Received: 8 September 2021 | Revised: 27 October 2022 | Accepted: 22 November 2022

DOI: 10.1002/sim.9620

RESEARCH ARTICLE

Statistics in Medicine WILEY

A systematic review and evaluation of statistical methods for group variable selection

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

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

[2] German Center for Cardiovascular Research (DZHK), partner site Rhine-Main, Mainz, Germany

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

Andreas Schulz, Preventive Cardiology and Preventive Medicine, Department of Cardiology, University Medical Center of the Johannes Gutenberg University Mainz, Langenbeckstr. 1, 55131 Mainz, Germany.

Email: andreas.schulz@unimedizin-mainz.de

Funding information

Bundesministerium für Bildung und Forschung, Grant/Award Numbers: 031L0217A, 01EO1003, 01EO1503; Preventive Cardiology and Preventive Medicine, Department of Cardiology, University Medical Center of the Johannes Gutenberg University Mainz

This review condenses the knowledge on variable selection methods implemented in R and appropriate for datasets with grouped features. The focus is on regularized regressions identified through a systematic review of the literature, following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines. A total of 14 methods are discussed, most of which use penalty terms to perform group variable selection. Depending on how the methods account for the group structure, they can be classified into knowledge and data-driven approaches. The first encompass group-level and bi-level selection methods, while two-step approaches and collinearity-tolerant methods constitute the second category. The identified methods are briefly explained and their performance compared in a simulation study. This comparison demonstrated that group-level selection methods, such as the *group minimax concave penalty*, are superior to other methods in selecting relevant variable groups but are inferior in identifying important individual variables in scenarios where not all variables in the groups are predictive. This can be better achieved by bi-level selection methods such as *group bridge*. Two-step and collinearity-tolerant approaches such as *elastic net* and *ordered homogeneity pursuit least absolute shrinkage and selection operator* are inferior to knowledge-driven methods but provide results without requiring prior knowledge. Possible applications in proteomics are considered, leading to suggestions on which method to use depending on existing prior knowledge and research question.

KEYWORDS

group variable selection, proteomics, simulation study, systematic review

1 | INTRODUCTION

Feature selection methods are an appropriate choice for identifying variables that are associated with a particular response. Their goal is to generate a model with the available variables that minimizes the error in predicting the dependent variable with as few variables as possible. Accordingly, only variables related to the dependent variable are included in the constructed model so that the selected variables can be considered relevant predictors.

Numerous methods for variable selection are now available, ranging from highly specialized to broadly applicable approaches. While the latter perform well in many situations, it might be worthwhile to consider more specialized methods in certain situations.[1,2] One such situation is the analysis of variables that are interrelated based on correlations or contextual similarities, since ignoring such group structure reduces the stability, consistency, and interpretability of the selection.[3] Methods accounting for group structures have been investigated since at least 1999[4] and have been used in a broad range of applications, including media classification,[5] disease prediction,[6] automotive engineering,[7] voting behavior analysis,[8] emotion recognition,[9] and credit risk analysis.[10] One of the most common applications is in omics research, such as gene expression microarray or single nucleotide polymorphism data.[11-15] Here, the interest is often in identifying genes that share a biological function or are involved in the same pathway related to a particular response. Since the interrelation of genes implies a natural group structure, the selection of variable groups is appealing. The same is valid for the analysis of other omics data, like proteomics data, where external sources, such as protein-protein interaction or pathway databases, are often used to enrich the analysis.[16] Such additional knowledge can be incorporated in various forms and steps into the research process,[17] notably in the selection process. This enables the identification of relevant variable groups, that is, variables that are interrelated and share a collective and traceable relationship with the response variable.

This review aimed to identify and evaluate group variable selection methods that are discussed in the scientific literature, sufficiently programmed and appropriate to select groups of important features in research on continuous outcome types. Our interest was focused on methods suitable for settings where the number of features is less than, equal to, and greater than the number of observations and that provide at least information about the direction of the relationship between selected and response variables. The review was restricted to non-Bayesian methods programmed in R that accounts for a natural group structure like those arising from biological processes. Since the majority of identified approaches are regularized regression techniques, the emphasis of this article is on such approaches.

This article updates previous reviews with novel and lesser-known approaches, so that the assessment of methods is more comprehensive. The following section describes the process for searching and selecting appropriate articles before addressing the strategies for assigning the identified methods to an implementation and evaluating them. Then, results of the literature research, an overview of the identified methods, and a summary of the results of a simulation study comparing the methods are presented. The final section discusses potential applications of the methods in proteomics research.

2 | METHODS

Methodology and reporting of this review follow the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines[18] as appropriate for a review of statistical methods.

2.1 | Search strategy

For the systematic review, the query "[GROUP* (VARIABLE OR FEATURE OR PROTE*) SELECTION] OR [(VARIABLE OR FEATURE OR PROTE*) GROUP SELECTION]" was used to search eight electronic databases: IEEE Xplore, JSTOR, MathSciNet, Project Euclid, Pubmed, Science Direct, Web of Science, and zbMATH, on February 19, 2020. An identical search was conducted on February 19, 2021 to update the results of the literature search. Query extensions with "group sparsity," "group selection," "group regularization" and combinations thereof were evaluated in a pretest but considered to be either too specific or too general.

2.2 | Selection strategy

To be included in the review, articles had to be published in English between the inception of the corresponding database and February 19, 2021 and had to meet the eligibility criteria of a screening and full-text phase. In the screening phase, the title, keywords, and abstract were screened to determine whether they contained “feature selection,” “variable selection,” “protein selection” or similar terms in content. For this review, articles dealing exclusively with Bayesian methods or with outcome types other than continuous, binomial, or time-to-event were excluded. Articles that passed the screening phase were assessed based on full text against a further eligibility criterion: articles had to propose a new method for selecting variable groups or to report on existing group variable selection methods. The latter included reviews, comparisons, and statements on group variable selection. For this purpose, a group variable selection method was defined as an approach that accounts for a group structure in the selection process by using the correlation structure or predefined group information. The latter were limited to non-overlapping group membership.

2.3 | Software identification strategy

All methods discussed in the articles under consideration that performed group variable selection were extracted. For all these methods, the following process was performed: If no reference to a code was given in the article, a suitable package or code was successively searched on the repositories “CRAN” (https://www.cran.r-project.org), “Bioconductor” (https://www.bioconductor.org), or “GitHub” (https://www.github.com). This was done with the internal search function on the respective websites and, if not successful, with the search engine “Google” (https://www.google.com). The name of the methods and their abbreviation together with the letter “R” were used as a search query. The package or code should be freely available and executable and the output should at least provide information about the direction of the relationship of the selected variable(s) and the dependent variable. If multiple implementations could be identified, the most frequently cited package was included in the review. Procedures that are combinations of several methods were only included in the review if a program could be identified that covered the entire sequence of methods.

2.4 | Evaluation strategy

All methods identified by the identification strategy were evaluated with a simulation study organized according to the ADEMP structure.[19] The simulations were performed in R (R Foundation for Statistical Computing, Vienna, Austria; https://www.R-project.org, version 4.0.3) with the packages given in Table 1 on a server with AMD EPYC 7542 32-core processor @2.9 GHz with an Ubuntu 22.04.1 LTS operating system and 126 GB RAM.

Aims: The aim of the simulation study was to evaluate the performance of the identified methods in selecting variables and groups of variables in different scenarios defined by varying numbers of predictive features within a fixed group structure. The simulation study was designed to highlight the differences between approaches in a setting where all methods are viable.

Data-generating mechanisms: We consider three different data-generating mechanisms. For all, generated datasets consist of $p = 500$ variables and $n = 500$ observations. The variables belonged to 10 non-overlapping groups that differ in size: two groups contain 2%, 6%, 10%, 14%, and 18% of all variables, respectively. A standard multivariate normal distribution with a block-wise correlation structure was used to create the predictors. The correlation blocks were randomly determined for each dataset, with $\text{corr}\,(\boldsymbol{x}_l, \boldsymbol{x}_m) \in \{0.95, 0.9, 0.85, \dots, 0.5\}$ for $l \neq m$ if $\boldsymbol{x}_l$ and $\boldsymbol{x}_m$ belong to the same variable group and $\text{corr}\,(\boldsymbol{x}_l, \boldsymbol{x}_m) \in \{0.4, 0.35, 0.3, \dots, 0\}$ for $l \neq m$ if $\boldsymbol{x}_l$ and $\boldsymbol{x}_m$ belong to different groups, with l and m denoting columns of the dataset $\boldsymbol{X}$.

In each of the 1000 simulated datasets, five groups were randomly chosen to contain variables related to a continuous dependent variable. In the first scenario 100%, in the second scenario 50%, and in the last scenario 10% of all variables of the five groups were assigned a nonzero association. Magnitude of nonzero effects was combinations of random group-specific and random variable-specific effects, with a fixed signal-to-noise ratio of 1. That is, $\boldsymbol{\beta}^{\mathrm{T}}\text{Var}(\boldsymbol{X})\boldsymbol{\beta}/\sigma^2 = 1$ with elements of $\boldsymbol{\beta}$ being generated using $J * k$, where $J \in \{1, 2, \dots, 10\}$ denotes the group index and $k \in \left\{1, 2, \dots, K_j\right\}$ with K_j denoting the size of group j. The effect direction was determined randomly.

Target: Target is the identifier indicating which variables and variable groups were selected.

Methods: All methods with suitable implementation were applied to the simulated data. The methods were provided with correct group information that they could use in the selection process or afterwards to identify which groups were selected. Once at least one variable in a group was estimated to have a nonzero effect, the group was considered to be selected by the method. Tuning parameters of the methods were set to default values or determined with 10-fold cross-validation (CV), as stated in Table 1.

