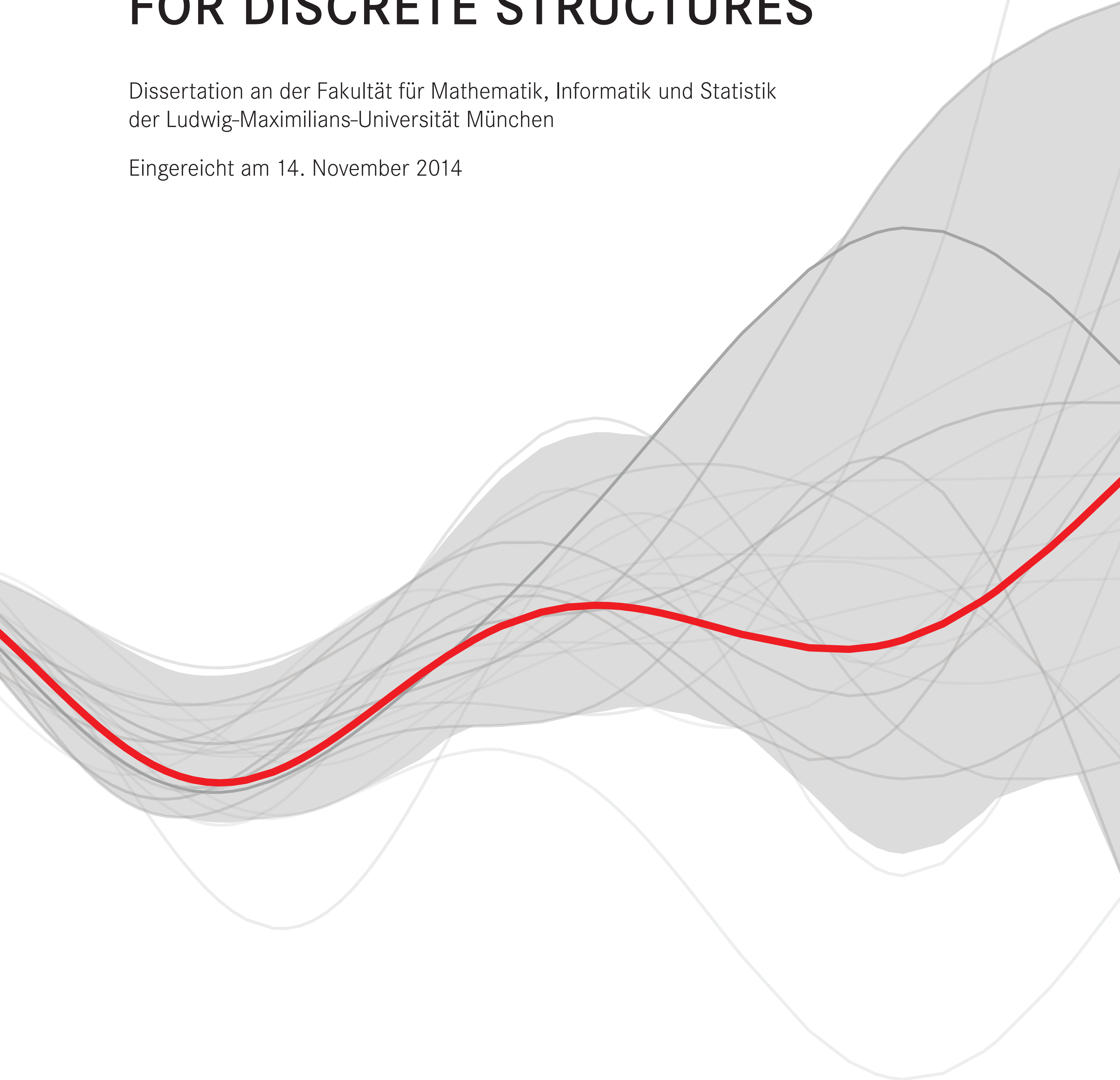
Margret-Ruth Oelker

PENALIZED REGRESSION FOR DISCRETE STRUCTURES

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Eingereicht am 14. November 2014



Margret-Ruth Oelker

PENALIZED REGRESSION FOR DISCRETE STRUCTURES

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Eingereicht am 14. November 2014

Erster Berichterstatter: Prof. Dr. Gerhard Tutz
Zweiter Berichterstatter: Prof. Dr. Thomas Kneib

Tag der Disputation: 08. Januar 2015

Ich bedanke mich bei allen, die diese Arbeit durch ihre Unterstützung ermöglicht haben.

Bei meinem Doktorvater Prof. Dr. Gerhard Tutz,
der mich in den letzten drei Jahren immer wieder gefordert und gefördert hat,
der mich – kurz um – hervorragend betreut hat.

Bei Prof. Dr. Thomas Kneib,
mit dem ich in einer sehr kurzweiligen Kooperation zusammen arbeiten durfte
und der sich bereit erklärt hat die Aufgabe des zweiten Berichterstatters zu übernehmen.

Bei meinen aktuellen und ehemaligen Kolleginnen und Kollegen,
die immer ein offenes Ohr für meine Fragen hatten (es waren viele),
die mich unterstützt und motiviert haben. Vielen Dank für die schönen Pausen!

Bei der Deutschen Forschungsgemeinschaft, dem Zentrum für empirische Studien
und dem LMU-Mentoring-Programm für finanzielle und ideelle Unterstützung.

Bei meiner Familie, meinen Freunden und vor allen Dingen,
bei meinem Freund.

Zusammenfassung

Penalisierte Regressionsmodelle stellen eine Möglichkeit dar die Selektion von Kovariablen in die Schätzung eines Modells zu integrieren. Penalisierte Ansätze eignen sich insbesondere dafür, komplexen Strukturen in den Kovariablen eines Modells zu berücksichtigen. Diese Arbeit beschäftigt sich mit verschiedenen Penalisierungsansätzen für diskrete Strukturen, wobei der Begriff „diskrete Struktur“ in dieser Arbeit alle Arten von kategorialen Einflussgrößen, von effekt-modifizierenden, kategorialen Einflussgrößen sowie von gruppenspezifischen Effekten in hierarchisch strukturierten Daten bezeichnet. Ihnen ist gemein, dass sie zu einer verhältnismäßig großen Anzahl an zu schätzenden Koeffizienten führen können. Deswegen besteht ein besonderes Interesse daran zu erfahren, welche Kategorien einer Einflussgröße die Zielgröße beeinflussen, und welche Kategorien unterschiedliche beziehungsweise ähnliche Effekte auf die Zielgröße haben. Kategorien mit ähnlichen Effekten können beispielsweise durch fused Lasso Penalties identifiziert werden. Jedoch beschränken sich einige, bestehende Ansätze auf das lineare Modell. Die vorliegende Arbeit überträgt diese Ansätze auf die Klasse der generalisierten linearen Regressionsmodelle. Das beinhaltet computationale wie theoretische Aspekte. Konkret wird eine fused Lasso Penalty für effekt-modifizierende kategoriale Einflussgrößen in generalisierten linearen Regressionsmodellen vorgeschlagen. Sie ermöglicht es, Einflussgrößen zu selektieren und Kategorien einer Einflussgröße zu fusionieren. Gruppenspezifische Effekte, die die Heterogenität in hierarchisch strukturierten Daten berücksichtigen, sind ein Spezialfall einer solchen effekt-modifizierenden, kategorialen Größe. Hier bietet der penalisierte Ansatz zwei wesentliche Vorteile: (i) Im Gegensatz zu gemischten Modellen, die stärkere Annahmen treffen, kann der Grad der Heterogenität sehr leicht reduziert werden. (ii) Die Schätzung ist effizienter als im unpenalisierten Ansatz. In orthonormalen Settings können Fused Lasso Penalties konzeptionelle Nachteile haben. Als Alternative wird eine L_0 Penalty für diskrete Strukturen in generalisierten linearen Regressionsmodellen diskutiert, wobei die sogenannte L_0 „Norm“ eine Indikatorfunktion für Argumente ungleich Null bezeichnet. Als Penalty ist diese Funktion so interessant wie anspruchsvoll. Betrachtet man eine Approximation der L_0 Norm als Verlustfunktion wird im Grenzwert der bedingte Modus einer Zielgröße geschätzt.

Summary

Penalties are an established approach to stabilize estimation and to select predictors in regression models. Penalties are especially useful when discrete structures matter. In this thesis, the term “discrete structure” subsumes all kinds of categorical effects, categorical effect modifiers and group-specific effects for hierarchical settings. Discrete structures can be challenging as they need to be coded, and as they can result in a huge number of coefficients. Moreover, users are interested in which levels of a discrete covariate are to be distinguished with respect to the response of a model, or in whether some levels have the same impact on the response. One wants to detect non-influential coefficients and to allow for coefficients with the same estimates. That requires carefully tailored penalization as, for example, provided by different variations of the fused Lasso. However, the reach of many existing methods is restricted as mostly, the response is assumed to be Gaussian. In this thesis, some efforts to extend these approaches are made. The focus is on appropriate penalization strategies for discrete structures in generalized linear models (GLMs). Lasso-type penalties in GLMs require special estimation procedures. In a first step, an existing Fisher scoring algorithm, that allows to combine different types of penalties in one model, is generalized. This algorithm provides the computational basis for the subsequent topics. In a second step, varying coefficients with categorical effect modifiers are considered. Existing methodology for linear models is extended to GLMs. In hierarchical settings, fixed effects models, which are also called group-specific models and which are a special case of categorical effect modifiers, are a common choice to account for the heterogeneity in the data. Applying the proposed penalization techniques for categorical effect modifiers to hierarchical settings offers some benefits: In comparison to mixed models, the approach is able to fuse second level units easily. In comparison to unpenalized group-specific models, efficiency is gained. In a third step, fused Lasso-type penalties for discrete structures are considered in more detail. Especially in orthonormal settings, Lasso-type penalties for categorical effects have some drawbacks regarding the clustering of the coefficients. To overcome these problems, an L_0 penalty for discrete structures is proposed. Again, computational issues are met by a quadratic approximation. This approximation is not only useful in the context of penalized regression for discrete structures, but also when an approximation of the L_0 norm is employed as a loss function. That is, it is useful for regression models that approximate the conditional mode of a response. For linear predictors, a close link to kernel methods allows to show that the proposed estimator is consistent and asymptotically normal. Regression models with semiparametric predictors are possible.

Contents

1. Introduction

In regression modeling, the impact of explanatory covariates on the conditional distribution of a variable of interest – the response – is investigated. The impact of explanatory covariates on the conditional distribution is captured by a predictor that has to be specified. In particular, one has to decide which covariates are considered in the predictor. If a covariate is incorporated in the predictor, its estimated effect has to be evaluated. As one wants to know which covariates are influential, questions of variable selection arise. Among other well-known methods, penalties are an established approach to select predictors in regression models: The least absolute shrinkage and selection operator (Lasso; Tibshirani, 1996) allows to select single covariates by applying the L_1 norm as a penalty. The group Lasso (Yuan and Lin, 2006) selects groups of somehow related effects simultaneously. When the transition among the coefficients shall be smooth, penalties on the differences of coefficients, so called fusion penalties, are popular (see, for example, Tibshirani et al., 2005), to mention only a few approaches.

Penalties are especially useful when discrete structures have to be considered in a model; whereat in this thesis, the term "discrete structure" subsumes all kinds of categorical effects, categorical effect modifiers and group-specific effects as often required for hierarchical settings. For such structures, the use of penalties has mainly three reasons: (i) Users are not only interested in the identification of influential discrete covariates. In fact, they want to know which levels of a discrete covariate affect the response of a model, and whether some levels of a discrete structure have the same impact on the response. The aim is to detect non-influential coefficients and to allow for coefficients with the same estimates. This demand can be met by fusion penalties even if the number of levels is relatively large. In contrast, the use of other methods like best subset selection (see, for example, Fahrmeir et al., 2013) is limited to cases with a low number of levels. (ii) Discrete structures can contain different information as, for example, a specific order or a spatial arrangement of the observed levels. These and related aspects can be considered by the use of penalties. For example, spatial structures can be represented by penalizing the differences of coefficients that belong to neighbored regions. (iii) Discrete structures require coded covariates. A categorical covariate will, for example, be considered by dummy variables that indicate its levels. Depending on the number of discrete covariates in a model and depending on their complexity, the number of coefficients to be estimated can be large. Depending on the number of observations per level, a large number of coefficients can render the estimation

unstable. In this case, penalties are advantageous as they stabilize the estimation procedure. Stable estimation and the selection of coefficients are obtained simultaneously.
Subject to the context and to the type of the discrete covariate, different penalization strategies are reasonable. However, when the fusion of levels or coefficients is intended, different variations of the fused Lasso are a common choice: Bondell and Reich (2009) employ a fused Lasso-type penalty for the selection of discrete factors and for the fusion of the corresponding levels in a analysis of variances. Gertheiss and Tutz (2010, 2012), for example, provide fused Lasso-type penalties for categorical covariates and categorical effect modifiers in a regression context. However, the reach of these methods is restricted, as most times, the response is assumed to be Gaussian which is not the case in many applications. In this thesis, some efforts to extend the mentioned approaches are made. The focus is on appropriate penalization strategies for discrete structures in generalized linear models (GLMs; see, for example, McCullagh and Nelder, 1983).

As a penalty term that selects coefficients, has to be singular at the origin (Fan and Li, 2001), GLMs with general penalties require special estimation procedures. Moreover, the combination of different penalties can be challenging. In **Chapter 2**, this thesis generalizes an algorithm of Ulbricht (2010) that is closely related to the ideas of Fan and Li (2001). The algorithm of Ulbricht (2010) approximates Lasso-type penalties by a quadratic term that can be easily added to conventional penalized iteratively re-weighted least squares (PIRLS) algorithms. The response can follow any simple exponential family. Different penalty terms subject to the L_1 norm can be combined in one model. The proposed generalization restructures the algorithm such that penalties subject to various norms can be combined. Only if necessary, the penalty terms are approximated. In contrast to Ulbricht (2010), the type of the approximation is not fixed. It is possible to employ penalties with vector based arguments like the group Lasso. Some advice on how to combine different penalties in one model is given. In short: Chapter 2 provides the computational basis for the subsequent chapters.

In **Chapter 3**, varying coefficients with categorical effect modifiers are considered. That is, in a regression model, continuous covariates are allowed to vary with the levels of discrete factors, the so called effect modifiers. One is interested in the identification of relevant effect modifiers and in the selection of covariates that are modified. For linear models, Gertheiss and Tutz (2010, 2012) approach this issue with a Lasso-type penalty. Chapter 3 extends this methodology to GLMs. Nominal and ordered effect modifiers are distinguished as their amount of information differs. The large sample properties of the penalized estimator are investigated. In numerical experiments, it is shown that the proposed approach performs well for finite samples.

In hierarchical settings such as in repeated measurement data, fixed effects models are a common choice to account for the heterogeneity in the data. Fixed effects models, which are also called group-specific models, are a special case of categorical effect modifiers. Exactly as for categorical effect modifiers in GLMs, there are some drawbacks regarding the group-specific models. For example, the increasing number of parameters that can render the estimation unstable. Thus, in **Chapter 4**, the proposed penalization techniques for categorical effect modifiers are extended to hierarchical settings. This is new and offers some benefits: In comparison to unpenalized group-specific models, the estimation is more stable; efficiency in terms of the degrees of freedom is gained. In comparison to random effects models (see, for example, Verbeke and Molenberghs, 2000), the approach allows to fuse second level units easily. In case of level 2 endogeneity, the estimator's bias is decreased.

In **Chapter 5**, fused Lasso-type penalties for discrete structures are considered in more detail. Especially for orthonormal settings, Lasso-type penalties for categorical effects have some drawbacks regarding the clustering of the coefficients. To overcome these problems, an L_0 penalty for discrete structures is proposed, where the L_0 "norm" is defined as the number of non-zero entries in a vector. Again, computational issues are met by quadratic approximations of the penalty. Numerical experiments with differently distributed responses are promising. Moreover, the proposed approach is not only an alternative to Lasso-type penalties. It fulfills the same requirements as best subset selection with information criteria for categorical effects while it is feasible for more complex models.

It stands out that the quadratic approximation of the L_0 norm is not only useful in the context of penalized regression for discrete structures, but also when an approximation of the L_0 norm is employed as a loss function, that is, for conditional mode regression. In **Chapter 6**, an iteratively reweighted least squares algorithm is proposed. It performs a regression that approximates the conditional mode of a response. For linear predictors, a close link to kernel methods allows to show that the proposed estimate is consistent and asymptotically normal. In contrast to existing approaches, the tuning parameters, and thus the accuracy of the algorithm, are adjusted while iterating. That makes the approach stable despite the complex loss function. As the employed approximations are quadratic, models can be combined with any quadratic penalty. Therefore, regression models with semiparametric predictors are possible. In practice, the proposed approximation can be combined with existing software for additive models such that a wide range of model components is available for conditional mode regression.

Apart from some cross-references, each chapter is self-contained and can be read separately. The title of the thesis subsumes a wide range of topics. Many important issues as, for example, variances of penalized estimates for finite samples or different cross-validation strategies are not discussed or only shortly sketched. Moreover, the thesis does not aim to give a comprehensive survey on penalized regression models for discrete structures – the focus is on the topics discussed above.

Contributing Manuscripts

Parts of this thesis have been published in peer reviewed journals, in proceedings of different conferences or as technical reports at the Department of Statistics of the Ludwig-Maximilians-Universität München. All manuscripts were written in cooperation with (supervising) co-authors. The manuscripts and the personal contributions of the authors to the respective subjects are listed below. Chapter by chapter, these are:

- **Chapter 2**

  Oelker and Tutz (2013). A general family of penalties for combining differing types of penalties in generalized structured models. Technical Report 139, Ludwig-Maximilians-Universität München, Department of Statistics. http://epub.ub.uni-muenchen.de/17664/.

  Margret Oelker restructured and extended the existing algorithm of Ulbricht (2010), she implemented the related `R` package `gvcm.cat` (R Core Team, 2014; Oelker, 2014) and performed the data analyses. The manuscript was written in close cooperation with Gerhard Tutz who initialized the project.
  Apart from minor modifications, Chapter 2 and Oelker and Tutz (2013) match.

- **Chapter 3**

  Oelker, Gertheiss, and Tutz (2014). Regularization and model selection with categorical predictors and effect modifiers in generalized linear models. *Statistical Modeling 14*(2), 157–177.

  Jan Gertheiss and Gerhard Tutz set up and supervised the project. Jan Gertheiss provided the boar data. Margret Oelker conducted the numerical experiments and the data analysis. Following the argumentation of Gertheiss and Tutz (2012) closely and thanks to the precise comments of Nils Lid Hjort (University of Oslo), she derived the asymptotic properties of the estimator. The manuscript was drafted in close collaboration of all coauthors.
  Chapter 3 and Oelker et al. (2014) differ by Sections 3.5–3.6. Apart from that, both manuscripts are very similar. Preliminary work on Chapter 3 is found in the Proceedings of COMPSTAT - 20th International Conference on Computational Statistics (Oelker et al., 2012a). The Technical Report 122 (Oelker et al., 2012b) is an early version of Chapter 3.

- **Chapter 4**

  Tutz and Oelker (2014). Modeling clustered heterogeneity: fixed effects, random effects and mixtures. Technical Report 156, Ludwig-Maximilians-Universität München, Department of Statistics. http://epub.ub.uni-muenchen.de/18987/.

Chapter 4 was initiated by Gerhard Tutz who conceptualized the theoretical framework and who investigated the literature. Margret Oelker conducted several numerical experiments and the data analyses; including conceptual work on the data generation. She contributed substantially to the presentation of the results.
Chapter 4 is a modified version of Tutz and Oelker (2014). While wording, notation and the arrangement of the sections differ for the most part, the message of both manuscripts is the same. Initial results can be found in the Proceedings of the 28th International Workshop on Statistical Modelling (Tutz and Oelker, 2013).

- **Chapter 5**

  Oelker, Pößnecker, and Tutz (2015). Selection and fusion of categorical predictors with L_0-type penalties. *Statistical Modeling 15*(4). Accepted for publication.

  Gerhard Tutz initiated the use of L_0-type penalties to regression models. Margret Oelker investigated the theory behind L_1-type penalties in orthonormal settings. She implemented the simulation study and the data analysis. Wolfgang Pößnecker contributed essentially to Section C.1 of the Appendix. The manuscript was drafted in close collaboration of Gerhard Tutz and Margret Oelker.
  Apart from minor modifications, Chapter 5 and Oelker et al. (2015) match.

- **Chapter 6**

  Oelker, Sobotka, and Kneib (2014). On (semiparametric) mode regression. In T. Kneib, F. Sobotka, J. Fahrenholz, and H. Irmer (Eds.), *Proceedings of the 29th International Workshop on Statistical Modelling*, Volume 1, pp. 243–248.

  Thomas Kneib focused on mode regression and came up with the idea of nested intervals. Margret Oelker derived the approximation of the loss function and showed that the asymptotic theory of Kemp and Santos Silva (2012) can be applied. She implemented the method in R (R Core Team, 2014). The numerical experiments and the data analyses are the product of a close cooperation of Margret Oelker and Fabian Sobotka, who had a strong impact on the manuscript through literature and data inquiries. Nadja Klein had strong influence on the structuring of the results; she commented on the assumptions for the asymptotic theory and proofread Appendix D. The manuscript was drafted in close collaboration of all coauthors.
  Oelker et al. (2014) is a very short version of Chapter 6. A slightly modified version of Chapter 6 is submitted. The contributions of Nadja Klein became relevant in the course of this project – she is the fourth co-author of this submission.

To provide a thesis that is easy to read, in the body of a chapter, the underlying publications are no longer cited although there are textual matches.

Software

All computations were done with the statistical program R (R Core Team, 2014; version 3.1.0, 2014-04-10) on two different platforms (x86_64-pc-linux-gnu, 64-bit; i386-w64-mingw32/i386, 32-bit). The results of Chapters 2–5 rely basically on the R package gvcm.cat (Oelker, 2014). It imports the R packages Matrix (Bates and Maechler, 2014), MASS (Venables and Ripley, 2002) and splines; unless indicated else wise, version 1.7 is employed. The functions for Chapter 6 are to be published in an R package and are available upon request. For comparisons with competing approaches, the R packages flexmix (Grün and Leisch, 2008a), glmnet (Friedman et al., 2010), grplasso (Meier, 2013), lars (Hastie and Efron, 2013), lme4 (Bates et al., 2014), mgcv (Wood, 2011), and nlme (Pinheiro et al., 2014) are employed. For some data in Section 3, the package catdata (Schauberger and Tutz, 2014) is needed. For the tables and some of the graphics, the R packages xtable (Dahl, 2014), EBImage (Pau et al., 2014), BayesX (Kneib et al., 2014), and the program GIMP (GIMP Team, 2012) are required.

2. A General Family of Penalties for Structured Regression

2.1. Introduction

In recent years, model selection and regularization in regression models has been an area of intensive research. Often, penalized approaches are the method of choice. Examples are Ridge regression (Hoerl and Kennard, 1970), the least absolute shrinkage and selection operator (Lasso; Tibshirani, 1996), the smoothly clipped absolute deviation penalty (SCAD; Fan and Li, 2001), the fused Lasso (Tibshirani et al., 2005), the elastic net (Zou and Hastie, 2005) and the (adaptive) group Lasso (Yuan and Lin, 2006; Wang and Leng, 2008), to mention only a few approaches. The number of applications is huge. In nonparametric regression, penalties smooth wiggly functions. Eilers and Marx (1996) work, for example, with Ridge penalties on higher order differences of B-spline coefficients. Meier et al. (2009) select splines with a group Lasso penalty. For wavelets and signals, L_0 penalties, or more general L_q penalties, $0 \leq q \leq 1$, are employed (Antoniadis and Fan, 2001; Rippe et al., 2012). Concerning categorical data, Bondell and Reich (2009) or Gertheiss and Tutz (2010) work with fused Lasso type penalties. Fahrmeir et al. (2010) offer a flexible framework for Bayesian regularization and variable selection, amongst others with spike and slab priors. Various efficient algorithms to solve the resulting optimization problems are available, be it in linear models, generalized linear models (GLMs), hazard rate models or others. Least angle regression (lars; Efron et al., 2004; Hastie and Efron, 2013) offers a conceptual framework to compute the entire regularization path of the Lasso by exploiting its piecewise linear coefficient profiles. Osborne and Turlach (2011) propose a homotopy algorithm for the quantile regression Lasso and related piecewise linear problems. Meier et al. (2008) propose a coordinate-descent algorithm for the group Lasso in logistic regression problems. Goeman (2010) solves Lasso, fused Lasso and Ridge-type problems in high-dimensional models by a

This chapter is a modified version of the Technical Report 139 (Oelker and Tutz, 2013). For more information on the contributions of the authors and on textual matches, see page 4.

combination of gradient ascent optimization and a Newton-Raphson algorithm. Friedman et al. (2010) use cyclical coordinate descent algorithms, computed along a regularization path, for the elastic net and related convex penalties. Ulbricht (2010) proposes a penalized iteratively re-weighted least squares (PIRLS) algorithm for Lasso-type penalties in GLMs. Wood (2011) offers a great PIRLS-based toolbox for generalized additive models and generalized Ridge regression (`R`-package `mgcv`).

In the mentioned approaches, penalties have one specific purpose, for example, the selection of variables in a linear predictor or the selection of smooth non-linear effects. However, in applications, a combination of penalties that serve different purposes, is needed frequently, for example, for the analysis of the rents of 1488 households in the city of Munich. To model the rent, continuous covariates like the flat's size and age, as well as some explanatory factors are collected. The effect of the age of a flat is known to be non-linear (see, for example, Fahrmeir and Tutz, 2001) and can be modeled by splines with a Ridge-type penalty on the function's curvature. When investigating whether the effect of the residential area is linear or not, an additional group Lasso penalty is helpful. As some levels of the categorical effects are only sparsely occupied, ordered effects like the number of rooms of a flat require regularization, too. This can be done by employing a fused Lasso penalty on the dummy coefficients of this effect. Hence, the overall penalty is a sum of Ridge-, group Lasso- and Lasso-type penalties. We will use a generalized structured regression model with gamma distributed response.

Although the algorithm of Friedman et al. (2010) covers Ridge- and Lasso-type penalties within one model via the elastic net, it does not allow for other penalties. The `R`-package `mgcv` allows for penalized smooth functions and penalized parametric effects, but the penalty terms for the parametric effects have to be quadratic. Even though the algorithm of Ulbricht (2010) works for a family of Lasso-type penalties, we found no algorithm obviously matching the requirements of our data.

As in Marx and Eilers (1998), many algorithms are based on Fisher scoring methods, which are the default approach for the estimation of GLMs. For quadratic penalties, a penalty matrix is added to the Fisher information matrix. See, for example, the PIRLS algorithm in the `R`-package `mgcv`. For non-quadratic penalties, approximations are available, and again PIRLS algorithms are applied. Fan and Li (2001) approximate the non-convex SCAD penalty quadratically. Ulbricht (2010) adopts this idea for Lasso-type penalties. Rippe et al. (2012) approximate the L_0 norm quadratically by a re-weighted Ridge penalty. Hence, to combine different penalties that employ different norms, quadratic approximations in PIRLS algorithms seem to be a natural choice.

In this chapter, it is shown how penalties, that are (semi-)norms of scalar linear transformations of the coefficient vector, can be approximated quadratically within a general model structure as in GLMs. The penalty is defined such that the Lasso, the fused Lasso, the Ridge, the SCAD, the elastic net and other regularization terms for categorical effects are embedded. The proposed approximation allows to combine all these penalties in one model. The estimation is based on conventional PIRLS algorithms and hence, easy to implement. The idea is based on and generalizes the approaches of Fan and Li (2001) and Ulbricht (2010). In contrast to Ulbricht (2010), it is not restricted to penalties that are based on (functions of) absolute values but allows for penalties with general norms. Each penalty term can be approximated differently, and differentiable penalty terms do not have to be approximated. The approach is extended to penalties like the group Lasso, that is, to penalties with norms of vectorial linear transformations of the coefficient vector.

Chapter 2 is organized as follows: Section 2.2 introduces the method. The approximation is derived, some technical remarks are given, and the approach is extended to vectorial linear transformations of the coefficient vector. Section 2.3 illustrates the performance of the approach by comparing established algorithms with the proposed approximation. In Section 2.4, the Munich rent data is analyzed.

2.2. Local Quadratic Approximations in PIRLS Algorithms

We consider a general model structure as in GLMs by assuming that the mean response $\mu_i = \mathbb{E}(y_i|x_i)$ is given by

$$\mu_i = h(\eta_i),$$

$i = 1, \dots, n$. Given the vector of covariates $\boldsymbol{x}_i \in \mathbb{R}^q$, for the response $y_i|\boldsymbol{x}_i$ a simple exponential family with log-likelihood $l_n(\boldsymbol{\beta})$ is assumed:

$$l_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{y_i \vartheta_i(\mu_i) - b(\vartheta_i(\mu_i))}{\varphi} + c(y_i, \varphi),$$

where $\vartheta_i(\mu_i)$ denotes the natural parameter, $b(\cdot)$ is a specific function corresponding on the type of the exponential family, $c(\cdot)$ is the log-normalization constant and φ the dispersion parameter (compare McCullagh and Nelder, 1983; Fahrmeir and Tutz, 2001). The conditional mean of the response y_i is linked to a linear predictor $\eta_i = \boldsymbol{x_i}^T\boldsymbol{\beta}$, where $\boldsymbol{\beta} \in \mathbb{R}^q$ is a coefficient vector and h is a twice continuously differentiable inverse link function, often referred to as response function. Vectors $\boldsymbol{x}_i$ build the design matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)^T \in \mathbb{R}^{n \times q}$, which represents $1, \dots, p$ covariates. By a general model structure, we mean that the covariates in the design matrix $\boldsymbol{X}$ can have any structure – provided that they can be parametrized as $\boldsymbol{x}_i^T\boldsymbol{\beta}$. In particular, we allow for nonparametric terms that represent unknown functions. For example, when a continuous covariate is modeled nonparametrically as $f(\boldsymbol{x}_j)$, we assume that $f(\boldsymbol{x}_j)$ is represented in $\boldsymbol{X}$ by the evaluations of some basis functions. Categorical covariates are assumed to be properly coded. The design matrix always contains an intercept; and, we assume that the structure of the coefficient vector is $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_0^T, \boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_p^T)$, where entries $\boldsymbol{\beta}_j \in \mathbb{R}^{k_j}$ are vectors that correspond to structures in the predictor space.

A vector $\boldsymbol{\beta}_j$ can, for example, contain the coefficients of the basis functions of a smoothly modeled covariate, or the coefficients linked to the dummies of a categorical covariate. In the penalized maximum likelihood (ML) framework considered here, the objective is

$$\mathcal{M}_{pen}(\boldsymbol{\beta}) = -l_n(\boldsymbol{\beta}) + P_{\lambda}(\boldsymbol{\beta}), \tag{2.1}$$

where $l_n(\boldsymbol{\beta})$ denotes the log-likelihood of the exponential family based on n observations and where $P_{\lambda}(\boldsymbol{\beta})$ denotes the penalty. Similar to Ulbricht (2010), the general penalty that we employ, has the form

$$P_{\lambda}(\boldsymbol{\beta}) = \sum_{l=1}^{L} \lambda_l p_l(\|\boldsymbol{a}_l^T \boldsymbol{\beta}\|_{\mathrm{N}_l}), \tag{2.2}$$

where functions p_l are penalty functions, λ_l are penalty parameters, and $\|\cdot\|_{\mathrm{N}_l}$ denotes any (semi-)norm, for example, $\|\xi\|_{\mathrm{N}_l} = |\xi|^r$, $r \geq 0$. $\|\cdot\|_{\mathrm{N}_l}$ is not restricted to (semi-)norms; it can be any term that is meaningful as a penalty, for example, an indicator for non-zero arguments, which is often called L_0 “norm” (Donoho and Elad, 2003). Vectors $\boldsymbol{a}_l \in \mathbb{R}^q$ build transformations of the coefficient vector, for example, differences of adjacent coefficients. In principal, there can be arbitrary many restrictions L. As proposed by Ulbricht (2010), for the penalty functions, we assume

1. $p_l : \mathbb{R}^+ \to \mathbb{R}^+$, $p_l(0) = 0$,
2. $p_l(\xi)$ is continuous and strictly monotone in ξ,
3. $p_l(\xi)$ is continuously differentiable for all $\xi \neq 0$, such that $p_l' = \mathrm{d}p_l(\xi)/\mathrm{d}\xi > 0$.

Together, p_l, $\|\cdot\|_{\mathrm{N}_l}$ and the vectors $\boldsymbol{a}_l$ define the type of the penalty. Note, that properties like the curvature of $p_l(\|\xi\|_{\mathrm{N}_l})$ depend on the properties of $p_l(\xi)$ and $\|\xi\|_{\mathrm{N}_l}$. For example, when $p_l(\xi)$ and $\|\xi\|_{\mathrm{N}_l}$ are convex for all l, and $p_l(\xi)$ is monotonically increasing as assumed, then the penalty is convex. The flexibility of the penalty lies in the possible choices of the three components. In the following, some examples are given:

Elastic net: To penalize a scalar effect β_j by the naïve elastic net penalty $\lambda_l \cdot |\beta_j| + \lambda_k \cdot \beta_j^2$ (Zou and Hastie, 2005), two penalty functions are needed; one denoted by $p_l(\xi) = \xi$, $\|\xi\|_{\mathrm{N}_l} = |\xi|$ and an indicator vector $\boldsymbol{a}_l$ such that $\boldsymbol{a}_l^T \boldsymbol{\beta} = \beta_j$; the other is $p_k(\xi) = \xi$ with $\|\xi\|_{\mathrm{N}_k} = \xi^2$ and with the same indicator vector $\boldsymbol{a}_l$ as before.

Adaptive Lasso: To penalize the effect of the j-th continuous covariate with the adaptive Lasso (Zou, 2006), $\boldsymbol{a}_l$ is an indicator vector for the position of β_j, $\|\xi\|_{\mathrm{N}_l}$ is the absolute value $|\xi|$ and $p_l(\xi) = |\boldsymbol{a}_l^T \hat{\boldsymbol{\beta}}^{ML}|^{-1} \cdot \xi$, where $\hat{\boldsymbol{\beta}}^{ML}$ denotes the ML estimate of $\boldsymbol{\beta}$.

Penalized B-splines: When the continuous covariate $\boldsymbol{x}_j$ is modeled nonparametrically, $\boldsymbol{\beta}_j$ is a sub-vector that represents coefficients on k_j basis functions, for example, of cubic B-splines. To penalize the roughness of $f(\boldsymbol{x}_j)$ as proposed by Eilers and Marx (1996), there are $k_j - 2$ penalty terms. Vectors $\boldsymbol{a}_l$ build all needed second order differences $(0, \dots, 0, 1, -2, 1, 0, \dots, 0)^T$. For all penalty terms, one employs $\|\xi\|_{\mathrm{N}_l} = \xi^2$, $p_l(\xi) = \xi$ and the same penalty parameter λ_l.

Typically, the penalty is structured as $P_\lambda(\boldsymbol{\beta}) = \sum_{j=0}^{p}\sum_{l=1}^{L_j}\lambda_{jl}p_{jl}(\left\|\boldsymbol{a}_{jl}^T\boldsymbol{\beta}_j\right\|_{\mathrm{N}_l})$. That is, the effects of each covariate are penalized separately. We will, however, use the general form (2.2), which uses one index to denote the specific terms.
In common GLMs, the unpenalized optimization problem is $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \; -l_n(\boldsymbol{\beta})$. The problem is solved iteratively by solving the linearized problem

$$\boldsymbol{s}^{lin}(\boldsymbol{\beta}) = \boldsymbol{s}(\boldsymbol{\beta}_{(k)}) + \boldsymbol{H}(\boldsymbol{\beta}_{(k)})(\boldsymbol{\beta} - \boldsymbol{\beta}_{(k)}) = \boldsymbol{0},$$

for a given vector $\boldsymbol{\beta}_{(k)}$ in each step, where $\boldsymbol{s}(\boldsymbol{\beta}) = \partial l_n(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ is the score function and $\boldsymbol{H}(\boldsymbol{\beta}) = \partial^2 l_n(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T$ is the Hessian matrix. Rearranging gives

$$\hat{\boldsymbol{\beta}}_{(k+1)} = \hat{\boldsymbol{\beta}}_{(k)} - \boldsymbol{H}(\hat{\boldsymbol{\beta}}_{(k)})^{-1}\boldsymbol{s}(\hat{\beta}_{(k)}),$$

which can be transformed to a Fisher scoring algorithm or an iteratively re-weighted least squares algorithm. In order to use a PIRLS algorithm for the penalized optimization problem (2.1), penalized versions of the score function $\boldsymbol{s}(\boldsymbol{\beta}_{(k)})$ and the Hessian matrix $\boldsymbol{H}(\boldsymbol{\beta}_{(k)})$ or the Fisher matrix are needed. In particular, derivatives of $\mathcal{M}_{pen}(\boldsymbol{\beta})$ or close approximations are needed. To this end, non-differentiable norms $\|\cdot\|_{\mathrm{N}_l}$ are approximated. We assume that an approximation $\mathcal{N}_l(\xi, \mathcal{T})$ to each employed norm $\|\cdot\|_{\mathrm{N}_l}$ exists, such that

$$\|\xi\|_{\mathrm{N}_l} = \lim_{\mathcal{T}\to\mathcal{B}} \; \mathcal{N}_l(\xi, \mathcal{T}),$$

where $\mathcal{T}$ denotes a set of possible tuning parameters and $\mathcal{B}$ denotes the corresponding set of boundary values with $\|\xi\|_{\mathrm{N}_l} = \mathcal{N}_l(\xi, \mathcal{B})$. $\mathcal{N}_l(\xi, \mathcal{T})$ is supposed to be at least twice continuously differentiable. We define $\mathcal{D}_l(\xi, \mathcal{T}) = \partial\mathcal{N}_l(\xi, \mathcal{T})/\partial\xi$ and assume that

$$\frac{\partial \|\xi\|_{\mathrm{N}_l}}{\partial\xi} = \lim_{\mathcal{T}\to\mathcal{B}} \; \mathcal{D}_l(\xi, \mathcal{T}),$$

for all l, $l = 1, \ldots, L$, and for all ξ for which $\partial \|\xi\|_{\mathrm{N}_l} / \partial\xi$ is defined. Moreover, we assume $\mathcal{D}_l(0, \mathcal{T}) = 0$. To keep the notation simple, we will write $\mathcal{N}_l(\xi)$ instead of $\mathcal{N}_l(\xi, \mathcal{T})$, and $\mathcal{D}_l(\xi)$ instead of $\mathcal{D}_l(\xi, \mathcal{T})$. Apart from this approximation, the schedule is the same as for the unpenalized case. The penalized score function $\boldsymbol{s}_{pen}(\boldsymbol{\beta})$ is linearized by a Taylor expansion. $\boldsymbol{s}_{pen}^{lin}(\boldsymbol{\beta}) = \boldsymbol{0}$ is solved iteratively.

2.2.1. Examples of Approximations

Table 2.1 gives an idea of the approximations of different norms. As in Koch (1996) and Ulbricht (2010), the L_1 norm is approximated by $\mathcal{N}_l(\xi) = \sqrt{\xi^2 + c}$ where c is a small positive number (in our experience $c \approx 10^{-5}$ works well) and controls how close the approximation and the L_1 norm are. For $c = 0$, we have $|\xi| = \sqrt{\xi^2}$. The first derivative of

Norm	$\mathcal{N}_l(\xi)$	$\mathcal{D}_l(\xi)$
$\lVert\xi\rVert_1 = \lvert\xi\rvert$	$\sqrt{\xi^2+c}$	$(\xi^2+c)^{-1/2}\cdot\xi$
$\lVert\xi\rVert_2^2 = \xi^2$	ξ^2	2ξ
$\lVert\xi\rVert_0 = I_{\xi\neq 0}$	$\frac{2}{1+\exp(-\gamma\lvert\xi\rvert)} - 1$	$\frac{2\gamma}{1+\exp(-\gamma\lvert\xi\rvert)}(1-\frac{1}{1+\exp(-\gamma\lvert\xi\rvert)})\frac{\xi}{\sqrt{\xi^2+c}}$
$\lVert\xi\rVert_r = \lvert\xi\rvert^r$	$(\xi^2+c)^{r/2}$	$r\xi(\xi^2+c)^{r/2-1}$

Table 2.1.: Examples for approximations of norms. Column $\mathcal{N}_l(\xi)$ depicts the approximations of the L_1 norm, of the quadratic term $\lVert\xi\rVert_2^2 = \xi^2$, of the L_0 norm and of the term $\lVert\xi\rVert_r$ which is needed for Bridge penalties. Column $\mathcal{D}_l(\xi)$ depicts the (approximated) derivatives of $\mathcal{N}_l(\xi)$. c is a small positive number, γ is a large integer.

the approximation $\mathcal{N}_l(\xi)$ is $\xi(\xi^2+c)^{-1/2}$ which is a continuous approximation for $\text{sign}(\xi)$, $\xi \neq 0$, the first derivative of the L_1 norm. There is no need to approximate $\lVert\xi\rVert_2^2 = \xi^2$ as it is quadratic. The approximation of the L_0 norm is motivated by the logistic function. We choose $\mathcal{N}_l(\xi) = 2(1+\exp(-\gamma\lvert\xi\rvert))^{-1} - 1$, where γ is a large integer. Accordingly, the derivative is $\mathcal{D}(\xi) = 2\gamma\xi/(1+\exp(-\gamma\lvert\xi\rvert))(1-1/(1+\exp(-\gamma\lvert\xi\rvert)))(\xi^2+c)^{-1/2}$, where the absolute value is approximated like defined above. Figure 2.1 illustrates these approximations (left column) and their derivatives (right column). On top, the approximation of the L_1 norm is shown. On bottom the L_0 norm is approximated. The dashed lines mark the exact norms and the exact derivatives based on sub-gradients at $\xi = 0$. For all plots, $c = 0.1$ and $\gamma = 5$ are employed for illustrative reasons.

Combining these approximations with different functions $p_l(\cdot)$ allows to approximate various known penalties. Table 2.2 illustrates the variety of penalties that can be approximated. The penalties are dissected in the underlying norm $\lVert\xi\rVert_{\mathrm{N}_l}$, functions p_l and expressions $\boldsymbol{a}_l^T\boldsymbol{\beta}$. Of course, other combinations of norms $\lVert\xi\rVert_{\mathrm{N}_l}$, functions p_l and vectors $\boldsymbol{a}_l^T\boldsymbol{\beta}$ are possible. For example, one could think of an adaptively weighted Ridge penalty, or L_1-type penalties for coefficients of splines. The Bridge penalty $\lvert\xi\rvert^r$, $r \geq 0$, for a metric covariate x_j corresponds to $\mathcal{N}_l(\xi) = \lVert\xi\rVert_r$ with $p_l(\xi) = \xi$ and $\boldsymbol{a}_l^T\boldsymbol{\beta} = \beta_j$ (Frank and Friedman, 1993). However, matters simplify a lot by employing the direct approximations for $r \in \{0, 1, 2\}$. The penalty of Gertheiss and Tutz (2012) for categorical effect modifiers fits in the proposed framework, too (see Section 3). In contrast to previous approaches, all these penalties can be combined in one model where the response $y_i|\boldsymbol{x}_i$ can follow any exponential family. Many models for other types of responses can be re-parametrized such that they fit in the framework of general structured models. The Cox model (Cox, 1972) for discrete time points can be written as a logit model (Fahrmeir and Tutz, 2001). Sequential models for ordinal response can be written as binary models, too (Tutz, 2012).

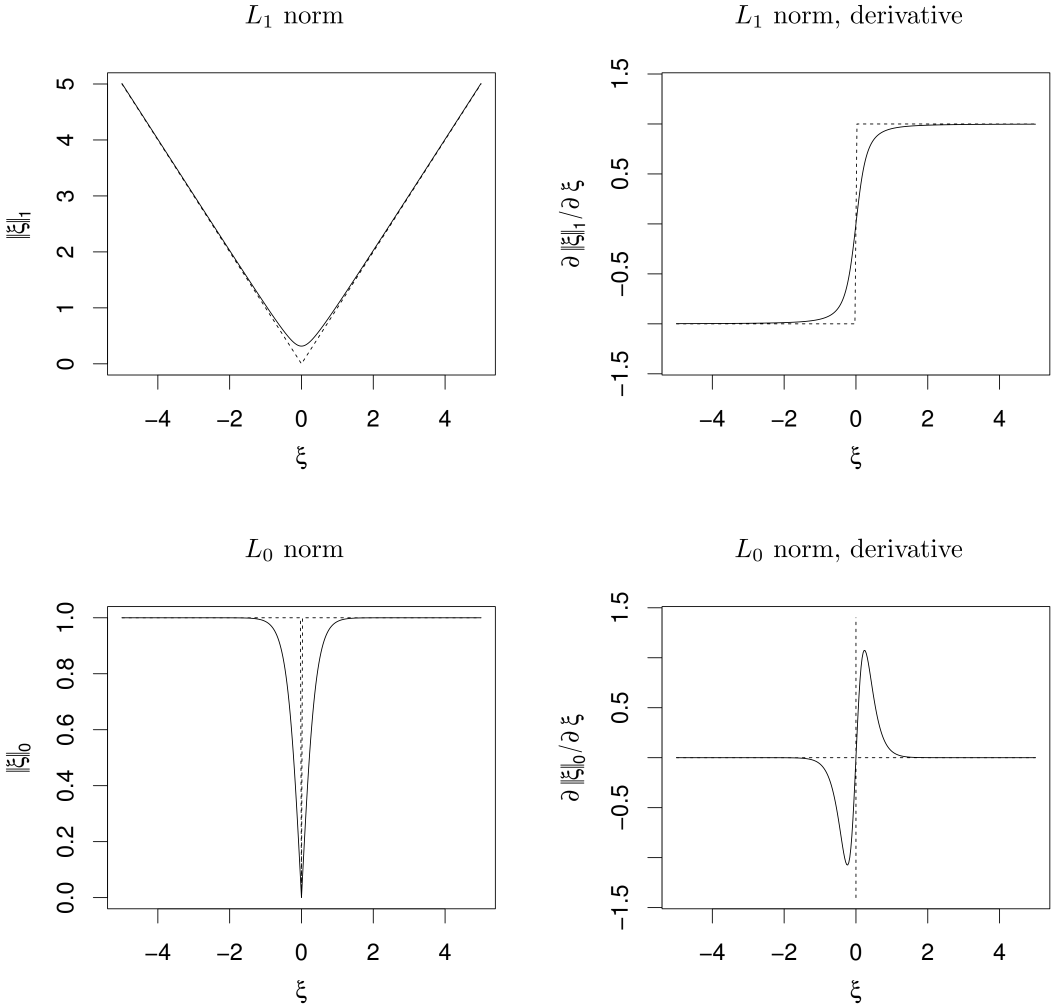


Figure 2.1.: Graphical illustration of the approximation of the L_1 and the L_0 norm (left column) and their derivatives (right column) with respect to $\xi = \boldsymbol{a}_l^T\boldsymbol{\beta}$. The upper row relates to the L_1 norm, the lower row to the L_0 norm. The dashed lines mark the exact norms and the exact derivatives based on sub-gradients at $\xi = 0$. $c = 0.1$, $\gamma = 5$ for graphical reasons. A similar figure can be found in Ulbricht (2010).

Penalty	Covariate	Penalty Terms		
		$\|\xi\|_{\mathrm{N}_l}$	Pen. Function	$\boldsymbol{a}_l^T\boldsymbol{\beta} =$
Lasso	metric $\boldsymbol{x}_j$	L_1	$p_l(\xi) = \xi$	β_j
Adaptive Lasso	metric $\boldsymbol{x}_j$	L_1	$p_l(\xi) = \xi/\|\boldsymbol{a}_l^T\hat{\boldsymbol{\beta}}^{ML}\|$	β_j
Ridge	metric $\boldsymbol{x}_j$	L_2^2	$p_l(\xi) = \xi$	β_j
SCAD	metric $\boldsymbol{x}_j$	L_1	$p_l'(\xi) = I_{\xi\leq\lambda_l} + \frac{(a\lambda_l-\xi)_+}{(a-1)\lambda_l} I_{\xi>\lambda_l}$	β_j
Elastic net	metric $\boldsymbol{x}_j$	1. L_1 2. L_2^2	$p_l(\xi) = \xi$ $p_k(\xi) = \xi$	β_j β_j
Fused Lasso	k_j ordered $\boldsymbol{x}_j$, $j = s, \dots, t$	L_1	$p_l(\xi) = \xi$	$\beta_j - \beta_{j-1}$, $j = s+1, \dots, t$
Penalized B-splines	$f(\boldsymbol{x}_j)$, $\boldsymbol{x}_j$ metric, parametrized by k_j coeff. β_{jk}	L_2^2	$p_l(\xi) = \xi$	$\beta_{jk} - 2\beta_{j,k-1} + \beta_{j,k-2}$, $k = 3, \dots, k_j$
Simultaneous factor selection (Bondell and Reich, 2009)	nominal factor $\boldsymbol{x}_j$ with k_j coeff. β_{jk}, $k = 1, \dots, k_j$	L_1	$p_l(\xi) = \xi$	$\beta_{jk} - \beta_{jr}$, $k > r \geq 0$
Sparse modeling of categ. variables (Gertheiss and Tutz, 2010)	ordinal factor $\boldsymbol{x}_j$ with k_j coeff. β_{jk}, $k = 1, \dots, k_j$	L_1	$p_l(\xi) = \xi$	$\beta_{jk} - \beta_{j,k-1}$, $k = 1, \dots, k_j$
L_0 penalty for signals (Rippe et al., 2012)	k_j ordered $\boldsymbol{x}_j$, $j = s, \dots, t$	L_0	$p_l(\xi) = \xi$	$\beta_j - \beta_{j-1}$, $j = s+1, \dots, t$

Table 2.2.: Examples of approximations of known penalties. The elastic net is made up by two terms with separate penalty parameters λ_l and λ_k. The other penalties are governed by one penalty parameter λ_l, even when they are defined by several terms. The fused Lasso consists out of $k_j - 1$ terms related to divers differences. The same holds for penalized B-splines ($k_j - 2$ penalty terms), the penalties in Bondell and Reich (2009) ($\frac{1}{2}(k_j + 1)k_j$ terms) and in Gertheiss and Tutz (2010) (k_j terms). In the penalty terms for factors, the coefficient $\beta_{j0} = 0$ relates to the reference category. The SCAD penalty (Fan and Li, 2001) is defined by its derivative; parameter a, $a > 2$, is an additional tuning parameter. Fan and Li (2001) recommend $a = 3.7$.

2.2.2. Approximation of the Penalty

In this section, the approximation of the penalty and the proposed algorithm are derived. As the approach extends the algorithms of Fan and Li (2001) and of Ulbricht (2010), emphasis is placed on what the approaches have in common and on how they differ. For the sake of simplicity, we write $\mathcal{N}_l(\cdot)$ and $\mathcal{D}_l(\cdot)$ for all penalty terms, even though not all norms have to be approximated.

In order to approximate the penalty, as in Fan and Li (2001), a Taylor expansion at $\boldsymbol{\beta}_{(k)}$ is employed:

$$P_\lambda(\boldsymbol{\beta}) \approx P_\lambda(\boldsymbol{\beta}_{(k)}) + \nabla P_\lambda(\boldsymbol{\beta}_{(k)})^T \cdot (\boldsymbol{\beta} - \boldsymbol{\beta}_{(k)}),$$

where

$$\nabla P_\lambda(\boldsymbol{\beta}_{(k)})^T \cdot (\boldsymbol{\beta} - \boldsymbol{\beta}_{(k)}) = \sum_{l=1}^{L} \lambda_l \nabla p_l(\left\|\boldsymbol{a}_l^T \boldsymbol{\beta}_{(k)}\right\|_{\mathrm{N}_l})^T (\boldsymbol{\beta} - \boldsymbol{\beta}_{(k)}).$$

In analogy to Fan and Li (2001), in what follows, we use the local approximation $\boldsymbol{a}_l^T\boldsymbol{\beta}/\boldsymbol{a}_l^T\boldsymbol{\beta}_{(k)} \approx 1$ for $\boldsymbol{\beta}_{(k)}$ close to $\boldsymbol{\beta}$. Moreover, $\boldsymbol{a}_l^T\boldsymbol{\beta}\boldsymbol{a}_l^T(\boldsymbol{\beta} - \boldsymbol{\beta}_{(k)})$ is approximated by $\frac{1}{2}(\boldsymbol{\beta}^T\boldsymbol{a}_l\boldsymbol{a}_l^T\boldsymbol{\beta} + \boldsymbol{\beta}_{(k)}^T\boldsymbol{a}_l\boldsymbol{a}_l^T\boldsymbol{\beta}_{(k)})$ via completing the square as proposed by Ulbricht (2010). That gives

$$\begin{aligned}
\nabla p_l(\left\|\boldsymbol{a}_l^T \boldsymbol{\beta}_{(k)}\right\|_{\mathrm{N}_l})^T (\boldsymbol{\beta} - \boldsymbol{\beta}_{(k)}) &= \frac{\partial p_l(\left\|\boldsymbol{a}_l^T \boldsymbol{\beta}_{(k)}\right\|_{\mathrm{N}_l})}{\partial \left\|\boldsymbol{a}_l^T \boldsymbol{\beta}_{(k)}\right\|_{\mathrm{N}_l}} \cdot \frac{\partial \left\|\boldsymbol{a}_l^T \boldsymbol{\beta}_{(k)}\right\|_{\mathrm{N}_l}}{\partial \boldsymbol{a}_l^T \boldsymbol{\beta}_{(k)}} \cdot \frac{\partial \boldsymbol{a}_l^T \boldsymbol{\beta}_{(k)}}{\partial \boldsymbol{\beta}_{(k)}^T} (\boldsymbol{\beta} - \boldsymbol{\beta}_{(k)}) \\
&\approx p_l'(\left\|\boldsymbol{a}_l^T \boldsymbol{\beta}_{(k)}\right\|_{\mathrm{N}_l}) \cdot \frac{\mathcal{D}_l(\boldsymbol{a}_l^T \boldsymbol{\beta}_{(k)})}{\boldsymbol{a}_l^T \boldsymbol{\beta}_{(k)}} \boldsymbol{a}_l^T \boldsymbol{\beta} \cdot \boldsymbol{a}_l^T \cdot (\boldsymbol{\beta} - \boldsymbol{\beta}_{(k)}), \\
&\approx \frac{1}{2} p_l'(\left\|\boldsymbol{a}_l^T \boldsymbol{\beta}_{(k)}\right\|_{\mathrm{N}_l}) \cdot \frac{\mathcal{D}_l(\boldsymbol{a}_l^T \boldsymbol{\beta}_{(k)})}{\boldsymbol{a}_l^T \boldsymbol{\beta}_{(k)}} (\boldsymbol{\beta}^T \boldsymbol{a}_l \boldsymbol{a}_l^T \boldsymbol{\beta} + \boldsymbol{\beta}_{(k)}^T \boldsymbol{a}_l \boldsymbol{a}_l^T \boldsymbol{\beta}_{(k)}) \\
&= \frac{1}{2} (\boldsymbol{\beta}^T \boldsymbol{A}_l \boldsymbol{\beta} + \boldsymbol{\beta}_{(k)}^T \boldsymbol{A}_l \boldsymbol{\beta}_{(k)}),
\end{aligned}$$

where

$$\boldsymbol{A}_l = p_l'(\left\|\boldsymbol{a}_l^T \boldsymbol{\beta}_{(k)}\right\|_{\mathrm{N}_l}) \cdot \frac{\mathcal{D}_l(\boldsymbol{a}_l^T \boldsymbol{\beta}_{(k)})}{\boldsymbol{a}_l^T \boldsymbol{\beta}_{(k)}} \cdot \boldsymbol{a}_l \boldsymbol{a}_l^T.$$

With $\boldsymbol{A}_\lambda = \sum_{l=1}^{L} \lambda_l \boldsymbol{A}_l$, the penalty is locally quadratically approximated by

$$P_\lambda(\boldsymbol{\beta}) \approx P_\lambda(\boldsymbol{\beta}_{(k)}) + \frac{1}{2}(\boldsymbol{\beta}^T \boldsymbol{A}_\lambda \boldsymbol{\beta} + \boldsymbol{\beta} a_{(k)}^T \boldsymbol{A}_\lambda \boldsymbol{\beta}_{(k)}). \tag{2.3}$$

Hence, the approximation's structure is the same as in Fan and Li (2001) and as in Ulbricht (2010). However, the proposed approach allows to approximate more general norms. In fact, the penalty terms to be approximated do not have to be norms as long as they are meaningful as a penalty. Differentiable penalty terms do not have to be approximated.

The penalized versions of the score function and of the Hessian matrix are $\boldsymbol{s}_{pen}(\boldsymbol{\beta}) = \boldsymbol{s}(\boldsymbol{\beta}) - \boldsymbol{A}_\lambda\boldsymbol{\beta}$ and $\boldsymbol{H}_{pen}(\boldsymbol{\beta}) = \boldsymbol{H}(\boldsymbol{\beta}) - \boldsymbol{A}_\lambda$. By employing the penalized score function, one obtains essentially the same optimization problem as for usual GLMs. When solving the linearized problem $\boldsymbol{s}_{pen}^{lin}(\boldsymbol{\beta}) = \boldsymbol{0}$ iteratively, one obtains $\hat{\boldsymbol{\beta}}_{(k+1)} = \hat{\boldsymbol{\beta}}_{(k)} - \boldsymbol{H}_{pen}(\hat{\boldsymbol{\beta}}_{(k)})^{-1}\boldsymbol{s}_{pen}(\hat{\boldsymbol{\beta}}_{(k)})$. To stabilize the estimation, we use the Fisher information matrix $F(\boldsymbol{\beta}) = -\mathbb{E}(\boldsymbol{H}(\boldsymbol{\beta}))$. The corresponding PIRLS algorithm with step length parameter ν, is

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{(k+1)} &= \hat{\boldsymbol{\beta}}_{(k)} - \nu \cdot (-\boldsymbol{F}(\hat{\boldsymbol{\beta}}_{(k)}) - \boldsymbol{A}_\lambda)^{-1}(\boldsymbol{s}(\hat{\boldsymbol{\beta}}_{(k)}) - \boldsymbol{A}_\lambda\hat{\boldsymbol{\beta}}_{(k)}) \\
&= \hat{\boldsymbol{\beta}}_{(k)} - \nu \cdot (\boldsymbol{F}(\hat{\boldsymbol{\beta}}_{(k)}) + \boldsymbol{A}_\lambda)^{-1}(-\boldsymbol{s}(\hat{\boldsymbol{\beta}}_{(k)}) + \boldsymbol{A}_\lambda\hat{\boldsymbol{\beta}}_{(k)}) \\
&= \hat{\boldsymbol{\beta}}_{(k)} - \nu \cdot (\boldsymbol{X}^T\boldsymbol{W}_{(k)}\boldsymbol{X} + \boldsymbol{A}_\lambda)^{-1}[-\boldsymbol{X}^T\boldsymbol{W}_{(k)} \\
&\quad \underbrace{(\boldsymbol{D}_{(k)}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_{(k)}) + \ \boldsymbol{X}\hat{\boldsymbol{\beta}}_{(k)}}_{\tilde{\boldsymbol{y}}_{(k)}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{(k)}) \ + \boldsymbol{A}_\lambda\hat{\boldsymbol{\beta}}_{(k)}] \\
&= (1-\nu) \cdot \hat{\boldsymbol{\beta}}_{(k)} + \nu \cdot (\boldsymbol{X}^T\boldsymbol{W}_{(k)}\boldsymbol{X} + \boldsymbol{A}_\lambda)^{-1}\boldsymbol{X}^T\boldsymbol{W}_{(k)}\tilde{\boldsymbol{y}}_{(k)}. \qquad (2.4)
\end{aligned}$$

Assuming a simple exponential family for $y_i|\boldsymbol{x}_i$, $i = 1, \ldots, n$, allows to define $\boldsymbol{F}(\hat{\boldsymbol{\beta}}_{(k)}) = \boldsymbol{X}^T\boldsymbol{D}_{(k)}\boldsymbol{\Sigma}_{(k)}^{-1}\boldsymbol{D}_{(k)}\boldsymbol{X} = \boldsymbol{X}^T\boldsymbol{W}_{(k)}\boldsymbol{X}$, $\boldsymbol{D}_{(k)} = \mathrm{diag}(\partial h(\eta_i(\hat{\boldsymbol{\beta}}_{(k)}))/\partial\boldsymbol{\eta})$, and $\boldsymbol{\Sigma}_{(k)} = \mathrm{diag}(\sigma_i^2(\hat{\boldsymbol{\beta}}_{(k)}))$, as well as $\boldsymbol{s}(\hat{\boldsymbol{\beta}}_{(k)}) = \boldsymbol{X}^T\boldsymbol{W}_{(k)}\boldsymbol{D}_{(k)}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_{(k)})$, $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, and $\boldsymbol{\mu}_{(k)} = (h(\boldsymbol{x}_1^T\hat{\boldsymbol{\beta}}_{(k)}), \ldots, h(\boldsymbol{x}_n^T\hat{\boldsymbol{\beta}}_{(k)}))^T$.

Starting with an initial value $\hat{\boldsymbol{\beta}}_{(0)}$, this algorithm is iterated until convergence. The algorithm is terminated when $|\hat{\boldsymbol{\beta}}_{(k+1)} - \hat{\boldsymbol{\beta}}_{(k)}|/|\hat{\boldsymbol{\beta}}_{(k)}| \leq \epsilon$, for a fixed value $\epsilon > 0$. Note that the step length parameter ν, $0 < \nu \leq 1$, equals one in unpenalized Fisher scoring algorithms. Only, if it is necessary, the step length is halved. When the objective function is nonstandard, it can be helpful to work with $\nu < 1$ in order to control the convergence of the algorithm and to avoid backfitting steps.

2.2.3. Some Technical Comments

This section comments on a few properties of the proposed PIRLS algorithm. A similar section with similar properties can be found in Ulbricht (2010). The results presented here, are adjusted and above all, restructured.

Newton-type algorithms are not unconditionally convergent. When the penalized Fisher information matrix $\boldsymbol{X}^T\boldsymbol{W}_{(k)}\boldsymbol{X} + \boldsymbol{A}_\lambda$ is positive definite, the optimization problem is strictly convex and a descent direction in each iteration of algorithm (2.4) is guaranteed. If a solution exists, the algorithm almost surely converges to the optimum, independently of the initial value $\hat{\boldsymbol{\beta}}_{(0)}$. The penalized Fisher information matrix is positive definite, when the Fisher information $\boldsymbol{X}^T\boldsymbol{W}_{(k)}\boldsymbol{X}$ and the penalty matrix $\boldsymbol{A}_\lambda$ are positive definite, or when one of the two matrices is positive and the other is positive semi-definite. In some cases, the penalized Fisher information will be positive definite for a positive semi-definite Fisher information $\boldsymbol{X}^T\boldsymbol{W}_{(k)}\boldsymbol{X}$ and a positive semi-definite penalty matrix $\boldsymbol{A}_\lambda$.

For simple exponential families, like assumed here, the negative log-likelihood $l_n(\boldsymbol{\beta})$ is convex. Hence, when the number of different observations is larger than the number of parameters $(n > q)$, the Fisher information is positive definite. The penalty matrix has to be at least positive semi-definite to assure the global convergence of the algorithm. The penalty is convex when the functions $p_l(\xi)$ and $\|\xi\|_{\mathrm{N}_l}$ are convex for all l, and when $p_l(\xi)$ is monotonically increasing as assumed. For example, this is the case for the (adaptive) Lasso, the fused Lasso or the Ridge penalty. It does not apply to the SCAD penalty or the L_0 penalty of Rippe et al. (2012).

In the $n < q$ case, the Fisher information will be positive semi-definite. In this case, the algorithm's convergence is assured if the penalty matrix is positive definite. When the penalized Fisher information matrix is positive semi-definite, algorithm (2.4) will find descent directions in each iteration. However, it can happen that there are several descent directions in one iteration.

