Electronic Journal of Statistics
Vol. 18 (2024) 4235–4291
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2296

Simultaneous factors selection and fusion of their levels in penalized logistic regression

Lea Kaufmann

Institute of Statistics, RWTH Aachen University, Germany,
e-mail: kaufmann@isw.rwth-aachen.de

Maria Kateri

Institute of Statistics, RWTH Aachen University, Germany,
e-mail: maria.kateri@rwth-aachen.de

Abstract: Nowadays, several data analysis problems are high-dimensional, requiring a complexity reduction for their modeling. Under the sparsity assumption, variable selection is feasible, removing the non-influential explanatory variables. When factors are present, with their levels being dummy coded, the number of parameters included in the model grows rapidly, leading to high-dimensional problems even in cases with moderate number of factors. This fact emphasizes the need for a drastical parameter reduction, not only through variable selection but also through fusion of levels of factors. The levels fused are those not differentiating significantly in terms of their influence on the response variable. Such fusions, beyond reducing the dimension of the model, propose scale adjustments for categorical predictors. In this work a new regularization technique is introduced, called L_0-fused group lasso (L_0-FGL) for binary logistic regression. It uses a group lasso penalty for factor selection and for the fusion part it applies a L_0 penalty on the differences among the levels' parameters of a categorical predictor. Using adaptive weights, the adaptive version of the L_0-FGL method is derived. Theoretical properties, such as existence, $\sqrt{n}$ consistency and oracle properties under certain conditions, are established. In addition, it is shown that even in the diverging case where the number of parameters p_n grows with the sample size n, $\sqrt{n}$ consistency and a consistency in variable selection result are achieved, as well as a respective result on asymptotic normality for an approximate L_0-FGL solution. Two computational methods, the penalized iteratively reweighted least squares (PIRLS) and a block coordinate descent (BCD) approach using quasi Newton, are developed and implemented. A simulation study supports the outstanding performance of L_0-FGL, especially in cases with a large number of factors. Finally, we apply our method on a real dataset corresponding to breast cancer recurrence events.

Keywords and phrases: High-dimensional statistics, lasso, group lasso, L_0 norm, L_1 norm, $\sqrt{n}$ consistency, PIRLS algorithm, block coordinate descent (BCD) method.

Received October 2023

.

Contents

1. Introduction

Regularization methods for generalized linear models (GLMs) have been in the center of interest for high-dimensional data analysis, especially in the last two decades. In this framework, factors deserve a special attention. Dimensionality becomes even higher in presence of factors with many levels, since such a factor, say $\mathcal{X}_j$ of $p_j + 1$ levels, brings p_j predictors in the model. Furthermore, factors allow-dimension reduction not only by eliminating non-significant predictors but also by fusing levels of a predictor that have the same influence on the response. Such fusions lead to sparser models strengthening simultaneously their interpretability and may propose scale adjustments for the categorical predictors. Procedures that allow factor selection and levels fusion at the same time are a powerful tool for meaningfully reducing the complexity of the model.

The most popular model selection and shrinkage estimation method for GLMs is the lasso that uses an L_1-type penalty and which was initially proposed for linear regression models ([33]). It is well-known that the lasso estimator is biased and its model selection can be inconsistent. An attractive alternative that enjoys selection consistency is the adaptive lasso, proposed by [42], which allows different shrinkage levels for different regression coefficients through the use of adaptive weights. Other methods leading to nearly unbiased estimators are the smoothly clipped absolute deviation (SCAD) penalty [6] and the minimax concave penalty (MCP) [40].

In a variable selection problem with categorical explanatory variables (i.e. factors), the method needs to be applicable factor-wise, i.e., to exclude or include in the model all levels of a factor. For this, a natural and suitable extension of the lasso is the group lasso, originally considered for linear regression ([15], [39]) and later adjusted for logistic regression ([22]). A review on group lasso is provided by [14]. The adaptive lasso has also been extended to the adaptive group lasso [34] while group SCAD [35] and group MCP [14] are the groupwise selection variants of the SCAD and MCP.

However, the above mentioned methods are not able to perform fusion among the levels of a categorical predictor. Such a fusion can be achieved in a penalized regression framework by applying the penalty on the differences of the parameters belonging to the same factor. For the L_1 penalty this was first considered by [1] in an ANOVA framework and by [9] for linear regression models. Since L_1-type penalties lead to biased estimates, penalties with adaptive weights could be considered. However, the performance of such adaptive methods depends on the quality of the adaptive weights used. This fact led [25] to consider the so called L_0 norm as penalty function on the parameter differences within a factor instead. The advantage of this L_0 based approach is that an L_0-type penalty just differentiates between an entry (hence a difference) being zero or nonzero and consequently does not depend on the absolute value of the coefficients' differences. A disadvantage of this approach is that the resulting optimization problem is non-convex and thus computationally more involved. Further, since the L_0 norm is not even continuous, it is difficult to investigate theoretical properties. In [25], the model was fitted with the penalized iteratively reweighted

least squares (PIRLS) algorithm while theoretical properties were not in the focus of the paper. This method performs indirectly factor selection as well, since a factor is excluded when all parameters corresponding to it are set equal to zero, i.e. to the value of the reference category. Levels fusion based on penalties imposed on the differences among the parameters of a factor, inducing also factor selection, has been considered by [29] as well. They intoduced the so called SCOPE methodology, which uses a non-convex penalty, the MCP.

Nevertheless, it is not clear whether such indirect factor selection procedures based on the differences of coefficients from the reference category perform well enough, comparable to a group variable selection approach. As the group lasso penalty is a natural choice for factor selection while the above described L_0 based approach is a convenient choice for levels fusion, this work introduces a new regularization technique, called L_0-fused group lasso (L_0-FGL), that combines these two penalties for capturing the two different sources of sparsity, namely variable selection (also called factor selection in the framework of factors) and fusion of levels for categorical predictors. The use of two penalty functions allows to set the focus on either factor selection or levels fusion, depending on the application context. If no focus is set, L_0-FGL will balance between factor selection and levels fusion performance. Here, L_0-FGL is developed and studied in the framework of penalized logistic regression with all explanatory variables being categorical. The method is adjustable to other types of GLMs and cases of co-existence of continuous and categorical explanatory variables. We will verify that in our setting the additional group lasso penalty consideration improves the selection performance compared to the approach based solely on the L_0 penalties on the differences, which justifies the consideration of an additional penalty term to enforce a stronger factor selection performance.

The rest of the paper is organized as follows. After introducing the new L_0-FGL method along with its adaptive variant and pointing out its main characteristics in Section 2, the theoretical properties of L_0-FGL and adaptive L_0-FGL are investigated in Section 3. In particular, the existence, a result on $\sqrt{n}$-consistency as well as a result on an asymptotic normality property, are proved. Also a result about consistency in variable selection is provided. All properties in Section 3 are considered (i) for fixed number of parameters, and (ii) for number of parameters growing with the sample size. The algorithms used for obtaining the L_0-FGL estimates are discussed in Section 4, where also coefficient paths for different computational methods are analyzed. The computational approaches of Section 4 are compared in Section 5 in terms of simulation studies and appropriate goodness of fit measures. A high-dimensional design is also included in the simulation studies, which underlines the outstanding performance of the new proposed approach. We close our work by applying L_0-FGL to a real dataset corresponding to breast cancer recurrence events in Section 6.

2. The L_0-fused group lasso for logistic regression

Consider a binary response variable Y and $J \in \mathbb{N}$ fixed candidate categorical explanatory variables, i.e. factors, denoted by $\mathcal{X}_1, \ldots, \mathcal{X}_J$. Y and $\mathcal{X}_1, \ldots, \mathcal{X}_J$ are

observed on a sample of size $n \in \mathbb{N}$. In general, some of the explanatory variables could also be continuous but since, besides variable selection, our goal is to perform fusion within the levels of each factor, here we will focus on factors and neglect the co-existence of continuous explanatory variables. The expansion of the setup and the results for this case is straightforward. Factor $\mathcal{X}_j$, $j \in \{1, \ldots, J\}$ has $p_j + 1$ levels, coded by $0, \ldots, p_j$, where 0 is chosen to be the reference category. This results in p_j dummy variables $\mathcal{X}_{j,k}$, $k \in \{1, \ldots, p_j\}$, taking values in $\{0, 1\}$, for each factor. It holds $\mathcal{X}_{j,+} = \sum_k \mathcal{X}_{j,k} = 1$, if for an item of the sample the level of $\mathcal{X}_j$ is in $\{1, \ldots, p_j\}$, or $\mathcal{X}_{j,+} = 0$, if its level is the reference category 0. Consequently, our resulting parameter vector is $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J)^T \in \mathbb{R}^{p+1}$, where $p := \sum_{j=1}^{J} p_j$, β_0 denotes the intercept and $\boldsymbol{\beta}_j = (\beta_{j,1}, \ldots, \beta_{j,p_j})$, $j \in \{1, \ldots, J\}$, is the parameter subvector corresponding to the j-th factor. The fixed design matrix is denoted by $\boldsymbol{X} \in \mathbb{R}^{n \times (p+1)}$ and is decomposed in sub-matrices, i.e, $\boldsymbol{X} = (\mathbf{1}^T; \boldsymbol{X}_1; \ldots; \boldsymbol{X}_J)$, where $\mathbf{1}$ is a $n \times 1$ column vector of ones, while $\boldsymbol{X}_j \in \mathbb{R}^{n \times p_j}$ is the sub-matrix corresponding to factor $\mathcal{X}_j$ and containing the associated dummy variable values for the n items of our sample. With $\boldsymbol{x}_i$, $i \in \{1, \ldots, n\}$, we denote the i-th row of the design matrix $\boldsymbol{X}$, hence the i-th observation of the factors $j = 1, \ldots, J$, expressed through the corresponding dummy variables. Defining $\mathcal{X} = (1, \mathcal{X}_{1,1}, \ldots, \mathcal{X}_{1,p_1}, \ldots, \mathcal{X}_{J,1}, \ldots, \mathcal{X}_{J,p_J})$ as the vector of all dummy variables corresponding to an item of the sample (the first entry being one refers to the intercept), we denote a realization of it as $\boldsymbol{x} \in \mathbb{R}^{p+1}$. Consequently, the rows $\boldsymbol{x}_i$ of the design matrix $\boldsymbol{X}$ arise as fixed realizations of $\mathcal{X}$. Using a generalized linear model (GLM) considering the canonical link function, the logistic regression model is given by

$$\mathbb{E}(Y|\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}\boldsymbol{\beta})}. \tag{1}$$

Throughout the whole work, as mentioned above, we consider all candidate categorical explanatory variables $\mathcal{X}_1, \ldots, \mathcal{X}_J$ as fixed. For an elaboration on the differences of considering a fixed or a random design, especially in terms of model misspecification, we refer to [4].

2.1. L_0-fused group lasso

In penalized regression, one minimizes the sum of the log-likelihood and an appropriate penalty function to obtain the resulting estimates. In particular,

$$M_{pen}(\boldsymbol{\beta}) := -L_n(\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta}) \tag{2}$$

is minimized, where $L_n(\boldsymbol{\beta})$ denotes the log-likelihood function and $P_\lambda(\boldsymbol{\beta})$ the penalty function of the chosen method depending on some tuning parameter $\lambda \geq 0$. The penalized regression estimator $\hat{\boldsymbol{\beta}}$ is then defined as a minimizer of $M_{pen}(\boldsymbol{\beta})$. We note that, depending on the chosen penalty function, a global minimizer of $M_{pen}(\boldsymbol{\beta})$ is not necessary uniquely determined. That is, if the objective function is not (strictly) convex, there may exist several local minimizers,

as for the well-established methods SCAD [6] and MCP [40], as well as the group variants group SCAD and group MCP [3]. This fact will also play a role in our setup (see Section 3). For the group lasso (see [39], [22]), $P_\lambda(\boldsymbol{\beta})$ and $M_{pen}(\boldsymbol{\beta})$ become

$$P_{\lambda_1}^{GL}(\boldsymbol{\beta}) := \lambda_1 \sum_{j=1}^{J} ||\boldsymbol{\beta}_j||_{\boldsymbol{K}_j}, \quad \text{and} \quad M_{pen}^{GL}(\boldsymbol{\beta}) := -L_n(\boldsymbol{\beta}) + P_{\lambda_1}^{GL}(\boldsymbol{\beta}),$$

respectively, with $\boldsymbol{K}_j$ for $j \in \{1, \dots, J\}$ being some positive definite and symmetric matrices and $\lambda_1 \geq 0$. Following [39], for some $\boldsymbol{\xi} \in \mathbb{R}^d$, $d \in \mathbb{N}$ and a positive definite and symmetric matrix $\boldsymbol{K} \in \mathbb{R}^{d \times d}$, the norm $||\boldsymbol{\xi}||_{\boldsymbol{K}}$ is defined as $||\boldsymbol{\xi}||_{\boldsymbol{K}} := (\boldsymbol{\xi}^T \boldsymbol{K} \boldsymbol{\xi})^{\frac{1}{2}}$.

The L_1 penalty applied on the differences among the parameters of a factor's levels (see [1], [9]) was initially referred to as CAS in [1]. Later, [25] considered the L_0 penalty for these differences. In a natural way, one can bring up the name CAS-L_0 for the corresponding L_0 penalty. In the sequel, we will refer to it simply as L_0, whenever needed for brevity of notation. In this case it holds for $\lambda_0 \geq 0$

$$P_{\lambda_0}^{L_0}(\boldsymbol{\beta}) = P_{\lambda_0}^{CAS-L_0}(\boldsymbol{\beta}) := \lambda_0 \sum_{j=1}^{J} \sum_{0 \leq r < s \leq p_j} w_0^{(j,rs)} ||\beta_{j,r} - \beta_{j,s}||_0,$$
$$M_{pen}^{L_0}(\boldsymbol{\beta}) = M_{pen}^{CAS-L_0}(\boldsymbol{\beta}) := -L_n(\boldsymbol{\beta}) + P_{\lambda_0}^{CAS-L_0}(\boldsymbol{\beta}).$$

To simultaneously perform factor selection and fusion of levels in case of factors, we propose the following penalty, called L_0-FGL. For $\lambda := (\lambda_0, \lambda_1) \in \mathbb{R}^{\geq 0} \times \mathbb{R}^{\geq 0}$, the L_0-FGL penalty function is given by

$$P_\lambda(\boldsymbol{\beta}) := \lambda_1 \sum_{j=1}^{J} ||\boldsymbol{\beta}_j||_{K_j} + \lambda_0 \sum_{j=1}^{J} \sum_{0 \leq r < s \leq p_j} w_0^{(j,rs)} ||\beta_{j,r} - \beta_{j,s}||_0, \tag{3}$$

which is an intersection between the well known group lasso and the L_0 fusion penalty (CAS-L_0). Thus, L_0-FGL is balancing factor selection and levels fusion. It confines from the L_0 fusion penalty [25] by adding a group lasso selection penalty and analogously from the group lasso for logistic regression [22] by adding an L_0 fusion penalty. In the sequel, we denote by $||\boldsymbol{t}||_2$ for some $\boldsymbol{t} \in \mathbb{R}^n$ the euclidean norm $||\boldsymbol{t}||_2 = \sqrt{\sum_{i=1}^n t_i^2}$ while sometimes we write $||\boldsymbol{t}|| = ||\boldsymbol{t}||_2$ for simplicity.

With regard to the choice of $\boldsymbol{K}_j$, we get with $\boldsymbol{K}_j := \tilde{w}_1^{(j)} \mathbf{Id}_{p_j \times p_j}$ for some weight $\tilde{w}_1^{(j)}$ and $w_1^{(j)} := \sqrt{\tilde{w}_1^{(j)}}$

$$P_\lambda(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{J} w_1^{(j)} ||\boldsymbol{\beta}_j||_2 + \lambda_0 \sum_{j=1}^{J} \sum_{0 \leq r < s \leq p_j} w_0^{(j,rs)} ||\beta_{j,r} - \beta_{j,s}||_0. \tag{4}$$

In particular, we set $w_1^{(j)} = \sqrt{p_j}$ using the convenient choice $\boldsymbol{K}_j = p_j \mathbf{Id}_{p_j \times p_j}$ (hence $\tilde{w}_1^{(j)} = p_j$). The use of adaptive weights leads us to the so called *adaptive L_0-FGL*, analogously to the adaptive group lasso. In the following, we will

use the latter choice of $\boldsymbol{K}_j$ and investigate theoretical properties both for the L_0-FGL and its adaptive version. The weights $w_0^{(j,rs)}$, $j \in \{1,\dots J\}, r,s \in \{1,\dots p_j\}, s \neq r$, of the L_0 part, as well as the particular choice of adaptive weights mentioned above, will be specified in Section 5.1.

Recall that the L_0 penalty term of the L_0-FGL method includes in the sum differences from the reference category $\beta_{j,0} = 0$, enforcing thus also factor selection. In this setting, factors selection refers to the case when all categories are fused with the reference category. Nevertheless, it is not clear whether this *indirect* factor selection is effective enough, fact that brought us to the idea of adding a group lasso part for improving factor selection. The L_0-FGL estimate $\hat{\boldsymbol{\beta}}$ is defined as (local) minimizer of

$$\begin{aligned} M_{pen}(\boldsymbol{\beta}) &:= -L_n(\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta}) \\ &= -L_n(\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{J} w_1^{(j)} ||\boldsymbol{\beta}_j||_2 + \lambda_0 \sum_{j=1}^{J} \sum_{0 \le r < s \le p_j} w_0^{(j,rs)} ||\beta_{j,r} - \beta_{j,s}||_0. \end{aligned} \tag{5}$$

Note that, even though $\hat{\boldsymbol{\beta}}$ depends on the sample size $n \in \mathbb{N}$, in general, we neglect to use a lower sub-index n in $\hat{\boldsymbol{\beta}}_n$ for simplicity. For the tuning parameters λ_0 and λ_1, we sometimes write λ_0^n and λ_1^n, respectively, to express the dependence on the sample size, which plays a significant role discussing asymptotic properties. As already mentioned, the expressions $w_1^{(j)}$ for the group lasso part and $w_0^{(j,rs)}$ for the L_0 fusion part are optional weights that allow to put the factors and their levels on a comparable scale. Using $w_1^{(j)} = \sqrt{p_j}$ in the group lasso part accounts for the fact that the factors may have a different number of levels.

Figure 1 shows the value of $||\boldsymbol{\beta}||_2$, corresponding to the penalty function for group lasso (left), the L_0 norm on the differences $||\beta_1 - \beta_2||_0$, corresponding to the L_0 penalty (middle), and the sum $||\boldsymbol{\beta}||_2 + ||\beta_1 - \beta_2||_0$, corresponding to L_0-FGL (right) for a factor of three levels, i.e. $p = 2$ and $\boldsymbol{\beta} = (\beta_1, \beta_2)$ (without intercept). It gets clear that L_0-FGL combines both shrinkage and fusion of levels in one penalty. Further, by adjusting the tuning parameters λ_0 and λ_1 we can put the focus on selection or fusion, depending on the application context.

Remark 2.1 (On tuning)**.** Since L_0-FGL has two tuning parameters, a stepwise procedure is proposed for their tuning, according to which

(1) optimal λ_1^* is determined first with cross-validation (CV) setting $\lambda_0 = 0$,
(2) for fixed $\lambda_1 = \lambda_1^*$, optimal λ_0^* is determined with CV.

This tuning approach is referred to as stepwise. We further consider an iterative tuning approach for L_0-FGL, performing first the procedure described above, while after (2), we fix $\lambda_0 = \lambda_0^*$ and, based on this fixed tuning parameter for the L_0 part, we determine the optimal λ_1^*. In the sequel, an analogue procedure is followed for λ_0, fixing $\lambda_1 = \lambda_1^*$. We iterate this procedure until a pre-specified number of iterations is reached, or the improvement in predictive deviance does not exceed a pre-specified tolerance. The use of the iterative tuning approach

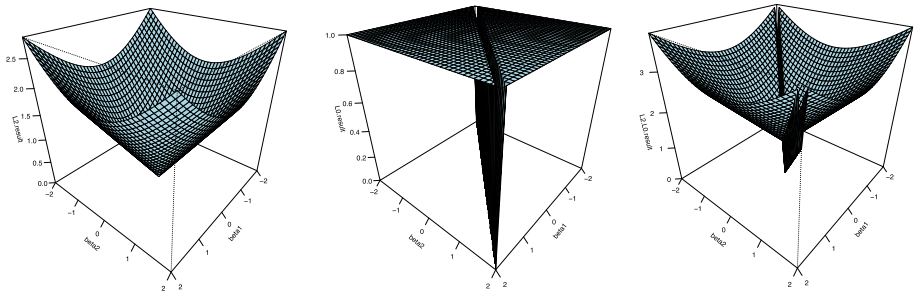

FIG 1. *Visualization of* $||\boldsymbol{\beta}||_2$ *(left, corresponding to group lasso),* L_0 *on the differences* $||\beta_1 - \beta_2||_0$ *(middle, corresponding to* L_0 *penalty) and the sum* $||\boldsymbol{\beta}||_2 + ||\beta_1 - \beta_2||_0$ *(right, corresponding to* L_0*-FGL), where* $\beta_1, \beta_2 \in [-2, 2]$ *and* $\lambda_0 = \lambda_1 = 1$.

for L_0-FGL in Section 5, is indicated by adding `iterative` at the name of the corresponding L_0-FGL procedure. Otherwise, the first described stepwise procedure is adopted. However, in both cases (iterative and non-iterative), the group lasso part (tuning of λ_1) is optimized first, since once a factor is excluded from the model, it has not to be investigated for fusion of categories.

3. Existence and theoretical properties of L_0-fused group lasso

Next, the existence and theoretical properties of L_0-FGL are investigated, including $\sqrt{n}$ consistency and asymptotic normality. Furthermore, consistency in variable selection is analyzed. The case of fixed p and that of diverging number of parameters, hence J_n and consequently p_n depending on the sample size n will be considered.

The next Theorem states the existence of L_0-FGL. The proof is provided for $p \leq n$ but Remark 3.2 in the sequel argues that the existence is also ensured in a high-dimensional setup with $p > n$. Notice that, in proving the existence, p is always considered fixed since this is not an asymptotic property.

Theorem 3.1 (Existence of L_0-FGL)**.** *Let* $\lambda_1 > 0$, $\lambda_0 \geq 0$ *and* $0 < \sum_{i=1}^{n} y_i < n$, *where* $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ *with* $y_i \in \{0, 1\}$, $i \in \{1 \ldots, n\}$ *is the vector of observed binary responses. Then, the set of (local) minimizers*

$$S := \left\{ \hat{\boldsymbol{\beta}} \,|\, \hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\arg\min} \left(-L_n(\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{J} ||\boldsymbol{\beta}_j||_{K_j} + \lambda_0 \sum_{j=1}^{J} \sum_{0 \leq r < s \leq p_j} w_0^{(j,rs)} ||\beta_{j,r} - \beta_{j,s}||_0 \right) \right\}$$

is nonempty. Moreover, the value of the objective function $M_{pen}(\cdot)$ *decreases if coefficients that are close enough to each other are fused.*

Proof. See Appendix A.1. □

Remark 3.2 (Existence in high-dimensional case $p > n$)**.** For the L_0-FGL estimator with $\lambda_1 > 0$ and $\lambda_0 \geq 0$, the proof above is not restricted to $p \leq n$, hence existence can be ensured in the high-dimensional case $p > n$. For $\lambda_1 = 0$, or both $\lambda_1 = \lambda_0 = 0$, the existence refers to existence of CAS-L_0 or of the maximum likelihood estimator (MLE), respectively, but this is not our focus here since we consider L_0-FGL.

For investigating the theoretical properties, some regularity conditions are required, which are provided next. We differentiate the cases of p being fixed and p_n allowed to grow with the sample size n. Further, let $\boldsymbol{\beta}^*$ be the true unknown parameter value.

3.1. Regularity conditions (fixed case)

(Reg1) The distribution of the response variable Y belongs to the exponential dispersion family, i.e., its probability density function (pdf) can be written as

$$f(\boldsymbol{v}, \theta, \phi) = \exp\left(\frac{y\theta - \varphi(\theta)}{a(\phi)} + c(y, \phi)\right),$$

for an observation $\boldsymbol{v} = (y, \boldsymbol{x}) \in \mathbb{R}^{p+2}$ and with $\theta = \theta(\boldsymbol{x}, \boldsymbol{\beta})$ for a given parameter vector $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$. For logistic regression $Y \sim \text{bin}(1, \pi)$, the natural parameter θ is the logit, hence $\theta = \log(\pi/(1-\pi))$, which equals the linear predictor $\boldsymbol{x}\boldsymbol{\beta}$ which we also denote by η, i.e. $\eta := \boldsymbol{x}\boldsymbol{\beta} = \theta$. Further, $\varphi(\eta) = \varphi(\boldsymbol{x}\boldsymbol{\beta}) = \log(1 + \exp(\boldsymbol{x}\boldsymbol{\beta}))$, $a(\phi) = 1$ and $c(y, \phi) = 1$ since $y \in \{0, 1\}$.

(Reg2) The Fisher information matrix $\boldsymbol{I}_F(\boldsymbol{\beta}) = \mathbb{E}\left(-\frac{\partial^2 L_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}\right)$ is finite and positive definite in $\boldsymbol{\beta} = \boldsymbol{\beta}^*$.