TABLE 1 Overview of methods included in the systematic review

	Method		R package			Article	
	Name (abbreviation)	Selection property	Name and purpose of selected control parameter (REC[a])	Name (version)	Supported response variables[b]	Main source for method	N. cited[c]
Group-level selection methods	Group least absolute shrinkage and selection operator (G-LASSO)	Selection in "all-in-all-out" fashion: Applies LASSO penalty to L_2-norm of groups	Lambda: Rate of penalization (CV[d])	grpreg (3.3.0)	C, B, T	Yuan and Lin[29]	6931
	Group least-angle regression (G-LARS)	Algorithm based selection in "all-in-all-out" fashion	sMax: Number of sequenced groups ($n/(2\overline{K})$)	robustHD (0.6.1)	C, B	Yuan and Lin[29]	6931
	Group smoothly clipped absolute deviation (G-SCAD)	Selection in "all-in-all-out" fashion with oracle property: Applies SCAD penalty to L_2-norm of groups	Lambda: Rate of penalization (CV) Gamma: Relaxation of penalization (4)	grpreg (3.3.0)	C, B, T	Breheny and Huang[32]	190
	Group minimax concave penalty (G-MCP)	Selection in "all-in-all-out" fashion with oracle property: Applies MCP to L_2-norm of groups	Lambda: Rate of penalization (CV) Gamma: Relaxation of penalization (3)	grpreg (3.3.0)	C, B, T	Breheny and Huang[32]	190
Bi-level selection methods	Group bridge (G-bridge)	Selection in hierarchical "bi-level" fashion: Applies bridge as outer and LASSO as inner penalty	Lambda: Rate of penalization (CV) Gamma: Exponent applied to LASSO regularized group-coefficients (0.5)	grpreg (3.3.0)	C, B	Huang et al[13]	318
	Sparse-group least absolute shrinkage and selection operator (SGL)	Selection in additive "bi-level" fashion: Additive combination of LASSO and G-LASSO penalty	Lambda: Rate of penalization (CV) Alpha: Balance between LASSO and G-LASSO (0.95)	SGL (1.3)	C, B, T	Simon et al[12]	1048
	Composite minimax concave penalty (cMCP)	Selection in hierarchical "bi-level" fashion with oracle property: Applies MCP as outer and inner penalty	Lambda: Rate of penalization (CV) Gamma: Relaxation of penalization (3)	grpreg (3.3.0)	C, B, T	Breheny and Huang[15]	224
	Group exponential least absolute shrinkage and selection operator (GEL)	Selection in hierarchical "bi-level" fashion with control over coupling effect: Applies EP[e] as outer and LASSO as inner penalty	Lambda: Rate of penalization (CV) Tau: Strength of coupling (1/3)	grpreg (3.3.0)	C, B, T	Breheny[11]	65
	Bi-level stagewise estimating equation (BiSEE)	Forward stagewise algorithm utilizing SGL-penalty to select in "bi-level" fashion	Epsilon: Step size (0.05) Alpha: Balance between LASSO and G-LASSO (0.5)	sgee (0.6-0)	C	Vaughan et al[45]	7
	Hierarchical stagewise estimating equation (HiSEE)	Forward stagewise algorithm utilizing penalty terms of G-LASSO and LASSO to select in hierarchically "bi-level" fashion	Epsilon: Step size (0.05)	sgee (0.6-0)	C	Vaughan et al[45]	7

(Continues)

TABLE 1 Continued

	Method		R package			Article	
	Name (abbreviation)	Selection property	Name and purpose of selected control parameter (REC[a])	Name (version)	Supported response variables[b]	Main source for method	N. cited[c]
Two-step approaches	Multi-layer group-least absolute shrinkage and selection operator (MLGL)	Hierarchical clustering to generate group formation, followed by an FWER[f] controlled selection of variable groups with G-LASSO	Lambda: Rate of penalization (CV) Alpha: Control level for testing procedure (0.05)	MLGL (0.6.1)	C	Grimonprez et al[49]	1
	Ordered homogeneity pursuit least absolute shrinkage and selection operator (OHPL)	PLS[g] and Fisher optimal partitions to generate group formation, followed by a selection of group representatives with LASSO	MaxComp: Max. PLS components G: Max. variable groups (CV) Gamma: Rate of penalization (CV)	OHPL (1.4)	C	Lin et al[51]	13
Collinearity tolerant methods	Elastic Net (E-Net)	Selection promoting similarity of highly correlated variables: Combination of LASSO and Ridge regression	Lambda: Rate of penalization (CV) Alpha: Balance between LASSO and ridge penalty (0.5)	glmnet (4.0-2)	C, B, T	Zou and Hastie[55]	13 768
	Octagonal selection and clustering algorithm in regression (OSCAR)	Selection promoting equality of highly correlated variables: Combination of LASSO and pair-wise L_∞-norm	Lambda: Rate of penalization (CV) Q: Degree of grouping (0.1)	SLOPE (0.3.0)	C, B	Bondell and Reich[58]	502

[a] Recommended value or process to determine parameter for continuous response variables. Based on information within cited article or corresponding package.

[b] C = continuous, B = binomial, T = time to event.

[c] Number of citations according to Google Scholar (May 21, 2021).

[d] CV = cross-validation.

[e] EP = exponential penalty.

[f] FWER = family-wise error rate.

[g] PLS = partial least squares regression.

Performance measures: For the comparison of the methods, the Matthews correlation coefficient (MCC) of both variable and group variable selection was computed, which can be interpreted as the correlation between the generated and true indication of which variable and variable group were involved in generating the response. A MCC value of one exhibits that the created model consists only of the groups and variables that defined the dependent variable when the data were generated. Random selection would result in a value of zero, while a value of minus one indicates that the model consists solely of groups and variables that played no role in generating the response. The MCC is defined according to (1).

$$\mathrm{MCC} = \frac{\mathrm{TP} * \mathrm{TN} - \mathrm{FP} * \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}. \tag{1}$$

With TP for true positive, TN for true negative, FP for false positive, and FN for false negative.

For methods that do not include group information in the selection process but group variables themselves, MCC was also used to evaluate group formation, according to the information provided in the Supplemental Appendix.

Other performance measures like sensitivity, specificity, mean squared error, and Rand index were calculated for supporting comparisons. Sensitivity analyses encompassed different signal-to-noise ratios (0.5 and 3), numbers of relevant variable groups (3 and 7), numbers of variables in the dataset (250 and 1000) and scenarios with variable groups of the same size and correlation structure (more details in the Supplemental Appendix).

3 | RESULTS

The structured literature research revealed 501 articles in the first round (until February 19, 2020) and 56 further articles in the second round (until February 19, 2021), which were processed as presented in the flowchart given in Figure 1. Among the distinct articles, about half were assessed for eligibility using the full text, which reduced the number of articles to 82, including three PhD theses. In about half of these articles, a new method was proposed, while most of the others investigated, compared, or further developed existing approaches. Only one out of the six identified review articles focused completely on group variable selection,[20] while the remaining articles addressed the topic only partially.[21-25] No method was included in the review based solely on the term "PROTE*" in the search query.

Since methods were only included in the review if a sufficient implementation could be identified, fewer methods were reviewed than were invented. As a result, 14 programmed methods for group variable selection were identified, implemented in eight different packages, all available on CRAN. The identified methods are summarized in Table 1 with a brief description of their selection properties, dedicated R packages, and primary reference.

3.1 | Categories of group variable selection

A major difference between the identified methods lies in their conception of which characteristic defines a group membership. The majority of the methods rely on prespecified group formations, which is particularly appropriate when variables belong together because of their meaning, such as proteins that are involved in the same pathway or single nucleotide polymorphisms in the same haplotype block. The alternative approach defined variable groups in a data-dependent manner by exploiting the correlation structure, assuming that high correlations imply similar information.[26] Methods belonging to the first concept can be referred to as knowledge-driven, while the remaining methods can be characterized as data-driven approaches.[23]

Within the framework of knowledge-driven approaches, a subdivision into two subcategories is possible, namely group-level and bi-level selection.[20] The first comprises four of the identified methods, which select in an "all-in or all-out" manner: When a group of variables is selected, all variables within that group will be considered relevant. This results in a sparse group selection, but does not provide sparsity within a group.[20] When it is desired to select important groups as well as important predictors within those groups, a bi-level selection is appropriate, which constitutes the second category. The six identified methods of this category yield sparse solutions at the group and within-group level.[13,15]

Data-driven methods for group variable selection can also be further subdivided. A distinction can be made between collinearity tolerant methods[27] and two-step approaches. For each type, two methods were identified. A two-step strategy uses a data-driven group variable formation technique, like cluster analysis, to generate variable groups in the first step.[28] Then, these constructed variable groups are used in the selection process to identify predictive variable groups in an "all-in or all-out" fashion.[23] In contrast, a collinearity tolerant method performs grouping and selection in one step by assigning comparable coefficients to highly correlated variables so that they can be considered as a group.

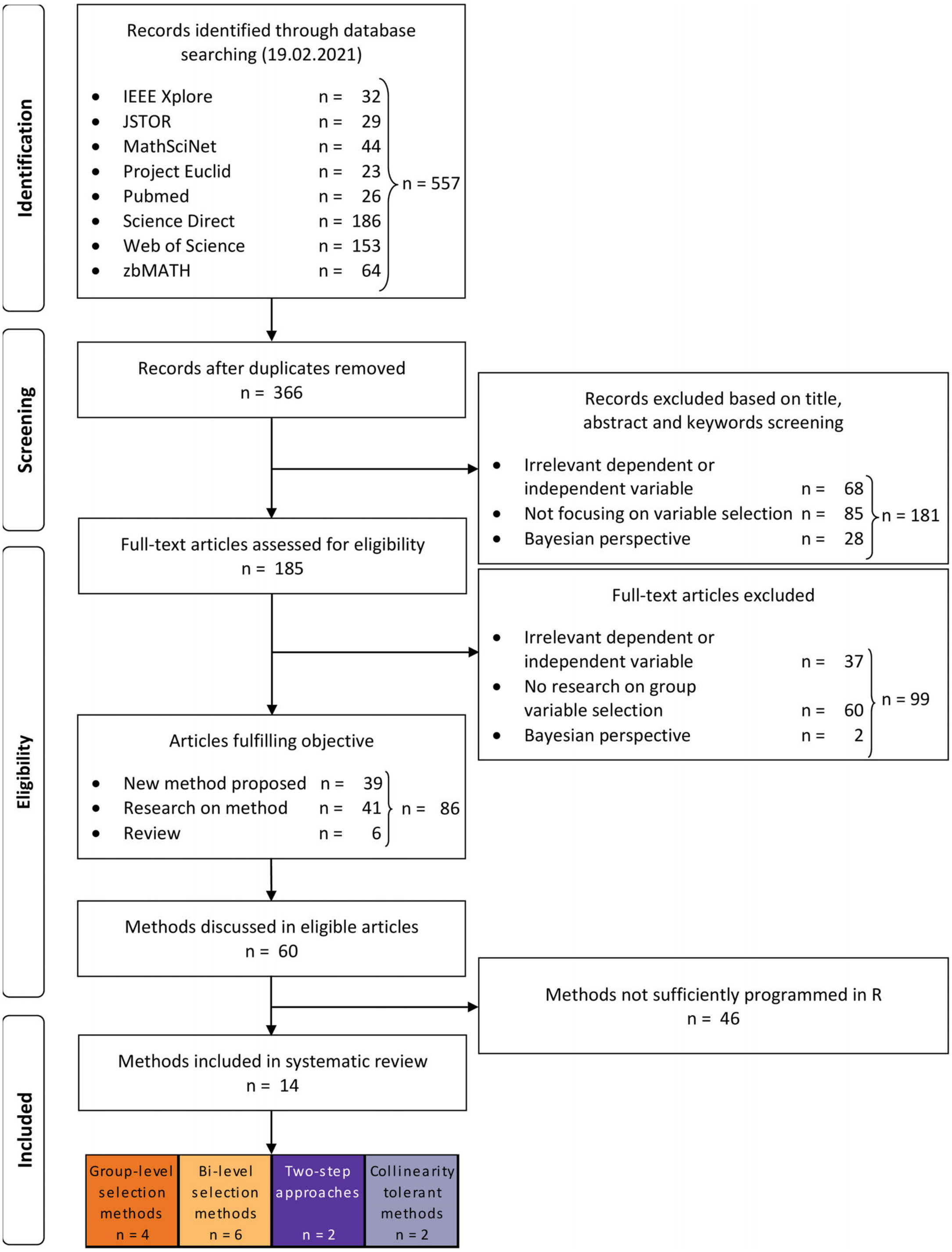

Identification
Screening
Eligibility
Included
Records identified through database searching (19.02.2021)
IEEE Xplore n = 32
JSTOR n = 29
MathSciNet n = 44
Project Euclid n = 23
Pubmed n = 26
Science Direct n = 186
Web of Science n = 153
zbMATH n = 64
n = 557
Records after duplicates removed n = 366
Records excluded based on title, abstract and keywords screening
Irrelevant dependent or independent variable n = 68
Not focusing on variable selection n = 85
Bayesian perspective n = 28
n = 181
Full-text articles assessed for eligibility n = 185
Full-text articles excluded
Irrelevant dependent or independent variable n = 37
No research on group variable selection n = 60
Bayesian perspective n = 2
n = 99
Articles fulfilling objective
New method proposed n = 39
Research on method n = 41
Review n = 6
n = 86
Methods discussed in eligible articles n = 60
Methods not sufficiently programmed in R n = 46
Methods included in systematic review n = 14
Group-level selection methods n = 4
Bi-level selection methods n = 6
Two-step approaches n = 2
Collinearity tolerant methods n = 2

