High-Dimensional Logistic Regression with Fusion-Type Penalties

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen University zur Erlangung des akademischen Grades einer Doktorin der Naturwissenschaften genehmigte Dissertation

vorgelegt von

Lea Johanna Kaufmann, M. Sc.

aus Essen

Berichter: Universitätsprofessorin Dr. Maria Kateri
Universitätsprofessorin Dr. Irini Moustaki
Universitätsprofessor Dr. Udo Kamps

Tag der mündlichen Prüfung: 09. April 2025

Diese Dissertation ist auf den Internetseiten der Universitätsbibliothek verfügbar.

Acknowledgments

Above all else, I am profoundly thankful to my supervisor *Prof. Dr. Maria Kateri* for the opportunity to do research in such an exciting and interesting field. Your expertise offered me valuable insights and new perspectives that substantially enhanced my work. Your guidance provided just the right balance of direction and independence and your trust gave me the confidence to grow and persevere, even in challenging periods of research. Thank you so much for being such a great supervisor.

I would like to thank *Prof. Dr. Irini Moustaki* and *Prof. Dr. Udo Kamps* for agreeing to be the co-referees. Thank you also for your helpful remarks on my work.

Furthermore, I thank all my current and former colleagues at the Institute of Statistics who have accompanied me during my PhD. Your moral support was a tremendous help and the positive and motivating atmosphere you provided made all the difference. In particular, my sincere thanks go to *Dr. Wolfgang Herff, Yao Lu, Dr. Nikolay Nikolov, Dr. Simon Pitzen, Despo Theocharous, Dr. Francesco Buono, Dr. Thomas van Bentum* and *Lene Grabowsky*.

I thank *Prof. Dr. Peter Kern* and *Prof. Dr. Axel Bücher* who significantly contributed to sparking my interest in the field of statistics and stochastic during my studies at the Heinrich-Heine University in Düsseldorf. Your excellent lectures and your encouragement during my master thesis played a key role for me to embark on a PhD.

Moreover, I want to thank my friends from Essen, Aachen and from my studies in Düsseldorf, who have unquestionably stood by me and showed me that they truly admire what I am doing, especially my dear best friend *Paulina Elbers*.

I want to express my deepest gratitude to my family, primarily my parents *Ralf Kaufmann* and *Gabriele Vossen* for their overwhelming support in my whole life. You always encouraged me to pursue my interests and provided invaluable advice when needed, which gave me the freedom to focus. Special thanks are due to my father Ralf for inspiring me with his enthusiasm for mathematics since my school days. Sharing your ability to continuously face new challenges and tackle them with a calm and composed mindset has helped me throughout my entire journey. Thanks also for your tireless perseverance in proofreading my thesis, providing constructive and helpful feedback. I thank my brother *Tom Kaufmann* for his support, given in that special way only a big brother can offer. I truly appreciate your backing. Further, I would like to thank my lovely grandparents *Erika* and *Helmut Burggraf*. Without all of you, this work would not have been possible.

Finally, I am deeply grateful to my love *Markus Stump* for being my companion since the first day of our studies. Your unconditional patience and understanding greatly helped me during tough times and your encouragement and motivation kept me going. Thank you for being there in all circumstances of life!

Contents

Introduction

Nowadays, large datasets arise in many application fields, where the notion "large" may either refer to the sample size, or to the number of observed attributes. In regression-type models, the goal is to obtain a model, which quantifies the dependence structure between some variable Y, which is called the response variable, and some predictors $X_1, X_2, \ldots, X_p$, based on their values for a sample of size $n \in \mathbb{N}$. It may happen that the number of predictors p exceeds the sample size n, which causes that classical methods to obtain estimates of the predictors' regression coefficients, e.g. maximum likelihood estimates, do not yield "accurate" estimates or even do not exist. If $p \leq n$ together with p being very large, similar issues may occur, even though e.g. maximum likelihood estimates may exist from the theoretical point of view. To be more precise, such cases where p is very large come with complex inversion tasks of large matrices, which cause problems in practice. In both cases, i.e. $p > n$ or $p \leq n$ and p very large, the terminology *high-dimensional* is used in the literature. In high-dimensional settings, the application of *penalized regression* techniques is able to resolve the discussed issues, yielding the desired estimates, where these technqies rely on the so-called *sparsity assumption*. That is, a model is called *sparse*, if only a relatively small fraction of the predictor variables has an influence on the response variable. The assumption that the true underlying model is sparse provides the possibility to obtain convenient estimates of the regression coefficients through penalized regression. The following example, which is taken from Hastie et al. (2015) (Chapter 1), illustrates this concept.

In medical research, scientists investigate the impact of genes on diseases. That is, given a patients' gene information, one aims to predict the probability of whether the patient will develop a certain disease or not. An example is considered where measurements of 4718 genes are given for $n = 349$ cancer patients. The goal is to predict the probability that the patient will develop a certain cancer class, based on the patients' given 4718 gene measurements. For the cancer classes, 15 different classes are considered, that is, bladder, breast, colon, kidney etc., see Hastie et al. (2015) (Chapter 1). Building a regression-type model with the elaborated purpose, the coefficient vector that needs to be estimated is of dimension $4718 \times 15 = 70770$, which is challenging with a sample size of only $n = 349$, thus known approaches like maximum likelihood estimation will not yield any results. However, it is known from former medical research, that only a few genes have an influence on the cancer class, i.e. the model is known to be *sparse*. Under the sparsity assumption, it is convenient to obtain coefficient estimates through penalized regression, which imposes a penalty function on the coefficients such that the resulting model balances the quality of the fit as well as the model complexity. There are plenty of different choices of penalty functions in the literature, which are presented at the beginning of this thesis. Coming back to the example, through penalized regression the (comparably) large amount of attributes (i.e. predictors) can be cut down to the essential parts that are needed to characterize the cancer class. To be more precise, Hastie et al. (2015) (Chapter 1) fitted a penalized regression method which yielded that only 254 genes have at least one nonzero coefficient, thus are influential at least for one cancer class. That is, out of 70770 coefficients to be estimated, only 254 are nonzero, which drastically decreases the model complexity and highly increases its interpretability, understanding what type of genes have an influence on the cancer class. The resulting (cross-validation) error rate of the obtained model was calculated as

10%, thus the grand majority of 90% of the samples were correctly classified by this model.

Motivated by the example of the previous paragraph, which shows the strength and importance of penalized regression methods, this thesis focuses on penalized *logistic* regression. That is, a binary response variable Y with an underlying logistic regression model is considered. For the predictor variables, the focus lies on *factors*, thus categorical predictors with a fixed number of levels. Another characteristic of this thesis is, that the notion *sparsity* is not only studied in the context of selecting predictors that have some influence on the response variable (referred to as *factor selection*), but also in the context of fusing the levels of a factor that have the *same* influence on the response variable (referred to as *levels fusion*), which further reduces the complexity of the considered regression model. The particular structure of this thesis is given below.

Chapter 1 starts with the introduction of the considered framework and provides an overview of existing penalized regression methods, where theoretical properties provided in the literature for linear regression are transferred to the considered setting of logistic regression. First, the focus lies on penalized regression methods for levels fusion, i.e. those introduced in Bondell and Reich (2009), Gertheiss and Tutz (2010b) incorporating an L_1-type penalty and Oelker et al. (2014a) incorporating an L_0-type penalty, to name the main references. Then, penalized regression methods introduced for the purpose of factor selection are considered, where the main references are the works of Yuan and Lin (2006), Meier et al. (2008) who consider the so-called group lasso penalty as well as Fan and Li (2001) and Wang et al. (2007) proposing (group) SCAD and, finally, Zhang (2010) and Huang et al. (2012) contributing (group) MCP. Details on the clear definitions, purposes and differences of the named methods are given at the respective point in the chapter. Computational methods employed for obtaining the corresponding estimates are discussed before characteristics of the different penalized regression approaches are underlined by comparing the corresponding coefficient paths. Finally, extensive simulation studies are conducted using the statistical software `R`, investigating the behavior of the presented methods in five different simulation designs, showing the advantages and disadvantages of the methods. These studies reveal that, on the one hand, the existing methods for performing levels fusion yield models that are not sparse enough on the factor level, i.e. it is desirable to improve their factor selection performance. On the other hand, it is demonstrated that the existing methods designed for factor selection are not able to conduct levels fusion, i.e. sparsity in the context of fusion of the levels of a factor with a similar influence on the response variable is neglected. These investigations underline that, so far, there exists no penalized regression approach for simultaneously performing factor selection and levels fusion. Since such an approach would be highly suitable especially (but not only) in the context of regression-type models incorporating factors, this gap will be closed in the upcoming chapters of this thesis.

In **Chapter 2**, a new approach is proposed to simultaneously perform factor selection and levels fusion with a penalized logistic regression technique, for which a new penalty function is introduced. To build a solid basis for the foregoing thesis, this chapter focuses on the theoretical investigation of the new method. In particular, theoretical properties are proven, e.g. consistency results, asymptotic normality along with factor selection consistency as well as fusion consistency. These properties are very valuable as they show that, e.g., under some suitable regularity conditions, the probability that the new method will exclude influential factors from the model goes to zero as the sample size goes to infinity (consistency in factor selection). A similar result for levels fusion is shown, i.e. it is proven that the probability that the method fuses the levels of the factors correctly goes to one as the sample size goes to infinity (fusion consistency). These properties justify that the new method is a suitable choice for the purpose

of obtaining sparse models in penalized logistic regression with factors.

Then, **Chapter 3** develops convenient algorithms to calculate the estimates for the new regularization technique introduced in Chapter 2. Since the optimization problem that needs to be solved to obtain estimates includes the minimization of a function that is not continuous caused by the structure of the novel penalty function, a suitable approximation will be applied. Subsequently, two algorithms to solve the resulting optimization problem will be considered, one of them being a Newton-type algorithm applied on penalized versions of the hessian and the score and one being a block-coordinate descent algorithm. Further, coefficient paths applying both algorithms are provided, which underline the functionality of the algorithms and emphasize that the new technique simultaneously performs factor selection as well as levels fusion, i.e. the purpose for which it was introduced. Having that, the focus of this chapter returns to the simulation designs provided in Chapter 1 and the behavior of the new method in the same five simulation designs is investigated. These simulations show that the new method is, on the one hand, able to improve the factor selection performance of those penalties for levels fusion and, on the other hand, able to perform *both* factor selection and levels fusion, closing the gap illustrated at the end of Chapter 1. Furthermore, it is demonstrated that both algorithms proposed at the beginning of this chapter perform well, one being more suitable in designs of lower dimensions, and the other in designs of higher dimensions.

Finally, **Chapter 4** provides statistical inference analysis for the proposed method of Chapter 2. In general, performing statistical inference in (high-dimensional) penalized regression is challenging, as thoroughly explained in this chapter. However, it is shown how statistical inference can be conducted in the framework of the novel penalized regression technique. In particular, a two-stage method is proposed, including a step for dimension reduction through factor selection and levels fusion (step 1), and an inferential step (step 2). Considering two different approaches for corrections for multiplicity of testing, a single and a multiple sample splitting approach is applied. The main difficulty ensuring a suitable theoretical basis applying the novel penalty function in step 1, compared to the existing approaches in the literature, e.g. the main references Wasserman and Roeder (2009) and Meinshausen et al. (2009), is the fact that factor selection *and* levels fusion is performed here. Consequently, in step 1, it is necessary to ensure the so-called screening properties for *both* factor selection and levels fusion. However, Chapter 2 justifies the assumption of these screening properties, enabling to proof error control properties, e.g. bounding the probability of even one false rejection of the null hypotheses of the tests considered in step 2. Generally speaking, the two-stage method first reduces the dimension in step 1 and, having that, those truly non-influential factors that are still included in the model are removed in step 2. The factor selection consistency shown in Chapter 2 yields that the probability that all the influential factors are included in the selected model of step 1 tends to one as the sample size increases, which is necessary to ensure that the two-stage method is reasonable, similar arguments apply for the respective fusion property. Consequently, the two-stage method further increases the sparsity level of the final selected model, while ensuring important and appropriate theoretical properties.

To sum up, this thesis contributes a novel penalty function for penalized logistic regression which is suitable especially for the incorporation of factors as it simultaneously performs factor selection and levels fusion. A thorough examination of theoretical properties and the behavior in simulation studies demonstrate that this technique is advantageous. Additionally, a two-stage approach enables statistical inference analysis.

Chapter 1

Penalized Logistic Regression

This chapter gives an overview of existing penalized (logistic) regression approaches, where its structure is as follows. Section 1.1 establishes the underlying regression-type model that is considered, including the explanatory variables, the random response variable and the underlying distribution of the latter. Having that, Section 1.2 proceeds with the framework and purpose of penalized regression. Since penalized regression techniques involve tuning parameters, the cross-validation procedure as a suitable tool for the determination of those is presented. Further, definitions of theoretical properties are introduced, which are desirable to be satisfied by the penalized regression approaches and with respect to which the approaches are compared, ensuring a comparable scale. Then, a selection of penalized regression approaches being suitable for the considered framework is presented, along with appropriate algorithms to solve the resulting optimization problems. Starting with penalties on differences in Section 1.3, those for grouped structures are discussed in Section 1.4 and finally non-convex penalties are examined in Section 1.5. Through the literature, theoretical properties for the majority of the considered approaches are proven for the linear regression model, while it is often stated that the results are extendable to logistic regression. However, it is not clear whether these extensions require some adjustments in the assumptions or proofs. Consequently, in this thesis, the results are transferred, if possible, to the case of logistic regression adjusting the assumptions and proofs, respectively. Finally, Section 1.7 introduces the framework of the simulation studies that were conducted to compare the considered penalized regression approaches from the computational side. The conducted simulation designs are presented, as well as details on tuning and measures with respect to which the methods are compared. The results are presented and discussed in Section 1.8. This chapter is closed by a final comparison of the considered methods in Section 1.9, including a discussion on improvement possibilities, which motivates the introduction of a new penalized regression approach in Chapter 2.

1.1 The Model

First of all, the underlying model that is considered throughout the whole thesis is introduced in this section, which is the logistic regression model as a special case of a generalized linear model (GLM). Before doing so, the necessary framework corresponding to the explanatory variables, the random response variable as well as distributional assumptions are specified. References for more detailed information on GLMs are given by McCullagh and Nelder (1989), Myers et al. (2010), as well as Agresti (2015), while further details on GLMs and corresponding treatments for categorical variables can be found in Kateri (2014), Tutz (2011) and Agresti (2002).

1.1.1 Explanatory Variables and the Response Variable

One considers $J \in \mathbb{N}$ categorical explanatory variables denoted by $\mathscr{X}_1, \ldots, \mathscr{X}_J$ and a binary random response variable Y, where the latter takes values in the set $\{0, 1\}$. The explanatory variables are also called *predictor variables* or *covariates*. In particular, a categorical variable is

a variable whose measurement scale consists of a set of categories (according to Agresti (2002), Section 1.1), where the categories are also called *levels*. The focus here lies on categorical explanatory variables, which are also called *factors*. Factors can be either *nominal* or *ordinal*, where the latter means that the levels can be naturally ordered - if this is not the case, the factor is called nominal. Further, if a factor has only two levels, it is called a *binary* factor and can be considered either as nominal or ordinal, depending on the modeling context.

Each factor $\mathscr{X}_j$, $j \in \{1, \dots, J\}$, is assumed to have $p_j + 1$ levels ($p_j \in \mathbb{N}$), which are coded by $0, \dots, p_j$. In this thesis, the commonly used dummy coding scheme is applied, that is, for each factor a series of dummy variables is introduced, which take values in the set $\{0, 1\}$, depending on the level of the corresponding factor, specified in (1.1). Consequently, for each factor $\mathscr{X}_j$, $j \in \{1, \dots, J\}$, p_j dummy variables $\mathscr{X}_{j,1}, \dots, \mathscr{X}_{j,p_j}$ are introduced, constructed as follows

$$\mathscr{X}_{j,k} = 1 \Leftrightarrow \mathscr{X}_j = k \ , j \in \{1, \dots, J\}, k \in \{1, \dots, p_j\}. \tag{1.1}$$

The reason why p_j dummy variables are introduced instead of $p_j + 1$, even though the number of levels of $\mathscr{X}_j$ is $p_j + 1$, is the following. If $\mathscr{X}_{j,k} = 0 \ \forall k \in \{1, \dots, p_j\}$, it is clear that $\mathscr{X}_j$ takes the first category, which is coded as zero, i.e. a dummy variable for this category is redundant. This category is the so-called *reference category*. Clearly, any other category of a factor can be chosen to be the reference category, however, it is common to use the first category, which is also done in this thesis. The total number of levels throughout all factors without counting the reference categories is defined as p, hence

$$p := \sum_{j=1}^{J} p_j.$$

Summing up, one sets

$$\mathscr{X}^* := (\underbrace{\mathscr{X}_{1,1}, \dots, \mathscr{X}_{1,p_1}}_{\mathscr{X}_1}, \dots, \underbrace{\mathscr{X}_{J,1}, \dots, \mathscr{X}_{J,p_J}}_{\mathscr{X}_J})$$

as the vector of the non-redundant dummy variables corresponding to the J factors under consideration and $\boldsymbol{x}^* \in \mathbb{R}^p$ as a realization of $\mathscr{X}^*$. One considers $n \in \mathbb{N}$ given observations of $\mathscr{X}^*$ denoted by $\boldsymbol{x}_1^*, \dots, \boldsymbol{x}_n^* \in \mathbb{R}^p$. These observations are written into a matrix $\boldsymbol{X}^* \in \mathbb{R}^{n \times p}$ yielding the following structure

$$\boldsymbol{X}^* = \begin{pmatrix} x_{1,1} & \dots & x_{1,p_1} & \dots & x_{1,\sum_{j=1}^{J-1} p_j} & \dots & x_{1,p} \\ x_{2,1} & \dots & x_{2,p_1} & \dots & x_{2,\sum_{j=1}^{J-1} p_j} & \dots & x_{2,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & \dots & x_{n,p_1} & \dots & x_{n,\sum_{j=1}^{J-1} p_j} & \dots & x_{n,p} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^* \\ \boldsymbol{x}_2^* \\ \vdots \\ \boldsymbol{x}_n^* \end{pmatrix}. \tag{1.2}$$

The (sub-)matrices $\boldsymbol{X}_j \in \mathbb{R}^{n \times p_j}$ are defined such that they contain the n observed samples of the dummy variables corresponding to covariate $\mathscr{X}_j$ as illustrated in (1.3), i.e. $\boldsymbol{X}^* = (\boldsymbol{X}_1; ...; \boldsymbol{X}_J)$. To be more precise, the structure of the sub-matrices is as follows

$$\boldsymbol{X}_1 = \begin{pmatrix} x_{1,1} & \dots & x_{1,p_1} \\ x_{2,1} & \dots & x_{2,p_1} \\ \vdots & \vdots & \vdots \\ x_{n,1} & \dots & x_{n,p_1} \end{pmatrix}, \ \boldsymbol{X}_j = \begin{pmatrix} x_{1,\sum_{i=1}^{j-1} p_i} & \dots & x_{1,\sum_{i=1}^{j} p_i} \\ x_{2,\sum_{i=1}^{j-1} p_i} & \dots & x_{2,\sum_{i=1}^{j} p_i} \\ \vdots & \vdots & \vdots \\ x_{n,\sum_{i=1}^{j-1} p_i} & \dots & x_{n,\sum_{i=1}^{j} p_i} \end{pmatrix} \text{ for } j \in \{2, \dots, J\}. \tag{1.3}$$

Remark 1.1.1. The above explained coding scheme for categorical covariates is commonly used and follows several works, e.g. Oelker et al. (2014b) and Gertheiss and Tutz (2010b). However, other coding schemes exist, as e.g. the so-called split coding scheme which is appropriate if ordinal covariates are considered, according to Tutz (2011) (Section 1.4.1). To be more precise, one sets $\mathscr{X}_{j,k} = 1 \Leftrightarrow \mathscr{X}_j > k$ and $\mathscr{X}_{j,k} = 0$ otherwise. This split coding scheme takes into account the ordering of the levels as the variable $\mathscr{X}_{j,k}$ indicates whether covariate $\mathscr{X}_j$ takes a level "above" k or not. Nevertheless, this thesis does not restrict the categorical covariates to be ordinal, i.e. the dummy coding scheme is applied. Further details on other coding schemes can also be found in Chiquet et al. (2016).

Finally, $\boldsymbol{y}$ is defined as the vector containing the observed sample of size n of the response variable, in particular

$$\boldsymbol{y} := (y_1, \ldots, y_n)^T \in \mathbb{R}^n. \tag{1.4}$$

1.1.2 Generalized Linear Model (GLM)

The goal now is to predict the expected value of the binomial distributed binary response variable Y given the predictor variables $\mathscr{X}_1, \ldots, \mathscr{X}_J$, where the particular assumed underlying model is introduced in the following. First, one needs to introduce the *design matrix* which is denoted by $\boldsymbol{X} \in \mathbb{R}^{n \times (p+1)}$. The design matrix $\boldsymbol{X}$ is given by an extension of $\boldsymbol{X}^*$ (1.2) adding a column of ones, which account for including an intercept in the model, thus

$$\boldsymbol{X} := (\boldsymbol{1}^T; \boldsymbol{X}^*). \tag{1.5}$$

The design matrix $\boldsymbol{X}$ is assumed to be *fixed*, i.e. this thesis considers a *fixed design*. With $\boldsymbol{x}_i$ the i-th row of the design matrix $\boldsymbol{X}$ is denoted, i.e. $\boldsymbol{x}_i = (1, \boldsymbol{x}_i^*)$, where $\boldsymbol{x}_i^*$ contains realization i of $\mathscr{X}^*$. Further, one defines

$$\mathscr{X} := (1, \mathscr{X}^*),$$

thus $\mathscr{X}$ is the vector of non-redundant dummy variables corresponding to the factors under consideration extended by the value one as the first entry.

In contrast to the explanatory variables, the response variable Y is assumed to be *random*. For the *random* response variable Y, the response vector is denoted by

$$\boldsymbol{Y} := (Y_1, \ldots, Y_n)^T$$

which contains a random sample of size n. The corresponding observed response vector is denoted by $\boldsymbol{y}$ and was introduced above in (1.4).

The vector of (unknown) regression coefficients is denoted by $\boldsymbol{\beta}$, which is also called coefficient vector for short in this thesis. In particular

$$\boldsymbol{\beta} := (\beta_{int}, \underbrace{\beta_{1,1}, \ldots, \beta_{1,p_1}}_{\boldsymbol{\beta}_1}, \ldots, \underbrace{\beta_{J,1}, \ldots, \beta_{J,p_J}}_{\boldsymbol{\beta}_J}),$$

hence $\boldsymbol{\beta}$ is of dimension $p+1$. Here, β_{int} represents the intercept of the model and $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J$ are defined as the parameter sub-vectors of dimensions $p_1, \ldots, p_J$ corresponding to the considered factors.

The classical linear model is given by

$$\mathbb{E}(Y|\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta}, \tag{1.6}$$

or, in matrix notation,

$$\mathbb{E}(\boldsymbol{Y}|\boldsymbol{X}) = \mathbf{1}^T\beta_0 + \sum_{j=1}^{J} \boldsymbol{X}_j\boldsymbol{\beta}_j = \boldsymbol{X}\boldsymbol{\beta}, \tag{1.7}$$

modeling the expectation of Y given the explanatory variables as a linear function, where the linearity corresponds to linearity in $\boldsymbol{\beta}$. With a binary random response variable Y, this model is in general not suitable, since the product $\boldsymbol{x}\boldsymbol{\beta}$, which is also called the *linear predictor* $\eta := \boldsymbol{x}\boldsymbol{\beta}$, in matrix notation $\boldsymbol{\eta} := \boldsymbol{X}\boldsymbol{\beta}$, may take values on the whole real line $\mathbb{R}$, while the expectation of a binary random response variable Y should be between zero and one.

For that reason, a generalized linear model (GLM) is considered next, which applies an appropriate function $g(\cdot)$ to the expectation on the left hand side of equation (1.6), and (1.7) respectively, one compares (1.11) and (1.12). Thus, the linear model (1.6) is a special GLM where $g(\cdot)$ is chosen to be the identity function. These type of models were introduced by Nelder and Wedderburn (1972) and provide a unified framework for different types of random response variables being modeled by explanatory variables. It is required that the distribution of the random response variable Y belongs to the exponential dispersion family, which is specified in the upcoming definition. Thus, before the model can be rigorously introduced, the following definition of the exponential dispersion family (EDF) is needed. The EDF is a general family of distributions for Y, that has continuous as well as discrete distributions as members, making it a broad class of distributions.

Definition 1.1.2 (Tutz (2011), Section 3.1)**.** A family of distributions $\mathscr{P} := \{P_{\theta,\phi} \,|\, \theta \in \mathbb{R},\, \phi > 0\}$ is said to belong to the exponential dispersion family, if and only if, for some given, real-valued functions $\varphi(\cdot)$, $c(\cdot,\cdot)$ and a non-negative real-valued function $a(\cdot)$, its probability density functions (pdf) or probability mass functions (pmf) can be written as

$$f(y|\theta,\phi) = \exp\left(\frac{y\theta - \varphi(\theta)}{a(\phi)} + c(y,\phi)\right). \tag{1.8}$$

The parameter $\phi > 0$ is called the dispersion parameter and $\theta \in \mathbb{R}$ the natural parameter.

More information on the EDF can be found in the work of Jorgensen (1997) and the references therein. With Definition 1.1.2, one further deduces for $i = 1, \ldots, n$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ as well as functions $a_1(\phi), \ldots, a_n(\phi)$

$$\mathbb{E}(Y|\boldsymbol{x}_i) = \varphi'(\theta_i),\ \mathrm{Var}(Y|\boldsymbol{x}_i) = a_i(\phi)\varphi''(\theta_i), \tag{1.9}$$

according to Tutz (2011) (Section 3.4.1) and McCullagh and Nelder (1989) (Section 2.2.2). As one can see, ensuring a positive variance one can conclude that $\varphi''(\cdot) > 0$ since the functions $a_i(\cdot)$ are non-negative according to Definition 1.1.2.

Remark 1.1.3 (Natural exponential family (NEF))**.** For a known dispersion parameter ϕ, (1.8) simplifies as follows

$$f(y|\theta) = \exp\left(y\theta - \varphi(\theta) + c(y)\right). \tag{1.10}$$

In this case, the family of distributions is called natural exponential family (NEF).

Generalized linear models consist of three different components (Agresti (2002), Section 4.1.1):

(i) Random component: the random response variable Y with an independent random sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ of size n, where the (conditional) response distribution of Y_i, $i \in \{1, \ldots, n\}$ given the predictors, is assumed to belong to the EDF (Definition 1.1.2) with natural parameter θ_i and conditional mean denoted by $\mu(\boldsymbol{x}_i) = \mathbb{E}(Y_i \,|\, \boldsymbol{x}_i)$.

(ii) Systematic component: the explanatory variables yielding the linear predictor $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$ ($\eta_i = \boldsymbol{x}_i\boldsymbol{\beta}$, $i \in \{1, \ldots, n\}$, respectively).

(iii) Link function: an invertible function $g(\cdot)$ linking the random and the systematic component, that is $g\left(\mathbb{E}(Y_i|\boldsymbol{x}_i)\right) = \eta_i$ for $i \in \{1, \ldots, n\}$, or, in matrix notation $g\left(\mathbb{E}(\boldsymbol{Y}|\boldsymbol{X})\right) = \boldsymbol{\eta}$, where in the latter case the link function is applied componentwise.

The following definition of a GLM is obtained.

Definition 1.1.4 (Generalized linear model (GLM), Nelder and Wedderburn (1972))**.** Y is assumed to be a random response variable and $\mathscr{X}_1, \ldots, \mathscr{X}_J$ J explanatory variables, measured on an observed sample of size n. Further, it is assumed that the components of the corresponding random response vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ are conditionally independent, given $\boldsymbol{x}_i$, $i \in \{1 \ldots, n\}$ and the distribution of $Y_i|\boldsymbol{x}_i$ belongs to the EDF. Then, for a monotonic, differentiable link function $g(\cdot)$, the GLM models the expected response as follows

$$g(\mathbb{E}(Y_i|\boldsymbol{x}_i)) = \boldsymbol{x}_i\boldsymbol{\beta}, \quad i \in \{1, \ldots, n\}. \tag{1.11}$$

Equivalently, in matrix notation applying the function $g(\cdot)$ componentwise it holds that

$$g(\mathbb{E}(\boldsymbol{Y}|\boldsymbol{X})) = \boldsymbol{X}\boldsymbol{\beta}. \tag{1.12}$$

It is noted that the linear predictor $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$ is linear in the parameter $\boldsymbol{\beta}$, that is, one can also include explanatory variables being non-linear functions of other underlying variables, such as interaction terms or polynomial terms.

Remark 1.1.5. In the framework of GLMs, the pmf or pdf $f(y|\theta, \phi)$ of Definition 1.1.2 may be written for simplicity reasons as $f(y|\boldsymbol{\beta})$.

Remark 1.1.6 (Random design and fixed design)**.** In this thesis, the explanatory variables are considered as fixed, which is denoted in the literature as *fixed design*. Nevertheless, it is also possible to consider the explanatory variables as random, called *random design*. For the latter, one can similarly set up a GLM as in the fixed case (Definition 1.1.4), but with the additional assumption that $(\boldsymbol{X}_1^*, Y_1), \ldots, (\boldsymbol{X}_n^*, Y_n)$ is an iid random sample of $(\mathscr{X}^*, Y)$, according to Bühlmann and Geer (2015). From the methodological point of view, there is no difference between these two cases as argued in Bühlmann and Geer (2011), but when it comes to model misspecification one needs to distinguish, as elaborated in Bühlmann and Geer (2015). References treating a random design in case of an underlying linear model in terms of inference analysis are given by Geer et al. (2014) and Javanmard and Montanari (2014b).

Definition 1.1.7 (Canonical link, Agresti (2002), Section 4.1.1)**.** If the link function in Definition 1.1.4 is chosen such that the conditional mean of the response is transformed to the natural parameter of the EDF, that is, if

$$\theta_i = g(\mu(\boldsymbol{x}_i)),$$

the function $g(\cdot)$ is called the *canonical link*.

In the case of a canonical link function, one can write by (1.11) that

$$\theta_i = \boldsymbol{x}_i\boldsymbol{\beta}, \text{ hence } \theta_i = \eta_i.$$

The choice of the canonical link function comes with some comfortable simplifications of the associated inferential procedure according to Agresti (2015) (Section 4.5.5), e.g. the observed and expected Fisher information matrices coincide, such that the Fisher scoring and Newton-Raphson algorithms are identical.

Having specified the underlying model, i.e. the GLM (1.12), the next natural step is to consider point estimation of the parameters. The predominant estimator in the frequentist approach is the maximum likelihood estimator (MLE), which is provided in Definition 1.1.9 after introducing the (log-)likelihood function in Definition 1.1.8. The MLE is given by the maximizer of the likelihood function, where the latter provides the probability of the observed sample $\boldsymbol{y}$ as a function of the parameter $\boldsymbol{\beta}$. In the following, $f(y|\boldsymbol{\beta})$ is assumed to be a pmf/pdf from the EDF (Definition 1.1.2 and Remark 1.1.5).

Definition 1.1.8 ((Log-)Likelihood function, Casella and Berger (2002), Definition 6.3.1)**.** One assumes $y_1, \ldots, y_n$ to be realizations of an iid sample from a pmf or pdf $f(y|\boldsymbol{\beta})$, where $\boldsymbol{\beta} := (\beta_0, \beta_1, \ldots, \beta_p) \in \mathbb{R}^{p+1}$, is some parameter vector and $\boldsymbol{X} \in \mathbb{R}^{n\times(p+1)}$. For $\boldsymbol{y} := (y_1, \ldots, y_n) \in \mathbb{R}^n$, the likelihood function is defined as

$$l_n(\boldsymbol{\beta}|\boldsymbol{y}) := \prod_{i=1}^{n} f(y_i|\boldsymbol{\beta}) \tag{1.13}$$

and the corresponding log-likelihood function

$$L_n(\boldsymbol{\beta}|\boldsymbol{y}) := \ln(l_n(\boldsymbol{\beta}|\boldsymbol{y})), \tag{1.14}$$

for which $L_n(\boldsymbol{\beta})$ and $l_n(\boldsymbol{\beta})$, respectively, is written for short. It is noted that the dependence on $\boldsymbol{X}$ is not expressed in $L_n(\boldsymbol{\beta}|\boldsymbol{y})$ and $l_n(\boldsymbol{\beta}|\boldsymbol{y})$, respectively, since the design matrix $\boldsymbol{X}$ is assumed as *fixed*.

Definition 1.1.9 (Maximum likelihood estimator, Casella and Berger (2002), Definition 7.2.4)**.** The same setting as in Definition 1.1.8 is assumed. The maximum likelihood estimator (MLE) is defined as the maximizer of the likelihood function, i.e.

$$\hat{\boldsymbol{\beta}}^{(\mathrm{ML})} := \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\arg\max}\, l_n(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\arg\min} -L_n(\boldsymbol{\beta}).$$

Under the choice of the canonical link function, the log-likelihood function is strictly concave if $\boldsymbol{X}$ is of full rank (cf. Agresti (2015), Section 4.5.1) which is thoroughly examined in the following remark.

Remark 1.1.10. This remark analyzes the definiteness of the Fisher information matrix, hence whether the log-likelihood function is strictly concave or not.

(i) First, the reason why the log-likelihood function is strictly concave under the choice of the canonical link function if $\boldsymbol{X} \in \mathbb{R}^{n\times(p+1)}$ is of full rank is obtained. Here, "full rank" means that $\mathrm{rank}(\boldsymbol{X}) = p + 1$, for which it is necessary to ensure $n \geq p + 1$ since it holds that

$$\mathrm{rank}(\boldsymbol{X}) \leq \min(n, p + 1).$$

By straightforward calculations, it can be seen that the Hessian matrix of the log-likelihood function, being equal to the expected Fisher information matrix $I_F(\boldsymbol{\beta})$ since a canonical

link is used (according to Agresti (2015), Section 4.5.5), is given by

$$\begin{aligned} I_F(\boldsymbol{\beta}) &:= \mathbb{E}\left(-\frac{\partial^2 L_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}\right) \\ &= -\frac{\partial^2 L_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} = -\boldsymbol{X}^T \operatorname{diag}\left(\frac{\varphi''(\boldsymbol{x}_1\boldsymbol{\beta})}{a_1(\phi)}, \ldots, \frac{\varphi''(\boldsymbol{x}_n\boldsymbol{\beta})}{a_n(\phi)}\right)\boldsymbol{X}, \end{aligned} \tag{1.15}$$

where $a_i(\cdot)$ and $\varphi(\cdot)$ are introduced in Definition 1.1.2. It holds that the matrix $\boldsymbol{W}$, defined by

$$\boldsymbol{W} := \operatorname{diag}\left(\frac{\varphi''(\boldsymbol{x}_1\boldsymbol{\beta})}{a_1(\phi)}, \ldots, \frac{\varphi''(\boldsymbol{x}_n\boldsymbol{\beta})}{a_n(\phi)}\right), \tag{1.16}$$

is strictly positive definite since $\frac{\varphi''(\boldsymbol{x}_i\boldsymbol{\beta})}{a_i(\phi)} > 0 \; \forall\, i = 1, \ldots, n$, one refers to the explanation right below (1.9). Having that, one can define $\boldsymbol{A} := \boldsymbol{W}^{1/2}\boldsymbol{X} \in \mathbb{R}^{(p+1)\times(p+1)}$ such that

$$\boldsymbol{A}^T\boldsymbol{A} = (\boldsymbol{W}^{1/2}\boldsymbol{X})^T(\boldsymbol{W}^{1/2}\boldsymbol{X}) = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} = -I_F(\boldsymbol{\beta}).$$

Now, assuming that $\boldsymbol{X}$ is of full rank, one can conclude that $\boldsymbol{A}$ is of full rank. Further, even though this is not used for the following lines, one can conclude that $\boldsymbol{A}^T\boldsymbol{A}$ and the Hessian matrix of the log-likelihood (1.15) are of full rank, i.e. $I_F(\boldsymbol{\beta})$ is invertible.
For some $\boldsymbol{z} \in \mathbb{R}^{p+1}\backslash\{\boldsymbol{0}\}$, one can deduce

$$\boldsymbol{z}^T\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\boldsymbol{z} = \boldsymbol{z}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{z} = (\boldsymbol{A}\boldsymbol{z})^T(\boldsymbol{A}\boldsymbol{z}) = ||\boldsymbol{A}\boldsymbol{z}||^2 > 0, \tag{1.17}$$

where the latter inequality uses that $\boldsymbol{A}$ is of full rank, hence $\boldsymbol{A}\boldsymbol{z} = \boldsymbol{0} \Leftrightarrow \boldsymbol{z} = \boldsymbol{0}$. Otherwise, if it would not be known that $\boldsymbol{A}$ is of full rank, one could just write $||\boldsymbol{A}\boldsymbol{z}||^2 \geq 0$ which would not be sufficient for *strict* concavity. Consequently, $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$ is strictly positive definite, such that the Hessian of the log-likelihood (1.15) is strictly negative definite, yielding that the log-likelihood is strictly concave.

(ii) It can be seen from (1.17) that the Fisher information matrix is always at least positive semi definite. Thus, it can always be ensured that the log-likelihood function in the setting considered here, i.e. using the canonical link function, is concave so the negative log-likelihood function is convex.

(iii) In Appendix B, regularity conditions are provided ensuring theoretical properties like consistency of the MLE, where the latter is discussed in Appendix C.

1.1.3 Logistic Regression

This section specifies the choice of the link function $g(\cdot)$ used in this thesis. Considering a binary random response variable, there are different appropriate choices for the link function, such as the logit link, probit link or complementary log-log link, which are provided in Definition 1.1.11.

Definition 1.1.11 (Agresti (2007), Section 3.2)**.** In case of a binomial random response variable, an appropriate link function should map the interval $[-\infty, \infty]$ to $[0, 1]$. Some well-known choices for link functions in case of a binomial random response variable are given by the following functions, where μ is written instead of $\mu(\boldsymbol{x})$ for simplicity.

(i) The *probit link* is defined as

$$g(\mu) := \Phi^{-1}(\mu),$$

where $\Phi(\cdot)$ is the cumulative density function (cdf) of the standard normal distribution $N(0, 1)$.

(ii) The *log-log link* is defined as

$$g(\mu) := -\log(-\log(\mu)),$$

in particular, this link function is the inverse of the cdf of the Gumbel distribution according to Agresti (2015), Section 5.6.3.

(iii) The *complementary log-log link* is defined as

$$g(\mu) := \log(-\log(1-\mu)).$$

It is called "complementary" log-log, since the log-log link (ii) is applied to $1-\mu$, the "complement" of μ.

(iv) The *logit link* is defined as

$$g(\mu) := \log\left(\frac{\mu}{1-\mu}\right), \tag{1.18}$$

in particular, this link function is the inverse of the cdf of the standard logistic distribution according to Agresti (2015), Section 5.1.3.

In general, it depends on the application context which link function is the most appropriate choice. Comparing the probit link, emerging from the inverse of the cdf of the standard normal distribution, and the logit link, emerging from the inverse of the cdf of the standard logistic distribution, one first notes that the standard logistic distribution looks similar to the standard normal distribution but the standard logistic distribution has slightly thicker tails and a standard deviation of 1.8 rather than one according to Agresti (2007) (Section 3.2.5). Consequently, the probit link function is a bit more steep than the logit link function. However, they both share the property of being symmetrical around 0.5 in terms of satisfying $g(\mu) = -g(1-\mu)$, and, with respect to goodness-of-fit measures, the resulting fits using the probit or the logit link are often similar (McCullagh and Nelder (1989), Section 4.3.1). In contrast, the log-log and the complementary log-log link do not satisfy the provided symmetry relation, i.e. they could be more suitable in cases where the probability that the response variable equals one approaches one at a different rate than it approaches zero, which could be reasonable for settings where the number of zeros and ones in the observed sample $\boldsymbol{y}$ significantly differ from each other (Ming-Hui Chen and Shao (1999)). However, one can say that the most common choice is the logit link (iv) as it is the canonical link function for a binary random response variable (shown below), yielding some comfortable simplifications, as explained earlier right below Definition 1.1.7.

In this thesis, the canonical link function for the considered binomial response is chosen, i.e. the logit link function (1.18). The fact that this is the canonical link can be directly seen as

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right) \overset{\mu=\pi}{=} \underbrace{\log\left(\frac{\pi}{1-\pi}\right)}_{=\theta, \text{ one compares Remark 1.1.12}} \Rightarrow g(\mu) = \theta. \tag{1.19}$$

Consequently, the logit link supplies the *logistic regression* model (cf. Agresti (2002), equation (5.9)), i.e.

$$\mathbb{E}(Y|\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}\boldsymbol{\beta})}, \tag{1.20}$$

which is the underlying model that is considered throughout this thesis. It is noted that, by the binomial distribution of the random response variable, it holds that

$$\mu(\boldsymbol{x}) = \mathbb{E}(Y|\boldsymbol{x}) = \mathbb{P}(Y = 1|\boldsymbol{x}) = \pi(\boldsymbol{x}), \tag{1.21}$$

where sometimes μ and π might be written for short as in (1.19).

The fact that the binomial distribution belongs to the EDF is needed for the rigorous application of Definition 1.1.2, for that reason the following remark is provided.

Remark 1.1.12. In Defintion 1.1.2, the first argument of the pmf $f(y|\theta, \phi)$ is the quantity y, while writing $\boldsymbol{v} := (y, \boldsymbol{x}) \in \mathbb{R}^{p+2}$ for an observation, one can similarly write $f(\boldsymbol{v}|\theta, \phi)$ for the pmf to emphasize that it also depends on $\boldsymbol{x}$, even though it is assumed *fixed and given*. For $Y \sim \text{Bin}(1, \pi)$, the following holds for *logistic regression* (Myers et al. (2010), Section 5.1)

$$\begin{aligned} \theta &= \log(\pi/(1-\pi)), \\ \varphi(\theta) &= \log(1 + \exp(\theta)), \\ a(\phi) &= 1, \\ c(y, \phi) &= 1. \end{aligned}$$

Consequently, the binomial distribution belongs to the EDF and NEF.
Considering an iid sample of size n and writing $\boldsymbol{v}_i = (y_i, \boldsymbol{x}_i) \in \mathbb{R}^{p+2}$ for $i = 1, \ldots, n$ and

$$\pi_i := \frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i\boldsymbol{\beta})}, \tag{1.22}$$

the log-likelihood function (1.14) considering logistic regression (1.20) is given by

$$L_n(\boldsymbol{\beta}) \quad = \quad L_n(\boldsymbol{\beta} \,|\, \boldsymbol{v}) = \sum_{i=1}^{n} \ln((1-\pi_i)^{1-y_i}\pi_i^{y_i}) = \sum_{i=1}^{n} y_i\boldsymbol{x}_i\boldsymbol{\beta} - \ln(1 + \exp(\boldsymbol{x}_i\boldsymbol{\beta})). \tag{1.23}$$

Next, the structure of the coefficient vector $\boldsymbol{\beta}$ is discussed. By (1.20), for every dummy variable $\mathscr{X}_{j,i}$, $j \in \{1, \ldots, J\}$, $i \in \{1, \ldots p_j\}$ in the model, one needs to estimate a coefficient $\beta_{j,i}$. Thus, one needs to obtain an estimate $\hat{\boldsymbol{\beta}}$ of the vector of coefficients $\boldsymbol{\beta}$ being of length $p + 1$ with parameter space $\Omega := \mathbb{R}^{p+1}$. As already shortly mentioned in Section 1.1.2, the structure of the unknown coefficient vector is as follows

$$\boldsymbol{\beta} = (\beta_{int}, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J)^T \in \mathbb{R}^{p+1}, \tag{1.24}$$

where β_{int} denotes the intercept and

$$\boldsymbol{\beta}_j = (\beta_{j,1}, \ldots, \beta_{j,p_j}) \in \mathbb{R}^{p_j}, \; j \in \{1, ..., J\}, \tag{1.25}$$

is the parameter sub-vector corresponding to the j-th factor. The estimates are denoted by $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_j$, respectively, where an upper index may be added indicating which method was used for the estimation, e.g. $\hat{\boldsymbol{\beta}}^{(\text{ML})}$ and $\hat{\boldsymbol{\beta}}_j^{(\text{ML})}$ for the MLE according to Definition 1.1.9.

Remark 1.1.13. Later in this thesis, when considering *asymptotic* properties of different estimation methods letting $n \to \infty$, the quantities p, and J respectively, can be either *fixed* or *diverging* with n. To be more precise, in the *diverging* case, the number of covariates J and the dimension of the parameter space $p + 1$, in particular p, may depend on the sample size n. Whenever it is necessary to emphasize that p or J are allowed to grow with n, the expressions

p_n and J_n, respectively, are employed. The differentiation between fixed p and diverging p_n will (only) play a role in theorems where asymptotic behavior is investigated, thus the differentiation will only be done when it comes to asymptotic properties, otherwise p is considered as *fixed*. Throughout the whole thesis, it is always clearly specified whether the case of p being fixed (denoted by *fixed p* or *fixed case*) or diverging with n (denoted by *diverging* p_n or *diverging case*) is observed. Clearly, writing *fixed p* also means that J is fixed, similarly for *diverging* p_n. Further, $\boldsymbol{X}$ and the estimate $\hat{\boldsymbol{\beta}}$ also depend on n, but since this is *always* the case the subscript is omitted for simplicity of notation, except for theorems where this is explicitly needed and mentioned.

To sum up, the previously introduced *fixed* explanatory variables and a binary random response variable Y are considered, where the expected value of the random response variable given the explanatory variables is modeled by a logistic regression model. The next section will address the question on the estimation of the coefficient vector $\boldsymbol{\beta}$.

1.2 Penalized Regression

The straightforward procedure to estimate $\boldsymbol{\beta}$ would be to use the maximum likelihood (ML) approach maximizing the (log-) likelihood function obtaining the MLE $\hat{\boldsymbol{\beta}}^{(\mathrm{ML})}$ by fitting (for example) a Newton-Raphson algorithm (cf. Agresti (2007), Section 3.5). Nevertheless, there may occur some problems with this classical approach, which were touched in the introduction of this thesis and are rigorously examined in the following.

1.2.1 High-Dimensional Models and Sparsity

One recalls that the parameter space of the estimate of the regression coefficient vector $\boldsymbol{\beta}$ is

$$\Omega := \mathbb{R}^{p+1},$$

hence the dimension of the parameter space is $p+1$. Further, the number of factors is J, where each covariate $j \in \{1, \dots, J\}$ contributes p_j dummy variables - and thus entries in the coefficient vector - to the model, one compares (1.24) and (1.25). Consequently, since $p = \sum_{j=1}^{J} p_j$, it is clear that the number of parameters to be estimated, which is $p+1$, may be very large compared to the sample size n, even though only a moderate size of factors is incorporated in the model.

In cases where p is "large", the model is called *high-dimensional*. High-dimensional models include those where (i) $p > n$ and also those where (ii) $p \leq n$ together with p being "large". In (i), the term of p being "large" refers to the fact that p is large compared to the sample size. To conclude, one can say for short that high-dimensional models are those where p is large, where the quotation marks are left out for simplicity reasons in the following. On the one hand, for (i), the MLE may be unreliable from a theoretical point of view, which is emphasized in Lederer (2022) (Section 1.2) for the case of an identity link function. On the other hand, for (ii), the MLE may exist from the theoretical point of view, however its calculation in practice comes with complex inversion tasks of large matrices, which may be problematic. As introduced below, penalized regression techniques can be able to resolve such issues, providing a convenient tool to obtain estimates of regression coefficients for high-dimensional models. These techniques rely on the so-called sparsity assumption, which is introduced next.

In this thesis, models are considered under the *sparsity assumption*, which means that it is assumed that the true underlying model is *sparse*. In particular, sparse models are models

where it is known in advance that *most* of the regression parameters are *equal to zero* (Lederer (2022), Section 2.2), i.e. most explanatory variables do not contribute significantly in modeling Y. In terms of factors, the term of sparsity is transferred to parameter sub-vectors being equal to zero, which is rigorously defined later in this chapter (Definition 1.2.4). Assuming that the model is sparse, it is known that the majority of the parameters in the model, in particular the sub-vectors corresponding to the covariates, are zero. In other words, it is assumed that only a few of the considered factors have a significant influence on the random response variable.

This sparsity assumption can be motivated with a variety of applications, one of them given in the introduction of this thesis. Under the sparsity assumption, the goal is to *identify* those factors having an influence on the response and *estimate* the effect, meaning the corresponding coefficient sub-vector, in *one* step. The identification of influential covariates is also referred to as *variable selection* or *factor selection*, where the latter is used in the framework of factors. However, considering factors, it is further desirable to identify those levels of a factor that have the same influence on Y, which further contributes to a more sparse model, as specified below.

1.2.2 Levels Fusion for Categorical Covariates

In the same way as understanding that there may be just a handful of covariates having an influence on the random response variable, one can also think about the impact of two levels of a factor on the random response variable. That is, with categorical covariates, it may happen that two levels of a factor have the same influence on the random response variable. In this case, the goal is to fuse those two levels, since it is unnecessary to keep both in the model. This procedure is called *levels fusion*. Clearly, this is not restricted to fusion of *two* levels, one can also fuse more than two levels corresponding to the same factor.

Variable/factor selection as well as levels fusion, while simultaneously performing parameter estimation, can be reached through *penalized regression* employing a penalty function $P_\lambda(\boldsymbol{\beta})$, which is the basis of this thesis and is introduced below.

1.2.3 The Objective Function

The general procedure of *penalized regression* is to minimize the sum of the negative log-likelihood function $-L_n(\boldsymbol{\beta}) : \mathbb{R}^{p+1} \to \mathbb{R}$ and an appropriate penalty function $P_\lambda(\boldsymbol{\beta}) : \mathbb{R}^{p+1} \to \mathbb{R}$ with $\lambda \geq 0$ being a tuning parameter which is specified below. One defines the *objective function* $M_{pen}(\boldsymbol{\beta}) : \mathbb{R}^{p+1} \to \mathbb{R}$ as

$$M_{pen}(\boldsymbol{\beta}) := -L_n(\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta}), \quad \boldsymbol{\beta} \in \mathbb{R}^{p+1}, \tag{1.26}$$

and the *penalized regression estimator* as

$$\hat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\arg\min}\, M_{pen}(\boldsymbol{\beta}). \tag{1.27}$$

Writing $\hat{\boldsymbol{\beta}}$ refers to (1.27), with no specific chosen penalty function. Otherwise, for specific penalty functions, an upper index is added indicating the specific applied penalty function $P_\lambda(\boldsymbol{\beta})$.
The procedure above is referred to as *penalized regression*, or *penalized logistic regression* if an underlying logistic regression model is considered. Further, this method is sometimes also referred to as *regularized regression* and the methods, thus the choices of the penalty function

$P_\lambda(\boldsymbol{\beta})$, are called *regularization methods*.

Clearly, choosing $\lambda = 0$ in (1.26) and minimizing $M_{pen}(\boldsymbol{\beta})$ yields the MLE $\hat{\boldsymbol{\beta}}^{(\text{ML})}$, where a non-significant effect is estimated by a very small value (in absolute value), but the coefficient of such an effect will not be *exactly* set to zero. Hence, the purpose of adding the penalty term $P_\lambda(\boldsymbol{\beta})$ is that these estimates of non-significant effects are set to zero to obtain a sparse model, where the tuning parameter λ controls the amount of imposed penalization. The effect of employing a penalty function can be demonstrated by considering the following popular penalized regression example, the so-called *lasso* introduced by Tibshirani (1996) which uses the L_1 norm as penalty function $P_\lambda(\boldsymbol{\beta})$, i.e.

$$P_\lambda^{(\text{Lasso})}(\boldsymbol{\beta}) := \lambda||\boldsymbol{\beta}||_1. \tag{1.28}$$

By penalizing the L_1 norm of the coefficient vector, this penalty function enforces that the entries of the coefficient vector $\boldsymbol{\beta}$ are shrunken (in absolute value) towards zero for increasing λ, while starting from some value of λ, they are exactly set to zero (Hastie et al. (2015), Section 2.2). A graphical illustration is provided later in this chapter (Section 1.4). The tuning parameter λ is required to decide on the level of shrinkage, balancing model fit and model complexity. In practice, the tuning parameter is determined using cross-validation (CV), where more details on the CV procedure are obtained below. Using the L_2 norm as penalty function $P_\lambda(\boldsymbol{\beta})$, that is,

$$P_\lambda^{(\text{Ridge})}(\boldsymbol{\beta}) := \lambda||\boldsymbol{\beta}||_2 \tag{1.29}$$

results in the *Ridge* penalty (Hoerl and Kennard (1970)), which applies shrinkage to the coefficients' estimates, but is not able to perform variable or factor selection. The latter fact is caused by the structure of the constraint region, which is explained in detail in Hastie et al. (2020) (Section 2.2, especially Figure 2.2). Consequently, with the Ridge penalty function, the coefficient estimates are not exactly set to zero, i.e. this method does *not* yield a sparse model. Another popular approach is the Elastic Net introduced by Zou and Hastie (2005), which is a convex combination of the Ridge and the lasso penalty. There exist also penalty functions that are applied to (pairwise or adjacent) differences of entries of the coefficient vector, which is used for shrinkage of the respective differences and levels fusion, more details are provided in Section 1.3.

The choice of the penalty function $P_\lambda(\boldsymbol{\beta})$ depends on the application context, the objective, as well as the structure of the covariates. Thus, by imposing different penalty functions, different objectives can be fulfilled, although there are multiple penalty functions even within the same objective, as discussed in more details in the course of this chapter.

1.2.4 Tuning: k-Fold Cross-Validation

The tuning parameter $\lambda \geq 0$ controls the impact of the penalty on the resulting estimates, in particular the amount of penalization that is imposed. For penalties performing factor/variable selection, the tuning parameter controls the sparsity in terms of factor/variable selection of the resulting model, whereas for fusion-type penalties, the tuning parameter controls the amount of performed fusions in the resulting model, which is also a type of sparsity. Clearly, there is a need for approaches to determine the "best" tuning parameter λ.

A general commonly used approach to determine λ is to execute *k-fold cross-validation* (CV) for $k \in \mathbb{N}$. That is, one randomly partitions the available dataset containing $\boldsymbol{y}$ and $\boldsymbol{X}$, denoted by $\mathscr{D}$, into k disjoint sets $\mathscr{D}_1, \ldots, \mathscr{D}_k$ such that $\mathscr{D} = \dot{\cup}_{i=1}^k \mathscr{D}_i$. This partition is carried out in a manner such that the derived subsets are roughly equal-sized ($\approx \frac{n}{k}$ samples contained

in each $\mathscr{D}_i$), of course it is not always possible that they are strictly equal-sized. With a predetermined grid for λ candidates, $\lambda \in \{\lambda_{min}, \ldots, \lambda_{max}\}$, the CV-error is calculated for each $\lambda \in \{\lambda_{min}, \ldots, \lambda_{max}\}$, explained below, and then one decides for the "best" λ_{opt}, that is, the one with the smallest CV-error.

It remains to explain how the CV-error is calculated for some λ. First, one fixes $\lambda \in \{\lambda_{min}, \ldots, \lambda_{max}\}$. Then, one goes through all the partitions $\mathscr{D}_i$, $i \in \{1, \ldots, k\}$, one fits the model with tuning parameter λ on the dataset $\mathscr{D} \backslash \mathscr{D}_i$ and validates the resulting fit in terms of the prediction error on the left-out dataset $\mathscr{D}_i$. To be more precise, the *predictive deviance* measure is used which is defined next.

Definition 1.2.1 (Predictive deviance, Agresti (2002) (Section 4.1.5))**.** $\boldsymbol{v} := (\boldsymbol{y}, \boldsymbol{X})$ is some given data, where $\boldsymbol{X}$ is a fixed design matrix and $\boldsymbol{y}$ is a realization from an iid sample from the random response variable Y. $\hat{\boldsymbol{\beta}}$ is the estimate of the coefficient vector from the GLM under consideration. Additionally, $L_n(\hat{\boldsymbol{\mu}}|\boldsymbol{v}) := L_n(\hat{\boldsymbol{\beta}}|\boldsymbol{v})$ is the log-likelihood of the GLM under consideration and $L_n(\boldsymbol{y}|\boldsymbol{v})$ the log-likelihood of the *saturated model*, i.e. the model of perfect fit. Then, the *predictive deviance* of the GLM under consideration is defined as

$$D(\boldsymbol{v}|\hat{\boldsymbol{\mu}}) := -2(L_n(\hat{\boldsymbol{\mu}}|\boldsymbol{v}) - L_n(\boldsymbol{y}|\boldsymbol{v})). \tag{1.30}$$

For short $D(\boldsymbol{v}|\hat{\boldsymbol{\mu}})$ is called *deviance*.

The deviance is based on the likelihood ratio statistic for nested models (cf. Appendix C) and it is an appropriate choice for measuring the goodness-of-fit having an underlying logistic regression model (cf. Tutz (2011) Sections 3.7.1 and 3.8.1). For the linear model, the deviance equals the sum of squared errors, that is $\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$.

Coming back to the CV technique, for each $\lambda \in \{\lambda_{min}, \ldots, \lambda_{max}\}$, the model is fitted k times, each time on the data subsample $\mathscr{D} \backslash \mathscr{D}_i$, $i = 1, \ldots, k$. The corresponding deviances are denoted by $D(\boldsymbol{v}|\hat{\boldsymbol{\mu}}^{(i)})$. Then the CV-error is defined by

$$\text{CV}(\lambda) := \frac{1}{k}\sum_{i=1}^{k} D(\boldsymbol{v}|\hat{\boldsymbol{\mu}}^{(i)}). \tag{1.31}$$

Finally, one chooses λ_{opt} as the one with the smallest CV-error. Clearly, the result (choice of λ_{opt}) depends on the number k of folds as well as the fineness of the grid for λ, i.e. the finer the grid, the more accurate the choice of λ_{opt}. Nevertheless, a finer grid for λ and a higher value for k comes with more involving computations. The most involving choice for k is $k = n$, which is known as *leave-one-out cross-validation* (LOOCV). As elaborated in James et al. (2014) (Section 5.4.1), LOOCV gives approximately unbiased estimates of the test error. Nevertheless, it produces a higher variance than k-fold CV with $k < n$, so one faces a bias-variance trade-off when choosing k for CV. In the literature, $k = 5$ and $k = 10$ are convenient, commonly used choices (e.g. Hastie et al. (2015) Section 3.5.1, James et al. (2014) Sections 5.1.3 and 5.1.4) balancing the computational costs as well as the bias-variance trade-off. In the simulation studies provided in this thesis, $k = 5$ is chosen.

1.2.5 Theoretical Properties: Definitions

Definitions concerning (possible) theoretical properties that are of interest and may be fulfilled by the penalized regression approaches introduced later in this chapter are provided.

The definition of *groupwise structure* is supplied, which is a property of the structure of the underlying explanatory variables in the regression model. In terms of factors, this structure was implicitly introduced in Section 1.1, as specified in Remark 1.2.3 below.

Definition 1.2.2 (Groupwise structure, Bühlmann and Geer (2011), Section 4.2)**.** One says that the parameter vector $\boldsymbol{\beta} = (\beta_{int}, \beta_1, \beta_2, \ldots, \beta_p)$ can be structured into groups, or that it shows a groupwise structure, if the index set $\{1, \ldots, p\}$ corresponding to the entries of $\boldsymbol{\beta}$ (except for the intercept) can be partitioned into $g \in \mathbb{N}$ disjoint groups. To be more precise, it is required that there exist pairwise disjoint index sets $\mathscr{G}_1, \ldots, \mathscr{G}_g$ such that

$$\mathscr{G}_1, \ldots, \mathscr{G}_g \subseteq \{1, \ldots, p\} \text{ with } \dot{\bigcup}_{j=1}^{g} \mathscr{G}_j = \{1, \ldots, p\}.$$

Further, it is required that there exists at least one group $\mathscr{G}_j$, $j \in \{1, \ldots, g\}$, for which $|\mathscr{G}_j| \geq 2$. One may also say that the *explanatory variables* show a groupwise structure.

Remark 1.2.3. In a natural way, categorical explanatory variables, i.e. factors, show a groupwise structure with $g = J$ and $|\mathscr{G}_j| = p_j$. For simplicity, one denotes the groups $\mathscr{G}_1, \ldots, \mathscr{G}_J$ by $1, \ldots, J$, one compares (1.24). In the following, this groupwise structure induced by the factor variables is considered.

In what follows, $\boldsymbol{\beta}^* \in \mathbb{R}^{p+1}$ denotes the *true* underlying coefficient vector and by $\boldsymbol{\beta}_j^* \in \mathbb{R}^{p_j}$ the true sub-vector corresponding to the j-th factor.

Definition 1.2.4 (True active sets and sparsity)**.**

(i) In the case of fixed p, the *true* underlying structure is said to be sparse if, without loss of generality, the *true* active set, which is given by

$$A^* := \{j \in \{1, \ldots J\} \,|\, \boldsymbol{\beta}_j^* \neq \mathbf{0}\} = \{j \in \{1, \ldots J\} \,|\, ||\boldsymbol{\beta}_j^*||_2 \neq 0\},$$

can be written as $A^* = \{1, ..., j_0\}$ with $j_0 \in \mathbb{N}$, $j_0 < J$. 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 $p_0 := \sum_{j=1}^{j_0} p_j$.

(ii) In the case of diverging p_n, the *true* underlying structure is said to be sparse if, without loss of generality, the *true* active set, which is given by

$$A^* := \{j \in \{1, \ldots J_n\} \,|\, \boldsymbol{\beta}_j^* \neq \mathbf{0}\} = \{j \in \{1, \ldots J_n\} \,|\, ||\boldsymbol{\beta}_j^*||_2 \neq 0\},$$

can be written as $A^* = \{1, ..., j_{0,n}\}$ with $j_{0,n} \in \mathbb{N}$, $j_{0,n} < J_n \; \forall\, n \in \mathbb{N}$. To avoid confusion, a lower index n in A^* is omitted, since the truth is fixed and does not really depend on the sample size. Additionally, it is assumed that departing from some $\tilde{n} \in \mathbb{N}$, all truly influential factors are included in the model, that is

$$\exists\, \tilde{n} \in \mathbb{N},\, j_0 \in \mathbb{N} \,:\, |A^*| = j_0 \;\forall n \geq \tilde{n}.$$

In other words, also for the diverging case, there exists a set A^* with $|A^*| = j_0$ containing all influential factors for n large enough. For asymptotic analyses, this means that one can operate with A^* also in the diverging case letting $n \to \infty$. For this j_0 introduced above it holds that $j_0 < J_n \,\forall n > \tilde{n}$. The Fisher information matrix can be written in the same way as above, but with $\boldsymbol{I}_F$, $\boldsymbol{I}_{11}$ and p_0 being replaced by $\boldsymbol{I}_{F,n}$, $\boldsymbol{I}_{11,n}$ and $p_{0,n}$, respectively.

Throughout the *whole thesis* it is assumed that the true underlying structure is sparse.

Analogously to the true active set (Definition 1.2.4), the estimated active set is defined as follows.

Definition 1.2.5 (Estimated active set)**.** One assumes that a penalized regression method is applied according to (1.26) yielding a penalized regression estimate $\hat{\boldsymbol{\beta}}$ given by (1.27). Then, the *estimated* active set is defined as

(i) case of fixed p

$$A_n := \{j \in \{1, \dots J\} \,|\, \hat{\boldsymbol{\beta}}_j \neq \mathbf{0}\} = \{j \in \{1, \dots J\} \,|\, ||\hat{\boldsymbol{\beta}}_j||_2 \neq 0\},$$

(ii) case of diverging p_n

$$A_n := \{j \in \{1, \dots J_n\} \,|\, \hat{\boldsymbol{\beta}}_j \neq \mathbf{0}\} = \{j \in \{1, \dots J_n\} \,|\, ||\hat{\boldsymbol{\beta}}_j||_2 \neq 0\}.$$

Since the estimated active set A_n depends on $\hat{\boldsymbol{\beta}}$ and on the chosen penalty function $P_\lambda(\boldsymbol{\beta})$, an upper index is added to A_n whenever necessary, indicating which penalty function was utilized to obtain $\hat{\boldsymbol{\beta}}$.

In a similar way, the following definition concerning fusion sets is provided.

Definition 1.2.6 (True active fusion set)**.** The set C^* is defined as the active set of true fusions, that is,

(i) case of fixed p

$$C^* := \left\{(j,k,m) \in \{1,\dots,J\} \times \{1,\dots,p_j\}^2 \,|\, \beta^*_{j,k} \neq \beta^*_{j,m}\right\},$$

(ii) case of diverging p_n

$$C^* := \left\{(j,k,m) \in \{1,\dots,J_n\} \times \{1,\dots,p_j\}^2 \,|\, \beta^*_{j,k} \neq \beta^*_{j,m}\right\}.$$

If a factor j is ordinal, rather than nominal, only the pairwise differences $\beta^*_{j,k} \neq \beta^*_{j,k-1}$ for $k \in \{2,\dots,p_j\}$ are considered in the definition of C^* above.

The reason why C^* is called *true active fusion set*, even though it include those indices for which $\beta^*_{j,k} \neq \beta^*_{j,m}$, is, that it can be understood as the true active set on the space of all pairwise (or adjacent) differences. Consequently, C^* includes the *active*, i.e. nonzero, differences. The same applies for the *estimated* active fusion set C_n defined next.

Definition 1.2.7 (Estimated active fusion set)**.** One assumes that a penalized regression method is applied according to (1.26) yielding a penalized regression estimate $\hat{\boldsymbol{\beta}}$ given by (1.27). Then, the *estimated* active set of fusions (also called the estimated active fusion set) is defined as

(i) case of fixed p

$$C_n := \left\{(j,k,m) \in \{1,\dots,J\} \times \{1,\dots,p_j\}^2 \,|\, \hat{\beta}_{j,k} \neq \hat{\beta}_{j,m}\right\}, \tag{1.32}$$

(ii) case of diverging p_n

$$C_n := \left\{(j,k,m) \in \{1,\dots,J_n\} \times \{1,\dots,p_j\}^2 \,|\, \hat{\beta}_{j,k} \neq \hat{\beta}_{j,m}\right\}. \tag{1.33}$$

Similar to Definition 1.2.6, the sets above need to be adjusted if a factor j is ordinal, since in this case one only compares the adjacent coefficients of a factor. Further, these sets can also be defined as estimated active sets (Definition 1.2.5) on the space of all pairwise or adjacent differences. Since C_n depends on a penalized regression estimator $\hat{\boldsymbol{\beta}}$, which depends on the chosen penalty function $P_\lambda(\boldsymbol{\beta})$, whenever needed, an upper index may be added to C_n indicating which penalty function was chosen to obtain $\hat{\boldsymbol{\beta}}$.

In addition, the following fusion sets F^* and F_n are introduced, which are restricted to the truly influential factors, reporting on the one hand the indices where the truth is fused (F^*) and on the other hand those where the estimates are fused (F_n). In contrast, the sets C^* and C_n are not restricted to the truly influential factors and the indices are reported where the truth and the estimates are *not* fused, which is the reason why C^* and C_n are called *active* sets of fusions.

Definition 1.2.8. One assumes that a penalized regression method is applied according to (1.26) yielding a penalized regression estimate $\hat{\boldsymbol{\beta}}$ given by (1.27). Then, the set of true fusions for the influential factors, as well as the corresponding set for the estimate are defined as

$$\begin{aligned} F^* &:= \left\{(j,k,m) \in \{1,\dots,J\} \times \{1,\dots,p_j\}^2 \,|\, \boldsymbol{\beta}_j^* \neq \mathbf{0},\, \beta_{j,k}^* = \beta_{j,m}^*,\, k < m\right\}, \\ F_n &:= \left\{(j,k,m) \in \{1,\dots,J\} \times \{1,\dots,p_j\}^2 \,|\, \boldsymbol{\beta}_j^* \neq \mathbf{0},\, \hat{\beta}_{j,k} = \hat{\beta}_{j,m},\, k < m\right\}. \end{aligned}$$

Remark 1.2.9. For some penalization methods *fusion consistency* is shown with respect to C_n and C^* and for some with respect to F_n and F^*, where it is always clarified on which set the fusion consistency is based on. Fusion consistency with respect to C_n and C^* is stronger, since $F_n \subseteq C_n^c$ and $F^* \subseteq (C^*)^c$. Nevertheless, starting from Chapter 2, the diverging case is examined, such that the focus lies on F_n and F^* since these sets are also bounded in the diverging case, which is the reason why no distinction is made in the definition above between the fixed and the diverging case.

For the following definition, one recalls that setting single entries of the coefficient vector to zero, i.e. removing single variables from the model, as done for example considering metric covariates, is referred to as *variable selection*. Considering categorical covariates, one aims to either remove the whole factor from the model or keep the whole factor in the model, which is referred to as *factor selection*. The following definition is about consistency in variable selection as well as about consistency in factor selection, where this thesis focuses on the latter since categorical explanatory variables are considered.

Definition 1.2.10 (Consistency in variable selection/factor selection)**.** A regularization method is said to be consistent in variable selection/factor selection, if the resulting active set A_n (resulting from the penalized estimator $\hat{\boldsymbol{\beta}}$) satisfies

$$\lim_{n\to\infty} \mathbb{P}(A_n = A^*) = 1. \tag{1.34}$$

This definition is valid for both cases of p being fixed and p_n diverging with n in the setup of this thesis. If for a penalized regression method it is known that

$$\lim_{n\to\infty} \mathbb{P}(A^* \subseteq A_n) = 1 \tag{1.35}$$

holds, this does not imply consistency in variable/factor selection as defined in (1.34). Nevertheless, if (1.35) is satisfied it corresponds also to a (weaker) kind of consistency in the sense that all significant variables or factors have been selected. Thus, it is also referred to as variable/factor selection consistency, but it is always clearly specified whether (1.34) or (1.35) holds.

For the next definition, the order in probability notation is used, i.e. O_p and o_p, where the first one refers to stochastic boundedness and the latter to convergence in probability. In particular, for a sequence of real-valued random variables $(Z_n)_{n\in\mathbb{N}}$ and a real-valued sequence $(a_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}$ one defines

$$Z_n = O_p(a_n) :\Leftrightarrow \forall \epsilon > 0 \; \exists \, M, N \in \mathbb{N} \, : \, \mathbb{P}\left(\left|\frac{Z_n}{a_n}\right| > M\right) < \epsilon \; \forall n > N,$$

$$Z_n = o_p(a_n) :\Leftrightarrow \operatorname*{p\,lim}_{n\to\infty} \frac{Z_n}{a_n} = 0, \text{ i.e.} \forall \epsilon > 0 \, : \, \lim_{n\to\infty} \mathbb{P}\left(\left|\frac{Z_n}{a_n}\right| > \epsilon\right) = 0.$$

Definition 1.2.11. A regularization method is said to be $\sqrt{n}$ consistent if, for the resulting penalized estimate $\hat{\boldsymbol{\beta}}$, it holds that

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

It is noted that this directly yields $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|| = o_p(1)$, which is referred to as consistency in the literature (Casella and Berger (2002), Section 10.1.1). However, consistency does not imply $\sqrt{n}$ consistency.

Definition 1.2.12. A regularization method is said to satisfy the asymptotic normality property if, for the resulting penalized estimate $\hat{\boldsymbol{\beta}}$ and a positive semi-definite, symmetric matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{j_0 \times j_0}$ with $\sigma_{j,j} > 0$ ($\sigma_{j,j}$ being the entries on the diagonal of Σ), it holds that

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{A^*} - \boldsymbol{\beta}^*_{A^*}) \to_d N(\mathbf{0}, \boldsymbol{\Sigma}) \quad (n \to \infty).$$

Here, $N(\mathbf{0}, \boldsymbol{\Sigma})$ is the j_0 dimensional normal distribution. This definition is valid for both cases of p being fixed and p_n diverging with n in the setup of this thesis. Further, $\hat{\boldsymbol{\beta}}_{A^*}$ and $\boldsymbol{\beta}^*_{A^*}$ are the sub-vectors of $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$, respectively, containing only the components that belong to the active set A^*.

Since the true active set A^* is not known in advance, the question that naturally arises is how the sub-vectors $\hat{\boldsymbol{\beta}}_{A^*}$ and $\boldsymbol{\beta}^*_{A^*}$ of the definition above can be identified. For this, the notion of oracle MLE and oracle properties is crucial, which is provided next. The following definition follows Fan et al. (2020) (Section 5.8) as well as Fan and Li (2001) (Section 3.2).

Definition 1.2.13 (Oracle MLE, oracle properties)**.**

(i) An *oracle MLE* $\left(\hat{\boldsymbol{\beta}}^{(\mathrm{ML})}\right)^{\mathrm{o}}$ is defined as a MLE knowing in advance which of the variables are influential and which are noise variables, i.e. knowing A^*. Specifically,

$$\left(\hat{\boldsymbol{\beta}}^{(\mathrm{ML})}\right)^{\mathrm{o}} = \left(\left(\hat{\boldsymbol{\beta}}^{(\mathrm{ML})}\right)^{\mathrm{o}}_{A^*}, \mathbf{0}\right),$$

where

$$\left(\hat{\boldsymbol{\beta}}^{(\mathrm{ML})}\right)^{\mathrm{o}}_{A^*} := \operatorname*{arg\,min}_{\boldsymbol{\beta}_{(A^*)^c}=\mathbf{0}} -L_n(\boldsymbol{\beta}).$$

Following further the lines of Fan et al. (2020) (Section 5.8), the oracle MLE supplies the MLE of the elements of the coefficient vector that are nonzero, assuming that A^* is known. Nevertheless, this oracle MLE is not a "real" estimator, since A^* is unknown. But, it is used as a reference evaluating the theoretical properties of an estimator, which is clarified next.

(ii) A penalized regression technique with some resulting estimator $\hat{\boldsymbol{\beta}}$ is said to satisfy the *oracle properties* if it is consistent in variable/factor selection in the sense of (1.34) and has the same asymptotic distribution as the oracle MLE. One also says that the estimator is an oracle estimator.

Remark 1.2.14 (On the oracle properties)**.** From the definition above, it is known that an oracle estimator needs to have the same asymptotic distribution as the oracle MLE. Under certain regularity conditions (Appendix B.1 and Appendix C.1), it is known that

$$\sqrt{n}\left(\left(\hat{\boldsymbol{\beta}}^{(\mathrm{ML})}\right)^{\mathrm{o}}_{A^*} - \boldsymbol{\beta}^*_{A^*}\right) \to_d N(\mathbf{0}, \boldsymbol{\Sigma}),$$

for a positive semi-definite, symmetric matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{j_0 \times j_0}$ with $\sigma_{j,j} > 0$ and $N(\mathbf{0}, \boldsymbol{\Sigma})$ being the j_0 dimensional normal distribution. This result can be found in Fan et al. (2020) (Section 5.8) as well as in Appendix C.1 of this thesis. Consequently, an estimator $\hat{\boldsymbol{\beta}}$ has the same asymptotic distribution as the oracle MLE if and only if the asymptotic normality property of Definition 1.2.12 is satisfied. To conclude, an estimator $\hat{\boldsymbol{\beta}}$ satisfies the oracle properties if it is consistent in variable/factor selection and satisfies the asymptotic normality property.

Next, different penalty functions being appropriate in the framework of categorical explanatory variables are investigated in detail. The motivation and characteristics of the penalty functions are provided, as well as associated computational methods and coefficient paths. Further, theoretical properties are shown, such as $(\sqrt{n})$ consistency, consistency in factor selection or fusion consistency, respectively, and asymptotic normality. To do so, the case of p *being fixed* is considered, whereas the more complex case allowing p_n to grow with n is treated later in this thesis for the new penalty function introduced in Chapter 2.

1.3 Penalties on Differences

This section introduces penalty functions $P_\lambda(\cdot)$ which are not directly applied to the coefficient vector, or coefficient sub-vector, rather they are applied on the differences of the (dummy) variables' coefficients. These type of penalties enforce fusion of levels belonging to the same factor that have the same influence on the random response variable. For a nominal factor, all pairwise differences of levels corresponding to the same factor are considered, while for an ordinal factor only adjacent ones are build. The reason for that is, for a nominal factor there is no ordering between the levels, hence all pairwise differences need to be considered. In contrast, for an ordinal factor, the fact that the levels can be reasonably ordered yields that only the adjacent differences need to be considered, since if two levels need to be fused, they have to be neighbored.

More specifically, the L_1 penalization of the differences (Section 1.3.1) is discussed, which is denoted by CAS-L_1. Moreover, the similarly structured L_0 penalization of the differences is investigated (Section 1.3.2), which is denoted by CAS-L_0. A computational method to calculate the estimates in practice is further presented in Section 1.3.3. This method is applied in Section 1.3.4 on a specific example, providing the associated coefficient paths which illustrate the dependence of the coefficients' estimates on the tuning parameter.

1.3.1 L_1 Penalization of Differences (CAS-L_1)

The L_1 norm as penalization was widely used for selection purposes, e.g. for variable selection in the lasso (Tibshirani (1996)). It is recalled that the L_1 norm of $\boldsymbol{t} \in \mathbb{R}^\kappa$, $\boldsymbol{t} = (t_1, \ldots, t_\kappa)$ for some $\kappa \in \mathbb{N}$ is given by $||\boldsymbol{t}||_1 := \sum_{i=1}^{\kappa} |t_i|$.

Considering an analysis of variance (ANOVA) model, Bondell and Reich (2009) introduced the CAS-ANOVA penalty ("Collapsing and Shrinkage in ANOVA") and Gertheiss and Tutz (2010b) observed this penalty in the linear model case. Bondell and Reich (2009) use the so-called sum to zero constraint for identifiability, i.e.

$$\sum_{k=1}^{p_j} \beta_{j,k} = 0 \;\; \forall\, j \in \{1, \ldots, J\},$$

while Gertheiss and Tutz (2010b) use zero as reference category

$$\beta_{j,0} = 0 \;\; \forall\, j \in \{1, \ldots, J\}, \tag{1.36}$$

where the latter is followed here (Section 1.1.1). The penalty function, which is called CAS-L_1 penalty or simply L_1 penalty for short in this thesis, is defined by

$$P_\lambda^{(\text{CAS-}L_1)}(\boldsymbol{\beta}) := \sum_{j=1}^{J} P_\lambda^{\text{CAS-}L_1}(\boldsymbol{\beta}_j), \tag{1.37}$$

where

$$\text{factor } j \in \{1, \ldots, J\} \text{ nominal:} \qquad P_\lambda^{\text{CAS-}L_1}(\boldsymbol{\beta}_j) := \lambda \sum_{0 \le r < s \le p_j} w_j^{(rs)} \, |\beta_{j,r} - \beta_{j,s}|, \tag{1.38}$$

$$\text{factor } j \in \{1, \ldots, J\} \text{ ordinal:} \qquad P_\lambda^{\text{CAS-}L_1}(\boldsymbol{\beta}_j) := \lambda \sum_{k=1}^{p_j} w_j^{(k)} \, |\beta_{j,k} - \beta_{j,k-1}|. \tag{1.39}$$

Here, in (1.38), $w_j^{(rs)}$ for $j \in \{1, \ldots, J\}$ and $r, s \in \{1, \ldots, p_j\}$, $r \neq s$, or $w_j^{(k)}$ in (1.39), respectively, are optional weights which are discussed below. If the tuning parameter λ is allowed to depend on the sample size n, a lower index is added, i.e. λ_n. In the same way as for the linear model, one can consider penalty function (1.37) for the setting of logistic regression, replacing the quadratic loss in the objective function (1) of Gertheiss and Tutz (2010b) by the negative log-likelihood function (1.23). However, in the mentioned papers, logistic regression or other GLMs are not considered.

This type of penalty function enforces coefficients of levels corresponding to the same factor to be set equal if they are close enough to each other, hence have a similar influence on the random response variable. Further, by choosing reference category zero (1.36) and including the reference category in the sum of the penalizing term, as done above in (1.38) and (1.39), respectively, these sums also contain the absolute values of the coefficients. In particular, for e.g. (1.38) the sum can be re-written as

$$\sum_{0 \le r < s \le p_j} w_j^{(rs)} \, |\beta_{j,r} - \beta_{j,s}| = \sum_{1 \le r < s \le p_j} w_j^{(rs)} \, |\beta_{j,r} - \beta_{j,s}| + \sum_{0 < s \le p_j} w_j^{(0s)} \, |\beta_{j,s}|, \tag{1.40}$$

which yields that not only the differences of the coefficients are penalized (first sum in (1.40)), but also the absolute values of the coefficients (second sum in (1.40)). Consequently, the coefficients (and not only their differences) are shrunken towards zero, similar to the lasso penalty (1.28) which (only) penalizes the absolute values of the coefficients. In the setting considering factors with reference category zero, shrinkage of a single entry of the coefficient vector refers to shrinkage of the difference between this particular entry (i.e. the corresponding category) and the reference category. Consequently, if a *single* coefficient is set to zero, this refers to fusion with the reference category. If a *whole* coefficient sub-vector of a factor is set to zero, this refers to factor selection. However, CAS-L_1 is not designed for *factor* selection.

Choice of Weights

The subsequent lines comment on the choice of the weights appearing in (1.38) and (1.39). In general, $w_j^{(rs)}$ (nominal case) and $w_j^{(k)}$ (ordinal case) are optional weights that can be appropriately chosen, or set to one. The goal of the weighting process is to transform all factors on the same scale, meaning that one takes into account a possible inhomogeneous distribution of the numbers of levels in each factor. Further, weights can account for the fact that the design may be unbalanced, i.e. if the number of observations for the levels of a factor differ from each other. Since the following choice of weights appear at several points throughout this thesis, it is discussed and justified here once.

With $n_r^{(j)}$ denoting the number of observations of level r of factor j, Bondell and Reich (2009) propose to use the following weights, which are also employed in Gertheiss and Tutz (2010b)

$$w_j^{(rs)} := \frac{2}{p_j+1} \cdot \sqrt{\frac{n_r^{(j)} + n_s^{(j)}}{n}} \qquad \text{(nominal factor)}, \tag{1.41}$$

$$w_j^{(k)} := \sqrt{\frac{n_k^{(j)} + n_{k-1}^{(j)}}{n}} \qquad \text{(ordinal factor)}. \tag{1.42}$$

These weights (1.41) and (1.42) are non-adaptive weights, where the adaptive versions are introduced below. Detailed justifications on the choice of the weights can be found in Gertheiss and Tutz (2010b) as well as in Bondell and Reich (2009), where both use the notion of "standardized predictors". A short form of the idea behind that is given in the subsequent paragraph, where the arguments are taken from these two references.

The CAS-L_1 penalty function is applied to the *differences* of the coefficients. Hence, these differences can be interpreted as predictors in an augmented regression model. Thus, to standardize these predictors, it is reasonable to consider an augmented design matrix $\boldsymbol{Z}$. To be more concrete, one considers $\boldsymbol{\delta}$ to contain all the coefficient vectors $\boldsymbol{\beta}_j$ and all the corresponding pairwise differences for $j = 1, \ldots, J$. That is, one considers for $j \in \{1, \ldots, J\}$ the vector

$$\boldsymbol{\gamma} := (\beta_{j,1} - \beta_{j,2}, \ldots, \beta_{j,1} - \beta_{j,p_j}, \ldots, \beta_{j,p_j-1} - \beta_{j,p_j}) \tag{1.43}$$

which contains all pairwise differences of factor j *except* for the differences with the reference category zero. Then the overparametrized model with parameter vector $\boldsymbol{\delta} = (\boldsymbol{\beta}_1, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\beta}_J, \boldsymbol{\gamma}_J)$ is considered, having $p + \frac{1}{2}\sum_{j=1}^{J} p_j(p_j - 1)$ parameters. $\boldsymbol{\delta} = \boldsymbol{M}\boldsymbol{\beta}$ is written for an appropriate matrix $\boldsymbol{M}$ consisting of blocks of identity matrices and blocks of matrices constructing all differences. Now, the corresponding design matrix for the augmented regression task is given by $\boldsymbol{Z}$, which is $n \times (p + \frac{1}{2}\sum_{j=1}^{J} p_j(p_j - 1))$, and satisfies $\boldsymbol{Z}\boldsymbol{\delta} = \boldsymbol{X}\boldsymbol{\beta}$, hence $\boldsymbol{Z}\boldsymbol{M} = \boldsymbol{X}$. Following further the arguments of Bondell and Reich (2009), one could also consider the regression model in the augmented parameter space using design matrix $\boldsymbol{Z}$, adding the set of constraints $\boldsymbol{\delta} = \boldsymbol{M}\boldsymbol{\beta}$ and directly standardizing the columns. However, they elaborate that this procedure comes with a loss of interpretation. In particular, the parameters that result from the explained procedure have a different interpretation than the differences of the parameters, i.e. the coefficients, as argued in Bondell and Reich (2009). Thus, the weights being the euclidean norm of the columns of $\boldsymbol{Z}$ that correspond to each of the differences are used, which directly yield the weights (1.41) and (1.42).

Applying the steps explained in the paragraph above, CAS-L_1 can be interpreted as lasso on an augmented parameter space, which is elaborated in Gertheiss and Tutz (2010b). This fact is

helpful to obtain convenient computational methods, as well as theoretical properties that are already known for lasso.

Having discussed the choice and the origin of non-adaptive weights, the focus is turned to the choice of *adaptive weights*. With an initial estimator $\tilde{\boldsymbol{\beta}}$, the adaptive weights are chosen as

$$\tilde{w}_j^{(rs)} := w_j^{(rs)} \cdot |\tilde{\beta}_{j,r} - \tilde{\beta}_{j,s}|^{-1} \qquad \text{(nominal covariates)}, \tag{1.44}$$

$$\tilde{w}_j^{(k)} := w_j^{(k)} \cdot |\tilde{\beta}_{j,k} - \tilde{\beta}_{j,k-1}|^{-1} \qquad \text{(ordinal covariates)}, \tag{1.45}$$

according to Gertheiss and Tutz (2010b) and Bondell and Reich (2009). By using these adaptive weights with the initial estimator being the MLE for example, one can see that weights for pairs where the MLEs are close to each other become inflated, such that the penalty will favor to set the corresponding difference to zero and fuse the coefficients. This makes intuitively sense, but one also needs to note that the result will depend on the quality of the initial estimate, so it needs to be chosen carefully. When CAS-L_1 is used with adaptive weights as given above, this method is simply called *adaptive CAS-*L_1. It is also possible to use other $\sqrt{n}$ consistent estimators as initial estimators in the adaptive weights, where the $\sqrt{n}$ consistency will ensure convenient theoretical properties, as discussed in the following.

Theoretical Properties of CAS-L_1

Bondell and Reich (2009) (Section 4) and Gertheiss and Tutz (2010b) (Appendix, Proposition 1) show that under some regularity conditions, the adaptive CAS-L_1 satisfies asymptotic normality and fusion consistency in terms of differences in their considered underlying model. It is assumed that δ_i is an entry of

$$\boldsymbol{\delta} := (\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_J), \tag{1.46}$$

where $\boldsymbol{\delta}_j = (\boldsymbol{\beta}_j, \boldsymbol{\gamma}_j)$, $j \in \{1, \ldots, J\}$ (cf. (1.43)) contains all possible pairwise/adjacent differences of the coefficient vector corresponding to factor j including the differences with the reference category zero. The number of entries in $\boldsymbol{\delta}_j$ is given by $\frac{1}{2}p_j(p_j-1)+p_j$ if the j-th factor is nominal and by $2p_j - 1$ if it is ordinal. $\boldsymbol{\delta}^*$ is the vector containing the true values of the differences. The true active set (of fusions) C^*, introduced in Definition 1.2.6, is given by the entries of the vector of the true differences $\boldsymbol{\delta}^*$ being truly zero, hence, written in terms of $\boldsymbol{\delta}^*$

$$C^* \quad = \quad \left\{ i \in \left\{ 1, \ldots, p + \frac{1}{2} \sum_{j=1}^{J} p_j(p_j - 1) \right\} \mid \delta_i^* \neq 0 \right\}. \tag{1.47}$$

In a similar way, with $\hat{\boldsymbol{\beta}}$ denoting a penalized regression estimator and $\hat{\boldsymbol{\delta}}$ the respective differences, the active set of the estimator is obtained (one compares Definition 1.2.7)

$$C_n = \left\{ i \in \left\{ 1, \ldots, p + \frac{1}{2} \sum_{j=1}^{J} p_j(p_j - 1) \right\} \mid \hat{\delta}_i \neq 0 \right\}. \tag{1.48}$$

Gertheiss and Tutz (2010b) show for CAS-L_1 in the *linear model* that, under certain regularity conditions, the estimator of the differences is asymptotically normal distributed and that the probability that C_n and C^* coincide converges to one (cf. Proposition 1 of this reference). They show this statement for $J = 1$ factor and argue that it can be directly generalized to the case of multiple categorical inputs, $J \geq 2$. A similar result can be found in Bondell and Reich (2009).

The natural question that arises is whether these properties similarly hold for the general case treating GLMs instead of the special case of linear regression. In particular, it is of interest whether the mentioned proposition in the paragraph above holds for *logistic regression*. Diving into the proof of Proposition 1 in Gertheiss and Tutz (2010b), one can see that the asymptotic normality part of the proposition can be adjusted for logistic regression, which is considered e.g. in Zou (2006) for adaptive lasso. To do so, one needs to ensure the consistency of the unpenalized MLE appearing in the adaptive weights, which can be done imposing the regularity conditions (Reg1)-(Reg3) which are given and discussed in Appendix B.1. The fact that these conditions ensure consistency of the MLE is provided in Theorem C.1.2 later in this thesis which particularly references Casella and Berger (2002), Theorem 10.1.6. In the linear model case, Gertheiss and Tutz (2010b) explain in their proof of Proposition 3 that consistency of the unpenalized least squares estimator is ensured by the condition $\frac{n_i}{n} \to c_i \in (0,1)$ $(n \to \infty)$. This condition is not required in case of logistic regression, as justified in Theorem 1.3.1 below. Further, to transfer Proposition 1 of Gertheiss and Tutz (2010b) to the logistic regression case, one needs to adjust the adaptive weights using the general unpenalized MLE instead of the particular unpenalized least squares estimate.
In Theorem 1.3.1 below, $\boldsymbol{\Sigma}$ is some covariance matrix, which is singular because of the over-parametrization coming from considering the differences, as explained above, where for further details reference is made to Bondell and Reich (2009). For the specific form of $\boldsymbol{\Sigma}$, one examines the proof of Proposition 1 in Gertheiss and Tutz (2010b). However, for the purposes of this thesis, no further analysis on the structure of this covariance matrix is necessary. Summing up, the following theorem is provided transferring Proposition 1 of Gertheiss and Tutz (2010b) to the case of logistic regression.

Theorem 1.3.1 (Asymptotic normality and fusion consistency of CAS-L_1 for *logistic regression*, fixed p)**.** One assumes that $\frac{\lambda_n}{\sqrt{n}} \to 0$, as well as $\lambda_n \to \infty$, for $n \to \infty$. Further, one sets $J = 1$ and $\hat{\boldsymbol{\beta}}^{(\mathrm{ML})}$ is the unpenalized MLE of the parameters of an underlying logistic regression model (Definition 1.1.9). One assumes (Reg1)-(Reg3) of Appendix B.1, ensuring that $\hat{\boldsymbol{\beta}}^{(\mathrm{ML})}$ is consistent (Theorem C.1.2). For the adaptive weights $\tilde{w}_1^{(rs)} := w_1^{(rs)} |\hat{\beta}_r^{(\mathrm{MLE})} - \hat{\beta}_s^{(\mathrm{MLE})}|^{-1}$ with $w_1^{(rs)} := \frac{1}{p_1+1}\sqrt{n_r + n_s}$ ((1.44) and (1.41)), respectively, one assumes that $w_1^{(rs)} \to q_{rs}$ $(n \to \infty)$ for some $q_{rs} \in (0, \infty)$ $\forall\, r, s \in \{1, \ldots, p_j\}$, $r \neq s$. C^* is defined as in (1.47), $\hat{\boldsymbol{\delta}}^{(\mathrm{CAS}-L_1)}$ as in (1.46) and $C_n^{(\text{CAS-}L_1)}$ as in (1.48), where the latter two quantities are based on the *penalized logistic regression* estimate $\hat{\boldsymbol{\beta}}^{(\mathrm{CAS}-L_1)}$ with penalty function (1.38). Then, it holds that

$$\begin{aligned} &\text{(i)} \quad \sqrt{n}(\hat{\boldsymbol{\delta}}_{C^*}^{(\text{CAS-}L_1)} - \boldsymbol{\delta}_{C^*}^*) \to_d N(\mathbf{0}, \boldsymbol{\Sigma}), \\ &\text{(ii)} \quad \lim_{n\to\infty} \mathbb{P}(C_n^{(\text{CAS-}L_1)} = C^*) = 1. \end{aligned}$$

Proof. Before starting with the proof, it is noted that in (i) above, one may ask the question how to identify the sub-vectors $\hat{\boldsymbol{\delta}}_{C^*}^{(\text{CAS-}L_1)}$ and $\boldsymbol{\delta}_{C^*}^*$ of $\hat{\boldsymbol{\delta}}^{(\text{CAS-}L_1)}$ and $\boldsymbol{\delta}^*$, respectively, since C^* is not known in advance. For that reason, reference is made to some previous Definitions 1.2.12 and 1.2.13 as well as to Remark 1.2.14 where this question is answered.

Referring to Zou (2006) (Theorem 4 with $\gamma = 1$), a similar result for adaptive lasso in GLMs is provided, i.e. logistic regression is covered. In Section 2.3 of Gertheiss and Tutz (2010b), it is shown that for the case of an ordinal factor, CAS-L_1 can be seen as lasso on an augmented parameter space of all adjacent differences, i.e. estimating $\boldsymbol{\delta}$. As already explained earlier introducing the weights for CAS-L_1, a transformation matrix $\boldsymbol{M}$ is obtained which satisfies $\boldsymbol{M}\boldsymbol{\beta} = \boldsymbol{\delta}$ and the design matrix $\boldsymbol{X}$ is similarly transformed $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{M}$. After the estimation by executing lasso with the transformed design matrix, the estimates of $\boldsymbol{\beta}$ are provided by a

back-transformation. Finally, if a nominal factor is considered, similar arguments apply by re-ordering the levels in an increasing manner. It is further noted that in this theorem *adaptive* weights are employed for CAS-L_1, consequently the mentioned results for *adaptive* lasso are suitable. To sum up, the claim follows applying Theorem 4 of Zou (2006) using the augmented parameter space as explained above according to Gertheiss and Tutz (2010b). □

Proposition 2 of Gertheiss and Tutz (2010b) provides consistency of the estimator $\hat{\boldsymbol{\beta}}^{(\text{CAS}-L_1)}$ in the *linear model*, in the sense that $||\hat{\boldsymbol{\beta}}^{(\text{CAS}-L_1)} - \boldsymbol{\beta}^*|| = o_p(1)$, however, it does not provide information about the rate of convergence, as $\sqrt{n}$ consistency does (Definition 1.2.11). The goal is to transfer this proposition to the case of logistic regression. As one can see in the proof of Theorem 1.3.2, uniqueness of the unpenalized MLE is needed, which is ensured if the log-likelihood function is strictly concave. For this, conditions on the definiteness of the Fisher information matrix are needed, provided in (Reg2), as well as $\text{rank}(\boldsymbol{X}) = p + 1$, for which one consults Remark 1.1.10 (i). Furthermore, the consistency of the unpenalized MLE is ensured by (Reg1)-(Reg3) and, as explained right before Theorem 1.3.1, the condition $\frac{n_i}{n} \to c_i \, (n \to \infty)$ imposed in the linear model case can be omitted.

Theorem 1.3.2 (Consistency of CAS-L_1 for *logistic regression*, fixed p)**.** One assumes a fixed tuning parameter $\lambda \in [0, \infty)$. Further, one sets $J = 1$ and (Reg1)-(Reg3) of Appendix B.1 are supposed to hold. Further one assumes that $\text{rank}(\boldsymbol{X}) = p + 1$. For the weights $w_1^{(rs)} := \frac{1}{p_1+1}\sqrt{n_r + n_s}$ (1.41), one assumes that $w_1^{(rs)} \to q_{rs} \, (n \to \infty)$ for some $q_{rs} \in (0, \infty) \, \forall \, r, s \in \{1, \ldots, p_j\}$, $r \neq s$. Then, the *penalized logistic regression* estimate $\hat{\boldsymbol{\beta}}^{(\text{CAS}-L_1)}$ with penalty chosen by (1.38) is consistent, that is $\forall \, \varepsilon > 0$ it holds

$$\lim_{n \to \infty} \mathbb{P}(||\hat{\boldsymbol{\beta}}^{(\text{CAS}-L_1)} - \boldsymbol{\beta}^*|| > \varepsilon) = 0.$$

Proof. Steps similar to Gertheiss and Tutz (2010b) (proof of Proposition 2) are provided, adjusting the loss function being the negative log-likelihood assuming an underlying logistic regression model. That is, the objective function is given by

$$M_{pen}^{(\text{CAS-}L_1)}(\boldsymbol{\beta}) = -L_n(\boldsymbol{\beta}) + \lambda \sum_{0 \leq r < s \leq p_1} w_1^{(rs)} \, |\beta_{1,r} - \beta_{1,s}|,$$

where λ is assumed fixed in this theorem. It is known that $-L_n(\boldsymbol{\beta})$ is strictly convex, so unique minimizers exist under (Reg1)-(Reg3) and $\text{rank}(\boldsymbol{X}) = p + 1$. $\hat{\boldsymbol{\beta}}^{(\text{ML})}$ is the uniquely determined unpenalized MLE, thus the minimizer of $-L_n(\boldsymbol{\beta})$ and $\hat{\boldsymbol{\beta}}^{(\text{CAS}-L_1)}$ the minimizer of $M_{pen}^{(\text{CAS-}L_1)}(\boldsymbol{\beta})$. Consequently, $\hat{\boldsymbol{\beta}}^{(\text{ML})}$ also uniquely minimizes $-\frac{1}{n}L_n(\boldsymbol{\beta})$ while $\hat{\boldsymbol{\beta}}^{(\text{CAS}-L_1)}$ minimizes $\frac{1}{n}M_{pen}^{(\text{CAS-}L_1)}(\boldsymbol{\beta})$. By Gertheiss and Tutz (2010b), it holds

$$\begin{aligned}
\frac{1}{n}M_{pen}^{(\text{CAS-}L_1)}(\hat{\boldsymbol{\beta}}^{(\text{CAS}-L_1)}) &\to_p -\frac{1}{n}L_n(\hat{\boldsymbol{\beta}}^{(\text{ML})}), \\
\frac{1}{n}M_{pen}^{(\text{CAS-}L_1)}(\hat{\boldsymbol{\beta}}^{(\text{CAS}-L_1)}) &\to_p -\frac{1}{n}L_n(\hat{\boldsymbol{\beta}}^{(\text{CAS}-L_1)}).
\end{aligned}$$

Consequently,

$$-\frac{1}{n}L_n(\hat{\boldsymbol{\beta}}^{(\text{CAS}-L_1)}) \to_p -\frac{1}{n}L_n(\hat{\boldsymbol{\beta}}^{(\text{ML})}).$$

Using that, as done in Gertheiss and Tutz (2010b) for the least squares estimate (LSE), the MLE $\hat{\boldsymbol{\beta}}^{(\text{ML})}$ is uniquely determined as the minimizer of $-\frac{1}{n}L_n(\boldsymbol{\beta})$ being strictly convex, it is deduced

$$\hat{\boldsymbol{\beta}}^{(\text{CAS}-L_1)} \to_p \hat{\boldsymbol{\beta}}^{(\text{ML})},$$

so the consistency of $\hat{\boldsymbol{\beta}}^{(\text{CAS}-L_1)}$ is provided by the consistency of $\hat{\boldsymbol{\beta}}^{(\text{ML})}$. □

The CAS-L_1 penalization can exhibit limitations, as discussed below.

