mathbf{W}_C(\mathbf{P}_C^{(X)})^T\right]^T + \mathbf{e}_i^{(X)}, \qquad (9)
\end{aligned}
$$

where $\mathbf{p}_m^{(g)}$ and $p_{0m}^{(g)}$ refer to the regression coefficients and the intercept that correspond to category m. By calculating $M-1$ sets of the regression coefficients, the log-odds of any pairs of response categories can be determined. As for the objective function, the negative log-likelihood function based on the baseline-category logit model replaces the negative log-likelihood concerning the binary classification provided in Eq. 7:

$$
\begin{aligned}
&L(\mathbf{W}_C, \mathbf{P}_C^{(X)}, \mathbf{p}_m^{(g)}, p_{0m}^{(g)}) \\
&= \frac{\alpha}{l_0}\left[-\sum_i^I\left\{\sum_m^{M-1} g_{im}(p_{0m}^{(g)} + {\mathbf{x}_C}_i^T\mathbf{W}_C\mathbf{p}_m^{(g)})\right.\right. \\
&\quad \left.\left. -\log(1+\sum_m^{M-1} e^{(p_{0m}^{(g)} + {\mathbf{x}_C}_i^T\mathbf{W}_C\mathbf{p}_m^{(g)})})\right\}\right] \\
&\quad + \frac{1-\alpha}{\|\mathbf{X}_C\|_2^2}\sum_i^I\left\|{\mathbf{x}_C}_i^T - {\mathbf{x}_C}_i^T\mathbf{W}_C(\mathbf{P}_C^{(X)})^T\right\|_2^2 \\
&\quad + \sum_r^R \lambda_{Lr}\,|{\mathbf{w}_C}_r|_1 + \sum_r^R\sum_k^K \lambda_{Gr}\sqrt{J_k}\left\|\mathbf{w}_r^{(k)}\right\|_2 \\
&\quad + \lambda_R\left\|\mathbf{p}^{(g)}\right\|_2^2 \qquad (10)
\end{aligned}
$$

where the loadings $\mathbf{P}_C^{(X)}$ are constrained to be column-orthogonal ($(\mathbf{P}_C^{(X)})^T\mathbf{P}_C^{(X)} = \mathbf{I}_R$) as done for the binary problem (Eq. 7). Other quantities and penalty terms are also defined the same. l_0 here refers to the negative log-likelihood of the null model $l_0 = -\sum_i^I \left[\sum_m^{M-1} g_{im} \log(\bar{p}_m) + g_{iM} \log(\bar{p}_M)\right]$ where $\bar{p}_m = \frac{1}{I}\sum_i^I g_{im}$ is the proportion of observations in the mth category. Hence, the negative log-likelihood and the sum of squared errors are also scaled in this objective function. The weighting parameter α can be rescaled to β in the same manner as for the binary classification problem (see Eq. 8). Furthermore, note that both the model (Eq. 9) and the objective function (Eq. 10) become equal to those of the binary problem (Eqs. 6 and 7) when the total number of categories M are set at two. To find the minimizing solution of Eq. 10, an alternating algorithm very similar to that for the binary problem is employed. The only difference is that the negative log-likelihood term with multiple categories is treated with partial quadratic approximation with respect to the category m where only $\mathbf{p}_m^{(g)}$ and $p_{0m}^{(g)}$ are allowed to vary at a time. This partial quadratic approximation has been used for treating a penalized multinomial logistic regression problem (Friedman et al., 2010b). Details on the algorithm are provided in the Appendix E.

Toy example

In order to provide a clearer picture of the goals that the method targets and the estimates it provides, we showcase the method on a toy example dataset for a binary classification problem in this section. We generated the dataset according to one of the conditions of the simulation study which follows later. The dataset is composed of two data blocks and its underlying model assumes three covariates. Two of these covariates represent processes that are distinctive to the first and the second data blocks, respectively, while the third covariate is a common process, affiliated with both data blocks. In addition, the model was defined such that the covariate distinctive to the second block is not relevant in the classification of the outcome variable. Each of the two data blocks consists of 15 predictors concerning the same set of 100 observation units. There is one binary outcome variable. Details of the data generation setup can be found in the simulation study section.

A few technicalities come with the application of the SCD-Cov-logR to data. First, it is important to note that the solution is influenced by several tuning parameters that need to be fine-tuned via model selection. Second, also different starting values may yield different solutions because the algorithm can converge to a local minimum. The model selection procedure we adopted to find the solutions presented in the following will be discussed in the next section, along with our consideration regarding multiple starting values. Third, a pre-processing step precedes method application. All of the predictor variables are centered and scaled to unit sum of squares. Subsequently, the different predictor blocks are weighted such that the sum of squares are equal across the blocks, in order to account for the differing block sizes.

The estimates retrieved by the method along with the population parameters used to generate the dataset are provided in Table 1. It first shows that the weights $\hat{\mathbf{W}}_C$ are found sparse and correctly reflect the population weights zero-nonzero structure. Most of the estimated weights are smaller in magnitude than the population weights because the lasso and group lasso penalties not only enforce sparsity but also shrink the coefficients towards zero. The weights are interpreted as the coefficients in the linear combination that forms the covariates from the predictor variables; $t_{ir} = \sum_j w_{jr} x_{ij}$. Therefore, the weights correctly represent that the first two covariates are distinctive for each of the data blocks while the third is common. The logistic regression coefficients and the intercept $\hat{\mathbf{p}}^{(g)}$ and $\hat{p}_0^{(g)}$ are also obtained and are in agreement with the population parameters; the covariate distinctive to the second data block is much less relevant than the other covariates in the classification problem. These coefficients can be combined with the covariates to yield the predicted log-odds; $\sum_r (\hat{p}_r^{(g)} \hat{t}_{ir}) + \hat{p}_0^{(g)} = \hat{y}_i$. The inverse-logistic function (Eq. 2) is used to transform the $\hat{y}_i$ log-odds into predicted probabilities for the categories of the outcome variable; if the probability is larger than 0.5, the class predicted by the model is 1. Let us take an example of the first observation $\mathbf{x}_{C1}$, the covariate scores of this observation $\hat{\mathbf{t}_1} = \mathbf{x}_{C}{}_1^T \hat{\mathbf{W}}_C = [2.875, 0.046, 3.384]^T$ are combined with the regression coefficients to get the predicted log odds $\sum_r (\hat{p}_r^{(g)} \hat{t}_{1r}) + \hat{p}_0^{(g)} = \log\left(\frac{p(\hat{g_1}=1)}{1-p(\hat{g_1}=1)}\right) = 0.862$. Applying the inverse logistic function, the predicted probability for this observation to be classified as 1 is $\frac{1}{1+e^{-0.862}} = 0.703$. Since this probability is larger than 0.5, we predict the observation as being in class 1, which is indeed true for the first observation in our toy example dataset.

Altogether, examining this solution, it would be concluded that there are two underlying predictor processes that exclusively involve predictor variables of only one of the two data blocks and one process that involves predictors from both data blocks. Predictors x9 to x15 and x24 to x30 are filtered out of the model; they are not related with any of these processes. Only two processes out of the three are important in classifying the binary outcome variable. The predictor process distinctive to the second data block is irrelevant for the classification problem. Concerning the performance of classifying the outcome, the method classified 92 in-sample observations. To gauge the quality of predicting the classes of unseen data, we applied the fitted

Table 1 Population weights, and the solution found by SCD-Cov-logR from the toy example dataset: weights and logistic regression coefficients

$\mathbf{W}_C$				$\hat{\mathbf{W}}_C$					
	D1	D2	C		D1	D2	C	Logistic regression coefficients	
Block 1				Block 1				Population	
x1	0.5	0	0	x1	0.358	0	0	D1	−0.600
x2	0.5	0	0	x2	0.391	0	0	D2	−0.010
x3	0.5	0	0	x3	0.463	0	0	C	0.800
x4	0.5	0	0	x4	0.475	0	0	intercept	0
x5	0	0	0.354	x5	0	0	0.359		
x6	0	0	0.354	x6	0	0	0.319	Estimated	
x7	0	0	0.354	x7	0	0	0.276	D1	−0.735
x8	0	0	0.354	x8	0	0	0.233	D2	−0.072
x9	0	0	0	x9	0	0	0	C	0.907
x10	0	0	0	x10	0	0	0	intercept	−0.090
x11	0	0	0	x11	0	0	0		
x12	0	0	0	x12	0	0	0		
x13	0	0	0	x13	0	0	0		
x14	0	0	0	x14	0	0	0		
x15	0	0	0	x15	0	0	0		
Block 2				Block 2					
x16	0	0	0.354	x16	0	0	0.358		
x17	0	0	0.354	x17	0	0	0.401		
x18	0	0	0.354	x18	0	0	0.342		
x19	0	0	0.354	x19	0	0	0.307		
x20	0	0.5	0	x20	0	0.483	0		
x21	0	0.5	0	x21	0	0.415	0		
x22	0	0.5	0	x22	0	0.381	0		
x23	0	0.5	0	x23	0	0.453	0		
x24	0	0	0	x24	0	0	0		
x25	0	0	0	x25	0	0	0		
x26	0	0	0	x26	0	0	0		
x27	0	0	0	x27	0	0	0		
x28	0	0	0	x28	0	0	0		
x29	0	0	0	x29	0	0	0		
x30	0	0	0	x30	0	0	0		

The column names D1, D2, and C indicate that the corresponding covariate is defined as being distinctive to block 1, distinctive to block 2 and common

model to 100 observations of out-of-sample data that were generated from the same population as the in-sample observations. The method was able to classify 92 out-of-sample observations correctly.

Model selection

The SCD-Cov-logR method involves several (usually) unknown parameters that govern the characteristics of the derived model; the number of covariates R, the weighting parameter α, the lasso and group lasso parameters λ_{Lr}, λ_{Gr} for the sparse weights and the ridge parameter λ_R for the logistic regression coefficients. These parameters are usually tuned in accordance with a certain optimality criterion such as prediction error. Several model selection strategies can be used for different model parameters, while we adopt cross-validation for all of the parameters except for the number of covariates. A straightforward way to administer cross-validation is the grid search that exhaustively compares all possible combinations of the ranges of values for the different parameters in optimizing the criterion of cross-validation error. However, as the current method entails many parameters to be tuned, such a scheme involves a very heavy computational load. Instead, a sequential approach where sets of parameters are tuned in turn can be considered as it was demonstrated to work well for model selection for PCovR (Vervloet, Van Deun, Van den Noortgate, & Ceulemans, 2016) and also for SCD-CovR (Park et al., 2020). In the following, we propose a sequential cross-validation model selection procedure and demonstrate it with the toy example dataset.

The first step of the sequential approach is to determine the number of covariates. This was recommended in a study that compares model selection strategies for PCovR (Vervloet et al., 2016). Park, Ceulemans, and Van Deun (2020) also selected the number of covariates first and obtained models with good performance in SCD-CovR. For finding the number of covariates in SCD-Cov-logR, we first perform PCA on the predictor variables with varying number of principal components. Instead of the well-known scree test that manually looks for an 'elbow' in the plot of eigenvalues (representing the amount of variance explained by each principal component) which involves an element of subjectivity, the acceleration factor technique proposed by Raîche, Walls, Magis, Riopel, and Blais (2013) is adopted. It finds the elbow by computing at which point the slope of the graph of eigenvalues change most sharply. The technique retains the principal components that derived prior to the principal component where the sharp change in slopes occurs. The R package "nFactors" is employed for this purpose (Raiche, Magis, & Raiche, 2020).

With the number of covariates fixed, cross-validation is administered to simultaneously select the optimal values of α and λ_R. For each combination of values, the mean of squared residuals is computed. These residuals are discrepancies between the binary outcome scores of the observations in held-out samples and their corresponding predicted probabilities computed by: $\frac{1}{n}\sum_i^n \left(g_i - 1/\left(1+e^{-(\mathbf{x}_{C_i}^T \hat{\mathbf{W}}_C \hat{\mathbf{p}}^{(g)} + \hat{p}_0^{(g)})}\right)\right)^2$ where n denotes the size of the held-out samples. In the case of the multiclass problem, the residuals are computed by $\frac{1}{n(M-1)} \sum_m^{M-1} \sum_i^n \left[g_{im} - e^{\mathbf{x}_{C_i}^T \hat{\mathbf{W}}_C \hat{\mathbf{p}}_m^{(g)} + \hat{p}_{0m}^{(g)}} / \left(1 + \sum_m^{M-1} e^{p_{0m}^{(g)} + \mathbf{x}_{C_i}^T \mathbf{W}_C \mathbf{p}_m^{(g)}}\right)\right]^2$. The one standard error rule (Friedman, Hastie, Tibshirani, & et al. 2001) is adopted, which selects the least complex model within one standard error of the best-performing model. For α, higher values are associated with model complexity and overfitting because it places a heavier emphasis on the prediction problem of the outcome which becomes prone to overfitting with increasing number of predictor variables (Babyak, 2004; McNeish, 2015). Similarly, lower values of λ_R are related with overfitting as it leads to high variance of parameter estimates across samples. Therefore, the one standard error rule aims to select the models with the lowest α and the highest λ_R values. When the two parameters are not in agreement, the model with lower α is preferred over the model with higher λ_R as the former is seen to exert more impact on the final model. Note that the rescaled parameter β can be tuned instead of directly tuning for α. Higher values of β are related to overfitting, in the same manner as for α. The one standard error rule would thus choose the models comprised with the lowest β and the highest λ_R values in this case.