(Reg3) There exists an open set $\mathcal{O} \subseteq \mathbb{R}^{p+1}$ with $\boldsymbol{\beta}^* \in \mathcal{O}$ such that for all $\boldsymbol{\beta} \in \mathcal{O}$ and observations $\boldsymbol{v}_i = (y, \boldsymbol{x}_i)$, $i = 1, \ldots, n$, there exists a function $M(\boldsymbol{v}) \in \mathbb{R}$ such that the following holds

$$\frac{\partial^3 \log(f(\boldsymbol{v}_i, \boldsymbol{\beta}))}{\partial\beta_j \partial\beta_k \partial\beta_l} \leq M(\boldsymbol{v}_i) < \infty,$$
$$\mathbb{E}(M(\boldsymbol{v}_i)) < \infty.$$

These regularity conditions are similar to several other approaches, such as [6] and [42] (appendix), being necessary for technical derivations. Further, as mentioned in the appendix of [6], they ensure the asymptotic normality of the unpenalized MLE.

Remark 3.3 (On the regularity conditions under the canonical link)**.** Under the use of the canonical link function, the natural parameter θ satisfies $\theta = \boldsymbol{x}\boldsymbol{\beta} = \eta$ (see (Reg1)). In this case we get the following simplifications.

1. The expected and observed Fisher information matrices coincide, such that we deduce with (Reg1) and $a(\phi) = 1$

$$\begin{aligned} \boldsymbol{I}_F(\boldsymbol{\beta}) &= \mathbb{E}\left(-\frac{\partial^2 L_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}\right) = -\frac{\partial^2 L_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} \\ &= \boldsymbol{X}^T \operatorname{diag}\left(\varphi''(\boldsymbol{x}_1\boldsymbol{\beta}), \ldots, \varphi''(\boldsymbol{x}_n\boldsymbol{\beta})\right) \boldsymbol{X}, \end{aligned} \tag{6}$$

   where we recall that $\boldsymbol{x}_i$, $i \in \{1, \ldots, n\}$ are the rows of the design matrix $\boldsymbol{X}$. Consequently, (Reg2) is satisfied if (6) is finite and positive definite in $\boldsymbol{\beta} = \boldsymbol{\beta}^*$.
   Nevertheless, whenever possible, we will continue working with the expected Fisher information in favor of generality and for applicability with other link functions. However, we emphasize that this work deals solely with logistic regression.
2. The second and third derivatives of the log-likelihood function no further depend on y. Consequently, (Reg3) is ensured if there exists a function $\mathfrak{M}(\boldsymbol{x}_i)$ such that

$$\begin{aligned} &|\varphi'''(\boldsymbol{x}_i\boldsymbol{\beta})| \leq \mathfrak{M}(\boldsymbol{x}_i) < \infty, \\ &\mathbb{E}(\mathfrak{M}(\boldsymbol{x}_i)|x_j x_k x_l|) < \infty \ \ \forall 1 \leq j, k, l \leq p. \end{aligned}$$

   The function $\mathfrak{M}(\cdot)$ replaces $M(\cdot)$ of (Reg3) that depends on $\boldsymbol{y}$, nevertheless, for simplicity of notation, we avoid using $\mathfrak{M}$ in the proofs and keep the $M(\cdot)$. The actual structure of the function plays no role, it is sufficient to ensure that there exists some function satisfying the above inequalities.

These simplifications for canonical link functions can also be found in [6] (Section 3.2).

3.2. Regularity conditions (diverging case)

For the diverging case, assume the following regularity conditions. Since, in this case, the dimension of the parameter space p_n is allowed to grow with the sample size n, the dimension of the Fisher information matrix will also depend on n, as well as the dimension of the open set $\mathcal{O}$ appearing in (div.Reg3). Clearly, the dimension of the truth $\boldsymbol{\beta}^*$ is also depending on n in the case of diverging p_n. Nevertheless, to keep notation clear, we will not use a lower index n for the truth $\boldsymbol{\beta}^*$. For all of the constants in the following regularity conditions we write C, even if they are not required to be the same.

(div.Reg1) The distributional assumption for Y is the same as in (Reg1) (3.1) with p being replaced by p_n and the corresponding pdf denoted by f_n.

(div.Reg2) The Fisher information matrix $\boldsymbol{I}_{F,n}(\boldsymbol{\beta})$ satisfies the same as in (Reg2) (3.1) with $\boldsymbol{I}_F$ being replaced by $\boldsymbol{I}_{F,n}$.

(div.Reg3) There exists an open set $\mathcal{O}_n \subseteq \mathbb{R}^{p_n+1}$ with $\boldsymbol{\beta}^* \in \mathcal{O}_n$ such that for all $\boldsymbol{\beta} \in \mathcal{O}_n$ and observations $\boldsymbol{v}_i, i = 1, \ldots, n$ there exist a function $M_{n,j,k,l}(\boldsymbol{v}) \in \mathbb{R}$ for which it holds

$$\frac{\partial^3 \log(f_n(\boldsymbol{v}_i, \boldsymbol{\beta}))}{\partial\beta_j \partial\beta_k \partial\beta_l} \leq M_{n,j,k,l}(\boldsymbol{v}_i) \ \forall \boldsymbol{\beta} \in \mathcal{O}_n \text{ and } \forall\, j,k,l = 1, \ldots, p_n$$

Additionally, we assume that for some constant $C < \infty$ it holds

$$\mathbb{E}(M_{n,j,k,l}(\boldsymbol{v}_i)) < C < \infty \ \forall\, j,k,l = 1, \ldots, p_n.$$

(div.Reg4) There exists a constant $C < \infty$ such that

$$\max_{1 \leq i \leq n} \max_{1 \leq j \leq J_n} ||\boldsymbol{x}_{j,i}||_2 \leq C,$$

where we recall that the sub-matrix $\boldsymbol{X}_j \in \mathbb{R}^{n \times p_j}$ of the design matrix $\boldsymbol{X}$ contains all n samples of the dummy variables corresponding to factor $\mathcal{X}_j$, so $\boldsymbol{x}_{j,i} \in \mathbb{R}^{p_j}$ denotes the i-th row of $\boldsymbol{X}_j$.

These regularity conditions are similar to [7] and [36] being necessary for technical reasons. For simplifications using the canonical link, as we do here considering logistic regression, we refer to Remark 3.3 which similarly applies in the diverging case.

Remark 3.4. The above regularity conditions and their consequences are discussed next.

1. Alternatively to (div.Reg2) we could have also assumed that all the eigenvalues of the Fisher information matrix are finite and strictly positive which ensures the positive definite property, see [7].
2. The fact that we assumed in (div.Reg2), and similarly in (Reg2), that the Fisher information matrix is finite means in particular that we have for a constant $C > 0$: $[\boldsymbol{I}_{F,n}(\boldsymbol{\beta})]^2_{j,k} < C < \infty \ \forall\, j,k = 1, \ldots, p_n$ and
$$[\boldsymbol{I}_{F,n}(\boldsymbol{\beta})]_{j,k} = \mathbb{E}\left(-\frac{\partial^2 \log(f_n(\boldsymbol{v}_1, \boldsymbol{\beta}))}{\partial\beta_j \partial\beta_k}\right) < C.$$
3. Condition (div.Reg4) is a technical condition needed for the oracle property of the (approximate) L_0-FGL which will be shown in Theorem 3.11. This condition is introduced in [36]. Since we are using a dummy coding scheme, each row $\boldsymbol{x}_{j,i}, j \in \{1, \ldots, J_n\}$, $i \in \{1, \ldots, n\}$ of each sub-matrix $\boldsymbol{X}_j$ consists of one entry being equal to one, where the others are zero. Hence, we have $||\boldsymbol{x}_{j,i}||_2 = 1$ independent of $j \in \{1, \ldots, J_n\}$, $i \in \{1, \ldots, n\}$, so in our setting observing factors in a dummy coding scheme, (div.Reg4) is not a restriction.

3.3. Asymptotic results

Having discussed the regularity conditions for both cases of p being fixed and p_n being allowed to diverge with n, we can state and proof asymptotic re-

sults. For $\sqrt{n}$ consistency (Theorems 3.5, 3.6) and selection consistency (Theorems 3.12, 3.13) notice that the results refer to a local minimizer of $M_{pen}(\cdot)$ which is the same for both consistencies ($\sqrt{n}$ and selection).
We start with a $\sqrt{n}$ consistency result for fixed p. Assuming that the amount of penalization for both, factor selection and levels fusion (expressed with the convergence properties of a_n^1 and a_n^0, respectively) can be controlled, i.e. $a_n^1/\sqrt{n} = o_p(1)$, $a_n^0 = O_p(1)$ (see below), we achieve that there exists a local minimizer of the L_0-FGL objective function satisfying $\sqrt{n}$ consistency. For factor selection, the corresponding assumption is similar to [34] where the adaptive group Lasso is considered.

Theorem 3.5 ($\sqrt{n}$ consistency for fixed p)**.** *Let the regularity conditions (Reg1)–(Reg3) from Section 3.1 hold. Furthermore, assume that p is fixed. Set $a_n^1 := \max\{\lambda_1^n w_1^{(j)}\,;\, j \in \{1,\dots J\}\}$ and $a_n^0 := \max\{\lambda_0^n w_0^{(j,rs)}\,;\, 0 \leq r < s \leq p_j, j \in \{1,\dots J\}\}$ and assume $a_n^1/\sqrt{n} = o_p(1)$, $a_n^0 = O_p(1)$. Then, there exists a local minimizer $\hat{\boldsymbol{\beta}}$ of $M_{pen}(\boldsymbol{\beta})$ satisfying*

$$||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2 = O_p\left(\frac{1}{\sqrt{n}}\right).$$

Proof. See Appendix A.1. □

Next, the Theorem above is extended to the case of non-fixed number of parameters, meaning that $J = J_n$ and $p = p_n$ depend on n. Consequently, in this case, the true active set depends on n as well and is given by $A_n^* := \{\boldsymbol{\beta}_j^* \in \{1,\dots J_n\} \,|\, ||\boldsymbol{\beta}_j^*|| \neq \mathbf{0}\}$. For proving Theorem 3.6, the regularity conditions of Section 3.1 need to be slightly modified and are provided in Section 3.2. The following Theorem shows $\sqrt{\frac{n}{p_n}}$ consistency in the case of a diverging number of parameters, or total number of levels, respectively. We further adjust the assumptions of Theorem 3.5 corresponding to the amount of penalization; in particular the number of factors (or levels, respectively) will be included in these assumptions. Remark 3.7, directly after Theorem 3.6, provides cases satisfying the assumptions, showing thus that they are not too restrictive. Lastly, we will require $p_n = o(n^{1/4})$ controlling the ratio of the sample size and the dimension of the parameter space, which is similar to [7].

Theorem 3.6 (Consistency in the diverging p_n case)**.** *Let the regularity conditions (div.Reg1)–(div.Reg3) of Section 3.2 hold. In addition, let a_n^1 and a_n^0 be defined analogously to Theorem 3.5. With $\alpha_n := \sqrt{\frac{p_n}{n}}$ we assume $\alpha_n a_n^1 J_n = o_p(1)$ and $a_n^0\, p_n(p_n - 1) = o_p(1)$. Lastly, we assume $p_n = o(n^{1/4})$. Then, there exists a local minimizer $\hat{\boldsymbol{\beta}}$ of $M_{pen}(\boldsymbol{\beta})$ satisfying*

$$||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2 = O_p(\alpha_n).$$

Proof. See Appendix A.1. □

Remark 3.7 (On the assumptions of Theorem 3.6)**.** The assumption $p_n = o(n^{1/4})$, hence $p_n^4/n \to 0$, implies $p_n^2/\sqrt{n} \to 0$. Consequently, the assumption

$\alpha_n a_n^1 J_n = o_p(1)$ holds for example if a_n^1 converges to some constant or is simply bounded, since $\alpha_n a_n^1 J_n = \sqrt{\frac{p_n}{n}} J_n a_n^1 \leq \sqrt{\frac{p_n}{n}} p_n a_n^1 = \frac{p_n^{3/2}}{\sqrt{n}} a_n^1 \leq \underbrace{\frac{p_n^2}{\sqrt{n}}}_{\to 0} a_n^1$. For the requirement that $a_n^0 p_n(p_n - 1) = o_p(1)$, we observe the case of weights chosen to be constant and equal to one for the L_0 part. Thus $a_n^0 = \lambda_0^n$ and choosing for example $\lambda_n^0 = o(1/p_n^2)$, $\lambda_0^n p_n(p_n - 1) = o_p(1)$ holds.

Having shown the consistency result for the cases of fixed and diverging number of parameters, oracle properties are investigated next. For this it is required that the true underlying model is sparse, as defined below.

Definition 3.8. In the case of fixed p, the true underlying structure is said to be sparse if, without loss of generality, the true active set $A^* := \{j \in \{1, \dots J\} \,|\, \boldsymbol{\beta}_j^* \neq \mathbf{0}\} = \{j \in \{1, \dots J\} \,|\, ||\boldsymbol{\beta}_j^*||_2 \neq 0\}$ can be written as $A^* = \{1, \dots, j_0\}$ with $j_0 < J$. In this case, the Fisher information matrix $\boldsymbol{I}_F(\boldsymbol{\beta}^*)$ is given in the following form

$$\boldsymbol{I}_F(\boldsymbol{\beta}^*) = \begin{bmatrix} \boldsymbol{I}_{11} & \boldsymbol{I}_{12} \\ \boldsymbol{I}_{21} & \boldsymbol{I}_{22} \end{bmatrix},$$

where $\boldsymbol{I}_{11} \in \mathbb{R}^{p_0 \times p_0}$ and $p_0 := \sum_{j=1}^{j_0} p_j$. In the case of diverging p_n, since the true active set may depend on the sample size, the underlying model is defined as sparse as above, but with A^*, J, j_0, $\boldsymbol{I}_F$, $\boldsymbol{I}_{11}$ and p_0 being replaced by A_n^*, J_n, $j_{0,n}$, $\boldsymbol{I}_{F,n}$, $\boldsymbol{I}_{11,n}$ and $p_{0,n}$.

The following Theorem states the asymptotic normality for L_0-FGL in the fixed p case. The requirements on the amount of penalization for the group lasso part are similar to those imposed by [42] for the tuning in adaptive lasso.

Theorem 3.9 (Existence of estimator satisfying the asymptotic normality property for the case fixed p)**.** *Assume that (Reg1)–(Reg3) of Section 3.1 hold and the true underlying structure is sparse (see Definition 3.8). For the group lasso part we choose the adaptive weights $w_1^{(j)} = ||\tilde{\boldsymbol{\beta}}_j||_2^{-\gamma}$ for some arbitrarily chosen $\gamma > 0$ where $\tilde{\boldsymbol{\beta}}$ is the unpenalized MLE. Furthermore, let $\lambda_1^n \cdot n^{-1/2} \to 0$ and $\lambda_1^n \cdot n^{(\gamma-1)/2} \to \infty$. For the tuning and weights of the L_0 part, we assume $a_n^0 = o_p(1)$. Then, there exists a local minimizer $\hat{\boldsymbol{\beta}}$ of $M_{pen}(\boldsymbol{\beta})$ satisfying*

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{A^*} - \boldsymbol{\beta}_{A^*}^*\right) \to_d N(0, \boldsymbol{I}_{11}^{-1}),$$

where $\hat{\boldsymbol{\beta}}_{A^*}$ and $\boldsymbol{\beta}_{A^*}^*$ denote the sub-vectors of $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$, respectively, containing only the components belonging in the true active set A^*.

Proof. See Appendix A.1. □

Remark 3.10 (Adaptive weights in the group lasso part)**.** In Theorem 3.9, we can also use any other initial estimator $\tilde{\boldsymbol{\beta}}$ (besides the MLE) satisfying consistency, as we will do in Theorem 3.11 below. Consequently, we could also

use an initial L_0-FGL estimator, setting $w_1^{(j)} = ||\tilde{\boldsymbol{\beta}}_j^{L_0-FGL}||_2^{-\gamma}$. Since L_0-FGL enforces factor selection and levels fusion, it may happen that $||\tilde{\boldsymbol{\beta}}_j^{L_0-FGL}||_2 = 0$. Thus, following [43] and [38], we set $w_1^{(j)} = \left(||\tilde{\boldsymbol{\beta}}_j^{L_0-FGL}||_2 + \frac{1}{n}\right)^{-\gamma}$. This will clearly not affect any shown asymptotic property.

For the extension of Theorem 3.9 to the diverging case, we need to approximate the L_0 part in L_0-FGL. In particular, we approximate as $||\xi||_0 \approx \frac{2}{1+\exp(-\gamma_0|\xi|)} - 1 =: N(\xi)$; see Section 4.1 for more details. The parameter $\gamma_0 > 0$ determines the steepness of the approximation and needs to be chosen, see [25]. With that, we define

$$\widetilde{M}_{pen}(\boldsymbol{\beta}) := -L_n(\boldsymbol{\beta}) + \lambda_1^n \sum_{j=1}^{J_n} w_1^{(j)} ||\boldsymbol{\beta}_j||_2 + \lambda_0^n \sum_{j=1}^{J_n} \sum_{0 \le r < s \le p_j} w_0^{(j,rs)} N(\beta_{j,r} - \beta_{j,s})$$

giving us an approximation of the L_0-FGL objective function, i.e. $M_{pen}(\boldsymbol{\beta}) \approx \widetilde{M}_{pen}(\boldsymbol{\beta})$. Theorem 3.11 on asymptotic normality in the diverging case shows that we can find a local minimizer of $\widetilde{M}_{pen}(\boldsymbol{\beta})$ (which we will call an approximate L_0-FGL solution) satisfying the asymptotic normality property. As mentioned above, the approximation will make it possible to obtain sub-differentials of the penalty function which we need for the proof. Using that $N(\xi) \le 1$, we can show consistency of an approximate L_0-FGL estimator under the same assumptions as in Theorem 3.6, since in step 2 of the proof of Theorem 3.6 we use the property of the L_0 norm that it is always less or equal to one, which holds in the same way for the approximation. To sum up, under the assumptions of Theorem 3.6, we know that there exists some local minimizer $\hat{\boldsymbol{\beta}}$ of $\widetilde{M}_{pen}(\boldsymbol{\beta})$ satisfying $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2 = O_p(\alpha_n)$. For this local minimizer, we will show the asymptotic normality property in the following Theorem 3.11. Since we will use the approximate L_0-FGL solution in this particular way only in Theorem, 3.11, we denote the approximate L_0-FGL by $\hat{\boldsymbol{\beta}}$ as well, avoiding the introduction of new notation. The idea of the following Theorem 3.11 originates from [36] who considered adaptive group lasso, while we adjust it for (approximate) L_0-FGL with two tuning parameters. Besides the regularity conditions (div.Reg1)–(div.Reg4) we need the assumptions of Theorem 3.6 to ensure that there exists an approximate L_0-FGL estimator $\hat{\boldsymbol{\beta}}$ satisfying consistency, i.e. $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2 = O_p(\alpha_n)$. Further, we use a condition controlling the size of the minimum of the true parameter, which corresponds to (A6) in [36] and (8) in [41]. Finally, the assumptions $\lambda_1^n n^{\frac{1}{2}(2-\delta+\frac{1}{4})} \to 0$ and $\lambda_1^n n^{1-\frac{1}{4}} = \lambda_1^n n^{\frac{3}{4}} \to \infty$ also correspond to the assumptions of Theorem 2.3 in [36].

Theorem 3.11 (Existence of an approximate L_0-FGL estimator satisfying the asymptotic normality property for the diverging p_n case)**.** *Assume that (div.Reg1)–(div.Reg4) and the assumptions of Theorem 3.6 hold. Let the true underlying structure be sparse (see Definition 3.8). With* $\beta^*_{min} := \min_{j=1,\dots j_{0,n}} ||\boldsymbol{\beta}_j^*||_2$ *we assume that there exists some* $\frac{3}{4} < \delta \le 1$ *and* $C > 0$ *such that* $n^{\frac{1}{2}(1-\delta)}\beta^*_{\min} \ge C$*. Additionally, we assume* $\lambda_n^1 n^{\frac{1}{2}(2-\delta+\frac{1}{4})} \to 0$ *and* $\lambda_n^1 n^{\frac{3}{4}} \to \infty$ *as* $n \to \infty$*. For*

the group lasso part we choose the adaptive weights $w_1^{(j)} = ||\tilde{\boldsymbol{\beta}}_j||_2^{-1}$, *where* $\tilde{\boldsymbol{\beta}}$ *is a* $\sqrt{p_n/n}$ *consistent initial estimator, hence* $||\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2 = O_p(\sqrt{p_n/n})$. *Finally, assume that* $\mathbb{E}(Y - \varphi'(\boldsymbol{x}_1\boldsymbol{\beta}^*))^4 < \infty$, *where* φ *is the cumulant function in the exponential dispersion family expression for the density function of* Y. *Then, there exists a local minimizer* $\hat{\boldsymbol{\beta}}$ *of* $\widetilde{M}_{pen}(\boldsymbol{\beta})$ *satisfying*

$$\boldsymbol{e}_n \boldsymbol{I}_{11,n}^{1/2} \left(\hat{\boldsymbol{\beta}}_{A_n^*} - \boldsymbol{\beta}_{A_n^*}^*\right) \to_d N(0,1),$$

where $\hat{\boldsymbol{\beta}}_{A_n^*}$ *and* $\boldsymbol{\beta}_{A_n^*}^*$ *denote the sub-vectors of* $\hat{\boldsymbol{\beta}}$ *and* $\boldsymbol{\beta}^*$, *respectively, containing only the components belonging in the true active set* A_n^*, *while* $\boldsymbol{e}_n$ *is a* $p_{0,n}$ *dimensional unit vector.*

Proof. See Appendix A.1. □

A consistency result concerning factor selection is discussed next. The desired method should, asymptotically, correctly detect the truly zero parameter vectors as well as the truly nonzero parameter vectors. The Theorem below is motivated by the work of [5] where the focus lies on L_1 and $L_1 + L_2$ penalization in linear and logistic regression, ignoring the presence of factors. Starting from the assumptions needed for the proof of $\sqrt{n}$-consistency of the estimator $\hat{\boldsymbol{\beta}}$, an asymptotic upper bound for the probability $\mathbb{P}(A^* \not\subseteq A_n)$ is derived, where A^* is the true active set in the case of fixed p (to be replaced by A_n^* for diverging p, see Defintion 3.8) and $A_n := \{j \in \{1, \ldots, J\} \,|\, ||\hat{\boldsymbol{\beta}}_j|| \neq 0\}$ is the active set of the estimate, depending on the sample size n (for diverging p, J has to be replaced by J_n). This is a result on the consistency of factor selection of our approach.

Theorem 3.12 (Selection consistency for fixed p)**.** *Assume that the conditions of Theorem 3.5 are satisfied and that the true underlying structure is sparse. Then, for the minimizer* $\hat{\boldsymbol{\beta}}$ *of Theorem 3.5 it holds that* $\forall \varepsilon > 0$ *one can find* $N \in \mathbb{N}$ *such that*

$$\mathbb{P}(A^* \not\subseteq A_n) < \varepsilon \quad \forall\, n \geq N. \tag{7}$$

Proof. See Appendix A.1. □

This result says that, depending on the sample size n, there exists an estimator for which the probability that it falsely sets factors to zero (meaning that we would delete influential factors from our model) can be made arbitrarily small. This is a property that is really useful in practice, especially for two-step procedures.

In the same way, an analogue result for selection consistency in case of a diverging number of parameters can be proved. An additional assumption controling the size of the minimum of the true parameter needs to be added.

Theorem 3.13 (Selection consistency in the diverging p_n case)**.** *Assume that the conditions of Theorem 3.6 are satisfied and that* $\forall n \in \mathbb{N}$ *it holds* $\min_{l \in A_n^*} ||\boldsymbol{\beta}_l^*|| \geq C$ *for some constant* $C > 0$. *Then, for the minimizer* $\hat{\boldsymbol{\beta}}$ *of Theorem 3.6 it holds that for* $\forall \varepsilon > 0$ *one can find* $N \in \mathbb{N}$ *such that*

$$\mathbb{P}(A_n^* \not\subseteq A_n) < \varepsilon \quad \forall\, n \geq N. \tag{8}$$

Proof. See Appendix A.1. □

4. Computational approaches

Two different computational approaches are considered, the penalized iteratively re-weighted least squares (PIRLS) algorithm and a block coordinate descent (BCD) procedure.