(i) Observing the CAS-L_1 penalty function (1.37) with (1.38) and (1.39), respectively, one can see that the penalization depends on the absolute value of the coefficients' differences. That is, larger differences are penalized more than smaller differences, so the shrinkage depends on the absolute value of the coefficients' differences. A similar phenomenon can be observed in the well-known lasso approach, where the penalization depends on the absolute value of the coefficients producing biased estimates (Fan and Li (2001)). To overcome this issue, as done for lasso in Zou (2006), one can use adaptive weights for CAS-L_1 introduced above. Further, asymptotic normality and fusion consistency for this choice of adaptive weights were stated in Theorem 1.3.1. However, the adaptive weights depend on the quality of the initial estimates, e.g. the unpenalized MLE. Consequently, one aims for an alternative to adaptive weights, especially for cases where a consistent initial estimate is not available.

(ii) Oelker et al. (2014b) state that the CAS-L_1 penalty does not always enforce fusion efficiently, since observing just one ordinal factor as an example (so $J = 1$) with $0 \leq \beta_1 \leq \beta_2 \leq \ldots \leq \beta_{p_1}$, the penalty becomes

$$\sum_{i=1}^{p_1} |\beta_i - \beta_{i-1}| = |\beta_{p_1} - \beta_0| = |\beta_{p_1}|$$

such that essentially the absolute value of the range of the coefficients is penalized.

(iii) One refers to Proposition 1 in Oelker et al. (2014b), which obtains an explicit structure of the penalized least squares estimate $\hat{\boldsymbol{\beta}}$ (hence in the linear model) with the CAS-L_1 penalty ((1.37) and (1.39)) for $J = 1$ ordinal predictor in an orthonormal design without an intercept. By the explicitly obtained structure of this proposition, the authors conclude that the outer categories of the factor are always fused first, no matter how close the coefficients in the middle are. Additionally, no shrinkage is applied for coefficients of levels that are not fused yet with another level, which is further visualized in Oelker et al. (2014b) (Figure 1).

These potential issues motivate the application of another norm on the differences, namely the L_0 "norm" which is discussed next. Actually the L_0 "norm" is not a norm, but in the literature this term is commonly used such that the quotation marks around "norm" are omitted for simplicity.

1.3.2 L_0 Penalization of Differences (CAS-L_0)

The following approach of penalizing with the L_0 norm was introduced by Oelker et al. (2014b). The L_0 norm is defined by

$$||t||_0 := 1_{\{\mathbb{R}\setminus 0\}}(t) \ , t \in \mathbb{R}, \tag{1.49}$$

where $1_S(\cdot)$ denotes the indicator function on a set $S \subseteq \mathbb{R}$. A visualization of the L_0 norm is given in Figure 1.1.

Applying the L_0 norm on the differences of the coefficients, that is $||\beta_{j,r} - \beta_{j,s}||_0$, $j \in \{1, \ldots, J\}$, $r, s \in \{1, \ldots, p_j\}$, $r \neq s$ and summing up, it counts the number of differences (of the entries of

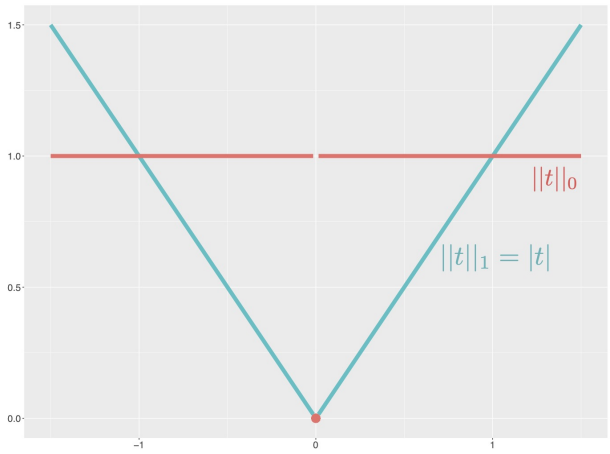

Figure 1.1: *For $t \in [-1.5, 1.5] \subseteq \mathbb{R}$, the L_0 norm (1.49) as well as the L_1 norm are visualized.*

$\boldsymbol{\beta}$) that are nonzero, meaning that the corresponding pairwise or adjacent coefficients are not equal. Each factor is penalized separately, that is

$$P_\lambda^{(\text{CAS-}L_0)}(\boldsymbol{\beta}) := \sum_{j=1}^{J} P_\lambda^{(\text{CAS-}L_0)}(\boldsymbol{\beta}_j), \tag{1.50}$$

similar to CAS-L_1 (1.37). The components $P_\lambda^{(\text{CAS-}L_0)}(\boldsymbol{\beta}_j)$, $j \in \{1, \ldots, J\}$ are again defined differently for nominal and ordinal factors, i.e.

$$\text{factor } j \in \{1, \ldots, J\} \text{ nominal:} \quad P_\lambda^{(\text{CAS-}L_0)}(\boldsymbol{\beta}_j) := \lambda \sum_{0 \leq r < s \leq p_j} w_j^{(rs)} \, ||\beta_{j,r} - \beta_{j,s}||_0, \tag{1.51}$$

$$\text{factor } j \in \{1, \ldots, J\} \text{ ordinal:} \quad P_\lambda^{(\text{CAS-}L_0)}(\boldsymbol{\beta}_j) := \lambda \sum_{k=1}^{p_j} w_j^{(k)} \, ||\beta_{j,k} - \beta_{j,k-1}||_0. \tag{1.52}$$

Here, $w_j^{(rs)}$ and $w_j^{(k)}$, respectively, are optional weights which are specified below. In the following this penalty is referred to as CAS-L_0, or simply L_0. Similarly to CAS-L_1, as explained in Section 1.3.1, this penalty enforces fusion of coefficients of the levels corresponding to the same factor by applying the L_0 norm on the differences of the coefficients. However, since CAS-L_0 just differentiates between a difference being zero or nonzero, this type of penalty applies no shrinkage to the (differences of the) coefficients, and it does not depend on the absolute value of the differences, which is different to CAS-L_1. This distinction will be further clarified and illustrated later in this thesis (Section 1.3.4).

Choice of Weights

Analogously to CAS-L_1 weights can be imposed, considering exactly the same choices as introduced in Section 1.3.1 for the adaptive and non-adaptive weighting schemes, according to Oelker et al. (2014b).

Theoretical Properties of CAS-L_0

To the best of one's knowledge, there are no theoretical properties developed so far for the presented CAS-L_0 approach, neither in penalized logistic regression, nor in the framework of

other GLMs. However, since a new penalty function is introduced in Chapter 2 which includes an L_0 part, $\sqrt{n}$ consistency can be obtained for CAS-L_0 as a special case of the penalty in the mentioned chapter. Thus, to avoid unnecessary redundancies, reference is made to a selection of theorems proven in Chapter 2 in the subsequent lines.

Specifically, under the regularity conditions (Reg1)-(Reg3) given in Appendix B.1, a result on $\sqrt{n}$ consistency is provided, which directly follows from Theorem 2.3.2, where particularly $\sqrt{n}$ consistency for this new penalty function is proven. It is noted that, since the objective function

$$M_{pen}^{(\text{CAS-}L_0)}(\boldsymbol{\beta}) = -L_n(\boldsymbol{\beta}) + P_\lambda^{(\text{CAS-}L_0)}(\boldsymbol{\beta})$$

is not convex, caused by the non-convexity $P_\lambda^{(\text{CAS-}L_0)}(\boldsymbol{\beta})$, the result is about a *local minimizer*, or more precisely a sequence of local minimizers.

Theorem 1.3.3 ($\sqrt{n}$ consistency for CAS-L_0, fixed p)**.** It is supposed that the regularity conditions (Reg1)-(Reg3) from Appendix B.1 hold. One sets $a_n^0 := \max\{\lambda_n w_j^{(rs)}; 0 \leq r < s \leq p_j, j \in \{1, \dots J\}\}$ and assumes $a_n^0 = O_p(1)$. Then, there exists a local minimizer $\hat{\boldsymbol{\beta}}^{(\text{CAS}-L_0)}$ of $M_{pen}^{(\text{CAS-}L_0)}(\boldsymbol{\beta})$ satisfying

$$||\hat{\boldsymbol{\beta}}^{(\text{CAS}-L_0)} - \boldsymbol{\beta}^*||_2 = O_p\left(\frac{1}{\sqrt{n}}\right).$$

Proof. Direct consequence of Theorem 2.3.2 choosing $\lambda_1^n = 0$, which yields $a_n^1 = o_p(1)$. □

With regard to asymptotic normality of CAS-L_0, no results are available for linear models, logistic regression or GLMs in general. The asymptotic normality results proven for the new penalty function in Chapter 2 (Theorem 2.3.5) are not transferable here, since, choosing $\lambda_1^n = 0$ in the referenced theorem, it cannot be guaranteed that $\lambda_1^n \cdot n^{(\gamma-1)/2} \to \infty$, which is a necessary assumption. Thus, further work needs to be investigated to obtain asymptotic normality for CAS-L_0 in logistic regression, as well as for linear regression and other GLMs.

For the same reasons as described above, reference is made in the following to another proof of a theorem presented later in this thesis to avoid unnecessary redundancies. In particular, the fusion consistency result, which is provided in Chapter 2, especially in Theorem 2.3.37, similarly holds for CAS-L_0 setting $\lambda_1^n = 0$ in the mentioned theorem. This directly gives the following result.

Theorem 1.3.4 (Fusion consistency for CAS-L_0, fixed p)**.** It is supposed that the regularity conditions (Reg1)-(Reg3) from Appendix B.1 hold. One sets $a_n^0 := \max\{\lambda_n w_j^{(rs)}; 0 \leq r < s \leq p_j, j \in \{1, \dots J\}\}$ and assumes $a_n^0 = O_p(1)$, thus, there exists a local minimizer $\hat{\boldsymbol{\beta}}^{(\text{CAS}-L_0)}$ of $M_{pen}^{(\text{CAS-}L_0)}(\boldsymbol{\beta})$ satisfying $||\hat{\boldsymbol{\beta}}^{(\text{CAS}-L_0)} - \boldsymbol{\beta}^*||_2 = O_p\left(\frac{1}{\sqrt{n}}\right)$ by Theorem 1.3.3. Further, one assumes that the conditions of Theorem 2.3.26 hold with $\lambda_1^n = 0$. Then it holds for $F_n^{(\text{CAS-}L_0)}$ being the fusion set corresponding to this particular $\sqrt{n}$ consistent $\hat{\boldsymbol{\beta}}^{(\text{CAS}-L_0)}$ that

$$\lim_{n\to\infty} \mathbb{P}(F^* = F_n^{(\text{CAS-}L_0)}) = 1.$$

Proof. Direct consequence of Theorem 2.3.37 choosing $\lambda_1^n = 0$. □

1.3.3 Computation of CAS-L_1 and CAS-L_0

For CAS-L_0, Oelker et al. (2014b) propose to use the PIRLS ("Penalized Iteratively Reweighted Least Squares") algorithm which was introduced in Oelker and Tutz (2013) and is examined for

a *general form* of penalties in Appendix A.1. By the flexibility of the PIRLS algorithm being applicable to a huge variety of penalties, this approach can be used to compute CAS-L_1 estimates as well. For the particular choices of the approximating functions introduced in Appendix A.1 (for short: N_l is an approximation of the applied norm in the penalty and D_l is the derivative), this thesis follows Oelker and Tutz (2013) (Table 1) using

$$\begin{aligned}
\text{CAS-}L_1 \ : \ & N_l(\xi) = \sqrt{\xi^2 + c}\,, \ D_l(\xi) = (\xi^2 + c)^{-1/2} \cdot \xi, \\
\text{CAS-}L_0 \ : \ & N_l(\xi) = \frac{2}{1 + \exp(-\gamma|\xi|)} - 1\,, \\
& D_l(\xi) = \frac{2\gamma}{1 + \exp(-\gamma|\xi|)} \left(1 - \frac{1}{1 + \exp(-\gamma|\xi|)}\right) \frac{\xi}{\sqrt{\xi^2 + c}},
\end{aligned}$$

where in the approximations of CAS-L_0, the absolute value $|\xi|$ is further approximated with $\sqrt{\xi^2 + c}$ as in CAS-L_1.

A visualization of these approximations and the impact of the parameters γ and c is provided in Figure 1.2, where $N_l(\xi)$ is named as $N(\xi)$ in this figure for simplicity. As described above, for the approximation of L_0 it is further used that $|\xi| \approx \sqrt{\xi^2 + c}$, which is not displayed in Figure 1.2 for the sake of simplicity, since the impact of c can already be seen in the plot on the left hand side. In general, c ensures the differentiability (for both L_0 and L_1) and γ determines the steepness of the logistic function (for L_0). For different choices of the parameters, which are displayed in different colors in Figure 1.2, the approximations show different qualities. Roughly speaking, one can see that c should be "small" whereas γ should be chosen "large", which is also supported by the fact that for $\gamma \to \infty$ and $c \to 0$, the approximation of L_0 converges to $||\xi||_0$, according to Oelker and Tutz (2013).

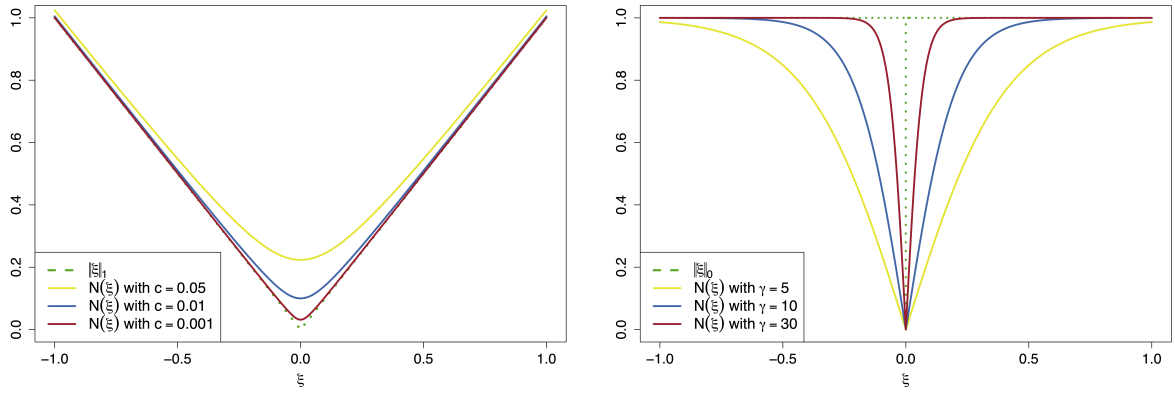

Figure 1.2: *Visualization of the approximations of the L_1 norm $||\xi||_1$ (left) and the L_0 norm $||\xi||_0$ (right) for $\xi \in [-1, 1] \subseteq \mathbb{R}$. In each plot, the dashed green line is the norm itself. However, it is noted that the L_0 norm on the right hand side (dashed green line) is not continuous at zero (one compares Figure 1.1), the dashed green vertical line is only drawn for visualization purposes. A similar figure can be found in Oelker and Tutz (2013).*

1.3.4 Coefficient Paths of CAS-L_1 and CAS-L_0

Coefficient paths are a visualization of the coefficient estimates applying a particular penalized regression method. On the horizontal axis, the tuning parameter is displayed while on the ver-

tical axis the corresponding resulting coefficient estimates are shown. To be more precise, the paths start at the unpenalized MLE (i.e. $\lambda = 0$) on the left. Proceeding to the right on the horizontal axis, the tuning parameter is increased until a pre-chosen maximum tuning parameter $\lambda_{\max}$ is reached. This value is chosen such that all factors are excluded from the model.

In Figure 1.3, coefficient paths of CAS-L_0 and CAS-L_1 for an example including $J = 2$ ordinal factors, where $p_1 = p_2 = 3$, are shown. The data, being of sample size $n = 400$, is simulated with multinomial distributions with equal probabilities and the truth is chosen as $\boldsymbol{\beta}^* = (1.5, 0.7, 1.4, 1.2, -0.3, -0.5, -1)$. In the following sections presenting other regularization methods, the same example is picked up to exhibit coefficient paths, such that all methods can be compared on the basis of a common example.

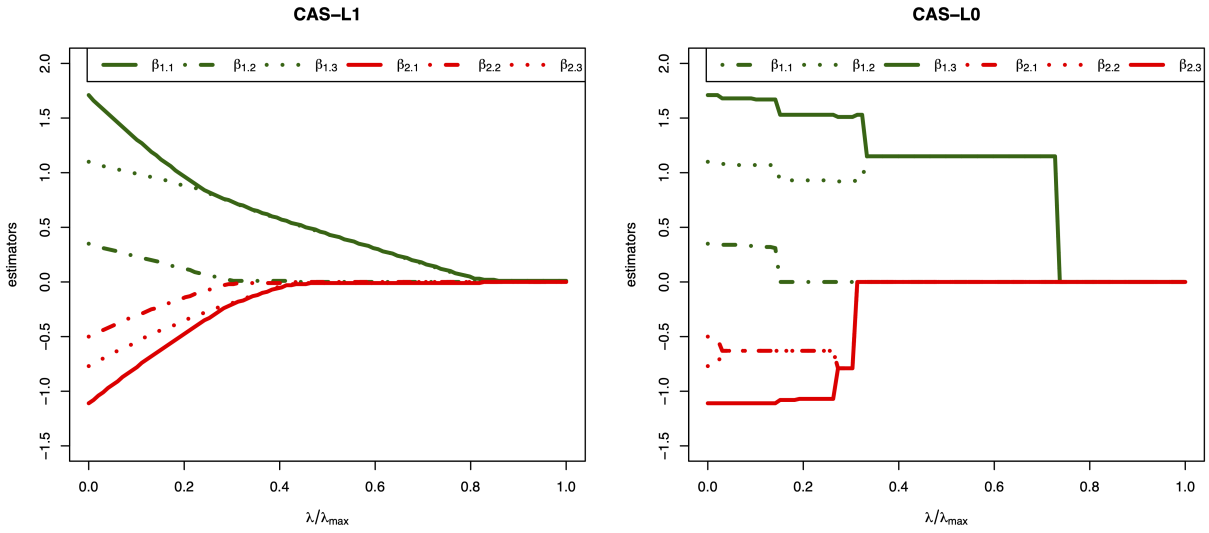

Figure 1.3: *Coefficient paths for CAS-L_1 (left) and CAS-L_0 (right). On the horizontal axis, the tuning parameter increases from left to right, where the coefficient estimates on the leftmost side in each plot are the MLE. In this example, $J = 2$ ordinal factors were chosen with $p_1 = p_2 = 3$ and $\boldsymbol{\beta}^* = (1.5, 0.7, 1.4, 1.2, -0.3, -0.5, -1)$.*

The shapes of the coefficient paths demonstrate the characteristics of both approaches. On the one hand, the L_1 approach shrinks the coefficients towards zero as the tuning parameter increases and fuses coefficients corresponding to the same level if they are "close enough" to each other. Further, it performs (indirect) selection, corresponding to fusion of all levels with the reference category. On the other hand, the L_0 approach performs no shrinkage of the coefficients, but it performs fusion, as well as (indirect) selection.

The fact that, in contrast to L_1, L_0 does not perform shrinkage is caused by the L_0 penalty which just differentiates between an element (a difference) being zero or non-zero as already explained earlier introducing L_0. Figure 1.4 represents this difference by displaying both penalty functions (CAS-L_1 and CAS-L_0) for *one* factor ($J = 1$, $p_1 = 2$). CAS-L_1 on the left hand side enforces shrinkage since the penalty function decreases as the value of the difference $|\beta_1 - \beta_2|$ decreases, whereas CAS-L_0 on the right hand side does not perform shrinkage and only examines whether the coefficients are equal or not.

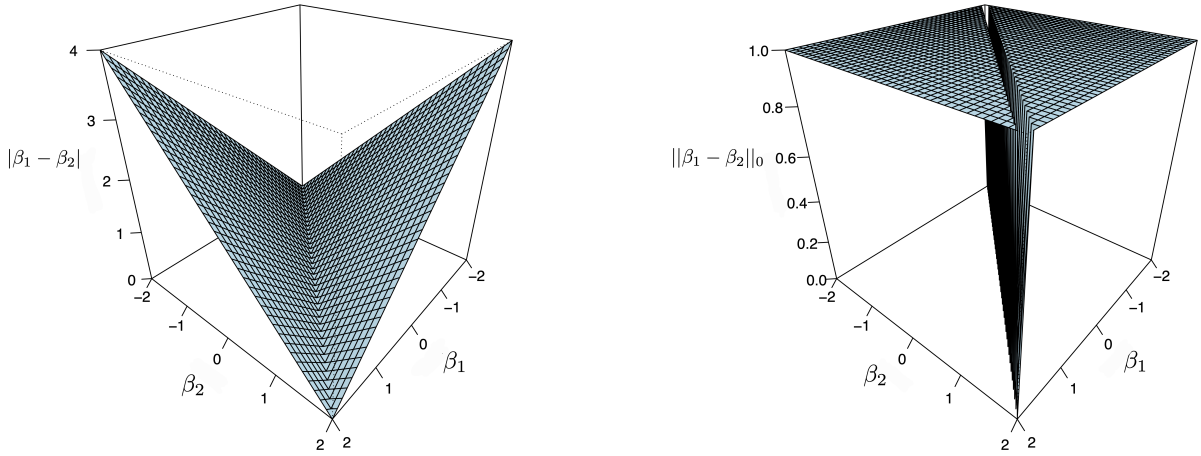

Figure 1.4: *$|\beta_1 - \beta_2|$, corresponding to the CAS-L_1 penalty function (left) and $||\beta_1 - \beta_2||_0$, corresponding to the CAS-L_0 penalty function (right) for $J = 1$, $\lambda = 1$, and weight $w_1 = 1$. It is noted that the L_0 norm is not continuous, the figure is only for visualization purposes.*

1.3.5 Other Penalties on Differences

Probably one of the most popular penalties using differences of the coefficients is the *fused lasso* introduced by Tibshirani et al. (2005). The reason why the fused lasso technique is not further analyzed here is that it does not account for groups, that is, the penalty constraint for the linear model is for tuning parameters $\lambda_1 \in \mathbb{R}^{\geq 0}$ and $\lambda_2 \in \mathbb{R}^{\geq 0}$ given by

$$\sum_{j=1}^{p} |\beta_j| \leq \lambda_1 \text{ and } \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| \leq \lambda_2.$$

By setting a constraint on *all* consecutive differences of the coefficients, it does not take into account the factor structure, making it not suitable for the setup of this thesis including factors. In contrast, CAS-L_1 and CAS-L_0 only build the differences of the coefficients that belong to the same factor. Further, the first constraint that is responsible for the selection tasks similarly ignores a groupwise structure.

The so-called *variable fusion* method introduced prior to the fused lasso by Land and Friedman (1996) takes the second constraint from the two above, that is $\sum_{j=2}^{p} |\beta_j - \beta_{j-1}| \leq \lambda$ for a tuning parameter $\lambda \in \mathbb{R}^{\geq 0}$. In this way, it similarly does not take into account the factor structure so it will not be further considered in this thesis.

Another penalty function applied on the coefficients' differences is the so-called SCOPE proposed by Stokell et al. (2021), which is given by the following penalty function

$$\sum_{j=1}^{J} \sum_{r=1}^{p_j} \rho_j(\beta_{j,(r)} - \beta_{j,(r-1)}).$$

Here, $\beta_{j,(1)} \leq \ldots \leq \beta_{j,(p_j)}$ are the order statistics of $\boldsymbol{\beta}_j$ and $\rho_j(\cdot)$ are concave, non-decreasing penalty functions. In a linear model with *hierarchical* categorical variables, the work of Zhao and Yang (2024) considers a combination of CAS-L_1 and SCOPE, however, this methodology is not

comparable to those presented in this thesis so far, as Zhao and Yang (2024) take into account the underlying hierarchical structure. The structure of the SCOPE penalty function mainly differs from the penalties on differences considered above, i.e. CAS-L_0 and CAS-L_1, in the aspect that the ordering in SCOPE is defined by the size of the coefficients and not by the ordering of the levels as for ordinal factors in CAS-L_1 and CAS-L_0 as argued in Gertheiss and Tutz (2023). To obtain the resulting SCOPE estimates, this ordering needs to be executed at each step of the optimization process, which can be computationally very demanding in high-dimensional settings, i.e. if p is large. In the framework of high-dimensional categorical data analysis in the linear model with multiple responses, the SCOPE is applied in Yang (2023), however, it is commented in the summary of the paper that for large sample sizes and dimensions, the computational cost of the method can not be ensured. Consequently, the SCOPE methodology will no longer be considered in the further course of this thesis.

1.4 Penalties for Grouped Structures

As explained in Section 1.2, in the presence of categorical covariates, one aims to perform *factor* selection rather than variable selection, i.e. either include or exclude a whole factor. For this purpose, a selective overview of penalty functions designed for factor selection is given, including group lasso (Section 1.4.1) as well as group SCAD (Section 1.5.1) and group MCP (Section 1.5.2), where the latter two are included in the next Section 1.5 about non-convex penalties, since they emerge from SCAD and MCP, respectively, being non-convex penalty functions.

1.4.1 Group Lasso

To achieve the goal of factor selection Yuan and Lin (2006) proposed the group lasso for the linear model. Kim et al. (2006) and Meier et al. (2008) investigated the group lasso for logistic regression, where Kim et al. (2006) named their approach "Blockwise Sparse Regression", given by the minimizer of the objective function (1.26) with the following choice of penalty function

$$P_\lambda^{(\mathrm{GL})}(\boldsymbol{\beta}) := \lambda \sum_{j=1}^{J} ||\boldsymbol{\beta}_j||_{\boldsymbol{K}_j}. \tag{1.53}$$

Here, $\boldsymbol{K}_j \in \mathbb{R}^{p_j \times p_j}$, $j = 1, ..., J$ are positive definite matrices and $||\boldsymbol{\beta}_j||_{\boldsymbol{K}_j} := (\boldsymbol{\beta}_j^T \boldsymbol{K}_j \boldsymbol{\beta}_j)^{1/2}$. In the following, the choice $\boldsymbol{K}_j = p_j \cdot \boldsymbol{I}_{p_j \times p_j}$ is made, following Yuan and Lin (2006) as well as Meier et al. (2008). This is a convenient choice resulting in weights which account for the fact that the factors may have a different number of levels, that is, the penalty function simplifies to

$$P_\lambda^{(\mathrm{GL})}(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{J} \sqrt{p_j}\, ||\boldsymbol{\beta}_j||_2. \tag{1.54}$$

In Figure 1.5, the group lasso penalty function for a simple example is displayed. The characteristics of the group lasso can be clearly seen in this figure, i.e. the penalty function decreases as the absolute value of the components β_1 and β_2 decrease, until the penalty function equals zero if $\beta_1 = \beta_2 = 0$. In practice, this means that the coefficients are shrunken towards zero and either *both* coefficients are set to zero, i.e. $\beta_1 = \beta_2 = 0$, or both are nonzero.

Theoretical Properties Group Lasso

For the *linear model*, asymptotic properties of the group lasso were studied in Nardi and Rinaldo (2008), as well as Wei and Huang (2010). With the penalty function introduced in (1.54), the

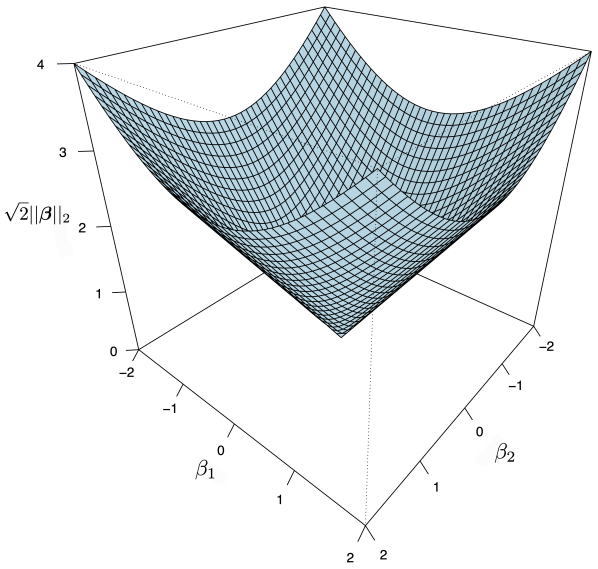

Figure 1.5: *Group lasso penalty function $P_\lambda(\boldsymbol{\beta})$ for $J = 1$, $p_1 = 2$ and $\lambda = 1$. In particular, the value of $\sqrt{2}||\boldsymbol{\beta}||_2$ is displayed on the vertical axis for $\boldsymbol{\beta} = (\beta_1, \beta_2)$ and $\beta_1, \beta_2 \in [-2, 2]$.*

(non-adaptive) group lasso objective function is given by

$$M_{pen}^{(\mathrm{GL})}(\boldsymbol{\beta}) := -L_n(\boldsymbol{\beta}) + \lambda \sum_{i=1}^{n} \sqrt{p_j} ||\boldsymbol{\beta}_j||_2. \tag{1.55}$$

However, the more general form

$$M_{pen}^{(\mathrm{GL})}(\boldsymbol{\beta}) = -L_n(\boldsymbol{\beta}) + \lambda \sum_{i=1}^{n} w_j ||\boldsymbol{\beta}_j||_2 \tag{1.56}$$

may also be used with some optional weights w_j, e.g. adaptive weights, where it will always be clarified what kind of weights are imposed.

A result on the existence concerning the minimum of the group lasso objective function is provided below. For this, it is required that $\sum_{i=1}^{n} y_i \in (0, n)$, which means that neither all observations y_i are equal to one nor equal to zero. Hence, at least one observation is equal to zero and at least one observation is equal to one, respectively.

Theorem 1.4.1 (Meier et al. (2008), Lemma 1, *logistic regression*)**.** It is supposed that for the given observed sample $\boldsymbol{y} = (y_1, \ldots, y_n)$ of the random response variable Y it holds that $\sum_{i=1}^{n} y_i \in (0, n)$ and that $\lambda > 0$. Then, the minimum of the group lasso optimization problem is attained, i.e. the set

$$S^{\mathrm{GL}} := \left\{ \hat{\boldsymbol{\beta}} \in \mathbb{R}^{p+1} \,\middle|\, \hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\arg\min}\, M_{pen}^{(\mathrm{GL})}(\boldsymbol{\beta}) \right\}, \tag{1.57}$$

where $M_{pen}^{(\mathrm{GL})}(\boldsymbol{\beta})$ is given by (1.55), is non-empty.

Proof. Meier et al. (2008), Appendix A.1. □

Theorem 1.4.1 ensures that at least one minimizer of the group lasso objective function exists. The next natural question is about conditions to ensure that the minimizer is unique, which is answered in the next corollary. The statement of this corollary is taken from Meier et al. (2008), however, in this reference it is not stated as a corollary and not rigorously proven whereas a proof is addressed here.

Corollary 1.4.2 (Meier et al. (2008), *logistic regression*)**.** It is supposed that the conditions of Theorem 1.4.1 hold. Then, it holds that the set of minimizers S^{GL} given by (1.57) is a convex set, where all elements in this set yield the same value of the objective function $M_{pen}(\boldsymbol{\beta})$. However, if the design matrix $\boldsymbol{X}$ is of full rank, the minimizer is unique, hence $|S^{\mathrm{GL}}| = 1$.

Proof. It is recalled that the canonical link function is utilized for logistic regression. Now, if the design matrix $\boldsymbol{X}$ is of full rank, it is known by Remark 1.1.10 that the log-likelihood function $L_n(\boldsymbol{\beta})$ is strictly concave, hence $-L_n(\boldsymbol{\beta})$ is *strictly* convex. Consequently, the objective function $M_{pen}^{(\mathrm{GL})}(\boldsymbol{\beta})$ (1.55) is strictly convex. Moreover, Theorem 1.4.1 yields that a minimizer exists, so it must be unique. In the other case, where one *cannot* ensure that $\boldsymbol{X}$ is of full rank, Remark 1.1.10 provides that the log-likelihood function is concave (*strict* concavity cannot be ensured in this case), yielding convexity for the negative log-likelihood function. Consequently, the resulting objective function $M_{pen}^{\mathrm{GL}}(\boldsymbol{\beta})$ is convex, so the set of minimizers S^{GL} is a convex set. □

In what follows, theorems on asymptotic properties of the group lasso in the considered setting are provided, which are, as some of the previous results for CAS-L_0, deduced from theorems of Chapter 2 and referenced here to avoid unnecessary redundancies. More specifically, the following two theorems are directly received by choosing $\lambda_0^n = 0 \ \forall\, n \in \mathbb{N}$ in Theorem 2.3.2 and Theorem 2.3.5, which are discussed later.

Theorem 1.4.3 ($\sqrt{n}$ consistency of group lasso in *logistic regression*, fixed p)**.** It is supposed that the regularity conditions (Reg1)-(Reg3) from Appendix B.1 hold. One sets $a_n^1 := \max\{\lambda_n w_j \,;\, j \in \{1, \dots J\}\}$ and assumes $a_n^1/\sqrt{n} = o_p(1)$. Then, for $\hat{\boldsymbol{\beta}}^{(\mathrm{GL})}$ being a global minimizer of (1.56), it holds that

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

Proof. Works completely analogous to the proof of Theorem 2.3.2, adjusting for the fact that, at the end of the proof, it is used that (1.56) is convex, thus $\hat{\boldsymbol{\beta}}^{(\mathrm{GL})}$ is a *global* minimizer. □

The next step is to provide an asymptotic normality result for group lasso in logistic regression.

Theorem 1.4.4 (Asymptotic normality of group lasso in *logistic regression*, fixed p)**.** One assumes that (Reg1)-(Reg3) of Appendix B.1 hold and the true underlying structure is sparse (Definition 1.2.4 (i)). For the group lasso with objective function (1.56), the adaptive weights $w_j := ||\hat{\boldsymbol{\beta}}_j^{(\mathrm{ML})}||_2^{-\gamma}$ are employed for some arbitrarily chosen $\gamma > 0$ where $\hat{\boldsymbol{\beta}}^{(\mathrm{ML})}$ is the unpenalized MLE. Furthermore, it is supposed that $\lambda_n \cdot n^{-1/2} \to 0$ and $\lambda_n \cdot n^{(\gamma-1)/2} \to \infty$. Then, it holds for a global minimizer $\hat{\boldsymbol{\beta}}^{(\mathrm{GL})}$ of (1.56) that

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

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

Proof. As mentioned earlier in this thesis, the question may arise how to identify $\hat{\boldsymbol{\beta}}_{A^*}^{(\mathrm{GL})}$ and $\boldsymbol{\beta}_{A^*}^*$, since A^* is unknown. For that reason, one refers to the previous Definitions 1.2.12, 1.2.13

as well as Remark 1.2.14 answering this question. The same applies for all upcoming theorems on asymptotic normality presented in this thesis.
The proof of this theorem works completely analogous to the proof of Theorem 2.3.5, adjusting for the fact that, at the end of the proof, it is used that (1.56) is convex, thus $\hat{\boldsymbol{\beta}}^{(\mathrm{GL})}$ is a *global* minimizer. □

Theorem 1.4.5 (Factor selection consistency of group lasso in *logistic regression*, fixed p)**.** One assumes that (Reg1)-(Reg3) of Appendix B hold and the true underlying structure is sparse (Definition 1.2.4 (i)). Further, one assumes that for $a_n^1 := \{\lambda_n w_j \,;\, j \in \{1, \dots, J\}$ it holds that $a_n^1/\sqrt{n} = o_p(1)$. Then, the $\sqrt{n}$ consistent estimator of Theorem 1.4.3 satisfies

$$\forall\, \epsilon > 0 \;\; \exists\, N \in \mathbb{N} \,:\; \mathbb{P}(A^* \not\subseteq A_n^{\mathrm{GL}}) < \epsilon \;\forall n \geq N.$$

Proof. Direct consequence of Theorem 2.3.8. □

By the popularity of lasso, group lasso is considered in several works deriving asymptotic results under different assumptions, which are discussed in comparison to this thesis in the following remark.

Remark 1.4.6 (Comparison to other results in the literature)**.** First, the work of Meier et al. (2008) considers a different kind of consistency than this thesis, making the results not comparable. Wang and Leng (2008) show similar results than those provided here, however, they consider the linear model rather than logistic regression. Similar arguments apply for Nardi and Rinaldo (2008), where both the fixed and the diverging case is considered for the linear model. The contribution of Wang and Tian (2019) show more general results considering GLMs for p_n being allowed to diverge. These results are stronger compared to those provided here, nevertheless, since the case of fixed p is considered in this chapter, it is desirable to avoid more complex regularity conditions that need to be imposed in the diverging case (cf. Appendix B). Zhang and Xiang (2015) consider a gaussian response variable, while their proving arguments are not extendable to the setting considered here, thus the results are not comparable. Finally, Wang et al. (2015) consider the diverging case in a general GLM framework under the so-called *sparse Riesz condition* (SRC). In particular, the SRC requires that the eigenvalues of the covariance matrices of all subsets of variables are bounded from above and from below according to Zhang and Huang (2008). However, in this thesis, the SRC is not imposed, where reference is made to Appendix B.5 for further details.
A selection of references treating group lasso in other frameworks different from the one considered here is given by e.g. Dahinden et al. (2006) and Nardi and Rinaldo (2012) in the context of log-linear models, while group lasso is considered from a non-parametric perspective in Bach (2008) and Ravikumar et al. (2009).

1.4.2 Computation of Group Lasso

A block coordinate descent (BCD) approach is applied to obtain a solution for the group lasso optimization problem, which is provided by Breheny and Huang (2015).

This algorithm, for linear *and* logistic regression, is similar to Yuan and Lin (2006), where group lasso in the *linear* model is considered in the latter reference. However, as Breheny and Huang (2015) pick up in their Section 2.4, applying the BCD algorithm to group lasso for *logistic* regression, the difficulty is that, other than in a linear model setup, there are no block-wise closed form solutions available. The authors solve this issue by employing a quadratic approximation of the log-likelihood function, which gives a loss function in the form of a weighted least squares

problem, such that a closed form solution can be obtained similar to the linear model case. The concrete form of this quadratic approximation is provided later in (1.63). In the work of Meier et al. (2008), a BCD is applied for group lasso (implemented in the `R` package `grplasso`), but since the comparison of `grplasso` and `grpreg` supplied in Breheny and Huang (2015) shows that `grpreg` tends to be faster, this thesis will follow the algorithm and `R` package of Breheny and Huang (2015). Different from that, the approaches proposed in these two referenced papers are similar.

The BCD algorithm for group lasso works analogously for the two other groupwise penalties considered in this chapter, which are not introduced yet. However, as announced at the beginning of this section, they will be presented in Section 1.5 such that, to avoid unnecessary repetitions, the application of BCD for group lasso is given in Section 1.5.3.

1.4.3 Coefficient Paths of Group Lasso

In the following, coefficient paths for lasso (1.28) and the previously discussed group variant of lasso, the group lasso, are investigated. To do so, the example provided in Section 1.3.4 is used, that is, $J = 2$ ordinal factors are considered with $p_1 = p_2 = 3$ and true coefficient vector chosen to be given by $\boldsymbol{\beta}^* = (1.5, 0.7, 1.4, 1.2, -0.3, -0.5, -1)$. Examining Figure 1.6, one can see that lasso (left) shrinks the coefficients towards zero and sets them to zero if they are close enough to zero. However, lasso completely ignores the underlying groupwise structure. The group lasso (right) takes into account the groupwise structure, that is, either the whole factor is removed from the model or all levels are kept. Further, as done in lasso, the coefficients are shrunken towards zero by group lasso. Additionally, one can observe that neither lasso nor group lasso perform levels fusion (in contrast to CAS-L_1 and CAS-L_0, one consults Figure 1.3), which one also expects by their construction. To conclude, the group lasso penalty enforces shrinkage of coefficients' values, as well as factor selection, whereas lasso is not performing factor selection.

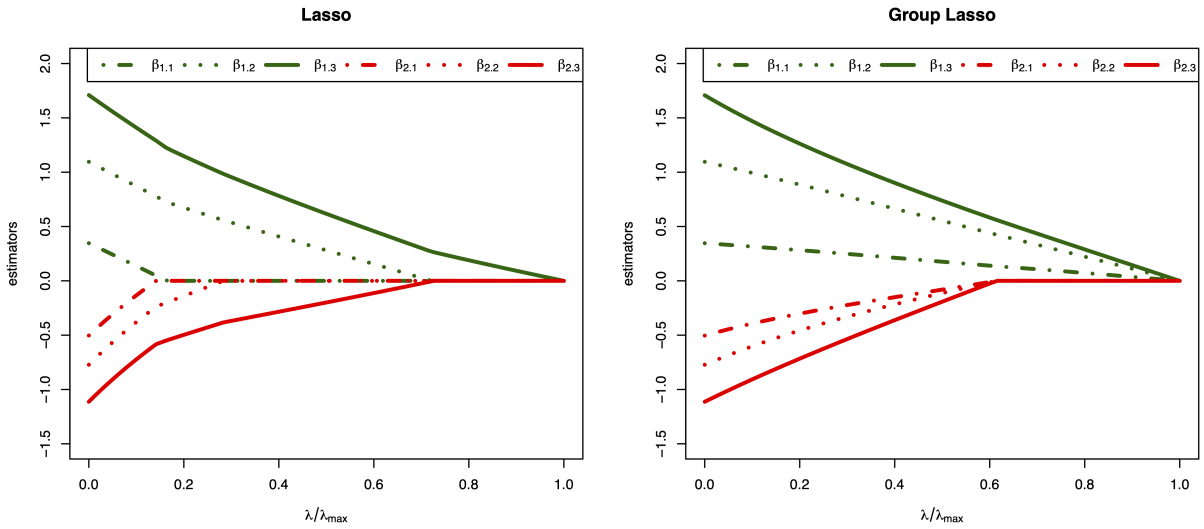

Figure 1.6: *Coefficient paths for lasso (left) and the groupwise variant group lasso (right). On the horizontal axis, the tuning parameter increases from left to right, where the coefficient estimates on the leftmost side in each plot are the unpenalized MLE. In this example, $J = 2$ ordinal factors were chosen with $p_1 = p_2 = 3$ and $\boldsymbol{\beta}^* = (1.5, 0.7, 1.4, 1.2, -0.3, -0.5, -1)$.*

1.5 Non-Convex Penalties

The domain of non-convex penalization is examined in this section, including the group SCAD and group MCP penalty where also the non-groupwise approaches SCAD and MCP are discussed, which build the basis for the corresponding groupwise versions. In the literature, these types of penalties are also called *folded concave penalties*, which is due to the fact that, even though they are neither convex nor concave on $\mathbb{R}$, they are concave on $\mathbb{R}^+$ and $\mathbb{R}^-$ and axially symmetric around zero. Starting with SCAD and group SCAD in Section 1.5.1 the MCP and group MCP are introduced in Section 1.5.2.

1.5.1 SCAD and Group SCAD

The abbreviation SCAD stands for "Smoothly Clipped Absolute Deviation Penalty", while group SCAD refers to its groupwise version. Before SCAD is introduced, it is noted that discussing simulation studies in Section 1.7, it is argued that the group SCAD is favored for the purposes of this thesis. Nevertheless, doe to the fact that group SCAD builds upon SCAD, it is of high importance to thoroughly introduce SCAD pointing out its purpose and characteristics. For that reason, the introduction of SCAD is more extensive compared to group SCAD as its major properties are directly transferrable from SCAD.

SCAD

The SCAD penalty function was proposed by Fan and Li (2001) targeting at fixing some (potential) drawbacks of the so-called Bridge penalty. To be more precise, the Bridge penalty is given by

$$P_\lambda^{(\text{Bridge})}(\boldsymbol{\beta}) := \lambda||\boldsymbol{\beta}||_q, \ q > 0,$$

and was introduced by Frank and Friedman (1993) (cf. Fu (1998) and Huang et al. (2008)). Depending on the choice of $q > 0$, the following consequences arise, which are described Fan and Li (2001). Choosing $q = 1$, $P_\lambda^{(\text{Bridge})}(\boldsymbol{\beta})$ leads to the lasso penalty $P_\lambda^{(\text{Lasso})}(\boldsymbol{\beta})$ (Section 1.2.3), which produces *biased estimates* since the penalization rate depends on the absolute value of the coefficients, for which one consults e.g. Table 4.1 in Fan et al. (2020) which summarizes some characteristics of different penalty functions. For $q > 1$, e.g. Ridge $P_\lambda^{(\text{Ridge})}(\boldsymbol{\beta})$ for $q = 2$ (Section 1.2.3), the resulting estimates are *not sparse*. Finally, for $q < 1$, Fan and Li (2001) argue that the solution is no longer *continuous in the data*, which may cause instability in the resulting estimate. Clearly, it depends on the application context whether these consequences are severe disadvantages or whether they are negligible, i.e. other advantages predominate. However, Fan and Li (2001) argue that a "good" penalty function should satisfy these three properties (unbiasedness, sparsity and continuity) and, as a consequence, they introduce SCAD.

The SCAD penalty function is given by

$$P_\lambda^{(\text{SCAD})}(\boldsymbol{\beta}) := \sum_{i=1}^{p} \rho_\lambda^{(\text{SCAD})}(|\beta_i|, \gamma), \tag{1.58}$$

where

$$\rho_\lambda^{(\text{SCAD})}(|\beta_i|, \gamma) := \begin{cases} \lambda|\beta_i| & \text{if } |\beta_i| \leq \lambda, \\ -\dfrac{(|\beta_i|^2 - 2\gamma\lambda|\beta_i| + \lambda^2)}{2(\gamma - 1)} & \text{if } \lambda < |\beta_i| < \gamma\lambda, \\ \dfrac{(\gamma + 1)\lambda^2}{2} & \text{if } |\beta_i| > \gamma\lambda, \end{cases} \tag{1.59}$$

for a parameter $\gamma > 2$ that needs to be chosen in practice. Even though $P_\lambda^{(\text{SCAD})}(\boldsymbol{\beta})$ depends on the parameter γ, a lower index γ is not added in $P_\lambda^{(\text{SCAD})}(\boldsymbol{\beta})$ to be consistent with the previously introduced penalty functions and the literature, e.g. Fan and Li (2001). The parameter γ is arbitrary but fixed, more details on its choice are given below.

To demonstrate the role of γ, for $\beta \in [-5, 5]$ the penalty function $\rho_1^{(\text{SCAD})}(|\beta|, \gamma)$ for different values of $\gamma > 2$ and fixed $\lambda = 1$ is visualized in Figure 1.7. It can be seen that for $|\beta| \leq \lambda$, the SCAD penalty function (1.59) equals the lasso penalty function applying the same amount of shrinkage. For $\lambda < |\beta| < \gamma\lambda$ the SCAD penalty function is quadratic while for $|\beta| > \gamma\lambda$ the penalization remains constant. Hence, for $|\beta| > \gamma\lambda$, the penalization rate does not depend on the size of the absolute value $|\beta|$, which is different from lasso. Thus, it is expected that SCAD removes the issue of biasedness for large coefficients in absolute value that one observes for the lasso penalty.

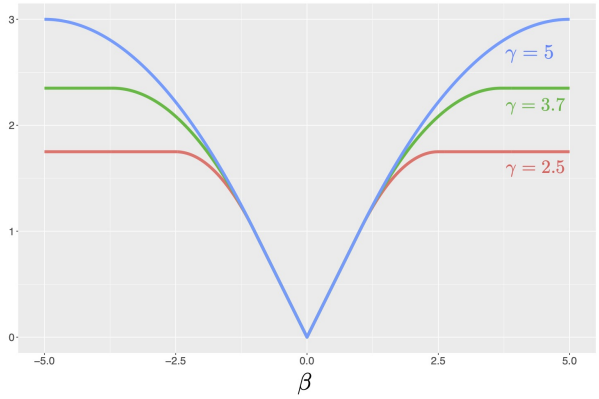

Figure 1.7: $\rho_1^{(SCAD)}(|\beta|, \gamma)$ *for different values of* $\gamma \in \{2.5, 3.7, 5\}$, $\lambda = 1$ *and* $\beta \in [-5, 5]$. *The value of* γ *controls how fast the penalty drops off.*

As one can further see in Figure 1.7, the parameter γ determines the concavity of the function, hence how fast the penalty function drops off until it remains constant. For larger values of γ, this drop off is more moderate than for smaller values. In practice, Fan and Li (2001) suggest to use $\gamma = 3.7$, which arises from Bayesian arguments corresponding to the minimization of the posterior risk function, where they further argue that the convenience of this choice is underlined by their simulation studies.

Remark 1.5.1 (Impact of γ on the resulting estimates)**.** It was demonstrated above that γ controls the concavity of the SCAD penalty function. That is, the choice of γ has an impact on the bias of the resulting estimates, meaning that lower values of γ give less biased estimates. However, the greater γ the "more concave" the penalty function, and consequently the greater the chance for multiple local minimizers such that the stability may decrease, hence the variability in the estimates increase, which results in a bias-variance trade-off.

The SCAD approach is an alternative to *adaptive* lasso which uses adaptive weights to account for the fact of biasedness, where the adaptive lasso depends on the quality of the MLE, which may be unstable especially in high-dimensional settings. Nevertheless, a drawback of SCAD compared to adaptive lasso is the fact that the resulting optimization problem is *non-convex* which makes it more involving from the computational point of view.

The other two potential issues of the Bridge penalty mentioned at the beginning of this section, i.e. the missing continuity of the solution for $q < 1$, as well as the missing sparsity for $q > 1$, are further addressed by SCAD. The sparsity property emerges from the sparsity property of lasso, since for $|\beta| \leq \lambda$, the penalization of SCAD equals those of lasso. Lastly, continuity of the SCAD penalty function is obviously given.

Discussing theoretical properties for SCAD, Theorem 2 of Fan and Li (2001) provides oracle properties (i.e. factor selection consistency and asymptotic normality, cf. Definition 1.2.13) for SCAD in the *logistic* regression setting under some regularity conditions for the case of fixed p. Oracle properties for the case where p_n is allowed to grow with n are investigated by Fan and Peng (2004), Fan and Lv (2011) and Kim et al. (2008), where the first two sources also include logistic regression. For one-step estimates using a local linear approximation (LLA) of SCAD, oracle properties are obtained by Zou and Li (2008). For varying coefficient models (models where the parameter vector $\boldsymbol{\beta}$ may depend on e.g. a time component), SCAD is analyzed in Wang et al. (2008) and for partially linear models by Xie and Huang (2009). However, since using the group variant *group* SCAD is favored instead of SCAD in this thesis, being more suitable for factor structures, no further details are provided on theoretical properties of SCAD.

Remark 1.5.2 (How to obtain a (potential) oracle solution?)**.** An important fact that needs to be mentioned working with such non-convex penalties as SCAD is, that in general, the properties (e.g. oracle properties) typically hold for a *local* minimizer. As a consequence, the question arises whether an algorithm, applied to solve the resulting minimization problem, is able to find *this particular* local minimizer, for which it is known that it satisfies the oracle properties. In the works of Zhang (2010), Fan and Lv (2011) and Zhang and Zhang (2012), conditions are found under which the optimization problem has a unique solution, however, as mentioned in Fan et al. (2014), these conditions may be too strong in practice. In their paper, they propose an algorithm for which it is shown that the probability that the local minimizer found by this algorithm equals the oracle estimator can be bounded from below. Nevertheless, the question whether one can find *the* local minimizer that satisfies certain properties in general is very challenging, and, following the focus of this thesis, this issue will not be discussed further, referring to the works of Fan et al. (2014), Loh and Wainwright (2013) as well as Loh and Wainwright (2017).

Group SCAD

The group SCAD was proposed by Wang et al. (2007) for time-varying coefficient models, i.e. models where the parameter $\boldsymbol{\beta}$ is allowed to depend on a time component. Huang et al. (2012) discussed the group SCAD for linear regression while in Breheny and Huang (2015), group SCAD is further considered for logistic regression, where the focus lies more on computational aspects rather than on asymptotic properties. The penalty function of group SCAD, denoted by $P_\lambda^{(\text{gSCAD})}(\boldsymbol{\beta})$, applies $\rho_\lambda^{(\text{SCAD})}(\cdot)$ for all factors $j \in \{1, \ldots, J\}$ to the L_2 norm of the sub-vector corresponding to the j-th factor. Thus, it is given by

$$P_\lambda^{(\text{gSCAD})}(\boldsymbol{\beta}) := \sum_{j=1}^{J} \rho_\lambda^{(\text{SCAD})}(||\boldsymbol{\beta}_j||_2, \gamma), \tag{1.60}$$

where $\rho_\lambda^{(\text{SCAD})}(\cdot, \gamma)$ is given by (1.59). Similar to SCAD, the parameter $\gamma > 2$ indicates how fast the penalization rate drops off. As further noted introducing SCAD, Fan and Li (2001) propose to choose $\gamma = 3.7$, consequently this is used in the simulation studies of this thesis for group SCAD. For $J = 1$, $p_1 = 2$, $\gamma = 3.7$ and $\lambda = 1$, the resulting group SCAD penalty function is displayed in Figure 1.8. By the singularity at the origin, one can see that the group SCAD

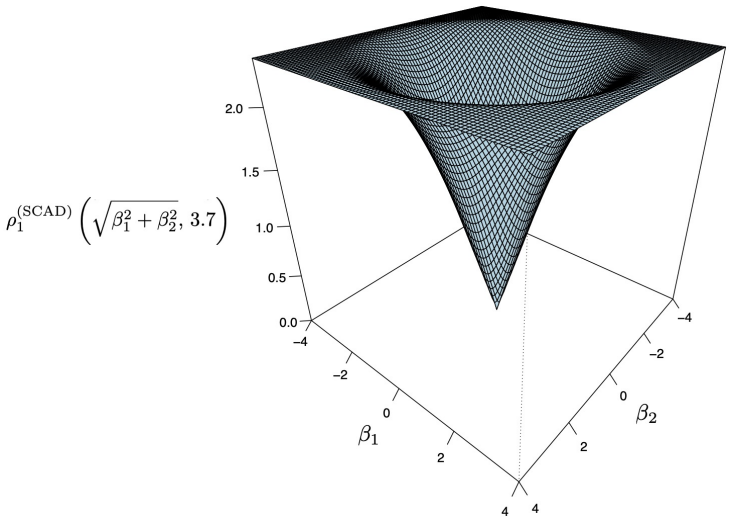

Figure 1.8: *Group SCAD for $\gamma = 3.7$, $\lambda = 1$ for one factor $J = 1$ with $p_1 = 2$, such that $P_1^{(gSCAD)}(\boldsymbol{\beta}) = \rho_1^{(SCAD)}(||\boldsymbol{\beta}_1||_2, 3.7) = \rho_1^{(SCAD)}(\sqrt{\beta_1^2+\beta_2^2}, 3.7)$ and $\boldsymbol{\beta}_1 = (\beta_1, \beta_2)$, $\boldsymbol{\beta} = (\beta_{int}, \boldsymbol{\beta}_1)$.*

penalty function results in a sparse model according to Fan and Li (2001), i.e. it performs factor selection. Additionally, this penalty function enforces shrinkage of the coefficients and for "large" values of $||\boldsymbol{\beta}_1||_2$, which means $||\boldsymbol{\beta}_1||_2 > \gamma\lambda = 3.7$, the penalization rate remains constant.

Group SCAD for logistic regression is not the most popular approach that is treated in the literature, especially from the theoretical point of view. As elaborated in the following, there are some references treating asymptotic properties for group SCAD in the *linear* model, but logistic regression is, to the best of one's knowledge, neglected from the *theoretical point of view*, i.e. asymptotic properties for group SCAD in the logistic regression model are missing so far. To be more precise, Fan and Li (2001) show oracle properties for SCAD, including linear and logistic regression. For the *linear model*, they similarly hold for group SCAD as shown by Wang et al. (2007) (fixed p case) and Guo et al. (2015) (diverging p_n case), respectively. However, these works do not treat other GLMs, or especially logistic regression. Asking whether (i) one can transfer the proofs for group SCAD in the *linear model* to group SCAD in the *logistic regression model*, or, (ii) one can transfer the proofs for *SCAD* in the logistic regression model to *group SCAD* in the logistic regression model, the following observations are made.
For (i), diving into the proof of Guo et al. (2015) (Theorem 1), the particular structure of the linear model is used, thus one cannot transfer this proof to group SCAD in logistic regression. For (ii), analyzing the requirements and the corresponding proofs of Theorem 1 and 2 provided by Fan and Li (2001), the authors require bounds on the derivatives of the SCAD penalty function needed for the Taylor approximation in the proof. However, for group SCAD, one receives an additional inner derivative from the L_2 norm, which complicates the Taylor approximation, especially from the second order term. To conclude, further work is required to close this gap and obtain theoretical properties for group SCAD in penalized logistic regression. Since, from Chapter 2 on, SCAD and group SCAD are not further considered, it is beyond the scope of this thesis to close this gap.

1.5.2 MCP and Group MCP

Having discussed the SCAD and group SCAD approaches, another penalty function that belongs to the category of folded concave penalty functions is investigated next, along with its groupwise variant. In particular, the MCP and group MCP are introduced, where the abbreviation MCP stands for "Minimax Concave Penalty". Generally speaking, the MCP penalty has several similar properties like SCAD, such that Section 1.5.1 will serve as reference to avoid unnecessary repetitions, however, the differences between these two approaches are pointed out. Further, by similar arguments as provided at the beginning of Section 1.5.1, it is important to introduce MCP in detail, even tough group MCP is favored for the purposes of this thesis.

MCP

The MCP penalty function was introduced in the work of Zhang (2010) with similar purposes as elaborated earlier for SCAD. For some parameter $\gamma > 0$ controlling the concavity, the MCP is defined as

$$P_\lambda^{(\mathrm{MCP})}(\boldsymbol{\beta}) := \sum_{i=1}^{p} \rho_\lambda^{(\mathrm{MCP})}(|\beta_i|, \gamma),$$

where

$$\rho_\lambda^{(\mathrm{MCP})}(|\beta_i|, \gamma) := \lambda \int_0^{|\beta_i|} \left(1 - \frac{x}{\gamma\lambda}\right)_+ \mathrm{d}x. \tag{1.61}$$

Here, the notation $(\cdot)_+$ refers to the positive part of the function, thus defining the function $f(x) := 1 - \frac{x}{\gamma\lambda}$, it holds that $(f(x))_+ := \max(f(x), 0)$.

In the following, the apostrophe denotes the derivative of a function. As the name of MCP indicates, ensuring an unbiasedness and selection property, which are introduced in (1.62) directly below, it minimizes the maximum concavity $\kappa(\rho_\lambda)$, which is defined as

$$\kappa(\rho_\lambda) := \sup_{0<\beta_1<\beta_2} \frac{\rho'_\lambda(\beta_1) - \rho'_\lambda(\beta_2)}{\beta_2 - \beta_1},$$

according to Zhang (2010). The unbiasedness and selection properties that are ensured are given by

$$\rho'_\lambda(\beta) = 0 \quad \forall \beta \geq \gamma\lambda, \qquad \lim_{\beta \to 0+} \rho'_\lambda(\beta) = \lambda, \tag{1.62}$$

where $\lim_{\beta\to 0+}$ denotes the right-hand limit. These conditions are taken from Zhang (2010) and can be similarly found in Fan et al. (2020) (Section 4.4.1). The unbiasedness property given on the left hand side of (1.62) ensures that the penalization rate remains constant starting from $\beta \geq \gamma\lambda$, which ensures that larger values of β are not penalized disproportionally more than smaller ones, thus it ensures unbiasedness. This conclusion is further explained in Fan and Li (2001) (Section 2). The selection property, or sparsity property, on the right hand side of (1.62), ensures that the penalty function is singular at the origin, which ensures that the penalty function can perform selection, further details on this conclusion can be found, again, in Fan and Li (2001) (Section 2). Referring to the work of Zhang (2010), among the penalty functions satisfying the unbiasedness and selection property (1.62), the MCP further minimizes the maximum concavity $\kappa(\rho_\lambda)$ which means that it is the "most convex" among these functions.

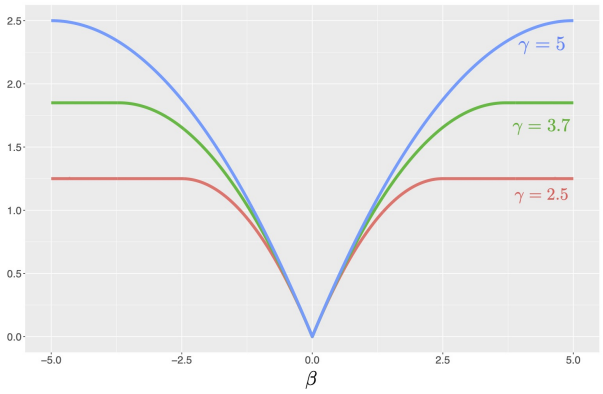

Figure 1.9: $\rho_1^{(MCP)}(|\beta|, \gamma)$ *for different values of* $\gamma \in \{2.5, 3.7, 5\}$, $\lambda = 1$ *for* $\beta \in [-5, 5]$.

For $\beta \in [-5, 5]$, the function $\rho_\lambda^{(\text{MCP})}(|\beta|, \gamma)$ is visualized in Figure 1.9. Being singular at the origin, the penalization rate of the MCP *directly* relaxes the penalization rate, until it remains constant. This is different from SCAD, which equals the lasso for $|\beta| \leq \lambda$ before the penalization rate drops off, which can be clearly seen in Figure 1.10, where both penalty functions are displayed with the same choices of parameters. However, for the role of γ, similar arguments as in SCAD apply, where it was discussed that the parameter γ controls the concavity of the function. Further, the impact of the choice of γ on the estimates, which was explained for SCAD in Remark 1.5.1, similarly apply for MCP. In practice, Zhang (2010) suggested to use $\gamma = 2.7$.

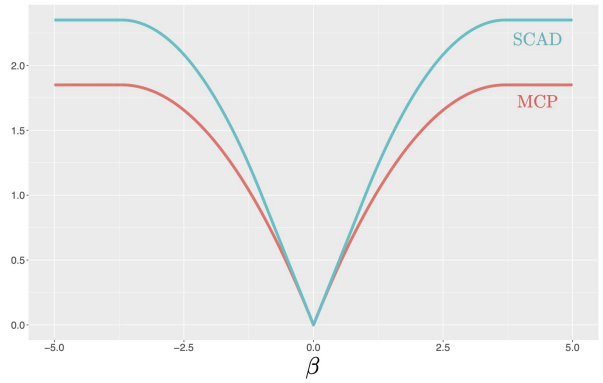

Figure 1.10: $\rho_1^{(SCAD)}(|\beta|, 3.7)$ *and* $\rho_1^{(MCP)}(|\beta|, 3.7)$, *both for* $\gamma = 3.7$, $\lambda = 1$ *and* $\beta \in [-5, 5]$.

Turning the focus to theoretical properties, Zhang (2010) prove oracle properties for MCP in penalized *linear* regression, however, logistic regression is not treated. Oracle properties for penalized likelihood models, incorporating logistic regression, in a more general framework of folded concave penalty functions, including SCAD and MCP, are provided in Fan and Lv (2011).

Based on the characteristics of MCP pointed out above, the extension to group MCP follows naturally and is provided in the following.

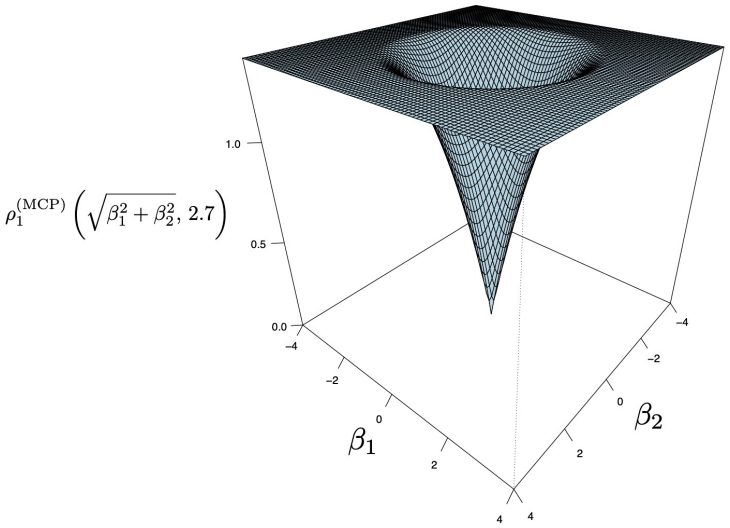

Figure 1.11: *Group MCP for $\gamma = 2.7$, $\lambda = 1$ for one factor $J = 1$ with $p_1 = 2$, such that $P_1^{(gMCP)}(\boldsymbol{\beta}) = \rho_1^{(MCP)}(||\boldsymbol{\beta}_1||_2, 2.7) = \rho_1^{(MCP)}(\sqrt{\beta_1^2+\beta_2^2}, 2.7)$ and $\boldsymbol{\beta}_1 = (\beta_1, \beta_2)$, $\boldsymbol{\beta} = (\beta_{int}, \boldsymbol{\beta}_1)$.*

Group MCP

The group MCP, which is considered in Huang et al. (2012) for the linear model, is, similar to group SCAD (1.60), given by the application of $\rho_\lambda^{(\text{MCP})}$ to all L_2 norms of the sub-vectors corresponding to factor j, where $j \in \{1, \ldots, J\}$. In particular

$$P_\lambda^{(\text{gMCP})}(\boldsymbol{\beta}) := \sum_{j=1}^{J} \rho_\lambda^{(\text{MCP})}(||\boldsymbol{\beta}_j||_2, \gamma),$$

where $\rho_\lambda^{(\text{MCP})}(||\boldsymbol{\beta}_j||_2, \gamma)$ is given by (1.61). In the same way as done for MCP in Figure 1.9, the resulting group MCP penalty function $P_\lambda^{(\text{gMCP})}(\boldsymbol{\beta})$ for $\gamma = 2.7$ and $\lambda = 1$ applied to $\boldsymbol{\beta} = (\beta_{int}, \boldsymbol{\beta}_1)$ with $\boldsymbol{\beta}_1 = (\beta_1, \beta_2) \in [-2, 2] \times [-2, 2]$ is displayed in Figure 1.11. It is recalled that the intercept is not penalized. The impact of γ in group MCP is similar to MCP and the optimization problem resulting from group MCP is non-convex. In Huang et al. (2012), the choice $\gamma = 2.7$ is suggested.

In Huang et al. (2012), the group MCP is considered for the *linear* model, where theoretical properties such as oracle properties are shown. They refer to group SCAD as "2-norm group MCP" which obviously comes from the fact that in the penalty function given above, the MCP (1.61) is applied to the L_2 norm of every sub-vector $\boldsymbol{\beta}_j$ for $j = 1, \ldots, J$. The corresponding proof can *not* be transferred to logistic regression without substantial adjustments, since it is used that the errors in the considered linear model follow a normal distribution, which is used to derive distributions of other quantities. In Breheny and Huang (2015), the group MCP approach is considered for logistic regression, where the focus lies on algorithms, in particular group descent algorithms, where the investigation of theoretical properties is neglected. Thus, analogously to group SCAD, further work is required to investigate theoretical properties for group MCP in penalized logistic regression.

1.5.3 Computation of Non-Convex Penalties

Breheny and Huang (2011) obtain coordinate descent algorithms for non-convex penalties including SCAD and MCP which was extended to the *groupwise* versions group SCAD and group

MCP (and group lasso) in Breheny and Huang (2015). Introducing first the coordinate descent (CD) approach for SCAD and MCP, the block coordinate descent (BCD) approach for group SCAD and group MCP (and group lasso) is subsequently discussed. The general idea of these two algorithms is provided in Appendix A.2.

Coordinate Descent for SCAD and MCP

In the logistic regression framework, Breheny and Huang (2011) start with a local quadratic approximation (LQA) of the loss function, which is the negative log-likelihood function $L_n(\boldsymbol{\beta})$. This local quadratic approximation (LQA) is done to obtain closed form solutions for MCP and SCAD in a single dimension being available in linear regression according to Breheny and Huang (2011) (Sections 2.1 and 2.2). Consequently, with the LQA, logistic regression is transformed to a weighted linear regression problem (Breheny and Huang (2011), Section 3). The LQA is updated in every iteration step k of the CD algorithm. In particular, according to Breheny and Huang (2011), the LQA at some given $\hat{\boldsymbol{\beta}}^{(k)}$ looks as follows

$$L_n(\boldsymbol{\beta}) \approx \frac{1}{2n}(\tilde{\boldsymbol{y}}^{(k)} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{W}^{(k)} (\tilde{\boldsymbol{y}}^{(k)} - \boldsymbol{X}\boldsymbol{\beta}), \tag{1.63}$$

where the working response $\tilde{\boldsymbol{y}}^{(k)}$ and the diagonal matrix with weights $\boldsymbol{W}^{(k)}$ are specified in Appendix A.2, spefically in (A.8) and (A.7), respectively. As one can see in (1.63), the LQA of the log-likelihood is divided by $2n$, the following remark comments on that.

Remark 1.5.3.

(i) In the literature, some authors multiply the log-likelihood (or the squared difference in the linear model, respectively) by the factor $\frac{1}{2n}$, e.g. Breheny and Huang (2011) and Guo et al. (2015). Since this multiplicative factor 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. One can also neglect this factor, in this case one has to be careful when comparing two solutions for different tuning parameters and different sample sizes. This factor is used employing a CD and BCD algorithm following Breheny and Huang (2011), as well as Breheny and Huang (2015), nevertheless it is *not* used in PIRLS which is specified in Appendix A.1.

(ii) In Breheny and Huang (2011), treating the particular case of logistic regression, they write that the diagonal entries of $\boldsymbol{W}^{(k)}$ are given by $\pi_i^{(k)}(1-\pi_i^{(k)})$, where $\boldsymbol{\pi}^{(k)} = (\pi_1^{(k)}, \ldots, \pi_n^{(k)})$ is evaluated at the current iteration $\hat{\boldsymbol{\beta}}^{(k)}$. It is noted that this is in line with Remark A.1.4 since $\boldsymbol{\mu}^{(k)} = \boldsymbol{\pi}^{(k)}$, whereby the mentioned remark specifies the matrix $\boldsymbol{W}^{(k)}$ for the case of logistic regression.

Having executed the LQA (1.63), Breheny and Huang (2011) propose the following algorithm, where it is noted that here, the notation $\boldsymbol{\beta} = (\beta_{int}, \beta_1, \beta_2, \ldots, \beta_p)$ instead of the factor-wise notation, i.e. $\boldsymbol{\beta} = (\beta_{int}, \beta_{1,1}, \ldots, \beta_{1,p_1}, \ldots, \beta_{J,1}, \ldots, \beta_{J,p_J}) = (\beta_{int}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J)$ is employed *once* being more convenient for CD approaches *not* proceeding factor-wise.

Algorithm 1.5.4 (CD for SCAD and MCP, Breheny and Huang (2011))**.**

(i) One sets the start value $\hat{\boldsymbol{\beta}}^{(0)}$. One sets $k = 1$.

(ii) One repeats the following steps until convergence is reached.

(ii.i) One updates the LQA of the loss function $L_n(\boldsymbol{\beta})$ according to (1.63).

(ii.ii) One updates the intercept $\hat{\beta}_{int}^{(k)}$.

(ii.iii) One cycles through all coordinates $i = 1, \ldots, p$ and updates coordinate i as follows

$$\hat{\beta}_i^{(k+1)} = \underset{\beta_i \in \mathbb{R}}{\arg\min}\, M_{pen}(\hat{\beta}_{int}^{(k+1)}, \hat{\beta}_1^{(k+1)}, \ldots, \hat{\beta}_{i-1}^{(k+1)}, \beta_i, \hat{\beta}_{i+1}^{(k)}, \ldots, \hat{\beta}_p^{(k)}).$$

(iii) Finally, one sets $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(k)}$.

For the update of the intercept term, Breheny and Huang (2011) propose to set $\lambda = 0$ in the considered framework using the LQA of the log-likelihood (i.e weighted least squares). To be more precise, in the most simple case of "classical" least squares, the weight matrix $\boldsymbol{W}^{(k)}$ is the identity matrix and the updating value of the intercept is simply the mean of the partial residual, i.e. $\hat{\beta}_{int}^{(k)} = \frac{1}{n}\boldsymbol{X}_{int}^T(\boldsymbol{y} - \boldsymbol{X}_{-int}\hat{\boldsymbol{\beta}}_{-int}^{(k-1)})$, cf. Breheny and Huang (2011). Here, $\boldsymbol{X}_{int}$ is the column of the design matrix belonging to the intercept, thus $\boldsymbol{X}_{int}^T = \mathbf{1}^T \in \mathbb{R}^{n\times 1}$ while the negative lower index indicates that the corresponding column/entry is removed. Updating the intercept in this way clearly cause no challenges if $n < p$. Considering the *weighted* least squares emerging from the LQA of the log-likelihood as considered here, a similar update incorporating the weight matrix $\boldsymbol{W}^{(k)}$ is used as derived in Breheny and Huang (2009) (equation (9)) and further obtained in Breheny and Huang (2015).

In the work of Breheny and Huang (2011), it is not specified which convergence criterion is applied for Algorithm 1.5.4. Further, they use Algorithm 1.5.4 to fit a whole solution path, that is, to obtain a minimizer for each value of the tuning parameter in some interval, hence $\lambda \in [\lambda_{min}, \lambda_{max}]$ with a given grid of values. They propose to start at $\hat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$ with tuning parameter λ_{max}, where λ_{max} is the smallest value for which all explanatory variables are excluded from the model. Then, proceeding towards λ_{min} (often chosen to be $\lambda_{min} = 0$, hence the unpenalized solution), they propose to take the estimate from the previous value of λ. The algorithm given above calculates the solution for one fixed value of λ, which is determined in practice by k-fold CV. Breheny and Huang (2011) implemented their approach in the `R` package `ncvreg`, which is used in the simulation studies of this thesis.

Remark 1.5.5 (Convergence of CD for SCAD and MCP)**.** As Breheny and Huang (2011) note, the algorithm described above is in general not guaranteed to converge, similar to the iteratively re-weighted least squares (IRLS) algorithm in the framework of GLMs, compare Hastie et al. (2015) (Section 4.4.1). However, Breheny and Huang (2011) argue that, ensuring that no model saturation occurs, they have not observed failures of convergence.

Block Coordinate Descent for Group SCAD, Group MCP and Group Lasso

The block coordinate descent (BCD) approach for group SCAD and group MCP, proposed by Breheny and Huang (2015) is considered. Furthermore, in the mentioned reference, the BCD approach is similarly applied to group lasso, such that the algorithm described in the following is applicable for group lasso, group SCAD and group MCP. For the *general idea* of BCD, one refers to Appendix A.2.

For the coefficient vector $\boldsymbol{\beta}$, one returns to the commonly used factor-wise notation, i.e. one writes $\boldsymbol{\beta} = (\beta_{int}, \beta_{1,1}, \ldots, \beta_{1,p_1}, \ldots, \beta_{J,1}, \ldots, \beta_{J,p_J}) = (\beta_{int}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J)$.

The BCD approach uses the same LQA of $L_n(\boldsymbol{\beta})$ as in (1.63). In Breheny and Huang (2015) closed form solutions are obtained which are useful for the factor-wise updates of step (ii.iii) in Algorithm 1.5.6, which is provided directly below.

Algorithm 1.5.6 (BCD for group SCAD, group MCP and group lasso, Breheny and Huang (2015))**.**

(i) One sets the start value $\hat{\boldsymbol{\beta}}^{(0)}$. One sets $k = 1$.

(ii) One repeats the following steps until convergence is reached.

(ii.i) One updates the approximation of $L_n(\boldsymbol{\beta})$.

(ii.ii) One updates the intercept $\hat{\beta}_{int}^{(k)}$.

(ii.iii) One cycles through all factors $j = 1, \ldots, J$ and updates for factor j as follows

$$\hat{\boldsymbol{\beta}}_j^{(k+1)} = \underset{\boldsymbol{\beta}_j \in \mathbb{R}^{p_j}}{\arg\min}\, M_{pen}(\hat{\beta}_{int}^{(k+1)}, \hat{\boldsymbol{\beta}}_1^{(k+1)}, \ldots, \hat{\boldsymbol{\beta}}_{j-1}^{(k+1)}, \boldsymbol{\beta}_j, \hat{\boldsymbol{\beta}}_{j+1}^{(k)}, \ldots, \hat{\boldsymbol{\beta}}_J^{(k)})$$

using the obtained closed form solutions.

(iii) Finally, one sets $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(k)}$.

For a discussion on the start values, one refers to the explanations given right below Algorithm 1.5.4, which similarly apply to Algorithm 1.5.6. The intercept term is updated analogously to Algorithm 1.5.4.

For group lasso, the resulting optimization problem is convex, which does not hold for group SCAD and group MCP since the penalty functions are non-convex. Thus, different results concerning convergence to global/local minima for group lasso on the one hand and group SCAD and group MCP on the other hand can be inferred. Breheny and Huang (2015) show in Proposition 2 the so called descent property, i.e. for $M_{pen}(\boldsymbol{\beta})$ being equal to $M_{pen}^{(\text{gMCP})}(\boldsymbol{\beta})$, $M_{pen}^{(\text{gSCAD})}(\boldsymbol{\beta})$, as well as $M_{pen}^{(\text{GL})}(\boldsymbol{\beta})$, they show that $M_{pen}(\boldsymbol{\beta}^{(k+1)}) \leq M_{pen}(\boldsymbol{\beta}^{(k)})$ holds for $\boldsymbol{\beta}^{(k)}$ being the estimate after iteration $k \in \mathbb{N}$. For group lasso this implies convergence to the global minimum because $M_{pen}^{(\text{GL})}(\boldsymbol{\beta})$ is convex. For group SCAD and group MCP, due to their non-convexity, only convergence to a local minimum is received.

Finally, references are provided that supply alternative possibilities to obtain estimates for the non-convex penalties SCAD, MCP and the corresponding groupwise versions. Fan and Li (2001) propose to use a local quadratic approximation of the penalty function and Newton-Raphson to obtain the estimate for SCAD, which is similar to the PIRLS algorithm (Appendix A.1). Wang et al. (2007) as well as Wang et al. (2008) applied this approach to group SCAD. Furthermore, Zou and Li (2008) propose a local linear approximation to obtain a lasso type optimization problem. Having that, they execute the least-angle regression (LARS, Efron et al. (2004)) algorithm. As elaborated in Breheny and Huang (2011), this approach requires the computation of a comparably huge amount of lasso paths for one approximation of SCAD and MCP path, making it computationally more involving. However, since closed form solutions are available as discussed above, the CD and BCD approaches are favored for SCAD, MCP and their groupwise variants, respectively.

1.5.4 Coefficient Paths of Non-Convex Penalties

Coefficient paths for SCAD and group SCAD, as well as for MCP and group MCP are visualized in Figures 1.12 and 1.13, respectively, exhibiting the characteristics of the respective method. To obtain a comparable scale, the same example as in Sections 1.3.4 and 1.4.3 is used.

Comparing the paths of SCAD on the left hand side of Figure 1.12 to the paths of lasso on the left hand side of Figure 1.6, one can see that for "smaller" values of the absolute value of the coefficients, the SCAD paths look similar to lasso, while from some point on, the penalization relaxes and is constant for "larger" values of the coefficients. Similar arguments apply for group SCAD, where it gets clear that group SCAD, compared to SCAD, includes or excludes a whole factor from the model, instead of single components, i.e. variables.

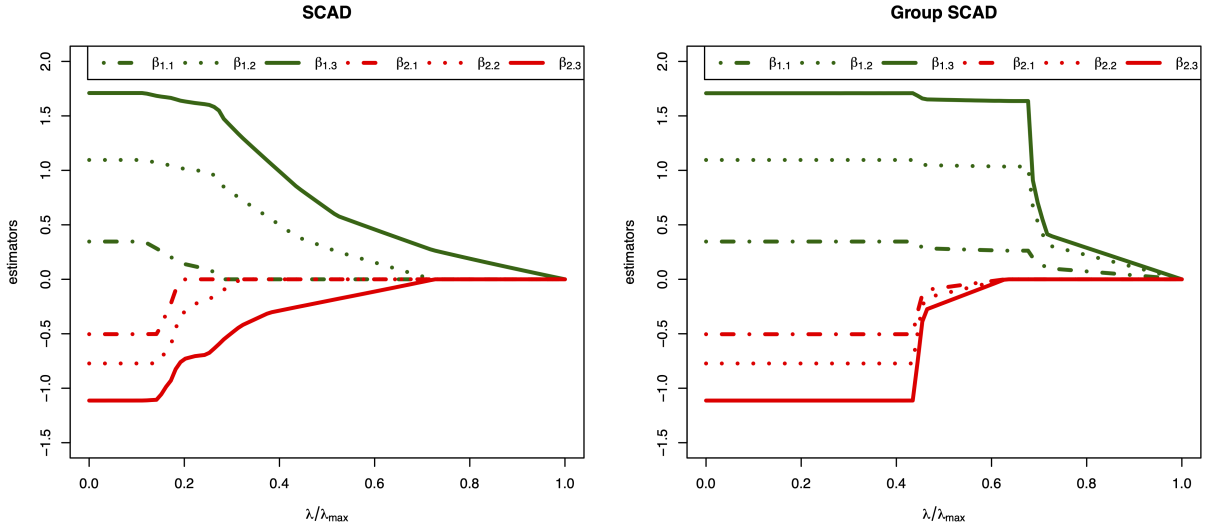

Figure 1.12: *Coefficient paths for SCAD (left) and the groupwise variant group SCAD (right). On the horizontal axis, the tuning parameter increases from left to right, where the coefficient estimates on the leftmost side in each plot are the MLE. In this example, $J = 2$ ordinal factors were chosen with $p_1 = p_2 = 3$ and $\boldsymbol{\beta}^* = (1.5, 0.7, 1.4, 1.2, -0.3, -0.5, -1)$.*

For MCP, displayed on the left hand side in Figure 1.13, compared to lasso, similar arguments apply, with the difference that MCP *directly* relaxes the penalization rate. Additionally, comparing group MCP (right hand side of Figure 1.13) to group SCAD (right hand side of Figure 1.12), one can clearly see that the paths, right before the respective factor is excluded from the model, are more steep for group MCP than for group SCAD. This is caused by the fact that (group) MCP directly relaxes the penalization rate of (group) lasso, while (group) SCAD applies the same penalization rate in this area (for $|\beta_i|$, $i \in \{1, \dots, p\}$ or $||\boldsymbol{\beta}_j||_2$, $j \in \{1, \dots, J\}$ being "small", respectively)

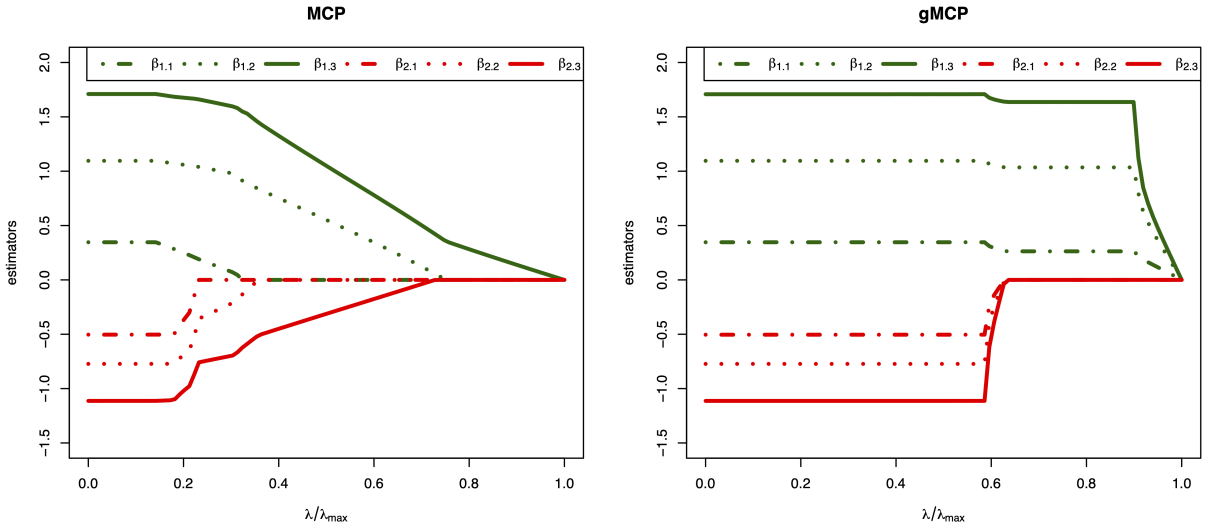

Figure 1.13: *Coefficient paths for MCP (left) and the groupwise variant group MCP (right). On the horizontal axis, the tuning parameter increases from left to right, where the coefficient estimates on the leftmost side in each plot are the MLE. In this example, $J = 2$ ordinal factors were chosen with $p_1 = p_2 = 3$ and $\boldsymbol{\beta}^* = (1.5, 0.7, 1.4, 1.2, -0.3, -0.5, -1)$.*

1.6 Overview

Having introduced several different penalized regression approaches, an overview of the theoretical properties that are, to the best of one's knowledge, shown so far in the existing literature or in this thesis is given in Table 1.1. However, it is noted that the regularity conditions needed may differ throughout the approaches, for more details one refers to the provided references. It is reported whether a specific property holds for the linear model (LM) or logistic regression (LogReg), since the linear model is treated in the majority of approaches while logistic regression is the model considered in this thesis.

Property	CAS-L_1		CAS-L_0		Group lasso		Group SCAD		Group MCP	
	LM	LogReg	LM	LogReg	LM	LogReg	LM	LogReg	LM	LogReg
$(\sqrt{n})$ consistency	Gertheiss and Tutz (2010b)	Thm. 1.3.2	n.s.	Thm. 1.3.3 $(\sqrt{n})$	Nardi and Rinaldo (2008) $(\sqrt{n})$	Thm. 1.4.3 $(\sqrt{n})$	Wang et al. (2007)	n.s.	Huang et al. (2012)	n.s.
consistency factor selection	-	-	-	-	Nardi and Rinaldo (2008)	Thm. 1.4.5	Wang et al. (2007)	n.s.	Huang et al. (2012)	n.s.
fusion consistency	Gertheiss and Tutz (2010b)	Thm. 1.3.1	n.s.	Thm. 1.3.4	-	-	-	-	-	-
asymp. normality	Gertheiss and Tutz (2010b)	Thm. 1.3.1	n.s.	n.s.	Nardi and Rinaldo (2008)	Thm. 1.4.4	Wang et al. (2007)	n.s.	n.s.	n.s.

Table 1.1: *Selected overview of theoretical properties shown so far of CAS-L_1, CAS-L_0, group lasso, group SCAD and group MCP. Here, "Thm." is an abbreviation of "Theorem". By $(\sqrt{n})$ consistency, it is either allowed that $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|| = o_p(1)$ or, knowing more about the rate $\sqrt{n}||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|| = O_p(1)$, where the latter implies the first. If $\sqrt{n}$ consistency is satisfied, thus $\sqrt{n}||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|| = O_p(1)$, the expression $\sqrt{n}$ is added in round brackets, otherwise only consistency, thus $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|| = o_p(1)$ holds. It is noted that fusion consistency can either rely on F_n and F^* or C_n and C^*, one compares Remark 1.2.9 and the referenced theorems in this table. The abbreviation "n.s." stands for "not shown", so the corresponding property is, to the best of one's knowledge, not shown so far in the considered setting.*

1.7 Simulation Studies

Having discussed the theoretical background of the considered methods, they are now compared with respect to their computational performance in different simulation studies. The goodness

of fit measures are supplied in Section 1.7.1, such that the quantities for the comparison are uniform and clear among all designs and methods.

Since *existing* approaches are compared in the following simulation studies, it is necessary to comment on the novelty of the studies presented in this thesis. In the literature, the majority of the existing studies consider only linear regression and neglect the area of logistic regression incorporating categorical explanatory variables. For example, Oelker et al. (2014b) present a simulation study where ML, CAS-L_1, CAS-L_0 and different choices of weights and choices of tuning with CV (and generalized CV) are compared. There is just one design where they assumed a binary random response variable, thus the question comes up how the approaches behave in different designs with a binary response. In addition, the high-dimensional case where $n < p$ is not covered in Oelker et al. (2014b) where these gaps are closed in the simulation studies of this thesis. More specifically, among others, a design including a factor with a rare but relevant category is considered, along with a design containing an interaction term of two explanatory variables and a design of high-dimension where $n < p$.
Gertheiss and Tutz (2010b) analyzed different approaches of CAS-L_1 for the linear model and not the logistic regression model, where the latter is treated in the following simulation studies. The work of Gertheiss and Tutz (2010a) deals with regularization in the area of categorical effect modifiers in the linear model which is generalized for GLMs in Oelker et al. (2014a). They both use a penalty that corresponds, in the case of ordinal data, to the fused lasso thus it is based on L_1. The observed study in Oelker et al. (2014a) assumes a binomial response incorporating only *continuous* covariates and one nominal effect modifier.
In the studies of Breheny and Huang (2015), linear and logistic regression were investigated but the comparison is limited to group lasso, group SCAD and group MCP. Guo et al. (2015) analyzed group lasso, group SCAD and SCAD but only for the linear model and not for logistic regression.
These references are some examples for simulation studies where the intensive analysis of penalized *logistic* regression for categorical covariates, including non-convex groupwise approaches, convex groupwise approaches and fusion-type penalties on differences of the coefficients are neglected, underlining the importance and novelty of the simulation studies conducted in this thesis. Different penalized regression approaches are compared in terms of goodness of fit measures which are provided in the next section.

1.7.1 Goodness of Fit Measures

One assumes that $\hat{\boldsymbol{\beta}}$ is some estimate arising from a particular penalized regression method and $\boldsymbol{\beta}^*$ the truth. Due to the focus of this thesis, modified versions of the "classical" false positive (FP) and false negative (FN) rates are considered to account for factor selection and levels fusion. It is noted that the sparsity measures, i.e. overall and practical sparsity introduced below, are not examined in the references discussed above, thus a new perspective for comparison with respect to the sparsity aspect is considered.

(i) FP/FN rates for factor selection, denoted by $FP_{s,\text{fac}}$ and $FN_{s,\text{fac}}$, respectively, that is

$$FP_{s,\text{fac}} := \frac{|\{j \in \{1,...,J\} \ : \ ||\hat{\boldsymbol{\beta}}_j|| \neq 0\,,\, ||\boldsymbol{\beta}_j^*|| = 0\}|}{|\{j \in \{1,...,J\} \ : \ ||\boldsymbol{\beta}_j^*|| = 0\}|}, \tag{1.64}$$

$$FN_{s,\text{fac}} := \frac{|\{j \in \{1,...,J\} \ : \ ||\hat{\boldsymbol{\beta}}_j|| = 0\,,\, ||\boldsymbol{\beta}_j^*|| \neq 0\}|}{|\{j \in \{1,...,J\} \ : \ ||\boldsymbol{\beta}_j^*|| \neq 0\}|}. \tag{1.65}$$

(ii) FP/FN concerning fusion with the truth being influential, measuring FP/FN fusion rates limited to those factors which are *truly* influential. These measures are denoted by

$FP_{f,\text{infl.truth}}$ and $FN_{f,\text{infl.truth}}$ and are defined as follows

$$FP_{f,\text{infl.truth}} := \frac{|\{(j,k,l) \,:\, \hat{\beta}_{j,k} \neq \hat{\beta}_{j,l}, \beta^*_{j,k} = \beta^*_{j,l}\,, \left(\sum_r |\beta^*_{j,r}|\right) \neq 0\}|}{|\{(j,k,l) : \beta^*_{j,k} = \beta^*_{j,l}\,, \left(\sum_r |\beta^*_{j,r}|\right) \neq 0\}|}, \tag{1.66}$$

$$FN_{f,\text{infl.truth}} := \frac{|\{(j,k,l) \,:\, \hat{\beta}_{j,k} = \hat{\beta}_{j,l}, \beta^*_{j,k} \neq \beta^*_{j,l}\,, \left(\sum_r |\beta^*_{j,r}|\right) \neq 0\}|}{|\{(j,k,l) : \beta^*_{j,k} \neq \beta^*_{j,l}\,, \left(\sum_r |\beta^*_{j,r}|\right) \neq 0\}|}. \tag{1.67}$$

In practice, the value zero is added at the beginning of every sub-vector of $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$ to ensure that fusion with the reference category is also taken into account. For ordinal factors the adjacent indices $(j,k,k-1)$ are compared.

(iii) Predictive deviance (compare Definition 1.2.1)

$$D(\boldsymbol{v}|\hat{\boldsymbol{\mu}}) := -2\sum_{i=1}^{n}\{y_i \log(\hat{\mu}_i) + (1-y_i)\log(1-\hat{\mu}_i)\}.$$

(iv) Mean squared error of the coefficients (MSEC)

$$\text{MSEC}(\hat{\boldsymbol{\beta}}) = \frac{1}{p}\sum_{j=1}^{p}(\beta^*_j - \hat{\beta}_j).$$

(v) Sparsity measures, in particular practical sparsity (PS) and overall sparsity (OS), which are defined as

$$\begin{aligned} \text{PS} &:= |\{j \in \{1,..,J\} : ||\hat{\boldsymbol{\beta}}_j||_2 \neq 0\}| \text{ and true value } \text{PS}^* := |\{j \in \{1,..,J\} : ||\boldsymbol{\beta}^*_j||_2 \neq 0\}| \\ \text{OS} &:= |\{k \in \{1,...,p\} : \hat{\beta}_k \neq 0\}| \text{ and true value } \text{OS}^* := |\{k \in \{1,...,p\} : \beta^*_k \neq 0\}| \end{aligned}$$

Having provided the goodness of fit measures with respect to which the methods are compared, the next step is to specialize which methods are analyzed, along with the choice of weights and computational details.

1.7.2 Methods

A broad comparison of the following methods in different design settings is obtained, where the respective abbreviation used in the studies is given in brackets

(i) unpenalized ML (`ML`) introduced in Section 1.1.2,

(ii) CAS-L_1 (`L1.CV`) introduced in Section 1.3.1,

(iii) CAS-L_0 (`L0.CV`) introduced in Section 1.3.2,

(iv) group lasso (`glasso.CV`) introduced in Section 1.4.1,

(v) group SCAD (`gSCAD.CV`) introduced in Section 1.5.1,

(vi) group MCP (`gMCP.CV`) introduced in Section 1.5.2.

Some fundamental methods discussed earlier in this thesis are excluded from the simulation studies for the following reasons. In the foregoing sections, Ridge regression, Elastic Net and the lasso were introduced. However, it is known that Ridge regression is not able to perform any variable selection. The lasso approach is able to perform variable selection, but it does not perform *factor* selection, thus it ignores an underlying groupwise structure. Similar arguments apply to Elastic Net as convex combination of lasso and Ridge. Thus, no further reporting and analyzing Ridge, lasso and Elastic Net is done here, since these approaches are not suitable for the purposes of this thesis.
Finally, SCAD and MCP are neglected by the same arguments that were given for lasso. Nevertheless, their groupwise variants are investigated.

The methods that are compared (listed above) are all (except for (i)) either designed for factor selection or for levels fusion, making them promising approaches for the presence of categorical explanatory variables. The unpenalized ML is only included for reference.

Choice of Weights

In Oelker and Tutz (2013), as well as in Gertheiss and Tutz (2010b), the weights given by (1.41) for nominal factors and (1.42) for ordinal factors are recommended, thus those weights are employed for the methods CAS-L_1 and CAS-L_0. For group lasso, Breheny and Huang (2015) (equation (2.5)) multiply the L_2 norm of the coefficient sub-vectors by the weight $\sqrt{p_j}$, which is in line with (1.54). The same is done for group SCAD and group MCP, in particular in their work they absorb the weight into the tuning parameter, that is, they write $\lambda_j := \lambda \cdot \sqrt{p_j}$ and use λ_j as tuning/penalization parameter for factor j, instead of explicitly defining some weighting factor. However, this is clearly equivalent. Consequently, the weight $\sqrt{p_j}$ is employed for every sub-vector in group lasso, group MCP and group SCAD.

All these weights mentioned above are *non-adaptive* weights. Non-adaptive weights are used for all considered methods to keep a comparable scale. However, adaptive weights are imposed in Chapter 3 considering the new introduced penalty function of Chapter 2.

Computational Details and Tuning

CAS-L_1 and CAS-L_0 were fitted with the `R` package `gvcm.cat` which obtains estimates executing the PIRLS algorithm. For the required parameters (one compares Appendix A.1) $c = 10^{-5}$ was chosen (for L_0 and L_1) as in Oelker et al. (2014b) which is needed for the approximation of the penalty in PIRLS. Furthermore, for L_0 an additional parameter γ for the approximation of the L_0 norm is needed, where $\gamma = 10$ was chosen following Oelker et al. (2014b). Finally, as step length parameter in PIRLS, $\nu = 0.05$ was used (for L_0 and L_1), again following Oelker et al. (2014b).

Group lasso, group SCAD and group MCP were fitted using the `R` package `grpreg` which obtains estimates executing BCD.

For every method, 5-fold CV was employed to determine the tuning parameter λ. The tuning ranges as well as the number of λ values to be fitted in this range are not specified, instead the default settings implemented in `gvcm.cat` and `grpreg`, respectively, were used. However, the authors explain that the minimum value λ_{min} of λ is typically chosen as being zero, while the maximum value λ_{max} is chosen such that all factors are excluded from the model. Then, λ is cross-validated over the interval $[\lambda_{min}, \lambda_{max}]$.

1.7.3 Simulation Designs

To compare the different approaches, several simulation designs were conducted which are presented below. Designs B8.1 and B8.2 are both *sparse* designs, with at least half of the candidate explanatory variables being noise variables, where B8.1 is unbalanced and in B8.2 the class probabilities were sampled in a specific range. Further, a design was incorporated with a rare but relevant category, called B6.rare. A design with an interaction term is obtained by design B6.inter.pos, which will sometimes be abbreviated as B6.inter. Finally, a design of high-dimension is included, denoted by design highdim, where $p = 171 > 100 = n$. In all designs introduced below, the random response variable Y is chosen to be binary, considering an underlying logistic regression model. Further, the datasets were sampled using the `R` function `simulation` contained in `gvcm.cat`.

Design B8.1

The idea of the following design, which is called B8.1, is taken from Gertheiss and Tutz (2010b) where in the given reference the design was used for the case of linear regression. It consists of $J = 8$ covariates $\mathscr{X}_1, \ldots, \mathscr{X}_8$, in particular $\mathscr{X}_5, \mathscr{X}_6, \mathscr{X}_7, \mathscr{X}_8$ are nominal with true parameter vectors $\boldsymbol{\beta}_5^*, ..., \boldsymbol{\beta}_8^*$ and the remaining variables $\mathscr{X}_1, ..., \mathscr{X}_4$ are ordinal with true parameter vectors $\boldsymbol{\beta}_1^*, ..., \boldsymbol{\beta}_4^*$. Among the four ordinal (nominal) covariates, two covariates have eight levels and two covariates have four levels, in particular $p_1 = p_2 = p_5 = p_6 = 7$ and $p_3 = p_4 = p_7 = p_8 = 3$. Four of the covariates are influential and the remaining four are non-influential noise variables. The covariates are drawn from a multinomial distribution with probabilities $(0.1, 0.1, 0.2, 0.05, 0.2, 0.1, 0.2, 0.05)$ for the eight-level covariates and $(0.1, 0.4, 0.2, 0.3)$ for the others. The true coefficient vector is given by

$$\begin{aligned}
&\boldsymbol{\beta}^* = (1, \boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*, \boldsymbol{\beta}_3^*, \boldsymbol{\beta}_4^*, \boldsymbol{\beta}_5^*, \boldsymbol{\beta}_6^*, \boldsymbol{\beta}_7^*, \boldsymbol{\beta}_8^*),\\
&\text{where } \boldsymbol{\beta}_1^* = (0, 1, 1, 2, 2, 4, 4),\ \boldsymbol{\beta}_2^* = (0, 0, 0, 0, 0, 0, 0),\ \boldsymbol{\beta}_3^* = (0, -2, -2),\ \boldsymbol{\beta}_4^* = (0, 0, 0),\\
&\qquad \boldsymbol{\beta}_5^* = (0, 1, 1, 1, 1, -2, -2),\ \boldsymbol{\beta}_6^* = (0, 0, 0, 0, 0, 0, 0),\ \boldsymbol{\beta}_7^* = (0, 2, 2),\ \boldsymbol{\beta}_8^* = (0, 0, 0).
\end{aligned}$$

The sample size is chosen to be $n = 1000$ and $R = 100$ replications were executed. The chosen number of replications follows the source of this design, i.e. Gertheiss and Tutz (2010b).

