LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Wolfgang Pößnecker, Gerhard Tutz

A General Framework for the Selection of Effect Type in Ordinal Regression

Technical Report Number 186, 2016
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

A General Framework for the Selection of Effect Type in Ordinal Regression

Wolfgang Pößnecker* & Gerhard Tutz

Ludwig-Maximilians-Universität München
Akademiestraße 1, 80799 München, Germany

January 18, 2016

Abstract

In regression models for ordinal response, each covariate can be equipped with either a simple, global effect or a more flexible and complex effect which is specific to the response categories. Instead of a priori assuming one of these effect types, as is done in the majority of the literature, we argue in this paper that effect type selection shall be data-based. For this purpose, we propose a novel and general penalty framework that allows for an automatic, data-driven selection between global and category-specific effects in all types of ordinal regression models. Optimality conditions and an estimation algorithm for the resulting penalized estimator are given. We show that our approach is asymptotically consistent in both effect type and variable selection and possesses the oracle property. A detailed application further illustrates the workings of our method and demonstrates the advantages of effect type selection on real data.

Keywords: Effect Type Selection, Effect Type Lasso, Proportional Odds Model, Partial Proportional Odds Model, Ordinal Regression, Regularization, Penalization.

1. Introduction

In his seminal paper, McCullagh (1980) propagated a wide class of regression models for ordinal response variables, which are nowadays used in many fields of statistics. For example, in social sciences and in psychological tests, study participants are commonly asked to asses statements or situations on a discrete scale, e.g. from 1 to 10 or from 'totally disagree' to 'totally agree'. Another common area of application are biostatistics and medicine, where, e.g., the degree of pain or the stage of a disease are assessed by an ordinal rating. Ordinal regression also arises naturally in the modeling of survival times that are measured in discrete time intervals.

*Corresponding author, email: Wolfgang.Poessnecker@stat.uni-muenchen.de.

Model classes that are in common use for ordinal regression are the *cumulative*, the *sequential* and the *adjacent categories* models. An overview of methods, applications and literature of ordinal regression is found in Agresti (2013) and Tutz (2012). For its use in discrete survival, see Tutz & Schmid (2016).

In all kinds of ordinal models, the information that is contained in the ordering of the response categories is typically exploited by specifying a single parameter per explanatory variable. The corresponding predictor therefore has a global effect that is not specific to the considered response category. Despite being parsimonious and easy to interpret, such a global effect might be too restrictive. Alternatively, one can also specify one parameter per response category, such that a more complex and flexible model is obtained. However, if such a category-specific effect is used for all covariates, one does not take advantage of the additional information of the ordinal response and uses a model that is equivalent in complexity to multinomial models for unordered responses.

Cumulative logit models that use a mix of both effect types are known in the literature as *partial proportional odds model* and have been investigated by Cox (1995), Brant (1990) and Peterson & Harrell (1990). Nonetheless, a common effect type for all covariates is frequently chosen a priori, which neglects the uncertainty associated with the choice of effect type. In this paper, we argue that one should individually chose an appropriate effect type for each available explanatory variable in order to obtain an ordinal model that is as simple as possible and as flexible as necessary. To pursue this task of *effect type selection*, we propose a novel and general penalty framework that allows for an automatic, data-driven selection between global and category-specific effects in all types of ordinal regression models.

Previous work on penalty approaches for ordinal regression is limited to variable selection penalties. Archer & Williams (2012) consider sequential logit models with a lasso penalty (Tibshirani, 1996) on coefficients that are a priori assumed to be global. In Archer et al. (2014), this approach is generalized to a method for fitting all types of ordinal model with a lasso penalty. Coefficients are again limited to be global. To the best of our knowledge, penalty approaches to effect type selection in ordinal regression have not been treated in the literature. In a Bayesian context, data-based effect type selection was recently tackled by McKinley et al. (2015) via a reversible-jump MCMC approach.

In Section 2, the basic classes of ordinal regression models are summarized and the different possible specifications of covariate effects are discussed. In Section 3, our penalty approach, called the "*Effect Type Lasso*" (ETL), is presented and investigated, including optimality conditions, details on tuning parameter selection and a review of related penalty concepts in the literature. In Section 4, effect type selection is considered in an asymptotic setting. An adaptively weighted version of our ETL penalty is suggested and its oracle property (Fan & Li, 2001) in terms of effect type selection is shown. Details on the computation of numerical estimates are given in Section 5. Section 6 contains a detailed real data application to the survival of newly founded firms that illustrates the workings of our method and demonstrates the advantages of allowing flexible effect types. Technical derivations and the proofs for all theorems are found in the Appendix.

2. Ordinal Response Models

2.1. Basic Models

Let $Y \in \{1, 2, \ldots, k\}$ denote a categorical response variable for ordered categories and let $\boldsymbol{x}$ be a vector of covariates. One may distinguish three families of ordinal regression models that are in common use. The first family are the *cumulative models*, which have the form

$$\mathbb{P}(Y \leq t|\boldsymbol{x}) = F(\beta_{t0} + \boldsymbol{x}^\top \boldsymbol{\alpha}), \quad t = 1, \ldots, k-1,$$

where $F(\cdot)$ is a strictly monotone cumulative distribution function and $-\infty = \beta_{00} < \beta_{10} < \cdots < \beta_{k0} = \infty$. The model can be motivated as a coarser version of a latent regression model $\tilde{Y}_i = -\boldsymbol{x}_i^T \boldsymbol{\gamma} + \epsilon$ with a noise variable ϵ that has distribution function $F(\cdot)$ (McCullagh (1980)). The most widely used model from this class of models is the cumulative logit model, which uses the logistic distribution $F(\eta) = \exp(\eta)/(1 + \exp(\eta))$. It is also called the *proportional odds model* and has the form

$$\log\left(\frac{\mathbb{P}(Y \leq t|\boldsymbol{x})}{\mathbb{P}(Y > t|\boldsymbol{x})}\right) = \beta_{t0} + \boldsymbol{x}^\top \boldsymbol{\alpha}, \quad t = 1, \ldots, k-1.$$

The second family of models are the *sequential models*, which have the form

$$\mathbb{P}(Y = t|Y \geq t, \boldsymbol{x}) = F(\beta_{t0} + \boldsymbol{x}^\top \boldsymbol{\alpha}), \quad t = 1, \ldots, k-1,$$

where again $F(\cdot)$ is a cumulative distribution function. The most prominent example is the sequential logit model, also called continuation ratio model. It results when using the logistic distribution function for $F(\cdot)$ and has the form

$$\log\left(\frac{\mathbb{P}(Y = t|\boldsymbol{x})}{\mathbb{P}(Y > t|\boldsymbol{x})}\right) = \beta_{t0} + \boldsymbol{x}^\top \boldsymbol{\alpha}, \quad t = 1, \ldots, k-1. \tag{1}$$

The third family of models are the *adjacent categories models*

$$\mathbb{P}(Y = t|Y \in \{t, t+1\}, \boldsymbol{x}) = F(\beta_{t0} + \boldsymbol{x}^\top \boldsymbol{\alpha}), \quad t = 1, \ldots, k-1,$$

which compare two adjacent categories given the response is in one of the categories. It is used in particular in psychometrics for the evaluation of latent traits. When $F(\cdot)$ is chosen as the logistic distribution model the corresponding latent trait model is the so-called partial credit model (Masters (1982)).

The models have different advantages and drawbacks. For example, an advantage of the cumulative model is that adjacent categories can be collapsed without changing the effect of the covariates, a drawback is that the intercepts have to be ordered, which sometimes raises convergence problems of estimates. The cumulative model has been used for all kinds of ordered responses. In contrast, the sequential model is appropriate primarily if the categories of the response are reached successively. The model (1) is actually a conditional binary model for the occurence of category t

given category t has been reached. It has a strong connection to discrete survival analysis. In discrete survival analysis the categories refer to months, weeks or, generally speaking, time intervals and $h(t|\boldsymbol{x}) = P(Y = t|Y \geq t|\boldsymbol{x})$ is the discrete hazard function.

Comments on the relation between the different models were already given in McCullagh (1980). More recently, a careful investigation of the relationship between families of categorical models was given by Peyhardi et al. (2015).

All models make use of the ordering of categories by assuming that the same effect strength $\boldsymbol{\alpha}$ is present in all of the binary decisions specified above. In this basic forms, the models allow for easy interpretation because all of them show a form of stochastic ordering. Let us consider two covariate vectors $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$. Then in the case of the proportional odds model it takes the form

$$\frac{\mathbb{P}(Y \leq r|\boldsymbol{x})/\mathbb{P}(Y > r|\boldsymbol{x})}{\mathbb{P}(Y \leq r|\tilde{\boldsymbol{x}})/\mathbb{P}(Y > r|\tilde{\boldsymbol{x}})} = \exp((\boldsymbol{x} - \tilde{\boldsymbol{x}})^T \boldsymbol{\alpha}).$$

Thus the comparison of populations in terms of *cumulative odds* $P(Y \leq r|\boldsymbol{x})/P(Y > r|\boldsymbol{x})$ does not depend on the category. If, for example, the cumulative odds in population $\boldsymbol{x}$ are twice the cumulative odds in population $\tilde{\boldsymbol{x}}$, this holds for all the categories.

All of the above models can be embedded in the framework of multivariate generalized linear models (GLMs). With $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_q)^\top$ denoting the vector of the $q = k-1$ 'free' response probabilities with components $\pi_r = P(Y = r|\boldsymbol{x})$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_q)^\top$ denoting the vector of corresponding linear predictors η_t, one obtains the following form:

$$\boldsymbol{g}(\boldsymbol{\pi}) = \boldsymbol{\eta} = \mathbf{Z}\boldsymbol{\theta} \quad \text{or} \quad \boldsymbol{\pi} = \boldsymbol{g}^{-1}(\mathbf{Z}\boldsymbol{\theta}),$$

where $\mathbf{Z}$ is a design matrix constructed from the explanatory variables and $\boldsymbol{\theta}$ is the overall parameter vector, $\boldsymbol{g} = (g_1, \ldots, g_q) : \mathbb{R}^q \to \mathbb{R}^q$ is a vector-valued *link function* and $\boldsymbol{g}^{-1}$ is the response function. Based on this representation, ML inference for ordinal models, for example computation of the loglikelihood and the score function, is available via standard techniques for multivariate GLMs. Further details are found in Fahrmeir & Tutz (2001) and Tutz (2012).

2.2. Category-Specific Effects Models

All the basic models use the predictor structure

$$\eta_t = \beta_{t0} + \boldsymbol{x}^\top \boldsymbol{\alpha}. \tag{2}$$

This structure yields simple interpretation of parameters but is often too simplistic to represent the underlying dependency between response and covariates. Therefore, models have been extended by using the more general form

$$\eta_t = \beta_{t0} + \boldsymbol{x}^\top \boldsymbol{\beta}_t = \beta_{t0} + \sum_{j=1}^{p} x_j \beta_{tj}, \qquad t = 1, \ldots, q, \tag{3}$$

which allows the effects to vary over categories. The corresponding models are much more flexible and typically yield a better fit. Interpretation is linked to the type of model. For the cumulative model it means that each split into categories $\{1,\ldots,t\}$ and $\{t+1,\ldots,k\}$ is determined by an effect that is specific for the split. For the sequential model it means that the transition to a higher category given the category is reached depends on the category. In applications to discrete survival, this means that the effect of an explanatory variable varies over time.

The structure (3) is very flexible, however in its general form typically too complex, in particular if many explanatory variables are available. The number of parameters in the general model is the same as in the multinomial logit model which is constructed for nominal responses. This implies that the general model does not exploit the ordinal response structure. A compromise that is as parsimonious as possible and as flexible / complex as necessary is the mixture of effect types in

$$\eta_t = \beta_{t0} + \boldsymbol{x}_1^\top \boldsymbol{\alpha} + \boldsymbol{x}_2^\top \boldsymbol{\beta}_t, \tag{4}$$

where $\boldsymbol{x}$ is partitioned into $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. The first covariates in vector $\boldsymbol{x}_1$ have *global effects* $\boldsymbol{\alpha}$ while the covariates in vector $\boldsymbol{x}_2$ have *category-specific effects* $\boldsymbol{\beta}_t$. The mixture of effect types leads to specific names for the corresponding models. For example, the cumulative logit model with mixed effect types is called the *partial proportional odds model* (Brant, 1990; Peterson & Harrell, 1990), since the proportional odds property still holds for the variables in $\boldsymbol{x}_1$.

The problem with mixed effect types is that one has to decide which variables have which effect. In a model with p covariates, 2^p different specifications of the model with mixed effect type are possible, which is the same as for the traditional variable selection problem (see, e.g., Hastie et al., 2009). Hence, even for moderate p, the number of such models becomes too big to reliably solve the problem of effect type selection by use of test statistics or by stepwise procedures based on model selection criteria. Moreover, when a large number of predictors is available, it is sensible to assume that some of them have no effect on the response. Then, the researcher has to choose, for each predictor, between a category-specific, a global and a zero effect. Hence, effect type selection extends to a 'three-way' selection problem with complexity 3^p, which aggravates issues of test-based or stepwise procedures and emphasizes the need for an automatic, data-driven solution. In the next section, such a solution is presented based on a penalized loglikelihood approach.