We tune the sparsity parameters for the weights at the final stage of the model selection procedure because they exert relatively small influences on the fit of the model with respect to both classification or reconstruction of the blocks of predictor variables (de Schipper & Van Deun, 2021; Park et al., 2020). In a paper that examined the efficacy of various model selection strategies for sparsity penalty parameters in sparse PCA that retrieves sparse weights like SCD-Cov-logR, it was reported that even a very sparse model yields good recovery of summary component scores (de Schipper & Van Deun, 2021). The authors advise using cross-validation with the one standard error rule to select the parameters, when the aim of the analysis includes understanding of underlying processes. For our proposed method, the one standard error rule is set up such that the model with the highest values of λ_{Lr} and λ_{Gr} are chosen within models with minimal cross-validation error. Between the two parameters, the model with higher λ_{Lr} is preferred over the model with higher λ_{Gr} because λ_{Lr} encourages the sparse solution in a more direct manner than λ_{Gr}. While different values of the parameters can be specified concerning the weights corresponding to each of the rth covariate, we usually adopt the same values across multiple covariates to ease the computational burden. Additionally, in choosing the ranges of sparsity parameters to be considered for model selection, values separated by a reasonable interval can be selected between a near-zero value and another value that leads to complete sparsity. One way to choose such an interval is by selecting a sequence of equally spaced values on the log scale, as done in de Schipper and Van Deun (2021) and recommended in Friedman, Hastie, and Tibshirani (2010b).

Model selection for the toy example We demonstrate the model selection procedure by applying it on the toy example dataset. First, PCA is administered to the concatenated set of centered and standardized predictor variables with various numbers of principal components. Figure 4 in Appendix F depicts the variance explained by each component. With the acceleration factor technique, the number of covariates is chosen to be three because the sharpest change in the slopes occurs at the fourth principal component. With the number of covariates fixed, we administered a five-fold cross-validation, simultaneously varying the values of β and

λ_R. Instead of directly controlling the values for α, we varied the values for its rescaled version β. The parameters λ_{Lr} and λ_{Gr} were fixed at zero for the cross-validation. We considered the values of [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9] for β and [0.1, 0.5, 1, 3, 5, 10, 30, 50] for λ_R. With the one standard error rule, a β value of 0.2 and λ_R of 1 was selected. Given these parameters, we finally conducted another five-fold cross-validation for λ_{Lr} and λ_{Gr}. The range of [0.5, 1, 5, 7, 10, 15, 30, 45, 100] was employed for λ_{Lr} and [0.1, 0.5, 1, 2, 5, 10] for λ_{Gr}. The one standard error rule selected the model with $\lambda_{Lr} = 45$ and $\lambda_{Gr} = 2$. The solution provided above in Table 1 was obtained by adopting these values for the analysis of the data. It is worth noting that using an exhaustive approach to cross-validation that considers all combinations of these ranges of parameters also resulted in models that are similar to this reported model. The results from this exhaustive approach can be found in Appendix G.

In the above model selection procedures, rational starting values (i.e., the PCovR solution) were used in initializing the SCD-Cov-logR algorithm. To account for the problem of local minima, 20 different sets of random starting values were generated. Using each set of starting values, we conducted the same model selection procedures to find the tuning parameters and the final model estimates. We found that the solution resulted from the rational starting values were associated with the lowest minimum, compared with the other starting values. Comparing the estimates obtained by different starting values, although some starting values yielded estimates that are quite different from those of the rational starting values, the starting values that resulted in smaller loss led to estimates that are very similar to those of the rational starting values. These estimates also correctly classified the same numbers of in-sample and out-of-sample observations as the estimates from the rational starting values. Since the rational starting values led to the lowest minimum, we reported these estimates in the previous section. It also seems sensible that the rational starting values from PCovR finds a lower minimum because the data was generated from a clear PCovR model structure (as seen in the Simulation Study section). However, in practice, it is recommended to adopt multiple random starting values and the rational values to initialize the algorithm and subsequently choose the solution that attains the lowest minimum. This applies especially if the underlying true model structure is unknown, unlike for the current toy example.

Related methods

SCD-Cov-logR is a classification method with three main objectives. It (a) classifies a categorical outcome, (b) recovers the underlying common and distinctive predictor processes via dimension reduction, and (c) derives sparse weights and therefore interpretable covariates. The method offers a solution that achieves all of these objectives in a flexible manner such that the user can emphasize one goal over another according to the research aim. In this section, we will present two methods that are related to SCD-Cov-logR, in the sense that they target a similar set of goals. Alongside, regularized logistic regression is also discussed as a benchmark method for classification with a large set of predictors.

PCR (logistic regression)

A commonly used method that aims both at classification and modeling the variation in the block of predictors is based on principal component regression (PCR; see Jolliffe, 1982). This method first performs PCA on the predictors and then, in a second and separate step, builds a classification model using the retrieved components as the predictor variables. In order to derive common and distinctive processes from multiblock data, the PCA step can be conducted with SCaDS (de Schipper & Van Deun, 2018). We will refer to this two-step approach of SCaDS followed by logistic regression by SCaDS-logR. As discussed above, this is the special case of SCD-Cov-logR with the weighting parameter α is specified at zero. Hence, it addresses the same research goals of SCD-Cov-logR, except that it does not take the outcome variable into consideration when deriving the components. Due to this, the underlying processes that play important roles for the outcome variable rather than the predictor variables may be omitted (Vervloet et al., 2016).

DIABLO

Data Integration Analysis for Biomarker discovery using a Latent component method for Omics (DIABLO; Singh et al., 2016) is a partial least squares (PLS)-based framework that addresses the multiple aims of prediction and sparse modeling of the variation in the predictors. PLS (Wold, 1982; Wold, Martens, & Wold, 1983) is a widely used method that has the same model structures as PCovR; it finds components that represent the underlying processes among the predictors while predicting the outcome variable. PLS can also be seen as an approach to structural equation modeling (SEM) when complex models are built without being mainly guided by theory (Tenenhaus, Tenenhaus, & Groenen, 2017). DIABLO is an extension of PLS that jointly analyzes multiple predictor blocks and obtains sparse components. Simultaneously, these sparse components explain the variation in the outcome variable. Therefore, DIABLO meets all of the research aims of SCD-Cov-logR. While our proposed method treats the multiblock problem by concatenating the predictor matrix to construct a single model that covers several data blocks, DIABLO derives one model separately for each data block; predictions from each model are accumulated via majority voting to give the overall classification. Therefore, DIABLO can be seen to only find components that are distinctive to each block. However, it is possible to specify how correlated these components built on each block would be. This would encourage capturing of the variance accounted for by common predictor processes, although they may not be explicitly obtained. Singh et al. (2016) demonstrated that when building a classification model for breast cancer subtypes with predictors from multiple data blocks (mRNA, miRNA, methylation and proteins) from The Cancer Genome Atlas (TCGA), DIABLO was able to select more variables that are strongly correlated with each other than elastic net regression.

Another core difference between SCD-Cov-logR and DIABLO lies with the parameter α that balances between reconstruction of the predictors and prediction of the outcome variable. PLS-based methods do not offer such an option and tend to lean closer to a PCovR model emphasizing prediction, this is α close to one (Vervloet et al., 2016; Van Deun et al., 2018). Furthermore, methods based on PLS are often more prone to overfitting than those derived from PCovR, which in turn results in a diminished quality of out-of-sample prediction. The results from Park et al. (2020) demonstrated this pattern of results in a multiblock regression setting.

Moreover, DIABLO does not adopt a generalized linear model framework to treat the classification of categorical outcome variables. Instead, when constructing a classification model, DIABLO adopts a simple heuristic where the categorical outcome is coded into a binary matrix with each column indicating the membership of the observation unit in a certain class. The classification model is then estimated in the same manner as the regression model by treating the binary matrix as continuous outcome variables. Among the fitted values given for each of the classes, the class that corresponds to the largest fitted value is the class determined by the DIABLO model. This approach of administering PLS for a classification problem has also been shown to be equivalent to performing discriminant analysis (Barker & Rayens, 2003). There are PLS methods that are formulated in combination with the generalized linear model framework such that a logistic regression model can be constructed (Ding & Gentleman, 2005; Chung & Keles, 2010), but these methods are only suitable for the analysis of a single data block. Additionally, Lê Cao, Boitard, and Besse (2011) reported that this approach performs comparatively with the binary indicator matrix approach of DIABLO.

Regularized logistic regression

Regularized logistic regression is a logistic regression method that performs variable selection (Friedman et al., 2010b). Due to the regularization penalties, the method can also be applied to high dimensional datasets. Hence, it can be considered as a benchmark method for classification in the setting of many predictors, being actively applied in behavioral sciences; for example to detect psychological symptom patterns from large-scale questionnaires (Tutun et al., 2019) and to classify different emotions using EEG signal patterns (Chen et al., 2020). However, since it does not extract covariates or factors, the method does not meet all of the aims of SCD-Cov-logR such as identifying the underlying processes governing the predictors.

Toy example illustration

In order to compare the two related methods that share the goals of SCD-Cov-logR, we administered them along with the benchmark of regularized logistic regression on the toy example dataset. As the population model parameters are known, we configured the methods such that they return the solutions that reflect the population model structure as closely as possible. For regularized logistic regression, the lasso penalty parameter was tuned by cross-validation, as it is not possible for the method to derive the covariate structures. For principal component (logistic) regression, we administered SCaDS (de Schipper & Van Deun, 2018) on the predictor matrix with three components. Lasso and group lasso parameters were chosen such that they reflect the population model. The outcome variable was regressed on the derived sparse principal components via logistic regression.

In order to fit the DIABLO model in accordance with the population model such that the common and distinctive predictor processes can be explicitly found, we fitted a one-component model separately from each of the two data blocks which would match the two distinctive covariates generated. For the common covariate, we constructed a one-component model from a supermatrix that concatenates the two data blocks. These components across the blocks were specified to be uncorrelated, as the true covariates were defined to be uncorrelated. As DIABLO allows the users to specify the number of non-zero weights per component, we specified these in correspondence with the number of non-zero weights in the true weights matrix.

Table 2 presents the estimates resulting from the different methods. The table shows that only the two-step principal component logistic regression approach of SCaDS-logR finds the covariates that perfectly represent the population

Table 2 Estimates provided by PCR, DIABLO and regularized logistic regression

	$\mathbf{W}_C$			SCaDS-logR			DIABLO			LogR
	D1	D2	C	D1	D2	C	D1	D2	C	b
Block 1										
x1	0.5	0	0	0.392	0	0	0	0	0	-0.198
x2	0.5	0	0	0.399	0	0	0	0	0	-0.304
x3	0.5	0	0	0.430	0	0	0	0	-013	-0.262
x4	0.5	0	0	0.496	0	0	0	0	0	-0.112
x5	0	0	0.354	0	0	0.328	0.606	0	0.480	0.265
x6	0	0	0.354	0	0	0.291	0.411	0	0.330	0.336
x7	0	0	0.354	0	0	0.262	0.636	0	0.502	0.333
x8	0	0	0.354	0	0	0.217	0.242	0	0.200	0.221
x9	0	0	0	0	0	0	0	0	0	0
x10	0	0	0	0	0	0	0	0	0	0
x11	0	0	0	0	0	0	0	0	0	0
x12	0	0	0	0	0	0	0	0	0	0
x13	0	0	0	0	0	0	0	0	0	0
x14	0	0	0	0	0	0	0	0	0	0
x15	0	0	0	0	0	0	0	0	0	0
Block 2										
x16	0	0	0.354	0	0	0.357	0	0.537	0.364	0.180
x17	0	0	0.354	0	0	0.370	0	0.533	0.353	0.189
x18	0	0	0.354	0	0	0.311	0	0.525	0.335	0.232
x19	0	0	0.354	0	0	0.281	0	0.389	0	0
x20	0	0.5	0	0	0.443	0	0	0	0	0
x21	0	0.5	0	0	0.424	0	0	0	0	0
x22	0	0.5	0	0	0.419	0	0	0	0	0
x23	0	0.5	0	0	0.479	0	0	0	0	0
x24	0	0	0	0	0	0	0	0	0	0
x25	0	0	0	0	0	0	0	0	0	0
x26	0	0	0	0	0	0	0	0	0	0
x27	0	0	0	0	0	0	0	0	0	0
x28	0	0	0	0	0	0	0	0	0	0
x29	0	0	0	0	0	0	0	0	0	0
x30	0	0	0	0	0	0	0	0	0	0

The true weights $\mathbf{W}_C$ is also provided as a reference

model structure. DIABLO can find the distinctive covariates, but does not perform well at correctly finding the non-zero parameters. It is difficult to interpret the regularized logistic regression coefficients as they do not go hand-in-hand with the population model. However, it can be seen that the predictors that do not have any relations with the covariates were filtered out, yet, also some of the predictors that do have a relation with the covariates were also filtered out.

With respect to the performance to classify the outcome variable, the number of correctly classified in-sample and out-of-sample observations for each of the methods are provided in Table 3. The results pertaining to SCD-Cov-logR are also given to offer comparison. It appears that SCD-Cov-logR and SCaDS-logR lead to comparable and good predictive performances, although the four methods don't exhibit large differences.

Extending this comparative evaluation of the related methods and SCD-Cov-logR to a simulation study requires comparison of the methods on all criteria that reflect the multiple research aims of SCD-Cov-logR. The benchmark regularized logistic regression does not meet this requirement since it fails to meet all of the research aims; it does not uncover underlying predictor processes via structures such

Table 3 Number of correctly classified observations provided by PCR, DIABLO and regularized logistic regression

	SCD-Cov-logR	SCaDS-logR	DIABLO	LogR
In-Sample	92	91	83	87
Out-of-Sample	92	92	84	88

as covariates. While both PCR (SCaDS-logR) and DIABLO address the aims, PCR has been compared in previous works against PCovR and showed underperformance in discovering the true covariate structure (Vervloet et al., 2016) and also in prediction of the outcome (Heij, Groenen, & van Dijk, 2007; Tu & Lee, 2019); the reason that PCR falls short is because its components are found without considering the outcome. Moreover, in the setting of multiple predictor blocks, PCovR resulted in better prediction of the outcome when some of the underlying predictor processes important for predicting the outcome only account for a small amount of variance in the predictors (Park et al., 2020). Therefore, in the simulation study section below, we evaluate the performance of our current method against the only competitor that accounts for all criteria, this is DIABLO.