FIGURE 1 Flowchart of systematic literature review

3.2 | Methods for group variable selection

Almost all methods identified by the literature research perform group variable selection by using a penalty term. Thereby, the methods minimize the objective function:

$$Q(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{y}) = L(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{y}) + P(\boldsymbol{\beta},\lambda), \tag{2}$$

which is a combination of a loss function, L, and a penalty term, P. While the loss function does not change across the methods, the penalty term differs. A loss function of the form $L(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{y}) = \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$ was considered in the following, which corresponds to linear regression. Since the data consist of different variable groups, $\boldsymbol{X}$ is a $n \times p$ matrix that composes J variable groups $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_j$ with K_j denoting the size of group j, so that the total number of variables $p = \sum_1^J K_j$.

Although the penalty terms differ between the methods, they all have in common that they introduce a tuning parameter λ, which is a weight multiplied by the penalty term. A λ of zero removes the penalization and reduces Equation (2) to that of classical regression, whereas a $\lambda > 0$ elevates the second part of Equation (2) for any nonzero coefficient in the model. Hence, minimizing the objective function (2) promotes that the $\boldsymbol{\beta}$-coefficients of variables without correlation to the dependent variable are shrunken towards zero. If the penalty term is not differentiable at the origin, the shrinkage process can set the $\boldsymbol{\beta}$-coefficients exactly to zero which corresponds to excluding them from the model. Commonly, λ is determined based on a CV procedure to optimize regularization, but information criteria like the AIC or BIC are also frequently used.

The penalty terms of the group variable selection models differ in their formalization and incorporate further parameters. In Figure 2, each penalty function occurring in the identified methods is visualized with different perspectives in three columns. The 3D visualization in the first column is common to illustrate penalties of group variable selection methods[11,20] and demonstrates the magnitude of penalization for a grid of two different coefficients belonging to two variables of the same group with fixed tuning parameters. The magnitude of penalization is zero if the two variables are removed from the model, and increases as soon as one or both coefficients are nonzero. Since the penalization rates in the 3D plots have been scaled so that the graphs are the same size, additional perspectives are useful to highlight different aspects of the penalty terms: In the second column, the contour plot of the same function represents the same rate of penalization for different pairs of coefficient values with contour lines. If a contour line has edges at the origin of a dimension, it means that the penalty here is non-differentiable, a property that encourages sparse solutions. Contrary, circular contours indicate a differentiable penalty, which discourage sparse solutions. The next column illustrates how the rate of penalization of a variable changes as its coefficient value varies. Note that the two-step methods are only indirectly represented in Figure 2, as they apply the penalty terms of other methods in their selection process.

3.2.1 | Group-level selection methods

Altogether, four group-level selection methods were identified that in- or exclude all variables of a prespecified variable group in their selection. This behavior is particularly desirable for selecting categorical variables, where the selection of either all levels of the categorical variable or none would be preferable to selecting only individual levels. However, these methods can also be valuable for non-categorical variables, if the "all-in or all-out" fashion fits the objective.

G-LASSO

The oldest group variable selection method identified within the scope of this review is group LASSO (G-LASSO),[4] developed by Yuan and Lin.[29] The idea behind the method is to penalize the square root of the sum of the squared coefficient values of a group (L_2-norm), rather than penalizing the coefficients of individual covariates. That is, the variables within a group are regularized similarly to ridge regression,[30] while the collective signal of a group is regularized as with LASSO.

$$P^{\text{G-LASSO}}(\boldsymbol{\beta}|\lambda) = \lambda \sum_{j=1}^{J} \left\|\boldsymbol{\beta}_j\right\|_2 . \tag{3}$$

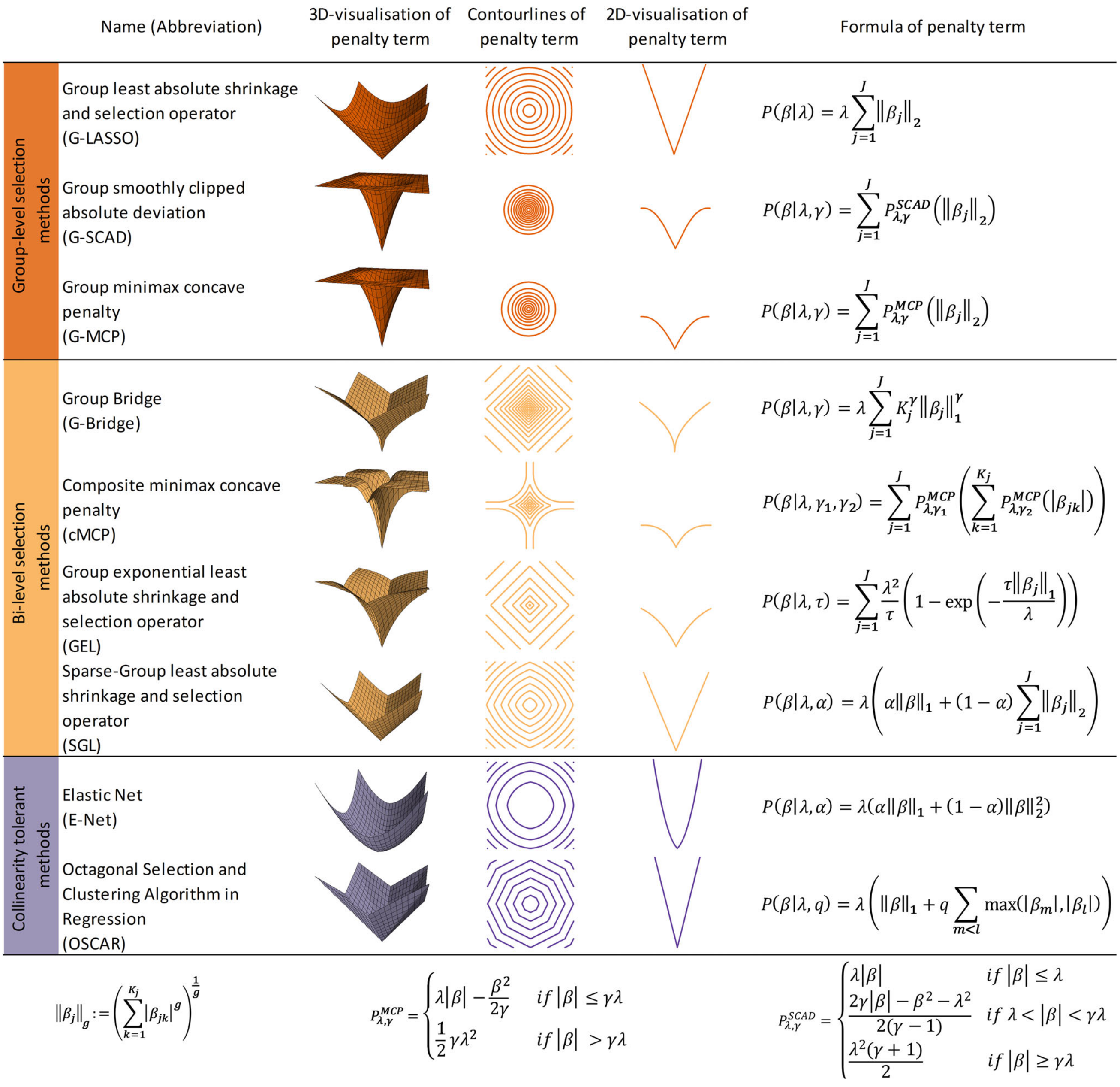
With $\boldsymbol{\beta}_j$ denoting the vector of coefficients for the jth variable group.

The circles in the contour plot in Figure 2 of G-LASSO visualize its "all-in or all-out" property: Since the regularization applied within a group is differentiable everywhere, the resulting contour lines are circular, and sparsity within a group is not introduced. Therefore, coefficients are only set to zero when the whole group is excluded from the model.

A suitable implementation of this method is in the grpreg package.[31] Here, the default is to weight $\left\|\boldsymbol{\beta}_j\right\|_2$ by $\sqrt{K_j}$, that is, the square root of the group size, to account for groups of different sizes.

G-SCAD and G-MCP

One disadvantage of G-LASSO received considerable attention in the literature: Like the classical LASSO, G-LASSO tends to over-shrink large coefficients, as the penalization does not change with the magnitude of coefficients.[32] Although