3. A Penalty Approach to Simultaneous Variable and Effect Type Selection

3.1. Penalized Loglikelihood

We consider the general model with the linear predictor

$$\eta_t = \beta_{t0} + \boldsymbol{x}^\top \boldsymbol{\beta}_t = \beta_{t0} + \sum_{j=1}^{p} x_j \beta_{tj}, \qquad t = 1,\ldots,q.$$

Hence, a catgegory-specific effect is principally allowed for all covariates. Our goal is to reduce this flexible specification to the more simple global one whenever possible, resulting in a model with a mix of effect types as in (4), with some of the global effects being set to zero.

For $j = 1, \ldots, p$, let $\boldsymbol{\beta}_{\bullet j} = (\beta_{1j}, \ldots, \beta_{qj})^\top$ denote the vector of all coefficients that are linked to the j-th covariate, let $\boldsymbol{\beta}_{\bullet 0} = (\beta_{10}, \ldots, \beta_{q0})^\top$ denote the vector of all intercept parameters and let $\boldsymbol{\beta}^\top = (\boldsymbol{\beta}_{\bullet 0}^\top, \boldsymbol{\beta}_{\bullet 1}^\top, \ldots, \boldsymbol{\beta}_{\bullet p}^\top)$ denote the overall parameter vector of the model. In general, penalized likelihood approaches compute the penalized estimator by

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \ \big(l(\boldsymbol{\beta}) - J(\boldsymbol{\beta})\big) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \ \big(-l(\boldsymbol{\beta}) + J(\boldsymbol{\beta})\big)\,, \tag{5}$$

where $l(\boldsymbol{\beta})$ denotes the log-likelihood of the model and $J(\boldsymbol{\beta})$ is a penalty term that regularizes the parameters and encourages solutions with a desired structure.

3.2. The Effect Type Lasso

Let $\lambda, \zeta \geq 0$ be tuning parameters and let $\boldsymbol{\Omega} = \boldsymbol{D}^\top \boldsymbol{D}$ denote the penalty matrix that is constructed from the first order difference matrix

$$\boldsymbol{D} = \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & & \\ & & & -1 & 1 \end{pmatrix}.$$

With this notation, we propose the following, combined penalty to achieve simultaneous selection of variables and effect type:

$$\begin{aligned} J(\boldsymbol{\beta}) &= \sum_{j=1}^{p} \left(\lambda \sqrt{\sum_{t=1}^{q} \beta_{tj}^2} \ + \ \zeta \sqrt{\sum_{t=2}^{q} (\beta_{tj} - \beta_{t-1,j})^2} \right) \\ &= \sum_{j=1}^{p} \left(\lambda \sqrt{\boldsymbol{\beta}_{\bullet j}^\top \boldsymbol{\beta}_{\bullet j}} \ + \ \zeta \sqrt{\boldsymbol{\beta}_{\bullet j}^\top \boldsymbol{\Omega} \boldsymbol{\beta}_{\bullet j}} \right) \\ &= \sum_{j=1}^{p} \left(\lambda \, ||\boldsymbol{\beta}_{\bullet j}||_2 \ + \ \zeta \, ||\boldsymbol{D} \boldsymbol{\beta}_{\bullet j}||_2 \right). \end{aligned} \tag{6}$$

Steered by λ, the first term in (6) enforces *variable selection* by use of a group lasso penalty (Yuan & Lin, 2006; Meier et al., 2008) with groups defined by all the parameters that belong to the same covariate.

The second term in (6), which is the main methodological contribution of this paper, is a penalty of the group lasso type which is applied to the vector of all differences between parameters for adjacent categories within $\boldsymbol{\beta}_{\bullet j}$. For large enough but finite values of ζ, this penalty yields solutions in which the estimated coefficients for some variables x_j satisfy $\sum_{t=2}^{q} (\hat{\beta}_{tj} - \hat{\beta}_{t-1,j})^2 = 0$, which

implies $\hat{\beta}_{1j} = \hat{\beta}_{2j} = \ldots = \hat{\beta}_{qj} := \hat{\alpha}_j$. In that case, variable x_j effectively has a global effect, so this penalty term is able to shrink category-specific effects to global effects and thus enforces *effect type selection*. Specific conditions for this shrinkage and details on what exactly constitutes a "large enough" ζ are presented in Section 3.3. For $\zeta \to \infty$, all variables obtain global effects, resulting in the model structure (2). Note that if a variable x_j is estimated to have a global effect, the first penalty term is $||\hat{\boldsymbol{\beta}}_{\bullet j}||_2 = \sqrt{q}\,|\hat{\alpha}_j|$ and thus reduces to the ordinary lasso. In cases in which a covariate's effect is selected to be category-specific, the penalty applies a smoothing type shrinkage to the within-group adjacent coefficient differences, see equation (11) in Section 3.3 for details.

The second penalty term in (6) is a grouped version of the fused lasso of Tibshirani et al. (2005) and combines the idea of parameter fusion with the all-or-nothing selection of the group lasso to achieve the goal of effect type selection. We refrain from calling it "group fused lasso" since this term has been used in the literature (see, for example, Heinzl & Tutz, 2014; Wytock et al., 2014) and typically refers to the fusion of vector-valued arguments that is conceptually of L_1-type. In generic notation, the penalty that is used in these papers is of the form $J(\boldsymbol{\Theta}) = \lambda \sum_{s=2}^{d} ||\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s-1}||$ or $J(\boldsymbol{\Theta}) = \lambda \sum_{s>r} ||\boldsymbol{\theta}_s - \boldsymbol{\theta}_r||$ and therefore provides a different kind of fusion / selection. To stress the different motivation and selection behavior of our penalty (6), we call it *"effect type lasso" (ETLasso* or *ETL)* since it selects between category-specific, global and zero effects.

3.3. Optimality Conditions for ETL

To understand the selection behaviour of the ETL penalty, the properties of the corresponding penalized estimator, that is, the solution of the estimation problem

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \; \big(l(\boldsymbol{\beta}) - J(\boldsymbol{\beta})\big) = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left(l(\boldsymbol{\beta}) \; - \; \sum_{j=1}^{p} \Big(\; \lambda \, ||\boldsymbol{\beta}_{\bullet j}||_2 - \zeta \, ||\boldsymbol{D}\boldsymbol{\beta}_{\bullet j}||_2 \; \Big) \right) \tag{7}$$

are investigated. It follows from the Karush-Kuhn-Tucker conditions that $\hat{\boldsymbol{\beta}}$ must satisfy $\mathbf{0} \in \nabla l(\hat{\boldsymbol{\beta}}) - \partial J(\hat{\boldsymbol{\beta}})$, where $\partial J(\hat{\boldsymbol{\beta}})$ is the subdifferential of the penalty at the point $\hat{\boldsymbol{\beta}}$. The gradient of the loglikelihood is simply the score vector of $\hat{\boldsymbol{\beta}}$, that is, $\nabla l(\hat{\boldsymbol{\beta}}) = \boldsymbol{s}(\hat{\boldsymbol{\beta}}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$. Accordingly, the score vector for one parameter group is $\boldsymbol{s}_j := \boldsymbol{s}(\hat{\boldsymbol{\beta}}_{\bullet j}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\bullet j}}\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$. With this notation and exploiting the block-separability of the penalty, the following optimality conditions for the ETL estimator can be derived:

Theorem 1. *Let $\boldsymbol{I}_{q-1}$ denote the $q-1$-dimensional identity matrix and let $\mathbf{1}_q$ be a vector of ones of length q. For $j = 1, \ldots, p$, let $\hat{\tau}_j$ denote the largest solution of the nonlinear equation*

$$\boldsymbol{s}_j^{\top} \boldsymbol{D}^{\top} \left(\zeta \boldsymbol{D}\boldsymbol{D}^{\top} + \tau_j \boldsymbol{I}_{q-1} \right)^{-2} \boldsymbol{D}\boldsymbol{s}_j = 1 \tag{8}$$

and let $\hat{\tau}_j^* = \max(0, \hat{\tau}_j)$. For $j = 1, \ldots, p$, the ETL estimator, i.e. the solution to (7), is characterized by the following:

a) ***Zero effect condition:*** $\hat{\boldsymbol{\beta}}_{\bullet j} = \mathbf{0}$ *if and only if* $\left|\left| \boldsymbol{s}_j - \zeta \boldsymbol{D}^{\top} \left(\zeta \boldsymbol{D}\boldsymbol{D}^{\top} + \hat{\tau}_j^* \boldsymbol{I}_{q-1} \right)^{-1} \boldsymbol{D}\boldsymbol{s}_j \right|\right|_2 \leq \lambda. \tag{9}$

b) ***Global effect condition:*** $\hat{\boldsymbol{\beta}}_{\bullet j} \neq \mathbf{0} \wedge \boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j} = \mathbf{0}$ *only if* (9) *does not hold and one has*

$$\left|\left|(\boldsymbol{D}\boldsymbol{D}^\top)^{-1}\boldsymbol{D}\boldsymbol{s}_j\right|\right|_2 = \sqrt{\boldsymbol{s}_j^\top \boldsymbol{D}^\top (\boldsymbol{D}\boldsymbol{D}^\top)^{-2}\boldsymbol{D}\boldsymbol{s}_j} \leq \zeta. \tag{10}$$

c) ***Category-specific effects:*** $\hat{\boldsymbol{\beta}}_{\bullet j} \neq \mathbf{0} \wedge \boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j} \neq \mathbf{0}$ *if and only if neither* (9) *nor* (10) *hold. Then,* $\hat{\boldsymbol{\beta}}_{\bullet j}$ *is characterized by*

$$\boldsymbol{s}_j - \lambda \frac{\hat{\boldsymbol{\beta}}_{\bullet j}}{||\hat{\boldsymbol{\beta}}_{\bullet j}||_2} - \zeta \frac{\boldsymbol{D}^\top \boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}}{||\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}||_2} = 0. \tag{11}$$

d) If (10) *holds for* $\hat{\boldsymbol{\beta}}_{\bullet j} = \hat{\alpha}_j \cdot \mathbf{1}_q$ *with arbitrary* $\hat{\alpha}_j \in \mathbb{R}$*, then* ***the global effect*** $\hat{\alpha}_j$ ***is lasso-regularized:***

i) The zero effect condition (9) *for* $\hat{\alpha}_j = 0$ *simplifies to*

$$\left|\mathbf{1}_q^\top \boldsymbol{s}_j\right| = |s(\hat{\alpha}_j)| \leq \lambda\sqrt{q}. \tag{12}$$

ii) Otherwise, $\hat{\alpha}_j$ *satisfies*

$$s(\hat{\alpha}_j) - \lambda\sqrt{q}\,\mathrm{sgn}(\hat{\alpha}_j) = \sum_{t=1}^{q} \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_{tj}}\bigg|_{\hat{\beta}_{tj}=\hat{\alpha}_j} - \lambda\sqrt{q}\frac{\hat{\alpha}_j}{|\hat{\alpha}_j|} = 0. \tag{13}$$

Note that the conditions given in Theorem 1 must hold simultaneously for all penalized parameter groups and are in general connected with each other through the nonlinear score function, which is affected by the shrinkage that depends on both λ and ζ. A zero effect can be obtained even if condition (10) is not satisfied, but if (10) holds, the ETL penalty applies a conventional lasso shrinkage with tuning parameter $\lambda\sqrt{q}$ as described in Theorem 1d). Since $s(\hat{\alpha}_j) = \sum_{t=1}^{q} \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_{tj}}\Big|_{\beta_{tj}=\hat{\alpha}_j}$, the implicit weighting factor $\sqrt{q}$ on global effects is appropriate (cf. Yuan & Lin, 2006).

Moreover, if $\lambda = 0$, which means that one is only interested in effect type selection, but not in variable selection, the condition in (13) simply becomes the ML equation $s(\hat{\alpha}_j) = 0$. Note, however, that the score function in that case still depends on the penalized estimates of the other parameters. The same is true for the unpenalized intercept parameters $\hat{\boldsymbol{\beta}}_{\bullet 0}$, which were left aside in Theorem 1 for the sake of simplicity.

Concerning the range of ζ-values that yield different models, the following can easily be derived from Theorem 1:

Corollary 1. *Let* $l^{global}(\boldsymbol{\beta}_{\bullet 0}, \boldsymbol{\alpha})$ *denote the loglikelihood of the model from* (2)*, that is, the model with all covariate effects a priori specified as global. For fixed* λ*, let*

$$(\hat{\boldsymbol{\beta}}_{\bullet 0}^{\lambda}, \hat{\boldsymbol{\alpha}}^{\lambda}) = \underset{(\boldsymbol{\beta}_{\bullet 0}, \boldsymbol{\alpha})}{\mathrm{argmax}} \left(l^{global}(\boldsymbol{\beta}_{\bullet 0}, \boldsymbol{\alpha}) - \sum_{j=1}^{p} \lambda\sqrt{q}\,|\alpha_j| \right)$$

denote the lasso-penalized estimator for this global-effects model. Then, with $\otimes$ *denoting the Kronecker product, the same global-effects model is also obtained from the ETL equation* (7) *when using the same* λ *and any* $\zeta \geq \zeta^{max}(\lambda)$*, where*

$$\zeta^{max}(\lambda) = \max_{j=1,\ldots,p} \left|\left| (\boldsymbol{D}\boldsymbol{D}^\top)^{-1}\boldsymbol{D} \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\bullet j}}\bigg|_{\boldsymbol{\beta}=\left(\hat{\boldsymbol{\beta}}_{\bullet 0}^{\lambda},\ \hat{\boldsymbol{\alpha}}^{\lambda}\otimes \mathbf{1}_q\right)} \right|\right|_2. \tag{14}$$

3.4. Effective Degrees of Freedom and Tuning Parameter Selection

The ETL penalty depends on two tuning parameters, λ that steers variable selection, and ζ that steers selection of category-specific versus global effects, and must be tuned over a two-dimensional grid of possible (λ, ζ)-combinations.

