Received: 30 November 2022 | Revised: 5 February 2024 | Accepted: 7 February 2024

DOI: 10.1002/bimj.202200334

Biometrical Journal

RESEARCH ARTICLE

Sparse Group Penalties for bi-level variable selection

Gregor Buch[1,2,3] | **Andreas Schulz**[1] | **Irene Schmidtmann**[2] | **Konstantin Strauch**[2] | **Philipp S. Wild**[1,3,4,5]

[1]Preventive Cardiology and Preventive Medicine, Department of Cardiology, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany

[2]Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany

[3]German Center for Cardiovascular Research (DZHK), Mainz, Germany

[4]Clinical Epidemiology and Systems Medicine, Center for Thrombosis and Hemostasis, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany

[5]Institute of Molecular Biology (IMB), Mainz, Germany

Correspondence
Gregor Buch and Philipp S. Wild, Preventive Cardiology and Preventive Medicine, Department of Cardiology, University Medical Center of the Johannes Gutenberg University Mainz, 55131 Mainz, Germany.
Email: Gregor.Buch@unimedizin-mainz.de and Philipp.Wild@unimedizin-mainz.de

Funding information
Federal Ministry of Education and Research (BMBF), Grant/Award Numbers: 031L0217A, 01EO1003, 01EO1503

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to computational complexity and data confidentiality issues.

Abstract

Many data sets exhibit a natural group structure due to contextual similarities or high correlations of variables, such as lipid markers that are interrelated based on biochemical principles. Knowledge of such groupings can be used through bi-level selection methods to identify relevant feature groups and highlight their predictive members. One of the best known approaches of this kind combines the classical *Least Absolute Shrinkage and Selection Operator* (LASSO) with the *Group LASSO*, resulting in the *Sparse Group LASSO*. We propose the Sparse Group Penalty (SGP) framework, which allows for a flexible combination of different SGL-style shrinkage conditions. Analogous to SGL, we investigated the combination of the *Smoothly Clipped Absolute Deviation* (SCAD), the *Minimax Concave Penalty* (MCP) and the *Exponential Penalty* (EP) with their group versions, resulting in the *Sparse Group SCAD*, the *Sparse Group MCP*, and the novel *Sparse Group EP* (SGE). Those shrinkage operators provide refined control of the effect of group formation on the selection process through a tuning parameter. In simulation studies, SGPs were compared with other bi-level selection methods (Group Bridge, composite MCP, and Group Exponential LASSO) for variable and group selection evaluated with the Matthews correlation coefficient. We demonstrated the advantages of the new SGE in identifying parsimonious models, but also identified scenarios that highlight the limitations of the approach. The performance of the techniques was further investigated in a real-world use case for the selection of regulated lipids in a randomized clinical trial.

KEYWORDS
bi-level selection, group variable selection, lipidomics, simulation study, Sparse Group LASSO

1 | INTRODUCTION

Omics (i.e., molecular) data, such as lipidomics or proteomics data, are frequently analyzed in medical science and biology. The aim of this research is to understand disease mechanisms, develop new diagnostics, or predict disease progression. Thereby, interest is often in identifying the most discriminating features, which justifies the use of variable selection methods.

An important characteristic of omics data sets is the intrinsic group structure. For example, proteoforms or single nucleotide polymorphisms can be grouped by genes or mapped to biological pathways. Consequently, the assumption that the data matrix consists of several variable groups seems reasonable, and it is necessary to consider this structure in the selection process. Since it is rarely the case that all members of a group are equally important, it is of limited use to reduce them to one dimension or to select solely at the group level. Instead, methods activating relevant groups and highlighting their predictive members are preferable. This can be achieved with so-called bi-level selection approaches (Huang et al., 2009).

Numerous bi-level selection methods have been introduced, which can be classified into two frameworks (Buch et al., 2023; Huang et al., 2012). In the first one, shrinkage operators acting at the variable and group level are combined hierarchically, while in the second they are combined additively. The hierarchical framework has led to several innovations and publicly available software (Breheny, 2015; Breheny & Huang, 2009; Huang et al., 2009), while the additive framework resulted only in the *Sparse Group LASSO* (SGL, where LASSO is Least Absolute Shrinkage and Selection Operator; Simon et al., 2013). The idea of combining two penalties additively (like in SGL) to unite their properties is not new (Zou & Hastie, 2005). However, in the context of bi-level selection procedures, alternative combinations have never been systematically investigated, so its full potential has not yet been exploited (Buch et al., 2023). More precisely, the limited approaches were studied independently (Bui et al., 2021; Liu et al., 2013), so their relationship was not formalized, and they were compared neither with each other nor with approaches of the hierarchical framework.

The aim of this work was to outline the possibilities of an additive combination of penalty functions for bi-level selection and to compare resulting approaches with established methods. Emphasis was on members of the additive framework applying the same operator at the group and the variable levels, which we referred to as *Sparse Group Penalties* (SGPs). We also introduce a novel approach of this kind, the *Sparse Group Exponential Penalty* (SGE), which additively combines an *Exponential Penalty* (EP) at the group and the variable levels. The SGE allows a refined control of the effect of a variable grouping on the selection process, which is outlined in the next section. An algorithm for solving the objective functions involving SGPs was subsequently considered, and simulation studies were conducted to evaluate the performance of various SGPs. Finally, the bi-level selection methods were applied to a real-world use case and the results discussed.

2 | PENALTIES FOR BI-LEVEL SELECTION

In regularized regressions, the motivation is to estimate the coefficient vector $\boldsymbol{\beta}$ by minimizing the objective function Q:

$$Q\ (\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{y}) = \ L\,(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{y}) + P(\boldsymbol{\beta}|\lambda, \theta). \tag{1}$$

Since Q consists of a loss function, L, and a penalty function, P, the minimum of Q is a compromise between both components. L quantifies the discrepancy between the response variable, $\boldsymbol{y}$, and predictions based on the explanatory variables in $\boldsymbol{X}$. P promotes setting $\boldsymbol{\beta}$-coefficients to zero, that is, removing variables from the model. The methods discussed in this paper are equal in terms of L, that is, $\frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$ for the linear regression, but they differ with respect to their definition of P. This part may contain several tuning parameters, denoted by θ, in addition to the tuning parameter λ that controls for the trade-off between L and P.

The approaches examined assume that $\boldsymbol{X}$ is composed of J variable groups, each consisting of k_j members. These groups can be of different sizes, but should be mutually exclusive. In addition, the coefficient vector $\boldsymbol{\beta}$ is assumed to be sparse, implying that only a subset of variables and variable groups should be estimated as nonzero.