Let $k*$ denote the final iteration of the proposed algorithm, then

$$\boldsymbol{H}_{(k*)} = \boldsymbol{W}_{(k*)}^{T/2}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}_{(k*)}\boldsymbol{X} + \boldsymbol{A}_\lambda)^{-1}\boldsymbol{X}^T\boldsymbol{W}_{(k*)}^{1/2},$$

allows to approximate the generalized hat matrix of a model. The approximated hat matrix is symmetric but not idempotent (Ulbricht, 2010). However, it allows to estimate the degrees of freedom as the trace of the approximated hat matrix:

$$\mathrm{df} = \mathrm{tr}(\boldsymbol{H}_{(k*)}).$$

Note that in contrast to Ulbricht (2010), the estimation of the hat matrix is based upon the algorithm's final iteration only.

In some exponential families, there is a scale parameter $\phi \neq 1$. As ϕ and $\boldsymbol{\beta}$ are orthogonal (see the mixed second derivatives $\frac{\partial l_n}{\partial\phi\partial\boldsymbol{\beta}}$ given in Claeskens and Hjort, 2008), an estimate $\hat{\phi}$ of ϕ can be plugged in with none but the usual restrictions (for example, consistent estimation of $\boldsymbol{\beta}$). That is, the proposed algorithm can be easily extended to quasi-likelihood approaches.

2.2.4. Tuning and Computational Issues

The performance of local quadratic approximations depends on the accuracy of the approximations $\mathcal{N}_l(\xi)$ and of the choice of the penalty parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_L)^T$. Of course, the more precise the approximations $\mathcal{N}_l(\xi)$ are, the more accurate is the proposed algorithm. The choice of the penalty parameters is more complex, because one has to find L possibly different parameters. However, penalized regression requires standardized data or data that is measured on comparable scales. Given that the data is standardized and that the penalty terms are comparable, many approaches employ one global penalty parameter.

Bondell and Reich (2009) illustrate that weighting the penalty terms adequately gives the same effect as standardization of the covariates. This is especially helpful for categorical covariates that are hard to standardize and that can result in many penalty terms. As proposed by Bondell and Reich (2009) and Gertheiss and Tutz (2010), the penalty terms linked to a covariate $\boldsymbol{x}_j$, are weighted such that they are of order k_j, the number of (free) coefficients related to $\boldsymbol{x}_j$. When, for example, a nominal factor with $k_j + 1$ categories is penalized by fused Lasso terms (Gertheiss and Tutz, 2010), all pairwise differences of the k_j related dummy coefficients and of the reference category are penalized absolutely. This results in $\frac{1}{2}(k_j + 1)k_j$ differences. Hence, the difference of the dummy coefficients β_{jl} and β_{jm} is weighted by the factor

$$\frac{2}{k_j + 1}\sqrt{\frac{n_{jl} + n_{jm}}{n}},$$

where n_{jl} and n_{jm} denote the number of observations on the levels l and m of the covariate $\boldsymbol{x}_j$. Weights for other penalties can be derived analogously.

To allow for comparisons with conventional methods, we choose the global penalty parameter λ by cross-validation. In numerical experiments, we employ K-fold cross-validation with the predictive deviance as loss criterion or a generalized cross-validation criterion (GCV) as, for example, defined by O'Sullivan et al. (1986) and as used in the `R` package `mgcv`. Thereby, the predictive deviance is defined as $\text{dev}(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = -\varphi(l_n(\boldsymbol{y}, \hat{\boldsymbol{\mu}}, \varphi) - l_n(\boldsymbol{y}, \boldsymbol{y}, \varphi))$, where $l_n(\cdot)$ denotes the log-likelihood. The GVC criterion is given by $\text{GCV} = n \cdot \text{dev}/(n - \text{df}(\text{model}))^2$, where "dev" denotes the deviance and where "df(model)" is estimated by the trace of the generalized hat matrix of the final iteration of the PIRLS algorithm. Both approaches seem to work reasonable for the proposed approximations.

Even though the proposed algorithm can combine a variety of penalties, it is easy to implement. In principle, it can be combined with any PIRLS algorithm – given, that additional quadratic penalties may depend on the estimates of the last iteration. Except for the approximation of the penalty, the computational complexity is the same as for Fisher scoring algorithms. The penalties mentioned here are implemented in the `R` package `gvcm.cat` (Oelker, 2014).

2.2.5. Extension to Vector-Valued Arguments

The penalties mentioned so far work with linear transformations $\boldsymbol{a}_l^T \boldsymbol{\beta}$ of the coefficient vector. That is, all norms $\|\xi\|_{\text{N}_l}$ have scalar arguments ξ. Penalties employing vectorial norms can be approximated in the same way. For illustration, in this section, the group Lasso (Yuan and Lin, 2006) is considered. The group Lasso penalty for a subvector of coefficients $\boldsymbol{\beta}_l$ is defined as

$$\lambda_l(\boldsymbol{\beta}_l^T \boldsymbol{K}_l \boldsymbol{\beta}_l)^{1/2} = \lambda_l \|\boldsymbol{\beta}_l\|_{\boldsymbol{K}_l}, \tag{2.5}$$

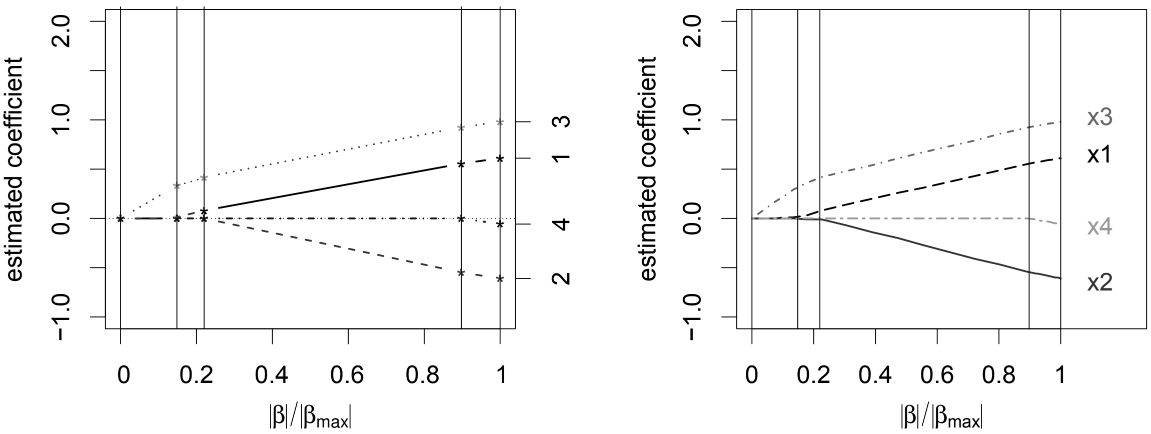


Figure 2.2.: Coefficient paths for a linear model with an intercept and four metric covariates penalized by the Lasso. On the left, the paths are computed by the lars algorithm; on the right, the penalty is approximated quadratically. In both panels, the path related to the intercept is omitted.

where the matrix $\boldsymbol{K}_l \in \mathbb{R}^{r \times r}$ is symmetric and positive (semi-)definite. Typically, it is an identity matrix. The penalty shrinks the coefficients in the vector $\boldsymbol{\beta}_l$ such that either none or the whole group of coefficients is selected. The group Lasso penalty (2.5) can be rewritten as

$$\lambda_l p_l(\|\boldsymbol{R}_l\boldsymbol{\beta}\|_2), \tag{2.6}$$

where $p_l(\xi) = \xi$. The norm $\|\cdot\|_2$ is the Euclidean norm and the matrix $\boldsymbol{R}_l \in \mathbb{R}^{q \times q}$ yields $\boldsymbol{\beta}^T\boldsymbol{R}_l^T\boldsymbol{R}_l\boldsymbol{\beta} = \boldsymbol{\beta}_l^T\boldsymbol{K}_l\boldsymbol{\beta}_l$. $\|\cdot\|_2$ corresponds to the norm $\|\xi\|_{\mathrm{N}_l}$ in penalty (2.2). It can be approximated by $\|\boldsymbol{\xi}\|_2 \approx (\boldsymbol{\xi}^T\boldsymbol{\xi} + c)^{1/2}$, where c is a small positive real number. Following the same schedule as in Subsection 2.2.2, an approximation of the penalty's gradient at $\boldsymbol{\beta} = \boldsymbol{\beta}_{(k)}$ is obtained by:

$$\begin{aligned}\nabla p_l(\left\|\boldsymbol{R}_l\boldsymbol{\beta}_{(k)}\right\|_2) &= \frac{\partial p_l(\left\|\boldsymbol{R}_l\boldsymbol{\beta}_{(k)}\right\|_2)}{\partial\left\|\boldsymbol{R}_l\boldsymbol{\beta}_{(k)}\right\|_2} \cdot \frac{\partial\left\|\boldsymbol{R}_l\boldsymbol{\beta}_{(k)}\right\|_2}{\partial(\boldsymbol{R}_l\boldsymbol{\beta}_{(k)})^T\boldsymbol{R}_l\boldsymbol{\beta}_{(k)}} \cdot \frac{\partial(\boldsymbol{R}_l\boldsymbol{\beta}_{(k)})^T\boldsymbol{R}_l\boldsymbol{\beta}_{(k)}}{\partial\boldsymbol{\beta}_{(k)}} \\ &\approx p_l'(\left\|\boldsymbol{R}_l\boldsymbol{\beta}_{(k)}\right\|_2) \cdot \frac{1}{2((\boldsymbol{R}_l\boldsymbol{\beta}_{(k)})^T\boldsymbol{R}_l\boldsymbol{\beta}_{(k)} + c)^{1/2}} \cdot 2\boldsymbol{R}_l^T\boldsymbol{R}_l\boldsymbol{\beta}.\end{aligned}$$

This yields the approximation

$$\nabla p_l(\left\|\boldsymbol{R}_l\boldsymbol{\beta}_{(k)}\right\|_2)^T(\boldsymbol{\beta} - \boldsymbol{\beta}_{(k)}) \approx \frac{1}{2}(\boldsymbol{\beta}^T\boldsymbol{A}_l^{gr}\boldsymbol{\beta} - \boldsymbol{\beta}_{(k)}^T\boldsymbol{A}_l^{gr}\boldsymbol{\beta}_{(k)}), \tag{2.7}$$

for $\boldsymbol{\beta}_{(k)}$ close to $\boldsymbol{\beta}$ and $\boldsymbol{A}_l^{gr} = p_l'(\left\|\boldsymbol{R}_l\boldsymbol{\beta}_{(k)}\right\|_2)((\boldsymbol{R}_l\boldsymbol{\beta}_{(k)})^T\boldsymbol{R}_l\boldsymbol{\beta}_{(k)} + c)^{-1/2}\boldsymbol{R}_l^T\boldsymbol{R}_l$.
In contrast to so far employed matrices $\boldsymbol{A}_l$, matrix $\boldsymbol{A}_l^{gr}$ is spanned by the product of matrices $\boldsymbol{R}_l^T\boldsymbol{R}_l$ and not by a product of vectors. Expression (2.7) fits exactly into the framework

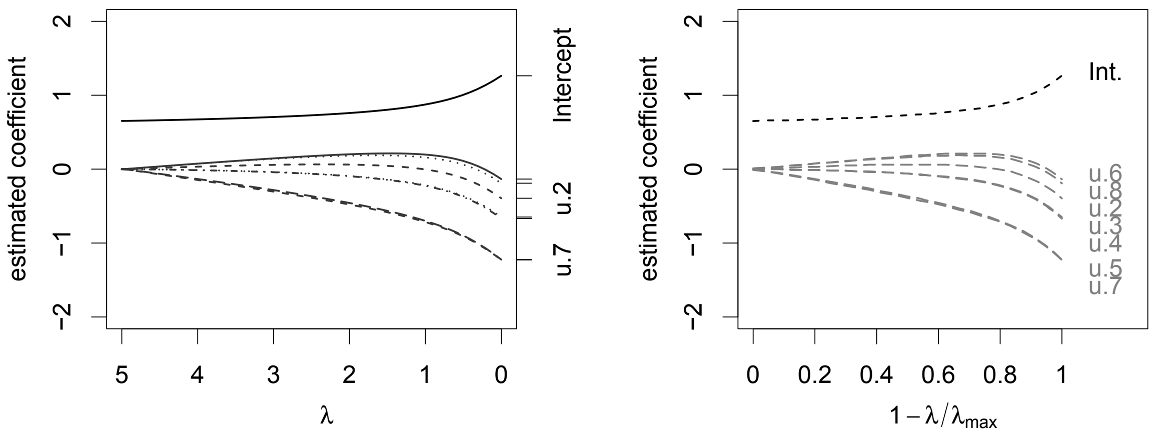


Figure 2.3.: Coefficient paths for a logistic model with an intercept and one ordered factor (eight levels) as covariate. The coefficients are penalized by a group Lasso penalty. On the left, the path is computed with `R` package `grplasso`; on the right, the proposed quadratic approximation is employed.

of approximation (2.3). Penalties of type (2.6) can be added to penalty (2.2) without any problem. To implement, for example, the penalty of Gertheiss et al. (2011), $\boldsymbol{R}_l\boldsymbol{\beta}$ is a vector of differences of coefficients related to an ordinal factor of the form $\beta_{jk} - \beta_{j,k-1}$, $k = 1, \ldots, k_j$. To obtain a penalty term of a comparable order, weights w_l are set to $\sqrt{r_l}$, where r_l denotes the number of differences in vector $\boldsymbol{R}_l\boldsymbol{\beta}$.

2.3. Illustrations

In order to show that the proposed approximations work well, we compare the results for different penalties, including L_0-type penalties and penalized smooth functions, computed by different algorithms with the results of the proposed method.

2.3.1. Comparison of Methods

When the penalty consists of one norm only, one can compare different algorithms with the proposed quadratic approximation. Yet, the results depend on many parameters: On the choice of the tuning parameters for the approximation, on the choice of the penalty parameter λ, on the criterion chosen for cross-validation, on the folds for cross-validation and so on. Hence, in order to judge how the proposed approximation works, we compare the coefficient paths of different penalties (the Lasso, the group Lasso, the elastic net) for different algorithms visually.

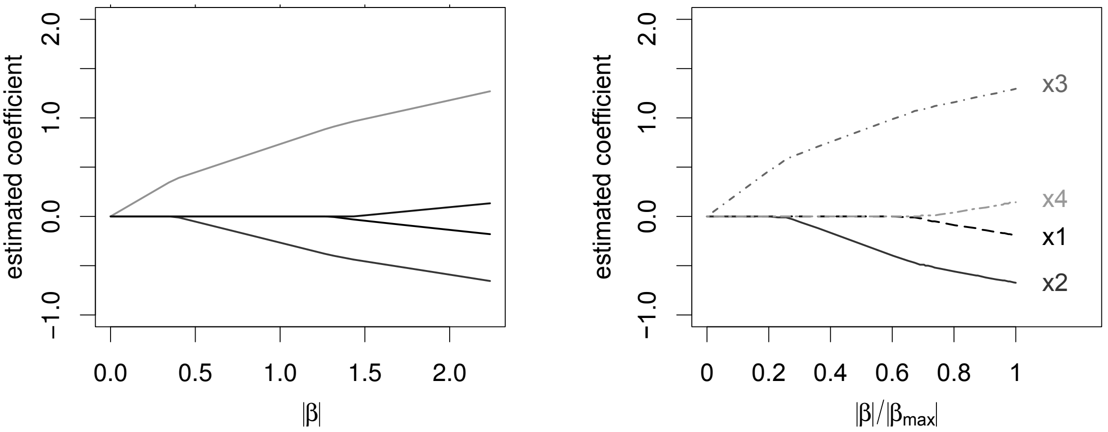


Figure 2.4.: Coefficient paths for a logistic model with an intercept and four metric covariates; each covariate is regularized by an elastic net penalty. On the left, `R` package `glmnet` is employed for computation; on the right, the elastic net penalty is quadratically approximated. In both panels, the path related to the intercept is omitted.

At first, the approximation of the Lasso is compared to the solution of the lars algorithm (`R` package `lars`, Hastie and Efron, 2013). We consider a linear model with four continuous covariates and $n = 400$ observations. The four covariates are drawn from an uniform distribution on $[0, 2]$. The predictor of the model for an observation i is denoted by

$$\eta_i^{lasso} = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + x_{i4}\beta_4, \tag{2.8}$$

where β_0 denotes the intercept of the model. The true vector of coefficients is $\boldsymbol{\beta}^{true} = (0.2, 0.7, -0.5, 1, 0)^T$. That is, there is one non-influential covariate to detect. Figure 2.2 shows the resulting coefficient paths. The left panel shows the solution computed by lars. There are four break points in the piecewise linear coefficient path, each marked by a vertical line. In the right panel, the coefficient path that is obtained with the proposed quadratic approximation is shown. The vertical lines mark the break points of the lars solution. They correspond to the break points of the quadratic approximation. However, the tuning parameter c impacts the penalty. The lars estimate and the PIRLS estimate for a certain value of λ may differ slightly.

In what follows, we assume a logistic model. The true predictor is

$$\eta_i^{logistic} = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + x_{i4}\beta_4 + \boldsymbol{u}_{i5}^T\boldsymbol{\beta}_5,$$

where $\boldsymbol{u}_5$ is an ordered factor with eight levels; it is dummy coded and drawn from a multinomial distribution with equal probabilities for each level. $\boldsymbol{x}_1, \dots, \boldsymbol{x}_4$ are continuous covariates drawn from a uniform distribution on $[0, 2]$. The data generating coefficients are $\boldsymbol{\beta}^{true} = (0.2, 0, -.5, 1, 0, 0.3, 0.7, 0.7, 0.4, 0.4, 0.4, 0.4)^T$. That is, $\boldsymbol{\beta}_5$ is a vector of seven coefficients corresponding to the dummies of $\boldsymbol{u}_5$.

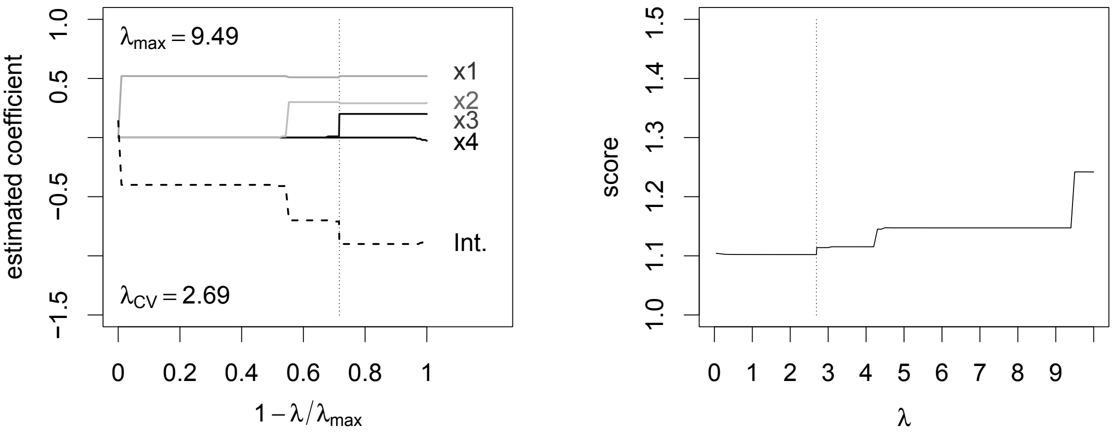


Figure 2.5.: Coefficient paths (left panel) and GCV score (right panel) for a Poisson model with an intercept and four metric covariates; each covariate is regularized by an L_0 penalty. As the GCV score has no unique minimum, λ_{CV} is the maximal penalization parameter that minimizes the GCV score.

We consider two models: In the first one, the predictor is $\eta_i^{group} = \beta_0 + \boldsymbol{u}_{i5}^T \boldsymbol{\beta}_5$. It contains the dummy coded ordered factor only. The dummy coefficients are penalized by a group Lasso penalty. We compare the solution of the coordinate-descent algorithm proposed by Meier et al. (2008) in the R package grplasso (Meier, 2013) with the quadratic approximation. Figure 2.3 shows the coefficient paths. In contrast to Figure 2.2, the path of the intercept is added. The x-axis depends on the (scaled) values of λ instead of $||\boldsymbol{\beta}||_1$. Again, the structure and the range of the two paths are almost identical.
In the second model, the predictor is $\eta_i^{elastic} = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + x_{i4}\beta_4$. That is, the influential factor $\boldsymbol{u}_5$ is ignored. All coefficients are penalized by the elastic net. Figure 2.4 illustrates the resulting coefficient paths. On the left, the paths are computed by the coordinate descent algorithm of Friedman et al. (2010) that is available in the R package glmnet. On the right, the paths that are obtained with the proposed local quadratic approximation are shown. Again, the two solutions coincide.

2.3.2. Penalties Based on the L_0 Norm

Apart from well known penalties like the Lasso or the elastic net that are based on the L_1 norm or on Ridge-type penalties, alternative penalties are made available by our approach. In this section, we consider a model with Poisson distributed responses. The model contains an intercept and four metric covariates. The covariate $\boldsymbol{x}_4$ is non-influential: $\beta^{true} = (-1, 0.5, 0.4, 0.2, 0)^T$. The ideal penalty should uncover that the effect of this co-

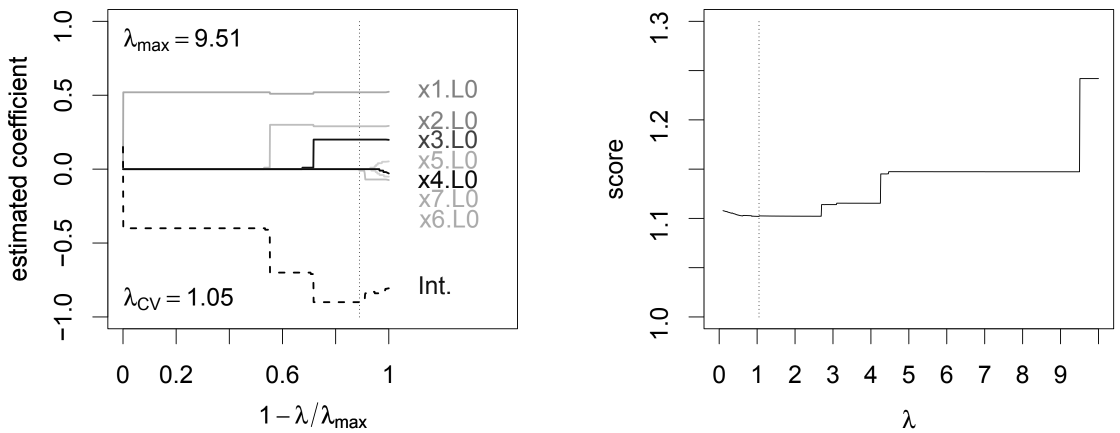


Figure 2.6.: Coefficient paths (left panel) and GCV score (right panel) for a Poisson model with an intercept and seven metric covariates; four covariates are truly non-influential. Each covariate is regularized by an L_0 penalty.

variate is zero – without any shrinkage effects on the other coefficients. Therefore, we use the L_0 penalty

$$P_\lambda(\beta) = \lambda \sum_{l=1}^{4} \|\beta_l\|_0 \,,$$

where $\|\xi\|_0$ denotes $I_{\xi \neq 0}$. This penalty is neither convex nor concave. The solution obtained for a set of initial values $\boldsymbol{\beta}_{(0)}$ does not have to be the global optimum. However, starting, for example, with $\boldsymbol{\beta}_{(0)} = \mathbf{0}^T$ works well. The tuning parameters of the approximation are set to $c = 10^{-5}$ and $\gamma = 50$. In the left panel of Figure 2.5, the coefficient paths for the considered model are shown. The dotted line marks the model chosen by the generalized cross-validation criterion of O'Sullivan et al. (1986). For $\lambda_{CV} = 2.69$, the coefficient related to $\boldsymbol{x}_4$ is zero. The remaining coefficients are not shrunken, the mean squared error is 0.0225 and hence, relatively small. The right panel of Figure 2.5 shows the GCV score. Like the coefficient paths, it is a step function. As the GCV score has no unique minimum, λ_{CV} is defined to be the maximal penalty parameter that minimizes the GCV score.

To challenge the proposed approximation, the above setting is extended by three more non-influential covariates $\boldsymbol{x}_5$, $\boldsymbol{x}_6$ and $\boldsymbol{x}_7$. $\boldsymbol{\beta}^{true}$ is $(-1, 0.5, 0.4, 0.2, 0, 0, 0, 0)^T$. That is, half of the coefficients is truly zero. Figure 2.6 shows the resulting coefficients paths (left) and the GCV score (right). For the optimal model, $\lambda_{CV} = 1.05$. All but one truly zero coefficient are detected ($\hat{\beta}_6 = -0.01$). The mean squared error is 0.0226. For $\lambda \in (1.05, 2.69]$, the true model is detected. As there are only marginal difference in the GCV score for $\lambda_{CV} = 1.05$ and $\lambda = 2.69$, one would probably choose $\lambda_{CV} = 2.69$, and hence, the right model.

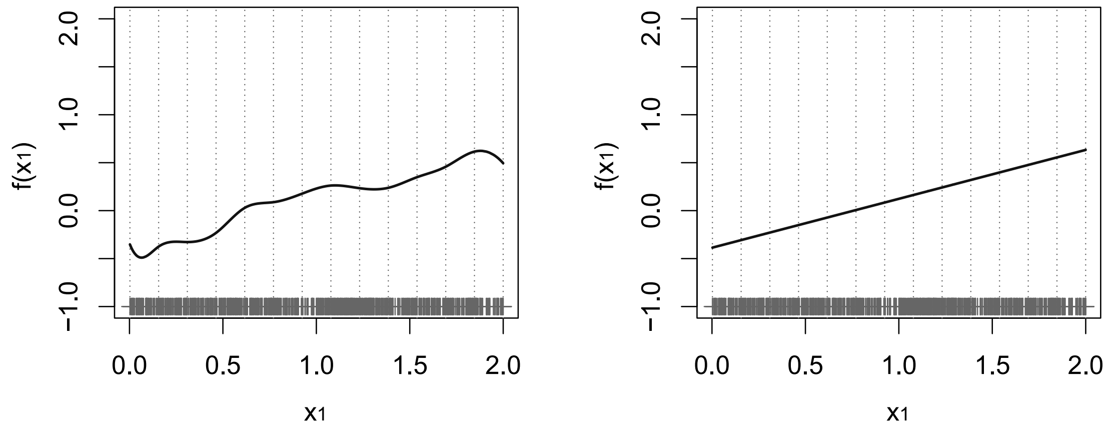


Figure 2.7.: Resulting estimate for the predictor $\eta_i = \beta_0 + f(x_1)$ in a model with Poisson distributed response. The data generating effect of x_1 is linear. $f(x_1)$ is represented by transformed B-spline basis evaluations with 20 knots and penalized like described in Section 2.3.3. On the left, $\lambda_1 = \alpha = 0$; on the right, λ and α are chosen by the GCV criterion. Dotted vertical lines mark the knots of the underlying B-spline basis evaluations. On bottom, the (jittered) observations are marked.

2.3.3. Nonparametric Terms

In many applications, the effect of a continuous covariate is non-linear. One wants to allow for unspecified smooth functions in the predictor. As it is a common choice, we assume that the smooth functions are modeled by penalized cubic B-splines with equidistant knots $\kappa_1, \ldots, \kappa_{M_j}$ as proposed by Eilers and Marx (1996). That is, we assume that $f_j(\boldsymbol{x}_j)$ is represented by $\boldsymbol{B}_j\boldsymbol{\beta}_j$ where $\boldsymbol{B}_j \in \mathbb{R}^{n\times(M_j-4)}$ is the matrix of basis function evaluations, and $\boldsymbol{\beta}_j$ is penalized by

$$\sum_{i=1}^{M_j-6} (\beta_{ji} - 2\beta_{j,i+1} + \beta_{j,i+2})^2 = \boldsymbol{\beta}_j^T(\boldsymbol{\Delta}^2)^T\boldsymbol{\Delta}^2\boldsymbol{\beta}_j, \tag{2.9}$$

where $\boldsymbol{\Delta}^2 \in \mathbb{R}^{(M_j-6)\times M_j}$ denotes the matrix of second order differences with full row rank $M_j - 6$. An attractive approach that centers the smooth function $f_j(\boldsymbol{x}_j) = \boldsymbol{B}_j\boldsymbol{\beta}_j$ for a given set of knots and that offers a decomposition of the function into a linear and a non-linear part is based on the representation of Fahrmeir et al. (2004). The coefficient vector $\boldsymbol{\beta}_j$ is decomposed into a linear part $\boldsymbol{\beta}_j^{lin} = (\beta_j^{int}, \beta_j^{slope})^T$ and into coefficients $\boldsymbol{\beta}_j^{nonlin}$ that model the deviation from the linear trend. One obtains

$$\boldsymbol{\beta}_j = \boldsymbol{\Psi}^{lin}\boldsymbol{\beta}_j^{lin} + \boldsymbol{\Psi}^{nonlin}\boldsymbol{\beta}_j^{nonlin},$$

with

$$\boldsymbol{\Psi}^{lin} = \begin{pmatrix} 1 & \kappa_1 \\ 1 & \kappa_2 \\ \vdots & \vdots \\ 1 & \kappa_{M_j-4} \end{pmatrix}$$

and with $\boldsymbol{\Psi}^{nonlin} = (\boldsymbol{\Delta}^2)^T \left(\boldsymbol{\Delta}^2(\boldsymbol{\Delta}^2)^T\right)^{-1}$. It holds, that $\boldsymbol{\Delta}^2\boldsymbol{\Psi}^{lin} = \mathbf{0}$ and that $\boldsymbol{\Psi}^{lin}\boldsymbol{\Delta}^2 = \mathbf{0}$. That is, $\boldsymbol{\Psi}^{nonlin} \in \mathbb{R}^{(M_j-4)\times(M_j-6)}$ represents the space of penalty (2.9), $\boldsymbol{\Psi}^{lin} \in \mathbb{R}^{(M_j-4)\times 2}$ represents its nullspace. β_j^{int} is incorporated in the (global) intercept β_0. β_j^{slope} and $\boldsymbol{\beta}_j^{nonlin}$ represent the centered smooth function $f_j(\boldsymbol{x}_j)$. Instead of second order differences, the penalty $\sum_{i=1}^{M_j-6}(\beta_{ji}^{nonlin})^2$ is sufficient. Hence, we obtain the same effect as Eilers and Marx (1996) by means of a structured representation with a less complex penalty. Moreover, the decomposition of Fahrmeir et al. (2004) allows to distinguish non-relevant, linear and nonlinear functions more easily by applying a group Lasso penalty on the coefficients $\boldsymbol{\beta}_j^{nonlin}$ and a Lasso penalty on the slope β_j^{slope}. If the total penalty is denoted by $P_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^{p} \lambda_j P_j(\boldsymbol{\beta}_j)$, the penalty that relates to the smooth function $f_j(\boldsymbol{x}_j)$, is given by

$$P_j(\boldsymbol{\beta}_j) = \alpha|\beta_j^{slope}| + (1-\alpha)\left\|\boldsymbol{\beta}_j^{nonlin}\right\|_2, \tag{2.10}$$

where α is an additional penalty parameter that allows to weigh the two parts of the penalty separately. Hence, depending on the tuning, the smooth function f_j can be estimated to be nonlinear, linear or non-influential. The proposed approximation allows for stable estimation of models with this novel penalty.

We consider the same Poisson data as in Section 2.3.2. The impact of all covariates is linear. Even though, we fit a model with the predictor

$$\eta_i = \beta_0 + f(x_{i1}),$$

and penalty (2.10) for $f(\boldsymbol{x}_1)$ which is represented by β^{slope} and $\boldsymbol{\beta}^{nonlin}$. Figure 2.7 shows the resulting functions $f(\boldsymbol{x}_1)$ for $\lambda_1 = \alpha = 0$ (left panel) and for cross-validated tuning (right panel). The effect is detected to be linear.

	Variable	Description
	rent	the rent of the flat, response
1	numbrooms	the number of rooms in the flat, ordered factor with 6 levels, dummy coded, flats with one room are the reference category
2	location	the urban district of the flat, nominal factor with 25 levels, dummy coded, reference is category 1, that is, the city center
3	age	the age (in years) of the flat in 2007, continuous covariate
4	residentialarea	the residential area in square meters, continuous covariate

Table 2.3.: Details on the covariates in the dataset on the rents in Munich (Fahrmeir et al., 2007).

2.4. Rents in Munich

Most major cities and many large communities in Germany conduct surveys in order to construct and publish rental guides. These guides are consulted to determine suitable rents for public and private properties. We are interested in data on the rents of 1488 households in the city of Munich that were collected in 2007 (Fahrmeir et al., 2007). The dataset contains continuous covariates like the flat's size and age, as well as some explanatory factors for a flat's quality and equipment. As the rent is positively skewed, a structured regression model with gamma distributed response and logarithmic link function is assumed. The effect of the age of a flat is known to be non-linear (see, for example, Fahrmeir and Tutz, 2001); it is considered by a spline with a Ridge-type penalty on the effects' curvature. We want to determine whether the effect of the residential area is influential or not. If it is influential, we want to know whether the effect is linear or not. This is reached by the penalty described in Section 2.3.3; it requires a Lasso and a group Lasso penalty. As some levels of the ordered factors are only sparsely occupied, these factors require regularization, too. There are, for example, only few flats with a high number of rooms. We want to employ an adaptive fused Lasso-type penalty on the dummy coefficients of these covariates. Table 2.3 gives the exact definitions of the employed covariates. For an observation i, the predictor is

$$\eta_i \quad = \quad \beta_0 + \boldsymbol{x}_{i1}^T \boldsymbol{\beta}_1 + \boldsymbol{x}_{i2}^T \boldsymbol{\beta}_4 + f_3(x_{i3}) + f_4(x_{i4}),$$

where transposed vectors $\boldsymbol{x}_i^T$ denote covariates that are related to more than one coefficient. The overall penalty is a sum of Ridge-, group Lasso- and Lasso-type penalties. It is denoted by

$$P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{4} P_j(\boldsymbol{\beta}_j),$$

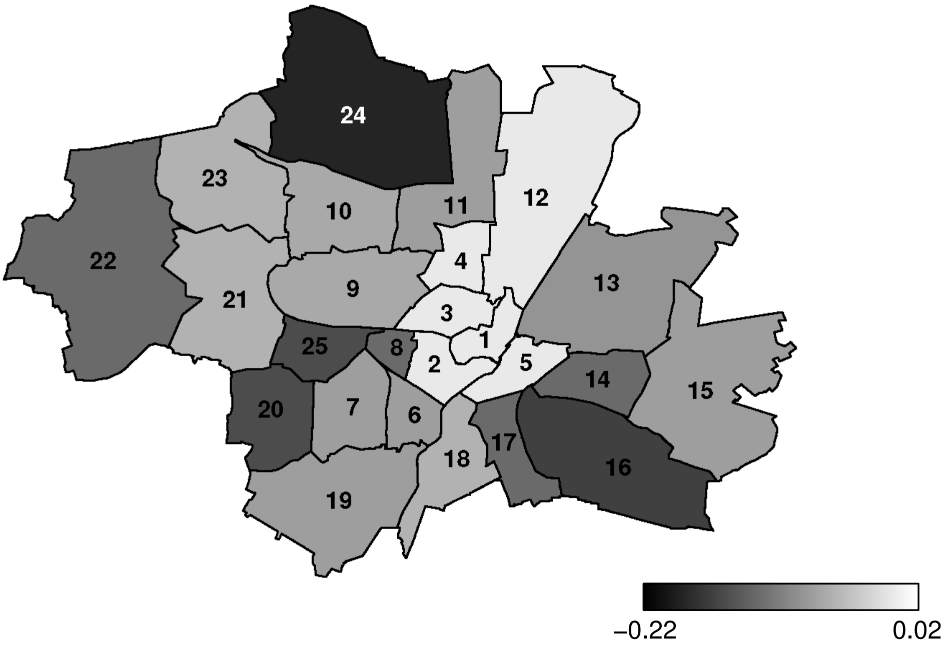


Figure 2.8.: Graphical illustration of the impact of the nominal factor location on the rent. The map depicts the 25 districts of the city of Munich. District 1 denotes the city center which is the reference category; the remaining districts correspond to one dummy coefficient each. The colors are allocated according to the regularized estimates. As the city center is the reference category, its color corresponds to an estimated effect of size zero. The color of the other districts is either the same (districts 2–5 and 12) or darker which corresponds to negative estimates. The fused Lasso penalty detects nine clusters of districts with similar effects. Districts in cluster 1: 18, 21, 23; in cluster 2: 9, 10; in cluster 3: 6, 7, 11, 15, 19; in cluster 4: 8, 14, 17, 22; in cluster 5: 20, 25. The map is provided by the `R` package `BayesX` (Kneib et al., 2014) and reworked with the GNU Image Manipulation Program (GIMP Team, 2012).

where $P_1(\boldsymbol{\beta}_1) = \sum_{r=2}^{6} w_{1r}|\beta_{1r} - \beta_{1,r-1}|$ is the fused Lasso penalty for the ordered factor "numbrooms" with reference $\beta_{11} = 0$. $P_2(\boldsymbol{\beta}_2) = \sum_{r>s} w_{2rs}|\beta_{2r} - \beta_{2s}|$ denotes the penalty for the flats' location. In contrast to P_1, and as the location is a nominal factor, all pairwise differences of coefficients are penalized. Weights w_{1r} and w_{2rs} contain both, the weights that account for the different number of levels and of observations on each level (see Section 2.2.4) and the adaptive weights (see Zou, 2006). The adaptive weights come along with quite huge penalty terms, when the according inverse differences are small. In this case, the penalty terms related to other covariates may become negligible. However, even with the adaptive weights, the penalty terms of different covariates should be comparable. To this end, one can abandon the idea of one global penalty parameter and introduce one penalty parameter λ_1 for the comparable but adaptively weighted penalty terms and one penalty parameter λ_2 for the penalty terms that are not adaptively weighted. $\lambda = (\lambda_1, \lambda_2)$ is then determined by cross-validation. However, to avoid multidimensional cross-validation,

Number	District
1	Altstadt, Lehel (city center)
2	Ludwigvorstadt, Isarvorstadt
3	Maxvorstadt
4	Schwabing West
5	Au, Haidhausen
6	Sendling
7	Sendling, Westpark
8	Schwanthaler Höhe
9	Neuhausen, Nymphenburg
10	Moosach
11	Milbertshofen, Am Hart
12	Schwabing - Freimann
13	Bogenhausen
14	Berg am Laim
15	Trudering, Riem
16	Ramersdorf, Perlach
17	Obergiesing
18	Untergiesing, Harlaching
19	Thalkirchen, Obersendling, Forstenried, Fürstenried, Solln
20	Hadern
21	Pasing, Obermenzing
22	Aubing, Lochhausen, Langwied
23	Allach, Untermenzing
24	Feldmoching, Hasenbergl
25	Laim

Table 2.4.: Overview on the districts in the city of Munich (Fahrmeir et al., 2007). The numbering corresponds to the labels in Figure 2.8 and to the names of the according dummy coefficients.

we propose to rescale the adaptively weighted penalty terms such that the overall penalty of one covariate is again of order k_j, the number of (free) coefficients related to x_j. For the covariate x_1, dedicating the number of rooms in a flat, we have, for example:

$$P_1(\boldsymbol{\beta}) = \frac{1}{\sum_{r=2}^{6} |\beta_{1r}^{ML} - \beta_{1,r-1}^{ML}|^{-1}} \sum_{r=2}^{6} w_{1r}^{level} \frac{|\beta_{1r} - \beta_{1,r-1}|}{|\beta_{1r}^{ML} - \beta_{1,r-1}^{ML}|},$$

where weights $w_{1r}^{level} = \sqrt{\frac{n_{1r}+n_{1,r-1}}{n}}$ adjust for the different number of observations on each level of $\boldsymbol{x}_1$. There is no need to adjust for the number of penalty terms as they are already of order five.

Functions f_3 and f_4 are represented by decomposed cubic B-spline functions based on 20 equidistant knots, see Section 2.3.3. The effect of the flats' age f_3 is penalized by $P_3(\boldsymbol{\beta}_3) = \sum_{r=1}^{14}(\beta_{3r}^{nonlin})^2$, that is, by a Ridge penalty on the curvature of the function. Due

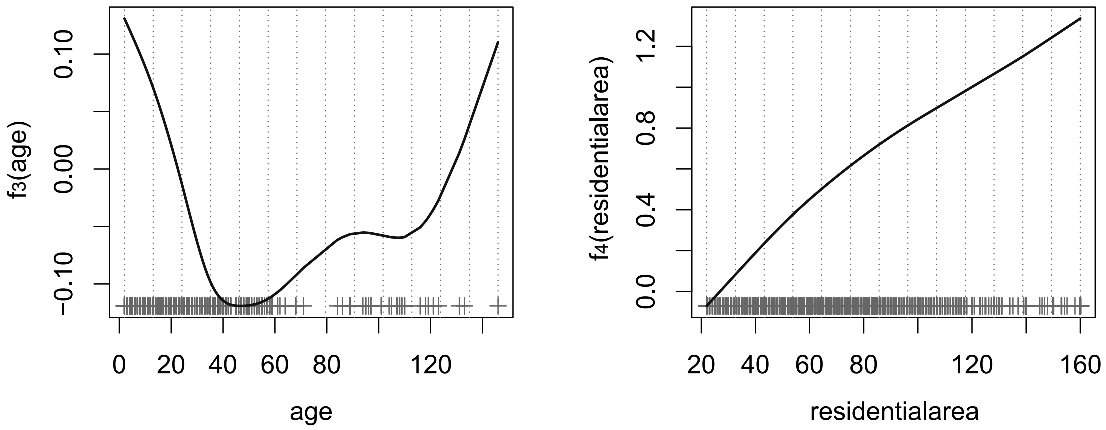


Figure 2.9.: Estimates of functions f_3 and f_4. In the left panel, the impact of the age of a flat is illustrated. The age is measured in years; a flat built in 2007 is aged zero. On the right, the effect of the residential area is shown. The residential area is measured in square meters. The y-axis corresponds to the effect of the age, the residential area, respectively, on the predictors η_i.

to the cubic decomposed B-splines with 20 knots, penalty P_3 relates to 14 coefficients and is of order 14. Hence, no additional weighting is needed. The coefficients related to f_4 are penalized by

$$P_4(\boldsymbol{\beta}_4) = \alpha|\beta_4^{slope}| + (1-\alpha)w_4\sqrt{\sum_{r=1}^{14}(\boldsymbol{\beta}_{4r}^{nonlin})^2},$$

as described in Section 2.3.3, that is, by a Lasso penalty on the linear effect and by a group Lasso penalty on the deviations from this linear effect. Weight w_4 guarantees that the group Lasso penalty is of the right order (see Section 2.2.5). Parameter α is an additional penalty parameter that allows to weight the two components of the penalty. In order to separate it strictly from the global penalty parameter, it is limited to the range $(0, 1)$. Like the global penalty parameter λ, it will be chosen by cross-validation.

In the resulting model, the penalty parameters are chosen by the GCV criterion and set to $(\lambda, \alpha) = (4.55, 0.3)$. It turns out that all covariates affect the response. Figure 2.8 shows how the districts of Munich are clustered by penalty P_2. The map depicts the 25 districts of the city of Munich that are itemized in Table 2.4. District 1 denotes the city center which is the reference category. The remaining districts correspond to one dummy coefficient each. The colors are allocated according to the regularized estimates. As the city center is the reference category, its color corresponds to an estimated effect of size zero. The color of the other districts is either the same (districts 2–5 and 12) or darker. That is, all estimated coefficients are negative. The reason for that is that the reference category "city center" is the most expensive district. With the employed fused Lasso penalty, five clusters of districts with similar effects are detected. The cluster correspond to what one would expect. Figure 2.9 depicts the estimates of the smooth functions f_3 and f_4. The estimated effect of the flats' age is actually non-linear (left panel). It captures the urban development

of Munich. After World War II, many flats were constructed. Flats build subsequent to the war (1945-1965), have a clearly negative impact on the rent. The more lately the flats are constructed the more expensive they become. Flats that where constructed in the beginning of the 20th century (1900-1930), seem to be of a higher value and outbalance the disadvantages of age. A few very old, extensively redecorated flats give the positive effect for flats build in the 19th century. The right panel of Figure 2.9 depicts the effect of the residential area. It is nearly linear. There are only small deviations from a linear trend with slope 0.01. The dummy coefficients for two and three rooms are fused with the reference category "one room". Four and more rooms have a negative impact on the response, the categories for five and six rooms are fused: $\beta_{14} = -0.01$, $\beta_{15} = -0.10$, $\beta_{16} = -0.10$.
Overall, the model seems to give a realistic picture of how the rents are arranged. Especially the effect of the flats' age has a close match in history. Of course, one could argue for many other models. One could spend more time on additional factors. One could think of different penalties, too. For example, the location is so far considered as a nominal factor; all pairwise differences of dummy coefficients are penalized. Instead, the penalty could take the spatial structure into account. One could consider only differences of neighbored districts or weight the differences by the length of their joint boundary.

2.5. Remarks

We propose a general approach to combine different types of penalties in one generalized structured regression model. For example, it allows for penalized smooth functions, Lasso-regularized covariates and categorical covariates, penalized by a group Lasso, in one model. The response can follow any exponential family. This is challenging because the objective function combines various potentially non-differentiable terms like quadratic terms, absolute values or indicators. To solve this problem, we employ a local quadratic approximation for the penalty that is based on ideas of Fan and Li (2001) and Ulbricht (2010). The approximation is iteratively updated in a PIRLS algorithm. That gives an algorithm of similar complexity as for usual GLMs. However, in order to obtain coefficient paths and cross-validation scores, the model has to be evaluated multiple times. An implementation of the algorithm is provided in the `R` package `gvcm.cat` (Oelker, 2014).
The penalty parameters are chosen by cross-validation. We propose a weighting scheme to adjust for differently weighted or scaled covariates. Alternatively, the penalty parameters could be estimated in a mixed model framework. Confidence regions could be constructed by bootstrap methods. As shown in Section 2.2.5, the algorithm can be easily extended to vector valued penalties like the group Lasso.

3. Varying Coefficients with Categorical Effect Modifiers

3.1. Introduction

In regression modeling, the researcher is often faced with categorical covariates, which are also called factors. Nevertheless, variable selection for discrete covariates and the connected problem which categories within one factor are to be distinguished has been somewhat neglected in the literature.

We analyze data from a consumer study on the acceptance of boar meat. As surgical castration of male piglets, as typically done, shall be banned by 2018 (European Declaration on alternatives to surgical castration of pigs, 2010), the production of so-called entire male pigs may become an alternative. To investigate whether this is indeed a suitable alternative (Meier-Dinkel et al., 2013), we consider meat from four different product groups: (1) castrate or gilt meat with label "pork", (2) castrate or gilt meat with label "young boar meat", (3) boar meat with label "pork", and (4) boar meat with label "young boar meat". The response is binary saying whether consumers liked the taste of the meat or not, see Meier-Dinkel et al. (2013). We investigate whether the probability of liking depends on the product group, and furthermore, if the influence of other variables like the gender, the age or the health status (sick: yes/no) on liking depends on the product group. Therefore, the product group is considered as an effect modifying factor. That is, we address model selection with discrete covariates in a slightly extended version of generalized linear models (GLMs), namely, in GLMs with varying coefficients and categorical effect modifiers.

This chapter is a modified version of Oelker, Gertheiss, and Tutz (2014). The original version of Oelker, Gertheiss, and Tutz (2014) is published in Statistical Modelling, Vol. 14, No. 2. Copyright © 2014 SAGE Publications. All rights reserved. Reproduced with the permission of the copyright holders and the publishers Sage Publications India Pvt. Ltd, New Delhi. Oelker, Gertheiss, and Tutz (2012b) is an early version of this chapter. For more information on the contributions of the authors and on textual matches, see page 4.

In a second application, we model the effects of pregnancy related covariates on the type of delivery, that is, whether birth was given vaginally or by means of a Cesarean. The data is presented by Boulesteix (2006). As medical standards typically change over time, modeling the type of delivery requires to consider (discrete) time-effects, and more importantly, to consider how the effects of the covariates change over the years. Thus, we are interested in the categorical effect modifier time in years in a logit model.

Varying-coefficient models (Hastie and Tibshirani, 1993) are a quite flexible tool to capture complex model structures and interactions. Regression coefficients β_j are allowed to vary with the value of other variables $\boldsymbol{u}_j$. Hence, the linear predictor $\boldsymbol{\eta}$ in a GLM has the form

$$\boldsymbol{\eta} = \beta_0(\boldsymbol{u}_0) + \boldsymbol{x}_1\beta_1(\boldsymbol{u}_1) + \ldots + \boldsymbol{x}_p\beta_p(\boldsymbol{u}_p), \tag{3.1}$$

where $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ are continuous covariates, and $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p$ are the so called effect modifiers, which modify the effects of the covariates in an unspecified, typically smooth, form $\beta_j(\cdot)$. Thus, the predictor is still linear in the regressors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$, but scalar coefficients β_j turn into functions depending on the effect modifiers $\boldsymbol{u}_j$, $j = 0, \ldots, p$. As common in GLMs, it is assumed that the predictor $\boldsymbol{\eta}$ is linked to the conditional mean of the response $\boldsymbol{y}$ by a known response function h. That is, $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{y}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) = h(\boldsymbol{\eta})$, and $\boldsymbol{y}$ follows a simple exponential family. Throughout Chapter 3, we assume that the covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ are measured on comparable scales or have been scaled.

For continuous effect modifiers, unknown functions $\beta_j(\cdot)$ are typically assumed to be smooth, and they have been modeled by splines (Hastie and Tibshirani, 1993; Hoover et al., 1998; Lu et al., 2008), using localizing techniques (Wu et al., 1998; Fan and Zhang, 1999; Kauermann and Tutz, 2000) or boosting methods (Hofner et al., 2013). Inference requires to distinguish between varying and non-varying coefficients, and between relevant and non-relevant terms. Hastie and Tibshirani (1993) propose to adopt techniques for additive models. Leng (2009) distinguishes between varying and non-varying coefficients by applying the component selection and smoothing operator (Cosso; Lin and Zhang, 2006), while Wang et al. (2008) obtain the selection of spline coefficients by groupwise SCAD penalization. Wang and Xia (2009) select covariates by local polynomial regression models penalized by the group Lasso (Yuan and Lin, 2006). However, apart from Hofner et al. (2013), the selection of covariates and the identification of smooth or constant functions is not reached simultaneously.

In contrast to most existing approaches, we consider categorical effect modifiers $\boldsymbol{u}_j \in \{1, \ldots, k_j\}$. In the boar data, for example, the effect modifier product group has four categories. Functions $\beta_j(\boldsymbol{u}_j)$ have the form $\sum_{r=1}^{k_j} \beta_{jr} I(\boldsymbol{u}_j = r)$, where $I(\cdot)$ denotes the indicator function and where $\beta_{j1}, \ldots, \beta_{jk_j}$ represent the regression parameters. Therefore, the linear predictor is given by

$$\boldsymbol{\eta} = \sum_{r=1}^{k_0} \beta_{0r} I(\boldsymbol{u}_0 = r) + \sum_{j=1}^{p} \boldsymbol{x}_j \sum_{r=1}^{k_j} \beta_{jr} I(\boldsymbol{u}_j = r).$$

The total coefficient vector is given by $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_0^T, \dots, \boldsymbol{\beta}_p^T)$, where a sub-vector $\boldsymbol{\beta}_j^T = (\beta_{j1}, \dots, \beta_{jk_j})$ contains the parameters of the jth covariate. With categorical effect modifiers, the number of parameters $q = \sum_{j=0}^{p} k_j$ can become very large, even for a moderate number of covariates p. Hence, usual maximum likelihood (ML) estimates may not exist. Alternative tools such as regularization techniques are needed. Moreover, it is desirable to reduce the model to the relevant terms. One wants to determine which coefficients are influential, and if so, which categories have to be distinguished.

The methods proposed here extend the work of Gertheiss and Tutz (2012), which is restricted to the classical linear model, and hence, cannot be used for analyzing data with non-normal response variables such as in the boar data. Hence, we present approaches that allow to model categorical effect modifiers within the GLM framework. In Section 3.2, we propose a penalized ML criterion. For the computation of the estimates, a different approach than in the classical linear model is needed; a penalized iteratively reweighted least squares (PIRLS) algorithm as described in Chapter 2 is employed. Moreover, large sample properties of the penalized estimator are derived. As a competing approach, we consider a forward selection procedure employing information criteria (Section 3.3). The proposed methods are shown to be competitive in numerical experiments (Section 3.4). In Section 3.5, the approaches are applied to the boar data. In Section 3.6, the birth data is analyzed. The special case of categorical effects is discussed in Section 3.7.

3.2. L_1 Penalized Estimation in GLMs

The main tool for regularization and model selection is the use of penalties. In GLMs, penalized estimation means to minimize

$$\mathcal{M}_n^{pen}(\boldsymbol{\beta}) = -l_n(\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta}) = -l_n(\boldsymbol{\beta}) + \lambda \cdot J_n(\boldsymbol{\beta}), \tag{3.2}$$

where $l_n(\boldsymbol{\beta})$ denotes the log-likelihood for n observations, and $P_\lambda(\boldsymbol{\beta})$ stands for a general penalty depending on the penalty parameter λ. The expression $\lambda \cdot J_n(\boldsymbol{\beta})$ breaks the penalty down to a product, underlining the dependency on only one scalar penalty parameter. With $\lambda = 0$, ordinary ML estimation is obtained.

The main issue is to choose an adequate penalty $J_n(\boldsymbol{\beta})$. The Ridge penalty (Hoerl and Kennard, 1970), for example, shrinks the coefficients, while the Lasso (Tibshirani, 1996) combines the shrinkage and the selection of coefficients. The fused Lasso (Tibshirani et al., 2005) applies the Lasso to differences of adjacent parameters. Thus, parameters are shrunk towards each other and potentially fused in order to gain a local consistent profile of ordered coefficients. The group Lasso (Yuan and Lin, 2006) selects whole groups of coefficients simultaneously. Although variable selection is implied, both the Lasso and its grouped version are off target as they do not enforce $\beta_{jr} = \beta_{js}$ for some $r \neq s$. The pure fused Lasso indeed leads to (piecewise) constant functions $\beta_j(\boldsymbol{u}_j)$, but disregards the selection of whole covariates. A combination of both allows not only for the shrinkage and the selection, but also for the gradual fusion of related coefficients – such that the effects of the group Lasso are embedded.

As nominal and ordinal effect modifiers in (3.1) contain a different amount of information, they should be treated differently. We consider the general penalty

$$J_n(\boldsymbol{\beta}) = \sum_{j=0}^{p} J_j(\boldsymbol{\beta}_j), \tag{3.3}$$

where $J_j(\boldsymbol{\beta}_j) = 0$ if covariate j is not modified, $J_j(\boldsymbol{\beta}_j) = J_j^{nom}(\boldsymbol{\beta}_j)$ for nominal effect modifiers and $J_j(\boldsymbol{\beta}_j) = J_j^{ord}(\boldsymbol{\beta}_j)$ for ordinal effect modifiers.

For a *nominal* effect modifier $\boldsymbol{u}_j$, we propose

$$J_j^{nom}(\boldsymbol{\beta}_j) = \sum_{r>s} |\beta_{jr} - \beta_{js}| + b_j \sum_{r=1}^{k_j} |\beta_{jr}|, \tag{3.4}$$

where b_j is an indicator that (de-)activates the second sum if wanted. The first sum in penalty (3.4) is equivalent to a fused Lasso penalty applied on all pairwise differences of coefficients belonging to $\beta_j(\boldsymbol{u}_j)$. Thus, not only adjacent coefficients but each subset of nominal categories can be collapsed. In the case of strong penalization, the effects $\beta_{j1}, \ldots, \beta_{jk_j}$ of covariate j are reduced to one constant coefficient, and do not depend on the categories of $\boldsymbol{u}_j$ any more. One obtains $\hat{\beta}_{j1} = \ldots = \hat{\beta}_{jk_j} = \hat{\beta}_j$. The second sum in (3.4) conforms to a Lasso penalty shrinking all coefficients belonging to $\beta_j(\boldsymbol{u}_j)$ individually toward zero. The effect is the selection of covariates. For strong penalization, $\hat{\beta}_{j1} = \ldots = \hat{\beta}_{jk_j} = 0$ is obtained, and covariate j is excluded. In most cases, a constant intercept shall remain in the model. Thus, we typically have $b_0 = 0$.

If $\boldsymbol{u}_j$ is *ordinal*, there is additional information. We propose to fuse the adjacent coefficients β_{jr} and $\beta_{j,r-1}$. Hence, for ordinal factors, we employ the penalty

$$J_j^{ord}(\boldsymbol{\beta}_j) = \sum_{r=2}^{k_j} |\beta_{jr} - \beta_{j,r-1}| + b_j \sum_{r=1}^{k_j} |\beta_{jr}|, \tag{3.5}$$

where b_j denotes the same indicator as above. Instead of all pairwise differences, now only differences of neighbored coefficients are penalized. That corresponds exactly to a fused Lasso-type penalty (Tibshirani et al., 2005). Again, with setting $b_0 = 0$, the intercept can be treated separately.

Apart from their different amount of information, J_j^{nom} and J_j^{ord} work similarly: One term leads to the fusion within the covariate, while a Lasso-type penalty selects the coefficients. Variable selection as well as the distinction of varying/non-varying coefficients is obtained.

If, for example, emphasis should be put on the selection of covariates, it may be advantageous to use weights for the two components of the penalty (compare Tibshirani et al., 2005). With parameter $\psi \in (0,1)$, the weighted penalty for a nominal effect modifier $\boldsymbol{u}_j$ is

$$J_j^{nom}(\boldsymbol{\beta}, \psi) = \psi \sum_{r>s} |\beta_{jr} - \beta_{js}| + (1-\psi) b_j \sum_{r=1}^{k_j} |\beta_{jr}|. \tag{3.6}$$

For an ordinal effect modifier, it is

$$J_j^{ord}(\boldsymbol{\beta}, \psi) = \psi \sum_{r=2}^{k_j} |\beta_{jr} - \beta_{j,r-1}| + (1-\psi) b_j \sum_{r=1}^{k_j} |\beta_{jr}|. \tag{3.7}$$

The parameter ψ is restricted to $(0,1)$ in order to separate it strictly from the penalty parameter λ. If the effect modifiers $\boldsymbol{u}_j$ have different numbers of categories, additional weighting of the penalty terms analogously to Bondell and Reich (2009) could be used in order to prevent a potential selection bias.

3.2.1. Computational Issues

Since penalty (3.3) contains absolute values, a convex but not continuously differentiable optimization problem has to be solved. In the classical linear model, quadratic programming can be used to tackle this problem, or the solution can be approximated by employing the lars algorithm (Efron et al., 2004), see Gertheiss and Tutz (2012) for details. In a GLM, however, a more general approach is needed. Non-differentiability can be evaded by approximating the penalty at the critical points, that is, in a neighborhood of $|\xi|$, $\xi = 0$. As, for example, in Koch (1996), the absolute values $|\xi|$ in the penalty are approximated by the differentiable function $\sqrt{\xi^2 + c}$, where c denotes a small, positive constant. Combining this approximation with a local approximation of Fan and Li (2001) and with an idea of Ulbricht (2010) allows to derive a PIRLS algorithm like described in Chapter 2.

The generalized hat matrix of the final iteration of this algorithm allows to estimate the model's degrees of freedom. However, the algorithm, is only locally convergent. Only if the objective function is strictly convex, the global optimum is found almost surely. Strict convexity implies that the penalized Fisher information matrix is positive definite. The penalty applied here leads to a positive semi-definite penalty matrix. Therefore, in the $n > q$ case, the quasi-Newton approach will find descent directions in each iteration; but for the $q > n$ case, it may happen that the solution is not unique (Ulbricht, 2010). In this case, we recommend to use several starting values and to check the likelihood scores of the according solutions. However, in our experience, this is a minor problem.

3.2.2. Large Sample Properties

For asymptotics, general assumptions have to hold and the number of observations has to grow in accordance with the requirements of categorical covariates: If the sample size n tends to infinity, it is assumed that the number of observations n_{jr} on level r of $\boldsymbol{u}_j$ tends to infinity for all j, r at the same rate. Practically, that means, that asymptotically the probability for an observation on level r of $\boldsymbol{u}_j$ must be positive and tend to a constant c_{jr} for all j, r. Let $\boldsymbol{\beta}^*$ denote the true value of $\boldsymbol{\beta}$. Then the following theorem holds:

Theorem 1. *Suppose* $0 \leq \lambda < \infty$ *has been fixed, and all class-wise sample sizes* n_r *satisfy* $n_{jr}/n \to c_{jr}$*, where* $0 < c_{jr} < 1$*. Then the estimate* $\hat{\boldsymbol{\beta}}$ *that minimizes (3.2) with* $J_n(\boldsymbol{\beta})$ *defined by (3.3), (3.4) and (3.5) is consistent, that is,* $\lim_{n\to\infty} \mathbb{P}(||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||^2 > \epsilon) = 0$ *for all* $\epsilon > 0$*.*

The proof is given in Appendix A. Employing the generalized versions (3.6) and (3.7) does not affect the consistency results.

As pointed out in Zou (2006), regularization as used so far does not ensure consistency in terms of variable selection. To gain selection consistency, Zou (2006) proposes an adaptive version of the original Lasso that has the so-called oracle properties. A corresponding modification for penalty (3.3) is available: Given effect modifiers $\boldsymbol{u}_j$, $j = 1, \ldots, p$, penalty (3.3) is modified to the adaptive penalty $J_n^{ad}(\boldsymbol{\beta})$ by employing

$$J_j^{ad,nom}(\boldsymbol{\beta}) = \sum_{r>s} w_{rs(j)} |\beta_{jr} - \beta_{js}| + b_j \sum_{r=1}^{k_j} w_{r(j)} |\beta_{jr}| \text{ and} \tag{3.8}$$

$$J_j^{ad,ord}(\boldsymbol{\beta}) = \sum_{r=2}^{k_j} w_{r,r-1(j)} |\beta_{jr} - \beta_{j,r-1}| + b_j \sum_{r=1}^{k_j} w_{r(j)} |\beta_{jr}|, \tag{3.9}$$

which replace (3.4) and (3.5), and by using the adaptive weights

$$w_{rs(j)} = \phi_{rs(j)}(n) |\hat{\beta}_{jr}^{ML} - \hat{\beta}_{js}^{ML}|^{-1} \text{ and} \tag{3.10}$$

$$w_{r(j)} = \phi_{r(j)}(n) |\hat{\beta}_{jr}^{ML}|^{-1}. \tag{3.11}$$

Here, $\hat{\beta}_{jr}^{ML}$ denotes the ML estimate of β_{jr}; functions $\phi_{rs(j)}(n)$ and $\phi_{r(j)}(n)$ are additional weights for the penalty terms, that are assumed to converge to fixed values: $\phi_{rs(j)}(n) \to q_{rs(j)}$ and $\phi_{r(j)}(n) \to q_{r(j)}$, with $0 < q_{rs(j)}$, $q_{r(j)} < \infty$. If $\phi_{rs(j)}(n) = \phi$ and $\phi_{r(j)}(n) = 1 - \phi$, $0 < \phi < 1$, are global constants, we obtain a generalization with the same structure as given in equations (3.6) and (3.7); $0 < \phi < 1$ or similar constraints for functions $\phi_{rs(j)}(n)$, $\phi_{r(j)}(n)$ guarantee that the effect of the weights and the effect of the global penalty parameter λ are separated. The adaptive weight of a penalty term becomes huge when the ML estimate of the penalty term is close to zero. The adaptive weight becomes the smaller, the bigger the ML estimate of the penalty term is. Thus, adaptive weights favor to set coefficients with small ML estimates to zero, to fuse coefficients with close ML estimates respectively. Technically, with some additional assumptions, this ensures selection consistency. First of all, the penalty parameter λ has to increase with the sample size n. One assumes that $\lambda = \lambda_n$ with $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n \to \infty$, and that all class-wise sample sizes n_r satisfy $n_r/n \to c_r$, where $0 < c_r < 1$.