Since the lasso-penalized estimator of the global-effects model can be computed easily, Corollary 1 provides an efficient way to determine suitable sequences of ζ-values given a sequence of λ-values. A similar result for the value $\lambda^{\max}$ (so that the null model is obtained for any $\lambda \geq \lambda^{\max}$) could be derived from (9), but since determining the optimal tradeoff parameters $\hat{\tau}_j^*$ according to (8) is computationally expensive, we suggest to use the more simple, well-known rule

$$\lambda^{\max} = \max_{j=1,\ldots,p} \left|\left| \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\bullet j}} \right|_{\boldsymbol{\beta}=(\hat{\beta}_{\bullet 0},\, \mathbf{0})} \right|\Big|_2,$$

which provides an upper bound on the range of relevant λ-values. Depending on ζ, this bound need not be tight, but it is computationally inexpensive and worked well in our empirical studies.

Given a list of estimated models for all considered (λ, ζ)-combinations, a concrete model is selected using a model selection criterion. The most common criteria are K-fold crossvalidation, the AIC and the BIC. Crossvalidation can be performed as usual, see, e.g., Hastie et al. (2009). The (approximated) AIC and BIC are given by

$$\widehat{\text{AIC}}(\hat{\boldsymbol{\beta}}) = -2l(\hat{\boldsymbol{\beta}}) + 2\widehat{\text{edf}}(\hat{\boldsymbol{\beta}}), \qquad \widehat{\text{BIC}}(\hat{\boldsymbol{\beta}}) = -2l(\hat{\boldsymbol{\beta}}) + \log(n)\widehat{\text{edf}}(\hat{\boldsymbol{\beta}}),$$

where n is the sample size and $\widehat{\text{edf}}(\hat{\boldsymbol{\beta}})$ denotes an estimate of the effective degrees of freedom of the considered model. With q unpenalized intercept parameters and p penalized parameter groups $\boldsymbol{\beta}_{\bullet j},\ j = 1, \ldots, p$, one has

$$\widehat{\text{edf}}(\hat{\boldsymbol{\beta}}) = q + \sum_{j=1}^{p} \widehat{\text{edf}}(\hat{\boldsymbol{\beta}}_{\bullet j}).$$

Let $\mathbb{1}$ denote the 0-1-indicator function. If only the variable selection penalty was used, that is, if $\zeta = 0$, Yuan & Lin (2006) argue that, for each parameter group $\boldsymbol{\beta}_{\bullet j}$, only

$$\widehat{\text{edf}}(\hat{\boldsymbol{\beta}}_{\bullet j}) = \mathbb{1}(||\hat{\boldsymbol{\beta}}_{\bullet j}||_2 > 0) \ + \ (q-1)\frac{||\hat{\boldsymbol{\beta}}_{\bullet j}||_2}{||\hat{\boldsymbol{\beta}}_{\bullet j}^{\text{ML}}||_2} \tag{15}$$

out of the q available degrees of freedom are effectively used by the estimate. Here, $\hat{\boldsymbol{\beta}}_{\bullet j}^{\text{ML}}$ denotes the ML estimate.

For each such parameter group, the effect type penalty reduces the effective degrees of freedom that are found in the $q-1$-dimensional vector of adjacent differences in the same fashion, although the shrinkage effects from the two different penalties partly overlap. Note the group lasso's shrinkage of the parameter group's $q-1$ degrees of freedom in (15) by a constant factor $||\hat{\boldsymbol{\beta}}_{\bullet j}||_2/||\hat{\boldsymbol{\beta}}_{\bullet j}^{\text{ML}}||_2$. Even if $\zeta = 0$, this group lasso induced shrinkage would reduce the quantity $||\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}||_2/||\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}^{\text{ML}}||_2$ by

the same factor. Thus, for $\zeta > 0$, the quantity $||\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}||_2/||\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}^{\mathrm{ML}}||_2$ gives the shrinkage factor that is applied to $q-2$ degrees of freedom by both the variable selection and the effect type penalty combined. Hence, we suggest the following formula to approximate the edf of ETL:

$$\widehat{\mathrm{edf}}(\hat{\boldsymbol{\beta}}_{\bullet j}) = \mathbb{1}(||\hat{\boldsymbol{\beta}}_{\bullet j}||_2 > 0) \; + \; \mathbb{1}(||\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}||_2 > 0)\frac{||\hat{\boldsymbol{\beta}}_{\bullet j}||_2}{||\hat{\boldsymbol{\beta}}_{\bullet j}^{\mathrm{ML}}||_2} \; + \; \max\big(0,(q-2)\big)\frac{||\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}||_2}{||\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}^{\mathrm{ML}}||_2}, \tag{16}$$

for $j = 1, \ldots, p$. Thus, if $\hat{\boldsymbol{\beta}}_{\bullet j} = \boldsymbol{0}$, one has $\widehat{\mathrm{edf}}(\hat{\boldsymbol{\beta}}_{\bullet j}) = 0$ and if a global effect is selected, which corresponds to $||\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}||_2 = 0$, one degree of freedom is obtained. If a category-specific effect is used, one of the remaining $q-1$ degrees of freedom is affected by the ridge-type shrinkage as in (15). If $q > 2$, the remaining $q-2$ degrees of freedom are affected by the combined shrinkage factor $||\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}||_2/||\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}^{\mathrm{ML}}||_2$.

If the ML estimator does not exist or is unstable, it can be replaced in the formulas above by a ridge estimate with small tuning parameter (cf. Tutz et al., 2015).

3.5. Related Concepts

The first term $||\boldsymbol{\beta}_{\bullet j}||_2$ in penalty (6) is similar in structure and concept to those recently proposed in Simon et al. (2013), Chen & Li (2013), Vincent & Hansen (2014) and Tutz et al. (2015) for multinomial logit models. We refer to these papers for a more thorough discussion of this penalty's behaviour.

In the following, we survey various penalization and regularization approaches in the literature that look similar to the second term in (6), the effect type penalty $\sqrt{\boldsymbol{\beta}_{\bullet j}^{\top}\boldsymbol{\Omega}\boldsymbol{\beta}_{\bullet j}}$, but are conceptually different. For example, Yuan & Lin (2006) defined the general group lasso penalty for a vector $\boldsymbol{u}$ as the matrix-weighted L_2-norm $||\boldsymbol{u}||_{\boldsymbol{M}} = \sqrt{\boldsymbol{u}^{\top}\boldsymbol{M}\boldsymbol{u}}$, with positive definite $\boldsymbol{M}$. Since our penalty matrix $\boldsymbol{\Omega}$ only has rank $q-1$, our effect type penalty does not technically match this definition. Moreover, elegant estimation via cholesky decomposition (see Huang et al., 2012) is not possible.

In Gertheiss et al. (2011), the focus is on the selection of ordinal predictors. A penalty looking very similar to our grouped fusion term is proposed for this purpose, but the first category of the ordinal variables is always specificed as the reference category with an implicit coefficient of zero. Therefore, the first rows of the difference matrix in Gertheiss et al. (2011) and in our $\boldsymbol{\Omega}$ differ. The consequence is that their penalty only enforces all parameter differences to be exactly zero when the parameters themselves are shrunk to zero aswell. Thus, it cannot perform effect type selection and just provides a more appropriate within-group shrinkage for selected ordinal predictors than the traditional group lasso. Additionally, the difference matrix employed by Gertheiss et al. (2011) allows for the elegant computation of numerical estimates by recoding the covariate values from dummy to split coding, which also crucially relies on the presence of a reference category. Despite optical similarities, Gertheiss et al. (2011) thus neither conceptualize effect type selection nor provide directly applicable tools for computing the ETL estimator.

In the context of spline-based high-dimensional additive models, Meier et al. (2009) considered variable selection combined with simultaneous smoothing, but their approach also cannot shrink

a parameter group to a constant but nonzero value. The same penalty is also used in Gertheiss et al. (2013) for variable selection in functional linear models. The approach in both papers cannot perform effect type selection and, again, relies on covariate retransformations for the computation of numerical estimates.

The only reference containing our effect type penalty that we found in the literature is Barbero & Sra (2011), who consider p-norm based discrete total variation penalties. Although motivated by applications in signal processing, their penalty is mathematically identical to our effect type penalty for one-dimensional signals and $p = 2$, which is referred to as the $\text{TV}_2^{1\text{D}}$-case. The exposition in Barbero & Sra (2011), however, focuses entirely on algorithmic solutions and neither conceptualizes effect type selection nor analyzes the properties of the resulting estimator. Empirical illustrations of the penalization *method* are missing in Barbero & Sra (2011). Moreover, only difference penalties, together with a quadratic loss function that uses no covariates, are considered. By contrast, we combine a difference penalty with a classical variable selection penalty and allow the negative loglikelihood of all types of ordinal regression models as loss function. Nonetheless, the algorithm in Barbero & Sra (2011) is a useful tool to solve a technical subproblem in our estimation algorithm, see Section 5.2.

4. Large Sample Properties of Effect Type Selection

In this section, we consider asymptotic properties of the Effect Type Lasso and, more generally, effect type selection in a large sample setting, that is, $n \to \infty$. Let $\boldsymbol{\beta}^*$ denote the true parameter vector and let $\hat{\boldsymbol{\beta}}$ denote an estimator of $\boldsymbol{\beta}^*$ which is computed using a sample size of n. In the following, let λ_n and ζ_n denote the tuning parameters that grow with n, let $l_n(\boldsymbol{\beta})$ denote the loglikelihood evaluated on a sample of size n and assume that the expected Fisher information $\boldsymbol{F}_n(\boldsymbol{\beta}) = \text{E}\left(-\frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}\right)$ of the true model converges to a positive definite limit $\boldsymbol{F}$: $\boldsymbol{F}_n(\boldsymbol{\beta}^*)/n \overset{n\to\infty}{\to} \boldsymbol{F}(\boldsymbol{\beta}^*)$.

The following theorem shows that simple consistency is obtained whenever the tuning parameters are kept fixed:

Theorem 2. *Suppose* $0 \leq \lambda < \infty$ *and* $0 \leq \zeta < \infty$ *have been fixed. Then, the estimate* $\hat{\boldsymbol{\beta}}$ *that maximizes* (7) *is consistent, that is,* $\lim_{n\to\infty} \mathbb{P}(||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2^2 > \epsilon) = 0$ *for all* $\epsilon > 0$.

As pointed out for the lasso in Zou (2006) and for the group lasso in Wang & Leng (2008), this is not enough to achieve selection consistency. In the following, we present an adaptively weighted ETLasso that possesses the oracle property in terms of effect type selection. Since we are considering a three-stage selection, the traditional notation has to be extended. We define the following index sets:

$$\begin{aligned}
\mathcal{A} &= \{j : \boldsymbol{\beta}^*_{\bullet j} \neq \mathbf{0}, \boldsymbol{D}\boldsymbol{\beta}^*_{\bullet j} \neq \mathbf{0}\}, & \mathcal{A}_n &= \{j : \hat{\boldsymbol{\beta}}_{\bullet j} \neq \mathbf{0}, \boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j} \neq \mathbf{0}\},\\
\mathcal{B} &= \{j : \boldsymbol{\beta}^*_{\bullet j} \neq \mathbf{0}, \boldsymbol{D}\boldsymbol{\beta}^*_{\bullet j} = \mathbf{0}\}, & \mathcal{B}_n &= \{j : \hat{\boldsymbol{\beta}}_{\bullet j} \neq \mathbf{0}, \boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j} = \mathbf{0}\},\\
\mathcal{C} &= \{j : \boldsymbol{\beta}^*_{\bullet j} = \mathbf{0}\}, & \mathcal{C}_n &= \{j : \hat{\boldsymbol{\beta}}_{\bullet j} = \mathbf{0}\}.
\end{aligned}$$

The sets $\mathcal{A}, \mathcal{B}$ and $\mathcal{C}$ are a disjoint partition of $\{0, 1, \ldots, p\}$ and correspond to the variables that have category-specific, global and zero effects, respectively. The sets $\mathcal{A}_n, \mathcal{B}_n$ and $\mathcal{C}_n$ contain the estimated effect types. Furthermore, let $\boldsymbol{\beta}^*_{\mathcal{A}}, \boldsymbol{\beta}^*_{\mathcal{B}}$ and $\boldsymbol{\beta}^*_{\mathcal{C}}$ denote the corresponding parameters, for example, $\boldsymbol{\beta}^*_{\mathcal{A}} = \{\beta^*_{tj} : j \in \mathcal{A}, t \in \{1, \ldots, q\}\}$, so that $\boldsymbol{\beta}^*_{\mathcal{A}}$ is a vector of length $q|\mathcal{A}|$. Without loss of generality, one hence obtains that

$$\boldsymbol{\beta}^* = \begin{pmatrix} \boldsymbol{\beta}^*_{\mathcal{A}} \\ \boldsymbol{\beta}^*_{\mathcal{B}} \\ \boldsymbol{\beta}^*_{\mathcal{C}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}^*_{\mathcal{A}} \\ \boldsymbol{\alpha}^* \otimes \mathbf{1}_q \\ \mathbf{0} \end{pmatrix}.$$

To achieve selection consistency, we propose the *adaptive ETL* penalty that is defined by

$$J_n(\boldsymbol{\beta}) = \sum_{j=1}^{p} \left(\lambda_n w_{1j} ||\boldsymbol{\beta}_{\bullet j}||_2 \; + \; \zeta_n w_{2j} ||\boldsymbol{D}\boldsymbol{\beta}_{\bullet j}||_2 \right), \tag{17}$$

with adaptive weights

$$w_{1j} = \frac{1}{||\hat{\boldsymbol{\beta}}^{\mathrm{ML}}_{\bullet j}||_2}, \qquad w_{2j} = \frac{\min\left(c, \; ||\hat{\boldsymbol{\beta}}^{\mathrm{ML}}_{\bullet j}||_2\right)}{c||\boldsymbol{D}\hat{\boldsymbol{\beta}}^{\mathrm{ML}}_{\bullet j}||_2}, \tag{18}$$

where $c > 0$ is a small constant. The first weight w_{1j} follows the ideas of Zou (2006) and Wang & Leng (2008). Due to the consistency of the ML estimator (in the considered large sample setting), w_{1j} becomes large when $\boldsymbol{\beta}^*_{\bullet j} = \mathbf{0}$ and otherwise converges to a constant value. Our construction of w_{2j} in (18) is novel and guarantees that $w_{2j} = \mathcal{O}_p(1)$ if $\boldsymbol{\beta}^*_{\bullet j} = \mathbf{0}$. As is seen from the proofs in Section A.3 in the appendix, this property of w_{2j} is necessary to obtain consistent effect type selection, which would not be ensured by usage of the "naive" adaptive weight $w_{2j} = 1/||\boldsymbol{D}\hat{\boldsymbol{\beta}}^{\mathrm{ML}}_{\bullet j}||$.