	Name (Abbreviation)	Formula of penalty term
Group-level selection methods	Group least absolute shrinkage and selection operator (G-LASSO)	$P(\beta\vert\lambda) = \lambda \sum_{j=1}^{J} \left\Vert\beta_j\right\Vert_2$
	Group smoothly clipped absolute deviation (G-SCAD)	$P(\beta\vert\lambda,\gamma) = \sum_{j=1}^{J} P_{\lambda,\gamma}^{SCAD}\left(\left\Vert\beta_j\right\Vert_2\right)$
	Group minimax concave penalty (G-MCP)	$P(\beta\vert\lambda,\gamma) = \sum_{j=1}^{J} P_{\lambda,\gamma}^{MCP}\left(\left\Vert\beta_j\right\Vert_2\right)$
Bi-level selection methods	Group Bridge (G-Bridge)	$P(\beta\vert\lambda,\gamma) = \lambda \sum_{j=1}^{J} K_j^{\gamma} \left\Vert\beta_j\right\Vert_1^{\gamma}$
	Composite minimax concave penalty (cMCP)	$P(\beta\vert\lambda,\gamma_1,\gamma_2) = \sum_{j=1}^{J} P_{\lambda,\gamma_1}^{MCP}\left(\sum_{k=1}^{K_j} P_{\lambda,\gamma_2}^{MCP}(\vert\beta_{jk}\vert)\right)$
	Group exponential least absolute shrinkage and selection operator (GEL)	$P(\beta\vert\lambda,\tau) = \sum_{j=1}^{J} \frac{\lambda^2}{\tau}\left(1 - \exp\left(-\frac{\tau\left\Vert\beta_j\right\Vert_1}{\lambda}\right)\right)$
	Sparse-Group least absolute shrinkage and selection operator (SGL)	$P(\beta\vert\lambda,\alpha) = \lambda\left(\alpha\Vert\beta\Vert_1 + (1-\alpha)\sum_{j=1}^{J}\left\Vert\beta_j\right\Vert_2\right)$
Collinearity tolerant methods	Elastic Net (E-Net)	$P(\beta\vert\lambda,\alpha) = \lambda(\alpha\Vert\beta\Vert_1 + (1-\alpha)\Vert\beta\Vert_2^2)$
	Octagonal Selection and Clustering Algorithm in Regression (OSCAR)	$P(\beta\vert\lambda,q) = \lambda\left(\Vert\beta\Vert_1 + q\sum_{m<l}\max(\vert\beta_m\vert,\vert\beta_l\vert)\right)$

$$\left\Vert\beta_j\right\Vert_g := \left(\sum_{k=1}^{K_j}\vert\beta_{jk}\vert^g\right)^{\frac{1}{g}}$$

$$P_{\lambda,\gamma}^{MCP} = \begin{cases} \lambda\vert\beta\vert - \frac{\beta^2}{2\gamma} & if\ \vert\beta\vert \le \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2 & if\ \vert\beta\vert > \gamma\lambda \end{cases}$$

$$P_{\lambda,\gamma}^{SCAD} = \begin{cases} \lambda\vert\beta\vert & if\ \vert\beta\vert \le \lambda \\ \frac{2\gamma\vert\beta\vert - \beta^2 - \lambda^2}{2(\gamma-1)} & if\ \lambda < \vert\beta\vert < \gamma\lambda \\ \frac{\lambda^2(\gamma+1)}{2} & if\ \vert\beta\vert \ge \gamma\lambda \end{cases}$$

3D-visualisation of penalty term
Contourlines of penalty term
2D-visualisation of penalty term

FIGURE 2 Overview of penalty terms for group variable selection. The $\left\Vert\boldsymbol{\beta}_j\right\Vert_2$ in the group-level selection methods are often weighted with $\sqrt{K_j}$, to account for different group sizes. The visualizations in the middle three columns show the magnitude of penalization for a grid of two different coefficients belonging to two variables of the same group. Values for γ and τ were set to the values recommended in the corresponding main source (see Table 1). Values for α were set to 0.5. Values for q and λ were set to 1. G-LARS is not shown because it is an algorithm that does not use a penalty function. BiSEE, HiSEE, and MLGL are not shown because the penalty functions used in these approaches (SGL and G-LASSO) are shown as separate methods. OHPL is not shown because the penalty function used in this approach (LASSO) does not lead to group variable selection by itself

shrinking coefficients towards zero is intended, the resulting bias can be reduced with more advanced methods. In addition to the generally applicable idea of adaptively weighting of penalties,[33,34] group minimax concave penalty (G-MCP) and group smoothly clipped absolute deviation (G-SCAD) optimize the trade-off between selection and introduction of bias.[35] They penalize within a group like G-LASSO, but between groups with the minimax concave penalty (MCP),[36] or smoothly clipped absolute deviation (SCAD),[37] respectively.

Both methods use thresholds triggered by the tuning parameter γ to change the amount of penalization according to the magnitude of the signal. They initially penalize like LASSO, but continuously relax this penalization until the rate drops to zero. The range where the penalization is reduced to zero can be determined with γ.

$$P^{\text{G-SCAD}}(\boldsymbol{\beta}|\lambda,\gamma) = \sum_{j=1}^{J} \begin{cases} \lambda \left\|\boldsymbol{\beta}_j\right\|_2, & \text{if } \left\|\boldsymbol{\beta}_j\right\|_2 \leq \lambda, \\ \frac{2\gamma\|\boldsymbol{\beta}_j\|_2 - \|\boldsymbol{\beta}_j\|_2^2 - \lambda^2}{2(\gamma-1)}, & \text{if } \lambda < \left\|\boldsymbol{\beta}_j\right\|_2 < \gamma\lambda, \\ \frac{\lambda^2(\gamma+1)}{2}, & \text{if } \left\|\boldsymbol{\beta}_j\right\|_2 \geq \gamma\lambda, \end{cases} \tag{4}$$

$$P^{\text{G-MCP}}(\boldsymbol{\beta}|\lambda,\gamma) = \sum_{j=1}^{J} \begin{cases} \lambda \left\|\boldsymbol{\beta}_j\right\|_2 - \frac{\|\boldsymbol{\beta}_j\|_2^2}{2\gamma}, & \text{if } \left\|\boldsymbol{\beta}_j\right\|_2 \leq \gamma\lambda, \\ \frac{1}{2}\gamma\lambda^2, & \text{if } \left\|\boldsymbol{\beta}_j\right\|_2 > \gamma\lambda. \end{cases} \tag{5}$$

For G-SCAD, a γ of 4 and for G-MCP, a γ of 3 is suggested.[32] Their contour plots in Figure 2 are similar to those from G-LASSO, as they all penalize the L_2-norm of the coefficients of a group. However, their rate of penalization differs, as visualized in the fourth column of Figure 2: while the penalization of G-LASSO increases strictly, it relaxes for G-SCAD and G-MCP to a constant rate for large coefficients. Both methods are implemented in the R package grpreg. As in G-LASSO, $\left\|\boldsymbol{\beta}_j\right\|_2$ is weighted by $\sqrt{K_j}$ in G-SCAD and G-MCP to account for the group size.

G-LARS

Another group-level selection method is the group LARS (G-LARS),[29] which is the group version of the LARS algorithm.[38] LARS is a gentler version of a forward stepwise regression method and was the first efficient algorithm for delivering the entire solution path of the LASSO. The G-LARS algorithm starts with a null model and includes in each iteration the predictor group with the highest correlation to the current residual. Depending on how the correlation criterion is defined, different G-LARS variations are possible.[22]

The G-LARS procedure from the RobustHD package[39] uses the R^2 measure as criterion[40] and offers options to prevent outliers from distorting the results. In this implementation the maximum number of groups that should be included in the final model has to be defined with the parameter sMax. The default is set according to formula (6)

$$\text{sMax} = \frac{n}{2 * \overline{K}}, \tag{6}$$

with $\overline{K}$ denoting the average group size.

3.2.2 | Bi-level selection methods

Bi-level selection methods can identify important groups, without including all of its members in the model.[15] Such methods are appropriate if group information is available but it is not assumed that all variables of a group are necessarily relevant.

The literature research conducted revealed four penalty-based methods and two algorithms based on generalized estimating equations for bi-level selection. The penalty-based methods belong to different frameworks, depending on how they introduce the bi-level sparsity. Breheny and Huang[15] introduced a hierarchical approach, where an outer penalty function is applied to the sum of an inner penalty function as given in formula (7).

$$P(\boldsymbol{\beta}|\lambda) = \sum_{j=1}^{J} P^{\text{Outer}}\left(\sum_{k=1}^{K_j} P^{\text{Inner}}\left(|\beta_{jk}|\right)\right), \tag{7}$$

where β_{jk} denotes coefficient for the kth variable in the jth variable group in $\boldsymbol{X}$.

The alternative approach combines different penalties in an additive fashion like in formula (8).[12]

$$P(\boldsymbol{\beta}|\lambda) = \sum_{j=1}^{J}\sum_{k=1}^{K_j} P^{\text{Variable}}\left(\beta_{jk}\right) + \sum_{j=1}^{J} P^{\text{Group}}\left(\left\|\boldsymbol{\beta}_j\right\|_2\right). \tag{8}$$

G-bridge

The pioneer method for bi-level selection is the group bridge (G-bridge),[13] which extends the bridge penalty to group variable selection.[41] It belongs to the hierarchical framework, applying the bridge penalty as an outer function and the LASSO penalty as an inner function.

If its tuning parameter γ equals 1, the method simplifies to the LASSO penalty weighted by group sizes and does not perform bi-level selection. A γ greater than 1 does not introduce sparsity, but with a γ between 0 and 1, bi-level selection can be performed.

$$P^{\text{G-bridge}}(\boldsymbol{\beta}|\lambda,\gamma) = \lambda \sum_{j=1}^{J} K_j^{\gamma} \left\| \boldsymbol{\beta}_j \right\|_1^{\gamma}. \tag{9}$$

A typical value for γ is 0.5,[13,42] which corresponds to penalizing the square roots of the sum over the absolute values of the regression coefficients of a group. This value is also the default in the implementation of the method in the grpreg package, which we used in the simulation study. Since the LASSO penalty is applied within groups, variables within a nonzero group can still be removed, which is visualized in the contour plots of Figure 2: like for all bi-level selection methods, the contour lines of G-bridge have edges at which coefficients can be set to zero.

cMCP

Similar behavior to G-bridge is exhibited by the composite MCP (cMCP),[15] which applies the MCP as an outer and inner penalty. As already mentioned, LASSO does not change the amount of penalization according to the magnitude of a coefficient. The same holds for the LASSO penalty within the G-bridge, so the penalization rate of G-bridge never flattens to zero. Since cMCP applies MCP at both levels of the selection hierarchy, the resulting approach relaxes not only the shrinkage at the group level, but also at the within-group level. This can be seen in the fourth column of Figure 2, where the rate of penalization relaxes, similar to G-SCAD and G-MCP.

$$P^{\text{cMCP}}\left(\boldsymbol{\beta}|\lambda,\gamma_1,\gamma_2\right) = \sum_{j=1}^{J} P^{\text{MCP}}_{\lambda,\gamma_1}\left(\sum_{k=1}^{K_j} P^{\text{MCP}}_{\lambda,\gamma_2}\left(\left|\beta_{jk}\right|\right) \right), \tag{10}$$

$$P^{\text{MCP}}_{\lambda,\gamma} = \sum \begin{cases} \lambda|\theta| - \frac{\theta^2}{2\gamma}, & \text{if } |\theta| \leq \gamma\lambda, \\ \frac{1}{2}\gamma\lambda^2, & \text{if } |\theta| > \gamma\lambda. \end{cases} \tag{11}$$

In the R package grpreg γ of the outer penalty is set as a function of the group sizes, while γ of the inner penalty has to be tuned. This tuning parameter handles the shrinkage reduction within a group. A small γ increases the region of constant penalization, whereas the opposite leads to a more LASSO-like penalization. Breheny, Huang[15] recommend setting the inner γ to 3.