2.1 | Sparse Group Penalties

To account for the group structure in $\boldsymbol{X}$, SGPs are considered for P, which are defined as follows:

$$P^{\mathrm{SGP}}(\beta|\lambda,\alpha,\theta) = \alpha \cdot \sum_{j=1}^{J} \sum_{k=1}^{K_j} P^V\left(\beta_{jk}|\lambda,\theta\right) + (1-\alpha) \cdot \sum_{j=1}^{J} P^G\left(\left\|\beta_{\mathrm{j}}\right\|_2|\lambda_j,\theta\right). \tag{2}$$

Such penalties consist of two components that are additively connected. In each component, a regularization term is used (P^V or P^G), weighted by either α or $(1-\alpha)$. The first operator, denoted P^V, is a shrinkage term applied to each element in the coefficient vector and should be designed to promote sparsity at the variable level. The second operator, P^G, is also a shrinkage term for model selection, but is applied only at the group level of the coefficients. Hence, the ℓ_2-norm of the coefficients of the J groups are the input of the function. Like indicated, λ and further tuning parameters, θ, can define the respective regularization terms and since P^V and P^G could be different functions, their θ may differ. For penalties acting on the group level, it is common to adjust the magnitude of the penalization of a group by its size (i.e., number of members) k_j. Concretely, groups are regularized by λ_j in P^G, with $\lambda_j = \sqrt{k_j} * \lambda$.

In addition to these tuning parameters, the parameter $\alpha \in [0,1]$ plays a central role as it modulates the combination of the two elements of (2). Setting α to 0 eliminates the first part of Equation (2) and leads to a pure group-level selection, while the other extreme simplifies the function to a single variable selection method. Values between these two extremes lead to a bi-level selection: Groups associated with the response and their predictive members are included in the model.

2.2 | Components of an SGP

The best known SGP is the SGL, which uses the LASSO (Tibshirani, 1996) at the variable level and the Group LASSO (Yuan & Lin, 2006) at the group level. Huang et al. (2012) pointed out that other shrinkage operators can be combined in the same way. They suggested replacing the LASSO components in SGL with alternative shrinkage operators, such as the *Smoothly Clipped Absolute Deviation* (SCAD; Fan & Li, 2001) or the *Minimax Concave Penalty* (MCP; Zhang, 2010), both of which also have a group-level selection version (Breheny & Huang, 2015). SCAD and MCP use an additional tuning parameter, γ, to relax the penalization of large coefficients. This can be advantageous because it provides a less biased estimate of the $\boldsymbol{\beta}$-coefficients compared to classical LASSO. When these penalty terms are combined in a manner analogous to SGL, the *Sparse Group SCAD* (SGS) and the *Sparse Group MCP* (SGM) are obtained. Alternatively, an EP could be applied to replace the LASSO components in (2). The EP has recently been considered in the context of hierarchical bi-level selection procedures (Breheny, 2015). An additive combination like the one used by SGL leads to the SGE.

To better understand the implications of such a modification, an examination of the resulting derivatives is instructive. For this purpose, the aforementioned penalties, their derivatives, and a visualization have been compiled in Figure 1. The derivative of a shrinkage operator describes the rate of penalization that a $\boldsymbol{\beta}$-coefficient receives and relates to the threshold that a coefficient must overcome to be set to a nonzero value (second column in the figure). In the simple case, this rate is fixed, as in the original LASSO. However, in many other penalty functions, the penalization changes upon certain conditions. For SCAD and MCP, the rate of penalization is set to zero for $\boldsymbol{\beta}$-coefficients larger than $\gamma * \lambda$, and for EP, the regularization gradually approaches 0 as the $\boldsymbol{\beta}$-coefficient becomes larger. This penalty relaxation reduces the bias in the estimation of the coefficients and was also extended to group-level selection. Here, the ℓ_2-norm of each group ($\|\boldsymbol{\beta}_j\|_2$) is penalized instead of shrinking the elements of the full coefficient vector independently. In the derivatives of the group-level selection methods, the penalization is constant until a group is activated. Then, the shrinkage effect applied to members of the activated group is weighted by $\frac{\beta_{jk}}{\|\beta_j\|_2}$, ensuring the major characteristic of group-level selection: an equitable distribution of the penalization across all members of a group. Variables with a relatively weak association with the response are penalized less, while predictors with a strong signal are penalized more. Thus, all coefficients in a group are estimated as nonzero or jointly removed from the model depending on whether their combined signal is greater than λ.

Combining variable-level and group-level penalties into an SGP results in an operator that shrinks a variable belonging to predictive group ($\left(\left\|\beta_j\right\|_2 > 0\right)$) less than the same variable belonging to a nonpredictive group ($\left(\left\|\beta_j\right\|_2 = 0\right)$). The assumption is that the association of a variable from a group already identified as relevant appears more reliable than a similar association but of a variable from a nonselected group.