4.1. PIRLS algorithm

This approach is suitable for a broad variety of existing penalty functions as discussed in [26]. It is introduced as an applicable algorithm for combinations of different penalties. Approximating the penalties quadratically, as done for the non-convex SCAD in [6], the computation is executed by a PIRLS algorithm, as described in the following. In general, PIRLS can be applied to penalty functions of the following form

$$P_\lambda^{gen}(\boldsymbol{\beta}) = \sum_{l=1}^{L} \lambda_l p_l(||\boldsymbol{a}_l^T \boldsymbol{\beta}||_{N_l}). \tag{9}$$

Here, $L \in \mathbb{N}$ is the number of penalizations with corresponding tuning parameter $\lambda_l \geq 0, ||\cdot||_{N_l}$ is a semi-norm, or at least some term that makes sense to be used as a penalty. Further, $p_l : \mathbb{R}^+ \to \mathbb{R}^+$ are penalty functions where $p_l(0) = 0$ holds and p_l is continuously differentiable on $\mathbb{R}^+$ with positive derivative. Additionally, $p_l(\xi)$ is assumed to be strictly monotonic in ξ. The vectors $\boldsymbol{a}_l^T$ transform the coefficient vector $\boldsymbol{\beta}$, for example in the case of fusion penalties for ordinal factors, this vector will form the differences of adjacent coefficients (if we have nominal factors we will form all pairwise differences). Most of the time, as explained in [26], the penalties are of the form $P_\lambda^{gen}(\boldsymbol{\beta}) = \sum_{j=1}^{J} \sum_{l=1}^{L_j} \lambda_{jl} p_{jl}(||\boldsymbol{a}_{jl}^T \boldsymbol{\beta}_j||_{N_l})$, meaning that we penalize each factor $j \in \{1, \ldots, J\}$ separately. Keeping this in mind, we will continue to use the more compact way of writing (9) where then the quantity L combines the different number of penalizations and the fact that we may penalize each factor separately. Note that here, $p_l(\cdot)$ or $p_{jl}(\cdot)$ respectively are functions and do not denote the number of categories of factor j which we also denoted by p_j. Since L_0-FGL has two penalization functions (and two tuning parameters), the PIRLS algorithm will have two penalty terms. In particular, we make the following choices

$$\text{group lasso part: } p_l(\zeta) = \sqrt{p_l} \cdot \zeta, \ \boldsymbol{R}_l \boldsymbol{\beta} = \boldsymbol{\beta}_j, \tag{10}$$

$$L_0 \text{ part: } p_l(\zeta) = w_0^{(j,km)} \zeta, \ \boldsymbol{a}_l^T \boldsymbol{\beta} = \beta_{j,k} - \beta_{j,m}, \ 0 \leq k < m \leq p_j. \tag{11}$$

The vectors $\boldsymbol{a}_l^T$ are responsible for picking the possible differences corresponding to factor j. The entries of these vectors are contained in the set $\{-1, 0, 1\}$. An extension of this linear transformations $\boldsymbol{a}_l^T \boldsymbol{\beta}$ to vector valued arguments (needed for group lasso) leads to the corresponding transformation matrices $\boldsymbol{R}_l$. In our

particular application, this matrix $\boldsymbol{R}_l$ picks the right sub-vector $\boldsymbol{\beta}_j$ out of the full vector $\boldsymbol{\beta}$. We refer to [26] (Section 2.5) for the detailed steps and more information. The semi-norms $||\cdot||_{N_l}$ appearing in the penalty function P_λ^{gen} above are approximated by some suitable function $N_l(\cdot)$, in case they are not twice continuously differentiable. The derivative of $N_l(\cdot)$ is denoted by $D_l(\cdot)$. As mentioned above, [26] extended the algorithm to penalties with vector valued arguments, hence $N_l(\boldsymbol{\xi})$ depends on a vector $\boldsymbol{\xi} \in \mathbb{R}^p$ instead of $\xi \in \mathbb{R}$, like it is the case for the group lasso penalty which we need for L_0-FGL. In particular, for the approximations of the "norms", following [26], we have

$$L_0 \text{ norm:} \qquad ||\xi||_0 \approx N_l(\xi) = \frac{2}{1+\exp(-\gamma|\xi|)} - 1\,, \tag{12}$$

$$\text{group lasso } ||\boldsymbol{\xi}||_2\text{:} \qquad ||\boldsymbol{\xi}||_2 \approx N_l(\boldsymbol{\xi}) = (\boldsymbol{\xi}^T\boldsymbol{\xi} + c)^{1/2}. \tag{13}$$

Note that in the approximation $N_l(\xi)$ of the L_0 norm, we further use $|\xi| \approx \sqrt{\xi^2 + c}$ whenever we need to ensure differentiability besides continuity. This was not the case in the proof of Theorem 3.11, since there sub-differentiability of $N_l(\xi)$ given in (12) was sufficient. Since we apply the same penalty to each factor (or difference of levels, respectively) we can also write $N(\cdot)$ instead of $N_l(\cdot)$. The PIRLS algorithm is sketched in Algorithm 4.1, see [26]. For more computational details (as details on $\boldsymbol{A}_\lambda, \widetilde{\boldsymbol{W}}, \tilde{\boldsymbol{y}}$ appearing in the following algorithm) and convergence analysis we refer to Appendix A.2.

Algorithm 4.1 (PIRLS for L_0-FGL)**.**

1. Set start value $\hat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$, if not specified otherwise. Set $k = 1$
2. While $\dfrac{||\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}^{(k-1)}||_2}{||\hat{\boldsymbol{\beta}}^{(k)}||_2} > \varepsilon$ and $k \leq$ maxsteps
   - 2.1 Update approximation of M_{pen} including updates of $\boldsymbol{A}_\lambda, \widetilde{\boldsymbol{W}}, \tilde{\boldsymbol{y}}$ as they depend on the value of the coefficient of the current iteration $\hat{\boldsymbol{\beta}}^{(k)}$
   - 2.2 Set $\hat{\boldsymbol{\beta}}^{(k+1)} = (1-\nu)\hat{\boldsymbol{\beta}}^{(k)} + \nu\left(\boldsymbol{X}^T\widetilde{\boldsymbol{W}}^{(k)}\boldsymbol{X} + \boldsymbol{A}_\lambda\right)^{-1}\boldsymbol{X}^T\widetilde{\boldsymbol{W}}^{(k)}\tilde{\boldsymbol{y}}^{(k)}$
3. Finally, set $\hat{\boldsymbol{\beta}}^{L_0-FGL} = \hat{\boldsymbol{\beta}}^{(k+1)}$.

Coefficient Paths (PIRLS)

Next we will compare coefficient paths for CAS-L_0, group lasso and L_0-FGL, all computed with the use of the PIRLS algorithm. Consider $J = 2$ factors where $\mathcal{X}_1$ has 4 and $\mathcal{X}_2$ has 3 categories with equal probabilities. All the factors are sampled from a multinomial distribution. The true coefficient vector is chosen to be $\boldsymbol{\beta}^* = (2, 1.2, 1, 0.5, -0.8, -0.5)$. We used the `simulation` function from the package `gvcm.cat` to simulate our dataset. For the tuning parameters we made the following choices: for L_0-FGL we chose $\lambda_{max,1} = 10$ for the group lasso

part and $\lambda_{max,0} = 20$ for the L_0 part. Furthermore, for the CAS-L_0 estimator we used $\lambda_{max} = 10$ and for the group lasso estimator we used $\lambda_{max} = 15$. Because of the different maximum tuning values and the fact that L_0-FGL needs two tuning parameters, the points where the fusion/selection occurs are not comparable among the methods. Further, we chose the unpenalized MLE as starting values.

The resulting coefficient paths for the chosen approaches are provided in Figure 2. We can directly verify that L_0-FGL (middle) is an intersection between CAS-L_0 (right) and the group lasso (left). The huge advantage of L_0-FGL compared to the other two is that it combines the ability of factor selection and fusion of levels while the group lasso itself executes factor selection and CAS-L_0 fusion of levels. Even if the CAS-L_0 approach is able to select factors since we include reference category zero, we can not be sure of its factor selection performance, the corresponding paths are not smooth. Consequently, the approach of using L_0-FGL seems to be an advantageous tool that combines both worlds, finding a compromise between factor selection and levels fusion.

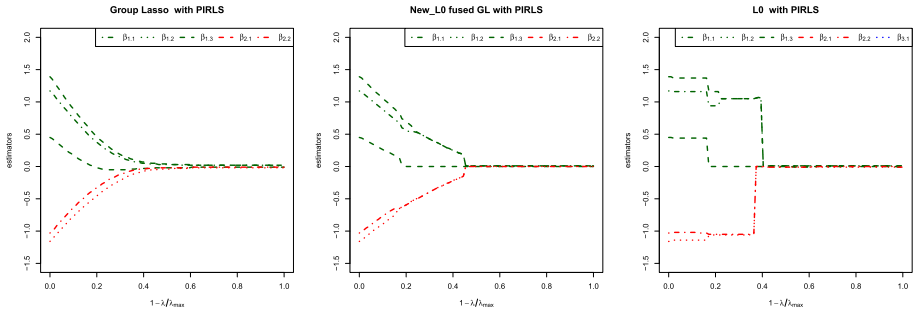

FIG 2. *Coefficient paths of two factors with* 4 *and* 3 *levels, respectively, for group lasso (left), L_0-FGL (middle) and CAS-L_0 (right). All methods are computed with PIRLS.*

4.2. Block coordinate descent

A block coordinate descent (BCD) approach with a quasi Newton step for obtaining the estimates is developed. The idea is to cycle through the factors, minimizing with respect to (wrt) one factor at a time while keeping the others fixed, as for example done in [22] and [3]. We start with an approximation of the objective function where the same function as in PIRLS is used for the L_0 part, whereas the group lasso part is added without approximating it. Details of the approximation of the first part can be found in Appendix A.3.1. As explained there, the approximation $g(\boldsymbol{\beta}_j, \hat{\boldsymbol{\beta}}^{(k)})$ is used for the log-likelihood and L_0 part of our penalty while the group lasso part is added afterwards. The resulting approximation of the L_0-FGL penalty function is denoted by

$$\tilde{g}(\boldsymbol{\beta}_j, \hat{\boldsymbol{\beta}}^{(k)}) := g(\boldsymbol{\beta}_j, \hat{\boldsymbol{\beta}}^{(k)}) + \lambda_1 \sqrt{p_j} ||\boldsymbol{\beta}_j||_2, \tag{14}$$

where $g(\boldsymbol{\beta}_j, \hat{\boldsymbol{\beta}}^{(k)})$ is given by (41). The function $\tilde{g}(\boldsymbol{\beta}_j, \hat{\boldsymbol{\beta}}^{(k)})$ is minimized wrt $\boldsymbol{\beta}_j$ while the remaining $\boldsymbol{\beta}_i$, $i \neq j$ are kept fixed. This works because of the separability property of the penalty function (in terms of factors), which ensures that building the derivative of $\tilde{g}$ wrt $\boldsymbol{\beta}_j$, $j \in \{1, \ldots, J\}$, makes the other terms depending on $\boldsymbol{\beta}_i$, $i \in \{1, \ldots, J\} \backslash \{j\}$, vanish. For more details on the separability property concerning BCD we refer to [13], Section 5.4.1. Note that, since we do not approximate the group lasso part, problems may occur with the derivative in zero. As in [22], a solution to that is to check in advance whether the minimum of the function to be minimized is at the non-differentiable point $\boldsymbol{\beta} = \mathbf{0}$. In the implementation of this algorithm in `R`, we employ the `optim` function that uses the `BFGS` (Broyden-Fletcher-Goldfarb-Shanno) method and this extra check is not required, as explained in Section A.3.2. The BCD quasi Newton algorithm for L_0-FGL is described in Algorithm 4.2 below.

Algorithm 4.2 (Block Coordinate Descent for L_0-FGL with quasi Newton)**.**

1. Set start value $\hat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$ if not specified otherwise. Set $k = 1$
2. While $\dfrac{||\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}^{(k-1)}||_2}{||\hat{\boldsymbol{\beta}}^{(k)}||_2} > \varepsilon$ and $k \leq$ maxsteps
   - 2.1 Update approximation of $M_{pen}^{CAS-L_0} \approx g$ including updates of $\boldsymbol{A}_\lambda, \widetilde{\boldsymbol{W}}$, $\tilde{\boldsymbol{y}}$ as they depend on the value of the coefficient of the current iteration $\hat{\boldsymbol{\beta}}^{(k)}$ which gives the approximation $g(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^{(k)})$.
     For $j = 1, \ldots, J$ execute the following:
     - 2.1a Set $\tilde{g}(\boldsymbol{\beta}_j, \hat{\boldsymbol{\beta}}^{(k)}) := g(\boldsymbol{\beta}_j, \hat{\boldsymbol{\beta}}^{(k)}) + \lambda_1 \sqrt{p_j} ||\boldsymbol{\beta}_j||_2$.
       Use quasi Newton to obtain $\hat{\boldsymbol{\beta}}_j^{(k+1)} = \arg\min_{\boldsymbol{\beta}_j} \tilde{g}(\boldsymbol{\beta}_j, \hat{\boldsymbol{\beta}}^{(k)})$
     - 2.1b Set $\hat{\boldsymbol{\beta}}^{(k+1)} = (\hat{\boldsymbol{\beta}}_1^{(k+1)}, \ldots, \hat{\boldsymbol{\beta}}_j^{(k+1)}, \hat{\boldsymbol{\beta}}_{j+1}^{(k)}, \ldots, \hat{\boldsymbol{\beta}}_J^{(k)})$

     Set $k = k + 1$
3. Finally, set $\hat{\boldsymbol{\beta}}^{L_0-FGL} = \hat{\boldsymbol{\beta}}^{(k+1)}$.

For the execution of the quasi Newton part of this algorithm in our applications we used the function `optim()` in `R`.

Coefficient paths (BCD)

Now we will show resulting coefficient paths for group lasso, CAS-L_0 and L_0-FGL where all are computed with the BCD quasi Newton procedure. Assume we have $J = 2$ factors with $p_1 = p_2 = 3$, hence 4 levels each, drawn from multinomial distribution with equal probabilities. The true parameter vector was chosen to be given by $\boldsymbol{\beta}^* = (-0.5, 2, -1, 2, -0.5, -1, 1)$. In Figure 3 the resulting coefficient paths are displayed. We can see that L_0-FGL connects the ability of group lasso to select variables and of CAS-L_0 to fuse coefficients when they are close enough to each other, analogously to the coefficient paths using PIRLS (Section 4.1). In fact we can see that using BCD and quasi Newton, the

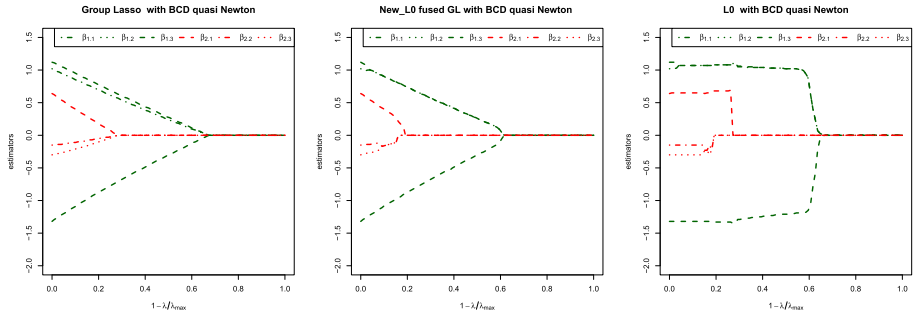

FIG 3. *Coefficient paths of two factors with* 4 *levels each, for group lasso (left),* L_0*-FGL (middle) and CAS-*L_0 *(right). All methods are computed with BCD.*

paths of group lasso (left) in Figure 3 look less smooth than those in Figure 2 where we used PIRLS (however, we note that the truth $\boldsymbol{\beta}^*$ is different for the coefficient paths for PIRLS and BCD, respectively, for visualization purposes). This is caused by the fact that PIRLS uses a quadratic approximation of the whole penalty function while the BCD quasi Newton approach does not use such an approximation for the group lasso part. To conclude, the methods of using a BCD approach with quasi Newton as well as PIRLS look promising and their performance will be investigated more detailed in the following simulation studies.

5. Simulation studies

L_0-FGL procedures, computed with the presented algorithms, have been compared with respect to their computational performance in practice for a representative selection of simulation designs. Here, we discuss in detail two of them, one with 8 factors (B8) and one high-dimensional (highdim), described in Section 5.3. The following methods in both their versions, adaptive and non-adaptive, are included in our comparison.

(i) CAS-L_0 estimator of [25], computed with PIRLS and the package `gvcm.cat`, for which the abbreviation (adaptive) L_0 is used (in the simulation studies we denote this approach by `L0.CV` and `L0.adap.CV`, respectively)

(ii) L_0-FGL (PIRLS) iterative, which is referred to as (adaptive) L_0-FGL (PIRLS) iterative, where the tuning parameters are determined in an iterative manner, see Remark 2.1 (in the simulation studies we denote this approach by `L0.FGL.PIRLS.iterative` and `L0.FGL.PIRLS.adap.iterative`, respectively).

(iii) L_0-FGL (BCD) and quasi newton, which is referred to as (adptive) L_0-FGL BCD, where the tuning parameter is determined with the stepwise procedure, see Remark 2.1 (in the simulation studies we denote this approach by `L0.FGL.BCD` and `L0.FGL.BCD.adap`, respectively)

For the sake of completeness, we also include the maximum likelihood approach in our studies, which we denote by `ML`.

5.1. Choice of the weights

Depending on whether the approach under consideration is non-adaptive or adaptive, the corresponding type of weights is used. The use of adaptive weights will always be explicitly mentioned. Otherwise, non-adaptive weights are used. Both types of weights are specified below.

Non-adaptive weights

Weights can be chosen in the group lasso and the L_0 part. As already explained, a common choice for the group lasso part is to set $\boldsymbol{K}_j$ in such a way that $w_1^{(j)} = \sqrt{p_j}$. The weights used for the L_0 fusion part should account for the number of observations per level, and are chosen along the lines of [9]. Let $n_j^{(r)}$ denote the number of observations of level r of the j-th factor, $j \in \{1, \ldots, J\}$. We have to distinguish between the cases of a nominal or ordinal factor, since in the latter only adjacent categories have to be compared. To sum up, our choice for non-adaptive weights is

(i) L_0-part

nominal $\quad w_0^{(j,rs)} = 2(p_j + 1)^{-1} \sqrt{\dfrac{n_j^{(r)} + n_j^{(s)}}{n}}$,

ordinal $\quad w_0^{(j,r)} = \sqrt{\dfrac{n_j^{(r)} + n_j^{(r-1)}}{n}}$

(ii) GL-part $\quad w_1^{(j)} = \sqrt{p_j}$

Note that, even for $p > n$, if we assume that for every j the number of levels p_j is bounded and in addition $n_j^{(r)}/n \to c_j^{(r)} \in (0, 1)$ for all $j \in \{1, \ldots, J\}$ and $r \in \{1, \ldots, p_j\}$, see [9], we can ensure that the weights for the L_0 part converge to a positive constant.

Adaptive weights

As for other penalties, we can also use adaptive weights to obtain the adaptive L_0-FGL method. This is done by choosing in the group lasso part the weights $w_1^{(j)} = ||\tilde{\boldsymbol{\beta}}_j||_2^{-1}$. A more general choice is $w_1^{(j)} = ||\tilde{\boldsymbol{\beta}}_j||_2^{-\gamma}$ for some chosen $\gamma > 0$ as in [42], where the oracle properties for this more general choice are proved. To keep the adaptive weights on a comparable scale for the group lasso and the L_0 part, we multiply the inverse of the norm of the ML estimate with the non-adaptive weight $\sqrt{p_j}$. In the L_0 part we multiply the weights chosen above with the inverse of the difference of the corresponding ML estimates $\tilde{\boldsymbol{\beta}}$, hence

we multiply with $|\tilde{\beta}_{j,r} - \tilde{\beta}_{j,s}|^{-1}$. Note that we take here the absolute value of the differences of the ML estimates as in [1]. Thus, we propose the following adaptive weights

(i) L_0-part

$$\text{adaptive nominal: } w_0^{(j,rs)} = \frac{1}{|\tilde{\beta}_{j,r} - \tilde{\beta}_{j,s}|} \cdot 2(p_j+1)^{-1} \sqrt{\frac{n_j^{(r)} + n_j^{(s)}}{n}} \tag{15}$$

$$\text{adaptive ordinal: } w_0^{(j,r)} = \frac{1}{|\tilde{\beta}_{j,r} - \tilde{\beta}_{j,r-1}|} \cdot \sqrt{\frac{n_j^{(r)} + n_j^{(r-1)}}{n}} \tag{16}$$

(ii) GL-part adaptive $w_1^{(j)} = \sqrt{p_j}||\tilde{\boldsymbol{\beta}}_j||_2^{-1}$

Remark 5.1. Analogously to the required adjustment of the initial estimator $\tilde{\boldsymbol{\beta}}$ discussed in Remark 3.10 for the adaptive group lasso weights, if $\tilde{\beta}_{j,r} = \tilde{\beta}_{j,s}$ for some j, r, s in (15), then the denominator $|\tilde{\beta}_{j,r} - \tilde{\beta}_{j,s}|$ of (15) is replaced by $|\tilde{\beta}_{j,r} - \tilde{\beta}_{j,s}| + \frac{1}{n}$. The same adjustment applies to the denominator of (16) for the ordinal case, if needed. However, in the following designs taking the unpenalized MLE as initial estimator, this problem did not occur so we used the adaptive weights as given above.

5.2. Goodness of fit measures

The approaches under investigation will be compared wrt the following measures

(i) mean squared error coefficients $MSEC(\hat{\boldsymbol{\beta}}) = \frac{1}{p}\sum_{j=1}^{p}(\beta_j^* - \hat{\beta}_j)^2$
(ii) predictive deviance $Dev(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = -2\sum_{i=1}^{n}\{y_i \log(\hat{\mu}_i) + (1-y_i)\log(1-\hat{\mu}_i)\}$
(iii) false positive (FP)/ false negative (FN) rates factor selection

$$FP_{s,\text{fac}}(\hat{\boldsymbol{\beta}}) = \frac{|\{j \in \{1,\dots,J\} \ : \ ||\hat{\boldsymbol{\beta}}_j|| \neq 0\,,\, ||\boldsymbol{\beta}_j^*|| = 0\}|}{|\{j \in \{1,\dots,J\} \ : \ ||\boldsymbol{\beta}_j^*|| = 0\}|}$$

$$FN_{s,\text{fac}}(\hat{\boldsymbol{\beta}}) = \frac{|\{j \in \{1,\dots,J\} \ : \ ||\hat{\boldsymbol{\beta}}_j|| = 0\,,\, ||\boldsymbol{\beta}_j^*|| \neq 0\}|}{|\{j \in \{1,\dots,J\} \ : \ ||\boldsymbol{\beta}_j^*|| \neq 0\}|}$$

(iv) FP/FN rates fusion, limited to truly influential factors to ensure that just levels fusion and no factor selection is measured

$$FP_{f,\text{infl.truth}} = \frac{|\{(j,k,l) \ : \ \hat{\beta}_{j,k} \neq \hat{\beta}_{j,l}, \beta_{j,k}^* = \beta_{j,l}^*\,,\, (\sum_r |\beta_{j,r}^*|) \neq 0\}|}{|\{(j,k,l) : \beta_{j,k}^* = \beta_{j,l}^*\,,\, (\sum_r |\beta_{j,r}^*|) \neq 0\}|}$$

$$FN_{f,\text{infl.truth}} = \frac{|\{(j,k,l) \ : \ \hat{\beta}_{j,k} = \hat{\beta}_{j,l}, \beta_{j,k}^* \neq \beta_{j,l}^*\,,\, (\sum_r |\beta_{j,r}^*|) \neq 0\}|}{|\{(j,k,l) : \beta_{j,k}^* \neq \beta_{j,l}^*\,,\, (\sum_r |\beta_{j,r}^*|) \neq 0\}|}$$

For ordinal factors we compare the adjacent indices $(j, k, k-1)$.

(v) practical sparsity (PS) := $|\{j \in \{1,..,J\} : ||\hat{\boldsymbol{\beta}}_j||_2 \neq 0\}|$ and overall sparsity (OS):= $|\{k \in \{1,\dots,p\} : \hat{\beta}_k \neq 0\}|$

5.3. Simulation designs

To investigate the performance of the approaches discussed above, a design of low and one of high-dimension are considered, as described in detail next.

Design B8

This design is taken from [25], where the sample size was $n = 400$ while we consider $n = 1000$. We have 8 ordinal factors with 4 levels each. Here, 4 factors are influential and 4 are non-influential. The probabilities for sampling the data where randomly sampled between 0.12 and 0.44. The true coefficient vector was chosen to be

$$\begin{aligned}\boldsymbol{\beta}^* = (2,\ \ 0, -0.8, -0.8,\ \ 1, 1, 0,\ \ 0.4, 0.6, 0.8,\ \ -0.7, -1, 0,\\ 0, 0, 0,\ \ 0, 0, 0,\ \ 0, 0, 0\ \ , 0, 0, 0)^T.\end{aligned}$$

The true overall sparsity is $OS^* = 9$ and the practical sparsity $PS^* = 4$. Hence 50% of the explanatory variables are not influential.

Design highdim

In this high-dimensional design, we observe 60 ordinal factors where the first 50 have 4 categories and the last 10 have 3 categories each. We draw them from a multinomial distribution with equal probabilities. We chose that just the first 5 factors are influential, hence just approximately 8% of the factors have influence on the response. We have $p = 171 > n = 100$. In particular, the true coefficient vector was chosen to be

$$\boldsymbol{\beta}^* = (2, -1, 0.5, 2, 1.5, 1.5, 0.5, 1, 2, 2.5, -0.5, -0.3, 0.5, 2, 1, 3, 0, \ldots, 0)^T.$$

The true overall and practical sparsity are given by $OS^* = 15$ and $PS^* = 5$.

5.4. Analysis of the results

For details on tuning, see Appendix A.4.

5.4.1. Results for design B8

Through simulation studies (also other which are not shown here) we verified that `L0.FGL.BCD` and `L0.FGL.BCD.adap` do not perform well, comparatively to the approaches `L0.FGL.PIRLS.iterative` and `L0.FGL.PIRLS.adap.iterative`, in designs of low to moderate dimension, while they are advantageous in high-dimensional designs. For this, `L0.FGL.BCD` and `L0.FGL.BCD.adap` will not be discussed for the design B8.