Toy example multiclass problem

As an additional demonstration for our current method under a multiclass classification problem, we generated a toy example dataset again with a categorical outcome variable with three categories. The characteristics of the data and the underlying model were kept the same as the toy example above, except for the definition of the regression parameters and the number of observation units ($I = 1000$). Appendix H provides further details on the data generating setup. Out of the three categories, the third category was taken as the baseline category in forming the log-odds models. We administered the sequential model selection procedure as done for the binary problem, employing fivefold cross-validation considering the same ranges of parameters as for the binary problem again (see "Model selection"). The following model parameters were selected: $R = 3$, $\beta = 0.1$, $\lambda_R = 0.1$, $\lambda_{Lr} = 100$ and $\lambda_{Gr} = 10$. Table 4 shows the solution together with the defined population parameters used to generate the data. It can be seen that the estimated weights correctly represent the true underlying weights. The logistic regression coefficients found are also in agreement with the population parameters; two covariates important for discerning the categories from the third (baseline) category are correctly picked out. Moreover, the constructed model classified 842 in-sample observations and 845 out-of-sample observations correctly (both out of 1000 total observations).

Simulation study

Through a simulation study, we study the performance of the SCD-Cov-logR and DIABLO with respect to retrieval of the underlying processes and the classification of a binary outcome variable. We focus on the binary classification problem as the multiclass problem is a direct extension of the binary problem; it is expected that the insights obtained from the binary problem to be applicable for the multiclass problem. We hypothesize that SCD-Cov-logR would be better at out-of-sample classification than DIABLO as it is less susceptible to overfitting. SCD-Cov-logR would also provide models that better reflect the true underlying predictor processes as it allows a good balance between explaining the predictors and the outcome via the weighting parameter.

Design and procedure

We relied on the data generating setup presented by Chung and Keles (2010) which was used for examining the performance of several variants of sparse PLS that were set up to address the classification problem. Fixing the number of observations I to 100, the setup was modified such that two blocks of predictor variables were generated from three underlying covariates. One distinctive covariate per each predictor block was defined, while the remaining covariate reflected a common process involving both of the blocks. The three covariates were defined to differ in relevance for predicting the outcome variable, in that only two of them were defined as being relevant. We generated $J = 200$ predictor variables (100 per data block) for the high dimensional setting and $J = 30$ (15 per data block) for the low dimensional. The following setup was used:

$$\begin{aligned}
&\mathbf{T} \sim \mathcal{MVN}(\mathbf{0}, \Sigma = 50^2\mathbf{I}_3)\\
&\mathbf{E} \sim \mathcal{MVN}(\mathbf{0}, \Sigma_E = \sigma^2\mathbf{I}_J)\\
&\mathbf{X}_C \leftarrow \mathbf{T}\mathbf{W}_C^T + \mathbf{E}\\
&\mathbf{z} \leftarrow 1/(1 + exp(-\mathbf{T}\mathbf{p}^{(g)}))\\
&g_i \sim Bernoulli(z_i)
\end{aligned} \tag{11}$$

$\mathbf{T}$ is a $I \times 3$ covariate scores matrix drawn from a multivariate normal distribution defined with the mean vector fixed to $\mathbf{0}$ and a diagonal covariance matrix Σ with all of its diagonal elements fixed at 50^2. The three covariates are therefore uncorrelated. The columns of the $J \times 3$ weights matrix $\mathbf{W}_C$ is defined such that they reflect the defined common or distinctive nature of the corresponding covariates. For example, weights corresponding to a covariate distinctive to the first predictor block, are non-zero only for predictors in the first block while the remaining weights corresponding to predictors in the second block are all zero. Likewise, for a common covariate, non-zero weights are defined for

Table 4 Population parameters and the solution found by SCD-Cov-logR from the toy example multiclass dataset

$\mathbf{W}_C$				$\hat{\mathbf{W}}_C$				Logistic regression coefficients		
	D1	D2	C		D1	D2	C		1	2
Block 1				Block 1				Population		
x1	0.5	0	0	x1	0.485	0	0	D1	0.600	0.950
x2	0.5	0	0	x2	0.485	0	0	D2	0.010	0.312
x3	0.5	0	0	x3	0.475	0	0	C	-0.800	0.010
x4	0.5	0	0	x4	0.476	0	0	intercept	0	0
x5	0	0	0.354	x5	0	0	0.345			
x6	0	0	0.354	x6	0	0	0.344	Estimated		
x7	0	0	0.354	x7	0	0	0.348	D1	1.843	2.865
x8	0	0	0.354	x8	0	0	0.338	D2	-0.026	0.941
x9	0	0	0	x9	0	0	0	C	-1.966	0.015
x10	0	0	0	x10	0	0	0	intercept	0.033	-0.025
x11	0	0	0	x11	0	0	0			
x12	0	0	0	x12	0	0	0			
x13	0	0	0	x13	0	0	0			
x14	0	0	0	x14	0	0	0			
x15	0	0	0	x15	0	0	0			
Block 2				Block 2						
x16	0	0	0.354	x16	0	0	0.350			
x17	0	0	0.354	x17	0	0	0.345			
x18	0	0	0.354	x18	0	0	0.348			
x19	0	0	0.354	x19	0	0	0.349			
x20	0	0.5	0	x20	0	0.482	0			
x21	0	0.5	0	x21	0	0.475	0			
x22	0	0.5	0	x22	0	0.480	0			
x23	0	0.5	0	x23	0	0.482	0			
x24	0	0	0	x24	0	0	0			
x25	0	0	0	x25	0	0	0			
x26	0	0	0	x26	0	0	0			
x27	0	0	0	x27	0	0	0			
x28	0	0	0	x28	0	0	0			
x29	0	0	0	x29	0	0	0			
x30	0	0	0	x30	0	0	0			

The column names D1, D2 and C indicate that the corresponding covariate is defined as being distinctive to block 1, distinctive to block 2 and common. The third category is chosen as the baseline category; the regression coefficients construct the log-odds of the first or the second category as opposed to the third

predictors in both blocks. On top of these zero weights that determine the common or distinctive nature of the covariates, further sparsity is added by defining more elements of $\mathbf{W}_C$ as zeros. The sparsity levels of the weights matrix is fixed at 82% and 85% for low and high dimensional settings, respectively. It is important to note that the weights matrix was constructed such that it is column-orthogonal: $\mathbf{W}_C^T\mathbf{W}_C = \mathbf{I}_R$. Together with the covariates $\mathbf{T}$ which are orthogonally defined, this model corresponds to the well-known PCA decomposition where the weights are equal to the loadings (Guerra-Urzola et al., 2021, for discussion;). This is why the weights $\mathbf{W}_C^T$ in Eq. 11 linearly combine the covariates $\mathbf{T}$ to generate the predictors $\mathbf{X}_C$ in the same manner as loadings in PCA decomposition. An example of the population weights matrix in a low dimensional setting is presented in "Toy example" (Table 1) along with the toy example dataset, and the weights are defined in a similar manner for a high dimensional setting.

The predictors $\mathbf{X}_C$ are generated by multiplying the covariate scores matrix with the weights matrix and adding random error on top. The residual matrix $\mathbf{E}$ is generated from a multivariate normal distribution with zero mean vector and a diagonal covariance matrix Σ_E such that the residuals are uncorrelated with each other and also with the covariate scores. The variance of the error variables is adjusted according to one of the manipulated design factors of the simulation study: proportion of variance in $\mathbf{X}_C$ explained by the underlying covariates. $\mathbf{p}^{(g)}$ indicates the regression coefficients. g_i is sampled from a Bernoulli distribution with the probability defined by the linear combination of $\mathbf{T}$ and $\mathbf{p}^{(g)}$ transformed by the inverse-logitic function (see Eq. 2).

Based on this data generating model, we manipulated three data characteristics which are listed in the overview below. The different levels taken by these manipulated factors are provided between square brackets.

Study setup

1. Number of predictors J_k in each block: [100], [15]
2. Covariates relevant to the response $\mathbf{g}$: [D1, D2], [D1, C]
3. Proportion of variance in $\mathbf{X}_C$ explained by the covariates: [0.8], [0.5], [0.2]

The number of predictors manipulated by the first design factor determines whether the dataset would be low or high-dimensional. The second design factor indicates which covariates are relevant for the classification of the binary outcome with D1, D2, and C denoting the two distinctive and the common covariate, respectively. The relevance of the covariates is manipulated by specification of regression coefficients $\mathbf{p}^{(g)}$, which equals $[0.60, -0.80, 0.01]$ and $[0.60, 0.01, -0.80]$ for the two levels respectively. For the first level, the two distinctive covariates are made relevant in explaining the outcome variable, while the covariate distinctive to the first block and the common covariate are relevant in the second level. As stated above, the proportion of variance in the predictors accounted for by the covariates is controlled by the variance of the error variables $\mathbf{E}$. Fully crossing these factors and generating 50 datasets per condition, $2 \times 2 \times 3 \times 50 = 600$ datasets were produced.

Two different analyses were administered to each of these datasets: SCD-Cov-logR and DIABLO. As done for DIABO for the toy example dataset, a one-component model was fitted for each of the two data blocks to match the two distinctive covariates generated. For the common covariate, we constructed a one-component model from a supermatrix that concatenates the two data blocks.

Model selection

As the true underlying structure of the datasets is already known, several tuning parameters were tailored to correspond to the true structure. For SCD-Cov-logR, the number of covariates was fixed at three. The weighting parameter α and the ridge penalty parameter λ_R that regularizes the logistic regression coefficients were tuned together via fivefold cross-validation. As done in the toy example in "Model selection", we used the rescaled weighting parameter β instead of α. The ranges of [0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9] and [0.5, 1, 5, 10, 30, 50] respectively were used for β and λ_R. We adopted the 1 standard error (SE) rule to select a set of parameters which provides the most general model among the set of parameters yielding errors within 1 SE from minimum cross-validation error. We chose the lowest β and the highest λ_R. For the toy example, the lasso λ_{Lr} and the group lasso λ_{Gr} parameters were fixed at zero while tuning β and λ_R. Instead, for the simulation study, they were fixed differently for various conditions of the simulation study to encourage retrieval of one common and two distinctive covariates (Appendix I).

Finally, with values of β and λ_R fixed, in order to find the parameters λ_{Lr} and λ_{Gr} that match the population weights structure the closest, we fitted the method with a range of values for λ_{Lr} and λ_{Gr}. The ranges of [3, 5, 10, 15, 20, 30, 50, 80] and [0.5, 1, 2, 3, 5, 10] were adopted respectively for λ_{Lr} and λ_{Gr}. As in the toy example dataset, the datasets have been generated such that a PCovR model underlies the true sparse model structure. This means that the rational starting values are likely to provide a more optimal solution than random starting values. Therefore, we only employed the rational starting values based on PCovR.

For DIABLO, we specified the number of nonzero weights according to the defined model structure. As done for the toy example dataset, the components from different blocks were fitted such that they are not correlated. This is sensible because the true covariates are generated to be uncorrelated from each other.

Evaluation criteria

Because the methods have several objectives, including recovery of the underlying processes and classification of a binary outcome, two measures are used to study performance of the methods in relation to each of these objectives. The performance measures are:

1. Out-of-sample balanced error rate (BER): (false positive rate + false negative rate)/2.
2. Correct weights classification rate: proportion of the weights correctly classified as zero and non-zero elements relative to the total number of coefficients.

An independent test set (of 100 observation units) needed for computing the out-of-sample BER was generated following the same data generating procedures as the data used for model-fitting. A BER equal to zero indicates a perfect classification. The correct weights classification rate represents the method's ability in retrieving the underlying processes. SCD-Cov-logR provides weights matrix $\hat{\mathbf{W}}_C$ of size $\sum_{k=1}^{2} J_k \times R$ which covers the entire set of the multiblock predictors. For the weights provided by SCD-Cov-logR, we first computed Tucker congruence (Tucker, 1951) between the columns of the true $\mathbf{W}_C$ matrix and those of the estimated $\hat{\mathbf{W}}_C$ matrix. After matching the columns that resulted in the highest Tucker congruence to account for the permutational freedom of the covariates, the correct classification rate was calculated from the matching pairs of true and estimated $\mathbf{W}_C$ columns.

On the other hand, for DIABLO, one component each was estimated for the two predictor blocks and the concatenated supermatrix. Components derived from the individual predictor blocks naturally correspond to the true distinctive covariates. In order to calculate the correct classification rate, the weights estimated for these estimated components were compared against true weights that correspond to the true distinctive covariates. Likewise, the weights found from the concatenated supermatrix were compared against the true weights corresponding to the common covariate.

Results

Out-of-sample BER

We first examine the performance of the two methods concerning the prediction for new data. The estimates obtained by the methods from the training dataset are applied on the out-of-sample test set generated under equal conditions. The results from our simulation study arranged for each condition are displayed in Fig. 1. It can first be seen that SCD-Cov-logR resulted in the smaller out-of-sample BER in almost all of the conditions. With regards to the manipulated design factors, the relevance of the covariates seems to have played an important role in different performances among the methods. When the two distinctive covariates are defined as being relevant, the discrepancy in the methods is smaller, but with the covariate distinctive to the first block and the common covariate relevant, the outperformance of SCD-Cov-logR stands out more prominently. The proportion of variance in $\mathbf{X}_C$ accounted for by the covariates resulted in the 'main effect' - with smaller proportion leading to higher BER for all of the methods. Finally, it appears that the discrepancy in the performance of the methods is larger when the dataset is high-dimensional. Overall, we conclude that SCD-Cov-logR outperforms DIABLO at predicting the classes of new observations. However, the methods present more comparable performance when the processes relevant for classification are distinctive, under low dimensionality.

Correct weights classification rate

Figure 2 presents the outcome of the correct weights classification rate. Across all of the conditions, SCD-Cov-logR resulted in the higher of correct classification. It is also noteworthy that the classification rate for the method is mostly above 0.95. The figure shows the influence of the relevance of the underlying covariates and its interaction with the other manipulated data circumstances. When the two distinctive covariates were relevant, regardless of the dimensionality, SCD-Cov-logR resulted in a much higher classification rate than DIABLO. On the other hand, when the covariate distinctive to the second data block was defined irrelevant, DIABLO's performance was closer to SCD-Cov-logR's in the conditions with more variance of the predictors explained and with 15 predictor variables per block. In conclusion, SCD-Cov-logR is better than DIABLO at correctly retrieving the underlying population weights.