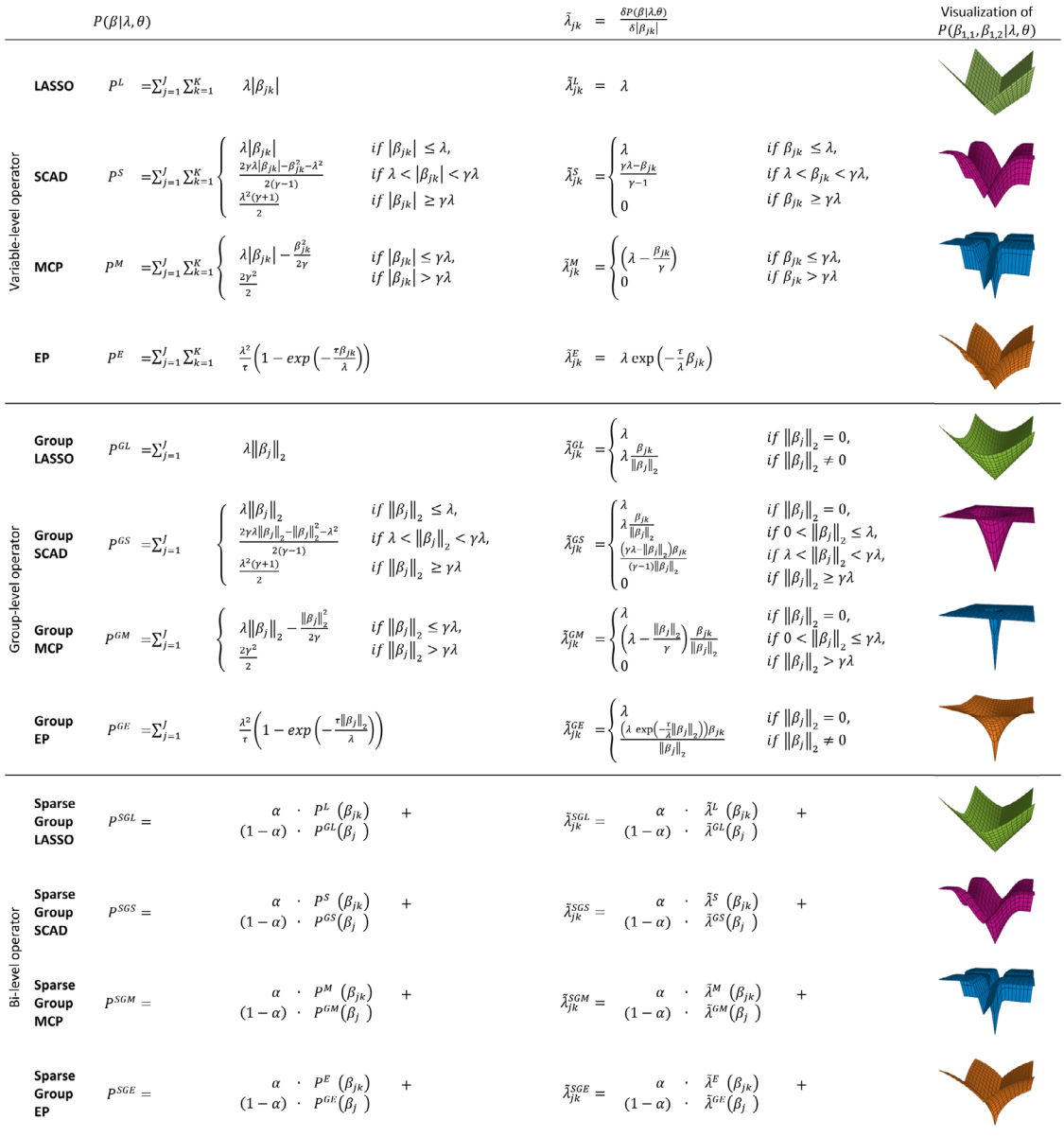

FIGURE 1 Sparse group penalties and their components. The formula and the derivative of penalty functions are given in the first and second columns. The visualizations in the third column show the magnitude of penalization for a grid of two coefficients belonging to two variables of the same group. Matrix X is a composite of J variable groups, each consisting of k_j members. Abbreviations: Smoothly Clipped Absolute Deviation Penalty (SCAD), Minimax Concave Penalty (MCP), Exponential Penalty (EP).

Comparing the 3D visualizations in the third column of the figure, the main differences between the penalties can be further illustrated. The height dimension of the plots corresponds to the penalization of an operator for a grid of values of two $\boldsymbol{\beta}$-coefficients belonging to the same group. At the center, both $\boldsymbol{\beta}$-coefficients are zero and so is the output of the penalty function. At the edges of the grid, the penalization is increased because the $\boldsymbol{\beta}$-coefficients are nonzero. For group-level operators, the penalty remains at a similar level when a $\boldsymbol{\beta}$-coefficient for a particular variable is zero as when the estimate is close to zero. This encourages that the $\boldsymbol{\beta}$-coefficients for variables of a selected group can be set to a nonzero value without increasing P substantially. This differs from variable-level operators, where a nonzero coefficient leads directly to a higher penalty. For bi-level operators, the $\boldsymbol{\beta}$-coefficient for a particular variable of an activated group can be estimated nonzero, with P increasing less than for a variable-level operator, but still not "for free" as with a group-level operator.

2.3 | Tuning parameters

SGPs have several parameters that need to be tuned (λ, α, and, depending on the penalty, γ or τ). Usually, λ is determined in regularization approaches using a cross-validation technique, which is also a recommended approach for SGPs (Simon et al., 2013). A grid search could be used for α, but since this parameter balances the impact of the group formation in the selection process, it may be better to define α based on the accuracy or relevance of the grouping. When $\alpha \neq 0$, there is a finite threshold that a variable itself must overcome to be included in the model. However, under the condition that $\alpha \neq 1$, variables in a group with a strong collective signal will be preferentially selected. This point is addressed in more detail in Section 4.1, where a simulation study is conducted to determine the recommended values as a function of confidence and relevance of available prior knowledge, and in Section 6, where the results of this study are discussed.

The literature proposed appropriate values for the tuning parameters that occur in addition when SCAD, MCP, or EP are part of the SGP. These penalties simplify to the original LASSO as $\gamma \to \infty$ or $\tau \to 0$. Accordingly, the suggested values for these tuning parameters try to be as far as possible from these extremes without risking convergence and discontinuity problems ($\gamma > 2$ for SCAD, $\gamma > 1$ for MCP, and $\tau > 0$ for EP). Otherwise, one would end up with LASSO-like results and should consider directly using the classical LASSO penalty instead. Therefore, tuning parameters were set to $\gamma = 4$ for SCAD, $\gamma = 3$ for MCP, and $\tau = 1$ for EP when they occur in P^V or P^G.

2.4 | Alternative approaches using a hierarchical framework

The previously addressed penalties are members of the additive framework for bi-level selection (Buch et al., 2023). An alternative, and historically older, approach to bi-level selection is the hierarchical framework (Breheny & Huang, 2009). Here, penalties acting on the group (P^{Outer}) and variable levels (P^{Inner}) are combined hierarchically, as given in Equation (3):

$$P^{Hierarchical}\left(\boldsymbol{\beta}|\lambda,\theta\right) = \sum_{j=1}^{J} P^{Outer}\left(\sum_{k=1}^{K_j} P^{Inner}\left(\beta_{jk}|\lambda,\theta\right)\right). \tag{3}$$

A pioneering method of this kind is *Group-Bridge* (G-Bridge; Breheny & Huang, 2009), which combines Bridge penalty (Frank & Friedman, 1993) at the group level and LASSO at the variable level. The *composite Minimax Concave Penalty* (cMCP; Breheny & Huang, 2009) uses MCP for both operators, and the *Group Exponential LASSO* (GEL; Breheny, 2015) uses EP as the outer penalty and LASSO as the inner penalty.

3 | COMPUTATIONAL ALGORITHM FOR THE SGP FRAMEWORK

As a next step, an algorithm for SGPs was developed. Gradient descent and coordinate descent have been proposed to solve the objective function (1) with an SGP (Liu et al., 2013; Simon et al., 2013). For the bi-level selection methods within the hierarchical framework, a Local Linear Approximated Coordinate Descent (LLACD) was used (Zou & Li, 2008). This algorithm uses the first derivative of the penalty function to approximate the rate of penalization for a coefficient, to update it in the style of a classical coordinate descent. Applying this strategy to SGPs, the resulting algorithm consists of two loops over which iteration occurs. An outer loop that iterates over the J variable groups and an inner loop that iterates over the K variables within a group, as illustrated in Algorithm 1.

Here, results of the previous update step are indicated by (s) and the current residuals are denoted by r. In classical LASSO, where no information about the group structure is used, one would update using the soft-thresholding operator given in Equation (4).

$$S\left(\beta,\lambda\right) = \begin{cases} \beta - \lambda & \text{if } \beta \geq \lambda, \\ 0 & \text{if } |\beta| < \lambda, \\ \beta + \lambda & \text{if } \beta \leq -\lambda \end{cases} \tag{4}$$

ALGORITHM 1 Local linear approximated coordinate descent for linear regression with a Sparse Group Penalty

repeat
  for $j = 1, 2, \ldots, J$ **do**
    for $k = 1, 2, \ldots, K$ **do**
      $z_{jk} = n^{-1}\boldsymbol{x}_{jk}^T r + \beta_{jk}^{(s)}$
      $\tilde{\lambda}_{jk} = \alpha \dot{P}^V\left(\beta_{jk}^{(s)}\lambda\right) + (1-\alpha)\dot{P}^G\left(\left\|\beta_j^{(s)}\right\|_2 \lambda\right)$
      $\beta_{jk}^{(s+1)} \leftarrow S(z_{jk}|\tilde{\lambda}_{jk})$
      $\boldsymbol{r}^{(s+1)} \leftarrow \boldsymbol{r}^{(s)} - \boldsymbol{x}_{jk}^T(\beta_{jk}^{(s+1)} - \beta_{jk}^{(s)})$
    end for
  end for
until convergence