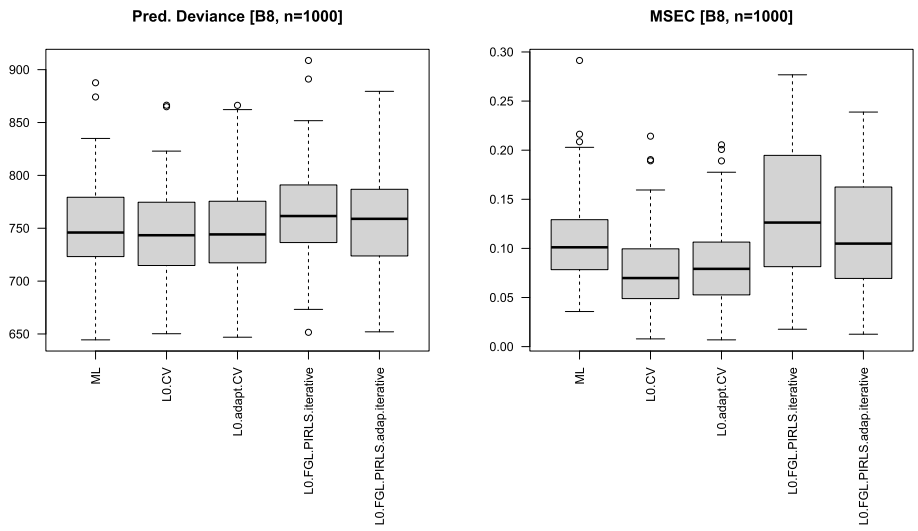

FIG 4. *Predictive deviance and MSEC results for design B8* ($n = 1000$).

We start by analyzing the predictive deviance and MSEC, shown in Figure 4. From the first view, we observe that in terms of predictive deviance, all approaches, including ML, can be ranked on a comparable scale. With respect to MSEC, the existing `L0.CV` or its corresponding adaptive version are preferable. Nevertheless, since the goal of our method is to achieve sparsity through factor selection and levels fusion, it is clear that the corresponding FP/FN rates as well as the overall and practical sparsity measures, which we will analyze next, are of high interest.

Keeping in mind that through L_0-FGL we aim to achieve a stronger factor selection performance than through the existing CAS-L_0 (`L0.CV` and `L0.adap.CV`), we turn our view to Table 1. Comparing the non-adaptive versions to each other, namely `L0.CV` to `L0.FGL.PIRLS.iterative`, we observe that the FP rate concerning factor selection is highly improved, at the cost of a comparably lower increase (in absolute differences) of the corresponding FN rate. That is, the decrease in FP rate from `L0.CV` to `L0.FGL.PIRLS.iterative` is $|0.62 - 0.18| = 0.44$, whereas the corresponding increase in FN rate is $|0.01 - 0.31| = 0.30$. We remark that, for the FN rate, the value for `L0.CV` is very low (0.01), which viewed jointly with the corresponding FP rate of 0.62 and the sparsity levels in Table 2 of `L0.CV`, indicates that this method performs weak factor selection. On the other side, `L0.FGL.PIRLS.iterative` exhibits a more balanced result between FP and FN rates in terms of factor selection. A similar, but less distinct, conclusion is derived for the corresponding adaptive versions `L0.adap.CV` and `L0.FGL.PIRLS.adap.iterative`. Consequently, the new method can clearly improve the factor selection rates. Coming to the FP/FN rates concerning fusion of `L0.CV` compared to `L0.FGL.PIRLS.iterative`, we are able to lower the FP

rate at the cost of a higher FN rate, but this is what we would also expect since `L0.CV` (and `L0.adap.CV`) are penalties directly designed for factor selection and our new method is balancing factor selection and levels fusion performance, finding a compromise between both tasks. Overall, in terms of FP/FN rates for factor selection and fusion, our new method clearly improves the factor selection performance of CAS-L_0.

TABLE 1
[B8, n=1000] FP/FN rates clustering and selection.

	ML	L0.CV	L0.adapt.CV	L0.FGL.PIRLS. iterative	L0.FGL.PIRLS. adap.iterative
$FP_{s,fac}$	1.00	0.62	0.45	0.18	0.10
$FN_{s,fac}$	0.00	0.01	0.03	0.31	0.33
$FP_{f,infl.truth}$	1.00	0.32	0.23	0.22	0.30
$FN_{f,infl.truth}$	0.00	0.21	0.26	0.48	0.46

Next, we analyze the sparsity measures displayed in Table 2. The true overall sparsity in this design is $OS^* = 9$ and the true practical sparsity $PS^* = 4$. We can directly see that both versions of the L_0-FGL with PIRLS algorithm clearly yield the most sparse model and the sparsity levels are near to the truth, especially for the non-adaptive version. Compared to the existing `L0.CV` and `L0.adap.CV`, respectively, we are able to evidently increase the sparsity of the resulting model. To sum up, we saw that L_0-FGL increases the factor

TABLE 2
[B8, n=1000] Overall/Practical Sparsity.

	ML	L0.CV	L0.adapt.CV	L0.FGL.PIRLS. iterative	L0.FGL.PIRLS. adap.iterative
OS	24.00	16.11	13.93	8.60	8.0
PS	8.00	6.46	5.68	3.49	3.1

selection performance of CAS-L_0 at the cost of a bit worse fusion performance and MSEC, but with the goal of balancing both factor selection and levels fusion and achieving a sparse model, near to the true sparsity level, our new method outperforms the other.

Remark 5.2 (Impact of initial values)**.** Since we apply Newton-type algorithms and our objective function is not strictly convex in general (requirements for strict convexity corresponding to PIRLS are given in A.2.2), the estimates may depend on the choice of initial values, which is similarly remarked in [25] applying PIRLS and facing non-convexity with the L_0 norm. They state that, in the majority of cases and for the tuning parameter being in a reasonable range, the results will not differ too much. Following their recommendations further, and due to the underlying sparsity assumption, it is reasonable to use $\boldsymbol{\beta}^{(0)} = \mathbf{0}$ as initial value, as we also did in our simulation studies. This choice is also recommended by [2] and [3] performing (group) coordinate descent with local linear approximations in terms of paths for the maximum tuning parameter.

5.4.2. Results of design highdim

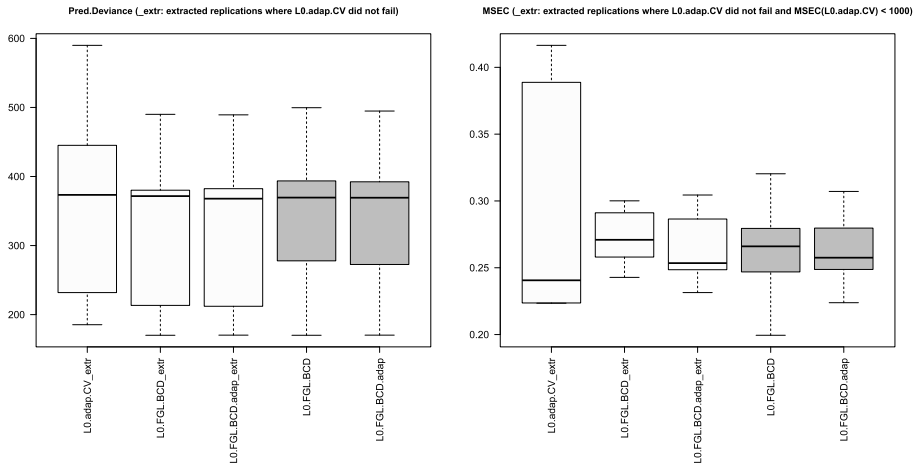

FIG 5. *Predictive deviance (left) and MSEC (right) results for the design highdim. The light colored boxes correspond to algorithms with the lower index* `extr` *and are based on part of the replications, for which the algorithm* `L0.adap.CV` *did not fail. In order to get a meaningful boxplot, two further replications for* `extr` *in MSEC are omitted (where MSEC value of* `L0.adap.CV` *is* ≥ 1000*).*

In a high-dimensional design we have by construction a high number of non-influential factors and, comparably, a very low sample size. Thus, we do not expect from the procedures to identify the true underlying model exactly, but to be able to differentiate between influential and non-influential factors and lower the complexity. A procedure being able to shift this high-dimensional design to a lower dimensional one, keeping (most of) the significant factors, would be a convenient choice for a two-stage procedure. In this design, it is also important to investigate the proportion of replications where the methods fail to yield an estimate. Since L_0 with PIRLS and both versions (adaptive/non-adaptive) of L_0-FGL with PIRLS failed in all $R = 100$ replications, we neglect these approaches in our analysis. Further, the adaptive L_0 with PIRLS (`L0.adap.CV`) algorithm failed in the majority of the replications (87 %). Thus, the corresponding results have to be interpreted with caution, since they are based on only 13 of the replications. The only approach that never failed in any replication is L_0-FGL with BCD (`L0.FGL.BCD`). The corresponding adaptive version (`L0.FGL.BCD.adap`) fails in 30% of the replications which can be explained by the fact that it uses the ML estimate which can cause problems especially in the high-dimensional setting. Figure 5 shows the predictive deviance and MSEC for the different approaches. The notation `extr` indicates that only the values of the replications for which the algorithm `L0.adap.CV` did not fail were extracted and used for the construction of the corresponding boxplot. We can see that L_0-FGL with BCD shows a lower variability in the MSEC from which we can

conclude that it seems to be less sensitive in changes in the data. The same conclusions are derived when observing the predictive deviance. Even though the median MSEC of `L0.adap.CV` is the lowest among all approaches, we observe that we have high variability within the replications, while these results need to be handled with caution, since they are based on very few replications (11 out of 13, since two replications had MSEC values ≥ 1000 and were ignored). Hence, based on our simulation study we strongly suggest the use of L_0-FGL with BCD for high-dimensional setups.

With respect to overall and practical sparsity (Table 3) we observe that L_0 adaptive selects the most sparse model, but since this approach fails in 87% of the replications, this outcome is not to be trusted. But, L_0-FGL BCD (adaptive and non adaptive), which do not fail in the great majority of replications, clearly reduce the number of predictors included in the model. Since it is important that the truly non influential predictors are excluded, we turn our view to Table 4.

TABLE 3
[highdim] Overall/Practical Sparsity ($OS^* = 15, PS^* = 5$).

	ML	L0.adapt. CV	L0.FGL. BCD	L0.FGL BCD.adap
OS	170.00	15.46	60.00	66.26
PS	60.00	10.00	24.93	27.01

TABLE 4
[highdim] FP/FN rates clustering and selection.

	ML	L0.adapt. CV	L0.FGL. BCD	L0.FGL BCD.adap
$FP_{s,fac}$	1.00	0.17	0.41	0.45
$FN_{s,factor}$	0.00	0.83	0.50	0.50
$FP_{f,infl.truth}$	1.00	0.08	0.23	0.27
$FN_{f,infl.truth}$	0.00	0.91	0.71	0.70

We observe that L_0-FGL with BCD (adaptive and non adaptive) has a FP and a FN factor selection rate of about 40% and 50%, respectively. In terms of fusion, we verify that it tends to be conservative in the sense that it has a high percentage (70 %) of falsely not fused levels. The very high corresponding FN rates of the adaptive L_0 with PIRLS are unsatisfactory but, as already mentioned, these results are not to be interpreted since the algorithm has convergence problems in high-dimensional setups leading to very few successful replications.

To sum up, the introduced L_0-FGL procedure, computed with the BCD approach using quasi Newton, is very convenient for such a high-dimensional design since it highly reduces the complexity of the problem. It is remarkable that it does not fail in any of the replications, even if the number of predictors highly exceeds the sample size.

5.4.3. L_0-FGL algorithms: PIRLS vs BCD

Based on conducted simulation studies, we verified that PIRLS is not suitable for computing the L_0-FGL estimates in high-dimensional setups. It is important to remark though, that the convergence of the PIRLS algorithm depends on the penalty function used. Thus, our observations hold for L_0-FGL and may not necessarily hold for other penalty functions. As explained in Section A.2, there are more requirements in high dimensions than in low dimensions for PIRLS to ensure convergence. This explains the observed performance of PIRLS in our high-dimensional design. The BCD approach outperforms PIRLS in high-dimensional designs, since it performs coordinate-wise. For more information on the convergence analysis of BCD we refer to Sections 4.2 and A.3. On the other hand, in low-dimensional sparse designs, as e.g., B8, PIRLS outperforms the BCD procedure and does not have any convergence problems. As we will see in the real data application in Section 6, if the design is low-dimensional, it may happen that also the BCD approach is suitable. However, the PIRLS approach (stepwise tuning) seems to perform slightly better for L_0-FGL (at least in this example). To sum up, we would recommend using PIRLS for L_0-FGL in low-dimensional designs and the BCD approach for L_0-FGL in high-dimensional designs.

6. Real data application

To investigate the performance of L_0-FGL, we applied the method to the "Breast Cancer" data set that has a binary response reporting breast cancer recurrence and is taken from the UCI Machine Learning Repository [24]. The dataset is of sample size $n = 286$ and the number of factors considered is $J = 9$, being binary, nominal or ordinal; see Table 5. The factors include patients' data, which are described in [24], where further details are given in [32]. After removing some cases with missing values, we end up with a complete sample of $n = 277$. Summing up the number of levels of all factors, after excluding levels of zero frequency, we obtain $p = 33$.

6.1. Model selection and quality of fit

After randomly splitting the data in training (70%) and test (30%) datasets, we conducted $R = 100$ replications fitting ML and L_0-FGL with PIRLS and BCD (adaptive and non-adaptive versions) on the training dataset. Both adaptive and non adaptive versions of the existing L_0 approach, fitted with `gvcm.cat`, failed in all replications, which is the reason why we do not show any results on CAS-L_0 in this application. For tuning details we refer to Appendix A.5. First, we observed that the adaptive versions of L_0-FGL PIRLS and L_0-FGL BCD perform worse than the corresponding non-adaptive versions, which can be explained by the dependence on the performance of the ML which is the worst in terms of predictive deviance (reported below). Consequently, we will

TABLE 5
Factors of breast cancer recurrence dataset. The abbreviation (n.i.d) stands for not in dataset, meaning that this level appeared in none of the cases. These levels are excluded from the analysis.

Factor	Levels
age (`x1`), ordinal	10-19 (n.i.d), 20-29, 30-39, 70-79, 80-89 (n.i.d), 90-99 (n.i.d)
menopause (`x2`), nominal	ge40, lt 40, premeno
tumor size (`x3`), ordinal	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59 (n.i.d)
inv nodes (`x4`), ordinal	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20 (n.i.d), 21-23 (n.i.d), 24-26, 27-29 (n.i.d), 30-32 (n.i.d), 33-35 (n.i.d), 36-39 (n.i.d)
node caps (`x5`), binary	yes, no
deg-malig (`x6`), ordinal	1,2,3
breast (`x7`), binary	left, right
breast quad (`x8`), nominal	left-up, left-low, right-up, right-low, central
irridat (`x9`), binary	yes, no

analyze the performance of L_0-FGL PIRLS (`L0.FGL.PIRLS`), iterative L_0-FGL PIRLS (`L0.FGL.PIRLS.iterative`) and L_0-FGL BCD (`L0.FGL.BCD`) with stepwise tuning (see Remark 2.1). In terms of convergence failures (out of $R = 100$ replications), we observed 1 failure for L_0-FGL PIRLS, 13 failures for L_0-FGL PIRLS iterative and no failure for ML and L_0-FGL BCD. We will compare the methods with respect to predictive deviance and misclassification error, i.e., the percentage of false predictions (false positive and false negative).

Figure 6 shows the boxplots for the misclassification errors. We can see that all considered approaches perform in a similar manner while L_0-FGL PIRLS has the lowest misclassification errors among them. Further, Figure 7 shows the predictive deviance, with the plot on the right hand side providing only the L_0-FGL approaches for better visibility of the values. It is remarkable that the ML has significantly higher predictive deviance values compared to the L_0-FGL approaches. Comparing L_0-FGL PIRLS to the corresponding BCD version, the L_0-FGL PIRLS is performing slightly better. The L_0-FGL PIRLS iterative versions has a bit higher predictive deviance values compared to the other approaches, but, as we will see in the fit on the full dataset below, it yields the most sparse model, which explains this fact. Nevertheless, all considered L_0-FGL approaches clearly outperform the existing ML approach. Further, as mentioned above, L_0 fitted with `gvcm.cat` failed in every replication so it is not an alternative for this real data application. Having verified the superiority of the newly proposed approach for this dataset, the L_0-FGL approaches will be applied next on the complete dataset.

6.2. Fit on full dataset

Since the L_0-FGL PIRLS, L_0-FGL PIRLS iterative and L_0-FGL BCD alorithms were the most promising ones in our foregoing analysis (Section 6.1), we proceed by applying them on the complete dataset. Thus, we obtain parameters' estimates for the levels of the influential factors after fusion of levels with statistically non-significant difference of their effects on the response variable. Table 6

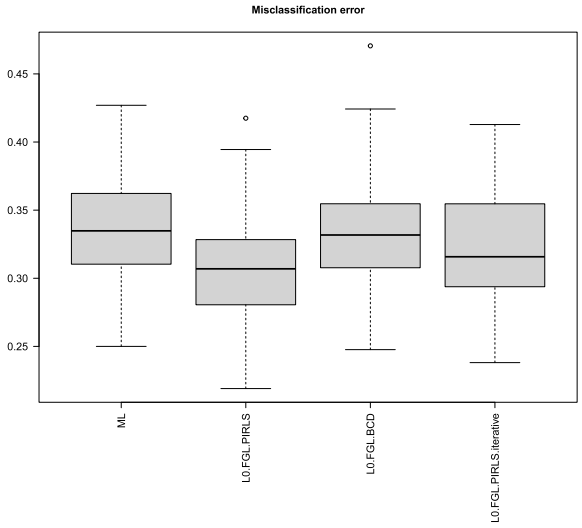

FIG 6. *Misclassification error for real data application.*

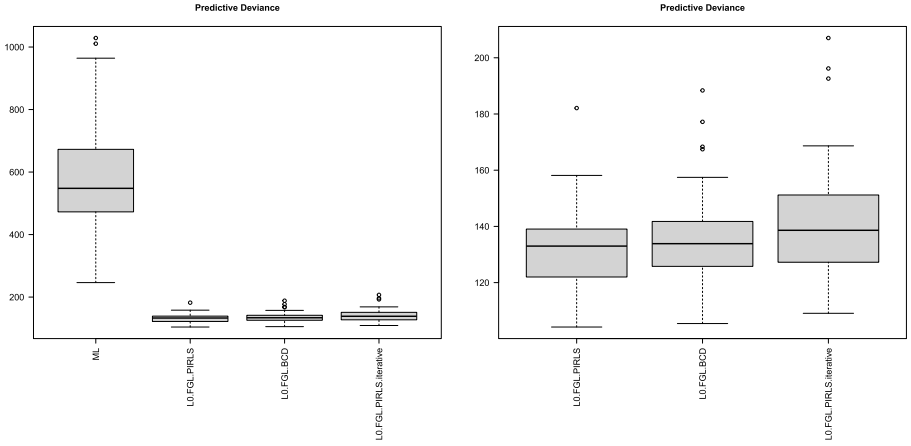

FIG 7. *Predictive deviance for real data application; left hand side: all considered approaches, right hand side: ML excluded.*

provides the coefficient estimates for the corresponding factor levels. L_0-FGL PIRLS and L_0-FGL BCD find all factors as influential on the response, so they perform only levels fusion in this example. In particular, L_0-FGL PIRLS fuses some levels of the factors x4 and x6 while L_0-FGL BCD leads to a sparser model, fusing more levels of x3, x4 and x8. The most sparse model is fitted by L_0-FGL PIRLS iterative. It performs selection of factors (eliminating x5, x7, x8 and x9) and levels' fusion for factors x1, x3, x4 and x6, reducing the

number of their parameters from 5, 10, 6 and 3 to 4, 3, 3 and 2, respectively (see Table 6). The three considered approaches perform similarly in terms of FP and FN missclassification errors, which are 0.63/0.67/0.68 and 0.06/0.06/0.04 for L_0-FGL PIRLS, L_0-FGL BCD and L_0-FGL PIRLS iterative, respectively. The sparsity of models fitted by L_0-FGL BCD and L_0-FGL PIRLS iterative is gained at the cost of slightly higher error measures. Thus the model simplification with the associated gain in interpretability, comes with a slight increase of misclassification error and predictive deviance.

It is notable that all approaches outperform the ML in terms of predictive deviance (see Section 6.1) and also the existing L_0 approach (implemented in `gvcm.cat`), which failed in every replication. This makes the new introduced method advantageous for real data applications. In some next steps one could perform statistical hypothesis tests to investigate the significance of the factors and the corresponding levels and proceed to associated interpretations. However, this is beyond the scopes of this work.

7. Conclusion

In this work, a new approach is introduced, the L_0-FGL, which performs both, factor selection and levels fusion of categorical predictors, combining two penalty terms, one for selection and one for fusion. The motivation for the introduction of L_0-FGL was to obtain an approach not only performing fusion with an L_0 type penalty (as does CAS-L_0), but also performing *direct* factor selection, balancing thus both tasks. First, we obtained theoretical results for the new proposed double penalized approach. Having proven the existence, it is further shown that, under certain regularity conditions, there exists an L_0-FGL estimator satisfying $\sqrt{n}$ consistency, even when the number of parameters grows with the sample size. In addition, this L_0-FGL estimator satisfies a result concerning consistency in variable selection in both cases, for fixed or sample size dependent number of parameters. Fixing p, there exists an adaptive L_0-FGL estimator satisfying asymptotic normality, while for the diverging p_n case we showed a similar result for the approximate L_0-FGL. Simulation studies verified that the new L_0-FGL approach implemented with PIRLS shows a superior performance in lower dimensional designs and tends to improve the selection performance of the classical L_0 method due to the incorporation of the group lasso part. Clearly, L_0-FGL balances factor selection and levels fusion performance. The performance of L_0-FGL computed with BCD in high dimensions outperformed the other approaches in the grand majority of replications. It is capable of identifying sparse models and reduces further the dimension of the problem through possible levels' fusion of categorical predictors, delivering thus sound interpretations for the associated effects on the response variable. The theoretical properties along with the simulation results make L_0-FGL a promising method for modeling high-dimensional data with factors, where sparsity is achieved not only through variable selection but also through levels' fusion. In our real data application we showed that L_0-FGL is also able to handle lower dimensional designs coming from real datasets.

TABLE 6
Estimates for fit on full dataset. The abbreviation (r.c.) stands for reference category. Pair or groups of parameter estimates of successive levels that are equal are in bold (or italics) and correspond to levels that are fused.

	Level	L_0-FGL PIRLS	L_0-FGL BCD	L_0-FGL PIRLS iterative
Intercept		−1.85	1.60	−2.69
age (x1)	20-29 (r.c.)	0.00	0.00	0.00
	30-39	0.14	−0.12	0.01
	40-49	−0.04	−0.57	**0.00**
	50-59	−0.15	−0.92	**0.00**
	60-69	0.12	−0.63	0.83
	70-79	−0.06	−1.11	−2.51
menopause (x2)	ge40 (r.c.)	0.00	0.00	0.00
	lt40	−0.20	−1.10	−2.92
	premeno	0.29	−0.43	0.84
tumor size (x3)	0-4 (r.c.)	0.00	0.00	0.00
	5-9	−0.05	−1.03	−1.71
	10-14	−0.21	−2.30	−1.73
	15-19	−0.05	**−0.84**	**0.48**
	20-24	0.05	**−0.84**	**0.48**
	25-29	0.10	−0.41	**0.48**
	30-34	0.20	**−0.33**	**0.48**
	35-39	−0.01	**−0.33**	**0.48**
	40-44	−0.05	−0.80	**0.48**
	45-49	−0.01	**−0.31**	**0.48**
	50-54	0.06	**−0.31**	**0.48**
inv nodes (x4)	0-2 (r.c.)	0.00	0.00	0.00
	3-5	0.03	0.30	**1.25**
	6-8	0.01	0.30	**1.25**
	9-11	0.02	0.38	**1.25**
	12-14	**0.00**	**0.17**	*0.68*
	15-17	**0.00**	**0.17**	*0.68*
	24-26	**0.00**	**0.17**	2.73
node caps (x5)	no (r.c.)	0.00	0.00	–
	yes	0.83	0.36	–
deg-malig (x6)	1 (r.c.)	**0.00**	0.00	**0.00**
	2	**0.00**	−1.03	**0.00**
	3	1.28	0.18	1.42
breast (x7)	left (r.c.)	0.00	0.00	–
	right	−0.03	−0.69	–
breast quad (x8)	central (r.c.)	0.00	0.00	–
	left-low	0.06	−0.68	–
	left-up	−0.08	**−0.69**	–
	right-low	−0.04	**−0.69**	–
	right-up	0.12	−0.04	–
irridat (x9)	no (r.c.)	0.00	0.00	–
	yes	0.54	0.07	–

From a theoretical point of view, [19] established the asymptotic normality of the maximum likelihood estimator of the parameter vector in the diverging p_n case under mild regularity conditions, allowing $p_n > n$. This procedure may be exploited in our case, opening the way for asymptotic inference. Approaches for goodness of fit testing for models estimated by L_0-FGL need to be devel-

oped and are currently under research, taking into account, among others, the methodology of [31], [30], [37] and [27]. Additionally, since our theoretical properties hold for a local minimizer, further work can be developed analyzing the connection to the minimizer obtained by the algorithms, as done e.g. in [8] for folded concave penalties.

Furthermore, it would be interesting to develop for high-dimensional logistic regression with fusion an alternative adaptive selection algorithm, of the type introduced by [28], and compare it with the results discussed here. Finally, the consideration of levels' fusion for categorical predictors is crucial also in a longitudinal context (see [18]) and, to the best of our knowledge, has been ignored so far. Extension of the approach proposed here towards this direction can be explored. Additionally, the L_0-FGL procedure can be extended for allowing for order restricted estimation of the levels' parameters of ordinal factors, which is of high interpretability value if ordered effects are expected.