Design B8.2

The idea of the second examined design, called B8.2, is taken from Oelker et al. (2014b). The simulation was done for $J = 8$ ordinal covariates $\mathscr{X}_1, \ldots, \mathscr{X}_8$ with four levels each, hence $p_i = 3\ \forall i \in \{1, \ldots, 8\}$, drawn from a multinomial distribution with sampled probabilities between 0.12 and 0.44. Among the eight covariates, four covariates are chosen to be influential and the remaining four are chosen to be non-influential noise variables. The true coefficient vector is given by

$$\begin{aligned}
&\boldsymbol{\beta}^* = (2, \boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*, \boldsymbol{\beta}_3^*, \boldsymbol{\beta}_4^*, \boldsymbol{\beta}_5^*, \boldsymbol{\beta}_6^*, \boldsymbol{\beta}_7^*, \boldsymbol{\beta}_8^*),\\
&\text{where } \boldsymbol{\beta}_1^* = (0, -0.8, -0.8),\ \boldsymbol{\beta}_2^* = (1, 1, 0),\ \boldsymbol{\beta}_3^* = (0.4, 0.6, 0.8),\ \boldsymbol{\beta}_4^* = (-0.7, -1, 0),\\
&\qquad \boldsymbol{\beta}_5^* = \boldsymbol{\beta}_6^* = \boldsymbol{\beta}_7^* = \boldsymbol{\beta}_8^* = (0, 0, 0).
\end{aligned}$$

As in design B8.1, the sample size is $n = 1000$ with $R = 100$ replications. As in design B8.1, the chosen number of replications follows the source of this design, i.e. Oelker et al. (2014b). Consequently, for the upcoming designs, the same number of replications is chosen to ensure comparability.

Design B6.rare

This design consists of $J = 6$ nominal covariates $\mathscr{X}_1, \ldots, \mathscr{X}_6$ where just $\mathscr{X}_1$ and $\mathscr{X}_2$ are chosen to have an influence on the random response variable. $\mathscr{X}_2$ includes a rare but relevant category, and further a rare but *not* relevant category. However, the focus here clearly lies on the detection of the rare but *relevant* category. The sampling is conducted as follows

$$\begin{aligned}
&\mathscr{X}_1 \sim \text{Bin}(1,\, 0.6),\\
&\mathscr{X}_2 \sim \text{Mult}(0.1, 0.01, 0.35, 0.01, 0.07, 0.2, 0.05, 0.06, 0.06, 0.09),\\
&\mathscr{X}_3 \sim \text{Bin}(0.3),\\
&\mathscr{X}_4 \sim \text{Mult}\left(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right),\\
&\mathscr{X}_5 \sim \text{Mult}\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right),\\
&\mathscr{X}_6 \sim \text{Bin}(0.3).
\end{aligned}$$

The true coefficient vector is given by

$$\begin{aligned}
&\boldsymbol{\beta}^* = (1, \boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*, \boldsymbol{\beta}_3^*, \boldsymbol{\beta}_4^*, \boldsymbol{\beta}_5^*, \boldsymbol{\beta}_6^*),\\
&\text{where } \boldsymbol{\beta}_1^* = \beta_{1,1}^* = -0.5,\ \boldsymbol{\beta}_2^* = (5, 1, 0, 1, -0.5, -0.5, 1.5, 2, 0.6),\ \boldsymbol{\beta}_3^* = \beta_{3,1}^* = 0,\\
&\qquad \boldsymbol{\beta}_4^* = (0, 0, 0, 0, 0),\ \boldsymbol{\beta}_5^* = (0, 0),\ \boldsymbol{\beta}_6^* = \beta_{6,1}^* = 0.
\end{aligned}$$

This design is chosen such that the first level of $\mathscr{X}_2$ (not the reference category) is rare (sample probability of 0.01) but relevant (entry of the coefficient vector equals 5).
As in designs above, the sample size is $n = 1000$ with $R = 100$ replications.

Design B6.inter.pos

This design is a design with an interaction (positive interaction) and it is called B6.inter.pos, sometimes abbreviated as B6.inter, including $J = 6$ covariates $\mathscr{X}_1, \ldots, \mathscr{X}_6$. Two binary variables $\mathscr{X}_1 \sim \text{Bin}(1,\, 0.6)$ and $\mathscr{X}_2 \sim \text{Bin}(1,\, 0.25)$ are simulated, thus with success probabilities 0.6 and 0.25, respectively. $\mathscr{X}_1$ and $\mathscr{X}_2$ are both chosen to be influential on the random response variable. Having that, an interaction variable $\mathscr{X}_3 := \mathscr{X}_1 \cdot \mathscr{X}_2$ is defined. Choosing $\beta_{1,1}^* = \beta_{2,1}^* = 0.5$ yields that both $\mathscr{X}_1$ and $\mathscr{X}_2$ have the same influence on the response while $\beta_{3,1}^* = 2$ means that the appearance of both yields a positive interaction. It is noted that, since $\mathscr{X}_1$, $\mathscr{X}_2$ and $\mathscr{X}_3$ are binary, their corresponding coefficient sub-vectors only have one entry. Choosing some noise variables $\mathscr{X}_4 \sim \text{Mult}(0.4, 0.2, 0.3, 0.1)$, $\mathscr{X}_5 \sim \text{Bin}(1,\, 0.7)$ another influential variable $\mathscr{X}_6 \sim \text{Mult}(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ is added which does not influence the interaction variable. All covariates are chosen to be ordinal and the true coefficient vector is given by

$$\begin{aligned}
&\boldsymbol{\beta}^* = (1, \boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*, \boldsymbol{\beta}_3^*, \boldsymbol{\beta}_4^*, \boldsymbol{\beta}_5^*, \boldsymbol{\beta}_6^*),\\
&\text{where } \boldsymbol{\beta}_1^* = \beta_{1,1}^* = 0.5,\ \boldsymbol{\beta}_2^* = \beta_{2,1}^* = 0.5,\ \boldsymbol{\beta}_3^* = \beta_{3,1}^* = 2,\\
&\qquad \boldsymbol{\beta}_4^* = (0, 0, 0),\ \boldsymbol{\beta}_5^* = \beta_{5,1}^* = 0,\ \boldsymbol{\beta}_6^* = (-0.4, -0.6).
\end{aligned}$$

As in the designs above, the sample size is $n = 1000$ with $R = 100$ replications.

Design highdim

In this design, $J = 60$ ordinal covariates $\mathscr{X}_1, \ldots, \mathscr{X}_{60}$ are considered where the first 50 have four categories, in particular

$$\mathscr{X}_j \sim \text{Mult}\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right),\ \forall\, j \in \{1, \ldots, 50\}$$

and the last ten covariates have three categories, that is

$$\mathscr{X}_j \sim \text{Mult}\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right), \ \forall\, j \in \{51, \ldots, 60\}.$$

The first five covariates $\mathscr{X}_1, \ldots, \mathscr{X}_5$ are chosen to be influential which is approximately 8% of the covariates in total. In contrast to the other designs, $n = 100$ observations are simulated such that this design with $p = 171$ and $p > n$ is high-dimensional. The true coefficient vector is given by

$$\begin{aligned}
&\boldsymbol{\beta}^* = (2, \boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*, \boldsymbol{\beta}_3^*, \boldsymbol{\beta}_4^*, \boldsymbol{\beta}_5^*, \boldsymbol{\beta}_6^*, \ldots, \boldsymbol{\beta}_{60}^*),\\
&\text{where } \boldsymbol{\beta}_1^* = (-1, 0.5, 2),\ \boldsymbol{\beta}_2^* = (1.5, 1.5, 0.5),\ \boldsymbol{\beta}_3^* = (1, 2, 2.5),\ \boldsymbol{\beta}_4^* = (-0.5, -0.3, 0.5),\\
&\qquad \boldsymbol{\beta}_5^* = (2, 1, 3),\ \boldsymbol{\beta}_j^* = (0, 0, 0)\ \forall j \in \{6, \ldots, 50\},\ \boldsymbol{\beta}_j^* = (0, 0)\ \forall j \in \{51, \ldots, 60\}.
\end{aligned}$$

$R = 100$ replications are executed.

1.8 Analysis of the Results

The results of the simulation designs presented in Section 1.7.3 with the methods listed in Section 1.7.2 are provided and analyzed in this section. It is recalled that L_1 is written as an abbreviation of CAS-L_1 and L_0 as an abbreviation of CAS-L_0.

For some simulation designs, further specific goodness of fit measures are of interest, which are provided in the following remark.

Remark 1.8.1.

(i) Sometimes it is necessary to add observations if data is analyzed where complete separation (Albert and Anderson (1984), Section 3.2) occurs. If this is the case, the augmented dataset is used for all the penalties such that they are on a comparable scale and the proportion of added observations is reported.

(ii) For design B6.inter.pos with an interaction term, the proportion of replications where the corresponding variables, incorporating the interaction term, are included in the model is reported.

(iii) Especially in the high-dimensional design highdim, it is important to report the proportion of failures of the considered methods. Here, failure refers to the fact that the conducted function in `R` did not yield a result, which may have different reasons, e.g. convergence issues or singular matrices.

(iv) In design B6.rare with a rare but relevant category the proportion of replications where the rare but relevant category was excluded from the model is measured.

(v) The values for $FP_{s,\text{fac}}$, $FN_{s,\text{fac}}$, $FP_{f,\text{infl.truth}}$, $FN_{f,\text{infl.truth}}$ as well as OS and PS provided in the tables presenting the results of the simulation studies are the *mean values* over all $R \in \mathbb{N}$ replications.

First, the proportion of observations that needed to be added for every replication is investigated. Calculating the mean value over all replications, the resulting mean proportions are displayed in Table 1.2. One can see that all of these values are comparably low (around 1% for highdim and $< 1\%$ for the other designs), so just a few observations needed to be added in some replications to avoid complete separation, consequently the results are representable.

Design	Mean value
B8.1 $(n = 1000)$	0.00056
B8.2 $(n = 1000)$	0
B6.rare $(n = 1000)$	0.00214
B6.inter.pos $(n = 1000)$	$6 \cdot 10^{-5}$
highdim $(n = 100)$	0.0108

Table 1.2: *Mean value over all $R = 100$ replications of proportion of added observations out of sample size n.*

1.8.1 Results of Design B8.1

In the conducted study, none of the methods failed in any replication. First, the sparsity levels obtained by the methods are examined, which are given in Table 1.3. In this table, the mean value over all $R = 100$ replications of the sparsity levels, overall sparsity (OS) and practical sparsity (PS), are displayed. Since ML performs no (factor) selection, one can see that OS = 40.00 and PS = 8.00 for this approach, which is what one expects and what is analogously observed in all other designs. However, looking at the values for L_1 and L_0, L_0 obtains a sparser model than L_1, indicated by smaller values of the sparsity levels, which is similarly observed in the simulation studies of Oelker et al. (2014b). Nevertheless, the sparsity levels of L_0 indicate that still too little selection is performed, since OS = 24.86 and PS = 6.61 where the true values are given by $\text{OS}^* = 16$ and $\text{PS}^* = 4$, respectively. As for L_1, the OS and PS of group lasso are comparably high, being near to ML, indicating that, in this design B8.1, L_1 and group lasso yield models that are not sparse enough. The sparsity levels of group SCAD and group MCP can be located on a similar level, yielding OS and PS values being nearest to the truth among the observed methods. Further, group MCP seems to select a slightly more sparse model, which is similarly observed by Breheny and Huang (2015) (Section 4).

	ML	L1.CV	L0.CV	gLasso.CV	gSCAD.CV	gMCP.CV
OS	40.00	35.09	24.86	37.29	20.75	20.33
PS	8.00	7.91	6.61	7.35	4.17	4.07

Table 1.3: *[B8.1, n=1000] Overall Sparsity (OS) and Practical Sparsity (PS), true values are given by $OS^* = 16$, $PS^* = 4$.*

Turning the focus to Table 1.4 analyzing the resulting FP/FN rates, the fact that L_1 selects models being not sparse enough in this design, which was observed in the paragraph above analyzing OS/PS, is confirmed since $FP_{s,\text{fac}}$ is given by 0.98 while $FN_{s,\text{fac}}$ is 0.00 so nearly no factor selection is performed. The corresponding fusion rates $FP_{f,\text{infl.truth}}$ and $FN_{f,\text{infl.truth}}$ show a similar pattern. The same applies for group lasso, which is related to L_1, both being based on a lasso-type penalty. Comparing both fusion penalties, thus L_1 and L_0, the $FP_{s,\text{fac}}$ for L_0 is 33% lower (in absolute difference), where the $FN_{s,\text{fac}}$ rate for L_0 is zero, so the factor selection performance of L_0 is improved against the L_1 penalization of the differences. Nevertheless, as discussed in the paragraph above, the (factor) selection of L_0 is still too weak. For L_0, comparing the fusion rates $FP_{f,\text{infl.truth}}$ and $FN_{f,\text{infl.truth}}$ to those of all other methods, L_0 shows the most satisfactory fusion performance. To be more precise, group lasso, group SCAD and group MCP perform no fusion at all, which can be seen by the $FP_{f,\text{infl.truth}}$ rate of 1.00 and correspondingly $FN_{f,\text{infl.truth}}$ rate of 0.00. Further, L_1 also shows a high value of $FP_{f,\text{infl.truth}}$, being near to 1.00, together with a value of $FN_{f,\text{infl.truth}}$ being near to 0.00, thus only weak fusion is performed. It remains to discuss the factor selection performance of group SCAD and group MCP. These

two approaches show an outstanding selection performance with both rates $FP_{s,\text{fac}}$ and $FN_{s,\text{fac}}$ being near to zero. As observed above in terms of OS/PS, group MCP seems to be slightly sparser compared to group SCAD.

	ML	L1.CV	L0.CV	gLasso.CV	gSCAD.CV	gMCP.CV
$FP_{s,\text{fac}}$	1.00	0.98	0.65	0.84	0.04	0.02
$FN_{s,\text{fac}}$	0.00	0.00	0.00	0.00	0.00	0.00
$FP_{f,\text{infl.truth}}$	1.00	0.81	0.39	1.00	1.00	1.00
$FN_{f,\text{infl.truth}}$	0.00	0.02	0.21	0.00	0.00	0.00

Table 1.4: *[B8.1, n=1000] FP/FN rates fusion and factor selection.*

To sum up, in terms of OS/PS and the discussed FP/FN rates, one can conclude that group SCAD and group MCP show the best selection performance, while they perform no fusion. Moreover, it is inferred that L_0 performs selection *and* fusion tasks, whereas the selection performance is comparably weak. The methods L_1 and group lasso achieve no satisfactory results in terms of OS/PS and the considered FP/FN rates. However, it remains to analyze the predictive deviance and MSEC, which is done next.

Remark 1.8.2 (Comparison to Gertheiss and Tutz (2010b))**.** In Gertheiss and Tutz (2010b), where the design B8.1 is taken from, a linear model is observed with a sample size of $n = 500$, while an underlying logistic regression model with $n = 1000$ was employed here, to be in line with the other designs that were considered. They focus on the L_1 penalty with different choices of weights, that is (i) without weights, (ii) with marginal class frequencies as weights, (iii) adaptive weights with the inverse of the LSE and (iv) adaptive weights with the inverse of the LSE and marginal class frequencies. The choice of weights in this thesis (one compares Section 1.7.2) corresponds to (ii). Looking at Figure 5 of Gertheiss and Tutz (2010b), one can see that the FP selection rate for the version weighted with marginal class frequencies (ii) are located around 0.80 and the corresponding FN rate around 0.00. For the fusion rate they report as FP value around 0.65 and for the FN value around 0.00. Since the rates are given in a bar chart, the exact values cannot be provided here. These rates show a similar pattern as the rates that were observed for L_1 in Table 1.4, with the latter results being more extreme. However, in Gertheiss and Tutz (2010b), they show that adaptive weights clearly improve the results, while such adaptive weights were not considered in the simulations of Chapter 1, to keep all methods on a comparable scale. The fact that adaptive weights improve the performance of L_1 is also what would be expected, since the penalization rate of L_1 depends on the absolute value of the (differences of the) coefficients. As mentioned earlier, adaptive weights are imposed in Chapter 3 considering the new introduced penalty function of Chapter 2.

According to Figure 1.14, it can be seen that the median of the MSEC is higher for L_0 compared to L_1 while L_0 also shows slightly more variance in the errors. Even though L_1 performs too little (factor) selection and fusion, as elaborated above, it shrinks the parameter estimates towards zero because of the properties of the lasso type penalty. This is a difference to L_0 which does not shrink estimates towards zero. This causes that even if the OS and PS values are "too high" for L_1, the MSEC is lower than for L_0 because of the shrinkage. Analogous arguments apply for group lasso. From both figures (MSEC and predictive deviance), one can infer that L_0 is more sensitive corresponding to changes in the data which yields a higher variance in predictive deviance and MSEC. In the simulation study B8 of Oelker et al. (2014b) (which is picked up in design B8.2), they similarly observe that L_0 is more sensitive with respect to variations in the data, compared to L_1. This may be caused by the fact that the penalty function is not

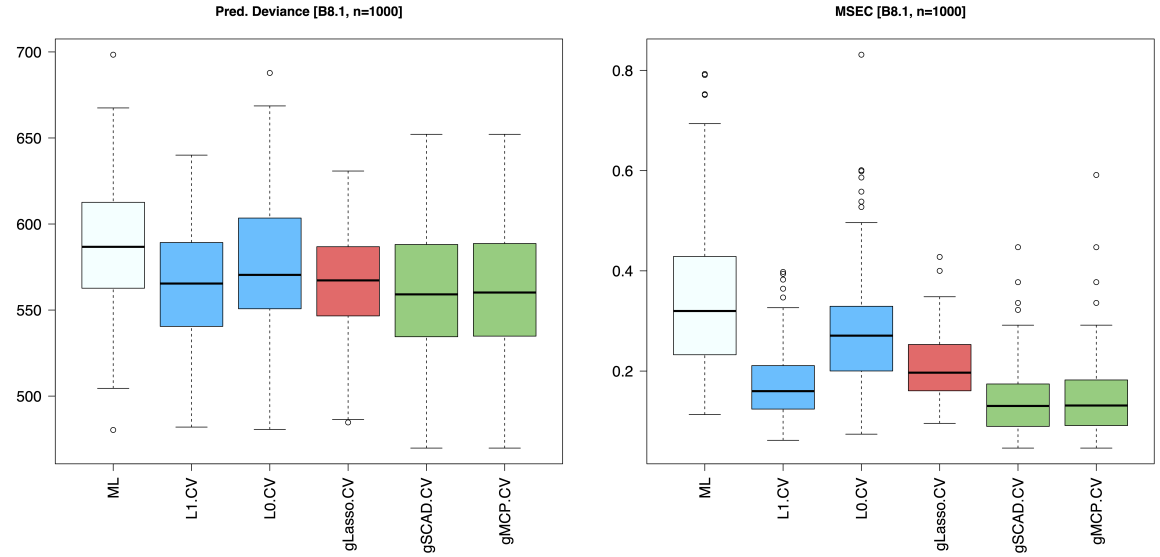

Figure 1.14: *Predictive deviance and mean squared error of coefficients (MSEC) for design B8.1,* $n = 1000$.

continuous. Group SCAD and group MCP can be located on a similar level in terms of MSEC and predictive deviance, which is in line with the observations made in Breheny and Huang (2015) (Section 4). In terms of the predictive deviance measure, all methods, except for ML, can be located on a comparable scale with their medians being similar.

To sum up the elaborations of the paragraphs above, one can argue that L_0 shows the best fusion performance, while group SCAD and group MCP show the best factor selection performance. However, even though group lasso cannot keep up with group SCAD and group MCP in terms of factor selection performance in this design, the shrinkage properties cause that the resulting predictive deviance and MSEC values are still competitive.

1.8.2 Results of Design B8.2

For all considered methods, no failures in any of the replications were observed. Table 1.5 represents the OS/PS sparsity levels of the fitted methods, in particular the mean value over all performed replications. A similar pattern as in the previous design B8.1 (Section 1.8.1) is observed, that is, L_1 and group lasso show values for OS and PS being the highest, indicating that the resulting models are not sparse enough. Recalling that L_1 and group lasso are both penalties based on the lasso explains their similarities in behavior. However, it is crucial to remark that the fact that the OS of L_0 is less than the OS of group lasso does *not* mean that the selection performance of L_0 is "better" (one consults $FP_{s,\text{fac}}$ and $FN_{s,\text{fac}}$ in Table 1.6), since the OS/PS values do not take into consideration whether the right entries/sub-vectors of the coefficient vector are set to zero. Further, also being similar to design B8.1 in Section 1.8.1, in terms of OS/PS, the methods group SCAD and group MCP can be located on a comparable scale, where the models selected by group MCP seem to be more sparse compared to group SCAD.

	ML	L1.CV	L0.CV	gLasso.CV	gSCAD.CV	gMCP.CV
OS	24.00	20.34	16.32	19.65	13.98	12.30
PS	8.00	7.66	6.58	6.55	4.66	4.10

Table 1.5: *[B8.2, n=1000] Overall Sparsity (OS) and Practical Sparsity (PS), true values are given by* $OS^* = 9$, $PS^* = 4$.

Next, the FP/FN rates obtained in Table 1.6 are analyzed. One can see that in this design, the factor selection performance of group lasso is better than the performance of L_1, indicated by a lower $FP_{s,\text{fac}}$ rate and similar $FN_{s,\text{fac}}$ rate. However, group lasso performs no fusion. Comparing both penalties designed for fusion, i.e. L_0 and L_1, L_1 shows a higher $FP_{f,\text{infl.truth}}$ rate, while the $FN_{f,\text{infl.truth}}$ rate is lower, where the differences in absolute values are $|0.60 - 0.37| = 0.23$ and $|0.09 - 0.21| = 0.12$. This indicates that the L_0 approach performs "more" fusion, but at the cost of some false fusions. However, for a more balanced and less conservative fusion result, one would prefer L_0.

	ML	L1.CV	L0.CV	gLasso.CV	gSCAD.CV	gMCP.CV
$FP_{s,\text{fac}}$	1.00	0.92	0.65	0.64	0.22	0.12
$FN_{s,\text{fac}}$	0.00	0.00	0.00	0.01	0.05	0.09
$FP_{f,\text{infl.truth}}$	1.00	0.60	0.37	1.00	0.97	0.95
$FN_{f,\text{infl.truth}}$	0.00	0.09	0.21	0.01	0.06	0.10

Table 1.6: *[B8.2, n=1000] FP/FN rates fusion and factor selection.*

Finally, one considers Figure 1.15. As in the last design (Section 1.8.1), the L_0 approach seems to be sensitive in variations in the data. The fact that the MSEC is the lowest for L_1 and group lasso can be explained by the fact that both penalties perform shrinkage of the coefficients. Thus, even though they may select too little factors as noise variables, it may happen that they shrink the corresponding coefficients towards zero, causing a competitive value for MSEC. In terms of predictive deviance, all approaches could be ranked on a similar scale.

To conclude, L_0 shows the most satisfactory fusion performance being less conservative than L_1. Group SCAD and group MCP show the best factor selection performance, however, in terms of MSEC, group lasso and L_1 outperform the other approaches.

1.8.3 Results of Design B6.rare

It is recalled that this design includes a category that is rare but relevant, hence dummy variable $\mathscr{X}_{2,1}$ has a comparably small success probability being 0.01 and a comparably high true coefficient vector of $\beta^*_{2,1} = 5$. Furthermore a non-influential rare category is included, in particular dummy variable $\mathscr{X}_{2,3}$, where the focus clearly lies on the detection of the rare but relevant category, which is of high importance in applications for example to detect rare diseases. Thus, besides the measures that were considered in the previous designs, measures showing whether the methods included or excluded the rare but (not) relevant categories in the model are examined, where it is noted that exclusion corresponds to fusion with the reference category.

For all considered methods, no failures in any of the replications were observed. First, the sparsity measures, thus OS/PS values given in Table 1.7 are analyzed. One can notice that L_1 selects the less sparse model, while L_0 and group lasso can be ranked on a similar scale. Group

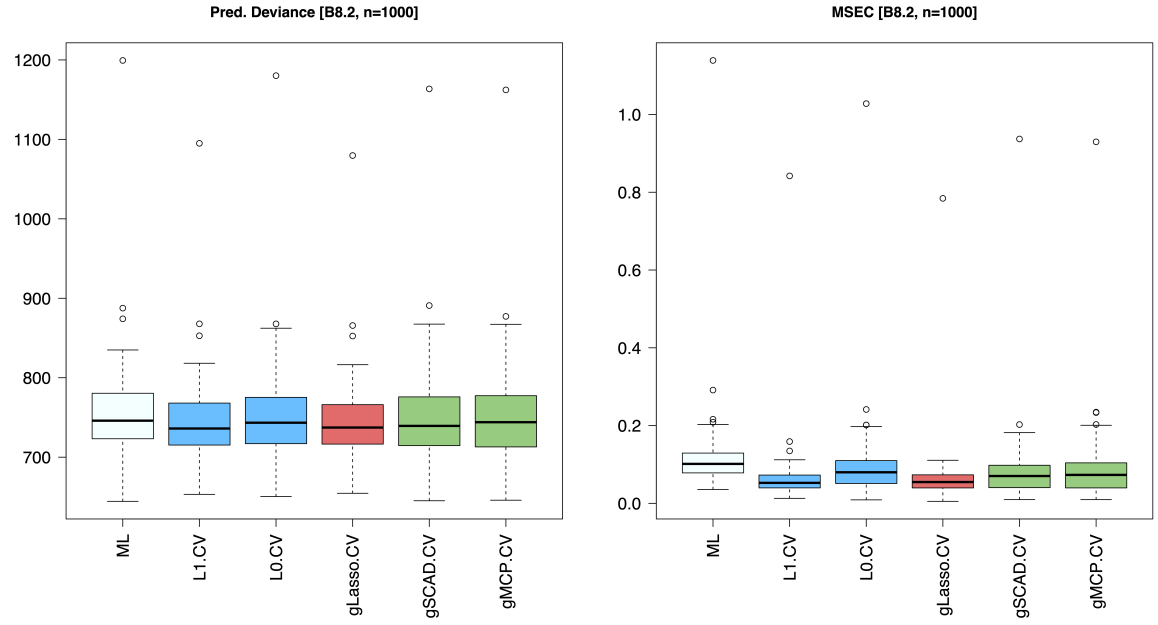

Figure 1.15: *Predictive deviance and mean squared error of coefficients (MSEC) for design B8.2, $n = 1000$.*

SCAD and group MCP yield the most sparse models, being nearest to the true sparsity levels. As in the previous designs, L_1 is less sparse than L_0, performing too little of dimension reduction.

	ML	L1.CV	L0.CV	gLasso.CV	gSCAD.CV	gMCP.CV
OS	19.00	18.23	15.55	15.10	10.61	9.89
PS	6.00	5.86	5.39	4.09	2.33	1.87

Table 1.7: *[B6.rare, n=1000] Overall Sparsity (OS) and Practical Sparsity (PS), true values are given by $OS^* = 9$, $PS^* = 2$.*

Turning the focus on Table 1.8 showing the FP/FN rates concerning factor selection and levels fusion, L_1 has the highest values of $FP_{s,\text{fac}}$, underlying the observations made above analyzing Table 1.7. L_0 decreases the $FP_{s,\text{fac}}$ rate compared to L_1, however, the factor selection performance of these two approaches is not satisfactory. Group lasso highly improves the factor selection performance compared to L_1 and L_0. The $FP_{s,\text{fac}}$ rates for group SCAD and group MCP are the lowest, fact that was similarly observed in the foregoing simulation designs. Nevertheless, group SCAD and group MCP, as well as group lasso, perform no fusion, showed by the $FP_{f,\text{infl.truth}}$ rates being equal to one and $FN_{f,\text{infl.truth}}$ rates being equal to zero. The fusion performance of L_1, measured in $FP_{f,\text{infl.truth}}$ and $FN_{f,\text{infl.truth}}$, indicates that not enough fusion is performed, while L_0 performs more fusion, coming with an increase in $FN_{f,\text{infl.truth}}$.

Table 1.9 shows the proportion of replications where the rare but relevant category was (falsely) excluded from the model and the corresponding value for the rare but not relevant category. Furthermore, a column was added showing the absolute difference of these two pro-

	ML	L1.CV	L0.CV	gLasso.CV	gSCAD.CV	gMCP.CV
$FP_{s,\text{fac}}$	1.00	0.96	0.85	0.53	0.10	0.02
$FN_{s,\text{fac}}$	0.00	0.00	0.00	0.01	0.03	0.10
$FP_{f,\text{infl.truth}}$	1.00	0.92	0.68	1.00	1.00	1.00
$FN_{f,\text{infl.truth}}$	0.00	0.05	0.25	0.00	0.00	0.00

Table 1.8: *[B6.rare, n=1000] FP/FN rates fusion and factor selection.*

portion values, where one aims for a high value such that the method can distinguish between rare but relevant and rare but not relevant. Here, exclusion from the model refers to fusion with the reference category for fusion-type penalties, while the factor-wise selection penalties group lasso, group SCAD and group MCP will either include or exclude the whole factor. The latter fact can be seen in Table 1.9, since both proportion values are equal. On the one hand, they do not exclude the rare but relevant category from the model, on the other hand they also cannot exclude the rare but not relevant one (hence fuse with the reference category). For L_1 and L_0, the absolute difference in the proportion values is around 10%, which is the highest value among the considered methods.

	rr excl.	rnr excl.	abs. diff.
ML	0.00	0.00	0.00
L1.CV	0.00	0.09	0.09
L0.CV	0.06	0.17	0.11
gLasso.CV	0.00	0.00	0.00
gSCAD.CV	0.00	0.00	0.00
gMCP.CV	0.00	0.00	0.00

Table 1.9: *[B6.rare, n=1000] Proportion of replications where the rare but relevant (rr) and rare but not relevant (rnr) category was excluded from the model. Further, in the last column, the absolute difference (abs. diff.) of the two proportions.*

Finally, the predictive deviance and the MSEC displayed in Figure 1.16 are observed, where all approaches can be ranked on a similar scale both in terms of predictive deviance as well as MSEC.

Such a design may appear in real data applications in detection of rare diseases, thus one would prefer the groupwise approaches group lasso, group SCAD and group MCP in terms of the measures discussed above in such examples. However, this design shows that it may be reasonable to use a penalty function that is able to perform *both* factor selection as well as fusion of levels. With that, one could obtain a sparse model and differentiate between rare but relevant and rare but not relevant categories.

1.8.4 Results of Design B6.inter.pos

This design is a design including an interaction term of the explanatory variables $\mathscr{X}_1$ and $\mathscr{X}_2$. Thus, it is analyzed whether the methods are able to detect this interaction. In particular, one sets $\mathscr{X}_3 = \mathscr{X}_1 \cdot \mathscr{X}_2$ and $\mathscr{X}_1, \mathscr{X}_2, \mathscr{X}_3$ are influential where the appearance of both $\mathscr{X}_1$ and $\mathscr{X}_2$ has significantly more influence on the response than the appearance of just one of the two, in terms of a higher true coefficient value.

No failures of the considered methods in any of the replications were observed. First, Table 1.10 is analyzed, which displays the proportion of replications where the respective factor is

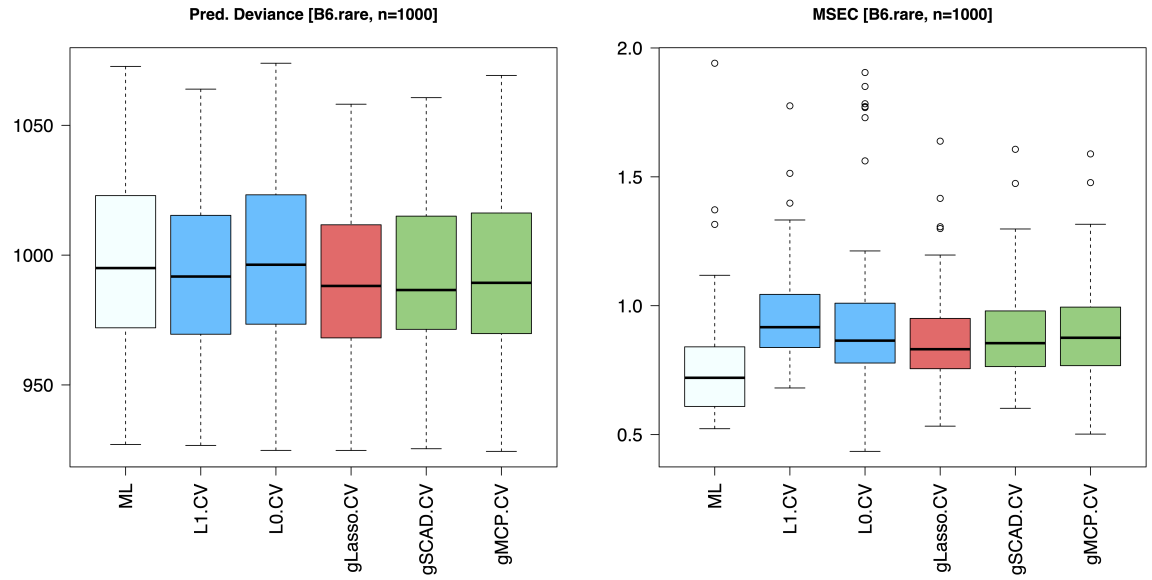

Figure 1.16: *Predictive deviance and mean squared error of coefficients (MSEC) for design B6.rare,* $n = 1000$.

	included factors		
	$\mathscr{X}_3$	$\mathscr{X}_1, \mathscr{X}_2$	$\mathscr{X}_1, \mathscr{X}_2, \mathscr{X}_3$
`ML`	1.00	1.00	1.00
`L1.CV`	1.00	0.95	0.95
`L0.CV`	0.93	0.68	0.61
`gLasso.CV`	1.00	0.90	0.90
`gSCAD.CV`	0.99	0.57	0.56
`gMCP.CV`	0.98	0.56	0.54

Table 1.10: *[B6.inter.pos, n=1000] Proportion of replications where* $\mathscr{X}_3$ *is included (left column),* $\mathscr{X}_1, \mathscr{X}_2$ *are included (middle column) and* $\mathscr{X}_1, \mathscr{X}, \mathscr{X}_3$ *are included (right column).*

included in the model. Here, methods are preferable where the *difference* of the values in the middle and the right column is small, that is, *if* the method includes $\mathscr{X}_1$ and $\mathscr{X}_2$ *then* factor $\mathscr{X}_3$ is further included. For L_0, the largest difference in the percentage of replications where $\mathscr{X}_1$ and $\mathscr{X}_2$ are included and where all three $\mathscr{X}_1, \mathscr{X}_2$ and $\mathscr{X}_3$ are included is noticed, where this difference is given by 7%. For L_1 and group lasso, the values in the middle column and in the right column of Table 1.10 are the same, meaning that if $\mathscr{X}_1$ and $\mathscr{X}_3$ are included in the model, the interaction term $\mathscr{X}_3 = \mathscr{X}_1 \cdot \mathscr{X}_2$ is also (correctly) included in the model. For group SCAD and group MCP, the detection of the interaction term seems to work better than for L_0, since the difference in proportion of replications detecting $\mathscr{X}_1, \mathscr{X}_2$ as influential and all three $\mathscr{X}_1, \mathscr{X}_2, \mathscr{X}_3$ is around 1% - 2%. With respect to these measures, L_1 and group lasso would be preferable, which further holds for the MSEC values displayed in Figure 1.17.

Now the FP/FN rates displayed in Table 1.12 are investigated. It is noted that in this design, no fusion of truly influential factors need to be performed, thus no fusion tasks occur such that

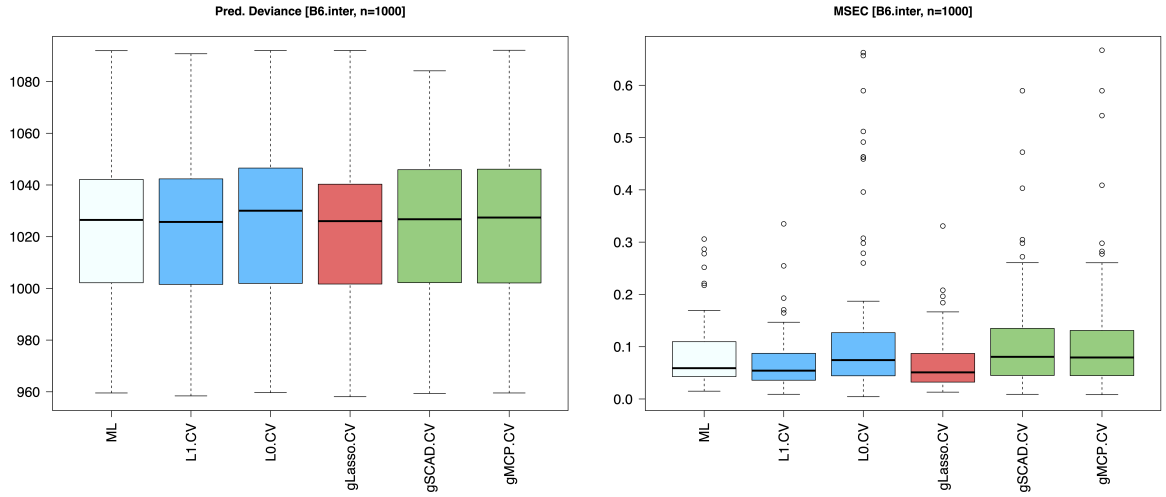

Figure 1.17: *Predictive deviance and mean squared error of coefficients (MSEC) for design B6.inter.pos, $n = 1000$.*

fusion measures are neglected. However, in terms of selection rates $FP_{s,\text{fac}}$ and $FN_{s,\text{fac}}$, one can notice that the model selected by L_1 is less sparse than the others, which is underlined by the sparsity measures OS/PS provided in Table 1.11.

	ML	L1.CV	L0.CV	gLasso.CV	gSCAD.CV	gMCP.CV
OS	9.00	8.22	6.70	7.71	5.97	5.32
PS	6.00	5.64	4.72	5.23	4.15	3.85

Table 1.11: *[B6.inter, n=1000] Overall Sparsity (OS) and Practical Sparsity (PS), true values $OS^* = 5$ and $PS^* = 4$.*

Group SCAD and group MCP select the most sparse models, however, they also show the highest values for $FN_{s,\text{fac}}$ in Table 1.12, thus this comes with some false selections. Even though the difference in performance of the methods is not that distinct in this design, group lasso may be ranked as best performing, since it balances sparsity, as well as MSEC and predictive deviance and, finally, it includes the influential variables $\mathscr{X}_1$ and $\mathscr{X}_2$ as well as their interaction $\mathscr{X}_3$ in around 90% of replications.

	ML	L1.CV	L0.CV	gLasso.CV	gSCAD.CV	gMCP.CV
$FP_{s,\text{fac}}$	1.00	0.84	0.62	0.68	0.41	0.33
$FN_{s,\text{fac}}$	0.00	0.01	0.13	0.03	0.17	0.20
$FP_{f,\text{infl.truth}}$						
$FN_{f,\text{infl.truth}}$	0.00	0.03	0.18	0.03	0.15	0.19

Table 1.12: *[B6.inter, n=1000] FP/FN rates fusion and factor selection.*

1.8.5 Results of Design highdim

At last, the results of the high-dimensional design are discussed. Here, the approaches ML, CAS-L_0 and CAS-L_1 failed in 100% of the replications with the chosen computational approach, which can be seen in Table 1.13. Thus, the results of group lasso, group SCAD and group MCP are analyzed, while ML, L_0 and L_1 are neglected.

	Proportion
ML	1.00
L1.CV	1.00
L0.CV	1.00
gLasso.CV	0.00
gSCAD.CV	0.00
gMCP.CV	0.00

Table 1.13: *[highdim] Proportion of replications where the methods failed.*

First, the values of OS/PS obtained in Table 1.14 are investigated. As given in this table, the true OS is given by $\text{OS}^* = 15$ and the true PS is given by $\text{PS}^* = 5$, thus group SCAD is the closest to the truth. The group MCP approach yields a too sparse model in this design, while the models selected by group lasso seem to be too large.

	gLasso.CV	gSCAD.CV	gMCP.CV
OS	36.98	15.08	6.84
PS	13.00	5.22	2.31

Table 1.14: *[highdim, n=100] Overall Sparsity (OS) and Practical Sparsity (PS), true values $OS^* = 15$ and $PS^* = 5$.*

The FP and FN rates in Table 1.15 underline the observations made concerning OS and PS. The $FP_{s,\text{fac}}$ rate for group MCP is nearly zero and the $FN_{s,\text{fac}}$ rate is the highest among all observed methods which underlines the fact that group MCP produces the most sparse models, which are too sparse for this design, the same applies for group SCAD. These rates of group MCP show that, approximately, none of the truly zero coefficients are set to nonzero ($FP_{s,\text{fac}}$) while approximately 60% of the truly nonzero coefficients are falsely set to zero ($FN_{s,\text{fac}}$). Group lasso shows the lowest $FN_{s,\text{fac}}$ value, at the cost of higher $FP_{s,\text{fac}}$, since the selected model is less sparse. However, especially when one is interested in two-step procedures, a method may be favored where the selected model *could* be sparser, but not too many truly influential factors are excluded, hence group lasso seems to be a reasonable choice.

	gLasso.CV	gSCAD.CV	gMCP.CV
$FP_{s,\text{fac}}$	0.17	0.04	0.00
$FN_{s,\text{fac}}$	0.27	0.43	0.59
$FP_{f,\text{infl.truth}}$	0.62	0.31	0.10
$FN_{f,\text{infl.truth}}$	0.26	0.41	0.56

Table 1.15: *[highdim, n=100] FP/FN rates fusion and factor selection.*

Looking at Figure 1.18, it can be noticed that all methods perform similarly in terms of MSEC

and predictive deviance for this design. To sum up, in this high-dimensional design the group lasso procedure is preferable.

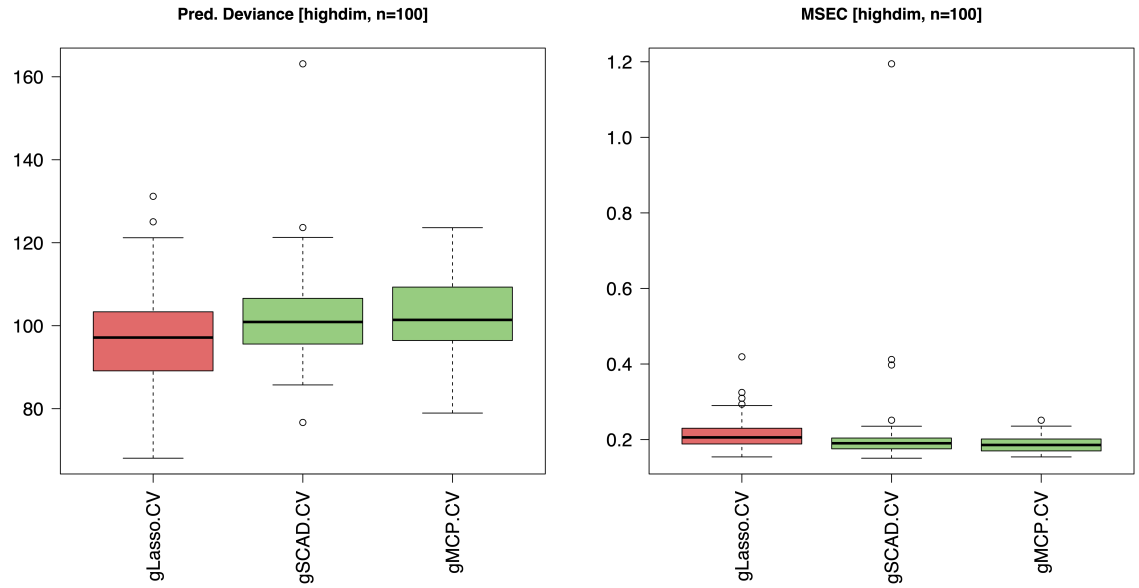

Figure 1.18: *Predictive deviance and mean squared error of coefficients (MSEC) for design highdim,* $n = 100$.

1.9 Conclusion

The previous sections gave an overview of methods for executing variable/factor selection and levels fusion in a framework of categorical explanatory variables with an underlying logistic regression model. Theoretical properties were investigated, as well as characteristics coming from the penalty function (e.g. convexity, non-convexity, shrinkage properties etc.) and computational methods. Finally, their performance was discussed in five different simulation designs.

The simulation studies conducted in this thesis showed that comparing the fusion-type penalties L_0 and L_1, the approach using the L_0 norm yields more sparse models, while the resulting models with L_1 were often not sparse enough. Further, compared to L_1, the L_0 approach does not shrink the difference of the coefficients (and the coefficients themselves) towards zero, which comes from the construction of the L_0 norm differentiating between zero and nonzero. However, this fact may prevent the need of using adaptive weights. Consequently, for the fusion-type penalties, one would prefer L_0 for the purposes of this thesis. Nevertheless, the computation with PIRLS offered some convergence issues in the high-dimensional design, such that there is a need for alternative computation methods for high-dimensions. Furthermore, it is of high interest to investigate whether the selection performance of L_0 could be improved, since L_0 only performs implicit selection including reference category zero in the differences, which does not respect groupwise structures.

The groupwise approaches that were investigated, group lasso, group SCAD and group MCP showed the most satisfactory factor selection performance, but they are not able to fuse levels of the same factor having a similar influence on the random response variable. However, their factor selection performance outperformed the factor selection performance of the fusion-type penalties L_1 and L_0. Among the groupwise approaches, it was noticed that group lasso produces less sparse models than group SCAD and group MCP, where group MCP produces the most sparse models among these considered approaches. Because of the theoretical nature of group lasso, it shrinks the coefficients more towards zero than group SCAD and group MCP, which caused that group lasso showed the lowest MSEC in the majority of the designs.

To increase the sparsity level of group lasso through levels fusion, it is proposed to combine the L_0 fusion penalty with the group lasso penalty for factor selection, which is introduced in the following Chapter 2. Group MCP or group SCAD were not chosen for the factor selection part for several reasons. First, these type of penalties include an additional tuning parameter γ, which would cause more involving tuning tasks as the new proposed penalty function of Chapter 2 already includes two tuning parameters λ_0 and λ_1, where one tuning parameter is needed for the fusion tasks, and one for the selection tasks. Thus, it is straightforward that it is desirable to avoid three tuning parameters. Second, the optimization problem coming from the L_0 approach is not convex, thus it is desirable to avoid another penalty that is not convex, e.g. group SCAD or group MCP. Third, for the group lasso penalty, in the simulation studies presented here, it was inferred that the MSEC was the lowest among the considered methods. Lastly, for group SCAD and group MCP, it was observed that there is a gap in the existing literature investigating theoretical properties for logistic regression models, hence the theoretical basis for this setting is missing. This justifies the idea of introducing a new penalty function combining L_0 and group lasso, which is introduced in the following chapter.

Chapter 2

Introduction and Theoretical Properties of L_0-Fused Group Lasso (L_0-FGL)

As the previous chapter gave an overview of existing penalty functions to perform penalized logistic regression, as well as the extension of some theoretical properties from linear to logistic regression, this chapter provides the introduction of a new penalized regression method. Motivated from the drawbacks of particular methods elaborated in the simulation studies in the previous chapter and some theoretical gaps, a new penalty function is proposed here. To be more precise, the penalty function to be introduced is called L_0-*Fused Group Lasso* (L_0-FGL), for the purpose of combining the ability of factor selection as well as levels fusion. To do so, an L_0 fusion-type penalty for levels fusion is combined with a group lasso penalty for factor selection. Besides L_0-FGL, the corresponding adaptive variant incorporating adaptive weights, the *adaptive* L_0-*FGL* is further introduced. As demonstrated in Section 1.3.2, using solely CAS-L_0 cannot ensure asymptotic normality, however, adding a group lasso penalty enables the asymptotic normality property for L_0-FGL, shown in the subsequent analyses. One of the main challenges is to show appropriate theoretical properties, including properties of the fusion performance, since including an L_0 norm causes that the penalty function is not continuous. Further, appropriate computational approaches are required, especially accounting for the fact that with two penalty functions, two tuning parameters are incorporated.

The chapter is organized as follows. Section 2.1 gives the motivation and intuition behind L_0-FGL, as well as the detailed structure of the penalty function and the resulting objective function to be minimized. Further, the penalty function is graphically illustrated and the impact of the tuning parameters is underlined. Having that, the investigation of the existence of L_0-FGL is exhibited in Section 2.2, further showing that the objective function decreases if two coefficients that are close enough to each other are fused, thus set to be equal. In Section 2.3, asymptotic properties of L_0-FGL are investigated, including estimation consistency results, asymptotic distributions as well as factor selection and levels fusion consistency.

It is noted that some earlier version of the content of Sections 2.1, 2.2 and 2.3 are partially included in Kaufmann and Kateri (2024) published in *Electronic Journal of Statistics*.

2.1 Motivation and Penalty Function

For the simultaneous performance of factor selection and fusion of levels corresponding to the same factor, the so-called L_0-*Fused Group Lasso*, abbreviated as L_0-FGL, is introduced, which is defined through the following penalty function

$$P_\lambda^{(L_0\text{-FGL})}(\boldsymbol{\beta}) := \lambda_1 \sum_{j=1}^{J} ||\boldsymbol{\beta}_j||_{\boldsymbol{K}_j} + \lambda_0 \sum_{j=1}^{J} \sum_{0 \leq r < s \leq p_j} w_0^{(j,rs)} ||\beta_{j,r} - \beta_{j,s}||_0. \tag{2.1}$$

Here, $\lambda := (\lambda_1, \lambda_0) \in \mathbb{R}^{\geq 0} \times \mathbb{R}^{\geq 0}$ is some (two-dimensional) tuning parameter and $\boldsymbol{K}_j$, $j \in \{1, \ldots, J\}$ are positive definite matrices. By the nature of the penalty function, one expects that L_0-FGL is a suitable choice for penalized regression tasks corresponding to discrete structures, that is, incorporating categorical explanatory variables. As introduced in Chapter 1, the focus of this thesis is logistic regression with factors, however, L_0-FGL can be further used for other underlying model schemes, e.g. in a probit model, as well as considering a mixture of categorical and continuous explanatory variables.

The penalty function (2.1) is an intersection between the group lasso penalty (Section 1.4.1) and the L_0 fusion penalty CAS-L_0 (Section 1.3.2). For $j \in \{1, \ldots, J\}$, it is supposed that some weight is given by $w_1^{(j)}$, which is specified below. Corresponding to the choice of the positive definite matrices $\boldsymbol{K}_j$, with $\boldsymbol{K}_j := \tilde{w}_1^{(j)} \mathbf{I}_{p_j \times p_j}$ and $\sqrt{\tilde{w}_1^{(j)}} =: w_1^{(j)}$, the penalty function (2.1) is of the following form

$$P_\lambda^{(L_0\text{-FGL})}(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{J} w_1^{(j)} ||\boldsymbol{\beta}_j||_2 + \lambda_0 \sum_{j=1}^{J} \sum_{0 \leq r < s \leq p_j} w_0^{(j,rs)} ||\beta_{j,r} - \beta_{j,s}||_0. \tag{2.2}$$

In particular, using the convenient choice $\boldsymbol{K}_j := p_j \mathbf{I}_{p_j \times p_j}$ (i.e. $\tilde{w}_1^{(j)} = p_j$) results in $w_1^{(j)} = \sqrt{p_j}$. These type of (non-adaptive) weights $w_1^{(j)}$ compensate for the fact that the number of levels of the factors may be distributed inhomogeneously. The weights for the L_0-part, namely $w_0^{(j,rs)}$ for $j \in \{1, \ldots, J\}$ and $r, s \in \{1, \ldots, p_j\}$, $r \neq s$, are discussed later.

The use of adaptive weights leads to the so-called *adaptive L_0-FGL*. In the following, the latter choice of $\boldsymbol{K}_j$ is used, that is $\boldsymbol{K}_j := p_j \mathbf{I}_{p_j \times p_j}$, and theoretical properties both for the L_0-FGL and the adaptive version are investigated.

The L_0-FGL penalty function includes an L_0 fusion-type penalty. From Chapter 1, it is known that this fusion penalty is convenient for the analyses of categorical explanatory variables, where including the reference category chosen to be zero, *indirect* selection is enforced. However, to *directly* enforce *factor* selection performance, a group lasso penalty is incorporated in the construction of L_0-FGL.

The L_0-FGL estimate $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$ is defined as a minimizer of the following objective function

$$\begin{aligned} M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta}) &:= -L_n(\boldsymbol{\beta}) + P_\lambda^{(L_0\text{-FGL})}(\boldsymbol{\beta}) \\ &= -L_n(\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{J} w_1^{(j)} ||\boldsymbol{\beta}_j||_2 + \lambda_0 \sum_{j=1}^{J} \sum_{0 \leq r < s \leq p_j} w_0^{(j,rs)} ||\beta_{j,r} - \beta_{j,s}||_0. \end{aligned} \tag{2.3}$$

The following remark discusses the question of local and global minimizers corresponding to the minimization of $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$.

Remark 2.1.1 (Local and global minimizers (L_0-FGL))**.** Caused by the incorporation of the L_0 norm in L_0-FGL, which is not even a continuous function (e.g. $||t||_0$ is not continuous in $t = 0$), the resulting optimization problem minimizing $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ is neither convex nor concave. The fact that the penalty function is not convex can be seen for example in Figure 2.1. That is, depending on the negative log-likelihood function $-L_n(\boldsymbol{\beta})$ which is added to the penalty function $P_\lambda^{(L_0\text{-FGL})}(\boldsymbol{\beta})$, several local minimizers of $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ may exist. Moreover, it *cannot* be ensured in general that a global minimizer of $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ is unique. Consequently,

the algorithms that are obtained as well as the theoretical properties throughout this thesis are about *local* minimizers. This comes from the natural structure of the objective function and is in line with other works investigating non-convex penalty functions, such as Fan and Li (2001), Fan and Peng (2004), Guo et al. (2015), Kim et al. (2008), Fan and Lv (2010).

The L_0-FGL estimate $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$ is defined as a minimizer of $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$, thus there may exist multiple estimates. However, to avoid notational complexity, one writes

$$\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\arg\min}\, M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$$

keeping in mind that *local and global* minimizers are allowed.

Figure 2.1 shows the value of the penalty function $P_{\lambda}^{(L_0\text{-FGL})}(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||_2 + ||\beta_1 - \beta_2||_0$ for $\beta_1, \beta_2 \in [-2, 2]$. Thus, the penalty tuning parameters were chosen to be equal $\lambda_0 = \lambda_1 = 1$. It becomes clear that L_0-FGL combines shrinkage, factor selection and fusion of levels in one penalty since the value of the penalty function decreases (i) if the absolute values of β_1 and β_2 decrease, (ii) if β_1 and β_2 are equal and (iii) if $\beta_1 = \beta_2 = 0$. In particular, (i) refers to the shrinkage property, (ii) to levels fusion and (iii) to factor selection.

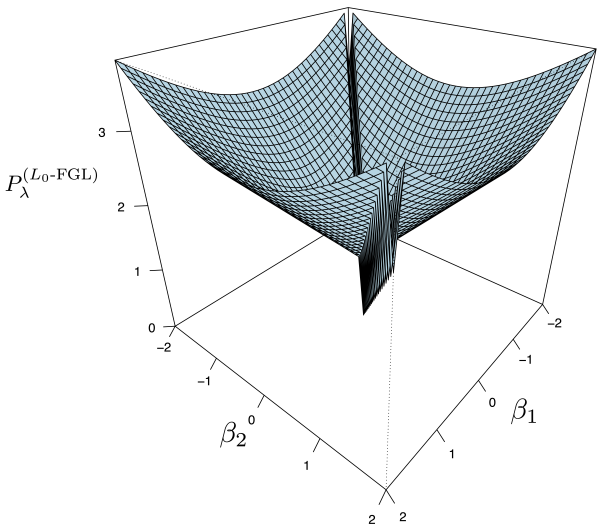

Figure 2.1: $||\boldsymbol{\beta}||_2 + ||\beta_1 - \beta_2||_0$ *is displayed for* $\beta_1, \beta_2 \in [-2, 2]$ *and* $\lambda_0 = \lambda_1 = 1$.

By varying the tuning parameters λ_0 and λ_1, the focus can either be set on selection or on fusion depending on the application context, where reference is made to Figure 2.2 being similar to Figure 2.1, where different tuning parameters, i.e. $\lambda_1 \neq \lambda_0$, are applied in the first figure. That is, on the left hand side of Figure 2.2, $\lambda_1 = 1$ and $\lambda_0 = 5$ are chosen, hence the focus is on levels fusion. On the right hand side, $\lambda_1 = 5$ and $\lambda_0 = 1$ are chosen, so the focus is on factor selection. This shows the flexibility of L_0-FGL in different applications and settings.

2.1.1 Tuning Methods

For the application of the L_0-FGL penalty function (2.1) in practice, one needs to determine the best pair $\lambda = (\lambda_0, \lambda_1) \in \mathbb{R}^{\geq 0} \times \mathbb{R}^{\geq 0}$ of the two-dimensional tuning parameter. For this, two different approaches to determine the best tuning parameter pair denoted by $\lambda^* = (\lambda_0^*, \lambda_1^*)$ are considered, a stepwise and an iterative approach.

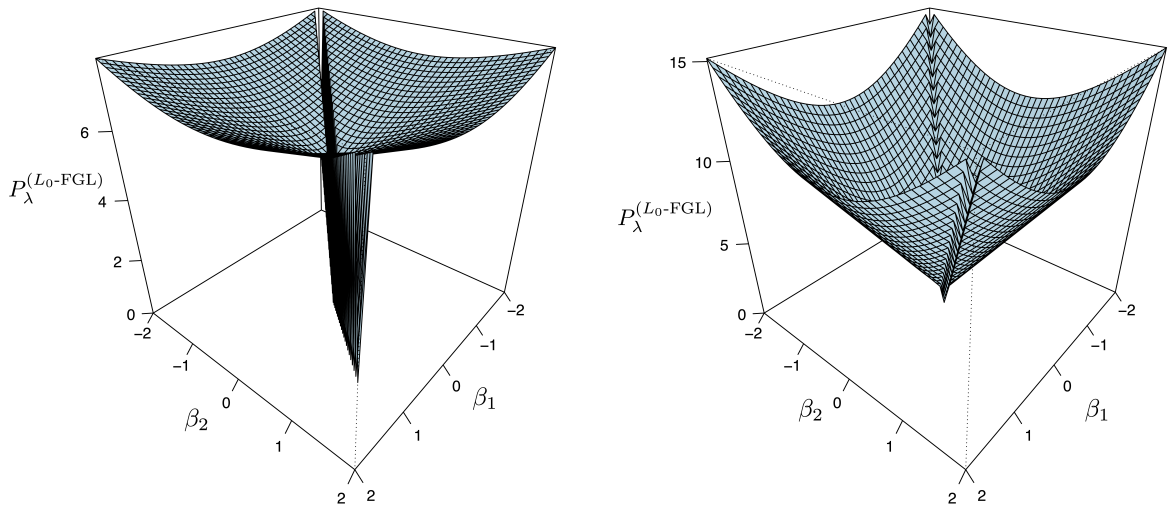

Figure 2.2: $||\boldsymbol{\beta}||_2 + ||\beta_1 - \beta_2||_0$ *is displayed for* $\beta_1, \beta_2 \in [-2, 2]$. *Left hand side:* $\lambda_0 = 1$, $\lambda_0 = 5$. *Right hand side:* $\lambda_1 = 5$, $\lambda_0 = 1$.

(i) *Stepwise* approach. In the stepwise approach, one proceeds as follows.

(a) $\lambda_0 := 0$ is fixed and the best tuning parameter for λ_1, denoted by λ_1^*, is determined applying (one-dimensional) CV (Section 1.2.4),

(b) $\lambda_1 := \lambda_1^*$ is fixed and the best tuning parameter for λ_0, denoted by λ_0^*, is determined applying (one-dimensional) CV (Section 1.2.4).

(ii) *Iterative* approach. In the iterative approach, one performs the procedure described above first, that is, (i)(a) and (i)(b), while after (i)(b) $\lambda_0 := \lambda_0^*$ is fixed and the optimal λ_1^* is determined with (one-dimensional) CV. In the sequel, an analogue procedure is followed for λ_0, fixing $\lambda_1 := \lambda_1^*$. This procedure is iterated until a pre-specified number of iterations is reached, or the improvement in predictive deviance does not exceed a pre-specified tolerance.

One could ask for performing CV on a two-dimensional grid to determine $\lambda^* = (\lambda_0^*, \lambda_1^*)$, hence two-dimensional CV, but this clearly results in a strongly computational intensive method. Thus, the techniques introduced above are applied here to avoid increasing the computational time. The reason that, for both approaches (i.e. stepwise and iterative), one starts with the group lasso part (tuning of λ_1) is caused by the fact that this saves computational time since once a factor is excluded from the model, one does not need to investigate this factor further for fusion of levels.

2.2 Existence

This section starts with a theorem about the existence of an L_0-FGL estimate, that is, the existence of a (local or global) minimum of the objective function. Further, it is shown that the objective function decreases if coefficients that are "close enough" to each other are fused.

What is meant by "close enough" is specified later in this thesis (Theorem 2.3.15), providing the existence of a fusion threshold. Moreover, the existence in the high-dimensional case where $p > n$ is provided in Remark 2.2.2 in the sequel.

Theorem 2.2.1 (Existence of L_0-FGL, Kaufmann and Kateri (2024))**.** One assumes that $\lambda_1 > 0$, $\lambda_0 \geq 0$ and $0 < \sum_{i=1}^{n} y_i < n$. Then, the set

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

is nonempty. Moreover, the value of the objective function $M_{pen}^{(L_0\text{-FGL})}(\cdot)$ decreases if coefficients that are close enough to each other are fused.

Proof.
(1) $S \neq \emptyset$**:** One sets $J = 1$, the proof for $J > 1$ works analogously. It is shown that for $J = 1$, the group lasso estimator, given by

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

fulfills $\hat{\boldsymbol{\beta}}^{(\mathrm{GL})} \in S$. By assumption, $0 < \sum_{i=1}^{n} y_i < n$ and by Meier et al. (2008) (Lemma 1) it can be concluded that the group lasso estimator $\hat{\boldsymbol{\beta}}^{(\mathrm{GL})}$ exists. It is noted that, for the group lasso estimator to be unique, further requirements are needed which are not imposed here, compare Corollary 1.4.2. However, by definition of $\hat{\boldsymbol{\beta}}^{(\mathrm{GL})}$, there exists an $\boldsymbol{\varepsilon}$-neighborhood of $\hat{\boldsymbol{\beta}}^{(\mathrm{GL})}$, where $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_p) \in \mathbb{R}^p$, such that $\hat{\boldsymbol{\beta}}^{(\mathrm{GL})}$ minimizes the sum $-L_n(\cdot) + \lambda_1 ||\cdot||_K$, that is

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

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

$$\begin{aligned}
&M_{pen}^{(L_0\text{-FGL})}(\hat{\boldsymbol{\beta}}^{(\mathrm{GL})} + \boldsymbol{\varepsilon}) \\
&= -L_n(\hat{\boldsymbol{\beta}}^{(\mathrm{GL})} + \boldsymbol{\varepsilon}) + \lambda_1 ||\hat{\boldsymbol{\beta}}^{(\mathrm{GL})} + \boldsymbol{\varepsilon}||_K + \lambda_0 \sum_{r,s} w_0^{(rs)} ||\hat{\beta}_r^{(\mathrm{GL})} - \hat{\beta}_s^{(\mathrm{GL})} + \varepsilon_r - \varepsilon_s||_0 \\
&\geq -L_n(\hat{\boldsymbol{\beta}}^{(\mathrm{GL})}) + \lambda_1 ||\hat{\boldsymbol{\beta}}^{(\mathrm{GL})}||_K + \lambda_0 \sum_{r,s} w_0^{(rs)} ||\hat{\beta}_r^{(\mathrm{GL})} - \hat{\beta}_s^{(\mathrm{GL})} + \varepsilon_r - \varepsilon_s||_0. \qquad (2.4)
\end{aligned}$$

For the group lasso estimate it either holds that $\hat{\boldsymbol{\beta}}^{(\mathrm{GL})} = \mathbf{0}$ or $\hat{\beta}_r^{(\mathrm{GL})} \neq 0 \, \forall r$, according to Meier et al. (2008). In the latter case it is further known that $\hat{\beta}_r^{(\mathrm{GL})} \neq \hat{\beta}_s^{(\mathrm{GL})} \; \forall \, r, s$ almost surely. For the case that $\hat{\beta}_r^{(\mathrm{GL})} \neq \hat{\beta}_s^{(\mathrm{GL})} \; \forall \, r, s$ one can choose $\boldsymbol{\varepsilon}$ small enough such that

$$\hat{\beta}_r^{(\mathrm{GL})} + \varepsilon_r \neq \hat{\beta}_s^{(\mathrm{GL})} + \varepsilon_s \; \forall \, r, s.$$

Consequently, one concludes that the L_0 norms of $\hat{\beta}_r^{(\mathrm{GL})} - \hat{\beta}_s^{(\mathrm{GL})}$ and $\hat{\beta}_r^{(\mathrm{GL})} - \hat{\beta}_s^{(\mathrm{GL})} + \varepsilon_r - \varepsilon_s$ coincide since all values of the differences are nonzero, i.e.

$$||\hat{\beta}_r^{(\mathrm{GL})} - \hat{\beta}_s^{(\mathrm{GL})} + \varepsilon_r - \varepsilon_s||_0 = ||\hat{\beta}_r^{(\mathrm{GL})} - \hat{\beta}_s^{(\mathrm{GL})}||_0.$$

Then one obtains that (2.4) equals

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

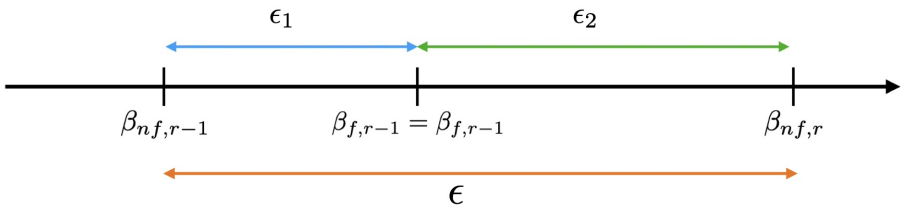

Figure 2.3: *Location of* $\beta_{nf,r}$, $\beta_{nf,r-1}$ *and the fused coefficients* $\beta_{f,r} = \beta_{f,r-1}$ *for the case* $\min\{\beta_{nf,r}, \beta_{nf,r-1}\} = \beta_{nf,r-1}$.

and thus $M_{pen}^{(L_0\text{-FGL})}(\hat{\boldsymbol{\beta}}^{(\text{GL})} + \boldsymbol{\varepsilon}) \geq M_{pen}^{(L_0\text{-FGL})}(\hat{\boldsymbol{\beta}}^{(\text{GL})})$ for a sufficiently small $\boldsymbol{\varepsilon}$. If $\hat{\boldsymbol{\beta}}^{(\text{GL})} = \mathbf{0}$, by applying the same arguments as above, it can be inferred

$$\begin{aligned} M_{pen}^{(L_0\text{-FGL})}(\hat{\boldsymbol{\beta}}^{(\text{GL})} + \boldsymbol{\varepsilon}) &\geq -L_n(\mathbf{0}) + \lambda_1 ||\mathbf{0}||_K + \lambda_0 \sum_{r,s} w_0^{(rs)} \underbrace{||\varepsilon_r - \varepsilon_s||_0}_{\geq 0} \\ &\geq -L_n(\mathbf{0}) + \lambda_1 ||\mathbf{0}||_K = M_{pen}^{(L_0\text{-FGL})}(\mathbf{0}) = M_{pen}^{(L_0\text{-FGL})}(\hat{\boldsymbol{\beta}}^{(\text{GL})}) \end{aligned}$$

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

(2) $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ decreases if coefficients that are close enough to each other are fused: Again, one assumes that $J = 1$ and one starts with the case of an ordinal factor comparing adjacent categories for fusion. The goal is to show that the objective function $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ decreases if coefficients that are close enough to each other are fused. It is noted that, since the reference category is chosen to be zero, there is no appearance of the reference category in the coefficient vector $\boldsymbol{\beta}$. One writes $\boldsymbol{\beta}_{nf}$ (not fused), $\boldsymbol{\beta}_f$ (fused) $\in \mathbb{R}^p$ with

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

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.
It is noted 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, it is assumed that $\min\{\beta_{nf,r}, \beta_{nf,r-1}\} = \beta_{nf,r-1}$. Since an ordinal factor is observed, this holds by definition, but when observing nominal factors one has to differentiate between these two cases. However, 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 (not to be mixed up with $\boldsymbol{\varepsilon}$) and $\beta_{nf,r} = \beta_{nf,r-1} + \epsilon$, one consults Figure 2.3 for a visualization of the setting. Now the goal is to show $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta}_f) < M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta}_{nf})$. Clearly, it depends on the design and the tuning etc. how small ϵ has to be such that the objective function decreases, which is further specified in the proof of Theorem 2.3.15. By construction one can write $\boldsymbol{\beta}_{nf} - \boldsymbol{\beta}_f = (0, ..., 0, -\epsilon_1, \epsilon_2, 0, ..., 0)$. 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

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

Because of the definition of $\boldsymbol{\beta}_{nf}$ (no categories fused) and $\boldsymbol{\beta}_f$ (category r and $r-1$ fused) it holds 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 $\sum_{i=1}^{p} w_0^{(i)} ||\beta_{f,i} - \beta_{f,i-1}||_0 = c - w_0^{(r)}$. 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 $||\boldsymbol{\beta}_{nf} - \boldsymbol{\beta}_f|| = \epsilon < \min\{\tilde{\epsilon}_1, \tilde{\epsilon}_2\}$ it can be inferred

$$
\begin{aligned}
& M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta}_{nf}) - M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta}_f) \\
&= -L_n(\boldsymbol{\beta}_{nf}) + \lambda_1 ||\boldsymbol{\beta}_{nf}||_{\boldsymbol{K}} + \lambda_0 \sum_{i=1}^{p} w_0^{(i)} ||\beta_{nf,i} - \beta_{nf,i-1}||_0 \\
&\quad + L_n(\boldsymbol{\beta}_f) - \lambda_1 ||\boldsymbol{\beta}_f||_{\boldsymbol{K}} - \lambda_0 \sum_{i=1}^{p} w_0^{(i)} ||\beta_{f,i} - \beta_{f,i-1}||_0 \\
&= -L_n(\boldsymbol{\beta}_{nf}) + \lambda_1 ||\boldsymbol{\beta}_{nf}||_{\boldsymbol{K}} + L_n(\boldsymbol{\beta}_f) - \lambda_1 ||\boldsymbol{\beta}_f||_{\boldsymbol{K}} + \lambda_0 \cdot w_0^{(r)} \\
&> -\delta_1 - \lambda_1 \delta_2 + \lambda_0 \cdot w_0^{(r)}.
\end{aligned} \tag{2.5}
$$

Now, if λ_0 (tuning for fusion) is chosen large enough and δ_1, δ_2 are chosen small enough such that $\lambda_0 \cdot w_0^{(r)} > \delta_1 + \lambda_1 \delta_2$, it can be argued from the above equation that

$$
M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta}_{nf}) - M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta}_f) > 0 \Leftrightarrow M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta}_{nf}) > M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta}_f).
$$

Consequently the value of the objective function $M_{pen}^{(L_0\text{-FGL})}(\cdot)$ in $\boldsymbol{\beta}_f$ is less than in $\boldsymbol{\beta}_{nf}$, hence the objective function decreases if coefficients that are close enough to each other are fused. The proof can directly be extended to the case where more categories are fused 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. □

Remark 2.2.2 (Existence in high-dimensional case where $p > n$)**.** For the L_0-FGL estimate with $\lambda_1 > 0$ and $\lambda_0 \geq 0$, the proof above is not restricted to the case $p \leq n$, hence existence can be ensured in the high-dimensional case where $p > n$, based on the existence of group lasso shown in Meier et al. (2008) and used in the proof above. Allowing for $\lambda_1 = 0$ or both $\lambda_1 = \lambda_0 = 0$, the existence of L_0-FGL corresponds to the existence of the maximum likelihood estimator (MLE) or CAS-L_0, respectively.

2.3 Asymptotic Properties

Asymptotic results for L_0-FGL are provided in the following, where both cases of p being fixed and p_n being allowed to diverge with n are considered. In particular, $\sqrt{n}$ consistency properties are shown first (Section 2.3.1), then asymptotic normality results are provided (Section 2.3.2) before factor selection consistency is exhibited (Section 2.3.3). Finally, fusion properties of L_0-FGL are obtained (Section 2.3.4).

2.3.1 $\sqrt{n}$ Consistency

First, the following lemma is provided being necessary for technical steps in the proof of Theorem 2.3.2. To be more precise, it is shown that if the objective function $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ satisfies that its functional value on the boundary of some $\mathbb{R}^p$ ball $\mathscr{D} := \{\boldsymbol{x}^* + \boldsymbol{u} \mid ||\boldsymbol{u}||_2 \leq c\}$ with center $\boldsymbol{x}^* \in \mathbb{R}^p$ and radius $c \in \mathbb{R}^{>0}$ is greater than its functional value in the center $\boldsymbol{x}^*$, then there exists a local minimum *inside* the $\mathbb{R}^p$ ball. Obviously, if the function $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ was continuous, the statement of the lemma would be clear since in a compact set a continuous function attains its minimum and maximum, hence in $\mathscr{D}$, and then one could use the assumption of the lemma

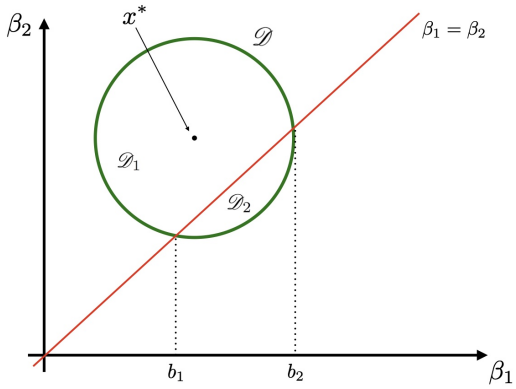

Figure 2.4: *Partition of the ball $\mathscr{D}$ into $\mathscr{D}_1$ and $\mathscr{D}_2$. The red line shows the 1-dimensional hyperplane where $f(\boldsymbol{\beta})$ is not continuous, hence $\beta_1 = \beta_2$.*

(i.e. (2.6) below) to show that the infimum (minimum) is not attained at the boundary of $\mathscr{D}$. However, to proof the following lemma, the explicit structure of $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$, especially the characteristics of the discontinuities, is crucial.

Lemma 2.3.1 (Kaufmann and Kateri (2024))**.** $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ is the objective function of L_0-FGL, given in (2.3). One assumes that for some $\boldsymbol{x}^* \in \mathbb{R}^p$ and $c \in \mathbb{R}^{>0}$ it holds

$$\inf_{||\boldsymbol{u}||_2=c} M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{x}^* + \boldsymbol{u}) > M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{x}^*). \tag{2.6}$$

Then, there exists at least one local minimum of $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ inside $\mathscr{D} := \{\boldsymbol{x}^* + \boldsymbol{u} \,|\, ||\boldsymbol{u}||_2 \leq c\}$, where inside means in the domain $\mathring{\mathscr{D}} = \{\boldsymbol{x}^* + \boldsymbol{u} \,|\, ||\boldsymbol{u}||_2 < c\}$.

Proof. For simplicity of notation, one writes $M_{pen}(\cdot)$ instead of $M_{pen}^{(L_0\text{-FGL})}(\cdot)$ throughout the following proof. Since the intercept is not penalized and only appears in the log-likelihood, it is neglected, i.e. $M_{pen}(\boldsymbol{\beta})$ is observed for $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{x}^* \in \mathbb{R}^p$ (instead of $\mathbb{R}^{p+1}$).

Now, it is shown that M_{pen} attains its infimum (i.e. minimum) in $\mathscr{D}$. Having that, (2.6) is used to show that the infimum is not attained at the boundary, hence it is in $\mathring{\mathscr{D}}$. The proof is conducted for the case $p = 2$ and $J = 1$, cases of higher dimensions work in a similar manner, although more possible cases arise for the infimum to occur, as specified below. The ball $\mathscr{D}$ is partitioned into two subsets in the following way

$$\begin{aligned} \mathscr{D}_1 &:= \mathscr{D} \backslash \{\boldsymbol{\beta} = (\beta_1, \beta_2) \,:\, \beta_1 < \beta_2\}, \\ \mathscr{D}_2 &:= \mathscr{D} \backslash \{\boldsymbol{\beta} = (\beta_1, \beta_2) \,:\, \beta_2 < \beta_1\}. \end{aligned}$$

So the hyperplane satisfying $\beta_1 = \beta_2$ is included in both subsets. Obviously it holds that $\mathscr{D} = \mathscr{D}_1 \cup \mathscr{D}_2$. This partition is displayed in Figure 2.4.

By definition, the objective function can be written as

$$M_{pen}(\boldsymbol{\beta}) = g(\boldsymbol{\beta}) + f(\boldsymbol{\beta}),$$

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

$$f(\boldsymbol{\beta}) := \lambda_0 w_0 ||\beta_1 - \beta_2||_0$$

is the L_0 part. By definition of the L_0 norm applied to differences, this norm is equal to zero if the object on which the norm is applied is zero and one otherwise, hence it is zero if the difference is zero and it is one if the difference is nonzero. Thus, with the considered partition, $f(\boldsymbol{\beta})$ is not continuous on the one-dimensional hyperplane $\beta_1 = \beta_2$, displayed in red in Figure 2.4. Actually, in the differences on which L_0 norm is applied in $f(\cdot)$, the differences with the reference category zero are omitted for simplicity, which is possible because these quantities can be treated similarly, adding the axes where $\beta_1 = 0, \beta_2 \in \mathbb{R}$, as well as $\beta_1 \in \mathbb{R}, \beta_2 = 0$ to the partition. Hence, with the same arguments provided for the generalization to more levels and more factors, it is sufficient to treat the functions as provided above.

For $g(\boldsymbol{\beta})$ being a continuous function, it is known that in the compact set $\mathscr{D}$ it attains a (local) minimum. One defines

$$\boldsymbol{\beta}_g = (\beta_{g,1}, \beta_{g,2}) := \arg\min_{\boldsymbol{\beta} \in \mathscr{D}} g(\boldsymbol{\beta}).$$

Without loss of generality, it is assumed that $\boldsymbol{\beta}_g \in \mathscr{D}_1$, the other case works analogously. Differentiating between the two possible cases (1) $\beta_{g,1} \neq \beta_{g,2}$ and (2) $\beta_{g,1} = \beta_{g,2}$ yields the following.

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

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

so the infimum of M_{pen} is either attained in $\boldsymbol{\beta}_g$ or in (b, b) for some $b \in [b_1, b_2]$. Thus, the infimum of M_{pen} is attained somewhere in $\mathscr{D}$.

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

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

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

In both cases, there exists some $\tilde{\boldsymbol{\beta}} \in \mathscr{D}$ for which the infimum of $M_{pen}(\boldsymbol{\beta})$ is attained, hence it equals the minimum, which yields

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

It is noted that, in Figure 2.4 the case may occur that the red hyperplane does not go through the domain $\mathscr{D}$ hence there is no intersection of the hyperplane and $\mathscr{D}$. If this is the case, the claim directly follows since then the function $f(\cdot)$ is equal to one everywhere, hence $M_{pen}(\cdot)$ would be continuous in the domain $\mathscr{D}$.

It remains to show that $\tilde{\boldsymbol{\beta}} \in \mathring{\mathscr{D}}$. One assumes that $\tilde{\boldsymbol{\beta}}$ is on the boundary of $\mathscr{D}$, hence $\tilde{\boldsymbol{\beta}} \in \mathscr{D} \backslash \mathring{\mathscr{D}}$. Consequently, it holds by definition of the infimum that

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

and this also holds for $\boldsymbol{\beta} = \boldsymbol{x}^* \in \mathscr{D}$ which is a contradiction to the assumption (2.6). Therefore, it holds that $\tilde{\boldsymbol{\beta}} \in \mathring{\mathscr{D}}$, hence there exists a local minimum of M_{pen} in $\mathring{\mathscr{D}}$. □

Next, a result concerning the existence of a $\sqrt{n}$ consistent (sequence of) L_0-FGL estimator(s) for the case of fixed p is exhibited. The convergence properties of a_n^1 and a_n^0 introduced below ensure that the amount of penalization for factor selection and levels fusion are controlled. The corresponding assumption for factor selection is taken from Wang and Leng (2008), where in their work the adaptive group Lasso is considered.

Theorem 2.3.2 ($\sqrt{n}$ consistency for fixed p, Kaufmann and Kateri (2024))**.** One assumes that the regularity conditions (Reg1)-(Reg3) from Appendix B.1 hold and that p is fixed. One sets $a_n^1 := \max\{\lambda_1^n w_1^{(j)}; j \in \{1,\dots J\}\}$ 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 assumes $a_n^1/\sqrt{n} = o_p(1)$, $a_n^0 = O_p(1)$. Then, it holds that there exists a local minimizer $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$ of $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ satisfying

$$||\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})} - \boldsymbol{\beta}^*||_2 = O_p\left(\frac{1}{\sqrt{n}}\right).$$

Proof. For simplicity of notation, one writes $M_{pen}(\boldsymbol{\beta})$ for $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$, $P_\lambda(\boldsymbol{\beta})$ for $P_\lambda^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ and $\hat{\boldsymbol{\beta}}$ for $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$ throughout the following proof. Following the ideas of Fan and Li (2001) and Wang and Leng (2008) it has to be shown that $\forall \varepsilon > 0$ one can find a suitable $c > 0$ such that the following holds

$$P\left(\inf_{\boldsymbol{u}\in\mathbb{R}^p, ||\boldsymbol{u}||_2 = c} M_{pen}\left(\boldsymbol{\beta}^* + \frac{1}{\sqrt{n}}\boldsymbol{u}\right) > M_{pen}(\boldsymbol{\beta}^*)\right) \geq 1 - \varepsilon. \tag{2.7}$$

In contrast to Fan and Li (2001), the sum of the negative log-likelihood and the chosen penalty is minimized where they maximize the negative objective function which is clearly equivalent. The idea of Fan and Li (2001) is transferred to the case of L_0-FGL including two penalties and two types of tuning parameters and weights. Having shown (2.7), one can deduce that there exists a local minimum inside the ball $\{\boldsymbol{\beta}^* + \frac{1}{\sqrt{n}}\boldsymbol{u} \text{ where } ||\boldsymbol{u}||_2 < c\}$ using Lemma 2.3.1. This yields that one can find a local minimizer such that $||\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}||_2 = O_p(1/\sqrt{n})$ which is the claim. Plugging in the definition of $M_{pen}(\boldsymbol{\beta})$ yields with $H_n(\boldsymbol{u}) := M_{pen}\left(\boldsymbol{\beta}^* + \frac{1}{\sqrt{n}}\boldsymbol{u}\right) - M_{pen}(\boldsymbol{\beta}^*)$ that

$$\begin{aligned} H_n(\boldsymbol{u}) &= -L_n\left(\boldsymbol{\beta}^* + \frac{1}{\sqrt{n}}\boldsymbol{u}\right) + L_n(\boldsymbol{\beta}^*) + \lambda_1^n \sum_{j=1}^{J} w_1^{(j)}(||\boldsymbol{\beta}_j^* + \frac{1}{\sqrt{n}}\boldsymbol{u}||_2 - ||\boldsymbol{\beta}_j^*||_2) \\ &\quad + \lambda_0^n \sum_{j=1}^{J} \sum_{0\leq r<s\leq p_j} w_0^{(j,rs)}\left(||\beta_{j,r}^* - \beta_{j,s}^* + \frac{1}{\sqrt{n}}(u_r - u_s)||_0 - ||\beta_{j,r}^* - \beta_{j,s}^*||_0\right). \end{aligned} \tag{2.8}$$

At first, the log-likelihood parts in $H_n(\boldsymbol{u})$ are analyzed, before the penalty parts are examined.

Step 1 (log-likelihood): Like in Zou (2006) (proof of Theorem 4) the behavior of the log-likelihood part $-L_n\left(\boldsymbol{\beta}^* + \frac{1}{\sqrt{n}}\boldsymbol{u}\right) + L_n(\boldsymbol{\beta}^*)$ is examined with a Taylor expansion of $g_n(\boldsymbol{u}) := -L_n\left(\boldsymbol{\beta}^* + \frac{1}{\sqrt{n}}\boldsymbol{u}\right) + L_n(\boldsymbol{\beta}^*)$ around $\boldsymbol{u} = 0$ which yields using $g_n(\mathbf{0}) = \mathbf{0}$

$$-L_n\left(\boldsymbol{\beta}^* + \frac{1}{\sqrt{n}}\boldsymbol{u}\right) + L_n(\boldsymbol{\beta}^*) = T_{1,n} + T_{2,n} + T_{3,n}. \tag{2.9}$$

In particular, it holds with $\alpha_n := \frac{1}{\sqrt{n}}$, which is only introduced here for consistency purposes with the proof of the respective theorem for the diverging case which is provided later (Theorem

2.3.3),

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

These equalities can be directly seen by straightforward calculations. Plugging in that $\alpha_n = \frac{1}{\sqrt{n}}$, the following asymptotic properties are obtained using (Reg1)-(Reg3), where the abbreviation CLT is used for central limit theorem (Casella and Berger (2002), Theorem 5.5.14), as well as LLN for the (weak) law of large numbers (Casella and Berger (2002), Theorem 5.5.2)

$$\begin{aligned}
T_{1,n} &= -\sum_{i=1}^{n}[y_i - \varphi^{'}(\boldsymbol{x}_i\boldsymbol{\beta}^*)]\frac{\boldsymbol{x}_i^T\boldsymbol{u}}{\sqrt{n}} \to_d N(\mathbf{0}, \boldsymbol{u}^T I_F(\boldsymbol{\beta}^*)\boldsymbol{u}) \text{ (using CLT)},\\
T_{2,n} &= \sum_{i=1}^{n}\frac{1}{2}\varphi''(\boldsymbol{x}_i\boldsymbol{\beta}^*)\boldsymbol{u}^T\frac{\boldsymbol{x}_i\boldsymbol{x}_i^T}{n}\boldsymbol{u} \to_p \frac{1}{2}\boldsymbol{u}^T I_F(\boldsymbol{\beta}^*)\boldsymbol{u} \text{ (using LLN)},\\
T_{3,n} &= n^{-1/2}\frac{1}{6}\underbrace{\frac{1}{n}\sum_{i=1}^{n}\varphi'''(\boldsymbol{x}_i\boldsymbol{\beta}^*)(\boldsymbol{x}_i^T\boldsymbol{u})^3}_{\to_p \mathbb{E}(M(\boldsymbol{x})|\boldsymbol{x}^T\boldsymbol{u}|^3) < \infty \text{ by (Reg3)}} \quad \text{(using LLN)}
\end{aligned} \tag{2.10}$$

thus $6\sqrt{n}T_{3,n} < \infty$ according to Zou (2006) (proof of Theorem 4). With these properties one can conclude that the likelihood part of the objective function, hence (2.9), is asymptotically dominated by (2.10), thus by the expression $\boldsymbol{u}^T I_F(\boldsymbol{\beta}^*)\boldsymbol{u}$.

Step 2 (penalty): By the triangle inequality it holds that

$$||\boldsymbol{\beta}_j^* - \frac{1}{\sqrt{n}}\boldsymbol{u}||_2 \leq ||\boldsymbol{\beta}_j^*||_2 + ||\frac{1}{\sqrt{n}}\boldsymbol{u}||_2 \Leftrightarrow ||\boldsymbol{\beta}_j^* - \frac{1}{\sqrt{n}}\boldsymbol{u}||_2 - ||\boldsymbol{\beta}_j^*||_2 \leq ||\frac{1}{\sqrt{n}}\boldsymbol{u}||_2,$$

which yields

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

which is similarly obtained in Wang and Leng (2008). For the L_0 part it holds

$$\lambda_0^n\sum_{j=1}^{J}\sum_{0\leq r<s\leq p_j}w_0^{(j,rs)}\left(||\beta_{j,r}^* - \beta_{j,s}^* + \frac{1}{\sqrt{n}}(u_r - u_s)||_0 - ||\beta_{j,r}^* - \beta_{j,s}^*||_0\right) \tag{2.11}$$
$$\leq \sum_{j=1}^{J}\sum_{0\leq r<s\leq p_j}\underbrace{w_0^{(j,rs)}\lambda_0^n}_{\leq a_n^0 = O_p(1)} \tag{2.12}$$

such that (2.11) is $O_p(1)$. It is noted that p and J are fixed in this theorem thus they do not

grow with the sample size n. All in all, one can write

$$\begin{aligned} & M_{pen}\left(\boldsymbol{\beta}^* + \frac{1}{\sqrt{n}}\right) - M_{pen}(\boldsymbol{\beta}) \\ = \; & T_{1,n} + T_{2,n} + T_{3,n} + \lambda_1^n \sum_{j=1}^{J} w_1^{(j)} \left(||\boldsymbol{\beta}_j^* + \frac{1}{\sqrt{n}}\boldsymbol{u}||_2 - ||\boldsymbol{\beta}_j^*||_2 \right) \\ & + \lambda_0^n \sum_{j=1}^{J} \sum_{0 \leq r < s \leq p_j} w_0^{(j,rs)} \left(||\beta_{j,r}^* - \beta_{j,s}^* + \frac{1}{\sqrt{n}}(u_r - u_s)||_0 - ||\beta_{j,r}^* - \beta_{j,s}^*||_0 \right) \\ = \; & \underbrace{T_{1,n}}_{\to_d N(\ldots)} + \underbrace{T_{2,n}}_{\to_p \frac{1}{2}\boldsymbol{u}^T I_F(\boldsymbol{\beta}^*)\boldsymbol{u}} + \underbrace{T_{3,n}}_{\text{bounded}} + o_p(1)||\boldsymbol{u}|| + O_p(1) \end{aligned} \tag{2.13}$$

It is concluded that the expression $H_n(\boldsymbol{u}) = M_{pen}(\boldsymbol{\beta}^* + \frac{1}{\sqrt{n}}) - M_{pen}(\boldsymbol{\beta})$ is asymptotically dominated 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, c can be chosen (large enough) such that (2.13) > 0 hence (2.7) holds so there exists a local minimizer $\hat{\boldsymbol{\beta}}$ being $\sqrt{n}$-consistent. □

Having proven the existence of a $\sqrt{n}$ consistent L_0-FGL estimator for the case of fixed p, this is extended to the case of diverging p_n. However, instead of imposing the regularity conditions of Appendix B.1, the respective conditions of Appendix B.2 are imposed. Further, some adjustments are needed on controlling the amount of penalization and, finally, an assumption on the growth of the ratio of n and p_n is needed, which is similar to Fan and Peng (2004). After the proof of the following theorem, a remark (Remark 2.3.4) is supplied discussing convenient cases where these assumptions are satisfied, demonstrating that they are not too restrictive.

Theorem 2.3.3 (Consistency in the diverging p_n case, Kaufmann and Kateri (2024))**.** One assumes that the regularity conditions (div.Reg1)-(div.Reg3) of Appendix B.2 hold and that a_n^1 and a_n^0 are defined analogously to Theorem 2.3.2. With $\alpha_n := \sqrt{\frac{p_n}{n}}$, it is assumed that $\alpha_n a_n^1 J_n = o_p(1)$ and $a_n^0 p_n(p_n - 1) = o_p(1)$. Lastly, it is supposed that $p_n = o(n^{1/4})$. Then, it holds that there exists a local minimizer $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$ of $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ satisfying

$$||\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})} - \boldsymbol{\beta}^*||_2 = O_p(\alpha_n).$$

Proof. For simplicity of notation, one writes $M_{pen}(\boldsymbol{\beta})$ for $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$, $P_\lambda(\boldsymbol{\beta})$ for $P_\lambda^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ and $\hat{\boldsymbol{\beta}}$ for $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$ throughout the following proof. The proof is related to the proof of Theorem 1 in Fan and Peng (2004), transferred to L_0-FGL including a sum of two penalties and incorporating two types of tuning parameters and weights. The first part of the proof, where the log-likelihood is observed, is similar to Fan and Peng (2004). As in the proof of Theorem 2.3.2, it is shown that for any given $\varepsilon > 0$, (2.7) is satisfied, where $\frac{1}{\sqrt{n}}$ is replaced by α_n. Similarly, one defines $H_n(\boldsymbol{u}) := M_{pen}(\boldsymbol{\beta}^* + \alpha_n\boldsymbol{u}) - M_{pen}(\boldsymbol{\beta}^*)$ and writes

$$\begin{aligned} H_n(\boldsymbol{u}) = & -L_n(\boldsymbol{\beta}^* + \alpha_n\boldsymbol{u}) + L_n(\boldsymbol{\beta}^*) + \lambda_1^n \sum_{j=1}^{J_n} (w_1^{(j)}||\boldsymbol{\beta}_j^* + \alpha_n\boldsymbol{u}_j||_2 - w_1^{(j)}||\boldsymbol{\beta}_j||_2) \\ & + \lambda_0^n \sum_{j=1}^{J_n} \sum_{0 \leq r < s \leq p_j} (w_j^{(j,rs)}||\beta_{j,r}^* - \beta_{j,s}^* + \alpha_n(u_r - u_s)||_0 - w_0^{(j,rs)}||\beta_{j,r}^* - \beta_{j,s}^*||_0). \end{aligned} \tag{2.14}$$

Again, the log-likelihood part and the penalty part of the objective function are analyzed separately.

Step 1 (log-likelihood): For the log-likelihood part a Taylor expansion as in the proof of Theorem 2.3.2 is performed, but since p_n is allowed to grow with n, the asymptotic behavior of the components of the Taylor expansion differ from the mentioned theorems. In particular, for the Taylor expansion of $g_n(\boldsymbol{u}) := -L_n(\boldsymbol{\beta}^* + \alpha_n \boldsymbol{u}) + L_n(\boldsymbol{\beta}^*)$ around $\boldsymbol{u} = \mathbf{0}$ using the fact that $g_n(\mathbf{0}) = 0$ it holds

$$-L_n(\boldsymbol{\beta}^* + \alpha_n \boldsymbol{u}) + L_n(\boldsymbol{\beta}^*) = T_{1,n} + T_{2,n} + T_{3,n}. \tag{2.15}$$

Here, the quantities $T_{1,n}, T_{2,n}, T_{3,n}$ are similar to the proof of Theorem 2.3.2 with general α_n. For $T_{1,n}$ using the Cauchy-Schwartz inequality (CS) and (div.Reg2) one deduces

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

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 it holds $T_{1,n} = O_p(p_n)||\boldsymbol{u}||_2$ since $\alpha_n^2 n = p_n = \alpha_n \sqrt{n p_n}$. For the second summand $T_{2,n}$ the following equalities are inferred as in Fan and Peng (2004)

$$\begin{aligned} T_{2,n} &= -\frac{1}{2} \boldsymbol{u}^T \nabla^2 L_n(\boldsymbol{\beta}^*) \boldsymbol{u} \alpha_n^2 \\ &= -\frac{1}{2} \boldsymbol{u}^T \nabla^2 L_n(\boldsymbol{\beta}^*) \boldsymbol{u} \alpha_n^2 \underbrace{+ \frac{1}{2} \boldsymbol{u}^T \boldsymbol{I}_F(\boldsymbol{\beta}^*) \boldsymbol{u} n \alpha_n^2 - \frac{1}{2} \boldsymbol{u}^T \boldsymbol{I}_F(\boldsymbol{\beta}^*) \boldsymbol{u} n \alpha_n^2}_{=0} \\ &= -\frac{1}{2} \boldsymbol{u}^T \left[\frac{1}{n} (\nabla^2 L_n(\boldsymbol{\beta}^*) + \boldsymbol{I}_F(\boldsymbol{\beta}^*)) \right] \boldsymbol{u} n \alpha_n^2 + \frac{1}{2} \boldsymbol{u}^T \boldsymbol{I}_F(\boldsymbol{\beta}^*) \boldsymbol{u} n \alpha_n^2 \\ &= -\frac{1}{2} \boldsymbol{u}^T \left[\frac{1}{n} (\nabla^2 L_n(\boldsymbol{\beta}^*) - \mathbb{E}(\nabla^2 L_n(\boldsymbol{\beta}^*)) \right] \boldsymbol{u} n \alpha_n^2 + \frac{1}{2} \boldsymbol{u}^T \boldsymbol{I}_F(\boldsymbol{\beta}^*) \boldsymbol{u} n \alpha_n^2 \\ &= \frac{n}{2} \alpha_n^2 \boldsymbol{u}^T o_p(1/p_n) \boldsymbol{u} + \frac{n}{2} \alpha_n^2 \boldsymbol{u}^T \boldsymbol{I}_F(\boldsymbol{\beta}^*) \boldsymbol{u}, \end{aligned}$$

where it was used that $\boldsymbol{I}_F(\boldsymbol{\beta}^*) = -\mathbb{E}(\nabla^2 L_n(\boldsymbol{\beta}^*))$ 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 Fan and Peng (2004) which needs the assumption $p_n^4/n \to 0$ as $n \to \infty$. Moreover, as elaborated in the previous sentence, it is known that

$$||\frac{1}{n} \nabla^2 L_n(\boldsymbol{\beta}^*) + \boldsymbol{I}_F(\boldsymbol{\beta}^*)|| \, p_n = o_p(1)$$

so using $p_n \geq 1$ yields

$$||\frac{1}{n} \nabla^2 L_n(\boldsymbol{\beta}^*) + \boldsymbol{I}_F(\boldsymbol{\beta}^*)|| = o_p(1).$$

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.