This function checks whether the absolute value of an estimate is greater than λ. If it is, the corresponding $\boldsymbol{\beta}$-coefficient is updated by shrinking toward zero by the amount of λ, and otherwise set to zero. The same idea is in LLACD, but at each iteration and for each variable, a weighted λ is used (denoted by $\tilde{\lambda}_{jk}$ in Algorithm 1). These $\tilde{\lambda}_{jk}$ are obtained based on the derivative of the penalty function applied at the variable and group levels and denoted $\dot{P}^V$ and $\dot{P}^G$ in the algorithm. The derivative of a shrinkage function at β is the rate of its penalization, that is, the $\tilde{\lambda}_{jk}$ required to update the $\boldsymbol{\beta}$-coefficient with the soft-thresholding operator. Since the $\boldsymbol{\beta}$-coefficient of a current iteration is not known, it can be approximated by the value of the previous iteration. Technically, also the ℓ_2-norm of the $\boldsymbol{\beta}$-coefficients of a variable group must be derived before the update step of the $\boldsymbol{\beta}$-coefficients of the group, due to the $\beta_{jk}\|\boldsymbol{\beta}_j\|_2$- term in the calculation of $\tilde{\lambda}_{jk}$. Again, an approximation, that is, using values from the previous iteration, provides a solution.

All SGPs and their derivatives discussed in this paper are shown in Figure 1. Exemplary, to determine the $\tilde{\lambda}_{jk}$ to perform SGE regularization, the derivative of the EP and the Group EP must be used as $\dot{P}^V$ and $\dot{P}^G$, leading to Equation (5).

$$\tilde{\lambda}_{jk}^{SGE} = \alpha\,\lambda exp\left(-\frac{\tau}{\lambda}\left|\beta_{jk}^{(s)}\right|\right) + (1-\alpha)\begin{cases}\lambda & \text{if } \left\|\boldsymbol{\beta}_j^{(s)}\right\|_2 = 0,\\ \lambda exp\left(-\frac{\tau}{\lambda}\left\|\boldsymbol{\beta}_j^{(s)}\right\|_2\right)\frac{\beta_{jk}^{(s)}}{\left\|\boldsymbol{\beta}_j^{(s)}\right\|_2} & \text{if } \left\|\boldsymbol{\beta}_j^{(s)}\right\|_2 \neq 0\end{cases} \tag{5}$$

For all SGP, Algorithm 1 was used to cycle over all variables and sequentially update their coefficients until convergence was achieved for a given set of tuning parameters (λ, α, θ). We consider convergence to be achieved when the highest absolute difference of a $\boldsymbol{\beta}$-coefficient between successive iterations is smaller than $0.0001 * SD(y)$. The code of the algorithm is available on the GitHub page: https://github.com/GregorBuch/SGPR. We implemented the algorithm for the SGP framework using the R-package Rcpp (Eddelbuettel & François, 2011) to take advantage of the speed benefits of C++.

In a simulation study, the gradient descent–based algorithm implemented in the SGL package was compared with the coordinate descent–based algorithm proposed here for the special case of SGL. The results and information on the mechanism of data generation can be found in the Supporting Information (simulation study S1). Both algorithms gave very similar prediction and selection performance. Differences occurred at smaller α values, indicating limits to the transferability of values of this tuning parameter between the two algorithms. For example, SGL with gradient descent activates more than 80% of a group at $\alpha = \frac{1}{5}$, whereas SGL with coordinate descent does so only at $\alpha = \frac{1}{10}$. Independently, the coordinate descent–based algorithm was about twice as fast as the gradient-based alternative and achieved a lower mean squared error.

3.1 | λ-sequence

Usually, one is not interested in solving Equation (1) for only one λ, but for several, for example, in a cross-validation procedure to find an appropriate value for λ. To do so, a suitable λ-sequence must be defined. The sequence should start with a value for λ at which the shrinkage effect is high enough to just generate the null-model, and with a next smaller

value for λ the first variable enters the model. This λ_{max} can be derived by cycling over all predictors like in the initial iteration of the LLACD, but without updating the residuals. The highest absolute $\boldsymbol{\beta}$-coefficient estimated by this strategy is the origin of the targeted λ-sequence. Form this value, a sequence of 100 values is formed along a logarithmic scale, where the minimal λ-value is $\lambda_{max} * 0.0001$. After the LLACD has converged for a given λ, the estimated $\boldsymbol{\beta}$-coefficients serve as a "warm start" for the next smaller λ-value.

3.2 | Logistic regression

The algorithm proposed in the previous section is suitable for linear regression; additional calculations are required for other loss functions. In the context of generalized linear models, the loss function in Equation (1) can be described by the negative log-likelihood (McCullagh & Nelder, 1989). This allows an approximation of the loss using current estimates of the classical linear predictors. The fitting process then additionally involves an iteratively reweighted least squares algorithm so that working residuals $\tilde{\boldsymbol{r}}$ are first derived before updating the $\boldsymbol{\beta}$-coefficients.

In case of a logistic regression, $\tilde{\boldsymbol{r}}$ can be obtained by adding the following lines before the outer loop that iterates over groups in Algorithm 1:

$$\begin{aligned} \boldsymbol{\eta} &\leftarrow \boldsymbol{\beta}_0^{(s)} + \boldsymbol{X}^T \boldsymbol{\beta}^{(s)} \\ \tilde{\boldsymbol{r}}^{(s)} &\leftarrow \left(\boldsymbol{y} - \frac{e^{\boldsymbol{\eta}}}{1+e^{\boldsymbol{\eta}}} \right) / w \end{aligned} \tag{6}$$

where w can be fixed to 0.25, that is, the upper bound of the Hessian matrix (Krishnapuram et al., 2005). In the remainder of the algorithm, w must be taken into account in two further places. First, in the determination of $\tilde{\lambda}_{jk}$, and second, when updating $\beta_{jk}^{(s+1)}$. In both cases, the value of α must be integrated to reflect the chosen mixture of variable and group operators. Specifically, this means that the following lines must be included in the inner loop of Algorithm 1:

$$\begin{aligned} \tilde{\lambda}_{jk} &= \alpha \dot{P}^V \left(\beta_{jk}^{(s)} | \lambda \right) + (1-\alpha) \dot{P}^G \left(w^G \left\| \boldsymbol{\beta}_j^{(s)} \right\|_2 | \lambda \right), \\ \beta_{jk}^{(s+1)} &\leftarrow S \left(w^V z_{jk} | \tilde{\lambda}_{jk} \right) / w^V \end{aligned} \tag{7}$$

where $w^V = 1 - \alpha + \alpha w$ and $w^G = \alpha + w - \alpha w$. The full algorithm for logistic regression is provided in the Supporting Information (Algorithm S1).

The effects of most tuning parameters in regularization methods are not scale independent, so standardization of all features is required prior to the fitting process. However, this approach is not sufficient for γ and τ when switching from a continuous to a binary dependent variable. The recommended values for these parameters mentioned above are only appropriate for linear regression. One way to apply these established values to other response types is to rescale the parameters by w (Breheny & Huang, 2011). For logistic regression, this means that we set $\gamma = 16$ for SCAD, $\gamma = 12$ for MCP, and $\tau = 1/4$ for EP when they occur in P^V or P^G.