Additionally, further investigation in compromising factor selection and levels fusion with L_0-FGL can be conducted by performing CV on a (fine) two-dimensional grid. However, especially by including an L_0 penalty, this will be computationally intensive and requires more research on fast algorithms to handle functions including an L_0 penalty on the differences of a factor's level parameters.

A possible alternative algorithm to solve the L_0-FGL optimization problem is the penalty decomposition (PD) method, which was introduced in [20] and applied to L_0 norm minimization problems in [21]. Generally speaking, after re-writing the optimization problem as a rank minimization problem, a penalty decomposition method is used to solve the optimization problem. Appearing sub-problems are solved with a BCD method. The optimization problem including a L_0 norm is formulated in [21]. However, further work has to be done to investigate the performance and computational details of the PD method applied to L_0-FGL on which we are currently working on.

Appendix

This appendix will provide the proofs of the theorems previously stated. In addition, details on the approximations used in the computational part will be obtained. Furthermore, we will provide some details on the algorithms used for fitting L_0-FGL, including some convergence analysis. Finally, details on the tuning used in our simulation studies and the real data application will be given.

A.1. Proofs

A.1.1. Proof of Theorem 3.1

(1) $S \neq 0$ **:** Set $J = 1$, the proof for $J > 1$ works analogously. We will show that for $J = 1$, the group lasso estimator, given by

$$\hat{\boldsymbol{\beta}}^{GL} := \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\arg\min} -L_n(\boldsymbol{\beta}) + \lambda_1 ||\boldsymbol{\beta}||_K,$$

fulfills $\hat{\boldsymbol{\beta}}^{GL} \in S$. By assumption, $0 < \sum_{i=1}^{n} y_i < n$ and by [22] (Lemma 1) we can follow that the group lasso estimator $\hat{\boldsymbol{\beta}}^{GL}$ exists. Further, as mentioned in [22], the function $-L_n(\boldsymbol{\beta}) + \lambda_1 ||\boldsymbol{\beta}||_K$ is convex and by Lemma 1 in their work, the minimum of this function is attained. Nevertheless, there are further requirements needed to ensure the uniqueness of the group lasso estimator, in particular the design matrix needs to be of full rank, which we do not require here. Now, it holds that there exists an $\boldsymbol{\varepsilon}$-neighborhood of $\hat{\boldsymbol{\beta}}^{GL}$, where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p) \in \mathbb{R}^p$, such that $\hat{\boldsymbol{\beta}}^{GL}$ minimizes the sum $-L_n(\cdot) + \lambda_1 ||\cdot||_K$. Hence

$$-L_n(\hat{\boldsymbol{\beta}}^{GL} + \boldsymbol{\varepsilon}) + \lambda_1 ||\hat{\boldsymbol{\beta}}^{GL} + \boldsymbol{\varepsilon}||_K \geq -L_n(\hat{\boldsymbol{\beta}}^{GL}) + \lambda_1 ||\hat{\boldsymbol{\beta}}^{GL}||_K.$$

Consequently, adding $\lambda_0 \sum_{r,s} w_0^{(rs)} ||\hat{\beta}_r^{GL} - \hat{\beta}_s^{GL} + \varepsilon_r - \varepsilon_s||_0$ on both sides of the inequality

$$\begin{aligned} & M_{pen}(\hat{\boldsymbol{\beta}}^{GL} + \boldsymbol{\varepsilon}) \\ &= -L_n(\hat{\boldsymbol{\beta}}^{GL} + \boldsymbol{\varepsilon}) + \lambda_1 ||\hat{\boldsymbol{\beta}}^{GL} + \boldsymbol{\varepsilon}||_K + \lambda_0 \sum_{r,s} w_0^{(rs)} ||\hat{\beta}_r^{GL} - \hat{\beta}_s^{GL} + \varepsilon_r - \varepsilon_s||_0 \\ &\geq -L_n(\hat{\boldsymbol{\beta}}^{GL}) + \lambda_1 ||\hat{\boldsymbol{\beta}}^{GL}||_K + \lambda_0 \sum_{r,s} w_0^{(rs)} ||\hat{\beta}_r^{GL} - \hat{\beta}_s^{GL} + \varepsilon_r - \varepsilon_s||_0. \end{aligned} \tag{17}$$

For the group lasso estimate we have either $\hat{\boldsymbol{\beta}}^{GL} = \mathbf{0}$ or $\hat{\beta}_r^{GL} \neq 0 \, \forall r$, see [22], in the latter case we further have that $\hat{\beta}_r^{GL} \neq \hat{\beta}_s^{GL} \; \forall \, r, s$ almost surely. For the case where we have $\hat{\beta}_r^{GL} \neq \hat{\beta}_s^{GL} \; \forall \, r, s$ we can choose $\boldsymbol{\varepsilon}$ small enough such that $\hat{\beta}_r^{GL} + \varepsilon_r \neq \hat{\beta}_s^{GL} + \varepsilon_s \; \forall \, r, s$. Consequently, we conclude that the L_0 norms of $\hat{\beta}_r^{GL} - \hat{\beta}_s^{GL}$ and $\hat{\beta}_r^{GL} - \hat{\beta}_s^{GL} + \varepsilon_r - \varepsilon_s$ coincide since all values of the differences are nonzero. Hence, $||\hat{\beta}_r^{GL} - \hat{\beta}_s^{GL} + \varepsilon_r - \varepsilon_s||_0 = ||\hat{\beta}_r^{GL} - \hat{\beta}_s^{GL}||_0$. Then we obtain that (17) equals

$$-L_n(\hat{\boldsymbol{\beta}}^{GL}) + \lambda_1 ||\hat{\boldsymbol{\beta}}^{GL}||_K + \lambda_0 \sum_{r,s} w_0^{(rs)} ||\hat{\beta}_r^{GL} - \hat{\beta}_s^{GL}||_0 = M_{pen}(\hat{\boldsymbol{\beta}}^{GL})$$

and thus $M_{pen}(\hat{\boldsymbol{\beta}}^{GL} + \boldsymbol{\varepsilon}) \geq M_{pen}(\hat{\boldsymbol{\beta}}^{GL})$ for a sufficiently small $\boldsymbol{\varepsilon}$. If $\hat{\boldsymbol{\beta}}^{GL} = \mathbf{0}$ we get with the same arguments as above

$$\begin{aligned} M_{pen}(\hat{\boldsymbol{\beta}}^{GL} + \varepsilon) &\geq -L_n(\mathbf{0}) + \lambda_1 ||\mathbf{0}||_K + \lambda_0 \sum_{r,s} w_0^{(rs)} \underbrace{||\varepsilon_r - \varepsilon_s||_0}_{\geq 0} \\ &\geq -L_n(\mathbf{0}) + \lambda_1 ||\mathbf{0}||_K = M_{pen}(\mathbf{0}) = M_{pen}(\hat{\boldsymbol{\beta}}^{GL}) \end{aligned}$$

thus $M_{pen}(\hat{\boldsymbol{\beta}}^{GL} + \varepsilon) \geq M_{pen}(\hat{\boldsymbol{\beta}}^{GL})$. Hence, the group lasso estimator $\hat{\boldsymbol{\beta}}^{GL}$ is a local minimizer of the L_0-FGL objective function, thus an element of the set S giving us that $S \neq \emptyset$ and the first part of the claim follows.

(2) $M_{pen}(\cdot)$ decreases if coefficients that are close enough to each other are fused : as we know that the group lasso estimator is one solution of L_0-FGL but without fusion, we have to show that the objective function $M_{pen}(\cdot)$ decreases if fusion occurs. Again, we assume that $J = 1$ and we start with the case of an

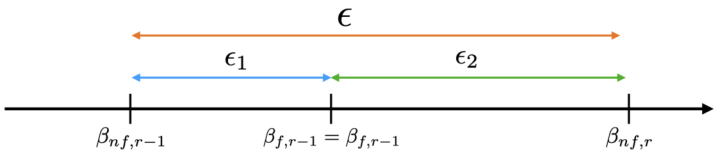

FIG 8. *Location of* $\beta_{nf,r}$, $\beta_{nf,r-1}$ *and the fused coefficients* $\beta_{f,r} = \beta_{f,r-1}$ *for the case* $\min\{\beta_{nf,r}, \beta_{nf,r-1}\} = \beta_{nf,r-1}$ *(other case works analogously).*

ordinal factor comparing adjacent categories for fusion. The goal is to show that the objective function $M_{pen}(\cdot)$ decreases if coefficients that are close enough to each other are fused. Note that, since we chose reference category zero, there is no appearance of the reference category in the coefficient vector $\boldsymbol{\beta}$. Write $\boldsymbol{\beta}_{nf}$ (not fused), $\boldsymbol{\beta}_f$ (fused) $\in \mathbb{R}^p$ with

$$\begin{aligned}
&\boldsymbol{\beta}_{nf} = (\beta_{nf,1}, \ldots, \beta_{nf,p}), \text{ where } \beta_{nf,i} \neq \beta_{nf,i-1}\, \forall i = 2, \ldots, p \text{ (not fused)},\\
&\boldsymbol{\beta}_f = (\beta_{f,1}, \ldots, \beta_{f,p}), \text{ where } \beta_{nf,i} = \beta_{f,i} \neq \beta_{f,i-1} = \beta_{nf,i-1}\\
&\qquad\qquad \forall i = 2, \ldots, r-1, r+1, \ldots p \text{ and } \beta_{f,r} = \beta_{f,r-1}.
\end{aligned}$$

So in $\boldsymbol{\beta}_f$ the categories r and $r-1$ are fused and, except for these categories, $\boldsymbol{\beta}_{nf}$ and $\boldsymbol{\beta}_f$ coincide.
Note that $\beta_{f,r} = \beta_{f,r-1} \in [\min\{\beta_{nf,r}, \beta_{nf,r-1}\}, \max\{\beta_{nf,r}, \beta_{nf,r-1}\}]$. Without loss of generality, we assume $\min\{\beta_{nf,r}, \beta_{nf,r-1}\} = \beta_{nf,r-1}$. Since we observe an ordinal factor, this holds by definition but observing nominal factors one has to differentiate between these two cases but the other case works in the same way. Thus it holds that $\beta_{nf,r} - \beta_{nf,r-1} = \epsilon_1 + \epsilon_2 = \epsilon > 0$ for some (small) ϵ and $\beta_{nf,r} = \beta_{nf,r-1} + \epsilon$, see Figure 8. Now we have to show that $M_{pen}(\boldsymbol{\beta}_f) < M_{pen}(\boldsymbol{\beta}_{nf})$. It depends on the design and the tuning etc. how small ϵ has to be such that the objective function decreases. We know that $\boldsymbol{\beta}_{nf} - \boldsymbol{\beta}_f = (0, \ldots, 0, -\epsilon_1, \epsilon_2, 0, \ldots, 0)$. Because of the continuity of the negative log-likelihood $-L_n(\boldsymbol{\beta})$ and the norm $||\boldsymbol{\beta}||_{\boldsymbol{K}}$ it holds that $\forall \delta_1, \delta_2,\ \exists \tilde{\epsilon}_1, \tilde{\epsilon}_2 > 0$ such that $\forall \boldsymbol{\beta}_{nf}, \boldsymbol{\beta}_f$ with $||\boldsymbol{\beta}_{nf} - \boldsymbol{\beta}_f|| < \min\{\tilde{\epsilon}_1, \tilde{\epsilon}_2\}$ it holds

$$\begin{aligned}
&|L_n(\boldsymbol{\beta}_{nf}) - L_n(\boldsymbol{\beta}_f)| < \delta_1,\\
&\left|||\boldsymbol{\beta}_{nf}||_{\boldsymbol{K}} - ||\boldsymbol{\beta}_f||_{\boldsymbol{K}}\right| < \delta_2.
\end{aligned}$$

Because of the definition of $\boldsymbol{\beta}_{nf}$ (no categories fused) and $\boldsymbol{\beta}_f$ (category r and $r-1$ fused) we know that $\sum_{i=1}^p w_0^{(i)} ||\beta_{nf,i} - \beta_{nf,i-1}||_0 = \sum_i w_0^{(i)} =: c$ and for the fused version we know $\sum_{i=1}^p w_0^{(i)} ||\beta_{f,i} - \beta_{f,i-1}||_0 = c - w_0^{(r)}$. Furthermore $L_n(\boldsymbol{\beta}_{nf}) - L_n(\boldsymbol{\beta}_f) > -\delta_1$ and $||\boldsymbol{\beta}_{nf}||_{\boldsymbol{K}} - ||\boldsymbol{\beta}_f||_{\boldsymbol{K}} > -\delta_2$. Thus, for $\boldsymbol{\beta}_{nf}, \boldsymbol{\beta}_f$ with $||\boldsymbol{\beta}_{nf} - \boldsymbol{\beta}_f|| = \epsilon < \min\{\tilde{\epsilon}_1, \tilde{\epsilon}_2\}$ we deduce

$$\begin{aligned}
&M_{pen}(\boldsymbol{\beta}_{nf}) - M_{pen}(\boldsymbol{\beta}_f)\\
&= -L_n(\boldsymbol{\beta}_{nf}) + \lambda_1 ||\boldsymbol{\beta}_{nf}||_{\boldsymbol{K}} + \lambda_0 \sum_{i=1}^p w_0^{(i)} ||\beta_{nf,i} - \beta_{nf,i-1}||_0
\end{aligned}$$

$$
\begin{aligned}
&+L_n(\boldsymbol{\beta}_f) - \lambda_1||\boldsymbol{\beta}_f||_{\boldsymbol{K}} - \lambda_0 \sum_{i=1}^{p} w_0^{(i)}||\beta_{f,i} - \beta_{f,i-1}||_0 \\
&= -L_n(\boldsymbol{\beta}_{nf}) + \lambda_1||\boldsymbol{\beta}_{nf}||_{\boldsymbol{K}} + L_n(\boldsymbol{\beta}_f) - \lambda_1||\boldsymbol{\beta}_f||_{\boldsymbol{K}} + \lambda_0 \cdot w_0^{(r)} \\
&> -\delta_1 - \lambda_1\delta_2 + \lambda_0 \cdot w_0^{(r)} \qquad (18)
\end{aligned}
$$

Now, if we choose λ_0 (tuning for fusion) large enough and δ_1, δ_2 small enough such that $\lambda_0 \cdot w_0^{(r)} > \delta_1 + \lambda_1\delta_2$, we get with the above equation

$$M_{pen}(\boldsymbol{\beta}_{nf}) - M_{pen}(\boldsymbol{\beta}_f) > 0 \Leftrightarrow M_{pen}(\boldsymbol{\beta}_{nf}) > M_{pen}(\boldsymbol{\beta}_f)$$

and consequently the value of the objective function in $\boldsymbol{\beta}_f$ is less than in $\boldsymbol{\beta}_{nf}$, hence the objective function decreases if we fuse coefficients that are close enough to each other. The proof can directly be extended to the case where we fuse more categories and also for the nominal case. It is clear that λ_0 controls fusion since larger λ_0 values enforce fusion for categories that are further apart.

A.1.2. Proof of Theorem 3.5

In the proof of Theorem 3.5, we will use the following Lemma.

Lemma A.1. *Let $M_{pen}(\boldsymbol{\beta})$ be the objective function of L_0-FGL see (5). Assume that we can show for some $\boldsymbol{x}^* \in \mathbb{R}^p$ and $c \in \mathbb{R}^{>0}$ that*

$$\inf_{||\boldsymbol{u}||_2 = c} M_{pen}(\boldsymbol{x}^* + \boldsymbol{u}) > M_{pen}(\boldsymbol{x}^*). \qquad (19)$$

Then, there exists at least one local minimum of $M_{pen}(\boldsymbol{\beta})$ inside $\mathcal{D} := \{\boldsymbol{x}^* + \boldsymbol{u} \,|\, ||\boldsymbol{u}||_2 \leq c\}$, where inside means in the domain $\mathring{\mathcal{D}} = \{\boldsymbol{x}^* + \boldsymbol{u} \,|\, ||\boldsymbol{u}||_2 < c\}$.

Proof. (of Lemma A.1)
Initial Remark: If the function M_{pen} was continuous, this would be clear since a continuous function attains its minimum and maximum in a compact set, hence in $\mathcal{D}$, and then we could use (19) to show that the infimum (minimum) is not attained at the boundary of $\mathcal{D}$. But, since M_{pen} consists, among other parts, of a L_0 part, it is not continuous. Since we do not penalize the intercept and the intercept just appears in the log-likelihood, we neglect it hence we observe $M_{pen}(\boldsymbol{\beta})$ for $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{x}^* \in \mathbb{R}^p$ (instead of $\mathbb{R}^{p+1}$). Consequently, we have to show that M_{pen} attains its infimum in $\mathcal{D}$. Having that, we use (19) to show that the infimum is not attained at the boundary, hence it is in $\mathring{\mathcal{D}}$.

Returning to the proof of Lemma A.1, we will prove it for the case $p = 2$ and $J = 1$. Cases of higher dimensions work in a similar manner, although we get more possible cases for the infimum to occur (see below). In this setting of choosing $p = 2$ and $J = 1$, we just have one weight w_0 in the L_0 part (see (5)). We start by partitioning $\mathcal{D}$ into two subsets in the following way

$$\mathcal{D}_1 \quad := \mathcal{D} \backslash \{\boldsymbol{\beta} = (\beta_1, \beta_2) \,:\, \beta_1 < \beta_2\},$$

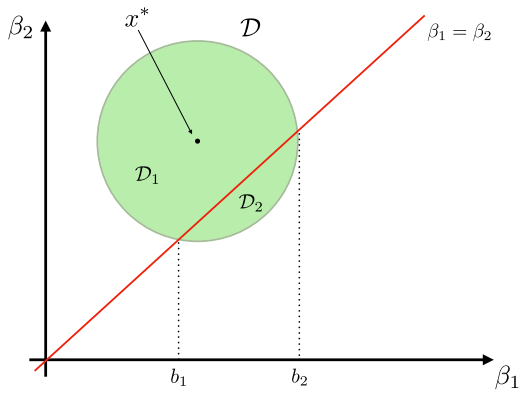

FIG 9. *Partition of the ball $\mathcal{D}$ into $\mathcal{D}_1$ and $\mathcal{D}_2$. The red line shows the 1-dimensional hyperplane where $f(\boldsymbol{\beta})$ is not continuous, hence $\beta_1 = \beta_2$.*

$$\mathcal{D}_2 \quad := \mathcal{D} \backslash \{\boldsymbol{\beta} = (\beta_1, \beta_2) \,:\, \beta_2 < \beta_1\}.$$

So the hyperplane satisfying $\beta_1 = \beta_2$ is included in both subsets. We clearly have that $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$. This partition is displayed in Figure 9.

We can write by definition of the objective function $M_{pen}(\boldsymbol{\beta}) = g(\boldsymbol{\beta}) + f(\boldsymbol{\beta})$ where, for $\boldsymbol{\beta} = (\beta_1, \beta_2)$, $g(\boldsymbol{\beta}) := -L_n(\boldsymbol{\beta}) + \lambda_1 ||\boldsymbol{\beta}||_{\boldsymbol{K}}$ is the sum of the log-likelihood and group lasso part and $f(\boldsymbol{\beta}) := \lambda_0 w_0 ||\beta_1 - \beta_2||_0$ the L_0 part (these abbreviations are just used within this proof). Note that, by definition of the L_0 norm applied to differences, this norm is equal to zero if the object on which we apply the norm is zero and one otherwise, hence it is zero if the difference is zero and it is one if the difference is nonzero. Keep in mind that we multiply the resulting value with the weight w_0. For $g(\boldsymbol{\beta})$ we know that it attains a (local) minimum in $\mathcal{D}$, we write $\boldsymbol{\beta}_g = (\beta_{g,1}, \beta_{g,2}) = \arg\min_{\boldsymbol{\beta} \in \mathcal{D}} g(\boldsymbol{\beta})$. Without loss of generality, we assume that $\boldsymbol{\beta}_g \in \mathcal{D}_1$, the other case works completely analogous. There are two possible cases that may occur.

Case (1): $\beta_{g,1} \neq \beta_{g,2}$
Here, we have that $f(\boldsymbol{\beta}_g) = f((\beta_{g,1}, \beta_{g,2})) = 1 \cdot w_0 = w_0$. Consequently, the infimum of the objective function either occurs in $\mathcal{D}_1$ without the hyperplane $(\beta_1 = \beta_2)$ or it occurs on this hyperplane. In particular, this means

$$\inf_{\boldsymbol{\beta} \in \mathcal{D}} M_{pen}(\boldsymbol{\beta}) \in \{g(\boldsymbol{\beta}_g) + w_0, \inf_{b \in [b_1, b_2]} g((b, b))\}$$

so the infimum of M_{pen} is either attained in $\boldsymbol{\beta}_g$ or in (b, b) for some $b \in [b_1, b_2]$. Later, we will show that with our additional assumption (19), we know that the infimum is not at the boundary hence $b \in (b_1, b_2)$ but this is not important at this point since we just want to show that the infimum is attained somewhere in $\mathcal{D}$.

Case (2): $\beta_{g,1} = \beta_{g,2}$
In this case we have that $f(\boldsymbol{\beta}_g) = f((\beta_{g,1}, \beta_{g,1})) = 0$. Consequently

$$\inf_{\boldsymbol{\beta}\in\mathcal{D}} M_{pen}(\boldsymbol{\beta}) = \inf_{\boldsymbol{\beta}\in\mathcal{D}} g(\boldsymbol{\beta})$$

and hence the infimum of M_{pen} is attained in $\boldsymbol{\beta}_g$.

In both cases, there exists some $\tilde{\boldsymbol{\beta}} \in \mathcal{D}$ for which the infimum is attained, hence it equals the minimum

$$\arg\min_{\boldsymbol{\beta}\in\mathcal{D}} M_{pen}(\boldsymbol{\beta}) = \tilde{\boldsymbol{\beta}}.$$

Note that, in Figure 9 it can of course also occur the case that the red hyperplane does not go through the domain $\mathcal{D}$ hence there is no intersection of the hyperplane and $\mathcal{D}$. If this is the case, we are finished since then the function f will be equal to one everywhere, hence M_{pen} would be continuous. It remains to show that $\tilde{\boldsymbol{\beta}} \in \mathring{\mathcal{D}}$. Assume that $\tilde{\boldsymbol{\beta}}$ is on the boundary of $\mathcal{D}$, hence $\tilde{\boldsymbol{\beta}} \in \mathcal{D}\backslash\mathring{\mathcal{D}}$. Consequently, it holds by definition of the infimum that

$$\inf_{||\boldsymbol{u}||_2=c} M_{pen}(\boldsymbol{x}^* + \boldsymbol{u}) = M_{pen}(\tilde{\boldsymbol{\beta}}) \leq M_{pen}(\boldsymbol{\beta}) \;\; \forall \boldsymbol{\beta} \in \mathcal{D}$$

and this also holds for $\boldsymbol{\beta} = \boldsymbol{x}^* \in \mathcal{D}$ which is a contradiction to the assumption (19). Therefore, it holds that $\tilde{\boldsymbol{\beta}} \in \mathring{\mathcal{D}}$, hence there exists a local minimum of M_{pen} in $\mathring{\mathcal{D}}$. □

Proof. (of Theorem 3.5) The L_0-FGL penalty function $P_\lambda(\boldsymbol{\beta})$ is given by (4), replacing λ_0 by λ_0^n and λ_1 by λ_1^n, respectively. $M_{pen}(\boldsymbol{\beta})$ is given by (5). Following the ideas of [6] and [34] we have to show that $\forall \varepsilon > 0$ we can find a suitable $c > 0$ such that the following holds

$$P\left(\inf_{\boldsymbol{u}\in\mathbb{R}^p, ||\boldsymbol{u}||_2=c} M_{pen}\left(\boldsymbol{\beta}^* + \frac{1}{\sqrt{n}}\boldsymbol{u}\right) > M_{pen}(\boldsymbol{\beta}^*)\right) \geq 1 - \varepsilon. \tag{20}$$

In contrast to [6], we minimize the sum of the negative log-likelihood and the chosen penalty where they maximize the negative objective function which is clearly equivalent. We will transfer the idea of [6] to our case of L_0-FGL. Having shown (20), we get that there exists a local minimum inside the ball $\{\boldsymbol{\beta}^* + \frac{1}{\sqrt{n}}\boldsymbol{u}$ where $||\boldsymbol{u}||_2 < c\}$ using Lemma A.1. This yields that we can find a local minimizer such that $||\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}||_2 = O_p(1/\sqrt{n})$ which is the claim. We start by plugging in the definition of $M_{pen}(\boldsymbol{\beta})$ giving us with $H_n(\boldsymbol{u}) := M_{pen}\left(\boldsymbol{\beta}^* + \frac{1}{\sqrt{n}}\boldsymbol{u}\right) - M_{pen}(\boldsymbol{\beta}^*)$

$$\begin{aligned} & H_n(\boldsymbol{u}) \\ = & -L_n\left(\boldsymbol{\beta}^* + \frac{1}{\sqrt{n}}\boldsymbol{u}\right) + L_n(\boldsymbol{\beta}^*) + \lambda_1^n \sum_{j=1}^{J} w_1^{(j)}(||\boldsymbol{\beta}_j^* + \frac{1}{\sqrt{n}}\boldsymbol{u}||_2 - ||\boldsymbol{\beta}_j^*||_2) \qquad (21) \\ & + \lambda_0^n \sum_{j=1}^{J} \sum_{0\leq r<s\leq p_j} w_0^{(j,rs)} \left(||\beta_{j,r}^* - \beta_{j,s}^* + \frac{1}{\sqrt{n}}(u_r - u_s)||_0 - ||\beta_{j,r}^* - \beta_{j,s}^*||_0\right). \end{aligned}$$