Illustration: 500 Family data

Dataset and pre-processing

We demonstrate an example use of SCD-Cov-logR by administering the method on an empirical dataset. We adopted the dataset from the 500 Family Study (Schneider & Waite, 2008) which investigated into how work impacts the well-being of parents and children in American middle-class families. Questionnaire data from different members of the same family were collected. We computed sum scores from questionnaire items that refer to the same construct. These scores concern the feelings of the family members, their recent mutual activities and how they perceive their relationship. 24 sum score variables were computed and are used as predictors in constructing the SCD-Cov-logR model. They can be found in Table 5. Eight of the predictors pertain to responses from the mother, another eight to responses from the father and lastly six predictors are based on the responses of the child. The dataset therefore is comprised of three blocks according to the member of the family, and each observation unit refers to a family. All of the predictors were centered and standardized. Since the blocks have different sizes, they were weighted such that the sum of squares is equal across blocks.

The families are categorized into two groups according to the child's most recent grade at school. The family with the child with a grade B or higher is classified as having academic overachievement (coded as 1), while grade C or lower is classified as underachievement (coded as 0). We excluded the families with missing values on any of the predictor variables, and made a random subset selection of 58 families in order to obtain a balance between the size

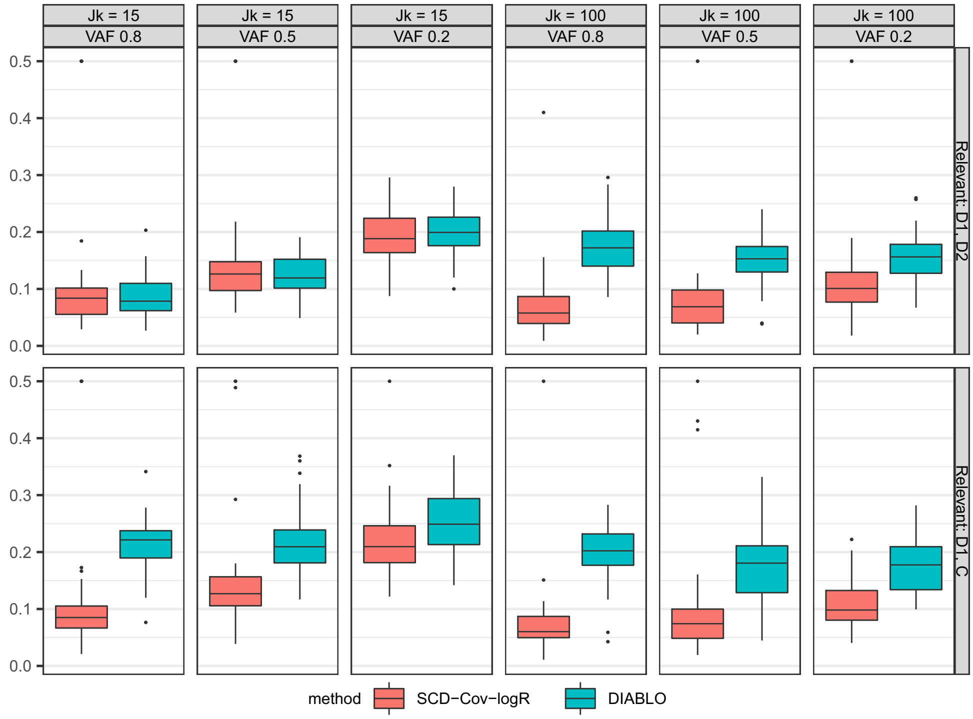

Fig. 1 Box plots of the out-of-sample BER; each panel corresponds to one of the 12 conditions. The *column panels* indicate the number of predictors in each data block and the proportion of variance accounted for by the underlying processes. The *row panels* indicate the two covariates relevant for the outcome variable; "D1", "D2" and "C" refer to the covariate distinctive to the first block, the covariate distinctive to the second block and the common covariate, respectively

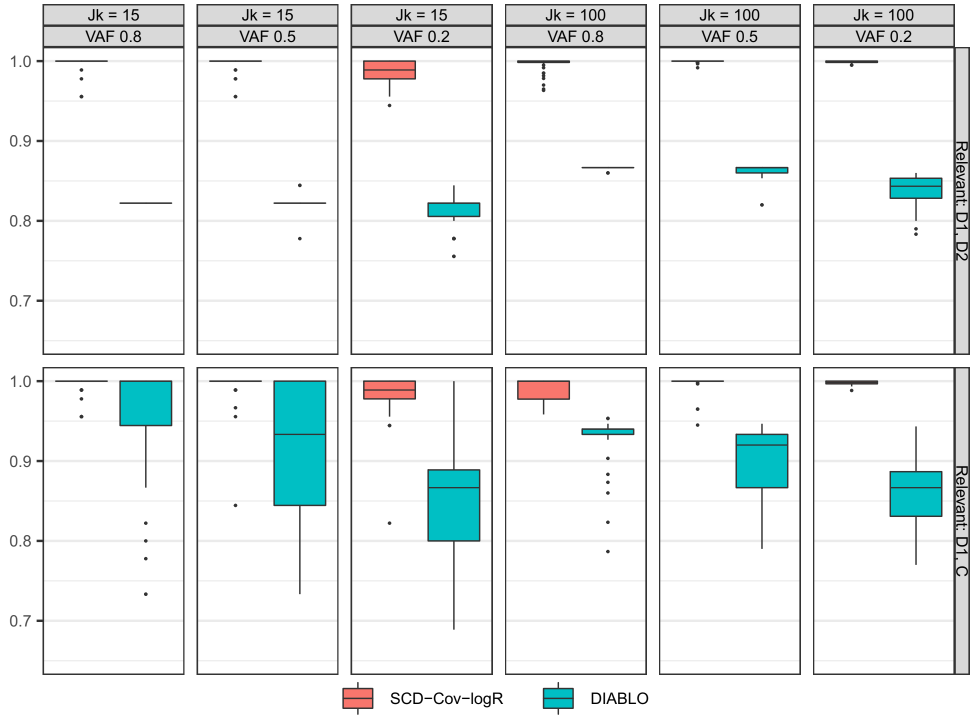

Fig. 2 Box plots of the correct weight classification rate; each panel corresponds to one of the 12 conditions. The column panels indicate the number of predictors in each data block and the proportion of variance accounted for by the underlying processes. The row columns refer to the two covariates relevant for the outcome variable; "D1", "D2" and "C" refer to the covariate distinctive to the first block, the covariate distinctive to the second block and the common covariate, respectively

Table 5 Weights and logistic regression coefficients derived by SCD-Cov-logR from the 500 family dataset

$\hat{\mathbf{W}}_C$				
	Child	Parents	Logistic regression coefficients	
Mother			Estimated	
Relationship with partners	0	0.276	Child	0.288
Argue with partners	0	0.269	Parents	0.034
Child's bright future	0	0	Intercept	-0.007
Activities with children	0	0		
Feeling about parenting	0	0.188		
Communication with children	0	0.357		
Argue with children	0	0.171		
Confidence about oneself	0	0.406		
Father				
Relationship with partners	0	0.091		
Argue with partners	0	0.183		
Child's bright future	0	0		
Activities with children	0	0		
Feeling about parenting	0	0		
Communication with children	0	0		
Argue with children	0	0.210		
Confidence about oneself	0	0.050		
Child				
Self-confidence/esteem	0.285	0		
Social life and extracurricular	0.336	0		
Importance of friendship	0.459	0		
Self Image	0.381	0		
Happiness	0.374	0		
Confidence about the future	0.281	0		

The covariate labels heading the columns of the table with weights and the rows of the table with logistic regression coefficients indicate which data blocks the corresponding covariate is associated with

of two categories. We conducted SCD-Cov-logR to target this classification problem of academic underachievement while simultaneously constructing a model that describes the underlying common and distinctive processes of the three predictor blocks.

Model selection

We employed the sequential cross-validation model selection strategy discussed in "Model selection" applied to the toy example dataset. Moreover, 50 sets of random starting values were employed alongside the rational starting values in conducting the model selection and final model fitting.

First, the number of covariates was found by administering PCA on the predictor matrix. By using the acceleration factor technique, we found that when going from 1 to 2 principal components, the amount of variance explained by the principal components changes the most drastically (Figure in the Appendix J). With the number of covariates determined at two, we carry out the cross-validation to select the other tuning parameters. The different sets of starting values were introduced at this stage. The complete process of model selection and model fitting was conducted for each set of starting values. The resulting solutions from 50 random starting values and 1 rational starting value were compared in terms of the value of the loss function: The solution with the smallest loss was retained as the final solution.

The cross-validation procedures administered for each of the starting values were as the following: first, 20-fold cross-validation was conducted with varying values of the rescaled weighting parameter β and λ_R. At this stage, the tuning parameters λ_{Lr} and λ_{Gr} were fixed at zero for the cross-validation. We considered the values of [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9] for β and [0.01, 0.05, 0.1, 0.5, 1, 2, 5, 10, 15, 20] for λ_R. Using the one standard error rule, values of β and λ_R are selected. Given these selected values, the second sequence of 20-fold cross-validation for λ_{Lr} and λ_{Gr} was conducted. With the ranges of [0, 0.05, 0.1, 0.3, 0.5, 1, 3, 5, 7, 10, 15, 20, 30, 50] adopted for both parameters, the same parameter value was used concerning the two covariates. We used the one standard error rule again to choose the values of λ_{Lr} and λ_{Gr}, completing the model selection procedure.

Similar to the toy example dataset, a smaller minimum was achieved by the set of rational starting values. The final values for the tuning parameters selected through the sequential procedure were: $\beta = 0.1$, $\lambda_R = 2$, $\lambda_{Lr} = 10$, $\lambda_{Gr} = 10$. The final model estimates obtained are presented in Table 5.

Results

The estimated weights matrix from Table 5 show that there are two predictive processes for the child's academic achievement. The first component is distinctive to the child block and is associated with all of the variables from the data block. It appears that all of the variables in the child block have an impact in the academic achievement. On the other hand, the second component is locally common, involving several variables from the mother and the father blocks but not from the child block. Observing the weights from the second covariate, it can be seen that parents' high confidence in the child's future and the amount of activities they partake with the child are not important in predicting the child's academic achievement. Also, according to this model, the father's positive feeling about parenting and his level of communication do not exert strong influence in the child's academic achievement. Moreover, the logistic regression coefficients suggest that the Child covariate is much more relevant in predicting child's academic achievement group. It appears that the attitudes that the children themselves have are the most important in leading to academic overachievement.

The covariate scores of the 58 families can be seen in Fig. 3 which presents a fair separation of the two categories of the families. With the observations separated along the X-axis, It can be seen that the Child covariate plays a more important role in separating the two groups. This is in line with the small magnitude of the coefficient corresponding to the Parents covariate. Out of the 58 families, the final model classifies 43 families correctly. In order to also examine the classification performance of the model on out-of-sample data, we performed a leave-one-out cross-validation which resulted in 40 families being correctly classified. Together, this implies that the model showed about 70% of classification accuracy for both in-sample and out-of-sample observations.

To obtain more comparative insight about the quality of the method under this empirical dataset, we administered the related methods discussed in the methods section; regularized logistic regression, PCR (SCaDS-logR) and DIABLO. The PCA step for the PCR was conducted with SCaDS to tackle the multiblock nature of the data, as demonstrated with the toy example dataset in "Related methods" The number of components for SCaDS was set at two, so that the model is comparable to the SCD-Cov-logR model constructed with two covariates. The lasso and group lasso parameters governing the sparseness of SCaDS weights were selected with 20-fold cross-validation with the one standard error rule. Similarly, a two-component model was estimated with DIABLO. The number of non-zero weights to be estimated per component was tuned via 20-fold cross-validation. Lastly, the lasso parameter for regularized logistic regression was also chosen with 20-fold cross-validation. Table 6 provides the number of correctly classified in-sample observations from each of the methods. As done for SCD-Cov-logR, leave-one-out cross-validation was conducted to gauge the out-of-sample classification quality. These results are also provided in the table. It can be seen that the four methods led to very comparable performances with respect to prediction. The estimates derived by the methods are provided in Appendix K and they can be inspected to understand the constructed models. It was found that only SCaDS-logR identified predictive processes concerted by several predictors, akin to the covariates of SCD-Cov-logR. Both regularized logistic regression and DIABLO found a very sparse model with only two non-zero coefficients.

In conclusion, our proposed method is capable in meeting its goals when applied to an empirical dataset. The method identifies common and distinctive covariates and weights that are interpretable. At the same time, the method is able to correctly classify both the samples used for fitting the model and new samples.

Discussion

A multitude of goals are of interest when building a classification model from a multiblock dataset. The common and distinctive predictor processes need to be identified in an interpretable manner while classifying the outcome variable. We have proposed the method of SCD-Cov-logR that fulfills these goals in a simultaneous manner. We have evaluated the method comparatively against DIABLO; a multiblock variant of PLS. It was found that the proposed method outperforms DIABLO in the objectives that the methods attain: quality of classification and retrieval of weights that are used to understand the underlying processes. Moreover, while DIABLO requires prior information for identifying the common and distinctive processes, our proposed method is able to explore these structures without explicit specification.