Using the weights from (18), the oracle property for effect type selection is indeed obtained:

Theorem 3. *Suppose* $\lambda_n/\sqrt{n}, \; \zeta_n/\sqrt{n} \to 0$ *and* $\lambda_n, \zeta_n \to \infty$. *Let* $\mathcal{D} = \mathcal{A} \cup \mathcal{B}$ *be the set of variables with nonzero effect. Then, the adaptive ETL estimator* $\hat{\boldsymbol{\beta}}$ *using penalty* (17) *with weights* (18) *satisfies*

(a) $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{D}} - \boldsymbol{\beta}^*_{\mathcal{D}}) \xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{F}^{-1}(\boldsymbol{\beta}^*_{\mathcal{D}})\right)$,

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

Part (b) of Theorem 3 implies that asymptotically, all variables are specified with the correct effect type. Since $\boldsymbol{\beta}^*_{\mathcal{B}} = \boldsymbol{\alpha}^* \otimes \mathbf{1}_q$, the entries of $\boldsymbol{F}^{-1}(\boldsymbol{\beta}^*_{\mathcal{D}})$ that belong to variables with a global effect have a block structure: $\mathrm{Cov}(\boldsymbol{\beta}^*_{\mathcal{B}}) = \boldsymbol{F}^{-1}(\boldsymbol{\beta}^*_{\mathcal{B}}) = \frac{1}{q}\boldsymbol{F}^{-1}(\boldsymbol{\alpha}^*) \otimes (\mathbf{1}_q \mathbf{1}_q^\top)$ and $\mathrm{Cov}(\boldsymbol{\beta}^*_{\mathcal{A}}, \boldsymbol{\beta}^*_{\mathcal{B}}) = \mathrm{Cov}(\boldsymbol{\beta}^*_{\mathcal{A}}, \boldsymbol{\alpha}^*) \otimes \mathbf{1}_q^\top$, where $\boldsymbol{F}^{-1}(\boldsymbol{\alpha}^*)$ is the inverse oracle Fisher matrix with respect to the global effects. Asymtotically, this structure of $\boldsymbol{F}^{-1}(\boldsymbol{\beta}^*_{\mathcal{D}})$ carries over to $\boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}}_{\mathcal{D}})$ with probability 1.

In the following, we show that the oracle property for effect type selection holds for a more

general class of penalties. We consider the following penalty term:

$$J_n(\boldsymbol{\beta}) = \sum_{j=1}^{p} \left(\lambda_n w_{1j} \rho(\boldsymbol{\beta}_{\bullet j}) \; + \; \zeta_n w_{2j} \rho(\boldsymbol{D}\boldsymbol{\beta}_{\bullet j}) \right) \tag{19}$$

where $\rho(\cdot)$ is a group penalty function with the following properties:

R1: For any input dimension d, $\rho : \mathbb{R}^d \to \mathbb{R}_0^+$ and ρ has its mimimum at $\mathbf{0}$: $\rho(\mathbf{0}) = 0 = \min_{\boldsymbol{\xi} \in \mathbb{R}^d} \rho(\boldsymbol{\xi})$.

R2: ρ is symmetric around $\mathbf{0}$: $\rho(\boldsymbol{\xi}) = \rho(|\boldsymbol{\xi}|) \;\; \forall \boldsymbol{\xi} \in \mathbb{R}^d$, where $|\boldsymbol{\xi}| = (|\xi_1|, \ldots, |\xi_d|)^\top$.

R3: ρ is continuously differentiable on $\mathbb{R}^d \setminus \{\mathbf{0}\}$ and $\rho'_t := \frac{\partial \rho(\boldsymbol{\xi})}{\partial \xi_t} \geq 0$ for any $\xi_t \in \mathbb{R}^+$.

R4: ρ satisfies

$$\liminf_{\boldsymbol{\xi} \to \mathbf{0}+} \left(\min_t \rho'_t \right) > 0. \tag{20}$$

Due to the symmetry assumption, (20) implies that ρ is nondifferentiable at $\mathbf{0}$ and thus performs selection of the quantity that is supplied as its argument (cf. Fan & Li, 2001).

It can be shown that the corresponding penalized estimator possesses the same oracle property as in Theorem 3 if the penalized negative loglikelihood is convex in a neighborhood around the true parameter vector $\boldsymbol{\beta}^*$:

Theorem 4. *Suppose $\mathcal{M}_n(\boldsymbol{\beta}) := -l_n(\boldsymbol{\beta}) + J_n(\boldsymbol{\beta})$ with penalty J_n from (19) has a unique minimum $\boldsymbol{\beta}^*$ and there exists an $\epsilon > 0$ such that $\mathcal{M}_n(\boldsymbol{\beta})$ is convex in $\mathcal{W}_\epsilon = \{\boldsymbol{\beta} : ||\boldsymbol{\beta} - \boldsymbol{\beta}^*|| < \epsilon\}$. Then, using the weights from (18), the results of Theorem 3 hold for the estimator $\hat{\boldsymbol{\beta}}$ that minimizes $\mathcal{M}_n$.*

Note that for data with small sample sizes, the ML estimator and thus the adaptive weights might (but need not) be unstable; or not exist at all. If one wants to use adaptive weights in this case, $\hat{\boldsymbol{\beta}}_{\bullet j}^{\mathrm{ML}}$ in (18) can be replaced by any $\sqrt{n}$-consistent estimator, for example the ridge estimator with a small (and asymptotically vanishing) λ_{ridge}, as is seen from the proof of Theorem 3 in the appendix.

5. Numerical Estimates and the Thresholding Operator of the ETL penalty

5.1. FISTA

To compute the penalized estimator $\hat{\boldsymbol{\beta}}$, the negative penalized loglikelihood $-l(\boldsymbol{\beta}) + J(\boldsymbol{\beta})$ has to be minimized, where $J(\boldsymbol{\beta})$ is the ETL penalty from (6) or (17). We suggest using the Fast Iterative Shrinkage Thresholding Algorithm (FISTA) of Beck & Teboulle (2009) for the optimization of this nonsmooth objective function. FISTA is an accelerated version of proximal gradient algorithms (for an overview, see Parikh & Boyd, 2013) that has quadratic convergence. It is based on the so-called proximal (or proximity) operator that, for a generic penalty $J(\cdot)$ with single tuning parameter λ and an arbitrary input vector $\boldsymbol{u}$, is defined by

$$\mathbf{Prox}_J(\boldsymbol{u} \,|\, \lambda) \;=\; \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\tfrac{1}{2} ||\boldsymbol{\beta} - \boldsymbol{u}||_2^2 + \lambda J(\boldsymbol{\beta}) \right). \tag{21}$$

Let $\nu^{(s)}$ be a stepsize for iteration s, let $a_0 = 0$ and $a_s = \left(1 + \sqrt{1 + 4a_{s-1}^2}\right)/2$ denote acceleration factors and let $\nabla l(\boldsymbol{\beta})$ denote the gradient of the loglikelihood, i.e. the score function. With this notation, iterations of FISTA are given by the following scheme ($s = 1, 2, \ldots$ until convergence):

$$\begin{aligned}
&\textbf{Extrapolate:} && \boldsymbol{\theta}^{(s)} = \hat{\boldsymbol{\beta}}^{(s)} + \frac{a_{s-1} - 1}{a_s}(\hat{\boldsymbol{\beta}}^{(s)} - \hat{\boldsymbol{\beta}}^{(s-1)}) \\
&\textbf{Search point:} && \boldsymbol{u}^{(s)} = \boldsymbol{\theta}^{(s)} + \tfrac{1}{\nu^{(s)}} \nabla l(\boldsymbol{\theta}^{(s)}) \\
&\textbf{Prox operator:} && \hat{\boldsymbol{\beta}}^{(s+1)} = \textbf{Prox}_J \left(\boldsymbol{u}^{(s)} \;\middle|\; \lambda^{(s)} = \tfrac{\lambda}{\nu^{(s)}} \right)
\end{aligned}$$

The search point $\boldsymbol{u}^{(s)}$ can be seen as a gradient step from the current estimate towards the ML estimate. Performing this step from the extrapolated point $\boldsymbol{\theta}^{(s)}$ instead of $\hat{\boldsymbol{\beta}}^{(s)}$ is the difference between FISTA and basic proximal gradient algorithms and allows for quadratic convergence. Applying the proximal operator to the search point $\boldsymbol{u}^{(s)}$ incorporates the nonsmooth penalty and yields sparse estimates. For technical details like convergence checks or a line search for $\nu^{(s)}$, we refer to Beck & Teboulle (2009) or, for the specific case of penalized loglikelihood estimation, to Tutz et al. (2015).

Using FISTA for the computation of ETL estimates requires three building blocks: computing the loglikelihood $l(\boldsymbol{\beta})$ and its gradient $\nabla l(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$, which can be carried out via the representation of ordinal models as multivariate GLMs as described in Section 2.1, and evaluating the proximal operator of the ETL penalty, which is considered next.

5.2. The Proximal Operator for ETL

Note that the block-separable structure of the ETL penalty, coupled with the atomic structure of the L_2^2-term in (21), allows to compute the proximal operator for ETL separately for each parameter group $\boldsymbol{\beta}_{\bullet j}$. Therefore, the key to computing ETL estimates with proximal gradient algorithms is being able to evaluate the associated proximal operator for one parameter group. Dropping the subscripts, this problem has, for an arbitrary input vector $\boldsymbol{u} \in \mathbb{R}^q$, the following form:

$$\textbf{Prox}_{\text{ETL}}(\boldsymbol{u} \mid \lambda, \zeta) \;=\; \underset{\boldsymbol{\beta} \in \mathbb{R}^q}{\operatorname{argmin}} \; \tfrac{1}{2}||\boldsymbol{\beta} - \boldsymbol{u}||_2^2 + \lambda||\boldsymbol{\beta}||_2 + \zeta||\boldsymbol{D}\boldsymbol{\beta}||_2.$$

The lemma below shows that the prox operator of the combined penalty can be computed by successively evaluating the prox operators of the two involved penalties:

Lemma 1. *With* $(\boldsymbol{v})_+ = \max(\boldsymbol{v}, \boldsymbol{0})$ *and* $\tilde{\boldsymbol{\beta}} = \boldsymbol{Prox}_{ETL}(\boldsymbol{u} \mid 0, \zeta)$ *, one has*

$$\begin{aligned}
\boldsymbol{Prox}_{ETL}(\boldsymbol{u} \mid \lambda, \zeta) \;&=\; \boldsymbol{Prox}_{ETL}\left(\boldsymbol{Prox}_{ETL}(\boldsymbol{u} \mid 0, \zeta) \;\middle|\; \lambda, 0 \right) \\
&=\; \boldsymbol{Prox}_{ETL}\left(\tilde{\boldsymbol{\beta}} \mid \lambda, 0 \right) \;=\; \left(1 - \frac{\lambda}{||\tilde{\boldsymbol{\beta}}||_2} \right)_+ \tilde{\boldsymbol{\beta}}.
\end{aligned}$$

The main technical difficulty therefore lies in computing the solution of the following problem:

$$\tilde{\boldsymbol{\beta}} \;:=\; \mathbf{Prox}_{\mathrm{ETL}}(\boldsymbol{u}|\,\lambda{=}0,\,\zeta) \;=\; \underset{\boldsymbol{\beta}\in\mathbb{R}^q}{\operatorname{argmin}}\; \tfrac{1}{2}||\boldsymbol{\beta}-\boldsymbol{u}||_2^2 + \zeta||\boldsymbol{D}\boldsymbol{\beta}||_2. \tag{22}$$

This problem has been analyzed by Barbero & Sra (2011), who developed and implemented a very efficient algorithm from a dual formulation that runs in $\mathcal{O}(q)$ time and is publicly available. Our implementation is based on the package `MRSP` (Pößnecker, 2014) for the statistical software `R` (R Development Core Team, 2014) and uses aforementioned implementation of the Barbero & Sra (2011) algorithm for solving subproblem (22).