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

where, using the Cauchy Schwarz inequality, one obtains

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

Following (div.Reg3), every component in $||(\log f_n(\boldsymbol{v}_l, \boldsymbol{\beta}^*))'''||_2$ can be bounded 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

$$(*) \leq \sum_{l=1}^{n} ||(\log f_n(\boldsymbol{v}_l, \boldsymbol{\beta}^*))'''||_2 \cdot ||\boldsymbol{u}||_2^3 \leq ||\boldsymbol{u}||_2^3 \sum_{l=1}^{n}\left(\sum_{i,j,k=1}^{p_n} M_{n,i,j,k}^2(\boldsymbol{x}_l)\right)^{1/2}. \tag{2.16}$$

Now the asymptotic behavior of the right hand side of the inequality (2.16) is observed. Using the Cauchy Schwarz inequality, it holds

$$\begin{aligned} &\left(\sum_{i,j,k=1}^{p_n} M_{n,i,j,k}^2(\boldsymbol{x}_l)\cdot 1\right)^2 \leq \left(\sum_{i,j,k=1}^{p_n} M_{n,i,j,k}^2(\boldsymbol{x}_l)\right) p_n^3 \\ &\Rightarrow \sum_{i,j,k=1}^{p_n} M_{n,i,j,k}^2(\boldsymbol{x}_l) \leq p_n^3 \Rightarrow \left(\sum_{i,j,k=1}^{p_n} M_{n,i,j,k}^2(\boldsymbol{x}_l)\right)^{1/2} \leq p_n^{3/2} \end{aligned} \tag{2.17}$$

With (2.17) using $\alpha_n = \sqrt{p_n/n}$ one can write

$$\begin{aligned} |T_{3,n}| &\leq \frac{1}{6}\alpha_n^3 ||\boldsymbol{u}||_2^3 \sum_{l=1}^{n}\left(\sum_{i,j,k=1}^{p_n} M_{n,i,j,k}^2(\boldsymbol{x}_l)\right)^{1/2} \\ &\leq \frac{1}{6}\alpha_n^3 ||\boldsymbol{u}||_2^3 \sum_{l=1}^{n} p_n^{3/2} = \frac{1}{6}\alpha_n^3 ||\boldsymbol{u}||_2^3 n p_n^{3/2} = \frac{1}{6}||\boldsymbol{u}||_2^3 \frac{p_n^3}{\sqrt{n}}. \end{aligned}$$

Since it was assumed that $\frac{p_n^4}{n} \to 0$, using $0 \leq \frac{p_n^2}{\sqrt{n}} = \sqrt{\frac{p_n^4}{n}} \leq \frac{p_n^4}{n} \to 0$ one deduces $\frac{p_n^2}{\sqrt{n}} \to 0$. 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

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

since by assumption $\alpha_n a_n^1 J_n = o_p(1)$. For the L_0 part, since

$$||\beta_{j,r}^* - \beta_{j,s}^* + \alpha_n(u_r - u_s)||_0 - ||\beta_{j,r}^* - \beta_{j,s}^*||_0 \leq 1,$$

one can deduce

$$\lambda_0^n \sum_{j=1}^{J_n} \sum_{0 \leq r < s \leq p_j} \left(w_0^{(j,rs)} ||\beta_{j,r}^* - \beta_{j,s}^* + \alpha_n(u_r - u_s)||_0 - w_0^{(j,rs)} ||\beta_{j,r}^* - \beta_{j,s}^*||_0 \right)$$
$$\leq \sum_{j=1}^{J_n} \sum_{0 \leq r < s \leq p_j} \lambda_0^n w_j^{(j,rs)} \leq a_n^0 \sum_{j=1}^{J_n} \sum_{0 \leq r < s \leq p_j} 1 =: (**).$$

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, i.e. whether ordinal or nominal factors or mixtures are observed. 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, by the assumptions $a_n^0 p_n(p_n - 1) = o_p(1)$, 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 $o_p(1)$.

It can be concluded by the steps above, as well as (2.14) and (2.15)

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

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

Remark 2.3.4 (On the assumptions of Theorem 2.3.3, Kaufmann and Kateri (2024))**.** 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 $a_n^0 p_n(p_n - 1) = o_p(1)$, 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 $\lambda_0^n = o(1/p_n^2)$ yields that $\lambda_0^n p_n(p_n - 1) = o_p(1)$ holds.

2.3.2 Asymptotic Normality

This section provides the existence of an L_0-FGL estimator being asymptotically normal distributed. As in the previous section, the case of p being fixed is investigated first, before the case of diverging p_n is covered. The assumptions that are needed to be imposed on the amount of penalization are, for the group lasso part, similar to those imposed by Zou (2006) for the adaptive lasso.

Theorem 2.3.5 (Existence of an estimator satisfying the asymptotic normality property for the case fixed p, Kaufmann and Kateri (2024))**.** One assumes that (Reg1)-(Reg3) of Appendix B.1 hold and that the true underlying structure is sparse according to Definition 1.2.4 (i). For the group lasso part the adaptive weights $w_1^{(j)} = ||\hat{\boldsymbol{\beta}}_j^{(\mathrm{ML})}||_2^{-\gamma}$ are employed for some arbitrarily chosen $\gamma > 0$ where $\hat{\boldsymbol{\beta}}^{(\mathrm{ML})}$ is the unpenalized MLE. Furthermore, it is supposed that $\lambda_1^n \cdot n^{-1/2} \to 0$ and $\lambda_1^n \cdot n^{(\gamma-1)/2} \to \infty$. For the tuning and weights of the L_0 part, one assumes $a_n^0 = o_p(1)$. Then, it holds that there exists a local minimizer $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$ of $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ satisfying

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{A^*}^{(L_0\text{-FGL})} - \boldsymbol{\beta}_{A^*}^*\right) \to_d N(0, \boldsymbol{I}_{11}^{-1}),$$

where $\hat{\boldsymbol{\beta}}_{A^*}^{(L_0\text{-FGL})}$ and $\boldsymbol{\beta}_{A^*}^*$ denote the sub-vectors of $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$ and $\boldsymbol{\beta}^*$, respectively, containing only the components belonging in the true active set A^*.

Proof. For simplicity of notation, one writse $M_{pen}(\boldsymbol{\beta})$ for $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$, $P_\lambda(\boldsymbol{\beta})$ for $P_\lambda^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ and $\hat{\boldsymbol{\beta}}$ for $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$ throughout the following proof. The proof follows Zou (2006) where the oracle properties for the adaptive lasso are shown. One writes $\boldsymbol{\beta} = \boldsymbol{\beta}^* + \frac{\boldsymbol{u}}{\sqrt{n}}$ and remember that by (2.8) it holds

$$\begin{aligned}
H_n(\boldsymbol{u}) &:= M_{pen}\left(\boldsymbol{\beta}^* + \frac{\boldsymbol{u}}{\sqrt{n}}\right) - M_{pen}(\boldsymbol{\beta}^*) \\
&= \underbrace{-L_n\left(\boldsymbol{\beta}^* + \frac{1}{\sqrt{n}}\boldsymbol{u}\right) + L_n(\boldsymbol{\beta}^*)}_{=:(*)} + \underbrace{\lambda_1^n \sum_{j=1}^{J} w_1^{(j)}(||\boldsymbol{\beta}_j^* + \frac{1}{\sqrt{n}}\boldsymbol{u}||_2 - ||\boldsymbol{\beta}_j^*||_2)}_{=:\mathfrak{P}_{1,n}(\boldsymbol{u})} \\
&\quad + \underbrace{\lambda_0^n \sum_{j=1}^{J} \sum_{0 \le r < s \le p_j} w_0^{(j,rs)} \left(||\beta_{j,r}^* - \beta_{j,s}^* + \frac{1}{\sqrt{n}}(u_r - u_s)||_0 - ||\beta_{j,r}^* - \beta_{j,s}^*||_0\right)}_{=:\mathfrak{P}_{2,n}(\boldsymbol{u})}.
\end{aligned}$$

It is noted that the quantities $\mathfrak{P}_{1,n}(\boldsymbol{u})$ and $\mathfrak{P}_{2,n}(\boldsymbol{u})$ are only defined and needed for simplicity of notation during this proof. One aims 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, a lower index in $\hat{\boldsymbol{u}}$ is omitted for simplicity, as for $\hat{\boldsymbol{\beta}}$. Since $H_n(\boldsymbol{u})$ is the same as (2.8) in the proof of Theorem 2.3.2, the first steps performing a Taylor expansion of the log-likelihood part $(*)$ resulting in $T_{1,n}, T_{2,n}, T_{3,n}$ (one compares (2.9)) and observing the asymptotic behavior of those using the regularity conditions are similar. Hence, it remains to analyze the asymptotic behavior of the functions $\mathfrak{P}_{1,n}(\boldsymbol{u})$ and $\mathfrak{P}_{2,n}(\boldsymbol{u})$.
It holds that

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

For the case that $\boldsymbol{\beta}_j^* \neq \mathbf{0}$, hence $j \in A^*$, one gets by the consistency of the MLE (compare Theorem C.1.2)

$$w_1^{(j)} = \frac{1}{||\hat{\boldsymbol{\beta}}_j^{(\mathrm{ML})}||_2^{\gamma}} \to_p ||\boldsymbol{\beta}_j^*||_2^{-\gamma} \tag{2.18}$$

and it holds further that

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

i.e., to conclude

$$-||\boldsymbol{u}_j||_2 \leq \sqrt{n}\{||\boldsymbol{\beta}_j^* + \frac{\boldsymbol{u}_j}{\sqrt{n}}||_2 - ||\boldsymbol{\beta}_j^*||_2\} \leq ||\boldsymbol{u}_j||_2. \tag{2.19}$$

In the following, denote the summands of $\mathfrak{P}_{1,n}(\boldsymbol{u})$ for $j \in A^*$ by $(\mathfrak{P}_{1,n}(\boldsymbol{u}))_{A^*}$, the analogous notation is used for the summands where $j \notin A^*$. Using Slutky's Theorem (Casella and Berger (2002), Theorem 5.5.17), it is inferred

$$\begin{aligned}(\mathfrak{P}_{1,n}(\boldsymbol{u}))_{A^*} &= \lambda_1^n \sum_{j \in A^*} \left[w_1^{(j)} ||\boldsymbol{\beta}_j^* + \frac{\boldsymbol{u}_j}{\sqrt{n}}||_2 - w_1^{(j)} ||\boldsymbol{\beta}_j^*||_2 \right] \\ &= \underbrace{\frac{\lambda_1^n}{\sqrt{n}}}_{\to 0} \sum_{j \in A^*} \underbrace{w_1^{(j)}}_{\to_p ||\boldsymbol{\beta}_j^*||_2^{-\gamma}} ||\boldsymbol{\beta}_j^*||^{-\gamma} \underbrace{\sqrt{n}\left[||\boldsymbol{\beta}_j^* + \frac{\boldsymbol{u}_j}{\sqrt{n}}||_2 - ||\boldsymbol{\beta}_j^*||_2 \right]}_{\text{bounded using (2.19)}} \to_p 0\end{aligned}$$

for the case that $\boldsymbol{\beta}_j^* \neq \mathbf{0}$. Now the case that $\boldsymbol{\beta}_j^* = \mathbf{0}$, hence $j \notin A^*$, is treated. With $w_1^{(j)} = ||\hat{\boldsymbol{\beta}}_j^{(\mathrm{ML})}||_2^{-\gamma} = n^{\gamma/2} ||\hat{\boldsymbol{\beta}}_j^{(\mathrm{ML})} \sqrt{n}||_2^{-\gamma}$, one can deduce

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

which goes to ∞ for $||\boldsymbol{u}_j||_2 \neq 0$ and equals zero otherwise. Hence one gets for the j-th summand of $\mathfrak{P}_{1,n}(\boldsymbol{u})$, denoted by $(\mathfrak{P}_{1,n}(\boldsymbol{u}))_j$, for $n \to \infty$

$$(\mathfrak{P}_{1,n}(\boldsymbol{u}))_j \to_p \begin{cases} 0 & \text{if } ||\boldsymbol{u}_j||_2 = 0,\ \boldsymbol{\beta}_j^* = \mathbf{0}, \\ \infty & \text{if } ||\boldsymbol{u}_j||_2 \neq 0,\ \boldsymbol{\beta}_j^* = \mathbf{0}, \\ 0 & \text{if } \boldsymbol{\beta}_j^* \neq \mathbf{0}. \end{cases} \tag{2.20}$$

This yields

$$\begin{aligned}\mathfrak{P}_{1,n}(\boldsymbol{u}) &= \sum_{j \in A^*} (\mathfrak{P}_{1,n}(\boldsymbol{u}))_j + \sum_{j \notin A^*} (\mathfrak{P}_{1,n}(\boldsymbol{u}))_j \\ &= \underbrace{\sum_{j \in A^*} (\mathfrak{P}_{1,n}(\boldsymbol{u}))_j}_{\to_p 0} + \underbrace{\sum_{j \notin A^*,\, ||\boldsymbol{u}_j||_2 = 0} (\mathfrak{P}_{1,n}(\boldsymbol{u}))_j}_{\to_p 0} + \sum_{j \notin A^*,\, ||\boldsymbol{u}_j||_2 \neq 0} (\mathfrak{P}_{1,n}(\boldsymbol{u}))_j,\end{aligned}$$

where the last summand converges in probability to ∞, however, if $||\boldsymbol{u}_j|| = 0$ for all $j \notin A^*$, the last summand is zero. Consequently it is concluded for $\mathfrak{P}_{1,n}(\boldsymbol{u})$ that

$$\mathfrak{P}_{1,n}(\boldsymbol{u}) \to_p \begin{cases} 0 & , \text{if } \boldsymbol{u}_j = \boldsymbol{0}\, \forall j \notin A^* \\ \infty & , \text{otherwise.} \end{cases}$$

For $\mathfrak{P}_{2,n}(\boldsymbol{u})$, using (2.11), it holds that

$$\mathfrak{P}_{2,n}(\boldsymbol{u}) \leq \underbrace{\sum_{j=1}^{J} \sum_{0 \leq r < s \leq p_j} \underbrace{w_0^{(j,rs)} \lambda_0^n}_{\leq a_n^0 = o_p(1)}}_{=o_p(1)} \; p \Rightarrow \mathfrak{P}_{2,n}(\boldsymbol{u}) \to_p 0.$$

To sum up

$$H_n(\boldsymbol{u}) = \underbrace{T_{1,n}}_{\to_d \boldsymbol{u}^T N(\boldsymbol{0}, I_F(\boldsymbol{\beta}^*))} + \underbrace{T_{2,n}}_{\to_p \frac{1}{2}\boldsymbol{u}^T \boldsymbol{I}_F(\boldsymbol{\beta}^*)\boldsymbol{u}} + \underbrace{\mathfrak{P}_{1,n}(\boldsymbol{u})}_{\text{see (2.20)}} + \underbrace{T_{3,n}}_{\to_p 0} + \underbrace{\mathfrak{P}_{2,n}(\boldsymbol{u})}_{\to_p 0}.$$

Finally, one infers

$$H_n(\boldsymbol{u}) \to_d H(\boldsymbol{u}) = \begin{cases} \frac{1}{2}\boldsymbol{u}_{A^*}^T I_{11} \boldsymbol{u}_{A^*} - \boldsymbol{u}_{A^*}^T \boldsymbol{W}_{A^*}, & , \text{if } \boldsymbol{u}_j = \boldsymbol{0}\, \forall j \notin A^* \\ \infty & , \text{otherwise} \end{cases}$$

where $\boldsymbol{W} \sim N(\boldsymbol{0}, \boldsymbol{I}_F(\boldsymbol{\beta}^*))$ and consequently $\boldsymbol{W}_{A^*} \sim N(\boldsymbol{0}, \boldsymbol{I}_{11})$. The unique minimum of $H(\boldsymbol{u})$ is clearly at $\boldsymbol{u}^{\min} = (\boldsymbol{u}_{A^*}^{\min}, \boldsymbol{u}_{(A^*)^c}^{\min}) = (\boldsymbol{I}_{11}^{-1}\boldsymbol{W}_{A^*}, \boldsymbol{0})^T$ which can be obtained by straightforward calculations. Consequently, there exists $\hat{\boldsymbol{u}} = \arg\min_{\boldsymbol{u}} H_n(\boldsymbol{u})$ satisfying $\hat{\boldsymbol{u}}_{A^*} \to_d I_{11}^{-1}\boldsymbol{W}_{A^*}$ and $\hat{\boldsymbol{u}}_{(A^*)^c} \to_d 0$ with the same arguments provided in Knight and Fu (2000) for the non-convex Bridge estimator and the claim follows. □

Besides the MLE, it is possible to use alternative adaptive weights in the group lasso part, discussed in the following remark.

Remark 2.3.6 (Adaptive weights in the group lasso part)**.** Besides the MLE used in Theorem 2.3.5, one could also use another *consistent* initial estimator $\tilde{\boldsymbol{\beta}}$, where consistency is understood in the sense of (2.18). That is, one could also use an L_0-FGL estimator as initial estimator with $w_1^{(j)} := ||\hat{\boldsymbol{\beta}}_j^{(L_0\text{-FGL})}||_2^{-\gamma}$. Since it may happen that $||\hat{\boldsymbol{\beta}}_j^{(L_0\text{-FGL})}||_2 = 0$ by the construction of L_0-FGL, one can follow Zou and Zhang (2009) and Xin et al. (2017) and set $w_1^{(j)} = \left(||\hat{\boldsymbol{\beta}}_j^{(L_0\text{-FGL})}||_2 + \frac{1}{n}\right)^{-\gamma}$, which clearly not affects any shown asymptotic property.

To extend Theorem 2.3.5 to the case of diverging p_n, it is noticed that an approximation of the L_0 part in L_0-FGL is necessary, since sub-differentiability of the penalty function is needed in the proof of the respective theorem. As in Appendix A.1, as well as displayed in Figure 1.2, the L_0 norm is approximated as follows

$$||\xi||_0 \approx \frac{2}{1 + \exp(-\gamma_0|\xi|)} - 1 =: N(\xi).$$

However, no further approximation of the absolute value in the denominator (as done in PIRLS, Appendix A.1) is needed since it is sufficient if the (approximation of the) penalty function satisfies sub-differentiability. Consequently, the approximation of $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ is defined as follows

$$\widetilde{M}_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta}) := -L_n(\boldsymbol{\beta}) + \lambda_1^n \sum_{j=1}^{J_n} w_1^{(j)} ||\boldsymbol{\beta}_j||_2 + \lambda_0^n \sum_{j=1}^{J_n} \sum_{0 \leq r < s \leq p_j} w_0^{(j,rs)} N(\beta_{j,r} - \beta_{j,s}),$$

satisfying $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta}) \approx \widetilde{M}_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$. In the same sense as for $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$, a minimizer of $\widetilde{M}_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ is called an *approximate* L_0-FGL estimator. The same notation $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$ is used for the approximate L_0-FGL, since it is *only* used in the following theorem, so there is no need to introduce a new notation.

Based on this approximation given above, the theorem that is provided next (Theorem 2.3.7) shows that in the diverging p_n case, one can find a local minimizer of $\widetilde{M}_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$, thus an approximate L_0-FGL estimator, being asymptotically normal distributed. Before doing so, the following comments on the respective proof and the imposed regularity conditions are essential.

For the proof, it is crucial to ensure that there exists an approximate L_0-FGL estimator satisfying consistency in the sense that $||\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})} - \boldsymbol{\beta}^*||_2 = O_p(\alpha_n)$ for $\alpha_n = \sqrt{\frac{p_n}{n}}$. However, using the fact that $N(\xi) \leq 1$, one can show the existence of such a $\sqrt{\frac{n}{p_n}}$ consistent approximate L_0-FGL estimator under the same assumptions as in Theorem 2.3.3, caused by the fact that in the second step of the respective theorem, it is used that $||\ldots||_0 \leq 1$, which similarly holds for the approximation. Consequently, under the same assumptions imposed in Theorem 2.3.3, there exists some local minimizer $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$ of $\widetilde{M}_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ satisfying $||\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})} - \boldsymbol{\beta}^*||_2 = O_p(\alpha_n)$ for $\alpha_n = \sqrt{\frac{p_n}{n}}$.

For this local minimizer being consistent, the asymptotic normality property is shown in the following Theorem 2.3.7. The idea of this theorem is taken from the reference Wang and Tian (2019), where the adaptive group lasso is investigated. However, it is transferred to the case of the approximate L_0-FGL penalty function, incorporating a group lasso and an L_0 part, as well as two tuning parameters. Clearly, some regularity conditions are necessary, starting with (div.Reg1)-(div.Reg4). Additionally, to ensure the existence of a consistent (local) minimizer, the assumptions of Theorem 2.3.3 are needed. Corresponding to (A6) in Wang and Tian (2019) and (8) in Zhao and Yu (2006), a further condition controlling the size of the minimum of the true parameter is needed. Lastly, as provided in the assumptions of Theorem 2.3 in Wang and Tian (2019), it is assumed that $\lambda_1^n n^{\frac{1}{2}(2-\delta+\frac{1}{4})} \to 0$ and $\lambda_1^n n^{1-\frac{1}{4}} = \lambda_1^n n^{\frac{3}{4}} \to \infty$ holds, where δ is specified below. Actually, the assumption including β_{min}^* in the theorem below is automatically satisfied with the sparsity assumption imposed here (Definition 1.2.4 (ii)). However, to keep the theorem as general as possible, this condition is kept.

Theorem 2.3.7 (Existence of an approximate L_0-FGL estimator satisfying the asymptotic normality property for the diverging p_n case, Kaufmann and Kateri (2024))**.** One assumes that (div.Reg1)-(div.Reg4) of Appendix B.2 and the assumptions of Theorem 2.3.3 hold. Further, one assumes that the true underlying structure is sparse according to Definition 1.2.4 (ii). With $\beta_{\min}^* := \min_{j=1,\ldots,j_{0,n}} ||\boldsymbol{\beta}_j^*||_2$ it is supposed that there exists some $\frac{3}{4} < \delta \leq 1$ and $C > 0$ such that $n^{\frac{1}{2}(1-\delta)}\beta_{\min}^* \geq C$. Additionally, one assumes that $\lambda_1^n n^{\frac{1}{2}(2-\delta+\frac{1}{4})} \to 0$ and $\lambda_n^1 n^{\frac{3}{4}} \to \infty$ as $n \to \infty$. For the group lasso part the adaptive weights $w_1^{(j)} = ||\tilde{\boldsymbol{\beta}}_j||_2^{-1}$ are employed, where $\tilde{\boldsymbol{\beta}}$ is a $\sqrt{n/p_n}$ consistent initial estimator, hence $||\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2 = O_p(\sqrt{p_n/n})$. Finally, one assumes that $\mathbb{E}(Y - \varphi'(\boldsymbol{x}_1\boldsymbol{\beta}^*))^4 < \infty$, where φ is the cumulant function in the EDF expression (Definition 1.1.2) for the density function of Y. Then, there exists a local minimizer $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$ of $\widetilde{M}_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ satisfying

$$\boldsymbol{e}_n \boldsymbol{I}_{11,n}^{1/2} \left(\hat{\boldsymbol{\beta}}_{A^*}^{(L_0\text{-FGL})} - \boldsymbol{\beta}_{A^*}^*\right) \to_d N(0,1),$$

where $\hat{\boldsymbol{\beta}}_{A^*}^{(L_0\text{-FGL})}$ and $\boldsymbol{\beta}_{A^*}^*$ denote the sub-vectors of $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$ and $\boldsymbol{\beta}^*$, respectively, containing

only the components belonging in the true active set A^*, while $\boldsymbol{e}_n$ is a $p_{0,n}$ dimensional unit vector.

Proof. During the following proof, one writes $\widetilde{M}_{pen}$ for $\widetilde{M}_{pen}^{(L_0\text{-FGL})}$, as well as $\hat{\boldsymbol{\beta}}$ for $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$ for simplicity of notation. The proof corresponds to Wang and Tian (2019) (Theorem 2.3), where a related theorem is shown for adaptive group lasso. First, introduce the following abbreviations

$$f_1(\boldsymbol{\beta}) := \lambda_1^n \sum_{j=1}^{J_n} w_1^{(j)} ||\boldsymbol{\beta}_j||_2 \text{ and } f_0(\boldsymbol{\beta}) := \lambda_0^n \sum_{j=1}^{J_n} \sum_{0 \leq r < s \leq p_j} w_0^{(j,rs)} N(\beta_{j,r} - \beta_{j,s}).$$

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

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

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

$$\begin{aligned} & \frac{\partial f_1(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}_{A^*}} + \frac{\partial f_0(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}_{A^*}} \\ = \; & \frac{\partial L_n(\boldsymbol{\beta}_{A^*}^*)}{\partial \boldsymbol{\beta}_{A^*}} + \nabla_{A^*} \left(\frac{\partial L_n(\boldsymbol{\beta}_{A^*}^*)}{\partial \beta_{A^*}} \right)^T (\hat{\boldsymbol{\beta}}_{A^*} - \boldsymbol{\beta}_{A^*}^*) \\ & + \frac{1}{2} (\hat{\boldsymbol{\beta}}_{A^*} - \boldsymbol{\beta}_{A^*}^*)^T \nabla_{A^*}^2 \left(\frac{\partial L_n(\boldsymbol{\xi}_{A^*})}{\partial \boldsymbol{\beta}_{A^*}} \right) (\hat{\boldsymbol{\beta}}_{A^*} - \boldsymbol{\beta}_{A^*}^*), \end{aligned} \tag{2.21}$$

where $\boldsymbol{\xi}_{A^*}$ is between $\hat{\boldsymbol{\beta}}_{A^*}$ and $\boldsymbol{\beta}_{A^*}^*$ Since the assumed regularity conditions of Wang and Tian (2019) also hold here (in particular (div.Reg2)-(div.Reg4) are required for the following equality) and $\hat{\boldsymbol{\beta}}$ is consistent (one consults Theorem 2.3.3 and the explanations right before Theorem 2.3.7), one can follow the lines of Wang and Tian (2019), equation (4.5), resulting in

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

Similarly, following Wang and Tian (2019) (below equation (4.5)), it is shown that for the sub-differential of the (adaptive) group Lasso part it holds

$$\left|\left| \frac{\partial f_1(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}_{A^*}} \right|\right|_2^2 = o_p\left(\frac{1}{n}\right).$$

The derivative of the approximation of the L_0 part leads to a sub-differential. Since the function $N(\xi)$ includes the absolute value function $|\xi|$, it is not differentiable in zero. Nevertheless, it is sub-differentiable with $|\frac{\partial |\xi|}{\partial \xi}| \leq 1$. Consequently, it holds

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

It is further inferred that

$$\begin{aligned}
\left\| \frac{\partial f_0(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}_{A^*}} \right\|_2^2 &= \left\| \lambda_0^n \sum_{j=1}^{J_n} \sum_{0 \leq r < s \leq p_j} w_0^{(j,rs)} \frac{\partial}{\partial \boldsymbol{\beta}_{A^*}} N(\beta_{j,r} - \beta_{j,s}) \right\|_2^2 \\
&= \left\| \lambda_0^n \sum_{j=1}^{j_{0,n}} \sum_{0 \leq r < s \leq p_j} w_0^{(j,rs)} \frac{\partial}{\partial \boldsymbol{\beta}_{A^*}} N(\beta_{j,r} - \beta_{j,s}) \right\|_2^2 \\
&\leq \left\| \frac{1}{2} p_{0,n}(p_{0,n} - 1) \cdot \gamma \cdot \max_{j \in \{1, \dots j_{0,n}\}, r,s \in \{0, \dots p_j\}} w_0^{(j,rs)} \lambda_0^n \right\|_2^2 \\
&\leq \left(|a_n^0| \gamma \frac{1}{2} p_{0,n}(p_{0,n} - 1) \right)^2 = o_p(1) o_p(1) = o_p(1),
\end{aligned}$$

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 2.3.3).

Additionally it holds that $\nabla_{A^*} \left(\frac{\partial L_n(\boldsymbol{\beta}_{A^*}^*)}{\partial \beta_{A^*}} \right)^T = -\boldsymbol{I}_{11,n}$, thus, putting all together one can deduce by (2.21)

$$\begin{aligned}
& o_p(1) + o_p(1) = \frac{\partial L_n(\boldsymbol{\beta}_{A^*}^*)}{\partial \boldsymbol{\beta}_{A^*}} - \boldsymbol{I}_{11,n}(\hat{\boldsymbol{\beta}}_{A^*} - \boldsymbol{\beta}_{A^*}^*) + o_p(1) \\
\Leftrightarrow \quad & o_p(1) + \boldsymbol{e}_n \boldsymbol{I}_{11,n}^{-1/2} \frac{\partial L_n(\boldsymbol{\beta}_{A^*}^*)}{\partial \boldsymbol{\beta}_{A^*}} = \boldsymbol{e}_n I_{11,n}^{1/2}(\hat{\boldsymbol{\beta}}_{A^*} - \boldsymbol{\beta}_{A^*}^*),
\end{aligned} \tag{2.22}$$

where it is known by (div.Reg2) that the Fisher information matrix is finite. It remains to show that the second summand on the left hand side of (2.22) converges in distribution to a standard normal distribution $N(0,1)$. Again, since the regularity conditions of Wang and Tian (2019) hold, it is referred to their work showing that by the CLT of Lindeberg-Feller

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

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

2.3.3 Consistency in Factor Selection

Having shown results on consistency and asymptotic normality, the focus of this section is on consistency in factor selection, which is also denoted as selection consistency for short. In particular, it is shown that, asymptotically, the probability that L_0-FGL excludes influential factors from the model can be made arbitrarily small. The theorem provided next (Theorem 2.3.8) was motivated by the work of Bunea (2008) where the focus of this reference lies on L_1 as well as L_1 *and* L_2 penalization in linear and logistic regression, however, the presence of factors is not addressed there. More specifically, Theorem 2.3.8 shows that, for the consistent estimator of Theorem 2.3.2, the probability that the true active set A^* is *not* contained in the estimated active set by L_0-FGL, that is $A_n^{(L_0-\text{FGL})}$, can be made arbitrarily small.

Theorem 2.3.8 (Selection consistency for fixed p, Kaufmann and Kateri (2024))**.** One assumes that the conditions of Theorem 2.3.2 are satisfied. Then, for the minimizer $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$ of Theorem 2.3.2 it holds that $\forall \varepsilon > 0$ one can find $N \in \mathbb{N}$ such that

$$\mathbb{P}(A^* \not\subseteq A_n^{(L_0\text{-FGL})}) < \varepsilon \quad \forall\, n \geq N, \tag{2.23}$$

where $A_n^{(L_0\text{-FGL})}$ is the estimated active set corresponding to $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$.

Proof. For simplicity of notation, one writes A_n for $A_n^{(L_0\text{-FGL})}$, as well as $\hat{\boldsymbol{\beta}}$ for $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$ throughout the following proof. The beginning of the proof follows Bunea (2008) (Proof of Lemma 3.1) but the proof is transferred to the more general case of $\boldsymbol{\beta}_j$ being a sub-vector instead of a single entry of the coefficient vector. Having that, the proven $\sqrt{n}$ consistency of the estimator $\hat{\boldsymbol{\beta}}$ of Theorem 2.3.2 is used to show the claim. Starting with the probability on the left hand side of the claim (2.23), one can deduce

$$\begin{aligned}
\mathbb{P}(A^* \not\subseteq A_n) \quad &\leq \mathbb{P}(j \notin A_n \text{ for some } j \in A^*) \\
&\leq \mathbb{P}(\hat{\boldsymbol{\beta}}_j = \mathbf{0} \text{ and } \boldsymbol{\beta}_j^* \neq \mathbf{0} \text{ for some } j \in A^*) \\
&\leq \mathbb{P}(||\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*||_2 = ||\boldsymbol{\beta}_j^*||_2 \text{ for some } j \in A^*) \\
&\leq \mathbb{P}(||\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*||_2 \geq \min_{l \in A^*} ||\boldsymbol{\beta}_l^*||_2 \text{ for some } j \in A^*) \\
&\leq \mathbb{P}(||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2 \geq \min_{l \in A^*} ||\boldsymbol{\beta}_l^*||_2). \qquad (2.24)
\end{aligned}$$

It is noted that $\min_{l \in A^*} ||\boldsymbol{\beta}_l^*||_2$ is a minimum over a finite set, thus it always exists. Now the goal is to bound (2.24) by some ε. Since it is known from Theorem 2.3.2 that $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2 = O_p(1/\sqrt{n})$ one gets that $\forall \varepsilon > 0$ there exist constants $M, \tilde{N} > 0$ such that

$$\mathbb{P}(||\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)||_2 > M) < \varepsilon \quad \forall\, n > \tilde{N}. \qquad (2.25)$$

Hence, for $n > \tilde{N}$ it holds that $\mathbb{P}\left(||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2 > \frac{M}{\sqrt{n}}\right) < \varepsilon$. Now, with $\varepsilon > 0$ and constants $M, \tilde{N} > 0$, one can always choose some $N' > 0$ such that $\frac{M}{\sqrt{n}} \leq \min_{l \in A^*} ||\boldsymbol{\beta}_l^*||_2$ for all $n \geq N'$. It is noted that by definition $||\boldsymbol{\beta}_l^*||_2 \neq 0\ \forall\, l \in A^*$. Now one can write expression (2.24) 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, $\forall \varepsilon > 0$ some $N := \max\{\tilde{N}, N'\}$ can be found such that

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

which completes the proof. □

As a next step, the respective result on consistency in factor selection for the case of diverging p_n is provided. Actually, in the theorem below, instead of imposing the sparsity assumption of Definition 1.2.4 (ii), it would be sufficient to require that there exists some $C > 0$ such that $\min_{l \in A^*} ||\boldsymbol{\beta}_l^*|| \geq C$, which can be seen during the respective proof.

Theorem 2.3.9 (Selection consistency in the diverging p_n case, Kaufmann and Kateri (2024))**.** One assumes that the conditions of Theorem 2.3.3 are satisfied. Then, for the minimizer $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$ of Theorem 2.3.3 it holds that for $\forall \varepsilon > 0$ one can find $N \in \mathbb{N}$ such that

$$\mathbb{P}(A^* \not\subseteq A_n^{(L_0\text{-FGL})}) < \varepsilon \ \ \forall\, n \geq N. \qquad (2.26)$$

Proof. For simplicity of notation, one writes A_n for $A_n^{(L_0\text{-FGL})}$, as well as $\hat{\boldsymbol{\beta}}$ for $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$ throughout the following proof. The idea of the proof works analogously to the proof of Theorem 2.3.8, hence showing $\mathbb{P}(A^* \not\subseteq A_n) \leq \mathbb{P}(||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2 \geq \min_{l \in A^*} ||\boldsymbol{\beta}_l^*||_2)$. Having that, (2.25) is modified using Theorem 2.3.3 (here, $\hat{\boldsymbol{\beta}}$ is the minimizer from Theorem 2.3.3) which yields

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

with $\alpha_n = \sqrt{\frac{p_n}{n}}$. By the sparsity assumption (Definition 1.2.4) the set A^* is finite, thus there exists some $C > 0$ such that $\min_{l \in A^*} ||\boldsymbol{\beta}_l^*|| \geq C \; \forall n \in \mathbb{N}$. Consequently, one can find $N' \in \mathbb{N}$ for which it holds that $0 < M\sqrt{\frac{p_n}{n}} \leq C \leq \min_{l \in A^*} ||\boldsymbol{\beta}_l^*|| \, \forall n \geq N'$, since $\sqrt{p_n/n} \to 0$ by the assumptions. The rest works analogously to the proof of Theorem 2.3.8. □

2.3.4 Fusion Properties

In the following examination, the fusion properties of L_0-FGL are analyzed. To be more precise, the existence of a fusion threshold corresponding to the (coefficients') difference of two levels from the same factor is proven. To do so, the behavior of the objective function $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ of L_0-FGL being minimized is investigated in two theorems, where it is analyzed how $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ behaves if entries of the coefficient vector $\boldsymbol{\beta}$ are fused. In particular, one of these results is an *asymptotic* result (Fusion property objective function II, Theorem 2.3.17) where the tuning parameter, as well as the number of factors, are allowed to grow with n, which further covers the case of p being fixed as a by-product. The other result is a *non-asymptotic* result (Fusion Property objective function I, Theorem 2.3.15), i.e. it holds for fixed n and fixed $p_n = p$. In particular, it is shown that, depending on the coefficients' difference, the objective function *decreases* if two levels corresponding to one factor are fused - that is, the corresponding coefficients are set to be equal. These Theorems 2.3.15 and 2.3.17 build the basis for Theorem 2.3.26, giving a statement about the resulting estimator $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$. Finally, this section is closed with a theorem showing an asymptotic screening property for fusion, provided in Theorem 2.3.37. For the majority of the theorems, corresponding corollaries are supplied showing a similar statement for sub-sequences, which are crucial in the proof of Theorem 2.3.37.

Recall that $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$ is the L_0-FGL estimator being of dimension $p + 1$. In the following, reduced versions of the estimator are introduced, accounting for the fact that levels fusion, as well as factor selection, may be performed. For this, an introduction of the quantities p_j^{af} and p^{af} in Notation 2.3.10 is necessary.

Notation 2.3.10. Remember that the number of levels of factor $\mathscr{X}_j$ is given by p_j for $j \in \{1, \ldots, J\}$. Now, *after* a fusion-type penalty as L_0-FGL is applied, the number of levels *after* fusion may be less or equal than before. That is, let p_j^{af} denote the number of levels of $\mathscr{X}_j$ *after* L_0-FGL was performed, where *af* is an abbreviation for *after fusion*. If factor $\mathscr{X}_j$ is identified as a non-influential noise variable, thus $j \notin A_n$ it holds that $p_j^{af} = 0$. In general $p_j^{af} \leq p_j \, \forall j \in \{1, \ldots, J\}$. Additionally, one defines

$$p^{af} := \sum_{j=1}^{J} p_j^{af} = \sum_{j \in A_n} p_j^{af},$$

where this quantity is defined similarly for diverging p_n replacing J by J_n, thus writing p_n^{af}.

After providing an example for motivation purposes, the notation of reduced versions of the estimates is introduced in Notation 2.3.12.

Example 2.3.11. Assume that $J = 2$ factors are considered as candidate explanatory variables in the L_0-FGL penalized logistic regression model. To be more precise, let $\mathscr{X}_1$ be an ordinal factor with five levels ($p_1 = 4$) and $\mathscr{X}_2$ a nominal factor with four levels ($p_1 = 3$), i.e. $p = 7$. To obtain $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$, one is looking for

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})} &= (\hat{\beta}_{int}, \hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2) \\ &= (\hat{\beta}_{int}, \;\; \hat{\beta}_{1,2}, \hat{\beta}_{1,3}, \hat{\beta}_{1,4}, \hat{\beta}_{1,5}, \;\; \hat{\beta}_{2,2}, \hat{\beta}_{2,3}, \hat{\beta}_{2,4}) \in \mathbb{R}^8. \end{aligned}$$

Now one assumes that, after L_0-FGL was performed, levels three and four of $\mathscr{X}_1$ are fused and factor $\mathscr{X}_2$ is identified as non-influential noise variable. This leads to the following structure

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})} &= (\hat{\beta}_{int}, \hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2) \\ &= (\hat{\beta}_{int},\ \hat{\beta}_{1,2}, \hat{\beta}_{1,3}, \hat{\beta}_{1,3}, \hat{\beta}_{1,5},\ 0,0,0) \in \mathbb{R}^8.\end{aligned}$$

It clearly holds that $p_1^{af} = 3$ and $p_2^{af} = 0$. Consequently, without loss of information, one can write $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$ in a *reduced* version, which is denoted by $\hat{\boldsymbol{\beta}}_{\text{red}}^{(L_0\text{-FGL})}$. That is, in $\hat{\boldsymbol{\beta}}_{\text{red}}^{(L_0\text{-FGL})}$ the excluded factors are left out, as well as the replicates for the conducted fusions. In particular, for this example one deduces

$$\hat{\boldsymbol{\beta}}_{\text{red}}^{(L_0\text{-FGL})} = (\hat{\beta}_{int},\ \hat{\beta}_{1,2}, \hat{\beta}_{1,3}, \hat{\beta}_{1,5}) \in \mathbb{R}^4,$$

with $p^{af} = p_1^{af} + p_2^{af} = 3$, thus $\hat{\boldsymbol{\beta}}_{\text{red}}^{(L_0\text{-FGL})} \in \mathbb{R}^{1+p^{af}}$.

Motivated by the provided example, the following notation is introduced.

Notation 2.3.12 (Extended/full and reduced version of the estimator/estimates)**.**

i) With $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$ being the (full) L_0-FGL estimator of dimension $p+1$, the *reduced* parameter vector is defined by

$$\hat{\boldsymbol{\beta}}_{\text{red}}^{(L_0\text{-FGL})} \in \mathbb{R}^{1+p^{af}}.$$

Clearly, depending on the given context, $\hat{\boldsymbol{\beta}}_{\text{red}}^{(L_0\text{-FGL})}$ can be considered as a random variable, as $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$, the same applies for the quantities introduced below. The *full* parameter vector (minimizer of (2.3)) including zeros for the removed factors and repetitions corresponding to the fused levels) is denoted by

$$\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})} \in \mathbb{R}^{p+1}.$$

To sum up, with some appropriate matrix $\boldsymbol{H} \in \{0,1\}^{(1+p)\times(1+p^{af})}$, it holds

$$\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})} = \boldsymbol{H} \cdot \hat{\boldsymbol{\beta}}_{\text{red}}^{(L_0\text{-FGL})}.$$

ii) In the same way, for every factor $\mathscr{X}_j$ and a suitable matrix $\boldsymbol{H}_j \in \{0,1\}^{p_j \times p_j^{af}}$, where "+1" is not needed for the intercept since this matrix $\boldsymbol{H}_j$ operates factor-wise, one gets the following

$$\hat{\boldsymbol{\beta}}_j^{(L_0\text{-FGL})} = \boldsymbol{H}_j \cdot \hat{\boldsymbol{\beta}}_{j,\text{red}}^{(L_0\text{-FGL})}.$$

Analogously this yields the corresponding reduced factor-wise sub-vector

$$\hat{\boldsymbol{\beta}}_{j,\text{red}}^{(L_0\text{-FGL})} \in \mathbb{R}^{p_j^{af}},$$

and the factor-wise full version

$$\hat{\boldsymbol{\beta}}_j^{(L_0\text{-FGL})} \in \mathbb{R}^{p_j}.$$

As continuation of Example 2.3.11, the following example concerning the extended/full version of the *truth* $\boldsymbol{\beta}^*$ is supplied.

Example 2.3.13. In the setting of Example 2.3.11, it is now assumed that the truth is given by

$$\boldsymbol{\beta}^* = (2,\ 1,1,1,8,\ 0,0,0),$$

so factor $\mathscr{X}_2$ is truly a noise variable and levels two, three and four of factor $\mathscr{X}_1$ are fused, hence $p_1^* = 2$, $p_2^* = 0$, $p^* = 2$. Then, the reduced version of the truth is given by

$$\boldsymbol{\beta}^*_{\text{red}} = (2,\ 1,8)$$

Consequently, Notation 2.3.14 is provided.

Notation 2.3.14 (Extended/full and reduced version of the truth)**.** As in Notation 2.3.12, one writes

$$\boldsymbol{\beta}^* \in \mathbb{R}^{p+1}$$

for the full vector of the truth and

$$\boldsymbol{\beta}^*_{\text{red}} \in \mathbb{R}^{p^*+1}$$

for the reduced version, where p^* is the dimension of the true parameter vector, including true levels fusions and true factor selections. In the same way for the factor-wise reduced versions for every j, one writes

$$\boldsymbol{\beta}^*_{j,\text{red}} \in \mathbb{R}^{p_j^*}$$

and for the full versions

$$\boldsymbol{\beta}^*_j \in \mathbb{R}^{p_j}.$$

It is recalled that, as a general assumption in the presented asymptotic theory, the number of levels p_j of a factor $j \in \{1, \ldots, J_n\}$ is *not* allowed to grow with n. Consequently, also in the case of a diverging number of parameters, the (possible) fusions are countable. It is further recalled that the weights $w_0^{(j,ik)}, w_1^{(j,i)}$ depend on n for all $j \in \{1, \ldots, J_n\}$, i, $k \in \{1, \ldots, p_j\}$, although it is not expressed in a sub-index. Analogously, the objective function $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ and $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$ depend on n. These considerations are crucial for the upcoming section.

Fusion Property of the Objective Function I and II

To investigate the behavior of the objective function $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ in case of fusion, the relation between $M_{pen}^{(L_0\text{-FGL})}(\tilde{\boldsymbol{\beta}}_f)$ and $M_{pen}^{(L_0\text{-FGL})}(\tilde{\boldsymbol{\beta}}_{nf})$ is revisited. It is noted that the quantities $\tilde{\boldsymbol{\beta}}_f$ and $\tilde{\boldsymbol{\beta}}_{nf}$ are *not* necessarily local minimizers of $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$, these are "candidate" coefficient vectors, and, at first, not random variables. Later, the developed results are used to show properties of the resulting estimators $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$ being random variables.

Different from Theorem 2.2.1, in the following theorem (Theorem 2.3.15) the existence of a fusion threshold is proven, specifying the fusion execution of L_0-FGL. The weight $w_0^{(\iota,r)}$ in the theorem below is treated as non-random, where it is commented on the case of $w_0^{(\iota,r)}$ being random after the proof in Remark 2.3.16. It is worth noting that Theorem 2.3.15 is *not* an asymptotic property, hence p is fixed. A continuity assumption on the log-likelihood function is required, which is introduced and specified in Appendix B.3 as well as in the proof of the following theorem.

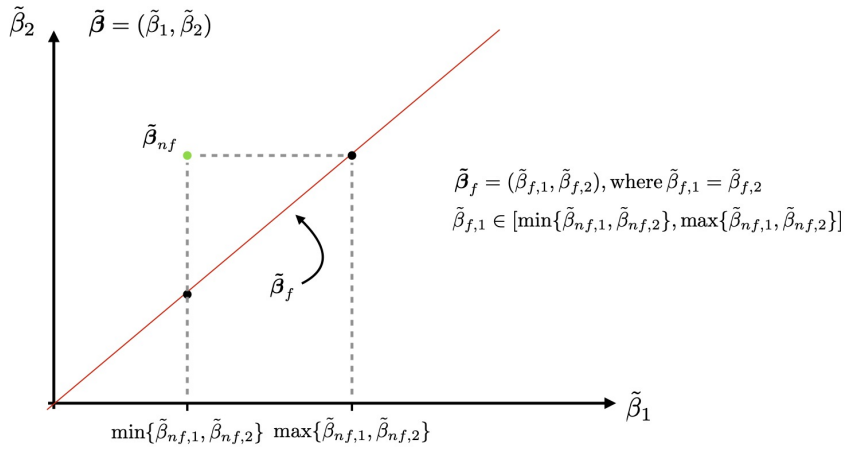

Figure 2.5: *Visualization of $\tilde{\boldsymbol{\beta}}_{nf}$ and the corresponding $\tilde{\boldsymbol{\beta}}_f$ in the simplified setting $J = \iota = 1$ and $p_\iota = p_1 = 2$.*

Theorem 2.3.15 (Fusion Property objective function I of L_0-FGL)**.** For a factor $\iota \in \{1, \dots, J\}$ (being ordinal without loss of generality), one defines $\tilde{\boldsymbol{\beta}}_{nf} \in \mathbb{R}^{p+1}$ and $\tilde{\boldsymbol{\beta}}_f \in \mathbb{R}^{p+1}$ such that for some (fixed) $r \in \{1, \dots, p_\iota\}$

$$\begin{aligned}
\tilde{\boldsymbol{\beta}}_{nf} &:= (\tilde{\beta}_{nf,int}, \tilde{\boldsymbol{\beta}}_{nf,1}, \tilde{\boldsymbol{\beta}}_{nf,2}, \dots, \tilde{\boldsymbol{\beta}}_{nf,\iota-1}, \tilde{\boldsymbol{\beta}}_{nf,\iota}, \tilde{\boldsymbol{\beta}}_{nf,\iota+1}, \dots \tilde{\boldsymbol{\beta}}_{nf,J}) \\
&\quad \tilde{\boldsymbol{\beta}}_{nf,\iota} := (\tilde{\beta}_{nf,\iota,1}, ..., \tilde{\beta}_{nf,\iota,p_\iota}), \text{ where } \tilde{\beta}_{nf,\iota,r} \neq \tilde{\beta}_{nf,\iota,r-1} \text{ (level } r \text{ and } r-1 \text{ of } \iota \textit{ not fused}), \\
\tilde{\boldsymbol{\beta}}_f &:= (\tilde{\beta}_{f,int}, \tilde{\boldsymbol{\beta}}_{f,1}, \tilde{\boldsymbol{\beta}}_{f,2}, \dots, \tilde{\boldsymbol{\beta}}_{f,\iota-1}, \tilde{\boldsymbol{\beta}}_{f,\iota}, \tilde{\boldsymbol{\beta}}_{f,\iota+1}, \dots \tilde{\boldsymbol{\beta}}_{f,J}) \\
&\quad \tilde{\boldsymbol{\beta}}_{f,\iota} := (\tilde{\beta}_{f,\iota,1}, ..., \tilde{\beta}_{f,\iota,p_\iota}), \text{ where } \tilde{\beta}_{nf,\iota,i} = \tilde{\beta}_{f,\iota,i} \, \forall i \in \{1, \dots, p_\iota\} \backslash \{r, r-1\} \\
&\qquad\qquad \text{and } \tilde{\beta}_{f,\iota,r} = \tilde{\beta}_{f,\iota,r-1} \text{ (level } r \text{ and } r-1 \text{ of } \iota \textit{ fused}).
\end{aligned}$$

For $j \in \{1, \dots, J\} \backslash \{\iota\}$, let $\tilde{\boldsymbol{\beta}}_{nf,j} = \tilde{\boldsymbol{\beta}}_{f,j}$. Further, one defines $\Delta := \tilde{\beta}_{nf,\iota,r} - \tilde{\beta}_{nf,\iota,r-1}$ and assumes that (Cont1) of Definition B.3.1 (Appendix B.3) is satisfied. Now, there exists some $\Delta_0 > 0$, depending on $\lambda_1, \lambda_0, w_0^{(\iota,r)}$, as well as on $\tilde{\boldsymbol{\beta}}_{nf}$ and n, such that for all $0 < \Delta < \Delta_0$, it holds that

$$M_{pen}^{(L_0\text{-FGL})}(\tilde{\boldsymbol{\beta}}_{nf}) > M_{pen}^{(L_0\text{-FGL})}(\tilde{\boldsymbol{\beta}}_f),$$

so the value of the objective function $M_{pen}^{(L_0\text{-FGL})}(\cdot)$ decreases if coefficients $\tilde{\beta}_{nf,\iota,r}$ and $\tilde{\beta}_{nf,\iota,r-1}$ are fused. This holds in the same way for ι being nominal and levels $r, s \in \{1, \dots, p_\iota\}$.

Proof. For simplicity of notation, $M_{pen}(\cdot)$ is written for $M_{pen}^{(L_0\text{-FGL})}(\cdot)$ throughout the following proof. Further, it is sufficient to prove this Theorem for the case of ι being ordinal while the nominal case works analogously.
By definition, in $\tilde{\boldsymbol{\beta}}_f$ the categories r and $r-1$ of factor ι are *fused* and except for these categories, $\tilde{\boldsymbol{\beta}}_{nf}$ and $\tilde{\boldsymbol{\beta}}_f$ coincide. It may occur that in $\tilde{\boldsymbol{\beta}}_{nf}$ levels of other factors are fused, but, as defined above, *not* level r and $r-1$ of factor ι. Hence, in $\tilde{\boldsymbol{\beta}}_{nf}$ there are $\xi \in \mathbb{N}_0$ fusions and in $\tilde{\boldsymbol{\beta}}_f$ there are $\xi + 1$ fusions. By the nature of fusion, it holds that

$$\tilde{\beta}_{f,\iota,r} = \tilde{\beta}_{f,\iota,r-1} \in [\min\{\tilde{\beta}_{nf,\iota,r}, \tilde{\beta}_{nf,\iota,r-1}\}, \max\{\tilde{\beta}_{nf,\iota,r}, \tilde{\beta}_{nf,\iota,r-1}\}],$$

see Figure 2.6 and Figure 2.5. Without loss of generality, one assumes $\min\{\tilde{\beta}_{nf,\iota,r}, \tilde{\beta}_{nf,\iota,r-1}\} = \tilde{\beta}_{nf,\iota,r-1}$. Since an ordinal covariate is observed, this holds by definition but observing nominal covariates one has to differentiate two cases, however, they can be treated analogously. Thus, for some Δ_1, $\Delta_2 > 0$ it holds that

$$\tilde{\beta}_{nf,\iota,r} - \tilde{\beta}_{nf,\iota,r-1} = \Delta_1 + \Delta_2 = \Delta > 0 \Rightarrow \tilde{\beta}_{nf,\iota,r} = \tilde{\beta}_{nf,\iota,r-1} + \Delta,$$

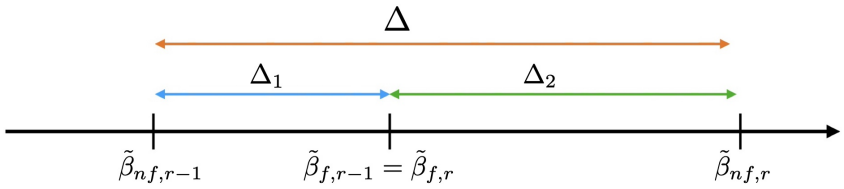

Figure 2.6: *Location of $\tilde{\beta}_{nf,\iota,r}$, $\tilde{\beta}_{nf,\iota,r-1}$ and the fused coefficients $\tilde{\beta}_{f,\iota,r-1} = \tilde{\beta}_{f,\iota,r}$ for the case $\min\{\tilde{\beta}_{nf,\iota,r}, \tilde{\beta}_{nf,\iota,r-1}\} = \tilde{\beta}_{nf,\iota,r-1}$, the other case works analogously. The sub-index ι is neglected in the figure for the sake of clarity.*

for which it is referred again to Figure 2.6 for a visualization. Now one has to show that $M_{pen}(\tilde{\boldsymbol{\beta}}_f) < M_{pen}(\tilde{\boldsymbol{\beta}}_{nf})$. By construction

$$\tilde{\boldsymbol{\beta}}_{nf} - \tilde{\boldsymbol{\beta}}_f = (0, ..., 0, \Delta_1, -\Delta_2, 0, ..., 0). \tag{2.27}$$

Because of the continuity of the negative log-likelihood $-L_n(\boldsymbol{\beta})$ by (Cont1) of Definition B.3.1 (Appendix B.3) and the norm $||\boldsymbol{\beta}||_{\boldsymbol{K}_\iota}$, one can find for all $\epsilon_1, \epsilon_2 > 0$ some $\delta(\epsilon_1, \tilde{\boldsymbol{\beta}}_{nf}, n), \delta(\epsilon_2, \tilde{\boldsymbol{\beta}}_{nf}, n) > 0$, such that for all $\tilde{\boldsymbol{\beta}}_f$ with $||\tilde{\boldsymbol{\beta}}_{nf} - \tilde{\boldsymbol{\beta}}_f||_2 < \min\{\delta(\epsilon_1, \tilde{\boldsymbol{\beta}}_{nf}, n), \delta(\epsilon_2, \tilde{\boldsymbol{\beta}}_{nf}, n)\}$ it holds that

$$|L_n(\tilde{\boldsymbol{\beta}}_f) - L_n(\tilde{\boldsymbol{\beta}}_{nf})| < \epsilon_1, \tag{2.28}$$

$$|||\tilde{\boldsymbol{\beta}}_{nf,\iota}||_{\boldsymbol{K}_\iota} - ||\tilde{\boldsymbol{\beta}}_{f,\iota}||_{\boldsymbol{K}_\iota}| < \epsilon_2. \tag{2.29}$$

For $\lambda_1 \neq 0$, one chooses ϵ_1 and ϵ_2 such that

$$0 < \epsilon_1 < \frac{1}{2}\lambda_0 w_0^{(\iota,r)} \quad \text{and} \quad 0 < \epsilon_2 < \frac{1}{2}\frac{\lambda_0 w_0^{(\iota,r)}}{\lambda_1}.$$

With these choices of ϵ_1, ϵ_2, one can deduce

$$\exists\, \delta(\epsilon_1, \tilde{\boldsymbol{\beta}}_{nf}, n),\ \delta(\epsilon_2, \tilde{\boldsymbol{\beta}}_{nf}, n)\ : \forall\, \tilde{\boldsymbol{\beta}}_f, \tilde{\boldsymbol{\beta}}_{nf} \text{ with } ||\tilde{\boldsymbol{\beta}}_{nf} - \tilde{\boldsymbol{\beta}}_f||_2 < \min\{\delta(\epsilon_1, \tilde{\boldsymbol{\beta}}_{nf}, n), \delta(\epsilon_2, \tilde{\boldsymbol{\beta}}_{nf}, n)\}\ :$$
$$|L_n(\tilde{\boldsymbol{\beta}}_f) - L_n(\tilde{\boldsymbol{\beta}}_{nf})| < \epsilon_1 \text{ and } |||\tilde{\boldsymbol{\beta}}_{nf,\iota}||_{\boldsymbol{K}_\iota} - ||\tilde{\boldsymbol{\beta}}_{f,\iota}||_{\boldsymbol{K}_\iota}| < \epsilon_2.$$

For $\lambda_1 = 0$, any $\epsilon_2 > 0$ can be chosen, since in this case, inequality (2.31) shown below, holds for all ϵ_2. Let $\delta(\epsilon_1, \tilde{\boldsymbol{\beta}}_{nf}, n), \delta(\epsilon_2, \tilde{\boldsymbol{\beta}}_{nf}, n)$ be given as above and let

$$||\tilde{\boldsymbol{\beta}}_{nf} - \tilde{\boldsymbol{\beta}}_f||_2 < \min\{\delta(\epsilon_1, \tilde{\boldsymbol{\beta}}_{nf}, n), \delta(\epsilon_2, \tilde{\boldsymbol{\beta}}_{nf}, n)\} =: \Delta_0(\tilde{\boldsymbol{\beta}}_{nf}, n, \epsilon_1, \epsilon_2).$$

Since ϵ_1, ϵ_2 depend on λ_0, λ_1 and $w_0^{(\iota,r)}$, the expression $\Delta_0(\tilde{\boldsymbol{\beta}}_{nf}, n, \epsilon_1, \epsilon_2)$ also depends on these quantities, however it is noted that n is fixed. Now, it remains to analyze the L_0 part of the objective function evaluated in $\tilde{\boldsymbol{\beta}}_{nf}$ and $\tilde{\boldsymbol{\beta}}_f$. Because of the structure of $\tilde{\boldsymbol{\beta}}_{nf}$ and $\tilde{\boldsymbol{\beta}}_f$ it can be inferred that

$$\begin{aligned}
&\sum_{j=1}^{J}\sum_{i=1}^{p_j} w_0^{(j,i)}||\tilde{\beta}_{nf,j,i} - \tilde{\beta}_{nf,j,i-1}||_0 \\
&= \sum_{i=1}^{p_\iota} w_0^{(\iota,i)}||\tilde{\beta}_{nf,\iota,i} - \tilde{\beta}_{nf,\iota,i-1}||_0 + \sum_{j=1,\, j\neq\iota}^{J}\sum_{i=1}^{p_j} w_0^{(j,i)}||\tilde{\beta}_{nf,j,i} - \tilde{\beta}_{nf,j,i-1}||_0 \\
&= \underbrace{\sum_{i=1,i\neq r}^{p_\iota} w_0^{(\iota,i)}||\tilde{\beta}_{nf,\iota,i} - \tilde{\beta}_{nf,\iota,i-1}||_0}_{=:\zeta_1} + w_0^{(\iota,r)} + \underbrace{\sum_{j=1,\, j\neq\iota}^{J}\sum_{i=1}^{p_j} w_0^{(j,i)}||\tilde{\beta}_{nf,j,i} - \tilde{\beta}_{nf,j,i-1}||_0}_{=:\zeta_2} \\
&= \zeta_1 + \zeta_2 + w_0^{(\iota,r)}
\end{aligned}$$

and analogously using that $\tilde{\beta}_{f,\iota,r} = \tilde{\beta}_{f,\iota,r-1} \Rightarrow ||\tilde{\beta}_{f,\iota,r} - \tilde{\beta}_{f,\iota,r-1}||_0 = 0$ $(*)$ one can write

$$\begin{aligned}
&\sum_{j=1}^{J}\sum_{i=1}^{p_j} w_0^{(j,i)}||\tilde{\beta}_{f,j,i} - \tilde{\beta}_{f,j,i-1}||_0 \\
&= \sum_{i=1}^{p_\iota} w_0^{(\iota,i)}||\tilde{\beta}_{f,\iota,i} - \tilde{\beta}_{f,\iota,i-1}||_0 + \underbrace{\sum_{j=1,\, j\neq\iota}^{J}\sum_{i=1}^{p_j} w_0^{(j,i)}||\tilde{\beta}_{f,j,i} - \tilde{\beta}_{f,j,i-1}||_0}_{=\zeta_2,\ \text{since } \tilde{\boldsymbol{\beta}}_{f,j}=\tilde{\boldsymbol{\beta}}_{nf,j}\ \forall j\neq\iota} \\
&\overset{(*)}{=} \underbrace{\sum_{i=1, i\neq r}^{p_\iota} w_0^{(\iota,i)}||\tilde{\beta}_{f,\iota,i} - \tilde{\beta}_{f,\iota,i-1}||_0}_{=\zeta_1,\ \text{since } \tilde{\boldsymbol{\beta}}_{f,\iota,i}=\tilde{\boldsymbol{\beta}}_{nf,\iota,i}\ \forall i\in\{2,\ldots,p_\iota\}\setminus\{r,r-1\}} + \zeta_2 \\
&= \zeta_1 + \zeta_2
\end{aligned}$$

By (2.28) and (2.29), for $||\tilde{\boldsymbol{\beta}}_{nf} - \tilde{\boldsymbol{\beta}}_f||_2 < \Delta_0(\tilde{\boldsymbol{\beta}}_{nf}, n, \epsilon_1, \epsilon_2)$, it holds that

$$L_n(\tilde{\boldsymbol{\beta}}_f) - L_n(\tilde{\boldsymbol{\beta}}_{nf}) > -\epsilon_1 \text{ and } ||\tilde{\boldsymbol{\beta}}_{nf}||_{\boldsymbol{K}_\iota} - ||\tilde{\boldsymbol{\beta}}_f||_{\boldsymbol{K}_\iota} > -\epsilon_2.$$

Putting all steps together, one obtains for $||\tilde{\boldsymbol{\beta}}_{nf} - \tilde{\boldsymbol{\beta}}_f||_2 < \Delta_0(\tilde{\boldsymbol{\beta}}_{nf}, n, \epsilon_1, \epsilon_2)$

$$\begin{aligned}
&M_{pen}(\tilde{\boldsymbol{\beta}}_{nf}) - M_{pen}(\tilde{\boldsymbol{\beta}}_f) \\
&= -L_n(\tilde{\boldsymbol{\beta}}_{nf}) + \lambda_1||\tilde{\boldsymbol{\beta}}_{nf,\iota}||_{\boldsymbol{K}_\iota} + \lambda_0\sum_{j=1}^{J}\sum_{i=1}^{p_j} w_0^{(j,i)}||\tilde{\beta}_{nf,j,i} - \tilde{\beta}_{nf,j,i-1}||_0 \\
&\quad + L_n(\tilde{\boldsymbol{\beta}}_f) - \lambda_1||\tilde{\boldsymbol{\beta}}_{f,\iota}||_{\boldsymbol{K}_\iota} - \lambda_0\sum_{j=1}^{J}\sum_{i=1}^{p_j} w_0^{(j,i)}||\tilde{\beta}_{f,j,i} - \tilde{\beta}_{f,j,i-1}||_0 \\
&= L_n(\tilde{\boldsymbol{\beta}}_f) - L_n(\tilde{\boldsymbol{\beta}}_{nf}) + \lambda_1||\tilde{\boldsymbol{\beta}}_{nf,\iota}||_{\boldsymbol{K}_\iota} - \lambda_1||\tilde{\boldsymbol{\beta}}_{f,\iota}||_{\boldsymbol{K}_\iota} + \lambda_0 \cdot w_0^{(\iota,r)} \qquad (2.30) \\
&> -\epsilon_1 - \lambda_1\epsilon_2 + \lambda_0 \cdot w_0^{(\iota,r)} \\
&> 0, \qquad (2.31)
\end{aligned}$$

where the latter holds by the choice of ϵ_1, ϵ_2. This yields in total, that with $\Delta_0(\tilde{\boldsymbol{\beta}}_{nf}, n, \epsilon_1, \epsilon_2)$ as given above, for all $\Delta < \Delta_0(\tilde{\boldsymbol{\beta}}_{nf}, n, \epsilon_1, \epsilon_2)$ with $\tilde{\beta}_{nf,\iota,r} - \tilde{\beta}_{nf,\iota,r-1} = \Delta < \Delta_0(\tilde{\boldsymbol{\beta}}_{nf}, n, \epsilon_1, \epsilon_2)$ it holds $||\tilde{\boldsymbol{\beta}}_{nf} - \tilde{\boldsymbol{\beta}}_f||_2 \leq \Delta_1 + \Delta_2 = \Delta < \Delta_0(\tilde{\boldsymbol{\beta}}_{nf}, n, \epsilon_1, \epsilon_2)$, so all the inequalities above hold and it can be concluded that

$$M_{pen}(\tilde{\boldsymbol{\beta}}_{nf}) > M_{pen}(\tilde{\boldsymbol{\beta}}_f) \ \ \forall\, \Delta < \Delta_0(\tilde{\boldsymbol{\beta}}_{nf}, n, \epsilon_1, \epsilon_2)$$

so the claim follows. Consequently the value of the objective function in $\tilde{\boldsymbol{\beta}}_f$ is less than in $\tilde{\boldsymbol{\beta}}_{nf}$, hence the objective function decreases if $\tilde{\beta}_{nf,\iota,r}$ and $\tilde{\beta}_{nf,\iota,r-1}$ are fused, given that they are close enough to each other, specified by $\Delta_0(\tilde{\boldsymbol{\beta}}_{nf}, n, \epsilon_1, \epsilon_2)$. It is recalled that the latter quantity depends on λ_1, λ_0 and $w_0^{(\iota,r)}$ through ϵ_1 and ϵ_2. □

Remark 2.3.16 (Theorem 2.3.15 for random variables)**.** Theorem 2.3.15 is stated with $\tilde{\boldsymbol{\beta}}_{nf}$ and $\tilde{\boldsymbol{\beta}}_f \in \mathbb{R}^{p+1}$, thus not for random variables. Further, the weight $w_0^{(\iota,r)}$ was considered and treated as non-random. However, one can similarly provide the theorem considering these quantities as random variables. For this, (Cont1) is needed, and a similar structure of $\tilde{\boldsymbol{\beta}}_f$ and $\tilde{\boldsymbol{\beta}}_{nf}$ is imposed, where the equalities are assumed to hold with probability one, hence almost surely (a.s.). Then, the statement of the theorem consequently holds with probability one and the proof is completely similar with the equalities and inequalities used in the proof holding a.s. Further, continuity is used as explained in Remark B.3.2. To sum up, Theorem 2.3.15 analogously holds for $\tilde{\boldsymbol{\beta}}_{nf}$,$\tilde{\boldsymbol{\beta}}_f$ and $w_0^{(\iota,r)}$ being random variables.

The following theorem (Theorem 2.3.17) discusses an *asymptotic property* of the objective function $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$, corresponding to Theorem 2.3.15. Throughout this theorem (Theorem 2.3.17) and its proof the case of diverging p_n is considered, however, it is obvious that it also holds for fixed p ($p_n = p$ constant). Similar to the previous theorem, the weight $w_0^{(\iota,r)}$ as well as the quantities $\tilde{\boldsymbol{\beta}}_{nf}$, $\tilde{\boldsymbol{\beta}}_f$, below being *sequences*, are treated as non-random, the case of those being random variables is picked up after providing and proving the theorem in a remark (Remark 2.3.19).

In some of the following theorems, for simplicity, c is used for *any* constant, which are of course not all necessarily equal.

Theorem 2.3.17 (Fusion Property objective function II of L_0-FGL, fixed p and diverging p_n)**.** Let c denote any constant. For any factor $\iota \in \{1, \ldots, J_n\}$ (ordinal without loss of generality, one compares Theorem 2.3.15), one defines *sequences* $\tilde{\boldsymbol{\beta}}_{nf} \subseteq \mathbb{R}^{p_n+1}$ and $\tilde{\boldsymbol{\beta}}_f \subseteq \mathbb{R}^{p_n+1}$ in the completely analogous way to Theorem 2.3.15, in particular these are sequences in n. It is assumed that with $\Delta_n := \tilde{\beta}_{nf,\iota,r} - \tilde{\beta}_{nf,\iota,r-1}$ it holds that $\Delta_n \to 0\,(n \to \infty)$ and for the weight $w_0^{(\iota,r)}$ it is supposed that $\inf\limits_{n\in\mathbb{N}} w_0^{(\iota,r)} = c > 0$. Further, let $\inf\limits_{n\in\mathbb{N}} \lambda_0^n = c > 0$ and $\sup\limits_{n\in\mathbb{N}} \lambda_1^n = c < \infty$. Finally, let (Cont2) of Definition B.3.1 (Appendix B.3) be satisfied for the case of fixed p and (Cont3) for the case of diverging p_n. Then, there exists some $n_0 \in \mathbb{N}$ such that the value of the objective function $M_{pen}^{(L_0\text{-FGL})}(\cdot)$ decreases $\forall n \geq n_0$ if coefficients $\tilde{\beta}_{nf,\iota,r}$ and $\tilde{\beta}_{nf,\iota,r-1}$ are fused, hence it holds that

$$M_{pen}^{(L_0\text{-FGL})}(\tilde{\boldsymbol{\beta}}_{nf}) > M_{pen}^{(L_0\text{-FGL})}(\tilde{\boldsymbol{\beta}}_f) \;\; \forall n \geq n_0.$$

Proof. For simplicity of notation, $M_{pen}(\cdot)$ is written for $M_{pen}^{(L_0\text{-FGL})}(\cdot)$ in the following proof. The crucial difference between this proof and the proof of Theorem 2.3.15 is, that for the diverging p_n case, the dimensions of $\tilde{\boldsymbol{\beta}}_{nf}$ and $\tilde{\boldsymbol{\beta}}_f$ grow. However, since these quantities coincide except for two *fixed* entries, one can similarly use the continuity of the log-likelihood, with the difference that one needs to ensure (Cont3), respectively. Further, it is noted that since the claim is an asymptotic property, the quantities $\tilde{\boldsymbol{\beta}}_{nf}$ and $\tilde{\boldsymbol{\beta}}_f$ are sequences in n, where the values of the entries may change with increasing n, as long as they satisfy the construction provided in Theorem 2.3.15, that is, they coincide except for two *fixed* entries.
By straightforward calculations using the construction of $\tilde{\boldsymbol{\beta}}_f$ and $\tilde{\boldsymbol{\beta}}_{nf}$ similar to the proof of Theorem 2.3.15 one can deduce

$$\begin{aligned}
& M_{pen}(\tilde{\boldsymbol{\beta}}_{nf}) - M_{pen}(\tilde{\boldsymbol{\beta}}_f) \\
= \;& -L_n(\tilde{\boldsymbol{\beta}}_{nf}) + \lambda_0^n \sum_{j=1}^{J_n} \sum_{0\leq t<s\leq p_j} ||\tilde{\beta}_{nf,j,t} - \tilde{\beta}_{nf,j,s}||_0 + \lambda_1^n \sum_{j=1}^{J_n} ||\tilde{\boldsymbol{\beta}}_{nf,j}||_{K_j} \\
& - \left(-L_n(\tilde{\boldsymbol{\beta}}_f) + \lambda_0^n \sum_{j=1}^{J_n} \sum_{0\leq t<s\leq p_j} ||\tilde{\beta}_{f,j,t} - \tilde{\beta}_{f,j,s}||_0 + \lambda_1^n \sum_{j=1}^{J_n} ||\tilde{\boldsymbol{\beta}}_{f,j}||_{K_j} \right) \\
= \;& \underbrace{L_n(\tilde{\boldsymbol{\beta}}_f) - L_n(\tilde{\boldsymbol{\beta}}_{nf})}_{\text{(I)}} + \underbrace{\lambda_0^n w_0^{\iota,r}}_{\text{(II)}} + \underbrace{\lambda_1^n \left(||\tilde{\boldsymbol{\beta}}_{nf,\iota}||_{K_\iota} - ||\tilde{\boldsymbol{\beta}}_f||_{K_\iota} \right)}_{\text{(III)}}.
\end{aligned}$$

Analyzing (I), by the continuity of the log-likelihood function given by (Cont3), it follows that

$$\forall \epsilon_1 > 0 \, \exists \, \delta(\epsilon_1, \tilde{\boldsymbol{\beta}}_{nf}) \, : \, \forall \tilde{\boldsymbol{\beta}}_f \, , \, ||\tilde{\boldsymbol{\beta}}_{nf} - \tilde{\boldsymbol{\beta}}_f||_2 < \delta(\epsilon_1, \tilde{\boldsymbol{\beta}}_{nf}) \, : \, ||L_n(\tilde{\boldsymbol{\beta}}_{nf}) - L_n(\tilde{\boldsymbol{\beta}}_f)||_2 < \epsilon_1.$$

It is important that this $\delta(\epsilon_1, \tilde{\boldsymbol{\beta}}_{nf})$ is valid for all functions $-L_n(\cdot)$, thus for all $n \in \mathbb{N}$, so $\delta(\epsilon_1, \tilde{\boldsymbol{\beta}}_{nf})$ does not depend on n, see Definition B.3.1 (Appendix B.3). For (II) it is known that

$$\inf_{n\in\mathbb{N}} \lambda_0^n w_0^{\iota,r} = c > 0.$$

Finally, for (III), the continuity of the norm $||\cdot||_{K_\iota} : \mathbb{R}^{p_\iota} \to \mathbb{R}$ is used which yields

$$\forall \epsilon_2 > 0 \, \exists \, \delta(\epsilon_2, \tilde{\boldsymbol{\beta}}_{nf,\iota}) \, : \, \forall \tilde{\boldsymbol{\beta}}_{f,\iota} \, , \, ||\tilde{\boldsymbol{\beta}}_{nf,\iota} - \tilde{\boldsymbol{\beta}}_{f,\iota}||_2 < \delta(\epsilon_2, \tilde{\boldsymbol{\beta}}_{nf,\iota}) \, : \, ||\tilde{\boldsymbol{\beta}}_{nf,\iota}||_{K_\iota} - ||\tilde{\boldsymbol{\beta}}_{f,\iota}||_{K_\iota} < \epsilon_2.$$

Now, let ϵ_1, ϵ_2 be chosen such that

$$0 < \epsilon_1 < \frac{1}{2} \inf_{n \in \mathbb{N}} \lambda_0^n \inf_{n \in \mathbb{N}} w_0^{(\iota,r)} \tag{2.32}$$

$$0 < \epsilon_2 < \frac{1}{2} \inf_{n \in \mathbb{N}} \lambda_0^n \inf_{n \in \mathbb{N}} w_0^{(\iota,r)} \frac{1}{\sup_{n \in \mathbb{N}} \lambda_1^n}, \tag{2.33}$$

where all expressions on the right hand side of the inequalities exist and are nonzero by assumption. For those ϵ_1, ϵ_2, it is known by the continuity that there exist $\delta(\epsilon_1, \tilde{\boldsymbol{\beta}}_{nf}), \delta(\epsilon_2, \tilde{\boldsymbol{\beta}}_{nf,\iota})$ such that for all $\tilde{\boldsymbol{\beta}}_f$ with $||\tilde{\boldsymbol{\beta}}_{nf} - \tilde{\boldsymbol{\beta}}_f||_2 < \min\{\delta(\epsilon_1, \tilde{\boldsymbol{\beta}}_{nf}), \delta(\epsilon_2, \tilde{\boldsymbol{\beta}}_{nf,\iota})\}$ one can infer

$$M_{pen}(\tilde{\boldsymbol{\beta}}_{nf}) - M_{pen}(\tilde{\boldsymbol{\beta}}_f) > -\epsilon_1 - \lambda_1^n \epsilon_2 + \lambda_0^n \cdot w_0^{(\iota,r)} \overset{(*)}{>} 0 \tag{2.34}$$

where $(*)$ holds by (2.32) and (2.33), that is, $\forall n \in \mathbb{N}$ it holds that

$$\begin{aligned}
-\epsilon_1 &> -\frac{1}{2} \inf_{n \in \mathbb{N}} \lambda_0^n \inf_{n \in \mathbb{N}} w_0^{(\iota,r)} \\
\epsilon_2 &< \frac{1}{2} \inf_{n \in \mathbb{N}} \lambda_0^n \inf_{n \in \mathbb{N}} w_0^{(\iota,r)} \frac{1}{\sup_{n \in \mathbb{N}} \lambda_1^n} \leq \frac{1}{2} \inf_{n \in \mathbb{N}} \lambda_0^n \inf_{n \in \mathbb{N}} w_0^{(\iota,r)} \frac{1}{\lambda_1^n} \\
\Rightarrow -\epsilon_2 &> -\frac{1}{2} \cdot \frac{1}{\lambda_1^n} \cdot \inf_{n \in \mathbb{N}} \lambda_0^n \inf_{n \in \mathbb{N}} w_0^{(\iota,r)},
\end{aligned}$$

and consequently

$$\begin{aligned}
-\epsilon_1 - \lambda_1^n \epsilon_2 + \lambda_0^n w_0^{(\iota,r)} &> -\frac{1}{2} \inf_{n \in \mathbb{N}} \lambda_0^n \inf_{n \in \mathbb{N}} w_0^{(\iota,r)} - \lambda_1^n \frac{1}{2} \cdot \frac{1}{\lambda_1^n} \cdot \inf_{n \in \mathbb{N}} \lambda_0^n \inf_{n \in \mathbb{N}} w_0^{(\iota,r)} + \lambda_0^n w_0^{(\iota,r)} \\
&= -\inf_{n \in \mathbb{N}} \lambda_0^n \inf_{n \in \mathbb{N}} w_0^{(\iota,r)} + \lambda_0^n w_0^{(\iota,r)} \\
&> 0,
\end{aligned}$$

so $(*)$ is justified.

Now, with this choice of ϵ_1, ϵ_2, inequality (2.34) holds for all $\tilde{\boldsymbol{\beta}}_f, \tilde{\boldsymbol{\beta}}_{nf}$ for which

$$||\tilde{\boldsymbol{\beta}}_f - \tilde{\boldsymbol{\beta}}_{nf}||_2 < \min\{\delta(\epsilon_1, \tilde{\boldsymbol{\beta}}_{nf}), \delta(\epsilon_2, \tilde{\boldsymbol{\beta}}_{nf,\iota})\}.$$

By assumption

$$\tilde{\beta}_{nf,\iota,r} - \tilde{\beta}_{nf,\iota,r-1} = \Delta_n \to 0$$

and by construction $||\tilde{\boldsymbol{\beta}}_f - \tilde{\boldsymbol{\beta}}_{nf}||_2 \leq \Delta_n \to 0$, so there exists $n_0 \in \mathbb{N}$ such that $||\tilde{\boldsymbol{\beta}}_f - \tilde{\boldsymbol{\beta}}_{nf}||_2 < \min\{\delta(\epsilon_1, \tilde{\boldsymbol{\beta}}_{nf}), \delta(\epsilon_2, \tilde{\boldsymbol{\beta}}_{nf,\iota})\} \; \forall \, n \geq n_0$, so inequality (2.34) holds $\forall \, n \geq n_0$ and the claim follows. The proof can directly be extended to the case of ι being a nominal factor. □

Remark 2.3.18 (Equicontinuity)**.** In the proof of Theorem 2.3.17, it is crucial that $\delta(\epsilon_1, \tilde{\boldsymbol{\beta}}_{nf})$ does not depend on n, otherwise it may happen that $\delta(\epsilon_1, \tilde{\boldsymbol{\beta}}_{nf})$ decreases for increasing n and this would cause problems if it decreases faster than $||\tilde{\boldsymbol{\beta}}_f - \tilde{\boldsymbol{\beta}}_{nf}||_2 \leq \Delta_n$ goes to zero. This justifies why it is necessary to assume that the family $\mathscr{F} = \{L_n(\boldsymbol{\beta}), \, n \in \mathbb{N}\}$ of functions is equicontinuous, one compares (Cont2) and (Cont3) of Definition B.3.1 (Appendix B.3). The quantity $\delta(\epsilon_2, \tilde{\boldsymbol{\beta}}_{nf})$ does not depend on n either, since it arises from the continuity of the norm $||\cdot||_{\boldsymbol{K}_\iota} : \mathbb{R}^{p_\iota} \to \mathbb{R}$. One can argue that $||\cdot||_{\boldsymbol{K}_\iota}$ may depend on n though some adaptive weight.

If this is the case, one needs to ensure equicontinuity of the family of functions in the same way as done for the log-likelihood functions in (Cont2) and the proof works completely similar. (Cont3) is not necessary, because the definition space of the function $||\cdot||_{\boldsymbol{K}_\iota} : \mathbb{R}^{p_\iota} \to \mathbb{R}$ that is considered does not depend on n. To conclude, since it is common for group lasso to impose $\sqrt{p_j}$ as a weight, or, using adaptive weights, the inverse of a consistent initial estimator, such as MLE (Theorem C.1.2), the continuity of the norm as used in the proof of Theorem 2.3.17 is sufficient.

Remark 2.3.19 (Theorem 2.3.17 for random variables)**.** In case of Theorem 2.3.17, one can analogously consider $\tilde{\boldsymbol{\beta}}_{nf}$ and $\tilde{\boldsymbol{\beta}}_f$, as well as $w_0^{(\iota,r)}$ as sequences of random variables. The requirements of the theorem are adjusted to hold a.s., that is $\mathbb{P}\left(\lim_{n\to\infty} \Delta_n = 0\right) = 1$, which is for short $\Delta_n \to_{a.s} 0$, and further $\mathbb{P}\left(\inf_{n\in\mathbb{N}} w_0^{(\iota,r)} = c\right) = 1$. The other requirements are not changed, since they do not include a random variable, so one similarly assumes $\inf_{n\in\mathbb{N}} \lambda_0^n = c > 0$, $\sup_{n\in\mathbb{N}} \lambda_1^n = c < \infty$ and (Cont2) or (Cont3), respectively. To sum up, all assumptions of Theorem 2.3.17 hold (at least) with probability one, and consequently one can deduce that the statement holds with probability one, in particular

$$\mathbb{P}\left(\exists n_0 \in \mathbb{N} : \forall n \geq n_0 : M_{pen}^{(L_0\text{-FGL})}(\tilde{\boldsymbol{\beta}}_{nf}) > M_{pen}^{(L_0\text{-FGL})}(\tilde{\boldsymbol{\beta}}_f)\right) = 1.$$

Remark 2.3.20 (On the assumptions of Theorem 2.3.15 and Theorem 2.3.17)**.**

i) In Theorem 2.3.15, it is only required that the continuity assumption (Cont1) holds for the *log-likelihood* function.

ii) In Theorem 2.3.17, where $n \to \infty$ is considered, it is required that the continuity assumptions for diverging n hold, i.e. the assumptions for the family of functions $\mathscr{F} = \{L_n(\boldsymbol{\beta}), n \in \mathbb{N}\}$, thus (Cont2) for fixed p and (Cont3) for diverging p_n.
Further, the sizes of the infimum and supremum of the tuning parameters λ_0^n, λ_1^n and the weight $w_0^{(\iota,r)}$ are controlled. These are straightforward and reasonable assumptions, since, if one would neglect to require the infimum of the tuning for fusion λ_0^n to be greater than zero, one could clearly *not* expect that the penalty performs some fusion. Further it is needed to ensure that the weight $w_0^{(\iota,r)}$ of the L_0 norm for the corresponding factor is bounded from below, ensuring the impact of the L_0 fusion part. Finally, the tuning part for factor selection needs to be bounded from above.

To conclude, the assumptions of Theorem 2.3.15 and Theorem 2.3.17 are reasonable and not restrictive.

Since in the proof of Theorem 2.3.17 the actual (numerical) value of the sample size n is not used, one can directly state the following Corollary 2.3.23 about a similar result for subsequences. To do so, the following notation is introduced.

Notation 2.3.21 (Subsequences)**.** In general, the quantity $\hat{\boldsymbol{\beta}}$ is defined as being a minimizer of the objective function $M_{pen}(\boldsymbol{\beta})$ with some penalty function $P_\lambda(\boldsymbol{\beta})$, where these quantities all depend on the sample size n. However, a lower index of n was neglected so far to avoid unnecessary notational complexity. Interpreting $\hat{\boldsymbol{\beta}}$ as a sequence of the sample size n, subsequences of $\hat{\boldsymbol{\beta}}$ are considered introducing the following notation.

i) *Square brackets.* First, write $[\hat{\boldsymbol{\beta}}]_n$ for the estimator $\hat{\boldsymbol{\beta}}$ based on a sample of size n. The same is done for the objective function $M_{pen}(\boldsymbol{\beta})$ based on a sample of size n, that is, $[M_{pen}(\boldsymbol{\beta})]_n$, as well as possibly for the weights, e.g. $[w_0^{(j,rs)}]_n$. Hence, the square brackets around the quantities indicate on which sample size the quantity is based on.

ii) *Round brackets.* Second, by writing round brackets around a sequence of quantities with a lower index n, it is indicated that the n-th element of this sequence is picked.

With this notation, one can construct sub-sequences $((\hat{\boldsymbol{\beta}}^{sub})_n)_{n\in\mathbb{N}}$ of $((\hat{\boldsymbol{\beta}})_n)_{n\in\mathbb{N}}$ as follows. Let $(a_n)_{n\in\mathbb{N}} \subseteq \mathbb{N}$ be a strictly monotonic increasing sequence. One defines the sub-sequence $\hat{\boldsymbol{\beta}}^{sub}$ of $\hat{\boldsymbol{\beta}}$ as

$$(\hat{\boldsymbol{\beta}}^{sub})_n := (\hat{\boldsymbol{\beta}})_{a_n} \ \ (= [\hat{\boldsymbol{\beta}}]_{a_n}),$$

hence the n-th element of the sub-sequence $\hat{\boldsymbol{\beta}}^{sub}$, denoted by $(\hat{\boldsymbol{\beta}}^{sub})_n$, picks the element of $\hat{\boldsymbol{\beta}}$ based on a sample of size a_n, hence, in this case, the a_n-th element $(\hat{\boldsymbol{\beta}})_{a_n}$. It is noted that for the "original" $\hat{\boldsymbol{\beta}}$ it holds that $(\hat{\boldsymbol{\beta}})_{a_n} = [\hat{\boldsymbol{\beta}}]_{a_n}$ so if one does not consider any subsequence, it does not matter if square or round brackets are used. However, this equality does not need to hold for the sub-sequence $\hat{\boldsymbol{\beta}}^{sub}$, since for $\hat{\boldsymbol{\beta}}^{sub}$ the sample size is not necessarily equal to the "number of the element" of the sequence. By construction it holds that $[\hat{\boldsymbol{\beta}}^{sub}]_{a_n} = (\hat{\boldsymbol{\beta}}^{sub})_n$, so the n-th element of the sub-sequence $\hat{\boldsymbol{\beta}}^{sub}$ is based on a_n samples. To clarify this notation, an example of the construction is given in Example 2.3.22. This notation is only used whenever sub-sequences are utilized or whenever it is needed for clarity. Otherwise, the index n and the brackets are omitted in all quantities except for the tuning parameters for simplicity of notation, as mentioned at the beginning of the thesis.

Example 2.3.22. One considers the strictly monotonic increasing sequence $(a_n)_{n\in\mathbb{N}} \subseteq \mathbb{N}$ given as

$$(a_n)_{n\in\mathbb{N}} := (2, 4, 5, 6, 7, 8, 11, \ldots)$$

and one defines $(\hat{\boldsymbol{\beta}}^{sub})_n := [\hat{\boldsymbol{\beta}}]_{a_n}$ following Notation 2.3.21. The construction of the sub-sequence is visualized in Table 2.1, where in the two respective columns on the right hand side of the sequence and the sub-sequence, the corresponding number of the element of the sequence (sub-sequence) as well as the sample size of the element of the sequence (sub-sequence) are given.

seq. $((\hat{\boldsymbol{\beta}})_n)_{n\in\mathbb{N}}$	number of element	sample size of element	sub-seq. $((\hat{\boldsymbol{\beta}}^{sub})_n)_{n\in\mathbb{N}}$	number of element	sample size of element
$(\hat{\boldsymbol{\beta}})_1$	1	1			
$(\hat{\boldsymbol{\beta}})_2$	2	2	$(\hat{\boldsymbol{\beta}}^{sub})_1$	1	2 ($a_1 = 2$)
$(\hat{\boldsymbol{\beta}})_3$	3	3			
$(\hat{\boldsymbol{\beta}})_4$	4	4	$(\hat{\boldsymbol{\beta}}^{sub})_2$	2	4 ($a_2 = 4$)
$(\hat{\boldsymbol{\beta}})_5$	5	5	$(\hat{\boldsymbol{\beta}}^{sub})_3$	3	5 ($a_3 = 5$)
$(\hat{\boldsymbol{\beta}})_6$	6	6	$(\hat{\boldsymbol{\beta}}^{sub})_4$	4	6 ($a_4 = 6$)
$(\hat{\boldsymbol{\beta}})_7$	7	7	$(\hat{\boldsymbol{\beta}}^{sub})_5$	5	7 ($a_5 = 7$)
$(\hat{\boldsymbol{\beta}})_8$	8	8	$(\hat{\boldsymbol{\beta}}^{sub})_6$	6	8 ($a_6 = 8$)
$(\hat{\boldsymbol{\beta}})_9$	9	9			
$(\hat{\boldsymbol{\beta}})_{10}$	10	10			
$(\hat{\boldsymbol{\beta}})_{11}$	11	11	$(\hat{\boldsymbol{\beta}}^{sub})_7$	7	11 ($a_7 = 11$)

Table 2.1: *Construction of sub-sequence $((\hat{\boldsymbol{\beta}}^{sub})_n)_{n\in\mathbb{N}}$ from the sequence $((\hat{\boldsymbol{\beta}})_n)_{n\in\mathbb{N}}$ with $(\hat{\boldsymbol{\beta}}^{sub})_n := [\hat{\boldsymbol{\beta}}]_{a_n}$ and $(a_n)_{n\in\mathbb{N}}$ given in Example 2.3.22.*

This example emphasizes that the following equalities hold with $(\hat{\boldsymbol{\beta}}^{sub})_n := [\hat{\boldsymbol{\beta}}]_{a_n}$

$$(\hat{\boldsymbol{\beta}})_n = [\hat{\boldsymbol{\beta}}]_n, \ (\hat{\boldsymbol{\beta}}^{sub})_n = [\hat{\boldsymbol{\beta}}^{sub}]_{a_n} = (\hat{\boldsymbol{\beta}})_{a_n}.$$

To conclude, with the sub-sequence $((\hat{\boldsymbol{\beta}}^{sub})_n)_{n\in\mathbb{N}}$, one picks the elements of the sequence $((\hat{\boldsymbol{\beta}})_n)_{n\in\mathbb{N}}$ with the sample sizes given by the elements of the strictly monotonic increasing sequence $(a_n)_{n\in\mathbb{N}} \subseteq \mathbb{N}$.

The following corollary gives an analogous statement to Theorem 2.3.17 applied to sub-sequences, with the distinction that in Corollary 2.3.23 it is only required that the difference $\tilde{\beta}_{nf,\iota,r} - \tilde{\beta}_{nf,\iota,r-1}$ (with objective function etc. based on a_n samples), denoted by Δ_{a_n}, goes to zero, hence the same requirement as in Theorem 2.3.17 applied to a sub-sequence. It is shown that in this case, the decrease in the objective function holds for this sub-sequence.

Corollary 2.3.23 (Fusion Property objective function II of L_0-FGL for sub-sequences, fixed p and diverging p_n)**.** Let c denote any constant. For any factor $\iota \in \{1, \ldots, J_n\}$ (ordinal without loss of generality, one compares Theorem 2.3.15), one defines $((\tilde{\boldsymbol{\beta}}_{nf})_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}^{p_n+1}$ and $((\tilde{\boldsymbol{\beta}}_{f})_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}^{p_n+1}$ completely analogous to Theorem 2.3.15. Let $(a_n)_{n\in\mathbb{N}} \subseteq \mathbb{N}$ be a strictly monotonic increasing sequence and, based on this sequence, one defines sub-sequences of $\tilde{\boldsymbol{\beta}}_{nf}, \tilde{\boldsymbol{\beta}}_{f}$ from Theorem 2.3.17 as follows

$$\left(\left(\tilde{\boldsymbol{\beta}}_{nf}^{sub}\right)_n\right)_{n\in\mathbb{N}} := \left(\left(\tilde{\boldsymbol{\beta}}_{nf}\right)_{a_n}\right)_{n\in\mathbb{N}}, \quad \left(\left(\tilde{\boldsymbol{\beta}}_{f}^{sub}\right)_n\right)_{n\in\mathbb{N}} := \left(\left(\tilde{\boldsymbol{\beta}}_{f}\right)_{a_n}\right)_{n\in\mathbb{N}}. \tag{2.35}$$

One assumes that

$$[\tilde{\beta}_{nf,\iota,r}^{sub}]_{a_n} - [\tilde{\beta}_{nf,\iota,r-1}^{sub}]_{a_n} =: \Delta_{a_n} \to 0 \, (n \to \infty \text{ so } a_n \to \infty)$$

and for the weight $[w_0^{(\iota,r)}]_n$ one supposes $\inf_{n\in\mathbb{N}} [w_0^{(\iota,r)}]_n = c > 0$. Further, let $\inf_{n\in\mathbb{N}} \lambda_0^n = c > 0$, $\sup_{n\in\mathbb{N}} \lambda_1^n = c < \infty$ and one assumes that (Cont2) of Definition B.3.1 (Appendix B.3) is satisfied for the case of fixed p and (Cont3) for the case of diverging p_n. Then, there exists some $n_0 \in \mathbb{N}$ such that $\forall a_n \geq a_{n_0}$, the value of the objective function $[M_{pen}^{(L_0\text{-FGL})}(\cdot)]_{a_n}$ decreases $\forall a_n \geq a_{n_0}$ if coefficients $[\tilde{\beta}_{nf,\iota,r}^{sub}]_{a_n}$ and $[\tilde{\beta}_{nf,\iota,r-1}^{sub}]_{a_n}$ are fused, hence

$$[M_{pen}^{(L_0\text{-FGL})}(\tilde{\boldsymbol{\beta}}_{nf}^{sub})]_{a_n} > [M_{pen}^{(L_0\text{-FGL})}(\tilde{\boldsymbol{\beta}}_{f}^{sub})]_{a_n} \; \forall a_n \geq a_{n_0}.$$

Proof. For simplicity of notation, $M_{pen}(\cdot)$ is written for $M_{pen}^{(L_0\text{-FGL})}(\cdot)$ in the following proof. Initially, the following comments are supplied.

i) By construction of the sub-sequences, the n-th element of the sub-sequence $((\tilde{\boldsymbol{\beta}}_{nf}^{sub})_n)_{n\in\mathbb{N}}$, denoted by $(\tilde{\boldsymbol{\beta}}_{nf}^{sub})_n$, is based on a_n samples (hence $(\tilde{\boldsymbol{\beta}}_{nf}^{sub})_n = [\tilde{\boldsymbol{\beta}}_{nf}^{sub}]_{a_n}$) and is of dimension $p_{a_n} + 1$. The same holds for $((\tilde{\boldsymbol{\beta}}_{f}^{sub})_n)_{n\in\mathbb{N}}$.

ii) It is only required that $[\tilde{\beta}_{nf,\iota,r}^{sub}]_{a_n} - [\tilde{\beta}_{nf,\iota,r-1}^{sub}]_{a_n} =: \Delta_{a_n} \to 0 \, (n \to \infty)$ for the *sub-sequence*, hence for every a_n-th element of $\tilde{\beta}_{nf,\iota,r}$ and $\tilde{\beta}_{nf,\iota,r-1}$, respectively. However, the assumptions on the weights and the tuning parameters are made for all $n \in \mathbb{N}$ and not only for sub-sequences. The reason for that is consistency with the other theorems, since in the proofs where this Corollary 2.3.23 is needed, the requirements are already given for all $n \in \mathbb{N}$, hence no sharper version of the assumptions is necessary.