4 | SIMULATION STUDIES

In this section, artificial data sets were used to evaluate SGL, SGS, SGM, and SGE, to optimize their tuning parameter α, and to compare these techniques with more established ones. For all scenarios, 1000 data sets were generated, denoted by $\boldsymbol{X}$, where n is the number of observations and p is the number of variables, fixed at 200. The features of $\boldsymbol{X}$ belong to 20 nonoverlapping groups of 10 members each. Columns in $\boldsymbol{X}$ were drawn independently from a standard multivariate normal distribution, of which only a subset was used to generate the response variable $\boldsymbol{y}$ with a fixed signal-to-noise ratio of 1.

For all methods, λ was determined by a 10-fold cross-validation procedure. The simulations were performed in R (R Foundation for Statistical Computing, Vienna Austria; https://www.R-project.org, version 4.0.3).

TABLE 1 Recommendable values for α.

Method	Optimal α for group selection	Optimal α for predictions	Optimal α for variable selection
SGL	1/10	1/4	9/10
SGS	1/9	1/3	1/2
SGM	1/8	1/3	1/2
SGE	1/7	1/3	1/3

Note: Results from a numerical study with data sets consisting of 400 observations and 20 variable groups, each with a size of 10. The continuous dependent variable was generated using 20%, 40%, 60%, or 80% of the variables from 10 groups. Each method was fitted with different values for $\alpha \in \{\frac{1}{10}, \frac{1}{9}, \frac{1}{8} \dots \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \dots, \frac{9}{10}\}$. The value of α that leads to the best performance in 1000 iterations on average is given. Group and variable selection performance was evaluated with the Matthews correlation coefficient of the identifier indicating which group/variable was involved in generating the response and the identifier indicating which group/variable was selected by a method. Prediction performance was evaluated by the correlation between true and predicted values in a hold-out data set generated by the same data generation procedure. Methods: Sparse Group LASSO (SGL), Sparse Group SCAD (SGS), Sparse Group MCP (SGM), Sparse Group EP (SGE).
Abbreviations: LASSO, Least Absolute Shrinkage and Selection Operator; SCAD, Smoothly Clipped Absolute Deviation; MCP, Minimax Concave Penalty; SGE, Sparse Group Exponential Penalty.

We used the *Matthews correlation coefficient* (MCC) to evaluate whether an approach selected the groups and variables that were involved in generating y without including the groups and variables that were not:

$$MCC = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}. \tag{8}$$

The MCC aggregates the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) to a single correlation value $\in [-1, 1]$. An MCC of 1 indicates that a method selected only relevant groups/variables and no irrelevant ones, while a value of -1 means the opposite. Random selection leads to values close to 0. The reduction to one value simplifies the interpretation, but further established evaluation criteria (sensitivity and specificity) may be helpful and are provided in the Supporting Information.

4.1 | Balancing the influence of the grouping on the selection

First, the number of observations was set to 400 and the Gaussian response was determined by half of the variable groups. Four different scenarios were considered, each with different proportions of group members that are predictive for the dependent variable. The proportions of predictive signals within the groups were 20%, 40%, 60%, or 80%. To evaluate the effect of different values of α and to find an appropriate default, each SGP with different values for α was fitted to each data set. The values for α followed the sequence $\alpha \in \{\frac{1}{10}, \frac{1}{9}, \frac{1}{8} \dots \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \dots, \frac{9}{10}\}$.

The performance in variable and group selection and prediction performance (evaluated on a hold-out data set) of the bi-level selection methods for different values of α are provided in the Supporting Information (Table S2). Lower values of α led to best prediction, variable and group selection performance, when the group information was very informative (80% of the variables of a group were predictive). In the opposite case (20% of the variables of a group were predictive), high values of α gave the best results. The values of α leading to the highest value for an evaluation criterion are given in Table 1 for each method. Small values of α like 1/10 yielded to the best results for group selection, while higher values such as 1/2 were preferable for variable selection. However, for such values for α, the penalty emphasizes only the group level or the variable level (the group level in case of $\alpha = 1/10$ and the variable level in case of $\alpha = 1/2$). This leads to an unfavorable performance on the other level if this is not consistent with the data generation mechanism. The optimal prediction performance was obtained with α values of 1/3. Since an α value of 1/3 was always in between the optimal values for group and variable selection performance, this value seems to balance the group and variable level best and is therefore suitable as default. In the further analyses, α was therefore fixed at 1/3.

4.2 | Comparison of bi-level selection methods

To compare the approaches of the additive framework (SGL, SGS, SGM, and SGE) with those from the hierarchical framework (G-Bridge, cMCP, and GEL), the previous data generating process was further extended. The predictive signals were