6. Real Data Application: The Munich Founder Study

To illustrate the Effect Type Lasso on real data, we consider a study on the survival of newly founded firms in Munich, Germany, that analyzes risk factors for the survival of such startups. Over a period of three years, data from $n = 1224$ business founders in and around Munich were collected. The firms' survival time, defined as the time to bankruptcy, is measured in intervals of six months, so, e.g., $y_i = 3$ means bankcruptcy occured between 12 and 18 months after the firm went into operation. Firms that are still in business after three years are considered to have survived their startup phase and are therefore pooled in a seventh response category.

To model the survival of the newly founded firms, 14 explanatory variables are available, for example the firm's starting capital, number of employees and clientele, but also information about the company founder, e.g. age and professional experience. The variables and their coding are summarized in Table 1. Since most covariates are categorical and are therefore dummy-coded, one obtains a model with $p = 21$.

We fitted a sequential logit model with category-specific effects and the adaptive ETL penalty from (17). The quantity $\mathbb{P}(Y = t|Y \geq t, \boldsymbol{x})$ can here be interpreted as a discrete hazard rate. Hence, category-specific and global effects here correspond to time-varying and (time)-constant effects, respectively. The tuning parameters λ and ζ, and thus the final model, are chosen via 10-fold crossvalidation over a two-dimensional grid.

The parameter estimates for the model are given in Table 2. Five variables/nine dummies are estimated to have a category-specific effect on the discrete hazard rate, five variables/six dummies plus *age* are equipped with a global effect and four variables/five dummies have been entirely removed from the model. To give an example for interpretation, the odds between failure and staying in business within a particular time period ceteris paribus decrease by a factor of $e^{-0.12} = 0.88$ for every ten years of age of the company founder. Since *age* has a global effect, this interpretation holds for every considered time period. By contrast, the presence of *debt* capital increases the hazard rate by a factor of $e^{0.4} = 1.49$ during the first period, but its effect steadily decreases over time. It is natural that starting conditions like *debt* have an effect that wears off over time. In the last two periods, startups that used debt capital ceteris paribus even have a lower risk of failure than those that did not raise external investment. The mostly negative coefficients for the number

Table 1: Description of the Munich Founder Data

Variable	Description	Coding
sector	Economic sector	1: industry, manufacturing and building sector 2: commerce 3: service industry
legal	Legal Form	1: small trade 2: one-man business 3: limited company (German: GmbH) 4: general partnership (German: OHG)
loc	Location	1: business or industrial district (0: residential district)
takeover	Type of foundation	1: firm is resulting from take-over (0: firm is entirely a startup)
secondary	Type of occupation	1: firm is secondary occupation of founder (0: main occupation)
startcap	Starting capital	1: startcap $\leq$ 20000 € 2: 20000 € $<$ startcap $\leq$ 60000 € 3: 60000 € $<$ startcap
debt	Usage of debt capital	1: yes (0: no)
national	Target market	1: national (0: local)
generalist	Clientele	1: broad customer market (0: few important customers)
degree	Degree of founder	1: none 2: High School 3: University or *Meister* (tertiary craftsmen degree)
male	Sex of founder	1: male (0: female)
profexp	Professional experience	1: profexp $<$ 10 years 2: 10 years $\leq$ profexp $<$ 20 years 3: 20 years $<$ profexp
employee	Number of employees	1: none or one 2: two or three 3: more than three
age	Age of founder	age in years (metric)

of *employees* imply that larger firms tend to be at a lower risk of failure than smaller ones.

To assess the uncertainty associated with effect type selection, we performed a nonparametric bootstrap ($B = 1000$) and, for each variable, computed the probabilities for the three possible effect types which are found in the last three columns of Table 2. The probability that corresponds to the estimated effect type is printed in italic. All variables with a nonzero effect on the firms' survival show a probability of at least 62% to be equipped with the estimated effect type. Among the variables that are estimated to have no effect, the uncertainty with regard to effect type specification tends to be higher. In particular, for modeling the influence of *professional experience*, all possible effect type specifications seem to be viable.

For illustration, Figure 1 shows the coefficient buildups for the variables *startcap2* and *generalist* against (logarithmized) λ and ζ, with the respective other tuning parameter fixed at its optimal value. The dashed vertical line indicates the optimal value of the tuning parameter which is varied in the plot. In Figure 1a, λ is fixed at the CV-optimal value $\lambda^* = 0.408$, for which both considered variables enter the model with a nonzero effect. It is seen that the ETLasso shrinks the variables' effect to a global, nonzero value when ζ becomes large. For the CV-optimal $\zeta^* = 7.348$ (vertical dashed line), variable *startcap2* retains its category-specific effect while a global effect is selected for *generalist*. Figure 1b shows the corresponding coefficient paths against λ. Since *generalist* is assigned a global effect for the fixed $\zeta^* = 7.348$, the buildup against λ only shows one line which corresponds to this global effect. As shown in Theorem 1, this bottom part of Figure 1b represents

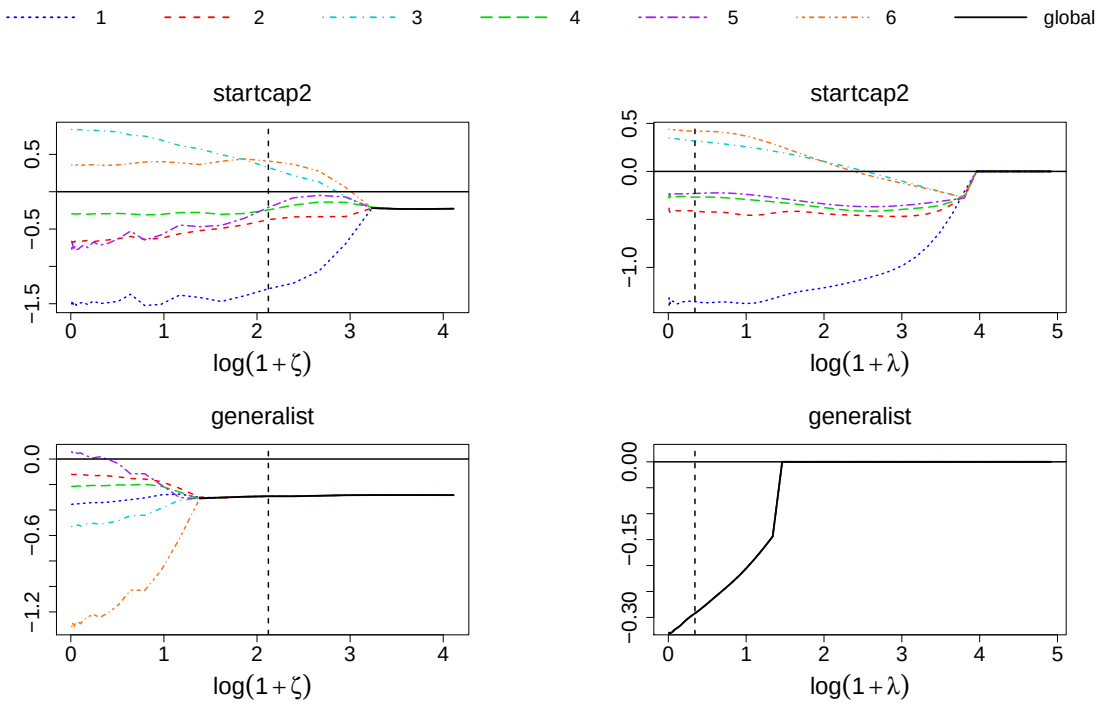
Table 2: ETL parameter estimates and selection probabilities for the Munich Founder Data

	Estimated parameters for time interval						*Probability for effect type*		
	1	**2**	**3**	**4**	**5**	**6**	**cat.-spec.**	**global**	**zero**
Intercept	-2.01	-2.14	-2.75	-2.62	-3.10	-2.72	*1.00*	0	0
legal2	-0.35	-0.39	0.15	-0.29	0.20	-0.95	*1.00*	0	0
legal3	-1.47	-1.03	-1.11	-1.30	-0.69	-1.10	*1.00*	0	0
legal4	-0.40	-0.26	0.16	0.04	0.85	-0.19	*1.00*	0	0
secondary	-0.36	0.04	0.20	0.46	0.31	0.75	*0.96*	0.03	0.01
startcap2	-1.30	-0.37	0.32	-0.25	-0.21	0.41	*1.00*	0	0
startcap3	-1.98	-0.69	-0.30	-0.07	-0.56	-0.27	*1.00*	0	0
debt	0.40	0.29	0.14	0.05	-0.09	-0.23	*0.68*	0.21	0.12
employee2	-0.23	0.19	-0.18	-0.11	-0.38	-0.58	*0.99*	0	0
employee3	-0.62	-0.05	-0.45	-0.41	0.03	-1.03	*0.99*	0	0
sector2				0.63			0.06	*0.94*	0
sector3				0.62			0.06	*0.94*	0
tm				-0.18			0.01	*0.74*	0.25
generalist				-0.29			0.01	*0.90*	0.09
degree2				-0.12			0.16	*0.64*	0.21
degree3				-0.21			0.16	*0.64*	0.21
age/10				-0.12			0.24	*0.62*	0.14
loc				0			0	0.40	*0.60*
takeover				0			0	0.15	*0.85*
male				0			0	0.11	*0.89*
profexp2				0			0.25	0.29	*0.46*
profexp3				0			0.25	0.29	*0.46*

an ordinary lasso path for a single variable. For *startcap2*, the coefficient paths in Figure 1b join to yield a global effect before it is shrunk to zero. This occurs because the shrinkage induced by the plain L_2-term $||\boldsymbol{\beta}_{\bullet j}||_2$ in our penalty also decreases the differences in the vector $\boldsymbol{D}\boldsymbol{\beta}_{\bullet j}$, so that an increasing λ can, for fixed ζ, affect whether the global effect condition (10) is satisfied or not.

Figure 2 shows the mean cross-validated deviance surface against the (λ, ζ)-grid. The plateau for $\log(\lambda) \geq 4$ corresponds to the null model. The peak at $\log(\lambda, \zeta) = (-4, -4)$ corresponds (approximately) to the ML estimator for the model with category-specific effects for all variables. This model's CV performance is close to that of the null model, which indicates severe overparameterization. From this peak, increasing the regularization in both directions improves the CV criterion, so that the CV surface forms a "curved hill". At the bottom of this hill lies a relatively flat "valley" in which the optimal model is located at $\log(\lambda, \zeta) = (-0.9, 2)$. For $\log(\zeta) \geq 3.2$, models with only global effects are obtained, for example, the ML estimator for the model with only global effects is found at (approximately) $\log(\lambda, \zeta) = (-4, 3.2)$. Note, in particular, that a range of ζ-values around $\log(\lambda, \zeta) = (-4, 2)$ exists that corresponds to models which use all variables, but with mixed effect types, and that are superior to both the unpenalized global and unpenalized category-specific model. This emphasizes the superiority of models with mixed effect type over modeling variants which are a priori restricted to one effect type - the ETLasso allows the researcher to find these superior models.

To quantify the differences in performance, Table 3 summarizes various performance measures of the ML- and the Lasso-estimator for the models with a priori global and category-specific effects, and our model that starts with category-specific effects and uses the ETL penalty. The penalized



(a) Coefficient buildups against ζ for fixed $\lambda^* = 0.408$.

(b) Coefficient buildups against λ for fixed $\zeta^* = 7.348$.

Figure 1: Coefficient buildups for two variables of the Munich Founder Data.

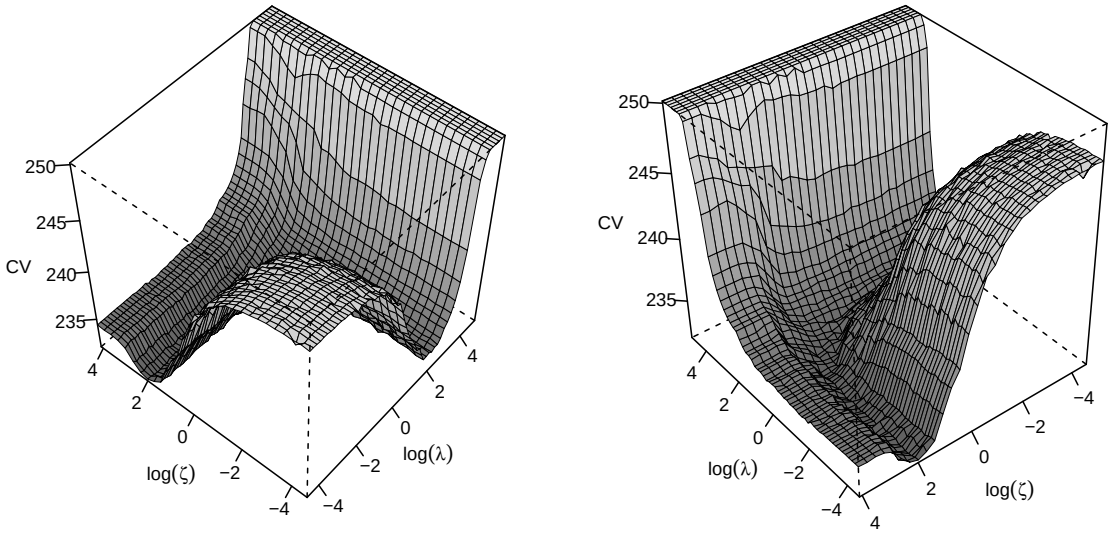


Figure 2: CV surface against tuning parameter grid, from two different angles.

model with only global effects is performing worse than its ML counterpart, so that an analysis which a priori assumes global effects would have concluded that a lasso penalty is not helpful

for the Munich Founder Data. As expected, the ML estimator of the model with category-specific effects yields the best fit to the data, but is excessively complex and outperformed in terms of CV, AIC and BIC. The category-specific model with a Grouped Lasso penalty is an improvement over the first three models, but is clearly outperformed by the ETL-based model in terms of fit, CV and AIC while the BICs of these two models are close.