The proof is similar to the proof of Theorem 2.3.17, nevertheless selected steps are executed again for adaption to notation (Notation 2.3.21). Let ϵ_1, ϵ_2 be given such that (2.32) and (2.33) hold and let $n \in \mathbb{N}$. For those ϵ_1, ϵ_2, it is known by the continuity, similar to the proof of

Theorem 2.3.17, that there exist $\delta(\epsilon_1, \tilde{\boldsymbol{\beta}}_{nf}^{sub}), \delta(\epsilon_2, \tilde{\boldsymbol{\beta}}_{nf,\iota}^{sub})$ such that for all $[\tilde{\boldsymbol{\beta}}_f^{sub}]_{a_n}, [\tilde{\boldsymbol{\beta}}_{nf}^{sub}]_{a_n}$ with $||[\tilde{\boldsymbol{\beta}}_{nf}^{sub}]_{a_n} - [\tilde{\boldsymbol{\beta}}_f^{sub}]_{a_n}||_2 < \min\{\delta(\epsilon_1, \tilde{\boldsymbol{\beta}}_{nf}^{sub}), \delta(\epsilon_2, \tilde{\boldsymbol{\beta}}_{nf,\iota}^{sub})\}$ one obtains

$$\begin{aligned}
& [M_{pen}(\tilde{\boldsymbol{\beta}}_{nf}^{sub})]_{a_n} - [M_{pen}(\tilde{\boldsymbol{\beta}}_f^{sub})]_{a_n} \\
\overset{(2.30)}{=} \; & [L_n(\tilde{\boldsymbol{\beta}}_f^{sub})]_{a_n} - [L_n(\tilde{\boldsymbol{\beta}}_{nf}^{sub})]_{a_n} + \lambda_{1,a_n} ||[\tilde{\boldsymbol{\beta}}_{nf,\iota}^{sub}]_{a_n}||_{\boldsymbol{K}_\iota} - \lambda_{1,a_n} ||[\tilde{\boldsymbol{\beta}}_{f,\iota}^{sub}]_{a_n}||_{\boldsymbol{K}_\iota} + \lambda_{0,a_n} \cdot [w_0^{(\iota,r)}]_{a_n} \\
> \; & -\epsilon_1 - \lambda_{1,a_n}\epsilon_2 + \lambda_{0,a_n} \cdot [w_0^{(\iota,r)}]_{a_n} \\
\overset{(*)}{>} \; & 0,
\end{aligned}$$

where $(*)$ holds by (2.32) and (2.33). To be more precise, the justification for $(*)$ is as follows: since

$$\begin{aligned}
(2.32) : 0 < \epsilon_1 < \frac{1}{2} \inf_{n\in\mathbb{N}} \lambda_0^n \inf_{n\in\mathbb{N}} [w_0^{(\iota,r)}]_n \leq \frac{1}{2} \inf_{n\in\mathbb{N}} \lambda_{0,a_n} \inf_{n\in\mathbb{N}} [w_0^{(\iota,r)}]_{a_n} \leq \frac{1}{2} \lambda_{0,a_n} [w_0^{(\iota,r)}]_{a_n} \\
(2.33) : 0 < \epsilon_2 < \frac{1}{2} \inf_{n\in\mathbb{N}} \lambda_0^n \inf_{n\in\mathbb{N}} [w_0^{(\iota,r)}]_n \frac{1}{\sup_{n\in\mathbb{N}} \lambda_1^n} \leq \frac{1}{2} \inf_{n\in\mathbb{N}} \lambda_{0,a_n} \inf_{n\in\mathbb{N}} [w_0^{(\iota,r)}]_{a_n} \frac{1}{\sup_{n\in\mathbb{N}} \lambda_{1,a_n}} \\
\leq \frac{1}{2} \lambda_{0,a_n} [w_0^{(\iota,r)}]_{a_n} \frac{1}{\lambda_{1,a_n}},
\end{aligned}$$

one infers ($\lambda_{1,a_n} > 0$)

$$\begin{aligned}
-\epsilon_1 \; &> \; -\frac{1}{2} \lambda_{0,a_n} [w_0^{(\iota,r)}]_{a_n} \\
-\lambda_{1,a_n}\epsilon_2 \; &> \; -\frac{1}{2} \lambda_{0,a_n} [w_0^{(\iota,r)}]_{a_n}
\end{aligned}$$

and finally

$$-\epsilon_1 - \lambda_{1,a_n}\epsilon_2 + \lambda_{0,a_n} \cdot [w_0^{(\iota,r)}]_{a_n} > -\frac{1}{2} \lambda_{0,a_n} [w_0^{(\iota,r)}]_{a_n} - \frac{1}{2} \lambda_{0,a_n} [w_0^{(\iota,r)}]_{a_n} + \lambda_{0,a_n} \cdot [w_0^{(\iota,r)}]_{a_n} > 0,$$

so $(*)$ is justified. Now, with this choice of ϵ_1, ϵ_2, it is known that

$$[M_{pen}(\tilde{\boldsymbol{\beta}}_{nf}^{sub})]_{a_n} - [M_{pen}(\tilde{\boldsymbol{\beta}}_f^{sub})]_{a_n} > 0 \quad (**)$$

holds for all $[\tilde{\boldsymbol{\beta}}_f^{sub}]_{a_n}, [\tilde{\boldsymbol{\beta}}_{nf}^{sub}]_{a_n}$ for which $||[\tilde{\boldsymbol{\beta}}_f^{sub}]_{a_n} - [\tilde{\boldsymbol{\beta}}_{nf}^{sub}]_{a_n}||_2 < \min\{\delta(\epsilon_1, \tilde{\boldsymbol{\beta}}_{nf}^{sub}), \delta(\epsilon_2, \tilde{\boldsymbol{\beta}}_{nf,\iota}^{sub})\}$. By assumption, it holds that $[\tilde{\beta}_{nf,\iota,r}^{sub}]_{a_n} - [\tilde{\beta}_{nf,\iota,r-1}^{sub}]_{a_n} = \Delta_{a_n} \to 0$ and by construction

$$||[\tilde{\boldsymbol{\beta}}_f^{sub}]_{a_n} - [\tilde{\boldsymbol{\beta}}_{nf}^{sub}]_{a_n}||_2 \leq \Delta_{a_n} \to 0,$$

so there exists $n_0 \in \mathbb{N}$ such that

$$||[\tilde{\boldsymbol{\beta}}_f^{sub}]_{a_n} - [\tilde{\boldsymbol{\beta}}_{nf}^{sub}]_{a_n}||_2 < \min\{\delta(\epsilon_1, \tilde{\boldsymbol{\beta}}_{nf}^{sub}), \delta(\epsilon_2, \tilde{\boldsymbol{\beta}}_{nf,\iota}^{sub})\} \;\; \forall\, a_n \geq a_{n_0}.$$

Hence, inequality $(**)$ holds $\forall\, a_n \geq a_{n_0}$. Consequently

$$[M_{pen}(\tilde{\boldsymbol{\beta}}_{nf}^{sub})]_{a_n} > [M_{pen}(\tilde{\boldsymbol{\beta}}_f^{sub})]_{a_n} \;\; \forall a_n \geq a_{n_0},$$

which was to be shown. The proof can directly be extended to the nominal case, see explanations in the proof of Theorem 2.3.15. □

Remark 2.3.24 (Corollary 2.3.23 for random variables)**.** Similarly to Remark 2.3.19, Corollary 2.3.23 can be transferred to the case considering $\tilde{\boldsymbol{\beta}}_{nf}$, $\tilde{\boldsymbol{\beta}}$ and $w_0^{(\iota,r)}$ as sequences of random variables assuming $\mathbb{P}\left(\lim_{n\to\infty} \Delta_{a_n} = 0\right) = 1$, which is for short $\Delta_{a_n} \to_{a.s.} 0$, and further $\mathbb{P}\left(\inf_{n\in\mathbb{N}} [w_0^{(\iota,r)}]_n = c\right) = 1$. The requirements that do not depend on random variables are not changed, i.e. $\inf_{n\in\mathbb{N}} \lambda_0^n = c > 0$, $\sup_{n\in\mathbb{N}} \lambda_1^n = c < \infty$ and (Cont2) or (Cont3), respectively, are assumed. Then, the statement of Corollary 2.3.23 holds with probability one, i.e.

$$\mathbb{P}\left(\exists\, n_0 \,:\, \forall a_n \geq a_{n_0} \,:\, [M_{pen}(\tilde{\boldsymbol{\beta}}_{nf}^{sub})]_{a_n} > [M_{pen}(\tilde{\boldsymbol{\beta}}_{f}^{sub})]_{a_n}\right) = 1.$$

Fusion Properties of the Estimates

After showing in Theorems 2.3.15 and 2.3.17 that the penalty function decreases when particular coefficient values are set to be equal, with the transfer to random variables given in Remarks 2.3.16, 2.3.19 and 2.3.24, results are inferred about $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$ being a *minimizer* of the objective function $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$.

Before the next theorem, which is Theoerm 2.3.26, is supplied, the following remark is provided on its assumptions.

Remark 2.3.25. In Theorem 2.3.26, the same assumptions as in Theorems 2.3.15 and 2.3.17 are adopted, adding a minimal neighborhood condition which is introduced and specified in Definition B.4.1 of Appendix B.4. The latter is needed to ensure that the radius of the neighborhood, with respect to which the local minimizer is considered, is bounded from below (over $n \in \mathbb{N}$).

Even though no sub-sequences are constructed in Theorem 2.3.26, Notation 2.3.21 is used, in particular the square brackets, to emphasize on which sample size the estimate is based on, since this clarifies the theorem and the steps of the proof. However, it holds that $\left([\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}]_n\right)_{n\in\mathbb{N}} = \left((\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})})_n\right)_{n\in\mathbb{N}}$, thus the n-th element of the sequence of estimates is based on n samples.

Finally, Theorem 2.3.26 below shows, knowing that for increasing n the estimates of two entries of an L_0-FGL estimate converge to each other, they are set to be equal at some $n_0 \in \mathbb{N}$. After the proof, Remark 2.3.27 explains the transfer from estimates to estimators, thus considering $([\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}]_n)_{n\in\mathbb{N}}$ and the weight $([w_0^{(\iota,r)}]_n)_{n\in\mathbb{N}}$ as sequences of random variables. However, in Theorem 2.3.26 the quantity $\left([\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}]_n\right)_{n\in\mathbb{N}} \subseteq \mathbb{R}^{p_n+1}$ (or $\subseteq \mathbb{R}^{p+1}$ in the fixed case, respectively) is considered as a sequence of estimates, further the sequence of weights $([w_0^{(\iota,r)}]_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}$ is treated as non-random.

Theorem 2.3.26 (Fusion properties of estimates, fixed p and diverging p_n)**.** Let c denote any constant and let $\iota \in \{1, \ldots, J_n\}$ be a factor (ordinal without loss of generality) and $r \in \{1, \ldots, p_\iota\}$. Let $\left([\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}]_n\right)_{n\in\mathbb{N}}$ be any sequence of (local) minimizers of $[M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})]_n$, respectively, based on $n \in \mathbb{N}$ samples, for which it is assumed that

$$[\hat{\beta}_{\iota,r}^{(L_0\text{-FGL})}]_n - [\hat{\beta}_{\iota,r-1}^{(L_0\text{-FGL})}]_n =: \Delta_n \to 0\,(n \to \infty).$$

Further, for the weight $[w_0^{(\iota,r)}]_n$ it is assumed that $\inf_{n\in\mathbb{N}} [w_0^{(\iota,r)}]_n = c > 0$ and additionally $\inf_{n\in\mathbb{N}} \lambda_0^n = c > 0$, as well as $\sup_{n\in\mathbb{N}} \lambda_1^n = c < \infty$. Finally, let (Cont2) be satisfied for the case of fixed p and

(Cont3) for the case of diverging p_n, as well as the minimal neighborhood condition of Definition B.4.1 (Appendix B.4). Then, for this given sequence $\left([\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}]_n\right)_{n\in\mathbb{N}}$, it holds that

$$\exists\, n_0 \in \mathbb{N} \,:\, \forall n \geq n_0 \,:\, [\hat{\beta}_{\iota,r}^{(L_0\text{-FGL})}]_n = [\hat{\beta}_{\iota,r-1}^{(L_0\text{-FGL})}]_n,$$

so departing from this n_0, levels r and $r-1$ of factor ι are fused.

Proof. For simplicity of notation, $[\hat{\boldsymbol{\beta}}]_n$ is written for $[\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}]_n$ and $[M_{pen}(\cdot)]_n$ for $[M_{pen}^{(L_0\text{-FGL})}(\cdot)]_n$ throughout this proof. It is noted that $([\hat{\boldsymbol{\beta}}]_n)_{n\in\mathbb{N}}$ is a sequence of (local) minimizers of $[M_{pen}^{(L_0\text{-FGL})}(\cdot)]_n$, respectively, for which $\Delta_n \to 0$. Hence, the sequence $([\hat{\boldsymbol{\beta}}]_n)_{n\in\mathbb{N}}$ is some given, fixed sequence of (local) minimizers.

It is supposed that the statement of the theorem is not true, that is, with $\nu_1 \in \mathbb{N}$

$$\forall n_0 \in \mathbb{N}\ \exists\, \nu_1 \geq n_0 \,:\, [\hat{\beta}_{\iota,r-1}]_{\nu_1} \neq [\hat{\beta}_{\iota,r}]_{\nu_1}. \tag{2.36}$$

Let some $n_0 \in \mathbb{N}$ be arbitrary but given. The following points are considered.

i) First, one defines some quantities based on sample size $\nu_1 \geq n_0$ satisfying (2.36). Let $[\hat{\boldsymbol{\beta}}]_{\nu_1}$ the ν_1-th element of the given sequence of estimators. Thus, $[\hat{\boldsymbol{\beta}}]_{\nu_1}$ is a (local) minimizer of $[M_{pen}(\cdot)]_{\nu_1}$. That is, there exists a neighborhood $\mathcal{N}_{\nu_1} \subseteq \mathbb{R}^{p_{\nu_1}+1}$ of $[\hat{\boldsymbol{\beta}}]_{\nu_1}$ with radius $\varepsilon_{\nu_1} > 0$ such that

$$[M_{pen}(\boldsymbol{\beta})]_{\nu_1} \geq [M_{pen}(\hat{\boldsymbol{\beta}})]_{\nu_1} \ \ \forall\, \boldsymbol{\beta} \text{ satisfying } ||\boldsymbol{\beta} - [\hat{\boldsymbol{\beta}}]_{\nu_1}||_2 < \epsilon_{\nu_1},$$

thus for all $\boldsymbol{\beta}$ in the neighborhood $\mathcal{N}_{\nu_1}$.

ii) Second, one defines the *fused* and the *not fused* quantities as in Theorem 2.3.17. In particular, one sets $[\tilde{\boldsymbol{\beta}}_{nf}]_{\nu_1} := [\hat{\boldsymbol{\beta}}]_{\nu_1}$ for the *not fused* version and $[\tilde{\boldsymbol{\beta}}_{f}]_{\nu_1}$ for the *fused* version (see Theorem 2.3.17). In other words, $[\tilde{\boldsymbol{\beta}}_{f}]_{\nu_1}$ completely coincides with $[\tilde{\boldsymbol{\beta}}_{nf}]_{\nu_1}$ except for the fact that for the *not fused* version $[\tilde{\boldsymbol{\beta}}_{nf}]_{\nu_1}$ it holds that

$$[\tilde{\beta}_{nf,\iota,r}]_{\nu_1} \neq [\tilde{\beta}_{nf,\iota,r-1}]_{\nu_1},$$

whereas for the *fused* version $[\tilde{\boldsymbol{\beta}}_{f}]_{\nu_1}$ it holds that

$$[\tilde{\beta}_{f,\iota,r}]_{\nu_1} = [\tilde{\beta}_{f,\iota,r-1}]_{\nu_1}.$$

By (2.36), one knows that $[\hat{\beta}_{\iota,r-1}]_{\nu_1} \neq [\hat{\beta}_{\iota,r}]_{\nu_1}$, thus the *not fused* version $[\tilde{\boldsymbol{\beta}}_{nf}]_{\nu_1}$ can be defined as $[\hat{\boldsymbol{\beta}}]_{\nu_1}$ since level r and $r-1$ of level ι are *not fused* for sample size ν_1.
Together with (i) this yields

$$\forall n_0 \in \mathbb{N}\ \exists\, \nu_1 \geq n_0,\, \mathcal{N}_{\nu_1} \subseteq \mathbb{R}^{p_{\nu_1}+1} \,:\, [M_{pen}(\boldsymbol{\beta})]_{\nu_1} \geq [M_{pen}(\tilde{\boldsymbol{\beta}}_{nf})]_{\nu_1}\ \forall \boldsymbol{\beta} \in \mathcal{N}_{\nu_1}. \tag{2.37}$$

By assumption and construction of the *not fused* version $\tilde{\boldsymbol{\beta}}_{nf}$, it holds

$$[\tilde{\beta}_{nf,\iota,r}]_n - [\tilde{\beta}_{nf,\iota,r-1}]_n = [\hat{\beta}_{\iota,r}]_n - [\hat{\beta}_{\iota,r-1}]_n = \Delta_n \to 0\,, \tag{2.38}$$

which further leads to

$$||[\tilde{\boldsymbol{\beta}}_{nf}]_n - [\tilde{\boldsymbol{\beta}}_{f}]_n||_2 \to 0.$$

Then, by Theorem 2.3.17

$$\exists\ \nu_2 \in \mathbb{N} \text{ such that } \forall\, n \geq \nu_2 \,:\, [M_{pen}(\tilde{\boldsymbol{\beta}}_{nf})]_n > [M_{pen}(\tilde{\boldsymbol{\beta}}_{f})]_n. \tag{2.39}$$

Now, n_0 from the beginning of the proof is set as $n_0 := \nu_2$. On the one hand, by (2.39) it holds that

$$\forall\, n \geq \nu_2 \,:\, [M_{pen}(\tilde{\boldsymbol{\beta}}_{nf})]_n > [M_{pen}(\tilde{\boldsymbol{\beta}}_f)]_n \qquad (*)$$

and, on the other hand, by (2.37)

$$\exists\, \nu_1 \geq \nu_2,\, \mathscr{N}_{\nu_1} \subseteq \mathbb{R}^{p_{\nu_1}+1} \,:\, [M_{pen}(\boldsymbol{\beta})]_{\nu_1} \geq [M_{pen}(\tilde{\boldsymbol{\beta}}_{nf})]_{\nu_1}\ \forall \boldsymbol{\beta} \in \mathscr{N}_{\nu_1}. \qquad (**)$$

Now one needs to show that $[\tilde{\boldsymbol{\beta}}_f]_{\nu_1} \in \mathscr{N}_{\nu_1}$ to get a contradiction. For this, the following cases may occur, where it is recalled that $||[\tilde{\boldsymbol{\beta}}_{nf}]_n - [\tilde{\boldsymbol{\beta}}_f]_n||_2 \leq \Delta_n$.

- *Case 1:* $\Delta_{\nu_1} < \epsilon_{\nu_1}$
  Then $[\tilde{\boldsymbol{\beta}}_f] \in \mathscr{N}_{\nu_1}$, hence $(**)$ yields

  $$[M_{pen}(\tilde{\boldsymbol{\beta}}_f)]_{\nu_1} \geq [M_{pen}(\tilde{\boldsymbol{\beta}}_{nf})]_{\nu_1},$$

  whereas by $(*)$ it holds

  $$[M_{pen}(\tilde{\boldsymbol{\beta}}_{nf})]_{\nu_1} > [M_{pen}(\tilde{\boldsymbol{\beta}}_f)]_{\nu_1},$$

  which is contradictory.

- *Case 2:* $\Delta_{\nu_1} \geq \epsilon_{\nu_1}$
  In this case, $\tilde{\boldsymbol{\beta}}_f$ and $\tilde{\boldsymbol{\beta}}_{nf}$ are yet not close enough. Thus, n_0 chosen above needs to be increased. However, even if $\Delta_n \to 0$, one needs to ensure that the radius of the neighborhood ϵ_n does *not* converge to zero. For this, the following cases may occur.

  – *Case 2.1:* $\min\limits_{n\in\mathbb{N}} \epsilon_n$ exists
    Then, let $\nu_{min,1}$ be this particular value for which $0 < \epsilon_{\nu_{min,1}} \leq \epsilon_n\ \forall n \in \mathbb{N}$. Further, since $\Delta_n \to 0$, there exists $\nu_{min,2}$ such that $\Delta_n < \epsilon_{\nu_{min,1}}\, \forall n \geq \nu_{min,2}$. Hence, for $\nu_{min} := \max\{\nu_{min,1}, \nu_{min,2}\}$, one knows that $\forall n \geq \nu_{min} \,:\, [\tilde{\boldsymbol{\beta}}_f]_n \in \mathscr{N}_n$, hence $[\tilde{\boldsymbol{\beta}}_f]_n$ is in the neighborhood of $[\tilde{\boldsymbol{\beta}}_{nf}]_n$ with radius ϵ_n (the neighborhood introduced above) for $n \geq \nu_{min}$. One sets $n_0 := \max\{\nu_2, \nu_{min}\}$, so ν_2 ensures that decrease of the objective function $(*)'$, whereas ν_{min} ensures that $\tilde{\boldsymbol{\beta}}_f$ is in the neighborhood of $\tilde{\boldsymbol{\beta}}_{nf}$ $(**)'$. Consequently, as above, by (2.39) it holds that

    $$\forall\, n \geq \max\{\nu_2, \nu_{min}\} \,:\, [M_{pen}(\tilde{\boldsymbol{\beta}}_{nf})]_n > [M_{pen}(\tilde{\boldsymbol{\beta}}_f)]_n \qquad (*)'$$

    and, on the other hand, by (2.37)

    $$\exists\, \nu_1 \geq \max\{\nu_2, \nu_{min}\},\, \mathscr{N}_{\nu_1} \subseteq \mathbb{R}^{p_{\nu_1}+1} \,:\, [M_{pen}(\boldsymbol{\beta})]_{\nu_1} \geq [M_{pen}(\tilde{\boldsymbol{\beta}}_{nf})]_{\nu_1}\ \forall \boldsymbol{\beta} \in \mathscr{N}_{\nu_1}. \quad (**)'$$

    By the elaboration above one infers $[\tilde{\boldsymbol{\beta}}_f]_n \in \mathscr{N}_n\ \forall n \geq \nu_{min}$, thus $[\tilde{\boldsymbol{\beta}}_f]_{\nu_1} \in \mathscr{N}_{\nu_1}$ since $\nu_1 \geq \max\{\nu_2, \nu_{min}\} \geq \nu_{min}$, so by $(**)'$ one can conclude

    $$[M_{pen}(\tilde{\boldsymbol{\beta}}_f)]_n > [M_{pen}(\tilde{\boldsymbol{\beta}}_{nf})]_n$$

    whereas by $(*)'$ one further knows

    $$[M_{pen}(\tilde{\boldsymbol{\beta}}_{nf})]_n > [M_{pen}(\tilde{\boldsymbol{\beta}}_f)]_n,$$

    which is contradictory.

  – *Case 2.2:* $\min\limits_{n\in\mathbb{N}} \epsilon_n$ does not exist. This case is excluded by the minimal neighborhood condition (Definition B.4.1, Appendix B.4).

Thus, the claim follows. □

Remark 2.3.27 (Theorem 2.3.26 for random variables)**.** Similar to the arguments provided in Remark 2.3.19, Theorem 2.3.26 similarly holds for sequences of random variables, satisfying the corresponding conditions a.s. That is, one assumes that $\mathbb{P}\left(\lim_{n\to\infty} \Delta_n = 0\right) = 1$, which is for short $\Delta_n \to_{a.s.} 0$, and further $\mathbb{P}\left(\inf_{n\in\mathbb{N}} [w_0^{(\iota,r)}]_n = c\right) = 1$. The conditions that do not depend on some random variable are the same, thus $\inf_{n\in\mathbb{N}} \lambda_0^n = c > 0$, $\sup_{n\in\mathbb{N}} \lambda_1^n = c < \infty$ and (Cont2) or (Cont3) are assumed hold, respectively. Further, it is supposed that the minimal neighborhood condition of Definition B.4.1 of Appendix B.4 holds (the one for random variables, respectively). Then, the statement of Theorem 2.3.26 holds a.s., that is

$$\mathbb{P}\left(\exists n_0 \in \mathbb{N} \,:\, \forall n \geq n_0 \,:\, [\hat{\beta}_{\iota,r}^{(L_0\text{-FGL})}]_n = [\hat{\beta}_{\iota,r-1}^{(L_0\text{-FGL})}]_n\right) = 1.$$

Now, as Theorem 2.3.17 was transferred to the case of subsequences in Corollary 2.3.23, the same is done with Theorem 2.3.26 in the following Corollary 2.3.28.

Corollary 2.3.28 (Fusion properties of estimates for subsequences, fixed p and diverging p_n)**.** Let c denote any constant and let $\iota \in \{1, \ldots, J_n\}$ be a factor (ordinal without loss of generality) and $r \in \{1, \ldots, p_\iota\}$. Let $\left([\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}]_n\right)_{n\in\mathbb{N}}$ be any sequence of (local) minimizers of $[M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})]_n$, respectively, based on $n \in \mathbb{N}$ samples, for which $([\hat{\boldsymbol{\beta}}]_n)_{n\in\mathbb{N}}$ is written for simplicity in this corollary. Let $(a_n)_{n\in\mathbb{R}} \subseteq \mathbb{N}$ be a strictly monotonic increasing sequence and, based on this sequence, a sub-sequence of $\hat{\boldsymbol{\beta}}$ is defined as follows

$$\left((\hat{\boldsymbol{\beta}}^{sub})_n\right)_{n\in\mathbb{N}} := \left((\hat{\boldsymbol{\beta}})_{a_n}\right)_{n\in\mathbb{N}}$$

analogously to Corollary 2.3.23, hence the n-th element of the sub-sequence $\hat{\boldsymbol{\beta}}^{sub}$ equals the a_n-th element of the L_0-FGL estimator sequence $\hat{\boldsymbol{\beta}}$, so $[\hat{\boldsymbol{\beta}}^{sub}]_{a_n} = (\hat{\boldsymbol{\beta}}^{sub})_n = (\hat{\boldsymbol{\beta}})_{a_n} = [\hat{\boldsymbol{\beta}}]_{a_n}$. It is assumed that $[\hat{\beta}_{\iota,r}^{sub}]_{a_n} - [\hat{\beta}_{\iota,r-1}^{sub}]_{a_n} =: \Delta_{a_n} \to 0\,(n \to \infty$, so $a_n \to \infty)$ and for the weight $[w_0^{(\iota,r)}]_n$ that $\inf_{n\in\mathbb{N}} [w_0^{(\iota,r)}]_n = c > 0$. Further, let $\inf_{n\in\mathbb{N}} \lambda_0^n = c > 0$ and $\sup_{n\in\mathbb{N}} \lambda_1^n = c < \infty$. Finally, let (Cont2) be satisfied for the case of fixed p and (Cont3) for the case of diverging p_n and let the minimal neighborhood condition be satisfied (Definition B.4.1, Appendix B.4). Then, for this given sub-sequence $\left([\hat{\boldsymbol{\beta}}^{sub}]_{a_n}\right)_{n\in\mathbb{N}}$ it holds that

$$\exists n_0 \in \mathbb{N} \,:\, \forall n \geq n_0 \,:\, [\hat{\beta}_{\iota,r}^{sub}]_{a_n} = [\hat{\beta}_{\iota,r-1}^{sub}]_{a_n}.$$

Proof. It is supposed that the statement is not true, that is

$$\forall n_0 \in \mathbb{N}\; \exists\, \nu_1 \geq n_0 \,:\, [\hat{\beta}_{\iota,r}^{sub}]_{a_{\nu_1}} \neq [\hat{\beta}_{\iota,s}^{sub}]_{a_{\nu_1}}. \qquad (*)$$

One defines the following quantities based on this sample size a_{ν_1}, similar to (i) and (ii) in the proof of Theorem 2.3.26. Thus $[\hat{\boldsymbol{\beta}}^{sub}]_{a_{\nu_1}}$ is a (local) minimizer of $[M_{pen}(\boldsymbol{\beta})]_{a_{\nu_1}}$ and one sets $[\tilde{\boldsymbol{\beta}}_{nf}]_{a_{\nu_1}} := [\hat{\boldsymbol{\beta}}^{sub}]_{a_{\nu_1}}$. Further $[\tilde{\boldsymbol{\beta}}_f]_{a_{\nu_1}}$ is defined as explained in Theorem 2.3.17. Hence, $[\hat{\boldsymbol{\beta}}^{sub}]_{a_{\nu_1}}$ (and $[\tilde{\boldsymbol{\beta}}_{nf}]_{a_{\nu_1}}$) is an L_0-FGL estimate (of dimension $p_{a_{\nu_1}} + 1$) based on a_{ν_1} samples. By assumption and construction, it holds that

$$[\tilde{\beta}_{nf,\iota,r}]_{a_n} - [\tilde{\beta}_{nf,\iota,r-1}]_{a_n} = [\hat{\beta}_{\iota,r}^{sub}]_{a_n} - [\hat{\beta}_{\iota,r-1}^{sub}]_{a_n} = \Delta_{a_n} \to 0\,(n \to \infty).$$

Then, by Corollary 2.3.23, it is known that

$$\exists\; \nu_2 \in \mathbb{N} \,:\, \forall\, a_n \geq a_{\nu_2} \,:\, [M_{pen}(\tilde{\boldsymbol{\beta}}_{nf})]_{a_n} > [M_{pen}(\tilde{\boldsymbol{\beta}}_f)]_{a_n}. \qquad (**)$$

Executing similar steps as in the proof of Theorem 2.3.26, the claim follows. □

Analogously to Remark 2.3.27, Corollary 2.3.28 similarly holds for random variables, as provided in the following remark.

Remark 2.3.29 (Corollary 2.3.28 for random variables)**.** Corollary 2.3.28 similarly holds for sequences of random variables, satisfying the corresponding conditions a.s. That is, one assumes that $\mathbb{P}\left(\lim_{n\to\infty} \Delta_{a_n} = 0\right) = 1$, which is for short $\Delta_{a_n} \to_{a.s} 0$, and further $\mathbb{P}\left(\inf_{n\in\mathbb{N}} [w_0^{(\iota,r)}]_n = c\right) = 1$. The other requirements are not changed hence assume $\inf_{n\in\mathbb{N}} \lambda_0^n = c > 0$, $\sup_{n\in\mathbb{N}} \lambda_1^n = c < \infty$ and let (Cont2) or (Cont3) hold, respectively. Further, one assumes that the minimal neighborhood condition of Definition B.4.1 (Appendix B.4) holds for random variables. Then, the statement of Corollary 2.3.28 holds a.s., that is

$$\mathbb{P}\left(\exists n_0 \in \mathbb{N} : \forall n \geq n_0 : [\hat{\beta}_{\iota,r}^{(L_0\text{-FGL})}]_{a_n} = [\hat{\beta}_{\iota,r-1}^{(L_0\text{-FGL})}]_{a_n}\right) = 1.$$

Screening Property Fusion

The goal of this subsection is to proof the screening property for fusion, in particular $\lim_{n\to\infty} \mathbb{P}(F_n = F^*) = 1$, which is provided in a theorem at the end of this section (Theorem 2.3.37). The Definition of F^* and F_n given in Definition 1.2.8 is recalled, where F_n is based on $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$, i.e.

$$\begin{aligned} F^* &= \left\{(j,r,s) \in \{1,\dots,J\} \times \{1,\dots,p_j\}^2 \,|\, \boldsymbol{\beta}_j^* \neq 0,\, \beta_{j,r}^* = \beta_{j,s}^*,\, r < s\right\}, \\ F_n &= \left\{(j,r,s) \in \{1,\dots,J\} \times \{1,\dots,p_j\}^2 \,|\, \boldsymbol{\beta}_j^* \neq 0,\, \hat{\beta}_{j,r}^{(L_0\text{-FGL})} = \hat{\beta}_{j,s}^{(L_0\text{-FGL})},\, r < s\right\}. \end{aligned}$$

For diverging p_n, $j \in \{1,\dots,J_n\}$ is written.

Remark 2.3.30 (Properties of Fusion sets F^*, F_n)**.** By construction and the sparsity assumption (see Definition 1.2.4), it is known that, even in the case of diverging p_n, the sets F^* and F_n are countable and finite $\forall n \in \mathbb{N}$. Further, there exists some

$$F_{max} := \{(j,r,s) \in \{1,\dots j_0\} \times \{1,\dots,p_j\}^2\}$$

with $|F_{max}| < \infty$, such that $F_n \subseteq F_{max}\, \forall n \in \mathbb{N}$ as well as $F^* \subseteq F_{max}$, where $j_0 \in \mathbb{N}$ is the number of truly influential factors. The sparsity assumption yields $j_0 < J$ (fixed case) and with $j_{0,n}$ being the number of truly influential factors with J_n candidate factors, it is known that $\exists n_0 \in \mathbb{N}$ such that $j_{0,n} = j_0\, \forall n \geq n_0$. Hence, it is concluded that in the fixed and in the diverging case, both sets F^*, F_n are countable and finite $\forall n \in \mathbb{N}$.

Since the assumptions ensuring the existence of some $\sqrt{n}$ consistent L_0-FGL estimator are needed more often, the following definition is introduced for brevity.

Definition 2.3.31. The assumptions for $\sqrt{n}$ consistency of the L_0-FGL estimator $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$ are said to be satisfied if

(i) Fixed p: (Reg1)-(Reg3) of Appendix B.1 hold and with a_n^0 and a_n^1 as defined in Theorem 2.3.2, it holds that $a_n^1/\sqrt{n} = o_p(1)$ and $a_n^0 = O_p(1)$.

(ii) Diverging p_n: (div.Reg1)-(div.Reg3) of Appendix B.2 hold and with $\alpha_n := \sqrt{\frac{p_n}{n}}$ and a_n^0, a_n^1 as defined in Theorem 2.3.2 it holds that $\alpha_n a_n^1 J_n = o_p(1)$, $a_n^0 p_n(p_n - 1) = o_p(1)$ and $p_n = o(n^{1/4})$.

If these assumptions are satisfied, there exists some sequence $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$ for which it holds for fixed p that $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2 = O_p\left(\frac{1}{\sqrt{n}}\right)$ and for diverging p_n that $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2 = O_p\left(\sqrt{\frac{p_n}{n}}\right)$, which was proven in Section 2.3.1.

Remark 2.3.32 (Consistency sub-sequences)**.** For short, $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})} = \hat{\boldsymbol{\beta}}$ is written below, where two important implications are presented for sub-sequences of $\hat{\boldsymbol{\beta}}$, assuming that $\hat{\boldsymbol{\beta}}$ is (i) $\sqrt{n}$ or (ii) $\sqrt{n/p_n}$ consistent, respectively.

(i) Fixed p: Assume that $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2 = O_p\left(\frac{1}{\sqrt{n}}\right)$, that is

$$\forall \epsilon > 0 \; \exists M, N > 0 \; : \; \mathbb{P}(\sqrt{n}||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2 > M) \leq \epsilon \; \forall \, n \geq N.$$

Then, for a strictly monotonic increasing sequence $(a_n)_{n\in\mathbb{N}} \subseteq \mathbb{N}$ it also holds that with $(\hat{\boldsymbol{\beta}})_{a_n}$ denoting the a_n-th element of the sequence, hence the estimate based on sample size a_n

$$\forall \epsilon > 0 \; \exists M, \, N > 0 \; : \; \mathbb{P}(\sqrt{a_n}||(\hat{\boldsymbol{\beta}})_{a_n} - \boldsymbol{\beta}^*||_2 > M) \leq \epsilon \; \forall \, a_n \geq N.$$

So for the sub-sequence $\left((\hat{\boldsymbol{\beta}}^{sub})_n\right)_{n\in\mathbb{N}} := \left((\hat{\boldsymbol{\beta}})_{a_n}\right)_{n\in\mathbb{N}}$, one can deduce that

$$\forall \epsilon > 0 \; \exists M, \, N > 0 \; : \; \mathbb{P}(\sqrt{a_n}||(\hat{\boldsymbol{\beta}}^{sub})_n - \boldsymbol{\beta}^*||_2 > M) \leq \epsilon \; \forall \, n \text{ with } a_n \geq N.$$

(ii) Diverging p_n: in the same way as for fixed p one can infer that

$$\forall \epsilon > 0 \; \exists M, \, N > 0 \; : \; \mathbb{P}\left(\sqrt{\frac{a_n}{p_{a_n}}}||(\hat{\boldsymbol{\beta}})_{a_n} - \boldsymbol{\beta}^*||_2 > M\right) \leq \epsilon \; \forall \, a_n \geq N.$$

The next two lemmata (Lemma 2.3.33 and Lemma 2.3.35) are needed as preparations for the upcoming theorems in this subsection.

Lemma 2.3.33. One assumes that for a real-valued sequence $(r_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}$ it holds that $\lim_{n\to\infty} r_n \neq r$ for some $r \in \mathbb{R}$. Then, there exists a sub-sequence $(\tilde{r}_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}$ of $(r_n)_{n\in\mathbb{N}}$ such that

$$\exists \, c > 0, n_0 \in \mathbb{N} \; : \; |\tilde{r}_n - r| \geq c \, \forall n \geq n_0.$$

Proof. Since $\lim_{n\to\infty} r_n \neq r$, it is known that

$$\exists \epsilon > 0 \; : \; \forall N \in \mathbb{N} \, \exists \, n \geq N \; : \; |r_n - r| \geq \epsilon.$$

This means that, for this ϵ, there exists for $N = 1$ some $n_1 \in \mathbb{N}$, $n_1 \geq N = 1$ such that $|r_{n_1} - r| \geq \epsilon$ and in the same way

$$\begin{aligned}
&|r_{n_1} - r| \geq \epsilon \;\; (n_1 \in \mathbb{N}, n_1 \geq 1), \\
&|r_{n_2} - r| \geq \epsilon \;\; (n_2 \in \mathbb{N}, n_2 \geq 2), \\
&|r_{n_3} - r| \geq \epsilon \;\; (n_3 \in \mathbb{N}, n_3 \geq 3), \\
&\ldots.
\end{aligned}$$

Now, from the sequence $(n_i)_{i\in\mathbb{N}}$ one needs to construct a strictly monotonic increasing sequence to be able to construct a sub-sequence $(\tilde{r}_n)_{n\in\mathbb{N}}$ of $(r_n)_{n\in\mathbb{R}}$. The set $\mathfrak{N} := \{n_1, n_2, n_3, n_4, \ldots\} \subseteq \mathbb{N}$ containing all elements of the sequence $(n_i)_{i\in\mathbb{N}}$ has a smallest element, since it is a subset of $\mathbb{N}$. Now, it may happen that for some $k, j \in \mathbb{N}$ it holds that $n_k = n_{k+j}$, so duplicates may appear in $\mathfrak{N}$. But, for this duplicate n_k, by construction it holds that $\exists \, i \in \mathbb{N} \; : \; n_k < n_i$ since $n_i \geq i$ so at least this holds for $i > n_k = n_{k+j}$. Hence, the number of duplicates and the successive elements in that are not strictly monotonic increasing in $\mathfrak{N}$ are countable and finite. Leaving out the duplicates one gets $\mathfrak{M} := \{n_i \,|\, i \in \mathbb{N}, n_i \neq n_j, \, \forall j \in \mathbb{N}\}$, which is a countably infinite set.

Further, with the same argument as above $\mathfrak{M}$ has a smallest element (note that $\mathfrak{M} \subseteq \mathfrak{N} \subseteq \mathbb{N}$). Hence, a sub-sequence $(\tilde{n}_i)_{i\in\mathbb{N}}$ can be constructed as follows

$$\begin{aligned}
&\tilde{n}_1 := \min\{\mathfrak{M}\},\\
&\tilde{n}_2 := \min\{\mathfrak{M}\backslash\{\tilde{n}_1\}\},\\
&\tilde{n}_3 := \min\{\mathfrak{M}\backslash\{\tilde{n}_1, \tilde{n}_2\}\},\\
&\dots\\
&\tilde{n}_i := \min\{\mathfrak{M}\backslash\{\tilde{n}_1, \dots, \tilde{n}_{i-1}\}\},\\
&\dots.
\end{aligned}$$

By definition, the sequence $(\tilde{n}_i)_{i\in\mathbb{N}}$ is strictly monotonic increasing. The sub-sequence of $(r_{\tilde{n}_i})_{i\in\mathbb{N}}$ is constructed as follows: $(\tilde{r}_i)_{i\in\mathbb{N}} := (r_{\tilde{n}_i})_{i\in\mathbb{N}}$ such that for $c := \epsilon$ and every $i \in \mathbb{N}$ it holds

$$|\tilde{r}_i - r| = |r_{\tilde{n}_i} - r| \geq \epsilon = c \;\forall i \in \mathbb{N} \text{ since } \tilde{n}_i \in \mathfrak{M} \subseteq \mathfrak{N}.$$

To sum up, the subsequence $(\tilde{r}_i)_{i\in\mathbb{N}}$ satisfies $|\tilde{r}_i - r| \geq c \;\forall i \in \mathbb{N}$ so the claim follows with $n_0 \geq 1$. □

Remark 2.3.34 (Lemma 2.3.33 for random variables)**.** Transferring the statement of Lemma 2.3.33 to the setting of sequences of random variables, let $\left((\hat{\boldsymbol{\zeta}})_n\right)_{n\in\mathbb{N}}$ be a sequence of random variables, and further $\boldsymbol{\zeta}$ some random variable. It is assumed that $\mathbb{P}\left(\lim_{n\to\infty} \hat{\boldsymbol{\zeta}} \neq \boldsymbol{\zeta}\right)$, thus $\hat{\boldsymbol{\zeta}} \not\to_{a.s.} \boldsymbol{\zeta}$, then there exists a sub-sequence $\left((\hat{\boldsymbol{\zeta}}^{sub})_n\right)_{n\in\mathbb{N}}$ of $\left((\hat{\boldsymbol{\zeta}})_n\right)_{n\in\mathbb{N}}$ such that

$$\mathbb{P}\left(\exists\, c > 0\,, n_0 \in \mathbb{N} : ||\hat{\boldsymbol{\zeta}}^{sub} - \boldsymbol{\zeta}|| \geq c \;\forall n \geq n_0\right) = 1.$$

In particular, one can further adjust the sub-sequence $\left((\hat{\boldsymbol{\zeta}}^{sub})_n\right)_{n\in\mathbb{N}}$ such that

$$\mathbb{P}\left(\exists\, c > 0 : ||\hat{\boldsymbol{\zeta}}^{sub} - \boldsymbol{\zeta}|| \geq c \;\forall n \in \mathbb{N}\right) = 1$$

holds.

Lemma 2.3.35 (Componentwise limit, fixed p and diverging p_n)**.** Let the conditions of Definition 2.3.31 be satisfied such that there exists some L_0-FGL estimator $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$, which is denoted for short by $\hat{\boldsymbol{\beta}}$ in this lemma, such that $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$ consistent (fixed p) or $\sqrt{n/p_n}$ consistent (diverging p_n), respectively. Then, for every $j \in \{1, \dots, J\}$ ($j \in \{1, \dots, J_n\}$, respectively) and every $r \in \{1, \dots, p_j\}$ it holds that

$$\operatorname*{p\,lim}_{n\to\infty} \hat{\beta}_{j,r} = \beta^*_{j,r},$$

hence the limit in probability exists and is equal to the true value.

Proof. The claim directly follows from $\sqrt{n}$ consistency, since using $\sqrt{n}||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|| = O_p(1)$ one infers $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|| = o_p(1)$ and $|\hat{\beta}_{j,r} - \beta^*_{j,r}| \leq ||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|| = o_p(1)$ so $|\hat{\beta}_{j,r} - \beta^*_{j,r}| = o_p(1)$. For $\sqrt{n/p_n}$ consistency the steps can be similarly executed using $\sqrt{n/p_n} \to \infty$. □

Corollary 2.3.36 (Componentwise limit for subsequences, fixed p and diverging p_n)**.** Lemma 2.3.35 holds in the same way for any sub-sequence $\hat{\boldsymbol{\beta}}^{sub}$ of $\hat{\boldsymbol{\beta}}$. In particular, let the requirements of Definition 2.3.31 hold and let $(a_n)_{n\in\mathbb{N}} \subseteq \mathbb{N}$ be an arbitrary strictly monotonic increasing sequence. Defining $\hat{\boldsymbol{\beta}}^{sub}$ as explained in Notation 2.3.21, that is $\left((\hat{\boldsymbol{\beta}}^{sub})_n\right)_{n\in\mathbb{N}} := \left((\hat{\boldsymbol{\beta}})_{a_n}\right)_{n\in\mathbb{N}}$

such that $(\hat{\boldsymbol{\beta}}^{sub})_n = [\hat{\boldsymbol{\beta}}^{sub}]_{a_n} = [\hat{\boldsymbol{\beta}}]_{a_n} = (\hat{\boldsymbol{\beta}})_{a_n}$, it holds for all $j \in \{1, \dots, J\}$ ($j \in \{1, \dots, J_n\}$, respectively) and $r \in \{1, \dots, p_j\}$ that

$$\operatorname*{p\,lim}_{n\to\infty} [\hat{\beta}_{j,r}^{sub}]_{a_n} = \beta_{j,r}^*.$$

Proof. Completely analogous to Lemma 2.3.35 using that any sub-sequence of a consistent estimator is itself consistent. □

In the next Theorem 2.3.37, it is proven that for a $\sqrt{n}$ consistent ($\sqrt{n/p_n}$, respectively) sequence of estimators, the probability that the corresponding fusion set F_n equals the true fusion set F^* tends to one for $n \to \infty$, which is called (asymptotic) screening property for fusion. It is noted that the fusion sets are those corresponding to the sequence of $\sqrt{n}$ consistent ($\sqrt{n/p_n}$, respectively) estimators. Since it is crucial which element of the sequence $\left((\hat{\boldsymbol{\beta}})_n\right) = \left([\hat{\boldsymbol{\beta}}]_n\right)_{n\in\mathbb{N}}$ is considered, the notation with round and square brackets is employed.

Theorem 2.3.37 ((Asymptotic) Screening Property Fusion, fixed p and diverging p_n)**.** Let the conditions of Theorem 2.3.26 hold. Further, the assumptions for $\sqrt{n}$ consistency are required to hold (Definition 2.3.31 (i) for the case of fixed p and (ii) for the case of diverging p_n), such that there exists some sequence of L_0-FGL estimators $\left((\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})})_n\right)_{n\in\mathbb{N}}$, denoted by $\left((\hat{\boldsymbol{\beta}})_n\right)_{n\in\mathbb{N}}$ for short in this theorem, such that $\left((\hat{\boldsymbol{\beta}})_n\right)_{n\in\mathbb{N}}$ is $\sqrt{n}$ consistent (fixed p) or $\sqrt{n/p_n}$ consistent (diverging p_n), respectively. Then, for both cases, for F_n being the fusion set corresponding to this particular $[\hat{\boldsymbol{\beta}}]_n$ it holds that

$$\lim_{n\to\infty} \mathbb{P}(F^* = F_n) = 1. \tag{2.40}$$

Proof. To show the claim, two subset relations are shown, first $\lim_{n\to\infty} \mathbb{P}(F_n \subseteq F^*) = 1$ and second $\lim_{n\to\infty} \mathbb{P}(F^* \subseteq F_n) = 1$.

(⊆) It is shown that $\lim_{n\to\infty} \mathbb{P}(F_n \not\subseteq F^*) = 0$. In particular, it is shown that $\forall \epsilon > 0$ there exists some $\tilde{N}_1 \in \mathbb{N}$ such that $\mathbb{P}(F_n \not\subseteq F^*) < \epsilon \ \forall n \geq \tilde{N}_1$. Starting with $\mathbb{P}(F_n \not\subseteq F^*)$, one deduces for some constant $c > 0$

$$\begin{aligned}
\mathbb{P}(F_n \not\subseteq F^*) &\leq \mathbb{P}\left((j,r,s) \notin F^* \text{ for some } (j,r,s) \in F_n\right) \\
&\leq \mathbb{P}\left(\beta_{j,r}^* \neq \beta_{j,s}^* \text{ and } [\hat{\beta}_{j,r}]_n = [\hat{\beta}_{j,s}]_n\right) \\
&\leq \mathbb{P}\left(|\beta_{j,r}^* - \beta_{j,s}^*| > c \text{ and either } |[\hat{\beta}_{j,r}]_n - \beta_{j,r}^*| \geq \frac{c}{2} \text{ or } |[\hat{\beta}_{j,s}]_n - \beta_{j,s}^*| \geq \frac{c}{2}\right) \\
&\leq \mathbb{P}\left(|[\hat{\beta}_{j,r}]_n - \beta_{j,r}^*| \geq \frac{c}{2} \text{ or } |[\hat{\beta}_{j,s}]_n - \beta_{j,s}^*| \geq \frac{c}{2}\right) \\
&\leq \mathbb{P}\left(||[\hat{\boldsymbol{\beta}}]_n - \boldsymbol{\beta}^*|| \geq \frac{c}{2}\right),
\end{aligned} \tag{2.41}$$

where in the last inequality it is used that if $|[\hat{\beta}_{j,r}]_n - \beta_{j,r}^*| \geq \frac{c}{2}$ or $|[\hat{\beta}_{j,s}]_n - \beta_{j,s}^*| \geq \frac{c}{2}$ it follows that $||[\hat{\boldsymbol{\beta}}]_n - \boldsymbol{\beta}^*|| \geq \frac{c}{2}$ since

$$\frac{c}{2} \leq \sqrt{|[\hat{\beta}_{j,r}]_n - \beta_{j,r}^*|^2 + |[\hat{\beta}_{j,s}]_n - \beta_{j,s}^*|^2} \leq ||[\hat{\boldsymbol{\beta}}]_n - \boldsymbol{\beta}^*||.$$

Now, one needs to show that there exists some $\tilde{N}_1 \in \mathbb{N}$ such that

$$\mathbb{P}\left(||[\hat{\boldsymbol{\beta}}]_n - \boldsymbol{\beta}^*|| \geq \frac{c}{2}\right) < \epsilon \ \forall n > \tilde{N}_1, \tag{2.42}$$

see (2.41). For this, the $\sqrt{n}$ consistency ($\sqrt{n/p_n}$, respectively) is used, similar to the proofs of Theorem 2.3.8 and 2.3.9. That is, since $\frac{c}{2}$ is some constant, it is known that n can be increased in a way such that $\frac{M}{\sqrt{n}} < \frac{c}{2}$ ($M\sqrt{\frac{p_n}{n}} < \frac{c}{2}$, respectively), for which it is refered to the previously referenced proofs for more details. To sum up, for both the case of p being fixed and for p_n diverging with n, one can find some $\tilde{N}_1 \in \mathbb{N}$ (large enough), such that (2.42) holds.

$\supseteq$ Now, it is shown that $\lim_{n\to\infty} \mathbb{P}(F^* \not\subseteq F_n) = 0$. In particular, it is shown that $\forall \epsilon > 0$ there exists some $\tilde{N}_2 \in \mathbb{N}$ such that $\mathbb{P}(F^* \not\subseteq F_n) < \epsilon \ \forall n \geq \tilde{N}_2$. Starting with $\mathbb{P}(F^* \not\subseteq F_n)$, one deduces

$$\begin{aligned}
\mathbb{P}(F^* \not\subseteq F_n) &\leq \mathbb{P}\left((j,r,s) \notin F_n \text{ for some } (j,r,s) \in F^*\right) \\
&\leq \mathbb{P}\left(\beta^*_{j,r} = \beta^*_{j,s},\ [\hat{\beta}_{j,r}]_n \neq [\hat{\beta}_{j,s}]_n\right) \\
&= \mathbb{P}\left(\beta^*_{j,r} = \beta^*_{j,s},\ |[\hat{\beta}_{j,r}]_n - [\hat{\beta}_{j,s}]_n| = c_n > 0\right) \\
&\leq \mathbb{P}\left(\beta^*_{j,r} = \beta^*_{j,s},\ |[\hat{\beta}_{j,r}]_n - [\hat{\beta}_{j,s}]_n| = c_n > 0,\ c_n \to_{a.s} 0\right) \qquad (2.43) \\
&\quad + \mathbb{P}\left(\beta^*_{j,r} = \beta^*_{j,s},\ |[\hat{\beta}_{j,r}]_n - [\hat{\beta}_{j,s}]_n| = c_n > 0,\ c_n \not\to_{a.s} 0\right). \qquad (2.44)
\end{aligned}$$

Observing the first summand (2.43), the fact that $c_n \to_{a.s.} 0$ yields, by using Theorem 2.3.26 and Remark 2.3.27, that $\mathbb{P}(\exists n_0 \ : \ \forall n \geq n_0 \ : \ [\hat{\beta}_{j,r}]_n = [\hat{\beta}_{j,s}]_n) = 1$, hence (2.43) equals

$$\begin{aligned}
&\mathbb{P}\left(\beta^*_{j,r} = \beta^*_{j,s},\ |[\hat{\beta}_{j,r}]_n - [\hat{\beta}_{j,s}]_n| = c_n > 0,\ c_n \to_{a.s} 0,\ \exists n_0 \ : \ \forall n \geq n_0 \ : \ [\hat{\beta}_{j,r}]_n = [\hat{\beta}_{j,s}]_n\right) \\
&\overset{(*)}{=} \mathbb{P}(\emptyset) = 0,
\end{aligned}$$

where in $(*)$ it was used that $[\hat{\beta}_{j,r}]_n = [\hat{\beta}_{j,s}]_n \, \forall n \geq n_0$ and $|[\hat{\beta}_{j,r}]_n - [\hat{\beta}_{j,s}]_n| = c_n > 0$ are contradictory. Thus, it remains to calculate (2.44), for which Remark 2.3.34 is used. In particular, this remark yields that a sub-sequence $\hat{\boldsymbol{\beta}}^{sub}$ of $\hat{\boldsymbol{\beta}}$ can be found such that for some constant $c > 0$ it holds $\mathbb{P}\left(|[\hat{\beta}^{sub}_{j,r}]_n - [\hat{\beta}^{sub}_{j,s}]_n| \geq c \ \forall n \in \mathbb{N}\right) = 1$. Consequently, for (2.44) it holds

$$\begin{aligned}
(2.44) &\leq \mathbb{P}\left(\beta^*_{j,r} - \beta^*_{j,s} = 0 \text{ and } \exists \text{ some sub-sequence } \hat{\boldsymbol{\beta}} \text{ for which } |[\hat{\beta}^{sub}_{j,r}]_n - [\hat{\beta}^{sub}_{j,s}]_n| \geq c > 0\right) \\
&\leq \mathbb{P}(||(\hat{\boldsymbol{\beta}}^{sub})_n - \boldsymbol{\beta}^*|| \geq c).
\end{aligned}$$

Now, for the same reason elaborated in the first step $\subseteq$, using that any sub-sequence is also consistent (see Remark 2.3.32), one can conclude that there exists some $\tilde{N}_2 \in \mathbb{N}$ such that $\forall n > \tilde{N}_2$ the following inequality holds

$$\mathbb{P}(||(\hat{\boldsymbol{\beta}}^{sub})_n - \boldsymbol{\beta}^*|| \geq c) < \epsilon.$$

To sum up, both relations are shown and the claim follows. □

Remark 2.3.38 (On the assumptions)**.** Since several requirements were imposed in Theorem 2.3.37, and similarly in other theorems of this section, it is demonstrated below that these assumptions do not exclude one another, in contrast, it may be even helpful to impose them together.

- Fixed p: the requirements $\inf_{n\in\mathbb{N}} \lambda^n_0 = c > 0$ and $\sup_{n\in\mathbb{N}} \lambda^n_1 = c < \infty$ are not contradictory to $a^1_n/\sqrt{n} = o_p(1)$ and $a^0_n = O_p(1)$, which are mutually imposed in Theorem 2.3.37. First,

one knows for every $\varepsilon > 0$

$$\left|\frac{a_n^1}{\sqrt{n}}\right| \leq \left|\frac{c \cdot \max\{w_1^{(j)} \,|\, j \in \{1,\dots,J\}\}}{\sqrt{n}}\right|$$
$$\Rightarrow \lim_{n\to\infty} \mathbb{P}\left(\left|\frac{a_n^1}{\sqrt{n}}\right| \geq \varepsilon\right) \leq \lim_{n\to\infty} \mathbb{P}\left(\left|\frac{c \cdot \max\{w_1^{(j)} \,|\, j \in \{1,\dots,J\}\}}{\sqrt{n}}\right| \geq \varepsilon\right). \quad (2.45)$$

Hence, $a_n^1/\sqrt{n} = o_p(1)$ would be e.g. satisfied if the weights $w_1^{(j)}$ are bounded a.s., or if $\max\{w_1^{(j)} \,|\, j \in \{1,\dots,J\}\} = o_p(\sqrt{n})$, thus bounding the supremum of λ_1^n over $n \in \mathbb{N}$ by some constant is helpful to control a_n^1. Second, the requirement $a_n^0 = O_p(1)$ means that a_n^0, which is defined as the maximum over $j \in \{1,\dots,J\}$ and $r, s \in \{1,\dots,p_j\}$ of the product $\lambda_0^n w_0^{(j,rs)}$, is stochastically bounded. Hence, bounding the tuning parameter λ_0^n from below is clearly not contradictory to this requirement. To sum up, one can see that bounding the tuning parameters for fusion and selection is not conflicting with $a_n^1/\sqrt{n} = o_p(1)$ and $a_n^0 = O_p(1)$, instead it can be even helpful, such that these conditions are reasonable to be imposed mutually.

- Diverging p_n: first, in the same way as for fixed p case one can write

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(|\alpha_n a_n^1 J_n| \geq \varepsilon) &= \lim_{n\to\infty} \mathbb{P}\left(\left|\sqrt{\frac{p_n}{n}} \max\{\lambda_1^n w_1^{(j)} \,|\, j \in \{j,\dots,J_n\}\} J_n\right| \geq \varepsilon\right) \\ &\leq \lim_{n\to\infty} \mathbb{P}\left(\left|c\sqrt{\frac{p_n}{n}} \max\{w_1^{(j)} \,|\, j \in \{j,\dots,J_n\}\} J_n\right| \geq \varepsilon\right), \quad (2.46)\end{aligned}$$

so to ensure that the first limit on the left hand side is zero, such that $\alpha_n a_n^1 J_n = o_p(1)$, it is sufficient to ensure that the last limit on the right hand side (2.46) is zero. As one can see, (2.46) takes advantage of the fact that the supremum of λ_1^n exists over $n \in \mathbb{N}$, such that this requirement is helpful to ensure $\alpha_n a_n^1 J_n = o_p(1)$. Second, coming to $a_n^0 p_n(p_n - 1) = o_p(1)$, this requirement is similarly not affected by $\inf_{n\in\mathbb{N}} \lambda_0^n > c > 0$ as in the fixed case. Third, assuming that the ratio of p_n and n can be controlled asymptotically, that is $p_n = o(n^{1/4})$, is obviously not affected by requirements on the tuning parameters. Summing up, as for fixed p, the assumptions made in Theorem 2.3.26 and the ones needed for the consistency Theorems in Kaufmann and Kateri (2024), which are mutually imposed in Theorem 2.3.37, do not exclude one another and are reasonable to be assumed mutually.

Chapter 3

Computational Methods and Simulation Studies for L_0-FGL

Having introduced and investigated the purpose of L_0-FGL in Chapter 2, altogether with solid theoretical properties, this chapter focuses on the application of L_0-FGL in practice. That is, computational methods are specified and resulting coefficient paths are exhibited. Moreover, the performance of L_0-FGL is examined in simulation studies.

To be more precise, in order to determine estimates $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$, optimization algorithms to minimize the objective function $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ are needed. Therefore, Section 3.1 presents the application of two different algorithms to L_0-FGL, specifically the PIRLS algorithm (Section 3.1.1) and a block coordinate descent (BCD) algorithm (Section 3.1.2). These types of algorithms are introduced in their general forms in Appendix A, however, their application to the objective function of L_0-FGL is specified here. Additionally, presenting the coefficient paths for both algorithms, the characteristics of L_0-FGL especially in comparison to CAS-L_0 and group lasso are demonstrated. Finally, Section 3.2 presents results from simulation studies that were conducted, showing that L_0-FGL fulfills the purpose for which it is introduced. To ensure a comparable scale, the goodness of fit measures, as well as the designs for the simulations are the same as presented in Section 1.7. An earlier version of the coefficient paths, the application of PIRLS and BCD for L_0-FGL, and a small selection of the conducted studies (i.e. parts of design B8.2 and highdim) can be similarly found in Kaufmann and Kateri (2024) published in *Electronic Journal of Statistics*.

3.1 Computational Methods

Analyzing the objective function $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ including an L_0 norm, the function is neither continuous (in $\mathbb{R}^{p+1}$), nor convex or concave. However, as discussed in Appendix A.1, especially in Example A.1.3, it is convenient to approximate the L_0 norm, which is applied for L_0-FGL in both algorithms presented below. For the group lasso part of $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$, the PIRLS algorithm applies a quadratic approximation, while in the BCD approach this approximation is neglected, which is one of the main differences of these two approaches. Clearly, another difference is that the BCD approach proceeds (block) coordinate-wise, thus the objective function is minimized with respect to each factor separately, keeping the others fixed. The subsequent sections provide details on these aspects, specifying the algorithms.

3.1.1 PIRLS for L_0-FGL

It is recalled from Appendix A.1, that the PIRLS algorithm can be applied to penalty functions of the following structure

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

which can be further written as in (A.3). Moreover, it is recalled that Oelker and Tutz (2013) extended PIRLS to penalties with vector-valued arguments, writing $\boldsymbol{R}_l\boldsymbol{\beta}$ for some matrix $\boldsymbol{R}_l$ instead of $\boldsymbol{a}_l^T\boldsymbol{\beta}$ for some vector $\boldsymbol{a}_l$, which is needed for the group lasso part in $P_\lambda^{(L_0\text{-FGL})}(\boldsymbol{\beta})$. Because of the fact that L_0-FGL has two tuning parameters λ_1 and λ_0, as well as two different penalty functions, one ends up with two penalty terms in the PIRLS algorithm. To be more precise, the following choices are made

$$\text{group lasso part:} \qquad p_l(\zeta) := \sqrt{p_l} \cdot \zeta, \ \boldsymbol{R}_l\boldsymbol{\beta} = \boldsymbol{\beta}_l, \tag{3.1}$$

$$L_0 \text{ part:} \qquad p_l(\zeta) := w_0^{(j,km)}\zeta, \ \boldsymbol{a}_l^T\boldsymbol{\beta} = \beta_{j,k} - \beta_{j,m} \text{ for } 0 \leq k < m \leq p_j. \tag{3.2}$$

The vectors $\boldsymbol{a}_l^T$ are responsible to pick the possible differences corresponding to the factors $j \in \{1, \ldots, J\}$, thus the entries of these vectors $\boldsymbol{a}_l$ are contained in the set $\{-1, 0, 1\}$ as justified in an example below (Example 3.1.1). For the group lasso part, the matrix $\boldsymbol{R}_l$ extracts the sub-vector $\boldsymbol{\beta}_l$ from the vector $\boldsymbol{\beta}$. Note that, for group lasso, the indices $l \in \{1, \ldots, L\}$ coincide with $j \in \{1, \ldots, J\}$, whereas for the L_0 part, $l \in \{1, \ldots, L\}$ combines both the number of factors, as well as the number of differences that need to be considered, where the latter depends on whether a factor is nominal or ordinal. This is the reason why in (3.1) l is written as index on the left and right hand side of $\boldsymbol{R}_l\boldsymbol{\beta} = \boldsymbol{\beta}_l$, while in (3.2) one needs to write $\boldsymbol{a}_l^T\boldsymbol{\beta} = \beta_{j,k} - \beta_{j,m}$.

Example 3.1.1. An example is exhibited to clarify the structure of the vectors $\boldsymbol{a}_l$ needed to produce all (pairwise/adjacent) differences of the coefficients of a factor. For this, $J = 2$ nominal factors with $p_1 = 1$, $p_2 = 3$ are considered. One defines the matrix $\mathfrak{A} \in \{-1, 0, 1\}^{\sum_{j=1}^{J} d_j \times (p+1)}$, where d_j is the number of differences corresponding to factor j. Hence one deduces

$$d_j = \begin{cases} p_j & \text{, factor } j \text{ ordinal,} \\ p_j + \frac{p_j(p_j-1)}{2} & \text{, factor } j \text{ nominal.} \end{cases}$$

The matrix $\mathfrak{A}$ is constructed such that the product $\mathfrak{A}\boldsymbol{\beta}$ gives a vector with entries being the corresponding differences. In the given example this matrix is

$$\mathfrak{A} = \begin{pmatrix} 0 & \mathbf{1} & 0 & 0 & 0 \\ 0 & 0 & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ 0 & 0 & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ 0 & 0 & \mathbf{0} & \mathbf{0} & \mathbf{1} \\ 0 & 0 & \mathbf{1} & \mathbf{-1} & \mathbf{0} \\ 0 & 0 & \mathbf{1} & \mathbf{0} & \mathbf{-1} \\ 0 & 0 & \mathbf{0} & \mathbf{1} & \mathbf{-1} \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \\ a_7 \end{pmatrix},$$

where the orange rows correspond to differences from the first factor and the blue rows to those from the second factor. Consequently, with $\boldsymbol{\beta} = (\beta_{int}, \beta_{1,1}, \beta_{2,1}, \beta_{2,2}, \beta_{2,3})$ one infers

$$\mathfrak{A}\boldsymbol{\beta} = (\beta_{1,1}, \ \beta_{2,1}, \ \beta_{2,2}, \ \beta_{2,3}, \ \beta_{2,1} - \beta_{2,2}, \ \beta_{2,1} - \beta_{2,3}, \ \beta_{2,2} - \beta_{2,3})^T.$$

Further it holds that $L = \sum_{j=1}^{J} d_j$ so in this example $L = 7$. One defines $\mathfrak{L}_j$ as the set of the rows in $\mathfrak{A}$ that build differences for factor j, thus $|\mathfrak{L}_j| = d_j$. In this example $\mathfrak{L}_1 = \{1\}$ (orange row) and $\mathfrak{L}_j = \{2, 3, 4, 5, 6, 7\}$ (blue rows) and it holds that $|\mathfrak{L}_j| = d_j$. Sub-matrices $\mathfrak{A}_j \in \{-1, 0, 1\}^{d_j \times p_j}$ are extracted from the matrix $\mathfrak{A}$ for each factor, such that $\mathfrak{A}_j \boldsymbol{\beta}_j$ gives the differences of factor j. Above, $\mathfrak{A}_1$ is given by the entries being orange *and* bold, while $\mathfrak{A}_2$ is given by the entries being blue *and* bold. Additionally, one sets $a_{l,j} \in \{-1, 0, 1\}^{p_j}$ for $l \in \{1, \dots, d_j\}$, $j \in \{1, \dots, J\}$ as the rows of $\mathfrak{A}_j$.

It remains to provide the choices of the approximations of the norms appearing in L_0-FGL. For this, one can follow the choice of Oelker and Tutz (2013), presenting an approximation for the L_0 norm, as well as the euclidean norm $||\cdot||_2$, that is,

$$\text{group lasso } ||\boldsymbol{\xi}||_2: \qquad N_l(\boldsymbol{\xi}) = (\boldsymbol{\xi}^T \boldsymbol{\xi} + c)^{1/2}, \tag{3.3}$$

$$L_0 \text{ norm:} \qquad N_l(\xi) = \frac{2}{1 + \exp(-\gamma|\xi|)} - 1\,, \tag{3.4}$$

$$D_l(\xi) = \frac{2\gamma}{1 + \exp(-\gamma|\xi|)} \left(1 - \frac{1}{1 + \exp(-\gamma|\xi|)}\right) \frac{\xi}{\sqrt{\xi^2 + c}}. \tag{3.5}$$

To sum up, all the quantities needed to apply the PIRLS algorithm to the objective function $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ are discussed, such that it can be proceeded to the resulting coefficient paths.

Coefficient Paths (PIRLS)

Coefficient paths for L_0-FGL are provided, together with paths for CAS-L_0 and group lasso, enabling a comparison with the approaches L_0-FGL is build from. All three approaches are computed with the PIRLS algorithm, using the approximations given in the paragraph above.

$J = 2$ categorical covariates are considered where $\mathscr{X}_1$ has three and $\mathscr{X}_2$ has two categories, drawn from a multinomial distribution with equal probabilities. The true coefficient vector is chosen to be

$$\boldsymbol{\beta}^* = (2, \;\; 1.2, 1, 0.5, \;\; -0.8, -0.5).$$

To simulate the dataset, the `simulation` function from the package `gvcm.cat` is used. For the fit of CAS-L_0, L_0-FGL and group lasso with PIRLS, dedicated `R` code was developed and used for this purpose. Note that the `R` package `grpreg`, which was chosen in Chapter 1 to obtain the group lasso estimates, does not employ the PIRLS algorithm.

The next step is to choose minimum and maximum values for the tuning parameters. As minimum values for the tuning parameters, zero was chosen for all methods which leads to the unpenalized MLE. The maximum values were chosen in a way such that all factors are excluded from the model, or fused with the reference category, respectively. The points in the paths where the fusion/selection occurs are not comparable among the methods, which is caused by the fact that L_0-FGL requires two tuning parameters and different maximum values were chosen here. However, the coefficient paths are only for visualization purposes, underlining the characteristics of the corresponding penalty, where their performance is evaluated later in this chapter.

Figure 3.1 displays the resulting coefficient paths. On the left hand side of each path plot, the unpenalized MLE can be seen, while proceeding to the right on the horizontal axis, the tuning parameter is increased until the maximum value is reached. For L_0-FGL, both tuning parameters are increased simultaneously.
On the left hand side (first row), the paths for group lasso are shown, on the right hand side

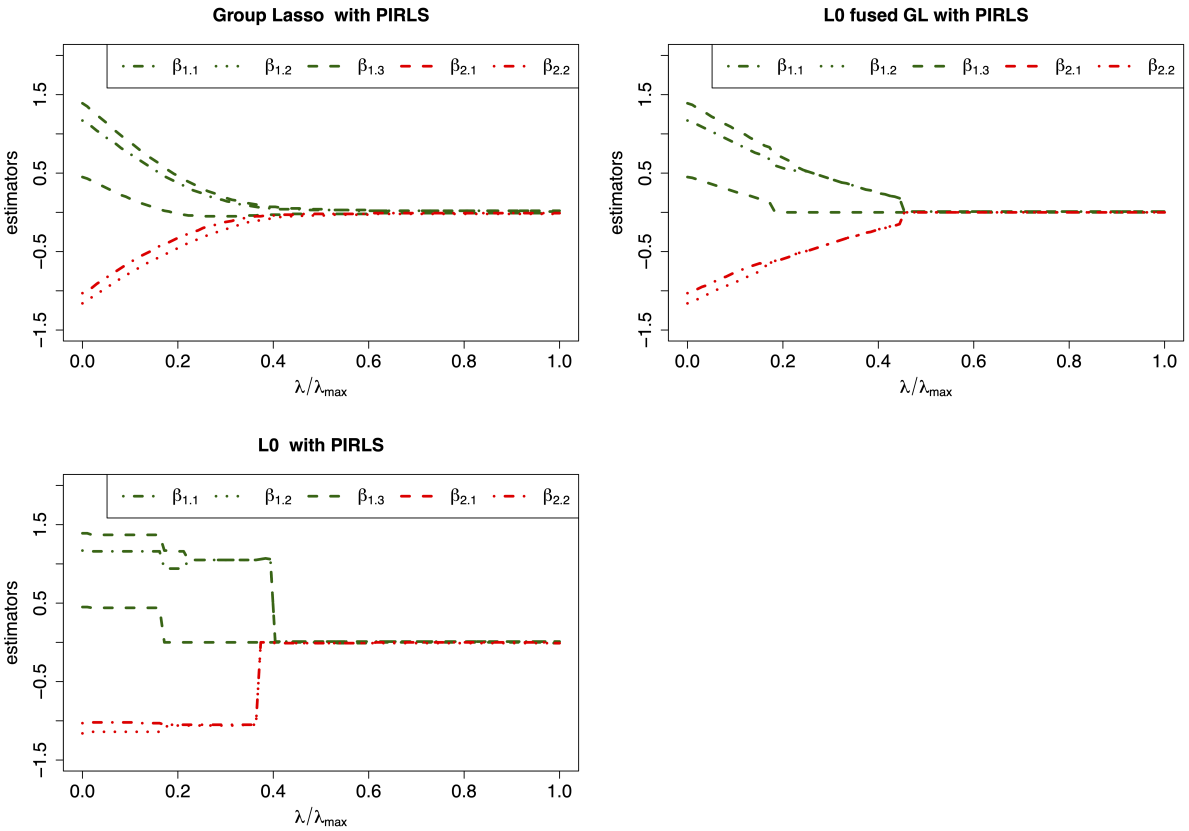

Figure 3.1: *Coefficient paths for several methods. In the first row on the left hand side, one can see group lasso and on the right hand side L_0-FGL, whereas in the second row one can see CAS-L_0. All methods are computed with PIRLS. This figure can similarly be found in Kaufmann and Kateri (2024).*

(first row) L_0-FGL is displayed while in the second row the paths for CAS-L_0 are provided. As elaborated in Chapter 1, group lasso performs *factor* selection, including or excluding *factors* from the model and performing shrinkage of coefficients. However, since PIRLS uses a quadratic approximation of group lasso, the paths in Figure 3.1 show that the coefficients of the factors are shrunken towards zero, while they are not completely set to zero, different from the paths shown in Figure 1.6. But, since PIRLS is only applied to group lasso for visualization purposes and not in the simulation studies, this is not an issue. Turning the focus to the paths of L_0-FGL, this issue does not occur since a further L_0 penalty on the differences is incorporated. The paths demonstrate that L_0-FGL performs shrinkage of coefficients, further it fuses levels of the same factor if they are close enough to each other, and it performs factor selection. In contrast, the CAS-L_0 penalty on the right hand side only performs levels fusion. To sum up, the coefficient paths underline the purpose for which L_0-FGL is introduced, simultaneously shrinking coefficients towards zero, performing factor selection as well as levels fusion. Further, they demonstrate that PIRLS is probably a convenient choice to compute the solutions, which is further investigated in the simulation studies in Section 3.2.

3.1.2 Block Coordinate Descent for L_0-FGL

As an alternative to the PIRLS algorithm, a block coordinate descent (BCD) approach applied to L_0-FGL is considered. The general idea of BCD is described in Appendix A.2, while one needs to specify how to deal with the L_0 norm, as well as how the block-wise updates including minimization problems in $\mathbb{R}^{p_j}$ are solved. Further, it needs to be commented on how the group lasso part, which was quadratically approximated in PIRLS, is treated in the BCD approach.

It is recalled that the BCD algorithm cycles through blocks of coordinates, minimizing the objective function with respect to the current block of coordinates while keeping the others fixed. The objective function is $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ and for the (block-wise) minimization of this function, an approximation for the L_0 part and the log-likelihood is employed. For this, it is recalled that $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$, given by (2.3) is defined as the sum of the negative log-likelihood function $-L_n(\boldsymbol{\beta})$, the group lasso penalty $P_\lambda^{(\text{GL})}(\boldsymbol{\beta})$ and the (CAS-)L_0 penalty $P_\lambda^{(\text{CAS-}L_0)}(\boldsymbol{\beta})$, compare (1.50) for the latter.

For $L_n(\boldsymbol{\beta})$ and $P_\lambda^{(\text{CAS-}L_0)}(\boldsymbol{\beta})$, the same approximations as used in PIRLS (Appendix A.1) are employed. The group lasso part $P_\lambda^{(\text{GL})}(\boldsymbol{\beta})$ is *not* approximated. In particular, for $P_\lambda^{(\text{CAS-}L_0)}(\boldsymbol{\beta})$ one approximates at some $\hat{\boldsymbol{\beta}}^{(k)}$ similar to (A.6)

$$P_\lambda^{(\text{CAS-}L_0)}(\boldsymbol{\beta}) \approx P_\lambda^{(\text{CAS-}L_0)}(\hat{\boldsymbol{\beta}}^{(k)}) + \frac{1}{2}(\boldsymbol{\beta}^T \boldsymbol{A}_\lambda \boldsymbol{\beta} + \hat{\boldsymbol{\beta}}^{(k),T} \boldsymbol{A}_\lambda \hat{\boldsymbol{\beta}}^{(k)}). \tag{3.6}$$

Actually, since one proceeds block coordinate-wise, which refers to factor-wise here, $P_\lambda^{(\text{CAS-}L_0)}(\boldsymbol{\beta})$ is approximated separately for each $j \in \{1, \ldots, J\}$ to ensure the possibility of coordinate-wise minimization. Thus, (3.6) is done separately for all $j \in \{1, ..., J\}$ so one obtains $\boldsymbol{A}_{\lambda,j}$ instead of $\boldsymbol{A}_\lambda$ as specified below. The log-likelihood function $L_n(\boldsymbol{\beta})$ is approximated by a Taylor expansion, similar to Breheny and Huang (2011), that is

$$-L_n(\boldsymbol{\beta}) \approx \frac{1}{2n}(\tilde{\boldsymbol{y}}^{(k)} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{W}^{(k)} (\tilde{\boldsymbol{y}}^{(k)} - \boldsymbol{X}\boldsymbol{\beta}). \tag{3.7}$$

Here, $\boldsymbol{W}^{(k)}$ is a diagonal matrix of weights, which is given by (A.7), for the special structure for logistic regression reference is made to Remark A.1.4. Further, $\tilde{\boldsymbol{y}}^{(k)}$ is the working response introduced in (A.8).

Proceeding factor-wise, for each $j \in \{1, ..., J\}$ one receives the following approximation of the penalty term in some $\hat{\boldsymbol{\beta}}_j^{(k)}$, analogously to (3.6)

$$P_\lambda^{(\text{CAS-}L_0)}(\boldsymbol{\beta}_j) \approx P_\lambda^{(\text{CAS-}L_0)}(\hat{\boldsymbol{\beta}}_j^{(k)}) + \frac{1}{2}(\boldsymbol{\beta}_j^T \boldsymbol{A}_{\lambda,j} \boldsymbol{\beta}_j + (\hat{\boldsymbol{\beta}}_j^{(k)})^T \boldsymbol{A}_{\lambda,j} \hat{\boldsymbol{\beta}}_j^{(k)}). \tag{3.8}$$

It is noted that the quantity $\boldsymbol{A}_{\lambda,j}$ depends on the iteration step k, but an upper index is neglected for simplicity. This factor-wise procedure is possible and reasonable since $P_\lambda^{(\text{CAS-}L_0)}(\boldsymbol{\beta})$ is constructed as the sum of the L_0 penalties, summing over all factors, cf. (1.50). Thus, $P_\lambda^{(\text{CAS-}L_0)}(\boldsymbol{\beta}_j)$ for some factor j only depends on this factor j and is independent of all other factors $i \neq j$, hence the penalty satisfies the separability property in terms of *factors* (cf. Appendix A.2).

Remark 3.1.2. [Details on $\boldsymbol{A}_{\lambda,j}$ for $P_\lambda^{(\text{CAS-}L_0)}(\boldsymbol{\beta}_j)$] First, for the structure of $\mathfrak{L}_j$ as well as $a_{l,j}$, one refers to Example 3.1.1. For factor $j \in \{1, ..., J\}$ with p_j+1 levels, being nominal without

loss of generality, the components of the approximation look as follows

$$\begin{aligned}\boldsymbol{A}_{\lambda,j} &= \lambda_0 \sum_{l=1}^{|\mathfrak{L}_j|} p_l'(||\boldsymbol{a}_{l,j}^T \hat{\boldsymbol{\beta}}_j^{(k)}||_0) \frac{D_l(\boldsymbol{a}_{l,j}^T \hat{\boldsymbol{\beta}}_j^{(k)})}{\boldsymbol{a}_{l,j}^T \hat{\boldsymbol{\beta}}_j^{(k)}} \boldsymbol{a}_{l,j} \boldsymbol{a}_{l,j}^T \\ &= \lambda_0 \sum_{l=1}^{|\mathfrak{L}_j|} \left(\frac{1}{1+\exp(-\gamma|\boldsymbol{a}_{l,j}^T \hat{\boldsymbol{\beta}}_j^{(k)}|)} \right) \left(1 - \frac{1}{1+\exp(-\gamma|\boldsymbol{a}_{l,j}^T \hat{\boldsymbol{\beta}}_j^{(k)}|)} \right) \\ &\quad \cdot \frac{2\gamma \boldsymbol{a}_{l,j} \boldsymbol{a}_{l,j}^T}{\sqrt{(\boldsymbol{a}_{l,j}^T \hat{\boldsymbol{\beta}}_j^{(k)})^2 + c}}. \end{aligned} \tag{3.9}$$

The vectors $\boldsymbol{a}_{l,j} \in \{-1, 0, 1\}^{p_j}$ build the differences of the coefficients corresponding to the same factor, cf. Example 3.1.1. 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.

Consequently, the function $g(\boldsymbol{\beta}_j, \hat{\boldsymbol{\beta}}^{(k)})$ given below provides the covariate-wise approximation of the log-likelihood and L_0 penalty part that is used in the BCD procedure, the remaining group lasso part is discussed below

$$\begin{aligned} g(\boldsymbol{\beta}_j, \hat{\boldsymbol{\beta}}^{(k)}) := \quad & \frac{1}{2n}(\tilde{\boldsymbol{y}}^{(k)} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{W}^{(k)} (\tilde{\boldsymbol{y}}^{(k)} - \boldsymbol{X}\boldsymbol{\beta}) + P_\lambda^{(\text{CAS-}L_0)}(\hat{\boldsymbol{\beta}}_j^{(k)}) \\ & + \frac{1}{2}(\boldsymbol{\beta}_j^T \boldsymbol{A}_{\lambda,j} \boldsymbol{\beta}_j + \hat{\boldsymbol{\beta}}_j^{(k),T} \boldsymbol{A}_{\lambda,j} \hat{\boldsymbol{\beta}}_j^{(k)}). \end{aligned} \tag{3.10}$$

For the group lasso part in L_0-FGL, it is recalled that the convenient choice of matrix $\boldsymbol{K}_j = p_j \cdot \boldsymbol{I}_{p_j \times p_j}$ was made such that $||\boldsymbol{\beta}||_{K_j} = \sqrt{p_j}||\boldsymbol{\beta}_j||_2$. The group lasso part $\sqrt{p_j}||\boldsymbol{\beta}_j||_2$ is sub-differentiable and even differentiable except in $\boldsymbol{\beta}_j = \boldsymbol{0}$. Thus, there is no need to employ an approximation of this function.

To sum up, one ends up with the following function $\tilde{g}$, being an approximation of the objective function $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$

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

For the minimization of $\tilde{g}$, a quasi Newton algorithm is used, for which the function `optim()` in `R` is applied. This leads to the BCD quasi Newton algorithm for L_0-FGL provided in the following algorithm.

Algorithm 3.1.3 (BCD for L_0-FGL with quasi Newton)**.**

(i) The start value is set as $\hat{\boldsymbol{\beta}}^{(0)} = 0$ if not specified otherwise. One sets $k = 1$.

(ii) While $|\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}^{(k-1)}| > \varepsilon$ and $k \leq$ maxsteps:
one updates approximation of $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$

(ii.a) for $j = 1, ..., J$
one sets $\tilde{g}(\boldsymbol{\beta}_j, \hat{\boldsymbol{\beta}}^{(k)}) := g(\boldsymbol{\beta}_j, \hat{\boldsymbol{\beta}}^{(k)}) + \lambda_1 \sqrt{p_j}||\boldsymbol{\beta}_j||_2$
one uses quasi Newton to obtain $\hat{\boldsymbol{\beta}}_j^{(k+1)} = \arg\min_{\boldsymbol{\beta}_j} \tilde{g}(\boldsymbol{\beta}_j, \hat{\boldsymbol{\beta}}^{(k)})$

(ii.b) one sets $\hat{\boldsymbol{\beta}}^{(k+1)} = (\hat{\boldsymbol{\beta}}_1^{(k+1)}, ..., \hat{\boldsymbol{\beta}}_j^{(k+1)}, \hat{\boldsymbol{\beta}}_{j+1}^{(k)}, ..., \hat{\boldsymbol{\beta}}_J^{(k)})$
one sets $k = k + 1$.

(iii) One sets $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})} = \hat{\boldsymbol{\beta}}^{(k+1)}$.

Coefficient Paths (BCD)

This subsection provides coefficient paths for group lasso, CAS-L_0 as well as L_0-FGL, all obtained with the BCD quasi Newton algorithm introduced above (Algorithm 3.1.3).

For this, $J = 2$ factors with four levels each are considered, thus $p_1 = p_2 = 3$, drawn from a multinomial distribution with equal probabilities. The truly underlying coefficient vector is chosen to be

$$\boldsymbol{\beta}^* = (-0.5, 2, 2, 1.8, 0.5, 1, 0.5).$$

For the minimum and maximum values of the tuning parameters, value zero was chosen as minimum, whereas the maximum value was chosen such that all factors are excluded from the model. The resulting coefficient paths are provided in Figure 3.2. On the left hand side in every plot, the unpenalized MLE is displayed, while proceeding to the right the tuning parameter is increased, where for L_0-FGL both tuning parameters are increased simultaneously, similar to the coefficient paths shown for PIRLS.

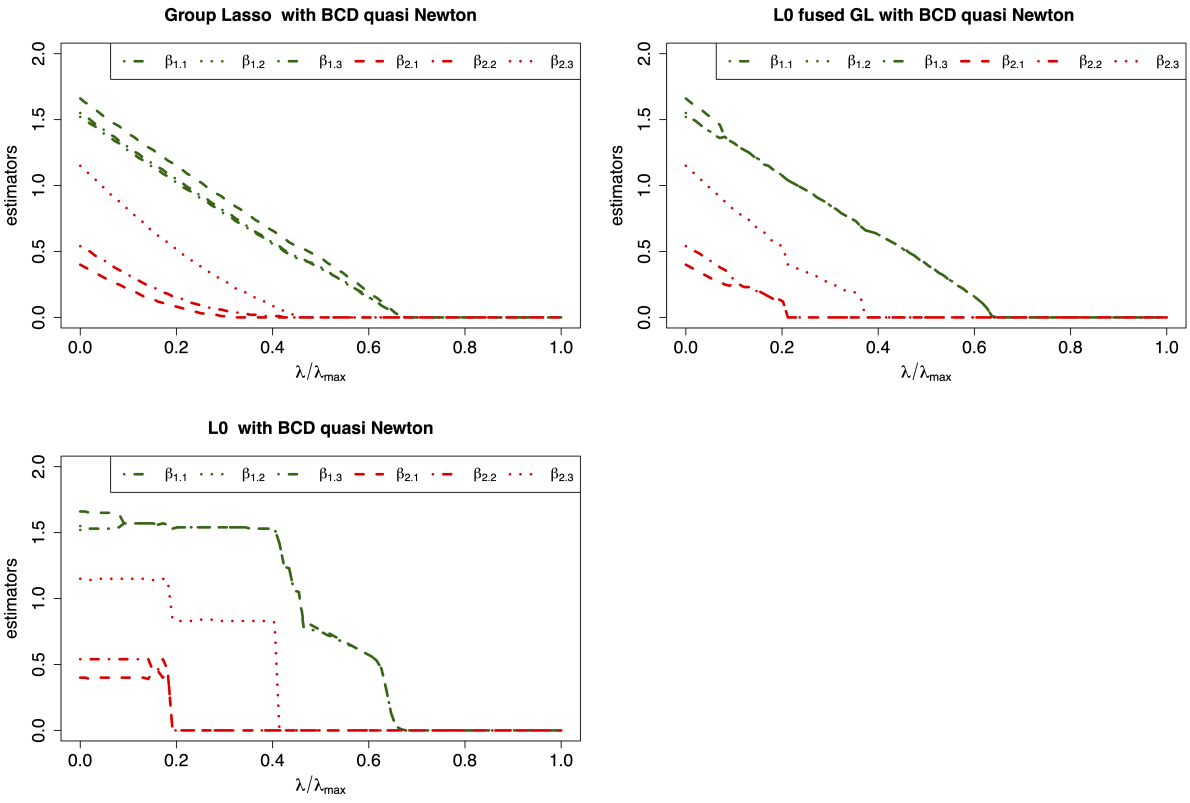

Figure 3.2: *Coefficient paths for several methods. In the first row on the left hand side, one can see group lasso and on the right hand side L_0-FGL, whereas in the second row one can see CAS-L_0. All methods are computed with BCD.*

Observing the paths, one can see that L_0-FGL connects the ability of group lasso to select factors and of CAS-L_0 to fuse coefficients corresponding to the same factor if they are close enough to each other. The same observations were made analyzing the coefficient paths using PIRLS. Comparing the paths of BCD and PIRLS, it is observed that using BCD and quasi

newton, the paths of group lasso (left) in Figure 3.2 look less smooth than those in Figure 3.1 where PIRLS was used. The reason for that is PIRLS uses a quadratic approximation of the *whole* penalty function, while BCD is not approximating the group lasso part.

One can conclude that both approaches, the PIRLS algorithm and the BCD algorithm look promising being applied to L_0-FGL. Next, in Section 3.2, these algorithms are examined and compared in different simulation studies, which show their potential as well as their limitations.

3.2 Simulation Studies

Now that algorithms to solve the resulting minimization problem of L_0-FGL are introduced, their performance is investigated in simulation studies. The goodness of fit measures, as well as details on the simulation designs can be found in Section 1.7.1 and Section 1.7.3, respectively. However, it needs to be clarified which methods are compared, that is, what kind of weights and tuning method are used, as well as which algorithm is applied.

3.2.1 Methods

The methods that are compared are provided in Table 3.1. For L_0-FGL including two tuning parameters, the *stepwise* and the *iterative* tuning method, introduced in Section 2.1.1 are compared. The iterative method is applied for L_0-FGL fitted with PIRLS and not with BCD, the reason for that is given analyzing the results of the simulation studies. Since L_0-FGL is introduced to increase the factor selection of CAS-L_0, the focus lies on comparing L_0-FGL with this method. However, it is also resorted to the results of group lasso, group SCAD and group MCP provided in the simulation studies of Chapter 1, comparing these results to L_0-FGL. Further, the unpenalized ML, referred to as `ML`, is included as reference similar to Chapter 1.

Method	Algorithm	Weights	Tuning	Referred to as
CAS-L_0	PIRLS	non-adaptive	CV	`L0.CV`
CAS-L_0	PIRLS	adaptive	CV	`L0.adap.CV`
L_0-FGL	PIRLS	non-adaptive	stepwise CV	`L0.FGL.PIRLS`
L_0-FGL	PIRLS	adaptive	stepwise CV	`L0.FGL.PIRLS.adap`
L_0-FGL	PIRLS	non-adaptive	iterative CV	`L0.FGL.PIRLS.iterative`
L_0-FGL	PIRLS	adaptive	iterative CV	`L0.FGL.PIRLS.adap.iterative`
L_0-FGL	BCD	non-adaptive	stepwise CV	`L0.FGL.BCD`
L_0-FGL	BCD	adaptive	stepwise CV	`L0.FGL.BCD.adap`

Table 3.1: *Methods compared in simulation studies of Chapter 3.*

Whenever needed for brevity, the expression `iterative` is written as `it`.

Choice of Weights

In Table 3.1, it is specified whether adaptive or non-adaptive weights are used in the corresponding method, which is specified in the subsequent paragraphs. For L_0-FGL, (adaptive) weights can be chosen in the L_0 part *and* in the group lasso part of the penalty function.

For the *non-adaptive weights* in the L_0 part of L_0-FGL, as well as for the non-adaptive weights in CAS-L_0, the weights (1.41) are used for a nominal factor while (1.42) are used for an ordinal factor. Further, for L_0-FGL in the group lasso part, the non-adaptive weight $w_1^{(j)} = \sqrt{p_j}$ is

chosen, accounting for the fact that the factors may have a different number of levels.

For the *adaptive weights* in the L_0 part of L_0-FGL, as well as for the adaptive weights in CAS-L_0, the non-adaptive weight is multiplied with the inverse of the difference of the corresponding unpenalized MLE $\hat{\boldsymbol{\beta}}^{(\mathrm{ML})}$, as in (1.44) and (1.45) with initial estimator chosen as $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(\mathrm{ML})}$. For the group lasso part of L_0-FGL, the non-adaptive weight is multiplied with the inverse of the norm of the sub-vector corresponding to factor j, that is $w_1^{(j)} = \sqrt{p_j}||\hat{\boldsymbol{\beta}}_j^{(\mathrm{ML})}||_2^{-1}$.

Remark 3.2.1. In the case of $|\hat{\beta}_{j,r}^{(\mathrm{ML})} - \hat{\beta}_{j,s}^{(\mathrm{ML})}| = 0$, there arise problems with the adaptive weights in the L_0 part dividing by zero. However, in this case one can replace $|\hat{\beta}_{j,r}^{(\mathrm{ML})} - \hat{\beta}_{j,s}^{(\mathrm{ML})}|$ by $|\hat{\beta}_{j,r}^{(\mathrm{ML})} - \hat{\beta}_{j,s}^{(\mathrm{ML})}| + \frac{1}{n}$, which is in line with Xin et al. (2017) and Zou and Zhang (2009). The same applies for the adaptive weights used for the group lasso part. However, in the simulation studies for this thesis this issue did not occur. It is expected that this issue may occur when using some *penalized* regression estimator as initial estimator on which the adaptive weights are based. This is caused by the fact that, depending on the method, the initial estimator may enforce that sub-vectors are set to zero or levels are fused, causing that the denominator of the adaptive weight may be zero.

Computational Details and Tuning

As in the simulation studies provided in Chapter 1, CAS-L_0 was fitted with the `R` package `gvcm.cat`, further $c = 10^{-5}$, $\gamma = 10$ and $\nu = 0.05$ were chosen. The same choices were made for all L_0-FGL approaches. For the CV range in CAS-L_0, again the default settings implemented in `gvcm.cat` were chosen. As upper and lower values for the L_0-FGL tuning parameter $\lambda = (\lambda_0, \lambda_1)$, the choice $\lambda_{lower} = (0, 0)$ was made and for $\lambda_{upper} = (\lambda_{upper,0}, \lambda_{upper,1})$ the values were chosen in a way that all factors are excluded from the model, where $\lambda_{upper,0} = \lambda_{upper,1}$ was used to ensure that no focus is set on factor selection or levels fusion. Since the tuning method incorporating two tuning parameters is much more computationally involving, $n_\lambda = 10$ different values were fitted for *both* tuning parameters in the specified tuning range for L_0-FGL in all considered designs. For all methods, $k = 5$ fold CV was applied.

3.2.2 Analysis of the Results

Having obtained all necessary details on the methods and the conducted simulation designs, this section discusses the corresponding results.

Results of Design B8.1

For all observed methods, no failures were observed in any of the replications. In Section 1.8.1, where the results of B8.1 for the methods considered in Chapter 1 are examined, it was inferred that the factor selection performance of `L0.CV` can still be improved, while keeping the ability to perform fusion tasks, for which L_0-FGL was introduced. Thus, it is investigated in the following whether L_0-FGL fulfills this purpose.

As a start, Table 3.2 is analyzed showing the FP/FN rates concerning factor selection as well as levels fusion. At first sight, which is also underlined by the OS/PS rates in Table 3.3 and the errors of Figure 3.3, one can see that the L_0-FGL fitted with BCD in both adaptive and non adaptive versions, does not show a satisfactory performance. Thus, `L0.FGL.BCD` and `L0.FGL.BCD.adap` are neglected in the further analysis of this design B8.1.

However, comparing `L0.CV` and `L0.FGL.PIRLS`, one can see that `L0.FGL.PIRLS` does not improve factor selection and performs worse levels fusion. In contrast, the corresponding adaptive version `L0.FGL.PIRLS.adap` clearly improves the factor selection performance, however, the fusion performance is worse. Observing L_0-FGL with PIRLS and the iterative tuning method, that is `L0.FGL.PIRLS.iterative`, there is a high improvement in $FP_{s,fac}$ compared to `L0.CV`, which is decreased from 0.67 to 0.09. Further, one can notice an improvement of the fusion performance, where both corresponding rates are significantly decreased, in particular from 0.43 to 0.11 and from 0.19 to 0.15, respectively. Similar arguments apply comparing the adaptive versions, that is `L0.adap.CV` and `L0.FGL.PIRLS.adap.iterative`, where the improvement is less distinct than for the non-adaptive versions. In terms of FP/FN rates, over all approaches, `L0.FGL.PIRLS.iterative` or the adaptive version `L0.FGL.PIRLS.adap.iterative` should be clearly favored over `L0.CV`. Furthermore, taking into account the results of FP/FN rates for group lasso, group SCAD and group MCP provided in Section 1.8.1, in particular Table 1.4, the new approaches `L0.FGL.PIRLS.iterative` and the adaptive version `L0.FGL.PIRLS.adap.iterative` also outperform group lasso, group SCAD and group MCP, with the factor selection performance being similar, but the fusion performance is considerably improved.

	ML	L0.CV	L0. adapt. CV	L0.FGL. PIRLS	L0.FGL. PIRLS. adap	L0.FGL. PIRLS. it	L0.FGL. PIRLS adap.it	L0.FGL. BCD	L0.FGL. BCD. adap
$FP_{s,fac}$	1.00	0.67	0.36	0.73	0.09	0.09	0.04	0.96	0.98
$FN_{s,fac}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$FP_{f,infl.truth}$	1.00	0.43	0.20	0.72	0.69	0.11	0.12	0.64	0.62
$FN_{f,infl.truth}$	0.00	0.19	0.19	0.04	0.07	0.15	0.20	0.05	0.04

Table 3.2: *[B8.1, n=1000] FP/FN rates fusion and factor selection.*

Coming to the sparsity measures of OS and PS in Table 3.3, a similar pattern is recognized. The resulting model of method `L0.CV` is not sparse enough, with an overall sparsity of 25.51 and a practical sparsity of 6.69 (mean value over all $R = 100$ replications). Also `L0.FGL.PIRLS` reaches not enough sparsity, where the values are even worse for `L0.CV`. In contrast, observing the iterative tuning method `L0.FGL.PIRLS.iterative` and the corresponding adaptive variant `L0.FGL.PIRLS.adap.iterative`, one can see that the resulting models are the most sparse ones, with the OS and PS values being nearest to the true values. Consequently, `L0.FGL.PIRLS.iterative` or the adaptive version `L0.FGL.PIRLS.adap.iterative` are similarly favored over `L0.CV` in terms of sparsity measures. Adding the results of OS/PS of group lasso, group SCAD and group MCP provided in Section 1.8.1, in particular Table 1.3, into the comparison, one can notice that the sparsity levels of `L0.FGL.PIRLS.iterative` and the adaptive version `L0.FGL.PIRLS.adap.iterative` are nearest to the truth.

	ML	L0.CV	L0. adapt. CV	L0.FGL. PIRLS	L0.FGL. PIRLS. adap	L0.FGL. PIRLS. it	L0.FGL. PIRLS adap.it	L0.FGL. BCD	L0.FGL. BCD. adap
OS	40.00	25.51	19.36	32.36	18.54	14.92	13.82	37.17	37.42
PS	8.00	6.69	5.42	6.93	4.36	4.34	4.17	7.86	7.92

Table 3.3: *[B8.1, n=1000] Overall Sparsity (OS) and Practical Sparsity (PS), true values are given by* $OS^* = 16$, $PS^* = 4$.

Finally, the error measures displayed in Figure 3.3 are considered. In terms of predictive deviance, which is displayed in the boxplot on the left hand side, all methods except for the

BCD approach can be located on a comparable scale. Also in terms of MSEC displayed in the boxplot on the right hand side, the differences in the methods, again except for BCD, are not that distinct, however the MSEC values of `L0.FGL.PIRLS.iterative` and the adaptive version `L0.FGL.PIRLS.adap.iterative` are the lowest among all methods.

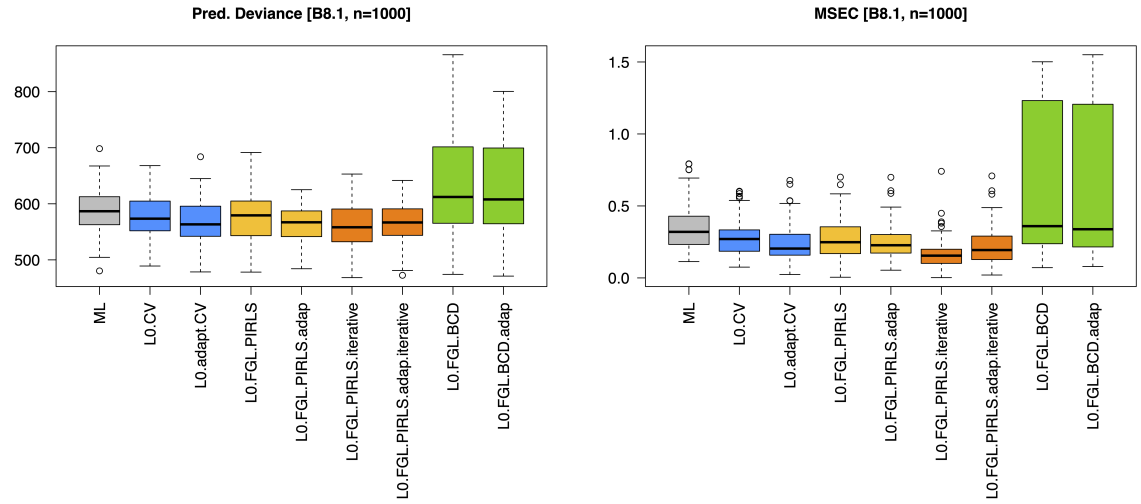

Figure 3.3: *Predictive deviance and mean squared error of coefficients (MSEC) for design B8.1,* $n = 1000$.

To sum up, in all considered error measures, rates and sparsity, the favorable methods are clearly `L0.FGL.PIRLS.iterative` and `L0.FGL.PIRLS.adap.iterative`, which outperformed the other methods and improved the factor selection performance of `L0.CV`, being the purpose for which it is introduced. The investigation also included a comparison to the methods group lasso, group SCAD and group MCP of Section 1.8.1, showing the advantages of `L0.FGL.PIRLS.iterative` and `L0.FGL.PIRLS.adap.iterative`. Further, in this design the stepwise tuning approach used in `L0.FGL.PIRLS` and `L0.FGL.PIRLS.adap` and the BCD approaches `L0.FGL.BCD` as well as `L0.FGL.BCD.adap` seem to be less convenient.

Results of Design B8.2

For all observed methods, no failures were observed in any of the replications. First, as in the previous design B8.1, it is noticed that L_0-FGL with BCD does not perform satisfactory, which can be seen for example in the high $FP_{s,\text{fac}}$ rates (Table 3.4) and OS and PS values (Table 3.5) indicating that nearly no factor selection is performed. Consequently, both adaptive and non-adaptive versions of L_0-FGL fitted with BCD are not considered further for B8.2.

One observes the FP/FN rates provided in Table 3.4. First, one can see that, comparing `L0.FGL.PIRLS` and `L0.FGL.iterative`, the FP rates $FP_{s,\text{fac}}$ and $FP_{f,\text{infl.truth}}$ are clearly improved, at the cost of a (comparably lower) increase in the corresponding FN rates $FN_{s,\text{fac}}$ and $FN_{f,\text{infl.truth}}$. This means that, applying the iterative method, more factor selection and levels fusion is performed, at the cost of some false selections and false fusions. Comparing the corresponding adaptive versions `L0.FGL.PIRLS.adap` and `L0.FGL.adap.iterative` a similar pattern is recognized, however not that distinct as for the non-adaptive versions. This observation is further underlined looking at the sparsity levels in Table 3.5, that is, the iterative tuning method yields a more sparse model, very close to the true sparsity level, than the non-iterative method.