GEL

In a bi-level selection method, a variable can enter the model because of its own signal or due to its membership in a group with a strong collective signal. While the influence of the group signal on the threshold for variables within a group is not transparent for G-bridge and cMCP, it can be directly controlled in group exponential LASSO (GEL)[11] and sparse group LASSO (SGL),[12] the latter of which is explained in the next subsection.

$$P^{\text{GEL}}(\beta|\lambda,\tau) = \sum_{j=1}^{J} \frac{\lambda^2}{\tau} \left(1 - \exp\left(-\frac{\tau \left\| \boldsymbol{\beta}_j \right\|_1}{\lambda} \right) \right). \tag{12}$$

GEL applies an outer exponential penalty to an inner LASSO penalty. Its tuning parameter τ controls the degree of "coupling," that is, the influence variables of a group have on the threshold of other variables within the same group. If τ is close to zero, the influence is small, meaning that variables have to overcome the threshold on their own. The closer τ gets to 1, the more the threshold is influenced by other variables of the same group. A suggested value for τ is 1/3,[11] which we used in the simulation study with the implementation of GEL in the grpreg package.

SGL

SGL combines the penalty term of LASSO with that of G-LASSO additively.[43,44] The tuning parameter α balances the sparsity between and within groups. An α of 1 reduces SGL to the LASSO, whereas an α of 0 corresponds to G-LASSO. Values in between those extremes perform bi-level sparsity by including more variables of selected groups the closer α gets to zero. If a variable is in a group where all other variables are not predictive, it receives the penalty of LASSO and G-LASSO. However, if the variable belongs to a group with other important variables, shrinkage can be reduced to be mainly determined by the rate of the LASSO penalty weighted by α.

$$P^{\mathrm{SGL}}(\boldsymbol{\beta}|\lambda,\alpha)=\lambda\left(\alpha\|\boldsymbol{\beta}\|_1+(1-\alpha)\sum_{j=1}^{J}\left\|\boldsymbol{\beta}_j\right\|_2\right). \tag{13}$$

Simon et al[12] implemented SGL in the R package SGL and suggested an α of 0.95, which we used in the simulation study. For the visualization of the method in Figure 2, α was set to 0.5 to highlight differences to other methods.

BiSEE and HiSEE

As the name suggests, bi-level stagewise estimation equation (BiSEE) and hierarchical stagewise estimation equation (HiSEE) are forward stage-wise algorithms, performing bi-level selection within the generalized estimation equation framework.[45] Those algorithms start with a full model with all coefficients set to zero, followed by sequentially updating the coefficients. In each iteration, the coefficients of the predictors that can best explain the current residuals are updated by a small step size. To perform bi-level selection, only a subset of coefficients within a particular group should be updated at each step. Therefore, BiSEE includes the SGL penalty in the update step, while HiSEE updates hierarchically: first the most relevant variable group is identified using the G-LASSO penalty, then the most relevant variables within that group are identified with the LASSO penalty. Compared to a direct application of the respective penalty methods, the implementation within the generalization estimation equation framework enables to account for clustered observations, such as those obtained in a longitudinal setting.

The default for α is set to 0.5 for BiSEE in the R package sgee,[46] which corresponds to the SGL visualization in Figure 2. For both BiSEE and HiSEE the step size was set to the package default of 0.05 in the simulation study.

3.2.3 | Two-step approaches

A common motivation of two-step approaches is to account for a high correlation in the dataset to stabilize the selection process.[47] The principal idea is to apply a data-driven group formation technique and a selection method sequentially. The data-driven grouping step alone can be performed in several ways,[48] and derived group formation could be used in all the knowledge-driven group variable selection methods, resulting in multiple different approaches. Therefore, this review focused only on approaches that perform the complete group formation and selection sequence automatically.

MLGL

Multi-layer group-LASSO (MLGL) is a sequence of methods combining hierarchical clustering and group-level selection.[49] It starts with hierarchically clustering the data, followed by a G-LASSO selection, where all levels of the hierarchical partition are used as group information. This is achieved by duplicating variables and then assigning them to multiple groups, so that selection across different levels of the hierarchy becomes possible. Since selected groups are potentially located at different levels of the hierarchy and may overlap, the selection may suffer from redundancy and inflated groups. Therefore, a hierarchical multiple testing procedure is necessary as a further step, to identify the smallest predictive groups. The testing procedure starts with reducing each group to its first principal component to overcome potential high dimensionality problems of testing. Those components are tested sequentially with partial F-tests, beginning at the top of the hierarchical tree moving downwards, while controlling for the family-wise error rate.[50] Significant groups are considered to be selected. The MLGL package performs the entire procedure automatically for the linear regression setting, with an alpha level of 0.05 as default for the testing strategy.

OHPL

Originally, ordered homogeneity pursuit LASSO (OHPL)[51] is a wavelength region selection method,[52] requiring ordered variables. However, OHPL can also be used in situations with a more general group structure.

The first step of OHPL is fitting a partial least squares regression (PLS) to the dataset. Thereby, the number of principal components has to be defined, for example, based on the standard errors of CV residuals. In the second step, Fisher's optimal partitions algorithm[53] is applied to the regression coefficients of the PLS model to define variable groups. This algorithm attempts to divide the values of the regression coefficients into groups, so that the values within a group are homogenous but significantly different from the values of other groups. For the algorithm, the number of groups has to be specified, which can be done with a grid search. In the third step, the variable with the highest correlation with the dependent variable of each group is identified. Those variables are considered as representatives of the groups. At last, LASSO is applied to the representative variables. The selected representatives, as well as their corresponding groups, are considered to be selected. For this step λ has to be defined, which can be done with CV. This approach is implemented in the R package OHPL, which automatically determines an appropriate number of principal components and the number of variable groups.

3.2.4 | Collinearity tolerant methods

When a variable selection method such as LASSO shrinks strongly correlated variables towards zero, one of them may be set to zero while others remain in the model. If the interest lies in interpreting the selection output, such behavior may be undesirable. In particular, if a high correlation is assumed to imply similar information, it would increase the interpretability if those variables were treated analogously. Collinearity tolerant methods achieve this by assigning similar coefficient values to correlated variables.

Compared to knowledge-driven approaches, their penalty acts directly at the variable level, without considering predefined groupings. They also do not group variables like the two-level approaches, so that their results do not provide information about which variables belong together per se. Only their tendency to treat correlated variables similarly justifies the widely held view in the literature that collinearity tolerant approaches are group variable selection methods.[26,47,54]

E-Net

The elastic net (E-Net)[55] combines the L_1 and L_2 norms in its penalty term, which corresponds to combining LASSO with Ridge regression. Since the L_1-norm introduces parsimonious models, while L_2-norm assigns similar coefficient values to correlated variables, the E-Net selects relevant variables and introduces similarity between correlated predictors. The combination of the two penalty terms is a weighting, controlled by the parameter α, ranging from 0 to 1. In glmnet[56] an α of 0 sets the total weight to the Ridge penalty, while the other extreme leads to LASSO. Setting α to 0.5 leads to a 50/50 mix of both methods, which is hereafter defined as an E-Net.

$$P^{\text{E-Net}}(\boldsymbol{\beta}|\lambda,\alpha) = \lambda\left(\alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_2^2\right). \tag{14}$$

The idea of including the L_2-norm into the penalty term to account for collinearity has been adopted to many penalties.[31,57] However, the LASSO-Ridge-mix was the only combination for which an article was found introducing the approach as a separate method.

OSCAR

An alternative approach to account for collinearity across predictors within the selection process is the octagonal shrinkage and clustering algorithm for regression (OSCAR),[58] which combines the L_1-norm with a pairwise L_∞-norm. The method is based on a stronger assumption than the E-Net, as it assumes that correlated variables will have the same absolute values of their coefficients. This tendency arises from the name-giving octagonal shape of its contour lines in the third column of Figure 2. While other methods introduce edges only at zero points, penalization with OSCAR proposes stronger restrictions: it forces correlated variables to have similar coefficients. This is achieved with the L_∞-norm, which promotes equal values of nonzero coefficients, by penalizing two equal coefficient values fewer than two different ones.

$$P^{\text{OSCAR}}(\boldsymbol{\beta}|\lambda,q) = \lambda\left(\|\boldsymbol{\beta}\|_1 + q\sum_{m<l}\max(|\beta_m|,|\beta_l|)\right). \tag{15}$$

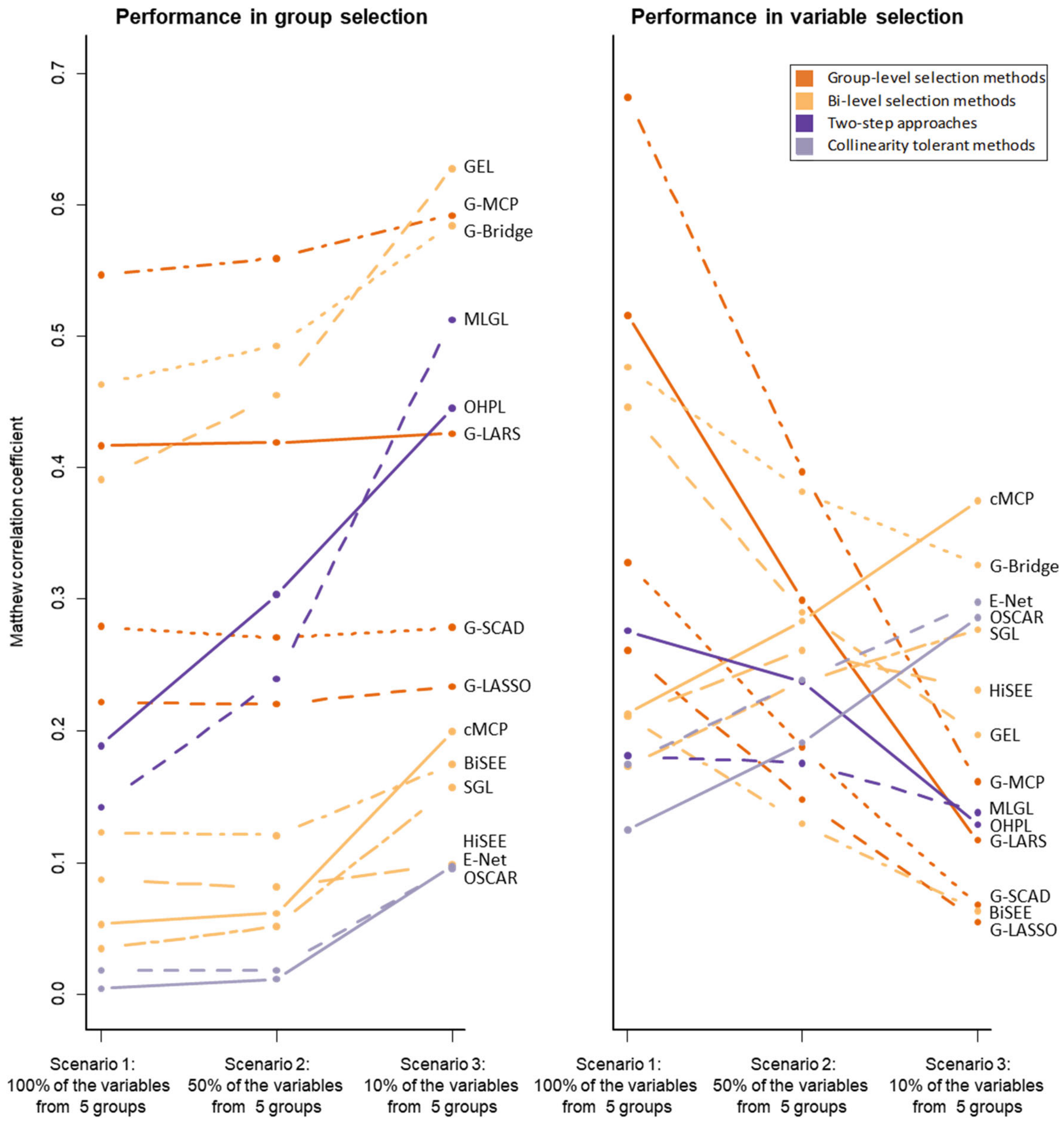


FIGURE 3 Simulation results. MCC for three scenarios. 5 out of 10 variable groups are related to the dependent variable. 100%, 50%, or 10% of their variables have nonzero effect. Number of observations: 500, total number of variables: 500. The performance of the methods is described by the median values derived from 1000 simulated replicates. MCC values close to 1 imply that relevant groups/variables are selected and irrelevant groups/variables are excluded. Random selection results in values close to 0

The parameter q is used in the R package SLOPE[59] to tune the grouping effect. Low values for q lead to LASSO-like selections, while larger values result in more equal coefficients. Values of 0.1 or lower are the default for OSCAR in the SLOPE package. For the visualization in Figure 2, the value for q was set to 1 to highlight differences from other penalties.