Table 3: Performance criteria of modeling variants for the Munich Founder Data.

Model	**Deviance**	**edf**	**CV**	**AIC**	**BIC**
global, ML	2298.41	30	235.13	2358.41	2511.71
global, Lasso	2405.89	18	239.27	2441.89	2533.87
cat.-spec., ML	2110.85	150	244.99	2410.85	3177.33
cat.-spec., Grouped Lasso	2248.58	29.01	235.38	2306.61	*2454.87*
cat.-spec, Effect Type Lasso	2210.19	35.24	**231.92**	**2280.68**	*2460.76*

7. Concluding Remarks

In this paper, the problem of effect type selection in ordinal regression is considered. We have argued in favor of allowing a mix of both global and category-specific effects. In a real data application, it was shown that such a model with mixed effect type can indeed be superior to all modeling variants which are a priori restricted to one effect type. To solve the task of effect type selection, we have proposed a novel penalty approach, called the "*effect type lasso*" (ETL). The ETL penalty is constructed to also perform classical variable selection on top of effect type selection. Optimality conditions for the ETL estimator were given and it was shown that a global effects only model with a conventional lasso penalty is included as a special case within the ETL framework. An algorithm for the computation of ETL estimates was presented and the asymptotic properties of the ETL estimator were investigated. It was shown that an adaptively weighted version of our ETL estimator asymptotically yields consistent variable and effect type selection and that it possesses the oracle property. These asymptotic results were extended to a general family of effect type penalties. The real data application illustrates the selection and shrinkage behavior of our penalty. Moreover, it demonstrates that the ETL approach can allow the researcher to find, in an automated and data-driven fashion, a model that is superior to standard modeling approaches for ordinal regression and that is as parsimonious as possible and as flexible as necessary.

In future research, a wider class of established group penalties (Huang et al., 2012) could be used within the general effect type selection framework that was formalized in Section 4 (see (19) for reference). A comparison of the resulting, possibly concave effect type penalties with our convex ETL penalty could be interesting. Another possible direction of future research is to apply the concept of the effect type lasso to other models than ordinal regression. For example, in finite mixture models, which are also known as clusterwise regression, covariate effects can be

cluster-specific or fixed across mixture components (for an overview, see McLachlan & Peel, 2000). By penalizing all pairwise parameter differences within a parameter group instead of only the adjacent ones, the basic idea of the effect type lasso can be transferred to finite mixture models. This will, however, lead to an even more complex penalty term and the estimation algorithm will have to be modified accordingly.

Acknowledgements

This work was partially supported by DFG project "Regularisierung für diskrete Datenstrukturen". We sincerely thank David Drießlein for his assistance with data preparation and for testing early versions of our R implementation.

A. Proofs

A.1. Proof of Theorem 1

Due to the block-seperability of the ETL penalty, the condition $\mathbf{0} \in \nabla l(\hat{\boldsymbol{\beta}}) - \partial J(\hat{\boldsymbol{\beta}})$ can be examined group-wise. Note, however, that the optimality conditions can only be *stated groupwise*, but have to *hold simultaneously* for all groups since they are connected with each other through the nonlinear score function. With $\boldsymbol{s}_j := \boldsymbol{s}(\hat{\boldsymbol{\beta}}_{\bullet j}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\bullet j}}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$ denoting the score vector of the j-th parameter group, the estimates for all penalized parameter groups must satisfy

$$\mathbf{0} \in \boldsymbol{s}_j - \lambda \mathbf{v}_{1j} - \zeta \boldsymbol{D}^\top \mathbf{v}_{2j}, \qquad j = 1, \ldots, p,$$

where

$$\mathbf{v}_{1j} = \begin{cases} \text{any } \mathbf{v}_1 \in \mathbb{R}^q \text{ with } ||\mathbf{v}_1|| \leq 1 & \text{if } \hat{\boldsymbol{\beta}}_{\bullet j} = \mathbf{0}, \\ \frac{\hat{\boldsymbol{\beta}}_{\bullet j}}{||\hat{\boldsymbol{\beta}}_{\bullet j}||} & \text{if } \hat{\boldsymbol{\beta}}_{\bullet j} \neq \mathbf{0}, \end{cases}$$

and

$$\mathbf{v}_{2j} = \begin{cases} \text{any } \mathbf{v}_2 \in \mathbb{R}^{q-1} \text{ with } |||\mathbf{v}_2|| \leq 1 & \text{if } \boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j} = \mathbf{0}, \\ \frac{\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}}{||\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}||} & \text{if } \boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j} \neq \mathbf{0}. \end{cases} \tag{23}$$

Here and in the following, $||\cdot||$ denotes the L_2-norm and $|\cdot|$ denotes the absolute value. The formula for $\mathbf{v}_{2j}$ follows from the chain rule for subdifferentials. Part c) of Theorem 1 follows immediately in the case that a category-specific effect is obtained, i.e. $\hat{\boldsymbol{\beta}}_{\bullet j} \neq \mathbf{0} \ \wedge \ \boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j} \neq \mathbf{0}$. To derive the optimality and shrinkage conditions given in Theorem 1a,b & d), we first consider only the cases with λ or ζ at zero, i.e. only one active penalty, and turn to the general case afterwards.

Case I (only Variable Selection): $\lambda > 0$ and $\zeta = 0$:

A condition for $\hat{\boldsymbol{\beta}}_{\bullet j} = \mathbf{0}$ can be derived as follows: This case is only possible if $\mathbf{0} \in \boldsymbol{s}_j - \lambda \mathbf{v}_1$ can be satisfied with $||\mathbf{v}_1|| \leq 1$. Obviously, $\mathbf{v}_1$ must point in the same direction as $\boldsymbol{s}_j$, so that we can set $\mathbf{v}_1 = \frac{\boldsymbol{s}_j}{||\boldsymbol{s}_j||} \cdot \kappa_j$ with $\kappa_j \in [0, 1]$. Hence,

$$\boldsymbol{s}_j - \lambda \kappa_j \frac{\boldsymbol{s}_j}{||\boldsymbol{s}_j||} \stackrel{!}{=} \mathbf{0} \quad \Leftrightarrow \quad \lambda \kappa_j \frac{1}{||\boldsymbol{s}_j||} = 1 \quad \Leftrightarrow \quad ||\boldsymbol{s}_j||_2 \leq \lambda.$$

Otherwise, $\boldsymbol{s}_j - \lambda \frac{\hat{\boldsymbol{\beta}}_{\bullet j}}{||\hat{\boldsymbol{\beta}}_{\bullet j}||} = 0$.

Case II (only Effect Type Selection): $\lambda = 0$ and $\zeta > 0$:

First, note that, for any vector $\boldsymbol{u} \in \mathbb{R}^{q-1}$, one has $\mathbf{1}_q^\top \boldsymbol{D}^\top \boldsymbol{u} = \sum_{r=1}^{q} [\boldsymbol{D}^\top \boldsymbol{u}]_r = 0$. Since one can safely and w.l.o.g. assume $\boldsymbol{s}_j \neq \mathbf{0}$, the global effect condition $\mathbf{0} \in \boldsymbol{s}_j - \zeta \boldsymbol{D}^\top \mathbf{v}_2$ can therefore only be satisfied if $\mathbf{1}_q^\top \boldsymbol{s}_j = \sum_{r=1}^{q} s_{rj} = 0$, so that the arithmetic mean $\bar{s}_j$ of $\boldsymbol{s}_j$ must be zero. If a global effect is selected, that is, if $\hat{\boldsymbol{\beta}}_{\bullet j} = \hat{\alpha}_j \cdot \mathbf{1}_q$, the score function for $\hat{\alpha}_j$ is $\sum_{r=1}^{q} s(\hat{\beta}_{rj} = \hat{\alpha}_j) = s(\hat{\alpha}_j) \stackrel{!}{=} 0$, so this condition simply means that $\hat{\alpha}_j$ must be the ML estimator for α_j in the case considered here, that is, a global effect is selected and no variable selection penalty is applied. This determines the value the unpenalized parameter $\hat{\alpha}_j$ must take on. Now, to derive a condition for $\boldsymbol{D}\boldsymbol{\beta}_{\bullet j} = \mathbf{0}$, consider the optimality equation

$$\mathbf{0} = \boldsymbol{s}_j - \zeta \boldsymbol{D}^\top \mathbf{v}_2 \quad \Leftrightarrow \quad \boldsymbol{D}^\top \mathbf{v}_2 = \frac{\boldsymbol{s}_j}{\zeta} \qquad \text{s.t. } ||\mathbf{v}_2|| \leq 1.$$

For the moment, we ignore the norm bound on $\mathbf{v}_2$ to derive the direction in which $\mathbf{v}_2$ must point. Let $\boldsymbol{D}^{\top g}$ denote the generalized inverse of $\boldsymbol{D}^\top$. In absence of a constraint, the equation above can be solved if and only if $\boldsymbol{D}^\top \boldsymbol{D}^{\top g} \frac{\boldsymbol{s}_j}{\zeta} = \frac{\boldsymbol{s}_j}{\zeta}$ (James, 1978). As it turns out, $\boldsymbol{D}^{\top g} = (\boldsymbol{D}\boldsymbol{D}^\top)^{-1}\boldsymbol{D}$ and $\boldsymbol{D}^\top \boldsymbol{D}^{\top g} = \boldsymbol{D}^\top (\boldsymbol{D}\boldsymbol{D}^\top)^{-1}\boldsymbol{D} = \boldsymbol{I}_q - \frac{1}{q}\mathbf{1}_q\mathbf{1}_q^\top = \boldsymbol{C}$ is the symmetric and idempotent centering matrix for which one has $\boldsymbol{C}\boldsymbol{u} = \boldsymbol{u} - \bar{u} \cdot \mathbf{1}_q$ and $\mathbf{1}_q^\top \boldsymbol{C}\boldsymbol{u} = 0$. Hence, when ignoring the norm bound, a solution exists since $\bar{s}_j = 0$ is required anyway and can always be satisfied by choosing the unpenalized global effect $\hat{\alpha}_j$ accordingly. Moreover, one has $\boldsymbol{D}^{\top g}\boldsymbol{D}^\top = \boldsymbol{I}_{q-1}$, so that the solution to the equation above is unique and given by $\mathbf{v}_2 = \boldsymbol{D}^{\top g} \cdot \frac{\boldsymbol{s}_j}{\zeta} = (\boldsymbol{D}\boldsymbol{D}^\top)^{-1}\boldsymbol{D}\frac{\boldsymbol{s}_j}{\zeta}$ (James, 1978), which yields the direction in which $\mathbf{v}_2$ must point. To include the norm bound, we proceed analogously to Case I and set

$$\mathbf{v}_2 = \frac{(\boldsymbol{D}\boldsymbol{D}^\top)^{-1}\boldsymbol{D}\boldsymbol{s}_j}{||(\boldsymbol{D}\boldsymbol{D}^\top)^{-1}\boldsymbol{D}\boldsymbol{s}_j||} \cdot \kappa_j, \qquad \kappa_j \in [0, 1].$$

Hence, the condition for $\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j} = \mathbf{0}$ becomes

$$\begin{aligned}
\boldsymbol{s}_j - \zeta\kappa_j \frac{\boldsymbol{D}^\top(\boldsymbol{D}\boldsymbol{D}^\top)^{-1}\boldsymbol{D}\boldsymbol{s}_j}{||(\boldsymbol{D}\boldsymbol{D}^\top)^{-1}\boldsymbol{D}\boldsymbol{s}_j||} &\stackrel{!}{=} \mathbf{0} && \Leftrightarrow \\
\boldsymbol{C}\boldsymbol{s}_j + \bar{s}_j \cdot \mathbf{1}_q - \zeta\kappa_j \frac{\boldsymbol{C}\boldsymbol{s}_j}{||(\boldsymbol{D}\boldsymbol{D}^\top)^{-1}\boldsymbol{D}\boldsymbol{s}_j||} &= \mathbf{0} && \stackrel{\bar{s}_j=0}{\Leftrightarrow} \\
\boldsymbol{s}_j - \zeta\kappa_j \frac{\boldsymbol{s}_j}{||(\boldsymbol{D}\boldsymbol{D}^\top)^{-1}\boldsymbol{D}\boldsymbol{s}_j||} &= \mathbf{0} && \stackrel{\kappa_j\in[0,1]}{\Leftrightarrow} \\
||(\boldsymbol{D}\boldsymbol{D}^\top)^{-1}\boldsymbol{D}\boldsymbol{s}_j|| &\leq \zeta.
\end{aligned}$$

Otherwise, $\boldsymbol{s}_j - \zeta \frac{\boldsymbol{D}^\top \boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}}{||\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}||} = 0$.