In addition, we define some vectors: $\hat{\boldsymbol{\beta}}^n$ denotes the estimate of $\boldsymbol{\beta}$; we emphasize that it is based on the sample size n. We define the block-diagonal matrix $\boldsymbol{A} = (\boldsymbol{a}_1, \dots, \boldsymbol{a}_L) \in \mathbb{R}^{q\times L}$, where vectors $\boldsymbol{a}_l$ are defined as in Chapter 2. Then, the vector $\hat{\boldsymbol{\theta}}^n = \boldsymbol{A}^T\hat{\boldsymbol{\beta}}^n$ contains the estimates of all the terms in penalty (3.3). That is, the estimated values of all penalized coefficients $\hat{\beta}_{rj}$ and – according to the level of measurement – the estimated values of their differences. Furthermore, we define $\mathcal{C}$ and $\mathcal{C}_n$. $\mathcal{C}$ denotes the set of indices corresponding to those entries of $\hat{\boldsymbol{\theta}}^n$ which are truly non-zero; whereas $\mathcal{C}_n$ denotes the estimate of $\mathcal{C}$ based on n observations. $\boldsymbol{\theta}^*_{\mathcal{C}}$ is the vector with the true values of the entries in $\mathcal{C}$; $\hat{\boldsymbol{\theta}}^n_{\mathcal{C}}$ denotes its estimate.

Previous assumptions concerning ML estimation are extended: The model must hold, the negative log-likelihood $-l_n(\boldsymbol{\beta})$ has to be convex, $l_n(\boldsymbol{\beta})$ has to be at least three times continuously differentiable, and the third moments of $\boldsymbol{y}$ have to be finite. Let $\boldsymbol{F}_n = \mathbb{E}\left(-\frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right)$ denote the expected information matrix, then $\boldsymbol{F}_n/n$ must have a positive definite limit $\boldsymbol{F}$. For the score function $\boldsymbol{s}_n(\boldsymbol{\beta}) = \frac{\partial l_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}$, we suppose $\mathbb{E}(\boldsymbol{s}_n(\boldsymbol{\beta})) = \boldsymbol{0}$. Then we obtain

Theorem 2. *Suppose $\lambda = \lambda_n$ with $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n \to \infty$, and all class-wise sample sizes n_{jr} satisfy $n_{jr}/n \to c_{jr}$, where $0 < c_{jr} < 1$. Then penalty $J_n^{ad}(\boldsymbol{\beta})$ employing terms (3.8) and (3.9) with weights (3.10) and (3.11), where $\hat{\beta}_{jr}^{ML}$, $\phi_{rs(j)}(n)$ and $\phi_{r(j)}(n)$ are defined as above, ensures that*

(a) $\sqrt{n}(\hat{\boldsymbol{\theta}}^n_{\mathcal{C}} - \boldsymbol{\theta}^*_{\mathcal{C}}) \xrightarrow{d} N(\boldsymbol{0}, \mathrm{Cov}(\boldsymbol{\theta}^*_{\mathcal{C}}))$,

(b) $\lim_{n\to\infty} \mathbb{P}(\mathcal{C}_n = \mathcal{C}) = 1$.

The proof uses ideas from Zou (2006) and Bondell and Reich (2009), and is given in Appendix A. The concrete form of $\mathrm{Cov}(\boldsymbol{\theta}^*_{\mathcal{C}})$ results from the asymptotic marginal distribution of a set of non-redundant truly non-zero regression parameters or differences thereof. Since all estimated differences are (deterministic) linear functions of estimated parameters, the covariance-matrix $\mathrm{Cov}(\boldsymbol{\theta}^*_{\mathcal{C}})$ is singular.

$\boldsymbol{F}_n/n \overset{n\to\infty}{\to} \boldsymbol{F}$ with positive definite $\boldsymbol{F}$ is typically assumed in observational studies, but it raises problems in experiments. In this case, the given proof can be extended to matrix normalization (see, for example, Fahrmeir and Kaufmann, 1985).

For $\lambda = 0$, the unpenalized likelihood is maximized. Therefore, asymptotic normality and consistency hold as shown by McCullagh (1983). Distributional properties for $n \to \infty$ for a fixed λ are not discussed as the penalty shall not vanish in proportion to $-l_n(\boldsymbol{\beta})$ for $n \to \infty$.

For the normality part of Theorem 2, the speed of convergence is $\lambda_n/\sqrt{n} \rightarrow 0$. Since $n^{-1/2}s_n(\boldsymbol{\beta}) \sim N(\boldsymbol{0}, \boldsymbol{F}_n(\boldsymbol{\beta})/n) + \mathcal{O}_p(n^{-1/2})$ and $\mathbb{P}(\sqrt{n}|\hat{\boldsymbol{\beta}}_{jr}^{ML}| \leq \lambda_n^{1/2}) \rightarrow 1$ like $c/\sqrt{n} \rightarrow 0$, the consistency part behaves the same. Thus, the overall speed of convergence is $\mathcal{O}_p(n^{-1/2})$.
In some cases, in particular for small sample sizes, the ML estimates that is required for the adaptive weighting may not exist. If necessary, the ML estimates can be replaced by other $\sqrt{n}$-consistent estimates, for example, Ridge estimates with a small, fixed penalty parameter. However, adaptive estimation is as good as the employed weights and hence, not recommended by all means.

3.3. Alternative Selection Strategies

Stepwise procedures are often used for the selection of variables. In particular, forward and backward selection methods based on information criteria like the Akaike information criterion (AIC; see, for example, Bozdogan, 1987) or the Bayesian information criterion (BIC; Schwarz, 1978) are popular. One tries to find the model that performs the best with respect to the chosen criterion. By construction, these strategies yield variable selection but not the fusion of categories. Gertheiss and Tutz (2012) obtain the fusion of categories by using an enlarged setting. For example, for a nominal effect modifier $\boldsymbol{u}_j$ with three categories that modifies the covariate $\boldsymbol{x}_j$, the varying coefficient $\beta_j(\boldsymbol{u}_j)$ corresponds to the sub-vector $(\beta_{j1}, \beta_{j2}, \beta_{j3})^T$ in the coefficient vector $\boldsymbol{\beta}$. All possible subsets of coefficients belonging to $\boldsymbol{x}_j$ would be: $\{()$, (β_{j1}), (β_{j2}), (β_{j3}), (β_{j1}, β_{j2}), (β_{j1}, β_{j3}), (β_{j2}, β_{j3}), $(\beta_{j1}, \beta_{j2}, \beta_{j3})\}$. Allowing for the fusion of coefficients increases the number of possibilities by $\{(\beta_{j1}, \beta_{j2} = \beta_{j3})$, $(\beta_{j2}, \beta_{j1} = \beta_{j3})$, $(\beta_{j3}, \beta_{j2} = \beta_{j1})$, $(\beta_{j1} = \beta_{j2} = \beta_{j3})\}$. When selecting a model, all possibilities to select and/or fuse coefficients must be considered.
Concretely, we start with a model containing an intercept only. In each step, the degrees of freedom of the model are enlarged by one until the chosen criteria (AIC or BIC) is not improved anymore – whereat the degrees of freedom are defined as the number of non-zero coefficient blocks in $\hat{\boldsymbol{\beta}}$ (Tibshirani et al., 2005). Hence, in each step, a former zero coefficient can be set to non-zero, or an entire group of zero coefficients can become non-zero, but with all coefficients within this group being equal. Alternatively, a group of non-zero but identical coefficients can be split into two groups of non-zero coefficients, with coefficients now being identical within each of both groups but different between the two groups.

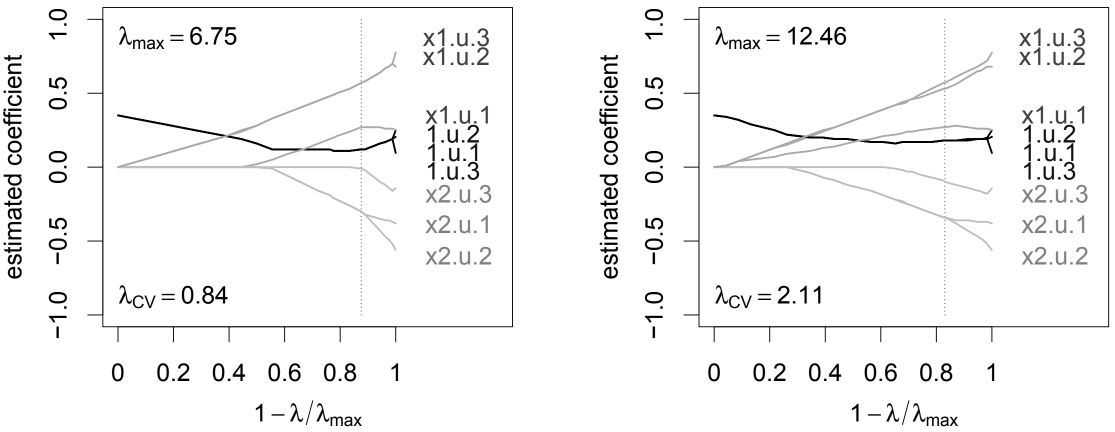


Figure 3.1.: Coefficient paths for the binary model (3.12) with predictor (3.13) – estimated with an adaptively weighted penalty (left panel) and with the default penalty (right panel).

3.4. Illustration and Numerical Experiments

For illustration, we start with a simple example. Assume a logistic regression model with two covariates $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and one nominal effect modifier $\boldsymbol{u}$ with categories 1, 2 and 3. $\boldsymbol{u}$ possibly impacts all covariates plus the intercept. Concretely, the predictor is

$$\begin{aligned}\boldsymbol{\eta}_{true} &= \beta_0(\boldsymbol{u}) + \boldsymbol{x}_1\beta_1(\boldsymbol{u}) + \boldsymbol{x}_2\beta_2(\boldsymbol{u}) \qquad (3.12)\\ &= \beta_0 + \boldsymbol{x}_1\,(\,\beta_{11}I(\boldsymbol{u}=1) + \beta_{12}I(\boldsymbol{u}=2) + \beta_{13}I(\boldsymbol{u}=3)\,) + \boldsymbol{x}_2\beta_2\\ &= 0.2 + \boldsymbol{x}_1\,(\,0.3I(\boldsymbol{u}=1) + 0.7I(\boldsymbol{u}=2) + 0.7I(\boldsymbol{u}=3)\,) - \boldsymbol{x}_2\cdot 0.5.\end{aligned}$$

That means, while the intercept and $\boldsymbol{x}_2$ do not depend on $\boldsymbol{u}$, covariate $\boldsymbol{x}_1$ varies with categories 1 and 2/3 of $\boldsymbol{u}$. Covariates $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are independently drawn from an uniform distribution $U[0,2]$; the effect modifier $\boldsymbol{u}$ is multinomial with probabilities 0.3, 0.4, 0.3 for categories 1, 2 and 3, respectively. For the response $\boldsymbol{y}$, $\boldsymbol{y} = h(\boldsymbol{\eta})$ holds, where $h^{-1}(\cdot)$ is the canonical link (logit) function. We generate $n = 400$ observations. When fitting the model, all coefficients are allowed to vary with effect modifier $\boldsymbol{u}$, that is, we have

$$\eta_{model} = \beta_0(\boldsymbol{u}) + \boldsymbol{x}_1\cdot\beta_1(\boldsymbol{u}) + \boldsymbol{x}_2\cdot\beta_2(\boldsymbol{u}). \qquad (3.13)$$

Figure 3.1 shows the resulting coefficient paths for the proposed estimator subject to the penalty parameter λ. λ is scaled as $1 - \lambda/\lambda_{max}$, where λ_{max} refers to the smallest value of the penalty parameter λ that already gives maximal penalization. That is, the smallest value of λ that sets all penalized coefficients to zero. Hence, we see the ML estimates at the right end of the two figures. The left end relates to the maximal penalization where only the intercept remains non-zero. In the left panel, the penalty is adaptive, the weights are fixed (see equation (3.8) with $b_0 = 0$, $\phi_{rs(j)} = \phi_{r(j)} = 0.5$). In the right panel, penalty (3.4) is

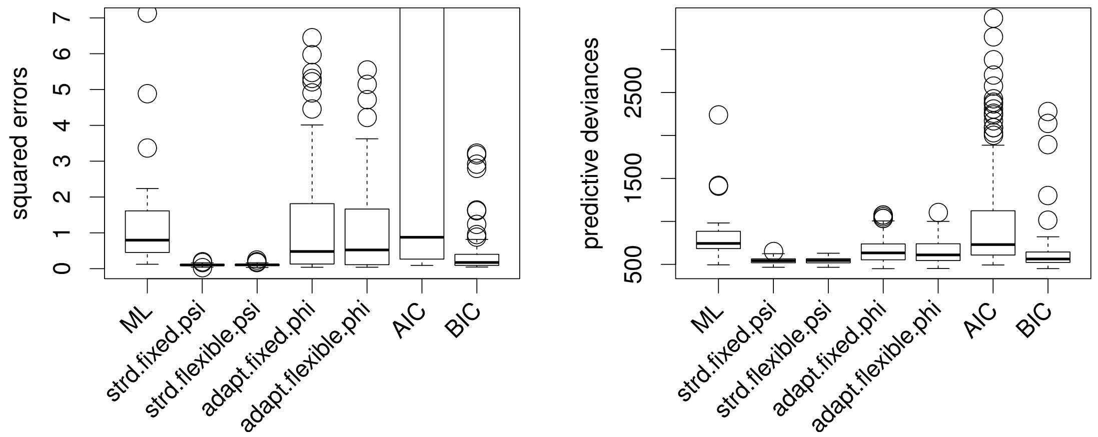


Figure 3.2.: Boxplots of the scaled squared errors (left panel) and deviances (right panel) for setting *S200*. The medians in the boxplots are robust estimates of the MSE and the MSEP. As the maximum of the squared errors for the method AIC is equal to 1319494, the figure is cropped.

employed. The paths show how the clustering and the selection of coefficients works. In the left panel, slight penalization discovers the intercept to be non-varying. The coefficients of the covariate $\boldsymbol{x}_1$ are fused such that only category 1 makes a difference. Concerning the covariate $\boldsymbol{x}_2$, the coefficients should be fused to one, non-varying scalar. However, a stronger penalty is necessary to make this happen. The dotted line marks the optimal model in terms of 5-fold cross-validation with the predictive deviance as loss function, see Section 2.2.4. It shrinks the coefficients slightly – in return, all but one relevant structures are identified. The absolute deviation to the true coefficients is small.

In the right panel of Figure 3.1, the unweighted penalty (3.4) is employed. Here, we see a different picture. While the structure of the coefficient paths remains basically the same, the coefficients are fused and/or selected for larger values of the penalty parameter λ. To reach the same effects as in the left panel of the figure, in the right panel, stronger penalization is needed. The cross-validated value of λ_{CV} is 2.11 now. The performance is worse than with the adaptively weighted penalty: In the model chosen by cross-validation (see dotted line), the coefficients of covariate $\boldsymbol{x}_1$ are not fused.

Settings To compare the proposed methods, various model features are systematically varied. Concretely, we consider a model with binomial response, two influential covariates, and six non-influential noise variables. The training data sets contain $n = 200$ and $n = 600$ observations, the test data sets $n = 600$ and $n = 1800$ observations. That is, we have two settings named *S200* and *S600*. For each setting, we generate 100 datasets. Therein, all covariates are continuous and independently drawn from an uniform distribution $U[-2, 2]$. There is one, known effect modifier. It is nominal, has four categories $1, \ldots, 4$ and is

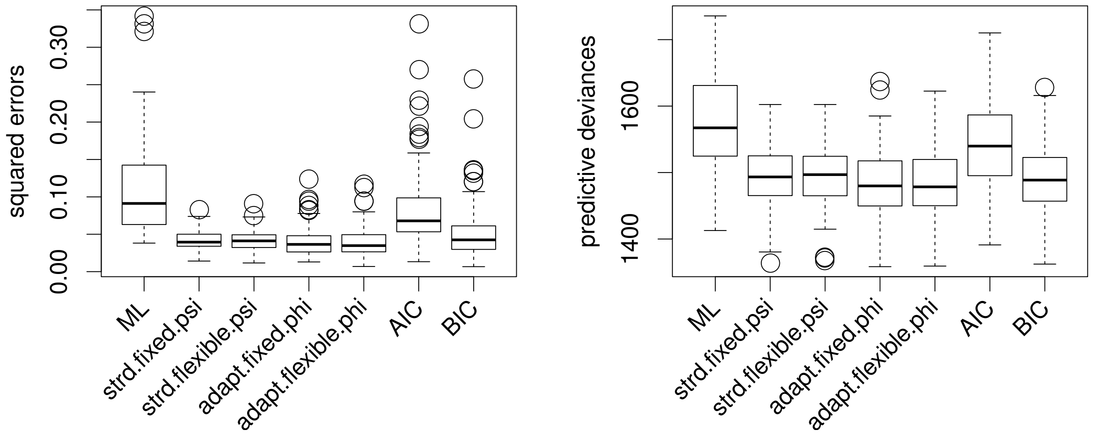


Figure 3.3.: Boxplots of the scaled squared errors (left panel) and deviances (right panel) for setting *S600*; the medians in the boxplots are robust estimates of the MSE and the MSEP.

independently drawn from a multinomial distribution with probability 0.25 per category. The true linear predictor is

$$\begin{aligned}\boldsymbol{\eta}_{true} &= \beta_0(\boldsymbol{u}) + \boldsymbol{x}_1\beta_1(\boldsymbol{u}) + \boldsymbol{x}_2\beta_2(\boldsymbol{u})\\ &= \qquad (\ 0.7\cdot I(\boldsymbol{u}=1) + 0.7\cdot I(\boldsymbol{u}=2) + 0.0\cdot I(\boldsymbol{u}=3) + 0.0\cdot I(\boldsymbol{u}=4)\ )\\ &\quad + \boldsymbol{x}_1\,(\ 1.0\cdot I(\boldsymbol{u}=1) - 1.5\cdot I(\boldsymbol{u}=2) - 1.5\cdot I(\boldsymbol{u}=3) + 0.5\cdot I(\boldsymbol{u}=4)\ )\\ &\quad + \boldsymbol{x}_2\,(\ 0.0\cdot I(\boldsymbol{u}=1) + 1.0\cdot I(\boldsymbol{u}=2) + 2.0\cdot I(\boldsymbol{u}=3) - 3.0\cdot I(\boldsymbol{u}=4)\ )\,.\end{aligned}$$

Since the truly varying coefficients are to be detected by the procedure, all coefficients are allowed to vary with effect modifier $\boldsymbol{u}$. As six non-influential noise variables $\boldsymbol{n}_3, \ldots, \boldsymbol{n}_8$ are added, the assumed predictor is

$$\boldsymbol{\eta}_{model} \;=\; \beta_0(\boldsymbol{u}) + \boldsymbol{x}_1\cdot\beta_1(\boldsymbol{u}) + \boldsymbol{x}_2\cdot\beta_2(\boldsymbol{u}) + \boldsymbol{n}_3\cdot\beta_3(\boldsymbol{u}) + \ldots + \boldsymbol{n}_8\cdot\beta_8(\boldsymbol{u}).$$

This model is estimated using by all the methods that we have discussed. We consider the estimates obtained with the proposed penalty for nominal effect modifiers and ...

- with weight ψ fixed at 0.5 (referred to as "strd.fixed.psi"),
- with flexible weight ψ (referred to as "strd.flexible.psi"),
- with adaptive weights and fixed $\phi_{rs(j)}$, $\phi_{r(j)}$ ($\phi_{rs(j)} = \phi_{r(j)} = 0.5$, "adapt.fixed.phi"),
- with adaptive weights and flexible $\phi_{rs(j)}$, $\phi_{r(j)}$ ($\phi_{rs(j)} + \phi_{r(j)} = 1$, "adapt.flexible.phi").

In addition, we consider the results of the forward selection strategies with the criteria AIC and BIC, and the usual ML estimate. For the ML estimates, neither regularization nor model selection is required. They are the benchmark for all the methods. The penalty parameter λ is chosen by 5-fold cross-validation. If the weights ψ and ϕ are flexible, they

Setting		ML	strd.fixed.psi	strd.flexible.psi	adapt.fixed.phi	adapt.flexible.phi	AIC	BIC
S200	FP$_s$	1	0.79	0.68	0.38	0.43	0.41	0.16
	FN$_s$	0	0.02	0.03	0.08	0.04	0.06	0.09
	FP$_c$	1	0.68	0.70	0.46	0.45	0.41	0.11
	FN$_c$	0	0.05	0.03	0.16	0.16	0.18	0.28
S600	FP$_s$	1	0.82	0.71	0.42	0.40	0.39	0.12
	FN$_s$	0	0.00	0.01	0.01	0.01	0.02	0.03
	FP$_c$	1	0.77	0.75	0.44	0.40	0.37	0.06
	FN$_c$	0	0.01	0.00	0.05	0.05	0.09	0.16

Table 3.1.: Estimates of false positive and false negative rates for the selection (s) and the clustering (c) process in settings *S200* and *S600*.

are cross-validated, too. Note, that the results for settings *S200* and *S600* are obtained with the version 1.4 of the package `gvcm.cat`.

Parameter Estimation To assess the parameter estimation by the mean squared error of coefficients (MSE), for each replication, we compute the squared errors

$$\widehat{\text{SSE}}(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) = \frac{1}{q} \sum_{j=1}^{q} \left(\boldsymbol{\beta}_j^* - \hat{\boldsymbol{\beta}}_j \right)^2 ,$$

where $q = \sum_{j=0}^{p} k_j$, $\boldsymbol{\beta}^*$ denotes the vector of true coefficients, and $\hat{\boldsymbol{\beta}}$ denotes its estimate. To judge the prediction accuracy by the mean predictive deviance (MSEP), the predictive deviances $\text{dev}(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ are considered. Figures 3.2 and 3.3 show the box plots of the squared errors and the predictive deviances for both settings. The squared errors of the penalized approaches and of the forward selection strategies tend to be smaller than those of the ML estimates. Moreover, setting *S200* is quite challenging for pure ML estimation. According to the criterion of Section 2.2.2, only 39 out of 100 models converged. The forward selection based on the AIC suffers from a high variability. Especially for $n = 200$, several extreme values are observed in Figure 3.2. For the penalized methods with adaptive weights, we see that the variability compared to the "standard" penalization becomes only smaller for an increasing sample size. This is due to the construction of the adaptive weights, which are the inverses of the ML estimates. The more observations we have, the more stable is the ML estimate and so is the corresponding weight. We clearly see that for small sample sizes, the oracle properties from Theorem 2 are not given at all, but with n becoming larger, the adaptive estimates become better and better.

Clustering and Selection Performance In addition, we evaluate the clustering and the selection performance of the proposed approaches. Non-influential covariates, especially pure noise variables, should be detected. That is, truly zero coefficients should not be selected. If some levels of an effect modifier have the same effect on the response, they should be detected, too. That is, truly non-varying coefficients should be fused. Thus, we consider the false negative (FN) and the false positive (FP) rates. False positive means that a truly zero coefficient is estimated to be non-zero. False negative means that truly non-zero coefficients are estimated to be zero. With # denoting "the number of coefficients" and with "s" standing for "selection", we have

$$\begin{aligned} \mathrm{FP}_s &= \frac{\#(\text{truly zero set to non-zero})}{\#(\text{truly zero})} \quad \text{and} \\ \mathrm{FN}_s &= \frac{\#(\text{truly non-zero set to zero})}{\#(\text{truly non-zero})}. \end{aligned}$$

FP_c and FN_c with "c" for "clustering" are defined analogously, but refer to differences of coefficients. Table 3.1 shows false positive and negative rates for both settings.

According to Theorem 2, the adaptive estimator with the selection consistency should yield better models in terms of the selection and of the clustering of the coefficients. This expectation seems to be confirmed. The adaptive penalty tends to perform better than the standard version of the penalty. The false positive rates are much smaller for the first one. For small n, however, the false negative rates are substantially larger when using the adaptive weights. This illustrates that the selection consistency is an asymptotic property that may not necessarily yield the best results for small sample sizes. With increasing n, the false negative rates are quite small for the adaptive version, too. The reason why the false positive rates are still rather high (for both, adaptive and non-adaptive weights) is that the penalty parameters are chosen by cross-validation, and cross-validation tends to select accurate estimates but a somewhat too large model. The AIC based forward selection performs similar. However, having the high variability of the scaled squared errors in mind, the previous recommendation for adaptive weights still holds. With the BIC, by contrast, typically a much smaller model is selected. That leads to smaller false positive but substantially larger false negative rates. So if the primary goal is a sparse model, and the analyst is willing to risk that a number of truly relevant coefficients or differences thereof are disregarded, the BIC based forward selection may be an alternative. Otherwise, sparseness and relatively low false negative rates are obtained by the proposed penalty with adaptive weights.

3.5. Application: Acceptance of Boar Meat and the Effect of Labeling

A known sensory problem with respect to boars is the occurrence of so-called boar taint, which may affect the consumer acceptance of boar meat, see, for example, Mörlein et al. (2012), Meier-Dinkel et al. (2013) and references therein. However, liking or disliking a food product does not only depend on the product's physicochemical properties but also on the consumers' expectations. Therefore, we are interested in whether the consumer acceptance is affected by solely labeling meat as "boar meat". In addition, various consumer characteristics may influence the individual liking or disliking of boar meat, such as the age or the gender of the consumer. The data considered here is a subset from Meier-Dinkel et al., 2013. Consumers tasted meat from four different product groups: (1) castrate or gilt meat (hereafter referred to as "control") with label "pork", (2) control with label "young boar meat", (3) real boar meat with label "pork", and (4) boar meat with label "young boar meat". We are interested in whether the probability of liking the taste of the product (binary response $y \in \{0, 1\}$) is affected by the product type, and in particular, in whether the acceptance of pork/boar meat differs between the labels "boar" and "pork". Hence, we include the product type as a categorical effect modifier in a logistic regression model and allow the influence of various covariates to change with the product type. The covariates that are modified by the product group are:

- the consumer's age,
- the consumer's gender,
- an indicator for smoking consumers (no/yes),
- an indicator for sickness (olfactory disability caused by sickness, in particular cold and allergy: no/yes), and
- a factor indicating whether the consumer knows what "boar meat" means (self-reported knowledge: no/yes).

In addition, we correct for the effect of contact to animal husbandry (contact: no/yes). Table 3.2 shows the coefficients estimated by pure maximum likelihood (block 1) and using the proposed penalty approach (block 2). The sample size is 133, which is small for a binary model with 28 parameters. This may explain the quite extreme ML estimates. Regularized estimates can be expected to be more reliable here.

Employing the proposed approach, we see that on average the probability of acceptance is estimated as equal for control and boar meat that is labeled as "pork", as in the intercept, it is not distinguished between these three groups. However, when we have a closer look at the consumers, this picture changes. In particular, if the consumer knows what "boar meat" means, the chance of accepting boar with label "pork" (product group (3)) decreases quite drastically. At first glance, this seems to contradict the hypothesis that

Coefficients	ML estimation				Penalized estimation				Forward selection AIC				Forward selection BIC			
	Product group				Product group				Product group				Product group			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
$\beta_0(t)$	9.2875	13.6364	10.9538	2.3709	2.28	2.28	2.28	1.49	1.54				1.72			
$\beta_{\text{know.lab}}(t)$	9.5848	9.2609	-9.9293	0.0791	1.60	1.60	-1.60		0.80	0.80		0.80	0.73	0.73		0.73
$\beta_{\text{gender}}(t)$	0.9793	8.8767	-0.3770	0.6693	0.56	2.18	-0.01	0.56	0.65	5.41		0.65	0.78	0.78		0.78
$\beta_{\text{contact}}(t)$	-0.8248	0.0709	9.3493	-0.1013			0.98									
$\beta_{\text{age}}(t)$	0.0252	0.2073	-0.0360	-0.0356		0.07	-0.02	-0.02		0.19						
$\beta_{\text{smoker}}(t)$	0.0222	0.4495	-0.2935	0.3936												
$\beta_{\text{sick}}(t)$	0.8016	0.3018	9.0947	0.1464	0.08		1.01									

Table 3.2.: Estimates for all methods fitted to the boar data. Excluded coefficients are omitted. Non-varying coefficients are represented by the remaining scalar only.

Coefficients	ML estimation				Penalized estimation				Forward Selection AIC				Forward Selection BIC			
	t				t				t				t			
	2001	2002	2003	2004	2001	2002	2003	2004	2001	2002	2003	2004	2001	2002	2003	2004
$\beta_0(t)$	25.0738	7.4078	14.8948	3.4140	7.85	7.85	7.85	3.97	6.98				5.55			
$\beta_{\text{term}}(t)$	15.4642	21.6350	-5.7049	-3.6013	9.77	9.77	-4.05	-3.56	-2.43							
$\beta_{\text{c.height}}(t)$	3.6217	-6.6733	-2.5879	0.7056	2.45	-1.95	-1.95									
$\beta_{\text{c.weight}}(t)$	9.6170	3.7054	-0.5718	0.0950	4.27	2.23	-0.05									
$\beta_{\text{m.age}}(t)$	-5.7380	1.9730	1.4235	1.1692	-2.73	1.12	1.12	1.12	0.73							
$\beta_{\text{m.height}}(t)$	-48.8037	-26.4130	-4.0494	-2.2466	-22.64	-18.94	-1.96	-1.96	-5.43				-5.54			
$\beta_{\text{m.bmi}}(t)$	-0.3888	1.5582	0.6445	0.1389		0.50	0.28									
$\beta_{\text{m.gain.w}}(t)$	-0.8593	0.0069	0.3428	-0.1781			0.01									
$\beta_{\text{m.prev}}(t)$	-4.4749	-5.7147	-0.6601	-0.4488	-1.53	-1.53	-0.53	-0.46	-0.53				-0.46			
$\beta_{\text{ind}}(t)$	1.8585	-0.4328	0.4599	0.1304	0.86	-0.16	0.39	0.11	0.25				0.26			
$\beta_{\text{memb}}(t)$	-0.1761	0.2289	-0.2716	-0.3278			-0.25	-0.25			-0.29	-0.29				
$\beta_{\text{rest}}(t)$	-0.4687	-2.8226	-0.1463	-0.0793	-0.02	-0.34			-0.37	-0.37	-0.13					
$\beta_{\text{cephalic}}(t)$	-0.2333	-7.4411	-5.2453	-0.4291		-1.97	-1.96	-0.42	-0.68				-0.67			

Table 3.3.: Estimates for all methods fitted to the birth data. Excluded coefficients are omitted. Non-varying coefficients are represented by the remaining scalar only.

consumers' expectations influence liking of products, but it may be explained by some sort of disappointment effect, as the consumer expected to taste pork, as labeled, but no boar. For control meat (group (1) and (2)), by contrast, there is a positive effect of knowledge which is constant over both labels. Only when the boar meat is labeled correctly (group (4)), there is no effect of knowledge as the coefficient is set to zero. The latter two findings, as well as the effect of gender, are rather difficult to explain, but there is another interesting effect of labeling: If boar is labeled as "pork" (product group (3)), being sick increases the chance that the consumer likes the taste of the product. A possible explanation is that sickness affects the sense of taste and the sense of smell, and sick people hence rather rely on the label saying that it's pork but no boar. Though also smoking might affect the sense of taste, the consumer's smoking status is fitted as completely irrelevant for the acceptance of the meat, as all coefficients are set to zero. If we look at the effect of age, we see that for older people, the chance of liking control meat labeled as boar increases, but the chance of liking boar – no matter which label is attached – decreases. Possibly, the expectation of a certain taste increases with age.

With the AIC/BIC-based forward selection strategies (block 3/4), the estimated model is much sparser. This may indicate that at least some of the effects found above are false positives. However, as seen in the simulations, forward selection strategies may have rather large false negative rates. To obtain more insight, further studies are necessary and currently conducted (Trautmann et al., 2014).

3.6. Application: Cesareans among Francophone Mothers

This data set contains various variables related to the pregnancy and the delivery of women recruited on French-speaking websites. The data is presented by Boulesteix (2006) and is available in the R package catdata (Schauberger and Tutz, 2014). As described in Section 3.1, we are interested in the type of the delivery, in whether birth was given vaginally or by means of a Cesarean. Between 2001 and 2004, 578 deliveries were observed, and modeling the type of the delivery requires to allow the covariate effects to vary with the time, since, for example, medical standards may have changed over time. As the time is measured discretely and on a rough grid, we consider the time in years as an ordinal effect modifier in a varying-coefficient model. The response is binary indicating the type of the delivery; 0 stands for a vaginal birth, 1 for a Cesarean. The model considers all covariates that were available and meaningful for all women. Details on the covariates are found in Table 3.4. To be on comparable scales, all covariates are scaled. As terms and delivery circumstances differ immensely for multiple births, these cases are excluded.

As we have no prior knowledge about the model's structure, the effect modifier t potentially impacts all coefficients. As there is a relatively large number of covariates, we are not only interested in the fusion of coefficients, but also in the selection of coefficients $\beta_j(t)$.

Variable	Description
cesarean	Type of the delivery (0: vaginal, 1: Cesarean), response
term	Term of the pregnancy in weeks form the last menstruation
c.height	Height of the child at birth in centimeter
c.weight	Weight of the child at birth in gram
m.age	Age of the mother before the pregnancy in years
m.height	Height of the mother in centimeter
m.bmi	BMI of mother beforethe pregnancy (mass(kg)/(height(m))^2)
m.gain.w	Gain in weight of the mother during the pregnancy in kg
m.prev	Number of previous pregnancies
ind	Was the labor induced?
memb	Did the membranes burst before the beginning of the throes?
rest	Was a strict bed rest ordered to the mother for at least one month during the pregnancy?
cephalic	Was the child in cephalic presentation before birth?
t	Year of the birth, effect modifier

Table 3.4.: Short description of the response, the covariates and the effect modifier for the birth data. The coding of the binary covariates is 0 for “no”, 1 for “yes”.

Table 3.3 shows the resulting estimates. The values of the ML estimates are quite extreme. To obtain a stable estimation procedure, that is able to distinguish among the covariates, regularization is required. As suggested by the numerical experiments in Section 3.4, we employ an adaptive penalty with fixed weights $\phi_{rs(j)} = \phi_{r(j)} = 0.5$. The penalty parameter λ is chosen by the generalized cross-validation criterion discussed in Section 2.2.4; it is set to 0.72. This is a relatively small value, but it stabilizes the estimation and shrinks the huge ML estimates. As an alternative, we consider the forward selection strategies that are presented in Section 3.3. Here, the forward selection strategies produce very sparse estimates. Only two (AIC), respectively none (BIC), coefficients are partly varying. The ML estimates, by contrast, argue for a strong dependency on time, see, for example, the intercept of the year 2001, which is ignored by the forward selection strategies. Penalized estimation gives a more differentiated picture. It selects some covariates and shows that not all coefficients are varying over time.

3.7. Special Case: Categorical Effects

So far, we have considered general categorical effect modifiers. We did not discuss categorical effects which are a special case of categorical effect modifiers. One obtains a coded categorical effect, when the effect modifier $\boldsymbol{u}_j$ is categorical and the modified covariate $\boldsymbol{x}_j$ is a constant vector. We have, for example, $1 \cdot \beta_j(\boldsymbol{u}_j) = 1 \cdot \sum_{r=1}^{k_j} \beta_{jr} I(\boldsymbol{u}_j = r)$. The penalization remains the same. The statements made for penalized varying coefficients hold for penalized categorical effects, too. Especially the large sample properties can be transferred. However, the devil is in the details: Unlike usual coding, the obtained coding does not contain a reference category. This implies at least two things: The design matrix is not of full rank and the interpretation changes. As the estimates are penalized and as the penalty parameter λ will be cross-validated, in most cases, the first aspect can be neglected. Concerning the interpretation, penalized estimates can be transformed, such that they correspond directly to conveniently coded categorical effects. However, the penalty that we use here is not designed for a reference category. In contrast to Gertheiss and Tutz (2010), all categories of a categorical effect are penalized in the same way.

3.8. Remarks

We investigate categorical effect modifiers within the framework of GLMs. When selecting a model with categorical effect modifiers, one wants to find out which covariates have an effect on the response, and if so, which categories have to be distinguished. In fact, this is a recoding of usual interactions between categorical and metric covariates, but the concept of effect modifiers allows for interpretable model selection strategies. We present two different approaches: On the one hand, we extended the ideas of Tibshirani et al. (2005) to varying-coefficient models with categorical effect modifiers. Thus, we are able to simultaneously identify varying coefficients and select covariates. The penalty adjusts for the different amount of information in nominal and ordinal effect modifiers. An adaptive version of the proposed penalty is shown to be asymptotically normal and consistent. These results remain valid when the scale parameter of the exponential family is estimated and plugged-in, which allows for quasi-likelihood approaches. On the other hand, we investigate a modified forward selection strategy: Start with a null-model and add one degree of freedom in each iteration until a chosen criterion is not improved anymore. Numerical experiments suggest both methods to be highly competitive: Penalized estimates and forward selection strategies perform distinctly better than non penalized ML estimates. Forward selection strategies, suffer from a immense variability, particularly when they are based on the AIC, which makes them less attractive. With the BIC, typically a smaller model than with L_1 penalties is selected, which leads to smaller false positive but higher false negative rates.

In practice, varying-coefficient models are highly relevant. We analyze data from a consumer study on boar meat. It can be confirmed that the chance of consumer acceptance is smaller for boar meat than for regular pork (castrate or gilt meat). In addition, some evidence for labeling effects are found: If wrong labeling causes too high expectations, the disappointment substantially reduces the chance of the acceptance of boar meat. If the sense of taste is affected by sickness, consumers seem to rely on the labeling.

Analyzing data on Cesareans among francophone mothers, we are interested in how the influence of various medical indicators changed over time. The data is quite challenging, standard approaches fail. However, penalized estimates give a coherent trend.

So far, we have employed a single penalty parameter λ only. For a modest number of effect modifiers, however, one penalty parameter per effect modifier could be advantageous. The proposed penalty's potential is apparent: For longitudinal studies its scope can be enlarged to marginal models. The approach can be further generalized. Varying coefficients may depend on more than one effect modifier. In this chapter, we assumed continuous covariates $\boldsymbol{x}_1, \dots, \boldsymbol{x}_p$, but of course, covariates can be categorical, too. Then, there are even more coefficients, and hence, there is an even stronger demand for regularization. In contrast, if the effect modifier is continuous, smooth functions can be specified and be added to the model with none but the usual restrictions.

4. Modeling Clustered Heterogeneity

4.1. Introduction

When observations are grouped within the data, this can cause additional heterogeneity which has to be considered adequately. Observations are, for example, grouped when there are repeated measurements over time as in longitudinal studies, or when there are subsamples of superior/primary sampling units as in cross-sectional studies. In such settings, it is usually too restrictive to assume that the groups behave the same. Strategies that take the heterogeneity of the effects into account have to be found.

Repeated measurements can be represented by $(y_{ij}, \boldsymbol{x}_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$, where y_{ij} denotes the response of unit i at measurement occasion j, and $\boldsymbol{x}_{ij}$ is a vector of covariates that potentially varies across measurements. In cross-sectional studies, the data has the same form if it is collected in groups. For example, in a multi-center treatment study, y_{ij} may denote the response of patient j in study center i. In the terminology of multilevel models, the patients are the first level units and the study centers the second level units.

An application, that will be considered in more detail later, deals with the effect of beta blockers on the mortality after myocardial infarction, see also Aitkin (1999), Grün and Leisch (2008a). In a 22-center clinical trial, for each center, the number of deceased/successfully treated patients in control/test groups was observed. Hence, the patients represent the first level units and the hospitals the second level units. The binary response (1 = deceased/0 = not deceased) suggests a logit model, which in its simplest form is given by

$$\text{logit } \mathbb{P}(y_{ij} = 1) = \beta_0 + \beta_T \cdot \text{Treatment}_{ij}, \quad i = 1, \ldots, 22 \text{ Centers}, \tag{4.1}$$

where $\text{Treatment}_{ij} \in \{-1, 1\}$ codes the treatment in hospital i for patient j (1: Treatment, -1: Control). Model (4.1) does not account for the heterogeneity among the hospitals. The

This chapter is a modified version of the Technical Report 156 (Tutz and Oelker, 2014); initial considerations can be found in Tutz and Oelker (2013). For more information on the contributions of the authors and on textual matches, see page 4.

treatment effect β_T, as well as the basic risk captured in β_0, are assumed to be the same for all hospitals. Of course, this is a very strong assumption that hardly holds.

The most popular model that incorporates heterogeneity is the *random effects model.* It replaces the intercept β_0 by $\beta_0 + b_{i0}$, yielding

$$\text{logit } \mathbb{P}(y_{ij} = 1) = \beta_0 + b_{i0} + \beta_T \cdot \text{Treatment}_{ij}, \quad i = 1, \ldots, 22 \text{ Centers}, \tag{4.2}$$

where b_{i0} is a random effect for which a distribution is assumed, typically a normal distribution: $b_{i0} \sim N(0, \sigma_b^2)$. Implicitly, the hospitals are considered as a random sample. The inference concerning the treatment effect should hold for the whole underlying set of hospitals. One can go one step further, and replace the treatment effect β_T by $\beta_T + b_{iT}$, allowing for heterogeneity of treatments over hospitals. Random effects models are a strong tool to model heterogeneity, and a wide body of literature is available (see, for example, Verbeke and Molenberghs, 2000; Molenberghs and Verbeke, 2005). However, there are drawbacks: Inference on the unknown distributional assumption is hard to obtain. And even though the choice of the distribution may not matter too much (McCulloch and Neuhaus, 2011), implicitly, it is assumed that all hospitals differ with respect to the basic risk and/or treatment effect. In other words: Hospitals with the same basic risk are not allowed, and the hospitals themselves are not clustered. Of course, one could fit a random effects model, look for similar effects in a second step, and refit the model in a third step. And of course, there are very sophisticated approaches to detect clusters of random effects, see, for example, Heinzl and Tutz (2013, 2014). However, it has to be assumed that there are no correlations between the covariates and the random effects. Again, there are methods that provide consistent estimates for the covariate effects in settings with such correlations, for example, conditional likelihood methods (Diggle et al., 2002, Section 9.2.1), or random effects models that decompose covariates into between- and within-group components (Neuhaus and McCulloch, 2006; Grilli and Rampichini, 2011). However, in the case of the beta blocker data, the aim is to account for the heterogeneity between the hospitals, to find potential clusters of hospitals and to obtain estimates that are not affected by potential correlations simultaneously. There is a special interest in the second level units.

An alternative approach, that accounts for the heterogeneity in the data, are *group-specific models* which are also known as *fixed effects models.* In this class of models, the effects of the groups are considered as unknown but fixed. In the beta blocker data, the intercept β_0 is replaced by the parameters β_{i0}, the treatment effect is (potentially) replaced by β_{iT}. In contrast to random effects models, the second level units are not considered as representatives of an underlying population. The inference refers to the given sample, that is, to the hospitals in the data set. Thus, group-specific models are especially useful when one is interested in the performance of specific second level units, that is, in the hospitals. Moreover, there are no assumptions on the correlation between the covariates and the group-specific effects. A disadvantage is that the number of parameters increases compared to random effects models. The second level units are not clustered. However, the parameters of the group-specific model can be seen as varying coefficients with categorical effect modifiers. And for categorical effect modifiers, carefully tailored regularization allows to reduce the number of parameters and to identify clusters of coefficients that share the same effect, see Chapter 3.

This is why, group-specific models are combined with the theory of Chapter 3. We try to show that regularized group-specific models are an attractive approach, when (i) the second level units themselves are of interest, (ii) the second level units are potentially clustered, and (iii) the estimates shall not be affected by potential correlations.

Chapter 4 is organized as follows: In Section 4.2, three conventional approaches to model heterogeneity are shortly introduced and discussed – namely, finite mixture models, random effects models and group-specific models. In Section 4.3, we propose regularized group-specific models that fulfill the requirements of the beta blocker data. Section 4.4 investigates the performance of regularized group-specific models. The data on the mortality after myocardial infarction is analyzed in Section 4.5. Section 4.6 gives some extensions on the simultaneous fusion of group-specific effects related to several covariates. Section 4.7 illustrates this idea by analyzing data on the math performance of pupils in ten different schools in the United States of America.

4.2. Modeling Heterogeneity

In what follows, different methods that model heterogeneity are shortly sketched. We start with finite mixture models that have not been mentioned yet.

4.2.1. Finite Mixture Models

In finite mixtures of generalized linear models, it is assumed that the density or mass function of observation y_i given $\boldsymbol{x}_i$ is a mixture:

$$f(y_i|\boldsymbol{x}_i) = \sum_{k=1}^{K} \pi_k f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}_k, \varphi_k), \tag{4.3}$$

where $f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}_k, \varphi_k)$ represents the k-th component of the mixture that follows a simple exponential family parametrized by the parameter vector $\boldsymbol{\beta}_k$ from the model $\mu_k = \mathbb{E}(y_i|\boldsymbol{x}_i, k) = h(\boldsymbol{x}_i^T\boldsymbol{\beta}_k)$. The exponential family is defined as $f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}_k, \varphi_k) = \exp\{(y_i\vartheta_k - \kappa(\vartheta_k))/\varphi_k + c(y_i, \varphi_k)\}$, where $\vartheta_k = \vartheta(\mu_k)$ denotes the natural parameter, $\kappa(\vartheta_k)$ is a specific function corresponding to the type of the exponential family, $c(\cdot)$ is the log-normalization constant and φ_k the dispersion parameter (see Fahrmeir and Tutz, 2001).

$h(\cdot)$ denotes a monotonic and continuously differentiable response function. The unknown component weights follow $\sum_{k=1}^{K} \pi_k = 1$, $\pi_k > 0$, $k = 1, \dots, K$. Based on the estimated values of π_k, the posteriori probability that an observation i is part of the component k is computed, and the observation is assigned to the component with the maximal posteriori probability.

For hierarchical settings, the components can be linked to the groups, the second level units respectively. Let $C = \{1, \dots, n\}$ denote the set of units that are observed. Then, one specifies one model per component k, for example,

$$g(\mu_{ij}) = \beta_{k(i)} + \boldsymbol{x}_{ij}^T \boldsymbol{\beta},$$

where $\beta_{k(i)}$ denotes that the effect β_k is the same for all second level unit i that are assigned to the component k. That is, $\beta_{k(i)} = \beta_k$ for all $i \in C_k$, where $C_1, \dots, C_K$ is a disjunct partition of C. Therefore, the units are clustered into subsets with identical intercepts with the total vector of coefficients being given by $\boldsymbol{\alpha}^T = (\beta_1, \dots, \beta_K, \boldsymbol{\beta}^T)$. For the estimation of finite mixture models, typically, the EM-algorithm is employed. Please note that the number of components for one finite mixture model is fixed. The optimal number of mixture components is chosen in a second step, for example, by information criteria.

Mixture models are, for example, considered by Follmann and Lambert (1989), and Aitkin (1999). An extensive treatment is given by Fruehwirth-Schnatter (2006). Follmann and Lambert (1989) investigate the identifiability of finite mixtures of binomial regression models and give sufficient identifiability conditions for mixing at the binary and the binomial level. Grün and Leisch (2008b) consider the identifiability of mixtures of multinomial logit models, and provide the `R` package `flexmix` with various applications (Grün and Leisch, 2008a).

Finite mixture models allow us to find second level units, hospitals respectively, with similar effects on the response. However, assuming a mixture of different components is a very strong assumptions. The estimation procedures are quite sophisticated. On account of this, finite mixture models will be a competing approach in the numerical experiments in Section 4.4, but the focus of this chapter will be on random effect models and on group-specific models.

4.2.2. Random Effects Models

The random effects model is probably the most popular model that accounts for heterogeneity in the data. Let the observations be given as y_{ij}, where j denotes an observation in the second level unit i, $i = 1, \dots, n$, $j = 1, \dots, n_i$. Let $\boldsymbol{x}_{ij}^T = (1, x_{ij1}, \dots, x_{ijp})$ be a covariate vector associated with fixed effects, and $\boldsymbol{z}_{ij}^T = (z_{ij1}, \dots, z_{ijq})$ be a covariate vector associated with random effects. In a generalized linear random effects model, the structural assumption specifies that the conditional means $\mu_{ij} = \mathbb{E}(y_{ij}|\boldsymbol{b}_i, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij})$ have the form

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i = \eta_{ij}^{par} + \eta_{ij}^{rand}, \tag{4.4}$$

where g is a monotonic and continuously differentiable link function and $\eta_{ij}^{par} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}$ is a linear parametric term with parameter vector $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \dots, \beta_p)$, which includes an intercept. The second term, $\eta_{ij}^{rand} = \boldsymbol{z}_{ij}^T\boldsymbol{b}_i$, contains the random effects that model the heterogeneity of the second level units. For the random effects, a distribution is assumed, typically a normal distribution with covariance matrix $\boldsymbol{Q}$: $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{Q})$. The random effects and the covariates observed per second level unit are assumed to be independent.
In a generalized linear random effects model, the distributional assumption for $y_{ij}|\boldsymbol{b}_i, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}$ is of the exponential family type, too: $f(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{b}_i) = \exp\{(y_{ij}\vartheta_{ij} - \kappa(\vartheta_{ij}))/\varphi + c(y_{ij}, \varphi)\}$. It is assumed that the observations y_{ij} are conditionally independent with means $\mu_{ij} = \mathbb{E}(y_{ij}|\boldsymbol{b}_i, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij})$ and variances $\mathbb{V}(y_{ij}|\boldsymbol{b}_i) = \varphi v(\mu_{ij})$, where $v(\cdot)$ is a known variance function.
Although, it is the most popular model that accounts for heterogeneity, it has some drawbacks: The focus of the random effects models is on the fixed effects, and not on the second level units as it is requested by the data on the mortality after myocardial infarction. The distribution of the random effects has to be specified. Often, it is questionable to assume that the vectors $\boldsymbol{b}_i$, $\boldsymbol{x}_{ij}$ are independent. If they are not, that is, in the case of so called level 2 endogeneity, the estimates related to the fixed effects may be biased. Grilli and Rampichini (2011) derive the exact impact of correlated vectors $\boldsymbol{b}_i$, $\boldsymbol{x}_{ij}$ in linear random intercept models. Moreover, assuming a continuous distribution prevents that the effects of the second level units can be the same. By construction, no clustering of second level units is available.

4.2.3. Group-Specific Models

The group-specific model is another approach that accounts for heterogeneity in the data. As before, the predictor η_{ij} contains the term $\boldsymbol{x}_{ij}^T\boldsymbol{\beta}$, but the heterogeneity is taken into account by the parameters $\boldsymbol{\beta}_i$ instead of by the random effects $\boldsymbol{b}_i$, $i = 1, \dots, n$. For the link between the explanatory variables and the mean $\mu_{ij} = E(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij})$, the group-specific model assumes

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{\beta}_i. \tag{4.5}$$

The model specifies that each group or second level unit has its own vector of coefficients $\boldsymbol{\beta}_i^T = (\beta_{i0}, \dots, \beta_{iq})$, $i = 1, \dots, n$, which represents weights on the vector $\boldsymbol{z}_{ij}^T = (1, z_{ij1}, \dots, z_{ijq})$. In order to avoid identifiability problems, we assume that $\boldsymbol{z}_{ij}$ is not a subset of $\boldsymbol{x}_{ij}$. As a representation of this form can always be obtained, this is only a minor limitation: Let $\boldsymbol{x}_{ij}$ be partitioned into $\boldsymbol{x}_{ij}^T = (\boldsymbol{z}_{ij}^T, \boldsymbol{w}_{ij}^T)$ and accordingly $\boldsymbol{\beta}$ into $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_z^T, \boldsymbol{\beta}_w^T)$. Then, the model with group-specific effects on $\boldsymbol{z}_{it}$ only is

$$g(\mu_{ij}) = \boldsymbol{z}_{ij}^T\boldsymbol{\beta}_z + \boldsymbol{w}_{ij}^T\boldsymbol{\beta}_w + \boldsymbol{z}_{ij}^T\tilde{\boldsymbol{\beta}}_i.$$

For identifiability, some constraint on the vectors $\tilde{\boldsymbol{\beta}}_i$ is needed, for example, $\sum_i \tilde{\boldsymbol{\beta}}_i = 0$. However, the model can also be given as

$$g(\mu_{ij}) = \boldsymbol{w}_{ij}^T\boldsymbol{\beta}_w + \boldsymbol{z}_{ij}^T(\boldsymbol{\beta}_z + \tilde{\boldsymbol{\beta}}_i) = \boldsymbol{w}_{ij}^T\boldsymbol{\beta}_w + \boldsymbol{z}_{ij}^T\boldsymbol{\beta}_i,$$

where $\boldsymbol{z}_{ij}$ is not a subset of $\boldsymbol{w}_{ij}$ and the parameters $\boldsymbol{\beta}_i = \boldsymbol{\beta}_z + \tilde{\boldsymbol{\beta}}_i$ are not restricted.

As already mentioned in the introduction of this chapter, the group-specific effects in model (4.5) can be seen as a varying coefficient term that represents the interaction between the variables in $\boldsymbol{z}_{ij}$ and the groups. Consider the model with group-specific intercepts,

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \beta_{i0},$$

where $\boldsymbol{z}_{ij} = 1$. In detail, let the second level units/groups in $C = \{1, \ldots, n\}$ be coded by the dummy variables $x_{C(1)}, \ldots, x_{C(n)}$, where $x_{C(i)} = 1$ if $C = i$, $x_{C(i)} = 0$, otherwise. Then, the model can be written as

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + x_{C(1)}\beta_{10} + \ldots + x_{C(n)}\beta_{n0}.$$

As the intercept depends on the second level units/groups, it is a model where the effect modifier is a factor. In this case, only the variable $\boldsymbol{z}_{ij} = 1$ is modified. In the general case, the model has the form

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + x_{C(1)}\boldsymbol{z}_{ij}\beta_{10} + \cdots + x_{C(n)}\boldsymbol{z}_{ij}\beta_{n0},$$

where the products of z-variables and the dummies for the second level units/groups represent the interaction terms – such that the factor “group” modifies the effects of all the z-variables.

The problem with group-specific models is the large number of parameters. They can render the estimates unstable and encourage overfitting. Typically, there is not enough information available to distinguish among all the units. So far, the second level units have not been clustered. One further issue of group-specific models is that the x-variables have to vary across the first level units. If the x-variables are split into $(\boldsymbol{x}_i^T, \boldsymbol{x}_{ij}^T)$ with the first component representing explanatory variables on the group level, in the corresponding model $g(\mu_{ij}) = \beta_{i0} + \boldsymbol{x}_i^T\boldsymbol{\beta}_1 + \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_2 + \boldsymbol{z}_{ij}^T\boldsymbol{\beta}_i$, the term $\boldsymbol{x}_i^T\boldsymbol{\beta}_1$ can be absorbed in the intercept β_{i0}. Thus, in the classical group-specific model, it is not possible to include explanatory variables that are group-constant. This is considered as a severe disadvantage.

4.2.4. Random Effects Models vs. Group-Specific Models

The comparison of random effects models and group-specific models, has a long tradition. Townsend et al. (2013) summarize much of the work that has been done concerning the choice between random effects models and group-specific models. There are various criteria that can be used when comparing the two approaches.

One advantage of group-specific models refers to the underlying assumptions. The assumptions in group-specific models are weaker because, in contrast to random effects models, conditional independence between the covariates and the groups does not have to be postulated. Although, this does not mean that the model is more robust to other violations of the model assumptions (see Townsend et al., 2013), it should suffer less from the violation of the conditional independence between the covariates and the groups.

What is often considered as a drawback of group-specific models is the reduced efficiency of estimates. The problem is that for a large number of groups, the number of degrees of freedom is consumed by the group-specific effects. With 22 study centers, the group-specific model with group-specific intercepts and one explanatory variable contains 23 parameters, whereas the random intercept model contains one intercept, one slope parameter and requires only one parameter for the heterogeneity, namely $\sigma_b^2 = \mathbb{V}(b_i)$. However, the effective degrees of freedom is typically larger. Ruppert et al. (2003) consider the linear random effects model

$$y_{ij} = \beta_0 + b_{i0} + \beta x_{ij} + \varepsilon_{ij},$$

with $\sigma_b^2 = \mathbb{V}(b_{i0})$ and $\sigma_\varepsilon^2 = \mathbb{V}(\varepsilon_{ij})$. Then, the vector of fitted values can be written as $\hat{\boldsymbol{y}} = \boldsymbol{H}_0\boldsymbol{y} + \boldsymbol{H}_b\boldsymbol{y} + \boldsymbol{H}_x\boldsymbol{y}$, where $\boldsymbol{H}_0$ refers to the intercept, $\boldsymbol{H}_b$ to the random intercepts and $\boldsymbol{H}_x$ to the covariate x_{ij}. The hat matrices yield the effective degrees of freedom for the components of the model as $df_0 = tr(\boldsymbol{H}_0)$, $df_b = tr(\boldsymbol{H}_b)$, $df_x = tr(\boldsymbol{H}_x)$. One obtains $df_0 = df_x = 1$. For balanced designs with $n_i = m$ for all i, it holds that

$$df_b = \frac{(n-1)m}{m + \sigma_\varepsilon^2/\sigma_b^2}.$$

Thus, the effective number of parameters depends on the ratio $\sigma_\varepsilon^2/\sigma_b^2$. In the extreme case $\sigma_b^2 = 0$, one obtains a model with two parameters, namely the intercept and the slope. In the case $\sigma_b^2 \to \infty$, one obtains the group-specific model with $n + 1$ parameters. Therefore, the random effects model can be seen as a compromise between these extreme cases, and the group-specific model itself represents an extreme case of the random effects model. The closeness to the group-specific model is determined by the ratio of within-group and between-group variance components. The possible large number of parameters of the group-specific model has led to several recommendations to use the group-specific model, in particular when there are few groups and moderately large numbers of observations in each group, see, for example, Goldstein (2011). However, this restriction does not hold for the approach advocated here. We penalize the group-specific model. Thereby, the number of parameters of the group-specific model is implicitly reduced; the parameters can be efficiently estimated. As the group-specific model can be seen as varying coefficients with a categorical effect modifier, the Lasso-type penalties of Chapter 3 provide an attractive way to cluster the second level units.

Moreover, the potential loss of efficiency has to be weighted against the bias reduction obtained by the group-specific model. By adding group-specific indicators as explanatory variables, the group-specific model controls for all sorts of confounders. That means, in the clinical trial example, it controls for all the confounding variables such as different sizes and different patient populations of the centers. Having the findings of Grilli and Rampichini (2011) in mind, the bias reduction does not only refer to potential confounders but also to biased estimates of the fixed effects $\boldsymbol{\beta}$ in the random effects model due to correlated vectors $\boldsymbol{b}_i$, $\boldsymbol{x}_{ij}$.

One further issue in the comparison of group-specific models and random effects models is that the former postulate that the x-variables have to vary across the first level units. However, if the estimates are regularized, also the effects of group-specific explanatory variables can be estimated. We will not consider this in detail here, but refer to Tutz and Schauberger (2014) for an example.

4.3. Regularized Estimation for Group-Specific Models

In the model for the beta blocker data, the hospitals form one cluster if the corresponding coefficients are estimated to be the same. As seen in Chapter 3, the basic concept to enforce the clustering of second level units according to their coefficients, is penalized maximum likelihood estimation. Let all the parameters be collected in $\boldsymbol{\alpha}^T = (\boldsymbol{\beta}^T, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_n^T)$, with $\boldsymbol{\beta}_i$, $i = 1, \ldots, n$, denoting the group-specific parameters. One maximizes

$$l_p(\boldsymbol{\alpha}) = l(\boldsymbol{\alpha}) - \lambda J(\boldsymbol{\alpha}),$$

where $l(\boldsymbol{\alpha})$ denotes the familiar unpenalized log-likelihood, the parameter λ is the penalty parameter, and $J(\boldsymbol{\alpha})$ denotes the penalty term that enforces the clustering of the second level units. The choice of the penalty is crucial because it determines the clusters to be found.

For simplicity, at first, we assume group-specific intercepts only. That is, the model is given by $g(\mu_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \beta_{i0}$, $i = 1, \ldots, n$. Then, a specific penalty term that enforces the clustering is given by the pairwise differences of group-specific coefficients:

$$J(\boldsymbol{\alpha}) = \sum_{r>m} |\beta_{r0} - \beta_{m0}|. \tag{4.6}$$

The effect of the penalty is seen by looking at the extreme values of the penalty parameter λ. If $\lambda = 0$, one obtains the unpenalized estimates of $\boldsymbol{\alpha}$ and each second level unit has its own intercept. If $\lambda \to \infty$, the penalty enforces that the estimates of all group-specific intercepts are the same. Then, the second level units form one cluster with the same intercept. Penalty (4.6) is a specific version of the fused lasso, which has been considered by Tibshirani et al. (2005) for ordered covariates. The use for categorical covariates has been proposed by Bondell and Reich (2009) for factorial designs, and as a selection tool by Gertheiss and Tutz (2010).

In the general case with q covariates $\boldsymbol{z}_{ij}^T = (1, z_{ij1}, \ldots, z_{ijq})$, one uses the pairwise differences of all group-specific coefficients

$$J(\boldsymbol{\alpha}) = \sum_{s=0}^{q} \sum_{r>m} |\beta_{rs} - \beta_{ms}|. \tag{4.7}$$

The penalty enforces that for $\lambda \to \infty$, all the estimated group-specific coefficients of a covariate s are the same, that is, $\beta_{1s} = \ldots = \beta_{ns} = \beta_s$. Hence, for $\lambda \to \infty$, there is one global parameter β_s per covariate.

Adaptive versions of Lasso-type penalties have been shown to have better properties in terms of variable selection. For the basic Lasso, this has been demonstrated by Zou (2006); for categorical covariates, similar results were obtained by Bondell and Reich (2009) and Gertheiss and Tutz (2010). With $w_{rms} = |\tilde{\beta}_{rs} - \tilde{\beta}_{ms}|^{-1}$ where $\tilde{\beta}_{rs}$ denotes an $\sqrt{n}$-consistent estimate as, for example, the maximum likelihood (ML) estimate, one obtains an adaptive version of the penalty:

$$J(\boldsymbol{\alpha}) = \sum_{s=0}^{q} \sum_{r>m} w_{rms} |\beta_{rs} - \beta_{ms}|. \tag{4.8}$$

The effect of the adaptive weights w_{rms} is that for a very small value $|\tilde{\beta}_{rs} - \tilde{\beta}_{ms}|$, the weights become very large, such that estimates of β_{rs} and β_{ms} have to be similar, because otherwise the penalty term itself becomes huge. As group-specific models can be seen as varying coefficient models, the adaptive weighting allows to prove asymptotically normal estimates and consistent variable selection (for details, see Chapter 3). That means for group-specific models, asymptotic properties are available for a fixed number of second level units whereas for random effects models, large sample theory requires an increasing number of second level units.

As in Chapter 3, computational issues are met by the local quadratic approximations that are discussed in Chapter 2.

Regarding the choice of the penalty parameter λ, the hierarchical structure of the data has to be considered: For example, ordinary cross-validation based on omitting vectors of observations $\boldsymbol{y}_i^T = (y_{i1}, \ldots, y_{in_i})$, will, not work as excluding second level observations changes the model. Assume a simple model with group-specific intercepts, $g(\mu_{ij}) = \beta_{i0} + \boldsymbol{x}_{ij}^T\boldsymbol{\beta}$. For example, if the second level observation $\boldsymbol{y}_n$ is excluded from the data set, there are only $n-1$ group-specific intercepts in the model fitted on the training data set. When one wants to predict the outcome for the omitted observation $\boldsymbol{y}_n$, no estimate of β_{n0} is available. Therefore, it is preferable to use cross-validation methods that allow to estimate the group-specific effects of all observations. One strategy is to exclude only parts of the measurements observed for unit i. When using the observation $\boldsymbol{y}_i$ for validation, one randomly selects components from the vector $\boldsymbol{y}_i^T = (y_{i1}, \dots, y_{in_i})$ to obtain sub-vectors $\boldsymbol{y}_{i_1}$ and $\boldsymbol{y}_{i_2}$. The first one is kept in the training sample while the last one is used in the test sample. For k-fold cross-validation, all the observations that are used for validation are split into sub vectors, and only the first one is used in the training sample. In order to obtain stable estimates, the first sub vectors to be used in the training sample have to be sufficiently long.

Alternatively, the penalty parameters can be estimated by a generalized cross-validation (GCV) criterion, as proposed, for example, by O'Sullivan et al. (1986), see Section 2.2.4. The criterion avoids that the data is split into a training and a test data set. However, it requires to estimate the degrees of freedom of the model.