	ML	L0.CV	L0. adapt. CV	L0.FGL. PIRLS	L0.FGL. PIRLS. adap	L0.FGL. PIRLS. it	L0.FGL. PIRLS adap.it	L0.FGL. BCD	L0.FGL. BCD. adap
$FP_{s,\text{fac}}$	1.00	0.62	0.45	0.47	0.16	0.31	0.10	0.87	0.87
$FN_{s,\text{fac}}$	0.00	0.01	0.03	0.28	0.22	0.32	0.30	0.01	0.00
$FP_{f,\text{infl.truth}}$	1.00	0.32	0.23	0.71	0.72	0.53	0.61	0.61	0.64
$FN_{f,\text{infl.truth}}$	0.00	0.21	0.26	0.31	0.26	0.40	0.35	0.13	0.11

Table 3.4: *[B8.2 n=1000] FP/FN rates fusion and factor selection.*

Thus, if a more sparse model is needed, the iterative tuning method should be recommended. However, if false selections or false fusions are not allowed, at the cost of higher false positives, the stepwise tuning method used in `L0.FGL.PIRLS` and `L0.FGL.PIRLS.adap` should be recommended. Additionally, the sparsity levels of e.g. `L0.FGL.PIRLS.iterative` are also closer to the true values than the corresponding levels that were observed for group lasso, group SCAD and group MCP in Table 1.5.

	ML	L0.CV	L0. adapt. CV	L0.FGL. PIRLS	L0.FGL. PIRLS. adap	L0.FGL. PIRLS. it	L0.FGL. PIRLS adap.it	L0.FGL. BCD	L0.FGL. BCD. adap
OS	24.00	16.13	13.93	13.04	10.66	10.40	9.00	20.79	20.78
PS	8.00	6.47	5.67	4.74	3.77	3.95	3.20	7.45	7.46

Table 3.5: *[B8.2, n=1000] Overall Sparsity (OS) and Practical Sparsity (PS), true values are given by* $OS^* = 9$, $PS^* = 4$.

Coming back to the rates table (Table 3.4) comparing `L0.CV` to the methods using L_0-FGL, e.g. `L0.FGL.PIRLS.iterative`, one observes that, on the one hand, the $FP_{s,\text{fac}}$ rate is decreased, at the cost of an increase in $FN_{s,\text{fac}}$ rate. But, the $FN_{s,\text{fac}}$ rate of `L0.CV` of 0.01 in combination with the comparably high $FP_{s,\text{fac}}$ rate indicates that too little factor selection is performed, which is further underlined by the sparsity measures of `L0.CV` given in Table 3.5. Thus, in terms of factor selection looking for a sparse model, `L0.FGL.PIRLS.iterative` would be recommended over `L0.CV`. On the other hand, the levels fusion performance of `L0.CV` is superior to the one of `L0.FGL.PIRLS.iterative`, but this is also what was expected since `L0.CV` is a penalty designed for levels fusion, while L_0-FGL performs both factor selection as well as levels fusion, thus it balances the performance of fusion and selection tasks. Adding a comparison to group lasso, group SCAD and group MCP, where the results are provided in Table 1.6, one can notice that the L_0-FGL approach, taking for example `L0.FGL.PIRLS.iterative`, comes with a moderately worse factor selection performance, but also with an improvement in levels fusion performance, hence, it can be seen once again that it balances both tasks, which is what is expected by its nature.

Turning the focus to the error plots displayed in Figure 3.4, in terms of predictive deviance all approaches are ranked on a similar scale. Moreover, with respect to MSEC, one can find that the adaptive versions of L_0-FGL improve the results of the non-adaptive versions, while in general one can see here that L_0-FGL seems to be a bit more sensitive with respect to changes in the data compared to L_0. Further outstanding is the fact that the adaptive version of L_0, thus `L0.adapt.CV`, does not improve the values for MSEC compared to `L0.CV`. However, for L_0-FGL the adaptive versions for the approaches fitted with PIRLS do improve the MSEC. This may

be explained by the fact that, since the L_0 norm only differentiates between a difference being zero or not, i.e. not depending on the absolute value of the difference, adaptive weights seem to have less of an impact. In contrast, L_0-FGL further incorporates a group lasso part, where the norms of the sub-vectors are calculated, which clearly depend on the actual value of the corresponding sub-vectors. Consequently, employing adaptive weights in L_0-FGL seem to have more of an impact on the MSEC compared to L_0.

Summing up the results of B8.2, the new regularization method L_0-FGL, especially fitted with PIRLS using iterative tuning and adaptive weights, thus `L0.FGL.PIRLS.adap.iterative` improves the factor selection results of L_0, increasing the sparsity level and balancing factor selection and levels fusion performance. The sparsity level of `L0.FGL.PIRLS.adap.iterative` is very close to the true sparsity level, thus this method is able to identify the true underlying sparsity structure, which is supported by the values of $FP_{s,\text{fac}}$ and $FN_{s,\text{fac}}$.

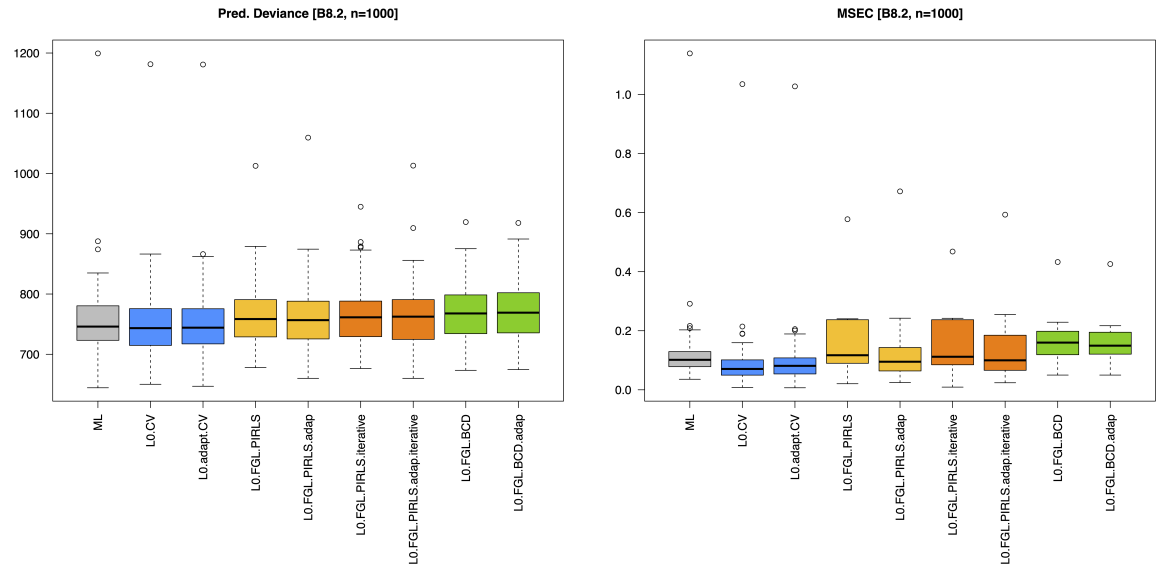

Figure 3.4: *Predictive deviance and mean squared error of coefficients (MSEC) for design B8.2,* $n = 1000$.

Results of Design B6.rare

For all considered methods, no failures were observed in any of the replications. As in the previous designs B8.1 and B8.2, the analysis of `L0.FGL.BCD` and its adaptive variant `L0.FGL.BCD.adap` is neglected since they show no satisfactory performance in this design B6.rare, meaning that the resulting models are not sparse, which further results in comparably high MSEC. It is recalled that the design B6.rare is a design with a rare but relevant category, thus the proportion of $R = 100$ replications, where this rare but relevant (rr) category was excluded from the model is provided.

At first, the FP/FN rates displayed in Table 3.6 are analyzed. Similar to the previous design B8.2, L_0-FGL PIRLS using stepwise tuning results in a less sparse model in terms of factor selection and levels fusion, than with iterative tuning. `L0.FGL.PIRLS.iterative` decreases the $FP_{s,\text{fac}}$ rate for factor selection of `L0.CV` from 0.85 to 0.15, which is outstanding. Nevertheless,

this is at the cost of a, comparably low, decrease in the corresponding $FN_{s,\text{fac}}$ rate from 0.00 to 0.27, where it should be noted that the rates of `L0.CV` indicate that only a slight factor selection is performed, which is underlined by the sparsity measures of Table 3.7. Comparing the fusion performances of these two methods, the $FP_{f,\text{infl.truth}}, FN_{f,\text{infl.truth}}$ rates of `L0.CV` are given by 0.68/0.25, while for `L0.FGL.PIRLS.iterative` one obtains 0.49/0.30, so the decrease in the $FP_{f,\text{infl.truth}}$ rate is clearly higher than the comparably slight increase in the corresponding $FN_{f,\text{infl.truth}}$ rate. For the corresponding adaptive versions, one can find that the model of `L0.FGL.PIRLS.adap.iterative` is more sparse than `L0.CV`, indicated by the smallest sparsity values for `L0.FGL.PIRLS.adap.iterative` in Table 3.7, which comes at the cost of more false selections in terms of higher $FN_{s,\text{fac}}$.

	ML	L0.CV	L0. adapt. CV	L0.FGL. PIRLS	L0.FGL. PIRLS. adap	L0.FGL. PIRLS. it	L0.FGL. PIRLS adap.it	L0.FGL. BCD	L0.FGL. BCD. adap
$FP_{s,fac}$	1.00	0.85	0.39	0.66	0.16	0.15	0.05	0.92	0.91
$FN_{s,fac}$	0.00	0.00	0.01	0.09	0.14	0.27	0.36	0.01	0.00
$FP_{f,infl.truth}$	1.00	0.68	0.42	0.87	0.83	0.49	0.39	0.82	0.84
$FN_{f,infl.truth}$	0.00	0.25	0.27	0.08	0.08	0.30	0.31	0.08	0.08

Table 3.6: *[B6.rare, n=1000] FP/FN rates fusion and factor selection.*

Further observing the sparsity measures (Table 3.7), one can notice again that the models fitted with methods `L0.FGL.PIRLS.iterative` and `L0.FGL.PIRLS.adap.iterative` obtain the most sparse model over all considered methods, which underlines the observations made in the paragraph above. However, the OS rate of `L0.adap.CV` is nearer to the true value than the one of `L0.FGL.PIRLS.adap.iterative`, thus the model obtained by `L0.FGL.PIRLS.adap.iterative` may be too sparse. But, method `L0.FGL.PIRLS.iterative` nearly hits the true sparsity levels OS* and PS*.

Finally, comparing the rates of Table 3.6 and the sparsity measures of Table 3.7 to the results provided in Section 1.8.3, L_0-FGL fitted with PIRLS using iterative tuning can keep up with the methods designed for factor selection (group lasso, group SCAD, group MCP), i.e. $FP_{s,\text{fac}}$ of for example `L0.FGL.PIRLS.iterative` being 0.15 is similar to the rate for `gSCAD.CV` which is 0.10 and the OS and PS values are competitive, with `L0.FGL.PIRLS.iterative` being a bit more sparse. The latter can further be recognized in the increased $FN_{s,\text{fac}}$ rate of `L0.FGL.PIRLS.iterative` (0.36) compared to `gSCAD.CV` (0.03). However, comparing the fusion rates $FP_{f,\text{infl.truth}}$ and $FN_{f,\text{infl.truth}}$ one can find that `L0.FGL.PIRLS.iterative` is able to perform levels fusion, while `gSCAD.CV` is not. Thus, once again, it can be recognized that L_0-FGL combines the ability of factor selection with levels fusion.

	ML	L0.CV	L0. adapt. CV	L0.FGL. PIRLS	L0.FGL. PIRLS. adap	L0.FGL. PIRLS. it	L0.FGL. PIRLS adap.it	L0.FGL. BCD	L0.FGL. BCD. adap
OS	19.00	15.14	10.54	14.47	10.45	8.32	7.23	17.13	17.03
PS	6.00	5.40	3.53	4.46	2.36	2.05	1.48	5.66	5.65

Table 3.7: *[B6.rare, n=1000] Overall Sparsity (OS) and Practical Sparsity (PS), the true values are given by $OS^* = 9$, $PS^* = 2$.*

Since this design includes a rr category, the focus is turned to Table 3.8, analyzing the pro-

portion of replications where the rr category, and the rare but not relevant (rnr) category respectively, were excluded from the model. Further, the absolute difference of these two proportions is obtained, resulting in a measure whether the method can differentiate between rare but relevant and rare but not relevant. The lowest values for (falsely) exlcluding rr are reached by `ML`, `L0.FGL.BCD` and its adaptive variant `L0.FGL.BCD.adap`, which can be explained by the fact that (nearly) no factor selection or levels fusion is performed, thus the rr category is not excluded, but, as one can see, the rnr is also not excluded. The goal is to find a method which can identify a rare but relevant category and obtain a sparse model, which is of high interest e.g. in medical applications researching rare diseases.
The maximum difference of (falsely) excluding the rr category and (correctly) excluding the rnr category is obtained by `L0.adap.CV`, thus this method seems to differentiate best between rnr and rr. However, `L0.adap.CV` performs less factor selection compared to `L0.FGL.PIRLS.adap.iterative`, which was found above when discussing $FP_{s,\text{fac}}$ and $FN_{s,\text{fac}}$. This is, once again, explained by the fact that by the nature of L_0-FGL and the iterative tuning method, L_0-FGL *balances* factor selection and levels fusion, while L_0 is not able to perform factor selection. Hence, performing L_0, the rr category can only be (falsely) excluded when it is fused with the reference category and not through (false) factor selection. However, this comes at the cost of a less sparse model with L_0, so it depends on the application context which method would be preferred, depending on whether the focus lies on a balance of sparsity and detecting a rr category, or whether one aims to strictly avoid the exclusion of the rr category, accepting a less sparse model. Further, referring to Section 1.8.3, L_0-FGL clearly improves the performance in detecting the rr category compared to group lasso, group SCAD and group MCP, which excluded both the rr and rnr category in none of the replications, caused by their nature of only performing factor selection.

	`ML`	`L0.CV`	`L0.adapt.CV`	`L0.FGL.PIRLS`	`L0.FGL.PIRLS.adap`	`L0.FGL.PIRLS.it`	`L0.FGL.PIRLS adap.it`	`L0.FGL.BCD`	`L0.FGL.BCD.adap`
rr excl.	0.00	0.09	0.05	0.06	0.05	0.21	0.18	0.00	0.01
rnr excl.	0.00	0.25	0.41	0.11	0.13	0.37	0.13	0.07	0.06
abs. diff.	0.00	0.16	0.36	0.06	0.08	0.19	0.13	0.06	0.05

Table 3.8: *[B6.rare, n=1000] Proportion of replications where the rare but relevant (rr) and rare but not relevant (rnr) category was excluded (excl.) from the model. Further, in the last row the absolute difference (abs. diff.) of the two proportions is reported.*

In terms of predictive deviance, displayed in the boxplot on the left hand side of Figure 3.5, all considered methods can be ranked on a comparable scale. Observing MSEC on the right hand side, `L0.CV` and `L0.adap.CV` outperform `L0.FGL.PIRLS` and `L0.FGL.PIRLS.iterative`, including their adaptive variants. This is explained by a similar fact as above, since if a rare but relevant category is excluded, the MSEC is increased since "relevant" in that sense corresponds to a high value of the corresponding coefficient.

Results of Design B6.inter.pos

Among the considered methods, no failures were observed in any of the replications. It is recalled that in this design, there is a positive interaction between two binary factors $\mathscr{X}_1$ and $\mathscr{X}_2$, which is denoted by $\mathscr{X}_3$. In particular, $\mathscr{X}_1$ and $\mathscr{X}_2$ have the same influence on the random

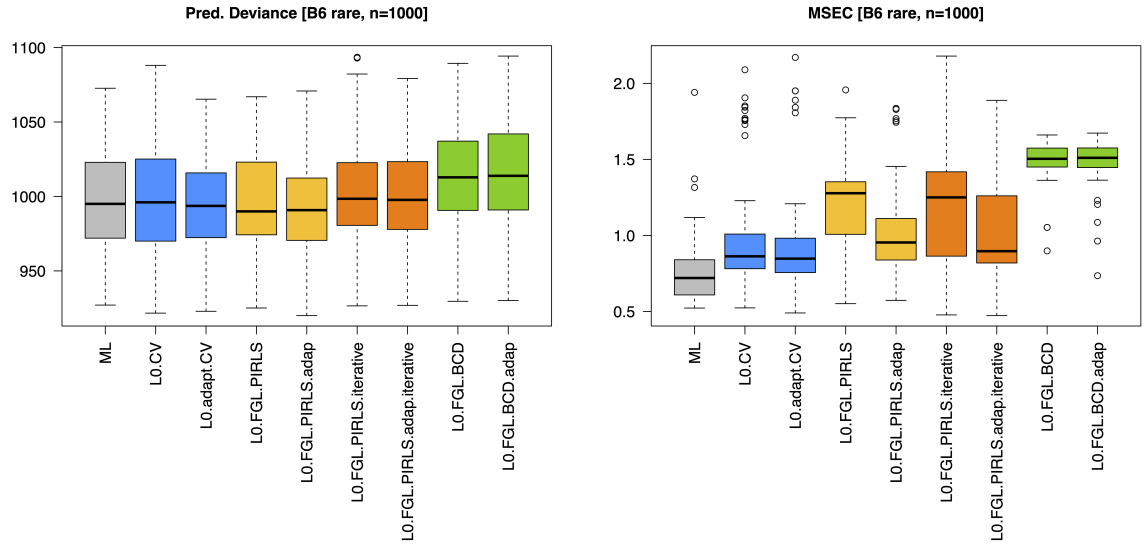

Figure 3.5: *Predictive deviance and mean squared error of coefficients (MSEC) for design B6.rare, $n = 1000$.*

response variable, with the coefficient being equal to 0.5, while the interaction of both $\mathscr{X}_3$ is positive, with coefficient being equal to two, thus the interaction has significantly more influence on the random response. Besides observing rates, as well as sparsity and error measures, the proportion of replications where $\mathscr{X}_1$, $\mathscr{X}_2$ and $\mathscr{X}_3$ are detected as influential by the examined methods is investigated. It is outlined that in this design, different from the designs considered above, `L0.FGL.BCD` as well as the corresponding adaptive version seem to be a convenient choice.

	`ML`	`L0.CV`	`L0. adapt. CV`	`L0.FGL. PIRLS`	`L0.FGL. PIRLS. adap`	`L0.FGL. PIRLS. it`	`L0.FGL. PIRLS adap.it`	`L0.FGL. BCD`	`L0.FGL. BCD. adap`
$FP_{s,fac}$	1.00	0.61	0.41	0.56	0.49	0.35	0.38	0.32	0.26
$FN_{s,fac}$	0.00	0.12	0.14	0.18	0.28	0.28	0.33	0.12	0.14
$FP_{f,infl.truth}$									
$FN_{f,infl.truth}$	0.00	0.17	0.21	0.20	0.30	0.31	0.35	0.23	0.26

Table 3.9: *[B6.inter, n=1000] FP/FN rates fusion and factor selection. By construction of this design, no fusions of truly influential factors need to be performed, thus $FP_{f,infl.truth}$ can not be obtained.*

The analysis is started by the examination of the FP/FN rates in Table 3.9. Since no fusion of truly influential factors needs to be performed, so no fusion tasks appear, the analysis of fusion measures is neglected. Coming to the $FP_{s,\text{fac}}, FN_{s,\text{fac}}$ factor selection rates, the iterative tuning method of L_0-FGL with PIRLS is favorable compared to the stepwise one. Further, comparing the selection performance of `L0.CV` and `L0.FGL.PIRLS.iterative`, similar observations as in the previous designs are made, that is, `L0.FGL.PIRLS.iterative` improves the factor selection performance of `L0.CV`, which is similarly observed looking at the sparsity measures in Table 3.10. Different from the designs considered so far, one can see that the method `L0.FGL.BCD` as well as the adaptive variant `L0.FGL.BCD.adap` further improves the factor selection performance

of the corresponding L_0-FGL methods computed with PIRLS. Hence, among all methods considered in this design, L_0-FGL fitted with BCD is favorable, both in the adaptive and in the non adaptive version. Drawing the connection to Chapter 1, especially Section 1.8.4, one can notice that these two approaches further outperform group lasso, group SCAD and group MCP in terms of $FP_{s,\text{fac}}$, $FN_{s,\text{fac}}$ rates.

	ML	L0.CV	L0. adapt. CV	L0.FGL. PIRLS	L0.FGL. PIRLS. adap	L0.FGL. PIRLS. it	L0.FGL. PIRLS adap.it	L0.FGL. BCD	L0.FGL. BCD. adap
OS	9.00	6.69	5.73	6.09	5.48	4.87	4.74	5.48	5.08
PS	6.00	4.73	4.25	4.41	3.88	3.60	3.44	4.15	3.97

Table 3.10: *[B6.inter, n=1000] Overall Sparsity (OS) and Practical Sparsity (PS), true values $OS^* = 5$, $PS^* = 4$.*

With respect to sparsity measures displayed in Table 3.10 a similar pattern can be recognized as described in the paragraph above. That is, the methods using PIRLS tend to select models that are more sparse than those using only L_0. In particular, the sparsity level resulting from L_0-FGL fitted with BCD and adaptive weights, `L0.FGL.BCD.adap`, is prominently near to the true values. These results are comparable to the OS/PS values of group SCAD and group MCP shown in Section 1.8.4, where these penalties are designed for factor selection, thus L_0-FGL is competitive.

	included factors		
	$\mathscr{X}_3$	$\mathscr{X}_1, \mathscr{X}_2$	$\mathscr{X}_1, \mathscr{X}_2, \mathscr{X}_3$
ML	1.00	1.00	1.00
L0.CV	0.94	0.72	0.67
L0.adapt.CV	0.97	0.62	0.60
L0.FGL.PIRLS	0.98	0.73	0.71
L0.FGL.PIRLS.adap	0.99	0.52	0.51
L0.FGL.PIRLS.it	0.99	0.56	0.55
L0.FGL.PIRLS.adap.it	0.99	0.43	0.42
L0.FGL.BCD	1.00	0.63	0.63
L0.FGL.BCD.adap	0.99	0.63	0.62

Table 3.11: *[B6.inter, n=1000] Proportion of replications where the mentioned factor is included in the model.*

By the nature of this design (B6.inter.pos), the proportion of replications where $\mathscr{X}_1$, $\mathscr{X}_2$ as well as the interaction term are included in the model are of interest. One might expect that, if $\mathscr{X}_1$ and $\mathscr{X}_2$ are included, the interaction $\mathscr{X}_3$ should also be included. For this, the focus is turned to Table 3.11, showing the proportion of replications where $\mathscr{X}_3$ (left column), $\mathscr{X}_1$ and $\mathscr{X}_2$ (middle column) and all three $\mathscr{X}_1, \mathscr{X}_2$ and $\mathscr{X}_3$ (right column) are included in the resulting model. One can notice that for all considered methods, the proportions given in the middle column compared to the right column do not differ significantly, meaning that if $\mathscr{X}_1, \mathscr{X}_2$ are included in the model, they all further include the interaction term $\mathscr{X}_3$ in the majority of cases. However, the value of the difference of the middle and the right column is the highest for `L0.CV`, a fact that was similarly observed in Section 1.8.4. Further, all methods show comparably high values of proportions including $\mathscr{X}_3$, which is also expected because, compared

to the other factors, it has a high influence on the random response in the sense that the true coefficient value is comparably high. However, looking at the column in the middle, one can recognize that `L0.FGL.PIRLS.adap`, `L0.FGL.PIRLS.it` and `L0.FGL.PIRLS.adap.it` show the worst performance in this example, coming with the lowest value of proportion including the influential factors $\mathscr{X}_1$ and $\mathscr{X}_2$. Nevertheless, the focus of this design is to analyze whether *if* $\mathscr{X}_1$ and $\mathscr{X}_2$ are evaluated as influential, *then* the interaction term $\mathscr{X}_3$ is also evaluated as influential, which is satisfied by all methods.

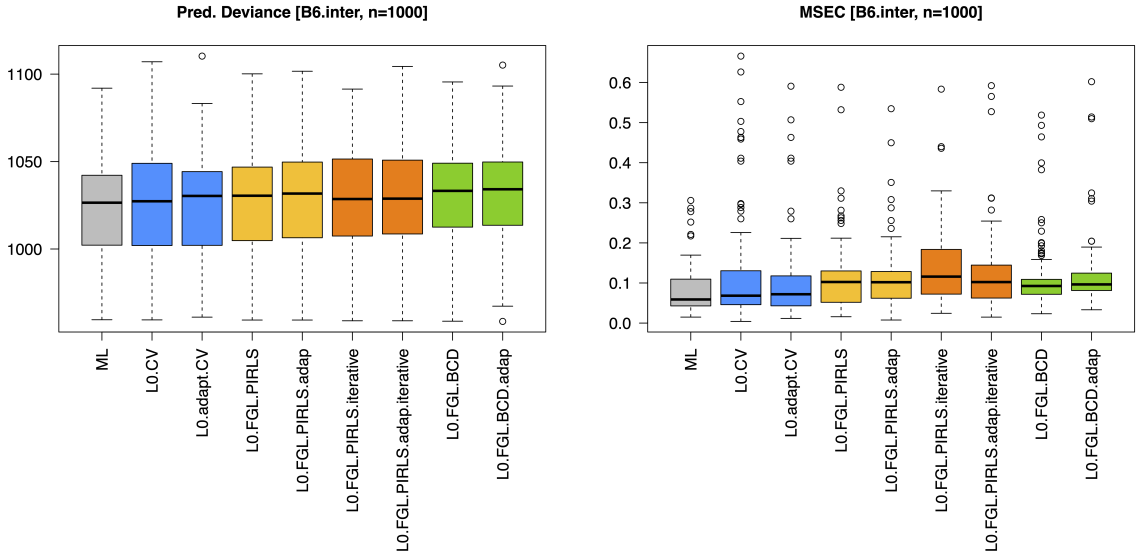

Figure 3.6: *Predictive deviance and mean squared error of coefficients (MSEC) for design B6.inter, $n = 1000$.*

Finally, exploiting the error measures displayed in Figure 3.6, all methods can be ranked on a similar level. Recalling the results above, L_0-FGL fitted with BCD (adaptive and non-adaptive version), which showed the best performance in terms of FP/FN rates and sparsity measures, would be the method of choice for this design.

Results of Design highdim

By construction, this design includes $p = 171$ parameters to be estimated while the sample size is given by $n = 100$. First, the proportion of replications where the methods failed are discussed, which are displayed in Table 3.12. Here, failure refers to the fact that the called function in `R` did not yield an estimate, which may be caused e.g. by convergence issues or by singular matrices.

All methods that include the PIRLS algorithm, except for `L0.adapt.CV` failed in every replication, and also `L0.adapt.CV` failed in the grand majority of replications (87%). Clearly, this is a disadvantage of the PIRLS algorithm and the reasons for these failures depend on the design and are discussed in Appendix A.1. The ML approach only failed in 16% of the replications, however, as noticed exploiting the values for MSEC and predictive deviance below, the results are not satisfactory, thus the ML approach is not a convenient option in this design. Lastly, fitting L_0-FGL with the BCD approach showed no failures, where the corresponding adaptive version failed in 30% of the replications, which can be explained by the missing quality of the ML estimates, which are used as adaptive weights. The fact that the number of failures for the BCD are much lower than for PIRLS comes from the fact that the BCD approach proceeds

	Proportion
ML	0.16
L0.CV	1.00
L0.adapt.CV	0.87
L0.FGL.PIRLS	1.00
L0.FGL.PIRLS.adap	1.00
L0.FGL.PIRLS.iterative	1.00
L0.FGL.PIRLS.adap.iterative	1.00
L0.FGL.BCD	0.00
L0.FGL.BCD.adap	0.30

Table 3.12: *[highdim, n=100] Proportion of replications where the methods failed.*

factor-wise, cycling through the factors and minimizing the objective function w.r.t. one factor, keeping all others fixed. Thus, the high-dimensionality has a significantly lower impact on this algorithm, since in every cycle $j \in \{1, \ldots, J\}$, a minimization problem in $\mathbb{R}^{p_j}$ instead of $\mathbb{R}^{p+1}$ is solved. As a consequence, in further analyses, only ML, L0.adapt.CV, L0.FGL.BCD and L0.FGL.BCD.adap are considered. For the results of L0.adapt.CV it is crucial to keep in mind that the method failed in the grand majority, that is 87%, of the replications, thus the results need to be interpreted carefully.

	ML	L0.adapt.CV	L0.FGL.BCD	L0.FGL.BCD.adap
$FP_{s,\text{fac}}$	1.00	0.17	0.34	0.34
$FN_{s,\text{fac}}$	0.00	0.83	0.63	0.70
$FP_{f,\text{infl.truth}}$	1.00	0.08	0.25	0.19
$FN_{f,\text{infl.truth}}$	0.00	0.91	0.76	0.81

Table 3.13: *[highdim, n=100] FP/FN rates fusion and factor selection.*

Table 3.13 shows the FP/FN rates for factor selection and levels fusion, respectively, whereas in Table 3.14 the sparsity levels OS/PS are displayed. From both of the tables, one can conclude that even though L0.adapt.CV nearly hits the true sparsity levels, in particular OS, the rates of $FP_{s,\text{fac}}$ and $FN_{s,\text{fac}}$ show that the true noise variables are not identified as noise and the other way around. This also results in a comparably high value of MSEC, shown in Figures 3.7 and more precisely in Figure 3.8, in both figures on the right hand side. In combination with the high proportion of failures, L0.adapt.CV is ranked as not convenient for this high-dimensional design.

	ML	L0.adapt.CV	L0.FGL.BCD	L0.FGL.BCD.adap
OS	170.00	15.46	49.66	49.79
PS	60.00	10.00	20.59	20.26

Table 3.14: *[highdim, n=100] Overall Sparsity (OS) and Practical Sparsity (PS), the true values are given by* $OS^* = 15$, $PS^* = 5$.

It remains to discuss the BCD approaches for L_0-FGL, namely L0.FGL.BCD and the adaptive variant L0.FGL.BCD.adap. The OS and PS values in Table 3.14 point out that the resulting models are not sparse enough, which is underlined by the rates of $FP_{s,\text{fac}}$ and $FN_{s,\text{fac}}$ of Table 3.13. However, even though the sparsity levels can still be improved, L_0-FGL shifted the high-dimensional problem to a non high-dimensional problem.

Coming back to Figures 3.7 and 3.8 showing the values for predictive deviance and MSEC of the considered methods, as mentioned above, ML shows no satisfactory performance with an outstanding high predictive deviance and MSEC. Further, in terms of MSEC, `L0.FGL.BCD` and `L0.FGL.BCD.adap` outperform `L0.adapt.CV`, whereas in terms of predictive deviance, they can be located on a similar scale.

Lastly, the connection to the results of the methods considered in Chapter 1 in the high-dimensional design (Section 1.8.5) is drawn. In terms of sparsity levels, it was recognized that group MCP yields a model that is too sparse. In terms of selection performance, group lasso and group SCAD perform better than L_0-FGL, however, this is what one expects by its nature, balancing fusion and selection tasks. It is concluded that L_0-FGL, fitted with BCD, is clearly advantageous compared to L_0, which failed in every replication of this high-dimensional design. Nevertheless, it is desirable that the factor selection performance of `L0.FGL.BCD` and `L0.FGL.BCD.adap` is further improved, especially in this high-dimensional design. For this purpose, a two-stage L_0-FGL is introduced in Chapter 4, which performs further model selection through statistical inference in a two-step procedure.

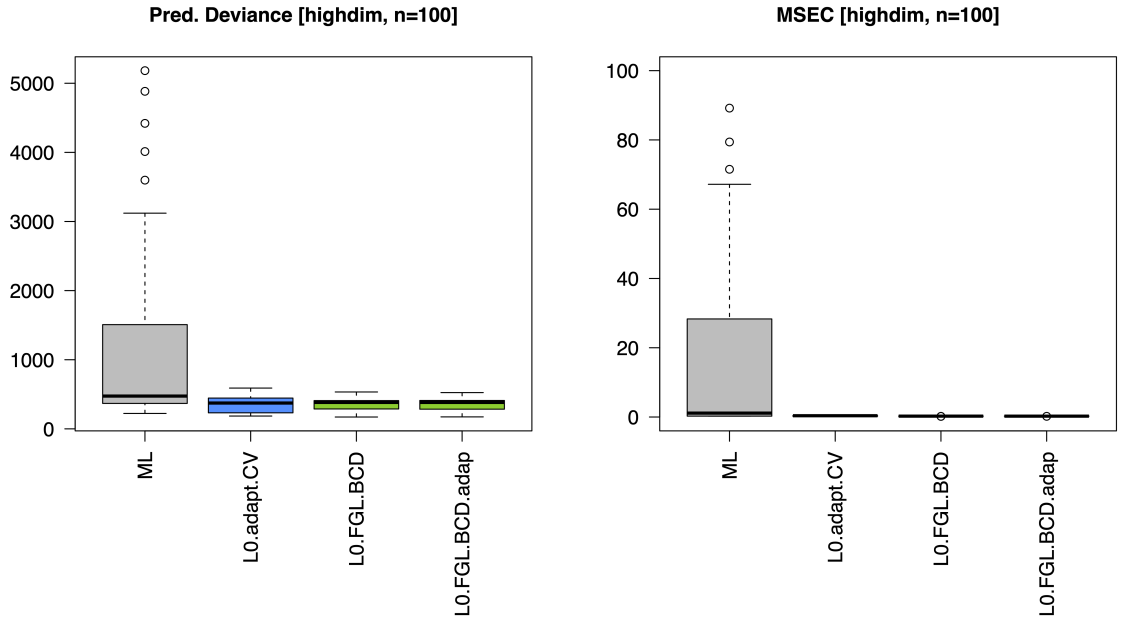

Figure 3.7: *Predictive deviance and mean squared error of coefficients (MSEC) for design highdim, $n = 100$.*

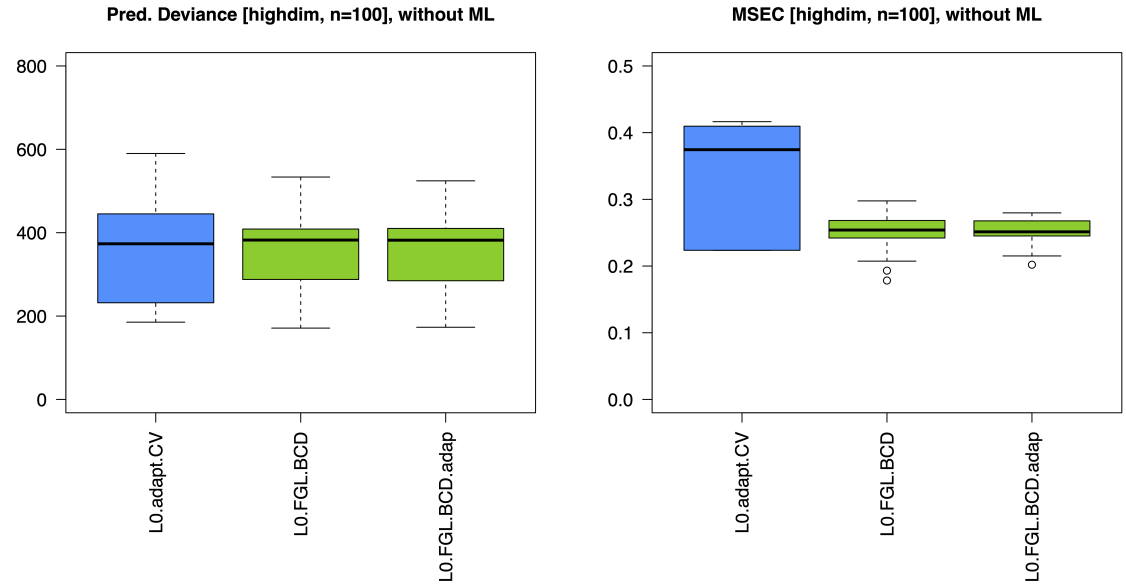

Figure 3.8: *Predictive deviance and mean squared error of coefficients (MSEC) for design highdim, $n = 100$, without ML.*

Chapter 4

Statistical Inference for L_0-FGL

Having introduced L_0-FGL in Chapter 2, along with asymptotic properties, and provided algorithms as well as the performance in simulation studies in Chapter 3, the next natural step is to investigate statistical inference analysis for L_0-FGL. Thus, this chapter provides the theoretical foundation of such a procedure. There are two major characteristics of (high-dimensional) penalized regression avoiding that one can *directly* apply standard procedures for statistical inference, e.g. likelihood ratio tests or score tests. First, the possibility of high-dimensionality may cause inconsistency of the MLE, which is discussed further in Appendix C providing a selection of important theorems in classical ML theory. In Ning and Liu (2017) (Section 2.2), the issues of score tests (Lehmann and Romano (2005), Section 12.4.3) coming with inconsistency of the MLE are considered. Second, since model selection *and* coefficient estimation are performed in one step in penalized regression, one only knows which hypotheses can be tested *after* performing penalized regression, since one needs to know the selected model. However, there are techniques to perform statistical inference in (high-dimensional) penalized regression, which are discussed in the following.

Going through the existing literature, one can recognize that most of the approaches, except for the so-called *sample splitting* approaches which are examined throughout this chapter, can be classified in two different groups: (i) conditional inference and (ii) unconditional inference. For the sake of completeness, the characteristics of these two approaches are briefly discussed in the paragraphs below before proceeding to sample splitting. In general, lasso for the linear model is examined in the majority of references, where there are just a few treating logistic regression and other penalties as lasso.

First, the difference between *conditional* and *unconditional* inference is clarified, before discussing the associated references. The common way of performing hypothesis tests is that the target hypothesis of interest is determined *before* the data for fitting the model is received and then the *pre-determined* hypothesis is tested. In particular, in the (unpenalized) logistic regression setting, a random response variable Y and J predictor variables are investigated, assuming an underlying logistic regression model. Given the underlying model, one can determine the target hypothesis, e.g. test whether a particular predictor variable is influential on the random response variable. If the setting is not high-dimensional, there are several approaches to test such hypotheses *knowing the model*. The issue that arises in the framework of penalized regression performing dimension reduction, in both non high-dimensional and high-dimensional cases, is that the model is not known in advance. For example, with L_0-FGL, factor selection and levels fusion are carried out based on the given data. As a consequence, one cannot determine a hypothesis in advance since the inference target being the penalized estimator $\boldsymbol{\beta}$ changes with the selected model, in particular the dimension of $\boldsymbol{\beta}$ changes. For that reason, one needs to condition on the selected model, which is referred to as *conditional inference*. Otherwise, if the target of inference is the truth $\boldsymbol{\beta}^*$, this problem obviously does not occur since the truth does not change with the selected model, even if the model is selected through penalized regression.

In this case, *unconditional inference* for the truth can be performed.

Now, selected approaches treating (i) conditional inference, including the works of Lee et al. (2016), as well as Lockhart et al. (2014) and finally Taylor and Tibshirani (2018) are examined with respect to their main purpose. First, the work of Lockhart et al. (2014) treats the linear model with a lasso penalization. They construct a covariance test statistic based on the fitted values and show that, assuming that the true model is linear, the asymptotic distribution of the test statistic under the null hypothesis is an exponential distribution. Their proposed test statistic is based on the paths of lasso depending on the tuning parameter λ, testing the significance of a variable entering the active set at a given knot of the path. Second, Lee et al. (2016) obtain post-selection inference for the lasso in the linear model, where they study the resulting conditional distribution using the fact that, for lasso, the selection events can be expressed through a set of linear inequalities in $\boldsymbol{y}$. Hence, the condition that a particular model is selected can be re-written through these linear inequalities. With the help of the so-called *polyhedral lemma*, this set of linear inequalities can be re-written in a particular way, simplifying the quantification of the conditional distribution. In Hastie et al. (2015) (Section 6.3.2) details about the polyhedral lemma and a graphical illustration are provided. Therefore, compared to Lockhart et al. (2014), the work of Lee et al. (2016) is not based on lasso paths. Third, Taylor and Tibshirani (2018) generalize the approach of Lee et al. (2016) explaining how one can treat the lasso in a more general penalized likelihood framework, including logistic regression. Using a Taylor expansion, they express the penalized likelihood problem as a weighted least squares regression (which was similarly done in Section 3.1.2) and then they also apply the polyhedral lemma.

Next, a short review of a selection of existing approaches treating (ii) unconditional inference is provided. In the work of Geer et al. (2014), a de-biased version of lasso is obtained by inverting the Karush-Kuhn-Tucker (KKT) conditions (the bias issue of the lasso was discussed in Section 1.5.1). It is shown that this unbiased lasso has a Gaussian limiting distribution, enabling statistical inference. The work of Zhang and Zhang (2014) is related to the approach of Geer et al. (2014), where in Zhang and Zhang (2014) relaxed projections are used, in particular low-dimensional projections (LDP), in a linear model framework with lasso. It is further generalized by Ma et al. (2020) through a weighted LDP method, referred to as generalized LDP, for de-biasing, making it applicable for logistic regression, where lasso is considered in their work. In Javanmard and Montanari (2014a), a de-biased version of lasso in high-dimensional linear regression is observed based on the sub-gradient of the lasso. Another approach for bias correction is Bühlmann (2013) using Ridge estimation. Other than that, Ning and Liu (2017) introduce the decorrelated score function approach.

The sample splitting approach, which is explained and applied in this chapter, cannot be classified to (i) conditional inference or (ii) unconditional inference, as justified in this chapter. Single sample splitting is proposed in Wasserman and Roeder (2009) and is generalized to multiple sample splitting in Meinshausen et al. (2009), in both cases in the context of linear regression models. These two works are the basis of the idea applying sample splitting to L_0-FGL in this thesis, where details on the differences and challenges are discussed rigorously throughout this chapter. Moreover, it is justified why the approach of sample splitting is the most convenient for L_0-FGL compared to those approaches presented in the paragraphs above.

This chapter is organized as follows. The two-stage L_0-FGL with single sample splitting is introduced in Section 4.1. After that, the extension to multiple sample splitting is provided

in Section 4.2. For both the single and multiple sample splitting two-stage L_0-FGL, not only the idea is introduced, but further convenient asymptotic error control properties are shown, yielding that two-stage L_0-FGL is a reasonable choice with a solid theoretical basis. It is noted that a selection of important theorems in classical ML theory is provided in Appendix C.

4.1 Two-Stage L_0-FGL with Single Sample Splitting

As explained above, one cannot *directly* perform statistical inference by applying the likelihood ratio test provided in Appendix C for two nested models $\mathscr{M}_0$ and $\mathscr{M}_1$ emerging from a *penalized* regression procedure. Consequently, a *two-stage L_0-FGL* procedure is introduced in the following, based on a sample splitting approach enabling statistical inference.

4.1.1 The Splitting Procedure (Single Sample Splitting)

Two-stage L_0-FGL for single sample splitting, as the name indicates, uses a single sample splitting method. The approach of sample splitting, in particular *single sample splitting*, has its roots in Wasserman and Roeder (2009) where it was applied to linear models. However, the authors do not test *factors* being influential, which requires testing whether *sub-vectors* of the coefficient vector are (simultaneously) equal to zero rather than single entries. It is noted that the designation *single* in *single sample splitting* refers to the fact that one performs a *single* data split as specified below (Algorithm 4.1.1), and *not* to testing only a *single* hypotheses. In the work of Meinshausen et al. (2009), the technique of single sample splitting was augmented to *multiple sample splitting* using a pre-specified number $B \in \mathbb{N}$ of splits, yielding more reproducible results.

It is noted that the differentiation between fixed p and diverging p_n is only needed when it comes to asymptotic properties letting $n \to \infty$. Thus, the following methods and quantities for L_0-FGL are introduced for fixed p, the case of diverging p_n is treated later in the provided asymptotic theory, as well as the corresponding regularity conditions that are imposed.

As the definitions of the reduced/full versions of the estimates (e.g. $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$, $\hat{\boldsymbol{\beta}}_{\text{red}}^{(L_0\text{-FGL})}$) and the truth, as well as p_j^{af}, introduced earlier in Section 2.3.4, are crucial in the following, one may recall Notation 2.3.12 and Example 2.3.11 (for the reduced/full versions of the estimates), Notation 2.3.14 and Example 2.3.13 (for the reduced/full versions of the truth) as well as Notation 2.3.10 (for p_j^{af}).

Let $\Omega_1 \subseteq \mathbb{R}^{p+1}$ be a given parameter space for the coefficient vector $\boldsymbol{\beta}$ and let D be a given dataset of sample size $n \in \mathbb{N}$. Then, one can proceed as given in Algorithm 4.1.1 below.

Algorithm 4.1.1 (Two-stage L_0-FGL with single sample splitting)**.**

(i) The data D is split randomly into two different independent parts of equal size, called D_1 and D_2.

(ii) *Step 1*: the L_0-FGL regularization method is performed on the initial parameter space $\Omega_1 \subseteq \mathbb{R}^{p+1}$, which yields $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})} \in \mathbb{R}^{p+1}$, $\hat{\boldsymbol{\beta}}_{\text{red}}^{(L_0\text{-FGL})} \in \mathbb{R}^{1+p^{af}}$ with $p^{af} = \sum_{j=1}^{J} p_j^{af}$ and a selected model denoted by $\tilde{S}_n$ coming with a selected set of factors A_n and a fusion set F_n. Through factor selection and levels fusion, the dimensionality of the parameter vector is reduced, thus the parameter space for step 2 (below) is reduced based on the parameter vector $\hat{\boldsymbol{\beta}}_{\text{red}}^{(L_0\text{-FGL})}$. This parameter space is denoted by Ω_2.

(iii) *Step 2*: MLE is performed on the dataset D_2 considering the parameter space Ω_2. The resulting estimate is denoted by $\hat{\boldsymbol{\beta}}^{\tilde{S}_n}_{\text{red}} \in \mathbb{R}^{p^{af}+1}$ and $\hat{\boldsymbol{\beta}}^{\tilde{S}_n} \in \mathbb{R}^{p+1}$, respectively (stated in Notation 4.1.2).

(iv) Based on the results of step 1 and step 2, likelihood ratio tests (LRT) are performed, testing whether the factors $j \in A_n$ selected by L_0-FGL in step 1 above are influential. In particular, only those factors $j \in A_n$ are kept in the (final) model, that are evaluated as influential by the executed tests, technical details are given during this section.

This procedure is called *two-stage L_0-FGL* for single sample splitting. For a visualization of two-stage L_0-FGL for single sample splitting it is referred to Figure 4.1.

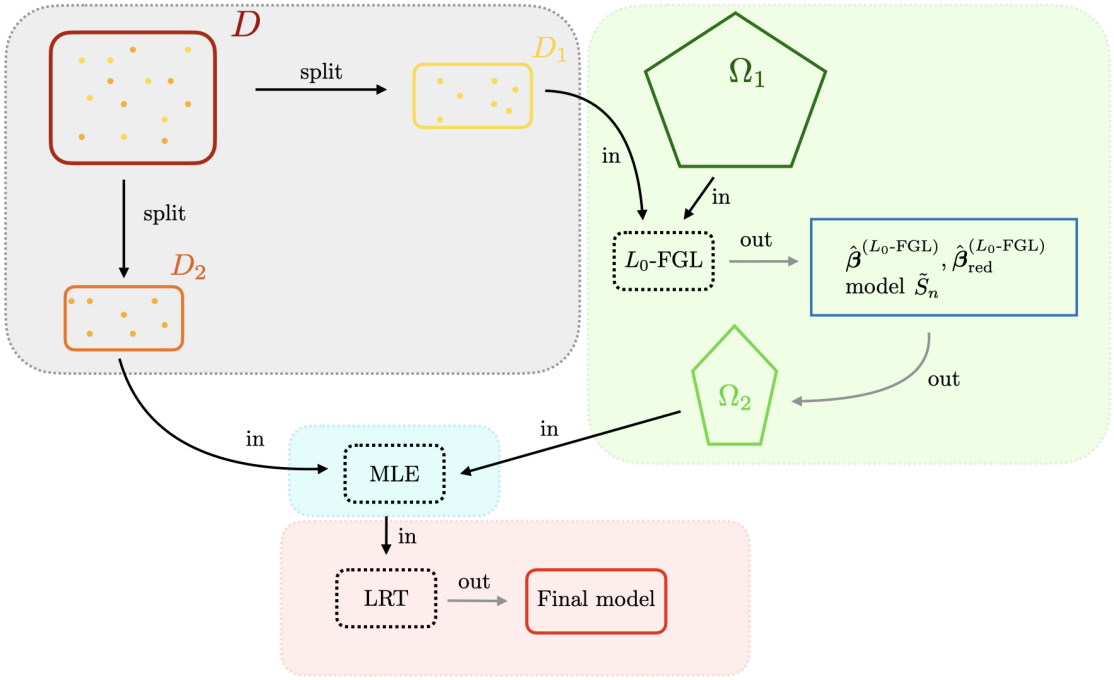

Figure 4.1: *Visualization of two-stage L_0-FGL for single sample splitting based on the LRT, introduced in Section 4.1.1. The light gray box visualizes (i) of Algorithm 4.1.1, the green one (ii), the blue one (iii) and, finally, the red one visualizes (iv).*

Notation 4.1.2. In the procedure described above (Algorithm 4.1.1), the model selected by L_0-FGL is denoted by $\tilde{S}_n$. The coefficient vectors based on the reduced parameter space Ω_2, thus the selected model $\tilde{S}_n$, are denoted by $\boldsymbol{\beta}^{\tilde{S}_n}_{\text{red}}$ (reduced version) and $\boldsymbol{\beta}^{\tilde{S}_n}$ (full version) and the corresponding estimates by $\hat{\boldsymbol{\beta}}^{\tilde{S}_n}_{\text{red}}$ (reduced version) and $\hat{\boldsymbol{\beta}}^{\tilde{S}_n}$ (full version). The corresponding factor-wise versions, thus the sub-vectors, are denoted by $\boldsymbol{\beta}^{\tilde{S}_n}_{j,\text{red}}, \boldsymbol{\beta}^{\tilde{S}_n}_{j}$ and $\hat{\boldsymbol{\beta}}^{\tilde{S}_n}_{j,\text{red}}, \hat{\boldsymbol{\beta}}^{\tilde{S}_n}_{j}$ respectively. The dimensions of full and reduced versions are induced by the dimension reduction through L_0-FGL similar to Notation 2.3.12 and Example 2.3.11. To ensure clarity, the following example is provided.

Example 4.1.3. The setting of Example 2.3.11 is recalled, i.e. $J = 2$ factors are considered as candidate explanatory variables, where $\mathscr{X}_1$ is an ordinal factor with five levels ($p_1 = 4$) and $\mathscr{X}_2$ a nominal factor with four levels ($p_1 = 3$) such that $p = 7$. It is assumed that fitting L_0-FGL in

step 1 of Algorithm 4.1.1 yields the following structure

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})} &= (\hat{\beta}_{int}, \hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2) \\ &= (\hat{\beta}_{int},\ \hat{\beta}_{1,2}, \hat{\beta}_{1,3}, \hat{\beta}_{1,3}, \hat{\beta}_{1,5},\ 0,0,0) \in \mathbb{R}^8.\end{aligned}$$

Thus, levels three and four of $\mathscr{X}_1$ are fused and factor $\mathscr{X}_2$ is identified as non-influential noise variable. The corresponding reduced version (cf. Example 2.3.11) is given by

$$\hat{\boldsymbol{\beta}}_{\text{red}}^{(L_0\text{-FGL})} = (\hat{\beta}_{int},\ \hat{\beta}_{1,2}, \hat{\beta}_{1,3}, \hat{\beta}_{1,5}) \in \mathbb{R}^4,$$

with $p^{af} = p_1^{af} + p_2^{af} = 3$, thus $\hat{\boldsymbol{\beta}}_{\text{red}}^{(L_0\text{-FGL})} \in \mathbb{R}^{1+p^{af}}$. Having that, in step 2 of Algorithm 4.1.1, the MLE on the parameter space Ω_2 (being of dimension $1+p^{af}$ induced by $\hat{\boldsymbol{\beta}}_{\text{red}}^{(L_0\text{-FGL})}$) is obtained, which is denoted by $\hat{\boldsymbol{\beta}}_{\text{red}}^{\tilde{S}_n}$ according to Notation 4.1.2. It is known that this estimate has the same structure as $\hat{\boldsymbol{\beta}}_{\text{red}}^{(L_0\text{-FGL})}$ by construction, i.e. in this example

$$\hat{\boldsymbol{\beta}}_{\text{red}}^{\tilde{S}_n} = (\hat{\beta}_{int},\ \hat{\beta}_{1,2}, \hat{\beta}_{1,3}, \hat{\beta}_{1,5}) \in \mathbb{R}^4,$$

where obviously the numerical values of $\hat{\boldsymbol{\beta}}_{\text{red}}^{\tilde{S}_n}$ and $\hat{\boldsymbol{\beta}}_{\text{red}}^{(L_0\text{-FGL})}$ are not necessarily identical. The full version $\hat{\boldsymbol{\beta}}^{\tilde{S}_n}$ is analogously to $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$ given by

$$\hat{\boldsymbol{\beta}}^{\tilde{S}_n} = (\hat{\beta}_{int},\ \hat{\beta}_{1,2}, \hat{\beta}_{1,3}, \hat{\beta}_{1,3}, \hat{\beta}_{1,5},\ 0,0,0) \in \mathbb{R}^8.$$

It is noted that the quantity $\hat{\boldsymbol{\beta}}_{\text{red}}^{\tilde{S}_n}$ is defined on the reduced parameter space Ω_2, so it is directly reduced by definition, however, the lower index "red" is kept to indicate that the quantity is of dimension $p^{af}+1$, while the full version $\hat{\boldsymbol{\beta}}^{\tilde{S}_n}$ is of dimension $p+1$. With that, notational consistency with the previous chapters (Notations 2.3.12 and 2.3.14) is ensured. The same applies for the other quantities of Notation 4.1.2.

The next section discusses the relation of sample splitting and conditional inference, in particular in terms of nominal and conditional type-I-errors.

4.1.2 Nominal and Conditional Type-I-Errors

In the introduction of this chapter, the concepts of conditional and unconditional inference and the role of sample splitting in statistical testing were touched, where this section supplies more technical details and underlines the advantage of applying sample splitting. The following arguments are taken from Fithian et al. (2017) and Schultheiss et al. (2021), adjusted for the setting used here. Further details on the hypothesis being tested and the employed test are given a subsequent section (Section 4.1.4), whereas here the focus lies on clarifying nominal and conditional type-I-errors along with their difference, for which no further details on the testing procedure are required at this point.

Starting with $J \in \mathbb{N}$ candidate (categorical) explanatory variables and a coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ to be estimated in the underlying *penalized* regression model, one aims to set the coefficients of non-influential factors to zero and for influential ones to fuse those levels having the same influence on the random response, with the help of the L_0-FGL regularization technique. Hence, one assumes that some regularization technique is applied, in fact L_0-FGL, which yields a selected model $\tilde{S}_n$ and an active set of selected predictors $A_n \subseteq \{1,\ldots,J\}$ with $|A_n| = j_0 < J$, for which Definition 1.2.5 is referenced. This dimension reduction induces a reduced parameter space of dimension $p^{af}+1$ (Notation 2.3.10).

In general, *post-selection inference* or *inference after selection*, as the name indicates, deals with inference on the estimator *after* having selected a model based on the data. Now, the question arises whether a method is applied where (i) the model is chosen (based on prior knowledge or data) and then inference is performed based on an independent dataset (independent from that used for model selection) or whether (ii) the model and the hypothesis are chosen adaptively. In the latter case one also chooses the hypothesis adaptively, since, if one does not know which model is selected, one also does not know which hypothesis is tested since statistical tests are only performed for the predictors included in the model, thus only for $j \in A_n$. As mentioned in the introduction of this chapter, the latter case (ii) is also referred to as *conditional inference*.

In particular, in two-stage L_0-FGL for $j \in A_n$, the following hypothesis is tested

$$H_{0,j}: \ \boldsymbol{\beta}_{j,\text{red}}^{\tilde{S}_n} = \mathbf{0} \quad \text{versus} \quad H_{1,j}: \ \boldsymbol{\beta}_{j,\text{red}}^{\tilde{S}_n} \neq \mathbf{0}, \tag{4.1}$$

where the upper index $\tilde{S}_n$ corresponds to the model selected by L_0-FGL (cf. Notation 4.1.2).

For testing a hypothesis $H_{0,j}$ for $j \in A_n$, a "classical" nominal level α-test satisfies that, under the null hypothesis $H_{0,j}$, the type-I-error can be bounded by the nominal level $\alpha \in (0,1)$, i.e.

$$\underbrace{\mathbb{P}_{H_{0,j},\tilde{S}_n}(\text{reject } H_{0,j})}_{\text{(nominal) type-I-error}} \leq \alpha. \tag{4.2}$$

However, for asymptotic tests, the above error-control is reached asymptotically. The lower index $\tilde{S}_n$ in the probability (4.2) indicates that this is a probability under the assumption that the data comes from the model $\tilde{S}_n$. Because of the fact that $\tilde{S}_n$ is not completely specified by the subset of selected predictors A_n since here fusion is allowed besides selection, $\tilde{S}_n$ instead of A_n is written in the lower index. Moreover, the other lower index $H_{0,j}$ indicates that this is a probability under the assumption that $H_{0,j}$ is true. This type-I-error (l.h.s. of (4.2)) is also referred to as *nominal type-I-error*.

In contrast, as Fithian et al. (2017) argue, if the model $\tilde{S}_n$ and the hypothesis being tested are chosen adaptively (described by (ii) above), one aims to control (asymptotically, if necessary) the *selective type-I-error*, also called *conditional type-I-error*, given by the left hand side of

$$\underbrace{\mathbb{P}_{H_{0,j},\tilde{S}_n}\left(\text{reject } H_{0,j} \,|\, (\tilde{S}_n, H_{0,j}) \text{ selected}\right)}_{\text{conditional type-I-error}} \leq \alpha. \tag{4.3}$$

It is noted that one only tests hypothesis $H_{0,j}$ for $j \in A_n$, hence to test $H_{0,j}$ one needs to ensure that j is in the selected subset A_n. Additionally, for the dimension p_j^{af} of the parameter vector being tested, one needs to know the fusions of levels so it is necessary to know the selected model $\tilde{S}_n$, hence the condition on the left hand side of (4.3) is crucial.

Following Fithian et al. (2017), if it can be ensured that model selection is performed on a dataset which is *independent* from the dataset where inference is performed, the *nominal and the conditional type-I-errors are identical*. Consequently, dividing the data D into two different independent parts D_1 and D_2 fitting the model on D_1 and then, having chosen the model, performing inference on D_2 (cf. Algorithm 4.1.1), it can be ensured that

$$\underbrace{\mathbb{P}_{H_{0,j},\tilde{S}_n}(\text{reject } H_{0,j})}_{\text{nominal type-I-error}} = \underbrace{P_{H_{0,j},\tilde{S}_n}\left(\text{reject } H_{0,j} \,|\, (\tilde{S}_n, H_{0,j}) \text{ selected}\right)}_{\text{conditional type-I-error}}. \tag{4.4}$$

To be more precise, the equality above holds since the randomness in selecting $(\tilde{S}_n, H_{0,j})$ emerges from the randomness in D_1, while the randomness in rejecting $H_{0,j}$ emerges from the randomness in D_2. It is noted that the randomness in D_1 and D_2 is caused by the respective included random sample of the random response variable. Thus, ensuring that D_1 and D_2 are independent, the conditional probability on the right hand side of (4.4) equals the unconditional probability on the corresponding left hand side by the definition of conditional probabilities (one compares e.g. Casella and Berger (2002), Definition 1.3.2).

To sum up, the *sample splitting* approach solves the problem of controlling the conditional type-I-error since it equals the nominal type-I-error. Applied to two-stage L_0-FGL this means that after step 1 of Algorithm 4.1.1 it can be assumed that $\tilde{S}_n$, hence A_n, and p_j^{af} $\forall j \in A_n$ are *known*. This makes it possible to derive (asymptotic) size α tests. Thus, ensuring that an independent split is availiable, performing the two-stage procedure with sample splitting yields (4.4) which is crucial in the further course of this chapter.

Next, in Section 4.1.3, the regularity conditions that need to be imposed for two-stage L_0-FGL are discussed.

4.1.3 Assumptions/Regularity Conditions

The conditions provided below are necessary to obtain asymptotic results for two-stage L_0-FGL. It is recalled that A^* is the true active set and A_n is the active set of the estimates, say the estimated active set, for which Definitions 1.2.4 and 1.2.5 are referenced. In what follows, the estimated active set A_n refers to $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$, that is

$$A_n = A_n^{(L_0\text{-FGL})} = \left\{ j \in \{1, \dots J\} \,|\, \hat{\boldsymbol{\beta}}_j^{(L_0\text{-FGL})} \neq \mathbf{0} \right\} \quad \left(= \{ j \in \{1, \dots J\} \,|\, \hat{\boldsymbol{\beta}}_{j,\text{red}}^{(L_0\text{-FGL})} \neq \mathbf{0} \} \right).$$

The following quantities based on the true active set A^* are defined

$$j_1 := |(A^*)^c|, \tag{4.5}$$
$$j_2 := |A^*| = J - j_1. \tag{4.6}$$

Further, one recalls the definition of the fusion sets (Definition 1.2.8), where F_n refers to L_0-FGL in the following, i.e.

$$\begin{aligned} F^* &= \left\{ (j,r,s) \in \{1,\dots,J\} \times \{1,\dots,p_j\}^2 \,|\, \boldsymbol{\beta}_j^* \neq 0,\, \beta_{j,r}^* = \beta_{j,s}^*,\, r < s \right\}, \\ F_n &= \left\{ (j,r,s) \in \{1,\dots,J\} \times \{1,\dots,p_j\}^2 \,|\, \boldsymbol{\beta}_j^* \neq 0,\, \hat{\beta}_{j,r}^{(L_0\text{-FGL})} = \hat{\beta}_{j,s}^{(L_0\text{-FGL})},\, r < s \right\}. \end{aligned}$$

An upper index indicating that the sets F_n and A_n are based on L_0-FGL is neglected for simplicity, since only L_0-FGL is considered in the following.

Remark 4.1.4. In the sets F^*, F_n above, the full/extended versions of $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$ and the truth $\boldsymbol{\beta}^*$ are used since in these sets the indices $r, s \in \{1, \dots, p_j\}$ are analyzed (and *not* just $r, s \in \{1, \dots, p_j^{af}\}$ or $r, s \in \{1, \dots, p_j^*\}$, respectively). However, investigating the requirement $\boldsymbol{\beta}_j^* \neq \mathbf{0}$ in F^* and F_n, it does not matter if one requires $\boldsymbol{\beta}_j^* \neq \mathbf{0}$ or $\boldsymbol{\beta}_{j,\text{red}}^* \neq \mathbf{0}$. The *latter* fact is also valid for the active sets A^* and A_n.

Finally, Notation 2.3.10 is recalled for the definition of p_n^{af}, which is $p_n^{af} := \sum_{j=1}^{J_n} p_j^{af}$. Clearly, for the fixed *and* the diverging case, p_j^{af} depends on n, however, similar to $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$, a lower index is neglected for simplicity. Further, one recalls that in both the fixed *and* the diverging

case, the levels corresponding to one factor, i.e. p_j, are *not* allowed to grow with n, they are assumed fixed $\forall j$.

One considers the following regularity conditions that are imposed for two-stage L_0-FGL.

(A1') It holds that $\lim_{n\to\infty} \mathbb{P}(A^* \not\subseteq A_n) = 0$ (screening property concerning factor selection). In particular one assumes that the assumptions of Theorems 2.3.8 (case of fixed p) and 2.3.9 (case of diverging p_n) are satisfied, ensuring that such an active set exists.

(A2') The resulting model is sparse, hence fitting L_0-FGL on a sample of size n results in an estimate $\hat{\boldsymbol{\beta}}_{\text{red}}^{(L_0\text{-FGL})} \in \mathbb{R}^{p_n^{af}+1}$, where $p_n^{af} < n$. In the case of fixed p, one analogously assumes $p^{af} < n$.

(A3') It holds that $\lim_{n\to\infty} \mathbb{P}(F^* = F_n) = 1$ (screening property concerning levels fusion). In particular, one assumes that the assumptions of Theorem 2.3.37 are satisfied, ensuring that such a fusion set exists.

(A4') The number of factors selected by L_0-FGL is bounded from above, in the sense that $\exists\, k \in \mathbb{N} : |A_n| < k \;\; \forall\, n \in \mathbb{N}$. Further, one assumes that departing from some $n_0 \in \mathbb{N}$, the number of executed fusions is constant $\forall j \in \{1, \ldots, J\}$ ($j \in \{1, \ldots, J_n\}$, respectively).

Remark 4.1.5 comments on the role and convenience of these conditions.

Remark 4.1.5 (On the regularity conditions)**.**

(i) Dividing the initial dataset with $n \in \mathbb{N}$ items into two datasets D_1 and D_2, the model in (A2') is fitted on $n/2$ samples which provides that the resulting dimension after regularization is less than $n/2$. For the case that n is odd, allocate the one remaining sampling item either to D_1 or to D_2. The selected active set is denoted by A_n rather than $A_{n/2}$, even if the model is based on $n/2$ items. Also later on, it is written $p_n^{af} < \frac{n}{2}$ rather than $p_{n/2}^{af} < \frac{n}{2}$ and similarly $p^{af} < \frac{n}{2}$, keeping in mind that the model is based on $\frac{n}{2}$ items.

(ii) In general, assumptions (A1') and (A2') are similar to those in Meinshausen et al. (2009), where the latter (A2') differs since fusion is performed in two-stage L_0-FGL and further grouped variables are treated here. If no grouped variables would be treated, and hence no levels fusion would be performed, one would assume that $|A_n| < \frac{n}{2}$ as in Meinshausen et al. (2009).

i) Conditions (A1') and (A3') are called *screening properties*, in particular (A1') concerning factor selection and (A3') concerning levels fusion. These conditions are crucial to ensure that two-stage L_0-FGL is a reasonable method as specified thoroughly in the course of this chapter. It is further underlined that these properties are proven for L_0-FGL in Chapter 2, which emphasizes the importance of the provided theoretical background.

(iv) First, (A4') ensures that the number of hypotheses being tested after step 2 (of Algorithm 4.1.1) is bounded from above, which is *not* a restriction on p_n or the growth of p_n, it only provides that the number of factors selected by L_0-FGL is bounded by some arbitrarily high constant being a reasonable assumption under the sparsity assumption (Definition 1.2.4). Second, (A4') assumes that for each j, there exists some $n_0 \in \mathbb{N}$ such that $p_j^{af} = p_{j,0}^{af}\ \forall n \geq n_0$ for some $p_{j,0}^{af} \in \{1, \ldots, p_j\}$. However, to avoid unnecessary notational complexity, p_j^{af} continues to be used in the proofs for $n \to \infty$, since it is only crucial that the quantity p_j^{af} is constant departing from some n_0. Without going into further details, this assumption

ensures that the degrees of freedom of the asymptotic distribution of the LRS is well-defined.

Remark 4.1.6 (On the truth $\boldsymbol{\beta}^*_{\text{red}}$ being an interior point of the parameter space Ω_2)**.** To ensure convenient asymptotic properties of the MLE, it needs to be ensured that the truth $\boldsymbol{\beta}^*_{\text{red}} \in \mathbb{R}^{p^*+1}$ is an interior point of the parameter space for MLE, i.e. Ω_2 with $\dim(\Omega_2) = 1 + \sum_{j=1}^{J} p_j^{af}$. However, the dimensions of $\boldsymbol{\beta}^*_{\text{red}} \in \mathbb{R}^{p^*+1}$ and Ω_2 may not be the same caused by the L_0-FGL regularization in step 1 (of Algorithm 4.1.1), such that this remark provides a technical explanation on how this requirement is still reasonable. The true parameter vector $\boldsymbol{\beta}^*_{\text{red}} \in \mathbb{R}^{p^*+1}$ has to be filled with zeros (for the true excluded factors and true fusions with reference category that are not detected by L_0-FGL) and repetitions (for the true fusions that are not detected by L_0-FGL). For example, in the same setting as considered in Example 4.1.3, it holds that $\dim(\Omega_2) = 4$ induced by $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}_{\text{red}} = (\hat{\beta}_{int},\ \hat{\beta}_{1,2}, \hat{\beta}_{1,3}, \hat{\beta}_{1,5}) \in \mathbb{R}^4$, where $\mathscr{X}_2$ was evaluated as noise variable and levels three and four of $\mathscr{X}_1$ were fused. Assume that the truth is given by $\boldsymbol{\beta}^*_{\text{red}} = (\beta^*_{int}, \beta^*_{1,2}, \beta^*_{1,3})$ where $\mathscr{X}_2$ is truly a noise variable but, in contrast to $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}_{\text{red}}$, levels three and five of $\mathscr{X}_1$ are fused. Then, it holds that $\dim(\boldsymbol{\beta}^*_{\text{red}}) = 3$ while $\dim(\Omega_2) = 4$, such that the entry $\beta^*_{1,3}$ of $\boldsymbol{\beta}^*_{\text{red}}$ is repeated, i.e. one writes $(\beta^*_{int}, \beta^*_{1,2}, \beta^*_{1,3}, \beta^*_{1,3}) \in \Omega_2$ such that this parameter vector can be an interior point of Ω_2. By the screening properties it is known that, asymptotically, $p^* \leq \sum_{j=1}^{J} p_j^{af}$, so this procedure can be executed ensuring that the dimension of Ω_2 is not smaller than those of $\boldsymbol{\beta}^*_{\text{red}}$.

As discussed in Appendix C, to guarantee desirable asymptotic behavior of the MLE and LRT, appropriate regularity conditions need to be satisfied. In the framework of two-stage L_0-FGL (Algorithm 4.1.1), this means particularly that the regularity conditions (Reg1)-(Reg3) for fixed p ((div.Reg1)-(div.Reg3) for diverging p_n) are needed for the corresponding parameter space where the MLE is executed on, i.e. Ω_2. Analogously, these conditions are needed for the initial parameter space to ensure theoretical properties of L_0-FGL (cf. Chapter 2). These considerations induce the following definition.

Definition 4.1.7 (GlobalReg)**.** In the given setting, the global regularity conditions, abbreviated as GlobalReg, are said to be satisfied if

(i) (Reg1) holds ((div.Reg1) for diverging p_n)

(ii) (Reg2) holds ((div.Reg2) for diverging p_n)

(iii) (A1')-(A4') hold

(iv) (Reg3) holds for both Ω_1 and Ω_2 ((div.Reg3) for diverging p_n)

Now that the regularity conditions are specified, one can proceed to the technical details of the test being executed in two-stage L_0-FGL.

4.1.4 Properties and Details of Likelihood Ratio Test and p-values for Two-Stage L_0-FGL

A likelihood ratio test is employed, taking into account factor selection and levels fusion performed in step 1 of Algorithm 4.1.1. One defines the following models that are compared for some $j \in A_n$

$$\begin{aligned} \mathscr{M}_0^{(j)} &: \text{ model where } \boldsymbol{\beta}^{\tilde{S}_n}_{j,\text{red}} = \mathbf{0} \text{ (nested in } \mathscr{M}_1^{(j)}), \\ \mathscr{M}_1^{(j)} &: \text{ model where } \boldsymbol{\beta}^{\tilde{S}_n}_{j,\text{red}} \neq \mathbf{0}. \end{aligned}$$

The following hypothesis is considered, testing the nested models given above, i.e.

$$H_{0,j} \ : \ \boldsymbol{\beta}_{j,\text{red}}^{\tilde{S}_n} = \mathbf{0} \qquad \text{versus} \qquad H_{1,j} \ : \ \boldsymbol{\beta}_{j,\text{red}}^{\tilde{S}_n} \neq \mathbf{0}. \tag{4.7}$$

Remark 4.1.8 (Role of the screening properties concerning the validity of the test)**.** The arguments provided in Schultheiss et al. (2021) (Section 2.1) concerning which hypothesis is tested and how the validity of the test for the truth (assumed screening) can be ensured are transferred to the setting considered here. On the one hand, one could test

$$H_{0,j} \ : \ \boldsymbol{\beta}_{j,\text{red}}^{*} = \mathbf{0} \ \text{ versus } \ H_{1,j} \ : \ \boldsymbol{\beta}_{j,\text{red}}^{*} \neq \mathbf{0} \tag{4.8}$$

for the sub-vectors of the truth, or, on the other hand, one could test (4.7) corresponding to the selected submodel $\tilde{S}_n$ and $j \in A_n$. Equivalently, in the hypothesis (4.7) one could write the full versions $\boldsymbol{\beta}_j^{\tilde{S}_n}$ of the parameter sub-vectors instead of the reduced ones $\boldsymbol{\beta}_{j,\text{red}}^{\tilde{S}_n}$, since the *independent* entries are crucial, and the full versions arise from the reduced ones by repeating fused entries and adding zeros for noise factors (cf. Notation 4.1.2), similar arguments apply to (4.8). Coming back to the hypotheses above, under the screening properties (A1') and (A3'), (A1') ensures that all truly active factors are included in the selected model while (A3') ensures that the fusions of the truly active variables are executed correctly. Thus, the true model S^* is nested in the selected model $\tilde{S}_n$, which provides by the given explanations according to Schultheiss et al. (2021) that

$$\boldsymbol{\beta}_j^{\tilde{S}_n} = \boldsymbol{\beta}_j^{*} \quad \forall j \in A_n.$$

Consequently, tests and inference statements valid for (4.7) are also valid for (4.8). The same applies in the multiple split case discussed later.

The difference of the number of parameters included in the models $\mathscr{M}_0^{(j)}$ and $\mathscr{M}_1^{(j)}$, particularly the number of entries in their coefficient vectors and *not* the number of factors, is given by

$$r_j \ := \ \text{param. included in model } \mathscr{M}_1^{(j)} - \text{param. included in model } \mathscr{M}_0^{(j)} = p_j^{af},$$

which is assumed to be known when performing the test (cf. Section 4.1.2). The test statistic that is observed for some $\boldsymbol{Y} = \boldsymbol{y}$ is given by (cf. (C.5))

$$T_{j,n}(\boldsymbol{y}) := T\left(\mathscr{M}_0^{(j)}, \mathscr{M}_1^{(j)}, n, \boldsymbol{y}\right) = -2\left(L_n(\hat{\boldsymbol{\mu}}_0^{(j),(\text{ML})}|\boldsymbol{y}) - L_n(\hat{\boldsymbol{\mu}}_1^{(j),(\text{ML})}|\boldsymbol{y})\right), \tag{4.9}$$

where $\hat{\boldsymbol{\mu}}_0^{(\text{j}),(\text{ML})}$ denotes the MLE of $\boldsymbol{\mu}$ under model $\mathscr{M}_0^{(j)}$, i.e. $\hat{\boldsymbol{\mu}}_0^{(\text{ML})}$ emerges from the MLE $\hat{\boldsymbol{\beta}}_0^{(\text{ML})}$ using the underlying model equation, similar to (C.4) of Appendix C.2 for the case of logistic regression. Analogous arguments apply to $\hat{\boldsymbol{\mu}}_1^{(\text{j}),(\text{ML})}$.

Theorem C.2.1 obtains the asymptotic distribution of $T_{j,n}(\boldsymbol{Y})$. One can conclude that, observing a *single* hypothesis, the decision rule for some given observed sample $\boldsymbol{Y} = \boldsymbol{y}$ is for $j \in A_n$

$$H_{0,j} \text{ is rejected} \Leftrightarrow T_{j,n}(\boldsymbol{y}) \geq \chi^2_{p_j^{af},\alpha}, \tag{4.10}$$

where $\chi^2_{p_j^{af},\alpha}$ is the *upper* α quantile of the χ^2 distribution with p_j^{af} degrees of freedom. Because of the structure of the χ^2 distribution and $T_{j,n}(\boldsymbol{y}) \geq 0$, one does not need to take the absolute value of $T_{j,n}(\boldsymbol{y})$ comparing it to the quantile in (4.10). If multiple tests are applied, some multiplicity corrections are necessary, which are introduced later (Bonferroni correction: Section 4.1.5 and Benjamini Yekutieli correction: Section 4.1.6).

The next remark supplies the asymptotic error control for the test provided above.

Remark 4.1.9 (On the type-I-errors and the asymptotics)**.** For the test statistic $T_{j,n}$ the *asymptotic* distribution is known, hence if the null hypothesis $H_{0,j}$ is rejected if and only if $T_{j,n} \geq \chi^2_{p_j^{af},\alpha}$, one obtains

$$\lim_{n\to\infty} \mathbb{P}_{H_{0,j},\tilde{S}_n}(\text{reject } H_{0,j}) = \alpha, \tag{4.11}$$

i.e. *asymptotic* control of the nominal type-I-error.

Next, a closer look is taken on the (large sample) p-values and their characteristics, which are crucial when discussing asymptotic properties of two-stage L_0-FGL.

Large Sample p-Values

$\Theta_{0,j}$ is the parameter space corresponding to null hypothesis $H_{0,j}$ provided in (4.7). In general, the corresponding *exact* p-value for the LRT is defined by

$$P^{exact}_{raw,j,n}(\boldsymbol{Y}) := \sup_{\boldsymbol{\beta}\in\Theta_{0,j}} \mathbb{P}_{\boldsymbol{\beta}}\left(T_{j,n}(\boldsymbol{Y}) \geq T_{j,n}(\boldsymbol{y})\right), \tag{4.12}$$

according to Casella and Berger (2002) (Theorem 8.3.27). The lower index "raw" corresponds to the fact that, later, adjusted p-values are introduced and used, hence this p-value is somehow "raw" in the sense that it is not adjusted. The computation of the exact p-value (4.12) calls for the distribution of the LRS, which is only known *asymptotically*. For this purpose, large sample p-values, also known as asymptotic p-values, are considered, based on the asymptotic distribution of the LRS (Theorem C.2.1). Of course, one could also work with the exact p-values (4.12) in the theoretical work, but it is not possible to calculate the exact values in applications, such that large sample p-values are considered here. One sets

$$\begin{aligned} P_{raw,j,n}(\boldsymbol{Y}) &:= \mathbb{P}_{H_{0,j}}\left(\tau_{p_j^{af}} \geq -2\left(L_n(\hat{\boldsymbol{\mu}}_0^{(j),MLE}|\boldsymbol{Y}) - L_n(\hat{\boldsymbol{\mu}}_1^{(j),MLE}|\boldsymbol{Y})\right)\right) \\ &= 1 - F_j(T_{j,n}(\boldsymbol{Y})), \end{aligned} \tag{4.13}$$

with $\tau_{p_j^{af}}$ being a $\chi^2_{p_j^{af}}$ distributed random variable with distribution function denoted by F_j.

In the following, solely large sample p-values (4.13) are utilized, where *p-value* is written instead of *large sample p-value* for simplicity.

Now, with expression (4.13) of $P_{raw,j}(\boldsymbol{y})$ one can calculate the p-value for each testing problem of the considered form. Basing the test decision on the p-value one proceeds as follows

$$\text{reject } H_{0,j} \Leftrightarrow P_{raw,j,n} \leq \alpha.$$

Remark 4.1.10 proves the equivalence of the test decision being based on the test statistic and on the p-value.

Remark 4.1.10. It can be easily seen that, on the one hand

$$\text{reject } H_0 \Leftrightarrow P_{raw,j,n}(\boldsymbol{y}) \leq \alpha \Leftrightarrow 1 - F_j(T_{j,n}(\boldsymbol{y})) \leq \alpha$$

and on the other hand, by using that for the upper quantile it holds by definition

$$F_j(\chi^2_{p_j,\alpha}) = \mathbb{P}(\tau_{p_j} \leq \chi^2_{p_j,\alpha}) = 1 - \mathbb{P}(\tau_{p_j} > \chi^2_{p_j,\alpha}) = 1 - \alpha.$$

Hence, using that $F_j(\cdot)$ is a monotonically increasing function being a cumulative distribution function, one can deduce

$$\begin{aligned} \text{reject } H_0 \quad &\Leftrightarrow \quad T_{j,n}(\boldsymbol{y}) \geq \chi^2_{p_j,\alpha} \Leftrightarrow F_j(T_{j,n}(\boldsymbol{y})) \geq F_j(\chi^2_{p_j,\alpha}) \Leftrightarrow F_j(T_{j,n}(\boldsymbol{y})) \geq 1-\alpha \\ &\Leftrightarrow \quad F_j(T_{j,n}(\boldsymbol{y})) - 1 \geq -\alpha \Leftrightarrow 1 - F_j(T_{j,n}(\boldsymbol{y})) \leq \alpha. \end{aligned}$$

Consequently, basing the test decision on the test statistic for a specified α is equivalent to employing the corresponding p-value.

For the random variable $P_{raw,j,n}(\boldsymbol{Y})$ one knows that it is *asymptotically* uniform distributed on $[0,1]$, since it is defined by the cumulative distribution function of a continuous random variable, namely $\tau_{p_j^{af}}$, evaluated at $T_{j,n}(\boldsymbol{Y})$, where the latter follows *asymptotically* the same distribution as $\tau_{p_j^{af}}$. In particular, one knows that for a nominal level $\alpha \in (0,1)$ under the null hypothesis $H_{0,j}$ it holds

$$\lim_{n\to\infty} \mathbb{P}_{H_{0,j}} \left(P_{raw,j,n}(\boldsymbol{Y}) \leq \alpha \right) = \alpha. \tag{4.14}$$

This can similarly be written as

$$\mathbb{P}_{H_{0,j}} \left(P_{raw,j,n}(\boldsymbol{Y}) \leq \alpha \right) = \alpha + o(1).$$

The definition of $P_{raw,j,n}$ was provided for selected factors, i.e. $j \in A_n$. Therefore, one defines $\tilde{P}_{j,n}$ which also takes into account those factors $j \notin A_n$. In particular, for $j \notin A_n$, hence the variables that are not selected by L_0-FGL in step 1 (of Algorithm 4.1.1), one sets $\tilde{P}_{j,n} = 1$. For the others, one calculates $\tilde{P}_{j,n} = P_{raw,j,n}$. To sum up

$$\tilde{P}_{j,n} := \begin{cases} P_{raw,j,n} & \text{if } j \in A_n, \\ 1 & \text{if } j \notin A_n. \end{cases} \tag{4.15}$$

Remark 4.1.11 (On setting p-values to one (or not) and the number of true null hypothesis)**.** In the single split, it is actually not necessary to set the p-value to one (cf. (4.15)) for the factors that are not selected in step 1 (of Algorithm 4.1.1), since one just takes into account the previously selected factors $j \in A_n$ in hypothesis testing (by definition of two-stage L_0-FGL in the single split). In contrast, considering multiple splitting later in this chapter (Section 4.2), setting the p-value to one for the factors that are not selected in step 1 is crucial, as justified in the appropriate section. However, it is important to note at this point that the number of *true* null hypotheses for two-stage L_0-FGL in the single split is given by $|(A^*) \cap A_n|$, while for the corresponding multiple split the number of *true* null hypotheses is $|(A^*)^c|$ for which justification is provided at the respective point. The reason why the latter is mentioned here even though the multiple split is not introduced yet is that two lemmata of this section (Lemma 4.1.19 and Lemma 4.1.20) are easily transferrable from single to multiple split and vice versa, adjusting for the number of true null hypotheses. Thus, to avoid repetitions it is shortly commented whenever there exist such a straightforward extension.

To sum up, technical details on the hypothesis being tested after step 2 of Algorithm 4.1.1 in two-stage L_0-FGL were provided. However, only the case of testing *one* hypothesis was considered, so the next step is to apply some multiplicity correction techniques. Consequently, the next two Sections 4.1.5 and 4.1.6 give further details on the Bonferroni and Benjamini Yekutieli corrections for *multiple testing*.

4.1.5 Bonferroni Correction

Recalling the particular structure of two-stage L_0-FGL given in Algorithm 4.1.1, only those factors are kept in the *final* model that are evaluated as influential by L_0-FGL in step 1 and by the executed test after step 2. Hence, having executed the likelihood ratio test and calculated the needed p-values $\tilde{P}_{j,n}$ (cf. (4.15) and Section 4.1.4) for all $j \in A_n$, the final selected set of predictors in the *single split* adjusted for multiple testing, denoted by $A_{n,f}(\alpha)$, is given by

$$A_{n,f}(\alpha) := \left\{ j \in A_n \,:\, \tilde{P}_{j,n}|A_n| \leq \alpha \right\} = \left\{ j \in A_n \,:\, P_{j,n} \leq \alpha \right\}, \tag{4.16}$$

where

$$P_{j,n} := \min\left(\tilde{P}_{j,n}|A_n|, 1 \right). \tag{4.17}$$

Thus, for the indices in $A_{n,f}(\alpha)$ there is statistical evidence that the corresponding factor has an influence on the random response variable. The multiplication of the p-value $\tilde{P}_{j,n}$ with the cardinality of A_n accounts for the fact that multiple tests (in particular $|A_n|$ tests, one for each factor $j \in A_n$) are executed. In other words, the nominal level with respect to which each of the tests is executed is divided by $|A_n|$, which is the well-known *Bonferroni correction* (Dunn (1961)). The Bonferroni correction ensures that the probability of at least one false rejection (among the considered family of hypotheses) is bounded by α, according to Lehmann and Romano (2005) (Theorem 9.1.1). This error is also called family wise error rate (FWER) and plays a crucial role in the subsequent part of this thesis.

It is important to remark that, at this point of the chapter, it is not clear whether Bonferroni correction applied to *two-stage L_0-FGL* satisfies FWER control caused by the preceding regularization step (cf. Algorithm 4.1.1). However, this question will be answered throughout this chapter.

Remark 4.1.12.

(i) If the rejection of $H_{0,j}$ is based on the test statistic $T_{j,n}$ instead of the p-values, as done subsequently in this chapter (Theorems 4.1.14 and 4.1.16), $A_{n,f}(\alpha)$ is written even though it is originally defined through p-values in (4.16). But, it is clear that in this case

$$A_{n,f}(\alpha) = \left\{ j \in A_n \,:\, T_{j,n} \geq c_{j,\alpha/|A_n|} \right\}, \tag{4.18}$$

where $c_{j,\alpha/|A_n|} = \chi^2_{p_j^{af},\alpha/|A_n|}$ is the upper $\alpha/|A_n|$ quantile of the χ^2 distribution with p_j^{af} degrees of freedom.

(ii) In (4.16) one could also write $A_{n,f}(\alpha) = \left\{ j \in \{1, ..., J\} \,:\, \tilde{P}_{j,n}|A_n| \leq \alpha \right\}$ since for $j \notin A_n$ it holds that $\tilde{P}_{j,n} = 1$ and consequently $\tilde{P}_{j,n}|A_n| = |A_n| > \alpha$ since $\alpha < 1$, one compares Remark 4.1.11.

Even though, in general, FWER control is achieved through the Bonferroni correction, it has its limitations as the nominal level with respect to which each of the hypotheses are tested can be very low depending on the number of executed tests. Consequently, for a "large" number of tests, this procedure may allow too little rejections according to Lehmann and Romano (2005) (Section 9.1), i.e. it may be too conservative. For this purpose, an alternative correction for multiple testing is discussed next.

4.1.6 Benjamini Yekutieli Correction

Caused by the possible limitations of the Bonferroni correction explained in the previous Section 4.1.5, the *Benjamini Yekutieli* correction introduced in Benjamini and Yekutieli (2001) and applied for sample splitting in Meinshausen et al. (2009) is further considered.

To weaken the boundedness of the probability of making even one false rejection (FWER), the Benjamini Yekutieli approach aims to control the false discovery rate (FDR). The FDR is the expected value of the ratio of the false rejections among the total rejections. This approach permits more rejections, making it less conservative, of course at the cost of probably more false rejections. In general, according to Lehmann and Romano (2005) (Section 9.1), it holds that FDR $\leq$ FWER where equality holds if all hypotheses are true. Thus, the FDR control is more liberal allowing more rejections. Nevertheless, it depends on the number of hypotheses to be tested which of the two approaches is more convenient so it depends on the particular application. For more on multiple testing and the FDR/FWER control it is referred to Finner and Roters (2001), Sarkar (2002) and Sarkar (2008).

The Benjamini Yekutieli correction introduced below emerges from the well-known Benjamini Hochberg correction obtained by Benjamini and Hochberg (1995), which was initially designed for controlling the FDR for *independent* p-values. Since p-values are random variables (cf. Section 4.1.4), the concept of independent p-values is defined as it is for random variables, where for the latter reference is made e.g. to Casella and Berger (2002) (Definition 4.6.5). The approach of Benjamini and Hochberg (1995) was generalized and adjusted in Benjamini and Yekutieli (2001) for p-values omitting the independence requirement, which is suitable for the setting of this thesis. In fact, both approaches execute the same ordering of the p-values and they reject the hypotheses in the "same" manner, but with a different level $q > 0$ (one compares Theorems 1.2 and 1.3 of Benjamini and Yekutieli (2001)), where the role of q is explained below.

With

$$P_{j,n} := \min\left(\tilde{P}_{j,n}|A_n|, 1\right)$$

as in (4.17), the p-values $P_{j,n}$ for $j = 1, \dots J$ are ordered corresponding to their size

$$P_{(1),n} \leq P_{(2),n} \leq \ldots \leq P_{(J),n}.$$

For some given $q > 0$, one calculates

$$k(q) := \max\left\{i \in \{1, \dots J\} \,:\, P_{(i),n} \leq i \cdot q\right\}$$

and rejects the null hypotheses up to hypothesis $H_{0,(k(q))}$, where the hypotheses are ordered due to their corresponding p-values, Thus, hypotheses $H_{0,(1)}, \dots, H_{0,(k(q))}$ are rejected. Hence one knows that

$$P_{(i),n} \leq k(q) \cdot q \;\; \forall i \in \{1, \dots, k(q)\}. \tag{4.19}$$

The final set of factors included in the model with the multiplicity correction as explained above, is denoted by $A_{n,f}^{BY}(q)$ and is given by

$$A_{n,f}^{BY}(q) := \left\{j \in A_n \,:\, P_{j,n} \leq P_{(k(q)),n}\right\}. \tag{4.20}$$

Clearly, using different corrections, in fact Bonferroni or Benjamini Yekutieli, to adjust for multiplicity of testing, results in different theoretical properties of two-stage L_0-FGL. Hence,

it is important to justify whether one examines the selected set $A_{n,f}(\alpha)$ using Bonferroni or the selected set $A^{BY}_{n,f}(q)$ using Benjamini Yekutieli. As analogously mentioned in the previous section, it is not clear at this point of this chapter whether Benjamini Yekutieli correction applied to *two-stage* L_0-FGL results in FDR control caused by the preceding regularization step (Algorithm 4.1.1), however, this will be clarified throughout this chapter.

4.1.7 Type-I-Error Control and Consistent Selection with Bonferroni Correction

In the following, theoretical properties for the two-stage L_0-FGL procedure using the Bonferroni correction for multiple testing (Section 4.1.5) are developed.

Remark 4.1.13 (Comparison to Wasserman and Roeder (2009))**.** Since the single split method, which is applied for two-stage L_0-FGL, was introduced by Wasserman and Roeder (2009), the main differences of this approach to two-stage L_0-FGL are emphasized here. For the concrete specification of the presented aspects, the corresponding upcoming theorems and proofs are linked.

(i) In Wasserman and Roeder (2009) the linear model with LSE is utilized rather than logistic regression with MLE as done in this thesis. Hence, they use some t-test statistic while here the LRS is used, i.e. the distribution of the test statistic is only known *asymptotically* rather than exact. Consequently, probabilities including the LRS can only be quantified asymptotically (i.e. $n \to \infty$) which impose challenges in the proofs of Theorem 4.1.14 as well as Theorem 4.1.16, as discussed at the appropriate point. To provide a brief example, considering sums of such probabilities where the summation index depends on n, challenges arise since the probability can only be quantified if the limit and the sum can be interchanged which is obviously not straightforward. In contrast, knowing the exact distribution of the test statistic, this challenge does not occur.

(ii) In this thesis L_0-FGL is proposed and used as regularization technique allowing for factor selection *and* levels fusion, while in their work the focus lies (only) on variable selection methods. Hence, it is necessary to add and ensure a screening property concerning fusion (A3') to ensure that the tests are also valid for testing the truth (cf. Remark 4.1.8). The challenge arising here is that this additional screening property (A3') cannot only be assumed to hold, it needs to be proven that it is valid under suitable assumptions. Such a theorem was rigorously presented and proven in Chapter 2, specifically Theorem 2.3.37, which requires several preparatory theorems and corollaries as provided in Section 2.3.4.

(iii) The focus of this thesis is on categorical covariates, i.e. factors, which leads to other hypotheses as in the work of Wasserman and Roeder (2009) since here testing is done on sub-vectors instead of single entries of the coefficient vector. In particular, in their work, the hypotheses are of the form $\beta_j = 0$ whereas in this investigation sub-vectors $\boldsymbol{\beta}_j = \mathbf{0}$ are tested. Consequently, since fusion is allowed here, the dimension of the sub-vector corresponding to a factor after step 1 (of Algorithm 4.1.1) may be of lower dimension, i.e. $p_j^{af} \leq p_j$. Thus, the distribution of the test statistic (in particular its degrees of freedom) and the critical value depend on the regularization of step 1. Since the critical value depends on p_j^{af}, each critical value may be different among the factors, which causes challenges as justified in the proof of Theorem 4.1.14, together with a possibility to overcome this issue.

Theorem 4.1.14 shows that, testing several hypotheses $H_{0,1}, H_{0,2}, ..., H_{0,|A_n|}$, also called a *family of hypotheses*, the probability of one false rejection (i.e. the FWER) does not exceed the level α. This FWER is also called type-I-error corresponding to the *global* null hypothesis. For

this theorem (Theorem 4.1.14), both cases of fixed p and diverging p_n are allowed, as long as GlobalReg (Definition 4.1.7) are ensured to hold. Hence, p without the lower index n is written for simplicity of notation.

Theorem 4.1.14 (Type-I-error control/FWER control of two-stage L_0-FGL in single split with Bonferroni, fixed p and diverging p_n)**.** One assumes that GlobalReg (Definition 4.1.7) hold and the LRT is executed at level $\alpha \in (0,1)$, where the correction for multiplicity is done by Bonferroni correction, thus for $j \in A_n$

$$\text{reject } H_{0,j} \Leftrightarrow T_{j,n}(\boldsymbol{y}) \geq \chi^2_{p_j^{af},\alpha/|A_n|}.$$

Then, it holds that two-stage L_0-FGL controls the type-I-error/FWER, in particular

$$\limsup_{n\to\infty} \mathbb{P}\left((A^*)^c \cap A_{n,f}(\alpha) \neq \emptyset\right) \leq \alpha.$$

Proof. The beginning of this proof roots in Wasserman and Roeder (2009). Setting

$$M_n := A_n \cap (A^*)^c$$

as the set of true noise factors that are (falsely) evaluated as influential by L_0-FGL and setting further

$$\begin{aligned} T_{j,n}(\boldsymbol{Y}) &:= T(\mathscr{M}_0^{(j)}, \mathscr{M}_j^{(j)}, n, \boldsymbol{Y}), \qquad \text{(LRS)} \\ c_{j,\alpha/|A_n|} &:= \chi^2_{p_j^{af},\alpha/|A_n|}, \end{aligned}$$

where $\chi^2_{p_j^{af},\alpha/|A_n|}$ is the upper $\frac{\alpha}{|A_n|}-$quantile of the $\chi^2_{p_j^{af}}$ distribution, it is shown that

$$\mathbb{P}\left(\exists\, j \in M_n \,:\, T_{j,n} > c_{j,\alpha/|A_n|}\right) \leq \alpha + o(1). \tag{4.21}$$

Here, $T_{j,n}$ denotes $T_{j,n}(\boldsymbol{Y})$ for brevity. The left hand side of the inequality (4.21) is the probability that at least one $j \in M_n$, hence estimated as influential in step 1 (of Algorithm 4.1.1) by L_0-FGL but truly non-influential, is (again) falsely evaluated as influential after step 2. It holds that, using $\mathbb{P}(A^* \not\subseteq A_n) = o(1)$ and $\mathbb{P}(F^* \neq F_n) = o(1)$ from GlobalReg (Definition 4.1.7), by applying the law of total probability twice

$$\begin{aligned} & \mathbb{P}\left(\exists\, j \in M_n \,:\, T_{j,n} > c_{j,\alpha/|A_n|}\right) && (4.22) \\ = \;& \mathbb{P}\left(\exists\, j \in M_n \,:\, T_{j,n} > c_{j,\alpha/|A_n|}, A^* \subseteq A_n\right) + \mathbb{P}\left(\exists\, j \in M_n \,:\, T_{j,n} > c_{j,\alpha/|A_n|}, A^* \not\subseteq A_n\right) \\ = \;& \mathbb{P}\left(\exists\, j \in M_n \,:\, T_{j,n} > c_{j,\alpha/|A_n|}, A^* \subseteq A_n, F^* = F_n\right) && (4.23) \\ & + \mathbb{P}\left(\exists\, j \in M_n \,:\, T_{j,n} > c_{j,\alpha/|A_n|}, A^* \subseteq A_n, F^* \neq F_n\right) \\ & + \mathbb{P}\left(\exists\, j \in M_n \,:\, T_{j,n} > c_{j,\alpha/|A_n|}, A^* \not\subseteq A_n, F^* = F_n\right) \\ & + \mathbb{P}\left(\exists\, j \in M_n \,:\, T_{j,n} > c_{j,\alpha/|A_n|}, A^* \not\subseteq A_n, F^* \neq F_n\right) \\ \leq \;& \mathbb{P}\left(\exists\, j \in M_n \,:\, T_{j,n} > c_{j,\alpha/|A_n|}, A^* \subseteq A_n, F^* = F_n\right) + \mathbb{P}\left(A^* \subseteq A_n, F^* \neq F_n\right) \\ & + \mathbb{P}\left(A^* \not\subseteq A_n, F^* = F_n\right) + \mathbb{P}\left(A^* \not\subseteq A_n, F^* \neq F_n\right) \\ \leq \;& \mathbb{P}\left(\exists\, j \in M_n \,:\, T_{j,n} > c_{j,\alpha/|A_n|}, A^* \subseteq A_n, F^* = F_n\right) + \mathbb{P}\left(F^* \neq F_n\right) + \mathbb{P}\left(A^* \not\subseteq A_n\right) \\ & + \mathbb{P}\left(A^* \not\subseteq A_n\right) \\ = \;& \mathbb{P}\left(\exists\, j \in M_n \,:\, T_{j,n} > c_{j,\alpha/|A_n|}, A^* \subseteq A_n, F^* = F_n\right) + o(1). \end{aligned}$$

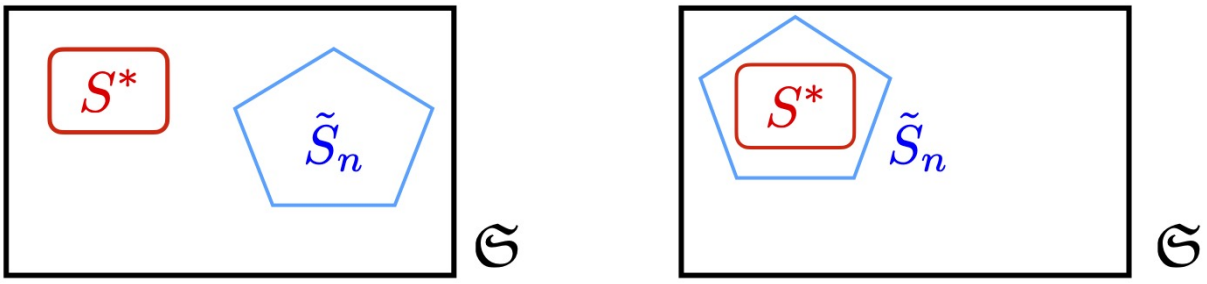

Figure 4.2: *Here, $\mathfrak{S}$ is the set containing all possible models. On the left hand side, one can see the case where the true model S^* is not nested in the selected model $\tilde{S}_n$, where the truth cannot be an interior point of the parameter space Ω_2, corresponding to selected submodel $\tilde{S}_n$. In contrast, on the right hand side, where the true model S^* is nested in $\tilde{S}_n$, one can ensure that the truth is an interior point of the parameter space Ω_2. It is noted that (A1') is needed for screening concerning factor selection and (A3') for screening concerning levels fusion.*

To sum up, one finds that

$$\begin{aligned} &\mathbb{P}\left(\exists\, j \in M_n \,:\, T_{j,n} > c_{j,\alpha/|A_n|}\right) \\ &= \mathbb{P}\left(\exists\, j \in M_n \,:\, T_{j,n} > c_{j,\alpha/|A_n|}, A^* \subseteq A_n, F^* = F_n\right) + o(1). \end{aligned} \tag{4.24}$$

At this point it has to be specified how the regularity conditions, especially the screening properties (A1') and (A3'), as well as the sparsity assumption (A2'), are employed in this proof. Assumptions (Reg1)-(Reg3) are necessary after step 2 (of Algorithm 4.1.1) to ensure that the statistic $T_{j,n}$ is asymptotically $\chi^2_{p_j^{af}}$ distributed. These conditions are given by GlobalReg (Definition 4.1.7), however, it has to be commented on whether it is reasonable to ensure (Reg3) ((div.Reg3), respectively) for Ω_2, in particular that the truth is an interior point of Ω_2, since this depends on the performance of L_0-FGL. The screening properties $A^* \subseteq A_n$ and $F^* = F_n$ (holding asymptotically a.s.) ensure that the truth is an interior point of Ω_2, since it is ensured that the true model is nested in the selected model $\tilde{S}_n$ due to the fact that no influential factors are excluded from the model and the fusions of the truly influential factors are detected correctly. For a visualization of the situation one observes Figure 4.2. The sparsity assumption (A2') is needed to ensure that the design in step 2 is not high-dimensional. Thus, (Reg1)-(Reg3) ((div.Reg1)-(div.Reg3), respectively) are ensured for step 2 and one can use that the asymptotic distribution of the considered test statistic is a $\chi^2_{p_j^{af}}$ distribution, where by (A4') p_j^{af} is well-defined for increasing n.