Case III (simultaneous Variable and Effect Type Selection): $\lambda > 0$ and $\zeta > 0$:

In this case, a zero effect, that is, $\hat{\boldsymbol{\beta}}_{\bullet j} = \mathbf{0}$, is only possible if

$$\mathbf{0} \in \boldsymbol{s}_j - \lambda\mathbf{v}_1 - \zeta\boldsymbol{D}^\top\mathbf{v}_2 \qquad \text{s.t. } ||\mathbf{v}_1|| \leq 1 \text{ and } ||\mathbf{v}_2|| \leq 1$$

can be satisfied. Since the difference penalty cannot reduce an effect to zero, the shrinkage to zero must come from the $\lambda\mathbf{v}_1$ term. Hence, the $\zeta\boldsymbol{D}^\top\mathbf{v}_2$ term must be chosen optimally for this purpose. Using the arguments from Case I, one obtains the condition

$$\begin{aligned}
&||\boldsymbol{s}_j - \zeta\boldsymbol{D}^\top\mathbf{v}_2|| \leq \lambda \qquad \text{s.t. } ||\mathbf{v}_2|| \leq 1 \qquad \Leftrightarrow \\
&||\boldsymbol{s}_j - \zeta\boldsymbol{D}^\top\mathbf{v}_2^*|| \leq \lambda \qquad \text{with} \qquad \mathbf{v}_2^* = \underset{||\mathbf{v}_2||\leq 1}{\operatorname{argmin}} ||\boldsymbol{s}_j - \zeta\boldsymbol{D}^\top\mathbf{v}_2||.
\end{aligned}$$

The term to be minimized over the L_2-norm unit ball $||\mathbf{v}_2|| \leq 1$ can be simplified as follows:

$$||\boldsymbol{s}_j - \zeta\boldsymbol{D}^\top\mathbf{v}_2|| \;=\; 2\zeta\left(\tfrac{1}{2}\zeta\mathbf{v}_2^\top\boldsymbol{D}\boldsymbol{D}^\top\mathbf{v}_2 - \mathbf{v}_2^\top\boldsymbol{D}\boldsymbol{s}_j\right) + \text{const.}$$

Therefore, $\mathbf{v}_2^*$ is the solution to the quadratically constrained quadratic problem (QCQP)

$$\mathbf{v}_2^* = \underset{\mathbf{v}_2}{\operatorname{argmin}} \tfrac{1}{2}\zeta\mathbf{v}_2^\top\boldsymbol{D}\boldsymbol{D}^\top\mathbf{v}_2 - \mathbf{v}_2^\top\boldsymbol{D}\boldsymbol{s}_j \qquad \text{s.t. } ||\mathbf{v}_2|| \leq 1.$$

Following Boyd & Vandenberghe (2004), p. 197, the solution to this QCQP is

$$\mathbf{v}_2^* = (\zeta\boldsymbol{D}\boldsymbol{D}^\top + \hat{\tau}_j^*\boldsymbol{I}_{q-1})^{-1}\boldsymbol{D}\boldsymbol{s}_j,$$

with the optimal tradeoff parameter $\hat{\tau}_j^*$ determined as the maximum between zero and the largest solution $\hat{\tau}_j$ to the nonlinear equation

$$\boldsymbol{s}_j^\top\boldsymbol{D}^\top(\zeta\boldsymbol{D}\boldsymbol{D}^\top + \tau_j\boldsymbol{I}_{q-1})^{-2}\boldsymbol{D}\boldsymbol{s}_j \stackrel{!}{=} 1.$$

Putting the pieces together, this yields part a) of Theorem 1. Now, assume additionally that $||(\boldsymbol{D}\boldsymbol{D}^\top)^{-1}\boldsymbol{D}\boldsymbol{s}_j|| \leq \zeta$ is satisfied, which is equivalent to $\boldsymbol{s}_j^\top \boldsymbol{D}^\top (\zeta \boldsymbol{D}\boldsymbol{D}^\top)^{-2}\boldsymbol{D}\boldsymbol{s}_j \leq 1$. It is immediately seen that $\hat{\tau}_j \leq 0$ for this case, so that $\hat{\tau}_j^* = 0$ and the condition for a zero effect from above reduces to the one given in Theorem 1d i):

$$||\boldsymbol{s}_j - \boldsymbol{D}^\top(\boldsymbol{D}\boldsymbol{D}^\top)^{-1}\boldsymbol{D}\boldsymbol{s}_j|| = ||\boldsymbol{s}_j - \boldsymbol{C}\boldsymbol{s}_j|| = ||\bar{s}_j \cdot \mathbf{1}_q|| = \sqrt{q}\,|\bar{s}_j| = \left|\frac{\sum_{r=1}^q s_{rj}}{\sqrt{q}}\right| = \left|\frac{s(\hat{\alpha}_j = 0)}{\sqrt{q}}\right| \leq \lambda.$$

Next, we show Theorem 1d ii), which covers the case of variable x_j having a global, nonzero effect. This corresponds to $\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j} = \mathbf{0}$ and $\hat{\boldsymbol{\beta}}_{\bullet j} = \hat{\alpha}_j \cdot \mathbf{1}_q \neq \mathbf{0}$. In that case, the following must be satisfied:

$$\mathbf{0} \in \boldsymbol{s}_j - \lambda \frac{\hat{\alpha}_j \cdot \mathbf{1}_q}{||\hat{\alpha}_j \cdot \mathbf{1}_q||} - \zeta \boldsymbol{D}^\top \mathbf{v}_2 \qquad \text{s.t. } ||\mathbf{v}_2|| \leq 1.$$

Due to the same arguments as in Case II ($\mathbf{1}_q^\top \boldsymbol{D}^\top \mathbf{v}_2 = 0$ for any $\mathbf{v}_2$), and since one has $||\hat{\alpha}_j \cdot \mathbf{1}_q|| = \sqrt{q}|\hat{\alpha}_j|$, this condition can only be satisfied if

$$\sum_{r=1}^q \left(s(\hat{\beta}_{rj=}\hat{\alpha}_j) - \lambda \frac{\hat{\alpha}_j}{\sqrt{q}|\hat{\alpha}_j|} \right) = s(\hat{\alpha}_j) - \lambda\sqrt{q}\frac{\hat{\alpha}_j}{|\hat{\alpha}_j|} = 0,$$

which implies that lasso shrinkage of strength $\lambda\sqrt{q}$ is applied to the global effect α_j. Together with the condition $s(\hat{\alpha}_j) \leq \lambda\sqrt{q}$ for $\hat{\alpha}_j = 0$ from above, this means that global effects obtained with ETL act as if they were simply lasso-regularized. Now that we know which value $\hat{\alpha}_j$ must take on, a condition for ETL to select a global effect is obtained by proceeding as in Case II:

$$\left|\left|(\boldsymbol{D}\boldsymbol{D}^\top)^{-1}\boldsymbol{D}\left(\boldsymbol{s}_j - \lambda\frac{\hat{\alpha}_j \cdot \mathbf{1}_q}{\sqrt{q}|\hat{\alpha}_j|}\right)\right|\right| = ||(\boldsymbol{D}\boldsymbol{D}^\top)^{-1}\boldsymbol{D}\boldsymbol{s}_j|| \leq \zeta,$$

where the constant shrinkage term $\lambda\hat{\alpha}_j \cdot \mathbf{1}_q/(\sqrt{q}|\hat{\alpha}_j|)$ canceled out due to building adjacent differences. This shows part b) of Theorem 1 and therefore completes the proof. □

A.2. Proof of Theorem 2

Theorem 2 follows since fixed and finite tuning parameters mean that the penalty terms will asymptotically vanish in comparison to l_n. Since the ML estimator maximizes l_n, one obtains $\hat{\boldsymbol{\beta}} \xrightarrow{\mathbb{P}} \hat{\boldsymbol{\beta}}^{\text{ML}}$ and consistency of $\hat{\boldsymbol{\beta}}$ follows from consistency of the ML estimator which holds under mild regularity assumptions that are given, for example, in Fahrmeir & Kaufmann (1985). □

A.3. Proof of Theorem 3

Our proof builds on and extends ideas that have evolved from Zou (2006) and Bondell & Reich (2009). In our proof, new arguments with regard to the limit behavior of the penalty terms and to selection consistency have to be used because we are dealing with grouped penalties.

• **Redefinition of the Objective Function:** The objective function is $\mathcal{M}_n(\boldsymbol{\beta}) = -l_n(\boldsymbol{\beta}) + J_n(\boldsymbol{\beta})$. Since $\mathcal{M}_n(\boldsymbol{\beta}^*)$ is a constant for any n, minimization of $\mathcal{M}_n(\boldsymbol{\beta})$ is equivalent to minimizing

$$V_n(\boldsymbol{\beta}) = \mathcal{M}_n(\boldsymbol{\beta}) - \mathcal{M}_n(\boldsymbol{\beta}^*) = -(l_n(\boldsymbol{\beta}) - l_n(\boldsymbol{\beta}^*)) + \tilde{J}_n(\boldsymbol{\beta})$$

with the modified penalty

$$\begin{aligned}\tilde{J}_n(\boldsymbol{\beta}) &= J_n(\boldsymbol{\beta}) - J_n(\boldsymbol{\beta}^*) \\ &= \sum_{j=1}^{p} \frac{\lambda_n}{\sqrt{n}} \frac{\sqrt{n}}{||\hat{\boldsymbol{\beta}}_{\bullet j}^{\mathrm{ML}}||} \left(||\boldsymbol{\beta}_{\bullet j}|| - ||\boldsymbol{\beta}_{\bullet j}^*||\right) + \sum_{j=1}^{p} \frac{\zeta_n}{\sqrt{n}} \frac{\sqrt{n} \cdot \min\left(c,\ ||\hat{\boldsymbol{\beta}}_{\bullet j}^{\mathrm{ML}}||\right)}{c||\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}^{\mathrm{ML}}||} \left(||\boldsymbol{D}\boldsymbol{\beta}_{\bullet j}|| - ||\boldsymbol{D}\boldsymbol{\beta}_{\bullet j}^*||\right).\end{aligned}$$

Let $\boldsymbol{b} = \sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$ so that $\boldsymbol{\beta} = \boldsymbol{\beta}^* + \frac{\boldsymbol{b}}{\sqrt{n}}$ and optimizing $\boldsymbol{V}_n$ in terms of $\boldsymbol{\beta}$ or $\boldsymbol{b}$ is equivalent, where

$$\begin{aligned}\tilde{J}_n(\boldsymbol{\beta}) = \tilde{J}_n(\boldsymbol{b}) \ :=\ & \sum_{j=1}^{p} \frac{\lambda_n}{\sqrt{n}} \frac{\sqrt{n}}{||\hat{\boldsymbol{\beta}}_{\bullet j}^{\mathrm{ML}}||} \left(\left|\left|\boldsymbol{\beta}_{\bullet j}^* + \frac{\boldsymbol{b}_j}{\sqrt{n}}\right|\right| - ||\boldsymbol{\beta}_{\bullet j}^*||\right) \\ & + \sum_{j=1}^{p} \frac{\zeta_n}{\sqrt{n}} \frac{\sqrt{n} \cdot \min\left(c,\ ||\hat{\boldsymbol{\beta}}_{\bullet j}^{\mathrm{ML}}||\right)}{c||\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}^{\mathrm{ML}}||} \left(\left|\left|\boldsymbol{D}\boldsymbol{\beta}_{\bullet j}^* + \frac{\boldsymbol{D}\boldsymbol{b}_j}{\sqrt{n}}\right)\right|\right| - ||\boldsymbol{D}\boldsymbol{\beta}_{\bullet j}^*||\right).\end{aligned}$$

• **Limit Behavior:** Consider the limiting behavior of $\tilde{J}_n(\boldsymbol{b})$:
Case I (True structure is nonzero):
If $\boldsymbol{\beta}_{\bullet j}^* \neq \mathbf{0}$, one has $||\hat{\boldsymbol{\beta}}_{\bullet j}^{\mathrm{ML}}|| \xrightarrow{\mathbb{P}} ||\boldsymbol{\beta}_{\bullet j}^*|| > 0$. A first order taylor expansion of function $||\xi||$ at point $\xi = \boldsymbol{\beta}_{\bullet j}^* + \frac{\boldsymbol{b}_j}{\sqrt{n}}$ around $\xi_0 = \boldsymbol{\beta}_{\bullet j}^*$ yields, with $\xi - \xi_0 = \frac{\boldsymbol{b}_j}{\sqrt{n}}$, that

$$\left|\left|\boldsymbol{\beta}_{\bullet j}^* + \frac{\boldsymbol{b}_j}{\sqrt{n}}\right|\right| \ = \ ||\boldsymbol{\beta}_{\bullet j}^*|| + \frac{\boldsymbol{\beta}_{\bullet j}^{*\top} \boldsymbol{b}_j}{\sqrt{n}||\boldsymbol{\beta}_{\bullet j}^*||} + \mathcal{O}_p\left(\frac{\boldsymbol{b}_j^\top \boldsymbol{b}_j}{n}\right).$$