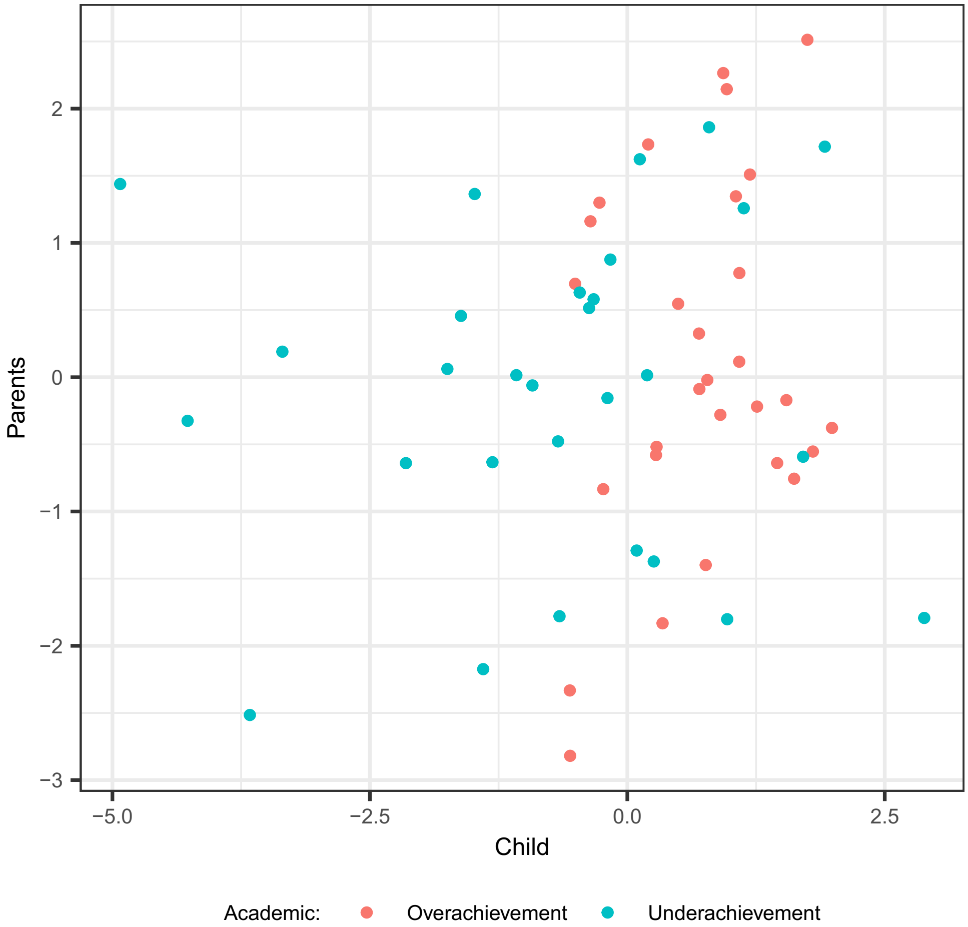

Fig. 3 Scatterplot of the two covariates found by SCD-Cov-logR. The colors represent the academic achievement of the child

In particular, SCD-Cov-logR was found to be considerably better than DIABLO in accurately retrieving the weights matrix. This finding is in line with existing literature that compares between the methodologies of PLS and PCovR. Methods based on PLS tend to place heavier focus on prediction of the outcome variables, as opposed to exploring the structure of the underlying predictor processes. In contrast, the weighting parameter α in the PCovR methods helps to attain a good balance between emphasizing the predictor or the outcome variables. In the current paper, all of the results were based on the rescaled parameter β tuned via cross-validation. This suggests that the parameter can be used effectively in a purely data-driven approach.

SCD-Cov-logR also has weaknesses. Model selection is an inherent challenge since the method requires many parameters to be tuned to meet its multiple research aims. There are in total 5 parameters to be selected and they all play an important role in shaping the retrieved model. Adopting the solution recommended by Vervloet, Van Deun, Van den Noortgate, and Ceulemans (2016), the current paper suggested a sequential model selection approach where sets of tuning parameters are chosen through cross-validation with the other parameters fixed. Models obtained by this approach led to good results in both simulation experiments and empirical study. We have not visited the model selection problem of our method in

Table 6 Number of correctly classified observations (out of the total 58) provided by SCD-Cov-logR, PCR, DIABLO and regularized logistic regression

	SCD-Cov-logR	SCaDS-logR	DIABLO	LogR
In-sample	43	43	44	43
Out-of-sample (leave-one-out CV)	40	41	38	40

The out-of-sample classification is computed via leave-out-out cross-validation great detail as the main purpose of this paper lies within the proposal and illustration of the novel method.

Another remark about the model selection procedure is the optimality criterion used for cross-validation. Throughout the paper, we adopted the sum of squared cross-validation errors concerning the binary outcome variable. This implies that the model selection procedure is conducted only considering the out-of-sample prediction quality. Since our method is not only used for classification of the outcome but also exploring the predictor processes, the optimality criterion for cross-validation can be changed to also include the errors pertaining to the predictor variables. This choice is in the same spirit of the weighting parameter α; if the user is interested more in the exploration of the predictor processes, it may be a viable option to look into such an optimality criterion different from what is used in this paper.

In our illustration of the toy data example and the simulation study, DIABLO was fitted in a peculiar manner to allow for derivation of the distinctive and common covariates. However, in practice, there may be other ways of specifying the method. For example, a supermatrix of concatenated blocks can be provided as the only input dataset and a single DIABLO model can be constructed on it.[3] We have explored into such a specification, and found that it results in consistent underperformance compared to SCD-Cov-logR with respect to prediction and retrieval of population parameters. It also has a tendency to only find common covariates.

Finally, the method and the current paper suggest several future directions of research. It would be a natural extension to broaden the method to encompass generalized linear models. This would allow modeling of outcome variables in diverse nature such as count data. Furthermore, such an extension would allow other related research questions to be addressed. For example, within the high-dimensional multiblock setting, it would be interesting to examine the impact of using a generalized linear model framework to model the categorical outcome, as opposed to the discriminant analysis approach adopted for DIABLO where the categorical outcome variable is simply changed into a dummy matrix and a linear regression model is fit. Although Lê Cao, Rossouw, Robert-Granié, and Besse (2008) compared the two approaches and reported that they show comparable performance in practice, the comparison has not been conducted in the multiblock data setting. Our proposed method SCD-Cov-logR can also be easily adapted into the linear regression approach using a dummy outcome matrix, if it is found to be useful in certain data circumstances.

[3] This would then be a single model of sPLS-DA (sparse partial least squares discriminant analysis).

Appendix A: SPCovR, SCaDS and SCD-CovR

A.1: SPCovR

For easier interpretation of the principal covariates and consistency of estimates in the high dimensional settings, regularization penalties have been imposed on the weights from Eq. 5 to lead to sparse PCovR (SPCovR; Van Deun et al., 2018). The method finds sparse weights by minimizing the following objective function:

$$\begin{aligned} L(\mathbf{W}_k, \mathbf{P}_k^{(X)}, \mathbf{p}^{(y)}) = \alpha \frac{\|\mathbf{y} - \mathbf{X}_k \mathbf{W}_k \mathbf{p}^{(y)}\|_2^2}{\|\mathbf{y}\|_2^2} \\ +(1-\alpha)\frac{\left\|\mathbf{X}_k - \mathbf{X}_k \mathbf{W}_k (\mathbf{P}_k^{(X)})^T\right\|_2^2}{\|\mathbf{X}_k\|_2^2} \\ +\lambda_L \left|\mathbf{W}_k\right|_1 + \lambda_R \left\|\mathbf{W}_k\right\|_2^2 \end{aligned} \quad (12)$$

such that $(\mathbf{P}_k^{(X)})^T \mathbf{P}_k^{(X)} = \mathbf{I}_R$ and with $\lambda_L \geq 0$, $\lambda_R \geq 0$ and $\alpha \geq 0$. The regularization parameters are the lasso, with $|\mathbf{W}_k|_1 = \sum_{j_k, r} |w_{j_k r}|$, and the ridge $\|\mathbf{W}_k\|_2^2 = \sum_{j_k, r} w_{j_k r}^2$, together forming the elastic net penalty (Zou & Hastie, 2005). The ridge penalty shrinks the magnitude of the estimates and encourages stable estimation for high-dimensional data, while the lasso penalty is involved in variable selection by shrinking and forcing the estimates to exactly zero. When both penalties are defined at 0, it can be seen that the PCovR formulation (Eq. 5) is retrieved.

A.2: SCA and SCD-CovR

SPCovR only targets data with a single predictor block and hence do not address the questions associated with multiple predictor blocks. A joint analysis of the K predictor blocks can be conducted by imposing a multiblock PCovR model, based on the SCA model (Kiers & Ten Berge, 1989):

$$\begin{aligned} \mathbf{X}_C &= \mathbf{X}_C \mathbf{W}_C (\mathbf{P}_C^{(X)})^T + \mathbf{E}^{(X)} \\ \mathbf{y} &= \mathbf{X}_C \mathbf{W}_C \mathbf{p}^{(y)} + \mathbf{e}^{(y)} \end{aligned} \quad (13)$$

where $\mathbf{X}_C = [\mathbf{X}_1, \ldots, \mathbf{X}_K]$ (of size $I \times \sum_{k=1}^{K} J_k$) denotes the supermatrix that concatenates the predictor blocks. Consequently, $\mathbf{W}_C$ and $\mathbf{P}_C^{(X)}$ are weight and loading matrices of size $\sum_{k=1}^{K} J_k \times R$. $\mathbf{p}^{(y)}$ indicates a vector of R regression coefficients.

When SCA is administered to study the processes underlying the variables without considering the regression problem, the concatenated weights matrix $\mathbf{W}_C$ is examined to understand the nature of the components. In order to allow SCA to explicitly distinguish common and distinctive processes and provide a sparse and interpretable solution from high dimensional multiblock datasets, de Schipper and Van Deun (2018) proposed SCaDS. Regularization penalties are imposed upon the weights to force certain elements to zero for handier interpretation, while the $\mathbf{W}_C$ matrix is further constrained such that certain components are a priori fixed as being common or distinctive.

Making use of the multiblock PCovR model (Eq. 13) and also combining with SCaDS, SCD-CovR extends SPCovR to allow multiblock analysis. It predicts the outcome, while providing sparse weights that capture the common and distinctive processes in the predictor blocks. SCD-CovR implies minimizing the following objective function:

$$
\begin{aligned}
L(\mathbf{W}_C, \mathbf{P}_C^{(X)}, \mathbf{p}^{(y)}) = {} & \alpha \frac{\left\| \mathbf{y} - \mathbf{X}_C \mathbf{W}_C \mathbf{p}^{(y)} \right\|_2^2}{\|\mathbf{y}\|_2^2} \\
& + (1-\alpha) \frac{\left\| \mathbf{X}_C - \mathbf{X}_C \mathbf{W}_C (\mathbf{P}_C^{(X)})^T \right\|_2^2}{\|\mathbf{X}_C\|_2^2} \\
& + \lambda_L \, |\mathbf{W}_C|_1 + \lambda_R \, \|\mathbf{W}_C\|_2^2
\end{aligned} \tag{14}
$$

such that $(\mathbf{P}_C^{(X)})^T \mathbf{P}_C^{(X)} = \mathbf{I}_R$, and subject to zero block constraints on $\mathbf{W}_C$ that fix weights that correspond to one or several predictor blocks to zero. This implies that the component is determined only by predictors of those blocks for which the weights have not been fixed to zero. Common components are obtained by not placing such zero block constraints on the component. The elastic net penalty and the constraints concerning the weights are the same as imposed in SCaDS. Also, as in SPCovR, the lasso penalty achieves sparseness within the common and distinctive covariates.

Appendix B: SCD-Cov-logR algorithm

The minimizing solution of Eq. 7 can be found by iteratively reweighted least squares which involves formulating the quadratic approximation of the negative log likelihood given the current estimates of the parameters (Friedman et al., 2010b). The negative log likelihood part of the objective function is as the following:

$$
\begin{aligned}
L_{logr}(\mathbf{W}_C, \mathbf{p}^{(g)}, p_0^{(g)}) = & -\sum_i^I g_i (p_0^{(g)} + \mathbf{x}_{C_i}^T \mathbf{W}_C \mathbf{p}^{(g)}) \\
& - \log(1 + e^{(p_0^{(g)} + \mathbf{x}_{C_i}^T \mathbf{W}_C \mathbf{p}^{(g)})})
\end{aligned} \tag{15}
$$

Quadratic approximation of Eq. 15 given the current estimates of the parameters is as the following.

$$
LQ_{logr}(\mathbf{W}_C, \mathbf{p}^{(g)}, p_0^{(g)}) = \frac{1}{2} \sum_i^I q_i (z_i - p_0^{(g)} - \mathbf{x}_{C_i}^T \mathbf{W}_C \mathbf{p}^{(g)})^2 \tag{16}
$$

where

$$
\begin{aligned}
q_i &= \tilde{p}_i (1 - \tilde{p}_i) \\
z_i &= \tilde{p}_0^{(g)} + \mathbf{x}_{C_i}^T \tilde{\mathbf{W}}_C \tilde{\mathbf{p}}^{(g)} + \frac{g_i - \tilde{p}_i}{\tilde{p}_i (1 + \tilde{p}_i)} \\
\tilde{p}_i &= e^{(p_0^{\tilde{(g)}} + \mathbf{x}_{C_i}^T \tilde{\mathbf{W}}_C \mathbf{p}^{\tilde{(g)}})} / (1 + e^{(p_0^{\tilde{(g)}} + \mathbf{x}_{C_i}^T \tilde{\mathbf{W}}_C \mathbf{p}^{\tilde{(g)}})})
\end{aligned} \tag{17}
$$

The parameters denoted with the ˜ symbol are the current parameters. With the quadratic approximation now replacing the negative log-likelihood in Eq. 7 and the rescaled weighting parameter β used instead of α (see Eq. 8), the objective function becomes:

$$
\begin{aligned}
& L(\mathbf{W}_C, \mathbf{P}_C^{(X)}, \mathbf{p}^{(g)}, p_0^{(g)}) \\
= {} & \frac{\beta}{2} \sum_i^I q_i (z_i - p_0^{(g)} - \mathbf{x}_{C_i}^T \mathbf{W}_C \mathbf{p}^{(g)})^2 \\
& + (1-\beta) \sum_i^I \left\| \mathbf{x}_{C\,i} - \mathbf{x}_{C_i}^T \mathbf{W}_C (\mathbf{P}_C^{(X)})^T \right\|_2^2 \\
& + \sum_r^R \lambda_{Lr} \, |\mathbf{w}_{Cr}|_1 + \sum_r^R \sum_k^K \lambda_{Gr} \sqrt{J_k} \left\| \mathbf{w}_r^{(k)} \right\|_2 \\
& + \lambda_R \left\| \mathbf{p}^{(g)} \right\|_2^2
\end{aligned} \tag{18}
$$

where q_i and z_i are defined as in Eq. 17. The optimization problem in Eq. 18 can be solved with an alternating procedure where the loadings $\mathbf{P}_C^{(X)}$ and the regression coefficients $\mathbf{p}^{(g)}$, $p_0^{(g)}$ are solved for conditional upon fixed values for the weights $\mathbf{W}_C$ and vice versa. The sparse group lasso problem for $\mathbf{W}_C$ is treated via coordinate descent (Friedman et al., 2010a), while closed-form solutions exist for the conditional updates of $\mathbf{p}^{(g)}$, $p_0^{(g)}$ and $\mathbf{P}_C^{(X)}$. The derivation of these updating rules can be found in Appendices C and D. After each run of conditional estimation of the parameters, the quadratic approximation in Eq. 18 is updated with new values of q_i and z_i calculated with the current parameters. To prevent the divergence of the coefficients, when the absolute difference between the current probability $\tilde{p}_i$ and 1 is less or equal to 10^{-5}, $\tilde{p}_i$ is fixed at 1. This follows the recommendation of Friedman et al. (2010b) which proposed a framework of combining regularization with GLM.