4.4. Numerical Experiments

The proposed penalized group-specific model is of special interest when there is clustered heterogeneity. However, penalized group-specific models are also an option when other assumptions, that are required for the random effects model, are not fulfilled. That is why we consider not only settings with clustered heterogeneity. In addition, we investigate settings with truly skew distributed random effects and with level 2 endogeneity. In order to extract the effect of the different types of model violations as good as possible, in all settings, the basic model is a random intercept model with one continuous covariate x_{ij}, one fixed effect β, respectively: $g(\mu_{ij}) = \beta_0 + b_{i0} + \beta x_{ij}$.

If the point of view is the group-specific model, the model contains the group-specific intercepts and the effect of one covariate x_{ij}: $g(\mu_{ij}) = \beta_{i0} + \beta x_{ij}$.

4.4.1. Data Generation

In order to generate data that contains any combination of skewly distributed random intercepts, level 2 endogeneity and clustered heterogeneity, different mechanisms are employed. If the random intercepts shall be distributed skewly, they are drawn from a χ^2_{df} distribution, where df denotes the degrees of freedom. In order to obtain comparable results for symmetrically and skewly distributed random intercepts, the realizations of the skew distributions are centered such that their expectation is zero.

Level 2 endogeneity means that a correlation between the random effects b_{i0}, the group-specific effects β_{i0} respectively, and the covariates x_{ij} is present. To incorporate such a correlation, we consider the joint distribution of $(b_{i0}, \boldsymbol{x}_i^T)$. Assume a multivariate normal distribution for $(b_{i0}, \boldsymbol{x}_i^T)$ with $\rho = corr(b_{0i}, x_{ij}) \neq 0$. Unfortunately, the covariance matrix of this distribution is not necessarily positive definite for arbitrary values of $corr(x_{ij}, x_{ik})$, $j \neq k$. Thus, we cannot simply draw random samples out of the joint distribution of b_{i0} and $\boldsymbol{x}_i^T$. Instead, we apply a sequential procedure that is based on two-dimensional distributions. In a first step, b_{i0} is generated by $b_{i0} \sim N(\mu_0 = 0, \sigma_0^2)$. In a second step, the bivariate normal distribution of b_{i0} and x_{ij} for a fixed value of j is considered. One realization of the univariate standard normal distribution is transformed such that it is a realization x_{ij} of the bivariate distribution of b_{i0} and x_{ij} – provided that there is already a realization of b_{i0}. This procedure is repeated for all j, $j = 1, \ldots, n_i$. The two steps are repeated for all i, $i = 1, \ldots, n$. In an exemplary setting with $\rho = 0.8$, the average empirical correlation of b_{i0} and x_{ij} based on 1000 replications is 0.8016 ($n = 30$, $n_i = 10$, $\sigma_0^2 = 4$, $\sigma_x^2 = 1$). The average range of $corr(x_{ij}, x_{ik})$, $j \neq k$, is $(0.4255, 0.8105)$. In an alternative setting, we consider skew distributions for b_{i0}. In this case, the joint distribution of $(b_{i0}, \boldsymbol{x}_i^T)$ is not multivariate normal but the sequential procedure can be applied with small modifications. Let, for example, b_{i0} be drawn from a χ^2_{df} distribution. The transformations to obtain x_{ij} are the same as before, but refer to the empirical counterparts of μ_0 and σ_0^2. With $b_{i0} \sim \chi^2_3$, b_{i0} centered such that $\mu_0 = 0$, and the same parameters as in the exemplary setting above, the average empirical correlations behave the same as for $b_{i0} \sim N(0, 4)$.

In order to construct clustered second level units, the random intercepts b_{i0}, the group-specific intercepts β_{i0} respectively, are ordered by size and assigned to clusters $C_1, \ldots, C_K$. If one considers $n = 30$ second level units in which $K = 5$ clusters are to be generated, each cluster contains six random effects; the six smallest in the first cluster, and so forth. Then, the mean of a cluster yields the cluster specific effect b_{k0}, β_{k0}, $k = 1, \ldots, K$. If one wants level 2 endogeneity as well as clustered second level units, the second level units are generated as described above and clustered afterwards.

4.4.2. Settings

With data of this kind, we consider different settings. In the first set of settings, the response is Gaussian: $\beta_0 = 1$, $\beta = 2$, $x_{ij} \sim N(0, 1)$. The distribution of the random intercepts is either symmetric or skew: $b_{i0} \sim N(0, 4)$ or $b_{i0} \sim \chi^2_3$. In all settings, $n = 30$. The number of clusters K in the second level units varies: $K \in \{30, 15, 5\}$. Settings with and without level 2 endogeneity are considered ($\rho = 0.8$ vs. $\rho = 0.0$). Moreover, the number of first level observations is varied; it is either $n_i = 10$ or $n_i = 5$. Since the variance of the responses determines the effective degrees of freedom, it has to be chosen carefully. We use the standard deviation $\sigma_\varepsilon = 6$, which yields effective degrees of freedom equal to 15.25 in the random effects model with $n_i = 10$, and equal to 10.36 in the random effects model with $n_i = 5$. Thus, one is not too close to the group-specific model ($\sigma_0^2 \to \infty$) but far away from the case without variation of the intercept.

For binary responses, we consider a logit model. The generation of the data is roughly the same, only two parameters differ: $\beta_0 = -0.3$, $\beta = 0.3$.

For each setting, we compare the performance of the unpenalized group-specific model, the proposed penalized group-specific model, the random effects model, and of the finite mixture models described in Section 4.2.1. For the penalized approaches, the penalty parameter λ is chosen by 5-fold cross-validation with the deviance as loss criterion or by the GCV criterion, see Section 2.2.4. The random effects models are estimated by the `R` package `lme4` (Bates et al., 2014). The finite mixture models are estimated by the `R` package `flexmix` (Grün and Leisch, 2008a).

Because for a binomial model, $n = 30$ is huge, estimates for the unrestricted group-specific model are quite unstable or do not exist; therefore, they are omitted. Accordingly, if adaptive weights are used, they do not rely on the unrestricted estimates but on an estimate obtained with a small ridge penalty.

As in Chapter 3, the estimation accuracy is measured in terms of the scaled squared errors/the mean squared error (MSE) of the coefficients of all $n_{rep} = 100$ replications. As we are especially interested in the potential bias of the estimates of β, the MSE is split into the contributions of the group-specific intercepts and into the contributions of the slopes β. In the case of the random effects models, the MSEs relate to the sum of the fixed and the random intercepts $\beta_0 + b_{i0}$. If the second level units are clustered, the "right" units should be merged. That is, the rate of falsely fused units should be low (false negatives/FN). The rate of units that should be in one cluster but are not (false positives/FP), should be small likewise. However, high FP rates are assessed less severe than high FN rates. Of course, when the second level units are not clustered, the FP rates are not defined.

4.4.3. Results

Figures 4.1–4.4 and Tables 4.1–4.2 present the corresponding results. The methods are abbreviated: "G" stands for the unpenalized group-specific model, "GL1" denotes the L_1-penalized estimates and "GL1a" the L_1-penalized estimates with adaptive weights. "R" stands for the random effects model, "AIC" and "BIC" for the finite mixture models with the respective model selection criterion. "CV" indicates 5-fold cross-validation; "GCV" denotes that the penalty parameter is chosen by the GCV criterion.

Gaussian Responses Figure 4.1 shows the boxplots of the squared errors for the setting with $n_i = 10$, without level 2 endogeneity and with normally distributed random intercepts. On top, there is no clustered heterogeneity; in the middle panel, there are 15 clusters of second level units whereas there are 5 clusters on bottom. Regarding the intercepts (left column), the unpenalized group-specific model performs slightly worse than the random effects model. The penalized group-specific models and the random effects model perform similarly. However, the adaptively weighted models perform slightly worse than the unweighted approaches and the random effects model. The smaller the number of clusters is, the more differ the results obtained with 5-fold cross-validation from those obtained with the GCV criterion. The slopes are estimated equally well by all methods (right column). It is remarkable that the random effects model performs very well, also in the case where clusters of parameters are present (second and third row). Moreover, the random effects model is not affected by truly skew distributed random intercepts (for illustration, see Appendix B). In contrast, the finite mixture model performs badly. As seen in Figure 4.2, the picture changes if there is level 2 endogeneity ($\rho = 0.8$). Now, the performance varies in terms of the squared errors: Regarding the intercepts, the unpenalized group-specific model and the random effects model perform equally well, but the variation of the group-specific model is large. The performance of the penalized group-specific model is distinctively better. Especially the adaptively weighted methods perform good. Regarding the slopes, it is seen that the random effects model performs as bad as the finite mixture models, whereas the group-specific model and its regularized versions perform very well. The combination of the non-adaptive regularized approach and of 5-fold cross-validation seems to have some issues. The same pattern is observed for $n_i = 5$.

Table 4.1 shows the FP and the FN rates for all settings with Gaussian responses with $n_i = 10$. The clustering performance of the regularized approaches is not overwhelming. Again, the combination “GL1, CV” has some issues. A possible explanation is that the folds for the cross-validation are so small that the penalty parameter that is needed for stable estimation is relatively large – such that the overall performance suffers when there is no additional correction as, for example, adaptive weights. Except for the finite mixture models, the FN rates are very small. Even though the BIC has the lowest FP rates of all methods that cluster second level units and in all settings, and even though the approach “GL1, CV” has the lowest FN rates, the adaptively penalized approaches are recommended the first due their estimation accuracy (especially in settings with level 2 endogeneity).

Binomial Responses The results for the settings with binomial responses, are shown in Figures 4.3, 4.4 and in Table 4.2. We focus on the setting with $n_i = 10$ and skewly distributed random effects. In Figure 4.3, there is no level 2 endogeneity. Here, the intercepts of the adaptively penalized approaches are estimated the best for $K \in \{15, 30\}$. The slopes are estimated equally well by all approaches. In Figure 4.4, there is level 2 endogeneity ($\rho = 0.8$). In contrast to Figure 4.3, we observe a clear trend regarding the intercepts if there is level 2 endogeneity. The pure L1 penalized approaches perform the best – followed by the adaptively weighted penalized approaches, the random effects models and the finite

mixture models. The same pattern is observed for the slopes. The results for the settings with symmetrically distributed random intercepts are presented in the Appendix: Without level 2 endogeneity and for Gaussian random intercepts, the intercepts are estimated similarly for all approaches except for the finite mixture models. Again, the slopes are estimated equally well by all approaches. In contrast to the settings with Gaussian responses, the results for binomial responses are affected by both, level 2 endogeneity and skewly distributed random intercepts. For $n_i = 5$, the approach "GL1, CV" has some issues. The other methods perform similar.
Table 4.2 shows the FP and the FN rates for symmetric and skew random intercepts, $n_i = 10$. Again, the adaptively weighted approaches perform the best - in the sense that these approaches are good compromise between a reasonable clustering performance and relatively high estimation accuracy. Overall, the clustering performance is better than for Gaussian responses. For $n_i = 5$, the FN rates are higher than for $n_i = 10$. This indicates that more second level units are fused than it is necessary. This is mainly the case when the optimal choice of the penalty parameter λ is relatively large. That is for $n_i = 5$, the penalty is really needed to stabilize the estimation.

Remark If one is especially interested in the detection of clusters of second level units, the clustering performance for both, Gaussian and binomial responses, can be considerably improved by an additional refit in the cross-validation procedure. However, with an additional refit, the estimation accuracy may suffer – it is only recommended in combination with adaptive weights and when the focus is on the clustering performance. Detailed results for all settings, with and without additional refit, can be found in Appendix B.

Random Intercepts				G	GL1, CV	GL1, GCV	GL1a, CV	GL1a, GCV	R	AIC	BIC
Gaussian	$K = 30$	$\rho = 0$	FP	-	-	-	-	-	-	-	-
			FN	0.00	0.06	0.04	0.12	0.11	0.00	0.60	0.81
		$\rho = 0.8$	FP	-	-	-	-	-	-	-	-
			FN	0.00	0.73	0.07	0.22	0.14	0.00	0.95	1.00
	$K = 15$	$\rho = 0$	FP	1.00	0.52	0.93	0.79	0.83	1.00	0.11	0.02
			FN	0.00	0.47	0.05	0.17	0.13	0.00	0.84	0.98
		$\rho = 0.8$	FP	1.00	0.25	0.92	0.74	0.83	1.00	0.04	0.00
			FN	0.00	0.74	0.07	0.21	0.14	0.00	0.95	1.00
	$K = 5$	$\rho = 0$	FP	1.00	0.70	0.94	0.80	0.84	1.00	0.19	0.04
			FN	0.00	0.27	0.04	0.13	0.11	0.00	0.68	0.92
		$\rho = 0.8$	FP	1.00	0.27	0.93	0.79	0.82	1.00	0.05	0.00
			FN	0.00	0.73	0.06	0.15	0.12	0.00	0.93	1.00
χ^2_3	$K = 30$	$\rho = 0$	FP	-	-	-	-	-	-	-	-
			FN	0.00	0.13	0.04	0.17	0.13	0.00	0.71	0.87
		$\rho = 0.8$	FP	-	-	-	-	-	-	-	-
			FN	0.00	0.60	0.06	0.15	0.12	0.00	0.93	1.00
	$K = 15$	$\rho = 0$	FP	1.00	0.73	0.94	0.81	0.84	1.00	0.15	0.04
			FN	0.00	0.25	0.05	0.15	0.13	0.00	0.75	0.93
		$\rho = 0.8$	FP	1.00	0.47	0.94	0.79	0.82	1.00	0.07	0.00
			FN	0.00	0.52	0.05	0.15	0.12	0.00	0.90	1.00
	$K = 5$	$\rho = 0$	FP	1.00	0.75	0.93	0.79	0.83	1.00	0.16	0.04
			FN	0.00	0.23	0.04	0.15	0.11	0.00	0.70	0.91
		$\rho = 0.8$	FP	1.00	0.39	0.93	0.78	0.83	1.00	0.09	0.00
			FN	0.00	0.61	0.06	0.18	0.12	0.00	0.88	1.00

Table 4.1.: Estimates of FP and FN rates for the settings with Gaussian response, $n_i = 10$.

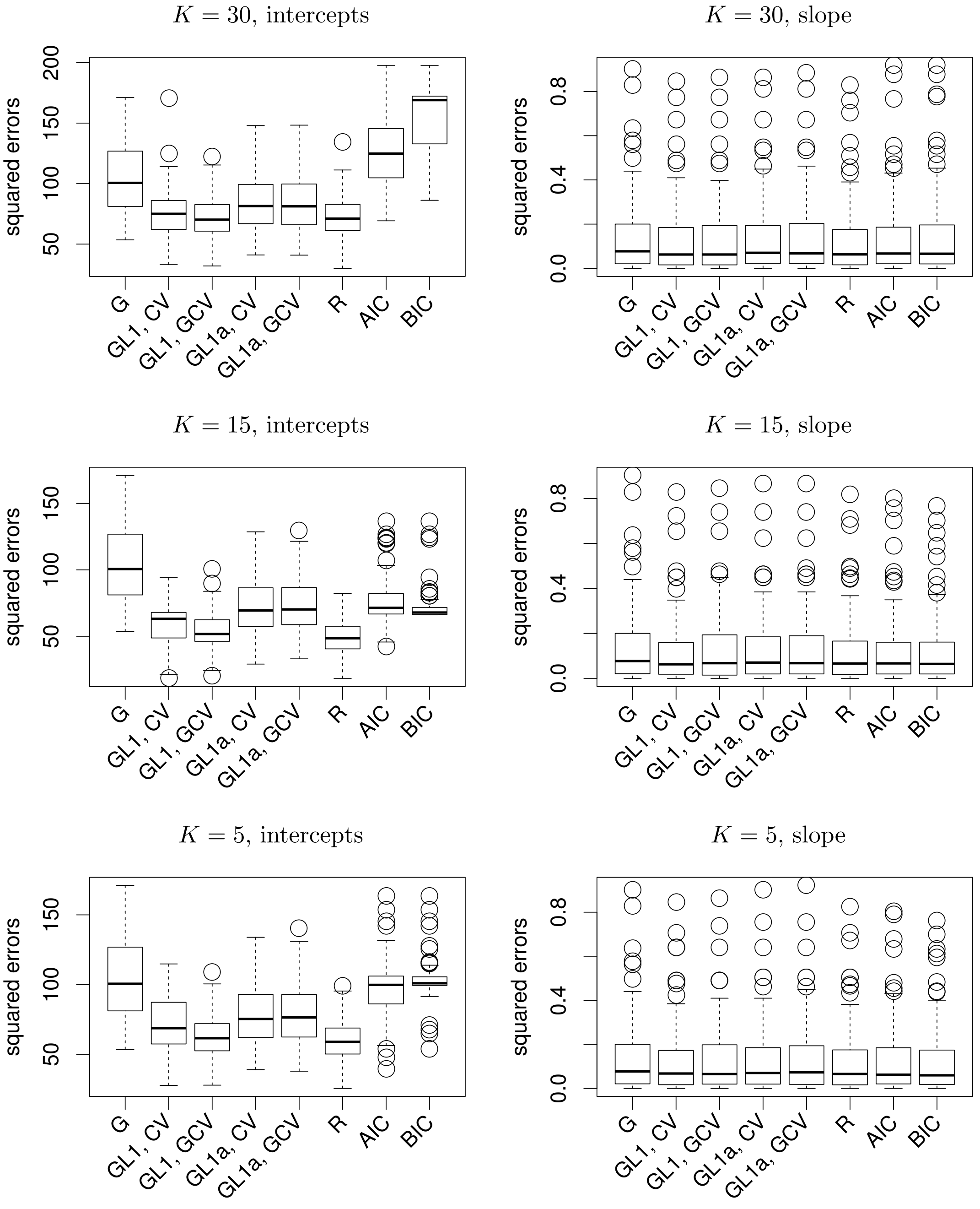


Figure 4.1.: Squared errors for the settings with Gaussian response and $b_{i0} \sim N(0, 4)$. The number of clusters K varies with the rows. The left panel relates to the intercepts, the right panel to the slopes. $\rho = 0.0$, $n_i = 10$. The medians in the boxplots are robust estimates of the MSE.

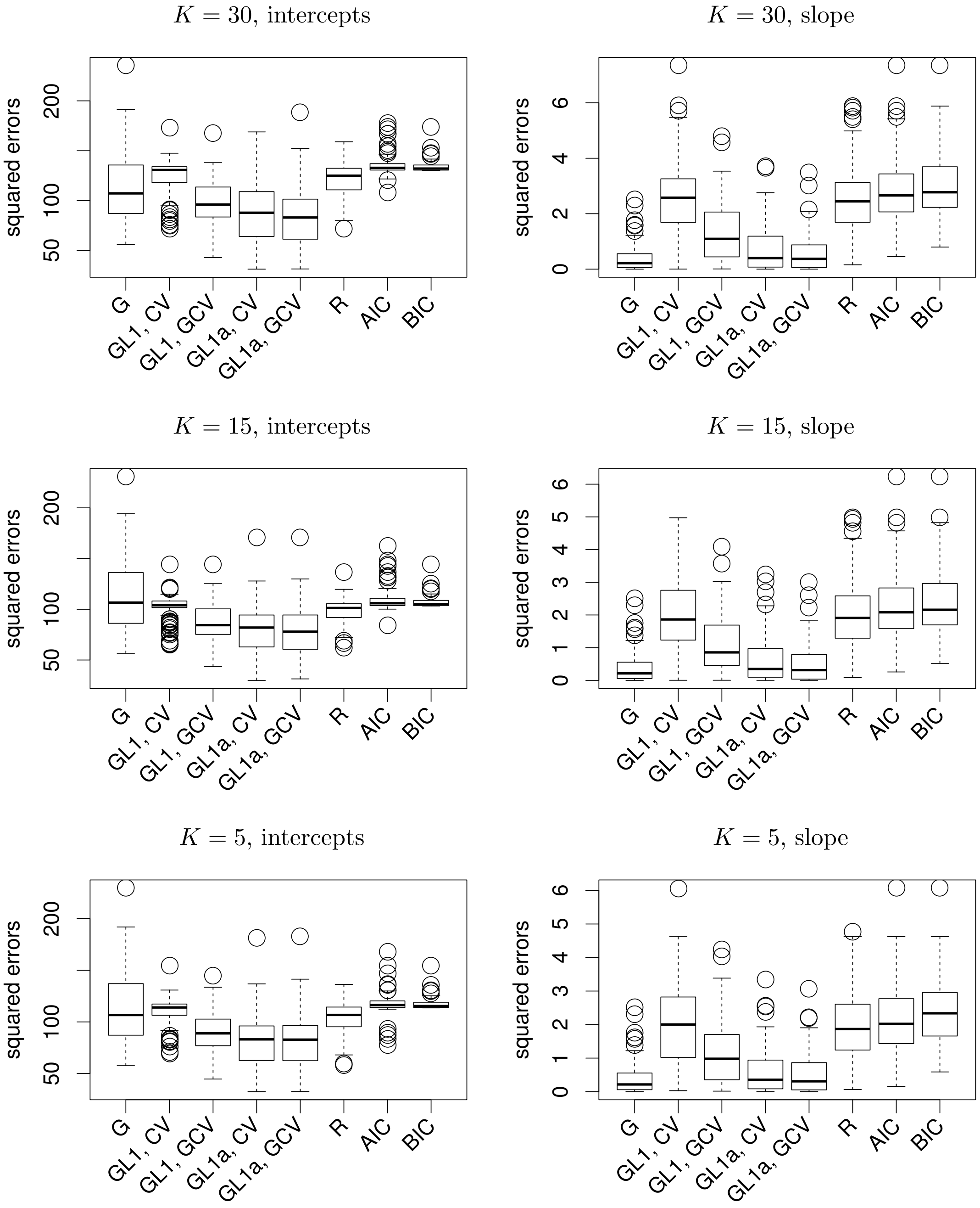


Figure 4.2.: Squared errors for the settings with Gaussian response and $b_{i0} \sim N(0, 4)$. The number of clusters K varies with the rows. The left panel relates to the intercepts, the right panel to the slopes. $\rho = 0.8$, $n_i = 10$. The medians in the boxplots are robust estimates of the MSE.

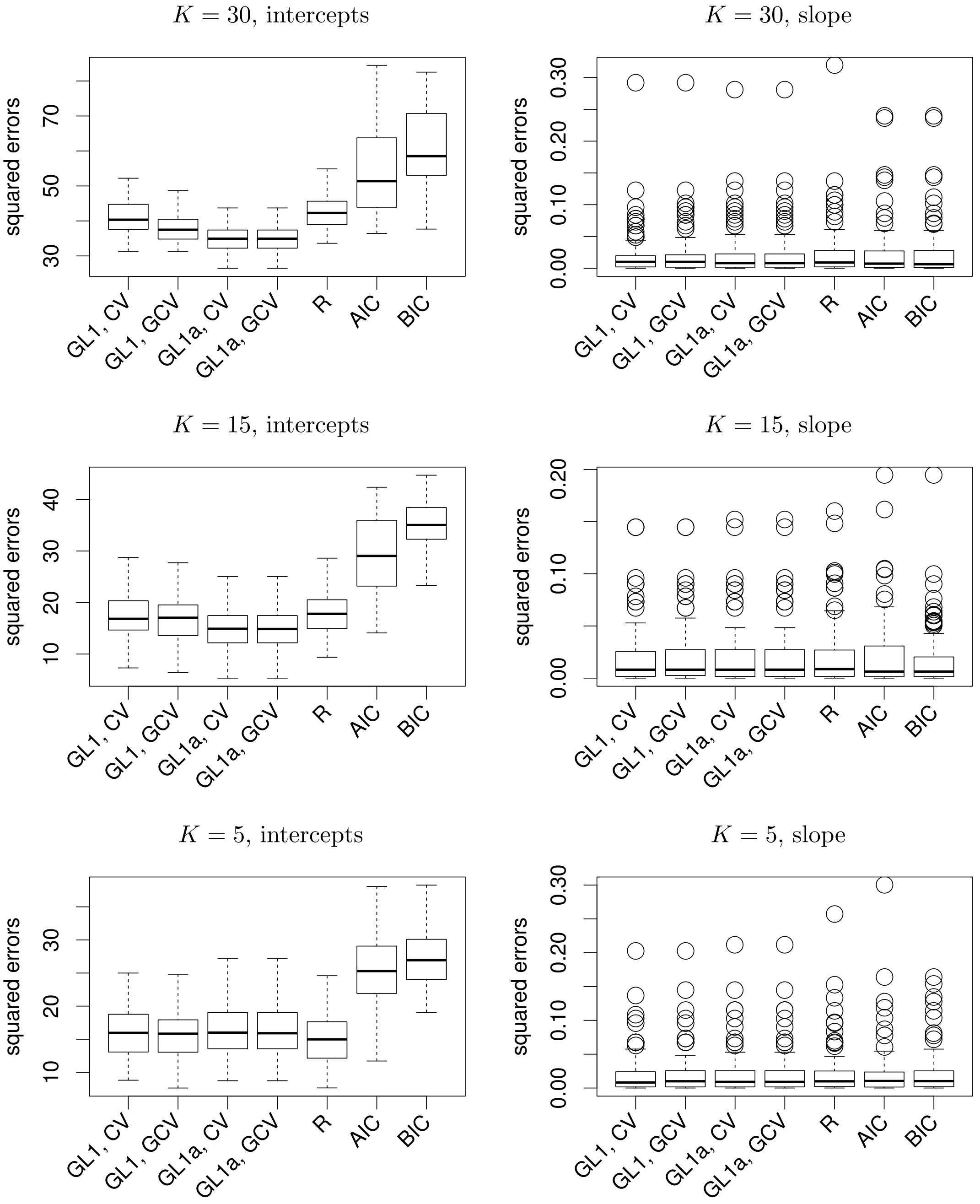


Figure 4.3.: Squared errors for the settings with binomial response and $b_{i0} \sim \chi^2_3$. The number of clusters K varies with the rows. The left panel relates to the intercepts, the right panel to the slopes. $\rho = 0.0$, $n_i = 10$. The medians in the boxplots are robust estimates of the MSE.

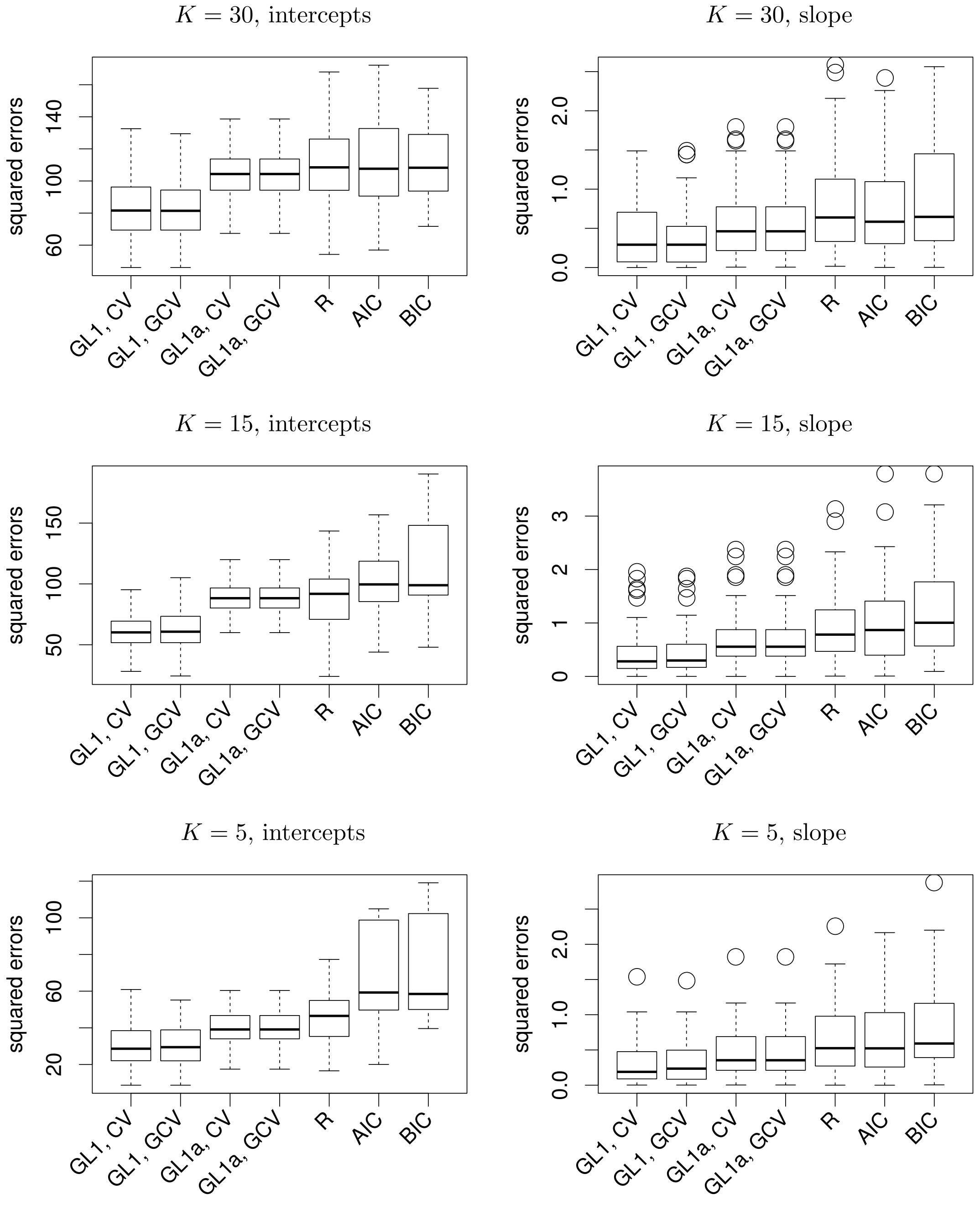


Figure 4.4.: Squared errors for the settings with binomial response and $b_{i0} \sim \chi^2_3$. The number of clusters K varies with the rows. The left panel relates to the intercepts, the right panel to the slopes. $\rho = 0.8$, $n_i = 10$. The medians in the boxplots are robust estimates of the MSE.

Random Intercepts				GL1, CV	GL1, GCV	GL1a, CV	GL1a, GCV	R	AIC	BIC
Gaussian	$K = 30$	$\rho = 0$	FP	-	-	-	-	-	-	-
			FN	0.04	0.04	0.08	0.08	0.00	0.39	0.47
		$\rho = 0.8$	FP	-	-	-	-	-	-	-
			FN	0.14	0.03	0.15	0.15	0.00	0.71	0.88
	$K = 15$	$\rho = 0$	FP	0.75	0.77	0.65	0.65	1.00	0.19	0.13
			FN	0.08	0.07	0.11	0.11	0.00	0.39	0.48
		$\rho = 0.8$	FP	0.87	0.89	0.67	0.67	1.00	0.20	0.12
			FN	0.04	0.02	0.13	0.13	0.00	0.49	0.66
	$K = 5$	$\rho = 0$	FP	0.82	0.83	0.70	0.71	1.00	0.21	0.12
			FN	0.03	0.02	0.05	0.05	0.00	0.27	0.36
		$\rho = 0.8$	FP	0.88	0.88	0.68	0.68	1.00	0.18	0.14
			FN	0.02	0.02	0.12	0.12	0.00	0.37	0.45
χ^2_3	$K = 30$	$\rho = 0$	FP	-	-	-	-	-	-	-
			FN	0.07	0.06	0.11	0.11	0.00	0.46	0.56
		$\rho = 0.8$	FP	-	-	-	-	-	-	-
			FN	0.03	0.03	0.14	0.14	0.00	0.50	0.59
	$K = 15$	$\rho = 0$	FP	0.83	0.85	0.72	0.72	1.00	0.19	0.14
			FN	0.06	0.05	0.10	0.10	0.00	0.42	0.50
		$\rho = 0.8$	FP	0.87	0.87	0.67	0.67	1.00	0.19	0.13
			FN	0.03	0.02	0.13	0.13	0.00	0.48	0.61
	$K = 5$	$\rho = 0$	FP	0.82	0.85	0.74	0.74	1.00	0.19	0.14
			FN	0.05	0.04	0.07	0.07	0.00	0.38	0.46
		$\rho = 0.8$	FP	0.89	0.89	0.71	0.71	1.00	0.19	0.13
			FN	0.02	0.02	0.11	0.11	0.00	0.45	0.58

Table 4.2.: Estimates of FP and FN rates for the settings with binomial response, $n_i = 10$.

4.5. Mortality after Myocardial Infarction

In the beta blocker example considered in the introduction, one models the mortality rate after myocardial infarctions depending on the study center and the treatment group assigned to the patients. A classical model that has been used on this data set is the random intercept model, which has the form logit $\mathbb{P}(y_{ij} = 1) = \beta_0 + b_{i0} + \beta_T \cdot \text{Treatment}_{ij}$, where the random intercepts b_{i0} are assumed to follow a normal distribution, and where $\text{Treatment}_{ij} \in \{-1, 1\}$ codes the treatment in hospital i for patient j. The model with group-specific intercepts has the form

$$\text{logit } \mathbb{P}(y_{ij} = 1) = \beta_{i0} + \beta_T \cdot \text{Treatment}_{ij}, \quad i = 1, \ldots, 22 \text{ Centers}, \tag{4.9}$$

where β_{i0} are fixed unknown parameters. Table 4.3 shows the results for the unpenalized group-specific model, the random intercept model, and the results of the group-specific model with the adaptively weighted penalty (4.8). For comparison, the results of the estimates that are obtained with the finite mixture model of Grün and Leisch (2008b) are added where the number of components K is chosen by the AIC and the BIC criterion. The intercept coefficients are ordered such that their structure becomes obvious. We see that the estimates of all approaches are quite similar. Moreover, the range of the coefficients corresponds to the number of effective parameters in the specific approaches. The model with the most parameters, the unpenalized group-specific model, is characterized by the largest range of coefficients. The range of the coefficients is smaller for the random intercept model. It decreases even more for the penalized group-specific model. It is the smallest for the finite mixture models where only four (AIC) and three (BIC) clusters of study centers are detected. The fitted treatment effect has approximately the same size in all models. Figure 4.5 shows the corresponding coefficient build-ups of the penalized approach; the dotted line denotes the penalty parameter selected by the GCV criterion with an additional refit ($\lambda_{CV} = 0.65$). There are basically five clusters of hospitals that are to be distinguished in terms of the basic risk captured by the intercepts, although numerically, one obtains more clusters. However, clusters with effects -2.35 and -2.36 can hardly be considered as having different effects. Note that the predictor in model (4.9) corresponds to a generalized linear model with the nominal covariates Center and Treatment. However, when the Centers are coded as a nominal covariate, the choice of the reference category affects the estimate of a penalized nominal covariate crucially. With group-specific intercepts for the hospitals, there is no need for a reference category; all hospitals are penalized in the same way. Moreover, there is an intrinsic interpretation for the group-specific intercepts, whereas a nominal covariate always draws comparisons with the reference category which is arbitrary.

Coefficients		Group-Specific Model	Random Intercept Model	Group-Specific, Pen. (4.8)	Finite Mixture Model	
					AIC	BIC
Center-specific Intercept	$\beta_{15,0}$	-1.4782	-1.5520	-1.70	-1.5688	-1.7388
	$\beta_{12,0}$	-1.5644	-1.6053			
	$\beta_{16,0}$	-1.5999	-1.6494			
	$\beta_{20,0}$	-1.6038	-1.6524			
	$\beta_{7,0}$	-1.8832	-1.8917	-1.92	-1.9171	
	$\beta_{17,0}$	-2.0801	-2.1065	-2.35	-2.3873	-2.3793
	$\beta_{9,0}$	-2.0910	-2.1079			
	$\beta_{8,0}$	-2.2083	-2.2133	-2.36		
	$\beta_{3,0}$	-2.2370	-2.2575	-2.37		
	$\beta_{21,0}$	-2.2832	-2.2859			
	$\beta_{2,0}$	-2.3059	-2.3097			
	$\beta_{6,0}$	-2.3113	-2.3162			
	$\beta_{10,0}$	-2.3840	-2.3832			
	$\beta_{11,0}$	-2.4278	-2.4239			
	$\beta_{1,0}$	-2.4798	-2.4144	-2.38		
	$\beta_{5,0}$	-2.5015	-2.4881			
	$\beta_{4,0}$	-2.5189	-2.5151			
	$\beta_{14,0}$	-2.7862	-2.7670	-2.72	-2.9626	-2.9628
	$\beta_{18,0}$	-3.0433	-2.8802	-2.87		
	$\beta_{22,0}$	-3.0610	-3.0122			
	$\beta_{13,0}$	-3.1155	-3.0020			
	$\beta_{19,0}$	-3.4942	-3.1536	-2.89		
Treatment	β_T	-0.1305	-0.1305	-0.13	-0.1292	-0.1291

Table 4.3.: Estimates for the beta blocker data. The intercept coefficients are ordered such that their structure becomes obvious. Presented intercept coefficients of the random effects model are the sum of the fixed and the random intercepts. Horizontal lines denote clusters of coefficients.

4.6. Extension: Models with Vector Fused Penalties

If more than one parameter is expected to be group-specific, the proposed penalties allow for different clusters of second level units in different components of the covariate $\boldsymbol{z}_{ij}$. Thus, for each component of the predictor, one obtains a disjunct partition $C_1^{(s)}, \dots, C_K^{(s)}$, where s refers to the component in $\boldsymbol{z}_{ij}$. It depends on the application, if this is desirable. If one wants one consistent partition that is based on the whole vector $\boldsymbol{z}_{ij}$, a modified penalty can be used. Let again variables that are in $\boldsymbol{z}_{ij}$, be excluded in $\boldsymbol{x}_{ij}$. Then, a penalty that fuses the second level units *simultaneously* is

$$J(\boldsymbol{\alpha}) = \sum_{r>m} \|\boldsymbol{\beta}_r - \boldsymbol{\beta}_m\|_2 , \tag{4.10}$$

where $\|\boldsymbol{\xi}\|_2 = \{\xi_1^2 + \dots + \xi_q^2\}^{1/2}$ denotes the L_2-norm of a q-dimensional vector $\boldsymbol{\xi}$. The penalty enforces that the whole vectors of parameters $\boldsymbol{\beta}_r$ and $\boldsymbol{\beta}_m$ are fused. In contrast to

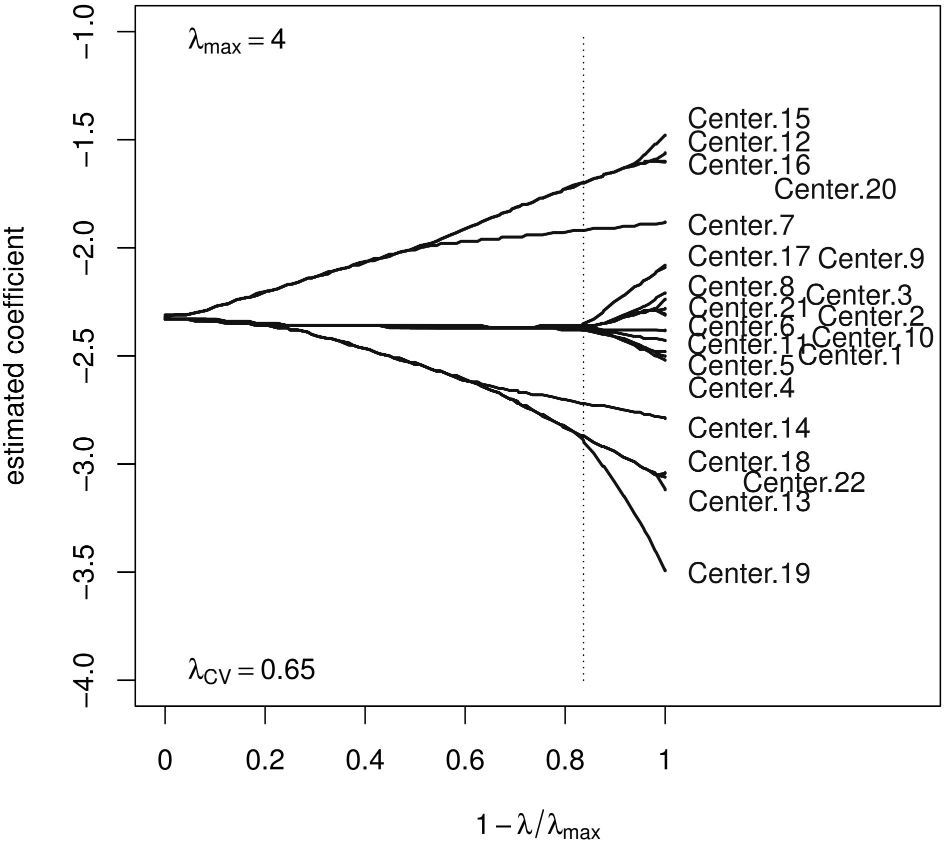


Figure 4.5.: Coefficient build-ups of the penalized model for the beta blocker data. The very right end of the figure relates to $\lambda = 0$, that is, to the ML estimate. The left end relates to the minimal value of λ giving maximal penalization; in this case $\lambda = 4$.

the componentwise penalty (4.7), the effect of the penalty (4.10) is that the group-specific coefficients $\boldsymbol{\beta}_r$ are shrunk towards each other and one obtains only one partition $C_1, \dots, C_K$ of C that is based on the whole vector $\boldsymbol{z}_{ij}$. The approach works in a similar way as the group Lasso by Yuan and Lin (2006). However, the group Lasso refers to the simultaneous selection of a group of parameters, whereas penalty (4.10) refers to the fusion of a set of coefficients.

As for componentwise fusion penalties, one can use an adaptive version of the penalty, see, for example, Wang and Leng (2008). It is given by $J(\boldsymbol{\alpha}) = \sum_{r>m} w_{rm} \, ||\boldsymbol{\beta}_r - \boldsymbol{\beta}_m||_2$, where $w_{rm} = ||\tilde{\boldsymbol{\beta}}_r - \tilde{\boldsymbol{\beta}}_m||_2^{-1}$ with $\sqrt{n}$-consistent estimates $\tilde{\boldsymbol{\beta}}_r$, $\tilde{\boldsymbol{\beta}}_m$. Interestingly, the componentwise fusion penalty (4.7) can be written as

$$J(\boldsymbol{\alpha}) = \sum_{r>m} ||\boldsymbol{\beta}_r - \boldsymbol{\beta}_m||_1 \, ,$$

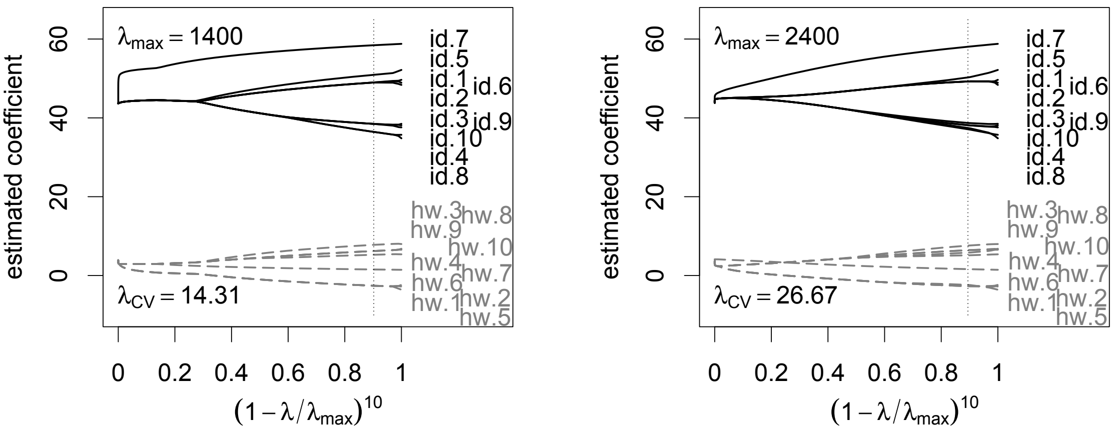


Figure 4.6.: Coefficient build-ups for the NELS:88 data with group-specific intercepts and group-specific covariate homework; left panel: componentwise penalty (4.8); right panel: simultaneous penalty (4.10) with adaptive weights.

where $\|\boldsymbol{\xi}\|_1 = |\xi_1| + \cdots + |\xi_q|$ denotes the L_1-norm of a q-dimensional vector $\boldsymbol{\xi}$. Thus, penalties (4.7) and (4.10) basically differ in the applied norm. Of course, it is possible to combine the penalties in specific applications. In the simplest case, let $\boldsymbol{\beta}_i$ be partitioned into $\boldsymbol{\beta}_i^T = (\boldsymbol{\beta}_{i1}^T, \boldsymbol{\beta}_{i2}^T)$ and use

$$J(\boldsymbol{\alpha}) = \sum_{r>m} \|\boldsymbol{\beta}_{r1} - \boldsymbol{\beta}_{m1}\|_s + \|\boldsymbol{\beta}_{r2} - \boldsymbol{\beta}_{m2}\|_t \,,$$

where $s, t \in \{1, 2\}$. One can, for example, employ the L_1-norm for the intercept and the L_2-norm for the remaining covariates. Again, computational issues are met by local quadratic approximations as described in Chapter 2.

4.7. Analysis of Selected Schools in the National Education Longitudinal Study of 1988

In order to illustrate the extended penalty (4.10), we analyze data on $n = 10$ selected schools out of the National Education Longitudinal Study of 1988 (NELS:88, Curtin et al., 2002). The data contains information on eighth-graders surveyed in 1988 in different schools in the United States and is a subsample of a nation-wide longitudinal study. The number of pupils per school varies. We consider the standardized mathematics score (y_{ij}, measured between 0 and 100, the higher the better), the time in hours spent on math homework weekly (hw) and the school (id) of 260 pupils. We investigate whether the math skills and the impact of the duration of the homework do differ over the schools when explaining the

		School (id)									
Penalty	Coeff.	1	2	3	4	5	6	7	8	9	10
none	β_{i0}	49.66	48.36	38.49	35.70	52.17	48.91	58.78	34.81	38.09	37.59
	β_{i1}	-2.80	-2.60	7.99	5.38	-3.60	-2.37	1.45	6.80	6.68	6.54
(4.8)	β_{i0}	48.98	48.97	38.47	36.44	50.92	48.98	58.43	36.44	38.47	38.47
	β_{i1}	-2.59	-2.59	7.79	5.27	-2.59	-2.59	1.54	6.21	6.21	6.21
(4.10)	β_{i0}	49.24	49.24	38.71	37.11	50.26	49.24	58.05	37.43	38.14	38.14
	β_{i1}	-2.69	-2.69	7.59	5.13	-2.38	-2.69	1.62	5.81	6.43	6.43

Table 4.4.: Estimated coefficients for the NELS:88 data obtained with model (4.11) and different penalization strategies.

mathematics score. The mathematics score is assumed to be Gaussian and we fit the linear model

$$y_{ij} = \beta_{i0} + \beta_{i1} \cdot \mathrm{hw}_{ij} + \varepsilon_{ij}, \quad i = 1, \dots, 10 \text{ Schools.} \tag{4.11}$$

In contrast to the example in Section 4.5, we assume group-specific intercepts β_{i0} and group-specific slopes β_{i1}. Hence, we can compare the componentwise penalty (4.8) and penalty (4.10) for simultaneous clusters, both adaptively weighted. Figure 4.6 shows the resulting coefficient build-ups. The left panel refers to penalty (4.8); that is, the pairwise differences of the group-specific intercepts are penalized by the adaptive Lasso and so are the differences of the group-specific slopes. In the right panel, the pairwise difference of the group-specific intercepts and the slopes are penalized simultaneously by penalty (4.10) with adaptive weights. In both panels, a dotted line marks the optimal results according to 5-fold cross-validation with the predictive deviance as a loss criterion. For the optimal penalty parameters, with both penalties, some schools are clustered. With penalty (4.8), one obtains two separate partitions of schools. Regarding the intercepts, schools $\{4, 8\}$, $\{3, 9, 10\}$ and $\{1, 6\}$ are merged to clusters; regarding the slopes, schools $\{1, 2, 5, 6\}$ and $\{8, 9, 10\}$ are merged while the other schools have individual effects. With penalty (4.10), one obtains one uniform partition: Schools $\{1, 2, 6\}$ and $\{9, 10\}$ are found to have similar effects for the intercepts and the slopes while the other schools form its own clusters. Table 4.4 gives the exact results. With both penalties, an interesting effect is seen: Schools with relatively low intercepts tend to have larger effects regarding the homework and vice versa. The effect is seen, for example, for the schools 3 and 5. In school 3, the average math skills without homework are relatively low while the effect of the homework is high. In school 5, the group-specific intercept is high; the impact of the homework is actually negative. A possible explanation is that in schools with higher average math skills, the time required for the homework might be an indicator for inertial pupils; while in schools with lower initial skills, the time spent on math homework might be a surrogate for the extent of the homework.

4.8. Remarks

In this chapter, we compare three different approaches that take the heterogeneity in hierarchical data into account: Finite mixture models, random effects models, and group-specific models. We are especially interested in situations where the second level units might be clustered and want to make as few assumptions as possible.

Finite mixture models assume that the effects of the second level units are drawn from an unknown finite set of effects. The performance is acceptable only for settings with very few clusters in the underlying data structure.

In random effects models, the effects of the observed second level units are assumed to be a random sample of a continuous distribution. Without modifications, random effects models do not allow for clustered second level units. When there is level 2 endogeneity in the data, the estimates of random effects model are biased.

In contrast, group-specific models rely only on the observed second level units, but the number of parameters is relatively large. The efficiency of the estimates is reduced. These problems can be solved by penalizing the group-specific model. Moreover, with the proposed fused Lasso penalty on group-specific coefficients, the second level units can be clustered – such that penalized group-specific models meet all our requirements for data with clustered heterogeneity. No distributional assumption as in random effects models are needed.

In numerical experiments, the estimation accuracy of the penalized approach is shown to be competitive. Although the identification of clusters can be enhanced, the estimation accuracy is much better than for the ML estimate. If the assumptions for the random effects model are not fulfilled for other reason, for example, due to level 2 endogeneity, the proposed Lasso-type penalty has substantial advantages.

5. Selection and Fusion of Categorical Covariates with L_0-Type Penalties

5.1. Introduction

In the majority of regression problems, at least some of the available covariates are categorical. A categorical covariate has to be coded. Depending on the number of categorical covariates and on the number of levels they have, the number of coefficients can become huge. Hence, the accuracy of estimates can be poor. Moreover, when including categorical variables, users want to know if and how these covariates determine the response, and, in particular, which categories have to be distinguished. Typically, there are subsets of categories that have the same effect on the response variable. Recently, various approaches to obtain selection and fusion of categories by regularized estimation have been proposed: Bondell and Reich (2009) propose to apply the fused Lasso (Tibshirani et al., 2005) to the coefficients of a nominal covariate; all pairwise differences of coefficients are penalized. For ordered factors, it is more appropriate to penalize differences of adjacent coefficients, see Gertheiss and Tutz (2010) and Tutz and Gertheiss (2014). However, Lasso-type penalties come with shrinkage effects that depend on the coefficients' absolute values (Fan and Li, 2001). As a consequence, there are often strong shrinkage effects for large (differences of) coefficients while small (differences of) coefficients are estimated to be non-zero. When the focus is on the fusion and the selection of categories, one wants to avoid such effects. To enhance Lasso-type penalties, Zou (2006) proposes adaptive weights; each penalty term is weighted by its inverse maximum likelihood (ML) estimate. It yields asymptotically normal and consistent model selection. However, the quality of the adaptive weights depends on the quality of the ML estimate that can be poor.

This chapter is a modified version of Oelker, Pößnecker, and Tutz (2015). The original version of Oelker, Pößnecker, and Tutz (2015) is published in Statistical Modelling, Vol. 15, No. 4. Copyright © 2015 SAGE Publications. All rights reserved. Reproduced with the permission of the copyright holders and the publishers Sage Publications India Pvt. Ltd, New Delhi. For more information on the contributions of the authors and on textual matches, see page 5.

As an alternative, we propose L_0 penalization for categorical effects, where the L_0 "norm" is defined as the number of non-zero entries in a vector. Like Bondell and Reich (2009) or Gertheiss and Tutz (2010), we consider differences of coefficients; but instead of the absolute value, the L_0 norm is applied to the differences of coefficients. The difference between unordered and ordered factors is taken into account by using all pairwise differences or only differences of adjacent categories. Computational issues are met by local quadratic approximations. The optimization problem is related to model selection with information criteria like the Akaike information criterion (AIC; see, for example, Bozdogan, 1987) or the Bayesian information criterion (BIC; Schwarz, 1978). As the proposed penalty allows for the fusion of categories, it extends this approach. L_0 penalization is an established approach in some fields of statistics: It is applied to wavelets (Antoniadis and Fan, 2001) and to signals (see Lu and Zhang, 2010, Rippe et al., 2012).

Moreover, minimizing (approximations of) constrained L_0 terms is employed to find sparse representations of signals; see, for example, Donoho and Elad (2003), Wipf and Rao (2005), Mancera and Portilla (2006) or Ge et al. (2011).

Chapter 5 is organized as follows: Section 5.2 motivates L_0 penalization for categorical effects in generalized linear models. In Section 5.3, we introduce the method; computational issues, the relation to best subset selection and some generalizations are discussed. Section 5.4 investigates the numerical properties of the proposed method. In Section 5.5, the unemployment rates in Germany between 2005 and 2010 are analyzed. We investigate which state-specific intercepts are clustered in a model with a global temporal trend.

5.2. Framework and L_1-Type Fusion Penalties

In what follows, we assume a generalized linear model (GLM) with response y_i for observation i, $i = 1, \dots, n$. As a start, we consider only one categorical covariate $\boldsymbol{x} = (x_i, \dots, x_n)^T$ with levels $0, \dots, k$. Let the rows of the design matrix $\boldsymbol{X}$ be given by $\boldsymbol{x}_i^T = (1,\ x_{i1}, \dots, x_{ik})$ with $x_{ir} = 1$ if x_i takes the value r and $x_{ir} = 0$ otherwise, $r = 1, \dots, k$. This representation refers to dummy coding with category 0 as reference category, $\beta_0 = 0$. The corresponding predictor is defined as $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_{int}, \beta_1, \dots, \beta_k)^T$ is the coefficient vector and β_{int} denotes the intercept. For the response $y_i|x_i$, a simple exponential family with log-likelihood $l_n(\boldsymbol{\beta})$ is assumed:

$$l_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{y_i \vartheta_i(\mu_i) - b(\vartheta_i(\mu_i))}{\varphi} + c(y_i, \varphi),$$

where $\vartheta_i(\mu_i)$ denotes the natural parameter, $b(\cdot)$ is a specific function corresponding to the type of the exponential family, $c(\cdot)$ is the log-normalization constant and φ the dispersion parameter (compare Fahrmeir and Tutz, 2001). The observations y_i are assumed to be conditionally independent. Response and predictor are linked by the response function $h(\eta_i)$ which is twice continuously differentiable with $\det(\partial h/\partial\eta_i) \neq 0 \; \forall i$. That is, we assume

$$\mu_i = \mathbb{E}(y_i|x_i) = h(\eta_i). \tag{5.1}$$

For more details on GLMs, see, for example, Fahrmeir and Tutz (2001). Estimates $\hat{\boldsymbol{\beta}}$ are obtained by minimizing the negative log-likelihood. Accounting for a penalty, the objective function is defined as

$$\mathcal{M}_{pen}(\boldsymbol{\beta}) = -l_n(\boldsymbol{\beta}) + \lambda \cdot P(\boldsymbol{\beta}), \tag{5.2}$$

where $P(\boldsymbol{\beta})$ denotes the penalty and where $\lambda \geq 0$ is the penalty parameter. The larger λ is, the stronger is the impact of the penalty. For $\lambda = 0$, the ML estimate is obtained.

The choice of the penalty $P(\boldsymbol{\beta})$ is crucial. The Lasso (Tibshirani, 1996) penalizes the absolute values of coefficients and enforces variable selection. One obtains sparse but shrunken estimates. For dummy-coded categorical covariates, this is not the best choice. Setting parameters to zero corresponds to the fusion with the reference category which can be chosen arbitrarily. Even though this problem can be handled by coding the categorical covariates differently – for example, as the deviation from a mean level ("effect coding") or as the deviation from adjacent categories ("split coding") – penalties that contain differences of parameters as proposed by Tibshirani et al. (2005), Bondell and Reich (2009) or Gertheiss and Tutz (2010) are a common choice. They encourage the fusion of coefficients, and thus of categories, irrespectively of the coding, and they allow one to fuse coefficients subject to more than k constraints. However, fusion-type penalties come along with some problems.

For an *ordered* effect, fusion-type penalties consider the differences of parameters that refer to adjacent categories, including the reference category 0. The corresponding Lasso-type penalty has the form

$$P(\boldsymbol{\beta}) = \sum_{r=1}^{k} |\beta_r - \beta_{r-1}|. \tag{5.3}$$

However, this penalty does not always enforce fusion efficiently. If the coefficients are ordered, for example, in the form $0 = \beta_0 \leq \beta_1 \leq \ldots \leq \beta_k$, and if one is close to the true values, that is, in the range where the estimated parameters are ordered, the effective penalty is $P(\boldsymbol{\beta}) = \sum_{r=1}^{k} |\beta_r - \beta_{r-1}| = |\beta_k - \beta_0| = |\beta_k|$. That means, the approach basically penalizes the range of the coefficients. The problem is even more obvious in an orthonormal linear model with one ordered effect and without an intercept. That is, $\boldsymbol{X}^T\boldsymbol{X}$ is the identity matrix $\mathbb{I}_{(k+1)\times(k+1)}$, each entry of $\boldsymbol{\beta}$ corresponds to one level of the ordered effect. Situations like this are, for example, typical for models with categorical effect modifiers or models with group specific intercepts. And in this case, one can derive an explicit solution of the objective function (5.2) with penalty (5.3):

Theorem 3. *Assume a penalized linear model with orthonormal design; that is $\boldsymbol{X}^T\boldsymbol{X} = \mathbb{I}_{(k+1)\times(k+1)}$ where $\boldsymbol{X} \in \mathbb{R}^{(k+1)\times(k+1)}$ denotes the design matrix without an intercept and where $\mathbb{I}$ denotes the identity matrix. Let the ML estimates be ordered $\hat{\beta}_0^{ML} < \ldots < \hat{\beta}_k^{ML}$ and employ penalty (5.3) with a fixed penalty parameter λ, $\lambda \geq 0$. Then for j, $\hat{\beta}_j^{ML} < \bar{\beta}^{ML}$, $\bar{\beta}^{ML} = \frac{1}{k+1}\sum_{j=0}^{k}\hat{\beta}_j^{ML}$, one obtains*

$$\hat{\beta}_j = \min\left\{\bar{\beta}^{ML},\ \max\{\hat{\beta}_l^{ML}, \hat{\beta}_j^{ML}\} + \frac{(\lambda-\lambda_l)I_{(l\geq j)}}{2(l+1)}\right\},$$

where $l = \max_{l=0,\ldots,k}(\lambda_l < \lambda)$, $\lambda_l = \sum_{u=1}^{l} 2u\left|\hat{\beta}_u^{ML} - \hat{\beta}_{u-1}^{ML}\right|$, and with indicator function I. For $\hat{\beta}_j^{ML} \geq \bar{\beta}^{ML}$, one obtains analogously

$$\hat{\beta}_j = \max\left\{\bar{\beta}^{ML},\ \min\{\hat{\beta}_l^{ML}, \hat{\beta}_j^{ML}\} - \frac{(\lambda-\lambda_l)I_{(k-l\geq j)}}{2(l+1)}\right\},$$

with $\lambda_l = \sum_{u=l}^{k-1} 2(k-u)\left|\hat{\beta}_{u+1}^{ML} - \hat{\beta}_u^{ML}\right|$ and l as before.

The proof of Theorem 3 is given in Appendix C. The structure of the explicit estimate reveals that the coefficients of the outer categories are always merged first. There is no shrinkage for coefficients that are not yet fused with one of the outer categories – no matter how close the corresponding ML estimates are. For the minimal penalty parameter that causes the fusion of all coefficients, the estimate of all coefficients is equal to $\bar{\beta}^{ML}$. For a fixed value of λ, the mean of the penalized estimate equals $\bar{\beta}^{ML}$ in the assumed setting.

The left panel of Figure 5.1 shows the (exact) coefficient path of an exemplary model with $k = 7$. One can see that a coefficient β_r is not fused with any other β_s, $r \neq s$, unless β_r is fused with one of the outer coefficients β_0, β_k. The right panel of Figure 5.1 shows the same situation, but the coefficient path is obtained with an L_0 norm instead of the L_1 norm in penalty (5.3). Categories with similar effects are fused – no matter which position they take in the order of the ML estimates. This is the main motivation to consider L_0 penalties as an alternative model selection strategy for categorical effects.

For a *nominal* effect (in the initially assumed setting with intercept β_{int} and where $\boldsymbol{x}$ is dummy coded with the reference category $\beta_0 = 0$), pairwise differences of coefficients are appropriate:

$$P(\boldsymbol{\beta}) = \sum_{r>s\geq 0} |\beta_r - \beta_s|. \tag{5.4}$$

Assume a fixed value of the penalty parameter λ and let $\beta_{(0)}, \ldots, \beta_{(k)}$ denote the arbitrary ordering of the solution (including $\beta_0 = 0$, without β_{int}). Then, a short transformation (see Appendix C) shows that $\sum_{r>s\geq 0}|\beta_{(r)} - \beta_{(s)}| = \sum_{r=1}^{k} w_{(r)}|\beta_{(r)} - \beta_{(r-1)}|$, where $w_{(r)} = r(k-r+1)$. For the "outer" differences $r \in \{1, k\}$, $w_{(r)} = k$; for medium values of r, the weights $w_{(r)}$ are higher. That is, penalty (5.4) can be represented as a weighted version of

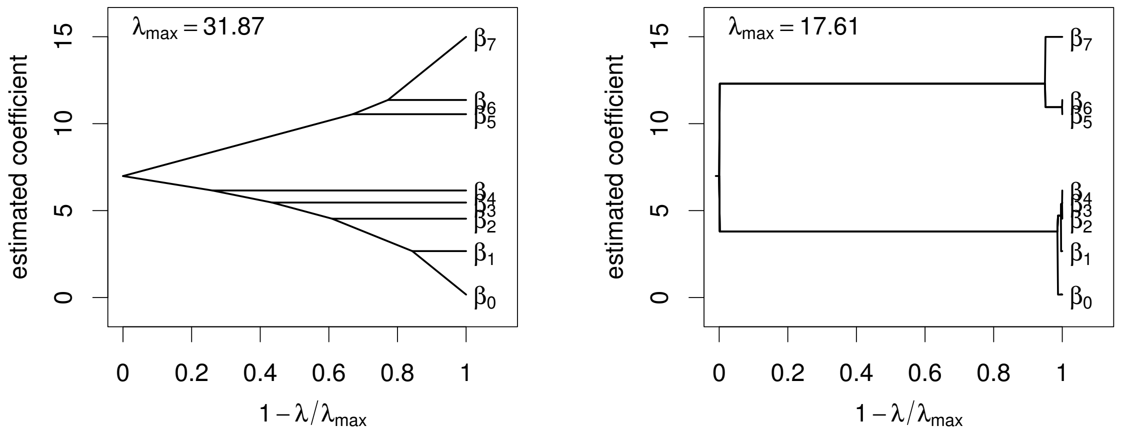


Figure 5.1.: Coefficient paths for an orthonormal linear model with one categorical covariate ($k = 7$), dummy coded without an intercept. In the left panel, penalty (5.3) is applied; in the right panel, the L_1 norm in penalty (5.3) is replaced by the L_0 norm.

penalty (5.3). It illustrates that the issues for nominal covariates are essentially the same as for ordered covariates. Similar to Theorem 3, one can show that the slopes of the coefficient path depend only on the order of the corresponding ML estimate – even though not only the "outer" coefficients are fused.

Efficiency of L_1-penalized estimates can be improved by using adaptive weights (Zou, 2006) that weigh each penalty term by its inverse ML estimate. This results in heavy weights on penalty terms with small ML estimates and in small weights on penalty terms with large ML estimates. When the adaptive weights of Zou (2006) are combined with fusion-type penalties, there is an incentive to fuse categories that have close ML estimates and one obtains asymptotically normal and consistent results (see, for example, Gertheiss and Tutz, 2010, Oelker et al., 2014). However, adaptive weighting requires ML estimates; its quality depends on the quality of the ML estimates.