The probability $\mathbb{P}\left(\exists\, j \in M_n \,:\, T_{j,n} > c_{j,\alpha/|A_n|}, A^* \subseteq A_n, F^* = F_n\right)$ is further quantified for $n \to \infty$, which is the first summand of (4.24). In Wasserman and Roeder (2009), the maximum over $j \in M_n$ of $T_{j,n}$ is considered, arguing that if the maximum over all test statistics $T_{j,n}$ exceeds some critical value, there exists at least one $j \in M_n$ for which $T_{j,n}$ exceeds this critical value. However, as shortly mentioned in Remark 4.1.13, this approach is not suitable here, since the critical value $c_{j,\alpha/|A_n|}$ depends on j. To overcome this challenge, a Bonferroni inequality is applied below using the asymptotic distribution of the test statistic and the imposed regularity conditions.

To be more precise, the following events are defined

$$\begin{aligned} B_{j,n} &:= \left\{T_{j,n} > c_{j,\alpha/|A_n|}, A^* \subseteq A_n, F^* = F_n\right\}, \qquad (j \in M_n) \\ B_n &:= \left\{\exists\, j \in M_n \,:\, T_{j,n} > c_{j,\alpha/|A_n|}, A^* \subseteq A_n, F^* = F_n\right\}. \end{aligned}$$

Clearly, it holds that $B_n \subset \bigcup_{j \in M_n} B_{j,n}$ and, since M_n is countable, using the Bonferroni inequality one gets

$$\mathbb{P}(B_n) \leq \sum_{j \in M_n} \mathbb{P}(B_{j,n}). \tag{4.25}$$

For every $j \in M_n$, one knows that $T_{j,n}$ asymptotically follows a $\chi^2_{p_j^{af}}$ distribution under the null hypothesis, which is left out as lower index in the probability for simplicity of notation. Hence, with $c_{j,\frac{\alpha}{|A_n|}}$ being the upper quantile of this limiting distribution of $T_{j,n}$, it holds that

$$\mathbb{P}(B_{j,n}) = \frac{\alpha}{|A_n|} + o(1). \tag{4.26}$$

By (A4'), it holds that $|A_n| = O(1)$ and by construction $M_n \subseteq A_n$, so the number of summands is bounded, hence

$$\mathbb{P}(B_n) \leq |M_n| \cdot \frac{\alpha}{|A_n|} + o(1) \leq \alpha + o(1) \Rightarrow \lim_{n\to\infty} \mathbb{P}(B_n) \leq \alpha.$$

Here, it is crucial to ensure that the term $o(1)$ in (4.26), coming from the asymptotic distribution of the test statistic, remains controlled going to zero even though the sum (4.25) depending on n is considered, which is ensured by (A4') as explained above. Now, by definition of B_n one can deduce

$$\lim_{n\to\infty} \mathbb{P}(B_n) = \lim_{n\to\infty} \mathbb{P}\left(\underbrace{\exists\, j \in M_n \,:\, T_{j,n} > c_{j,\alpha/|A_n|}, A^* \subseteq A_n, F^* = F_n}_{=B_n} \right) \leq \alpha. \tag{4.27}$$

Consequently, combining (4.24) and (4.27) yields

$$\lim_{n\to\infty} \mathbb{P}\left(\exists\, j \in M_n \,:\, T_{j,n} > c_{j,\alpha/|A_n|} \right) \leq \alpha + o(1), \tag{4.28}$$

so the claim follows. □

For the subsequent thesis, the notion of *consistency* (of the test) is required, for which the following definition is supplied.

Definition 4.1.15. In the considered setting, let for all factors $j \in A^*$ exist a sequence of significance levels $(\alpha_n)_{n\in\mathbb{N}} \subseteq (0,1)$ at a rate such that

$$\lim_{n\to\infty} \rho_{n,j} = 0 \;\; \forall j \in A^*, \tag{4.29}$$

where

$$\rho_{n,j} := \mathbb{P}\left(T_{j,n} < c_{j,\frac{\alpha_n}{k}}, A^* \subseteq A_n, F^* = \tilde{F}_n \right) \tag{4.30}$$

and k given by (A4'). Then, the executed test is said to be *consistent*

Hence, according to the definition provided above, consistency of the test means that for truly influential factors $j \in A^*$, the probability of *not* rejecting $H_{0,j}$, i.e. the probability of a false rejection, tends to zero as $n \to \infty$, where the corresponding significance level is allowed to go to zero. Probably, to ensure this consistency, one needs to control the rate of $\alpha_n \to 0$, to ensure that this convergence is not too fast compared to the increase of the sample size. However,

this is not further investigated in this thesis, i.e., being in line with Meinshausen et al. (2009) (Section 3.4), it is assumed that such a rate exists.

The next theorem shows factor selection consistency using the Bonferroni correction. Hence, using the two-stage L_0-FGL with Bonferroni correction it is achieved that, asymptotically, the *true noise* factors are selected as noise and the *true influential* factors as influential. The following theorem holds for both cases, i.e. fixed p and diverging p_n, as long as GlobalReg (Definition 4.1.7) and the consistency of the test (Definition 4.1.15) are ensured.

Theorem 4.1.16 (Consistency of two-stage L_0-FGL in single split using Bonferroni correction, fixed p and diverging p_n)**.** One assumes that GlobalReg (Definition 4.1.7) hold and that the LRT is executed at level α_n, where the correction for multiplicity is done by Bonferroni correction and for the sequence $(\alpha_n)_{n\in\mathbb{N}} \subseteq (0,1)$ it holds $\alpha_n \to 0$ for $n \to \infty$, such that consistency in the sense of Definition 4.1.15 is ensured. Then, it holds that the two-stage L_0-FGL in the single split is factor selection consistent, i.e.

$$\lim_{n\to\infty} \mathbb{P}\left(A^* = A_{n,f}(\alpha_n)\right) = 1. \tag{4.31}$$

Proof. The idea of this proof roots in Wasserman and Roeder (2009), transferred to the case of two-stage L_0-FGL (cf. Remark 4.1.13). By Theorem 4.1.14, using $\alpha_n \to 0$ for $n \to \infty$, it is known that

$$\lim_{n\to\infty} \mathbb{P}\left((A^*)^c \cap A_{n,f}(\alpha_n) \neq \emptyset\right) = 0,$$

which directly yields

$$\lim_{n\to\infty} \mathbb{P}\left(A_{n,f}(\alpha_n) \subseteq A^*\right) = 1.$$

Now, it remains to show that $\lim_{n\to\infty} \mathbb{P}\left(A^* \subseteq A_{n,f}(\alpha_n)\right) = 1$.

For some $j \in A^*$, it is shown that

$$\lim_{n\to\infty} \mathbb{P}\left(j \notin A_{n,f}(\alpha_n)\right) \overset{!}{=} 0.$$

By the law of total probability (applied twice) and the regularity conditions of GlobalReg (Definition 4.1.7), it holds that

$$\begin{aligned}
\mathbb{P}\left(j \notin A_{n,f}(\alpha_n)\right) &= \mathbb{P}\left(j \notin A_{n,f}(\alpha_n), A^* \subseteq A_n\right) + \mathbb{P}\left(j \notin A_{n,f}(\alpha_n), A^* \not\subseteq A_n\right) \\
&= \mathbb{P}\left(j \notin A_{n,f}(\alpha_n), A^* \subseteq A_n, F^* = F_n\right) \\
&\quad + \mathbb{P}\left(j \notin A_{n,f}(\alpha_n), A^* \subseteq A_n, F^* \neq F_n\right) \\
&\quad + \mathbb{P}\left(j \notin A_{n,f}(\alpha_n), A^* \not\subseteq A_n, F^* = F_n\right) \\
&\quad + \mathbb{P}\left(j \notin A_{n,f}(\alpha_n), A^* \not\subseteq A_n, F^* \neq F_n\right) \\
&\leq \mathbb{P}\left(j \notin A_{n,f}(\alpha_n), A^* \subseteq A_n, F^* = F_n\right) + 2\underbrace{\mathbb{P}\left(F^* \neq \tilde{F}_n\right)}_{=o(1)} + \underbrace{\mathbb{P}\left(A^* \not\subseteq A_n\right)}_{=o(1)}.
\end{aligned}$$

The first summand on the right hand side of the inequality is the probability of not rejecting $H_{0,j}$ for $j \in A^*$ being truly influential and can be re-written as

$$\mathbb{P}\left(j \notin A_{n,f}(\alpha_n), A^* \subseteq A_n, F^* = F_n\right) = \mathbb{P}\left(T_{j,n} < c_{j,\frac{\alpha_n}{|A_n|}}, A^* \subseteq A_n, F^* = F_n\right). \tag{4.32}$$

Quantifying this probability (4.32) further requires steps different from Wasserman and Roeder (2009) since a LRT is applied in this thesis (cf. Remark 4.1.13). Consequently, the proof here is continued by utilizing the consistency of the LRT rather than the concrete distribution of the test statistic as done in the mentioned reference. The latter would not be suitable here as the distribution of the LRS is just known asymptotically.

Thus, (4.32) is analyzed with the help of $\rho_{n,j}$ defined in (4.30) and for which one knows that (4.29) holds. However, one needs to get rid of the dependence on $|A_n|$ in the quantile on the right hand side of (4.32). The requirements $A^* \subseteq A_n, F^* = F_n$ ensure the theoretical properties of the LRT and they hold asymptotically a.s. by the GlobalReg (Definition 4.1.7) assumption, one further regards the explanations in the proof of Theorem 4.1.14.

For $j \in A^*$ with $c_{j,\frac{\alpha_n}{|A_n|}} = \chi^2_{p_j^{af},\frac{\alpha_n}{|A_n|}}$ and $c_{j,\frac{\alpha_n}{k}} = \chi^2_{p_j^{af},\frac{\alpha_n}{k}}$ (k comes from (A4')), one deduces

$$\begin{aligned}\frac{\alpha_n}{|A_n|} > \frac{\alpha_n}{k} &\Rightarrow c_{j,\frac{\alpha_n}{|A_n|}} < c_{j,\frac{\alpha_n}{k}} \\ &\Rightarrow \mathbb{P}\left(T_{j,n} > c_{j,\frac{\alpha_n}{k}} A^* \subseteq A_n, F^* = F_n\right) \leq \mathbb{P}\left(T_{j,n} > c_{j,\frac{\alpha_n}{|A_n|}} A^* \subseteq A_n, F^* = F_n\right). \quad (4.33)\end{aligned}$$

Thus, by (4.29) it holds

$$\lim_{n\to\infty} \mathbb{P}\left(T_{j,n} > c_{j,\frac{\alpha_n}{k}} A^* \subseteq A_n, F^* = F_n\right) = \lim_{n\to\infty} 1 - \rho_{n,j} = 1.$$

By (4.33) one can infer

$$\lim_{n\to\infty} \mathbb{P}\left(T_{j,n} > c_{j,\frac{\alpha_n}{|A_n|}}, A^* \subseteq A_n, F^* = F_n\right) = 1.$$

Thus, for $j \in A^*$ this yields

$$\lim_{n\to\infty} \mathbb{P}\left(T_{j,n} < c_{j,\frac{\alpha_n}{|A_n|}}, A^* \subseteq A_n, F^* = F_n\right) = 0. \quad (4.34)$$

To sum up, it was shown that for each $j \in A^*$ that

$$\lim_{n\to\infty} \mathbb{P}\left(j \notin A_{n,f}(\alpha_n)\right) = 0,$$

i.e.

$$\lim_{n\to\infty} \mathbb{P}\left(A^* \subseteq A_{n,f}(\alpha_n)\right) = 1$$

and the claim follows for both the fixed p and the diverging p_n case. □

4.1.8 FDR Control with Benjamini Yekutieli Correction

The main objective of the following examination is Theorem 4.1.18 and its proof, showing false discovery rate (FDR) control of the two-stage L_0-FGL procedure using the Benjamini Yekutieli (BY) correction for multiplicity of testing.

Remark 4.1.17. At first, a comparison of two-stage L_0-FGL with the main references where this idea emerged from is provided in the following. For the concrete specification of the presented aspects, the corresponding upcoming theorems and proofs are linked. Even though single sample splitting is considered at the moment, this remark is further valid for the multiple sample splitting applied for two-stage L_0-FGL introduced later in this chapter (Section 4.2), which is mentioned here to avoid repetitions later.

(a) Comparison to Meinshausen et al. (2009)

(i) In their work, the linear model with LSE is utilized rather than logistic regression with MLE as done in this thesis. Consequently, similar challenges arise as described in Remark 4.1.13 (i), while the difference is that considering the BY adjustment, the focus lies on p-values rather than on the test statistic. However, only the *asymptotic* distribution of the the p-values (and the test statistic) is available such that probabilities including these quantities can only be determined asymptotically. Further, considering e.g. the expectation of a quantity incorporating the p-value (as done in the proof of Theorem 4.2.4), it is necessary to utilize the weak convergence of the probability measures that come with the fact that the considered p-value is asymptotically uniform distributed. Justifications for these challenges are given in the respective theorems and proofs (single split: proof of Theorem 4.1.18, multiple split: proofs of Theorems 4.2.3, 4.2.4 and 4.2.6).

(ii) A screening property concerning levels fusion is required rather than only for selection, which is already explained in detail in Remark 4.1.13 (ii) and is analogously valid for this comparison.

(iii) In their work, the hypotheses are of the form $\beta_j = 0$ whereas in this investigation subvectors $\boldsymbol{\beta}_j = \mathbf{0}$ are tested because of the group structure of categorical covariates, which results in the fact that the distribution of the test statistic depends on the dimension of $\boldsymbol{\beta}_j$, i.e. on the L_0-FGL regularization of step 1 (Algorithm 4.1.1). Reference is made to Remark 4.1.13 (iii), where this aspect is discussed.

(b) Comparison to Benjamini and Yekutieli (2001)
They neither observe a two step procedure, nor particular tests, in their work the focus rather lies (only) on the correction for multiple testing with FDR control based on p-values. Thus, this approach has a substantially different purpose, however, it serves as a valuable tool for two-stage L_0-FGL. For that reason, some of their results concerning properties of the BY procedure are picked up in the following examination as they are of high importance in several proofs.

For the explanation of the Benjamini Yekutieli correction it is referred to Section 4.1.6. The definition of the quantities j_1 and j_2 given in (4.5) and (4.6) is recalled, where j_1 denoted the number of true null hypotheses (truth $\boldsymbol{\beta}_j^* = 0$ is a noise variable), thus $j_1 = |(A^*)^c|$. Further, $j_2 = J - j_1$ denoted the number of truly false null hypotheses, thus $j_2 = |A^*|$. For the single split case, one considers these quantities with the intersection of A_n (Remark 4.1.11), i.e. one defines

$$\begin{aligned}
&\tilde{J} := |A_n| && \text{(total number of hypotheses in single split)},\\
&\tilde{j}_1 := |(A^*)^c \cap A_n| && \text{(number of true null hypotheses in single split)},\\
&\tilde{j}_2 := |A^* \cap A_n| && \text{(number of truly false null hypotheses in single split)}.
\end{aligned}$$

In addition, the following quantities are defined

$$V^{(BY)}(q) := |A_{n,f}^{(BY)}(q) \cap (A^*)^c|, \qquad \text{(number of false rejections)} \tag{4.35}$$

$$R^{(BY)}(q) := |A_{n,f}^{(BY)}(q)|, \qquad \text{(number of total rejections)} \tag{4.36}$$

recalling that $A_{n,f}^{(BY)}(q)$ is given by (4.20) and denotes the final set of factors included in the model, applying BY correction. Consequently, $V^{(BY)}(q)$ is the number of false rejections of the

null hypotheses, whereas $R^{(BY)}(q)$ is the number of total rejections of the null hypotheses. In addition, one sets

$$W(q) := \frac{V^{(BY)}(q)}{\max\{1, R^{(BY)}(q)\}}, \tag{4.37}$$

which is the ratio of false rejections and total rejections. Clearly, the maximum in the denominator is taken to avoid the division by zero for the case that none of the hypotheses is rejected. The expectation of the ratio (4.37), i.e. $\mathbb{E}(W(q))$ is called *false discovery rate* (FDR).

Theorem 4.1.18 (FDR control for two-stage L_0-FGL in single split with Benjamini Yekutieli, fixed p and diverging p_n)**.** It is assumed that GlobalReg (Definition 4.1.7) hold. Further, $\tilde{q} > 0$ and $A^{(BY)}_{n,f}(q)$ is defined as in (4.20) with $q := \tilde{q}\left(\sum_{i=1}^{\tilde{J}} \frac{1}{i}\right)^{-1}$, thus the LRT of two-stage L_0-FGL is executed at level q, where the correction for multiplicity is done by BY. Then it holds that, asymptotically, the two-stage L_0-FGL in the single split controls the FDR by $\tilde{q}$, in particular

$$\limsup_{n\to\infty} \mathbb{E}(W(q)) \leq \tilde{q}. \tag{4.38}$$

The proof of Theorem 4.1.18 requires the following lemmata on properties of the BY procedure (Lemmata 4.1.19 and 4.1.20), which can similarly be shown by replacing j_1, j_2 and J by $\tilde{j}_1, \tilde{j}_2$ and $\tilde{J}$ (cf. Remark 4.1.11). However, since they are also applied in the multiple split, the proof is conducted for the more general case taking j_1, j_2 and J. Finally, it is noted that these two lemmata are not asymptotic properties, i.e. p is obviously fixed. The following lemma and its proof can be found in Benjamini and Yekutieli (2001) (Lemma 4.1).

Lemma 4.1.19 (Benjamini and Yekutieli (2001), Lemma 4.1)**.** j_1 is the number of true null hypotheses among the hypotheses being tested based on the p-values $P_{i,n}$, $i \in \{1, \ldots, j_1 + j_2\}$, while j_2 is the number of false null hypotheses. By defining the event $E_{v,s}(q)$ that the Benjamini Yekutieli (BY) procedure rejects exactly v true null hypotheses (falsely) and s false null hypotheses (correctly) for given $q > 0$, it holds that

$$\mathbb{P}(E_{v,s}(q)) = \frac{1}{v}\sum_{i=1}^{j_1} \mathbb{P}(\{P_{i,n} \leq q_{v+s}\} \cap E_{v,s}(q)), \tag{4.39}$$

where

$$q_{v+s} := q \cdot (v + s). \tag{4.40}$$

This also holds when replacing $P_{i,n}$ by aggregated p-values $Q_i(\gamma)$ and Q_i^{adap} in the multiple split rigorously introduced later in this chapter but mentioned at this point for the sake of completeness.

Proof. This lemma is shown for "general" p-values $P_{i,n}$, i.e. no particular form or structure of the p-values is used, it is only used that the test decision rejecting the null hypothesis $H_{0,i}$ or not is based on the p-values $P_{i,n}$. Therefore, the proof is also valid replacing $P_{i,n}$ by $Q_i(\gamma)$ or Q_i^{adap} in the multiple split.

By definition of the event $E_{v,s}(q)$, exactly v null hypotheses are falsely rejected. Since one recognizes that j_1 of the null hypotheses are truly correct (i.e. those should not be rejected), there are different possibilities for these v wrong rejections to occur among the truly correct null

hypotheses. Hence, $w \subseteq \{1, \dots, j_1\}$ is a subset of the truly correct null hypotheses where $|w| = v$ for some $v \in \mathbb{N}$, $v \leq j_1$. Thus, w is the set of the (indices of) the true null hypotheses that are falsely rejected. Now, one defines the event $E_{v,s}^w(q)$ as the event in $E_{v,s}(q)$ that the v falsely rejected null hypotheses are exactly those null hypotheses in $w \subseteq \{1, \dots, j_1\}$. Consequently, $E_{v,s}(q)$ is the (general) event that exactly v true null and s wrong null hypothesis are rejected while the event $E_{v,s}^w(q)$ specifies the particular hypotheses that are falsely rejected. With these definitions one can write by the law of total probability

$$\mathbb{P}(E_{v,s}(q)) = \sum_{w \subseteq \{1,\dots,j_1\}, |w|=v} \mathbb{P}(E_{v,s}^w(q)) \tag{4.41}$$

and further

$$\mathbb{P}\left(\{P_{i,n} \leq q_{v+s}\} \cap E_{v,s}(q)\right) = \sum_{w \subseteq \{1,\dots,j_1\}, |w|=v} \mathbb{P}(\{P_{i,n} \leq q_{v+s}\} \cap E_{v,s}^w(q)). \tag{4.42}$$

For $i \in w$ one knows that $P_{i,n} \leq q \cdot (v+s) = q_{v+s}$ since $v+s$ null hypotheses are rejected and hypothesis i as well, since $i \in w$ (cf. (4.19)). For $i \in \{1, \dots, j_1\} \backslash w$ it holds that $P_{i,n} > q_{v+s}$ since this hypothesis is not rejected by construction. This yields

$$\mathbb{P}(\{P_{i,n} \leq q_{v+s}\} \cap E_{v,s}^w(q)) = \begin{cases} \mathbb{P}(E_{v,s}^w(q)), & \text{if } i \in w \\ 0, & \text{if } i \in \{1, \dots, j_1\} \backslash w. \end{cases} \tag{4.43}$$

Combining all steps from above one can deduce

$$\begin{aligned}
\sum_{i=1}^{j_1} \mathbb{P}(\{P_{i,n} \leq q_{v+s}\} \cap E_{v,s}(q)) &\overset{(4.42)}{=} \sum_{i=1}^{j_1} \sum_{w \subseteq \{1,\dots,j_1\},\, |w|=v} \mathbb{P}(\{P_{i,n} \leq q_{v+s}\} \cap E_{v,s}^w(q)) \\
&= \sum_{w \subseteq \{1,\dots,j_1\},\, |w|=v} \sum_{i=1}^{j_1} \mathbb{P}(\{P_{i,n} \leq q_{v+s}\} \cap E_{v,s}^w(q)) \\
&\overset{(4.43)}{=} \sum_{w \subseteq \{1,\dots,j_1\},\, |w|=v} \sum_{i=1}^{j_1} \mathbf{1}_{\{i \in w\}} \mathbb{P}(E_{v,s}^w(q)) \\
&\overset{(*)}{=} \sum_{w \subseteq \{1,\dots,j_1\},\, |w|=v} v \cdot \mathbb{P}(E_{v,s}^w(q)) \\
&\overset{(4.41)}{=} v \cdot \mathbb{P}(E_{v,s}(q)),
\end{aligned}$$

where in $(*)$ above it is used that $\sum_{i=1}^{j_1} \mathbf{1}_{\{i \in w\}} \mathbb{P}(E_{v,s}^w(q)) = v \cdot \mathbb{P}(E_{v,s}^w(q))$ since by construction $w \subseteq \{1, \dots, j_1\}$ and $|w| = v$. Consequently, the claim follows. □

The following lemma and its proof can be found in Section 4 of Benjamini and Yekutieli (2001). Similarly, the statement is picked up in Meinshausen et al. (2009) in their proof of Theorem 3.3.

Lemma 4.1.20 (Benjamini and Yekutieli (2001) Section 4, expressions (9), (10) and (28))**.** One assumes the same setting as in Lemma 4.1.19. $C_{v,s}^{(i)}(q)$ is the event in which *if* hypothesis $H_{0,(i)}$ is rejected, *then* $v-1$ true null hypotheses are (falsely) rejected and s false null hypotheses are (correctly) rejected alongside with it. J is the number of total hypotheses. One defines

$$C_k^{(i)}(q) := \bigcup_{t=1}^{k} \{C_{v,s}^{(i)}(q) \,|\, v + s = t\}, \tag{4.44}$$

$$p_{ijk}(q) := \mathbb{P}(\{P_{i,n} \in [(j-1)q, jq]\} \cap C_k^{(i)}(q)). \tag{4.45}$$

Then, the FDR can be written in the following way

$$\mathbb{E}(W(q)) = \sum_{i\in(A^*)^c} \sum_{k=1}^{J} \frac{1}{k} \sum_{j=1}^{k} p_{ijk}(q). \tag{4.46}$$

This also holds when replacing $P_{i,n}$ by aggregated p-values $Q_i(\gamma)$ and Q_i^{adap} in the multiple split rigorously introduced later in this chapter but mentioned at this point for the sake of completeness.

Before proving Lemma 4.1.20, the following remark is supplied.

Remark 4.1.21. The outer sum in (4.46) is taken over all true null hypotheses, which can be adjusted if just a subset as in the single split is taken, i.e. replacing $i \in (A^*)^c$ by $i \in (A^*)^c \cap A_n$ (cf. Remark 4.1.11). Further, the inner sum is taken over all hypotheses being tested (which is J here) and this sum can also be adjusted by taking $\tilde{J} = |A_n| \leq J$ instead of J restricting the tested hypotheses to the selected factors of step 1 (of Algorithm 4.1.1) in the single split. Consequently, the lemma above and its proof is valid in the same way for the explained adjustments, replacing in the proof j_1, j_2 and J by $\tilde{j}_1, \tilde{j}_2$ and $\tilde{J}$, respectively.

Proof. (of Lemma 4.1.20) First, the fact that this Lemma also holds when replacing $P_{i,n}$ by the aggregated p-values $Q_i(\gamma)$ or Q_i^{adap} is explained at the beginning of the proof of Lemma 4.1.19 and this applies in the same way to this proof.
Since $|(A^*)^c| = j_1$, hence one has j_1 true null hypotheses, the number of false rejections v satisfies $v \leq j_1$, and, without loss of generality, one assumes that $(A^*)^c = \{1, \ldots, j_1\}$ for simplicity of notation. The number of true rejections, which is denoted by s, satisfies $s \leq j_2$ with $j_2 = |A^*|$. It is noted that for the number of false rejections one does not include zero to avoid that the denominator is zero, one recalls the definition of $W(q)$. The number of total rejections is then $v + s$. Now, with the event $E_{v,s}(q)$ defined in Lemma 4.1.19, the FDR, i.e. $\mathbb{E}(W(q))$, can be expressed more precisely using the definition of the expectation and the explanations above

$$\begin{aligned}
\mathbb{E}(W(q)) &= \sum_{s=0}^{j_2} \sum_{v=1}^{j_1} \frac{v}{v+s} \mathbb{P}(E_{v,s}(q)) && (4.47)\\
&= \sum_{s=0}^{j_2} \sum_{v=1}^{j_1} \frac{v}{v+s} \left(\frac{1}{v} \sum_{i=1}^{j_1} \mathbb{P}(\{P_{i,n} \leq q_{v+s}\} \cap E_{v,s}(q)) \right) \quad \text{(using Lemma 4.1.19)} \\
&= \sum_{i=1}^{j_1} \left(\sum_{s=0}^{j_2} \sum_{v=1}^{j_1} \frac{1}{v+s} \mathbb{P}(\{P_{i,n} \leq q_{v+s}\} \cap E_{v,s}(q)) \right). && (4.48)
\end{aligned}$$

The next goal is to get rid of the direct dependency on v and s by reconstructing $E_{v,s}(q)$ such that it only depends on i and the total number of rejections $k = v + s$ (and the chosen level q).

For all $i \in \{1, \ldots, j_1\} = (A^*)^c$ the corresponding hypotheses are truly correct. Hence, if $P_{i,n} \leq q_{v+s}$ for $i \in \{1, \ldots, j_1\}$ one has a false rejection. With that, for $i \in \{1, \ldots, j_1\}$, one deduces that $\{P_{i,n} \leq q_{v+s}\} \cap C_{v,s}^{(i)}(q)$ is the event that hypothesis i is falsely rejected and, given this rejection, one has $v - 1$ other false rejections and s correct rejections because of the intersection with $C_{v,s}^{(i)}(q)$. Hence in total there are $v - 1 + 1 = v$ false rejections and s correct rejections. Consequently, by definition, this is the same event as $\{P_{i,n} \leq q_{v+s}\} \cap E_{v,s}(q)$ for $i \in \{1, \ldots, j_1\}$ where also the true null hypothesis i is rejected (since $P_{i,n} \leq q_{v+s}$) and one has v false and s correct rejections in total. To sum up, it holds that

$$\{P_{i,n} \leq q_{v+s}\} \cap E_{v,s}(q) = \{P_{i,n} \leq q_{v+s}\} \cap C_{v,s}^{(i)}(q) \ , \ i \in \{1, \ldots, j_1\}. \tag{4.49}$$

Recalling the definition of $C_k^{(i)}(q)$, that is

$$C_k^{(i)}(q) := \bigcup_{t=1}^{k} \{C_{v,s}^{(i)}(q) \,|\, v+s=t\},$$

one can see that $C_k^{(i)}(q)$ for $i \in \{1,\dots,j_1\}$ is the event that *if* one (falsely) rejects hypothesis i, one further rejects another $t \in \{0,\dots,k\}$ hypotheses where for the latter it is not specified which of them are correct and which of them are wrong rejections. By construction, it holds that for each $i \in \{1,\dots,j_1\}$, the $C_k^{(i)}(q)$ are disjoint. As a consequence, replacing $v+s$ by k and using the definition of $C_k^{(i)}(q)$, one ends up with

$$\sum_{s=0}^{j_2}\sum_{v=1}^{j_1} \frac{1}{v+s}\mathbb{P}(\{P_{i,n} \leq q_{v+s}\} \cap C_{v,s}^{(i)}(q)) = \frac{1}{k}\sum_{k=1}^{J} \mathbb{P}(\{P_{i,n} \leq q_k\} \cap C_k^{(i)}(q)). \tag{4.50}$$

Combining the equation (4.50) above with (4.48) and (4.49)

$$\mathbb{E}(W(q)) \quad = \quad \sum_{i=1}^{j_1}\sum_{k=1}^{J} \frac{1}{k}\mathbb{P}(\{P_{i,n} \leq q_k\} \cap C_k^{(i)}(q)). \tag{4.51}$$

With $q_k = q \cdot k$ by definition, one can write

$$\mathbb{P}(\{P_{i,n} \leq q_k\} \cap C_k^{(i)}(q)) = \sum_{j=1}^{k} \mathbb{P}(P_{i,n} \in [(j-1)q, jq] \cap C_k^{(i)}(q)),$$

which, in combination with (4.51) and $j_1 = |(A^*)^c|$, provides

$$\begin{aligned}
\mathbb{E}(W(q)) \quad &= \quad \sum_{i=1}^{j_1}\sum_{k=1}^{J} \frac{1}{k}\sum_{j=1}^{k} \mathbb{P}(P_{i,n} \in [(j-1)q, jq] \cap C_k^{(i)}(q)) \\
&= \quad \sum_{i=1}^{j_1}\sum_{k=1}^{J} \frac{1}{k}\sum_{j=1}^{k} p_{ijk}(q) \\
&= \quad \sum_{i \in (A^*)^c}\sum_{k=1}^{J} \frac{1}{k}\sum_{j=1}^{k} p_{ijk}(q)
\end{aligned}$$

so the claim follows. □

Having all necessary tools available, the proof of Theorem 4.1.18 can be executed.

Proof. (of Theorem 4.1.18) The proof roots in Meinshausen et al. (2009) (Theorem 3.1) and Benjamini and Yekutieli (2001) adapted for two-stage L_0-FGL (cf. Remark 4.1.17). At first, adjusting the approach of Meinshausen et al. (2009) with respect to the additional screening property concerning fusion (A3') proven in Chapter 2, the following quantities are defined

$$\begin{aligned}
\kappa_{j,n} := \quad & P_{j,n} \cdot \mathbf{1}\{A^* \subseteq A_n, F^* = F_n\} + \mathbf{1}\{A^* \not\subseteq A_n, F^* = F_n\} \\
& + \mathbf{1}\{A^* \not\subseteq A_n, F^* \neq F_n\} + \mathbf{1}\{A^* \not\subseteq A_n, F^* = F_n\},
\end{aligned}$$

where $\mathbf{1}\{\}$ are indicator functions being zero or one. Consequently, $\kappa_{j,n}$ is given by the adjusted p-value $P_{j,n}$ if $A^* \subseteq A_n$ and $F^* = F_n$, and by one otherwise. Since by GlobalReg (Definition 4.1.7) $\lim_{n\to\infty} \mathbb{P}(A^* \not\subseteq A_n) = 0$ and $\lim_{n\to\infty} \mathbb{P}(F^* \neq F_n) = 0$, it holds that $\kappa_{j,n} = P_{j,n}$ almost surely for

$n \to \infty$. This step is needed to ensure the regularity conditions for the convergence properties of MLE and LRS, one refers to the explanations provided in the proof of Theorem 4.1.14. However, to avoid confusion and keep the notation simple, one continues writing $P_{j,n}$ instead of $\kappa_{j,n}$ in the following and omits the conditions $A^* \subseteq A_n$, $F^* = F_n$ in the probabilities. This is completely analogous to the steps below the lines starting from (4.22).
Further, without loss of generality, one omits the $\min(1, \cdot)$ function in the definition of $P_{j,n}$ and one assumes that $P_{j,n} \leq 1$. This does clearly not affect the selection process of the test.

To prove the claim, one needs to show that

$$\limsup_{n\to\infty} \mathbb{E}(W(q)) \leq \tilde{q}.$$

Using Lemma 4.1.20 the expression $\mathbb{E}(W(q))$ can be re-written as follows, noting that the single split is considered hence the intersection with the selected set A_n is taken, i.e. $\tilde{j}_1 = |(A^*)^c \cap A_n|$ are the true null hypothesis instead of j_1, analogously $\tilde{j}_2 = |A^* \cap A_n|$ is used instead of $j_2 = |A^*|$ and $\tilde{J} = |A_n|$, respectively, whereby reference is made to Remarks 4.1.11 and 4.1.21.

$$\begin{aligned}
\mathbb{E}(W(q)) &= \sum_{i\in(A^*)^c\cap A_n} \sum_{k=1}^{\tilde{J}} \frac{1}{k} \sum_{j=1}^{k} p_{ijk}(q) \overset{(*)}{=} \sum_{i\in(A^*)^c\cap A_n} \sum_{j=1}^{\tilde{J}} \sum_{k=j}^{\tilde{J}} \frac{1}{k} p_{ijk}(q) \\
&\leq \sum_{i\in(A^*)^c\cap A_n} \sum_{j=1}^{\tilde{J}} \sum_{k=j}^{\tilde{J}} \frac{1}{j} p_{ijk}(q) = \sum_{i\in(A^*)^c\cap A_n} \sum_{j=1}^{\tilde{J}} \frac{1}{j} \sum_{k=j}^{\tilde{J}} p_{ijk}(q) \\
&\leq \sum_{i\in(A^*)^c\cap A_n} \sum_{j=1}^{\tilde{J}} \frac{1}{j} \sum_{k=1}^{\tilde{J}} p_{ijk}(q) \ \ (\text{since } p_{ijk}(q) \geq 0) \\
&= \sum_{j=1}^{\tilde{J}} \frac{1}{j} \underbrace{\sum_{i\in(A^*)^c\cap A_n} \sum_{k=1}^{\tilde{J}} p_{ijk}(q)}_{=:f(j)} = \sum_{j=1}^{\tilde{J}} \frac{1}{j} f(j),
\end{aligned} \tag{4.52}$$

where $(*)$ above can be seen by explicitly writing down the summands, that is

$$\begin{aligned}
&\sum_{k=1}^{\tilde{J}} \frac{1}{k} \sum_{j=1}^{k} p_{ijk}(q) \\
&= p_{i11}(q) + \frac{1}{2}(p_{i12}(q) + p_{i22}(q)) + \frac{1}{3}(p_{i13}(q) + p_{i23}(q) + p_{i33}(q)) \\
&+ \frac{1}{4}(p_{i14}(q) + p_{i24}(q) + p_{i34}(q) + p_{i44}(q)) + \ldots + \frac{1}{\tilde{J}}(p_{i1\tilde{J}}(q) + \ldots + p_{i\tilde{J}\tilde{J}}(q)) \\
&= p_{i11}(q) + \frac{1}{2}p_{i12}(q) + \frac{1}{3}p_{i13}(q) + \frac{1}{4}p_{i14}(q) + \ldots + \frac{1}{\tilde{J}}p_{i1\tilde{J}}(q) \\
&+ \frac{1}{2}p_{i22}(q) + \frac{1}{3}p_{i23}(q) + \frac{1}{4}p_{i24}(q) + \ldots + \frac{1}{\tilde{J}}p_{i2\tilde{J}}(q) \\
&+ \frac{1}{3}p_{i33}(q) + \frac{1}{4}p_{i34}(q) + \ldots + \frac{1}{\tilde{J}}p_{i3\tilde{J}}(q) \\
&+ \frac{1}{4}p_{i44}(q) + \ldots + \frac{1}{\tilde{J}}p_{i4\tilde{J}}(q) \\
&+ \ldots \\
&+ \frac{1}{\tilde{J}}p_{i\tilde{J}\tilde{J}}(q)
\end{aligned}$$

$$= \sum_{j=1}^{\tilde{J}} \sum_{k=j}^{\tilde{J}} \frac{1}{k} p_{ijk}(q).$$

It is noted that the quantity $f(j)$ defined in (4.52) further depends on other quantities rather than j, e.g. on q. Nevertheless, further indices are omitted for simplicity of notation. The next step is to re-write the sum $\sum_{j=1}^{\tilde{J}} \frac{1}{j} f(j)$ of (4.52) with straightforward steps as

$$\begin{aligned}
&\sum_{j=1}^{\tilde{J}-1} \left(\frac{1}{j} - \frac{1}{j+1} \right) \sum_{j'=1}^{j} f(j') + \frac{1}{\tilde{J}} \sum_{j'=1}^{\tilde{J}} f(j') \\
&= \underbrace{\sum_{j=1}^{\tilde{J}-1} \frac{1}{j} \sum_{j'=1}^{j} f(j')}_{\text{sum 1}} - \underbrace{\sum_{j=1}^{\tilde{J}-1} \frac{1}{j+1} \sum_{j'=1}^{j} f(j')}_{\text{sum 2}} + \underbrace{\frac{1}{\tilde{J}} \sum_{j'=1}^{\tilde{J}} f(j')}_{\text{sum 3}} \\
&= \underbrace{1f(1) + \frac{1}{2}(f(1) + f(2)) + \frac{1}{3}(f(1) + f(2) + f(3)) + \ldots + \frac{1}{\tilde{J}-1}(f(1) + \ldots + f(\tilde{J}-1))}_{\text{sum 1}} \\
&\quad - \underbrace{\left[\frac{1}{2} f(1) + \frac{1}{3}(f(1) + f(2)) + \frac{1}{4}(f(1) + f(2) + f(3)) + \ldots + \frac{1}{\tilde{J}}(f(1) + \ldots + f(\tilde{J}-1)) \right]}_{\text{sum 2}} \\
&\quad + \underbrace{\frac{1}{\tilde{J}}(f(1) + f(2) + \ldots + f(\tilde{J}))}_{\text{sum 3}} \\
&= f(1) + \frac{1}{2} f(2) + \frac{1}{3} f(3) + \ldots + \frac{1}{\tilde{J}} f(\tilde{J}) \\
&= \sum_{j=1}^{\tilde{J}} \frac{1}{j} f(j),
\end{aligned}$$

where all summands, except for the colored ones, cancel out. Having that, in combination with (4.52) one can deduce

$$\mathbb{E}(W(q)) \leq \sum_{j=1}^{\tilde{J}-1} \left(\frac{1}{j} - \frac{1}{j+1} \right) \sum_{j'=1}^{j} f(j') + \frac{1}{\tilde{J}} \sum_{j'=1}^{\tilde{J}} f(j'). \tag{4.53}$$

Since by definition $f(j) = \sum_{i \in (A^*)^c \cap A_n} \sum_{k=1}^{\tilde{J}} p_{ijk}(q)$, the sum $\sum_{k=1}^{\tilde{J}} p_{ijk}(q)$ is further specified as follows

$$\begin{aligned}
\sum_{k=1}^{\tilde{J}} p_{ijk}(q) \quad &= \sum_{k=1}^{\tilde{J}} \mathbb{P}(\{P_{i,n} \in [(j-1)q, jq] \cap C_k^{(i)}(q)\}) \quad \text{(by definition, one compares (4.45))} \\
&= \mathbb{P}\left(\{P_{i,n} \in [(j-1)q, jq]\} \cap \left(\bigcup_{k=1}^{\tilde{J}} C_k^{(i)}(q) \right) \right) \qquad (4.54) \\
&= \mathbb{P}(P_{i,n} \in [(j-1)q, jq]), \qquad (4.55)
\end{aligned}$$

where in (4.54) it was used that $C_k^{(i)}(q)$ are disjoint for $i \in (A^*)^c \cap A_n$ (or $i \in (A^*)^c$, respectively, all true null hypotheses), whereby reference is made to the proof of Lemma 4.1.20. In equation

(4.55) it was used that by construction the union $\bigcup_{k=1}^{\tilde{J}} C_k^{(i)}(q)$ is the event that *if* hypothesis i is rejected (at level q), another $t \in \{0, \dots, \tilde{J}-1\}$ hypotheses are rejected as well, resulting in $k \in \{1, \dots, \tilde{J}\}$ total rejections. Hence, following the arguments of Benjamini and Yekutieli (2001) (right after equation (11)), this union is simply the whole space of events that possibly occur. In consequence, using the definition of $f(j)$ and equation (4.55) one can write

$$f(j) = \sum_{i \in (A^*)^c \cap A_n} \sum_{k=1}^{\tilde{J}} p_{ijk}(q) = \sum_{i \in (A^*)^c \cap A_n} \mathbb{P}(P_{i,n} \in [(j-1)q, jq]). \tag{4.56}$$

For the defined p-values $P_{i,n} = \tilde{P}_{i,n}|A_n|$ (4.17), one deduces for $i \in (A^*)^c \cap A_n$

$$\mathbb{P}(P_{i,n} \leq jq) = \mathbb{P}_{H_{0,i}}(P_{i,n} \leq jq) = \mathbb{P}_{H_{0,i}}(\tilde{P}_{i,n}|A_n| \leq jq) = \mathbb{P}_{H_{0,i}}\left(\tilde{P}_{i,n} \leq \frac{jq}{|A_n|}\right), \tag{4.57}$$

where for $i \in (A^*)^c \cap A_n$ the lower index $H_{0,i}$ was omitted so far for simplicity. Moreover, by the asymptotic uniform distribution (4.14) it holds

$$\mathbb{P}_{H_{0,i}}\left(\tilde{P}_{i,n} \leq \frac{jq}{|A_n|}\right) \leq \min\left(1, \frac{jq}{|A_n|} + o(1)\right) \quad \forall i \in A_n \cap (A^*)^c. \tag{4.58}$$

One notes that here, the fact that the distribution of the p-values is only known asymptotically is crucial (cf. Remark 4.1.17 (i)) as the *sum* of the probability that is bounded in (4.58) is considered right below, where the summation index depends on n, in particular the number of summands. Thus, one needs to ensure that the term $o(1)$ going to zero for $n \to \infty$ keeps controlled even though the term is summed over a quantity depending on n. To be more precise, the bound (4.58) which is independent of $i \in (A^*)^c \cap A_n$ yields (again, under $H_{0,i}$)

$$\begin{aligned} \sum_{i \in (A^*)^c \cap A_n} \mathbb{P}(P_{i,n} \leq jq) &= \sum_{i \in (A^*)^c \cap A_n} \mathbb{P}_{H_{0,i}}(P_{i,n} \leq jq) \\ &\leq \min\left(1, \frac{jq}{|A_n|} + o(1)\right) \cdot |(A^*)^c \cap A_n| \\ &\leq \left(\frac{jq}{|A_n|} + o(1)\right) \cdot |(A^*)^c \cap A_n|. \end{aligned} \tag{4.59}$$

Consequently, it holds that

$$\begin{aligned} \sum_{j'=1}^{j} f(j') &\overset{(4.56)}{=} \sum_{j'=1}^{j} \sum_{i \in (A^*)^c \cap A_n} \mathbb{P}(P_{i,n} \in [(j'-1)q, j'q]) \\ &= \sum_{i \in (A^*)^c \cap A_n} \sum_{j'=1}^{j} \mathbb{P}(P_{i,n} \in [(j'-1)q, j'q]) \\ &= \sum_{i \in (A^*)^c \cap A_n} \mathbb{P}(P_{i,n} \leq jq) \\ &\overset{(4.59)}{\leq} \left(\frac{jq}{|A_n|} + o(1)\right) \cdot |(A^*)^c \cap A_n| \\ &\leq jq + o(1), \end{aligned}$$

where in the last step it was used that $|(A^*)^c \cap A_n|$ is bounded by (A4') (i.e. properties of L_0-FGL), hence the term $o(1)$ remains controlled going to zero. Finally, the result is

$$\begin{aligned}
\mathbb{E}(W(q)) &\leq \sum_{j=1}^{\tilde{J}-1} \left(\frac{1}{j} - \frac{1}{j+1}\right) \underbrace{\sum_{j'=1}^{j} f(j')}_{\leq jq+o(1)} + \frac{1}{\tilde{J}} \underbrace{\sum_{j'=1}^{\tilde{J}} f(j')}_{\leq \tilde{J}q+o(1)} \\
&\leq \left\{ \sum_{j=1}^{\tilde{J}-1} \left(\frac{1}{j} - \frac{1}{j+1}\right) jq \right\} + \frac{1}{\tilde{J}} \tilde{J}q + o(1) \\
&= \left\{ \sum_{j=1}^{\tilde{J}-1} \frac{1}{j+1} q \right\} + q + o(1) = \underbrace{q \sum_{j=1}^{\tilde{J}} \frac{1}{j}}_{=\tilde{q}} + o(1) \\
&= \tilde{q} + o(1),
\end{aligned}$$

which yields

$$\limsup_{n\to\infty} \mathbb{E}(W(q)) \leq \tilde{q},$$

so the claim follows. □

4.2 Extension: Two-Stage L_0-FGL with Multiple Sample Splitting

In this section, the investigations of Section 4.1 for single sample splitting are extended to multiple sample splitting. As done for the single split, the differentiation between fixed p and diverging p_n is only needed when it comes to asymptotic properties, as well as regularity conditions needed for those. Hence, two-stage L_0-FGL for the multiple split is introduced for fixed p, the case of diverging p_n is treated later when asymptotic theory is provided.

The purpose for which multiple sample splitting was proposed in Meinshausen et al. (2009) is to provide a tool that is more stable in terms of reproducibility compared to the single split. More precisely, performing the single split method only once, the result depends on the particular data split, while for multiple sample splitting, the idea is to execute the single sample splitting method $B \in \mathbb{N}$ times for some *fixed* $B > 0$ resulting not in single p-values but in a collection of p-values for each hypothesis. For that reason, the two-stage L_0-FGL for multiple sample splitting is introduced in Section 4.2.1.

4.2.1 The Splitting Procedure (Multiple Sample Splitting)

Let $\Omega_1 \subseteq \mathbb{R}^{p+1}$ be a given parameter space for the coefficient vector $\boldsymbol{\beta}$ and let D be a given dataset of sample size $n \in \mathbb{N}$, as in the single split (Algorithm 4.1.1). The following procedure is considered.

Algorithm 4.2.1 (Two-stage L_0-FGL with multiple sample splitting)**.** For $b = 1, \ldots, B$, the following steps are executed.

(i) The data D is split randomly into two independent parts of equal size, called D_1^b and D_2^b.

(ii) *Step 1*: L_0-FGL regularization method is performed on the initial parameter space $\Omega_1 \subseteq \mathbb{R}^{p+1}$, which yields $\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL}),b} \in \mathbb{R}^{p+1}$ and $\hat{\boldsymbol{\beta}}_{\text{red}}^{(L_0\text{-FGL}),b} \in \mathbb{R}^{1+p^{af,b}}$. Here, $p^{af,b} := \sum_{j=1}^{J} p_j^{af,b}$

and $p_j^{af,b}$ is the number of levels of factor j after fusion in split b. The selected submodel is called $\tilde{S}_n^b$ coming with a selected set of variables $A_n^b \subseteq \{1,\dots,J\}$ and a fusion set F_n^b. Through factor selection and levels fusion, the dimensionality of the parameter vector is reduced, thus the parameter space for step 2 (below) is reduced based on the parameter vector $\hat{\boldsymbol{\beta}}_{\text{red}}^{(L_0\text{-FGL}),b}$. This parameter space is denoted by $\Omega_2^b \subseteq \mathbb{R}^{1+p^{af,b}}$.

(iii) *Step 2*: MLE is performed on the dataset D_2^b with parameter space Ω_2^b. The resulting estimate is denoted by $\hat{\boldsymbol{\beta}}_{\text{red}}^{\tilde{S}_n^b} \in \mathbb{R}^{1+p^{af,b}}$ and $\hat{\boldsymbol{\beta}}^{\tilde{S}_n^b} \in \mathbb{R}^{p+1}$, respectively (stated in Notation 4.2.2).

(iv) Based on the results of step 1 and step 2, LRTs are performed, testing whether factors $j \in A_n^b$ selected by L_0-FGL are influential, in particular p-values for these hypotheses are obtained resulting in B p-values for each hypothesis.

Aggregating the B p-values for each hypothesis, for which details are given during this section, the final test decision keeping a factor or not is based on the aggregated p-values.

This procedure is called *two-stage L_0-FGL* for multiple sample splitting, where it may be abbreviated as two-stage L_0-FGL whenever needed for simplicity. As in the single split case, elaborated in Section 4.1.2, for each $b = 1,\dots,B$, one can assume $\tilde{S}_n^b, A_n^b$ and $p_j^{af,b}\ \forall j \in \{1,\dots,J\}$ to be known after step 1 (of Algorithm 4.2.1). Further, the fact that nominal and conditional type-I-errors coincide apply in the same way for the multiple split case for each split (cf. Section 4.1.2).

Notation 4.2.2. Similar to Notation 4.1.2, the notation $\boldsymbol{\beta}^{\tilde{S}_n^b}$, $\boldsymbol{\beta}_{\text{red}}^{\tilde{S}_n^b}$, $\boldsymbol{\beta}_j^{\tilde{S}_n^b}$ and $\boldsymbol{\beta}_{j,\text{red}}^{\tilde{S}_n^b}$ for the coefficient vectors and $\hat{\boldsymbol{\beta}}^{\tilde{S}_n^b}$, $\hat{\boldsymbol{\beta}}_{\text{red}}^{\tilde{S}_n^b}$, $\hat{\boldsymbol{\beta}}_j^{\tilde{S}_n^b}$ and $\hat{\boldsymbol{\beta}}_{j,\text{red}}^{\tilde{S}_n^b}$ for the corresponding estimates for each split $b = 1,\dots,B$ is introduced.

In terms of regularity conditions, no further conditions are needed in the multiple split case as imposed in the single split case. GlobalReg (Definition 4.1.7) are assumed for each split $b = 1,\dots,B$, where $B \in \mathbb{N}$ is fixed, which is no further restriction since GlobalReg do not impose a structure on the dataset or the split.

4.2.2 Details of Likelihood Ratio Test and p-Values

For a broad discussion on properties and details on the LRT and the p-values for two stage L_0-FGL in the single split, reference is made to Section 4.1.4, which is valid for the multiple split case as well. However, details on some differences in the multiple split are given below.

After performing MLE in step 2 (of Algorithm 4.2.1), one obtains p-values $\tilde{P}_{j,n}^b$ for each $b = 1,\dots,B$ based on testing whether factor j is influential, i.e.

$$H_{0,j}^b : \boldsymbol{\beta}_{\text{red}}^{\tilde{S}_n^b} = \mathbf{0} \quad \text{versus} \quad H_{1,j}^b : \boldsymbol{\beta}_{\text{red}}^{\tilde{S}_n^b} \neq \mathbf{0}. \tag{4.60}$$

for all $j \in \{1,\dots,J\}$. As in the single split, under the GlobalReg (Definition 4.1.7) assumption, the p-values obtained for the testing problem above are also valid for testing the truth

$$H_{0,j} : \boldsymbol{\beta}_{j,\text{red}}^* = \mathbf{0} \quad \text{versus} \quad H_{1,j} : \boldsymbol{\beta}_{j,\text{red}}^* \neq \mathbf{0}, \tag{4.61}$$

for all $j \in \{1,\dots,J\}$. In particular, it is referred to Remark 4.1.8 which similarly applies in the multiple splitting case. Thus, for every $b = 1,\dots B$, one gets a valid p-value for testing (4.61). The final test decision is based on an aggregated p-value, which is discussed more detailed in

the upcoming section (Section 4.2.3).

With A_n^b denoting the active set of factors selected by L_0-FGL in split $b \in \{1, \dots, B\}$, analogously to (4.15), one sets

$$\tilde{P}_{j,n}^b := \begin{cases} P_{raw,j,n}^b & \text{if } j \in A_n^b, \\ 1 & \text{if } j \notin A_n^b. \end{cases}$$

The p-value $P_{raw,j,n}^b$ for split $b = 1, \dots, B$ is the same as $P_{raw,j,n}$ in the single split case, one compares Section 4.1.4, based on the current split.

For the p-values $\tilde{P}_{j,n}^b$ for which $j \in A_n^b$, one needs to adjust for the multiplicity as in the previous section dealing with single splits. For this, one applies the Bonferroni correction for every split $b \in \{1, \dots, B\}$

$$P_{j,n}^b := \min\left(\tilde{P}_{j,n}^b \cdot |A_n^b|, 1\right) \ \forall j \in \{1, \dots, J\}, \tag{4.62}$$

cf. Section 4.1.5. Further, the Benjamini Yekutieli correction introduced in Section 4.1.6 is applied for the multiple split case in some subsequent section (Section 4.2.5).

Most of the quantities for the multiple split case corresponding to those in the single split case have been introduced. However, what is still missing is the aggregation method of the resulting p-values for each split, which is provided next.

4.2.3 Aggregation of p-Values

The execution of two-stage L_0-FGL in the multiple split case yields B p-values $P_{j,n}^1, \dots, P_{j,n}^B$ for all $j \in \{1, \dots, J\}$, thus it has to be decided how to further treat/aggregate these p-values for each factor j. This is done by following the choice made in Meinshausen et al. (2009), where they propose to use empirical quantiles. For this, $q_\gamma(\cdot)$ denotes the empirical γ quantile function. For some $\gamma \in (0, 1)$, one sets

$$Q_j(\gamma) := \min\left\{1, q_\gamma\left(\left\{\frac{P_{j,n}^b}{\gamma}, b = 1, \dots, B\right\}\right)\right\}. \tag{4.63}$$

Then, for some given $\gamma \in (0, 1)$, there is *one* aggregated p-value $Q_j(\gamma)$ for each factor $j = 1, \dots, J$, which is based on the p-values calculated in each split $b = 1, \dots, B$. To ensure that for each factor $j \in \{1, \dots, J\}$, the same number B of p-values are aggregated, it is crucial to set the p-value to one if a factor is not selected as influential in step 1 (Algorithm 4.2.1) in some cycle $b = 1, \dots, B$. Otherwise, this would result in a loss of information. Thus, in the multiple split case, the number of true null hypotheses is given by $|A^*|$ since all factors are tested after step 2 (Algorithm 4.2.1) based on the aggregated p-values, a fact that was briefly mentioned in Remark 4.1.11.

The remaining problem here is the choice of $\gamma \in (0, 1)$. It is proposed in Meinshausen et al. (2009) to use an adaptive version of $Q_j(\gamma)$, which selects a suitable value of γ for the quantile based on the data. In particular, for γ_{min} being some lower bound for γ, one defines

$$Q_j^{adap} := \min\left\{1, (1 - \log(\gamma_{min})) \inf_{\gamma \in (\gamma_{min}, 1)} Q_j(\gamma)\right\}. \tag{4.64}$$

This extra correction factor $(1 - \log(\gamma_{min}))$ accounts for the fact that the FWER remains controlled at level α, which is justified in the proof of some subsequent theorem (Theorem 4.2.4). In brief, the origin of this factor is lead back to the asymptotic uniform distribution of the p-values $\tilde{P}^b_{j,n}$.

Analogously to the single split case, the Bonferroni correction as well as the Benjamini Yekutieli correction are applied to adjust for multiplicity of testing. Only the resulting final selected sets need to be given, as the procedures were already explained in Sections 4.1.5 and 4.1.6, respectively. The cases of using $Q_j(\gamma)$ for some fixed γ, or the adaptive version Q_j^{adap}, as aggregated p-value are differentiated.

4.2.4 Bonferroni Correction

The resulting sets of final selected predictors for the multiple sample splitting using the aggregation (4.63) and (4.64), respectively, with Bonferroni correction are given by

$$A_{n,f,mult}(\alpha|\gamma) := \{j \in \{1, \dots J\} : Q_j(\gamma) \leq \alpha\}, \tag{4.65}$$

$$A_{n,f,,mult}(\alpha) := \{j \in \{1, \dots, J\} : Q_j^{adap} \leq \alpha\}. \tag{4.66}$$

It is recalled that in $Q_j(\gamma)$ and Q_j^{adap} p-values were aggregated that already include the multiplication with the number of hypotheses, which is $|A_n^b|$ in split $b \in \{1, \dots, B\}$, for which (4.62) is consulted.

If the test decision is based on using the multiplicity correction as given in (4.65), one gets the following decision rule, where null hypothesis $H_{0,j}$ given by (4.61)

$$\text{reject } H_{0,j} \Leftrightarrow Q_j(\gamma) \leq \alpha.$$

The analogous decision rule is applied using the multiplicity correction as given in (4.66) replacing $Q_j(\gamma)$ above by Q_j^{adap}.

4.2.5 Benjamini Yekutieli Correction

For the Benjamini Yekutieli correction in the multiple split case, one proceeds as follows. Having obtained the aggregated p-values $Q_j(\gamma)$ for some given fixed $\gamma \in (0, 1)$ or Q_j^{adap}, respectively, they are ordered according their size and, as in the single split method, for $q \in (0, 1)$, the value

$$k(q|\gamma) := \max\left\{i \in \{1, \dots J\} \,|\, Q_{(i)}(\gamma) \leq iq\right\}, \tag{4.67}$$

is determined, or, using the adaptive version

$$k(q) := \max\{i \in \{1, \dots, J\} \,|\, Q_{(i)}^{adap} \leq iq\}. \tag{4.68}$$

Then, all null hypotheses for which $i \leq k(q)$ (or $i \leq k(q|\gamma)$, respectively) are rejected, hence the final selected sets of predictors for the multiple sample splitting using the Benjamini Yekutieli correction are given by

$$A_{n,f}^{(BY)}(q|\gamma) := \{i \in \{1, \dots J\} \,|\, Q_i(\gamma) \leq P_{(k(q|\gamma))}\}, \tag{4.69}$$

$$A_{n,f}^{(BY)}(q) := \{i \in \{1, \dots J\} \,|\, Q_i^{adap} \leq P_{(k(q))}\}. \tag{4.70}$$

4.2.6 Type-I-error Control and Consistent Selection with Bonferroni Correction for Multiple Split

In Theorem 4.2.3 provided below, FWER control is shown using the aggregated p-values $Q_j(\gamma)$ based on the Bonferroni correction to adjust for multiplicity. In particular, it is shown that, asymptotically, the probability of at least one false rejection can be bounded from above by the level α, so type-I-error control is obtained.

Theorem 4.2.3 (Type-I-error control/FWER control of two-stage L_0-FGL in multiple split with Bonferroni applied on $Q_j(\gamma)$, fixed p and diverging p_n)**.** It is assumed that GlobalReg (Definition 4.1.7) hold for every split $b \in \{1, \dots B\}$, where $B \in \mathbb{N}$ is fixed and $\gamma \in (0, 1)$ is given. Further, it is assumed that the LRT is executed at some given level $\alpha \in (0, 1)$, resulting in $A_{n,f,mult}(\alpha|\gamma)$ defined as in (4.65). Then, it holds that two-stage L_0-FGL (multiple split) controls the type-I-error/FWER, in particular

$$\limsup_{n\to\infty} \mathbb{P}(A_{n,f,mult}(\alpha|\gamma) \cap (A^*)^c \neq \emptyset) \leq \alpha. \tag{4.71}$$

Proof. Since this theorem holds for p being fixed and p_n being allowed to diverge, J and p instead of J_n and p_n are written in this proof for simplicity. The proof roots in Meinshausen et al. (2009) (proof of Theorem 3.1) transferred to two-stage L_0-FGL (cf. Remark 4.1.17). To show the claim, the goal is to show that

$$\limsup_{n\to\infty} \mathbb{P}\left(\exists\, j \in (A^*)^c : Q_j(\gamma) \leq \alpha\right) \leq \alpha. \tag{4.72}$$

For all $b = 1, \dots, B$, it holds by GlobalReg (Definition 4.1.7)

$$\begin{aligned} \mathbb{P}(A^* \not\subseteq A_n^b) &= o(1), \\ \mathbb{P}(F^* \neq F_n^b) &= o(1). \end{aligned}$$

It is noted that here screening properties concerning selection *and* fusion are required (cf. Remark 4.1.17 (ii)). Under these properties one deduces by the Bonferroni inequality

$$\mathbb{P}\left(A^* \not\subseteq A_n^b \text{ for at least one } b \in \{1, \dots, B\}\right) \leq \sum_{b=1}^{B} \mathbb{P}(A^* \not\subseteq A_n^b) = o(1), \tag{4.73}$$

$$\mathbb{P}\left(F^* \neq F_n^b \text{ for at least one } b \in \{1, \dots, B\}\right) \leq \sum_{b=1}^{B} \mathbb{P}(F^* \neq F_n^b) = o(1), \tag{4.74}$$

since

$$\begin{aligned} \{A^* \not\subseteq A_n^b \text{ for at least one } b \in \{1, \dots, B\}\} &\subseteq \bigcup_{b=1}^{B} \{A^* \not\subseteq A_n^b\}, \\ \{F^* \neq F_n^b \text{ for at least one } b \in \{1, \dots, B\}\} &\subseteq \bigcup_{b=1}^{B} \{F^* \neq F_n^b\}. \end{aligned}$$

Consequently one can write (abbreviating "for at least one" by "f.a.l.o.") by applying the law

of total probability twice

$$\begin{aligned}
& \mathbb{P}\left(\exists\, j \in (A^*)^c : Q_j(\gamma) \leq \alpha\right) \qquad (4.75)\\
= \;& \mathbb{P}\left(\exists\, j \in (A^*)^c : Q_j(\gamma) \leq \alpha | A^* \subseteq A_n^b \text{ and } F^* = F_n^b \;\forall\, b \in \{1,\dots,B\}\right)\\
& + \mathbb{P}\left(\exists\, j \in (A^*)^c : Q_j(\gamma) \leq \alpha | A^* \not\subseteq A_n^b \text{ f.a.l.o. } b \in \{1,\dots,B\} \text{ and } F^* = F_n^b \;\forall\, b \in \{1,\dots,B\}\right)\\
& + \mathbb{P}\left(\exists\, j \in (A^*)^c : Q_j(\gamma) \leq \alpha | F^* \neq F_n^b \text{ f.a.l.o. } b \in \{1,\dots,B\} \text{ and } A^* \subseteq A_n^b \;\forall\, b \in \{1,\dots,B\}\right)\\
& + \mathbb{P}\left(\exists\, j \in (A^*)^c : Q_j(\gamma) \leq \alpha | A^* \not\subseteq A_n^b \text{ and } F^* \neq F_n^b \text{ f.a.l.o. } b \in \{1,\dots,B\}\right)\\
\leq \;& \mathbb{P}\left(\exists\, j \in (A^*)^c : Q_j(\gamma) \leq \alpha | A^* \subseteq A_n^b \text{ and } F^* = F_n^b \;\forall\, b = 1,\dots,B\right)\\
& + \mathbb{P}\left(A^* \not\subseteq A_n^b \text{ f.a.l.o. } b \in \{1,\dots,B\}\right)\\
& + \mathbb{P}\left(F^* \neq F_n^b \text{ f.a.l.o. } b \in \{1,\dots,B\}\right) + \mathbb{P}\left(A^* \not\subseteq A_n^b \text{ f.a.l.o. } b \in \{1,\dots,B\}\right)\\
\leq \;& \underbrace{\mathbb{P}\left(\exists\, j \in (A^*)^c : Q_j(\gamma) \leq \alpha | A^* \subseteq A_n^b \text{ and } F^* = F_n^b \;\forall\, b = 1,\dots,B\right)}_{=:(*)} + o(1). \qquad (4.76)
\end{aligned}$$

where in (4.76) inequalities (4.73) and (4.74) are applied. For simplicity of notation, the latter condition in the probabilities and expectations are written in the following steps as

$$(+) := \{A^* \subseteq A_n^b,\, F^* = F_n^b \;\forall\, b = 1,\dots,B\}.$$

As a next step, the first summand $(*)$ of (4.76) is further analyzed. Throughout the whole proof, the $\min(1,\cdot)$ function in the definition of $P_{j,n}^b$ and $Q_j(\gamma)$ is omitted such that it is assumed that $P_{j,n} \leq 1$ and $Q_j(\gamma) \leq 1$. This does clearly not affect the selection process of the test.

Further, one defines the following quantity $\pi_j(u)$ representing the proportion of p-values $P_{j,n}^b$ among the splits $b = 1,\dots,B$ being less than or equal to $u \in \mathbb{R}$, in particular

$$\pi_j(u) := \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{(-\infty,u]}\left(P_{j,n}^b\right). \qquad (4.77)$$

It can be recognized that the events $\{Q_j(\gamma) \leq \alpha\}$ and $\{\pi_j(\alpha\gamma) \geq \gamma\}$ are equivalent since, on the one hand, starting with $\{Q_j(\gamma) \leq \alpha\}$ and the definition of $Q_j(\gamma)$ one deduces

$$Q_j(\gamma) \leq \alpha \Leftrightarrow q_\gamma\left(\left\{\frac{1}{\gamma}P_{j,n}^b, b = 1,\dots,B\right\}\right) \leq \alpha, \qquad (4.78)$$

where the latter means that, by the definition of the empirical quantile, $q_\gamma(\cdot)$ is an element for which it holds that

$$\frac{1}{\gamma}P_{j,n}^b \leq q_\gamma\left(\left\{\frac{1}{\gamma}P_{j,n}^b, b = 1,\dots,B\right\}\right) \quad \text{for at least } \gamma \cdot B \text{ items.} \qquad (4.79)$$

Equation (4.79) holds because for a random sample $\boldsymbol{S} := (S_1, S_2, \dots, S_B)$, the empirical quantile $q_\gamma(\boldsymbol{S})$ is by definition an element which fulfills that a fraction γ of the sample is less than or equal to this quantile $q_\gamma(\boldsymbol{S})$ and a fraction of $1-\gamma$ is greater than or equal to this quantile $q_\gamma(\boldsymbol{S})$. Using that $\boldsymbol{S}$ is of size B, it holds for at least $\gamma \cdot B$ items (S_b is some entry of $\boldsymbol{S}$, $b = 1,\dots,B$) that $S_b \leq q_\gamma(\boldsymbol{S})$ and for at least $(1-\gamma)\cdot B$ items that $S_b \geq q_\gamma(\boldsymbol{S})$, $b = 1,\dots,B$. Now, this consideration is applied to $S_b = \frac{1}{\gamma}P_{j,n}^b$, hence $\boldsymbol{S} = \left(\frac{1}{\gamma}P_{j,n}^1, \dots, \frac{1}{\gamma}P_{j,n}^B\right)$ which yields (4.79).
One can write

$$(4.79) \Leftrightarrow P_{j,n}^b \leq \gamma q_\gamma\left(\left\{\frac{1}{\gamma}P_{j,n}^b, b = 1,\dots,B\right\}\right) \quad \text{for at least } \gamma \cdot B \text{ items.} \qquad (4.80)$$

Hence, the inequality on the right hand side of (4.78) means that

$$P_{j,n}^b \leq \gamma q_\gamma \left(\left\{ \frac{1}{\gamma} P_{j,n}^b, b = 1, \ldots, B \right\} \right) \leq \gamma\alpha \text{ for at least } \gamma \cdot B \text{ items.} \tag{4.81}$$

On the other hand, starting with $\pi_j(\alpha\gamma) \geq \gamma$, one can infer

$$\pi_j(\alpha\gamma) \geq \gamma \Leftrightarrow \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}_{(-\infty,\alpha\gamma]} \left(P_{j,n}^b \right) \geq \gamma \Leftrightarrow \sum_{b=1}^{B} \mathbf{1}_{(-\infty,\alpha\gamma]} \left(P_{j,n}^b \right) \geq B \cdot \gamma,$$

where the latter means that $P_{j,n}^b \leq \alpha\gamma$ for at least $\gamma \cdot B$ items, which is the same as (4.81). To sum up, one can deduce

$$Q_j(\gamma) \leq \alpha \Leftrightarrow \pi_j(\alpha\gamma) \geq \gamma. \tag{4.82}$$

Having shown this equivalence (4.82), one can re-write $(*)$ in (4.76) as

$$\begin{aligned}
(*) \quad &= \quad \mathbb{P}\left(\exists\, j \in (A^*)^c : Q_j(\gamma) \leq \alpha | (+)\right) \\
&\leq \quad \sum_{j \in (A^*)^c} \mathbb{P}\left(Q_j(\gamma) \leq \alpha | (+)\right) \\
&= \quad \sum_{j \in (A^*)^c} \mathbb{E}\left(\mathbf{1}_{(-\infty,\alpha]}(Q_j(\gamma)) | (+) \right) \\
&\overset{(4.82)}{=} \quad \sum_{j \in (A^*)^c} \mathbb{E}\left(\mathbf{1}_{[\gamma,\infty)}(\pi_j(\alpha\gamma)) | (+) \right).
\end{aligned} \tag{4.83}$$

In the next step, the Markov inequality is applied, which says that for $h(\cdot)$ being a monotonic increasing function and $Z : \Omega \to \mathbb{R}$ being a real-valued random variable, it holds that

$$h(a)\mathbb{P}(Z \geq a) \leq \mathbb{E}(h(Z)).$$

Choose the function h to be the identity and $a := \gamma$ resulting in the following inequality

$$\mathbb{E}\left(\mathbf{1}_{[\gamma,\infty)}(\pi_j(\alpha\gamma)) \,|\, (+) \right) = \mathbb{P}\left(\pi_j(\alpha\gamma) \geq \gamma \,|\, (+)\right) \leq \frac{1}{\gamma} \cdot \mathbb{E}\left(\pi_j(\alpha\gamma) \,|\, (+)\right). \tag{4.84}$$

To sum up, one can infer for $(*)$ the following inequalities and equalities starting with (4.83), where explanations of the steps can be found below

$$\begin{aligned}
(*) \quad &\overset{(4.83)}{\leq} \sum_{j \in (A^*)^c} \mathbb{E}\left(\mathbf{1}_{[\gamma,\infty)}(\pi_j(\alpha\gamma)) | (+) \right) \\
&\overset{(4.84)}{\leq} \frac{1}{\gamma} \sum_{j \in (A^*)^c} \mathbb{E}(\pi_j(\gamma\alpha) \,|\, (+)) \\
&\overset{\text{Def.}\pi_j}{=} \frac{1}{\gamma} \sum_{j \in (A^*)^c} \mathbb{E}\left(\frac{1}{B} \sum_{b=1}^{B} \mathbf{1}_{(-\infty,\alpha\gamma]}(P_{j,n}^b) \,|\, (+) \right) \\
&\overset{\text{Lin.}\mathbb{E}}{=} \frac{1}{\gamma} \frac{1}{B} \sum_{j \in (A^*)^c} \sum_{b=1}^{B} \mathbb{E}\left(\mathbf{1}_{(-\infty,\alpha\gamma]}(P_{j,n}^b) \,|\, (+) \right) \\
&= \frac{1}{\gamma} \frac{1}{B} \sum_{b=1}^{B} \sum_{j \in (A^*)^c \cap (A_n^b)} \mathbb{E}\left(\mathbf{1}_{(-\infty,\alpha\gamma]}(P_{j,n}^b) \,|\, (+) \right)
\end{aligned} \tag{4.85}$$

$$
\begin{aligned}
&= \frac{1}{\gamma}\frac{1}{B}\sum_{b=1}^{B}\sum_{j\in(A^*)^c\cap(A_n^b)} \mathbb{P}\left(P_{j,n}^b \leq \alpha\gamma \,|\, (+)\right) \\
&\leq \frac{1}{\gamma}\frac{1}{B}\sum_{b=1}^{B}\sum_{j\in(A^*)^c\cap(A_n^b)} \left(\frac{\alpha\gamma}{|A_n^b|} + o(1)\right) \qquad (4.86) \\
&= \frac{1}{\gamma}\cdot\frac{1}{B}\cdot\sum_{b=1}^{B}\left(|(A^*)^c\cap(A_n^b)|\frac{\alpha\gamma}{|A_n^b|} + o(1)\cdot|(A^*)^c\cap(A_n^b)|\right) \\
&\leq \frac{1}{\gamma}\cdot\frac{1}{B}\cdot\sum_{b=1}^{B}\left(|A_n^b|\frac{\alpha\gamma}{|A_n^b|} + o(1)\right) \qquad (4.87) \\
&\leq \alpha + \frac{1}{\gamma}\,o(1) = \alpha + o(1). \qquad (4.88)
\end{aligned}
$$

The following details are provided for the inequalities and equalities above:

- In (4.85) it is used that for $j \in (A^*)^c\backslash(A_n^b)$ it holds $\tilde{P}_j^b = 1$ hence $P_{j,n}^b = 1$ and since $\alpha\gamma < 1$ one deduces $\mathbf{1}_{(-\infty,\alpha\gamma]}(P_{j,n}^b) = 0$ for those $j \in (A^*)^c\backslash(A_n^b)$, hence the sum can be written over $j \in (A^*)^c \cap (A_n^b)$.
- In (4.86) it is used that all the probabilities are conditional on $A^* \subseteq A_n^b$ and $F^* = F_n^b \;\forall\, b = 1,\ldots,B$ (one recalls the definition of $(+)$), and the p-values $\tilde{P}_{j,n}^b$ are chosen as *valid* p-values corresponding to the LRT being *asymptotically* uniform, similar to (4.58) in the single split case.
- In (4.87) it is used that by (A4') the number of factors selected by L_0-FGL in step 1 (of Algorithm 4.2.1) is bounded, thus $|(A^*)^c \cap A_n^b| \cdot o(1) = O(1)\cdot o(1) = o(1)$. This step is crucial since due to the fact that the p-values are *asymptotically* uniform distributed, the term $o(1)$ appears in (4.86). Thus, one needs to ensure that this term remains controlled going to zero (i.e. $o(1)$), even though the sum depending on n is considered in (4.86).

Now, it can be concluded that $(*) \leq \alpha + o(1)$, thus

$$
\lim_{n\to\infty}(*) \overset{\text{Def}}{=} \lim_{n\to\infty}\mathbb{P}\left(\exists\, j \in (A^*)^c : Q_j(\gamma) \leq \alpha | (+)\right) \leq \alpha,
$$

which, using (4.75) and (4.76), further yields

$$
\mathbb{P}\left(\exists\, j \in (A^*)^c : Q_j(\gamma) \leq \alpha\right) \leq \alpha + o(1),
$$

so finally

$$
\limsup_{n\to\infty}\mathbb{P}\left(\exists\, j \in (A^*)^c : Q_j(\gamma) \leq \alpha\right) \leq \alpha.
$$

□

Similar to Theorem 4.2.3, type-I-error control with Bonferroni correction is provided using the adaptive version of p-values Q_j^{adap}, hence the selected set given by (4.66). In particular, this error control is provided in Theorem 4.2.4.

Theorem 4.2.4 (Type-I-error control/FWER control of two-stage L_0-FGL in multiple split with Bonferroni applied on Q_j^{adap}, fixed p and diverging p_n)**.** One assumes that GlobalReg (Definition 4.1.7) hold for every split $b \in \{1,\ldots,B\}$, where $B \in \mathbb{N}$ is fixed. Further, one assumes that the LRT is executed at some given level $\alpha \in (0,1)$ resulting in $A_{n,f,mult}(\alpha)$ defined as in (4.66). Then, it holds that two-stage L_0-FGL (multiple split) controls the type-I-error/FWER, in particular

$$
\limsup_{n\to\infty}\mathbb{P}\left(A_{n,f,mult}(\alpha)\cap(A^*)^c \neq \emptyset\right) \leq \alpha. \qquad (4.89)
$$

Proof. To show the claim, the goal is to show

$$\limsup_{n\to\infty} \mathbb{P}\left(\exists\, j \in (A^*)^c \,:\, Q_j^{adap} \leq \alpha\right) \leq \alpha. \tag{4.90}$$

The first part of the proof is analogous to the proof of Theorem 4.2.3, especially the steps around equation (4.75), i.e. it roots in Meinshausen et al. (2009) (Theorem 3.2), adjusted for two-stage L_0-FGL (cf. Remark 4.1.17). Adapting the first steps from Theorem 4.2.3 using the law of total probability twice, the remaining goal is to show

$$\limsup_{n\to\infty} \mathbb{P}\left(\exists\, j \in (A^*)^c \,:\, Q_j^{adap} \leq \alpha \;\middle|\; \underbrace{A^* \subseteq A_n^b \text{ and } F^* = F_n^b \;\forall\, b = 1,\ldots,B}_{=:(+)}\right) \leq \alpha. \tag{4.91}$$

Throughout the whole proof, as in the proof of Theorem 4.2.3, without loss of generality, the $\min(1,\cdot)$ function in the definition of $P_{j,n}^b$ and Q_j^{adap} is omitted and it is assumed that $P_{j,n}^n \leq 1$ and $Q_j^{adap} \leq 1$, the same can be applied for $\tilde{P}_{j,n}^b$.

Under the null hypothesis, hence for $j \in (A^*)^c \cap A_n^b$, for each b, it holds that the p-values $P_{j,n}^b$ are asymptotically uniform distributed on $[0,1]$. Moreover, for all $\gamma \in (0,1)$ (later $\gamma \in (\gamma_{min},1)$ is observed) one has $\alpha\cdot\gamma \in (0,1)$ and further $P_{j,n}^b = 1$ for $j \notin A_n^b$, which yields

$$\mathbf{1}_{(-\infty,\alpha\gamma]}\left(P_{j,n}^b\right) = 0 \quad \text{for } j \notin A_n^b.$$

Next, an expression for the expectation of the supremum of some uniform distributed random variable is derived, which is later replaced by the p-values $\tilde{P}_{j,n}^b$ which are *asymptotically* uniform distributed given $(+)$. In particular, for some continuous real-valued random variable, denoted by U, with support $\text{supp}(U) = [0,1] \subseteq \mathbb{R}$, it holds (justification is provided below)

$$\sup_{\gamma\in(\gamma_{min},1)} \frac{\mathbf{1}_{(-\infty,\alpha\gamma]}(U)}{\gamma} = \begin{cases} 0 & U \geq \alpha, \\ \frac{\alpha}{U} & \alpha\gamma_{min} \leq U < \alpha, \\ \frac{1}{\gamma_{min}} & U < \alpha\gamma_{min}. \end{cases} \tag{4.92}$$

The following lines justify (4.92).

- Case 1: $U \geq \alpha$. Then, since $\gamma \in (0,1)$, it follows that $U > \alpha\gamma$, i.e. $\mathbf{1}_{(-\infty,\alpha\gamma]}(U) = 0$ so the supremum is clearly zero.
- Case 2: $\alpha\gamma_{min} \leq U \leq \alpha$. Then, the supremum is clearly attained at $\gamma = \frac{U}{\alpha} \in (\gamma_{min},1)$, which lies between γ_{min} and one by the definition of case 2. With that, one gets $\mathbf{1}_{(-\infty,\alpha\gamma]}(U) = \mathbf{1}_{(-\infty,U]}(U) = 1$ and the supremum of the whole expression equals $\frac{1}{\gamma} = \frac{\alpha}{U}$.
- Case 3: $U < \alpha\gamma_{\min}$. Then, it is known that $U \leq \alpha\gamma$ holds for all $\gamma \in (\gamma_{min},1)$. Consequently the supremum is attained at $\gamma = \gamma_{min}$, thus the supremum equals $\frac{1}{\gamma_{min}}$.

With (4.92), one can further deduce

$$\mathbb{E}\left(\sup_{\gamma\in(\gamma_{min},1)} \frac{\mathbf{1}_{(-\infty,\alpha\gamma]}(U)}{\gamma}\right) = \int_0^{\gamma_{min}\cdot\alpha} \gamma_{min}^{-1}\mathrm{d}x + \int_{\alpha\gamma_{min}}^{\alpha} \alpha x^{-1}\mathrm{d}x = \alpha(1-\log(\gamma_{min})). \tag{4.93}$$

In contrast to Meinshausen et al. (2009), due to the fact that the p-values $\tilde{P}_{j,n}^b$ considered here are *asymptotically* uniform distributed (cf. Remark 4.1.17), further steps according to the weak

convergence of the corresponding probability measure of the p-values are required, thus (4.93) cannot be applied directly.

It holds that the conditional distribution of $\tilde{P}^b_{j,n}$ (conditional on $A^* \subseteq A^b_n$ and $F^* = F^b_n$) converges to a uniform distribution on $[0,1]$ for all $j \in (A^*)^c$. Denote the probability measure of $\tilde{P}^b_{j,n}$ conditional on $A^* \subseteq A^b_n$ and $F^* = F^b_n$ by $\mathbb{P}_{\tilde{P}^b_{j,n}}$, where it is noted that $\mathbb{P}_{\tilde{P}^b_{j,n}}$ depends on n. The probability measure of the uniform distribution on $[0,1]$ is denoted by $\mathbb{P}_U$. It holds $\tilde{P}^b_{j,n} \to_d U$, so the corresponding probability measures converge weakly, i.e.

$$\mathbb{P}_{\tilde{P}^b_{j,n}} \to_w \mathbb{P}_U. \tag{4.94}$$

Further $\lim_{n\to\infty} \mathbb{P}(A^* \not\subseteq A^b_n) = o(1)$ and $\lim_{n\to\infty} \mathbb{P}(F^* \neq F^b_n) = o(1)$ by GlobalReg (Definition 4.1.7), hence it holds

$$\begin{aligned}
& \lim_{n\to\infty} \mathbb{E}\left(\sup_{\gamma\in(\gamma_{min},1)} \frac{\mathbf{1}_{(-\infty,\alpha\gamma]}(\tilde{P}^b_{j,n})}{\gamma} \right) \\
\leq\ & \lim_{n\to\infty} \mathbb{E}\left(\sup_{\gamma\in(\gamma_{min},1)} \frac{\mathbf{1}_{(-\infty,\alpha\gamma]}(\tilde{P}^b_{j,n})}{\gamma} \,\middle|\, (+) \right) + o(1) \\
\overset{(4.92)}{=}\ & \lim_{n\to\infty} \mathbb{E}\left(\frac{\alpha}{\tilde{P}^b_{j,n}} \mathbf{1}_{[\alpha\gamma_{\min},\alpha)}(\tilde{P}^b_{j,n}) + \frac{1}{\gamma_{\min}} \mathbf{1}_{(-\infty,\alpha\gamma_{min})}(P^b_j) \,\middle|\, (+) \right) + o(1) \\
=\ & \lim_{n\to\infty} \int_0^{\alpha\cdot\gamma_{min}} \frac{1}{\gamma_{\min}} \mathrm{d}\mathbb{P}_{\tilde{P}^b_{j,n}}(x) + \lim_{n\to\infty} \int_{\alpha\gamma_{\min}}^{\alpha} \alpha x^{-1} \mathrm{d}\mathbb{P}_{\tilde{P}^b_{j,n}}(x) \\
\overset{(4.94)}{=}\ & \int_0^{\alpha\cdot\gamma_{min}} \frac{1}{\gamma_{\min}} \mathrm{d}\mathbb{P}_U(x) + \int_{\alpha\gamma_{\min}}^{\alpha} \alpha x^{-1} \mathrm{d}\mathbb{P}_U(x) \qquad (4.95) \\
\overset{(4.93)}{=}\ & \alpha(1 - \log(\gamma_{min})), \qquad (4.96)
\end{aligned}$$

Consequently, one infers

$$\mathbb{E}\left(\sup_{\gamma\in(\gamma_{min},1)} \frac{\mathbf{1}_{(-\infty,\alpha\gamma]}(\tilde{P}^b_{j,n})}{\gamma} \right) \leq \alpha(1 - \log(\gamma_{min})) + o(1). \tag{4.97}$$

Moreover, using that $\alpha\gamma \in (0,1)$ and $P^b_{j,n} = 1$ for $j \notin A^b_n$, one receives the following equalities and inequalities

$$\begin{aligned}
& \sum_{j\in(A^*)^c} \mathbb{E}\left(\sup_{\gamma\in(\gamma_{min},1)} \frac{\mathbf{1}_{(-\infty,\alpha\gamma]}(P^b_{j,n})}{\gamma} \right) \\
=\ & \sum_{j\in(A^*)^c\cap A^b_n} \mathbb{E}\left(\sup_{\gamma\in(\gamma_{min},1)} \frac{\mathbf{1}_{(-\infty,\alpha\gamma]}(P^b_{j,n})}{\gamma} \right) \\
=\ & \sum_{j\in(A^*)^c\cap A^b_n} \mathbb{E}\left(\sup_{\gamma\in(\gamma_{min},1)} \frac{\mathbf{1}_{\left(-\infty,\frac{\alpha\gamma}{|A^b_n|}\right]}(\tilde{P}^b_{j,n})}{\gamma} \right) \quad \left(\text{since } P^b_{j,n} = \tilde{P}^b_{j,n}|A^b_n|\right) \\
\leq\ & |(A^*)^c \cap A^b_n| \cdot \max_{j\in(A^*)^c\cap A^b_n} \mathbb{E}\left(\sup_{\gamma\in(\gamma_{min},1)} \frac{\mathbf{1}_{\left(-\infty,\frac{\alpha\gamma}{|A^b_n|}\right]}(\tilde{P}^b_{j,n})}{\gamma} \right) \qquad (4.98) \\
\overset{(4.97)}{\leq}\ & |(A^*)^c \cap A^b_n| \frac{\alpha}{|A^b_n|}(1 - \log(\gamma_{min})) + |(A^*)^c \cap A^b_n| o(1) \\
\leq\ & \alpha(1 - \log(\gamma_{min})) + o(1). \qquad (4.99)
\end{aligned}$$

It is noted that in (4.99) it was used that by (A4') it is known that $|(A^*)^c \cap A_n^b| = O(1)$ and consequently $o(1) \cdot |(A^*)^c \cap A_n^b| = o(1)$.

Since the inequality (4.99) holds for every $b = 1, \dots, B$ it also holds for the average over $b = 1, \dots, B$. In particular, since the supremum of the sum of two functions can be bounded from above by the sum of their suprema, one deduces

$$\begin{aligned}\sum_{j\in(A^*)^c} \mathbb{E}\left(\sup_{\gamma\in(\gamma_{min},1)} \frac{\frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{(-\infty,\alpha\gamma]}(P_{j,n}^b)}{\gamma}\right) &\leq \frac{1}{B}\sum_{b=1}^{B}\sum_{j\in(A^*)^c} \mathbb{E}\left(\sup_{\gamma\in(\gamma_{min},1)} \frac{\mathbf{1}_{(-\infty,\alpha\gamma]}(P_{j,n}^b)}{\gamma}\right) \\ &\overset{(4.99)}{\leq} \frac{1}{B}\sum_{b=1}^{B}[\alpha(1-\log(\gamma_{min})) + o(1)] \\ &= \alpha(1-\log(\gamma_{min})) + o(1). \end{aligned} \tag{4.100}$$

The definition of the function $\pi_j(\cdot)$ yields $\pi_j(\alpha\gamma) = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{(-\infty,\alpha\gamma]}(P_{j,n}^b)$ which, using (4.100), allows to infer

$$\sum_{j\in(A^*)^c} \mathbb{E}\left(\sup_{\gamma\in(\gamma_{min},1)} \frac{\pi_j(\alpha\gamma)}{\gamma}\right) \leq \alpha(1-\log(\gamma_{min})) + o(1). \tag{4.101}$$

Using equation (4.101) and the Markov inequality (similar to (4.84)) one can see that

$$\begin{aligned}\sum_{j\in(A^*)^c} \mathbb{E}\left(\sup_{\gamma\in(\gamma_{min},1)} \mathbf{1}_{[\gamma,\infty)}(\pi_j(\alpha\gamma))\right) &\leq \sum_{j\in(A^*)^c} \frac{1}{\gamma}\mathbb{E}\left(\sup_{\gamma\in(\gamma_{min},1)} \pi_j(\alpha\gamma)\right) \\ &\overset{(4.101)}{\leq} \alpha(1-\log(\gamma_{min})) + o(1). \end{aligned} \tag{4.102}$$

Since it was previously shown that $Q_j(\gamma) \leq \alpha \Leftrightarrow \pi_j(\alpha\gamma) \geq \gamma$ (proof of Theorem 4.2.3), one finds

$$\begin{aligned}\sum_{j\in(A^*)^c} \mathbb{P}\left(\inf_{\gamma\in(\gamma_{min},1)} Q_j(\gamma) \leq \alpha\right) &= \sum_{j\in(A^*)^c} \mathbb{P}\left(\inf_{\gamma\in(\gamma_{min},1)} \pi_j(\alpha\gamma) \geq \gamma\right) \\ &\leq \sum_{j\in(A^*)^c} \mathbb{P}\left(\sup_{\gamma\in(\gamma_{min},1)} \pi_j(\alpha\gamma) \geq \gamma\right) \\ &= \sum_{j\in(A^*)^c} \mathbb{E}\left(\sup_{\gamma\in(\gamma_{min},1)} \mathbf{1}_{[\gamma,\infty)}(\pi_j(\alpha\gamma))\right) \\ &\overset{(4.102)}{\leq} \alpha(1-\log(\gamma_{min})) + o(1). \end{aligned} \tag{4.103}$$

Applying the inequality (4.103) above to a level $\tilde{\alpha} := \dfrac{\alpha}{1-\log(\gamma_{min})} \in (0,1)$ (since $\alpha, \gamma_{min} \in (0,1)$ hence $1 - \log(\gamma_{min}) \geq 1$) one receives

$$\sum_{j\in(A^*)^c} \mathbb{P}\left(\underbrace{\inf_{\gamma\in(\gamma_{min},1)} Q_j(\gamma)(1-\log(\gamma_{min}))}_{=Q_j^{adap}} \leq \alpha\right) \leq \alpha + o(1). \tag{4.104}$$

To conclude, by the definition of Q_j^{adap} it holds that

$$\begin{aligned} (4.104) \quad &\Leftrightarrow \quad \sum_{j \in (A^*)^c} \mathbb{P}\left(Q_j^{adap} \leq \alpha\right) \leq \alpha + o(1) \\ &\Rightarrow \quad \mathbb{P}\left(\min_{j \in (A^*)^c} Q_j^{adap} \leq \alpha\right) \leq \alpha + o(1) \\ &\Rightarrow \quad \limsup_{n \to \infty} \mathbb{P}\left(\min_{j \in (A^*)^c} Q_j^{adap} \leq \alpha\right) \leq \alpha, \end{aligned} \tag{4.105}$$

thus (4.90) is proven. Hence, the probability that there exists at least one $j \in (A^*)^c$ (true noise) for which it holds $j \in A_{n,f,mult}(\alpha)$, hence (falsely) evaluated as influential by two-stage L_0-FGL with multiple split, is asymptotically bounded by α and the claim follows. □

Selection consistency for two-stage L_0-FGL in multiple split using Bonferroni correction and the aggregated p-values Q_j^{adap} is investigated in the next theorem (Theorem 4.2.5).

Theorem 4.2.5 (Selection consistency of two-stage L_0-FGL in multiple split using Bonferroni correction with Q_j^{adap}, fixed p and diverging p_n)**.** It is assumed that GlobalReg (Definition 4.1.7) hold for every split $b \in \{1, \ldots, B\}$, where $B \in \mathbb{N}$ is fixed. Further, it is assumed that the single split is factor selection consistent, that is, for a sequence $(\alpha_n)_{n \in \mathbb{N}} \subseteq (0, 1)$ with $\alpha_n \to 0$ for $n \to \infty$ it holds

$$\lim_{n \to \infty} \mathbb{P}\left(A^* = A_{n,f}(\alpha_n)\right) = 1. \tag{4.106}$$

Then, for any choice of $\gamma_{min} \in (0, 1)$, the two-stage L_0-FGL with multiple split based on $A_{n,f,mult}(\alpha_n)$, hence based on the aggregated p-values Q_j^{adap}, is factor selection consistent, that is

$$\lim_{n \to \infty} \mathbb{P}\left(A^* = A_{n,f,mult}(\alpha_n)\right) = 1. \tag{4.107}$$

Proof. The main idea of the proof roots in Meinshausen et al. (2009) (proof of Corollary 3.1), which is brief, but is adapted thoroughly for two-stage L_0-FGL (cf. Remark 4.1.17). As in the previous proofs, the $\min(1, \cdot)$ function in the definitions of $P_{j,n}^b, Q_j^{adap}$ and $Q_j(\gamma)$ can be omitted and it can be assumed that these quantities are all ≤ 1.

In this theorem it is assumed that the two-stage L_0-FGL in the *single split* is factor selection consistent (cf. Theorem 4.1.16), thus (4.106) is satisfied. By definition it holds that $A_{n,f}(\alpha_n) = \{j \in A_n : \tilde{P}_{j,n}|A_n| \leq \alpha_n\}$ hence by (4.106) one deduces for the *single split*

$$\lim_{n \to \infty} \mathbb{P}\left(\tilde{P}_{j,n}|A_n| \leq \alpha_n\right) = 1 \quad \forall j \in A^*. \tag{4.108}$$

Actually, as a step in between to receive (4.108), one needs to apply the law of total probability twice conditioning on the screening properties. However, since this is completely analogous to the previous proofs, it is left out here and in the further steps for simplicity. Since $|A^*|$ is bounded from above by the sparsity assumption for both cases of p being fixed and p_n being allowed to diverge (Definition 1.2.4) it is known that the maximum over all p-values for $j \in A^*$ exists. Thus, (4.108) yields

$$\lim_{n \to \infty} \mathbb{P}\left(\max_{j \in A^*} \tilde{P}_{j,n}|A_n| \leq \alpha_n\right) = 1, \tag{4.109}$$

i.e. (4.109) holds for every p-value $\tilde{P}^b_{j,n}$ in split $b \in \{1,\dots,B\}$

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{j\in A^*} \tilde{P}^b_{j,n}|A^b_n| \leq \alpha_n\right) = 1 \ \forall b \in \{1,\dots,B\}.$$

Since $B \in \mathbb{N}$ is fixed, one further deduces

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{j\in A^*} \max_{b\in\{1,\dots,B\}} \tilde{P}^b_{j,n}|A^b_n| \leq \alpha_n\right) = 1. \tag{4.110}$$

The steps above are used in the following lines, where it is shown that

$$\lim_{n\to\infty} \mathbb{P}\left(Q_j^{adap} \leq (1-\log(\gamma_{min}))\cdot\alpha_n\right) = 1 \ \ \forall j \in A^*. \tag{4.111}$$

By definition, for $\gamma = 1$

$$Q_j(1) = q_1\left(\left\{\tilde{P}^b_{j,n}|A^b_n|, b = 1,..,B\right\}\right) =: q_{1,j},$$

for which $q_{1,j}$ is used as abbreviation in the following. By the definition of the empirical quantile for $\gamma = 1$, using $P^b_{j,n} = \tilde{P}^b_{j,n}|A^b_n|$ one can deduce

$$q_{1,j} = \max_{b\in\{1,\dots,B\}} P^b_{j,n} = \max_{b\in\{1,\dots,B\}} \tilde{P}^b_{j,n}|A^b_n| \ \forall j \in A^*, \tag{4.112}$$

since for $\gamma = 1$ the empirical quantile is the value for which all other elements of the set $\left\{\frac{1}{\gamma}P^1_{j,n},\dots,\frac{1}{\gamma}P^B_{j,n}\right\} = \left\{P^1_{j,n}\dots,P^B_{j,n}\right\}$ are smaller (or equal). Since this equality (4.112) holds $\forall\, j \in A^*$, it also holds for the maximum, which exists by the sparsity assumption, thus

$$\max_{j\in A^*} q_{1,j} = \max_{j\in A^*} \max_{b\in\{1,\dots,B\}} \tilde{P}^b_{j,n}|A^b_n|. \tag{4.113}$$

Combining (4.110), (4.113) and $Q_j(1) = q_{1,j}$, one can infer

$$\begin{aligned} \lim_{n\to\infty} \mathbb{P}\left(\max_{j\in A^*} Q_j(1) \leq \alpha_n\right) &= \lim_{n\to\infty} \mathbb{P}\left(\max_{j\in A^*} q_{1,j} \leq \alpha_n\right) \qquad (4.114)\\ &\overset{(4.113)}{=} \lim_{n\to\infty} \mathbb{P}\left(\max_{j\in A^*} \max_{b\in\{1,\dots,B\}} \tilde{P}^b_{j,n}|A^b_n| \leq \alpha_n\right)\\ &\overset{(4.110)}{=} 1, \qquad (4.115) \end{aligned}$$

which is used further below. Moreover, it holds

$$\begin{aligned} \inf_{\gamma\in(\gamma_{min},1)} Q_j(\gamma) &\overset{\text{(Def.)}}{=} \inf_{\gamma\in(\gamma_{min},1)} \frac{1}{\gamma} q_\gamma\left(\{\tilde{P}^b_{j,n}|A^b_n|, b = 1,\dots,B\}\right)\\ &\overset{(4.112)}{\leq} \inf_{\gamma\in(\gamma_{min},1)} \frac{1}{\gamma} q_{1,j} = q_{1,j} \inf_{\gamma\in(\gamma_{min},1)} \frac{1}{\gamma} = q_{1,j} = Q_j(1). \end{aligned}$$

The steps above yield

$$\inf_{\gamma\in(\gamma_{min},1)} Q_j(\gamma) \leq Q_j(1) \leq \max_{j\in A^*} Q_j(1)$$

such that if $\max_{j\in A^*} Q_j(1) \leq \alpha_n$, this also holds for $Q_j(1)$ and for $\inf Q_j(\gamma)$, but the other direction(s) does not hold in general. Consequently

$$\left\{\max_{j\in A^*} Q_j(1) \leq \alpha_n\right\} \subseteq \{Q_j(1) \leq \alpha_n\} \subseteq \left\{\inf_{\gamma\in(\gamma_{min},1)} Q_j(\gamma) \leq \alpha_n\right\},$$

thus

$$\mathbb{P}\left(\max_{j\in A^*} Q_j(1) \leq \alpha_n\right) \leq \mathbb{P}\left(Q_j(1) \leq \alpha_n\right) \leq \mathbb{P}\left(\inf_{\gamma\in(\gamma_{min},1)} Q_j(\gamma) \leq \alpha_n\right). \tag{4.116}$$

Hence, since the inequality (4.116) also holds for the limit, this yields in combination with (4.115)

$$\lim_{n\to\infty} \mathbb{P}\left(\inf_{\gamma\in(\gamma_{min},1)} Q_j(\gamma) \leq \alpha_n\right) = 1 \quad \forall j \in A^*. \tag{4.117}$$

Multiplying by the factor $1-\log(\gamma_{min})$ on both sides in (4.117) and using the definition of Q_j^{adap} provides the following

$$\begin{aligned}
& \lim_{n\to\infty} \mathbb{P}\left(Q_j^{adap} \leq (1-\log(\gamma_{min}))\cdot\alpha_n\right) \\
= \quad & \lim_{n\to\infty} \mathbb{P}\left(\underbrace{(1-\log(\gamma_{min}))}_{\geq 1}\cdot \inf_{\gamma\in(\gamma_{min},1)} Q_j(\gamma) \leq (1-\log(\gamma_{min}))\cdot\alpha_n\right) \\
\leq \quad & \lim_{n\to\infty} \mathbb{P}\left(\inf_{\gamma\in(\gamma_{min},1)} Q_j(\gamma) \leq (1-\log(\gamma_{min}))\cdot\alpha_n\right) \\
\overset{(4.117)}{=} \quad & 1, \text{ since } (1-\log(\gamma_{min}))\alpha_n \to 0 \text{ at the same rate as } \alpha_n.
\end{aligned}$$

Finally (4.111) is valid and the proof is continued.