3.3 | Results of simulation studies

The performance of the identified methods was evaluated using simulated data. For this purpose, the methods of each category were compared with each other to identify the best among them. The primary evaluation criterion was the MCC for variable and group variable selection, which are shown in Figure 3 for all scenarios. An alternative visualization in which the results are stratified by simulation scenario is provided in the Supplemental Appendix.

3.3.1 | Selection performance of knowledge-driven approaches

The first result to note is that across the three scenarios, different methods demonstrated the best performance, yet they always belong to the knowledge-driven approaches. Thereby, group-level selection methods comprised the method that outperformed the others in identifying relevant groups and variables in the first two scenarios: G-MCP. Its superiority is particularly evident in the first scenario, while it achieved less outstanding results in the second situation. Even though the method was still leading here, its performance in variable selection dropped substantially, while that of group selection even slightly increased. In the third scenario, a critical disadvantage of group-level selection methods becomes apparent, as they are prone to include irrelevant variables when not all variables of a group are truly related to the dependent variable. In this scenario, only 10% of the variables should be selected out of the predictive groups, so the “all-in or all-out” approach is too extreme. In such cases, bi-level selection methods demonstrated better performance in terms of variable and group variable selection.

Methods of this category achieved heterogeneous results in both group and variable selection. First, there is a substantial difference between GEL and G-bridge compared to the other bi-level selection methods for the selection of variable groups: GEL and G-bridge performed remarkably well, while all other bi-level selection methods were inferior to the two-step methods, which had not even received the true group information. Second, a comparison regarding variable selection performance showed also distinct patterns of behavior: the MCC of BiSEE, GEL, and G-bridge decreased with fewer relevant variables per group, whereas it increased for SGL and cMCP.

This heterogeneous behavior can be attributed to how parsimoniously the methods perform selection at the group or variable level. The average number of selected groups (Supplemental Appendix, Tables S1-S3) demonstrates that GEL and G-bridge were parsimonious in the selection of groups, while the other bi-level selection methods selected an average of nine or more groups, which does not correspond to the idea of highlighting only relevant signals. However, although GEL and G-bridge selected fewer groups, cMCP was even more parsimonious at the variable level in scenarios two and three. This implies that cMCP tends to select variables with a strong signal among many groups, while GEL and G-bridge include fewer groups but more of their variables in the model.

Regarding group selection, G-bridge achieved the best performance of the bi-level selection methods in the first two scenarios, but remained below the performance of G-MCP. GEL had slightly lower MCC values in scenarios 1 and 2, but then outperformed all methods in scenario 3 in terms of group selection. Despite this performance, GEL always fell behind G-bridge in terms of variable selection, which was among the top three for this performance measure in all scenarios. In scenario 3, it is only outperformed by cMCP, which, however, was less successful in the other scenarios. So overall, G-bridge seems to be the most consistent bi-level selection performer in both group and variable selection, while GEL and cMCP are specialized to certain selection tasks.

3.3.2 | Selection performance of data-driven approaches

As already described, data-driven methods performed inferior to knowledge-driven methods over the scenarios. As such, OHPL and MLGL operated poorly in identifying relevant variables across all scenarios. Only when 100% of a group had nonzero effects (scenario 1) or few groups were relevant (Supplemental Appendix, Table S4) did they outperform collinearity tolerant methods. However, in scenario 3, they were superior to all group-level selection methods except for G-MCP in group and variable selection. Although other methods performed even better, this achievement is remarkable, as two-step approaches perform the grouping of variables on their own, while group-level selection methods receive prior information and thus have an advantage.

The ability to both group and select variables gives MLGL and OHPL a special position, as they provide additional information that other methods do not reveal. Therefore, the MCC of group formation was determined for both methods (Supplemental Appendix, Tables S1-S3). The correlation suggests a better group formation by MLGL compared to OHPL, but the opposite holds for selecting variables and variable groups in scenarios 1 and 2. Here, OHPL outperformed MLGL, mainly because the parsimonious models that MLGL built (Supplemental Appendix, Tables S1 and S2) are inappropriate here.

Considering the group variable selection performance of collinearity tolerant methods, it becomes evident that this technique is unsuitable for selecting variable groups. OSCAR and E-Net selected on average between nine and ten variable groups, which was rated with MCC values close to zero (Supplemental Appendix, Tables S1-S3). On the other hand, they were quite convincing in the third scenario regarding the selection of variables: Followed by OSCAR, E-Net was the best method after G-bridge and cMCP concerning this evaluation criterion. So they performed relatively well here, even though they have received no group information. Part of the reason for this is that in the third scenario, the group information was less valuable than in other scenarios, as only small parts of the groups were truly relevant. Another reason is that compared to group-level and two-step methods, collinearity tolerant methods are independent of the “all in or all out” constraint, which allows them to select only a fraction of variables out of a group, like bi-level selection methods. In the scenarios considered here, E-Net always performed better than OSCAR in both evaluation criteria, although the differences were steadily diminishing from scenarios 1 to 3.

3.3.3 | Performance on supporting measures

To further compare the methods, other evaluation criteria, such as mean squared error, sensitivity and specificity, were computed. The corresponding performance can be found in Table 2. Regarding the prediction performance of the models, group-level selection and two-step approaches show quite stable performance. Across the three scenarios, they achieve similar results, while that of the bi-level and collinear tolerant methods improved from scenario 1 to 3. Since variables within a group can be highly correlated (up to 0.95), a good prediction model can be generated without activating all members of a group. For example, in scenario 1, it may be sufficient to select only part of a group to reflect the main signal of it in the model. Hence, also the classical LASSO performs well on this evaluation criterion and is partially superior to the data-driven approaches (Supplemental Appendix, Tables S1-S3). Only the knowledge-driven methods can increase the prediction substantially (G-SCAD, G-MCP, cMCP). Thereby, G-MCP seems to be recommendable when many features of a group are associated with the response, while cMCP is preferable when few variables of the groups have a predictive effect. However, since in such situations, collinear tolerant methods, such as the E-Net, perform almost equally well, the benefit of using prior information seems limited.

Comparing the approaches regarding their sensitivity and specificity, it is noticeable that some methods optimize only one of the two criteria. The methods performing best in sensitivity are often the methods performing worst in specificity of the scenario, such as OSCAR in scenario 1, which has the best sensitivity but also the worst specificity on group selection. This is because OSCAR tends to include almost all groups in the model (Supplemental Appendix, Table S1). Likewise, the highest values in variable and group specificity are regularly accompanied by moderate sensitivity within the same scenario. Corresponding models are often too sparse, like that of G-LARS in the first scenario, which activates only 1.6 groups on average (Supplemental Appendix, Table S2). Regarding group and variable selection, G-LASSO, G-SCAD, BiSEE, OSCAR and mostly OHPL achieve higher sensitivity than specificity, whereas the opposite is true for G-LARS, G-bridge, GEL, MLGL, and largely G-MCP. The selection procedures SGL, cMCP, E-Net, and HiSEE are sensitive in group selection and specific in variable selection. To identify methods with a balanced specificity and sensitivity in variable and group selection, the mean squared error provides a first indication. But this evaluation criterion does not differentiate precisely in some settings, like in scenario 1, where G-SCAD and G-MCP achieved both values of 0.67 but had very different performances in sensitivity and specificity. The MCC discriminated much better here and was therefore chosen as primary evaluation criterion.

The computing time required by the methods for the selection was also measured within the simulation study. The slowest methods by far were OSCAR and SGL, with runtimes of over an hour in some cases. Both approaches were 10 to 20 times slower than for example, two-step methods. Thereby, changing the sample size (Supplemental Appendix, Tables S8 and S9) or simplifying the correlation structure (Supplemental Appendix, Tables S10-S12) leads to strong speed advances for OSCAR but not for SGL. The fastest method was G-LARS with a runtime of only 4 to 5 seconds. Only in the situation with 1000 observations (Supplemental Appendix, Table S9) did G-LARS slow down and G-MCP and G-SCAD were slightly faster.

The effect of different signal-to-noise ratios (Supplemental Appendix, Tables S4 and S5), number of predictive variable groups (Supplemental Appendix, Tables S6 and S7), and total number of observations (Supplemental Appendix, Tables S8 and S9) were investigated in further scenarios presented in the supplement and support the main findings. In general, methods perform better in terms of MCC for variable and group variable selection when the signal-to-noise ratio or the number of observations is increased, or the number of predictive variable groups is decreased.