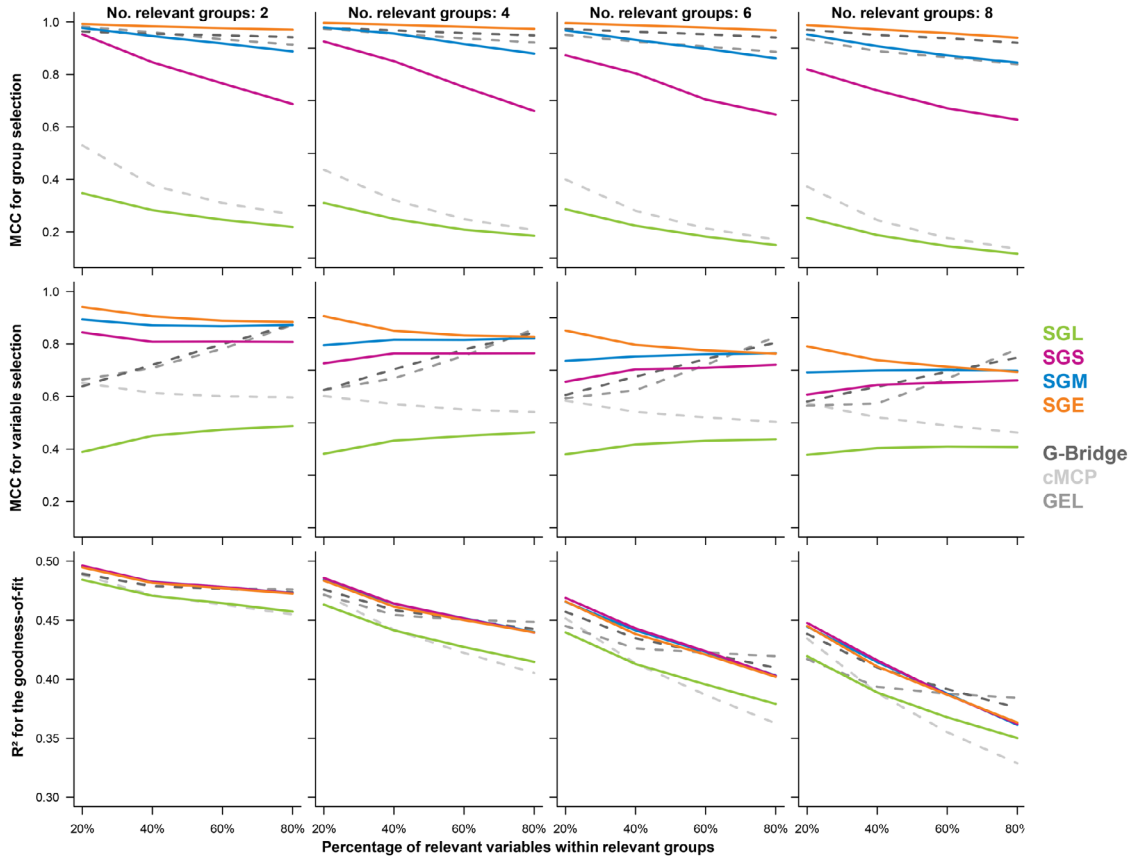

FIGURE 2 Results of the simulation study with N=400. Different panels for 2,4,6 or 8 of 20 variable groups related to a continuous dependent variable. In each panel, 20%, 40%, 60% or 80% of the members of the relevant groups have nonzero effect. Number of observations: 400, total number of variables: 200. The performance of the methods is described by the mean values derived from 1000 simulated replicates. A Matthews Correlation Coefficient (MCC) close to 1 implies that relevant groups / variables are selected and irrelevant groups / variables are excluded. Random selection results in values close to 0. Methods: Sparse Group LASSO (SGL), Sparse Group SCAD (SGS), Sparse Group MCP (SGM), Sparse Group EP (SGE), Group Exponential LASSO (GEL), composite Minimax Concave Penalty (cMCP), Group-Bridge (G-Bridge).

now in two, four, six, or eight variable groups. As before, different scenarios were simulated, in which 20%, 40%, 60%, or 80% of the members of a group were involved in generating the response. Tuning parameters of the hierarchical approaches were set to their default values as defined in the grpreg package (Breheny, 2015; Breheny & Huang, 2009).

The results are visualized in Figure 2, with separate panels for the number of groups with predictive members. Almost always, an SGP was superior to the hierarchical techniques, with the SGE being the best approach for variable and group selection. Its performance in group selection was almost flawless with a mean MCC > 0.94 and superior to the classical SGL by up to 0.82 points. In variable selection, SGE achieved the highest mean MCC in variable selection overall (0.94, SGL's best performance: 0.49), but was narrowly outperformed by G-Bridge and GEL in the settings where many variables were related to the response, like in the last situation, where about 1/3 of the features were relevant. The second best SGP in terms of variable and group selection was SGM, followed by SGS and SGL as the worst.

With SGL as an exemption, the SGPs achieved a high prediction performance. Similar performance was reached by G-Bridge, while GEL and cMCP results convinced only in certain settings. Additional comparison based on the Bias and MSE can be found in the Supporting Information (Table S3) and are in line with these findings. Overall, the difference in predictive performance between the methods was often marginal, but they achieved this with varying model sizes. The most parsimonious selection was produced by SGE (4.71 variables of 2.04 groups), and the largest model built by SGL (110.24 variables of 19.21 groups). SGM and SGE always selected fewer variables than their related methods from the hierarchical framework, namely, cMCP and GEL. The sparsity of SGM, SGE, and GEL at the group level showed only minor differences, whereas cMCP always activated too many groups. For group selection, all methods achieved perfect or near-perfect results in sensitivity, while SGE was superior to all other methods in specificity. This was particularly evident in the scenarios with two or four predictive variable groups, where all approaches achieved a sensitivity of one in group

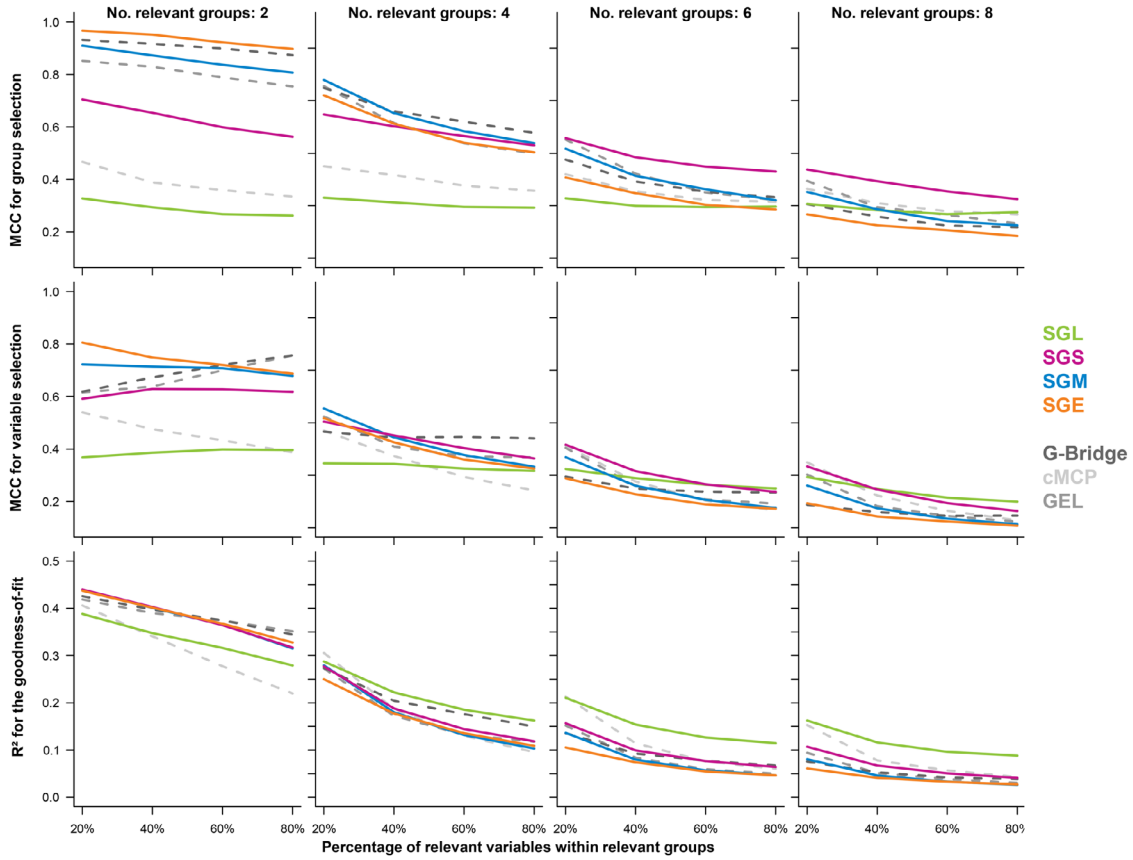

FIGURE 3 Results of the simulation study with N=100. Different panels for 2,4,6 or 8 of 20 variable groups related to a continuous dependent variable. In each panel, 20%, 40%, 60% or 80% of the members of the relevant groups have nonzero effect. Number of observations: 100, total number of variables: 200. The performance of the methods is described by the mean values derived from 1000 simulated replicates. A Matthews Correlation Coefficient (MCC) close to 1 implies that relevant groups / variables are selected and irrelevant groups / variables are excluded. Random selection results in values close to 0. Methods: Sparse Group LASSO (SGL), Sparse Group SCAD (SGS), Sparse Group MCP (SGM), Sparse Group EP (SGE), Group Exponential LASSO (GEL), composite Minimax Concave Penalty (cMCP), Group-Bridge (G-Bridge).

selection but differed in their specificity: The worst specificity was always achieved by SGL, the best always by SGE. In terms of variable selection, SGE was the most specific (>0.953, performance of SGL: >0.587) and GEL the most sensitive (>0.92, performance of SGL: >0.844). With the same performance in variable selection sensitivity as in the scenario where 20% of two groups of variables had to be selected, SGL achieved the lowest specificity (0.872) and SGE the highest (0.996).