We will observe the three parts from the right hand side of the equation seperately. Like in [42] (proof of Theorem 4) we will investigate the behavior of the first part $-L_n\left(\boldsymbol{\beta}^* + \frac{1}{\sqrt{n}}\boldsymbol{u}\right) + L_n(\boldsymbol{\beta}^*)$ with a Taylor expansion of $f_n(\boldsymbol{u}) := -L_n\left(\boldsymbol{\beta}^* + \frac{1}{\sqrt{n}}\boldsymbol{u}\right) + L_n(\boldsymbol{\beta}^*)$ around $\boldsymbol{u} = 0$ which gives us using $f_n(\mathbf{0}) = \mathbf{0}$

$$-L_n\left(\boldsymbol{\beta}^* + \frac{1}{\sqrt{n}}\boldsymbol{u}\right) + L_n(\boldsymbol{\beta}^*) = T_{1,n} + T_{2,n} + T_{3,n}. \tag{22}$$

In particular, it holds with $\alpha_n := \frac{1}{\sqrt{n}}$

$$\begin{aligned}
T_{1,n} &= -\alpha_n \nabla^T L_n(\boldsymbol{\beta}^*)\boldsymbol{u} = -\sum_{i=1}^{n}[y_i - \varphi'(\boldsymbol{x}_i\boldsymbol{\beta}^*)]\boldsymbol{x}_i^T\boldsymbol{u}\,\alpha_n \\
T_{2,n} &= -\frac{1}{2}\boldsymbol{u}^T\nabla^2 L_n(\boldsymbol{\beta}^*)\boldsymbol{u}\alpha_n^2 = \sum_{i=1}^{n}\frac{1}{2}\varphi''(\boldsymbol{x}_i\boldsymbol{\beta}^*)\boldsymbol{u}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{u}\,\alpha_n^2 \\
T_{3,n} &= -\frac{1}{6}\sum_{i,j,k=1}^{p}\frac{\partial L_n(\boldsymbol{\beta}^*)}{\partial\beta_i\partial\beta_j\partial\beta_k}u_iu_ju_k\alpha_n^3 = \alpha_n^3\sum_{i=1}^{n}\frac{1}{6}\varphi'''(\boldsymbol{x}_i\boldsymbol{\beta}^*)(\boldsymbol{x}_i^T\boldsymbol{u})^3.
\end{aligned}$$

These equalities can be directly seen by straightforward calculations. Plugging in that $\alpha_n = \frac{1}{\sqrt{n}}$ (the quantity α_n is just introduced here for consistency with the proof of Theorem 3.6), we obtain the following asymptotics using (Reg1)–(Reg3)

$$\begin{aligned}
T_{1,n} &= -\sum_{i=1}^{n}[y_i - \varphi'(\boldsymbol{x}_i\boldsymbol{\beta}^*)]\frac{\boldsymbol{x}_i^T\boldsymbol{u}}{\sqrt{n}} \to_d N(\mathbf{0}, \boldsymbol{u}^T I_F(\boldsymbol{\beta}^*)\boldsymbol{u}) \text{ (using CLT)}, \\
T_{2,n} &= \sum_{i=1}^{n}\frac{1}{2}\varphi''(\boldsymbol{x}_i\boldsymbol{\beta}^*)\boldsymbol{u}^T\frac{\boldsymbol{x}_i\boldsymbol{x}_i^T}{n}\boldsymbol{u} \to_p \frac{1}{2}\boldsymbol{u}^T I_F(\boldsymbol{\beta}^*)\boldsymbol{u} \text{ (using LLN)}, \\
T_{3,n} &= n^{-1/2}\frac{1}{6}\underbrace{\frac{1}{n}\sum_{i=1}^{n}\varphi'''(\boldsymbol{x}_i\boldsymbol{\beta}^*)(\boldsymbol{x}_i^T\boldsymbol{u})^3}_{\to_p \mathbb{E}(M(\boldsymbol{x})|\boldsymbol{x}^T\boldsymbol{u}|^3) < \infty \text{ by (Reg3)}} \quad \text{(using LLN)}
\end{aligned} \tag{23}$$

thus $6\sqrt{n}T_{3,n} < \infty$, see also [42] (proof of Theorem 4). With these properties we can conclude that the likelihood part of the objective function, hence (22), is asymptotically dominated by (23), thus by the expression $\boldsymbol{u}^T I_F(\boldsymbol{\beta}^*)\boldsymbol{u}$.
Since by the triangle inequality $||\boldsymbol{\beta}_j^* - \frac{1}{\sqrt{n}}\boldsymbol{u}||_2 \leq ||\boldsymbol{\beta}_j^*||_2 + ||\frac{1}{\sqrt{n}}\boldsymbol{u}||_2$ we obtain $||\boldsymbol{\beta}_j^* - \frac{1}{\sqrt{n}}\boldsymbol{u}||_2 - ||\boldsymbol{\beta}_j^*||_2 \leq ||\frac{1}{\sqrt{n}}\boldsymbol{u}||_2$. Consequently we admit

$$\lambda_1^n\sum_{j=1}^{J}w_1^{(j)}\left(||\boldsymbol{\beta}_j^* + \frac{1}{\sqrt{n}}\boldsymbol{u}||_2 - ||\boldsymbol{\beta}_j^*||_2\right) \leq a_n^1\frac{1}{\sqrt{n}}||\boldsymbol{u}||_2 J.$$

Therefore, $\lambda_1^n\sum_{j=1}^{J}w_1^{(j)}\left(||\boldsymbol{\beta}_j^* + \frac{1}{\sqrt{n}}\boldsymbol{u}||_2 - ||\boldsymbol{\beta}_j^*||_2\right) = o_p(1)||\boldsymbol{u}||$, which is also

obtained in [34]. Further it holds

$$\lambda_0^n \sum_{j=1}^{J} \sum_{0\leq r<s\leq p_j} w_0^{(j,rs)} \left(||\beta_{j,r}^* - \beta_{j,s}^* + \frac{1}{\sqrt{n}}(u_r - u_s)||_0 - ||\beta_{j,r}^* - \beta_{j,s}^*||_0 \right) \quad (24)$$

$$\leq \sum_{j=1}^{J} \sum_{0\leq r<s\leq p_j} \underbrace{w_0^{(j,rs)} \lambda_0^n}_{\leq a_n^0 = O_p(1)} \quad (25)$$

giving us that (24) is $O_p(1)$. Note that p and J are fixed in this theorem thus they do not grow with the sample size n. All in all, we can write

$$\begin{aligned}
& M_{pen}\left(\boldsymbol{\beta}^* + \frac{1}{\sqrt{n}}\right) - M_{pen}(\boldsymbol{\beta}) \\
= \; & T_{1,n} + T_{2,n} + T_{3,n} + \lambda_1^n \sum_{j=1}^{J} w_1^{(j)} \left(||\boldsymbol{\beta}_j^* + \frac{1}{\sqrt{n}}\boldsymbol{u}||_2 - ||\boldsymbol{\beta}_j^*||_2 \right) \\
& + \lambda_0^n \sum_{j=1}^{J} \sum_{0\leq r<s\leq p_j} w_0^{(j,rs)} \left(||\beta_{j,r}^* - \beta_{j,s}^* + \frac{1}{\sqrt{n}}(u_r - u_s)||_0 - ||\beta_{j,r}^* - \beta_{j,s}^*||_0 \right) \\
= \; & \underbrace{T_{1,n}}_{\to_d N(\ldots)} + \underbrace{T_{2,n}}_{\to_p \frac{1}{2}\boldsymbol{u}^T I_F(\boldsymbol{\beta}^*)\boldsymbol{u}} + \underbrace{T_{3,n}}_{\text{bounded}} + o_p(1)||\boldsymbol{u}|| + O_p(1) \quad (26)
\end{aligned}$$

We conclude that the expression $H_n(\boldsymbol{u}) = M_{pen}(\boldsymbol{\beta}^* + \frac{1}{\sqrt{n}}) - M_{pen}(\boldsymbol{\beta})$ is dominated (asymptotically) by $\frac{1}{2}\boldsymbol{u}^T I_F(\boldsymbol{\beta}^*)\boldsymbol{u} > 0$ where this expression is positive since the Fisher information matrix was assumed to be positive definite at $\boldsymbol{\beta}^*$. Hence, for n large enough, we can choose c in such a way (in particular it has to be large enough) that (26) > 0 hence (20) holds so there exists a local minimizer $\hat{\boldsymbol{\beta}}$ being $\sqrt{n}$-consistent. □

A.1.3. Proof of Theorem 3.6

Remark A.2 (Initial Remark on Theorem 3.6)**.** In the assumptions of Theorem 3.6, instead of $a_n^0 J_n p_n(p_n - 1) = o_p(1)$, one could also assume that (i) the number of levels is bounded, hence $\max\{p_j | j = 1, \ldots, J_n\} = c_{\text{levels}} < \infty$ for some constant $c_{\text{levels}} > 0$ and (ii) $a_n^0 J_n = o_p(1)$, see $(**)$ in the following proof.

Proof. The proof is related to the proof of Theorem 1 in [7], where such a theorem is shown for nonconcave penalties as SCAD. We will transfer the idea to our case of L_0-FGL. In [7] (Theorem 1) the weights are chosen to be equal to one. The first part of the proof, where the log-likelihood is observed, is similar to [7]. As in the proof of Theorem 3.5, we will show that for any given $\varepsilon > 0$, (20) is satisfied, where we replace $\frac{1}{\sqrt{n}}$ by α_n. As in the proof of Theorem 3.9, we define $H_n(\boldsymbol{u}) := M_{pen}(\boldsymbol{\beta}^* + \alpha_n \boldsymbol{u}) - M_{pen}(\boldsymbol{\beta}^*)$, where in the proof of Theorem 3.9

α_n corresponds to $\frac{1}{\sqrt{n}}$, and obtain

$$H_n(\boldsymbol{u}) = -L_n(\boldsymbol{\beta}^* + \alpha_n \boldsymbol{u}) + L_n(\boldsymbol{\beta}^*) + \lambda_1^n \sum_{j=1}^{J_n} (w_1^{(j)} ||\boldsymbol{\beta}_j^* + \alpha_n \boldsymbol{u}_j||_2 - w_1^{(j)} ||\boldsymbol{\beta}_j||_2)$$
$$+\lambda_0^n \sum_{j=1}^{J_n} \sum_{0 \le r < s \le p_j} (w_j^{(j,rs)} ||\beta_{j,r}^* - \beta_{j,s}^* + \alpha_n (u_r - u_s)||_0 - w_0^{(j,rs)} ||\beta_{j,r}^* - \beta_{j,s}^*||_0). \tag{27}$$

We will observe the log-likelihood part and the penalty part of the objective function separately.

Step 1 (Log Liklelihood): for the log-likelihood part we perform a Taylor expansion as in the proofs of Theorems 3.5 and 3.9 but since we are in the case that p_n grows with n, the asymptotic behavior of the components of the Taylor expansion will differ from the mentioned theorems. In particular, we get for the Taylor expansion of $f_n(\boldsymbol{u}) := -L_n(\boldsymbol{\beta}^* + \alpha_n \boldsymbol{u}) + L_n(\boldsymbol{\beta}^*)$ around $\boldsymbol{u} = \boldsymbol{0}$ using the fact that $f_n(\boldsymbol{0}) = 0$

$$-L_n(\boldsymbol{\beta}^* + \alpha_n \boldsymbol{u}) + L_n(\boldsymbol{\beta}^*) = T_{1,n} + T_{2,n} + T_{3,n}. \tag{28}$$

Here, the quantities $T_{1,n}, T_{2,n}, T_{3,n}$ are similar to the proof of Theorem 3.5 with general α_n. For $T_{1,n}$ we get using the Cauchy-Schwartz inequality (C.S.) and (div.Reg2)

$$\begin{aligned} |T_{1,n}| &= |\alpha_n \nabla^T L_n(\boldsymbol{\beta}^*) \boldsymbol{u}| \overset{\text{(C.S.)}}{\le} \alpha_n ||\nabla^T L_n(\boldsymbol{\beta}^*)||_2 \, ||\boldsymbol{u}||_2 \\ &\overset{\text{(div.Reg2)}}{=} O_p(\alpha_n \sqrt{n p_n}) ||\boldsymbol{u}||_2 = O_p(\alpha_n^2 n) ||\boldsymbol{u}||_2 = O_p(p_n) ||\boldsymbol{u}||_2 \end{aligned}$$

since

$$\begin{aligned} ||\nabla^T L_n(\boldsymbol{\beta}^*)||_2^2 &= \sum_{j=1}^{p_n} \frac{\partial L_n(\boldsymbol{\beta}^*)}{\partial \beta_j} \frac{\partial L_n(\boldsymbol{\beta}^*)}{\partial \beta_j} \\ &= n \sum_{j=1}^{p_n} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \frac{\partial \log f_n(v_i, \boldsymbol{\beta}^*)}{\partial \beta_j} \frac{\partial \log f_n(v_i, \boldsymbol{\beta}^*)}{\partial \beta_j}}_{\to_p \, \mathbb{E}\left(\frac{\partial \log f_n(v_i, \boldsymbol{\beta}^*)}{\partial \beta_j} \frac{\partial \log f_n(v_i, \boldsymbol{\beta}^*)}{\partial \beta_j} \right) = [\boldsymbol{I}_F(\boldsymbol{\beta}^*)]_{j,j} < C} \\ &= p_n \, n \, O_p(1). \end{aligned}$$

Hence $||\nabla^T L_n(\boldsymbol{\beta}^*)||_2 = O_p(\sqrt{n \, p_n})$. In particular we can write $T_{1,n} = O_p(p_n)$ $||\boldsymbol{u}||_2$ since $\alpha_n^2 n = p_n = \alpha_n \sqrt{n p_n}$. For the second summand $T_{2,n}$ it holds as in [7]

$$\begin{aligned} T_{2,n} &= -\frac{1}{2} \boldsymbol{u}^T \nabla^2 L_n(\boldsymbol{\beta}^*) \boldsymbol{u} \alpha_n^2 \\ &= -\frac{1}{2} \boldsymbol{u}^T \nabla^2 L_n(\boldsymbol{\beta}^*) \boldsymbol{u} \alpha_n^2 \underbrace{+ \frac{1}{2} \boldsymbol{u}^T \boldsymbol{I}_F(\boldsymbol{\beta}^*) \boldsymbol{u} n \alpha_n^2 - \frac{1}{2} \boldsymbol{u}^T \boldsymbol{I}_F(\boldsymbol{\beta}^*) \boldsymbol{u} n \alpha_n^2}_{=0} \end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{2}\boldsymbol{u}^T\left[\frac{1}{n}(\nabla^2 L_n(\boldsymbol{\beta}^*) + \boldsymbol{I}_F(\boldsymbol{\beta}^*))\right]\boldsymbol{u}n\alpha_n^2 + \frac{1}{2}\boldsymbol{u}^T\boldsymbol{I}_F(\boldsymbol{\beta}^*)\boldsymbol{u}n\alpha_n^2 \\
&= -\frac{1}{2}\boldsymbol{u}^T\left[\frac{1}{n}(\nabla^2 L_n(\boldsymbol{\beta}^*) - E(\nabla^2 L_n(\boldsymbol{\beta}^*))\right]\boldsymbol{u}n\alpha_n^2 + \frac{1}{2}\boldsymbol{u}^T\boldsymbol{I}_F(\boldsymbol{\beta}^*)\boldsymbol{u}n\alpha_n^2 \\
&= \frac{n}{2}\alpha_n^2\boldsymbol{u}^T o_p(1/p_n)\boldsymbol{u} + \frac{n}{2}\alpha_n^2\boldsymbol{u}^T\boldsymbol{I}_F(\boldsymbol{\beta}^*)\boldsymbol{u}
\end{aligned}
$$

where we used that $\boldsymbol{I}_F(\boldsymbol{\beta}^*) = -E(\nabla^2 L_n(\boldsymbol{\beta}^*))$ and $||\frac{1}{n}\nabla^2 L_n(\boldsymbol{\beta}^*) + \boldsymbol{I}_F(\boldsymbol{\beta}^*)|| = o_p(\frac{1}{p_n})$ following Lemma 8 of [7] which needs the assumption $p_n^4/n \to 0$ as $n \to \infty$. Since we have $||\frac{1}{n}\nabla^2 L_n(\boldsymbol{\beta}^*) + \boldsymbol{I}_F(\boldsymbol{\beta}^*)|| = o_p(\frac{1}{p_n})$ we know that, by definition, $||\frac{1}{n}\nabla^2 L_n(\boldsymbol{\beta}^*) + \boldsymbol{I}_F(\boldsymbol{\beta}^*)||\, p_n$ converges to zero in probability so using $p_n \geq 1$ we get $||\frac{1}{n}\nabla^2 L_n(\boldsymbol{\beta}^*) + \boldsymbol{I}_F(\boldsymbol{\beta}^*)||\, p_n \geq ||\frac{1}{n}\nabla^2 L_n(\boldsymbol{\beta}^*) + \boldsymbol{I}_F(\boldsymbol{\beta}^*)||$ hence $||\frac{1}{n}\nabla^2 L_n(\boldsymbol{\beta}^*) + \boldsymbol{I}_F(\boldsymbol{\beta}^*)|| = o_p(1)$. Consequently,

$$
T_{2,n} = \frac{n}{2}\alpha_n^2\boldsymbol{u}^T o_p(1)\boldsymbol{u} + \frac{n}{2}\alpha_n^2\boldsymbol{u}^T\boldsymbol{I}_F(\boldsymbol{\beta}^*)\boldsymbol{u} = \frac{1}{2}p_n\boldsymbol{u}^T(\boldsymbol{I}_F(\boldsymbol{\beta}^*) + o_p(1))\boldsymbol{u}.
$$

The last summand $T_{3,n}$ is treated as follows.

$$
\begin{aligned}
|T_{3,n}| &= \frac{1}{6}\left|\sum_{i,j,k=1}^{p_n}\frac{\partial^3 L_n(\boldsymbol{\beta}^*)}{\partial\beta_i\partial\beta_j\partial\beta_k}u_iu_ju_k\alpha_n^3\right| = \frac{1}{6}\left|\sum_{l=1}^{n}\sum_{i,j,k}^{p_n}\frac{\partial^3 \log f_n(\boldsymbol{v}_l,\boldsymbol{\beta}^*)}{\partial\beta_i\partial\beta_j\partial\beta_k}u_iu_ju_k\alpha_n^3\right| \\
&\leq \frac{1}{6}\alpha_n^3\underbrace{\sum_{l=1}^{n}\left|\sum_{i,j,k}^{p_n}\frac{\partial^3 \log f_n(\boldsymbol{v}_l,\boldsymbol{\beta}^*)}{\partial\beta_i\partial\beta_j\partial\beta_k}u_iu_ju_k\right|}_{(*)},
\end{aligned}
$$

where, using the Cauchy Schwarz inequality we obtain

$$
(*) = \sum_{l=1}^{n}\left|\sum_{i,j,k}^{p_n}\frac{\partial^3 \log f_n(\boldsymbol{v}_l,\boldsymbol{\beta}^*)}{\partial\beta_i\partial\beta_j\partial\beta_k}u_iu_ju_k\right| \overset{\text{(C.S.)}}{\leq} \sum_{l=1}^{n}||(\log f_n(\boldsymbol{v}_l,\boldsymbol{\beta}^*))'''||_2 \cdot ||\boldsymbol{u}||_2^3.
$$

Following (div.Reg3), we know that we can bound every component in $||(\log f_n(\boldsymbol{v}_l,\boldsymbol{\beta}^*))'''||_2$ by some function $M_{n,i,j,k}(\boldsymbol{x}_l)$, hence

$$
\sum_{l=1}^{n}||(\log f_n(\boldsymbol{v}_l,\boldsymbol{\beta}^*))'''||_2 \leq \sum_{l=1}^{n}\left(\sum_{i,j,k=1}^{p_n}M_{n,i,j,k}^2(\boldsymbol{x}_l)\right)^{1/2}
$$

so consequently

$$
(*) \leq \sum_{l=1}^{n}||(\log f_n(\boldsymbol{v}_l,\boldsymbol{\beta}^*))'''||_2 \cdot ||\boldsymbol{u}||_2^3 \leq ||\boldsymbol{u}||_2^3\sum_{l=1}^{n}\left(\sum_{i,j,k=1}^{p_n}M_{n,i,j,k}^2(\boldsymbol{x}_l)\right)^{1/2}. \tag{29}
$$

Now we have to observe the asymptotic behavior of the r.h.s. of the inequality (29). Using the Cauchy Schwarz inequality we obtain

$$\left(\sum_{i,j,k=1}^{p_n} M^2_{n,i,j,k}(\boldsymbol{x}_l)\cdot 1\right)^2 \leq \left(\sum_{i,j,k=1}^{p_n} M^2_{n,i,j,k}(\boldsymbol{x}_l)\right) p_n^3$$

$$\Rightarrow \sum_{i,j,k=1}^{p_n} M^2_{n,i,j,k}(\boldsymbol{x}_l) \leq p_n^3 \Rightarrow \left(\sum_{i,j,k=1}^{p_n} M^2_{n,i,j,k}(\boldsymbol{x}_l)\right)^{1/2} \leq p_n^{3/2} \quad (30)$$

With (30) we can write using $\alpha_n = \sqrt{p_n/n}$

$$|T_{3,n}| \leq \frac{1}{6}\alpha_n^3 ||\boldsymbol{u}||_2^3 \sum_{l=1}^{n}\left(\sum_{i,j,k=1}^{p_n} M^2_{n,i,j,k}(\boldsymbol{x}_l)\right)^{1/2}$$

$$\leq \frac{1}{6}\alpha_n^3 ||\boldsymbol{u}||_2^3 \sum_{l=1}^{n} p_n^{3/2} = \frac{1}{6}\alpha_n^3 ||\boldsymbol{u}||_2^3 n p_n^{3/2} = \frac{1}{6}||\boldsymbol{u}||_2^3 \frac{p_n^3}{\sqrt{n}}$$

Since we assumed $\frac{p_n^4}{n} \to 0$ we get using $0 \leq \frac{p_n^2}{\sqrt{n}} = \sqrt{\frac{p_n^4}{n}} \leq \frac{p_n^4}{n} \to 0$ that $\frac{p_n^2}{\sqrt{n}} \to 0$. Consequently, it holds that $\frac{p_n^3}{\sqrt{n}} = o_p(p_n)$. So $T_{3,n} = o_p(p_n)||\boldsymbol{u}||_2^3$.