One further deduces for a sequence $\alpha_n \to 0$ and $j \in A^*$ (in particular one defines $\tilde{\alpha}_n := \alpha_n \cdot (1-\log(\gamma_{\min})) \to 0$ and, instead of $\tilde{\alpha}_n$, one continues to write α_n for simplicity of notation)

$$\lim_{n\to\infty} \mathbb{P}\left(Q_j^{adap} \leq \alpha_n\right) = 1 \quad \forall j \in A^*. \tag{4.118}$$

It is noted that to ensure consistency in the single split (Theorem 4.1.16), it is necessary that α_n converges to zero at a particular rate, such that consistency is ensured in the sense of (4.29). However, the rate at which $\tilde{\alpha}_n$ converges to zero if it embarks from α_n multiplied by a constant, that is $\tilde{\alpha}_n := \alpha_n \cdot (1-\log(\gamma_{\min}))$, is the same. Thus, for the single split case, (4.29) is also satisfied replacing α_n by $\tilde{\alpha}_n$ in (4.30).

Having shown the equation (4.118), one proceeds as in the single split case (Theorem 4.1.16). From Theorem 4.2.4 it is known that it holds

$$\limsup_{n\to\infty} \mathbb{P}\left(A_{n,f,mult}(\alpha_n) \cap (A^*)^c \neq \emptyset\right) \leq \alpha_n,$$

i.e. for $\alpha_n \to 0$ this yields $\limsup_{n\to\infty} \mathbb{P}\left(A_{n,f,mult}(\alpha_n) \cap (A^*)^c \neq \emptyset\right) = 0$, and consequently $\mathbb{P}\left(A_{n,f,mult}(\alpha_n) \subseteq A^*\right) \to 1$. Hence, it remains to show that $\mathbb{P}\left(A^* \subseteq A_{n,f,mult}(\alpha_n)\right) \to 1$. It is noted that $A_{n,f,mult}(\alpha_n)$ is based on the aggregated p-values Q_j^{adap}. As in Theorem 4.1.16, for $j \in A^*$ it holds that

$$\mathbb{P}\left(j \notin A_{n,f,mult}(\alpha_n)\right) \leq \mathbb{P}\left(Q_j^{adap} > \alpha_n\right) = 1 - \mathbb{P}\left(Q_j^{adap} \leq \alpha_n\right).$$

Consequently, with (4.118) one can infer for all $j \in A^*$ that

$$\lim_{n\to\infty} \mathbb{P}\left(j \notin A_{n,f,mult}(\alpha_n)\right) = 0,$$

which yields

$$\lim_{n\to\infty} \mathbb{P}\left(A^* \subseteq A_{n,f,mult}(\alpha_n)\right) = 1$$

and the claim of Theorem 4.2.5 follows. □

4.2.7 FDR Control with Benjamini Yekutieli Correction for Multiple Split

Having investigated FWER control and selection consistency for two-stage L_0-FGL in the multiple split case with Bonferroni correction, the next step is to analyze asymptotic properties applying Benjamini Yekutieli correction. To be more precise, Theorem 4.2.6 provides FDR control with Benjamini Yekuteli correction using aggregated p-values from the multiple sample splitting approach resulting in the final selected set (4.70).

Theorem 4.2.6 (FDR control for two-stage L_0-FGl in multiple split with Benjamini Yekutieli, fixed p and diverging p_n)**.** It is assumed that $\tilde{q} > 0$ and $A_{n,f}^{(BY)}(q)$ is given as in (4.70) with $q := \tilde{q}\left(\sum_{i=1}^{J} \frac{1}{i}\right)^{-1}$. Then for the two-stage L_0-FGL it holds that

$$\limsup_{n\to\infty} \mathbb{E}(W(q)) \leq \tilde{q}, \tag{4.119}$$

hence the FDR of two-stage L_0-FGL in the multiple split case can be controlled with Benjamini Yekutieli correction by $\tilde{q}$.

Proof. Adapting for two-stage L_0-FGL, the idea of the following proof roots in Meinshausen et al. (2009) (cf. Remark 4.1.17). Since this theorem holds for p being fixed and p_n being allowed to diverge, in this proof J and p, instead of J_n and p_n is written for simplicity of notation. The proof works similar to the proof of Theorem 4.1.18, while here the aggregated p-values Q_j^{adap} are used instead of $P_{j,n}$. Since the multiple split case is considered, the total number of hypotheses being tested is given by $|(A^*)^c|$, instead of $|(A^*)^c \cap A_n|$ as in the single split. Hence, as in the proof of Theorem 4.1.18 one knows that

$$\mathbb{E}(W(q)) = \sum_{i\in(A^*)^c} \sum_{k=1}^{J} \sum_{j=1}^{k} \frac{1}{k} p_{ijk} \leq \ldots \text{ (proof of Theorem 4.1.18) } \ldots = \sum_{j=1}^{J} \frac{1}{j} f(j).$$

Additionally, as deduced in the proof Theorem 4.1.18

$$\sum_{j=1}^{J-1} \left(\frac{1}{j} - \frac{1}{j+1}\right) \sum_{j'=1}^{j} f(j') + \frac{1}{J} \sum_{j'=1}^{J} f(j') = \sum_{j=1}^{J} \frac{1}{j} f(j),$$

which results in

$$\mathbb{E}(W(q)) \leq \sum_{j=1}^{J-1} \left(\frac{1}{j} - \frac{1}{j+1}\right) \sum_{j'=1}^{j} f(j') + \frac{1}{J} \sum_{j'=1}^{J} f(j').$$

Further, again executing the same steps as in the proof of Theorem 4.1.18, writing $i \in (A^*)$ instead of $i \in (A^*)^c \cap A_n$ and Q_i^{adap} instead of $P_{i,n}$, it holds

$$f(j) = \sum_{i\in(A^*)^c} \mathbb{P}(Q_i^{adap} \in [(j-1)q, jq]). \tag{4.120}$$

As a next step, (4.105) is used to show

$$\sum_{i\in(A^*)^c} \mathbb{P}(Q_i^{adap} \leq jq) \leq jq + o(1). \tag{4.121}$$

Since in the derivation of (4.105), it was used that $\alpha < 1$ and it cannot be ensured that $jq < 1$, one has to execute some steps to clarify why (4.121) holds. Clearly, for $jq < 1$, (4.121) follows

directly. Now one considers $jq \geq 1$. As explained previously, the function $\min(1, \cdot)$ can be omitted for simplicity since it does not affect the test decision. With that, it can be assumed that $Q_i^{adap} \leq 1$. From (4.105) it is known that (4.121) holds for $j = 1$ since then $jq < 1$ (since $q \in (0, 1)$). Proceeding by induction over $j \in \mathbb{N}$, one assumes that (4.121) holds and shows (4.121) for $j + 1$.

$$\begin{aligned}
\sum_{i \in (A^*)^c} \mathbb{P}(Q_i^{adap} \leq (j+1)q) &= \sum_{i \in (A^*)^c} \left\{ \mathbb{P}(Q_i^{adap} \leq jq) + \mathbb{P}(jq < Q_i^{adap} \leq (j+1)q) \right\} \\
&= \sum_{i \in (A^*)^c} \mathbb{P}(Q_i^{adap} \leq jq) \qquad (4.122) \\
&\leq jq + o(1) \quad \text{(by induction assumption)} \\
&\leq (j+1)q + o(1),
\end{aligned}$$

where in (4.122) it was used that $jq \geq 1$ and $Q_i^{adap} \leq 1$, hence $\mathbb{P}\left(jq < Q_i^{adap} \leq (j+1)q\right) = 0$. It is noted, once again, that the term $o(1)$ appearing above and below emerges due to the fact that the distribution of the p-values is only known asymptotically similar to the proofs presented previously. Consequently (4.121) holds and the proof continues.
Next, by (4.120) and (4.121) one obtains

$$\begin{aligned}
\sum_{j'=1}^{j} f(j') &= \sum_{j'=1}^{j} \sum_{i \in (A^*)^c} \mathbb{P}(Q_i^{adap} \in [(j'-1)q, j'q]) \\
&= \sum_{i \in (A^*)^c} \mathbb{P}(Q_i^{adap} \leq jq) \leq jq + o(1).
\end{aligned}$$

This finally yields

$$\begin{aligned}
\mathbb{E}(W(q)) &\leq \sum_{j=1}^{J-1} \left(\frac{1}{j} - \frac{1}{j+1} \right) \underbrace{\sum_{j'=1}^{j} f(j')}_{\leq jq + o(1)} + \frac{1}{J} \underbrace{\sum_{j'=1}^{J} f(j')}_{\leq Jq + o(1)} \\
&\leq \sum_{j=1}^{J-1} \left(\frac{1}{j} - \frac{1}{j+1} \right) (jq + o(1)) + \frac{1}{J} Jq + o(1) \\
&\overset{(*)}{=} \left(\sum_{j=1}^{J-1} \frac{1}{j+1} q \right) + q + o(1) = q \sum_{j=1}^{J} \frac{1}{j} + o(1) = \tilde{q} + o(1),
\end{aligned}$$

where in $(*)$ it was used that

$$\sum_{j=1}^{J-1} \left(\frac{1}{j} - \frac{1}{j+1} \right) o(1) = \sum_{j=1}^{J-1} \left(\frac{1}{j(j+1)} \right) o(1) \leq \sum_{j=1}^{J} \left(\frac{1}{j^2} \right) o(1) = o(1)$$

since $\sum_{j=1}^{J} \left(\frac{1}{j^2} \right)$ is finite, even if $J = J_n$ grows with n.

To sum up, it was shown that

$$\limsup_{n \to \infty} \mathbb{E}(W(q)) \leq \tilde{q}$$

and the claim follows. □

Chapter 5

Conclusion and Further Research

In this last chapter of the thesis, a summary of the main results is provided, along with an outline of possible further research directions corresponding to this topic.

5.1 Summary of Main Results

In **Chapter 1**, an overview of existing methods suitable for penalized logistic regression, especially incorporating categorical explanatory variables, i.e. those performing factor selection and/or levels fusion was supplied. First, theoretical properties, such as ($\sqrt{n}$) consistency, selection or fusion consistency, and asymptotic normality were transferred from the linear model case to logistic regression, if possible. Second, it was asserted that the fusion-type penalties (only) perform indirect selection tasks, whereas this selection ignores the underlying groupwise structure. Additionally, the penalties designed for factor selection, i.e. group lasso, group SCAD and group MCP do not perform levels fusion. These observations were underlined in simulation studies, incorporating five different simulation designs. These issues motivated the introduction of a new penalty function for penalized logistic regression, especially for categorical explanatory variables, giving the possibility to perform factor selection and levels fusion *simultaneously*.

Caused by the observations described in the paragraph above, a new penalty function, called L_0-FGL was introduced in **Chapter 2**, combining a group lasso penalty for factor selection with an L_0 penalty on the differences for levels fusion. Under suitable regularity conditions, asymptotic properties, such as $\sqrt{n}$ consistency, selection consistency and asymptotic normality were shown for the case of p being fixed, as well as for the case where p_n is allowed to diverge with n. Further, also for both cases, the fusion performance of L_0-FGL was specified by analyzing the behavior of the objective function when fusion occurs, as well as the resulting impact on the local minimizers, hence on the L_0-FGL estimates. Moreover, the existence of a fusion threshold was obtained, i.e. if the difference of two entries of a coefficient vector undercuts this fusion threshold, these two entries will be fused by L_0-FGL. Hence, it was intensively analyzed how fusion in L_0-FGL works, where selection properties such as factor selection consistency are shown to be valid at the same time, along with $\sqrt{n}$ consistency and asymptotic normality.

Then, **Chapter 3** focused on the computational approaches to obtain L_0-FGL in practice. For this, the PIRLS algorithm and a BCD algorithm were applied to L_0-FGL. Coefficient paths were allocated for both algorithms applied to CAS-L_0, L_0-FGL and group lasso, which underlined the characteristics of L_0-FGL combining these two approaches. Finally, the provided simulation studies showed that that L_0-FGL fulfills the purpose for which it was introduced, i.e. performing factor selection as well as levels fusion simultaneously, balancing both selection and fusion tasks.

Finally, **Chapter 4** investigated statistical inference for L_0-FGL. Introducing a two-stage L_0-FGL, first for single sample splitting and then for the extension using multiple sample splitting, the difficulty of statistical inference in (high-dimensional) penalized regression was conquered.

In particular, ensuring equality of conditional and nominal type-I-error in sample splitting, the need of conditioning on a particular selected model was eliminated. The existing single and multiple sample splitting approaches were transferred to L_0-FGL. In particular, it was investigated how the screening properties need to be adjusted if one allows for fusion and what type of tests one needs to execute when testing whether a factor is influential, especially if the dimension of the corresponding coefficient vector depends on the fusion by L_0-FGL that is executed beforehand. Considering two different types of multiplicity corrections, the Bonferroni and Benjamini Yekutieli correction, FWER control and selection consistency was shown for the first one, while for the latter FDR control was provided.

Overall, the new introduced penalty function was intensively investigated both from the theoretical, as well as from the computational perspective. The next section provides an outline of possible further research directions linked to L_0-FGL.

5.2 Outline of Further Research Directions

Computational Investigation of Two-Stage L_0-FGL

The novel L_0-FGL regularization technique was extended to two-stage L_0-FGL in Chapter 4, enabling statistical inference. Since the theoretical foundation of two-stage L_0-FGL was provided in the mentioned chapter, it would be interesting to study its performance in simulation studies as a next step. By the nature of two-stage L_0-FGL, the regularization step is not executed once, but $B \in \mathbb{N}$ times, each including the described stepwise or iterative CV procedure for the two-dimensional vector of tuning parameters. It is well-known that optimization problems including an L_0 norm are computationally challenging, as is the CV procedure of a two-dimensional tuning parameter, which both needs to be performed for *one* execution of L_0-FGL. Consequently, it is expected that the computation of two-stage L_0-FGL is computationally very demanding. To decrease the computational costs, further research on computational simplifications linked to two-stage L_0-FGL could be advantageous.

Local and Global Minimizers of the L_0-FGL Objective Function

Since the L_0-FGL penalty function is neither convex nor concave, minimizer(s) of the objective function $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ are not guaranteed to be unique and they are allowed to be local minimizers as well as global ones, as elaborated in this thesis. It is recalled that this fact similarly applies to other approaches employing a non-convex penalty function, e.g. SCAD, MCP or Bridge. The provided theoretical results for L_0-FGL in Chapter 2 have shown that there exists, at least one, local minimizer of $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ satisfying a certain property. However, it is desirable to investigate whether possible other (local) minimizers of $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ satisfy similar theoretical properties. In particular, it would be interesting to investigate whether the (multiple) local minimizers of $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ may share some (distributional) properties. For constrained M-estimators, this was done by Geyer (1994) requiring, among other conditions, the so-called *Clarke regularity* of the constrained set. It would be beyond the scope of this thesis to go into more details here, but it would be interesting to investigate whether the provided theorems could be applied in some sense to L_0-FGL. Furthermore, it is desirable to analyze how to obtain this local minimizer for which the corresponding theoretical properties are shown, as investigated in Fan et al. (2014) for SCAD and MCP. Another reference studying local optima in penalized regression is Loh and Wainwright (2017). In general, this is very difficult to analyze as there are papers only considering this specific topic. It is expected that the fact that

the L_0 norm is not continuous will be challenging, which requires further investigations on the properties of the objective function.

Tuning and Computation of L_0-FGL Estimates

In this thesis, a PIRLS and a BCD algorithm were implemented, using a stepwise and an iterative tuning approach. The simulation studies showed that these are both convenient approaches, while the advantages and disadvantages were analyzed. It could be advantageous to apply the tuning approach over a two-dimensional grid considering more possible pairs (λ_0, λ_1), which, however, makes the method much more computationally involving. It is noted that in the classical penalized regression approaches where only one tuning parameter is used, CV is clearly less demanding from the computational point of view. To ensure that CV on a two-dimensional grid is advantageous and does not "explode" with respect to computational time, one could try to further decrease the computational time of the considered algorithms, or even develop other algorithms which are faster. However, it is commonly known that optimization problems including an L_0 norm are very demanding and complex to solve, i.e. further work in this direction could be beneficial, as similarly mentioned in the first paragraph of this section.

Appendix A

Algorithms

Appendix A covers the topic of how minimization problems in penalized regression can be solved. Of course, it depends on the considered penalty function which algorithm is suitable to solve the resulting minimization problem. In particular, the PIRLS algorithm is presented in Appendix A.1, while coordinate descent approaches, especially block coordinate descent, is provided in Appendix A.2. For the exact application of these algorithms, one refers to the corresponding computational subsections provided for each penalty approach.

A.1 Penalized Iteratively Re-weighted Least Squares (PIRLS)

The PIRLS algorithm (Oelker and Tutz (2013)), where PIRLS is an abbreviation of "Penalized Iteratively Re-weighted Least Squares", is a penalized version of the well-known IRLS algorithm. The IRLS algorithm is suitable for calculating the unpenalized MLE of the parameter vector $\boldsymbol{\beta}$, thus minimizing the negative log-likelihood function $-L_n(\boldsymbol{\beta})$. To do so, IRLS applies a Newton algorithm (cf. Fan et al. (2020) Section 5.1.4 and Hastie et al. (2009) Section 4.4.1). To be more precise, starting at some initial value $\boldsymbol{\beta}^{(0)}$ (e.g. $\boldsymbol{\beta}^{(0)} = \mathbf{0}$ as recommended in Hastie et al. (2009)), one updates the current value $\boldsymbol{\beta}^{(k)}$ at iteration step $k \in \mathbb{N}$ as follows

$$\boldsymbol{\beta}^{(k)} = \boldsymbol{\beta}^{(k-1)} - \left(\boldsymbol{H}(\boldsymbol{\beta}^{(k-1)})\right)^{-1} \boldsymbol{s}(\boldsymbol{\beta}^{(k)}), \tag{A.1}$$

where $\boldsymbol{\beta}^{(k-1)}$ denotes the value at the previous iteration step $k-1$. Here,

$$\boldsymbol{H}(\boldsymbol{\beta}) := \frac{-\partial^2 L_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}, \quad \boldsymbol{s}(\boldsymbol{\beta}) := -\frac{\partial L_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}},$$

are the Hessian matrix of the negative log-likelihood function and the score function, respectively. In particular, one obtains by straightforward calculations as in (1.15) in the case of considering a canonical link function

$$\begin{aligned} \boldsymbol{s}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{1}{a_i(\phi)} \left(y_i \boldsymbol{x}_i - \varphi'(\boldsymbol{x}_i \boldsymbol{\beta}) \boldsymbol{x}_i\right), \\ \boldsymbol{H}(\boldsymbol{\beta}) &= -\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} = \boldsymbol{I}_F(\boldsymbol{\beta}), \end{aligned}$$

where $\boldsymbol{W}$ is given by (1.16). It is noted that in logistic regression, it holds $a_i(\phi) = 1 \, \forall i \in \{1 \ldots, n\}$, for which one refers to Remark 1.1.12.

Remark A.1.1. The updating equation (A.1) is executed until the algorithm (i.e. IRLS) converges, as described in Fan et al. (2020) (Section 5.1.4). For canonical link functions, one can ensure that the Hessian matrix $\boldsymbol{H}(\boldsymbol{\beta})$ is non-negative definite, such that the function to be minimized $-L_n(\boldsymbol{\beta})$ is concave. However, for strict concavity ensuring the uniqueness of a global minimizer, it is needed that the Hessian matrix is strictly positive definite, for which it is further required that $\boldsymbol{X}$ is of full rank, i.e. $\operatorname{rank}(\boldsymbol{X}) = p + 1$ hence $n \geq p + 1$, where reference is made to Remark 1.1.10. In Hastie et al. (2009) (Section 4.4.1), the authors also comment on the fact that one cannot ensure convergence of IRLS in general.

To obtain the *penalized* MLE, *penalized* versions of the Hessian matrix and the score are needed. However, one cannot ensure that every penalty function $P_\lambda(\boldsymbol{\beta})$ is twice (continuously) differentiable, so one cannot directly build derivatives of $M_{pen}(\boldsymbol{\beta})$, which consist of $P_\lambda(\boldsymbol{\beta})$ and $L_n(\boldsymbol{\beta})$. To overcome this issue, PIRLS employs a local quadratic approximation (LQA) of the penalty function $P_\lambda(\boldsymbol{\beta})$. Thus, the main idea of the PIRLS algorithm is to apply a Newton algorithm to the LQA (specified below), which is updated in every iteration step. Before further details on PIRLS are given, the general form of the penalty functions for which the PIRLS algorithm is applicable is provided.

Structure of the Penalty Function

The following explanations according to the PIRLS algorithm can be found in Oelker and Tutz (2013). One considers $L \in \mathbb{N}$ penalty functions $p_l : \mathbb{R}^+ \to \mathbb{R}^+$ satisfying $p_l(0) = 0$ and one supposes that $p_l(\cdot)$ is continuously differentiable on $\mathbb{R}^+$ with positive derivative, $l \in \{1, \dots, L\}$. Further, $p_l(\zeta)$ is assumed to be strictly monotonic in $\zeta \in \mathbb{R}^+$. It is noted that $p_l(\cdot)$ should not be mixed up with p_j introduced at the beginning of this thesis (denoting the number of levels of factor j). For every of these L penalty functions $p_l(\cdot)$, a separate tuning parameter $\lambda_l \geq 0$ is allowed. The penalty functions $p_l(\cdot)$ are applied to some function of $\boldsymbol{\beta}$, in particular to $||\boldsymbol{a}_l^T\boldsymbol{\beta}||_{N_l}$, where $||\cdot||_{N_l}$ is assumed to be a semi-norm, or at least some term that makes sense to be used as a penalty. 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 forms the differences of adjacent coefficients. An example to clarify these quantities is given below (Example A.1.2).

Finally, the PIRLS algorithm can be applied to penalty functions of the following form

$$P_\lambda^{gen}(\boldsymbol{\beta}) = \sum_{l=1}^{L} \lambda_l p_l(||\boldsymbol{a}_l^T\boldsymbol{\beta}||_{N_l}). \tag{A.2}$$

Most of the time, as explained in Oelker and Tutz (2013), the penalties are of the form

$$P_\lambda^{gen}(\boldsymbol{\beta}) = \sum_{j=1}^{J} \sum_{l=1}^{L_j} \lambda_{jl} p_{jl}(||\boldsymbol{a}_{jl}^T\boldsymbol{\beta}_j||_{N_l}) \tag{A.3}$$

meaning that each covariate $j \in \{1, ..., J\}$ is penalized separately with L_j penalty functions per covariate. Keeping this in mind, the more compact way of writing (A.2) is used where the parameter L combines on the one hand the different number of penalizations (for each factor) and on the other hand the separate penalizations for each covariate.

Example A.1.2. For CAS-L_0 (Section 1.3.2), the following choices are made

$$\begin{aligned} &p_l(\zeta) = w_0^{(j,km)}\zeta, \\ &\boldsymbol{a}_l^T\boldsymbol{\beta} = \beta_{j,k} - \beta_{j,m} \ \text{ for } 0 \leq k < m \leq p_j, \end{aligned}$$

where in the last equality, the vectors $\boldsymbol{a}_l^T$ are responsible for picking up the possible differences corresponding to factor j. The entries of these vectors $\boldsymbol{a}_l$ are contained in the set $\{-1, 0, 1\}$.

As underlined in Example A.1.2 above, the vectors $\boldsymbol{a}_l^T$ produce the differences corresponding to factor j. An extension of the linear transformations $\boldsymbol{a}_l^T\boldsymbol{\beta}$ to vector valued arguments, which is needed in Chapters 2, 3 and 4 applying this to group lasso, leads to the corresponding transformation matrices $\boldsymbol{R}_l$. For group lasso, this matrix $\boldsymbol{R}_l$ picks the right sub-vector $\boldsymbol{\beta}_j$ out of the full vector $\boldsymbol{\beta}$. Reference is made to Oelker and Tutz (2013) (Section 2.5) for the detailed steps

of the explained extension.

Finally, the (semi-)norms $||\cdot||_{N_l}$ appearing in the penalty function P_λ^{gen} above need to be specified. For simplicity, $||\cdot||_{N_l}$ is called "norm " even though, as e.g. for CAS-L_0 mentioned earlier, these quantities do not need to be real norms. These norms $||\cdot||_{N_l}$ 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 $D_l(\cdot)$. Considering penalties with vector valued arguments, the function $N_l(\boldsymbol{\xi})$ depends on a *vector* $\boldsymbol{\xi} \in \mathbb{R}^p$ instead of $\xi \in \mathbb{R}$.

Example A.1.3. For the approximation of the L_0 norm $||\cdot||_0 = ||\cdot||_{N_l}$, following Oelker and Tutz (2013), the following choices are made

$$||\xi||_0 \approx N_l(\xi) = \frac{2}{1+\exp(-\gamma|\xi|)} - 1\,, \tag{A.4}$$

$$D_l(\xi) = \frac{2\gamma}{1+\exp(-\gamma|\xi|)}\left(1 - \frac{1}{1+\exp(-\gamma|\xi|)}\right) \cdot \frac{\xi}{\sqrt{\xi^2+c}}. \tag{A.5}$$

In practice, the absolute value is further approximated $|\xi| \approx \sqrt{\xi^2+c}$ in $N_l(\xi)$ and $D_l(\xi)$ for some (small) $c > 0$ (as done in (A.5) in the last multiplicative factor). Since in CAS-L_0, the same penalty is applied to each factor (or difference of levels, respectively) one can also write $N(\cdot)$ instead of $N_l(\cdot)$, the same applies to $D_l(\cdot)$. For more details on the approximation used for CAS-L_0, including explanations on the choices for γ and c and a visualization one refers to Section 1.3.3.

To conclude, the necessary quantities contained in the penalty functions for which the PIRLS algorithm can be applied were discussed. Thus, the next step is to take a closer look at the local quadratic approximations (LQA) of the penalty function.

LQA of the Penalty Function

The arguments of this subsection follow Oelker and Tutz (2013) and Oelker et al. (2014b). Similar to Fan and Li (2001) (LQA for SCAD), the first step is to obtain a quadratic approximation of $P_\lambda^{gen}(\boldsymbol{\beta})$ (A.2) at some $\hat{\boldsymbol{\beta}}^{(k)}$. The resulting LQA is given by

$$P_\lambda^{gen}(\boldsymbol{\beta}) \approx P_\lambda^{gen}(\hat{\boldsymbol{\beta}}^{(k)}) + \frac{1}{2}\left(\boldsymbol{\beta}^T\boldsymbol{A}_\lambda\boldsymbol{\beta} + \hat{\boldsymbol{\beta}}^{(k)}\boldsymbol{A}_\lambda(\hat{\boldsymbol{\beta}}^{(k)})^T\right), \tag{A.6}$$

where

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

For further details on the derivation of these quantities, reference is made to Oelker and Tutz (2013) and Oelker et al. (2014b). Now, based on (A.6), penalized versions of the Hessian and the score, as well as the Fisher information matrix, can be obtained building the derivatives of $M_{pen}(\boldsymbol{\beta})$ replacing $P_\lambda(\boldsymbol{\beta})$ in (1.26) by the LQA given in (A.6). In particular, the following equations are deduced

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

It is recalled that $\boldsymbol{I}_F(\boldsymbol{\beta}) = -\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$ is the unpenalized Fisher information matrix and $\boldsymbol{W}$ is given by (1.16). The update of matrix $\boldsymbol{W}$ at the current iteration step k, denoted by $\boldsymbol{W}^{(k)}$, is derived by replacing $\boldsymbol{\beta}$ in $\boldsymbol{W}$ by $\hat{\boldsymbol{\beta}}^{(k)}$. Thus, one defines

$$\boldsymbol{W}^{(k)} := \operatorname{diag}\left(\frac{\varphi''(\boldsymbol{x}_1\hat{\boldsymbol{\beta}}^{(k)})}{a_1(\phi)}, \ldots, \frac{\varphi''(\boldsymbol{x}_n\hat{\boldsymbol{\beta}}^{(k)})}{a_n(\phi)}\right), \tag{A.7}$$

$$\tilde{\boldsymbol{y}}^{(k)} := (\boldsymbol{W}^{(k)})^{-1}(\boldsymbol{y} - \boldsymbol{\mu}^{(k)}) + \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(k)}, \tag{A.8}$$

where $\tilde{\boldsymbol{y}}^{(k)}$ is called the *working response*. Further, the current estimate of the conditional mean of the random response variable $\boldsymbol{\mu}^{(k)}$ is given by

$$\boldsymbol{\mu}^{(k)} = \left(g(\boldsymbol{x}_1\hat{\boldsymbol{\beta}}^{(k)}), \ldots, g(\boldsymbol{x}_n\hat{\boldsymbol{\beta}}^{(k)})\right).$$

Remark A.1.4 (Specific equations for logistic regression)**.** For logistic regression, $\boldsymbol{\mu}^{(k)}$ is given by

$$\boldsymbol{\mu}^{(k)} = \left(\frac{\exp(\boldsymbol{x}_1\hat{\boldsymbol{\beta}}^{(k)})}{1+\exp(\boldsymbol{x}_1\hat{\boldsymbol{\beta}}^{(k)})}, \ldots, \frac{\exp(\boldsymbol{x}_n\hat{\boldsymbol{\beta}}^{(k)})}{1+\exp(\boldsymbol{x}_n\hat{\boldsymbol{\beta}}^{(k)})}\right) = \frac{\exp(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(k)})}{1+\exp(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(k)})}.$$

Further, it holds by Remark 1.1.12 that $\varphi(\theta) = \log(1+\exp(\theta))$, such that for every $i = 1, \ldots, n$

$$\varphi''(\boldsymbol{x}_i\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{(1+\exp(\boldsymbol{x}_i\boldsymbol{\beta}))^2} = \frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i\boldsymbol{\beta})}\left(1 - \frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i\boldsymbol{\beta})}\right).$$

With these equations, one can write the matrix $\boldsymbol{W}^{(k)}$ for the logistic regression setting alternatively as $\boldsymbol{W}^{(k)} = \operatorname{diag}(\boldsymbol{\mu}^{(k)}(\boldsymbol{1} - \boldsymbol{\mu}^{(k)}))$.

Now, all quantities and approximations are available to obtain the PIRLS algorithm. In particular, as in the unpenalized case one receives at iteration step k, replacing the (penalized) Hessian by the (penalized) Fisher information, the following updating equation

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

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

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

according to Oelker et al. (2014b). Finally, the detailed steps of the PIRLS algorithm are given in Algorithm A.1.5, where the quantities maxsteps $\in \mathbb{N}$ and $\varepsilon \in \mathbb{R}^{>0}$ are pre-chosen values controlling the termination of the algorithm.

Algorithm A.1.5 (PIRLS, Oelker et al. (2014b))**.**

(i) One sets the start value $\hat{\boldsymbol{\beta}}^{(0)} = 0$, if not specified otherwise. One sets $k = 1$ and assigns ϵ and maxsteps.

(ii) While $\dfrac{||\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}^{(k-1)}||_2}{||\hat{\boldsymbol{\beta}}^{(k)}||_2} > \varepsilon$ and $k \leq$ maxsteps

(ii.i) One updates approximation of M_{pen} including updates of $\boldsymbol{A}_\lambda, \boldsymbol{W}, \tilde{\boldsymbol{y}}$ as they depend on the current value of the coefficient, that is, $\hat{\boldsymbol{\beta}}^{(k-1)}$

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

(iii) Finally, one sets $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(k)}$.

Convergence and Possible Limitations of PIRLS

As IRLS in the unpenalized case (Remark A.1.1), the PIRLS algorithms has some limitations, as the algorithm may not always converge. In particular, by arguments presented in Oelker and Tutz (2013) which are summarized below, the convergence of PIRLS depends on the penalized Fisher information matrix $\boldsymbol{I}_{F,pen}$. If it can be ensured that $\boldsymbol{I}_{F,pen}$ is strictly positive definite, the convergence of the algorithm, independently of the starting point, is guaranteed, since in this case the optimization problem is strictly convex. Now, the question that naturally arises is in which cases $\boldsymbol{I}_{F,pen}$ might be strictly positive definite. To investigate that, a distinction between the cases $n < p$ and $n \geq p$ is needed.

(i) $n \geq p$: in this case it is known that, since by assumption the response distribution belongs to the exponential dispersion family, the unpenalized Fisher information matrix $\boldsymbol{I}_F$ is strictly positive definite if $\boldsymbol{X}$ is of full rank (Remark 1.1.10), consequently the penalty matrix $\boldsymbol{A}_\lambda$ has to be at least positive semi-definite to ensure that $\boldsymbol{I}_{F,pen}$ is strictly positive definite so ensuring convergence of the algorithm.

(ii) $n < p$: analogously to (i) above, the unpenalized Fisher information matrix $\boldsymbol{I}_F$ is positive semi-definite (Remark 1.1.10 (ii)) and thus the penalty matrix $\boldsymbol{A}_\lambda$ has to be strictly positive definite to ensure that $\boldsymbol{I}_{F,pen}$ is strictly positive definite.

If it cannot be ensured that $\boldsymbol{I}_{F,pen}$ is strictly positive definite, it may happen that the algorithm finds non-unique descent directions. One can see above that, as expected, the convergence of PIRLS clearly depends on the chosen penalty function. Further, for the case that $n < p$, there are more restrictions that the penalty matrix has to fulfill.

A.2 Coordinate Descent

The general idea of the coordinate descent (CD) algorithm is introduced, before discussing the block coordinate descent (BCD) algorithm.

As elaborated in Hastie et al. (2015) (Section 5.4.1), the application of CD is convenient for objective functions for which the penalty part has an additive decomposition, that is, if one can write

$$P_\lambda(\boldsymbol{\beta}) = \sum_{i=1}^{p} h_i(\beta_i),$$

for some suitable functions $h_i : \mathbb{R} \to \mathbb{R}$, $i = 1, \ldots, p$. This property is also referred to as *separability property* and generalizes in a canonical way to the case of performing *block* coordinate descent. As an example, the penalty function of SCAD, that is $P_\lambda^{(\text{SCAD})}(\boldsymbol{\beta})$, given by (1.58), clearly satisfies the separability property, whereas the fused lasso with penalty function

$$\lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p} |\beta_i - \beta_{i-1}|,$$

introduced by Tibshirani et al. (2005), does not, due to the (pairwise) differences.

It is recalled that the goal is to find a minimizer of an objective function $M_{pen}(\boldsymbol{\beta})$, where $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$. Coordinate descent is an iterative procedure, which cycles through all coordinates of $\boldsymbol{\beta}$ (that is β_i for $i = 1, \ldots, p$ without the intercept) performing a univariate minimization with

respect to one coordinate β_i, keeping all others β_j, $j \in \{1, \ldots, p\} \backslash \{i\}$ fixed. The intercept is updated separately employing the unpenalized solution. It is assumed that the current value at iteration step $k \in \mathbb{N}$ is $\hat{\boldsymbol{\beta}}^{(k)} = (\hat{\beta}_{int}^{(k)}, \hat{\beta}_1^{(k)}, \ldots, \hat{\beta}_p^{(k)})$, whereby for simplicity, the factor-wise notation $\boldsymbol{\beta} = (\beta_{int}, \beta_{1,1}, \ldots, \beta_{1,p_1}, \ldots, \beta_{J,1}, \ldots, \beta_{J,p_J}) = (\beta_{int}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J)$ is neglected *once* for these explanations about CD. The update for the next iteration step for coordinate β_i for some $i = 1, \ldots, p$ is given by

$$\hat{\beta}_i^{(k+1)} = \underset{\beta_i \in \mathbb{R}}{\arg\min}\, M_{pen}(\hat{\beta}_{int}^{(k)}, \hat{\beta}_1^{(k)}, \ldots, \hat{\beta}_{i-1}^{(k)}, \beta_i, \hat{\beta}_{i+1}^{(k)}, \ldots, \hat{\beta}_p^{(k)}).$$

Clearly, cycling through the coordinates from the first to the last one, the update would be as follows

$$\hat{\beta}_i^{(k+1)} = \underset{\beta_i \in \mathbb{R}}{\arg\min}\, M_{pen}(\hat{\beta}_{int}^{(k+1)}, \hat{\beta}_1^{(k+1)}, \ldots, \hat{\beta}_{i-1}^{(k+1)}, \beta_i, \hat{\beta}_{i+1}^{(k)}, \ldots, \hat{\beta}_p^{(k)}).$$

CD is applied in the literature for lasso (e.g. Friedman et al. (2010)), as well as for non-convex penalties such as SCAD and MCP (Breheny and Huang (2011)). Since CD procedures, as well as BCD procedures which are discussed next, are very general algorithms, not only applied in the penalized regression framework, the detailed forms of the algorithms are provided in the corresponding sections about the method to which CD (and BCD, respectively) is applied. Thus, for details on the concrete application of CD for SCAD and MCP in the logistic regression framework, as well as the concrete form of the algorithm, it is referred to Section 1.5.3.

It is not only possible to minimize with respect to one coordinate at a time, but also with respect to (non-overlapping) blocks of coordinates, which leads to the BCD algorithm, being a suitable choice for minimization problems with a groupwise structure, e.g. group lasso.

Block Coordinate Descent

Similar to the CD algorithm, the BCD algorithm, which is also called "group coordinate descent", cycles through *blocks* of coordinates and minimizes the objective function with respect to this block of coordinates, keeping all others fixed. The groups/blocks need to be non-overlapping. One considers as blocks the groupwise structure naturally induced by the factors, thus the blocks are of sizes p_j, $j \in \{1, \ldots, J\}$. Having that, one minimizes the objective function with respect to one *factor* at a time, keeping all other factors fixed. As mentioned above introducing CD, the BCD approach is suitable for penalties satisfying the separability property, adjusted for the fact that separability is needed with respect to the considered blocks.

Returning to the factor-wise notation for $\boldsymbol{\beta}$, i.e.

$$\boldsymbol{\beta} = (\beta_{int}, \beta_{1,1}, \ldots, \beta_{1,p_1}, \ldots, \beta_{J,1}, \ldots, \beta_{J,p_J}) = (\beta_{int}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J),$$

the update at iteration step k for factor $j \in \{1, \ldots, J\}$ is given as follows

$$\hat{\boldsymbol{\beta}}_j^{(k+1)} = \underset{\boldsymbol{\beta}_j \in \mathbb{R}^{p_j}}{\arg\min}\, M_{pen}(\hat{\beta}_{int}^{(k+1)}, \hat{\boldsymbol{\beta}}_1^{(k+1)}, \ldots, \hat{\boldsymbol{\beta}}_{j-1}^{(k+1)}, \boldsymbol{\beta}_j, \hat{\boldsymbol{\beta}}_{j+1}^{(k)}, \ldots, \hat{\boldsymbol{\beta}}_J^{(k)}). \tag{A.9}$$

Group lasso, group SCAD and group MCP clearly satisfy the separability property with respect to factors, thus BCD is a convenient choice for these penalties. In practice Breheny and Huang (2015) apply BCD to these penalty approaches, where for the block-wise updates (A.9), they obtain closed form solutions for group lasso, group MCP and group SCAD. Section 1.5.3 gives more details on the application of this algorithm to group lasso, group MCP and group SCAD in the logistic regression framework.

Appendix B

Regularity Conditions

Regularity conditions which are needed throughout the whole thesis are obtained in this segment. First, regularity conditions emerging from classical ML theory are given in Appendix B.1 for the case of fixed p, which are transferred to the case of diverging p_n in Appendix B.2. These two sections of Appendix B can also be found in Kaufmann and Kateri (2024). In Appendix B.3 some continuity assumptions imposed on the log-likelihood function are discussed. Finally, Appendix B.4 introduces a regularity condition on the nature of (local) minimizers of $M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$ for increasing sample size n.

Thus, a collection of regularity conditions is provided in the following, however, they are not all imposed throughout the whole thesis. Whenever some of the following regularity conditions are imposed, it is clearly specified.

B.1 Fixed p Case

For the case of p being fixed, one considers the following regularity conditions (Reg1)-(Reg3).

(Reg1) The distribution of the response variable Y belongs to the exponential dispersion family (Definition 1.1.2).

(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

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

These regularity conditions are similar to several other approaches, such as Fan and Li (2001) and Zou (2006), being necessary for technical derivations, as well as to ensure the asymptotic normality of the classical MLE, cf. Appendix C.

In (Reg3), it is assumed that there exists an open set $\mathcal{O} \subseteq \mathbb{R}^{p+1}$ such that for the true value it holds that $\boldsymbol{\beta}^* \in \mathcal{O}$. One may ask the question what happens if the truth lies at the boundary of the parameter space, which is answered next following the arguments of Lehmann and Casella (1998) (Section 10.6). The proofs of the theorems providing consistency and asymptotic normality of the MLE (Lehmann and Casella (1998) Theorems 3.7 and 3.10 of Section 6) use this regularity condition in the following way: knowing that the truth is an interior point of

the parameter space, it holds that for increasing n, the MLE is a root of the derivative of the likelihood function. However, if the truth is at the boundary of the parameter space, this can obviously not be ensured as the maximum occurs at the boundary. As explained in Lehmann and Casella (1998) (Section 10.6), in this case it may be possible to provide other asymptotic distributions, as supplied e.g. in Self and Liang (1987) and Andrews (1999). An approach considering an enlarged parameter space is discussed in Feng and McCulloch (1992). However, in this thesis, this regularity condition is not a restriction, as the parameter space is $\mathbb{R}^{p+1}$.

Remark B.1.1 (On the regularity conditions (Reg1)-(Reg3) under the canonical link)**.** Under the use of the canonical link function, the natural parameter θ equals the linear predictor $\boldsymbol{x}\boldsymbol{\beta}$, which is also denoted by η, i.e. $\eta = \boldsymbol{x}\boldsymbol{\beta}$. In this case, one receives the following simplifications.

(i) The expected and observed Fisher information matrices coincide, so from (Reg1) it can be inferred

$$\begin{aligned} \boldsymbol{I}_F(\boldsymbol{\beta}) &= \mathbb{E}\left(-\frac{\partial^2 L_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}\right) = -\frac{\partial^2 L_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} \\ &= \boldsymbol{X}^T \operatorname{diag}\left(\varphi''(\boldsymbol{x}_1\boldsymbol{\beta}), \ldots, \varphi''(\boldsymbol{x}_n\boldsymbol{\beta})\right) \boldsymbol{X}. \end{aligned} \tag{B.1}$$

It is recalled that $\boldsymbol{x}_i$, $i \in \{1, \ldots, n\}$ are the rows of the design matrix $\boldsymbol{X}$ and $\eta_i = \boldsymbol{x}_i\boldsymbol{\beta}$, $i = 1, \ldots, n$. Consequently, (Reg2) is satisfied if (B.1) is finite and positive definite in $\boldsymbol{\beta} = \boldsymbol{\beta}^*$.
Even though logistic regression is considered in this thesis, i.e. the canonical link is employed, whenever possible, the expected Fisher information is used to keep the theory more general and easily applicable for other link functions.

(ii) The second and third derivatives of the log-likelihood function $L_n(\boldsymbol{\beta})$ do not further depend on y. Consequently, (Reg3) is ensured if there exists a function $\mathfrak{M}(\boldsymbol{x}_i)$ such that

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

It is clear that the functions $\mathfrak{M}(\cdot)$ and $M(\cdot)$ (one compares (Reg3)) are not the same, nevertheless it is neglected to write $\mathfrak{M}$ in the proofs, hence $M(\cdot)$ is written for simplicity of notation. The actual structure of the function plays no role, it is sufficient that here exists such a function.

These simplifications for canonical link functions can also be found in Fan and Li (2001) (Section 3.2).

B.2 Diverging p_n Case

For the diverging p_n case, one assumes the following regularity conditions. In this scenario, the dimension of the parameter space (which is $1 + p_n$) is allowed to grow with the sample size n, thus the dimension of the Fisher information matrix also depends on n, as well as the dimension of the open set $\mathcal{O}$ appearing in (div.Reg3). Clearly, the dimension of the truth $\boldsymbol{\beta}^*$ also depends on n in this case. Nevertheless, to keep notation clear, a lower index n for the truth $\boldsymbol{\beta}^*$ is omitted. For all of the constants in the following regularity conditions C is written, even if they are not required to be the same.

(div.Reg1) The distributional assumption for Y is the same as in (Reg1) (Section B.1) with p being replaced by p_n and the corresponding pdf denoted by f_n instead of f.

(div.Reg2) The Fisher information matrix $\boldsymbol{I}_{F,n}(\boldsymbol{\beta})$ satisfies the same as in (Reg2) (Section B.1) with $\boldsymbol{I}_F$ being replaced by $\boldsymbol{I}_{F,n}$.

(div.Reg3) There exists an open set $\mathscr{O}_n \subseteq \mathbb{R}^{p_n+1}$ with $\boldsymbol{\beta}^* \in \mathscr{O}_n$ such that for all $\boldsymbol{\beta} \in \mathscr{O}_n$ 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

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

Additionally, one assumes that for some constant $C < \infty$ it holds

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

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

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

It is recalled 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 $\mathscr{X}_j$, 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 Fan and Peng (2004) and Wang and Tian (2019). For simplifications using the canonical link, as done here considering logistic regression, it is referred to Remark B.1.1 which similarly applies in the diverging case.

Remark B.2.1. The above regularity conditions and their consequences are discussed next.

i) Alternatively to (div.Reg2) one could have also assumed that all the eigenvalues of the Fisher information matrix are finite and strictly positive which ensures the positive definite property, cf. Fan and Peng (2004).

ii) The fact that in (div.Reg2), and similarly in (Reg2), it was assumed that the Fisher information matrix is finite means in particular that for a constant $C > 0$ it holds that $[\boldsymbol{I}_{F,n}(\boldsymbol{\beta})]^2_{j,k} < C < \infty \; \forall\, j,k = 1, ..., p_n$ and

$$[\boldsymbol{I}_{F,n}(\boldsymbol{\beta})]_{j,k} = \mathbb{E}\left(-\frac{\partial^2 \log(f_n(\boldsymbol{v}_1|\boldsymbol{\beta}))}{\partial\beta_j\partial\beta_k}\right) < C.$$

iii) Condition (div.Reg4) is a technical condition needed for the oracle property of the (approximate) L_0-FGL which is shown in Theorem 2.3.7. This condition is taken from Fan and Peng (2004). Since a dummy coding scheme is used in this thesis, 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. Consequently $||\boldsymbol{x}_{j,i}||_2 = 1$ independent of $j \in \{1, \ldots, J_n\}$, $i \in \{1, \ldots, n\}$, so observing factors in a dummy coding scheme as done here, (div.Reg4) is not a restriction.

B.3 Log-Likelihood

One considers the following continuity assumptions for the log-likelihood, that are imposed in some of the theorems of this thesis.

Definition B.3.1. $L_n(\boldsymbol{\beta})$ is defined as in Definition 1.1.9. The following cases are considered:

(i) Fixed p, fixed n
The continuity assumption 1, (Cont1) for short, is said to be satisfied if the following holds $\forall \boldsymbol{\zeta}_0 \in \mathbb{R}^{p+1}$ and $\forall n \in \mathbb{N}$

$$\forall \varepsilon > 0 \, \exists \, \delta = \delta(\varepsilon, \boldsymbol{\zeta}_0, n) > 0 \; : \; \forall \boldsymbol{\zeta} \in \mathbb{R}^{p+1}, \, ||\boldsymbol{\zeta}_0 - \boldsymbol{\zeta}||_2 < \delta \; : \; ||L_n(\boldsymbol{\zeta}_0) - L_n(\boldsymbol{\zeta})||_2 < \varepsilon, \tag{B.2}$$

thus if the log-likelihood function $-L_n(\cdot)$ is continuous for every $n \in \mathbb{N}$ and in every $\boldsymbol{\zeta}_0 \in \mathbb{R}^{p+1}$. It is noted that δ may depend on ε, on $\boldsymbol{\zeta}_0$ as well as on the function $-L_n(\cdot)$, in particular on n.

(ii) Fixed p, diverging n
Let $\mathscr{F}$ be the family of functions $\mathscr{F} := \{-L_n(\boldsymbol{\beta}), n \in \mathbb{N}\}$. The continuity assumption 2, (Cont2) for short, is said to be satisfied if the following holds $\forall \boldsymbol{\zeta}_0 \in \mathbb{R}^{p+1}$

$$\begin{aligned} \forall \varepsilon > 0 \, \exists \, \delta(\varepsilon, \boldsymbol{\zeta}_0) > 0 \; : \; \forall \boldsymbol{\zeta} \in \mathbb{R}^{p+1}, \, ||\boldsymbol{\zeta}_0 - \boldsymbol{\zeta}||_2 < \delta \\ \text{and} \, \forall \, \text{functions} - L_n(\cdot) \in \mathscr{F} \; : \; ||L_n(\boldsymbol{\zeta}_0) - L_n(\boldsymbol{\zeta})||_2 < \varepsilon, \end{aligned}$$

thus, if the family $\mathscr{F}$ is equicontinuous $\forall \boldsymbol{\zeta}_0 \in \mathbb{R}^{p+1}$. Compared to (Cont1), where δ may depend on n, ε and $\boldsymbol{\zeta}_0$ (and a function $-L_n(\cdot)$ for *fixed* n is observed and not a family of functions), in (Cont2), δ may only depend on ε and $\boldsymbol{\zeta}_0$. Thus, this "rate of continuity" δ, is the same for all functions in the family $\mathscr{F}$.

(iii) Diverging p_n, diverging n
Suppose that $\mathscr{F}$ is defined as in (ii). If p_n diverges with n, the dimension of the definition space of $-L_n(\boldsymbol{\beta})$ grows. However, for every $\boldsymbol{\zeta}_0 \in \mathbb{R}^{p_n+1}$ only those $\boldsymbol{\zeta} \in \mathbb{R}^{p_n+1}$ are considered that coincide with $\boldsymbol{\zeta}_0$ in every entry except for one entry, without loss of generality entry $r \in \{1, \dots p_n + 1\}$ such that $\boldsymbol{\zeta}_0 - \boldsymbol{\zeta} = (0, \dots, 0, \zeta_{0,r} - \zeta_r, 0, \dots, 0)$ and $||\boldsymbol{\zeta}_0 - \boldsymbol{\zeta}|| = |\zeta_{0,r} - \zeta_r|$. Thus, this case can be transferred to (ii). The continuity assumption 3, (Cont3) for short, is said to be satisfied if the following holds $\forall \boldsymbol{\zeta}_0 \in \mathbb{R}^{p_n+1}$

$$\begin{aligned} \forall \varepsilon > 0 \, \exists \, \delta(\varepsilon, \boldsymbol{\zeta}_0) > 0 \; : \; \forall \boldsymbol{\zeta} \in \mathbb{R}^{p_n+1}, \, ||\boldsymbol{\zeta}_0 - \boldsymbol{\zeta}||_2 = |\zeta_{0,r} - \zeta_r| < \delta \\ \text{and} \, \forall \, \text{functions} - L_n(\cdot) \in \mathscr{F} \; : \; ||L_n(\boldsymbol{\zeta}_0) - L_n(\boldsymbol{\zeta})||_2 < \varepsilon. \end{aligned}$$

The next remark obtains an expansion to random variables.

Remark B.3.2. Since (Cont1)-(Cont3) are provided for deterministic values $\boldsymbol{\zeta}$, these conditions are also applied to random variables. $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$ are supposed to be two real-valued random variables. Then, the following holds

(i) If (Cont1) is ensured, then it holds $\forall n \in \mathbb{N}$

$$\forall \epsilon > 0 \, \exists \delta = \delta(\epsilon, \boldsymbol{\beta}_0, n) > 0 \; : \; \text{if} \, \mathbb{P}(||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_2 < \delta) = 1, \; \text{then} \, \mathbb{P}(|L_n(\boldsymbol{\beta}) - L_n(\boldsymbol{\beta}_0)| < \epsilon) = 1.$$

(ii) If (Cont2) is ensured, then it holds

$$\begin{aligned} \forall \varepsilon > 0 \, \exists \, \delta(\varepsilon, \boldsymbol{\zeta}_0) > 0 \; : \; \text{if} \, \mathbb{P}(||\boldsymbol{\beta}_0 - \boldsymbol{\beta}||_2 < \delta) = 1 \\ \text{then it holds} \; \forall \, \text{functions} - L_n(\cdot) \in \mathscr{F} \; : \; \mathbb{P}(|L_n(\boldsymbol{\beta}_0) - L_n(\boldsymbol{\beta})| < \varepsilon) = 1. \end{aligned}$$

(iii) Analogously to (ii), an implication of (Cont3) for random variables can be formulated.

B.4 Local Minimizers

For some theorems concerning fusion properties, the following regularity condition is needed, which can be formulated for fixed p and diverging p_n in the same manner, while below it is formulated for fixed p. For diverging p_n, the condition is similar replacing p by p_n. Considering sequences below, one writes *round brackets* around these sequences and add the sequence index *outside of the round brackets*, which is further explained and introduced in Notation 2.3.21. For now, no further details on this notation are necessary.

Definition B.4.1. One considers $((\boldsymbol{\zeta})_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}^{p+1}$ being a sequence of local minimizers of the objective function $((M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\zeta}))_n)_{n\in\mathbb{N}}$, meaning that $(\boldsymbol{\zeta})_n$ minimizes $(M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\zeta}))_n$ for $n \in \mathbb{N}$. Thus, for each $n \in \mathbb{N}$, there exists a neighborhood $\mathcal{N}_n \subseteq \mathbb{R}^{p+1}$ of $(\boldsymbol{\zeta})_n$ of radius $\epsilon_n > 0$ such that

$$(M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\zeta}))_n \leq (M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\zeta}))_n \;\; \forall\, \boldsymbol{\zeta} \text{ satisfying } \;\; ||\boldsymbol{\zeta} - (\boldsymbol{\zeta})_n|| < \epsilon_n.$$

The minimal neighborhood condition, for short (MinNeigh), is said to be satisfied if $\min_{n\in\mathbb{N}} \epsilon_n$ exists.
If $((\hat{\boldsymbol{\beta}})_n)_{n\in\mathbb{N}}$ is considered as a sequence of random variables being a sequence of local minimizers as above, it holds that

$$(M_{pen}^{(L_0\text{-FGL})}(\hat{\boldsymbol{\beta}}))_n \leq (M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta}))_n \;\; \forall\, \boldsymbol{\beta} \text{ satisfying } \;\; ||\boldsymbol{\beta} - (\hat{\boldsymbol{\beta}})_n|| < \epsilon_n,$$

where the difference is that ϵ_n may depend on the random sample, as does $(\hat{\boldsymbol{\beta}})_n$. Thus, in this case, it is required that the minimum of ϵ_n exists almost surely, that is $\mathbb{P}\left(\min_{n\in\mathbb{N}} \epsilon_n \text{ exists}\right) = 1$. It is not necessary to further specify the size or structure of ϵ_n since the actual value does not play any role, the existence of the minimum is sufficient.

The minimal neighborhood condition ensures that the radius of the neighborhood, with respect to which the local minimizer is a minimizer, does not converge to zero for increasing n. It is noted that, following Notation 2.3.21 introduced in Chapter 2, one can similarly write $[\cdot]_n$ instead of $(\cdot)_n$ above.

B.5 Discussion On Different Regularity Conditions

In the literature, theoretical properties for penalized regression, in the grand majority penalized *linear* regression, are investigated under a variety of different regularity conditions. These conditions include restricted eigenvalue conditions, restricted isometry properties, compatibility conditions, sparse Riesz conditions, as well as the irrepresentable condition (Huang et al. (2021)), to give some examples. However, which conditions need to be imposed highly depends on the considered setting, e.g. distributional assumptions, the underlying model assumption, as well as the penalty functions. Thus, these conditions are not comparable *in general* over different frameworks. For the Lasso considered in a linear model setting, a broad comparison of several conditions imposed in the literature is given by Geer and Bühlmann (2009). Another reference giving more details on these conditions for different penalty functions is Bühlmann and Geer (2011).

Besides the technical conditions that were presented above, throughout the *whole thesis* it is assumed that the true underlying model is sparse (Defintion 1.2.4), where this kind of sparsity assumption is also referred to as L_0 sparsity in the literature, e.g. in Zhang and Zhang (2012).

In particular, it is required that, departing from some p_n (in the diverging case), all *influential* factors are included in the model. However, one can impose other sparsity conditions, such as

$$||\boldsymbol{\beta}^*|| = o\left(\sqrt{\frac{n}{\ln(p_n)}}\right),$$

according to Bühlmann and Geer (2011) (Section 2.4.2).

To conclude, there are various ways to impose regularity conditions that lead to different theoretical properties. However, it is beyond the scope of this thesis to compare all these conditions for penalized logistic regression, since the comparisons not only depend on the underlying model and the distributional assumptions, but also on the considered penalty function. Consequently, this thesis focusses on the regularity conditions introduced above, which follow the references that are given at the respective point. But, one should keep in mind that one can impose alternative regularity conditions, stronger or weaker, that may lead to different, or similar, theoretical results.

Appendix C

Selection of General Results on MLE and LRT

Let Ω denote the parameter space of the parameter vector $\boldsymbol{\beta}$ to be estimated in the GLM framework introduced in Section 1.1.2, e.g. $\Omega = \mathbb{R}^{p+1}$. Starting with theoretical properties of the (unpenalized) MLE in Appendix C.1, the area of likelihood ratio tests (LRT) is covered in Appendix C.2.

C.1 Theoretical Properties of the Maximum Likelihood Estimator

First, Definition 1.1.9 of Chapter 1 is recalled for the definition of the MLE in a GLM framework. To be able to obtain asymptotic results for the MLE, one needs to impose some regularity conditions, given in Appendix B.1. Since the main source of the theorems given in this Appendix C is Casella and Berger (2002), the following remark on the regularity conditions of Appendix B.1 and those given in Casella and Berger (2002) (Section 10.6.2) is provided.

Remark C.1.1. The regularity conditions that are given in Appendix B.1 and Appendix B.2, respectively, ensure that (A1)-(A6) in Casella and Berger (2002) are satisfied. However, the conditions provided in Casella and Berger (2002) are more general, but for the purposes of this thesis (Reg1)-(Reg3) for fixed p ((div.Reg1)-(div.Reg3) for diverging p_n) are assumed, so that the conditions for testing are consistent with those used for selection/fusion and estimation.

Consistency and Asymptotic Normality

At first, a theorem treating consistency of the MLE is provided. Here, consistency refers to estimation consistency, meaning that the difference between the estimator $\hat{\boldsymbol{\beta}}$ and the true parameter vector $\boldsymbol{\beta}$ converges in probability to zero (cf. Section 1.2.5).

Theorem C.1.2 (Consistency of MLE, Casella and Berger (2002), Theorem 10.1.6)**.** One assumes that regularity conditions (Reg1)-(Reg3) of Appendix B.1 hold. Further, $\tau(\boldsymbol{\beta})$ is assumed to be a continuous function of $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}^{(\mathrm{ML})}$ its MLE (Definition 1.1.9). Then, for all $\varepsilon > 0$ and all $\boldsymbol{\beta} \in \Omega$ it holds that

$$\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\beta}}(|\tau(\hat{\boldsymbol{\beta}}^{(\mathrm{ML})}) - \tau(\boldsymbol{\beta})| \geq \varepsilon) = 0. \tag{C.1}$$

Consequently, $\tau(\hat{\boldsymbol{\beta}}^{(\mathrm{ML})})$ is a consistent estimator of $\tau(\boldsymbol{\beta})$.

The next theorem yields a result on the asymptotic distribution of the MLE, in particular it states asymptotic normality.

Theorem C.1.3 (Asymptotic normality of MLE, Casella and Berger (2002), Theorem 10.1.12)**.** It is assumed that (Reg1)-(Reg3) of Appendix B.1 hold, that $\tau(\boldsymbol{\beta})$ is a continuous function of $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}^{(\mathrm{ML})}$ its MLE (Definition 1.1.9). Then it holds that

$$\sqrt{n}(\tau(\hat{\boldsymbol{\beta}}^{(\mathrm{ML})}) - \tau(\boldsymbol{\beta})) \to_d N(\mathbf{0}, v(\boldsymbol{\beta})), \tag{C.2}$$

where

$$v(\boldsymbol{\beta}) := \frac{(\tau'(\boldsymbol{\beta}))^2}{\mathbb{E}\left(\frac{\partial}{\partial \boldsymbol{\beta}} f(y|\boldsymbol{\beta})\right)} \tag{C.3}$$

is the Cramér-Rao lower bound.

If $\tau(\cdot)$ is the identity function, i.e. $\tau(\boldsymbol{\beta}) = \boldsymbol{\beta}$, the Cramér-Rao lower bound is simply the inverse of the expected Fisher information matrix, which coincides with the observed Fisher information matrix when the canonical link is used, which is the case for logistic regression.

C.2 Likelihood Ratio Test

To evaluate the fit of a model, calculating the (predictive) deviance, introduced in Definition 1.2.1, is a convenient tool. For the purposes of this thesis, a comparison of the fit resulting from two *nested* models is of high interest, which is specified in the following. It is noted that $\mathscr{M}_0$ and $\mathscr{M}_1$ are defined as general GLMs at this moment, however, in Chapter 4 these are models resulting from L_0-FGL in penalized logistic regression.

$\mathscr{M}_0$ is assumed to be a GLM with $q_0 \in \mathbb{N}$ parameters (i.e. $\dim(\boldsymbol{\beta}) = q_0$) and $\mathscr{M}_1$ a GLM with $q_1 \in \mathbb{N}$, $q_0 < q_1$ parameters (i.e. $\dim(\boldsymbol{\beta}) = q_1$). In particular, it is assumed that $\mathscr{M}_0$ is a "simpler" GLM evoking from $\mathscr{M}_1$ by eliminating $r = q_1 - q_0$ of its parameters, thus it is said that $\mathscr{M}_0$ is *nested* in $\mathscr{M}_1$.

$\hat{\boldsymbol{\mu}}_0^{(\mathrm{ML})}$ denotes the MLE of $\boldsymbol{\mu}$ under model $\mathscr{M}_0$, i.e. $\hat{\boldsymbol{\mu}}_0^{(\mathrm{ML})}$ emerges from the MLE $\hat{\boldsymbol{\beta}}_0^{(\mathrm{ML})}$ using the underlying model equation, which is e.g. for logistic regression applying (1.20)

$$\hat{\boldsymbol{\mu}}_0^{(\mathrm{ML})} = \frac{\exp(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0^{(\mathrm{ML})})}{1 + \exp(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0^{(\mathrm{ML})})}. \tag{C.4}$$

Analogously, $\hat{\boldsymbol{\mu}}_1^{(\mathrm{ML})}$ denotes the MLE of $\boldsymbol{\mu}$ under model $\mathscr{M}_1$, where it is assumed that both $\hat{\boldsymbol{\beta}}_0^{(\mathrm{ML})}$ and $\hat{\boldsymbol{\beta}}_1^{(\mathrm{ML})}$ exist, which is obviously needed for further steps. Corresponding to Definition 1.2.1, one sets

$$D\left(\boldsymbol{y}|\hat{\boldsymbol{\mu}}_i^{(\mathrm{ML})}\right) := -2 \cdot \left(L_n(\hat{\boldsymbol{\mu}}_i^{(\mathrm{ML})} \,|\, \boldsymbol{y}) - L_n(\boldsymbol{y} \,|\, \boldsymbol{y})\right) \quad , i \in \{1, 2\}.$$

It is recalled that, since $\boldsymbol{X}$ is assumed fixed and $\boldsymbol{v} := (\boldsymbol{y}, \boldsymbol{X})$, one could equivalently write $D\left(\boldsymbol{v}|\hat{\boldsymbol{\mu}}_i^{(\mathrm{ML})}\right)$ instead of $D\left(\boldsymbol{y}|\hat{\boldsymbol{\mu}}_i^{(\mathrm{ML})}\right)$, favoring the latter for the sake of simplicity. Similar arguments apply to the log-likelihood function $L_n(\hat{\boldsymbol{\mu}}_i^{(\mathrm{ML})} \,|\, \boldsymbol{y})$ and the corresponding version of the saturated model $L_n(\boldsymbol{y}|\, \boldsymbol{y})$.

The difference of the deviances of models $\mathscr{M}_0$ and $\mathscr{M}_1$ is given by

$$\begin{aligned} T\left(\mathscr{M}_0, \mathscr{M}_1, n, \boldsymbol{y}\right) &:= D\left(\boldsymbol{y}|\hat{\boldsymbol{\mu}}_0^{(\mathrm{ML})}\right) - D\left(\boldsymbol{y}|\hat{\boldsymbol{\mu}}_1^{(\mathrm{ML})}\right) \\ &= -2\left(L_n(\hat{\boldsymbol{\mu}}_0^{(\mathrm{ML})}|\, \boldsymbol{y}) - L_n(\hat{\boldsymbol{\mu}}_1^{(\mathrm{ML})}|\, \boldsymbol{y})\right). \end{aligned} \tag{C.5}$$

The statistic $T(\mathscr{M}_0, \mathscr{M}_1, n, \boldsymbol{Y})$ is the *likelihood ratio statistic* (LRS) for testing

$$H_0 : \text{model } \mathscr{M}_0 \text{ holds} \quad \text{versus} \quad H_1 : \text{model } \mathscr{M}_1 \text{ holds},$$

compare Casella and Berger (2002) (Section 8.2.1). The next Theorem C.2.1 provides the asymptotic distribution of the LRS, which is a χ^2_r distribution with $r := q_1 - q_0$ degrees of freedom.

Theorem C.2.1 (Asymptotic distribution of likelihood ratio statistic, Casella and Berger (2002), Theorem 10.3.3)**.** One assumes that the regularity conditions (Reg1)-(Reg3) of Appendix B.1 hold and $\hat{\boldsymbol{\beta}}_0^{(\mathrm{ML})}$ and $\hat{\boldsymbol{\beta}}_1^{(\mathrm{ML})}$ are given as in Definition 1.1.9 for models $\mathscr{M}_0$ and $\mathscr{M}_1$, respectively. For testing

$$H_0\text{: model } \mathscr{M}_0 \text{ holds} \qquad \text{versus} \qquad H_1\text{: model } \mathscr{M}_1 \text{ holds}$$

as described above, under the null hypothesis H_0 it holds for $n \to \infty$

$$T(\mathscr{M}_0, \mathscr{M}_1, n, \boldsymbol{Y}) \to_d \chi_r^2 \tag{C.6}$$

with $r := q_1 - q_0$.

Now the question can be answered in which case the null hypothesis H_0 is rejected at a given *nominal level* $\alpha \in (0, 1)$ to ensure error control as in (C.7) below. In particular, the test decision is as follows

$$H_0 \text{ is rejected} \Leftrightarrow T\left(\mathscr{M}_0, \mathscr{M}_1, n, \boldsymbol{y}\right) \geq \chi_{r,\alpha}^2,$$

where $\chi_{r,\alpha}^2$ is defined as the *upper quantile* of the χ_r^2 distribution.

Following Casella and Berger (2002) (p. 490 f.), under the null hypothesis H_0, which is expressed by the lower index in the probability below, the type-I-error probability is asymptotically equal to α, the nominal level, i.e.

$$\lim_{n\to\infty} \mathbb{P}_{H_0}(\text{reject } H_0) = \alpha. \tag{C.7}$$

Notation

Part (I): General mathematical symbols

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of real-valued random variables on some probability space $(\Omega, \mathfrak{A}, \mathbb{P})$ and let X be a real-valued random variable on $(\Omega, \mathfrak{A}, \mathbb{P})$. Let F_n denote the cumulative distribution function (cdf) of X_n for $n \in \mathbb{N}$ and F the cdf and X, respectively. Further, let $(a_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}$ be a real-valued sequence.

Symbol	Description
$\mathbb{R}$	real numbers
$\mathbb{R}^k$	k-dimensional cartesian product of $\mathbb{R}$, $k \in \mathbb{N}$
$\mathbb{N}$	natural numbers
$1_{\{S\}}(\cdot)$	indicator function for some set $S \subseteq \mathbb{R}$
$\|\|\boldsymbol{x}\|\|_q$	L_q norm, that is $\|\|\boldsymbol{x}\|\|_q := \left(\sum_{i=1}^k x_i^q\right)^{1/q}$ for $\boldsymbol{x} \in \mathbb{R}^k$ and $q > 0$, for $q = 2$ one uses the abbreviation $\|\|\boldsymbol{x}\|\|$, see row below
$\|\|\boldsymbol{x}\|\|$	Euclidean norm of $\boldsymbol{x} \in \mathbb{R}^k$, that is $\|\|\boldsymbol{x}\|\| = \sqrt{\sum_{i=1}^k x_i^2}$
$\|\|\boldsymbol{x}\|\|_0$	L_0 norm, that is $\|\|\boldsymbol{x}\|\|_0 := \|\{x_i \neq 0,\, i \in \{1,\ldots,k\}\}\|$ where $\boldsymbol{x} = (x_1,\ldots,x_k) \in \mathbb{R}^k$
$(f(x))_+$	positive part of a real-valued function $f(x)$, hence $(f(x))_+ := \max(f(x), 0)$
$\mathbb{E}(X)$	expected value of a random variable X
$\mathbb{P}(A)$	probability of some event $A \in \mathfrak{A}$
$\Phi(\cdot)$	cdf of the standard normal distribution
$\text{rank}(\boldsymbol{M})$	rank of a matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$, $d \in \mathbb{N}$
$X \sim F$	X is distributed according to the cdf F
$\dot\cup$	disjoint union
$\exp(\cdot)$	exponential function
$\ln(\cdot)$	natural logarithm

(continued)

(continued)

Symbol	Description
Stochastic convergence/stochastic boundeness	
$X_n = O_p(a_n)$	$\forall \epsilon > 0 \;\; \exists\, M, N \in \mathbb{N}$ such that $\mathbb{P}\left(\left\lvert \frac{X_n}{a_n} \right\rvert > M\right) < \epsilon \;\; \forall\, n > N.$
$X_n = o_p(a_n)$	$\operatorname{p\,lim}_{n\to\infty} \frac{X_n}{a_n} = 0$, thus $\forall \epsilon > 0 : \lim_{n\to\infty} \mathbb{P}\left(\left\lvert \frac{X_n}{a_n} \right\rvert > \epsilon\right) = 0$
$X_n \to_d X$ or $X_n \to_d F$	$\lim_{n\to\infty} F_n(x) = F(x)\, \forall x \in \mathbb{R}$ where F is continous.
$X_n \to_p X$ or $\operatorname{p\,lim}_{n\to\infty} X_n = X$	$\forall \epsilon > 0 : \lim_{n\to\infty} \mathbb{P}(\lvert X_n - X \rvert > \epsilon) = 0$
$X_n \to_{a.s.} X$	$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1$
Distributions	
$\mathrm{Bin}(k, p)$	binomial distribution with $k \in \mathbb{N}$ independent Bernoulli trials and success probability $p \in (0, 1)$ for each trail
$\mathrm{Mult}(p_1, \ldots, p_k)$	multinomial distribution with success event probabilities $p_i \in (0, 1)$, $i \in \{1, \ldots, k\}$, $k \in \mathbb{N}$ and $\sum_{i=1}^{k} p_i = 1$
$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	(multi-dimensional) normal distribution with expectation $\boldsymbol{\mu} \in \mathbb{R}^k$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{k\times k}$, $k \in \mathbb{N}$
χ^2_r	χ^2 distribution with $r \in \mathbb{N}$ degrees of freedom

Part (II): Mathematical symbols in penalized regression/L_0-FGL

In what follows, a list of used symbols concerning penalized regression, especially L_0-FGL, is provided, ordered by their respective topics for clarity.

Symbol	Description
Estimators	
$\hat{\boldsymbol{\beta}}^{(\mathrm{ML})}$	maximum likelihood estimator of (full) dimension $p+1$
$\hat{\boldsymbol{\beta}}^{(\mathrm{ML})}_{\mathrm{red}}$	maximum likelihood estimator of (reduced) dimension $p^{af}+1$
$\left(\hat{\boldsymbol{\beta}}^{(\mathrm{ML})}\right)^{\mathrm{o}}$	oracle maximum likelihood estimator
$\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}$	L_0-FGL estimator of (full) dimension $p+1$
$\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}_{\mathrm{red}}$	L_0-FGL estimator of (reduced) dimension $p^{af}+1$
$\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}_{j}$	sub-vector corresponding to factor j of L_0-FGL estimator of (full) dimension p_j
$\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL})}_{j,\mathrm{red}}$	sub-vector corresponding to factor j of L_0-FGL estimator of (reduced) dimension p_j^{af}
$\hat{\boldsymbol{\beta}}$	penalized regression estimator (no penalty specified, see (1.27))
$\hat{\boldsymbol{\beta}}^{(\mathrm{CAS}-L_1)}$	CAS-L_1 estimator (dimension $p+1$)
$\hat{\boldsymbol{\beta}}^{(\mathrm{CAS}-L_0)}$	CAS-L_0 estimator (dimension $p+1$)
$\hat{\boldsymbol{\beta}}^{(\mathrm{SCAD})}$	SCAD estimator (dimension $p+1$)
$\hat{\boldsymbol{\beta}}^{(\mathrm{gSCAD})}$	group SCAD estimator (dimension $p+1$)
$\hat{\boldsymbol{\beta}}^{(\mathrm{MCP})}$	MCP estimator (dimension $p+1$)
$\hat{\boldsymbol{\beta}}^{(\mathrm{gMCP})}$	group MCP estimator (dimension $p+1$)
$\hat{\boldsymbol{\beta}}^{(\mathrm{GL})}$	group Lasso estimator (dimension $p+1$)
$\boldsymbol{\beta}^*$	true coefficient vector (dimension $p+1$)
$\boldsymbol{\beta}^*_{\mathrm{red}}$	true reduced coefficient vector (reduced dimension p^*+1)
$\boldsymbol{\beta}^*_{j}$	sub-vector corresponding to factor j of true coefficient vector (full dimension p_j)
$\boldsymbol{\beta}^*_{j,\mathrm{red}}$	sub-vector corresponding to factor j of true reduced coefficient vector (reduced dimension p_j^*)

(continued)

(continued)

Symbol	Description
Functions	
$l_n(\boldsymbol{\beta})$	likelihood function
$L_n(\boldsymbol{\beta})$	log-likelihood function
$M_{pen}(\boldsymbol{\beta})$	objective function
$P_\lambda(\boldsymbol{\beta})$	penalty function
$P_\lambda^{(\text{SCAD})}(\boldsymbol{\beta})$	SCAD penalty function
$P_\lambda^{(\text{Lasso})}(\boldsymbol{\beta})$	lasso penalty function
$P_\lambda^{(\text{Ridge})}(\boldsymbol{\beta})$	Ridge penalty function
$P_\lambda^{(\text{gSCAD})}(\boldsymbol{\beta})$	group SCAD penalty function
$P_\lambda^{(\text{MCP})}(\boldsymbol{\beta})$	MCP penalty function
$P_\lambda^{(\text{gMCP})}(\boldsymbol{\beta})$	group MCP penalty function
$P_\lambda^{(L_0\text{-FGL})}(\boldsymbol{\beta})$	L_0-FGL penalty function
$P_\lambda^{(\text{CAS-}L_1)}(\boldsymbol{\beta})$	CAS-L_1 penalty function
$P_\lambda^{(\text{CAS-}L_0)}(\boldsymbol{\beta})$	CAS-L_0 penalty function
$P_\lambda^{(\text{Bridge})}(\boldsymbol{\beta})$	Bridge penalty function
$M_{pen}^{(\text{SCAD})}(\boldsymbol{\beta})$	SCAD objective function
$M_{pen}^{(\text{gSCAD})}(\boldsymbol{\beta})$	group SCAD objective function
$M_{pen}^{(\text{MCP})}(\boldsymbol{\beta})$	MCP objective function
$M_{pen}^{(\text{gMCP})}(\boldsymbol{\beta})$	group MCP objective function
$M_{pen}^{(L_0\text{-FGL})}(\boldsymbol{\beta})$	L_0-FGL objective function
$M_{pen}^{(\text{CAS-}L_1)}(\boldsymbol{\beta})$	CAS-L_1 objective function
$M_{pen}^{(\text{CAS-}L_0)}(\boldsymbol{\beta})$	CAS-L_0 objective function
Active Sets	
C^*	true active set of fusions
C_n	estimated active set of fusions
F^*	truly influential factors, true fusions
F_n	truly influential factors, estimated (performed) fusions
A^*	true active set
A_n	estimated active set
FP/FN sets in simulation studies	
$FP_{f,\text{infl.truth}}$	false positive rate concerning fusion
$FN_{f,\text{infl.truth}}$	false negative rate concerning fusion
$FP_{s,\text{fac}}$	false positive rate concerning factor selection
$FN_{s,\text{fac}}$	false negative rate concerning factor selection

(continued)

(continued)

Symbol	Description
Two-Stage L_0-FGL: Single Split	
$\tilde{S}_n$	selected model by L_0-FGL in step 1
$\hat{\boldsymbol{\beta}}^{\tilde{S}_n}_{\text{red}}$	reduced MLE of step 2, dimension $p^{af}+1$
$\hat{\boldsymbol{\beta}}^{\tilde{S}_n}_{j,\text{red}}$	reduced coeff. sub-vector corresponding to $\hat{\boldsymbol{\beta}}^{\tilde{S}_n}_{\text{red}}$, dimension p_j^{af}
$\hat{\boldsymbol{\beta}}^{\tilde{S}_n}$	MLE of step 2, dimension $p+1$
$\hat{\boldsymbol{\beta}}^{\tilde{S}_n}_j$	coeff. sub-vector corresponding to $\hat{\boldsymbol{\beta}}^{\tilde{S}_n}$, dimension p_j
$\boldsymbol{\beta}^{\tilde{S}_n}$	coeff. vector of step 2, dimension $p+1$
$\boldsymbol{\beta}^{\tilde{S}_n}_j$	coeff. sub-vector corresponding to $\boldsymbol{\beta}^{\tilde{S}_n}$, dimension p_j
$\boldsymbol{\beta}^{\tilde{S}_n}_{\text{red}}$	reduced coeff. vector step 2, dimension $p^{af}+1$
$\boldsymbol{\beta}^{\tilde{S}_n}_{j,\text{red}}$	reduced coeff. sub-vector corresponding to $\boldsymbol{\beta}^{\tilde{S}_n}_{\text{red}}$, dimension p_j^{af}
Two-Stage L_0-FGL: Multiple Split, quantities for split $b \in \{1,\ldots,B\}$	
$\tilde{S}_n^b$	Selected model by L_0-FGL in step 1
F_n^b	truly influential factors, estimated (performed) fusions
A_n^b	estimated active set
$\hat{\boldsymbol{\beta}}^{\tilde{S}_n^b}_{\text{red}}$	reduced MLE of step 2, dimension $p^{af}+1$
$\hat{\boldsymbol{\beta}}^{\tilde{S}_n^b}_{j,\text{red}}$	reduced Coeff.sub-vector corresponding to $\hat{\boldsymbol{\beta}}^{\tilde{S}_n^b}_{\text{red}}$, dimension p_j^{af}
$\hat{\boldsymbol{\beta}}^{\tilde{S}_n^b}$	MLE of step 2, dimension $p+1$
$\hat{\boldsymbol{\beta}}^{\tilde{S}_n^b}_j$	coeff. sub-vector corresponding to $\hat{\boldsymbol{\beta}}^{\tilde{S}_n^b}$, dimension p_j
$\boldsymbol{\beta}^{\tilde{S}_n^b}$	coeff. vector of step 2, dimension $p+1$
$\boldsymbol{\beta}^{\tilde{S}_n^b}_j$	coeff. sub-vector corresponding to $\boldsymbol{\beta}^{\tilde{S}_n^b}$, dimension p_j
$\boldsymbol{\beta}^{\tilde{S}_n^b}_{\text{red}}$	reduced coeff. vector step 2, dimension $p^{af}+1$
$\boldsymbol{\beta}^{\tilde{S}_n^b}_{j,\text{red}}$	reduced coeff. sub-vector corresponding to $\boldsymbol{\beta}^{\tilde{S}_n^b}_{\text{red}}$, dimension p_j^{af}
$\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL}),b}$	L_0-FGL estimator of (full) dimension $p+1$
$\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL}),b}_{\text{red}}$	L_0-FGL estimator of (reduced) dimension $p^{af}+1$
$\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL}),b}_j$	sub-vector corresponding to factor j of L_0-FGL estimator of (full) dimension p_j
$\hat{\boldsymbol{\beta}}^{(L_0\text{-FGL}),b}_{j,\text{red}}$	sub-vector corresponding to factor j of L_0-FGL estimator of (reduced) dimension p_j^{af}

Part (III): Abbreviations

In what follows, a list of used abbreviations in alphabetical order is provided.

Abbreviation	Description
a.s.	Almost surely
Adap.	Adaptive
ANOVA	Analysis of variance
BCD	Block coordinate descent
BH	Benjamini Hochberg
Bin	Binomial distribution
BY	Benjamini Yekutieli
cdf	Cummulative distribution function
CLT	Central limit theorem
CS	Cauchy-Schwartz inequality
CV	Cross validation
EDF	Exponential dispersion family
FDR	False discovery rate
FN	False negative
FP	False positive
FWER	Family wise error rate
GLM	Generalized Linear Model
iid	Independent and identically distributed
It.	Iterative
IRLS	Iteratively re-weighted least squares
KKT	Karush-Kuhn-Tucker
L_0-FGL	L_0-fused group lasso
LARS	Least angle regression
LDP	Low dimensional projections
LLA	Local linear approximation
LLN	Law of large numbers
LM	Linear model

(continued)

(continued)

Abbreviation	Description
LogReg	Logistic regression
LOOCV	Leave one out cross validation
LQA	Local quadratic approximation
LRS	Likelihood ratio statistic
LRT	Likelihood ratio test
LS	Least squares
LSE	Least squares estimate/estimator
MCMC	Markov chain Monte Carlo
ML	Maximum likelihood
MLE	Maximum likelihood estimator/estimate
MSEC	Mean squared error of coefficients
Mult	Multinomial distribution
NEF	Natural exponential family
OS	Overall sparsity
PIRLS	Penalized iteratively reweighted least squares
pmf	Pobability mass function
Pred.	Predictive
PS	Practical sparsity
SRC	Sparse Riesz condition
Thm.	Theorem

List of Figures

List of Tables

References

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley & Sons.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. John Wiley & Sons.

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. Wiley.

Albert, A. and Anderson, J. A. (1984). "On the existence of maximum likelihood estimates in logistic regression models". In: *Biometrika* 71.1, pp. 1–10.

Andrews, D. W. K. (1999). "Estimation when a parameter is on a boundary". In: *Econometrica* 67.6, pp. 1341–1383.

Bach, F. R. (2008). "Consistency of the group lasso and multiple kernel learning". In: *Journal of Machine Learning Research* 9, pp. 1179–1225.

Benjamini, Y. and Hochberg, Y. (1995). "Controlling the false discovery rate: A practical and powerful approach to multiple testing". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 57.1, pp. 289–300.

Benjamini, Y. and Yekutieli, D. (2001). "The control of the false discovery rate in multiple testing under dependency". In: *The Annals of Statistics* 29.4, pp. 1165–1188.

Bondell, H. D. and Reich, B. J. (2009). "Simultaneous factor selection and collapsing levels in ANOVA". In: *Biometrics* 65.1, pp. 169–177.

Breheny, P. and Huang, J. (2009). "Penalized methods for bi-level variable selection". In: *Statistics and its interface* 2, pp. 369–380.

Breheny, P. and Huang, J. (2011). "Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection". In: *The Annals of Applied Statistics* 5.1, pp. 232–253.

Breheny, P. and Huang, J. (2015). "Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors". In: *Statistics and Computing* 25.2, pp. 173–187.

Bühlmann, P. (2013). "Statistical significance in high-dimensional linear models". In: *Bernoulli* 19.4, pp. 1212–1242.

Bühlmann, P. and Geer, S. van de (2011). *Statistics for High-Dimensional Data*. Springer.

Bühlmann, P. and Geer, S. van de (2015). "High-dimensional inference in misspecified linear models". In: *Electronic Journal of Statistics* 9.1, pp. 1449–1473.

Bunea, F. (2008). "Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization". In: *Electronic Journal of Statistics* 2, pp. 1153–1194.

Casella, G. and Berger, R. (2002). *Statistical Inference*. 2nd ed. Duxbury.

Chiquet, J., Grandvalet, Y., and Rigaill, G. (2016). "On coding effects in regularized categorical regression". In: *Statistical Modelling* 16, pp. 228–237.

Dahinden, C., Parmigiani, G., Emerick, M., and Bühlmann, P. (2006). *Sparse contingency tables and high-dimensional log-linear models for alternative splicing in full-length cDNA libraries*. Technical Report 132. ETH Zürich.

Dunn, O. J. (1961). "Multiple comparisons among means". In: *Journal of the American Statistical Association* 56.293, pp. 52–64.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). "Least angle regression". In: *The Annals of Statistics* 32.2, pp. 407–499.

Fan, J. and Li, R. (2001). "Variable selection via nonconcave penalized likelihood and its oracle properties". In: *Journal of the American Statistical Association* 96.456, pp. 1348–1360.

Fan, J., Li, R., Zhang, C.-H., and Zou, H. (2020). *Statistical Foundations of Data Science*. CRC Press.

Fan, J. and Lv, J. (2010). "A selective overview of variable selection in high dimensional feature space". In: *Statistica Sinica* 20.1, pp. 101–148.

Fan, J. and Lv, J. (2011). "Nonconcave penalized likelihood with NP-dimensionality". In: *IEEE Transaction on Information Theory* 57, pp. 5467–5484.

Fan, J. and Peng, H. (2004). "Nonconcave penalized likelihood with a diverging number of parameters". In: *The Annals of Statistics* 32.3, pp. 928–961.

Fan, J., Xue, L., and Zou, H. (2014). "Strong oracle optimality of folded concave penalized estimation". In: *The Annals of Statistics* 42.3, pp. 819–849.

Feng, Z. and McCulloch, C. E. (1992). "Statistical inference using maximum likelihood estimation and the generalized likelihood ratio when the true parameter is on the boundary of the parameter space". In: *Statistics & Probability Letters* 13.4, pp. 325–332.

Finner, H. and Roters, M. (2001). "On the false discovery rate and expected type I errors". In: *Biometrical Journal* 43.8, pp. 985–1005.

Fithian, W., Sun, D., and Taylor, J. (2017). "Optimal inference after model selection". In: *arXIv:1410.2597 [Statistics Theory]*, pp. 1–22.

Frank, l. and Friedman, J. (1993). "A statistical view of some chemometrics regression tools". In: *Technometrics* 35.2, pp. 109–135.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). "Regularization paths for generalized linear models via coordinate descent". In: *Journal of Statistical Software* 33.1.

Fu, W. J. (1998). "Penalized regressions: the bridge versus the lasso". In: *Journal of Computational and Graphical Statistics* 7.3, pp. 397–416.

Geer, S. van de, Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). "On asymptotically optimal confidence regions and tests for high-dimensional models". In: *The Annals of Statistics* 42.3, pp. 1166–1202.

Geer, S. A. van de and Bühlmann, P. (2009). "On the conditions used to prove oracle results for the Lasso". In: *Electronic Journal of Statistics* 3, pp. 1360–1392.

Gertheiss, J. and Tutz, G. (2010a). *Regularization and model selection with categorial effect modifiers*. Technical Report 73. LMU Munich.

Gertheiss, J. and Tutz, G. (2010b). "Sparse modeling of categorial explanatory variables". In: *Annals of Applied Statistics* 4.4, pp. 2150–2180.

Gertheiss, J. and Tutz, G. (2023). "Trends and challenges in categorical data analysis". In: ed. by M. Kateri and I. Moustaki. Springer. Chap. 7, pp. 199–232.

Geyer, C. J. (1994). "On the asymptotics of constrained M-estimation". In: *The Annals of Statistics* 22.4, pp. 1993–2010.

Guo, X., Zhang, H., Wang, Y., and Wu, J.-L. (2015). "Model selection and estimation in high dimensional regression models with group SCAD". In: *Statistics & Probability Letters* 103, pp. 86–92.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.

Hastie, T., Tibshirani, R., and Tibshirani, R. (2020). "Best subset, forward stepwise or lasso? Analysis and recommendations based on extensive comparisons". In: *Statistical Science* 35.4, pp. 579–592.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity. The Lasso and Generalizations*. Chapman and Hall/CRC.

Hoerl, A. E. and Kennard, R. W. (1970). "Ridge regression: Biased estimation for nonorthogonal problems". In: *Technometrics* 12, pp. 55–67.

Huang, J., Breheny, P., and Ma, S. (2012). "A selective review of group selection in high-dimensional models". In: *Statistical Science* 27.4, pp. 481–499.

Huang, J., Horowitz, J. L., and Ma, S. (2008). "Asymptotic properties of bridge estimators in sparse high-dimensional regression models". In: *The Annals of Statistics* 36.2, pp. 587–613.

Huang, J., Jiao, Y., Kang, L., and Liu, Y. (2021). "Fitting sparse linear models under the sufficient and necessary condition for model identification". In: *Statistics & Probability Letters* 168, p. 108925.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2014). *An Introduction to Statistical Learning: With Applications in R*. Springer.

Javanmard, A. and Montanari, A. (2014a). "Confidence intervals and hypothesis testing for high-dimensional regression". In: *Journal of Machine Learning Research* 15, pp. 2869–2909.

Javanmard, A. and Montanari, A. (2014b). "Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory". In: *IEEE Transactions on Information Theory* 60.10, pp. 6522–6554.

Jorgensen, B. (1997). *The Theory of Dispersion Models*. Chapman & Hall.

Kateri, M. (2014). *Contingency Table Analysis: Methods and Implementation Using R*. Birkhäuser.

Kaufmann, L. and Kateri, M. (2024). "Simultaneous factors selection and fusion of their levels in penalized logistic regression". In: *Electronic Journal of Statistics* 18.2, pp. 4235–4291.

Kim, Y., Choi, H., and Oh, H.-S. (2008). "Smoothly clipped absolute deviation on high dimensions". In: *Journal of the American Statistical Association* 103.484, pp. 1665–1673.

Kim, Y., Kim, J., and Kim, Y. (2006). "Blockwise sparse regression". In: *Statistica Sinica* 16, pp. 375–390.

Knight, K. and Fu, W. (2000). "Asymptotics for lasso-type estimators". In: *The Annals of Statistics* 28.5, pp. 1356–1378.

Land, S. R. and Friedman, J. H. (1996). *Variable fusion: A new adaptive signal regression method*. Technical Report 656. Department of Statistics, Carnegie Mellon University.

Lederer, J. (2022). *Fundamentals of High-Dimensional Statistics: With Exercises and R Labs*. Springer.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). "Exact post-selection inference, with application to the lasso". In: *The Annals of Statistics* 44.3, pp. 907–927.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. 2nd ed. Springer.

Lehmann and Romano (2005). *Testing Statistical Hypotheses*. 3rd ed. Springer.

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). "A significance test for the lasso". In: *The Annals of Statistics* 42.2, pp. 413–468.

Loh, P.-L. and Wainwright, M. (2013). "Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima". In: *Journal of Machine Learning Research* 16, pp. 559–616.

Loh, P.-L. and Wainwright, M. J. (2017). "Support recovery without incoherence: A case for nonconvex regularization". In: *The Annals of Statistics* 45.6, pp. 2455–2482.

Ma, R., Cai, T. T., and Li, H. (2020). "Global and simultaneous hypothesis testing for high-dimensional logistic regression models". In: *Journal of the American Statistical Association* 116.534, pp. 984–998.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd ed. Chapman & Hall.

Meier, L., Geer, S. van de, and Bühlmann, P. (2008). "The group lasso for logistic regression". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.1, pp. 53–71.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). "P-values for high-dimensional regression". In: *Journal of the American Statistical Association* 104, pp. 1671–1681.

Ming-Hui Chen, D. K. D. and Shao, Q.-M. (1999). "A new skewed link model for dichotomous quantal response data". In: *Journal of the American Statistical Association* 94.448, pp. 1172–1186.

Myers, R. H., Montgomery, D. C., Vining, G. G., and Robinson, T. J. (2010). *Generalized Linear Models. With Applications in Engineering and the Sciences*. 2nd ed. John Wiley & Sons, Inc.

Nardi, Y. and Rinaldo, A. (2008). "On the asymptotic properties of the group lasso estimator for linear models". In: *Electronic Journal of Statistics* 2, pp. 605–633.

Nardi, Y. and Rinaldo, A. (2012). "The log-linear group-lasso estimator and its asymptotic properties". In: *Bernoulli* 18.3, pp. 945–974.

Nelder, J. A. and Wedderburn, R. W. M. (1972). "Generalized linear models". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 135.3, pp. 370–384.

Ning, Y. and Liu, H. (2017). "A general theory of hypothesis tests and confidence regions for sparse high dimensional models". In: *The Annals of Statistics* 45.1, pp. 158–195.

Oelker, M.-R., Gertheiss, J., and Tutz, G. (2014a). "Regularization and model selection with categorical predictors and effect modifiers in generalized linear models". In: *Statistical Modelling* 14.2, pp. 157–177.

Oelker, M.-R., Pößnecker, W., and Tutz, G. (2014b). "Selection and fusion of categorical predictors with L0-type penalties". In: *Statistical Modelling* 15.5, pp. 389–410.

Oelker, M.-R. and Tutz, G. (2013). *A general family of penalties for combining differing types of penalties in generalized structured models*. Technical Report 139. LMU Munich.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). "Sparse additive models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.5, pp. 1009–1030.

Sarkar, S. K. (2002). "Some results on false discovery rate in stepwise multiple testing procedures". In: *The Annals of Statistics* 30.1, pp. 239–257.

Sarkar, S. K. (2008). "On methods controlling the false discovery rate". In: *Sankhyā: The Indian Journal of Statistics* 70-A.2, pp. 135–168.

Schultheiss, C., Renaux, C., and Bühlmann, P. (2021). "Multicarving for high-dimensional post-selection inference". In: *Electronic Journal of Statistics* 15.1, pp. 1695–1742.

Self, S. G. and Liang, K.-Y. (1987). "Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions". In: *Journal of the American Statistical Association* 82.398, pp. 605–610.

Stokell, B. G., Shah, R. D., and Tibshirani, R. J. (2021). "Modelling high-dimensional categorical data using nonconvex fusion penalties". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 83.3, pp. 579–611.

Taylor, J. and Tibshirani, R. (2018). "Post-selection inference for L1-penalized likelihood models". In: *The Canadian Journal of Statistics* 46.1, pp. 41–61.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 58.1, pp. 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). "Sparsity and smoothness via the fused lasso". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.1, pp. 91–108.

Tutz, G. (2011). *Regression for Categorical Data*. Cambridge University Press.

Wang, H. and Leng, C. (2008). "A note on adaptive group Lasso". In: *Computational Statistics & Data Analysis* 52, pp. 5277–5286.

Wang, L., Chen, G., and Li, H. (2007). "Group SCAD regression analysis for microarray time course gene expression data". In: *Bioinformatics* 23.12, pp. 1486–1494.

Wang, L., You, Y., and Lian, H. (2015). "Convergence and sparsity of lasso and group lasso in high-dimensional generalized linear models". In: *Statistical Papers* 56, pp. 819–828.

Wang, L., Li, H., and Huang, J. Z. (2008). "Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements". In: *Journal of the American Statistical Association* 103.484, pp. 1556–1569.

Wang, M. and Tian, G.-L. (2019). "Adaptive group Lasso for high-dimensional generalized linear models". In: *Statistical Papers* 60.5, pp. 1469–1486.

Wasserman, L. and Roeder, K. (2009). "High-dimensional variable selection". In: *The Annals of Statistics* 37.5A, pp. 2178–2201.

Wei, F. and Huang, J. (2010). "Consistent group selection in high-dimensional linear regression". In: *Bernoulli* 16.4, pp. 1369–1384.

Xie, H. and Huang, J. (2009). "SCAD-penalized regression in high-dimensional partially linear models". In: *The Annals of Statistics* 37.2, pp. 673–696.

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". In: *Journal of Multivariate Analysis* 162, pp. 16–31.

Yang, Y. (2023). “Dimension reduction of high-dimension categorical data with two or multiple responses considering interactions between responses”. In: *Expert Systems with Applications* 221.

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

Zhang, C. and Xiang, Y. (2015). “On the oracle property of adaptive group lasso in high-dimensional linear models”. In: *Statistical Papers* 57, pp. 249–265.

Zhang, C.-H. (2010). “Nearly unbiased variable selection under minimax concave penalty”. In: *The Annals of Statistics* 38.2, pp. 894–942.

Zhang, C.-H. and Huang, J. (2008). “The sparsity and bias of the lasso selection in high-dimensional linear regression”. In: *The Annals of Statistics* 36.4, pp. 1567–1594.

Zhang, C.-H. and Zhang, S. S. (2014). “Confidence intervals for low-dimensional parameters in high-dimensional linear models”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76, pp. 217–242.

Zhang, C.-H. and Zhang, T. (2012). “A general theory of concave regularization for high-dimensional sparse estimation problems”. In: *Statistical Science* 27.4, pp. 576–593.

Zhao, P. and Yu, B. (2006). “On model selection consistency of lasso”. In: *Journal of Machine Learning Research* 7, pp. 2541–2563.

Zhao, Z. and Yang, Y. (2024). “Nonconvex fusion penalties for high-dimensional hierarchical categorical variables”. In: *Information Sciences* 680, p. 121143.

Zou, H. (2006). “The adaptive Lasso and its oracle properties”. In: *Journal of the American Statistical Association* 101, pp. 1418–1429.

Zou, H. and Hastie, T. (2005). “Regularization and variable selection via the elastic net”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2, pp. 301–320.

Zou, H. and Li, R. (2008). “One-step sparse estimates in nonconcave penalized likelihood models”. In: *The Annals of Statistics* 36.4, pp. 1509–1533.

Zou, H. and Zhang, H. H. (2009). “On the adaptive elastic-net with a diverging number of parameters”. In: *The Annals of Statistics* 37.4, pp. 1733–1751.