5.3. L_0-Type Fusion Penalties

In what follows, the fusion of categories is enforced by penalizing differences of coefficients as in the approaches discussed previously, but to overcome the drawbacks of Lasso-type penalties, the L_0 norm is employed. For an *ordered* effect, we propose

$$P_{ord}(\boldsymbol{\beta}) = \sum_{r=1}^{k} \left\|\beta_r - \beta_{r-1}\right\|_0 , \tag{5.5}$$

where $\left\|\xi\right\|_0 = I_{\xi \neq 0}$ and where I denotes the indicator function. In contrast to Lasso-type penalties, it does not matter whether a difference is small or huge; the penalty is reduced

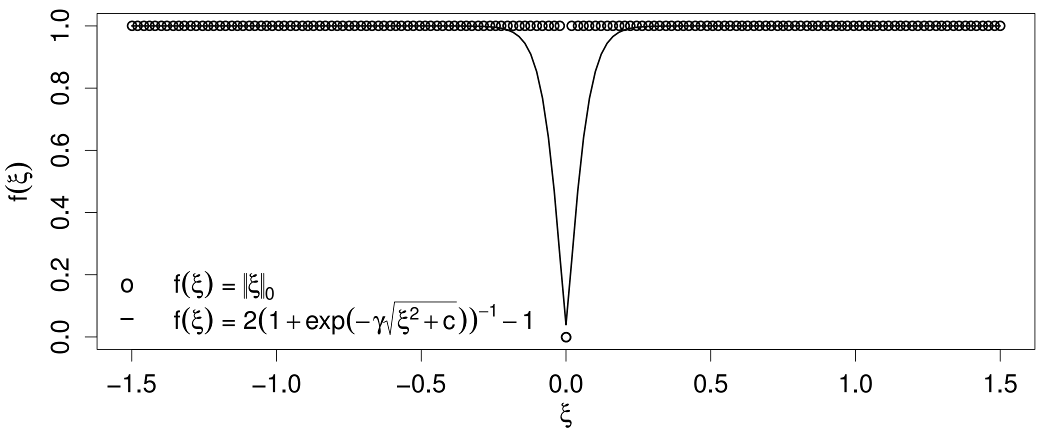


Figure 5.2.: Graphical illustration of the approximation of the L_0 norm. $\gamma = 25$, $c = 10^{-5}$.

only if one of the differences equals zero. As a consequence, when two different values of λ, for example $\lambda_1 > \lambda_2$, yield solutions with the same set of zero and non-zero differences, it holds that $P_{ord}(\hat{\boldsymbol{\beta}}_{\lambda_1}) = P_{ord}(\hat{\boldsymbol{\beta}}_{\lambda_2})$. The set of zero/non-zero differences changes for specific thresholds.

When the effect $\boldsymbol{x}$ is *nominal*, an appropriate coefficient profile does not only relate to the coefficients of adjacent categories, but to the comparison of all coefficients. The penalty considers all pairwise differences of coefficients:

$$P_{nom}(\boldsymbol{\beta}) = \sum_{r>s\geq 0} \|\beta_r - \beta_s\|_0 \,. \tag{5.6}$$

Penalty (5.6) is more complex. With k levels, there are $k(k+1)/2$ pairwise differences, but apart from that, the effect of the penalty is the same as for ordered covariates.

Note that for large values of the penalty parameter λ, the coefficients $\beta_1, \ldots, \beta_k$ are set to zero as the differences in penalties (5.5) and (5.6) include the difference to the reference category $\beta_0 = 0$. As it will be seen in Section 5.5, it can be useful to weight the penalty terms. To this end, we introduce general weights w_r, $w_{r,s}$ respectively, and use the modified penalties

$$P_{ord}(\boldsymbol{\beta}) = \sum_{r=1}^{k} w_r \|\beta_r - \beta_{r-1}\|_0 \quad \text{and} \quad P_{nom}(\boldsymbol{\beta}) = \sum_{r>s\geq 0} w_{r,s} \|\beta_r - \beta_s\|_0 \,. \tag{5.7}$$

To enhance the performance, we will, for example, combine the L_0 approach with adaptive weights as employed for the L_1 penalties in Section 5.4.1. When analyzing the unemployment rates in Germany in Section 5.5, the weights allow one to account for the spatial structure of the federal states. Therefore, we define the weights $w_{r,s}$ as an indicator for a common border of two states – such that we obtain a coefficient profile that is consistent with geography.

5.3.1. Computational Issues

In the literature, there is a wide range of strategies to handle optimization problems that contain L_0 norms. For example, in order to represent signals sparsely, the objective $\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_0$ subject to $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$ has to be optimized. With some assumptions on $\boldsymbol{X}$ and assuming that there is a sufficiently sparse representation of $\boldsymbol{y}$, Donoho and Elad (2003) find this representation by solving a convex optimization problem instead. Wipf and Rao (2005) derive a method based on sparse Bayesian learning including local optimality conditions to solve the same problem. In the framework of wavelets, the L_0 norm acts as penalty. The problem $\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_0$ is, for example, solved by penalty decomposition methods that are based on rank optimization procedures (see Lu and Zhang, 2010; Lu and Zhang, 2013). Rippe et al. (2012) and Johnson (2013) minimize $\sum_{i=1}^n (y_i - \beta_i)^2 + \lambda \sum_{i=2}^n \|\beta_i - \beta_{i-1}\|_0$ to smooth segmented observations $y_1, \ldots, y_n$. While Johnson (2013) proposes a dynamic programming algorithm, Rippe et al. (2012) solve the problem iteratively employing a weighted Ridge penalty.

In order to minimize the objective function (5.2), we propose to approximate the L_0 norm by a modified logistic function and to derive a quasi Newton method for the approximated objective function. As proposed in Chapter 2, the L_0 norm is approximated by

$$\|\xi\|_0 \approx \frac{2}{1 + \exp(-\gamma\sqrt{\xi^2 + c})} - 1, \tag{5.8}$$

where γ is a relatively large scalar and where c is a small, positive constant. Figure 5.2 gives some illustration: The circles denote the L_0 norm for a scalar argument ξ. The continuous line denotes the approximation of the L_0 norm. For $\gamma \to \infty$ and $c \to 0$, the approximation approaches the L_0 norm.

To obtain a penalized iteratively re-weighted least squares (PIRLS) algorithm, in addition to approximation (5.8), we employ a local quadratic approximation if $\hat{\boldsymbol{\beta}}_{(k+1)}$ is close to $\hat{\boldsymbol{\beta}}_{(k)}$ as proposed by Fan and Li (2001) and as described in Chapter 2, by Oelker and Tutz (2013) respectively.

Concerning the tuning, the constant $c > 0$ guarantees differentiability, γ determines the steepness of the logistic function. Both parameters have to be determined subject to the scale of the (coded) covariate $\boldsymbol{x}$. However, in penalized regression models, the covariates are usually scaled and/or standardized ($\mathbb{E}(\boldsymbol{x}) = 0$ and $\mathbb{V}(\boldsymbol{x}) = 1$). In such settings, $c = 10^{-5}$ works quite well in our experience. When γ is sufficiently large, the coefficients paths look like step functions. The steps occur when the coefficients are merged and as the shift of the estimates is relatively large compared to the change of λ. As long as the approximation is close enough to the L_0 norm, the concrete choice of γ has no major impact on the result's quality. However, for different tuning parameters, the scale of λ changes. If γ is too large, there may be convergence problems.

The structure of the objective function is not trivial. As the penalty is not convex, there is no guarantee that the proposed algorithm finds the global minimum of the objective function. However, the results are very plausible. Given that the penalty parameter λ is in a realistic range, the results for different initial values do not differ essentially in a majority of cases. We recommend $\hat{\boldsymbol{\beta}}_{(0)} = \mathbf{0}^T$ or to combine the default approach of the `R` function `glm()` (for example $\boldsymbol{\mu}_{(0)} = \boldsymbol{y}$ for Gaussian responses for the loss function) with the initial value $\hat{\boldsymbol{\beta}}_{(0)} = \mathbf{1}^T$ for the approximation of the penalty (referred to as "default set of initial values"). Furthermore, the results for different initial values should be checked. Comparisons with a simulated annealing algorithm (Xiang et al., 2013) that is appropriate for complex optimization problems, show that the deviations of the PIRLS algorithm from the simulated annealing are small for the relevant range of λ. The L_0 approach of Rippe et al. (2012) for signals works with a different approximation but also with a PIRLS algorithm and obtains similar results. Fan and Li (2001) propose to approximate the SCAD penalty that has a similar curvature, by a PIRLS algorithm; comparisons with the exact estimate in an orthonormal setting approve this procedure.

5.3.2. The General Case with Multiple Covariates

So far, we have assumed that there is only one predictive factor $\boldsymbol{x}$. Of course, this is not the standard case and, in what follows, we assume that there are p nominal and/or ordered predictive factors $\boldsymbol{x}_j$ with $k_j + 1$ levels each. The design matrix is still denoted by $\boldsymbol{X}$, but now $\boldsymbol{X} \in \mathbb{R}^{n \times q}$, where $q = 1 + \sum_{j=1}^{p} k_j$; $\boldsymbol{X}$ contains p dummy coded effects and an intercept. The according penalty is defined as

$$P(\boldsymbol{\beta}) = \sum_{j=1}^{p} J_j(\boldsymbol{\beta}_j),$$

where J_j equals penalty (5.5) for ordered factors and penalty (5.6) for nominal factors. The parameter $\boldsymbol{\beta}_j = (\beta_{j1}, \dots, \beta_{jk_j})^T$ denotes the vector of coefficients linked to the j-th covariate. The computational issues are not affected by this generalization.

However, the tuning should be adjusted as the penalty terms of several factors with different numbers of levels and measured on different scales (nominal/ordered) should be comparable. For example, Bondell and Reich (2009) argue that there is a bijective relation between the standardization of the data and weighting the penalty terms if one penalty term relates to one covariate and if the penalty is a norm; for example, in case of the Lasso for p continuous covariates. Gertheiss and Tutz (2010) transfer this idea to penalties that contain pairwise differences related to nominal covariates; they propose to weigh each difference by $k_j^{-1}\sqrt{(n_j^{(l)} + n_j^{(m)})/n}$, where $n_j^{(l)}$ denotes the number of observations on level l of covariate j. For ordered covariates, $\sqrt{(n_j^{(l)} + n_j^{(m)})/n}$ is appropriate. The weights consider the number of observations per level and the number of differences in the penalty. They can be combined with the L_0 penalty easily. Depending on the concrete context, different weighting schemes can be reasonable; alternatively, one could, for example, think of $J_j(\boldsymbol{\beta}_j) = const. \forall j$.

5.3.3. L_0 Penalization and Information Criteria

There is a relation between L_0 penalization and model selection by information criteria like the AIC or the BIC. One minimizes

$$\mathcal{IC}(\boldsymbol{\beta}) = -l_n(\boldsymbol{\beta}) + \lambda \cdot \mathrm{df(model)}, \tag{5.9}$$

where $\lambda_{AIC} = 1$ for the AIC and $\lambda_{BIC} = \log(n)/2$ for the BIC. The degrees of freedom df(model) are the number of influential parameters in the model and therefore equal $\sum_{j=0}^{p} \|\beta_r\|_0$. Let us first consider a model with binary effects only. Then, the proposed L_0 penalty has the form $P_{bin}(\boldsymbol{\beta}) = \sum_{j=1}^{p} \|\beta_j\|_0$. Hence, it holds that

$$P_{bin}(\boldsymbol{\beta}) + 1 = \mathrm{df(model)}.$$

That is, when the penalty parameter of the proposed L_0 approach is fixed to the values λ_{AIC} or λ_{BIC}, the objectives of the L_0 approach and of model selection based on information criteria AIC/BIC coincide. However, the computational approach differs: For model selection based on information criteria, unconstrained models with all possible subsets of coefficients are compared by criterion (5.9). The L_0 approach optimizes the approximated, constrained objective and does it without the subsets.

Selection problems are much more complex if one has p categorical covariates with $k_j + 1$ levels each, because then, in addition to the simple selection of relevant variables, one also wants to investigate which categories of the categorical effect have to be distinguished. In best subset selection with categorical effects, all models that can be built by the fusion of categories must be considered as candidate models. It is well known from cluster theory that the number of such candidate models increases strongly with the number of categories per effect (Jain and Dubes, 1988). For a nominal covariate with 3 coefficients $\beta_1, \beta_2, \beta_3$, this already results in 15 combinations: $\{(), \{\beta_1\}, \{\beta_2\}, \{\beta_3\}, \{\beta_1, \beta_2\}, \{\beta_1, \beta_3\}, \{\beta_2, \beta_3\}, \{\beta_1 = \beta_2\}, \{\beta_1 = \beta_3\}, \{\beta_2 = \beta_3\}, \{\beta_1 = \beta_2\ , \beta_3\}, \{\beta_1 = \beta_3, \beta_2\}, \{\beta_2 = \beta_3, \beta_1\}, \{\beta_1, \beta_2, \beta_3\}, \{\beta_1 = \beta_2 = \beta_3\}\}$. Thus, model selection based on candidate models as it is used by AIC and BIC, is restricted to cases with very few categories in the effect(s).

In this general case, the degrees of freedom are defined as the number of coefficients in a model that are different from another and unequal zero. They are given by $\mathrm{df(model)} = 1 + \sum_{j=1}^{p} \sum_{r=1}^{k_j} \|\beta_{jr}\|_0 \prod_{s<r} \|\beta_{jr} - \beta_{js}\|_0$ – which is unequal to the proposed penalties. However, the proposed penalties can be applied to the same situations as best subset selection

for categorical covariates. In contrast to best subset selection, the proposed penalties do not need candidate models and – as we will see later on – they are nevertheless feasible for more complex models. As the penalty parameter λ can be varied, information on the order of the fusions of coefficients is obtained. Hence, the proposed penalty is not only an attractive alternative to Lasso-type penalties, but as well an alternative to model selection based on information criteria.

5.4. Illustration and Numerical Experiments

In this section, we investigate some aspects of the proposed approach by numerical experiments. We start with an illustrative example followed by some experiments on the estimation accuracy and on the clustering/selection performance when the penalty parameter is chosen according to cross-validation criteria.

5.4.1. Illustrative Example

Consider a linear model with the two ordered effects $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Both effects have four levels and are drawn from a multinomial distribution with equal probabilities for each level. The response is Gaussian; $\boldsymbol{\beta}^{true} = (\beta_{int}, \beta_{12}, \beta_{13}, \beta_{14}, \beta_{22}, \beta_{23}, \beta_{24})^T = (2, 0.8, 0.6, 0.4, -0.6, -0.6, -0.4)^T$ and $\mathbb{V}(y_i|\boldsymbol{X}) = 1 \ \forall i$. All levels of $\boldsymbol{x}_1$ have a different impact on the response whereas the levels 2 and 3 of $\boldsymbol{x}_2$ are influential but do not need to be distinguished. We generate $n = 100$ observations and consider two models: The proposed L_0 penalty, that is, penalty (5.5) for the two ordered factors, and a penalty with the same differences but with the L_1 norm instead of the L_0 norm. There is one global penalty parameter λ in both models, which is chosen by the generalized cross-validation criterion (GCV) that is, for example, introduced in Section 2.2.4. The GVC criterion is given by $\text{GCV} = n \cdot \text{dev}/(n - \text{df(model)})^2$, where the deviance is defined as $\text{dev}(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = -\varphi(l_n(\boldsymbol{y}, \hat{\boldsymbol{\mu}}, \varphi) - l_n(\boldsymbol{y}, \boldsymbol{y}, \varphi))$, where $l_n(\cdot)$ denotes the log-likelihood; df(model) is estimated by the trace of the generalized hat matrix of the final iteration of the PIRLS algorithm (see Section 2.2.3). Figure 5.3 (top) shows the resulting coefficients paths and the GCV score for the L_0 penalization with tuning parameters $c = 10^{-5}$, $\nu = 0.05$, $\gamma = 60$. At the very right end of the coefficient paths, the ML estimates are displayed. At the very left end, the estimate for the minimal penalty parameter λ that gives maximal penalization is shown. As there are no shrinkage effects, for some values of λ, the estimates are the same. The coefficient path looks like a horizontal tree. The GCV score in the right panel is a step function that jumps when the estimate changes. This happens because the GVC criterion does not depend on the penalty parameter λ. For identical estimates $\hat{\beta}_{\lambda_1} = \hat{\beta}_{\lambda_2}$, the GCV scores are the same. Hence, the function

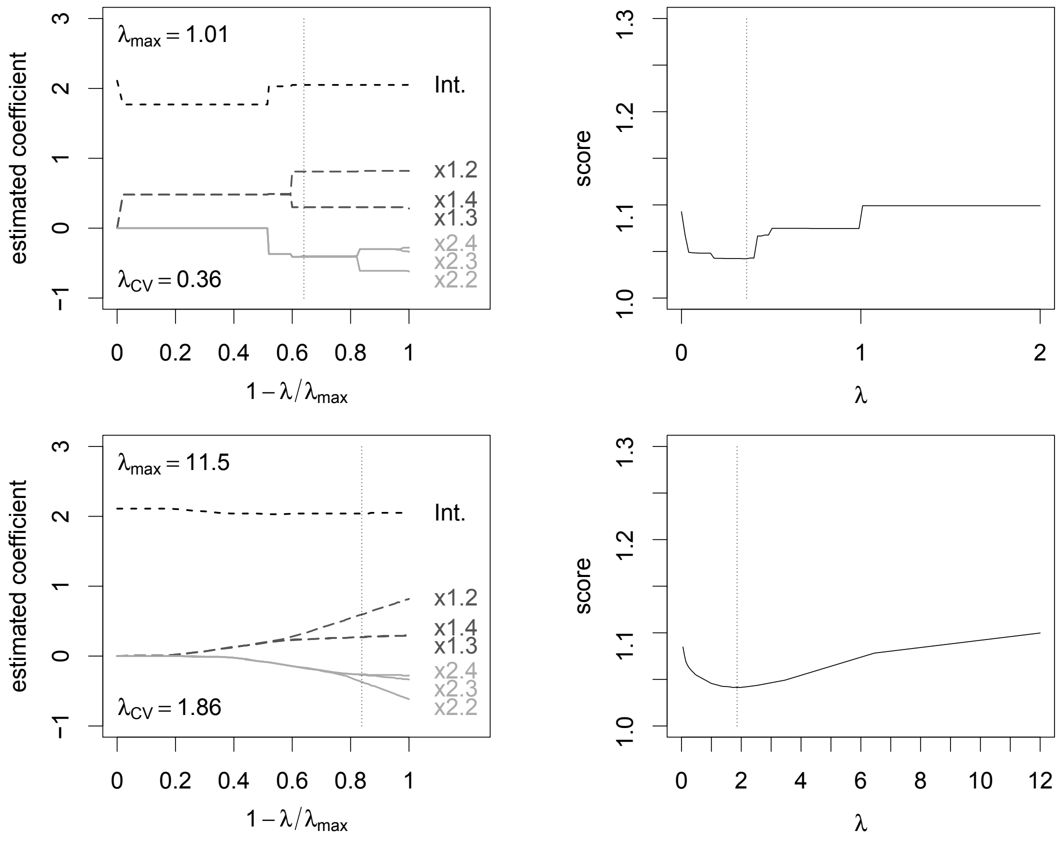


Figure 5.3.: Illustration of L_0 penalization (top) and L_1 penalization (bottom) in a linear model with two ordered effects. The left column shows the resulting coefficient path. The right column shows the corresponding GCV score. The tuning for the approximation of the L_0 norm is $\gamma = 60$, $c = 10^{-5}$ and $\nu = 0.05$. The tuning for the approximation of the L_1 norm is $c = 10^{-5}$. In all panels, the dotted line marks the optimal models.

has no clear minimum. We choose the maximal value of λ with minimal GCV score as λ_{CV}. The optimal model is marked by a dotted line at $\lambda_{CV} = 0.36$. In this model, levels 2 and 3 of covariate $\boldsymbol{x}_2$ have the same impact on the response ($\hat{\beta}_{22} = \hat{\beta}_{23} = -0.41$, $\hat{\beta}_{24} = -0.40$). Levels 3 and 4 of covariate $\boldsymbol{x}_1$ are falsely fused ($\hat{\beta}_{13} = \hat{\beta}_{14} = 0.30$). The estimates of the remaining effects are: $\hat{\beta}_{int} = 2.05$, $\hat{\beta}_{11} = 0.81$. In the optimal model with the same differences but with a L_1 norm in the penalty, two coefficients are falsely fused ($\hat{\beta}^{L1} = (2.04, 0.60, 0.27, 0.27, -0.37, -0.27, -0.26)^T$). Figure 5.3 (bottom) shows the according coefficient paths and the GCV score. In contrast to the L_0 penalty, the path is characterized by steady shrinkage effects; the GCV score is a continuous function with a clear minimum. For the L_1 penalty, the shrinkage effect is slightly bigger as for the L_0 penalty: The sum of squared errors are $\widehat{\text{SSE}}_{L_1} = \sum_{i=1}^{7}(\hat{\beta}_i^{L1} - \beta_i^{true})^2 = 0.3488 \; > \; 0.1748 = \widehat{\text{SSE}}_{L_0}$.

5.4.2. The Choice of the Penalty Parameter λ

As for every penalized approach, the choice of the penalty parameter λ is a crucial issue for L_0 penalization. In the illustrative example, we employ a GCV criterion that requires an estimate of df(model) and that gives concurrent jumps in the coefficient paths and in the GCV score. An alternative, frequently used approach to choose the penalty parameter is K-fold cross-validation with the predictive deviance as loss criterion. K-fold cross-validation relies on models estimated on different training/test data sets for different values of λ. As we approximate the L_0 norm which is not continuous, the estimate can change abruptly even if the tuning varies only slightly. The values of λ at which the estimate changes, will not be the same for all training data set. Thus, depending on the chosen folds for K-fold cross-validation, the overall cross-validation score can be quite wiggly. Therefore, we compare the performance of the GCV criterion and of 5-fold cross-validation in Section 5.4.3.

5.4.3. Performance

To evaluate the overall performance of the penalties, we consider the estimation accuracy, the prediction accuracy and the error rates of the selection and clustering process. The estimation accuracy is assessed by the squared errors in terms of coefficients: $\widehat{\text{SSE}} = \frac{1}{q}\sum_{j=1}^{q}(\beta_j^{true} - \hat{\beta}_j)^2$, where $\boldsymbol{\beta}^{true}$ denotes the vector of true coefficients and $\hat{\boldsymbol{\beta}}$ the estimate of the current simulation run. The median of all squared errors is the robust estimate for the mean squared error (MSE) of a method. The prediction accuracy is assessed by the predictive deviances. To judge the model selection process, we consider the selection and the clustering of coefficients separately; the selection of coefficients refers to the coefficients ($\beta_{jl} = 0$) whereas the clustering process refers to the differences of coefficients ($\beta_{jl} = \beta_{jm}$). We distinguish between false positive rates (fraction of truly zero coefficients that are set to non-zero, FP) and false negative rates (fraction of truly non-zero coefficients that are set to zero, FN). We focus on four settings. A setting similar to the illustrative example of Section 5.4.1 is considered in more detail, it is referred to as *G3*. In addition, a setting with Gaussian response and 50 nominal covariates is investigated (*G50*). Settings with Poisson distributed and with binomial distributed response are analyzed (*P8*, *B8*). For each setting, 100 replications are considered; for each replication, we compute the ML estimate, the estimate obtained with the L_1 penalty, the estimate obtained with the adaptively weighted L_1 penalty and the estimate obtained with the proposed L_0 penalty. Moreover, we combine the proposed L_0 penalty with the same adaptive weights as employed for the adaptively weighted L_1 penalty. For all penalized approaches, the penalty parameter is chosen by the GCV criterion and by 5-fold cross-validation with the predictive deviance as loss criterion (CV). In addition, a model selection method for categorical covariates is implemented. The

Setting		ML	L1, CV	L1, GCV	L1, adapt, CV	L1, adapt, GCV	L0, CV	L0, GCV	L0, adapt, CV	L0, adapt, GCV	AIC	BIC
G3	FP$_s$	1.00	0.85	0.80	0.50	0.50	0.60	0.40	0.36	0.35	0.27	0.12
	FN$_s$	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01
	FP$_c$	1.00	0.90	0.88	0.53	0.58	0.61	0.42	0.30	0.34	0.26	0.14
	FN$_c$	0.00	0.02	0.03	0.07	0.06	0.09	0.11	0.15	0.12	0.15	0.22
G50	FP$_s$	1.00	0.74	0.72	0.37	0.47	0.13	0.51	0.23	0.44	-	-
	FN$_s$	0.00	0.00	0.00	0.01	0.01	0.02	0.01	0.02	0.01	-	-
	FP$_c$	1.00	0.77	0.76	0.42	0.52	0.21	0.60	0.28	0.50	-	-
	FN$_c$	0.00	0.00	0.00	0.01	0.01	0.02	0.00	0.02	0.01	-	-
P8	FP$_s$	1.00	0.74	0.69	0.40	0.41	0.17	0.33	0.15	0.31	-	-
	FN$_s$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	-
	FP$_c$	1.00	0.78	0.75	0.42	0.44	0.20	0.39	0.15	0.33	-	-
	FN$_c$	0.00	0.01	0.01	0.02	0.02	0.05	0.02	0.06	0.03	-	-
B8	FP$_s$	1.00	0.74	0.83	0.53	0.69	0.37	0.65	0.35	0.56	-	-
	FN$_s$	0.00	0.07	0.04	0.17	0.09	0.40	0.15	0.40	0.20	-	-
	FP$_c$	1.00	0.55	0.71	0.36	0.54	0.16	0.42	0.15	0.34	-	-
	FN$_c$	0.00	0.18	0.12	0.32	0.20	0.57	0.28	0.54	0.35	-	-

Table 5.1.: Estimates of false positive (FP) and false negative (FN) rates for the selection (s) and clustering (c) performance for all considered settings.

method is based on the information criteria AIC and BIC and compares not only all possible subsets of coefficients, but as well all possibilities to fuse different numbers of levels of a categorical covariate. This method is referred to as AIC, BIC respectively.
For all settings, the tuning parameters $c = 10^{-5}$ and $\nu = 0.05$ are fixed; we employ the default set of initial values. γ is empirically chosen as described in Section 5.3.1 ($\gamma_{G3} = 20$, $\gamma_{G50} = 10$, $\gamma_{P8} = 20$, $\gamma_{B8} = 10$).

Gaussian Responses In setting *G3*, there are 3 nominal covariates with four levels each; $\boldsymbol{\beta}^{true} = (\beta_{int}, \beta_1^T, \beta_2^T, \beta_3^T)^T = (1,\ (0, -1.5, -1.5),\ (0, 0, 2),\ (-3, -3.5, 4))^T$. In each replication, $n = 50$ observations are generated. The upper left panel of Figure 5.4 shows the boxplots of the squared errors. Apart from some outliers, the estimation accuracy of all considered approaches is approximately the same. This is typical: In standard situations, (adaptive) L_1 and L_0 penalization do not show substantial differences. However, as seen in Table 5.1, the L_0 approach produces more parsimonious and interpretable models. The methods based on information criteria are characterized by the highest FN rates. Comparing the L_1 penalization with and without adaptive weights, the adaptive weighting improves the FP rates substantially. Comparing the adaptively weighted L_1 and the adaptively weighted L_0 penalty, the clustering performance is substantially enhanced by the L_0

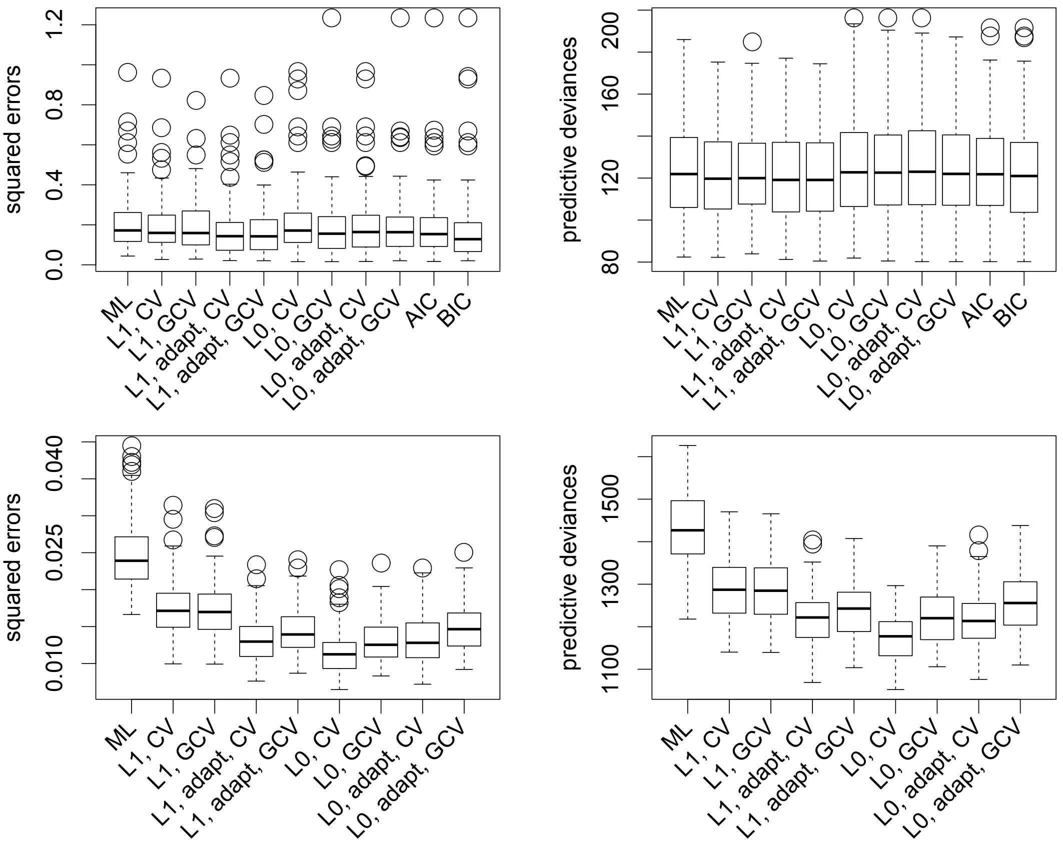


Figure 5.4.: Results for settings *G3* (top) and *G50* (bottom): Boxplots of squared errors of coefficients (left panel) and of the predictive deviances (right panel).

penalty. Again, this is typical: With the L_0 penalty, the false positive rates are lower while it can happen that the FN rates increase slightly in comparison with L_1 penalization.

In setting *G50*, there are 50 nominal covariates with four levels each. $\boldsymbol{\beta}^{true}$ is a vector of length 151, it contains 72 non-influential coefficients and 40 truly different effects. $n = 500$. In contrast to setting *G3*, for this and the following settings, model selection based on information criteria is not feasible anymore on a default computer. The lower panel of Figure 5.4 depicts the squared errors of setting *G50*. It stands out that the L_0 penalized models perform slightly better than the pure L_1 penalized approaches. Regarding the FP/FN rates in Table 5.1, it is even more obvious that the proposed L_0 approach generates more parsimonious models. Overall, the approach "L_0, CV" performs the best.

Poisson Distributed Response In setting *P8*, there are four influential nominal covariates; $\boldsymbol{\beta}^{true} = (\beta_{int}, \beta_1^T, \beta_2^T, \beta_3^T)^T = (2,\ (0, -1.2, -1.2),\ (1.4, 1.4, 0),\ (0.4, 0.6, 0.8), (-0.7, -1, -1.3))^T$. We assume four more non-influential, nominal covariates which are to be detected. For an observation i, the assumed predictor is $\eta_i^{model} = \beta_{int} + \sum_{j=1}^{8} x_{ij}^T \beta_j$;

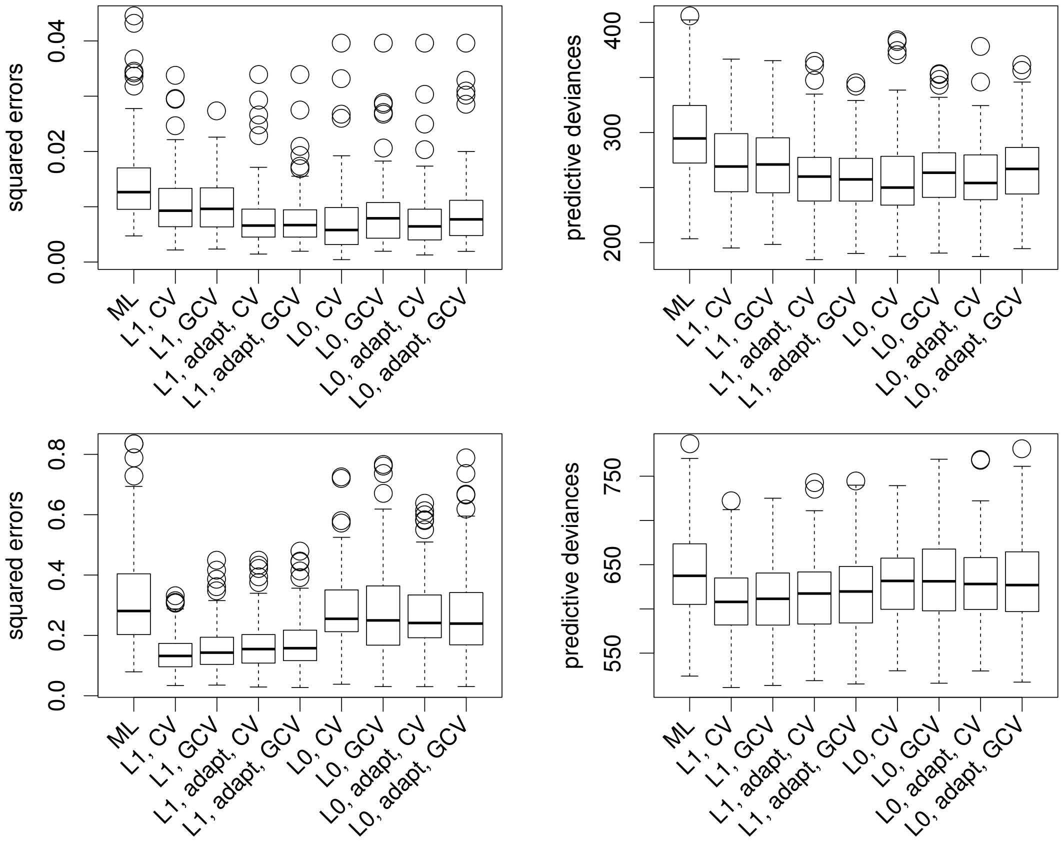


Figure 5.5.: Results for settings *P8* (top) and *B8* (bottom): Boxplots of the squared errors of coefficients (left panel); boxplots of predictive deviances (right panel).

$n = 100$. In Figure 5.5, the squared errors and the predictive deviances of the penalized methods are distinctively smaller than those of the ML estimates. In Table 5.1, the L_0 approach reduces the false positive rates even more than the adaptively weighted L_1 penalty. However, for the L_0 approach, the CV performs better than the GCV criterion. A possible explanation is that df(model) is not estimated as good as before. In the optimal model obtained by L_0 penalization and GCV, df(model) is estimated adequately in only eight of 100 cases (was 52 in setting *G3*).

Binomial Response In setting *B8*, there are four influential and four non-influential ordered covariates. In contrast to the previous settings, the distribution of the categorical covariates is not balanced; for the $n = 400$ observations, the covariates are drawn from a multinomial distribution with sampled probabilities between 0.12 and 0.44. In this setting, it can happen that the unpenalized estimate is quite extreme. As the adaptive weights depend on the quality of the ML estimate, they rely on estimates with a slight Ridge penalty for all coefficients (in the PIRLS algorithm $\boldsymbol{A}_\lambda$ is replaced by $\boldsymbol{A}_\lambda^{ridge} = \text{diag}(0.2)$). As seen in Figure 5.5, the estimation accuracy of the approximated L_0 penalty is slightly worse

Abbreviation	Federal State	2005	2006	2007	2008	2009	2010
BB	Brandenburg	18.2	17.0	14.9	13.0	12.3	11.1
BE	Berlin	19.0	17.5	15.5	13.9	14.1	13.6
BW	Baden-Würtenberg	7.0	6.3	4.9	4.1	5.1	4.9
BY	Bayern	7.8	6.8	5.3	4.2	4.8	4.5
HB	Hansestadt Bremen	16.8	14.9	12.7	11.4	11.8	12.0
HE	Hessen	9.7	9.2	7.6	6.6	6.8	6.4
HH	Hansestadt Hamburg	11.3	11.0	9.2	8.1	8.6	8.2
MV	Mecklenburg-Vorpommern	20.3	19.0	16.5	14.1	13.6	12.7
NI	Niedersachsen	11.6	10.5	8.9	7.7	7.8	7.5
NRW	Nordrhein-Westfalen	12.0	11.4	9.5	8.5	8.9	8.7
RP	Rheinland-Pfalz	8.8	8.0	6.5	5.6	6.1	5.7
SA	Sachsen	18.3	17.0	14.7	12.8	12.9	11.9
SH	Schleswig-Holstein	11.6	10.0	8.4	7.6	7.8	7.5
SL	Saarland	10.7	9.9	8.4	7.3	7.7	7.5
ST	Sachsen-Anhalt	20.2	18.3	16.0	14.0	13.6	12.5
TH	Thüringen	17.1	15.6	13.2	11.3	11.4	9.8

Table 5.2.: Unemployment rates for the federal states of Germany in 2005 to 2010.

than that of the Lasso penalties. Concerning the selection and clustering performance in Table 5.1, it stands out that the FN rates are quite high when the L_0 penalty is combined with the CV criterion. For these approaches, the optimal values of λ are relatively large, too. Apparently, the sample size of the training data sets is too small for differentiated estimates. Hence, in settings like *B8*, the CV criterion is not recommended for L_0-type penalties. In contrast, with the GCV criterion, the clustering and the selection performance of the L_0/GCV penalized models is better than that of the corresponding L_1 penalized approaches.

Remark In general, L_0 penalized models seem to be sparser; the L_1 penalized models tend to have smaller FN rates. Even though there is less shrinkage in the coefficients paths obtained with L_0 penalization, in general, the MSE of the L_0 approach is not smaller than the MSE of the L_1 approach as the estimates obtained with the L_0 approach are more sensitive to variations in the data. In standard situations, the adaptively weighted L_1 penalty and the L_0 approach perform comparably in terms of the estimation accuracy. In terms of variable selection, the L_0 approach has a higher incentive to cluster categories and reaches smaller FP rates while slightly enlarged FN rates are possible. Combining the L_0 approach with adaptive weights enhances the clustering and variable selection performance distinctly. With the L_0 penalization, we obtain stable results in settings where the computation of all possible subsets of coefficients which is needed for model selection based on information criteria is not possible or efficient.

5.5. Unemployment Rates in Germany

We analyze the unemployment rates of the federal states of Germany in the years 2005 to 2010 (Weise et al., 2011). The data is given in Table 5.2. For each of the 16 federal states, there are six annual unemployment rates observed: $(\text{state}_{it}, \text{rate}_{it})$, $i = 1, \ldots, 16$, $t = 2005, \ldots, 2010$. The aim is to find states with the same trends in the unemployment rates while accounting for the heterogeneity amongst the 16 units. Random effects models are the default approach to such data, see, for example, Molenberghs and Verbeke (2005). If one wants to model the unemployment rates by a random effects model, a potential predictor is $\eta_{it} = \beta_{int} + b_{int,i} + \beta_1 \cdot \text{time}$; where $b_{int,i}$, $i = 1, \ldots, 16$, are random effects for which a distribution is assumed, typically a normal distribution with variance σ_b^2: $b_{int,i} \sim N(0, \sigma_b^2)$. Clustering federal states with similar effects relates to identical random effects and requires sophisticated distributional assumptions; for example, a mixture distribution of Gaussian components. This in turn requires elaborate estimation theory, see, for example, Heinzl (2013). Moreover, the data is positively skewed. There are high unemployment rates, but they occur rarely. The response seems to be rather gamma than Gaussian distributed. Hence, we assume a group-specific model with gamma distributed response and a logarithmic link function. The predictor contains one intercept per federal state and a global temporal trend:

$$\eta_{it} = \beta_{int,i} + \beta_1 \cdot \text{time}, \quad \text{with} \quad i = 1, \ldots, 16. \tag{5.10}$$

To cluster the federal states, in a first model, the subject-specific intercepts are penalized by penalty (5.6), whereat differences to the reference category are omitted as there is none. That is, all pairwise differences of intercepts are penalized by the L_0 norm:

$$\lambda \cdot P(\boldsymbol{\beta}) = \lambda \cdot \sum_{r>s>0} \|\beta_{int,r} - \beta_{int,s}\|_0 \,. \tag{5.11}$$

In a second model with the same predictor, the spatial structure of the federal states is considered. Weights $w_{r,s}$ are defined as indicators for states with a common border ($w_{r,s} = 1$ if neighbored, $w_{r,s} = 0$ else). We will refer to this model as the "spatial" model. For both models, the tuning of the algorithm is similar: $\gamma = 36$, $\gamma^{spatial} = 26$, $\nu = 0.5$. The penalty parameter λ is chosen by the generalized cross-validation criterion of O'Sullivan et al. (1986). It yields $\lambda_{CV} = 0.14$ and $\lambda_{CV}^{spatial} = 0.05$. Figure 5.6 shows the coefficient paths of the subject specific intercepts for both models. The left panel relates to the first model with penalty (5.11), the right one to the spatial model. In both models, there are basically two clusters of federal states. The upper cluster contains the states of the former German Democratic Republic (GDR) including Berlin plus the city state Bremen. Interestingly, with the spatial weights, the city state Bremen switches the cluster for a relatively large value of λ. Figure 5.7 illustrates the resulting clusters for

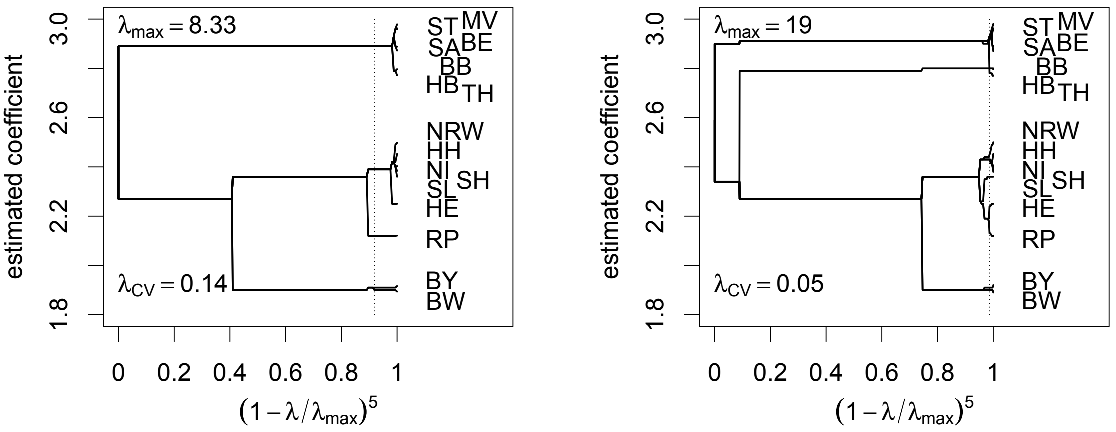


Figure 5.6.: Unemployment rates in Germany – coefficients paths for L_0 penalization considering all pairwise differences (left panel) and differences of coefficients related to neighbored federal states only (right panel).

the optimal choice of λ in a map of Germany. The darker the coloring, the bigger is the subject-specific intercept and the higher is the unemployment rate over the time. In the left panel, the ML estimates are illustrated. Even though all estimates differ, the pattern of the former GDR in the north-east is clearly seen. In the middle, the subject specific intercepts are clustered by the pairwise penalty (5.11). Here, the states of the former GDR plus Berlin and Bremen form one cluster with the biggest impact on the response ($\hat{\beta}_{int,i} = 2.89$). The effects of the southern states Baden-Würtenberg and Bayern are the lowest ($\hat{\beta}_{int,BW} = 1.90$, $\hat{\beta}_{int,BW} = 1.91$). The remaining states except for Rheinland-Pfalz ($\hat{\beta}_{int,RP} = 2.12$) are clustered; the according intercept is $\hat{\beta}_{int,i} = 2.39$. The right panel of Figure 5.7 illustrates the results of the spatial model. The results resemble the middle panel; however, the picture is more differentiated. The states of the former GDR form one cluster ($\hat{\beta}^{spatial}_{int,i} = 2.93$), but there are slightly different estimates for the states Thüringen and Bremen ($\hat{\beta}^{spatial}_{int,TH} = 2.78$, $\hat{\beta}^{spatial}_{int,HB} = 2.80$). The effects of Baden-Würtenberg and Bayern are the same as before, but in the west, only Hamburg, Niedersachen and Schleswig-Holstein are clustered ($\hat{\beta}^{spatial}_{int,i} = 2.43$). The other states have individual intercepts ($\hat{\beta}^{spatial}_{int,RP} = 2.13$, $\hat{\beta}^{spatial}_{int,HE} = 2.24$, $\hat{\beta}^{spatial}_{int,SL} = 2.36$, $\hat{\beta}^{spatial}_{int,NRW} = 2.45$). The estimates for the global temporal trend are approximately the same in all models: $\hat{\beta}^{ML}_t = -0.0875$, $\hat{\beta}_t = -0.09$, $\hat{\beta}^{spatial}_t = -0.09$. In the considered time period, the unemployment rates decreased in all states. Interestingly, Heinzl (2013) obtains similar clusters for the same data by fitting a linear mixed model with Dirichlet process mixtures using the EM algorithm. However, the computational burden for such models is higher.

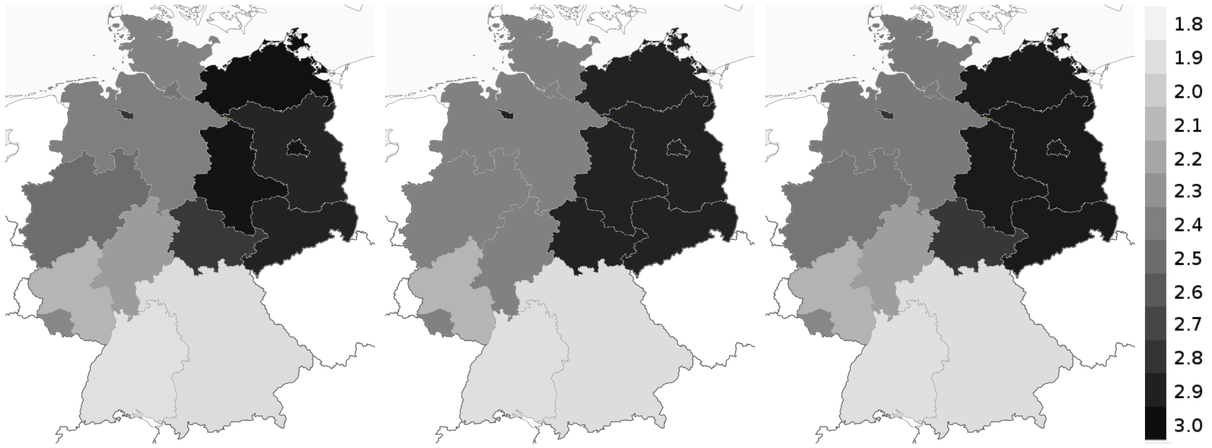


Figure 5.7.: Unemployment rates in Germany – visualization of the effects of the federal states on the unemployment rates. In the very left panel, the ML estimates are shown; in the middle, all pairwise differences of coefficients are penalized by an L_0 penalty; in the very right panel, only the differences of coefficients of neighbored states are penalized. The maps are based on a figure of Wikipedia User NordNordWest (2008); they are manipulated with the GNU Image Manipulation Program (GIMP Team, 2012) and with the `R` package `EBImage` (Pau et al., 2014).

5.6. Remarks

In Chapter 5, we propose L_0 penalization for categorical effects in GLMs. The penalty works on differences of coefficients and accounts for the different amount of information contained in nominal and ordered factors. Unlike Rippe et al. (2012), we provide a classical regression framework for L_0 penalization. Computational issues are met by a local quadratic approximation which can be traced back to Fan and Li (2001). The approximation allows one to derive a PIRLS algorithm; that is, all features of Fisher scoring algorithms are sustained. The coefficient paths are obtained easily.

Applying L_0 penalization to plain coefficients with a fixed penlty parameter has a close relation to best subset selection. As the L_0 approach allows for more flexible terms such as difference in the penalty, and as it works for more complex models, L_0 penalization is an attractive alternative to model selection based on information criteria. In an illustrative example and several numerical experiments, the proposed method is competitive. However, it requires carefully tailored tuning. The range of applications for L_0 penalization is huge. The application to subject-specific intercepts is convincing. The computational framework allows one to combine L_0 penalization easily with other types of (smooth) covariates.

6. Semiparametric Mode Regression

6.1. Introduction

Recent years have seen a tremendous increase in interest related to regression beyond the mean of the conditional distribution of a response given covariates. The most prominent examples are quantile regression and the special case of median regression. They are particularly attractive alternatives to mean regression due to two reasons: Due to their inherent robustness with respect to outliers, and due to the general information they provide concerning distributional features such as heteroscedasticity or skewness. Surprisingly, regression models for the conditional mode of the response distribution given covariates have received far less attention. This may partially be explained by the inherent difficulty to determine an estimate for the mode based on samples from a continuous distribution, where in theory each sampled value should appear only once almost surely, and therefore, there will be multiple "empirical modes". Still, estimating conditional modes is of high interest as

- the mode is by far the visually most prominent feature of a density as compared to the mean and the median,
- the mode is extremely robust with respect to outliers,
- the mode provides a location measure that is easily communicated to practitioners such that mode regression will be of high interest in applied regression situations,
- there may be situations where the dependence of the mode on covariates may be quite different from the dependence of the median and/or the mean,
- mode regression allows to deal with truncated dependent variables. It can still be estimated and interpreted as long as the modal part of the distribution is not truncated. This can, for example, be relevant in applications on income where quite often the upper part of the response distribution is truncated due to non-participation of the high income part of a society.

This chapter is an extended version of Oelker, Sobotka, and Kneib (2014) with contributions of Nadja Klein. For more information on the contributions of the authors and on textual matches, see page 5.

Consider the regression specification

$$y = \boldsymbol{x}^T\boldsymbol{\beta} + \varepsilon, \tag{6.1}$$

where y is the response variable of interest, $\boldsymbol{x} \in \mathbb{R}^q$ is a vector of covariates supplemented with regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^q$ and ε is the error term. Unlike in mean regression, we do not assume $\mathbb{E}(\varepsilon) = 0$ which leads to regression effects on the mean of the response variable, but

$$\arg\max_{\xi} f_{\varepsilon|\boldsymbol{x}}(\xi|\boldsymbol{x}) = 0. \tag{6.2}$$

That is, the conditional density of the error terms $f_\varepsilon(\cdot|\boldsymbol{x})$ is assumed to have a global mode at zero. In turn, this implies that the predictor $\boldsymbol{x}^T\boldsymbol{\beta}$ is the conditional mode of the response distribution $f_y(\cdot|\boldsymbol{x})$. The mode regression coefficient is obtained as

$$\boldsymbol{\beta} = \arg\max_{\boldsymbol{b}} f_\varepsilon(y - \boldsymbol{x}^T\boldsymbol{b}|\boldsymbol{x}). \tag{6.3}$$

An equivalent approach is based on the step loss function $\mathcal{L}_\epsilon(\xi) = \mathbb{1}(-\epsilon \leq \xi \leq \epsilon)$, where ϵ is a positive constant that defines a local environment around zero. With this loss function, we obtain

$$\boldsymbol{\beta}_\epsilon = \arg\min_{\boldsymbol{b}} \mathbb{E}\left[\mathcal{L}_\epsilon(y - \boldsymbol{x}^T\boldsymbol{b})|\boldsymbol{x}\right]. \tag{6.4}$$

In the limiting case $\epsilon \to 0$, $\mathcal{L}_\epsilon(\xi)$ approaches

$$L(\xi) = \mathbb{1}(\xi \neq 0), \tag{6.5}$$

and $\boldsymbol{\beta}_\epsilon$ approaches $\boldsymbol{\beta}$ from equation (6.3) (Manski, 1991). Held (2008, page 158) proves the equivalence of the two approaches in more detail. However, based on n independent observations $(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$, from model (6.1) subject to the condition (6.2), an estimate for the mode regression coefficient can not be determined by an empirical analogue to (6.3) unless specific assumptions are made for the error density $f_\varepsilon(\cdot|\boldsymbol{x})$. In contrast, (6.4) is empirically minimized by

$$\hat{\boldsymbol{\beta}}_\epsilon = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left[\mathcal{L}_\epsilon(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})\right],$$

which does not require any other further assumptions than independence of the observations. However, for the limiting case $\epsilon \to 0$, this criterion is not useful for modal regression based on data with continuous error distributions since in general, there will be no unique solution – even if the density of the errors ε_i has a global mode. As a consequence, earlier attempts to mode regression usually either rely on nonparametric kernel methods from which the mode is then derived in a second step, or on different types of approximations (6.5).

Collomb et al. (1987) follow the first of these two routes and show the uniform convergence of the mode determined from a kernel density estimate to the conditional mode function for a certain class of processes. Lee (1989) approaches the estimation of conditional modes by an empirical approximation to the theoretical loss function defining the mode based on a rectangular kernel. Lee (1989) also shows identification and strong consistency of the resulting estimate, but this requires quite strong assumptions on the error distribution, which has either to be symmetric around the mode (in which case median or mean regression would be obvious alternatives to determine the mode) or – if assumed to be asymmetric – all error distributions have to be identical leading to an i.i.d. model. Lee (1993) extends his approach from 1989 by replacing the rectangular kernel with a quadratic kernel. This allows to construct a more efficient estimate, but it also requires stronger assumptions on the error term such as local symmetry around the mode. Yu and Aristodemou (2012) introduce Bayesian mode regression relying on a working likelihood corresponding to either a uniform or a triangular density.

Einbeck and Tutz (2006) again rely on a kernel regression estimate to implicitly derive the mode in a regression model, but they extend the linear regression specification to a semiparametric predictor. This allows for the nonlinear dependence of the conditional mode on the covariate of interest, but the approach is limited to one single predictive variable. A multivariate extension based on a product kernel for the multivariate covariate vector is outlined in Taylor and Einbeck (2011). However, the resulting estimate is hard to interpret beyond two-dimensional covariates since no additivity assumption can be placed on the predictor. Gannoun et al. (2010) follow a different approach by noting that for many distributions there exists a simple parametric relationship between mode, median and mean. As a consequence, once estimates for the mean and the median are available, the conditional mode can be derived based on this parametric relationship. Their approach is motivated by a forecasting problem in financial time series such that no interpretability for the regression effects on the mode is required, which would be difficult to achieve when combining mean and median estimates.

Kemp and Santos Silva (2012) return to the idea of Lee (1989, 1993). They use a modified kernel to approximate the limiting case (6.5) and to derive a consistent, asymptotically normal estimator for linear mode regression models. In this chapter, we build upon Kemp and Santos Silva (2012) and

- provide a differentiable approximation of the limiting case (6.5) that is based on nested intervals such that an iteratively re-weighted least squares (IRLS) algorithm can be used to estimate the mode regression coefficients (Section 6.2),
- show the consistency and the asymptotic normality of the obtained estimator,
- investigate the practical performance of the approach in a simulation study,
- extend the purely linear mode regression model to additive models by combining nonparametric effects of several covariates in one penalized IRLS framework (Section 6.3),

- provide an extended analysis of the evolution of the BMI in England that has been already studied in Kemp and Santos Silva (2010) and where the polynomial specification of the effect of the age is replaced by a nonparametric specification (Section 6.4).
- perform a geoadditive analysis of the rents in the city of Munich combining penalized spline smoothing with spatial effects (Section 6.5).

The main advantage of this Nested Interval Least Squares (NILS) framework is that it allows to easily include extended regression functionality from (generalized) additive models which also rely on IRLS estimation. In fact, we can further exploit this connection by determining the smoothing parameters within the IRLS framework such that the proposed semiparametric mode regression is fully data-driven.

6.2. The Nested Interval Least Squares Approach

As seen in the introduction of this chapter, there are two equivalent approaches to mode regression: maximizing the conditional density $f_\varepsilon(\cdot|\boldsymbol{x})$ and minimizing the expectation of the step loss function $\mathcal{L}_\epsilon(\xi)$ for the limiting case $\epsilon \to 0$. The reasoning behind the latter can be illustrated based on a set of simulated standard normal data: Iteratively reducing the environment $[-\epsilon, \epsilon]$ allows to determine the mode via nested intervals that contain the largest fraction of observations. Stacking these intervals upon each other allows to graphically indicate how reducing the width of the intervals captures the mode of the distribution. For comparison, in Figure 6.1, a kernel density estimate is added.

6.2.1. Construction of the Estimator

Our approach to mode regression follows a similar reasoning: The limiting case $L(\xi)$ is approximated such that it is zero not only for $\xi = 0$ but in a surrounding of $\xi = 0$. The approximation – denoted by $\mathcal{L}(\xi)$ – will replicate the nested interval approach, that is, $\mathcal{L}(\xi)$ will have a very broad minimum in the early iterations and it will be very close to $L(\xi)$ for the final iteration of the proposed algorithm. However, $L(\xi)$ is approximated by a continuously differentiable function. This has two important advantages: (i) The approximation $\mathcal{L}(\xi)$ can be linked to iteratively re-weighted least squares estimation, and (ii) the smooth approximation allows to determine asymptotic properties such as consistency and asymptotic normality.

In detail, we employ the function

$$\mathcal{L}(\xi) = 1 - \exp(c^{\frac{1}{2g}} - ((k\xi)^{2g} + c)^{\frac{1}{2g}}), \tag{6.6}$$

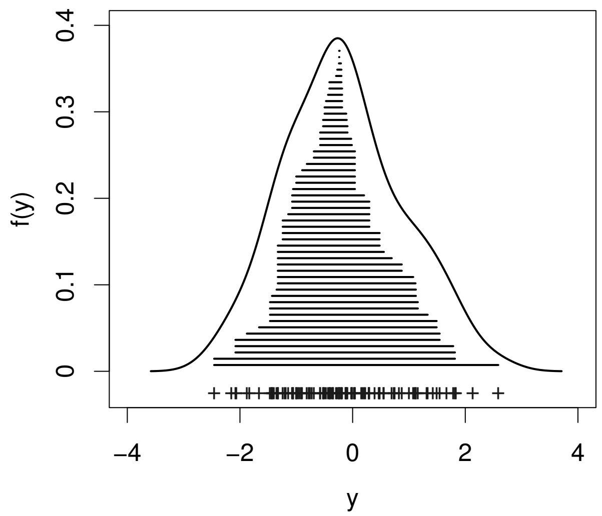


Figure 6.1.: Determining the mode based on nested intervals: Based on 100 realizations from a standard normal distribution ("+"), nested intervals are constructed such that the interval covers the largest possible fraction of data points given a fixed width. Stacking these intervals upon each other allows to graphically indicate how reducing the width of the intervals captures the mode of the distribution. For comparison, a kernel density estimate is added.

depending on the set of tuning parameters $\mathcal{T} = \{g, k, c\}$ with $\lim_{\mathcal{T}\to T} \mathcal{L}(\xi) = L(\xi)$ for some set of limiting values T. The approximation $\mathcal{L}(\xi)$ is constructed as the scaled composition of the two functions $f(\xi)$ and $h(\xi)$. The former is given by $f(\xi) = 1 - \exp(-\xi)$. Let k be a positive number, then $f(k \cdot \xi)$ actually approximates the indicator $L(\xi)$ with the approximation being closer to $L(\xi)$ the larger k is. The latter function is defined as $h(\xi) = (\xi^{2g} + c)^{\frac{1}{2g}}$, where g is as a positive integer and c is a small, positive constant. As illustrated in Figure 6.2, $h(\xi)$ accounts for the broad minimum needed to imitate the nested interval approach. For the limiting value $g = 1$, $h(\xi)$ simply approximates the absolute value function. Due to the constant c, it is a continuously differentiable approximation of the absolute value. Scaling the composition $f(h(k \cdot \xi))$ gives function (6.6). As $\mathcal{L}(\xi)$ is continuously differentiable, an iteratively re-weighted least squares algorithm is derived. The approximated objective

$$\mathcal{M}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathcal{L}(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})$$

is minimized by iterating

$$\hat{\boldsymbol{\beta}}_{(l+1)} = (1-\nu)\hat{\boldsymbol{\beta}}_{(l)} + \nu \boldsymbol{A}_{(l)}^{-1} \boldsymbol{a}_{(l)} \tag{6.7}$$

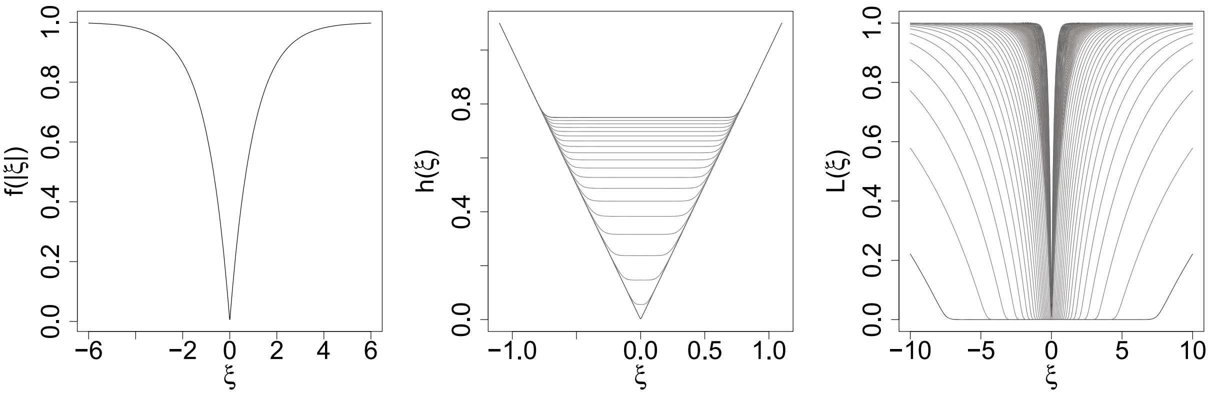


Figure 6.2.: Illustration of the employed loss function. The left panel shows function $f(|\xi|)$. The panel in the middle depicts function $h(\xi)$, where the tuning parameter $c = 10^{-5}$ is fixed. Parameters k and g vary as follows: $g = 20, \ldots, 1$, $k = 0.1, \ldots, 6$ in 99 steps. The right panel shows the scaled composition $\mathcal{L}(\xi)$ for the same tuning parameters.

until convergence. Thereby

$$\begin{aligned} \boldsymbol{a}_{(l)} &= \boldsymbol{X}^T \mathrm{diag}\left(\frac{\mathcal{D}(y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(l)})}{y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(l)}} \right) \boldsymbol{y}, \\ \boldsymbol{A}_{(l)} &= \boldsymbol{X}^T \mathrm{diag}\left(\frac{\mathcal{D}(y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(l)})}{y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(l)}} \right) \boldsymbol{X}, \end{aligned} \tag{6.8}$$

and $\mathcal{D}(\xi) = \frac{\partial \mathcal{L}(\xi)}{\partial \xi}$ denotes the derivative of the employed loss. The design matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T \in \mathbb{R}^{n \times q}$ comprises the covariate vectors $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{i,q-1})^T$, $i = 1, \ldots, n$, and the step length $\nu > 0$ controls the speed of convergence. The algorithm is terminated when

$$\frac{\sum_{j=0}^{q} |\boldsymbol{\beta}_{(l+1)} - \boldsymbol{\beta}_{(l)}|}{\sum_{j=0}^{q} |\boldsymbol{\beta}_{(l)}|} \leq \tau,$$

where τ is a small, positive constant. For a detailed derivation of (6.7), see Appendix D.1. To imitate the idea of nested intervals, the tuning parameters have to be chosen such that g is relatively large in the early iterations of the IRLS algorithm while it should equal one for the final iteration. In contrast, k is relatively small in the beginning of the algorithm and as large as possible for the final iteration. The constant c is as small as possible. To allow for a smooth transition $\mathcal{T} \to T$ and thus reliable results, the algorithm will have a small step length ν (for example, $\nu = 0.25$) and thus relatively many iterations until convergence. In Section 6.2.3, the (data-driven) choice of the tuning parameters is discussed in more detail.