A schematic outline of the algorithm is provided in what follows. The optimization procedure that we propose here closely follows those proposed for SCaDS and SPCovR (de Schipper & Van Deun, 2018; Van Deun et al., 2018). This procedure boils down to solving for all components together (unlike deflation methods that solve for each component in turn). The alternating routine continues until the algorithm converges to a stationary point, usually a local minimum. To avoid local minima problems, we recommend using multiple random and a rational starting value based on PCovR.

```
Inputs:
    $\mathbf{X}_C$ and $\mathbf{g}$, number of components $R$, rescaled weighting parameter $\beta$, regularization parameters $\lambda_{Lr}$, $\lambda_{Gr}$ and $\lambda_R$, maximum number of iterations $T$, convergence threshold $\epsilon \geq 0$
Initialize:
    $\mathbf{W}_C \leftarrow \mathbf{W}_C{}^{(0)}$, $\mathbf{P}_C^{(X)} \leftarrow \mathbf{P}_C^{(X)\,(0)}$, $\mathbf{p}^{(g)} \leftarrow \mathbf{p}^{(g)\,(0)}$, $p_0{}^{(g)} \leftarrow p_0{}^{(g)\,(0)}$, $L_0 \leftarrow$ Initial loss,
    Loss difference $d \leftarrow 1$, Iteration counter $t \leftarrow 1$
while $t < T$ or $\epsilon < d$ do
    Update of $q_i, z_i$ given $\mathbf{W}_C{}^{(t-1)}$, $\mathbf{P}_C^{(X)\,(t-1)}$, $\mathbf{p}^{(g)\,(t-1)}$, and $p_0^{(g)\,(t-1)}$
    Conditional estimation of $\mathbf{W}_C{}^{(t)}$ given $\mathbf{P}_C^{(X)\,(t-1)}$, $\mathbf{p}^{(g)\,(t-1)}$ and $p_0^{(g)\,(t-1)}$
    Update of $q_i, z_i$ given $\mathbf{W}_C{}^{(t)}$, $\mathbf{P}_C^{(X)\,(t-1)}$, $\mathbf{p}^{(g)\,(t-1)}$, and $p_0^{(g)\,(t-1)}$
    Conditional estimation of $\mathbf{P}_C^{(X)\,(t)}$, $\mathbf{p}^{(g)\,(t)}$ and $p_0^{(g)\,(t)}$ given $\mathbf{W}_C{}^{(t)}$
    $L_u \leftarrow$ updated loss given $\mathbf{W}_C{}^{(t)}$, $\mathbf{P}_C^{(X)\,(t)}$, $\mathbf{p}^{(g)\,(t)}$ and $p_0^{(g)\,(t)}$
    $d \leftarrow L_0 - L_u$
    $t \leftarrow t + 1$
    $L_0 \leftarrow L_u$
end while
```

Algorithm 1 SCD-Cov-logR.

Appendix C: Estimation of $\mathbf{W}_C$

Conditional estimation of $\mathbf{W}_C$ given the other parameters $\mathbf{P}^{(X)}$, $\mathbf{p}^{(g)}$ and $p_0^{(g)}$ pertains to a sparse group lasso problem. The SCD-Cov-logR objective function with the quadratic approximation of the negative log-likelihood (Eq. 18) is first arranged with respect to the weights corresponding to predictor block k and component r^*:

$$\begin{aligned}
&L(\mathbf{w}_{r^*}^{(k)}, \mathbf{P}_C^{(X)}, \mathbf{p}^{(g)}, p_0^{(g)}) = \\
&\frac{\beta}{2}\sum_i^I q_i \left(z_i - p_0^{(g)} - \sum_r^R \sum_{l \neq k}^K p_r^{(g)} \mathbf{x}_i^{(l)\,T} \mathbf{w}_r^{(l)} \right. \\
&\qquad \left. - \sum_{r \neq r^*}^R p_r^{(g)} \mathbf{x}_i^{(k)\,T} \mathbf{w}_r^{(k)} - p_{r^*}^{(g)} \mathbf{x}_i^{(k)\,T} \mathbf{w}_{r^*}^{(k)} \right)^2 \\
&+(1-\beta)\sum_i^I \left\| \mathbf{x}_{C\,i} - \sum_r^R \sum_{l \neq k}^K \mathbf{w}_r^{(l)\,T} \mathbf{x}_i^{(l)} \mathbf{p}_{C\,r}^{(X)} \right. \\
&\qquad \left. - \sum_{r \neq r^*} \mathbf{w}_r^{(k)\,T} \mathbf{x}_i^{(k)} \mathbf{p}_{C\,r}^{(X)} - \mathbf{w}_{r^*}^{(k)\,T} \mathbf{x}_i^{(k)} \mathbf{p}_{C\,r^*}^{(X)} \right\|_2^2 \\
&+\lambda_L \left| \mathbf{w}_{r^*}^{(k)} \right|_1 + \lambda_G \sqrt{J_k} \left\| \mathbf{w}_{r^*}^{(k)} \right\|_2
\end{aligned} \tag{19}$$

Taking the derivative with respect to $\mathbf{w}_{r^*}^{(k)}$ we get:

$$\begin{aligned}
&-\ \beta \sum_i^I q_i\, p_{r^*}^{(g)} (Z_i^{(k)} - p_{r^*}^{(g)} \mathbf{x}_i^{(k)\,T} \mathbf{w}_{r^*}^{(k)}) \mathbf{x}_i^{(k)} \\
&-\ 2(1-\beta) \sum_i^I (Y_i^{(k)} - \mathbf{w}_{r^*}^{(k)\,T} \mathbf{x}_i^{(k)}) \mathbf{x}_i^{(k)} \\
&+\ \lambda_L \partial \left| \mathbf{w}_{r^*}^{(k)} \right|_1 + \lambda_G \sqrt{J_k} \partial \left\| \mathbf{w}_{r^*}^{(k)} \right\|_2
\end{aligned} \tag{20}$$

where

$$\begin{aligned}
Z_i^{(k)} &= z_i - p_0^{(g)} - \sum_r^R \sum_{l \neq k}^K p_r^{(g)} \mathbf{x}_i^{(l)\,T} \mathbf{w}_r^{(l)} \\
&\quad - \sum_{r \neq r^*}^R p_r^{(g)} \mathbf{x}_i^{(k)\,T} \mathbf{w}_r^{(k)} \\
Y_i^{(k)} &= \mathbf{x}_{C\,i}^{\,T} \mathbf{p}_{C\,r^*}^{(X)} - \sum_{l \neq k}^K \mathbf{w}_{r^*}^{(l)\,T} \mathbf{x}_i^{(l)}
\end{aligned} \tag{21}$$

The subdifferential of $\left\| \mathbf{w}_{r^*}^{(k)} \right\|_2$ is defined as the following:

$$\partial \left\| \mathbf{w}_{r^*}^{(k)} \right\|_2 = \begin{cases} \dfrac{\hat{\mathbf{w}}_{r^*}^{(k)}}{\left\| \hat{\mathbf{w}}_{r^*}^{(k)} \right\|_2}, & \text{if } \hat{\mathbf{w}}_{r^*}^{(k)} \neq \mathbf{0} \\ \in \{\mathbf{u} : \|\mathbf{u}\|_2 \leq 1\}, & \text{if } \hat{\mathbf{w}}_{r^*}^{(k)} = \mathbf{0} \end{cases} \tag{22}$$

where $\mathbf{u}$ is a vector of equal length as $\mathbf{w}_{r^*}^{(k)}$.

The jth element of the subdifferential of $\partial \left| \mathbf{w}_{r^*}^{(k)} \right|_1$ is defined as the following:

$$\partial \left(\left| \mathbf{w}_{r^*}^{(k)} \right|_1 \right)_j = \begin{cases} \operatorname{sign}\left(\hat{w}_{jr^*}^{(k)} \right), & \text{if } \hat{w}_{jr^*}^{(k)} \neq 0 \\ \in \{v : |v| \leq 1\}, & \text{if } \hat{w}_{jr^*}^{(k)} = 0 \end{cases} \tag{23}$$

where v is a scalar.

By equating Eq. 20 to zero and rearranging, the condition that an optimal solution satisfies with $\hat{\mathbf{w}}_{r^*}^{(k)} = \mathbf{0}$ is the following:

$$\begin{aligned}
&\left\| S\left(\sum_i^I \left(\beta q_i\, p_{r^*}^{(g)} Z_i^{(k)} + 2(1-\beta) Y_i^{(k)} \right) \mathbf{x}_i^{(k)}, \lambda_L \right) \right\|_2 \\
&\leq \lambda_G \sqrt{J_k}
\end{aligned} \tag{24}$$

where S(.) is a element-wise soft-thresholding operator.

In the case that Eq. 24 is not satisfied and thus $\hat{\mathbf{w}}_{r*}^{(k)} \neq \mathbf{0}$, we find the conditions for an optimal solution for the hth element of the weights concerning predictor block k and component r^*; $w_{hr*}^{(k)}$. We first write the objective function with respect to $w_{hr*}^{(k)}$.

$$\begin{aligned}
&L(w_{hr*}^{(k)}, \mathbf{P}_C^{(X)}, \mathbf{p}^{(g)}, p_0^{(g)}) \\
&= \frac{\beta}{2} \sum_i^I q_i (z_i - p_0^{(g)} - \sum_r^R \sum_l^K \sum_{j \neq h}^{J_k} p_r^{(g)} x_{ij}^{(l)} w_{jr}^{(l)} \\
&- \sum_{r \neq r*}^R \sum_{l \neq k}^K p_r^{(g)} x_{ih}^{(l)} w_{hr}^{(l)} - p_{r*}^{(g)} x_{ih}^{(k)} w_{hr*}^{(k)})^2 \\
&+(1-\beta) \sum_i^I \Bigg\| \mathbf{x}_{Ci} - \sum_r^R \sum_l^K \sum_{j \neq h}^{J_k} \mathbf{p}_r^{(X)} x_{ij}^{(l)} w_{jr}^{(l)} \\
&- \sum_{r \neq r*}^R \sum_{l \neq k}^K \mathbf{p}_r^{(X)} x_{ih}^{(l)} w_{hr}^{(l)} - \mathbf{p}_{r*}^{(X)} x_{ih}^{(k)} w_{hr*}^{(k)} \Bigg\|_2^2 \\
&+\lambda_L \left| w_{hr*}^{(k)} \right| + \lambda_G \sqrt{J_k} \left\| \mathbf{w}_{r*}^{(k)} \right\|_2
\end{aligned} \tag{25}$$

Taking the derivative with respect to $w_{hr*}^{(k)}$:

$$\begin{aligned}
&-\beta \sum_i^I q_i p_{r*}^{(g)} x_{ih}^{(k)} (Z_i - p_{r*}^{(g)} x_{ih}^{(k)} w_{hr*}^{(k)}) \\
&-2(1-\beta) \sum_i^I x_{ih}^{(k)} (Y_i - x_{ih}^{(k)} w_{hr*}^{(k)}) \\
&+\lambda_L \partial \left| w_{hr*}^{(k)} \right| + \lambda_G \sqrt{J_k} \partial \left\| \mathbf{w}_{r*}^{(k)} \right\|_2
\end{aligned} \tag{26}$$

where

$$\begin{aligned}
Z_i &= z_i - p_0^{(g)} - \sum_r^R \sum_l^K \sum_{j \neq h}^{J_k} p_r^{(g)} x_{ij}^{(l)} w_{jr}^{(l)} \\
&\quad - \sum_{r \neq r*}^R \sum_{l \neq k}^K p_r^{(g)} x_{ih}^{(l)} w_{hr}^{(l)} \\
Y_i &= \mathbf{x}_{C\,i}^{\,T} \mathbf{p}_{C\,r*}^{\,(X)} - \sum_l^K \sum_{j \neq h}^{J_k} x_{ij}^{(l)} w_{jr*}^{(l)} - \sum_{l \neq k}^K x_{ih}^{(l)} w_{hr*}^{(l)}
\end{aligned} \tag{27}$$

The subdifferential of $\left\| \mathbf{w}_{r*}^{(k)} \right\|_2$ with respect to $w_{hr*}^{(k)}$ is provided in Eq. 22; it is the hth element of $\frac{\hat{\mathbf{w}}_{r*}^{(k)}}{\left\| \hat{\mathbf{w}}_{r*}^{(k)} \right\|_2}$. The subdifferential of $\partial \left| w_{hr*}^{(k)} \right|$ is defined as the following:

$$\partial \left| w_{hr*}^{(k)} \right| = \begin{cases} \text{sign}\left(\hat{w}_{hr*}^{(k)} \right), & \text{if } \hat{w}_{hr*}^{(k)} \neq 0 \\ \in \{v : |v| \leq 1\}, & \text{if } \hat{w}_{hr*}^{(k)} = 0 \end{cases} \tag{28}$$

where v is a scalar.