It is noteworthy that the relative ranking of the methods in terms of performance hardly changes across the scenarios in the Supplemental Appendix. The best methods in regard to variable or group selection as well as prediction performance were always G-MCP or G-bridge. Moreover, with the following exceptions, both methods are always the best in their category. In scenario 10, G-LARS is the best group-level selection method in variable and group selection, and G-LASSO

TABLE 2 Results of the simulation study

		Scenario 1					Scenario 2					Scenario 3				
		Variable level		Group level			Variable level		Group level			Variable level		Group level		
	Name (abbreviation)	SE	SP	SE	SP	MSE	SE	SP	SE	SP	MSE	SE	SP	SE	SP	MSE
Group-level selection methods	Group least absolute shrinkage and selection operator (G-LASSO)	**0.93** **(0.11)**	0.30 (0.29)	0.86 (0.16)	0.33 (0.26)	0.69 (0.05)	0.93 (0.12)	0.21 (0.2)	0.87 (0.16)	0.32 (0.25)	0.69 (0.05)	0.95 (0.09)	0.16 (0.16)	0.91 (0.14)	0.28 (0.24)	0.70 (0.05)
	Group least-angle regression (G-LARS)	0.43 (0.18)	**1.00** **(0.01)**	0.31 (0.15)	**0.99** **(0.03)**	0.76 (0.1)	0.43 (0.18)	0.86 (0.05)	0.32 (0.15)	**0.99** **(0.03)**	0.76 (0.11)	0.42 (0.18)	0.81 (0.08)	0.34 (0.17)	0.99 (0.05)	0.78 (0.11)
	Group smoothly clipped absolute deviation (G-SCAD)	0.92 (0.12)	0.37 (0.29)	0.85 (0.16)	0.40 (0.26)	**0.67** **(0.05)**	0.91 (0.13)	0.28 (0.22)	0.84 (0.17)	0.39 (0.26)	0.68 (0.05)	0.94 (0.1)	0.19 (0.17)	0.89 (0.14)	0.34 (0.24)	0.68 (0.05)
	Group minimax concave penalty (G-MCP)	0.70 (0.19)	0.95 (0.1)	0.58 (0.21)	0.92 (0.13)	**0.67** **(0.05)**	0.71 (0.2)	0.73 (0.12)	0.59 (0.22)	0.92 (0.13)	**0.67** **(0.05)**	0.75 (0.19)	0.61 (0.13)	0.66 (0.21)	0.90 (0.14)	0.68 (0.05)
Bi-level selection methods	Group bridge (G-bridge)	0.40 (0.12)	0.99 (0.05)	0.38 (0.16)	0.98 (0.06)	0.72 (0.07)	0.42 (0.13)	**0.91** **(0.05)**	0.41 (0.16)	0.99 (0.05)	0.68 (0.07)	0.50 (0.15)	0.92 (0.05)	0.52 (0.19)	**0.99** **(0.04)**	0.60 (0.05)
	Sparse-group least absolute shrinkage and selection operator (SGL)	0.39 (0.1)	0.77 (0.09)	0.95 (0.11)	0.07 (0.14)	0.80 (0.08)	0.46 (0.1)	0.79 (0.07)	0.96 (0.1)	0.08 (0.14)	0.73 (0.07)	0.62 (0.12)	0.85 (0.05)	0.96 (0.09)	0.15 (0.17)	0.61 (0.04)
	Composite minimax concave penalty (C-MCP)	0.27 (0.11)	0.90 (0.05)	0.92 (0.14)	0.12 (0.18)	0.82 (0.09)	0.35 (0.1)	0.90 (0.04)	0.93 (0.12)	0.11 (0.16)	0.73 (0.08)	0.53 (0.13)	**0.94** **(0.03)**	0.95 (0.1)	0.21 (0.19)	**0.58** **(0.04)**
	Group exponential least absolute shrinkage and selection operator (GEL)	0.36 (0.18)	0.99 (0.03)	0.42 (0.2)	0.91 (0.15)	0.78 (0.12)	0.40 (0.17)	0.87 (0.05)	0.48 (0.2)	0.91 (0.14)	0.76 (0.1)	0.41 (0.15)	0.86 (0.08)	0.63 (0.19)	0.95 (0.11)	0.71 (0.08)
	Bi-level stagewise estimating equation (BiSEE)	0.91 (0.1)	0.26 (0.18)	0.92 (0.12)	0.17 (0.17)	0.74 (0.07)	**0.93** **(0.08)**	0.19 (0.12)	0.92 (0.12)	0.17 (0.17)	0.71 (0.07)	**0.97** **(0.06)**	0.14 (0.1)	0.97 (0.08)	0.16 (0.16)	0.66 (0.05)
	Hierarchical stagewise estimating equation (HiSEE)	0.33 (0.07)	0.85 (0.06)	0.94 (0.1)	0.12 (0.15)	0.83 (0.09)	0.42 (0.08)	0.84 (0.03)	0.95 (0.1)	0.11 (0.15)	0.77 (0.08)	0.65 (0.12)	0.80 (0.03)	**0.98** **(0.06)**	0.09 (0.13)	0.68 (0.06)

(Continues)

TABLE 2 Continued

		Scenario 1					Scenario 2					Scenario 3				
		Variable level		Group level			Variable level		Group level			Variable level		Group level		
	Name (abbreviation)	SE	SP	SE	SP	MSE	SE	SP	SE	SP	MSE	SE	SP	SE	SP	MSE
Two-step approaches	Multi-layer group-least absolute shrinkage and selection operator (MLGL)	0.18 (0.21)	0.96 (0.11)	0.14 (0.16)	0.96 (0.1)	0.98 (0.05)	0.29 (0.24)	0.88 (0.11)	0.22 (0.19)	0.95 (0.1)	0.98 (0.05)	0.63 (0.22)	0.67 (0.13)	0.53 (0.21)	0.93 (0.12)	0.90 (0.08)
	Ordered homogeneity pursuit least absolute shrinkage and selection operator (OHPL)	0.74 (0.23)	0.51 (0.33)	0.68 (0.26)	0.49 (0.32)	0.99 (0.03)	0.69 (0.22)	0.56 (0.26)	0.59 (0.26)	0.67 (0.31)	0.97 (0.03)	0.59 (0.21)	0.68 (0.16)	0.47 (0.22)	0.90 (0.18)	0.89 (0.06)
Collinearity tolerant methods	Elastic Net (E-Net)	0.37 (0.12)	0.79 (0.09)	0.97 (0.1)	0.04 (0.13)	0.81 (0.07)	0.44 (0.12)	0.80 (0.07)	0.97 (0.09)	0.04 (0.12)	0.74 (0.07)	0.60 (0.13)	0.88 (0.05)	0.98 (0.07)	0.09 (0.13)	0.61 (0.04)
	Octagonal selection and clustering algorithm in regression (OSCAR)	0.62 (0.22)	0.50 (0.25)	**0.98** **(0.07)**	0.02 (0.1)	0.80 (0.07)	0.58 (0.19)	0.62 (0.21)	**0.99** **(0.06)**	0.02 (0.09)	0.74 (0.07)	0.62 (0.14)	0.84 (0.1)	0.98 (0.07)	0.09 (0.14)	0.61 (0.04)

Note: Results of the simulation study. The generated datasets consist of 10 non-overlapping groups with different correlation structure and size. 5 groups are relevant, 100% (scenario 1), 50% (scenario 2), or 10% (scenario 3) of their variables are related to the response, number of observations: 500, total number of variables: 500, signal-to-noise ratio: 1. The performance in sensitivity (SE), specificity (SP), and mean squared error (MSE) are described by the mean values and the SD derived from 1000 iterations. MSE was evaluated on a hold-out dataset of the same data generating process. The best results are shown in bold.

and G-SCAD share the first rank in prediction performance in scenario 12. Both scenarios differ from the others in having a simplified group and correlation structure, which is also beneficial for some bi-level selection methods: in Scenario 11 and 12, G-bridge is replaced by BiSEE as the favorite in terms of predictive performance. This approach also outperforms G-bridge in variable and group selection when the dataset consists of fewer observations than predictors, as in Scenario 8, where SGL achieves the best predictive performance of all bi-level selection techniques.

4 | DISCUSSION

The objective of this work was to summarize and evaluate the current knowledge on established group variable selection methods appropriate for research on continuous outcome types. A systematic review of the literature identified 14 candidate methods, which were classified into knowledge-driven and data-driven approaches. The first category encompasses group-level and bi-level selection methods, while two-step and collinearity tolerant approaches belong to the second category. A simulation study was then conducted to reveal relevant strengths and weaknesses of the methods in the linear regression situation. The results show that group-level selection methods comprise approaches that are superior to other methods in selecting relevant variable groups, but inferior in identifying important individual variables, once not all variables in the groups are predictive. In such situations, an identification of important individual variables can be better achieved through bi-level selection methods. Interestingly, two-step approaches and collinearity tolerant methods provide comparable results to knowledge-driven methods without prior knowledge when only a few variables per group are predictive. Hence, methods in all four categories are suitable for analyzing data with variables exhibiting a natural group structure, such as proteomics data.

Our systematic approach was able to complement the earlier overview[20] with less visible (G-LARS), recently emerged (GEL, HiSEE, BiSEE) and data-driven approaches (OHPL, MLGL, OSCAR, E-Net). Also novel is the comparison of the methods via a simulation study conducted by independent authors, that is, authors who have not already introduced a group variable selection approach. This may make the results less biased by conflicts of interest than in previous evaluations.[60] These results can be used to consider which method might be appropriate for a particular application. Of relevance is the research question and whether prior information about the group structure is available. The research question should clarify what the primary goal of the investigation is: to identify interrelated variables associated with a response variable or to identify variables associated with a particular response from a structure of related variables. The former might correspond to the selection of relevant pathways or functional groups and may be best addressed using group-level selection methods or two-step approaches. The latter refers to the identification of the most relevant variables from these structures, such as highlighting the most relevant proteins of the important pathways, in which case the use of bi-level selection or collinearity tolerant methods should be considered. Since the results of the simulation study indicate that incorporation of prior information improves the selection process, such information should be used if available. Therefore, knowledge-driven methods should be preferred over data-driven methods. This presumes that prior information is accurate and meaningful, as in the simulation study described here. In practice, prior knowledge may be incorrect, incomplete, or difficult to convert into overlap-free grouping. In proteomics, for example, databases may be incomplete or erroneous and measured proteins have multiple and overlapping functions. This is where bi-level selection methods can provide a solution, as they are not forced to treat all variables in a group equally. This feature allows them to handle group information more flexibly than group-level selection methods, so bi-level selection methods should be considered when the accuracy of prior information is in doubt.

When data-driven approaches are used, underlying assumptions imply that highly correlated variables contain related information and should therefore be treated similarly in the selection process. However, high correlation does not guarantee similarity in content. In a real-world application, it would therefore be more advisable not to trust the automatism blindly, but to assess each step of the pipeline manually.

Typically, the use of selection methods is motivated by improved prediction performance and increased interpretability. The strength of the group variable selection approaches appears to lie more in variable selection performance (ie, increased interpretability) than in building good predictive models. In particular, the data-driven approaches are little or no better than classical LASSO in terms of predictive performance, and the additional benefit obtained with knowledge-driven methods is rather moderate.

To build a good prediction model, it may be sufficient to include only a few representatives of a group in order to account for the group information adequately. However, when exploring biological processes, this behavior can lead to a spotty picture that is hard to interpret. As such, one protein may be responsible for multiple tasks, and its function, central to the object of study, may still be unclear even after selection. This is why, in proteomics, selection results are often enriched with information from databases, such as Metascape[61] to map the identified markers into interpretable structures. Knowledge-based group variable selection methods can optimize this approach. The prior knowledge could be used to group the proteins ahead of the selection process so that relevant functional groups can be identified. In this way, the interpretability of the results increases and sensitivity is enhanced, as the identification of weak signals that have high explanatory power only when considered as a joint entity becomes possible.

According to the results of the simulation study, G-MCP can be recommended when prior information on group structure is available and the goal of an analysis is to identify variable groups associated with a response variable. In case where prior information is available but primary interest is in selecting predictive variables from these groups, G-bridge is appropriate. When no prior information is available but yet there is interest in selecting variable groups, two-step approaches such as OHPL can be useful. If highly correlated variables are to be treated similarly in the selection process, E-Net is an appropriate choice. These recommendations are based on a general comparison between methods. Depending on the research objective, additional properties should be considered to refine the choice of method. For example, G-MCP and G-bridge are more specific than E-NET and OHPL in selecting groups of variables, which is appropriate when the research interest is related to the most relevant groups. The sensitivity of E-NET and OHPL, on the other hand, is helpful in obtaining a broader and more exploratory view.