6.2.2. Asymptotic Properties

For the final iteration of the IRLS algorithm, it holds that $g = 1$ and that k is relatively large. Hence, to show asymptotic properties, we assume $g = 1$ and consider the properties of

$$\hat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}} \mathcal{M}(\boldsymbol{\beta}) \tag{6.9}$$

for $k_n \to \infty$ and $n \to \infty$ at appropriate rates. The index n emphasizes the dependence on the sample size n. With $g = 1$, minimizing $\mathcal{M}(\boldsymbol{\beta})$ is equivalent to the minimization of $1 - K(u)$ where

$$K : \mathbb{R} \to \mathbb{R}, \quad u \mapsto K(u) = \frac{1}{2} \exp\left\{-\sqrt{u^2 + c}\right\}, \ 0 < c \leq 1, \tag{6.10}$$

and where $u = k_n \cdot (y - \boldsymbol{x}^T\boldsymbol{\beta})$. The kernel $K(u)$ in turn is an approximation of $\frac{1}{2}\exp(-|u|)$ which is the density of a Laplace distributed random variable U with mean $\mathbb{E}(U) = 0$ and variance $\mathbb{V}(U) = 2$. That is, for the final iteration, the proposed approximation can be interpreted as one minus a rounded (and thus, differentiable) Laplace kernel. As discussed in Section 6.1, approaching mode regression with kernel methods is well established and investigated. Kemp and Santos Silva (2012) derive asymptotic properties for mode regression for a general kernel $K(u)$, where $u = (y - \boldsymbol{x}^T\boldsymbol{\beta})/\delta_n$ with positive bandwidth δ_n depending on the sample size n and with the objective function $1 - \mathcal{M}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}(\delta_n^{-1}(y - \boldsymbol{x}^T\boldsymbol{\beta}))$. One can easily see that function (6.10) structurally fits in this framework as the tuning parameter k_n relates inversely to the bandwidth δ_n. Moreover, we show in Lemma 1 and Lemma 2 that a scaled version of function (6.10) meets all requirements needed to prove asymptotically consistent and normal estimates.

Consistency For proving consistency, we make the following extended assumptions following Kemp and Santos Silva (2012):

A1 $\{(\varepsilon_i, \boldsymbol{x}_i)\}_{i=1}^{\infty}$ is an independent and identically distributed (i.i.d.) sequence, where ε_i takes values in $\mathbb{R}$ and $\boldsymbol{x}_i$ takes values in $\mathbb{R}^q$ for some finite q.

A2 The parameter space $\mathcal{B}$ is a compact subset of $\mathbb{R}^q$ and contains the true value $\boldsymbol{\beta}^*$.

A3 The distribution of $\boldsymbol{x}$ is such that:

(i) $\mathbb{E}(\|\boldsymbol{x}_i\|) < \infty$, where $\|\boldsymbol{a}\|$ denotes the Euclidean norm of $\boldsymbol{a}$ for any scalar or finite-dimensional vector $\boldsymbol{a}$,

(ii) $\mathbb{P}(\boldsymbol{x}_i^T\boldsymbol{c} = 0) < 1$ for all fixed $\boldsymbol{c} \neq \boldsymbol{0}$.

A4 There exists a version of the conditional density of ε given $\boldsymbol{x}$, denoted $f_{\varepsilon|\boldsymbol{x}}(\cdot|\cdot)$: $\mathbb{R} \times \mathbb{R}^q \to \mathbb{R}$, such that:
 (i) $\sup_{\varepsilon \in \mathbb{R}, \boldsymbol{x} \in \mathbb{R}^q} f_{\varepsilon|\boldsymbol{x}}(\varepsilon|\boldsymbol{x}) \leq \infty$,
 (ii) $f_{\varepsilon|\boldsymbol{x}}(\varepsilon|\boldsymbol{x})$ is continuous for all ε and $\boldsymbol{x}$. In addition, there exists a set $A \subseteq \mathbb{R}^q$ such that $\mathbb{P}(\boldsymbol{x}_i \in A) = 1$ and $f_{\varepsilon|\boldsymbol{x}}(\varepsilon|\boldsymbol{x}) \leq f_{\varepsilon|\boldsymbol{x}}(0|\boldsymbol{x})$ for all $\varepsilon \neq 0$ and $\boldsymbol{x} \in A$.

A5 $\{k_n\}_{n=1}^{\infty}$ is a strictly positive sequence such that:
 (i) $k_n \to \infty$,
 (ii) $n\,(k_n \ln(n))^{-1} \to \infty$.

Assumptions A1 and A3 are standard assumptions. Together with A4 and Lemma 1, they are needed to prove that the objective function $\mathcal{M}$ has a global minimum at $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ which is unique (compare Lemma 1, Kemp and Santos Silva, 2012). Note that A1 does imply an i.i.d. assumption for $\varepsilon_i|\boldsymbol{x}_i$, but not for ε_i. Furthermore, A1 could be relaxed even further, but then stronger assumptions on the distribution of $\boldsymbol{x}_i$ would be required. Assumptions A2 and A5 are required to prove that the objective function satisfies a uniform law of large numbers (Lemma 2, Kemp and Santos Silva, 2012). Assumption A4 (ii) imposes that the conditional density has a global mode at zero. Assumption A5 ensures that k_n is increasing with a moderate rate. This is the crucial factor in the algorithm since no symmetry assumptions are made on the conditional density. We come back to the choice of k_n in Section 6.2.3. The kernel needs to be a bounded density that can be normalized having a bounded derivative:

Lemma 1. *The kernel function $K : \mathbb{R} \to \mathbb{R}$ defined in* (6.10) *is differentiable and fulfills the following conditions:*

(i) $\int_{-\infty}^{\infty} K(u)\mathrm{d}u = 1$,

(ii) $\sup_{u \in \mathbb{R}} |K(u)| = c_0 < \infty$,

(iii) $\sup_{u \in \mathbb{R}} |K'(u)| = c_1 < \infty$, *where* $K'(u) = \mathrm{d}K(u)/\mathrm{d}u$.

The proof of Lemma 1 can be found in Appendix D.2.2.
With these assumptions, we obtain the consistency of the mode regression estimate:

Theorem 4. *If assumptions A1–A5 hold, the IRLS-based mode regression estimate is consistent, that is*

$$\hat{\boldsymbol{\beta}}_n \xrightarrow{\mathbb{P}} \boldsymbol{\beta}^*.$$

Theorem 4 is a direct consequence of Kemp and Santos Silva (Theorem 1, 2012) and Lemma 1.

Asymptotic Normality To obtain asymptotic normality, we need the following additional assumptions:

B1 $\mathbb{E}(|\boldsymbol{x}_i|^{5+\xi}) < \infty$ for some $\xi > 0$.
B2 $\boldsymbol{\beta}^*$ belongs to the interior of $\mathcal{B}$.
B3 $f_{\varepsilon|\boldsymbol{x}}(\varepsilon|\boldsymbol{x})$ is three times differentiable with respect to ε for all $\boldsymbol{x}$ such that:
(i) $f^{(j)}_{\varepsilon|\boldsymbol{x}}(\varepsilon|\boldsymbol{x}) = \partial^j f_{\varepsilon|\boldsymbol{x}}(\varepsilon|\boldsymbol{x})/\partial\varepsilon^j$ is uniformly bounded for $j = 1, 2, 3$,
(ii) $\mathbb{E}\left[f^{(2)}_{\varepsilon|\boldsymbol{x}}(0|\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^T\right]$ is negative definite.
B4 The sequence $\{k_n\}_{n=1}^{\infty}$ is such that:
(i) $n/k_n^7 = o(1)$,
(ii) $n\left(k_n^5 \ln(n)\right)^{-1} \to \infty$.

As expected, each of these assumptions is a stronger version of the assumptions A1–A5. In particular, further moments of the distribution of $\boldsymbol{x}_i$ are required to be finite (B1) and the true parameter has to be in the interior of the parameter space $\mathcal{B}$ (B2). The latter assumption is standard in maximum-likelihood estimation. Assumption B3 guarantees the existence of a Taylor expansion of the first derivative $f^{(1)}_{\varepsilon|\boldsymbol{x}}(u/k_n|\boldsymbol{x})$ around $u = 0$. Note that no smoothness in $\boldsymbol{x}_i$ is required such that the theory also holds for categorical covariates. Finally, assumptions B4(i) and B4(ii) imply more constrained rates on k_n compared to assumption A5. We will see in Theorem 5 that B4(ii) implies that the speed of convergence of the estimate is at most $n^{2/7}$. For the kernel function, the following stronger assumptions about its smoothness are needed:

Lemma 2. *The kernel function $K : \mathbb{R} \to \mathbb{R}$ defined in* (6.10) *is three times differentiable and fulfills the following conditions:*

(i) $\int_{-\infty}^{\infty} uK(u)\mathrm{d}u = 0$,

(ii) $\lim_{u\to\pm\infty} K(u) = 0$,

(iii) $\int_{-\infty}^{\infty} u^2|K(u)|\mathrm{d}u = M_0 < \infty$,

(iv) $\int_{-\infty}^{\infty} |K'(u)|^2\mathrm{d}u = M_1 < \infty$,

(v) $\sup_{u\in\mathbb{R}} |K''(u)| = M_2 < \infty$,

(vi) $\sup_{u\in\mathbb{R}} |K'''(u)| = M_3 < \infty$,

(vii) $\int_{-\infty}^{\infty} |K''(u)|^2\mathrm{d}u = M_4 < \infty$.

With Lemma 2 and Theorems 2 and 3 from Kemp and Santos Silva (2012) asymptotic normality follows:

Theorem 5. *Under Assumptions A1–A5 and B1–B4, the IRLS-based mode regression estimate is asymptotically normal, that is*

$$\left(\frac{n}{k_n^3}\right)^{1/2}\left[\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}^*\right]\xrightarrow{d}N(0,\boldsymbol{\Omega}^*),$$

where the asymptotic covariance matrix is given by

$$\begin{aligned}
\boldsymbol{\Omega}^* &= \boldsymbol{C}^{*-1}\boldsymbol{B}^*\boldsymbol{C}^{*-1},\\
\boldsymbol{B}^* &= \lim_{n\to\infty}\mathbb{V}\left(\left(\frac{n}{k_n^3}\right)^{1/2}\left(-\frac{\partial\mathcal{M}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}^*}\right)\right)=M_1\mathbb{E}\left(f_{\varepsilon|\boldsymbol{x}}(0|\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^T\right),\\
M_1 &= \int_{-\infty}^{\infty}|K'(u)|^2\mathrm{d}u<\frac{1}{4},\\
K'(u) &= \mathrm{d}K(u)/\mathrm{d}u,\\
\boldsymbol{C}^* &= \lim_{n\to\infty}\mathbb{E}\left(-\frac{\partial^2\mathcal{M}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\bigg|_{\boldsymbol{\beta}^*}\right)=\mathbb{E}\left(f^{(2)}_{\varepsilon|\boldsymbol{x}}(\varepsilon|\boldsymbol{x})(0|\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^T\right).
\end{aligned}$$

A consistent estimate for the asymptotic covariance matrix is obtained by

$$\hat{\boldsymbol{\Omega}}_n=\hat{\boldsymbol{C}}_n^{-1}\hat{\boldsymbol{B}}_n\hat{\boldsymbol{C}}_n^{-1}\xrightarrow{\mathbb{P}}\boldsymbol{\Omega}^*,$$

where

$$\begin{aligned}
\hat{\boldsymbol{B}}_n &= n^{-1}\sum_{i=1}^{n}k_n\left[K'\left(k_n\left(y_i-\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_n\right)\right)\right]^2(\boldsymbol{x}_i\boldsymbol{x}_i^T);\\
\hat{\boldsymbol{C}}_n &= n^{-1}\sum_{i=1}^{n}k_n^3K''\left(k_n\left(y_i-\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_n\right)\right)(\boldsymbol{x}_i\boldsymbol{x}_i^T).
\end{aligned}$$

Remark The close connection to the approach of Kemp and Santos Silva (2012) provides not only the asymptotic theory for the NILS approach. Kemp and Santos Silva (2012) argue that their approach has two limiting cases: As they employ the Gaussian kernel, mean regression for $\delta_n\longrightarrow\infty$, and mode regression for $\delta_n\longrightarrow 0$. Considering $\mathcal{L}(\xi)$ as one minus a rounded Laplace kernel yields a similar interpretation for the NILS approach: The loss function $\mathcal{L}_\epsilon$ corresponds to the loss function of median regression for $k_n=g=1$. For $g=1$ and $k_n\longrightarrow\infty$, $\mathcal{L}(\xi)\longrightarrow L(\xi)$. Hence, depending on the choice of k_n, the NILS approach is closer to mode or to median regression. As $\bar{x}>\tilde{x}_{median}>\tilde{x}_{mode}$ for positively skewed distributions and $\bar{x}<\tilde{x}_{median}<\tilde{x}_{mode}$ for negatively skewed distributions, the NILS approach seems to be a natural choice to approximate mode regression.

6.2.3. Adaptive Tuning

As indicated in Section 6.2.1, the NILS approach requires tuning. The constant $c > 0$ guarantees that the loss function $\mathcal{L}(\xi)$ is differentiable. As long as it is sufficiently small, it has a minor impact on the performance and in our experience, $c = 10^{-5}$ works well. The integer g governs how broad the minimum of $\mathcal{L}(\xi)$ is. It should be large enough to guarantee $\mathcal{D}(\xi) \neq 0$ for the initial iteration and decreases towards 1 within the natural numbers while iterating. As the value of k affects the width of the minimum of $\mathcal{L}(\xi)$ for $g > 1$, it is possible to choose a fixed sequence for g (we propose to choose the fixed sequence from 10 to 1 for g) and to address all issues of tuning by a properly chosen sequence k_n of k. Since k determines how close $\mathcal{L}(\xi)$ and $L(\xi)$ are, it has to be chosen carefully and its impact on the asymptotic variance of the estimates has to be controlled. Thus, we propose to choose the sequence of values for k_n driven by the data and by the asymptotic theory. The initial value for k_n is determined as $k_{initial} = (n/const)^{1/7}$ where $const$ is chosen such that n fulfills both assumptions B4(i) and B4(iii): $const = n^{5/12} \cdot \log(n)^{7/12}/(n^{7/12})$. Then, $k_{initial}$ is increased up to $k_{final} = k_{initial} n^{1/7}/sd$ in 10 steps while iterating and where sd is the standard deviation of the residuals of a fitted median regression. After that, the final value for k_n is kept until convergence. In order to obtain smooth transitions and to reach a value of k_n that is sufficiently large, the step length is set to $\nu = 0.25$.

6.2.4. Numerical Experiments

To evaluate the performance of the NILS approach in finite samples, we consider the estimation accuracy and the applicability of the asymptotic results in a linear model. Concretely, we generate $n_{rep} = 100$ replications of the model

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \\ &= 1 + 0.2 x_1 - 2 x_2 + 3 x_3 + \varepsilon, \end{aligned} \tag{6.11}$$

where x_1, x_2, x_3 are drawn from the continuous uniform distribution on $[0, 2]$. Thereby, different *model features* are systematically varied:

- The distribution of the errors ε is either Gaussian $\varepsilon \sim N(0,1)$, log-normal $\varepsilon \sim LN(0,1)$ or gamma $\varepsilon \sim Ga(s = 2, r = 2)$, where s and r denote the shape and the inverse scale parameter. That is, we consider a symmetric scenario where mean, median and mode coincide and two skew scenarios with differently shaped error distributions. As the mode of the skew distributions is unequal zero, they are shifted accordingly.
- Different sample sizes are considered: $n \in \{100, 500, 1000, 10000\}$.

For each replication of model (6.11), four different *methods* are compared:

- the NILS approach with adaptive tuning as proposed in Section 6.2.3,
- the approach of Kemp and Santos Silva (2012),
- mean regression and
- median regression.

According to Kemp and Santos Silva (2012), the bandwidth δ_n of their approach is chosen based on the median of the absolute deviation from the median least squares residual of a preceding mean regression: $\text{MAD} = \text{med}_i \left\{ \left| (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}) - \text{med}_j (y_j - \boldsymbol{x}_j^T \boldsymbol{\beta}) \right| \right\}$ and $\delta_n = 1.2 \cdot \text{MAD} \cdot n^{1/7}$. The mean regression estimates are employed as starting values.
The results of the median regression are obtained by an IRLS algorithm that approximates the absolute loss function $|\xi|$ by $\sqrt{\xi + c}$, where c denotes a small positive constant, for example, $c = 10^{-5}$. This is advantageous as it allows for exactly the same computational structure for both, median and mode regression.
As the speed of convergence of the asymptotic theory in Section 6.2.2 is rather slow, we expect that the coverage rates of the confidence intervals (CI) based on the asymptotic covariance matrix $\hat{\boldsymbol{\Omega}}_n$ are reliable only for a rather large number of observations n. Hence, beside the CIs derived from asymptotic normality, we evaluate $(1-\alpha)$ CIs based on bootstrap samples of the residuals $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_n$. For each sample $(\boldsymbol{y}_b^*, \boldsymbol{X})_{b=1,\dots,B}$ with $\boldsymbol{y}_b^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}_n + \hat{\varepsilon}_b^*$, the according model is estimated and we obtain the bootstrap estimates $\hat{\boldsymbol{\beta}}^*_{b=1,\dots,B}$. The pointwise $(1-\alpha)$ CI for the estimated coefficient $\hat{\boldsymbol{\beta}}_n$ is then defined by the $\alpha/2$ and the $1-\alpha/2$ quantile of the empirical distribution of the bootstrap estimates $\hat{\boldsymbol{\beta}}^*_{b=1,\dots,B}$. This approach assumes that the functional form of the regression model is correctly specified and that the errors are identically distributed (Fox, 2008, page 598). While this may seem rather restrictive, nonparametric bootstrap samples are not a good choice for mode regression as the samples $(\boldsymbol{y}, \boldsymbol{X})^*_{b=1,\dots,B}$ contain duplicated observations. Duplicated or even multiplied observations imply a mode of $\varepsilon|\boldsymbol{X}$ and can therefore render the estimation procedure unstable. Following Efron and Tibshirani (1993), we choose $B = 1000$ to determine the bootstrap CIs.

Results To judge the results, the estimation accuracy and the coverage rates of 95% CIs are considered. The left panel of Figure 6.3 shows boxplots of the resulting coefficients for $n = 100$ observations, $\varepsilon \sim N(0,1)$ (top) and $\varepsilon \sim LN(0,1)$ (bottom). The results can be summarized as follows:

- In the most simple scenario with standard normal errors, the estimation accuracy of the NILS approach is not as precise as the results of mean and median regression which was to be expected since the error distribution is symmetric. Due to some outliers, the variations of the approach of Kemp and Santos Silva (2012) are slightly larger.

Error Distribution:		$N(0,1)$				$LN(0,1)$				$Ga(2,2)$			
	n:	100	500	1000	10000	100	500	1000	10000	100	500	1000	10000
Mode	β_0	0.04	0.10	0.09	0.29	0.26	0.28	0.36	0.45	0.05	0.06	0.09	0.11
regression	β_1	0.04	0.06	0.13	0.32	0.32	0.48	0.53	0.79	0.08	0.15	0.12	0.26
NILS	β_2	0.09	0.10	0.07	0.29	0.32	0.50	0.57	0.78	0.09	0.18	0.12	0.21
	β_3	0.08	0.08	0.11	0.28	0.35	0.50	0.53	0.65	0.04	0.12	0.16	0.20
Mode	β_0	0.76	0.87	0.76	0.90	0.63	0.54	0.49	0.37	0.70	0.71	0.67	0.40
regression	β_1	0.83	0.79	0.80	0.93	0.80	0.85	0.76	0.83	0.79	0.75	0.84	0.85
NILS BS	β_2	0.73	0.79	0.84	0.86	0.81	0.83	0.83	0.79	0.73	0.72	0.80	0.83
	β_3	0.84	0.81	0.83	0.90	0.78	0.82	0.80	0.75	0.84	0.73	0.79	0.89
Mode	β_0	0.50	0.77	0.72	0.81	0.58	0.35	0.13	0.01	0.33	0.46	0.53	0.60
regression	β_1	0.59	0.77	0.69	0.86	0.83	0.87	0.80	0.88	0.39	0.59	0.46	0.64
Kemp	β_2	0.55	0.70	0.76	0.81	0.77	0.88	0.88	0.81	0.39	0.51	0.59	0.60
	β_3	0.57	0.74	0.76	0.86	0.71	0.84	0.81	0.83	0.39	0.44	0.55	0.57
Mode	β_0	0.75	0.83	0.74	0.85	0.67	0.46	0.21	0.03	0.64	0.77	0.78	0.78
regression	β_1	0.75	0.78	0.69	0.80	0.89	0.91	0.83	0.91	0.83	0.81	0.80	0.91
Kemp BS	β_2	0.78	0.72	0.74	0.77	0.88	0.91	0.90	0.83	0.74	0.80	0.79	0.85
	β_3	0.74	0.70	0.77	0.90	0.89	0.89	0.89	0.85	0.83	0.82	0.83	0.80

Table 6.1.: Coverage rates of the confidence intervals estimated for different sample sizes and different error distributions; BS denotes that the results rely on $B = 1000$ bootstrap samples.

- In the scenario with log-normal errors, mean, median and mode of the error distribution differ by a location shift. The lower left plot of Figure 6.3 illustrates that this shift is captured by the estimates of the intercept β_0. The results of mean and median regression are clearly scattered around a value different from the true value which is indicated by a horizontal line. Again, this was to be expected as the structure of the errors is additive. Both, the NILS approach and the proposal of Kemp and Santos Silva (2012) are biased slightly regarding the intercept while the remaining coefficients are estimated equally well by all methods.
- The middle panel of Figure 6.3 shows the widths of the confidence intervals for each coefficient obtained with the asymptotic theory of Section 6.2.2. One sees that the interval widths for the NILS and the Kemp approach differ substantially as they depend on the choice of the tuning parameters k_n and δ_n. For the NILS approach, k_n is increased steadily while iterating whereas Kemp and Santos Silva (2012) choose a fixed bandwidth. Table 6.1 gives the corresponding coverage rates for normally, log-normally and gamma distributed errors, $n \in \{100, 500, 1000, 10000\}$. Both approaches perform differently well for different error distributions. Beside the different tunings employed, another reason for the partly insufficient results is the slow rate of convergence of at most $n^{2/7}$. In practice, we advise to apply bootstrap methods to assess the estimate's variance. The coverage rates relying on $B = 1000$ bootstrap samples in Table 6.1 seem to confirm this recommendation for both approaches.

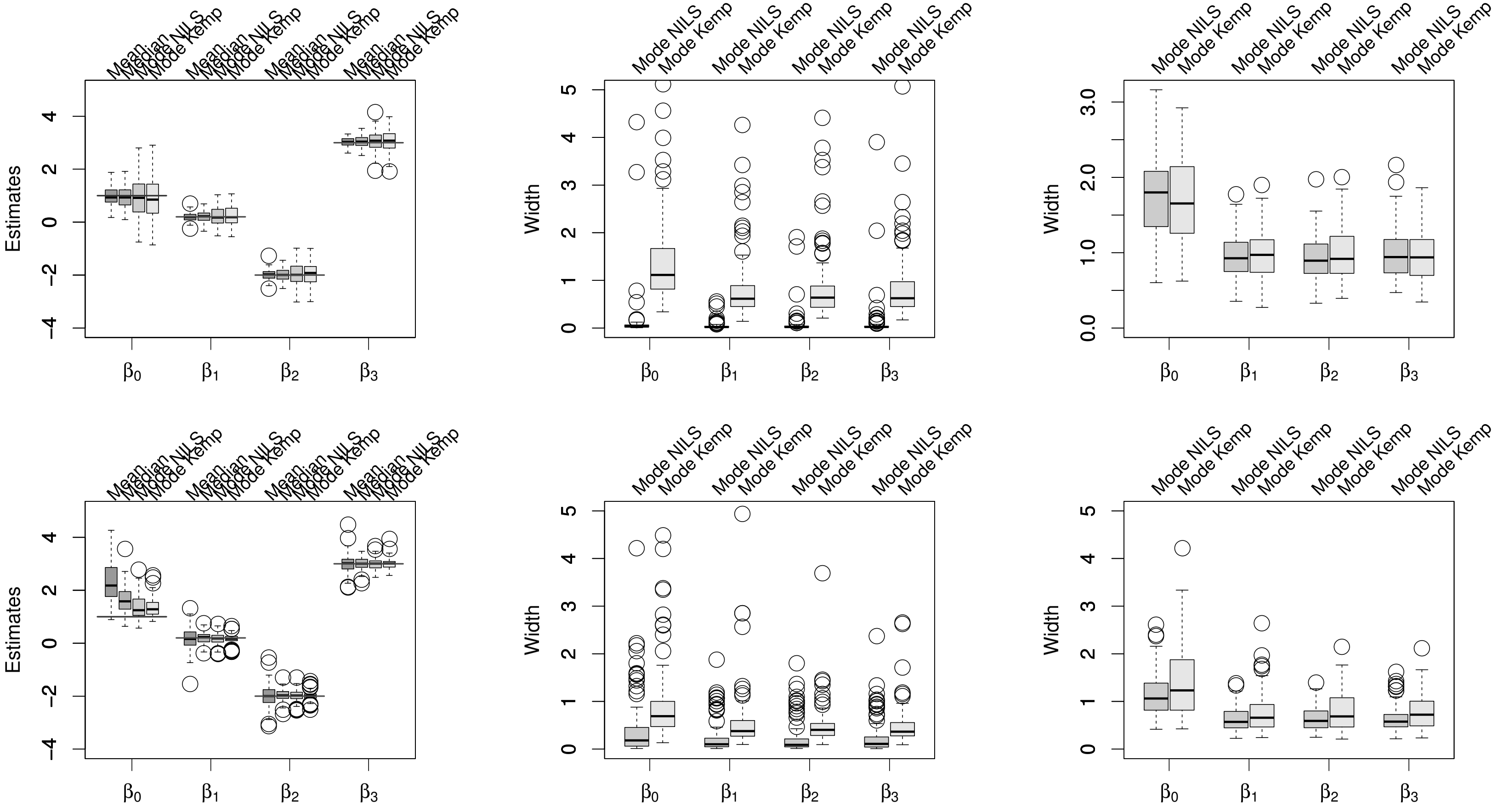


Figure 6.3.: Estimated regression coefficients (left), estimated widths of the asymptotic confidence intervals (middle) and of the bootstrap confidence intervals (right); $n = 100$ observations; $\varepsilon \sim N(0,1)$ (top) and $\varepsilon \sim LN(0,1)$ (bottom). The true coefficients are indicated by horizontal lines.

6.3. Semiparametric Mode Regression

6.3.1. Semiparametric Modeling

So far, for mode regression, predictors have been restricted to parametric effects – either due to methodical reasons or to ensure numerical stability. In contrast, the NILS approach allows to easily augment the linear predictor in model (6.1) to semiparametric predictors of the form

$$y = \boldsymbol{x}^T\boldsymbol{\beta} + \sum_{j=1}^{r} f_j(z_j) + f_{geo}(s) + \varepsilon,$$

where as before, $\boldsymbol{x}^T\boldsymbol{\beta}$ represent the linear effects. The functions f_j represent nonlinear smooth effects of continuous covariates z_j, $j = 1, \ldots, r$, modeled by penalized B-splines (Eilers and Marx, 1996) of degree 3 and with 20 outer knots as a default option. The effect f_{geo} allows to include spatial information which will be relevant in our application on the Munich rent index in Section 6.5.

Semiparametric models – which are also known as generalized additive (mixed) models (Hastie and Tibshirani, 1990; Wood, 2006) – are an established tool in many fields of regression modeling. And in fact, the predictor above is not the most general form. For mean regression, Fahrmeir et al. (2013) give an extensive overview of generic predictor representations where further effect types such as interactions between two continuous covariates or random effects can be included into the predictor. The general assumption is that each function f (independent of the type of the covariate x) can be written as a linear combination of appropriate basis functions, that is, $f(x) = \sum_{k=1}^{d} B_k(x)\beta_k$ which allows to write a vector of n function evaluations in matrix notation as $\boldsymbol{f} = \boldsymbol{X\beta}$. To achieve specific properties such as the smoothness of a function f, estimation is regularized by additional penalty terms. Specifically, we assume quadratic penalties of form $P_\lambda(\boldsymbol{\beta}) = \lambda\boldsymbol{\beta}^T\boldsymbol{K\beta}$, where $\boldsymbol{K} \in \mathbb{R}^{q\times q}$ is an appropriate penalty matrix and $\lambda \geq 0$ is a penalty parameter that determines the strength of the regularization. Estimation in mode regression is then enabled by augmenting matrix $\boldsymbol{A}$ defined in equation (6.8):

$$\boldsymbol{A} \quad = \quad \boldsymbol{X}^T \text{diag}\left(\frac{\mathcal{D}(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(l)})}{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(l)}}\right)\boldsymbol{X} + \lambda\boldsymbol{K}.$$

Like a modular system and with none but the usual restrictions, mode regression can be combined with any quadratic penalty and/or smooth component. As we do work with an IRLS algorithm, the proposed approximation can be combined with the `R` package `mgcv` (Wood, 2011) such that a wide range of smooth components and several options to choose the penalty parameter λ are available. Note that the asymptotic theory of Section 6.2.2 does not include penalized estimates. While for fixed smoothing parameters one might argue

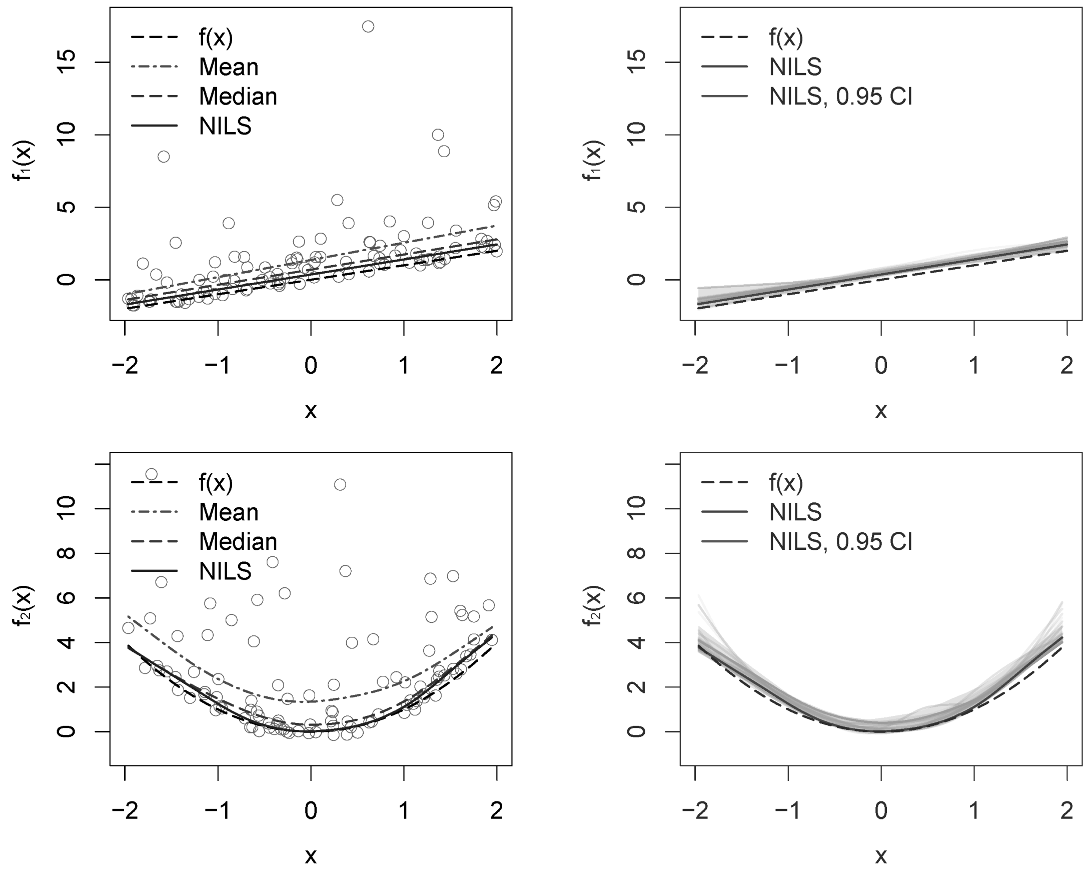


Figure 6.4.: Examples for data fits of functions $f_1(x)$ and $f_2(x)$ (by rows) for $\varepsilon \sim LN(0,1)$, $n = 100$. On the left, the results of mean, median and mode regression are compared. On the right, the bootstrap confidence intervals for the NILS approach are illustrated. The penalty parameter λ is chosen by the REML criterion.

that the asymptotic theory may carry over to penalized estimation, a much more careful investigation would be required for data-driven smoothing parameter estimates. Anyway, to achieve reliable results, a moderate number of observations relative to the model complexity was already required in a parametric setting, compare Section 6.2.4. Therefore, bootstrap methods turned out to be an attractive alternative. We will likewise employ bootstrap methods to asses how stable the estimated effects are in semiparametric mode regression. More specifically, we consider the pointwise $\alpha/2$ and $1 - \alpha/2$ quantiles of the functions fitted on the bootstrap samples in order to judge the variability of a fitted function.

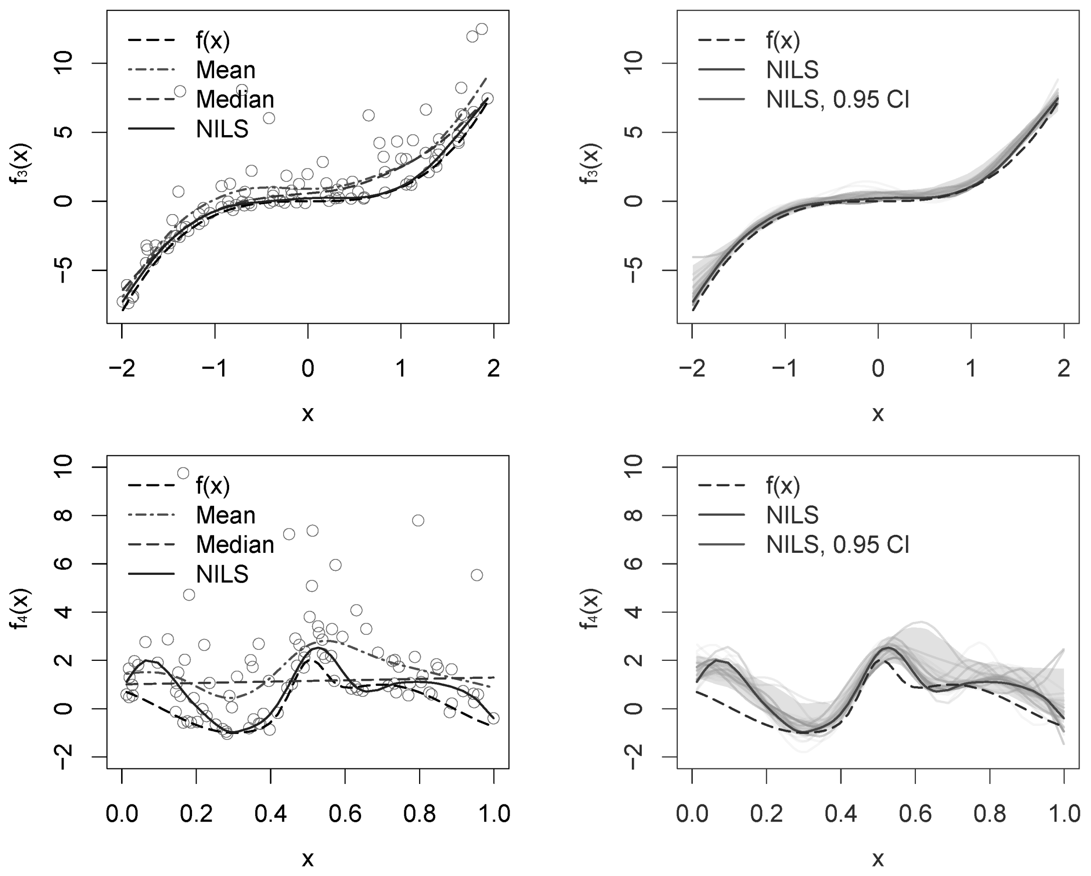


Figure 6.5.: Examples for data fits of functions $f_3(x)$ and $f_4(x)$ (by rows) for $\varepsilon \sim LN(0,1)$, $n = 100$. On the left, the results of mean, median and mode regression are compared. On the right, the bootstrap confidence intervals for the NILS approach are illustrated. The penalty parameter λ is chosen by the REML criterion.

6.3.2. Numerical Experiments

We investigate the performance of the proposed methods empirically. In contrast to the previous settings, penalized smooth components require to choose the penalty parameter(s) λ adequately. For mean regression, different strategies such as k-fold cross-validation with a specific loss criterion or the generalized cross-validation criterion of O'Sullivan et al. (1986) are available. Often, these criteria are based on rank estimation, that is, on the estimated hat matrix, or estimated degrees of freedom. It is not clear whether this makes sense for the employed loss function and if so, how data-sensitive the proposed method is. Moreover, combining the estimation with the `R` package `mgcv` implies that the estimation of the coefficient vector $\boldsymbol{\beta}$ and of the penalty parameter λ are interlaced. Hence, we consider not only the performance of semiparametric mode regression but compare the performance of different strategies for the choice of λ. Concretely, we consider (i) the generalized cross-validation criterion of O'Sullivan et al. (1986) where $\boldsymbol{\beta}$ and λ are estimated separately (referred to as “CV”), (ii) the same generalized cross-validation criterion with interlaced estimation (“GCV”) and (iii) the negative log restricted likelihood criterion with interlaced estimation (“REML”); whereat (ii) and (iii) are implemented in `mgcv`. As a benchmark, we combine the approach of Kemp and Santos Silva (2010) with quadratic penalties, too.

The adaptive tuning depends on the assumptions for the asymptotic theory for parametric models and requires a preceding median regression. As in some semiparametric settings the results of median regression differ substantially from those of the mode regression, we avoid adaptive tuning in the following. Instead, tuning parameters that summarize the experiences of the data-adaptive choice of k in the parametric settings in Section 6.2.4 are employed. For samples sizes $n = 100$ and $n = 500$, we consider $n_{rep} = 100$ replications of the model $y = f(x) + \varepsilon$. The errors ε are normally or log-normally distributed as in Section 6.2.4. In order to consider an extreme scenario where the mode should be found at the lower boundary of the data, exponential errors $\varepsilon \sim Exp(0.5)$ are included. The data generating function $f(x)$ is chosen as either

- a linear effect $f_1(x) = x$,
- a parabola $f_2(x) = x^2$,
- a cubic polynomial $f_3(x) = x^3$
- or a trigonometric function $f_4(x) = \sin(2(4x - 2)) + 2\exp(-16^2 \cdot (x - 0.5)^2)$.

The covariate x is uniformly distributed on $[-2, 2]$ for $f_1(x)$, $f_2(x)$, $f_3(x)$ and on $[0, 1]$ for $f_4(x)$. The functions are modeled with cubic B-spline bases with 20 equally spaced outer knots and second order differences in the penalty matrix.

Figures 6.4 and 6.5 show the fitted functions for exemplary data sets with sample size $n = 100$. In the left panels, the results of mean, median and mode regression are compared while in the right panels, $1 - \alpha = 0.95$ confidence intervals based on $B = 100$ bootstrap samples illustrate the variability of the estimation procedure.

Results To evaluate the results, the root mean squared errors (RMSE) for the fitted values are shown in Figures 6.6 and 6.7 for $n = 100$. We conclude:

- Combining penalized splines and the NILS approach seems to be very reasonable for the purpose of a nonlinear mode regression, especially, when the penalty parameter λ is chosen by GCV or REML. With skew errors and GCV/REML, the NILS approach results in the lowest RMSE in nearly all settings. For normal errors, it is obviously less efficient, but the loss is about of the same magnitude as from mean to median regression.
- The performance of the approach of Kemp and Santos Silva (2010) depends strongly on the set of starting values. Even though the boxplots are based on the best starting values we found (a preceding mean regression), the approach of Kemp and Santos Silva (2012) comes along with a distinctively larger RMSE in nearly all settings.

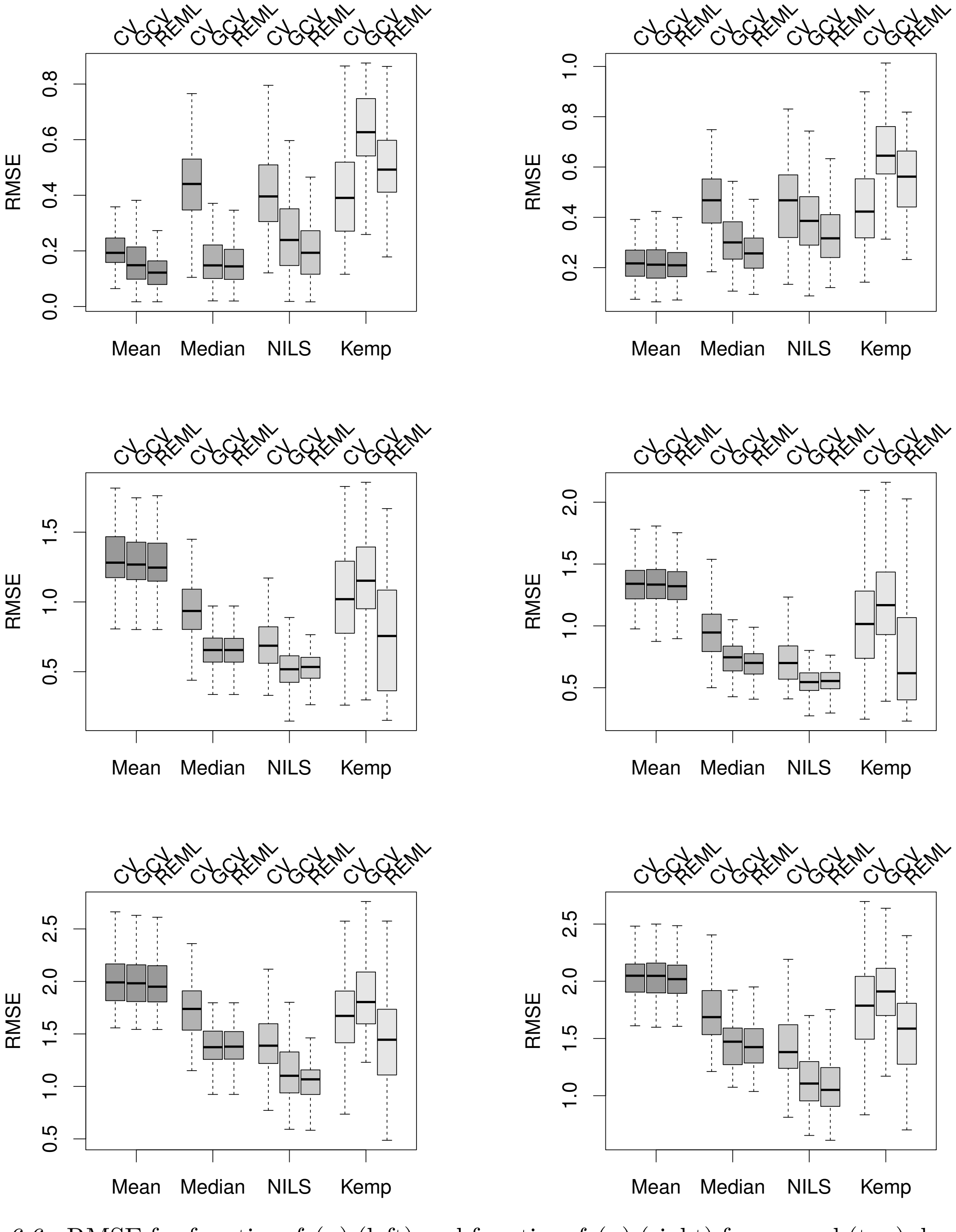


Figure 6.6.: RMSE for function $f_1(x)$ (left) and function $f_2(x)$ (right) for normal (top), log-normal (middle) and exponential (bottom) errors; $n = 100$.

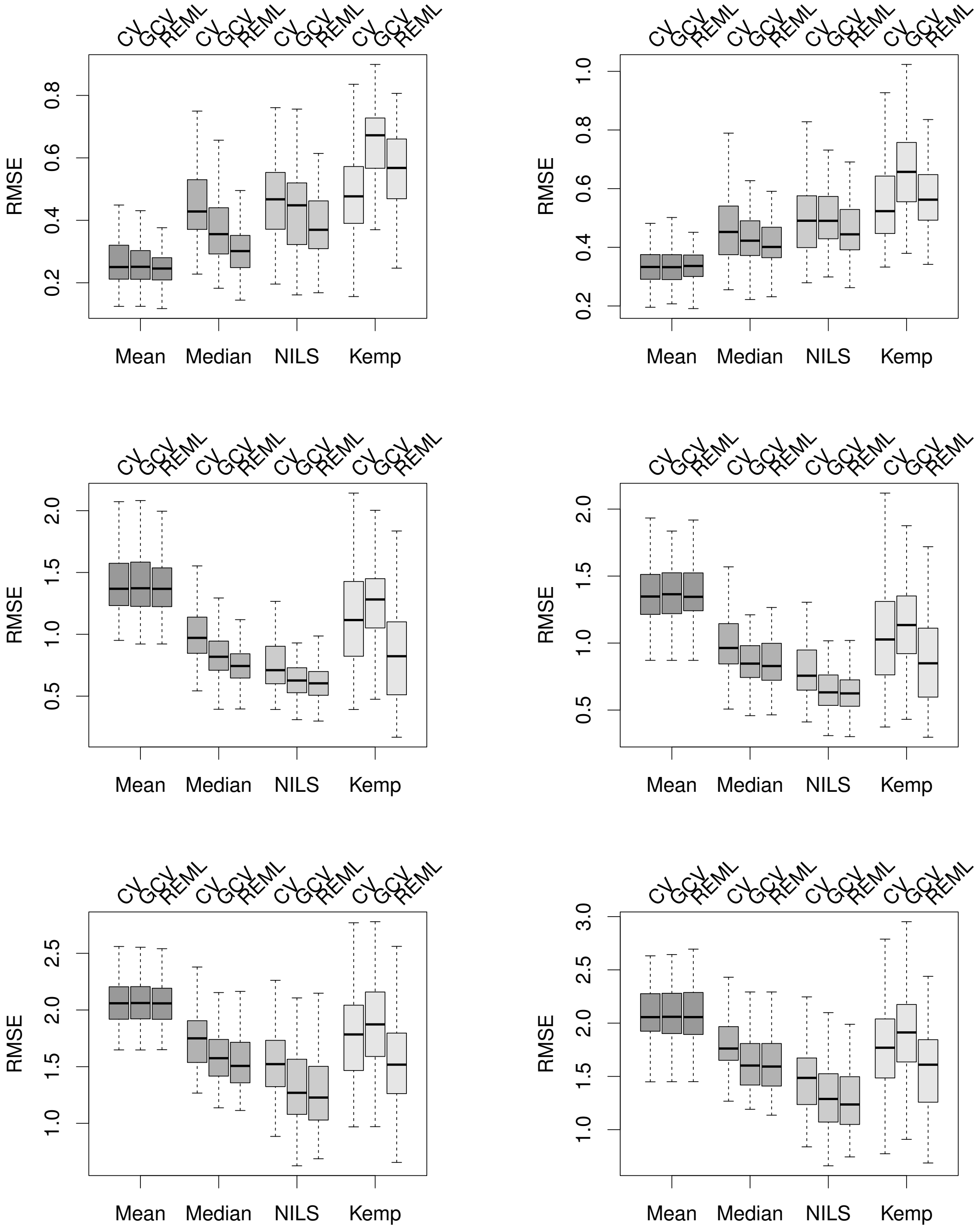


Figure 6.7.: RMSE for function $f_3(x)$ (left) and function $f_4(x)$ (right) for normal (top), log-normal (middle) and exponential (bottom) errors; $n = 100$.

	Model (6.12)			Model (6.13)			Model (6.14)		
	Mean	Median	Mode Kemp	Mean	Median	Mode NILS	Mean	Median	Mode NILS
β_0	26.610	25.380	23.846	26.437	25.303	23.088	26.533	25.382	23.049
β_n	0.074	0.426	-0.354	0.074	0.431	0.009	0.075	0.444	-0.084
β_y	0.064	0.052	-0.028	0.064	0.051	-0.024	–	–	–
β_{a1}	3.051	3.549	4.095	–	–	–	–	–	–
β_{a2}	-0.342	0.565	0.088	–	–	–	–	–	–
β_{a3}	0.733	0.839	-2.059	–	–	–	–	–	–

Table 6.2.: Estimated parametric effects for the mean, median and mode of the BMI in the considered models.

6.4. Mode Regression for the BMI Distribution in England

To explain the development of the body mass index (BMI) in England, we reanalyze a data set already used in Kemp and Santos Silva (2010) with a focus on non-pregnant women between the ages of 18 and 65 observed in the period between 1997 and 2006. This yields a data set of 44,651 observations with the age, the calendar year of the study and a binary factor indicating non-white women as available covariates. Our first model is in accordance with Kemp and Santos Silva (2010) where the effect of age is modeled by a polynomial while the other covariates are treated linearly:

$$\text{BMI} = \beta_0 + \beta_n \text{non-white} + \beta_y \text{year} + \beta_{a1}\log(\text{age}) + \beta_{a2}\log(\text{age})^2 + \beta_{a3}\log(\text{age})^3 + \varepsilon. \quad (6.12)$$

As seen in Table 6.2, one finds a slightly negative effect of the calendar year in mode regression while in mean regression, the effect of the calendar year is positive. In the left panel of Figure 6.8, the estimated effect of the age is shown.

A more flexible way to model the effect of the age is to replace the linear predictor above with

$$\text{BMI} = \beta_0 + \beta_n \text{non-white} + \beta_y \text{year} + f(\text{age}) + \varepsilon, \quad (6.13)$$

where $f(\text{age})$ is modeled by penalized cubic B-splines with 14 knots as the default choice of 20 knots caused some numerical instabilities. The set of tuning parameters is chosen as described in Section 6.3.2, and the penalty parameter λ is chosen by the REML criterion. Estimates of the parametric effects are given in Table 6.2, the estimate of the smooth function is shown in the right panel of Figure 6.8. At first sight, the effect of the age seems to be wigglier, but the trend does perfectly fit to the routines of a typical lifestyle and to typical biological changes: The effect of the BMI is relatively constant in early adulthood and increases around the age of 30. The second increase of the effect coincides with the typical age of the climacteric period.

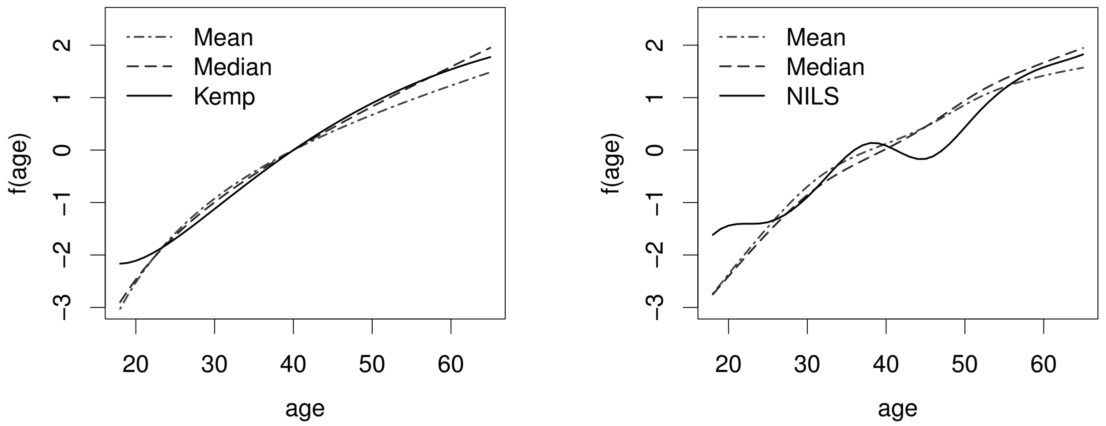


Figure 6.8.: The estimated effect of age in model (6.12) (left panel) and in model (6.13) (right panel). For comparison, the results of mean and median regression are added.

In a third model, not only the effect of the age but of the age and of the calendar year are modeled smoothly:

$$\text{BMI} = \beta_0 + \beta_n \text{non-white} + f_1(\text{year}) + f_2(\text{age}) + \varepsilon, \tag{6.14}$$

Again, the estimates of parametric effects are given in Table 6.2. In Figure 6.9 (top), the estimates of f(age) and f(year) are plotted. For both effects, there is a clear difference between the fitted functions for mean, median and mode regression. For mean and median regression, the estimated effect of the age has the same functional form as in model (6.13), while the estimated effect of the calendar year is an increasing function suggesting an increasing BMI over time. However, for mode regression, the effect is ambiguous: There is a positive effect in the first two years of the study, but a negative one in the last two years. In Figure 6.9 (bottom), the results of the models fitted on $B = 50$ bootstrap samples are added confirming the shape of the effect of the age on the mode of the response (left panel). As before, the effect of the year cannot be clearly classified (right panel).

As seen in the empirical evaluation in Section 6.2.4, the differences in the estimated intercepts $\hat{\beta}_0$ as seen in Table 6.2 indicate a general skewness in the conditional distribution of the BMI. However, in Section 6.2.4, the effects of the covariates are estimated equally well for mean, median and mode regression even when the errors are skewed. As the estimated effects in Table 6.2 differ, this may be an indication for a more complex error structure.

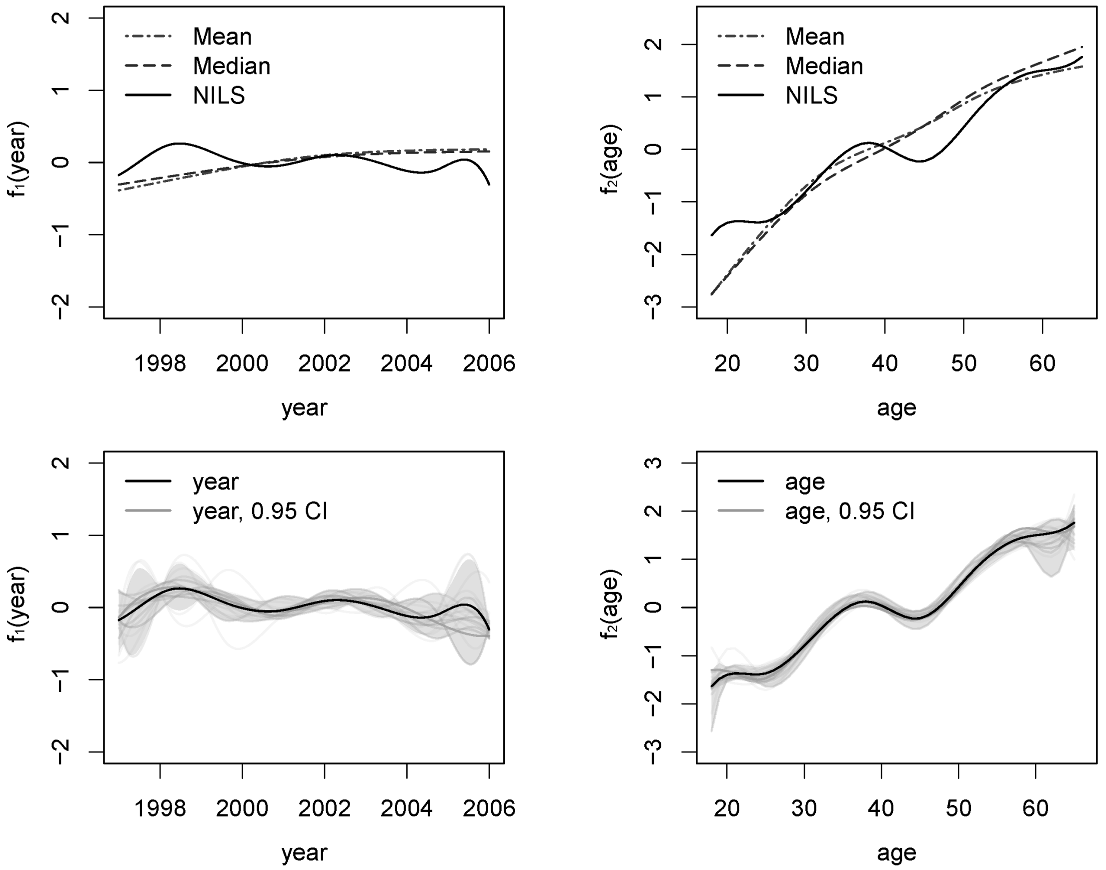


Figure 6.9.: The estimated effects of the calendar year (left panel) and the age (right panel) in model (6.14). On top, the results of mean and median regression are added. On bottom, fitted functions for $B = 50$ bootstrap samples are added.

	Mean		Median		NILS	
Intercept	8.84	(8.81, 8.94)	8.91	(8.87, 9.01)	8.93	(8.87, 9.05)
Absence of bathroom	0.89	(0.52, 1.18)	0.71	(0.35, 1.20)	0.23	(-0.16, 0.85)
Presence of second bathroom	-0.64	(-0.79, -0.48)	-0.87	(-1.07, -0.70)	-0.97	(-1.28, -0.81)
Special features of bathroom	0.54	(0.37, 0.78)	0.53	(0.30, 0.72)	0.71	(0.47, 0.90)
Normal quality kitchen	0.70	(0.57, 0.93)	0.64	(0.56, 0.86)	0.70	(0.59, 0.80)
Good quality kitchen	1.05	(0.86, 1.20)	1.13	(0.91, 1.29)	1.03	(0.68, 1.16)
Absence of intercom	-0.47	(-0.66, -0.41)	-0.58	(-0.77, -0.51)	-0.62	(-0.86, -0.58)
Simple floor cover	-1.10	(-1.33, -0.96)	-1.05	(-1.28, -0.90)	-1.00	(-1.21, -0.84)
Absence of warm water supply	-1.39	(-1.81, -1.02)	-1.42	(-1.92, -1.01)	-1.82	(-2.70, -1.27)
Absence of central heating system	-1.16	(-1.29, -0.86)	-1.31	(-1.56, -0.96)	-1.31	(-1.50, -0.67)
Presence of storage heating system	-0.79	(-1.10, -0.57)	-0.66	(-0.96, -0.50)	-0.38	(-0.68, 0.09)
Simple and old building	-0.71	(-0.96, -0.54)	-0.83	(-1.18, -0.65)	-0.67	(-0.94, -0.27)
Simple and post world war building	-0.64	(-0.83, -0.31)	-0.84	(-1.05, -0.51)	-1.18	(-1.37, -0.74)

Table 6.3.: List of categorical covariates in model (6.15) and the corresponding estimated effects for mean, median and mode regression. In parenthesis, there are 95% confidence intervals based on $B = 50$ bootstrap samples.

6.5. The Munich Rent Index

In a second application, we analyze data on the rents in Munich with mode regression. The data has been collected in 2003 and gives detailed information on the living conditions and the associated costs of 3051 flats in Munich. Previous analyses of this data set show strong nonlinear and spatial effects on the expected net rent as dependent variable, but also reveal the presence of heteroscedasticity and skewness (Kneib, 2013). Hence, we compare the results of mean, median and mode regression for the model equation

$$\text{rent} = \boldsymbol{x}^T\boldsymbol{\beta} + f_1(\text{year}) + f_2(\text{size}) + f_{\text{spat}}(\text{district}) + \varepsilon, \tag{6.15}$$

where $\boldsymbol{x}$ consists of 12 categorical covariates indicating several quality attributes of a flat such as the kitchen equipment or the type of heating (see Table 6.3 for a complete list). Functions f_1(year) and f_2(size) represent nonlinear effects of the year of construction and of the size of the flat (in square meters). They are approximated by penalized cubic splines with 14 outer knots and with a second order difference penalty. The spatial effect f_{spat}(district) is defined by 100 districts in Munich and estimated by a Markov random field (Rue and Held, 2005).

The estimated coefficients for the categorical covariates obtained with mode, mean and median regression are given in Table 6.3. While, for the mean, a second bathroom makes the flat about 0.89 € per square meter more expensive, the effect of this covariate is stronger for mode regression. We also see that flats in simple post-war buildings are generally cheaper, but there is a clear difference between a price reduction of 0.64 € on average and a re-

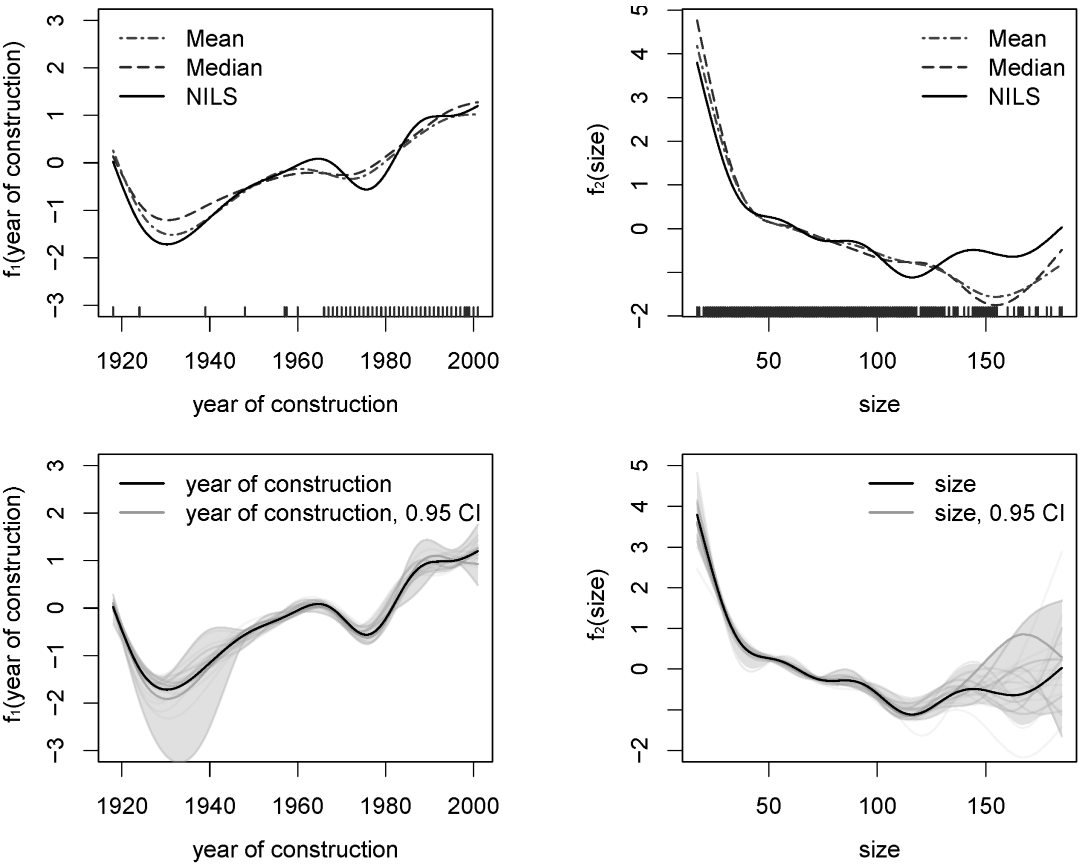


Figure 6.10.: Estimated effects of year of construction and size in square meters in model (6.15).

duction of 1.18€ for mode regression. Overall, the table shows that the estimates for the mean and the median are very similar while we can find stronger differences in comparison to the mode regression estimates. Therefore, we might distinguish between average rents and typical rents. This is supported by the estimated nonlinear effect of the size of a flat in square meters as shown in Figure 6.10. Again, we find strong similarities between the mean and median while the estimated effect for mode regression is less extreme, especially for flats larger than $140\,m^2$. Similarly, the estimated spatial effects of mean and mode regression depicted in Figure 6.11 show similar patterns even though the results of the median regression are less extreme. The results of mode regression show that the typical rents of a few outlying districts differ immensely from those estimated by the other location measures. In fact, the variability of the spatial effect is the largest for mode regression and the smallest for median regression.