Step 2 (Penalty): it holds that

$$|\lambda_1^n \sum_{j=1}^{J_n}(w_1^{(j)}||\boldsymbol{\beta}_j^* + \alpha_n \boldsymbol{u}_j||_2 - w_1^{(j)}||\boldsymbol{\beta}_j^*||_2)|$$

$$\leq \lambda_1^n \sum_{j=1}^{J_n} w_1^{(j)} \alpha_n ||\boldsymbol{u}_j||_2 \leq ||\boldsymbol{u}||_2 \, \alpha_n \sum_{j=1}^{J_n} \lambda_1^n w_1^{(j)} \leq ||\boldsymbol{u}||_2 \, \alpha_n a_n^1 J_n = o_p(1)||\boldsymbol{u}||_2$$

since by assumption $\alpha_n a_n^1 J_n = o_p(1)$. Lastly, since $||\beta^*_{j,r} - \beta^*_{j,s} + \alpha_n(u_r - u_s)||_0 - ||\beta^*_{j,r} - \beta^*_{j,s}||_0 \leq 1$, we obtain

$$\lambda_0^n \sum_{j=1}^{J_n} \sum_{0\leq r<s\leq p_j} (w_0^{(j,rs)}||\beta^*_{j,r} - \beta^*_{j,s} + \alpha_n(u_r - u_s)||_0 - w_0^{(j,rs)}||\beta^*_{j,r} - \beta^*_{j,s}||_0)$$

$$\leq \sum_{j=1}^{J_n} \sum_{0\leq r<s\leq p_j} \lambda_0^n w_j^{(j,rs)} \leq \; a_n^0 \sum_{j=1}^{J_n} \sum_{0\leq r<s\leq p_j} 1 := (**)$$

The quantity $\sum_{j=1}^{J_n}\sum_{0\leq r<s\leq p_j} 1$ is equal to the number of differences including all J_n predictors of the model. Of course, this depends on the design whether we observe ordinal or nominal factors, or mixtures. The highest number of possible differences occurs when all factors are nominal, hence it can be bounded by $\frac{p_n(p_n-1)}{2}$. Additionally, we assumed that $a_n^0 p_n(p_n-1) = o_p(1)$, hence

$$(**) = a_n^0 \sum_{j=1}^{J_n} \sum_{0\leq r<s\leq p_j} 1 \leq a_n^0 \frac{p_n(p_n-1)}{2} = o_p(1)$$

so the L_0 part of the penalty function is $o_p(1)$.
We conclude with the steps above, as well as (27) and (28)

$$H_n(\boldsymbol{u}) = \underbrace{T_{1,n}}_{=O_p(p_n)||\boldsymbol{u}||} + \underbrace{T_{2,n}}_{=\frac{1}{2}p_n\boldsymbol{u}^T(\boldsymbol{I}_F(\boldsymbol{\beta}^*)+o_p(1))\boldsymbol{u}.} + \underbrace{T_{3,n}}_{=o_p(p_n)||\boldsymbol{u}||^2} + o_p(1)||\boldsymbol{u}||^2 + o_p(1)$$

We can see that all the summands are dominated by $\frac{1}{2}p_n\boldsymbol{u}^T\boldsymbol{I}_F(\boldsymbol{\beta}^*)\boldsymbol{u} > 0$ where the last inequality holds since we assumed that the Fisher information matrix is positive definite in $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ in (div.Reg2). So choosing c large enough, we can ensure that $H_n(\boldsymbol{u}) > 0$. □

A.1.4. Proof of Theorem 3.9

Proof. The proof follows [42] where the oracle properties for the adaptive lasso are shown. We write $\boldsymbol{\beta} = \boldsymbol{\beta}^* + \frac{\boldsymbol{u}}{\sqrt{n}}$ and remember that $H_n(\boldsymbol{u}) := M_{pen}\left(\boldsymbol{\beta}^* + \frac{\boldsymbol{u}}{\sqrt{n}}\right) - M_{pen}(\boldsymbol{\beta}^*)$. We aim to minimize $\hat{\boldsymbol{u}} = \arg\min_{\boldsymbol{u}} H_n(\boldsymbol{u})$, then $\hat{\boldsymbol{u}} = \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$. Although $\hat{\boldsymbol{u}}$ depends on n, we neglect to use a lower index in $\hat{\boldsymbol{u}}$ for simplicity, as we do with $\hat{\boldsymbol{\beta}}$. Since $H_n(\boldsymbol{u})$ is the same as (21) in the proof of Theorem 3.5, the first steps performing a Taylor expansion of the log-likelihood resulting in $T_{1,n}, T_{2,n}, T_{3,n}$ and observing the asymptotic behavior of those using the regularity conditions are similar. Hence, it remains to analyze the asymptotic behavior of the penalties, which we denote by $P^{GL}_{\lambda_1^n}$ and $P^{L_0}_{\lambda_0^n}$ in this proof for simplicity. Writing

$$\begin{aligned} P^{GL}_{\lambda_1^n} &= \lambda_1^n \sum_{j=1}^{J}\left[w_1^{(j)}||\boldsymbol{\beta}_j^* + \frac{\boldsymbol{u}_j}{\sqrt{n}}||_2 - w_1^{(j)}||\boldsymbol{\beta}_j^*||_2\right] \\ &= \lambda_1^n \sum_{j\in A^*}\left[w_1^{(j)}||\boldsymbol{\beta}_j^* + \frac{\boldsymbol{u}_j}{\sqrt{n}}||_2 - w_1^{(j)}||\boldsymbol{\beta}_j^*||_2\right] \\ &\quad + \lambda_1^n \sum_{j\notin A^*}\left[w_1^{(j)}||\boldsymbol{\beta}_j^* + \frac{\boldsymbol{u}_j}{\sqrt{n}}||_2 - w_1^{(j)}||\boldsymbol{\beta}_j^*||_2\right] \end{aligned}$$

we get for the case that $\boldsymbol{\beta}_j^* \neq \boldsymbol{0}$, hence $j \in A^*$, the following (recall that $\tilde{\boldsymbol{\beta}}$ is the unpenalized MLE)

$$w_1^{(j)} = \frac{1}{||\tilde{\boldsymbol{\beta}}_j||_2^{\gamma}} \to_p ||\boldsymbol{\beta}_j^*||_2^{-\gamma}$$

$$\sqrt{n}\{||\boldsymbol{\beta}_j^* + \frac{\boldsymbol{u}_j}{\sqrt{n}}||_2 - ||\boldsymbol{\beta}_j^*||_2\} \leq \sqrt{n}\{||\boldsymbol{\beta}_j^*||_2 + \frac{1}{\sqrt{n}}||\boldsymbol{u}_j||_2 - ||\boldsymbol{\beta}_j^*||_2\} = ||\boldsymbol{u}_j||_2 < \infty$$

$$\sqrt{n}\{||\boldsymbol{\beta}_j^* + \frac{\boldsymbol{u}_j}{\sqrt{n}}||_2 - ||\boldsymbol{\beta}_j^*||_2\} \geq \sqrt{n}\{||\boldsymbol{\beta}_j^*||_2 - ||\frac{\boldsymbol{u}_j}{\sqrt{n}}||_2 - ||\boldsymbol{\beta}_j^*||_2\} = -||\boldsymbol{u}_j||_2 > -\infty$$

This gives us

$$-||\boldsymbol{u}_j||_2 \leq \sqrt{n}\{||\boldsymbol{\beta}_j^* + \frac{\boldsymbol{u}_j}{\sqrt{n}}||_2 - ||\boldsymbol{\beta}_j^*||_2\} \leq ||\boldsymbol{u}_j||_2. \tag{31}$$

Using Slutsky we end up with (note that we call the summands of $P^{GL}_{\lambda_1^n}$ for $j \in A^*$ in the following $\left(P^{GL}_{\lambda_1^n}\right)_{A^*}$, the analogous notation is used for the summands where $j \notin A^*$)

$$\begin{aligned}\left(P^{GL}_{\lambda_1^n}\right)_{A^*} &= \lambda_1^n \sum_{j\in A^*}\left[w_1^{(j)}||\boldsymbol{\beta}_j^* + \frac{\boldsymbol{u}_j}{\sqrt{n}}||_2 - w_1^{(j)}||\boldsymbol{\beta}_j^*||_2\right] \\ &= \underbrace{\frac{\lambda_1^n}{\sqrt{n}}}_{\to 0}\sum_{j\in A^*}\underbrace{w_1^{(j)}}_{\to_p ||\boldsymbol{\beta}_j^*||_2^{-\gamma}}||\boldsymbol{\beta}_j^*||^{-\gamma}\underbrace{\sqrt{n}\left[||\boldsymbol{\beta}_j^* + \frac{\boldsymbol{u}_j}{\sqrt{n}}||_2 - ||\boldsymbol{\beta}_j^*||_2\right]}_{\text{bounded using (31)}} \to_p 0\end{aligned}$$

for the case that $\boldsymbol{\beta}_j^* \neq \mathbf{0}$. Now we come to the case that $\boldsymbol{\beta}_j^* = \mathbf{0}$, hence $j \notin A^*$. In this case we get with $w_1^{(j)} = ||\tilde{\boldsymbol{\beta}}_j||_2^{-\gamma} = n^{\gamma/2}||\tilde{\boldsymbol{\beta}}_j\sqrt{n}||_2^{-\gamma}$

$$\begin{aligned}\left(P^{GL}_{\lambda_1^n}\right)_{(A^*)^c} &= \lambda_1^n \sum_{j\in(A^*)^c} w_1^{(j)}||\frac{\boldsymbol{u}_j}{\sqrt{n}}||_2 = \frac{\lambda_1^n}{\sqrt{n}}\sum_{j\in(A^*)^c}\sqrt{n}w_1^{(j)}||\frac{\boldsymbol{u}_j}{\sqrt{n}}||_2 \\ &= \underbrace{\lambda_1^n n^{(\gamma-1)/2}}_{\to\infty}\sum_{j\in(A^*)^c}\underbrace{||\tilde{\boldsymbol{\beta}}_j\sqrt{n}||_2^{-\gamma}}_{O_p(1)}||\boldsymbol{u}_j||_2,\end{aligned}$$

which goes to ∞ for $||\boldsymbol{u}_j||_2 \neq 0$ and equals 0 otherwise. Hence we get for the j-th summand of $P^{GL}_{\lambda_1^n}$, denoted by $P^{GL}_{\lambda_1^n,j}$, for $n \to \infty$

$$P^{GL}_{\lambda_1^n,j} \to_p \begin{cases} 0 & \text{if } ||\boldsymbol{u}_j||_2 = 0,\ \boldsymbol{\beta}_j^* = \mathbf{0}, \\ \infty & \text{if } ||\boldsymbol{u}_j||_2 \neq 0,\ \boldsymbol{\beta}_j^* = \mathbf{0}, \\ 0 & \text{if } \boldsymbol{\beta}_j^* \neq \mathbf{0}. \end{cases} \tag{32}$$

As in the proof of Theorem 3.5 using (24) we get (in this theorem we require $a_n^0 = o_p(1)$ where in Theorem 3.5 we require $a_n^0 = O_p(1)$)

$$P^{L_0}_{\lambda_0^n} \leq \underbrace{\sum_{j=1}^{J}\sum_{0\leq r<s\leq p_j}\underbrace{w_0^{(j,rs)}\lambda_0^n}_{\leq a_n^0 = o_p(1)}}_{=o_p(1)}\ p \Rightarrow P^{L_0}_{\lambda_0^n} \to_p 0.$$

To sum up

$$H_n(\boldsymbol{u}) = \underbrace{T_{1,n}}_{\to_d \boldsymbol{u}^T N(\mathbf{0},I_F(\boldsymbol{\beta}^*))} + \underbrace{T_{2,n}}_{\to_p \frac{1}{2}\boldsymbol{u}^T\boldsymbol{I}_F(\boldsymbol{\beta}^*)\boldsymbol{u}} + \underbrace{P^{GL}_{\lambda_1^n}}_{\text{see (32)}} + \underbrace{T_{3,n}}_{\to_p 0} + \underbrace{P^{L_0}_{\lambda_0^n}}_{\to_p 0}$$

holds which yields (note that for $j \in A^*$ it holds that $||\boldsymbol{\beta}_j^*||_2 \neq \mathbf{0}$, thus $\boldsymbol{\beta}_j^* \neq \mathbf{0}$, and for $j \notin A^*$ it holds that $||\boldsymbol{\beta}_j^*||_2 = \mathbf{0}$, thus $\boldsymbol{\beta}_j^* = \mathbf{0}$, by definition of the active set A^*)

$$H_n(\boldsymbol{u}) \to_d H(\boldsymbol{u}) = \begin{cases} \boldsymbol{u}_{A^*}^T I_{11}\boldsymbol{u}_{A^*} - 2\boldsymbol{u}_{A^*}^T\boldsymbol{W}_{A^*}, & \text{if } \boldsymbol{u}_j = \mathbf{0}\,\forall j \notin A^* \\ \infty, & \text{otherwise} \end{cases}$$

where $\boldsymbol{W} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_F(\boldsymbol{\beta}^*))$ and consequently $\boldsymbol{W}_{A^*} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{11})$. The unique minimum of $H(\boldsymbol{u})$ is clearly at $\boldsymbol{u}^{\min} = (\boldsymbol{u}^{\min}_{A^*}, \boldsymbol{u}^{\min}_{(A^*)^c}) = (\boldsymbol{I}_{11}^{-1}\boldsymbol{W}_{A^*}, \mathbf{0})^T$ which can be obtained by straightforward calculations. We conclude that there exists an $\hat{\boldsymbol{u}} = \arg\min_{\boldsymbol{u}} H_n(\boldsymbol{u})$ satisfying $\hat{\boldsymbol{u}}_{A^*} \to_d \boldsymbol{I}_{11}^{-1}\boldsymbol{W}_{A^*}$ and $\hat{\boldsymbol{u}}_{(A^*)^c} \to_d 0$ with the same arguments provided in [16] for the non-convex Bridge estimator. Now, as we can see above, $\boldsymbol{W}_{A^*} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{11})$, so the asymptotic normality follows. □

A.1.5. Proof of Theorem 3.11

Proof. The proof corresponds to [36] (Theorem 2.3), where a related theorem is shown for adaptive group lasso. First, we introduce the following abbreviations, solely for this proof

$$f_1(\boldsymbol{\beta}) := \lambda_n^1 \sum_{j=1}^{J_n} w_1^{(j)} ||\boldsymbol{\beta}_j||_2 \text{ and } f_0(\boldsymbol{\beta}) := \lambda_n^0 \sum_{j=1}^{J_n} \sum_{0 \le r < s \le p_j} w_0^{(j,rs)} N(\beta_{j,r} - \beta_{j,s}).$$

Using an approximate L_0-FGL estimator $\hat{\boldsymbol{\beta}}$ being a minimum of $\widetilde{M}_{pen}(\boldsymbol{\beta})$ by definition, we know that $\nabla \widetilde{M}_{pen}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$, giving us

$$\frac{\partial L_n(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}_{A_n^*}} = \frac{\partial f_1(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}_{A_n^*}} + \frac{\partial f_0(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}_{A_n^*}}.$$

A Taylor expansion of the left hand side $\frac{\partial L_n(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}_{A_n^*}}$ and re-arranging gives us

$$\begin{aligned} &\frac{\partial f_1(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}_{A_n^*}} + \frac{\partial f_0(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}_{A_n^*}} \\ &= \frac{\partial L_n(\boldsymbol{\beta}^*_{A_n^*})}{\partial \boldsymbol{\beta}_{A_n^*}} + \nabla_{A_n^*} \left(\frac{\partial L_n(\boldsymbol{\beta}^*_{A_n^*})}{\partial \beta_{A_n^*}} \right)^T (\hat{\boldsymbol{\beta}}_{A_n^*} - \boldsymbol{\beta}^*_{A_n^*}) \\ &\quad + \frac{1}{2} (\hat{\boldsymbol{\beta}}_{A_n^*} - \boldsymbol{\beta}^*_{A_n^*})^T \nabla^2_{A_n^*} \left(\frac{\partial L_n(\boldsymbol{\xi}_{A_n^*})}{\partial \boldsymbol{\beta}_{A_n^*}} \right) (\hat{\boldsymbol{\beta}}_{A_n^*} - \boldsymbol{\beta}^*_{A_n^*}), \end{aligned} \tag{33}$$

where $\boldsymbol{\xi}_{A_n^*}$ is between $\hat{\boldsymbol{\beta}}_{A_n^*}$ and $\boldsymbol{\beta}^*_{A_n^*}$ Since the assumed regularity conditions of [36] also hold in our case, in particular (div.Reg2)–(div.Reg4) for the following equality, and $\hat{\boldsymbol{\beta}}$ is consistent (see Theorem 3.6 and the explanations right before Theorem 3.11), we can follow the lines of [36], equation (4.5), giving us

$$\left|\left| \frac{1}{2} (\hat{\boldsymbol{\beta}}_{A_n^*} - \boldsymbol{\beta}^*_{A_n^*})^T \nabla^2_{A_n^*} \left(\frac{\partial L_n(\boldsymbol{\xi}_{A_n^*})}{\partial \boldsymbol{\beta}_{A_n^*}} \right) (\hat{\boldsymbol{\beta}}_{A_n^*} - \boldsymbol{\beta}^*_{A_n^*}) \right|\right|_2^2 = o_p(1).$$

Similarly, we follow [36] (below equation (4.5)), where it is shown that for the sub-differential of the (adaptive) group Lasso part it holds $\left|\left| \frac{\partial f_1(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}_{A_n^*}} \right|\right|_2^2 = o_p\left(\frac{1}{n}\right)$. For the approximation of the L_0 part we also get a sub-differential. At this point

it gets clear why the approximation of the L_0 part is necessary, since otherwise it is not differentiable and not even continuous. Since $N(\xi)$ includes $|\xi|$, the function is not differentiable in zero. Nevertheless, it is sub-differentiable with $|\frac{\partial|\xi|}{\partial\xi}| \leq 1$. Consequently, it holds

$$\frac{\partial N(\xi)}{\partial \xi} = \frac{\gamma \exp(-\gamma|\xi|)\frac{\partial|\xi|}{\partial\xi}}{(1+\exp(-\gamma|\xi|))^2} \Rightarrow \left|\left|\frac{\partial N(\xi)}{\partial \xi}\right|\right|_2^2 \leq \gamma.$$

This gives us

$$\begin{aligned}
\left|\left|\frac{\partial f_0(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}_{A_n^*}}\right|\right|_2^2 &= \left|\left|\lambda_{0,n}\sum_{j=1}^{J_n}\sum_{0\leq r<s\leq p_j} w_0^{(j,rs)}\frac{\partial}{\partial \boldsymbol{\beta}_{A_n^*}} N(\beta_{j,r}-\beta_{j,s})\right|\right|_2^2 \\
&= \left|\left|\lambda_{0,n}\sum_{j=1}^{j_{0,n}}\sum_{0\leq r<s\leq p_j} w_0^{(j,rs)}\frac{\partial}{\partial \boldsymbol{\beta}_{A_n^*}} N(\beta_{j,r}-\beta_{j,s})\right|\right|_2^2 \\
&\leq \left|\left|\frac{1}{2}p_{0,n}(p_{0,n}-1)\cdot\gamma\cdot \max_{j\in\{1,\dots j_{0,n}\},r,s\in\{0,\dots p_j\}} w_0^{(j,rs)}\lambda_{0,n}\right|\right|_2^2 \\
&\leq \left(|a_n^0|\gamma\frac{1}{2}p_{0,n}(p_{0,n}-1)\right)^2 = o_p(1)o_p(1) = o_p(1),
\end{aligned}$$

where the latter equalities hold since $p_{0,n} < p_n$ and $a_n^0 p_n(p_n-1) = o_p(1)$ by the assumptions of the consistency theorem (Theorem 3.6). We further see that $\nabla_{A_n^*}\left(\frac{\partial L_n(\boldsymbol{\beta}_{A_n^*}^*)}{\partial \beta_{A_n^*}}\right)^T = -I_{11,n}$. Putting all together we deduce by (33)

$$o_p(1) + \boldsymbol{e}_n\, I_{11,n}^{-1/2}\frac{\partial L_n(\boldsymbol{\beta}_{A_n^*}^*)}{\partial \boldsymbol{\beta}_{A_n^*}} = \boldsymbol{e}_n I_{11,n}^{1/2}(\hat{\boldsymbol{\beta}}_{A_n^*} - \boldsymbol{\beta}_{A_n^*}^*), \tag{34}$$

where we know by (div.Reg2) that the Fisher information matrix is finite. It remains to show that the second summand on the l.h.s. of (34) converges in distribution to a standard normal. Again, since the regularity conditions of [36] hold, we refer to their work showing that by Lindeberg-Feller

$$\boldsymbol{e}_n\, I_{11,n}^{-1/2}\frac{\partial L_n(\boldsymbol{\beta}_{A_n^*}^*)}{\partial \boldsymbol{\beta}_{A_n^*}} \to_d N(0,1)$$

as $n\to\infty$. Now the claim follows. □

A.1.6. Proof of Theorem 3.12

Proof. The beginning of the proof follows [5] (Proof of Lemma 3.1) but we will transfer the proof to the more general case of $\boldsymbol{\beta}_j$ being a sub-vector instead

of a single entry. Having that, we will use the proven $\sqrt{n}$-consistency of the estimator $\hat{\boldsymbol{\beta}}$ of Theorem 3.5 to show the inequality. We have that

$$\begin{aligned}\mathbb{P}(A^* \not\subseteq A_n) &\leq \mathbb{P}(j \notin A_n \text{ for some } j \in A^*)\\ &\leq \mathbb{P}(\hat{\boldsymbol{\beta}}_j = \mathbf{0} \text{ and } \boldsymbol{\beta}_j^* \neq \mathbf{0} \text{ for some } j \in A^*)\\ &\leq \mathbb{P}(||\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*||_2 = ||\boldsymbol{\beta}_j^*||_2 \text{ for some } j \in A^*)\\ &\leq \mathbb{P}(||\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*||_2 \geq \min_{l \in A^*} ||\boldsymbol{\beta}_l^*||_2 \text{ for some } j \in A^*)\\ &\leq \mathbb{P}(||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2 \geq \min_{l \in A^*} ||\boldsymbol{\beta}_l^*||_2). \end{aligned} \tag{35}$$

Note that $\min_{l \in A^*} ||\boldsymbol{\beta}_l^*||_2$ is a minimum over a bounded set, since we assumed that the true underlying structure is sparse, thus the minimum always exists. Now our goal is to bound (35) by some ε. Since we know from Theorem 3.5 that $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2 = O_p(1/\sqrt{n})$ we get that $\forall \varepsilon > 0$ there exists constants $M, \tilde{N} > 0$ such that

$$\mathbb{P}(||\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)||_2 > M) < \varepsilon \quad \forall n > \tilde{N}. \tag{36}$$

Hence, for $n > \tilde{N}$ we have $\mathbb{P}\left(||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2 > \frac{M}{\sqrt{n}}\right) < \varepsilon$. Now, with $\varepsilon > 0$ and constants $M, \tilde{N} > 0$, we can always choose some $N' > 0$ such that $\frac{M}{\sqrt{n}} \leq \min_{l \in A^*} ||\boldsymbol{\beta}_l^*||_2$ for all $n \geq N'$. Note that by definition we have that $||\boldsymbol{\beta}_l^*||_2 \neq 0 \; \forall l \in A^*$. Now we can write expression (35) as

$$\mathbb{P}(||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2 \geq \min_{l \in A^*} ||\boldsymbol{\beta}_l^*||_2) \leq \mathbb{P}\left(||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2 > \frac{M}{\sqrt{n}}\right) < \varepsilon \;\; \forall n > \max\{\tilde{N}, N'\}.$$

Consequently, $\forall \varepsilon > 0$ we can find some $N := \max\{\tilde{N}, N'\}$ such that

$$P(A^* \not\subseteq A_n) < \varepsilon \;\; \forall n > N$$

which completes the proof. □

A.1.7. Proof of Theorem 3.13

Proof. The idea of the proof works analogously to the proof of Theorem 3.12 replacing A^* by A_n^*, hence showing $\mathbb{P}(A_n^* \not\subseteq A_n) \leq \mathbb{P}(||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2 \geq \min_{l \in A_n^*} ||\boldsymbol{\beta}_l^*||_2)$. Having that, (36) is modified using Theorem 3.6 (here, $\hat{\boldsymbol{\beta}}$ is the minimizer from Theorem 3.6)

$$\mathbb{P}(||\alpha_n^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)||_2 > M) < \varepsilon \quad \forall n > \tilde{N}.$$

with $\alpha_n = \sqrt{\frac{p_n}{n}}$. Now, since by assumption there exists some $C > 0$ such that $\min_{l \in A_n^*} ||\boldsymbol{\beta}_l^*|| \geq C \; \forall n \in \mathbb{N}$, we can find $N' \in \mathbb{N}$ for which $0 < M\sqrt{\frac{p_n}{n}} \leq C \leq \min_{l \in A_n^*} ||\boldsymbol{\beta}_l^*|| \, \forall n \geq N'$ holds, since $\sqrt{p_n/n} \to 0$ by the assumptions. The rest works analogously to Theorem 3.12. □

A.2. Computational details and convergence of PIRLS

In the following, we will provide some further details on the PIRLS algorithm, concerning the approximations used and its convergence.

A.2.1. Details on approximation used in PIRLS

Some computational details of the PIRLS algorithm are provided next, aiming at understanding of its chances and limitations, especially in the high-dimensional setting. The arguments of this subsection follow [26] and [25]. In the unpenalized case ($\lambda_0 = \lambda_1 = 0$), when the goal is to minimize the negative log-likelihood function, a common approach is to use a Taylor approximation and then solve at iteration step $k \in \mathbb{N}$ a linearized problem, which can be solved by a Fisher scoring or Newton Raphson algorithm. Since we chose the canonical link function in our application, the Fisher scoring and Newton Raphson algorithms are equivalent since the observed and the expected Fisher information matrices are equal. Nevertheless, we will keep it more general and use in the sequel the expected Fisher information matrix, to allow the applicability of the results for other link functions as well. Now, since we want to minimize a *penalized* objective function, the goal is to construct penalized versions of the score function $s_{pen}(\cdot)$ and the Hessian $\boldsymbol{H}_{pen}(\cdot)$, or the Fisher information $\boldsymbol{I}_{F,pen}(\cdot)$, respectively, meaning that we need (approximations of) the derivatives of the objective function. Having that, we just need to solve the linearized problem as explained above. To obtain these penalized versions, the first step is, as in [6], to obtain a quadratic approximation of $P_\lambda^{gen}(\boldsymbol{\beta})$ (see (9)) at some $\hat{\boldsymbol{\beta}}^{(k)}$ resulting in the following, for details on the derivation we refer to [26], [25]

$$P_\lambda^{gen}(\boldsymbol{\beta}) \approx P_\lambda^{gen}(\hat{\boldsymbol{\beta}}^{(k)}) + \frac{1}{2}\left(\boldsymbol{\beta}^T \boldsymbol{A}_\lambda \boldsymbol{\beta} + \hat{\boldsymbol{\beta}}^{(k)} \boldsymbol{A}_\lambda (\hat{\boldsymbol{\beta}}^{(k)})^T\right),$$

where

$$\boldsymbol{A}_\lambda := \sum_{l=1}^{L} \lambda_l \boldsymbol{A}_l \text{ and } \boldsymbol{A}_l := p_l'(||\boldsymbol{a}_l \hat{\boldsymbol{\beta}}^{(k)}||_{N_l}) \cdot \frac{D_l(\boldsymbol{a}_l^T \hat{\boldsymbol{\beta}}^{(k)})}{\boldsymbol{a}_l^T \hat{\boldsymbol{\beta}}^{(k)}} \cdot \boldsymbol{a}_l \boldsymbol{a}_l^T.$$

Since the objective function $M_{pen}(\boldsymbol{\beta})$ to be minimized is given by the sum of the penalty function $P_\lambda(\boldsymbol{\beta})$ and the negative log-likelihood, the penalized versions of the score and the Hessian, and Fisher information, respectively, based on the approximation of $P_\lambda^{gen}(\boldsymbol{\beta})$ above are given by

$$\begin{aligned} s_{pen}(\boldsymbol{\beta}) &= s(\boldsymbol{\beta}) - \boldsymbol{A}_\lambda \boldsymbol{\beta},\ \boldsymbol{H}_{pen}(\boldsymbol{\beta}) = \boldsymbol{H}(\boldsymbol{\beta}) - \boldsymbol{A}_\lambda, \\ \boldsymbol{I}_{F,pen}(\boldsymbol{\beta}) &= -\mathbb{E}(\boldsymbol{H}(\boldsymbol{\beta}) - \boldsymbol{A}_\lambda) = \boldsymbol{X}^T \widetilde{\boldsymbol{W}}^{(k)} \boldsymbol{X} + \boldsymbol{A}_\lambda, \end{aligned}$$

where the latter equation, corresponding to the penalized Fisher information matrix, can be derived by straightforward calculations. $\boldsymbol{I}_F(\hat{\boldsymbol{\beta}}^{(k)}) = \boldsymbol{X}^T \widetilde{\boldsymbol{W}}^{(k)} \boldsymbol{X}$ is the unpenalized Fisher information matrix. The matrix $\widetilde{\boldsymbol{W}}^{(k)}$ is a diagonal

matrix with weights, in particular $\widetilde{\boldsymbol{W}}^{(k)} = \text{diag}(\boldsymbol{\mu}^{(k)}(\mathbf{1} - \boldsymbol{\mu}^{(k)}))$ with $\boldsymbol{\mu}^{(k)} = \exp(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(k)})/(1 + \exp(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(k)}))$. Further, $\tilde{\boldsymbol{y}}^{(k)} = (\widetilde{\boldsymbol{W}}^{(k)})^{-1}(\boldsymbol{y} - \boldsymbol{\mu}^{(k)}) + \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(k)}$ is the working response. Now, as in the unpenalized case we get at iteration step $k+1$ replacing the (penalized) Hessian by the (penalized) Fisher information

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{I}_{F,pen}(\hat{\boldsymbol{\beta}}^{(k)})^{-1} s_{pen}(\hat{\boldsymbol{\beta}}^{(k)}).$$

Introducing a stepsize $\nu \in (0, 1]$ and executing straightforward calculations, the iteration step can be re-written as follows, see [25]

$$\hat{\boldsymbol{\beta}}^{(k+1)} = (1-\nu)\hat{\boldsymbol{\beta}}^{(k)} + \nu \left(\boldsymbol{X}^T \widetilde{\boldsymbol{W}}^{(k)} \boldsymbol{X} + \boldsymbol{A}_\lambda \right)^{-1} \boldsymbol{X}^T \widetilde{\boldsymbol{W}}^{(k)} \tilde{\boldsymbol{y}}^{(k)}.$$

A.2.2. Convergence of PIRLS

Now, we can understand the limitations of the PIRLS algorithm. In particular, as explained in [26], the convergence of PIRLS is not always ensured, it depends on the penalized Fisher information matrix $\boldsymbol{I}_{F,pen}$. If we can ensure that $\boldsymbol{I}_{F,pen}$ is positive definite, the convergence of the algorithm, independently of the starting point, is guaranteed if a solution exists, since in this case the optimization problem is strictly convex. We differentiate between the following cases.