Results of the same study, but with a binary dependent variable, are also provided in the Supporting Information (Table S4). The additional information supports the main finding, with a few differences. G-Bridge often did not converge in the case of a binary response (Table S5) and more situations occurred, in which the SGPs performed only comparably to the hierarchical approaches.

Often, there is interest in analyzing high-dimensional data sets, that is, with more predictors than observations. Therefore, the previous simulation was repeated with $n = 100$. The results are shown in Figure 3, and further information is provided in the Supporting Information (Table S6). Compared to the setting with 400 observations, no method was clearly superior to the others in terms of variable and group selection performance. Although an SGP was often the best approach, different techniques achieved the highest performance, depending on true parsimony. In the scenario where two groups contain the relevant predictors, the results are as in the setting with $n = 400$ and SGE performed best on all evaluation criteria. As the number of relevant variables increases, the performance of the methods decreases and starts to equalize, until the performance of SGE falls below that of SGL in the last scenario. In this scenario, SGE's MCC for group selection was 0.18 compared to SGL's 0.27, and SGE's MCC for variable selection was 0.11 compared to SGL's 0.2. SGS was even better at group selection in scenarios with six or more relevant groups, with stable performance between scenarios ranging from 0.32 to 0.56.

TABLE 2 Results of the real-world use case.

Method	No. selected variables	No. selected groups	Test set R^2
SGL	73	8	0.531
SGS	08	4	0.537
SGM	05	3	0.547
SGE	03	2	0.548
GEL	01	1	0.545
cMCP	06	4	0.568
G-Bridge	01	1	0.545

Note: Model size and performance of the bi-level selection approaches in a real-world use case. The data set was split into a training set and a test set ($N = 100$, 30), each with 272 variables belonging to 14 groups of different sizes. Number of (No.) selections in the training set and R^2 of the generated model predicting the response in the test set are reported. Methods: Sparse Group LASSO (SGL), Sparse Group SCAD (SGS), Sparse Group MCP (SGM), Sparse Group EP (SGE), Group Exponential LASSO (GEL), composite Minimax Concave Penalty (cMCP), Group-Bridge (G-Bridge).
Abbreviations: LASSO, Least Absolute Shrinkage and Selection Operator; SCAD, Smoothly Clipped Absolute Deviation; MCP, Minimax Concave Penalty; EP, Exponential Penalty.

SGL frequently achieved the highest sensitivity in variable (>0.36) and group (>0.78) selection, while SGE led in specificity in variable (>0.98) and group (>0.98) selection (Supporting Information, Table S6). Compared to the simulations with $n = 400$, a superior method regarding prediction performance could be identified: In situations with more than two relevant variable groups, SGL led to the lowest MSE and the highest R^2.

5 | ANALYZING THE LIPIDOME IN A RANDOMIZED CLINICAL TRIAL

To gain further insights into the methods, they were applied to the data of a randomized clinical, phase IV trial in patients with diabetes mellitus (EmDia study, ClinicalTrials.gov Identifier: NCT02932436). The objective was to identify lipids and lipid groups that regulated E/E', a continuous marker of the left ventricular diastolic function. Bi-level selection was used to highlight the regulated biochemical processes and their specific components to increase the interpretability of the selection.

A total of 133 patients were analyzed, for which 272 lipid variables were available, profiled by liquid chromatography mass spectrometry. Prior knowledge was used to classify the measured lipid variables according to chemical and biochemical principles. Based on this information, the 272 variables were assigned into 14 groups of different sizes. Here, some predictors formed a group by themselves, while the largest group included 73 members. An overview of all predictors and their group membership can be found in the Supporting Information (Table S7).

The data set was randomly split into a training and a test set ($n = 100$, $n = 33$). To evaluate the goodness of fit, the R^2 of the generated models predicting the response in the test data set was calculated. Results of this real use case and information about how many variables and variable groups were selected are shown in Table 2.

The highest R^2 was achieved by cMCP, followed by SGE with a model half the size of that of cMCP. The worst performance in terms of R^2 was obtained by SGL, with more variables and groups included in the model than by any other approach. The different models are subsets of the large model produced by SGL, with the smallest model produced by GEL and G-Bridge (Figure 4).

6 | DISCUSSION

In this work, we have focused on penalty-based techniques that select feature groups associated with a response and simultaneously identify their most relevant members. Such a task can be addressed with bi-level selection methods, as they incorporate a predefined group structure of predictors in their model building process (Huang et al., 2009). This is accomplished by combining different penalties operating at the variable and group levels, either hierarchically or additively (Huang et al., 2012). The latter leads to the proposed SGPs, which comprise SGL, SGS, SGM, as well as SGE that was introduced here. SGE combines an exponential term additively at the variable and group levels. This combination was best in simulation studies in variable and group selection when more observations than predictors were available. However,

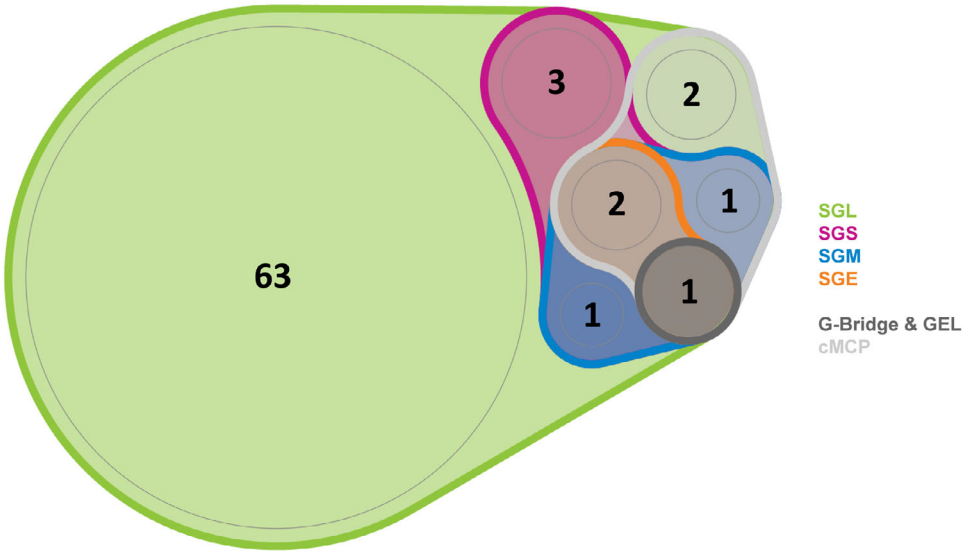

FIGURE 4 Overlap of the generated models in the real-world use case. The models generated by the different methods are represented by colored areas, where the size of an area indicates the model size and overlaps indicate that the same variables were selected by different methods. Methods: Sparse Group LASSO (SGL), Sparse Group SCAD (SGS), Sparse Group MCP (SGM), Sparse Group EP (SGE), Group Exponential LASSO (GEL), composite Minimax Concave Penalty (cMCP), Group-Bridge (G-Bridge).

in situations where the features outnumbered the observations, SGE was outperformed by other techniques when many predictors had an association with the response. Techniques like SGS performed superior in such cases.