Overall, the estimated tendencies are roughly the same for mean, median and mode regression – despite some effects that are remarkably different for average and for typical flats. The results of mode regression point out effects that should be investigated carefully in order to understand the pricing mechanism.

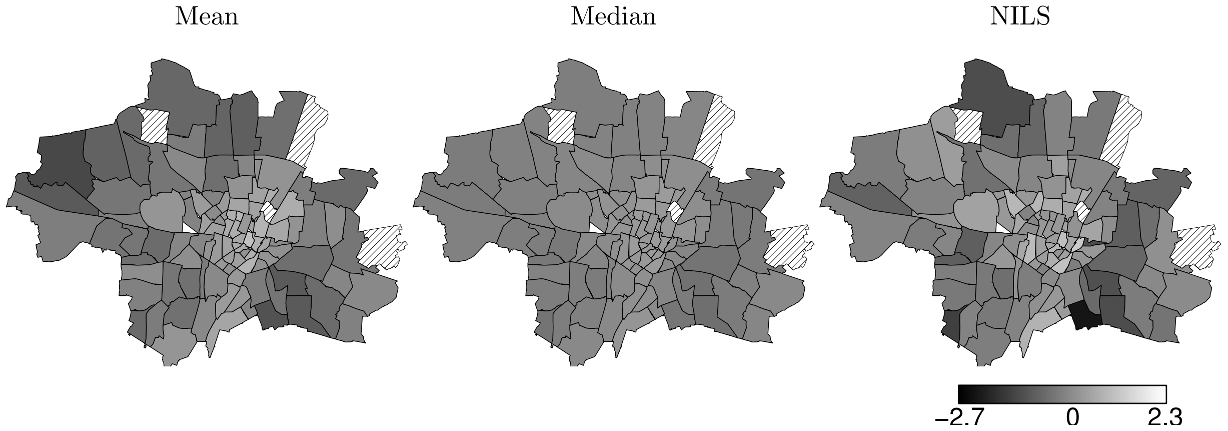


Figure 6.11.: Estimated spatial effects of the districts in model (6.15) in mean, median and mode regression. For the hatched areas, there are no observations. The figure is created with the R package BayesX (Kneib et al., 2014).

6.6. Remarks

We developed a new estimator for the conditional mode based on a local quadratic approximation $\mathcal{L}$ of the limiting case in (6.5) which can be determined iteratively with a nested interval approach. The properties of our kernel function allow to adapt asymptotic properties of the estimator in a parametric setting. However, similar to Kemp and Santos Silva (2012), the rate of convergence is rather slow such that depending on the error structure, confidence intervals are much to narrow. In most situations, this problem can be reduced with bootstrap methods.

The main advantage of our approach is that it can easily be extended to semiparametric predictor structures, yielding considerably expanded flexibility in the specification of conditional mode regression. The penalized IRLS framework also allows to borrow existing inferential tools from mean regression, for example, for the determination of smoothing parameters. An open question for future research is the extension of the asymptotic results to such semiparametric specifications, especially when including data-driven estimates for the smoothing parameter and/or basis dimensions increasing with the sample size.

In cases, where the error structure is additive, mean, median and mode regression should only differ in a shift of the intercepts such that mode regression (in addition to the appeals mentioned in the introduction) can be a helpful tool to draw conclusions about the underlying error structure. Although Taylor and Einbeck (2011) show that the true multivariate mode regression is hard to interpret due to the non-additivity of the predictor, extending our approach to bivariate problems would allow to study the mode of a joint bivariate distribution and is conceptually straight forward.

7. Concluding Remarks

In this thesis, different aspects of penalized regression models for discrete structures are discussed. The term “discrete structure” subsumes a wide range of concepts. Here, it denotes all kinds of categorical effects, of categorical effect modifiers, and of group-specific effects in hierarchical settings. If such structures are part of a regression model, each level of a categorical covariate or of a categorical effect modifier and each second level unit in a hierarchical setting results in one parameter that has to be estimated. In order to obtain less complex models and more efficient estimates, different strategies are possible. In this thesis, penalized approaches are the method of choice:

- L_1-type penalties are employed to fuse the categories of categorical effects and of categorical effect modifiers in generalized linear models (Chapter 3).
- Applying similar penalties to group-specific models, allows to detect clustered heterogeneity in hierarchical data sets (Chapter 4).
- There are situations where L_1-type penalties for discrete structures have some drawbacks. Therefore, the performance of L_0-type penalties is investigated (Chapter 5).
- As the proposed penalty terms are singular or non-continuous, the estimation is not an easy task. A general family of penalties is approximated quadratically such that the computational issues can be met by a conventional algorithm (Chapter 2).

As the L_0 norm is not restricted to penalty terms, this thesis goes beyond the scope of penalties. In Chapter 6, the L_0 norm is employed as a loss function. Regression models that approximate the conditional mode of a response are considered.

At first sight, a categorical effect modifier is just an alternative coding of the interaction of a categorical effect and a continuous covariate. However, the study on the acceptance of boar meat illustrates that categorical effect modifiers allow to meet the special requirements of the data, in this case, of the study design (see Section 3.5). Thus, with regards to the context, the notion of an effect modifier can give meaningful interpretations. With regards to the theory, the concept of categorical effect modifiers is advantageous as categorical effect modifiers do not require a reference category. This property allows for penalties that are capable of model selection and that do not depend on the choice of a reference category which is arbitrary. Concretely, a combination of the Lasso (least absolute shrinkage and selection operator; Yuan and Lin, 2006) and the fused Lasso (Tibshirani et al., 2005) in the framework of generalized linear models (GLM, see, for example, McCullagh and Nelder, 1983) is proposed. This penalty allows to select varying coefficient terms and to fuse categories that have the same effect on the response. The different amount of information of nominal and ordinal effects is considered. For the former, the fused Lasso penalty consists of all pairwise differences of coefficients. For the latter, only adjacent differences of coefficients are penalized. The approach is shown to have both, nice asymptotic properties and a reliable performance for finite samples. The weak point of the approach is that the number of penalty terms grows quadratically for nominal effect modifiers. This is problematic when there are effect modifiers with a large number of categories. It remains to be investigated how the penalty's complexity can be reduced for such settings. One could, for example, consider the pairwise differences of coefficients in a certain neighborhood of coefficients. However, such neighborhoods have to be chosen carefully. If they rely, for example, on the maximum likelihood estimate, the order of the estimate will affect the results. Moreover, one has to consider that the relevant neighborhood of coefficients may change with an increasing penalty parameter. The proposed approach can be extended to various model classes: Zhao et al. (2014) apply the same penalty to quantile varying coefficient models. For longitudinal studies, the penalty can be applied to marginal models. For different data situations, further generalizations arise. For example, it may be eligible that the categories of different effects and interactions are fused in a specific order.

The proposed penalization techniques for categorical effect modifiers can be extended to hierarchical settings. As seen in Section 4.2.4, there is an ongoing discussion on the choice between group-specific models – more generally, fixed effects models – and random effects models for hierarchical data. Group-specific models may suffer from their relatively large number of parameters, whereas the random effects models are criticized for their relatively strong assumptions. Combining group-specific models with the theory of Chapter 3 allows to reduce the model complexity. Like random effects models, the penalized group-specific model can be seen as a compromise between the unpenalized group-specific model and the naive model that does not account for the heterogeneity in the data. However, in contrast to the random effects model, less assumptions are required. Especially when the random effects and the covariates are not independent, the group-specific model has substantial advantages. Moreover, with the fused Lasso-type penalty, the penalized group-specific model is able to detect clusters of second level units. The fused Lasso-type penalty can result in different partitions of the second level units for different explanatory variables. In situations where this is not desirable, the proposed approach can be generalized: With a group Lasso penalty, one obtains consistent partitions (see Section 4.6). A disadvantage of group-specific models is that the explanatory variables have to vary across the second level units. Group-constant variables as, for example, characteristics of the hospital in the data set on the mortality after myocardial infarction (Section 4.5), cannot be considered. However, as group-specific models with constant explanatory effects can be estimated by penalized approaches, this does not hold for the approach advocated in Chapter 4. In future work, it has to be examined whether or not penalized group-specific models with constant explanatory variables can compete with random effects models.

Lasso-type penalties on the differences of coefficients reach their limit in specific orthonormal settings. In Section 5.2, it is shown that the clustering performance of the fused Lasso can be poor. As an alternative, L_0-type penalties for categorical effects and categorical effect modifiers are considered. The so called L_0 "norm" is an indicator for non-zero arguments. Applied to a vector, it counts the number of entries that are unequal to zero. That is, in Chapter 5, the term "discrete structure" is amended by a another aspect: If the evaluations of the L_0 norm are considered as discrete, equally spaced points, a discrete penalty is proposed. As discussed in Section 5.3.3, the L_0 penalty has a close connection to model selection based on information criteria like the Akaike information criterion (AIC; see, for example, Bozdogan, 1987) or the Bayesian information criterion (BIC; Schwarz, 1978). But, the computational approach differs. Model selection procedures that are based on information criteria compare all unpenalized models with all possible subsets of covariates. In contrast, the L_0 approach does not need subsets and optimizes the penalized objective directly – whereat the penalty is approximated by a continuous function. As there is no guarantee that the proposed approximation finds the global optimum of the objective, the comparison of the proposed computational approach with competing methods should be extended systematically in future work. For example, Dicker et al. (2013) consider L_0-type penalties for continuous covariates. In contrast to Chapter 5, they employ a coordinate descent algorithm. They are able to prove that the proposed method is consistent and asymptotically normal. It should be investigated whether or not these properties can be transferred to the methods that are proposed in Chapter 5.

Throughout the thesis, singular or non-continuous penalties are employed which are not differentiable at some points and thus, computationally challenging. Therefore, in Chapter 2, a general family of penalties for GLMs is defined that allows nevertheless for reasonable estimation procedures for the proposed approaches. The family is characterized by the components of a typical penalty: (i) a linear transformation of the coefficients, (ii) a (semi-) norm, and (iii) an additional function – such that the norm can be amended by weights or be rearranged as it is, for example, required for the smoothly clipped absolute deviation penalty (SCAD; Fan and Li, 2001). The approach approximates non-differentiable penalty terms by local quadratic approximations such that a penalized iteratively re-weighted least squares algorithm can be derived. The approach is very general. This is both, a blessing and a curse. On the one hand, it is less efficient than other algorithms that exploit the structure of a concrete problem as, for example, the least angle regression (lars; Efron et al., 2004). To obtain coefficient paths or in order to determine the optimal value of the penalty parameter, the model has to be estimated multiple times. If there are several penalty parameters, multidimensional cross-validation is required. This drawback can be compensated: In order to obtain efficient estimates for the penalty parameters, one could exploit the close connection of penalized likelihood approaches and random effects models. On the other hand, the general design of the penalty allows to plug in different approximations for the same norm. It is possible to apply penalties depending on different norms in the same model. The approach is capable of vector-valued arguments as, for example, needed for the group Lasso (Yuan and Lin, 2006). An extension to cumulative logit models, or more precisely to Bradley-Terry-type models, has been proposed by Tutz and Schauberger (2014).

The proposed approximations are not restricted to penalty terms. In Chapter 6, an approximation of the “discrete” L_0 norm is employed as a “discrete” loss function. Regression models that approximate the conditional mode of a response are considered. This has been done before. For example, Kemp and Santos Silva (2012) approximate the according loss function by different kernels. They are able to prove the asymptotic properties of their approach. However, the loss function is challenging. For an increasing number of parameters, the approach of Kemp and Santos Silva (2012) reaches its limits. In order to stabilize the estimation, in Section 6.2, a nested interval approach (NILS) is proposed. The tuning parameters of the approximation are adjusted with the iterations of the derived iteratively re-weighted least squares algorithm. The adjustment is data driven. It is shown that the obtained estimator is consistent and asymptotically normal – just as the estimator of Kemp and Santos Silva (2012). The novelty is that the approach is relatively stable and that it allows for semiparametric mode regression with quadratic penalties. The approach can be combined with the well known `R` package `mgcv` (R Core Team, 2014; Wood, 2011) – such that the range of possible semiparametric predictors is large. For example, a Markov random field is employed for the analysis of the rents in Munich (see Section 6.5). There are Bayesian approaches to mode regression (Yu and Aristodemou, 2012). Future work should examine the similarities and the differences of the NILS approach and this competing Bayesian approach. Of course, further work on the data adaptive tuning of the NILS approach could be conducted.

Overall, penalized regression for discrete structures is a broad topic. There are numerous interactions with related subjects such as conditional mode regression. This thesis gives an insight into a few aspect of this wide field. I really hope that the thesis is a contribution to the questions that are still to be answered.

Appendices

A. Proofs for Chapter 3

Theorem 1. *Suppose $0 \leq \lambda < \infty$ has been fixed, and all class-wise sample sizes n_r satisfy $n_{jr}/n \to c_{jr}$, where $0 < c_{jr} < 1$. Then the estimate $\hat{\boldsymbol{\beta}}$ that minimizes (3.2) with $J_n(\boldsymbol{\beta})$ defined by (3.3), (3.4) and (3.5) is consistent, that is, $\lim_{n\to\infty} \mathbb{P}(||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||^2 > \epsilon) = 0$ for all $\epsilon > 0$.*

Proof. If $\hat{\boldsymbol{\beta}}$ minimizes $\mathcal{M}_n^{pen}(\boldsymbol{\beta})$ with $J_n(\boldsymbol{\beta})$ as defined by $J_n(\boldsymbol{\beta})$, with $J_j^{nom}(\beta_j)$ and $J_j^{ord}(\beta_j)$, then it also minimizes $\mathcal{M}_n^{pen}(\boldsymbol{\beta})/n$. The ML estimate $\hat{\boldsymbol{\beta}}^{ML}$ minimizes $\mathcal{M}_n(\boldsymbol{\beta}) = -l_n(\boldsymbol{\beta})$, respectively $\mathcal{M}_n(\boldsymbol{\beta})/n$. Since λ is fixed, $\mathcal{M}_n^{pen}(\hat{\boldsymbol{\beta}})/n \xrightarrow{\mathbb{P}} \mathcal{M}_n(\hat{\boldsymbol{\beta}}^{ML})/n$ and $\mathcal{M}_n^{pen}(\hat{\boldsymbol{\beta}})/n \xrightarrow{\mathbb{P}} \mathcal{M}_n(\hat{\boldsymbol{\beta}})/n$, $\mathcal{M}_n(\hat{\boldsymbol{\beta}})/n \xrightarrow{\mathbb{P}} \mathcal{M}_n(\hat{\boldsymbol{\beta}}^{ML})/n$ hold as well. Since $\hat{\boldsymbol{\beta}}^{ML}$ is the unique minimizer of $\mathcal{M}_n(\boldsymbol{\beta})/n$, and $\mathcal{M}_n(\boldsymbol{\beta})/n$ is convex, we have $\hat{\boldsymbol{\beta}} \xrightarrow{\mathbb{P}} \hat{\boldsymbol{\beta}}^{ML}$; and consistency follows from consistency of the ML estimate $\hat{\boldsymbol{\beta}}^{ML}$, under assumptions given, for example, by Fahrmeir and Kaufmann (1985). □

Theorem 2. *Suppose $\lambda = \lambda_n$ with $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n \to \infty$, and all class-wise sample sizes n_{jr} satisfy $n_{jr}/n \to c_{jr}$, where $0 < c_{jr} < 1$. Then penalty $J_n^{ad}(\boldsymbol{\beta})$ employing terms (3.8) and (3.9) with weights (3.10) and (3.11), where $\hat{\beta}_{jr}^{ML}$, $\phi_{rs(j)}(n)$ and $\phi_{r(j)}(n)$ are defined as above, ensures that*

(a) $\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathcal{C}}^n - \boldsymbol{\theta}_{\mathcal{C}}^*) \xrightarrow{d} N(\mathbf{0}, \mathrm{Cov}(\boldsymbol{\theta}_{\mathcal{C}}^*))$,

(b) $\lim_{n\to\infty} \mathbb{P}(\mathcal{C}_n = \mathcal{C}) = 1$.

Proof. A proof with a similar line of argumentation can be found in Gertheiss and Tutz (2012), where the asymptotic properties for categorical effect modifiers in linear models are considered. In order to prove the same asymptotic properties in a general likelihood context, some arguments differ. To prove asymptotic normality, these arguments are the redefinition of the objective function, the limit behavior of truly zero coefficients and the arguments needed for consistency. To prove consistency, again, different arguments related to the consistency of ML estimates are needed.
Due to the additivity of arguments, a predictor of the following form can be assumed without loss of generality:

$$\eta_i = \beta_0(u) + x_1\beta_1(u) + \ldots + x_p\beta_p(u).$$

That is, only one effect modifier u is assumed. In addition, let $\boldsymbol{Z}$ denote the design matrix given by $\boldsymbol{Z} = (\boldsymbol{Z}_0, \dots, \boldsymbol{Z}_p)$, where

$$\boldsymbol{Z}_j = \begin{pmatrix} x_{1j}I(u_{1j} = 1) & \cdots & x_{1j}I(u_{1j} = k_j) \\ \vdots & \ddots & \vdots \\ x_{nj}I(u_{nj} = 1) & \cdots & x_{nj}I(u_{nj} = k_j) \end{pmatrix}.$$

(a) Normality

- *Redefinition of the Objective Function* Redefine the optimization problem $\mathcal{M}_n^{pen}(\boldsymbol{\beta})$ as $\arg\min_{\boldsymbol{\beta}} \ \Psi_n(\boldsymbol{\beta})$, where $\Psi_n(\boldsymbol{\beta}) = -l_n(\boldsymbol{\beta}) + \frac{\lambda_n}{\sqrt{n}} J_n(\boldsymbol{\beta})$. $J_n(\boldsymbol{\beta})$ denotes the penalty term. For the proof, the penalty parameter λ is divided by the factor $\sqrt{n}$. In turn, the penalty $J_n(\boldsymbol{\beta})$ is multiplied by the same factor:

  $$J_n(\boldsymbol{\beta}) = \sqrt{n} \left(\sum_{j=0}^{p} \sum_{r>s} w_{rs(j)} |\beta_{jr} - \beta_{js}| + \sum_{j=1}^{p} \sum_{r=1}^{k} w_{r(j)} |\beta_{jr}| \right).$$

  The log-likelihood is defined as

  $$l_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{y_i \vartheta_i(\mu_i) - b(\vartheta_i(\mu_i))}{\varphi_i} = \sum_{i=1}^{n} \frac{y_i \vartheta_i(h(\boldsymbol{z}_i^T \boldsymbol{\beta})) - b(\vartheta_i(h(\boldsymbol{z}_i^T \boldsymbol{\beta})))}{\varphi_i}.$$

  That is, $l_n(\boldsymbol{\beta})$ is determined by a simple exponential family where $\vartheta_i \in \Theta \subset \mathbb{R}$ is the natural parameter of the family depending on expectation μ_i; φ_i is a scale or dispersion parameter, $b(\cdot)$ and $c(\cdot)$ are specific functions corresponding to the type of the family. For given φ_i, one assumes Θ to be the natural parameter space. That is, the set of all ϑ_i satisfying $0 < \int \exp(y_i \vartheta_i / \varphi_i + c(y_i, \varphi_i)) \mathrm{d}y_i < \infty$. Then, Θ is convex, and in the nonempty interior Θ^0, all derivatives of $b(\vartheta_i)$ and all moments of y_i exist (see Fahrmeir and Tutz, 2001). Hence, it is equivalent to solve

  $$\arg\min_{\boldsymbol{\beta}} \ 2\left(\Psi_n(\boldsymbol{\beta}) - \Psi_n(\boldsymbol{\beta}^*)\right) = \arg\min_{\boldsymbol{\beta}} \ V_n(\boldsymbol{\beta}),$$

  where $\boldsymbol{\beta}^*$ denotes the true coefficient vector and where

  $$\begin{aligned} V_n(\boldsymbol{\beta}) &= -2\left(l_n(\boldsymbol{\beta}) - l_n(\boldsymbol{\beta}^*)\right) + 2\frac{\lambda_n}{\sqrt{n}}\left(J_n(\boldsymbol{\beta}) - J_n(\boldsymbol{\beta}^*)\right) \\ &= -2\left(l_n(\boldsymbol{\beta}) - l_n(\boldsymbol{\beta}^*)\right) + 2\frac{\lambda_n}{\sqrt{n}}\tilde{J}_n(\boldsymbol{\beta}). \end{aligned}$$

- *Limit Behavior* Following Bondell and Reich (2009) closely, $\tilde{J}_n(\boldsymbol{\beta})$ with respect to $\boldsymbol{b}$ is considered, where $\boldsymbol{b} = \sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$:

$$\begin{aligned}
\tilde{J}_n(\boldsymbol{\beta}) &= J_n(\boldsymbol{\beta}) - J_n(\boldsymbol{\beta}^*) \\
\Rightarrow \tilde{J}_n(\boldsymbol{b}) &= J_n(\boldsymbol{b}) - J_n(\mathbf{0}) \\
&= \sum_{j=0}^{p} \sum_{r>s} \sqrt{n} \frac{\phi_{rs(j)}(n)}{|\hat{\beta}_{jr}^{ML} - \hat{\beta}_{js}^{ML}|} \left| \beta_{jr}^* - \beta_{js}^* + \frac{b_{jr} - b_{js}}{\sqrt{n}} \right| \\
&+ \sum_{j=1}^{p} \sum_{r=1}^{k} \sqrt{n} \frac{\phi_{r(j)}(n)}{|\hat{\beta}_{jr}^{ML}|} \left| \beta_{jr}^* + \frac{b_{jr}}{\sqrt{n}} \right| \\
&- \left(\sum_{j=0}^{p} \sum_{r>s} \sqrt{n} \frac{\phi_{rs(j)}(n)}{|\hat{\beta}_{jr}^{ML} - \hat{\beta}_{js}^{ML}|} \left| \beta_{jr}^* - \beta_{js}^* \right| + \sum_{j=1}^{p} \sum_{r=1}^{k} \sqrt{n} \frac{\phi_{r(j)}(n)}{|\hat{\beta}_{jr}^{ML}|} \left| \beta_{jr}^* \right| \right) \\
&= \sum_{j=0}^{p} \sum_{r>s} \sqrt{n} \frac{\phi_{rs(j)}(n)}{|\hat{\beta}_{jr}^{ML} - \hat{\beta}_{js}^{ML}|} \left(\left| \beta_{jr}^* - \beta_{js}^* + \frac{b_{jr} - b_{js}}{\sqrt{n}} \right| - \left| \beta_{jr}^* - \beta_{js}^* \right| \right) \\
&+ \sum_{j=1}^{p} \sum_{r=1}^{k} \sqrt{n} \frac{\phi_{r(j)}(n)}{|\hat{\beta}_{jr}^{ML}|} \left(\left| \beta_{jr}^* + \frac{b_{jr}}{\sqrt{n}} \right| - \left| \beta_{jr}^* \right| \right).
\end{aligned}$$

Distinction of cases:

- **Case 1:** $\beta_{jr}^* \neq \beta_{js}^*$ or $\beta_{jr}^* \neq 0$, that is, if $\theta_i^* \neq 0$.
As given in Zou (2006), we consider the limit behavior of $(\lambda_n/\sqrt{n})\tilde{J}_n(\boldsymbol{b})$.
If $\beta_{jr}^* \neq \beta_{js}^*$, then

$$|\hat{\beta}_{jr}^{ML} - \hat{\beta}_{js}^{ML}| \xrightarrow{\mathbb{P}} |\beta_{jr}^* - \beta_{js}^*|,$$

and

$$\sqrt{n} \left(\left| \beta_{jr}^* - \beta_{js}^* + \frac{b_{jr} - b_{js}}{\sqrt{n}} \right| - \left| \beta_{jr}^* - \beta_{js}^* \right| \right) \to (b_{jr} - b_{js}) \mathrm{sgn}(\beta_{jr}^* - \beta_{js}^*).$$

Similarly, if $\beta_{jr}^* \neq 0$, then

$$|\hat{\beta}_{jr}^{ML}| \xrightarrow{\mathbb{P}} |\beta_{jr}^*|,$$

and

$$\sqrt{n} \left(\left| \beta_{jr}^* + \frac{b_{jr}}{\sqrt{n}} \right| - \left| \beta_{jr}^* \right| \right) \longrightarrow b_{jr} \mathrm{sgn}(\beta_{jr}^*).$$

Since by assumption $\phi_{rs(j)}(n) \to q_{rs(j)}$ and $\phi_{r(j)}(n) \to q_{r(j)}$ $(0 < q_{rs(j)}, q_{r(j)} < \infty)$ and $\lambda_n/\sqrt{n} \to 0$, by Slutsky's theorem, we have

$$\frac{\lambda_n}{\sqrt{n}} \sqrt{n} \frac{\phi_{rs(j)}(n)}{|\hat{\beta}_{jr}^{ML} - \hat{\beta}_{js}^{ML}|} \left(\left| \beta_{jr}^* - \beta_{js}^* + \frac{b_{jr} - b_{js}}{\sqrt{n}} \right| - \left| \beta_{jr}^* - \beta_{js}^* \right| \right) \xrightarrow{\mathbb{P}} 0,$$

and

$$\frac{\lambda_n}{\sqrt{n}}\sqrt{n}\frac{\phi_{r(j)}(n)}{|\hat{\beta}_{jr}^{ML}|}\left(\left|\beta_{jr}^* + \frac{b_{jr}}{\sqrt{n}}\right| - |\beta_{jr}^*|\right) \xrightarrow{\mathbb{P}} 0,$$

respectively. That means, if $\theta_i^* \neq 0$, we have $\frac{\lambda_n}{\sqrt{n}}\tilde{J}(\boldsymbol{b}) \xrightarrow{\mathbb{P}} 0$.

- **Case 2:** $\beta_{jr}^* = \beta_{js}^*$ or $\beta_{jr}^* = 0$, that is, if $\theta_i^* = 0$.
Here, it holds that

$$\sqrt{n}\left(\left|\beta_{jr}^* - \beta_{js}^* + \frac{b_{jr} - b_{js}}{\sqrt{n}}\right| - \left|\beta_{jr}^* - \beta_{js}^*\right|\right) = |b_{jr} - b_{js}|,$$

and

$$\sqrt{n}\left(\left|\beta_{jr}^* + \frac{b_{jr}}{\sqrt{n}}\right| - \left|\beta_{jr}^*\right|\right) = |b_{jr}|.$$

Moreover, due to the consistency of the ML estimates, we have

$$\hat{\boldsymbol{\beta}}^{ML} - \boldsymbol{\beta}^* = \boldsymbol{F}_n^{-1}(\boldsymbol{\beta}^*)\boldsymbol{s}_n(\boldsymbol{\beta}^*) + \mathcal{O}_p(n^{-1}),$$

where the expected information matrix is denoted by $\boldsymbol{F}_n(\boldsymbol{\beta}) = \mathbb{E}\left(-\frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right)$ and where the score function is defined as $\boldsymbol{s}_n(\boldsymbol{\beta}) = \frac{\partial l_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}$. $\boldsymbol{x}_n = \mathcal{O}_p(r_n) \Leftrightarrow \mathbb{P}(\|\boldsymbol{x}_n\|/r_n < C) \geq 1 - \epsilon$, $n \geq N$, denotes the Landau notation for $\epsilon > 0$, some constant C and a sufficiently large N. Therewith, $\hat{\boldsymbol{\beta}}^{ML} - \boldsymbol{\beta}^* = \mathcal{O}_p(n^{-1/2})$, see McCullagh (1983). As a conclusion, it holds that

$$\lim_{n\to\infty}\mathbb{P}\left(\sqrt{n}|\hat{\beta}_{jr}^{ML} - \hat{\beta}_{js}^{ML}| \leq \lambda_n^{1/2}\right) = 1,$$

or

$$\lim_{n\to\infty}\mathbb{P}\left(\sqrt{n}|\hat{\beta}_{jr}^{ML}| \leq \lambda_n^{1/2}\right) = 1,$$

respectively, since $\lambda_n \to \infty$ by assumption. Hence,

$$\frac{\lambda_n}{\sqrt{n}}\sqrt{n}\frac{\phi_{rs(j)}(n)}{|\hat{\beta}_{jr}^{ML} - \hat{\beta}_{js}^{ML}|}\left(\left|\beta_{jr}^* - \beta_{js}^* + \frac{b_{jr} - b_{js}}{\sqrt{n}}\right| - \left|\beta_{jr}^* - \beta_{js}^*\right|\right) \xrightarrow{\mathbb{P}} \infty,$$

or

$$\frac{\lambda_n}{\sqrt{n}}\sqrt{n}\frac{\phi_{r(j)}(n)}{|\hat{\beta}_{jr}^{ML}|}\left(\left|\beta_{jr}^* + \frac{b_{jr}}{\sqrt{n}}\right| - \left|\beta_{jr}^*\right|\right) \xrightarrow{\mathbb{P}} \infty,$$

if $b_{jr}^* \neq 0$, respectively $b_{jr}^* \neq b_{js}^*$. That means, if for any r, s, j with $\beta_{jr}^* = 0$ $(j > 0)$ or $\beta_{jr}^* = \beta_{js}^*$ $(j \geq 0)$, $b_{jr} \neq 0$ or $b_{jr} \neq b_{js}$, respectively, then we have $\frac{\lambda_n}{\sqrt{n}}\tilde{J}(\boldsymbol{b}) \xrightarrow{\mathbb{P}} \infty$.

- *Normality* Following Bondell and Reich (2009), let $\boldsymbol{\theta}_{\mathcal{C}}$ denote the vector of $\boldsymbol{\theta}$ entries which are truly non zero, that is, from $\mathcal{C}$. And let $\boldsymbol{\beta}_{\mathcal{C}}$ be the subset of entries of $\boldsymbol{\theta}_{\mathcal{C}}$ which are part of $\boldsymbol{\beta}$. By contrast, $\boldsymbol{\theta}_{\mathcal{C}^c}$ denotes the vector of $\boldsymbol{\theta}$ entries which are truly zero and therefore not from $\mathcal{C}$ but from $\mathcal{C}^c$. Analogously to $\boldsymbol{\beta}_{\mathcal{C}}$, $\boldsymbol{\beta}_{\mathcal{C}^c}$ is defined as the subset of entries of $\boldsymbol{\theta}_{\mathcal{C}^c}$ which are part of $\boldsymbol{\beta}$.

  Let now $l_n(\boldsymbol{\beta}_{\mathcal{C}})$ denote the likelihood of the oracle model, that is, the model where truly zero $\boldsymbol{\beta}$ entries are fixed to zero. Analogously to usual ML theory, an expansion of the (oracle) ML equations $\boldsymbol{s}_n(\boldsymbol{\beta}_{\mathcal{C}}^{ML}) = \mathbf{0}$ about $\boldsymbol{\beta}_{\mathcal{C}}^*$ gives

$$\boldsymbol{s}_n(\boldsymbol{\beta}_{\mathcal{C}}^*) = \left.\frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_{\mathcal{C}}^*} (\boldsymbol{\beta}_{\mathcal{C}}^{ML} - \boldsymbol{\beta}_{\mathcal{C}}^*) + \mathcal{O}_p(n^{-1}),$$

  with $\boldsymbol{\beta}_{\mathcal{C}}^{ML}$ denoting the ML estimate of the oracle model, and $\boldsymbol{\beta}_{\mathcal{C}}^*$ being the vector of corresponding true $\boldsymbol{\beta}$ coefficients. Hence, it holds that

$$\boldsymbol{\beta}_{\mathcal{C}}^{ML} - \boldsymbol{\beta}_{\mathcal{C}}^* = \left.\frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_{\mathcal{C}}^*} \boldsymbol{s}_n(\boldsymbol{\beta}_{\mathcal{C}}^*) = \boldsymbol{F}_n^{-1}(\boldsymbol{\beta}_{\mathcal{C}}^*)\boldsymbol{s}_n(\boldsymbol{\beta}_{\mathcal{C}}^*) + \mathcal{O}_p(n^{-1}).$$

  Multiplying both sides by $n^{1/2}$, using $\boldsymbol{F}_n(\boldsymbol{\beta}_{\mathcal{C}}^*)/n \overset{n\to\infty}{\to} \boldsymbol{F}(\boldsymbol{\beta}_{\mathcal{C}}^*)$ and $n^{-1/2}\boldsymbol{s}_n(\boldsymbol{\beta}_{\mathcal{C}}^*) \overset{d}{\to} N(\mathbf{0}, \boldsymbol{F}(\boldsymbol{\beta}_{\mathcal{C}}^*))$, one obtains

$$n^{1/2}(\boldsymbol{\beta}_{\mathcal{C}}^{ML} - \boldsymbol{\beta}_{\mathcal{C}}^*) \overset{d}{\to} N(\mathbf{0}, \boldsymbol{F}(\boldsymbol{\beta}_{\mathcal{C}}^*)^{-1}),$$

  as in usual generalized linear models (McCullagh, 1983).

  Back to the given varying-coefficient model, with $n \to \infty$, and employing the redefined objective, the limit behavior and Slutsky's theorem, we have $V_n(\boldsymbol{\beta}) \overset{d}{\to} V(\boldsymbol{\beta})$ for each $\boldsymbol{\beta}$, where

$$V(\boldsymbol{\beta}) = \begin{cases} -2\left(l_n(\boldsymbol{\beta}_{\mathcal{C}}) - l_n(\boldsymbol{\beta}_{\mathcal{C}}^*)\right) & \text{if } \boldsymbol{\theta}_{\mathcal{C}^c} = 0, \\ \infty & \text{otherwise.} \end{cases}$$

  Since $V_n(\boldsymbol{\beta})$ is convex, and the unique minimizer of $V(\boldsymbol{\beta})$ is $(\boldsymbol{\beta}_{\mathcal{C}}^{ML}, \mathbf{0})^T$, we have

$$\hat{\boldsymbol{\beta}}_{\mathcal{C}} \overset{d}{\to} \boldsymbol{\beta}_{\mathcal{C}}^{ML},$$

  (see Zou, 2006; Bondell and Reich, 2009), and eventually

$$n^{1/2}(\hat{\boldsymbol{\beta}}_{\mathcal{C}} - \boldsymbol{\beta}_{\mathcal{C}}^*) \overset{d}{\to} N(\mathbf{0}, \boldsymbol{F}(\boldsymbol{\beta}_{\mathcal{C}}^*)^{-1}).$$

  Via a reparametrization of $\boldsymbol{\beta}$ as, for example, $\check{\boldsymbol{\beta}} = (\check{\boldsymbol{\beta}}_0^T, ..., \check{\boldsymbol{\beta}}_p^T)^T$, with $\check{\boldsymbol{\beta}}_j = (\beta_{jr} - \beta_{j1}, ..., \beta_{jr}, ..., \beta_{jr} - \beta_{jk})^T$, that is, changing the subset of entries of $\boldsymbol{\theta}$ which are part of $\boldsymbol{\beta}$, asymptotic normality can be proved for all entries of $\boldsymbol{\theta}_{\mathcal{C}}$.

(b) Consistency: $\lim_{n\to\infty} \mathbb{P}(\mathcal{C}_n = \mathcal{C}) = 1$

To show consistency, it has to be shown that $\lim_{n\to\infty} \mathbb{P}(\mathcal{J} \in \mathcal{C}_n) = 1$ if $\mathcal{J} \in \mathcal{C}$ and that $\lim_{n\to\infty} \mathbb{P}(\mathcal{J} \in \mathcal{C}_n) = 0$ if $\mathcal{J} \notin \mathcal{C}$, where $\mathcal{J}$ denotes a triple of indices (j, s, r) or pair (j, r).

- *Selection of Influential Coefficients:* $\lim_{n\to\infty} \mathbb{P}(\mathcal{J} \in \mathcal{C}_n) = 1$ *if* $\mathcal{J} \in \mathcal{C}$
  Follows directly from part (a) of the proof.

- *Exclusion of Non-Influential Coefficients:* $\lim_{n\to\infty} \mathbb{P}(\mathcal{J} \in \mathcal{C}_n) = 0$ *if* $\mathcal{J} \notin \mathcal{C}$
  A similar proof is found in Bondell and Reich (2009). Let $\mathcal{B}_n$ denote the (nonempty) set of indices $\mathcal{J}$ which are in $\mathcal{C}_n$ but not in $\mathcal{C}$. Without loss of generality, we assume that the largest $\hat{\boldsymbol{\theta}}$ entry corresponding to indices from $\mathcal{B}_n$ is $\hat{\beta}_{lq} > 0$, $l \geq 0$. If a certain difference $\hat{\beta}_{lr} - \hat{\beta}_{ls}$ is the largest $\hat{\boldsymbol{\theta}}$ entry included in $\mathcal{B}_n$, we just need to reparameterize $\boldsymbol{\beta_l}$ in an adequate way by $\breve{\boldsymbol{\beta}}_l$ as given in part (a) of the proof. Since all coefficients and differences thereof are penalized in the same way, this can be done without any problems. Moreover, we may order the categories such that $\hat{\beta}_{l1} \leq \ldots \leq \hat{\beta}_{lz} \leq 0 \leq \hat{\beta}_{l,z+1} \leq \ldots \leq \hat{\beta}_{lk}$. Defining the set $\mathcal{B}_n$ and ordering the coefficients by size, allows for a proof by contradiction.
  Estimating $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}\ \Psi(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}}\ -l_n(\boldsymbol{\beta}) + \frac{\lambda_n}{\sqrt{n}} J_n(\boldsymbol{\beta})$ like defined in part (a) of the proof is equivalent to

$$\arg\min_{\mathfrak{B}}\ -l_n(\boldsymbol{\beta}) + \lambda_n \sum_j J_j(\boldsymbol{\beta}),$$

with

$$\begin{aligned}
\mathfrak{B} &= \{\boldsymbol{\beta} : \beta_{0,1}, \ldots, \beta_{l-1,k}, \beta_{l,1} \leq \ldots \\
&\qquad \leq \beta_{l,z} \leq 0 \leq \beta_{l,z+1} \leq \ldots \leq \beta_{l,k}, \beta_{l+1,1}, \ldots, \beta_{p,k}\}, \\
J_j(\boldsymbol{\beta}) &= \sum_{r>s} \frac{\phi_{rs(j)}(n)}{|\hat{\beta}_{jr}^{ML} - \hat{\beta}_{js}^{ML}|} \, |\beta_{jr} - \beta_{js}| + I(j \neq 0) \sum_{r=1}^{k} \frac{\phi_{r(j)}(n)}{|\hat{\beta}_{jr}^{ML}|} \, |\beta_{jr}|\,,\ j \neq l, \\
J_l(\boldsymbol{\beta}) &= \sum_{r>s} \frac{\phi_{rs(l)}(n)}{|\hat{\beta}_{lr}^{ML} - \hat{\beta}_{ls}^{ML}|} \, (\beta_{lr} - \beta_{ls}) + \sum_{r \geq z+1} \frac{\phi_{r(l)}(n)}{|\hat{\beta}_{lr}^{ML}|} \, (\beta_{lr}) \\
&\quad - \sum_{r \leq z} \frac{\phi_{r(l)}(n)}{|\hat{\beta}_{lr}^{ML}|} \, (\beta_{lr})\,.
\end{aligned}$$

Since $\hat{\beta}_{lq} \neq 0$ is assumed, at the solution $\hat{\boldsymbol{\beta}}$, this optimization criterion is differentiable with respect to β_{lq}. We may consider this derivative in a neighborhood of the solution where coefficients which are set equal/to zero remain equal/zero. That means, terms corresponding to pairs/triples of indices which are not in $\mathcal{C}_n$ can be omitted, since they will vanish in $J(\hat{\boldsymbol{\beta}}) = \sum_j J_j(\hat{\boldsymbol{\beta}})$. If $x_{(l)q}$ denotes the column of design matrix $\boldsymbol{Z}$ which belongs to β_{lq}, due to differentiability, the estimate $\hat{\boldsymbol{\beta}}$ must satisfy

$$\frac{\boldsymbol{s}_n(\boldsymbol{\beta})}{\sqrt{n}} = \boldsymbol{A}_n + \boldsymbol{D}_n,$$

with

$$\boldsymbol{A}_n = \frac{\lambda_n}{\sqrt{n}} \left(\sum_{s<q;(l,q,s)\in\mathcal{C}} \frac{\phi_{qs(l)}(n)}{|\hat{\beta}^{ML}_{lq} - \hat{\beta}^{ML}_{ls}|} - \sum_{r>q,(l,r,q)\in\mathcal{C}} \frac{\phi_{rq(l)}(n)}{|\hat{\beta}^{ML}_{lr} - \hat{\beta}^{ML}_{lq}|} \right) \text{and}$$

$$\boldsymbol{D}_n = \frac{\lambda_n}{\sqrt{n}} \left(\sum_{s<q;(l,q,s)\in\mathcal{B}_n} \frac{\phi_{qs(l)}(n)}{|\hat{\beta}^{ML}_{lq} - \hat{\beta}^{ML}_{ls}|} + \frac{\phi_{q(l)}(n)}{|\hat{\beta}^{ML}_{lq}|} \right).$$

From part (a) of the proof, we know that $n^{-1/2}\boldsymbol{s}_n(\boldsymbol{\beta}) \overset{d}{\to} N(\boldsymbol{0}, \boldsymbol{F}(\boldsymbol{\beta}))$, where $\boldsymbol{F}_n(\boldsymbol{\beta})/n \leftrightarrow \boldsymbol{F}(\boldsymbol{\beta})$. Hence, for any $\epsilon > 0$, we have

$$\lim_{n\to\infty} \mathbb{P}(\frac{\boldsymbol{s}_n(\boldsymbol{\beta})}{\sqrt{n}} \leq \lambda_n^{1/4} - \epsilon) = 1.$$

Since $\lambda_n/\sqrt{n} \to 0$, we also know $\exists\, \epsilon > 0$ such that $\lim_{n\to\infty} \mathbb{P}(|\boldsymbol{A}_n| < \epsilon) = 1$. By assumption $\lambda_n \to \infty$. Due to consistency of the ordinary ML estimate, we know that

$$\lim_{n\to\infty} \mathbb{P}(\sqrt{n}|\hat{\boldsymbol{\beta}}^{ML}_{lq}| \leq \lambda_n^{1/2}) = 1,$$

if $(l, q) \in \mathcal{B}_n$. Hence,

$$\lim_{n\to\infty} \mathbb{P}(\boldsymbol{D}_n \geq \lambda_n^{1/4}) = 1.$$

As a consequence,

$$\lim_{n\to\infty} \mathbb{P}(\frac{\boldsymbol{s}_n(\boldsymbol{\beta})}{\sqrt{n}} = \boldsymbol{A}_n + \boldsymbol{D}_n) = 0.$$

That means if $\mathcal{J} \notin \mathcal{C}$, it also holds that

$$\lim_{n\to\infty} \mathbb{P}(\mathcal{J} \in \mathcal{C}) = 0.$$

□

B. Numerical Results for Chapter 4

Appendix B presents the detailed results of the numerical experiments conducted for Section 4.4. Tables B.1–B.4 show the results for Gaussian responses. Tables B.5–B.8 the results for binomial responses. The methods are labeled as described on page 62. If there is an additional "R", the penalty parameter is chosen with an additional refit in the cross-validation procedure.

		$\rho = 0.0$				$\rho = 0.8$			
K	Method	Intercepts	Slope	FP	FN	Intercepts	Slope	FP	FN
30	G	100.59	0.08		0.00	107.23	0.21		0.00
	GL1, CV	75.03	0.06		0.06	130.65	2.58		0.73
	GL1, GCV	70.21	0.06		0.04	96.00	1.09		0.07
	GL1a, CV	81.42	0.07		0.12	87.83	0.40		0.22
	GL1a, GCV	81.22	0.07		0.11	83.03	0.37		0.14
	GL1, CV, R	115.57	0.08		0.39	132.14	2.69		0.91
	GL1, GCV, R	119.71	0.06		0.27	123.08	2.46		0.41
	GL1a, CV, R	81.10	0.06		0.23	81.98	0.97		0.40
	GL1a, GCV, R	99.54	0.06		0.25	95.49	1.49		0.29
	R	71.04	0.06		0.00	124.92	2.44		0.00
	Finite AIC	124.80	0.07		0.60	132.84	2.66		0.95
	Finite BIC	169.01	0.07		0.81	132.11	2.77		1.00
15	G	100.60	0.08	1.00	0.00	106.59	0.21	1.00	0.00
	GL1, CV	63.19	0.06	0.52	0.47	103.97	1.86	0.25	0.74
	GL1, GCV	51.74	0.07	0.93	0.05	84.38	0.86	0.92	0.07
	GL1a, CV	69.42	0.07	0.79	0.17	81.99	0.35	0.74	0.21
	GL1a, GCV	70.21	0.07	0.83	0.13	77.95	0.31	0.83	0.14
	GL1, CV, R	67.64	0.07	0.14	0.85	104.51	2.15	0.02	0.98
	GL1, GCV, R	57.68	0.06	0.57	0.36	99.80	1.90	0.55	0.42
	GL1a, CV, R	56.22	0.06	0.61	0.35	77.14	0.87	0.55	0.39
	GL1a, GCV, R	56.86	0.06	0.62	0.31	81.58	1.08	0.61	0.30
	R	48.51	0.07	1.00	0.00	101.27	1.91	1.00	0.00
	Finite AIC	71.43	0.07	0.11	0.84	105.97	2.08	0.04	0.95
	Finite BIC	67.90	0.06	0.02	0.98	105.06	2.16	0.00	1.00
5	G	100.60	0.08	1.00	0.00	106.70	0.21	1.00	0.00
	GL1, CV	68.73	0.07	0.70	0.27	113.98	2.00	0.27	0.73
	GL1, GCV	61.54	0.07	0.94	0.04	88.91	0.98	0.93	0.06
	GL1a, CV	75.39	0.07	0.80	0.13	83.09	0.35	0.79	0.15
	GL1a, GCV	76.42	0.07	0.84	0.11	82.85	0.31	0.82	0.12
	GL1, CV, R	99.82	0.07	0.28	0.71	114.76	2.15	0.10	0.90
	GL1, GCV, R	84.30	0.06	0.63	0.27	107.10	2.02	0.61	0.36
	GL1a, CV, R	69.10	0.06	0.64	0.27	84.10	0.94	0.56	0.37
	GL1a, GCV, R	79.37	0.06	0.61	0.27	90.09	1.33	0.61	0.27
	R	58.95	0.07	1.00	0.00	106.76	1.87	1.00	0.00
	Finite AIC	99.80	0.06	0.19	0.68	116.29	2.02	0.05	0.93
	Finite BIC	100.96	0.06	0.04	0.92	115.35	2.34	0.00	1.00

Table B.1.: Results for the settings with Gaussian response, $b_{i0} \sim N(0, 4)$, $n_i = 10$.

		$\rho = 0.0$				$\rho = 0.8$			
K	Method	Intercepts	Slope	FP	FN	Intercepts	Slope	FP	FN
30	G	100.59	0.08		0.00	107.58	0.21		0.00
	GL1, CV	63.26	0.07		0.13	152.42	2.69		0.60
	GL1, GCV	61.28	0.06		0.04	104.45	1.10		0.06
	GL1a, CV	70.25	0.07		0.17	84.49	0.40		0.15
	GL1a, GCV	71.23	0.07		0.13	84.18	0.31		0.12
	GL1, CV, R	96.71	0.07		0.56	153.56	3.12		0.87
	GL1, GCV, R	81.78	0.06		0.32	141.95	2.86		0.39
	GL1a, CV, R	64.20	0.07		0.32	94.60	1.11		0.36
	GL1a, GCV, R	67.74	0.06		0.29	103.27	1.56		0.29
	R	67.49	0.07		0.00	142.41	2.85		0.00
	Finite AIC	95.60	0.07		0.71	154.66	3.09		0.93
	Finite BIC	141.13	0.07		0.87	154.41	3.23		1.00
15	G	100.59	0.08	1.00	0.00	107.47	0.21	1.00	0.00
	GL1, CV	59.45	0.07	0.73	0.25	182.86	3.28	0.47	0.52
	GL1, GCV	57.94	0.06	0.94	0.05	114.49	1.45	0.94	0.05
	GL1a, CV	70.89	0.07	0.81	0.15	86.28	0.31	0.79	0.15
	GL1a, GCV	72.59	0.06	0.84	0.13	86.71	0.29	0.82	0.12
	GL1, CV, R	99.71	0.07	0.25	0.73	210.42	4.02	0.19	0.80
	GL1, GCV, R	75.40	0.07	0.58	0.32	181.28	3.46	0.65	0.31
	GL1a, CV, R	62.93	0.06	0.61	0.31	91.39	0.99	0.59	0.33
	GL1a, GCV, R	65.92	0.05	0.63	0.27	104.79	1.45	0.62	0.27
	R	59.48	0.07	1.00	0.00	172.74	3.54	1.00	0.00
	Finite AIC	98.97	0.07	0.15	0.75	211.47	4.03	0.07	0.90
	Finite BIC	100.45	0.07	0.04	0.93	211.69	4.48	0.00	1.00
5	G	100.59	0.08	1.00	0.00	106.72	0.21	1.00	0.00
	GL1, CV	63.66	0.07	0.75	0.23	81.90	0.76	0.39	0.61
	GL1, GCV	57.58	0.06	0.93	0.04	66.92	0.44	0.93	0.06
	GL1a, CV	70.31	0.07	0.79	0.15	74.80	0.20	0.78	0.18
	GL1a, GCV	73.07	0.06	0.83	0.11	77.55	0.16	0.83	0.12
	GL1, CV, R	100.62	0.07	0.29	0.69	82.75	0.92	0.08	0.92
	GL1, GCV, R	77.21	0.06	0.61	0.28	72.18	0.86	0.61	0.35
	GL1a, CV, R	62.05	0.07	0.63	0.28	63.62	0.43	0.59	0.35
	GL1a, GCV, R	67.88	0.06	0.61	0.28	68.55	0.58	0.62	0.29
	R	59.78	0.07	1.00	0.00	69.25	0.70	1.00	0.00
	Finite AIC	100.53	0.07	0.16	0.70	85.09	0.89	0.09	0.88
	Finite BIC	101.74	0.07	0.04	0.91	83.34	0.98	0.00	1.00

Table B.2.: Results for the settings with Gaussian response, $b_{i0} \sim \chi^2_3$, $n_i = 10$.

		$\rho = 0.0$				$\rho = 0.8$			
K	Method	Intercepts	Slope	FP	FN	Intercepts	Slope	FP	FN
30	G	211.68	0.14		0.00	227.37	0.38		0.00
	GL1, CV	116.26	0.10		0.41	123.60	2.21		0.80
	GL1, GCV	110.70	0.11		0.05	138.87	1.16		0.07
	GL1a, CV	140.18	0.11		0.18	135.57	0.50		0.23
	GL1a, GCV	150.01	0.11		0.13	149.85	0.40		0.14
	GL1, CV, R	126.61	0.11		0.90	123.04	2.40		0.99
	GL1, GCV, R	104.03	0.11		0.35	122.90	2.12		0.41
	GL1a, CV, R	114.44	0.12		0.38	119.18	0.91		0.43
	GL1a, GCV, R	111.48	0.10		0.31	116.53	1.31		0.34
	R	96.58	0.11		0.00	122.92	2.27		0.00
	Finite AIC	135.56	0.12		0.85	125.04	2.35		0.96
	Finite BIC	127.63	0.11		0.99	123.33	2.46		1.00
15	G	211.68	0.14	1.00	0.00	227.55	0.38	1.00	0.00
	GL1, CV	125.50	0.12	0.59	0.40	121.54	2.13	0.19	0.81
	GL1, GCV	116.01	0.12	0.94	0.05	138.83	1.18	0.95	0.07
	GL1a, CV	145.39	0.12	0.81	0.16	139.17	0.52	0.74	0.24
	GL1a, GCV	149.66	0.12	0.85	0.13	149.10	0.43	0.85	0.14
	GL1, CV, R	139.60	0.12	0.11	0.89	120.62	2.31	0.01	0.99
	GL1, GCV, R	118.63	0.11	0.62	0.33	119.27	2.16	0.60	0.40
	GL1a, CV, R	121.62	0.10	0.66	0.31	118.65	1.16	0.49	0.48
	GL1a, GCV, R	119.03	0.11	0.65	0.29	114.72	1.29	0.63	0.31
	R	99.81	0.13	1.00	0.00	120.46	2.14	1.00	0.00
	Finite AIC	148.00	0.12	0.14	0.81	122.16	2.28	0.03	0.97
	Finite BIC	140.59	0.12	0.01	0.99	120.83	2.33	0.00	1.00
5	G	211.68	0.14	1.00	0.00	226.37	0.38	1.00	0.00
	GL1, CV	91.91	0.12	0.40	0.59	118.93	2.10	0.18	0.82
	GL1, GCV	103.26	0.11	0.94	0.06	134.74	1.08	0.94	0.07
	GL1a, CV	134.73	0.12	0.78	0.19	141.33	0.45	0.76	0.21
	GL1a, GCV	144.08	0.12	0.83	0.14	146.58	0.40	0.84	0.14
	GL1, CV, R	90.76	0.11	0.04	0.96	117.41	2.25	0.00	0.99
	GL1, GCV, R	83.61	0.12	0.58	0.37	117.09	1.88	0.60	0.39
	GL1a, CV, R	101.19	0.11	0.59	0.37	113.66	0.91	0.53	0.43
	GL1a, GCV, R	97.70	0.11	0.65	0.30	113.06	1.25	0.63	0.31
	R	80.51	0.12	1.00	0.00	115.93	1.95	1.00	0.00
	Finite AIC	98.20	0.12	0.09	0.88	119.91	2.25	0.04	0.96
	Finite BIC	90.09	0.11	0.00	1.00	117.35	2.25	0.00	1.00

Table B.3.: Results for the settings with Gaussian response, $b_{i0} \sim N(0, 4)$, $n_i = 5$.

		$\rho = 0.0$				$\rho = 0.8$			
K	Method	Intercepts	Slope	FP	FN	Intercepts	Slope	FP	FN
30	G	211.68	0.14		0.00	226.58	0.38		0.00
	GL1, CV	98.35	0.11		0.36	287.29	4.84		0.67
	GL1, GCV	106.86	0.11		0.05	203.37	2.24		0.05
	GL1a, CV	136.76	0.09		0.18	160.75	0.72		0.18
	GL1a, GCV	149.79	0.09		0.13	160.97	0.63		0.12
	GL1, CV, R	150.20	0.12		0.89	289.72	5.81		0.94
	GL1, GCV, R	101.17	0.13		0.34	266.65	4.80		0.33
	GL1a, CV, R	107.07	0.10		0.33	172.88	1.99		0.39
	GL1a, GCV, R	104.50	0.10		0.33	181.41	2.76		0.28
	R	101.08	0.12		0.00	269.26	5.30		0.00
	Finite AIC	158.18	0.12		0.83	291.85	5.72		0.93
	Finite BIC	153.81	0.13		0.98	290.34	6.04		1.00
15	G	211.67	0.14	1.00	0.00	225.78	0.38	1.00	0.00
	GL1, CV	101.66	0.11	0.66	0.30	305.92	4.75	0.42	0.58
	GL1, GCV	113.22	0.11	0.94	0.05	207.59	2.27	0.94	0.05
	GL1a, CV	140.25	0.10	0.78	0.18	171.49	0.74	0.76	0.20
	GL1a, GCV	150.37	0.10	0.85	0.13	169.86	0.70	0.84	0.12
	GL1, CV, R	175.34	0.11	0.14	0.85	308.14	5.93	0.09	0.91
	GL1, GCV, R	111.80	0.13	0.61	0.34	278.77	5.11	0.64	0.35
	GL1a, CV, R	113.10	0.09	0.58	0.38	166.74	1.73	0.56	0.37
	GL1a, GCV, R	106.85	0.10	0.64	0.30	190.39	2.92	0.64	0.29
	R	113.89	0.12	1.00	0.00	272.15	4.95	1.00	0.00
	Finite AIC	176.66	0.11	0.14	0.80	311.39	5.84	0.08	0.91
	Finite BIC	176.88	0.14	0.02	0.96	310.41	6.02	0.00	1.00
5	G	211.68	0.14	1.00	0.00	225.95	0.38	1.00	0.00
	GL1, CV	129.26	0.10	0.66	0.32	176.74	3.01	0.27	0.73
	GL1, GCV	118.10	0.11	0.94	0.04	163.23	1.51	0.94	0.06
	GL1a, CV	146.52	0.12	0.81	0.15	154.50	0.69	0.76	0.19
	GL1a, GCV	152.55	0.12	0.85	0.12	157.10	0.49	0.83	0.12
	GL1, CV, R	151.30	0.12	0.10	0.89	178.26	3.28	0.03	0.97
	GL1, GCV, R	132.98	0.13	0.62	0.31	170.91	2.82	0.59	0.39
	GL1a, CV, R	126.97	0.10	0.65	0.28	151.50	1.30	0.51	0.43
	GL1a, GCV, R	131.66	0.11	0.64	0.28	153.66	1.78	0.63	0.29
	R	109.32	0.11	1.00	0.00	171.66	2.77	1.00	0.00
	Finite AIC	156.39	0.10	0.14	0.79	181.35	3.12	0.05	0.95
	Finite BIC	153.51	0.10	0.02	0.97	178.47	3.28	0.00	1.00

Table B.4.: Results for the settings with Gaussian response, $b_{i0} \sim \chi^2_3$, $n_i = 5$.

		$\rho = 0.0$				$\rho = 0.8$			
K	Method	Intercepts	Slope	FP	FN	Intercepts	Slope	FP	FN
30	GL1, CV	17.04	0.01		0.04	31.89	0.28		0.14
	GL1, GCV	16.98	0.01		0.04	32.26	0.27		0.03
	GL1a, CV	16.18	0.01		0.08	39.64	0.37		0.15
	GL1a, GCV	16.34	0.01		0.08	39.64	0.37		0.15
	GL1, CV, R	68.76	0.01		0.30	76.16	1.16		0.96
	GL1, GCV, R	54.37	0.01		0.12	65.49	0.94		0.21
	GL1a, CV, R	23.10	0.00		0.17	55.15	0.70		0.47
	GL1a, GCV, R	24.23	0.01		0.13	49.32	0.60		0.22
	R	16.61	0.01		0.00	51.31	0.74		0.00
	Finite AIC	32.54	0.01		0.39	63.88	0.80		0.71
	Finite BIC	34.22	0.01		0.47	76.08	1.09		0.88
15	GL1, CV	26.64	0.01	0.75	0.08	45.66	0.34	0.87	0.04
	GL1, GCV	23.15	0.01	0.77	0.07	44.85	0.34	0.89	0.02
	GL1a, CV	25.85	0.01	0.65	0.11	71.92	0.61	0.67	0.13
	GL1a, GCV	25.62	0.01	0.65	0.11	71.92	0.61	0.67	0.13
	GL1, CV, R	86.84	0.01	0.37	0.39	156.30	1.82	0.04	0.95
	GL1, GCV, R	61.80	0.01	0.66	0.11	102.77	0.90	0.78	0.10
	GL1a, CV, R	37.98	0.01	0.49	0.22	98.54	0.96	0.44	0.32
	GL1a, GCV, R	35.74	0.01	0.60	0.14	92.54	0.91	0.58	0.18
	R	23.86	0.01	1.00	0.00	76.12	0.84	1.00	0.00
	Finite AIC	45.20	0.01	0.19	0.39	84.88	0.83	0.20	0.49
	Finite BIC	53.95	0.01	0.13	0.48	85.77	0.95	0.12	0.66
5	GL1, CV	17.80	0.01	0.82	0.03	51.64	0.27	0.88	0.02
	GL1, GCV	17.51	0.01	0.83	0.02	51.64	0.27	0.88	0.02
	GL1a, CV	17.30	0.01	0.70	0.05	78.55	0.61	0.68	0.12
	GL1a, GCV	17.30	0.01	0.71	0.05	78.55	0.61	0.68	0.12
	GL1, CV, R	82.77	0.02	0.41	0.19	204.17	2.10	0.17	0.79
	GL1, GCV, R	60.97	0.01	0.67	0.06	93.85	0.86	0.83	0.05
	GL1a, CV, R	22.64	0.01	0.57	0.10	108.19	1.06	0.45	0.25
	GL1a, GCV, R	23.80	0.01	0.65	0.06	95.01	0.84	0.61	0.14
	R	17.25	0.01	1.00	0.00	67.50	0.60	1.00	0.00
	Finite AIC	32.43	0.01	0.21	0.27	93.57	0.71	0.18	0.37
	Finite BIC	37.21	0.01	0.12	0.36	93.57	0.80	0.14	0.45

Table B.5.: Results for the settings with binomial response, $b_{i0} \sim N(0, 4)$, $n_i = 10$.

		$\rho = 0.0$				$\rho = 0.8$			
K	Method	Intercepts	Slope	FP	FN	Intercepts	Slope	FP	FN
30	GL1, CV	40.34	0.01		0.07	81.59	0.29		0.03
	GL1, GCV	37.48	0.01		0.06	81.41	0.29		0.03
	GL1a, CV	34.92	0.01		0.11	104.34	0.46		0.14
	GL1a, GCV	34.92	0.01		0.11	104.34	0.46		0.14
	GL1, CV, R	105.21	0.01		0.49	196.96	1.74		0.87
	GL1, GCV, R	74.15	0.01		0.13	147.88	0.98		0.12
	GL1a, CV, R	42.24	0.01		0.23	132.70	0.90		0.33
	GL1a, GCV, R	42.71	0.01		0.14	122.15	0.67		0.20
	R	42.28	0.01		0.00	108.48	0.64		0.00
	Finite AIC	51.39	0.01		0.46	107.62	0.58		0.50
	Finite BIC	58.52	0.01		0.56	108.26	0.65		0.59
15	GL1, CV	16.84	0.01	0.83	0.06	60.17	0.28	0.87	0.03
	GL1, GCV	17.04	0.01	0.85	0.05	60.73	0.30	0.87	0.02
	GL1a, CV	14.90	0.01	0.72	0.10	88.25	0.56	0.67	0.13
	GL1a, GCV	14.87	0.01	0.72	0.10	88.25	0.56	0.67	0.13
	GL1, CV, R	57.34	0.01	0.36	0.41	182.64	1.90	0.04	0.95
	GL1, GCV, R	39.48	0.01	0.70	0.12	114.34	0.99	0.79	0.07
	GL1a, CV, R	18.79	0.01	0.56	0.20	116.55	1.02	0.46	0.32
	GL1a, GCV, R	19.70	0.01	0.66	0.13	106.71	0.83	0.60	0.17
	R	17.81	0.01	1.00	0.00	91.84	0.78	1.00	0.00
	Finite AIC	29.05	0.01	0.19	0.42	99.60	0.87	0.19	0.48
	Finite BIC	35.04	0.01	0.14	0.50	98.93	1.00	0.13	0.61
5	GL1, CV	15.96	0.01	0.82	0.05	28.61	0.19	0.89	0.02
	GL1, GCV	15.83	0.01	0.85	0.04	29.45	0.24	0.89	0.02
	GL1a, CV	16.00	0.01	0.74	0.07	39.12	0.35	0.71	0.11
	GL1a, GCV	15.91	0.01	0.74	0.07	39.12	0.35	0.71	0.11
	GL1, CV, R	51.72	0.01	0.28	0.58	116.37	1.42	0.14	0.81
	GL1, GCV, R	32.71	0.01	0.71	0.09	76.81	0.83	0.77	0.11
	GL1a, CV, R	21.02	0.01	0.53	0.20	60.72	0.69	0.51	0.26
	GL1a, GCV, R	20.47	0.01	0.67	0.10	55.76	0.60	0.64	0.16
	R	14.98	0.01	1.00	0.00	46.50	0.52	1.00	0.00
	Finite AIC	25.30	0.01	0.19	0.38	59.27	0.52	0.19	0.45
	Finite BIC	26.94	0.01	0.14	0.46	58.48	0.59	0.13	0.58

Table B.6.: Results for the settings with binomial response, $b_{i0} \sim \chi^2_3$, $n_i = 10$.