We can equate the derivate to zero to find the optimality conditions for $\hat{w}_{hr*}^{(k)}$, which can be summarized by the following:

$$\hat{w}_{hr*}^{(k)} = \frac{S(\sum_i^I x_{ih}^{(k)} (\beta p_{r*}^{(g)} q_i Z_i + 2(1-\beta) Y_i), \lambda_L)}{\beta {p_{r*}^{(g)}}^2 \sum_i^I q_i {x_{ih}^{(k)}}^2 + 2(1-\beta) \sum_i^I {x_{ih}^{(k)}}^2 + \lambda_G \sqrt{J_k} / \left\| \mathbf{w}_{r*}^{(k)} \right\|_2} \tag{29}$$

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

```
for $r^*$ in $1 : R$ do
    for $k$ in $1 : K$ do
        if $\left\| S(\sum_i^I (\beta q_i p_{r*}^{(g)} Z_i^{(k)} + 2(1-\beta) Y_i^{(k)}) \mathbf{x}_i^{(k)}, \lambda_L) \right\|_2 \leq \lambda_G \sqrt{J_k}$ then
            $\hat{\mathbf{w}}_{r*}^{(k)} \leftarrow \mathbf{0}$
        for $h$ in $1 : J_k$ do
            $\hat{w}_{hr*}^{(k)} \leftarrow \frac{S(\sum_i^I x_{ih}^{(k)} (\beta p_{r*}^{(g)} q_i Z_i + 2(1-\beta) Y_i), \lambda_L)}{\beta {p_{r*}^{(g)}}^2 \sum_i^I q_i {x_{ih}^{(k)}}^2 + 2(1-\beta) \sum_i^I {x_{ih}^{(k)}}^2 + \lambda_G \sqrt{J_k} / \left\| \mathbf{w}_{r*}^{(k)} \right\|_2}$
```

Algorithm 2 Coordinate descent for sparse group lasso.

Appendix D: Estimation of $\mathbf{p}^{(g)}$, $p_0^{(g)}$ and $\mathbf{P}_C^{(X)}$

Closed-form solutions exist for the regression coefficients and the intercept.

$$\begin{aligned}
\hat{\mathbf{p}}^{(g)} &= [(\mathbf{X}_C \mathbf{W}_C)^T \mathbf{Q} \mathbf{X}_C \mathbf{W}_C + (2/\alpha) \lambda_R \mathbf{I}_R]^{-1} \\
&\quad [(\mathbf{X}_C \mathbf{W}_C)^T \mathbf{Q} \mathbf{z} - p_0^{(g)} (\mathbf{X}_C \mathbf{W}_C)^T \mathbf{q}]
\end{aligned} \tag{30}$$

$$\hat{p}_0^{(g)} = \left(\sum_i^I q_i \left(z_i - \mathbf{x}_{C\,i}^{\,T} \mathbf{W}_C \mathbf{p}^{(g)} \right) \right) / \left(\sum_i^I q_i \right) \tag{31}$$

where $\mathbf{Q}$ is a diagonal matrix with the ith diagonal element being q_i. $\mathbf{q}$ and $\mathbf{z}$ are vectors with the elements being q_i and z_i respectively, which are defined in Eq. 17.

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

Appendix E: SCD-Cov-logR multiclass algorithm

Like for the binary problem, the solution to Eq. 10 is found by iteratively reweighted least squares. Partial quadratic approximation can be conducted such that only parameters that concern the mth category can vary at a time. With the quadratic approximation replacing the negative log-likelihood in Eq. 10 and the rescaled weighting parameter β used instead of α (see Eq. 8), the objective function becomes:

$$
\begin{aligned}
&L(\mathbf{W}_C, \mathbf{P}_C^{(X)}, \mathbf{p}_m^{(g)}, p_{0m}^{(g)}) \\
&= \frac{\beta}{2}\sum_i^I q_i (z_i - p_{0m}^{(g)} - \mathbf{x}_{C_i}^T \mathbf{W}_C \mathbf{p}_m^{(g)})^2 \\
&+(1-\beta)\sum_i^I \left\| \mathbf{x}_{Ci} - \mathbf{x}_{C_i}^T \mathbf{W}_C (\mathbf{P}_C^{(X)})^T \right\|_2^2 \\
&+\sum_r^R \lambda_{Lr} \left| \mathbf{w}_{Cr} \right|_1 + \sum_r^R \sum_k^K \lambda_{Gr} \sqrt{J_k} \left\| \mathbf{w}_r^{(k)} \right\|_2 \\
&+\lambda_R \left\| \mathbf{p}_m^{(g)} \right\|_2^2 \qquad (32)
\end{aligned}
$$

where

$$
\begin{aligned}
q_i &= \tilde{p}_i (1 - \tilde{p}_i) \\
z_i &= \tilde{p}_{0m}^{(g)} + \mathbf{x}_{C_i}^T \tilde{\mathbf{W}}_C \tilde{\mathbf{p}}_m^{(g)} + \frac{g_i - \tilde{p}_i}{\tilde{p}_i (1 + \tilde{p}_i)} \\
\tilde{p}_i &= e^{(\tilde{p}_{0m}^{(g)} + \mathbf{x}_{C_i}^T \tilde{\mathbf{W}}_C \tilde{\mathbf{p}}_m^{(g)})} / (1 + \sum_m^{M-1} e^{(\tilde{p}_{0m}^{(g)} + \mathbf{x}_{C_i}^T \tilde{\mathbf{W}}_C \tilde{\mathbf{p}}_m^{(g)})}) \qquad (33)
\end{aligned}
$$

the parameters denoted with the ˜ symbol are the current parameters. The loadings are constrained to be column-orthogonal: $(\mathbf{P}_C^{(X)})^T \mathbf{P}_C^{(X)} = \mathbf{I}_R$. This optimization problem can be solved with an alternating procedure similar to that of the binary classification. In fact, the conditional estimation of the parameters is done in the same way as for the binary problem (shown in Appendices C and D) with a small tweak on the definition of certain quantities. We can first notice that this objective function with quadratic approximation with respect to category m can be considered as a binary problem between category m and the baseline category M. It can be seen that the only difference between the functions for the multiclass (Eqs. 32 and 33) and the binary (Eqs. 18 and 17) problems is the definition of the current parameter $\tilde{p}_i$. Therefore, from the binary objective function (Eq. 18), computing $\tilde{p}_i$ by following (Eq. 33) and replacing the regression coefficients $\mathbf{p}^{(g)}$, $p_0^{(g)}$ into $\mathbf{p}_m^{(g)}$, $p_{0m}^{(g)}$ specific for category m would enable us to rely on the same solutions for the conditional updates of the quantities $\mathbf{W}_C$, $\mathbf{p}_m^{(g)}$, $p_{0m}^{(g)}$ and $\mathbf{P}_C^{(X)}$. The algorithm for the multiclass problem however cycles over the $M-1$ categories on top of the conditional updates of the quantities. After each run of conditional estimation of the quantities, the quadratic approximation in Eq. 32 is updated with new values of q_i and z_i calculated with the current parameters.

A schematic outline of the algorithm is provided below. The alternating routine continues until the algorithm converges to a stationary point, usually a local minimum. To avoid local minima problems, we recommend using multiple random and a rational starting value based on PCovR.

```
Inputs:
    $\mathbf{X}_C$ and $\mathbf{G}$, number of components $R$, rescaled weighting parameter $\beta$, regularization parameters $\lambda_{Lr}$, $\lambda_{Gr}$ and $\lambda_R$, maximum
    number of iterations $T$, convergence threshold $\epsilon \geq 0$
Initialize:
    $\mathbf{W}_C \leftarrow \mathbf{W}_C^{(0)}$, $\mathbf{P}_C^{(X)} \leftarrow \mathbf{P}_C^{(X)(0)}$, $\mathbf{p}_m^{(g)} \leftarrow \mathbf{p}_m^{(g)(0)}$, $p_{0m}^{(g)} \leftarrow p_{0m}^{(g)(0)}$, $L_0 \leftarrow$ Initial loss,
    Loss difference $d \leftarrow 1$, Iteration counter $t \leftarrow 1$
while $t < T$ or $\epsilon < d$ do
    for $m \leftarrow 1$ to $M-1$ do
        Update of $q_i, z_i$ given $\mathbf{W}_C^{(t-1)}$, $\mathbf{P}_C^{(X)(t-1)}$, $\mathbf{p}_m^{(g)(t-1)}$, and $p_{0m}^{(g)(t-1)}$
        Conditional estimation of $\mathbf{W}_C^{(t)}$ given $\mathbf{P}_C^{(X)(t-1)}$, $\mathbf{p}_m^{(g)(t-1)}$ and $p_{0m}^{(g)(t-1)}$
        Update of $q_i, z_i$ given $\mathbf{W}_C^{(t)}$, $\mathbf{P}_C^{(X)(t-1)}$, $\mathbf{p}_m^{(g)(t-1)}$, and $p_{0m}^{(g)(t-1)}$
        Conditional estimation of $\mathbf{P}_C^{(X)(t)}$, $\mathbf{p}_m^{(g)(t)}$ and $p_{0m}^{(g)(t)}$ given $\mathbf{W}_C^{(t)}$
        $L_u \leftarrow$ updated loss given $\mathbf{W}_C^{(t)}$, $\mathbf{P}_C^{(X)(t)}$, $\mathbf{p}_m^{(g)(t)}$ and $p_{0m}^{(g)(t)}$
    $d \leftarrow L_0 - L_u$
    $t \leftarrow t + 1$
    $L_0 \leftarrow L_u$
end while
```

Algorithm 3 SCD-Cov-logR for multiclass classification.

Appendix F: The scree test with acceleration factor conducted to determine the number of covariates for the toy example dataset

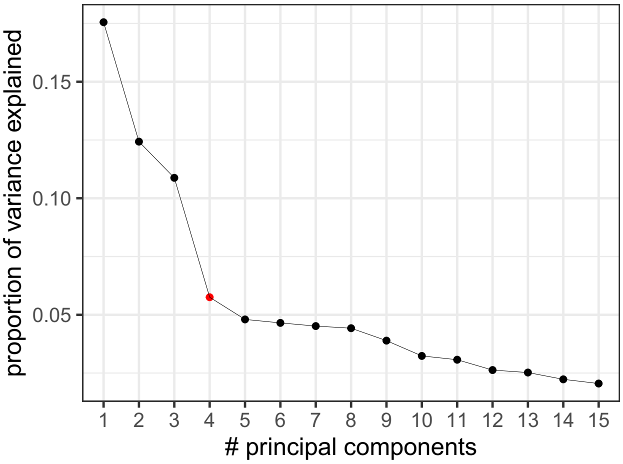

Fig. 4 It can be seen that the sharpest change of slopes occurs at four principal components. Three components are therefore retained in the model

Appendix G: Toy example dataset: model selection via exhaustive grid search of all parameters

Instead of the sequential model selection procedure adopted in the toy example dataset ("Toy example"), we have conducted cross-validation (CV) in which all of the possible parameters are crossed exhaustively. The ranges of the parameters considered were the same as in the sequential procedure:

- β: [0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9]
- λ_R: [0.1,0.5,1,3,5,10,30,50]
- λ_L: [0.5,1,5,7,10,15,30,45,100]
- λ_G: [0.1,0.5,1,2,5,10]

For the number of covariates R, we adopted the range of [1,2,3,4] because PCA on the predictor data matrix revealed that from the fifth component onwards, the proportion of explained variance is smaller than 5% (this has been depicted in Appendix F). Crossing all of the possible parameters, we administered fivefold CV to 15552 models in total.

The model with the smallest CV error was characterized by the parameters: $R = 3, \beta = 0.6, \lambda_R = 0.5, \lambda_L = 10$, $\lambda_G = 5$. The estimated weights and regression coefficients are reported in Table 7. It can be seen that the estimates are very similar to the ones found by the model obtained through the sequential approach of CV.

If we apply the one standard error rule to select the simplest model among those within 1 SE from the minimum CV error, we would need to make a choice regarding which parameter to look consider first. The number of covariates R can be considered as the most influential parameter, followed by the weighting parameter β. Prioritizing these two parameters, the one standard error rule selects the model: $R = 2, \beta = 0.5, \lambda_R = 0.5, \lambda_L = 30, \lambda_G = 1$. Table 8 shows the estimates of this two-covariate model. It can be seen that the covariate which is distinctive to the second predictor block (D2 in Table 1) is excluded from this model. This is sensible because this covariate was defined to have a very small predictive influence on the outcome variable when generating the data: population value of the logistic regression weight was set at -0.01. Hence, it is natural that the exhaustive CV approach that only considers the prediction error could result in omitting this covariate. The two covariates extracted are in agreement to the covariates found by the sequential approach of CV.