The results show that SGPs can be superior to techniques of the hierarchical framework. Their strength lies in their better interpretability (i.e., they select more accurately at the group and variable levels) than in their ability to generate better predictive models than established techniques from the hierarchical framework. However, SGPs come with a relatively large number of tuning parameters, which can be considered a disadvantage. Special attention should be paid to the tuning parameter α, which allows control over the influence of the group information on the selection process. A small value for α, such as $1/10$, seems recommended when there is high confidence in group information and greater interest in the group level. If prior knowledge of the grouping is available but confidence in its relevance or accuracy is limited, as well as when the variable level is more important than the group level, larger values of α, like $1/2$, are appropriate. This reduces the impact of the group signal on the selection process, so that predominantly variables that are predictive on their own enter the model. In cases where there is at least a moderate confidence in the accuracy and relevance of the grouping, or a balanced interest in both the group- and variable levels, an α of $1/3$ is advisable. This value seems recommendable for most applications. For sensitivity analyses, it may be reasonable to set α to a higher value to limit the selection to those variables that have a group-independent relationship with the dependent variable. However, it should be kept in mind that these suggestions are based on limited simulation studies. They may not be optimal for settings that strongly differ from the data generation mechanism used here.

The performance of SGE seems to stem in particular from the fact that the method is parsimonious at both the group and variable levels. Established methods tend to be parsimonious on only one of the two levels, making SGE a valuable addition to the methodological landscape. However, this pronounced parsimony can also be a disadvantage if a large proportion of the features or groups are relevant. Therefore, further research is needed to complete the comparison between SGPs and hierarchical approaches. Additional investigations could also focus on the effects of the tuning parameters. By adjusting γ and τ, results like those of SGL might be achieved with SGM, SGS, and SGE, which may be beneficial in situations where a high sensitivity is desired (Belhechmi et al., 2020; Boulesteix et al., 2017). Alternatively, different types of penalties could be combined at the variable and group levels (Huang et al., 2012). So far, the shrinkage operators were set equally for the variable and group levels. This is not strictly necessary, and in the case of hierarchical techniques, the combination of unequal penalties has been shown to be advantageous (Breheny, 2015; Huang et al., 2009). The results of our research suggest that different penalties are recommendable depending on the dimensionality of the data set or whether interest is in an optimal sensitivity or specificity. Therefore, a mixture of two operators might be useful to get closer to an all-around method.

The algorithm proposed here to solve the objective function of SGPs is based on approximations and cannot guarantee to find the global optimum. It would be worthwhile to research more accurate algorithms, such as those already proposed for SGL (Simon et al., 2013). To further facilitate research on this topic, the code for the proposed approach is available on GitHub (https://github.com/GregorBuch/SGPR). We plan to integrate the code in an R package in a further step and extend the methods to also apply to time-to-event analysis.

In summary, this work has highlighted the strengths of SGPs and of SGE in particular. The approaches presented seem particularly promising for exploratory analysis of grouped omics data like transcriptome and proteome data, where biological grouping information is available in advance.

ACKNOWLEDGMENTS

This work is part of the dissertation of Gregor Buch and was supported by the DIASyM project, funded by the Federal Ministry of Education and Research (BMBF, Grant No. 031L0217A) and the Center of Preventive Cardiology and Preventive Medicine of the University Medical Centre of the Johannes Gutenberg-University Mainz. Philipp. S. Wild is principal investigator of the German Center for Cardiovascular Research (DZHK), principal investigator of the DIASyM research core (BMBF DIASyM research core [BMBF 031L0217A]), principal investigator of the Institute of Molecular Biology, and was funded by the Federal Ministry of Education and Research (BMBF 01EO1003 and 01EO1503). We are indebted to the study participants and all coworkers of the EmDia Study for their support and commitment.

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

All authors declare to have nothing to disclose that could be perceived as conflict of interest in the context of the present work.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in the Supporting Information of this paper.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to computational complexity and data confidentiality issues.

ORCID

Gregor Buch https://orcid.org/0000-0002-9963-1245

REFERENCES

Belhechmi, S., De Bin, R., Rotolo, F., & Michiels, S. (2020). Accounting for grouped predictor variables or pathways in high-dimensional penalized Cox regression models. *BMC Bioinformatics*, *21*, 1–20. https://doi.org/10.1186/s12859-020-03618-y

Boulesteix, A.-L., De Bin, R., Jiang, X., & Fuchs, M. (2017). IPF-LASSO: Integrative-penalized regression with penalty factors for prediction based on multi-omics data. *Computational Mathematical Methods in Medicine*, *2017*, 2017. https://doi.org/10.1155/2017/7691937

Breheny, P. (2015). The group exponential LASSO for bi-level variable selection. *Biometrics*, *71*, 731–740. https://doi.org/10.1111/biom.12300

Breheny, P., & Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and Its Interface*, *2*, 369. https://doi.org/10.4310/SII.2009.v2.n3.a10

Breheny, P., & Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, *5*, 232. https://doi.org/10.1214/10-AOAS388

Breheny, P., & Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, *25*, 173–187. https://doi.org/10.1007/s11222-013-9424-2

Buch, G., Schulz, A., Schmidtmann, I., Strauch, K., & Wild, P. S. (2023). A systematic review and evaluation of statistical methods for group variable selection. *Statistics in Medicine*, *42*, 331–352. https://doi.org/10.1002/sim.9620

Bui, K., Park, F., Zhang, S., Qi, Y., & Xin, J. (2021). Structured sparsity of convolutional neural networks via nonconvex sparse group regularization. *Frontiers in Applied Mathematics Statistics*, *6*, 529564 https://doi.org/10.3389/fams.2020.529564

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*, 1–18. https://doi.org/10.18637/jss.v040.i08

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360. https://doi.org/10.1198/016214501753382273

Frank, L. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, *35*, 109–135. https://doi.org/10.1080/00401706.1993.10485033

Huang, J., Breheny, P., & Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *27*, 481–499. https://doi.org/10.1214/12-STS392

Huang, J., Ma, S., Xie, H., & Zhang, C.-H. (2009). A group bridge approach for variable selection. *Biometrika*, *96*, 339–355. https://doi.org/10.1093/biomet/asp020

Krishnapuram, B., Carin, L., Figueiredo, M. A., & Hartemink, A. (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis Machine Intelligence*, *27*, 957–968. https://doi.org/10.1109/TPAMI.2005.127

Liu, J., Huang, J., & Ma, S. (2013). Integrative analysis of multiple cancer genomic datasets under the heterogeneity model. *Statistics in Medicine*, *32*, 3509–3521. https://doi.org/10.1002/sim.5780

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*, Chapman and Hall.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2013). A sparse-group LASSO. *Journal of Computational Graphical Statistics*, *22*, 231–245. https://doi.org/10.1080/10618600.2012.681250

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

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, *68*, 49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*, 894–942. https://doi.org/10.1214/09-AOS729

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, *67*, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Zou, H., & Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, *36*, 1509.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Buch, G., Schulz, A., Schmidtmann, I., Strauch, K., & Wild, P. S. (2024). Sparse Group Penalties for bi-level variable selection. *Biometrical Journal*, *66*, 2200334. https://doi.org/10.1002/bimj.202200334