(i) $n > p$: in this case we know that, since by assumption the response distribution belongs to the exponential dispersion family, the unpenalized Fisher information matrix $\boldsymbol{I}_F$ is positive definite, consequently the penalty matrix $\boldsymbol{A}_\lambda$ has to be at least positive semi-definite to ensure that $\boldsymbol{I}_{F,pen}$ is positive definite so ensuring convergence of the algorithm

(ii) $n \leq p$: analogously to (i) above, the unpenalized Fisher information matrix $\boldsymbol{I}_F$ is positive semi-definite and thus the penalty matrix $\boldsymbol{A}_\lambda$ has to be at least positive definite to ensure that $\boldsymbol{I}_{F,pen}$ is positive definite.

If we can not ensure that $\boldsymbol{I}_{F,pen}$ is positive definite, it may happen that the algorithm finds non-unique descent directions. Naturally arises thus the question, what happens if strict convexity can not be ensured and multiple local minimizers may exist. As the name PIRLS indicates, we are using a *penalized* iteratively re-weighted least-squares (IRLS) step by using a Taylor approximation of the log-likelihood function. Hence, following the arguments of [12] (Section 4.4.1) as well as [13] (Section 5.4.3) analyzing the IRLS algorithm, it is natural that we can not guarantee convergence of PIRLS in general. Nevertheless, PIRLS is applying similar steps as done in [6] for SCAD using local quadratic approximations and the Newton algorithm. Hence, when the algorithm converges, the resulting estimator $\hat{\boldsymbol{\beta}}$ satisfies the penalized likelihood equations (with the approximations) for nonzero elements of $\hat{\boldsymbol{\beta}}$ so we have (at least) a local minimizer. As we can see above, the convergence of PIRLS clearly depends on the chosen penalty function, which is also what we expect. This can be seen in the real data application in Section 6, where the PIRLS algorithm (with `gvcm.cat`) failed in every replication for the L_0 penalty, while the PIRLS algorithm worked

in this example for L_0-FGL (with our own code). As we can see above, in the high-dimensional case, there are more restrictions that the penalty matrix has to fulfill, which explains the problems we faced with PIRLS in our simulation studies. Another problem that occurred in our simulation studies is that, especially in the high-dimensional case, some matrices that need to be inverted during the iteration process may not be invertible, causing failure of the whole algorithm. Nevertheless, in lower dimensional problems, our simulation studies show that this approach is suitable for L_0-FGL. To sum up, even though we can not ensure that PIRLS finds a global optimum by the nature of the chosen penalty function, we can find local minima by applying the Newton algorithm with local quadratic approximations.

A.3. Computational details and convergence of BCD

A.3.1. Details on approximation used in BCD

For the approximation of the objective function $\tilde{g}(\boldsymbol{\beta}_j, \boldsymbol{\beta}^{(k)})$ used in the factor-wise BCD approach, we provide next the details on the derivation of the function $g(\boldsymbol{\beta}_j, \boldsymbol{\beta}^{(k)})$, which is part of $\tilde{g}(\boldsymbol{\beta}_j, \boldsymbol{\beta}^{(k)})$. In general, we use the following quadratic approximation of the L_0 part at some $\hat{\boldsymbol{\beta}}^{(k)}$ (see also [25])

$$P_\lambda^{L_0}(\boldsymbol{\beta}) \approx P_\lambda^{L_0}(\hat{\boldsymbol{\beta}}^{(k)}) + \frac{1}{2}(\boldsymbol{\beta}^T \boldsymbol{A}_\lambda \boldsymbol{\beta} + \hat{\boldsymbol{\beta}}^{(k),T} \boldsymbol{A}_\lambda \hat{\boldsymbol{\beta}}^{(k)}), \tag{37}$$

where details on the construction of $\boldsymbol{A}_\lambda$ can be found in [25]. For the factor-wise approach we obtain $\boldsymbol{A}_{\lambda,j}$, on which details can be found in Remark A.4. We proceed as follows: since our penalty function shows a separable structure, we obtain the solution coodinate-wise. With the help of a Taylor approximation of the log-likelihood, we approximate the L_0 penalty function $P_\lambda^{L_0}(\boldsymbol{\beta}_j)$ separately for each $j \in \{1, \ldots, J\}$ such that it is possible to follow a coordinate-wise procedure for minimization. So, we will obtain an approximation as in (37) for $P_\lambda^{L_0}(\boldsymbol{\beta}_j)$ for each $j \in \{1, \ldots, J\}$.
Now it remains to obtain an approximation of the log-likelihood. In particular, we approximate the negative log-likelihood with Taylor as in [2] yielding an approximation of $-L_n(\boldsymbol{\beta})$ given by

$$-L_n(\boldsymbol{\beta}) \approx \frac{1}{2n}(\tilde{\boldsymbol{y}}^{(k)} - \boldsymbol{X}\boldsymbol{\beta})^T \widetilde{\boldsymbol{W}}^{(k)} (\tilde{\boldsymbol{y}}^{(k)} - \boldsymbol{X}\boldsymbol{\beta}). \tag{38}$$

Here, $\widetilde{\boldsymbol{W}}^{(k)}$ is a diagonal matrix of weights, see below for details.

Remark A.3 (On the factor of $\frac{1}{2n}$ in the log-likelihood)**.** In the literature, the log-likelihood, or the squared difference in the linear model case respectively, is sometimes divided by the factor $\frac{1}{2n}$, as for example in [2] and [11]. Since it does not change the solution of the minimum of the log-likelihood it is a convenient choice because it stabilizes the algorithm and it ensures that the impact of the tuning parameter λ does not depend on the sample size n. Note that one can

also neglect this factor but in this case one has to be careful when comparing two solutions for different tuning parameters and different sample sizes respectively, but basically it works the same way. We will use this factor in the sections about computation with block coordinate descent (BCD) and keep in mind that it is not used by [25] in PIRLS.

For the matrix with weights, already introduced in Appendix A.2, it holds $\widetilde{\boldsymbol{W}}^{(k)} = \text{diag}(w_i) \in \mathbb{R}^{n\times n}$, $w_i = \pi_i(1-\pi_i)$ for $i = 1,\dots,n$ where $\pi_i = \exp(\eta_i)/(1+\exp(\eta_i))$ and $\eta_i = (\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(k)})_i$ thus π_i is evaluated at the current iteration k. Furthermore, the working response $\tilde{\boldsymbol{y}}^{(k)}$, also introduced in Appendix A.2, is given by $\tilde{\boldsymbol{y}}^{(k)} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(k)} + \left(\widetilde{\boldsymbol{W}}^{(k)}\right)^{-1}(\boldsymbol{y}-\boldsymbol{\pi})$ where, $\boldsymbol{\pi} = (\pi_1,\dots,\pi_n)$ is evaluated at $\hat{\boldsymbol{\beta}}^{(k)}$. For each $j \in \{1,\dots,J\}$ we get the following approximation of the penalty term in some iteration step $\hat{\boldsymbol{\beta}}_j^{(k)}$, analogously to (37)

$$P_\lambda^{L_0}(\boldsymbol{\beta}_j) \approx P_\lambda^{L_0}(\hat{\boldsymbol{\beta}}_j^{(k)}) + \frac{1}{2}(\boldsymbol{\beta}_j^T \boldsymbol{A}_{\lambda,j}\boldsymbol{\beta}_j + (\hat{\boldsymbol{\beta}}_j^{(k)})^T \boldsymbol{A}_{\lambda,j}\hat{\boldsymbol{\beta}}_j^{(k)}). \tag{39}$$

Note that, in particular one has to write $\boldsymbol{A}_{\lambda,j}^{(k)}$ instead of $\boldsymbol{A}_{\lambda,j}$ since this quantity depends on the iteration step k, but we will leave the upper index out for simplicity.

Remark A.4. [Details on $\boldsymbol{A}_{\lambda,j}$ for L_0] For factor $j \in \{1,\dots,J\}$ with p_j +1 levels (including the reference category), the components of the approximation look as follows where we assume observing a nominal factor. For an ordinal one the value for $|L_j|$ will change as we just compare adjacent categories for ordinal factors. Let L_j be the set containing the row numbers of the matrix $\mathfrak{A}$ with rows $\boldsymbol{a}_l$ that correspond to the (pairwise) differences for factor j. We have with $\boldsymbol{a}_{l,j}$ ($l \in \{1,\dots,|L_j|\}$) being the columns of some matrix $\mathfrak{A}_\text{j}$ that produces the differences of the entries in $\boldsymbol{\beta}_j^{(k)}$ that

$$\begin{aligned}
|L_j| &= p_j + \binom{p_j}{2} = \frac{p_j(p_j-1)}{2} \quad \text{(number of differences of entries in } \boldsymbol{\beta}_j^{(k)}\text{)} \\
\boldsymbol{A}_{\lambda,j} &= \lambda_0 \sum_{l=1}^{|L_j|} p_l'(||\boldsymbol{a}_{l,j}^T\hat{\boldsymbol{\beta}}_j^{(k)}||_0)\frac{D_l(\boldsymbol{a}_{l,j}^T\hat{\boldsymbol{\beta}}_j^{(k)})}{\boldsymbol{a}_{l,j}^T\hat{\boldsymbol{\beta}}_j^{(k)}}\boldsymbol{a}_{l,j}\boldsymbol{a}_{l,j}^T \\
&= \lambda_0 \sum_{l=1}^{|L_j|}\left(\frac{1}{1+\exp(-\gamma|\boldsymbol{a}_{l,j}^T\hat{\boldsymbol{\beta}}_j^{(k)}|)}\right)\left(1-\frac{1}{1+\exp(-\gamma|\boldsymbol{a}_{l,j}^T\hat{\boldsymbol{\beta}}_j^{(k)}|)}\right) \\
&\qquad \cdot \frac{2\gamma\boldsymbol{a}_{l,j}\boldsymbol{a}_{l,j}^T}{\sqrt{(\boldsymbol{a}_{l,j}^T\hat{\boldsymbol{\beta}}_j^{(k)})^2 + c}}
\end{aligned} \tag{40}$$

We have that the columns $\boldsymbol{a}_{l,j} \in \mathbb{R}^{p_j\times 1}$ so they produce the differences of the coefficients and since they also include a columns in the shape of $(1,0,\dots,0),(0,1,0,\dots,0),\dots,(0,\dots,0,1)$, they also build the differences of each coefficient with reference category zero. It holds that $\boldsymbol{A}_{\lambda,j} \in \mathbb{R}^{p_j\times p_j}$ and $\boldsymbol{A}_{\lambda,j}$ is symmetric. $\boldsymbol{A}_{\lambda,j}$ depends on $\hat{\boldsymbol{\beta}}_j^{(k)}$ at iteration step k.

So the following function $g(\boldsymbol{\beta}_j, \boldsymbol{\beta}^{(k)})$ will be the factor-wise approximation of the log-likelihood and L_0 penalty part that we use in the BCD procedure

$$g(\boldsymbol{\beta}_j, \hat{\boldsymbol{\beta}}^{(k)}) := \frac{1}{2n}(\tilde{\boldsymbol{y}}^{(k)} - \boldsymbol{X}\boldsymbol{\beta})^T \widetilde{\boldsymbol{W}}^{(k)}(\tilde{\boldsymbol{y}}^{(k)} - \boldsymbol{X}\boldsymbol{\beta}) + P_\lambda^{L_0}(\hat{\boldsymbol{\beta}}_j^{(k)}) + \frac{1}{2}(\boldsymbol{\beta}_j^T \boldsymbol{A}_{\lambda,j}\boldsymbol{\beta}_j + \boldsymbol{\beta}_j^{(k),T} \boldsymbol{A}_{\lambda,j}\hat{\boldsymbol{\beta}}_j^{(k)}). \tag{41}$$

A.3.2. Convergence of BCD

Recall that, proceeding factor-wise, we use a quasi Newton method for minimization (BFGS method). In general, the theory for the BFGS method ensuring convergence is developed for twice continuously differentiable, convex functions [23]. It is generalized in [10] for non-smooth functions under reasonable conditions, including analyses on the euclidean norm. This exactly suits our application, since the part of the objective function $\tilde{g}$ (14) to be minimized which is not differentiable in $\mathbf{0}$ is the group lasso part being the euclidean norm of the corresponding factor. Consequently, by [10], the BFGS method is applicable for our function to be optimized $\tilde{g}$, since it is the sum over a quadratic function g (including the quadratic approximations of the L_0 part and the log-likelihood), which is clearly twice continuous differentiable and convex and the euclidean norm, which is convex, but not differentiable in $\mathbf{0}$. To sum up, following [10], we can ensure convergence of the quasi Newton algorithm in the blockwise updates. We will see in our simulation studies that this approach is convenient for high-dimensional problems. There also exist low-dimensional problems where the BCD approach is competetive to PIRLS, see the real data application in Section 6. Further analyses on the behavior of quasi Newton methods applied to non-smooth (and not necessarily convex) functions can be found in [17].

A.4. Details on simulation study

A.4.1. Details on tuning

We used cross-validation (CV) for all penalties to determine the tuning parameters λ_0 and $\lambda = (\lambda_0, \lambda_1)$ for the L_0-FGL approach, respectively. In particular, we used $k = 5$ fold CV, where we used $\lambda_{lower} = (\lambda_{lower,1}, \lambda_{lower,0}) = (0, 0)$ and for λ_{upper} we chose a value which excludes all variables from the model. Note that, as for the lower values, for L_0-FGL we need two upper values for lambda, hence $\lambda_{upper} = (\lambda_{upper,1}, \lambda_{upper,0})$. So for the case of two tuning parameters, we chose them in a way such that for $\lambda_{upper} = (\lambda_{upper,1}, \lambda_{upper,0})$ all parameters are excluded from the model where we took $\lambda_{upper,1} = \lambda_{upper,0}$ to avoid that we set the focus on selection or fusion. Between these two values, the CV procedure fitted the model for $n_\lambda = 10$ different values of λ_0 and λ_1 in highdim design and $n_\lambda = 30$ different values of λ_0 and λ_1 in B8 design. For the CV of L_0 with PIRLS, we used the stored functions in `gvcm.cat`. As explained at the beginning of this work (Remark 2.1), for the cross validation procedure for L_0-FGL, which

includes two tuning parameters, we considered two approaches: a stepwise and an iterative procedure. In highdim, we used the stepwise procedure for the L_0-FGL approaches, while in B8 we used the iterative procedure for the L_0-FGL approaches. The parameters for the approximation of the L_0 part, which is used in all of our considered methods, were chosen equally in all approaches $c = 10^{-5}$ and $\gamma = 10$ (recommended in [25]). Even if L_0-FGL with BCD do not require a stepsize, we used a stepsize of $\nu = 0.05$ for all considered approaches. This is done to stabilize the algorithm and to obtain comparable results.

A.5. Details on real data application

For tuning of L_0-FGL (L_0-FGL PIRLS, L_0-FGL PIRLS iterative, L_0-FGL BCD) we used a total number of $n_\lambda = 10$ values for λ_1 and λ_0. Since we divide the log-likelihood in the BCD computing approach by $2n$, we use different tuning ranges for L_0-FGL PIRLS and L_0-FGL PIRLS iterative as for L_0-FGL BCD, in particular we used $\lambda_{0,lower} = \lambda_{1,lower} = 0$ for L_0-FGL BCD, L_0-FGL PIRLS and L_0-FGL PIRLS iterative and as maximum values for L_0-FGL PIRLS and L_0-FGL PIRLS iterative we chose $\lambda_{0,upper} = \lambda_{1,upper} = 1$ and for the corresponding BCD approach $\lambda_{0,upper} = \lambda_{1,upper} = 0.01$.

Acknowledgments

We would like to express our sincere thanks to the associate editor and the reviewers for their useful and constructive comments which helped us improve the paper.

References

[1] Bondell, H. D. and Reich, B. J. (2009). Simultaneous factor selection and collapsing levels in ANOVA. *Biometrics* **65** 169–177. https://doi.org/10.1111/j.1541-0420.2008.01061.x MR2665858

[2] Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics* **5** 232–253. https://doi.org/10.1214/10-aoas388 MR2810396

[3] Breheny, P. and 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 MR3306699

[4] Bühlmann, P. and van de Geer, S. (2015). High-dimensional inference in misspecified linear models. *Electronic Journal of Statistics* **9**. https://doi.org/10.1214/15-EJS1041 MR3367666

[5] Bunea, F. (2008). Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electronic Journal of Statistics* **2** 1153–1194. https://doi.org/10.1214/08-EJS287 MR2461898

[6] Fan, J. and 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 MR1946581

[7] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* **32** 928–961. https://doi.org/10.1214/009053604000000256 MR2065194

[8] Fan, J., Xue, L. and Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics* **42** 819–849. https://doi.org/10.1214/13-AOS1198 MR3210988

[9] Gertheiss, J. and Tutz, G. (2010). Sparse modeling of categorial explanatory variables. *Annals of Applied Statistics* **4** 2150–2180. https://doi.org/10.1214/10-AOAS355 MR2829951

[10] Guo, J. and Lewis, A. S. (2018). Nonsmooth Variants of Powell's BFGS Convergence Theorem. *SIAM Journal on Optimization* **28** 1301–1311. https://doi.org/10.1137/17M1121883 MR3799065

[11] Guo, X., Zhang, H., Wang, Y. and Wu, J.-L. (2015). Model selection and estimation in high dimensional regression models with group SCAD. *Statistics & Probability Letters* **103** 86–92. https://doi.org/10.1016/j.spl.2015.04.017 MR3350866

[12] Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning. Springer Series in Statistics*. Springer, New York, NY. https://doi.org/10.1007/978-0-387-84858-7 MR2722294

[13] Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical Learning with Sparsity*, 1 ed. Chapman and Hall/CRC, New York, NY. https://doi.org/10.1201/b18401 MR3616141

[14] Huang, J., Breheny, P. and Ma, S. (2012). A Selective Review of Group Selection in High-Dimensional Models. *Statistical Science* **27** 481–499. https://doi.org/10.1214/12-sts392 MR3025130

[15] Kim, Y., Kim, J. and Kim, Y. (2006). Blockwise sparse regression. *Statistica Sinica* **16** 375–390. MR2267240

[16] Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28** 1356–1378. https://doi.org/10.1214/aos/1015957397 MR1805787

[17] Lewis, A. and Overton, M. (2012). Nonsmooth Optimization via Quasi-Newton Methods. *Mathematical Programming* **141**. https://doi.org/10.1007/s10107-012-0514-2 MR3097282

[18] Li, M., Kong, L. and Su, Z. (2021). Double fused Lasso regularized regression with both matrix and vector valued predictors. *Electronic Journal of Statistics* **15** 1909–1950. https://doi.org/10.1214/21-EJS1829 MR4255315

[19] Liang, H. and Du, P. (2012). Maximum likelihood estimation in logistic regression models with a diverging number of covariates. *Electronic Journal of Statistics* **6** 1838–1846. https://doi.org/10.1214/12-ejs731 MR2988466

[20] Lu, Z. and Zhang, Y. (2012). Penalty Decomposition Methods for *L*0-

Norm Minimization. *arXiv*. MR3359809

[21] Lu, Z. and Zhang, Y. (2013). Sparse approximation via penalty decomposition methods. *SIAM Journal on Optimization* **23** 2448–2478. https://doi.org/10.1137/100808071 MR3138116

[22] Meier, L., van de Geer, S. and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70** 53–71. https://doi.org/10.1111/j.1467-9868.2007.00627.x MR2412631

[23] Powell, M. J. D. (1976). Some global convergence properties of a variable metric algorithm for minimization without exact line searches. *Nonlinear Programming*. MR0426428

[24] Newman, D. J., Hettich, S., Blake, C. L. and Merz, C. J. (1998). UCI Repository of machine learning databases.

[25] Oelker, M.-R., Pössnecker, W. and Tutz, G. (2014). Selection and fusion of categorical predictors with L0-type penalties. *Statistical Modelling: An International Journal* **15** 389-410. https://doi.org/10.1177/1471082X14553366 MR3403123

[26] Oelker, M.-R. and Tutz, G. (2013). A General Family of Penalties for Combining Differing Types of Penalties in Generalized Structured Models. 2013. https://doi.org/10.5282/ubm/epub.17664

[27] Schultheiss, C., Renaux, C. and Bühlmann, P. (2021). Multicarving for high-dimensional post-selection inference. *Electronic Journal of Statistics* **15** 1695–1742. https://doi.org/10.1214/21-EJS1825 MR4255311

[28] Staerk, C., Kateri, M. and Ntzoufras, I. (2021). High-dimensional variable selection via low-dimensional adaptive learning. *Electronic Journal of Statistics* **15** 830–879. https://doi.org/10.1214/21-EJS1797 MR4203346

[29] Stokell, B. G., Shah, R. D. and Tibshirani, R. J. (2021). Modelling high-dimensional categorical data using nonconvex fusion penalties. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **83** 579-611. https://doi.org/10.1111/rssb.12432 MR4294545

[30] Tang, W. and Ye, Y. (2020). The existence of maximum likelihood estimate in high-dimensional binary response generalized linear models. *Electronic Journal of Statistics* **14** 4028–4053. https://doi.org/10.1214/20-EJS1766 MR4168791

[31] Taylor, J. and Tibshirani, R. (2018). Post-selection inference for L1-penalized likelihood models. *The Canadian Journal of Statistics* **46** 41–61. MR3767165

[32] Tibrewal, A. O. (2020). Developement of a nomogram for predicting recurrence in breast cancer patients using a machine learning method. *International Journal of Community Medicine and Public Health* **7** 2661–2666. https://www.ijcmph.com/index.php/ijcmph/article/view/6405

[33] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B* **58** 267–288. https://doi.org/10.1111/j.1467-9868.2011.00771.x MR1379242

[34] Wang, H. and Leng, C. (2008). A note on adaptive group Lasso. *Com-*

putational Statistics & Data Analysis **52** 5277–5286. https://doi.org/10.1016/j.csda.2008.05.006 MR2526593

[35] Wang, L., Chen, G. and Li, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics* **23** 1486–1494. https://doi.org/10.1093/bioinformatics/btm125

[36] Wang, M. and Tian, G.-L. (2019). Adaptive group Lasso for high-dimensional generalized linear models. *Statistical Papers* **60** 1469–1486. https://doi.org/10.1007/s00362-017-0882-z MR4017019

[37] William Fithian, J. T. Dennis Sun (2017). Optimal Inference After Model Selection. *arXiv*.

[38] Xin, X., Hu, J. and Liu, L. (2017). On the oracle property of a generalized adaptive elastic-net for multivariate linear regression with a diverging number of parameters. *Journal of Multivariate Analysis* **162** 16–31. https://doi.org/10.1016/j.jmva.2017.08.005 MR3719332

[39] Yuan, M. and 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 MR2212574

[40] 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 MR2604701

[41] Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Mashine Learning Research* **7** 2541–2563. MR2274449

[42] Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429. https://doi.org/10.1198/016214506000000735 MR2279469

[43] Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics* **37** 1733–1751. https://doi.org/10.1214/08-AOS625 MR2533470