Table 7 Weights and regression coefficients provided by the three-covariate model with the smallest cross-validation error

Weights				Logistic regression coefficients	
Block 1					
x1	0.420	0	0	1	-1.272
x2	0.420	0	0	2	-0.096
x3	0.439	0	0	3	1.499
x4	0.486	0	0	intercept	-0.206
x5	0	0	0.330		
x6	0	0	0.324		
x7	0	0	0.288		
x8	0	0	0.261		
x9	0	0	0		
x10	0	0	0		
x11	0	0	0		
x12	0	0	0		
x13	0	0	0		
x14	0	0	-0.021		
x15	0	0	0		
Block 2					
x16	0	0	0.343		
x17	0	0	0.361		
x18	0	0	0.316		
x19	0	0	0.256		
x20	0	0.437	0		
x21	0	0.429	0		
x22	0	0.439	0		
x23	0	0.470	0		
x24	0	0	0		
x25	0	0	0		
x26	0	-0.085	0		
x27	0	0	0		
x28	0	0	0		
x29	0	0	0		
x30	0	0	0		

Appendix H: Data generation for multiclass toy example dataset

The data generating setup employed for our simulation study is adapted slightly such that it can generate more than two categories, in generating the toy example dataset for the multiclass classification problem. As for the simulation study, two blocks of predictor variables were generated from three underlying covariates; one distinctive covariate per each predictor block and one common covariate. Each predictor block comprised of 15 variables ($J = 30$ in total), and $I = 1000$ observation units were generated. With the population weights and logistic regression coefficients provided in Table 4, the toy example dataset was generated via the following setup:

$$\mathbf{T} \sim \mathcal{MVN}(\mathbf{0}, \Sigma = 50^2\mathbf{I}_3)$$

$$\mathbf{E} \sim \mathcal{MVN}(\mathbf{0}, \Sigma_E = \sigma^2\mathbf{I}_J)$$

$$\mathbf{X}_C \leftarrow \mathbf{T}\mathbf{W}_C^T + \mathbf{E}$$

Table 8 Weights and regression coefficients provided by the two-covariate model with the one standard error rule

Weights			Logistic regression coefficients	
Block 1			1	−1.263
x1	0.369	0	2	1.481
x2	0.380	0	intercept	−0.199
x3	0.507	0		
x4	0.427	0		
x5	0	0.345		
x6	0	0.311		
x7	0	0.265		
x8	0	0.212		
x9	0	0		
x10	0	0		
x11	0	0		
x12	0	0		
x13	0	0		
x14	0	0		
x15	0	0		
Block 2				
x16	0	0.337		
x17	0	0.378		
x18	0	0.302		
x19	0	0.206		
x20	0	0		
x21	0	0		
x22	0	0		
x23	0	0		
x24	0	0		
x25	0	0		
x26	0	0		
x27	0	0		
x28	0	0		
x29	0	0		
x30	0	0		

$$\begin{aligned}
&\mathbf{z}_m \leftarrow exp(\mathbf{T}\mathbf{p}_m^{(g)})/(1+exp(\sum_{m'}^{M-1}\mathbf{T}\mathbf{p}_{m'}^{(g)}))\\
&\quad \text{for } m = 1,\ldots,M-1\\
&\mathbf{z}_M \leftarrow 1/(1+exp(\sum_{m'}^{M-1}\mathbf{T}\mathbf{p}_{m'}^{(g)}))\\
&g_{im} \sim Multinoulli(z_{im}) \text{ for } m = 1,\ldots,M
\end{aligned} \tag{34}$$

where $\mathbf{T}$, Σ and $\mathbf{W}_C$ are all defined in the same manner as in the simulation study (see "Design and procedure"). The predictors $\mathbf{X}_C$ are generated by multiplying the covariate scores matrix with the weights matrix and adding random error. The diagonal covariance matrix Σ_E that governs the variance of error variables $\mathbf{E}$ is specified such that the covariates $\mathbf{T}$ account for 50% of variance in $\mathbf{X}_C$. $\mathbf{p}_m^{(g)}$ indicates the logistic regression coefficients for the log-odds of the mth category as opposed to the baseline category $M = 3$. The statements in the fourth and the fifth lines together specify the ($I = 1000 \times M = 3$) matrix $\mathbf{Z}$; z_{im} denotes the probability of the ith observation belonging to mth category, defined according to the baseline-category logit model (Agresti, 2003). g_{im} is therefore sampled from a Multinoulli distribution defined by the prescribed probabilities z_{im}.

Appendix I

Table 9 Lasso and Group lasso penalty parameters initially fixed in the simulation study, per each condition

Dimensions	Relevant	VAF	λ_{G1}	λ_{G2}	λ_{G3}	λ_{L1}	λ_{L2}	λ_{L3}
low	D1,D2	0.8	0.5	0.5	0.5	20	10	20
low	D1,D2	0.5	0.5	0.5	0.5	30	15	30
low	D1,D2	0.2	0.5	0.5	0.5	30	15	30
low	D1,C	0.8	0.5	0.5	0.5	30	15	30
low	D1,C	0.5	0.5	0.5	0.5	30	15	30
low	D1,C	0.2	0.5	0.5	0.5	30	15	30
high	D1,D2	0.8	3	3	3	15	7.5	15
high	D1,D2	0.5	2	2	2	30	15	30
high	D1,D2	0.2	1	1	1	20	10	20
high	D1,C	0.8	1	1	1	10	10	10
high	D1,C	0.5	1	1	1	30	15	30
high	D1,C	0.2	1	1	1	10	10	10

Appendix J: The scree test with acceleration factor conducted to determine the number of covariates for the 500 Family dataset

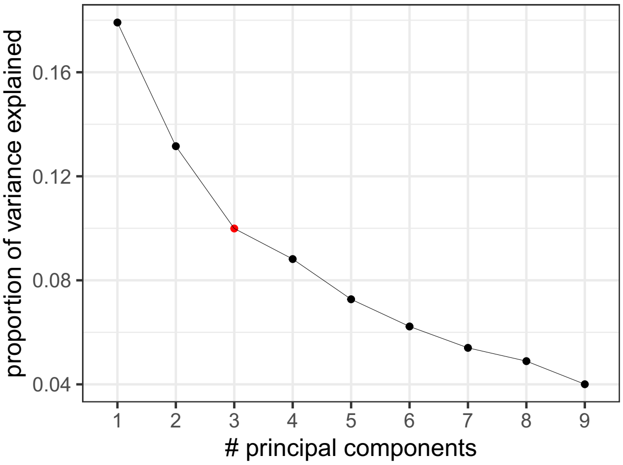

Fig. 5 It can be seen that the sharpest change of slopes occurs at three components. Two components are therefore retained in the model

Appendix K: Models constructed from the 500 Family dataset using the related methods

Table 10 Estimates provided by PCR (SCaDS-logR), DIABLO and regularized logistic regression for the 500 Family dataset

	SCaDS-logR		DIABLO		LogR
	1	2	1	2	b
Mother					
Relationship with partners	0	0.243	0	0	0
Argue with partners	0	0.247	0	0	0
Child’s bright future	0	0	0	0	0
Activities with children	0	0	0	0	0
Feeling about parenting	0	0.175	0	0	0
Communication with children	0	0.338	0	0	0
Argue with children	0	0.152	0	0	0
Confidence about oneself	0	0.382	0	0	0
Father					
Relationship with partners	0	0.097	0	0	0
Argue with partners	0	0.208	0	0	0
Child’s bright future	0	0	0	1	0.058
Activities with children	0	0	0	0	0
Feeling about parenting	0	0	0	0	0
Communication with children	0	0	0	0	0
Argue with children	0	0.255	0	0	0
Confidence about oneself	0	0.047	0	0	0
Child					
Child self-confidence/esteem	0.274	0	0	0	0
Social life and extracurricular	0.333	0	0	0	0
Importance of friendship	0.460	0	0	0	0
Self-image	0.360	0	1	0	0.278
Happiness	0.371	0	0	0	0
Confidence about the future	0.275	0	0	0	0

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

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.

Data Availability Code used to generate the results reported in the manuscript is available on GitHub: https://github.com/soogs/SCD-Cov-logR. Publicly available dataset from the 500 Family Study (Schneider & Waite, 2008) was used for empirical illustration.

Code Availability The implementation of SCD-Cov-logR was done in R and Rcpp, which can be found on GitHub: https://github.com/soogs/SCD-Cov-logR.

References

Agresti, A. (2003). *Categorical Data Analysis*. Hoboken: Wiley.

Babor, T. F., Higgins-Biddle, J., Saunders, J., & Monteiro, M. (2001). The alcohol use disorders identification test: Guidelines for use in. World Health Organization. Recuperado de https://apps.who.int/iris/handle/10665/67205.

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, *66*(3), 411–421.

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *17*(3), 166–173.

Barnes, D., Covinsky, K., Whitmer, R., Kuller, L., Lopez, O., & Yaffe, K. (2009). Predicting risk of dementia in older adults: The late-life dementia risk index. *Neurology*, *73*(3), 173–179.

Botella, J., Huang, H., & Suero, M. (2015). Meta-analysis of the accuracy of tools used for binary classification when the primary studies employ different references. *Psychological Methods*, *20*(3), 331.

Chen, D.-W., Miao, R., Deng, Z.-Y., Lu, Y.-Y., Liang, Y., & Huang, L. (2020). Sparse logistic regression with l1/2 penalty for emotion recognition in electroencephalography classification. *Frontiers in Neuroinformatics*, *14*, 29.

Chung, D., & Keles, S. (2010). Sparse partial least squares classification for high dimensional data. *Statistical Applications in Genetics and Molecular Biology 9*(1).

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., & Van Deun, K. (2018). Revealing the joint mechanisms in traditional data linked with big data. Zeitschrift für Psychologie.

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

Ding, B., & Gentleman, R. (2005). Classification using generalized partial least squares. *Journal of Computational and Graphical Statistics*, *14*(2), 280–298.

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

Friedman, J., Hastie, T., & Tibshirani, R. (2010b). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1.

Friedman, J., Hastie, T., Tibshirani, R., & et al. (2001). *The elements of statistical learning* Vol. 1. New York: Springer Series in Statistics New York.

Gizer, I. R., Ficks, C., & Waldman, I. D. (2009). Candidate gene studies of ADHD: A meta-analytic review. *Human Genetics*, *126*(1), 51–90.

Grizenko, N., Fortier, M.-E., Zadorozny, C., Thakur, G., Schmitz, N., Duval, R., & Joober, R. (2012). Maternal stress during pregnancy, ADHD symptomatology in children and genotype: Gene–environment interaction. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, *21*(1), 9.

Guerra-Urzola, R., Van Deun, K., Vera, J. C., & Sijtsma, K. (2021). A guide for sparse PCA: Model comparison and applications. Psychometrika, 1–27.

Heij, C., Groenen, P. J., & van Dijk, D. (2007). Forecast comparison of principal component regression and principal covariate regression. *Computational Statistics & Data Analysis*, *51*(7), 3612–3625.

Hill, L. S., Reid, F., Morgan, J. F., & Lacey, J. H. (2010). Scoff, the development of an eating disorder screening questionnaire. *International Journal of Eating Disorders*, *43*(4), 344–351.

Jia, J., & Yu, B. (2010). On model selection consistency of the elastic net when p $\gg$ n. Statistica Sinica, 595–611.

Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *31*(3), 300–303.

Kahn, R. S., Khoury, J., Nichols, W. C., & Lanphear, B. P. (2003). Role of dopamine transporter genotype and maternal prenatal smoking in childhood hyperactive-impulsive, inattentive, and oppositional behaviors. *The Journal of Pediatrics*, *143*(1), 104–110.

Kawano, S., Fujisawa, H., Takada, T., & Shiroishi, T. (2018). Sparse principal component regression for generalized linear models. *Computational Statistics & Data Analysis*, *124*, 180–196.

Kiers, H. A., & Ten Berge, J. M. (1989). Alternating least squares algorithms for simultaneous components analysis with equal component weight matrices in two or more populations. *Psychometrika*, *54*(3), 467–473.

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. 7*(1).

Måge, I., Smilde, A. K., & Van der Kloet, F. M. (2019). Performance of methods that separate common and distinct variation in multiple data blocks. *Journal of Chemometrics*, *33*(1), e3085.

McFadden, D., et al. (1973). Conditional logit analysis of qualitative choice behavior.

McNeish, D. M. (2015). Using lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences. *Multivariate Behavioral Research*, *50*(5), 471–484.

Mioshi, E., Dawson, K., Mitchell, J., Arnold, R., & Hodges, J. R. (2006). The Addenbrooke's cognitive examination revised (ACE-R): A brief cognitive test battery for dementia screening. *International Journal of Geriatric Psychiatry: A Journal of the Psychiatry of Late Life and Allied Sciences*, *21*(11), 1078–1085.

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

Raiche, G., Magis, D., & Raiche, M. G. (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*.

Schneider, B., & Waite, L. J. (2008). The 500 family study [1998–2000: United States]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/ICPSR04549.v1

Schouteden, M., Van Deun, K., Pattyn, S., & Van Mechelen, I. (2013). SCA with rotation to distinguish common and distinctive information in linked data. *Behavior Research Methods*, *45*(3), 822–833.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, *22*(2), 231–245.

Singh, A., Gautier, B., Shannon, C. P., Vacher, M., Rohart, F., Tebbutt, S. J., & Le Cao, K.-A. (2016). Diablo–an integrative, multi-omics, multivariate method for multi-group classification. BioRxiv, page 067611.

Tenenhaus, M., Tenenhaus, A., & Groenen, P. J. (2017). Regularized generalized canonical correlation analysis: A framework for sequential multiblock component methods. *Psychometrika*, *82*(3), 737–777.

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

Tu, Y., & Lee, T.-H. (2019). Forecasting using supervised factor models. *Journal of Management Science and Engineering*, *4*(1), 12–27.

Tucker, L. R. (1951). A method for synthesis of factor analysis studies. Technical report. Educational Testing Service Princeton NJ.

Tutun, S., Ahmed, A. A., Irgil, S., Yesilkaya, I., Analytics, D., & Khasawneh, M. T. (2019). Detecting psychological symptom patterns using regularized multinomial logistic regression. In *2019 Institute of industrial and systems engineers annual conference and expo, IISE 2019, p 967087. Institute of Industrial and Systems Engineers, IISE.*

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

Van Mechelen, I., & Smilde, A. K. (2010). A generic linked-mode decomposition model for data fusion. *Chemometrics and Intelligent Laboratory Systems*, *104*(1), 83–94.

Vervloet, M., Van Deun, K., Van den Noortgate, W., & Ceulemans, E. (2013). On the selection of the weighting parameter value in principal covariates regression. *Chemometrics and Intelligent Laboratory Systems*, *123*, 36–43.

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.

Wold, H. (1982). Soft modeling: The basic design and some extensions. *Systems Under Indirect Observation*, *2*, 343.

Wold, S., Martens, H., & Wold, H. (1983). The multivariate calibration problem in chemistry solved by the PLS method. In *Matrix pencils*, (pp. 286–293): Springer.

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.

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.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.