However, it should be kept in mind that this research was not exhaustive and that further investigations are needed. Especially the simulation study could be elaborated by tuning control parameters of the methods or extending the scenarios to include noncontinuous distributed response variables. The literature search could have been expanded with a more general search query to broaden the scope of this review. For example, methods that consider overlapping groups[62] and approaches that model the functional form of predictors using group variable selection methods[63] are missing.

Further research on group variable selection methods would be desirable to follow up on the results of this article. The findings from this literature review revealed that the diverse possibilities of the two-step approaches fall short of their potential. There are a variety of group formation techniques,[64,65] each of which can be combined with one of the bi-level selection methods to circumvent the “all-in or all-out” limitation. Research on bi-level selection methods still offers many additional possibilities, even though most of the methods identified in this literature review belong to this category. The framework of additive combination of penalty terms, as with SGL, offers promising but as yet unexplored possibilities. For example, the additive combination of G-MCP and MCP has already been considered in the literature,[20,66] but a suitable implementation is still missing.

In summary, this review identified and compared a series of different methods for selecting related variables. Different methods are appropriate depending on prior knowledge and research question, which has been highlighted here in order to provide guidance regarding application of these methods in practice.

ACKNOWLEDGEMENTS

This work is part of Gregor Buch’s dissertation 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]) and was funded by the Ministry of Education and Research (BMBF 01EO1003 and 01EO1503). Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

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 from the corresponding author upon reasonable request.

ORCID

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

REFERENCES

1. Desboulets LDD. A review on variable selection in regression analysis. *Econometrics*. 2018;6(4):45.

2. Heinze G, Wallisch C, Dunkler D. Variable selection—a review and recommendations for the practicing statistician. *Biom J*. 2018;60(3):431-449.
3. Tian S, Wang C, Wang B. Incorporating pathway information into feature selection towards better performed gene signatures. *Biomed Res Int*. 2019;2019:2497509.
4. Bakin S. *Adaptive Regression and Model Selection in Data Mining Problems*. Canberra: Australian National University; 1999.
5. Zaharieva M, Breiteneder C, Hudec M. Unsupervised group feature selection for media classification. *Int J Multimed Inf Retr*. 2017;6(3):233-249.
6. Lin H, Wang C, Liu P, Holtkamp D. Construction of disease risk scoring systems using logistic group lasso: application to porcine reproductive and respiratory syndrome survey data. *J Appl Stat*. 2013;40(4):736-746.
7. Masoud HI, Reed MP, Paynabar K, et al. Predicting subjective responses from human motion: application to vehicle ingress assessment. *J Manuf Sci Eng Trans ASME*. 2016;138(6):061001.
8. Tutz G, Pößnecker W, Uhlmann L. Variable selection in general multinomial logit models. *Comput Stat Data Anal*. 2015;82:207-222.
9. Pérez-Espinosa H, Reyes-García CA, Villaseñor-Pineda L. Acoustic feature selection and classification of emotions in speech using a 3D continuous emotion model. *Biomed Signal Process Contr*. 2012;7(1):79-87.
10. Bozkurt Gönen G, Gönen M, Gürgen F. Probabilistic and discriminative group-wise feature selection methods for credit risk analysis. *Expert Syst Appl*. 2012;39(14):11709-11717.
11. Breheny P. The group exponential lasso for bi-level variable selection. *Biometrics*. 2015;71(3):731-740.
12. Simon N, Friedman J, Hastie T, Tibshirani R. A sparse-group lasso. *J Comput Graph Stat*. 2013;22(2):231-245.
13. Huang J, Ma S, Xie H, Zhang C-H. A group bridge approach for variable selection. *Biometrika*. 2009;96(2):339-355.
14. Zeng L, Xie J. Group variable selection methods and their applications in analysis of genomic data. *J Stat Comput Simul*. 2012;82(1):95-106.
15. Breheny P, Huang J. Penalized methods for bi-level variable selection. *Stat Interf*. 2009;2(3):369-380.
16. Doncheva NT, Morris JH, Gorodkin J, Jensen LJ. Cytoscape StringApp: network analysis and visualization of proteomics data. *J Proteome Res*. 2018;18(2):623-632.
17. Swan AL, Mobasheri A, Allaway D, Liddell S, Bacardit J. Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology. *Omics J Integr Biol*. 2013;17(12):595-610.
18. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med*. 2009;6(7):e1000097.
19. Morris TP, White IR, Crowther M. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38(11):2074-2102.
20. Huang J, Breheny P, Ma S. A selective review of group selection in high-dimensional models. *Stat Sci Rev J Inst Math Stat*. 2012;27(4):481-499.
21. Khaire UM, Dhanalakshmi R. Stability of feature selection algorithm: a review. *J King Saud Univ Comput Inf Sci*. 2019;34(4):1060-1073.
22. Hesterberg T, Choi NH, Meier L, Fraley C. Least angle and $\ell 1$ penalized regression: a review. *Stat Surv*. 2008;2:61-93.
23. He Z, Yu W. Stable feature selection for biomarker discovery. *Comput Biol Chem*. 2010;34(4):215-225.
24. Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. *Stat Sin*. 2010;20(1):101-148.
25. AlNuaimi N, Masud MM, Serhani MA, Zaki N. Streaming feature selection algorithms for big data: a survey. *Appl Comput Inform*. 2022;18(1/2):113-135.
26. Xie J, Zeng L. Group variable selection methods and their applications in analysis of genomic data. In: Feng J, Fu W, Sun F, eds. *Frontiers in Computational and Systems Biology*. London: Springer; 2010:231-248.
27. Dormann CF, Elith J, Bacher S, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*. 2013;36(1):27-46.
28. Reid S, Tibshirani R. Sparse regression and marginal testing using cluster prototypes. *Biostatistics*. 2016;17(2):364-376.
29. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *J Royal Stat Soc Ser B*. 2006;68(1):49-67.
30. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*. 1970;12(1):55-67.
31. Breheny P. R Package “grpreg”; 2020.
32. Breheny P, Huang J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Stat Comput*. 2015;25(2):173-187.
33. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc*. 2006;101(476):1418-1429.
34. Belhechmi S, De Bin R, Rotolo F, Michiels S. Accounting for grouped predictor variables or pathways in high-dimensional penalized Cox regression models. *BMC Bioinform*. 2020;21(1):277.
35. Wei F, Zhu H. Group coordinate descent algorithms for nonconvex penalized regression. *Comput Stat Data Anal*. 2012;56(2):316-326.
36. Zhang C-H. Nearly unbiased variable selection under minimax concave penalty. *Ann Stat*. 2010;38(2):894-942.
37. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc*. 2001;96(456):1348-1360.
38. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Ann Stat*. 2004;32(2):407-499.
39. Alfons A. Package “robustHD”; 2016.
40. Alfons A, Croux C, Gelper S. Robust groupwise least angle regression. *Comput Stat Data Anal*. 2016;93:421-435.
41. Frank LE, Friedman JH. A statistical view of some chemometrics regression tools. *Technometrics*. 1993;35(2):109-135.
42. Zhou N, Zhu J. Group variable selection via a hierarchical lasso and its oracle property. *Stat Interf*. 2010;3(4):557-574.
43. Wu TT, Lange K. Coordinate descent algorithms for lasso penalized regression. *Ann Appl Stat*. 2008;2(1):224-244.
44. Friedman J, Hastie T, Tibshirani R. A note on the group lasso and a sparse group lasso. arXiv preprint arXiv:10010736; 2010.

45. Vaughan G, Aseltine R, Chen K, Yan J. Stagewise generalized estimating equations with grouped variables. *Biometrics*. 2017;73(4):1332-1342.
46. Vaughan G, Chen K, Yan J. R package "sgee"; 2018.
47. Kamkar I, Gupta SK, Phung D, Venkatesh S. Stabilizing l1-norm prediction models by supervised feature grouping. *J Biomed Inform*. 2016;59:149-168.
48. Rodriguez MZ, Comin CH, Casanova D, et al. Clustering algorithms: a comparative approach. *PLoS One*. 2019;14(1):e0210236.
49. Grimonprez Q, Blanck S, Celisse A, Marot G. MLGL: an R package implementing correlated variable selection by hierarchical clustering and group-lasso. Preprint; 2018.
50. Meinshausen N. Hierarchical testing of variable importance. *Biometrika*. 2008;95(2):265-278.
51. Lin Y-W, Xiao N, Wang L-L, Li C-Q, Xu Q-S. Ordered homogeneity pursuit lasso for group variable selection with applications to spectroscopic data. *Chemom Intel Lab Syst*. 2017;168:62-71.
52. Zhang R, Zhang F, Chen W, et al. A new strategy of least absolute shrinkage and selection operator coupled with sampling error profile analysis for wavelength selection. *Chemom Intel Lab Syst*. 2018;175:47-54.
53. Fisher WD. On grouping for maximum homogeneity. *J Am Stat Assoc*. 1958;53(284):789-798.
54. Kim Y, Kim SB. Collinear groupwise feature selection via discrete fusion group regression. *Pattern Recognit*. 2018;83:1-13.
55. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J Royal Stat Soc Ser B*. 2005;67(2):301-320.
56. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1-22.
57. Breheny P. Package "ncvreg"; 2020.
58. Bondell HD, Reich BJ. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*. 2008;64(1):115-123.
59. Bogdan M, Van Den Berg E, Sabatti C, Su W, Candès E. SLOPE—adaptive variable selection via convex optimization. *Ann Appl Stat*. 2015;9(3):1103-1140.
60. Buchka S, Hapfelmeier A, Gardner PP, Wilson R, Boulesteix A-L. On the optimistic performance evaluation of newly introduced bioinformatic methods. *Genome Biol*. 2021;22(1):152.
61. Zhou Y, Zhou B, Pache L, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun*. 2019;10(1):1523.
62. Jacob L, Obozinski G, Vert J-P. Group lasso with overlap and graph lasso. Paper presented at: Proceedings of the 26th Annual International Conference on Machine Learning; 2009.
63. Lin Y, Zhang HH. Component selection and smoothing in multivariate nonparametric regression. *Ann Stat*. 2006;34(5):2272-2297.
64. Park MY, Hastie T, Tibshirani R. Averaged gene expressions for regression. *Biostatistics*. 2007;8(2):212-227.
65. Loscalzo S, Yu L, Ding C. Consensus group stable feature selection. Paper presented at: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2009.
66. Liu J, Huang J, Ma S. Integrative analysis of multiple cancer genomic datasets under the heterogeneity model. *Stat Sci*. 2013;32(20):3509-3521.

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 PS. A systematic review and evaluation of statistical methods for group variable selection. *Statistics in Medicine*. 2023;42(3):331-352. doi: 10.1002/sim.9620