		$\rho = 0.0$				$\rho = 0.8$			
K	Method	Intercepts	Slope	FP	FN	Intercepts	Slope	FP	FN
30	GL1, CV	43.14	0.03		0.11	75.43	1.21		0.50
	GL1, GCV	39.94	0.03		0.09	52.66	0.78		0.05
	GL1a, CV	42.54	0.02		0.16	68.20	1.21		0.36
	GL1a, GCV	42.54	0.02		0.16	68.22	1.22		0.37
	GL1, CV, R	118.05	0.02		0.92	84.68	1.56		0.98
	GL1, GCV, R	89.33	0.02		0.17	76.71	1.37		0.25
	GL1a, CV, R	49.83	0.01		0.19	82.89	1.39		0.68
	GL1a, GCV, R	60.92	0.02		0.17	73.12	1.23		0.40
	R	44.03	0.03		0.00	75.26	1.48		0.00
	Finite AIC	63.98	0.02		0.51	84.96	1.46		0.89
	Finite BIC	63.02	0.03		0.52	84.71	1.52		0.98
15	GL1, CV	22.83	0.01	0.80	0.10	97.05	1.07	0.57	0.37
	GL1, GCV	23.14	0.01	0.82	0.08	77.70	0.81	0.88	0.04
	GL1a, CV	21.96	0.01	0.70	0.15	100.36	1.13	0.52	0.33
	GL1a, GCV	21.96	0.01	0.70	0.15	99.77	1.14	0.52	0.32
	GL1, CV, R	66.25	0.01	0.04	0.93	126.95	1.60	0.02	0.98
	GL1, GCV, R	46.36	0.01	0.68	0.17	114.40	1.30	0.67	0.23
	GL1a, CV, R	23.48	0.01	0.62	0.21	122.43	1.40	0.24	0.67
	GL1a, GCV, R	27.93	0.01	0.67	0.17	108.69	1.21	0.49	0.36
	R	23.41	0.01	1.00	0.00	106.91	1.51	1.00	0.00
	Finite AIC	34.08	0.02	0.18	0.50	126.51	1.50	0.11	0.83
	Finite BIC	34.59	0.02	0.16	0.54	126.96	1.55	0.02	0.97
5	GL1, CV	36.35	0.02	0.71	0.07	53.08	0.80	0.64	0.30
	GL1, GCV	32.50	0.03	0.73	0.07	42.45	0.62	0.89	0.04
	GL1a, CV	35.19	0.02	0.59	0.12	53.95	0.79	0.58	0.30
	GL1a, GCV	35.19	0.02	0.59	0.12	53.27	0.79	0.58	0.30
	GL1, CV, R	128.95	0.02	0.07	0.85	74.81	1.22	0.02	0.97
	GL1, GCV, R	88.56	0.02	0.58	0.12	63.84	1.00	0.68	0.23
	GL1a, CV, R	45.18	0.02	0.51	0.17	67.98	1.02	0.29	0.64
	GL1a, GCV, R	56.21	0.02	0.57	0.12	58.37	0.93	0.55	0.33
	R	36.61	0.03	1.00	0.00	58.37	1.10	1.00	0.00
	Finite AIC	64.74	0.03	0.19	0.43	74.96	1.14	0.14	0.78
	Finite BIC	60.74	0.03	0.17	0.48	74.84	1.20	0.02	0.96

Table B.7.: Results for the settings with binomial response, $b_{i0} \sim N(0, 4)$, $n_i = 5$.

		$\rho = 0.0$				$\rho = 0.8$			
K	Method	Intercepts	Slope	FP	FN	Intercepts	Slope	FP	FN
30	GL1, CV	35.13	0.02		0.11	144.52	1.20		0.17
	GL1, GCV	32.05	0.03		0.10	137.26	0.99		0.06
	GL1a, CV	33.44	0.02		0.17	166.16	1.42		0.28
	GL1a, GCV	33.44	0.02		0.17	166.16	1.42		0.28
	GL1, CV, R	105.03	0.01		0.90	219.45	2.22		0.95
	GL1, GCV, R	66.66	0.01		0.17	193.53	1.55		0.17
	GL1a, CV, R	41.92	0.02		0.27	191.15	1.90		0.48
	GL1a, GCV, R	47.38	0.02		0.19	184.65	1.59		0.31
	R	36.84	0.03		0.00	172.23	1.93		0.00
	Finite AIC	54.99	0.03		0.53	218.36	2.01		0.73
	Finite BIC	54.71	0.03		0.55	218.69	2.11		0.85
15	GL1, CV	86.90	0.03	0.71	0.11	200.60	1.35	0.74	0.13
	GL1, GCV	79.55	0.03	0.74	0.10	196.30	1.28	0.81	0.06
	GL1a, CV	81.39	0.03	0.60	0.16	230.92	1.61	0.48	0.27
	GL1a, GCV	81.39	0.03	0.60	0.16	231.56	1.72	0.47	0.28
	GL1, CV, R	202.00	0.03	0.06	0.87	315.01	2.77	0.03	0.97
	GL1, GCV, R	145.39	0.03	0.61	0.16	260.34	2.04	0.71	0.14
	GL1a, CV, R	97.63	0.02	0.48	0.26	273.13	2.19	0.25	0.57
	GL1a, GCV, R	103.46	0.03	0.58	0.18	259.38	1.92	0.45	0.29
	R	88.96	0.03	1.00	0.00	238.12	2.52	1.00	0.00
	Finite AIC	114.89	0.04	0.20	0.50	262.16	2.34	0.17	0.63
	Finite BIC	114.76	0.04	0.18	0.53	311.54	2.41	0.08	0.76
5	GL1, CV	23.64	0.02	0.71	0.13	93.98	1.09	0.71	0.17
	GL1, GCV	23.12	0.02	0.74	0.11	88.04	0.95	0.80	0.09
	GL1a, CV	18.99	0.02	0.60	0.18	109.30	1.25	0.48	0.32
	GL1a, GCV	18.99	0.02	0.60	0.18	109.30	1.23	0.49	0.31
	GL1, CV, R	84.81	0.02	0.07	0.81	171.34	1.99	0.05	0.94
	GL1, GCV, R	37.82	0.01	0.64	0.16	124.69	1.35	0.71	0.17
	GL1a, CV, R	23.27	0.01	0.47	0.30	137.24	1.59	0.28	0.58
	GL1a, GCV, R	23.99	0.01	0.57	0.20	125.81	1.39	0.46	0.35
	R	26.68	0.02	1.00	0.00	126.30	1.88	1.00	0.00
	Finite AIC	34.78	0.02	0.14	0.53	170.65	1.86	0.16	0.71
	Finite BIC	33.58	0.02	0.11	0.55	172.05	1.94	0.05	0.89

Table B.8.: Results for the settings with binomial response, $b_{i0} \sim \chi^2_3$, $n_i = 5$.

C. Proofs for Chapter 5

C.1. Proof of Theorem 3

Lemma 3. *Consider the estimate* $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda \cdot P(\boldsymbol{\beta})$ *of a penalized linear model with orthonormal design* $\boldsymbol{X}^T\boldsymbol{X} = \mathbb{I}_{(k+1)\times(k+1)}$ *and the general penalty* $P(\boldsymbol{\beta}) = \sum_{r\in\mathcal{I}_1, s\in\mathcal{I}_2} g(|\beta_r - \beta_s|)$*, where* $\mathcal{I}_1$*,* $\mathcal{I}_2$ *denote nonempty sets of indices, and where* $g : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ *denotes a monotonically increasing function. Then, it holds that* $\sum_{r=0}^k \hat{\beta}_r = \sum_{r=0}^k \hat{\beta}_r^{ML}$ *and thus,* $\bar{\hat{\boldsymbol{\beta}}} = \bar{\boldsymbol{\beta}}^{ML}$*.*

Proof. Consider

$$\boldsymbol{\beta}^* = \arg \min_{\boldsymbol{\beta}\in\mathbb{R}^{(k+1)}} \left(\mathcal{M}(\boldsymbol{\beta}) := ||\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}||_2^2 + \lambda P(\boldsymbol{\beta})\right), \tag{C.1}$$

for any input vector $\tilde{\boldsymbol{\beta}} \in \mathbb{R}^{(k+1)}$, for any $\lambda \geq 0$ and for the penalty $P(\boldsymbol{\beta})$ that is defined in Lemma 3. The penalty P and thus the objective function $\mathcal{M}$ can be non-convex such that $\boldsymbol{\beta}^*$ is not unique. By definition, P and thus $\mathcal{M}$ are bounded by 0 such that $\mathcal{M}$ has a unique minimum nonetheless. The proof relies only on the uniqueness of this minimum and can be applied to all solutions of (C.1).

Let $m \in \mathbb{R}$ be a scalar and let $\mathbb{1}_{k+1}$ denote a vector of ones of length $k+1$. Consider the point $\boldsymbol{u} := \boldsymbol{\beta}^* - m \cdot \mathbb{1}_{k+1}$ and compare $\mathcal{M}(\boldsymbol{\beta}^*)$ with $\mathcal{M}(\boldsymbol{u})$. First of all, note that, for any $m \in \mathbb{R}$,

$$\begin{aligned} P(\boldsymbol{u}) = P(\boldsymbol{\beta}^* - m \cdot \mathbb{1}_{k+1}) &= \sum_{r\in\mathcal{I}_1}\sum_{s\in\mathcal{I}_2} g\Big(\big|(\beta_r^* - m) - (\beta_s^* - m)\big|\Big) \\ &= \sum_{r\in\mathcal{I}_1}\sum_{s\in\mathcal{I}_2} g\big(|\beta_r^* - \beta_s^*|\big) \\ &= P(\boldsymbol{\beta}^*). \end{aligned}$$

Hence, the penalty is irrelevant for the comparison of $\mathcal{M}(\boldsymbol{\beta}^*)$ and $\mathcal{M}(\boldsymbol{u})$. Differentiation of the L_2^2-term in $\mathcal{M}(\boldsymbol{u})$ with respect to m shows that

$$m^* = \arg\min_{m\in\mathbb{R}} ||\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}} - m \cdot \mathbb{1}_{k+1}||_2^2 = \frac{1}{k+1}\sum_{r=0}^{k}(\beta_r^* - \tilde{\beta}_r).$$

For $\boldsymbol{u}^* = \boldsymbol{\beta}^* - m^* \cdot \mathbb{1}_{k+1}$, it holds that

$$\begin{aligned}
\mathcal{M}(\boldsymbol{u}^*) - \mathcal{M}(\boldsymbol{\beta}^*) &= \Big(||\boldsymbol{u}^* - \tilde{\boldsymbol{\beta}}||_2^2 + \lambda P(\boldsymbol{u}^*)\Big) - \Big(||\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}||_2^2 + \lambda P(\boldsymbol{\beta}^*)\Big) \\
&= ||\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}} - m^* \cdot \mathbb{1}_{k+1}||_2^2 - ||\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}||_2^2 \\
&\leq 0 \\
\Leftrightarrow \mathcal{M}(\boldsymbol{u}^*) &\leq \mathcal{M}(\boldsymbol{\beta}^*).
\end{aligned}$$

As the the L_2^2-terms are strictly convex, $\mathcal{M}(\boldsymbol{u}^*) = \mathcal{M}(\boldsymbol{\beta}^*)$ holds if and only if $\boldsymbol{u}^* = \boldsymbol{\beta}^*$. Hence, as $\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{(k+1)}} \mathcal{M}(\boldsymbol{\beta})$, any $\boldsymbol{u}^* \neq \boldsymbol{\beta}^*$ is a contradiction. Thus, it holds that

$$\begin{aligned}
\boldsymbol{u}^* &= \boldsymbol{\beta}^* - m^* \cdot \mathbb{1}_{k+1} = \boldsymbol{\beta}^* \\
\Leftrightarrow m^* &= \frac{1}{k+1}\sum_{r=0}^{k}(\beta_r^* - \tilde{\beta}_r) = 0.
\end{aligned}$$

As $\boldsymbol{X}^T\boldsymbol{X} = \mathbb{I}_{(k+1)\times(k+1)}$, $\hat{\boldsymbol{\beta}}^{ML} = \boldsymbol{X}^T\boldsymbol{y}$.
According to Fan and Li (2001), in this case, the objective can be rewritten as

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda P(\boldsymbol{\beta}) \quad = \quad \left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{ML}\right\|_2^2 + \lambda P(\boldsymbol{\beta}) + \text{const.}$$

Hence, the results obtained above can be applied to the assumed orthonormal setting with $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{ML}$; thus, Lemma 3 holds. □

Theorem 3. *Assume a penalized linear model with orthonormal design; that is $\boldsymbol{X}^T\boldsymbol{X} = \mathbb{I}_{(k+1)\times(k+1)}$ where $\boldsymbol{X} \in \mathbb{R}^{(k+1)\times(k+1)}$ denotes the design matrix without an intercept and where $\mathbb{I}$ denotes the identity matrix. Let the ML estimates be ordered $\hat{\beta}_0^{ML} < \ldots < \hat{\beta}_k^{ML}$ and employ penalty (5.3) with a fixed penalty parameter λ, $\lambda \geq 0$. Then for j, $\hat{\beta}_j^{ML} < \bar{\beta}^{ML}$, $\bar{\beta}^{ML} = \frac{1}{k+1}\sum_{j=0}^{k}\hat{\beta}_j^{ML}$, one obtains*

$$\hat{\beta}_j = \min\left\{\bar{\beta}^{ML},\ \max\{\hat{\beta}_l^{ML}, \hat{\beta}_j^{ML}\} + \frac{(\lambda - \lambda_l)I_{(l\geq j)}}{2(l+1)}\right\},$$

where $l = \max_{l=0,\ldots,k}(\lambda_l < \lambda)$, $\lambda_l = \sum_{u=1}^{l} 2u\left|\hat{\beta}_u^{ML} - \hat{\beta}_{u-1}^{ML}\right|$, and with indicator function I. For $\hat{\beta}_j^{ML} \geq \bar{\beta}^{ML}$, one obtains analogously

$$\hat{\beta}_j = \max\left\{\bar{\beta}^{ML},\ \min\{\hat{\beta}_l^{ML}, \hat{\beta}_j^{ML}\} - \frac{(\lambda - \lambda_l)I_{(k-l\geq j)}}{2(l+1)}\right\},$$

with $\lambda_l = \sum_{u=l}^{k-1} 2(k-u)\left|\hat{\beta}_{u+1}^{ML} - \hat{\beta}_u^{ML}\right|$ and l as before.

Proof. According to Fan and Li (2001), the objective and the estimate are defined by

$$\begin{aligned} \mathcal{M}_{pen}(\boldsymbol{\beta}) &= \left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{ML}\right\|_2^2 + \lambda\left\|\boldsymbol{R}\boldsymbol{\beta}\right\|_1, \qquad \text{(C.2)} \\ \hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}} \mathcal{M}_{pen}(\boldsymbol{\beta}), \end{aligned}$$

where λ denotes the penalty parameter of the penalty, and where $\boldsymbol{R}\boldsymbol{\beta}$ with

$$\boldsymbol{R} = \begin{pmatrix} -1 & 1 & 0 & & \ldots & 0 \\ 0 & -1 & 1 & 0 & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & 0 & -1 & 1 & 0 \\ 0 & \ldots & & 0 & -1 & 1 \end{pmatrix} \in \mathbb{R}^{k\times(k+1)},$$

builds the adjacent differences of coefficients.

As the objective (C.2) is convex, the Karush-Kuhn-Tucker conditions (KKT; Boyd and Vandenberghe, 2004, p. 243-244) are sufficient for a solution. The necessary background on subdifferential calculus for the following proof can be found in Hiriart-Urruty and Lemaréchal, 2004. With $\nabla\mathcal{M}_{pen}$ denoting the subdifferential or, depending on context, the gradient of $\mathcal{M}_{pen}$, each solution $\hat{\boldsymbol{\beta}}$ is characterized by the condition

$$0 \ \in\ \nabla\mathcal{M}_{pen}(\hat{\boldsymbol{\beta}}).$$

Hence, $\hat{\boldsymbol{\beta}}$ is obtained by solving the following equation for $\boldsymbol{\beta}$:

$$\begin{aligned} 0 &\in \nabla \mathcal{M}_{pen}(\boldsymbol{\beta}) = 2(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{ML}) + \lambda \cdot \nabla \left\| \boldsymbol{R\beta} \right\|_1 \\ \Leftrightarrow \quad \hat{\boldsymbol{\beta}}^{ML} - \boldsymbol{\beta} &\in \tfrac{\lambda}{2} \nabla \left\| \boldsymbol{R\beta} \right\|_1 , \end{aligned} \tag{C.3}$$

In order to obtain $\hat{\beta}_j$, start with $\lambda = 0$ and increase λ gradually. For $\lambda = 0$, $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{ML}$. For $\lambda > 0$, let λ_1 denote the value of λ for which the first pair of coefficients is fused. That is, for $0 < \lambda < \lambda_1$, all differences in $\boldsymbol{R\beta}$ are unequal zero; the penalty term is differentiable:

$$\left[\nabla \left\| \boldsymbol{R\beta} \right\|_1\right]_j = \begin{cases} \frac{\partial}{\partial \beta_j} \left(|\beta_j - \beta_{j-1}| + |\beta_{j+1} - \beta_j| \right) = 1 - 1 = 0 & \text{for } 0 < j < k, \\ \frac{\partial}{\partial \beta_j} \left(|\beta_{j+1} - \beta_j| \right) = -1 & \text{for } j = 0, \\ \frac{\partial}{\partial \beta_j} \left(|\beta_j - \beta_{j-1}| \right) = 1 & \text{for } j = k. \end{cases} \tag{C.4}$$

Hence, for $\lambda > 0$, a distinction of cases is helpful. As the ML estimate is assumed to be ordered and due to Lemma 3, distinguish coefficients with $\hat{\beta}_j^{ML} < \bar{\beta}^{ML}$ and with $\hat{\beta}_j^{ML} \geq \bar{\beta}^{ML}$.

- **Case 1:** β_j with $\hat{\beta}_j^{ML} < \bar{\beta}^{ML}$
  Due to (C.4), for $0 < \lambda \leq \lambda_1$, shrinkage only affects β_0. There is no shrinkage for $j > 0$; the first fusion of coefficients at $\lambda = \lambda_1$ must affect β_0, β_1. If the coefficients are fused, it holds that $|\beta_1 - \beta_0| = 0$. Therefore, define the subdifferential v of $|\xi|$:

$$v \begin{cases} \in [-1, 1] & \text{for } \xi = 0, \\ = \text{sign}(\xi) & \text{else wise.} \end{cases}$$

  Thus, for $0 < \lambda \leq \lambda_1$,

$$\begin{aligned} \left[\nabla \left\| \boldsymbol{R\beta} \right\|_1\right]_0 &= \frac{\partial}{\partial \beta_0} |\beta_1 - \beta_0| \\ &= -v. \end{aligned}$$

  With (C.3), it follows that

$$\begin{aligned} \hat{\beta}_j &= \hat{\beta}_j^{ML}, \quad j > 0 \\ \hat{\beta}_0 &= \begin{cases} \hat{\beta}_0^{ML} + \frac{1}{2}\lambda & \text{for } \lambda < 2(\hat{\beta}_1^{ML} - \hat{\beta}_0^{ML}), \\ \beta_1 & \text{for } \lambda = 2(\hat{\beta}_1^{ML} - \hat{\beta}_0^{ML}). \end{cases} \end{aligned}$$

  That is, the first fusion takes place for $\lambda \geq \lambda_1 = 2(\hat{\beta}_1^{ML} - \hat{\beta}_0^{ML})$; for $\lambda = \lambda_1$, it holds that $\hat{\beta}_0 = \hat{\beta}_1 = \hat{\beta}_1^{ML}$. Let λ_2 denote the value of λ for which the second pair of coefficients is fused. Consider now the case $\lambda_1 = 2(\hat{\beta}_1^{ML} - \hat{\beta}_0^{ML}) < \lambda \leq \lambda_2$, where it holds that

$$\begin{aligned}
[\nabla \|\boldsymbol{R\beta}\|_1]_1 &= \frac{\partial}{\partial \beta_1} \left| \beta_2 - \frac{\beta_0 + \beta_1}{2} \right| \\
&= -\frac{v}{2}, \\
[\nabla \|\boldsymbol{R\beta}\|_1]_2 &= 0.
\end{aligned}$$

With the same arguments as above, we obtain

$$\hat{\beta}_1 = \begin{cases} \hat{\beta}_1^{ML} + \frac{1}{4}(\lambda - \lambda_1) & \text{for } \lambda < \lambda_1 + 4(\hat{\beta}_2^{ML} - \hat{\beta}_1^{ML}), \\ \beta_2 & \text{for } \lambda = \lambda_1 + 4(\hat{\beta}_2^{ML} - \hat{\beta}_1^{ML}). \end{cases}$$

That is, the estimates of β_0, β_1, β_2 are the same for $\lambda \geq \lambda_2 = \lambda_1 + 4(\hat{\beta}_2^{ML} - \hat{\beta}_1^{ML})$; and it holds that $\hat{\beta}_0 = \hat{\beta}_1 = \hat{\beta}_2 = \hat{\beta}_2^{ML}$ for $\lambda = \lambda_2$. Recursive application gives

$$\hat{\beta}_j = \min \left\{ \bar{\beta}^{ML},\ \max\{\hat{\beta}_l^{ML}, \hat{\beta}_j^{ML}\} + \frac{(\lambda - \lambda_l) I_{(l \geq j)}}{2(l+1)} \right\},$$

with $l = \max_{l=0,\dots,k} (\lambda_l < \lambda)$, $\lambda_l = \sum_{u=1}^{l} 2u \left| \hat{\beta}_u^{ML} - \hat{\beta}_{u-1}^{ML} \right|$, and with indicator function I.

- **Case 2:** β_j with $\hat{\beta}_j^{ML} \geq \bar{\beta}^{ML}$
  Analogously, one obtains

  $$\hat{\beta}_j = \max \left\{ \bar{\beta}^{ML},\ \min\{\hat{\beta}_l^{ML}, \hat{\beta}_j^{ML}\} - \frac{(\lambda - \lambda_l) I_{(k-l \geq j)}}{2(l+1)} \right\},$$

  with $\lambda_l = \sum_{u=l}^{k-1} 2(k-u) \left| \hat{\beta}_{u+1}^{ML} - \hat{\beta}_u^{ML} \right|$ and l as before.

Note that, with λ_{max} denoting the minimal value of λ that yields maximal penalization, we have $\hat{\beta}_j = \bar{\beta}^{ML}$ for all j for $\lambda \geq \lambda_{max}$. Due to Lemma 3, for $\lambda = \lambda_{max}$, at least two (groups of) coefficients are fused which have estimates $\hat{\beta}_j \neq \bar{\beta}^{ML}$ for $\lambda < \lambda_{max}$. □

C.2. Representing Pairwise Fusion Penalties as Weighted Sum of Adjacent Differences

On page 80, it says: “Assume a fixed value of the penalty parameter λ and let $\beta_{(0)}, \ldots, \beta_{(k)}$ denote the arbitrary ordering of the solution (including $\beta_0 = 0$, without β_{int}). Then a short transformation (see Appendix C) shows that $\sum_{r>s\geq 0} |\beta_{(r)} - \beta_{(s)}| = \sum_{r=1}^{k} w_{(r)} |\beta_{(r)} - \beta_{(r-1)}|$, where $w_{(r)} = r(k-r+1)$.”

Proof. The ordering of the coefficients implies (for $r > s$) that

$$|\beta_{(r)} - \beta_{(s)}| = \sum_{l=s+1}^{r} |\beta_{(l)} - \beta_{(l-1)}|.$$

With

$$d_{(r)} \quad = \quad \left|\beta_{(r)} - \beta_{(r-1)}\right|,$$

one obtains

$$\begin{aligned} \sum_{r>s\geq 0} |\beta_{(r)} - \beta_{(s)}| &= \sum_{s=1}^{k} \sum_{l=s}^{k} \sum_{r=s}^{l} d_{(r)}, \\ \sum_{r=1}^{k} w_{(r)} |\beta_{(r)} - \beta_{(r-1)}| &= \sum_{r=1}^{k} w_{(r)} d_{(r)}. \end{aligned}$$

Hence, it is to show that

$$\sum_{s=1}^{k}\sum_{l=s}^{k}\sum_{r=s}^{l} d_{(r)} = \sum_{r=1}^{k} w_{(r)} d_{(r)}.$$

$$\begin{array}{rcccccccccccccc}
\sum_{s=1}^{k}\sum_{l=s}^{k}\sum_{r=s}^{l} d_{(r)} & = & \sum_{s=1}^{k} & \sum_{l=s}^{k} & \sum_{r=s}^{l} & d_{(r)} & & & & & & & & \\
 & & \overbrace{s=1} & \overbrace{l=1} & \overbrace{r=1} & d_{(1)} & & & & & & & & \\
 & & & l=2 & r=1,2 & d_{(1)} & + & d_{(2)} & & & & & & \\
 & & & l=3 & r=1,2,3 & d_{(1)} & + & d_{(2)} & + & d_{(3)} & & & & \\
 & & & \vdots & \vdots & \vdots & & \vdots & & \vdots & & & & \\
 & & & l=k & r=1,\ldots,k & d_{(1)} & + & d_{(2)} & + & d_{(3)} & + & \ldots & + & d_{(k)} \\
 & & s=2 & l=2 & r=2 & & & d_{(2)} & & & & & & \\
 & & & l=3 & r=2,3 & & & d_{(2)} & + & d_{(3)} & & & & \\
 & & & \vdots & \vdots & & & \vdots & & \vdots & & & & \\
 & & & l=k & r=2,\ldots,k & & & d_{(2)} & + & d_{(3)} & + & \ldots & + & d_{(k)} \\
 & & s=3 & l=3 & r=3 & & & & & d_{(3)} & & & & \\
 & & & \vdots & \vdots & & & & & \vdots & & & & \\
 & & & l=k & r=3,\ldots,k & & & & & d_{(3)} & + & \ldots & + & d_{(k)} \\
 & & & & & & & & & & & & & \\
 & & \vdots & \vdots & \vdots & & & & & & & & & \\
 & & s=k & l=k & r=k & & & & & & & & & d_{(k)} \\
 & & & & & \rotatebox{90}{k terms} & & \rotatebox{90}{2(k-1) terms} & & \rotatebox{90}{3(k-2) terms} & & & & \rotatebox{90}{k terms}
\end{array}$$

$$\begin{aligned}
&= k \cdot d_{(1)} + 2 \cdot (k-1) \cdot d_{(3)} + 3 \cdot (k-2) \cdot d_{(3)} + \ldots + k \cdot d_{(k)} \\
&= \sum_{r=1}^{k} r \cdot (k-r+1) \cdot d_{(r)} \\
&= \sum_{r=1}^{k} w_{(r)} d_{(r)}.
\end{aligned}$$

If the ordering of the solution is not bijective as there are fused categories, the number of categories k has to be reduced accordingly and the procedure is the same as described above. □

D. Derivations and Proofs for Chapter 6

D.1. Derivation of the IRLS Algorithm

In what follows, the iteratively re-weighted least squares (IRLS) algorithm from page 101 is derived. We start with a first order Taylor expansion of $\mathcal{M}(\boldsymbol{\beta})$ around $\boldsymbol{\beta}_{(l)}$:

$$\begin{aligned}
\mathcal{M}(\boldsymbol{\beta}) &\approx \mathcal{M}(\boldsymbol{\beta}_{(l)}) + \nabla\mathcal{M}(\boldsymbol{\beta}_{(l)})^T(\boldsymbol{\beta} - \boldsymbol{\beta}_{(l)}) \\
\nabla\mathcal{M}(\boldsymbol{\beta}_{(l)})^T(\boldsymbol{\beta} - \boldsymbol{\beta}_{(l)}) &= \sum_{i=1}^{n} \left(\nabla\mathcal{L}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)})\right)^T (\boldsymbol{\beta} - \boldsymbol{\beta}_{(l)}) \\
\nabla\mathcal{L}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)}) &= \frac{\partial\mathcal{L}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)})}{\partial(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)})} \frac{\partial(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)})}{\partial\boldsymbol{\beta}} \\
&= \mathcal{D}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)}) \cdot (-\boldsymbol{x}_i) \\
\text{with}\mathcal{D}(\xi) &= \frac{\partial\mathcal{L}(\xi)}{\partial\xi} = k((k\xi)^{2g} + c)^{\frac{1}{2g}-1}(k\xi)^{2g-1} \\
&\quad \cdot \exp(c^{\frac{1}{2g}} - ((k\xi)^{2g} + c)^{\frac{1}{2g}}).
\end{aligned}$$

The local trick of Fan and Li (2001) gives

$$(\nabla\mathcal{L}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)}))^T(\boldsymbol{\beta} - \boldsymbol{\beta}_{(l)}) \approx \mathcal{D}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)}) \cdot \frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}}{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)}} \cdot (-\boldsymbol{x}_i^T) \cdot (\boldsymbol{\beta} - \boldsymbol{\beta}_{(l)}).$$

With the quadratic approximation of Ulbricht (2010), we obtain

$$\begin{aligned}
(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})(-\boldsymbol{x}_i^T)(\boldsymbol{\beta} - \boldsymbol{\beta}_{(l)}) &= -y_i\boldsymbol{x}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}_{(l)}) + \boldsymbol{x}_i^T\boldsymbol{\beta}\boldsymbol{x}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}_{(l)}) \\
&\approx -y_i\boldsymbol{x}_i^T\boldsymbol{\beta} + y_i\boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)} + \frac{1}{2}(\boldsymbol{\beta}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{\beta}_{(l)}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)}).
\end{aligned}$$

Overall, we have

$$\begin{aligned}(\nabla\mathcal{L}(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)}))^T(\boldsymbol{\beta}-\boldsymbol{\beta}_{(l)}) &\approx -\boldsymbol{a}_i^T\boldsymbol{\beta}+\boldsymbol{a}_i^T\boldsymbol{\beta}_{(l)}+\frac{1}{2}(\boldsymbol{\beta}^T\boldsymbol{A}_i\boldsymbol{\beta}+\boldsymbol{\beta}_{(l)}^T\boldsymbol{A}_i\boldsymbol{\beta}_{(l)})\\ \text{with } \boldsymbol{a}_i^T &= \frac{\mathcal{D}(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)})}{y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)}}y_i\boldsymbol{x}_i^T\\ \text{and } \boldsymbol{A}_i &= \frac{\mathcal{D}(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)})}{y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)}}\boldsymbol{x}_i\boldsymbol{x}_i^T.\end{aligned}$$

Hence, the approximated objective function can be written as

$$\mathcal{M}(\boldsymbol{\beta}) \approx \mathcal{M}(\boldsymbol{\beta}_{(l)})-\boldsymbol{a}^T\boldsymbol{\beta}+\boldsymbol{a}^T\boldsymbol{\beta}_{(l)}+\frac{1}{2}(\boldsymbol{\beta}^T\boldsymbol{A}\boldsymbol{\beta}+\boldsymbol{\beta}_{(l)}^T\boldsymbol{A}\boldsymbol{\beta}_{(l)})=\mathcal{M}^{app}.$$

$\mathcal{M}^{app}$ has the derivatives

$$\begin{aligned}s(\boldsymbol{\beta}) &= \frac{\mathcal{M}^{app}}{\partial\boldsymbol{\beta}}=-\boldsymbol{a}+\boldsymbol{A}\boldsymbol{\beta}\\ \text{and } H(\boldsymbol{\beta}) &= \frac{\mathcal{M}^{app}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}=\boldsymbol{A}.\end{aligned}$$

Hence, the update of the current estimate $\boldsymbol{\beta}_{(l)}$ in iteration (l) is

$$\begin{aligned}\boldsymbol{\beta}_{(l+1)} &= \boldsymbol{\beta}_{(l)}-\nu H(\boldsymbol{\beta}_{(l)})^{-1}s(\boldsymbol{\beta}_{(l)})\\ &= \boldsymbol{\beta}_{(l)}-\nu\boldsymbol{A}_{(l)}^{-1}(-\boldsymbol{a}_{(l)}+\boldsymbol{A}_{(l)}\boldsymbol{\beta}_{(l)})\\ &= \boldsymbol{\beta}_{(l)}+\nu\boldsymbol{A}_{(l)}^{-1}\boldsymbol{a}_{(l)}-\nu\boldsymbol{A}_{(l)}^{-1}\boldsymbol{A}_{(l)}\boldsymbol{\beta}_{(l)}\\ &= (1-\nu)\boldsymbol{\beta}_{(l)}+\nu\boldsymbol{A}_{(l)}^{-1}\boldsymbol{a}_{(l)}.\end{aligned}$$

Apart from the update of the derivative $\mathcal{D}(\xi)$ of the approximated loss, the algorithm is as complex as usual IRLS algorithms. With $\boldsymbol{A}_{(l)}$ positive definite for all iterations l, the algorithm converges almost surely. However, as $\mathcal{L}(\xi)=1=const.$ and therewith $\mathcal{D}(\xi)=0$ for sufficiently large $|\xi|$, the initial values of $\beta_{(l)}$ have to be chosen carefully. The algorithm can be easily implemented; we employ an `R` (R Core Team, 2014) implementation.

D.2. Proofs for Section 6.2.2

D.2.1. Characteristics of K and Some Auxiliary Functions

Note that

$$\exp(-\sqrt{u^2+c}) \quad < \quad \exp(-|u|), \tag{D.1}$$

as $\sqrt{u^2+c} > |u|$ for all $u \in \mathbb{R}$.

$$\int_{-\infty}^{\infty} \frac{1}{2}\exp(-|u|)\mathrm{d}u \quad = \quad 1, \tag{D.2}$$

since $\frac{1}{2}\exp(-|u|)$ is the density of the Laplace distribution with $\mathbb{E}(u) = 0$, $\mathbb{V}(u) = 2$. Moreover,

$$\frac{u}{\sqrt{u^2+c}} \quad < \quad 1. \tag{D.3}$$

The derivatives of K are

$$K'(u) \quad = \quad -\frac{1}{2}\exp(-\sqrt{u^2+c})\frac{u}{\sqrt{u^2+c}}, \tag{D.4}$$

$$K''(u) \quad = \quad \frac{1}{2}\exp(-\sqrt{u^2+c})\left(\frac{u^2}{u^2+c} - \frac{c}{(u^2+c)^{3/2}}\right), \tag{D.5}$$

$$K'''(u) \quad = \quad u\exp(-\sqrt{u^2+c})\left(\frac{1.5c}{(u^2+c)^{5/2}} - \frac{u^2\sqrt{u^2+c}-3c}{2(u^2+c)^2}\right). \tag{D.6}$$

Note: $|K'|$, $|K''|$ and $|K'''|$ are symmetric around zero.

$$\int_0^{\infty} u^a \exp(-bu)\mathrm{d}u \quad = \quad \frac{a!}{b^{a+1}}, \tag{D.7}$$

for $a \in \mathbb{Z}$, $b > 0$, see Gradshteyn and Ryzhik (2007; Section 3.326, equation 2.^{10}).
Define

$$s : \mathbb{R}_0^+ \to \mathbb{R},\ u \mapsto s(u) = \left(\frac{u^2}{u^2+c} - \frac{c}{(u^2+c)^{3/2}}\right), \tag{D.8}$$

with $|s(u)| \leq c^{-1/2}$ as:

- $$\frac{\partial s(u)}{\partial u} = \frac{cu(2\sqrt{u^2+c}+3)}{(u^2+c)^{5/2}},$$

with roots at $u = 0$ and $u = \frac{1}{2}\sqrt{9-4c}$.

- $|\min_{u\geq 0} s(u)| = |s(0)| = |-c^{-1/2}| > |\max_{u\geq 0} s(u)| = |s(\frac{1}{2}\sqrt{9-4c})| = |1 - \frac{20}{27}c|$.

$$\begin{aligned} \sup_{u\geq 0} \exp(-u) &= 1. && \text{(D.9)} \\ u\exp(-|u|) &< 1, \quad \text{for all} u \in \mathbb{R}. && \text{(D.10)} \end{aligned}$$

Define

$$t : \mathbb{R}_0^+ \to \mathbb{R},\ u \mapsto t(u) = \frac{1.5c}{(u^2+c)^{5/2}} - \frac{u^2\sqrt{u^2+c} - 3c}{2(u^2+c)^2}, \tag{D.11}$$

Later on, we need $|t(u)| \leq \frac{3}{2}(c^{-1} + c^{-3/2})$. This follows from

-
$$\begin{aligned} t(u) &= t_{(+)}(u) - t_{(-)}(u) \quad \text{with} \\ t_{(+)} : \mathbb{R}_0^+ \to \mathbb{R},\ u \mapsto t_{(+)}(u) &= \frac{1.5c}{(u^2+c)^{5/2}} + \frac{3c}{2(u^2+c)^2}, \\ t_{(-)} : \mathbb{R}_0^+ \to \mathbb{R},\ u \mapsto t_{(-)}(u) &= \frac{u^2\sqrt{u^2+c}}{2(u^2+c)^2}. \end{aligned}$$

-
$$\frac{\partial t_{(+)}(u)}{\partial u} = -u\left(\frac{6c}{(u^2+c)^3} + \frac{7.5}{(u^2+c)^{7/2}}\right),$$
with one root at $u = 0$.

- $0 \leq t_{(+)}(u) \leq \max_{u\geq 0} t_{(+)}(u) = \frac{3}{2}(c^{-1} + c^{-3/2})$.

-
$$\frac{\partial t_{(-)}(u)}{\partial u} = \frac{2cu - u^3}{2(u^2+c)^{5/2}}$$
with roots at $u = 0$, $u = \sqrt{2c}$.

- $0 = \min_{u\geq 0} t_{(-)}(u) \leq t_{(-)}(u) \leq \max_{u\geq 0} t_{(-)}(u) = t_{(-)}(\sqrt{2c}) = 3^{-3/2}c^{-1/2}$.

- $\max_{u\geq 0} t_{(+)}(u) > \max_{u\geq 0} t_{(-)}(u) \Rightarrow |t(u)| \leq \frac{3}{2}(c^{-1} + c^{-3/2})$.

D.2.2. Proofs

Lemma 1. *The kernel function $K : \mathbb{R} \to \mathbb{R}$ defined in* (6.10) *is differentiable and fulfills the following conditions:*

(i) $\int_{-\infty}^{\infty} K(u)\mathrm{d}u = 1$*,*

(ii) $\sup_{u\in\mathbb{R}} |K(u)| = c_0 < \infty$*,*

(iii) $\sup_{u\in\mathbb{R}} |K'(u)| = c_1 < \infty$*, where* $K'(u) = \mathrm{d}K(u)/\mathrm{d}u$.

Proof.

(i) $\int_{-\infty}^{\infty} K(u)\mathrm{d}u \overset{(D.1),(D.2)}{<} 1$ and $K(u) > 0$.
Hence, it exists a constant $\delta \in \mathbb{R}^+$ such that $\int_{-\infty}^{\infty} \delta K(u)\mathrm{d}u = 1$.

(ii) Using (D.1) and (D.9), it follows that

$$\sup_{u\in\mathbb{R}} |K(u)| < \max_{u\in\mathbb{R}} \frac{1}{2} \exp\left(-|u|\right) = \frac{1}{2} < \infty.$$

(iii) $|K'(u)|$ is symmetric around zero, see (D.4); hence, it is to prove that $\sup_{u\in\mathbb{R}_0^+} |K'(u)| < \infty$.

$$\begin{aligned} |K'(u)| &= \left|-\frac{1}{2}\exp(-\sqrt{u^2+c})\frac{u}{\sqrt{u^2+c}}\right| \\ &\overset{(D.3)}{<} \frac{1}{2}\exp(-\sqrt{u^2+c}) \\ &\overset{(D.1)}{<} \frac{1}{2}\exp(-u) \\ &< \frac{1}{2}. \end{aligned}$$

□

Lemma 2. *The kernel function $K : \mathbb{R} \to \mathbb{R}$ defined in* (6.10) *is three times differentiable and fulfills the following conditions:*

(i) $\int_{-\infty}^{\infty} uK(u)\mathrm{d}u = 0,$

(ii) $\lim_{u\to\pm\infty} K(u) = 0,$

(iii) $\int_{-\infty}^{\infty} u^2|K(u)|\mathrm{d}u = M_0 < \infty,$

(iv) $\int_{-\infty}^{\infty} |K'(u)|^2\mathrm{d}u = M_1 < \infty,$

(v) $\sup_{u\in\mathbb{R}} |K''(u)| = M_2 < \infty,$

(vi) $\sup_{u\in\mathbb{R}} |K'''(u)| = M_3 < \infty,$

(vii) $\int_{-\infty}^{\infty} |K''(u)|^2\mathrm{d}u = M_4 < \infty.$

Proof. For differentiability, see (D.4), (D.5), (D.6).

(i)

$$\begin{aligned}\int_{-\infty}^{\infty} uK(u)\mathrm{d}u &= \left[\frac{1}{2}\exp(-\sqrt{u^2+c})(-\sqrt{u^2+c}-1)\right]_{-\infty}^{\infty}\\ &= 0.\end{aligned}$$

(ii) Follows directly from the definition.

(iii)

$$\begin{aligned}\int_{-\infty}^{\infty} u^2|K(u)|\mathrm{d}u &= 2\cdot\frac{1}{2}\int_{0}^{\infty} u^2\exp(-\sqrt{u^2+c})\mathrm{d}u\\ &\overset{(D.1)}{<} \int_{0}^{\infty} u^2\exp(-u)\mathrm{d}u\\ &\overset{(D.7)}{=} 2.\end{aligned}$$

(iv)

$$\begin{aligned}
\int_{-\infty}^{\infty} |K'(u)|^2 \mathrm{d}u &\overset{(D.4)}{=} \int_{-\infty}^{\infty} \frac{1}{4} \exp(-2\sqrt{u^2+c}) \frac{u^2}{u^2+c} \mathrm{d}u \\
&= \frac{1}{2} \int_{0}^{\infty} \exp(-2\sqrt{u^2+c}) \frac{u^2}{u^2+c} \mathrm{d}u \\
&\overset{(D.3),(D.1)}{<} \frac{1}{2} \int_{0}^{\infty} \exp(-2u) \mathrm{d}u \\
&\overset{(D.7)}{=} \frac{1}{4}.
\end{aligned}$$

(v) $|K''(u)|$ is symmetric around zero, see (D.5); therefore, it is to prove that

$$\begin{aligned}
\sup_{u \in \mathbb{R}_0^+} |K''(u)| &< \infty. \\
|K''(u)| &= \left| \frac{1}{2} \exp(-\sqrt{u^2+c}) s(u) \right| \\
&\overset{(D.8)}{\leq} \frac{1}{2} \exp(-\sqrt{u^2+c}) c^{-1/2} \\
&\overset{(D.1)}{<} \frac{1}{2} \exp(-u) c^{-1/2} \\
&\overset{(D.9)}{\leq} \frac{1}{2\sqrt{c}}.
\end{aligned}$$

(vi) $|K'''(u)|$ is symmetric around zero, see (D.6); therefore, it is to prove that

$$\begin{aligned}
\sup_{u \in \mathbb{R}_0^+} |K'''(u)| &< \infty. \\
|K'''(u)| &\overset{(D.6)}{=} \left| u \exp(-\sqrt{u^2+c}) t(u) \right| \\
&\overset{(D.11)}{\leq} u \exp(-\sqrt{u^2+c}) \frac{3}{2} (c^{-1} + c^{-3/2}) \\
&\overset{(D.1)}{<} u \exp(-u) \frac{3}{2} (c^{-1} + c^{-3/2}) \\
&\overset{(D.10)}{<} \frac{3}{2} (c^{-1} + c^{-3/2}).
\end{aligned}$$

(vii)

$$\begin{aligned}
\int_{-\infty}^{\infty} |K''(u)|^2 \mathrm{d}u &\overset{(D.5)}{=} 2\int_0^{\infty} |K''(u)|^2 \mathrm{d}u \\
&\overset{(D.5)}{=} 2 \cdot \frac{1}{4} \int_0^{\infty} \exp(-2\sqrt{u^2+c}) \left(\frac{u^2}{u^2+c} - \frac{c}{(u^2+c)^{3/2}} \right)^2 \mathrm{d}u \\
&= \frac{1}{2} \int_0^{\infty} \exp(-2\sqrt{u^2+c}) \left(s(u)\right)^2 \mathrm{d}u \\
&\overset{(D.8)}{\leq} \frac{1}{2} \int_0^{\infty} \exp(-2\sqrt{u^2+c}) c^{-1} \mathrm{d}u \\
&\overset{(D.1)}{<} \frac{1}{2c} \int_0^{\infty} \exp(-2u) \mathrm{d}u \\
&\overset{(D.7)}{=} \frac{1}{4c}.
\end{aligned}$$

□

References

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics 55*(1), 117–128.

Antoniadis, A. and J. Fan (2001). Regularization of wavelet approximations. *Journal of the American Statistical Association 96*(455), 939–967.

Bates, D. and M. Maechler (2014). *Matrix: sparse and dense matrix classes and methods.* R package versions 1.1-3/1.1-4.

Bates, D., M. Maechler, B. Bolker, and S. Walker (2014). *lme4: linear mixed-effects models using Eigen and S4.* R package versions 1.1-6/1.1-7.

Bivand, R. and N. Lewin-Koh (2014). *maptools: tools for reading and handling spatial objects.* R package version 0.8-30.

Bivand, R. S., E. J. Pebesma, and V. Gomez-Rubio (2013). *Applied Spatial Data Analysis with R* (second ed.). New York, USA: Springer. R package version 1.0-15.

Bondell, H. D. and B. J. Reich (2009). Simultaneous factor selection and collapsing levels in ANOVA. *Biometrics 65*(1), 169–177.

Boulesteix, A.-L. (2006). Maximally selected chi-square statistics for ordinal variables. *Biometrical Journal 48*(3), 451–462.

Boyd, S. and L. Vandenberghe (2004). *Convex Optimization.* New York, USA: Cambridge University Press.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika 52*(3), 345–370.

Claeskens, G. and N. L. Hjort (2008). Minimizing average risk in regression models. *Econometric Theory 24*(2), 493–527.

Collomb, G., W. Härdle, and S. Hassani (1987). A note on prediction via estimation of the conditional mode function. *Journal of Statistical Planning and Inference 15*, 227–236.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B. Statistical Methodology 34*(2), 187–220.

Curtin, T., S. Ingels, S. Wu, and R. Heuer (2002). *National education longitudinal study of 1988: base-year to fourth follow-up data file user's manual (NCES 2002-323).* Washington DC, USA: U.S. Department of Education, National Center for Education Statistics.

Dahl, D. B. (2014). *xtable: export tables to LaTeX or HTML.* R package version 1.7-3.

Dicker, L., B. Huang, and X. Lin (2013). Variable selection and estimation with the seamless-L_0 penalty. *Statistica Sinica 23*(2), 929–962.

Diggle, P. J., P. Heagerty, K.-Y. Liang, and S. L. Zeger (2002). *Analysis of Longitudinal Data* (second ed.). New York, USA: Oxford University Press.

Donoho, D. and M. Elad (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization. In *Proceedings of the National Academy of Sciences*, Volume 100, pp. 2197–2202. National Academy of Sciences.

Eddelbuettel, D. (2013). *Seamless R and C++ integration with Rcpp.* New York, USA: Springer. R package version 0.11.2.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics 32*(2), 407–451.

Efron, B. and R. Tibshirani (1993). *An Introduction to the Bootstrap.* New York, USA: Chapman & Hall.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with B-splines and penalties. *Statistical Science 11*(2), 89–102.

Einbeck, J. and G. Tutz (2006). Modelling beyond regression functions: An application of multimodal regression to speed-flow data. *Journal of the Royal Statistical Society. Series C. Applied Statistics 55*(4), 461–475.

Fahrmeir, L., C. Belitz, C. Biller, A. Brezger, S. Heim, A. Hennerfeind, and A. Jerak (2007). Statistische Analyse der Nettomieten. In *Mietspiegel für München 2007. Statistik, Dokumentation und Analysen. Landeshauptstadt München, Sozialreferat, Amt für Wohnen und Migration.*

Fahrmeir, L. and H. Kaufmann (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized liner models. *The Annals of Statistics 13*(1), 342–368.

Fahrmeir, L., T. Kneib, and S. Konrath (2010). Bayesian regularisation in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection. *Statistics and Computing 20*(2), 203–219.

Fahrmeir, L., T. Kneib, and S. Lang (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica 14*(3), 715–745.

Fahrmeir, L., T. Kneib, S. Lang, and B. Marx (2013). *Regression – Models, Methods and Applications*. Heidelberg, Germany: Springer.

Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models* (second ed.). New York, USA: Springer.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Fan, J. and W. Zhang (1999). Statistical estimation in varying coefficient models. *The Annals of Statistics 27*(5), 1491–1518.

Follmann, D. A. and D. Lambert (1989). Generalizing logistic regression by nonparametric mixing. *Journal of the American Statistical Association 84*(405), 295–300.

Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models*, Volume 2. Thousand Oaks, California, USA: SAGE Publications.

Frank, l. E. and J. H. Friedman (1993). A statistical view of some chemometrics regression tools. *Technometrics 35*(2), 109–135.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1–22. R package version 1.9-8.

Fruehwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. New York, USA: Springer.

Gannoun, A., J. Saracco, and K. Yu (2010). On semiparametric mode regression estimation. *Communications in Statistics. Theory and Methods 39*(7), 1141–1157.

Ge, D., X. Jiang, and Y. Ye (2011). A note on the complexity of L_p minimization. *Mathematical Programming, Series B 129*(2), 285–299.

Gertheiss, J., S. Hogger, C. Oberhauser, and G. Tutz (2011). Selection of ordinally scaled independent variables with applications to international classification of functioning core sets. *Journal of the Royal Statistical Society. Series C. Applied Statistics 60*(3), 377–395.

Gertheiss, J. and G. Tutz (2010). Sparse modeling of categorial explanatory variables. *The Annals of Applied Statistics 4*(4), 2150–2180.

Gertheiss, J. and G. Tutz (2012). Regularization and model selection with categorial effect modifiers. *Statistica Sinica 22*(3), 957–982.

GIMP Team (2012). *GNU image manipulation program.* http://www.gimp.org.

Goeman, J. J. (2010). L_1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal 52*(1), 70–84.

Goldstein, H. (2011). *Multilevel Statistical Models.* Chichester, West Sussex, UK: Wiley.

Gradshteyn, I. S. and I. M. Ryzhik (2007). *Table of Integrals, Series, and Products.* New York, USA: Elsevier Science.

Grilli, L. and C. Rampichini (2011). The role of sample cluster means in multilevel models: a view on endogeneity and measurement error issues. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences 7*(4), 121–133.

Grün, B. and F. Leisch (2008a). FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software 28*(4), 1–35. R package version 2.3-11.

Grün, B. and F. Leisch (2008b). Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. *Journal of Classification 25*(2), 225–247.

Hastie, T. and B. Efron (2013). *lars: least angle regression, Lasso and forward stagewise.* R package version 1.2.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models.* London, UK: Chapman & Hall.

Hastie, T. and R. Tibshirani (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B. Statistical Methodology 55*(4), 757–796.

Heinzl, F. (2013). *Clustering in Linear and Additive Mixed Models.* Ph. D. thesis, Department of Statistics, Ludwig-Maximilians-Universität München. Cuvillier Verlag, Göttingen.

Heinzl, F. and G. Tutz (2013). Clustering in linear mixed models with approximate Dirichlet process mixtures using EM algorithm. *Statistical Modelling 13*(1), 41–67.

Heinzl, F. and G. Tutz (2014). Clustering in linear-mixed models with a group fused Lasso penalty. *Biometrical Journal 56*(1), 44–68.

Held, L. (2008). *Methoden der statistischen Inferenz: Likelihood und Bayes.* Heidelberg, Germany: Spektrum Akademischer Verlag.

Hiriart-Urruty, J.-B. and C. Lemaréchal (2004). *Fundamentals of Convex Analysis.* Berlin, Germany: Springer.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

Hofner, B., T. Hothorn, and T. Kneib (2013). Variable selection and model choice in structured survival models. *Computational Statistics 28*(3), 1079–1101.

Hoover, D. R., J. A. Rice, C. O. Wu, and L.-P. Yang (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika 85*(4), 809–822.

Jain, A. and R. Dubes (1988). *Algorithms for Clustering Data.* New Jersey, USA: Prentice Hall.

Johnson, N. A. (2013). A dynamic programming algorithm for the fused Lasso and L_0-segmentation. *Journal of Computational and Graphical Statistics 22*(2), 246–260.

Kauermann, G. and G. Tutz (2000). Local likelihood estimation in varying-coefficient models including additive bias correction. *Journal of Nonparametric Statistics 12*(3), 343–371.

Kemp, G. C. and J. Santos Silva (2010). Regression towards the mode. Economics Discussion Papers 686, Department of Economics, University of Essex.

Kemp, G. C. and J. Santos Silva (2012). Regression towards the mode. *Journal of Econometrics 170*(1), 92–101.

Kneib, T. (2013). Beyond mean regression. *Statistical Modelling 13*(4), 275–303.

Kneib, T., F. Heinzl, A. Brezger, D. S. Bove, and N. Klein (2014). *BayesX: R utilities accompanying the software package BayesX.* R package versions 0.2-8/0.2-9.

Koch, I. (1996). On the asymptotic performance of median smoothers in image analysis and nonparametric regression. *The Annals of Statistics 24*(4), 1648–1666.

Lee, M. (1989). Mode regression. *Journal of Econometrics 42*(3), 337–349.

Lee, M. (1993). Quadratic mode regression. *Journal of Econometrics 57*(1–3), 1–19.

Leng, C. (2009). A simple approach for varying-coefficient model selection. *Journal of Statistical Planning and Inference 139*(7), 2138–2146.

Lin, Y. and H. H. Zhang (2006). Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics 34*(5), 2272–2297.

Lu, Y., R. Zhang, and L. Zhu (2008). Penalized spline estimation for varying-coefficient models. *Communications in Statistics. Theory and Methods 37*(14), 2249–2261.

Lu, Z. and Y. Zhang (2010). Penalty decomposition methods for l_0-norm minimization. *arXiv:1008.5372*.

Lu, Z. and Y. Zhang (2013). Sparse approximation via penalty decomposition methods. *SIAM Journal on Optimization 23*(4), 2448–2478.

Mancera, L. and J. Portilla (2006). L0-norm-based sparse representation through alternate projections. In *International Conference on Image Processing*, pp. 2089–2092. IEEE.

Manski, C. F. (1991). Regression. *Journal of Economic Literature 29*(1), 34–50.

Marx, B. D. and P. H. C. Eilers (1998). Direct generalized additive modeling with penalized likelihood. *Journal of Computational Statistics & Data Analysis 28*(2), 193–209.

McCullagh, P. (1983). Quasi-likelihood functions. *The Annals of Statistics 11*(1), 59–67.

McCullagh, P. and J. A. Nelder (1983). *Generalized Linear Models*. London, UK: Chapman & Hall.

McCulloch, C. E. and J. M. Neuhaus (2011). Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. *Statistical Science 26*(3), 388–402.

Meier, L. (2013). *grplasso: fitting user specified models with group Lasso penalty*. R package version 0.4-3.

Meier, L., S. van de Geer, and P. Bühlmann (2008). The group Lasso for logistic regression. *Journal of the Royal Statistical Society. Series B. Statistical Methodology 70*(1), 53–71.

Meier, L., S. van de Geer, and P. Bühlmann (2009). High-dimensional additive modeling. *The Annals of Statistics 37*(6B), 3779–3821.

Meier-Dinkel, L., J. Trautmann, L. Frieden, E. Tholen, C. Knorr, A. R. Sharifi, M. Bücking, M. Wicke, and D. Mörlein (2013). Consumer perception of boar meat as affected by labelling information, malodorous compounds and sensitivity to androstenone. *Meat Science 93*(2), 248–256.

Molenberghs, G. and G. Verbeke (2005). *Models for Discrete Longitudinal Data*. New York, USA: Springer.

Mörlein, D., A. Grave, A. R. Sharifi, M. Bücking, and M. Wicke (2012). Different scalding techniques do not affect boar taint. *Meat Science 91*(4), 435–440.

Neuhaus, J. M. and C. E. McCulloch (2006). Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *Journal of the Royal Statistical Society. Series B. Statistical Methodology 68*(5), 859–872.

Oelker, M.-R. (2014). *gvcm.cat: regularized categorial effects/categorial effect modifiers in GLMs.* R package version 1.7.

Oelker, M.-R., J. Gertheiss, and G. Tutz (2012a). Categorial effect modiffers in generalized linear models. In A. Colubi, K. Fokianos, G. Gonzalez-Rodriguez, and E. J. Kontoghiorghes (Eds.), *Proceedings of COMPSTAT - 20th International Conference on Computational Statistics*, pp. 665–676.

Oelker, M.-R., J. Gertheiss, and G. Tutz (2012b). Regularization and model selection with categorial predictors and effect modifiers in generalized linear models. Technical Report 122, Department of Statistics, Ludwig-Maximilians-Universität München. http://epub.ub.uni-muenchen.de/13082/.

Oelker, M.-R., J. Gertheiss, and G. Tutz (2014). Regularization and model selection with categorical predictors and effect modifiers in generalized linear models. *Statistical Modelling 14*(2), 157–177.

Oelker, M.-R., W. Pößnecker, and G. Tutz (2015). Selection and fusion of categorical predictors with L_0-type penalties. *Statistical Modelling 15*(4). Accepted for publication.

Oelker, M.-R., F. Sobotka, and T. Kneib (2014). On (semiparametric) mode regression. In T. Kneib, F. Sobotka, J. Fahrenholz, and H. Irmer (Eds.), *Proceedings of the 29th International Workshop on Statistical Modelling*, Volume 1, pp. 243–248.

Oelker, M.-R. and G. Tutz (2013). A general family of penalties for combining differing types of penalties in generalized structured models. Technical Report 139, Department of Statistics, Ludwig-Maximilians-Universität München. http://epub.ub.uni-muenchen.de/17664/.

Osborne, M. R. and B. A. Turlach (2011). A homotopy algorithm for the quantile regression Lasso and related piecewise linear problems. *Journal of Computational and Graphical Statistics 20*(4), 972–987.

O'Sullivan, F., B. S. Yandell, and W. J. Raynor, Jr. (1986). Automatic smoothing of regression functions in generalized linear models. *Journal of the American Statistical Association 81*(393), 96–103.

Pau, G., A. Oles, M. Smith, O. Sklyar, and W. Huber (2014). *EBImage: image processing toolbox for R.* R package version 4.6.0.

Peng, R. D., with contributions from, D. Murdoch, B. Rowlingson, and GPC library by Alan Murta (2013). *gpclib: general polygon clipping library for R.* R package version 1.5-5.

Pinheiro, J., D. Bates, S. DebRoy, D. Sarkar, and R Core Team (2014). *nlme: linear and nonlinear mixed effects models.* R package version 3.1-117.

R Core Team (2014). *R: a language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. R version 3.1.0 (2014-04-10).

Revolution Analytics (2014a). *doMC: foreach parallel adaptor for the multicore package.* R package version 1.3.3.

Revolution Analytics (2014b). *iterators: iterator construct for R.* R package version 1.0.7.

Revolution Analytics and S. Weston (2014). *foreach: foreach looping construct for R.* R package version 1.4.2.

Rippe, R. C. A., J. J. Meulman, and P. H. C. Eilers (2012). Visualization of genomic changes by segmented smoothing using an L_0 penalty. *PloS One 7*(6), 1–14.

Rue, H. and L. Held (2005). *Gaussian Markov Random Fields: Theory and Applications.* Monographs on Statistics & Applied Probability. Boca Raton, USA: Chapman & Hall.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression.* Cambridge, UK: Cambridge University Press.

Sarkar, D. (2008). *lattice: multivariate data visualization with R.* New York, USA: Springer. R package version 0.20-29.

Schauberger, G. and G. Tutz (2014). *catdata: categorical data.* R package version 1.2.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*(2), 461–464.

Stabler, B. (2013). *shapefiles: read and write ESRI shapefiles.* R package version 0.7.

Taylor, J. and J. Einbeck (2011). Multivariate regression smoothing through the ‘fallling net’. In D. Conesa, A. Forte, A. Lopez-Quilez, and F. Munoz (Eds.), *Proceedings of the 26th International Workshop on Statistical Modelling*, pp. 597–602.

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

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005). Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology 67*(1), 91–108.

Townsend, Z., J. Buckley, M. Harada, and M. A. Scott (2013). The choice between fixed and random effects. In B. D. M. Marc A. Scott, Jeffrey S. Simonoff (Ed.), *The SAGE Handbook of Multilevel Modeling*. Thousand Oaks, California, USA: SAGE Publications.

Trautmann, J., J. Gertheiss, M. Wicke, and D. Mörlein (2014). How olfactory acuity affects the sensory assessment of boar fat: a proposal for quantification. *Meat Science 98*(2), 255–262.

Tutz, G. (2012). *Regression for Categorical Data*. New York, USA: Cambridge University Press.

Tutz, G. and J. Gertheiss (2014). Rating scales as predictors – the old question of scale level and some answers. *Psychometrika 79*(3), 357–376.

Tutz, G. and M.-R. Oelker (2013). Modeling heterogeneity by fixed effects models. In V. M. R. Muggeo, V. Capursi, G. Boscaino, and G. Lovison (Eds.), *Proceedings of the 28th International Workshop on Statistical Modelling*, Volume 2, pp. 807–811.

Tutz, G. and M.-R. Oelker (2014). Modeling clustered heterogeneity: fixed effects, random effects and mixtures. Technical Report 156, Department of Statistics, Ludwig-Maximilians-Universität München. http://epub.ub.uni-muenchen.de/18987/.

Tutz, G. and G. Schauberger (2014). Extended ordered paired comparison models with application to football data from german Bundesliga. *AStA Advances in Statistical Analysis*, 1–19. Published online: October 10, 2014. Accepted for publication.

Ulbricht, J. (2010). *Variable Selection in Generalized Linear Models*. Ph. D. thesis, Department of Statistics, Ludwig-Maximilians-Universität München. Verlag Dr. Hut, München.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (fourth ed.). New York, USA: Springer. ISBN 0-387-95457-0, R package versions 7.3-31/7.3-33.

Verbeke, G. and G. Molenberghs (2000). *Linear Mixed Models for Longitudinal Data*. New York, USA: Springer.

Wang, H. and C. Leng (2008). A note on adaptive group Lasso. *Computational Statistics and Data Analysis 52*(12), 5277–5286.

Wang, H. and Y. Xia (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association 104*(486), 747–757.

Wang, L., H. Li, and J. Z. Huang (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association 103*(484), 1556–1569.

Weise, F.-J., H. Alt, and R. Becker (Eds.) (2011). *Arbeitsmarkt in Zahlen*, Nürnberg, Germany. Statistik der Bundesagentur für Arbeit.

Wikipedia User NordNordWest (2008). *Federal states of Germany.* `http://commons.wikimedia.org/wiki/File:Germany_location_map.svg`. Licenses: GNU Free Documentation License, Version 1.2 `http://commons.wikimedia.org/wiki/Commons:GNU_Free_Documentation_License,_version_1.2`, Creative Commons Attribution-Share Alike 3.0 Unported `http://creativecommons.org/licenses/by-sa/3.0/deed.en`.

Wipf, D. and B. Rao (2005). l_0-norm minimization for basis selection. In *Advances in Neural Information Processing Systems 17*, pp. 1513–1520.

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society. Series B. Statistical Methodology 73*(1), 3–36. R package versions 1.7-29/1.8-2.

Wood, S. N. (2006). *Generalized Additive Models : An Introduction with R.* New York, USA: Chapman & Hall.

Wu, C. O., C.-T. Chiang, and D. R. Hoover (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of the American Statistical Association 93*(444), 1388–1402.

Xiang, Y., S. Gubian, B. Suomela, and J. Hoeng (2013). Generalized simulated annealing for global optimization: the GenSA package. *The R Journal 5*(1), 13–29. R package version 1.1.4.

Yu, K. and K. Aristodemou (2012). Bayesian mode regression. Technical report. `http://arxiv.org/abs/1208.0579v1`.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B. Statistical Methodology 68*(1), 49–67.

Zhao, W., R. Zhang, and J. Liu (2014). Regularization and model selection for quantile varying coefficient model with categorical effect modifiers. *Computational Statistics and Data Analysis 79*, 44–62.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B. Statistical Methodology 67*(2), 301–320.

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12. Juli 2011, § 8 Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, 14. November 2014

Margret-Ruth Oelker