Hence, one obtains

$$\sqrt{n}\left(\left|\left|\boldsymbol{\beta}_{\bullet j}^* + \frac{\boldsymbol{b}_j}{\sqrt{n}}\right|\right| - ||\boldsymbol{\beta}_{\bullet j}^*||\right) \ = \ \frac{\boldsymbol{\beta}_{\bullet j}^{*\top} \boldsymbol{b}_j}{||\boldsymbol{\beta}_{\bullet j}^*||} + \mathcal{O}_p\left(\frac{\boldsymbol{b}_j^\top \boldsymbol{b}_j}{\sqrt{n}}\right) \ \xrightarrow{\mathbb{P}} \ \frac{\boldsymbol{\beta}_{\bullet j}^{*\top} \boldsymbol{b}_j}{||\boldsymbol{\beta}_{\bullet j}^*||}.$$

Since this term is linear in $\boldsymbol{b}_j$ and since $\lambda_n/\sqrt{n} \to 0$ by assumption, one obtains with Slutsky's Theorem that

$$\frac{\lambda_n}{\sqrt{n}} \frac{\sqrt{n}}{||\hat{\boldsymbol{\beta}}_{\bullet j}^{\mathrm{ML}}||} \left(\left|\left|\boldsymbol{\beta}_{\bullet j}^* + \frac{\boldsymbol{b}_j}{\sqrt{n}}\right|\right| - ||\boldsymbol{\beta}_{\bullet j}^*||\right) \xrightarrow{\mathbb{P}} 0.$$

If $\boldsymbol{D}\boldsymbol{\beta}_{\bullet j}^* \neq \mathbf{0}$, one has $||\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}^{\mathrm{ML}}|| \xrightarrow{\mathbb{P}} ||\boldsymbol{D}\boldsymbol{\beta}_{\bullet j}^*|| > 0$ and also $||\hat{\boldsymbol{\beta}}_{\bullet j}^{\mathrm{ML}}|| \xrightarrow{\mathbb{P}} ||\boldsymbol{\beta}_{\bullet j}^*|| > 0$. Again using a first order taylor expansion of function $||\xi||$ at point $\xi = \boldsymbol{D}\boldsymbol{\beta}_{\bullet j}^* + \frac{\boldsymbol{D}\boldsymbol{b}_j}{\sqrt{n}}$ around $\xi_0 = \boldsymbol{D}\boldsymbol{\beta}_{\bullet j}^*$ yields, with

$\xi - \xi_0 = \frac{\boldsymbol{D}\boldsymbol{b}_j}{\sqrt{n}}$ and $\boldsymbol{D}^\top \boldsymbol{D} = \boldsymbol{\Omega}$, that

$$\left|\left|\boldsymbol{D}\boldsymbol{\beta}^*_{\bullet j} + \frac{\boldsymbol{D}\boldsymbol{b}_j}{\sqrt{n}}\right|\right| \;=\; ||\boldsymbol{D}\boldsymbol{\beta}^*_{\bullet j}|| + \frac{\boldsymbol{\beta}^{*\top}_{\bullet j}\boldsymbol{\Omega}\boldsymbol{b}_j}{\sqrt{n}||\boldsymbol{D}\boldsymbol{\beta}^*_{\bullet j}||} + \mathcal{O}_p\left(\frac{\boldsymbol{b}_j^\top \boldsymbol{\Omega}\boldsymbol{b}_j}{n}\right).$$

Hence, one obtains

$$\sqrt{n}\Big(\left|\left|\boldsymbol{D}\boldsymbol{\beta}^*_{\bullet j} + \frac{\boldsymbol{D}\boldsymbol{b}_j}{\sqrt{n}}\right|\right| - ||\boldsymbol{D}\boldsymbol{\beta}^*_{\bullet j}||\Big) \;=\; \frac{\boldsymbol{\beta}^{*\top}_{\bullet j}\boldsymbol{\Omega}\boldsymbol{b}_j}{||\boldsymbol{D}\boldsymbol{\beta}^*_{\bullet j}||} + \mathcal{O}_p\left(\frac{\boldsymbol{b}_j^\top \boldsymbol{\Omega}\boldsymbol{b}_j}{\sqrt{n}}\right) \xrightarrow{\mathbb{P}} \frac{\boldsymbol{\beta}^{*\top}_{\bullet j}\boldsymbol{\Omega}\boldsymbol{b}_j}{||\boldsymbol{D}\boldsymbol{\beta}^*_{\bullet j}||}.$$

Since this term is again linear in $\boldsymbol{b}_j$ and since $\zeta_n/\sqrt{n} \to 0$ by assumption, one obtains with Slutsky's Theorem that

$$\frac{\zeta_n}{\sqrt{n}} \frac{\sqrt{n}\cdot\min\big(c,\ ||\hat{\boldsymbol{\beta}}^{\mathrm{ML}}_{\bullet j}||\big)}{c||\boldsymbol{D}\hat{\boldsymbol{\beta}}^{\mathrm{ML}}_{\bullet j}||}\Big(\left|\left|\boldsymbol{D}\boldsymbol{\beta}^*_{\bullet j} + \frac{\boldsymbol{D}\boldsymbol{b}_j}{\sqrt{n}}\right|\right| - ||\boldsymbol{D}\boldsymbol{\beta}^*_{\bullet j}||\Big) \xrightarrow{\mathbb{P}} 0.$$

Hence, the penalty on structures that are truly nonzero asymptotically converges to zero.
Case II (True structure is zero):
In case of $\boldsymbol{\beta}^*_{\bullet j} = \boldsymbol{0}$ or $\boldsymbol{D}\boldsymbol{\beta}^*_{\bullet j} = \boldsymbol{0}$, one has

$$\sqrt{n}\Big(\left|\left|\boldsymbol{\beta}^*_{\bullet j} + \frac{\boldsymbol{b}_j}{\sqrt{n}}\right|\right| - ||\boldsymbol{\beta}^*_{\bullet j}||\Big) = ||\boldsymbol{b}_j|| \qquad \text{or} \qquad \sqrt{n}\Big(\left|\left|\boldsymbol{D}\boldsymbol{\beta}^*_{\bullet j} + \frac{\boldsymbol{D}\boldsymbol{b}_j}{\sqrt{n}}\right|\right| - ||\boldsymbol{D}\boldsymbol{\beta}^*_{\bullet j}||\Big) = ||\boldsymbol{D}\boldsymbol{b}_j||.$$

It follows from the $\sqrt{n}$-consistency of the ML estimator that $\hat{\boldsymbol{\beta}}^{\mathrm{ML}} - \boldsymbol{\beta}^* = \mathcal{O}_p(n^{-1/2})$. Since $\lambda_n \to \infty$ and $\zeta_n \to \infty$ by assumption, it therefore follows that

$$\lim_{n\to\infty} \mathbb{P}\Big(\sqrt{n}||\hat{\boldsymbol{\beta}}^{\mathrm{ML}}_{\bullet j}|| \le \lambda^{1/2}\Big) = 1 \qquad \text{or} \qquad \lim_{n\to\infty} \mathbb{P}\Big(\sqrt{n}||\boldsymbol{D}\hat{\boldsymbol{\beta}}^{\mathrm{ML}}_{\bullet j}|| \le \zeta^{1/2}\Big) = 1.$$

Now, if $\boldsymbol{\beta}^*_{\bullet j} = \boldsymbol{0}$, then $\frac{\min\left(c,\ ||\hat{\boldsymbol{\beta}}^{\mathrm{ML}}_{\bullet j}||\right)}{c||\boldsymbol{D}\hat{\boldsymbol{\beta}}^{\mathrm{ML}}_{\bullet j}||} = \mathcal{O}(1)$, so that one obtains for $\boldsymbol{b}_j \neq \boldsymbol{0}$ that

$$\frac{\lambda_n}{\sqrt{n}}\frac{\sqrt{n}}{||\hat{\boldsymbol{\beta}}^{\mathrm{ML}}_{\bullet j}||}\Big(\left|\left|\boldsymbol{\beta}^*_{\bullet j} + \frac{\boldsymbol{b}_j}{\sqrt{n}}\right|\right| - ||\boldsymbol{\beta}^*_{\bullet j}||\Big) \xrightarrow{\mathbb{P}} \infty; \quad \frac{\zeta_n}{\sqrt{n}}\frac{\sqrt{n}\min\big(c, ||\hat{\boldsymbol{\beta}}^{\mathrm{ML}}_{\bullet j}||\big)}{c||\boldsymbol{D}\hat{\boldsymbol{\beta}}^{\mathrm{ML}}_{\bullet j}||}\Big(\left|\left|\boldsymbol{D}\boldsymbol{\beta}^*_{\bullet j} + \frac{\boldsymbol{D}\boldsymbol{b}_j}{\sqrt{n}}\right|\right| - ||\boldsymbol{D}\boldsymbol{\beta}^*_{\bullet j}||\Big) \xrightarrow{\mathbb{P}} 0.$$

Conversely, if $\boldsymbol{D}\boldsymbol{\beta}^*_{\bullet j} = \boldsymbol{0}$ while $\boldsymbol{\beta}^*_{\bullet j} \neq \boldsymbol{0}$, i.e. x_j has a global effect, one obtains for $\boldsymbol{D}\boldsymbol{b}_j \neq \boldsymbol{0}$ that

$$\frac{\lambda_n}{\sqrt{n}}\frac{\sqrt{n}}{||\hat{\boldsymbol{\beta}}^{\mathrm{ML}}_{\bullet j}||}\Big(\left|\left|\boldsymbol{\beta}^*_{\bullet j} + \frac{\boldsymbol{b}_j}{\sqrt{n}}\right|\right| - ||\boldsymbol{\beta}^*_{\bullet j}||\Big) \xrightarrow{\mathbb{P}} 0; \quad \frac{\zeta_n}{\sqrt{n}}\frac{\sqrt{n}\min\big(c, ||\hat{\boldsymbol{\beta}}^{\mathrm{ML}}_{\bullet j}||\big)}{c||\boldsymbol{D}\hat{\boldsymbol{\beta}}^{\mathrm{ML}}_{\bullet j}||}\Big(\left|\left|\boldsymbol{D}\boldsymbol{\beta}^*_{\bullet j} + \frac{\boldsymbol{D}\boldsymbol{b}_j}{\sqrt{n}}\right|\right| - ||\boldsymbol{D}\boldsymbol{\beta}^*_{\bullet j}||\Big) \xrightarrow{\mathbb{P}} \infty.$$

To sum up the limit behavior for this case, $\tilde{J}_n(\boldsymbol{b}_j) \to \infty$ whenever $\boldsymbol{b}_j \neq \boldsymbol{0}$ while $\boldsymbol{\beta}^*_{\bullet j} = \boldsymbol{0}$, or when $\boldsymbol{D}\boldsymbol{b}_j \neq \boldsymbol{0}$ while $\boldsymbol{D}\boldsymbol{\beta}^*_{\bullet j} = \boldsymbol{0}$.
• **Normality:** With $\mathcal{D} = \mathcal{A} \cup \mathcal{B}$ denoting be the set of variables with nonzero effect, we define, with a slight abuse of notation, the oracle ML estimator as $\hat{\boldsymbol{\beta}}^{\mathrm{ML}}_{\mathcal{D}} = (\hat{\boldsymbol{\beta}}^{\mathrm{ML}\top}_{\mathcal{A}}, (\hat{\boldsymbol{\alpha}}^{\mathrm{ML}} \otimes \boldsymbol{1}_q)^\top)^\top$, which is a priori

restricted to the truly nonzero effects and that a priori utilizes a global effect for those variables that truly have a global effect. Following Oelker et al. (2014) closely, expansion the oracle ML equation $\boldsymbol{s}_n(\hat{\boldsymbol{\beta}}_{\mathcal{D}}^{\text{ML}}) = \mathbf{0}$ around $\boldsymbol{\beta}_{\mathcal{D}}^*$ yields, in analogy to usual ML theory, that $\hat{\boldsymbol{\beta}}_{\mathcal{D}}^{\text{ML}} - \boldsymbol{\beta}_{\mathcal{D}}^* = \boldsymbol{F}_n^{-1}(\boldsymbol{\beta}_{\mathcal{D}}^*)\boldsymbol{s}_n(\boldsymbol{\beta}_{\mathcal{D}}^*) + \mathcal{O}_p(n^{-1})$ and hence, with $\boldsymbol{F}_n(\boldsymbol{\beta}_{\mathcal{D}}^*)/n \overset{n\to\infty}{\to} \boldsymbol{F}(\boldsymbol{\beta}_{\mathcal{D}}^*)$, one obtains that $n^{-1/2}\boldsymbol{s}_n(\boldsymbol{\beta}_{\mathcal{D}}^*) \xrightarrow{d} N\big(\mathbf{0}, \boldsymbol{F}(\boldsymbol{\beta}_{\mathcal{D}}^*)\big)$ and also $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{D}}^{\text{ML}} - \boldsymbol{\beta}_{\mathcal{D}}^*) \xrightarrow{d} N\big(\mathbf{0}, \boldsymbol{F}^{-1}(\boldsymbol{\beta}_{\mathcal{D}}^*)\big)$. Furthermore, it can be shown that $-2(l_n(\boldsymbol{\beta}) - l_n(\boldsymbol{\beta}^*)) = \boldsymbol{s}_n(\boldsymbol{\beta}^*)^\top \boldsymbol{F}_n^{-1}(\boldsymbol{\beta}^*)\boldsymbol{s}_n(\boldsymbol{\beta}^*) + \mathcal{O}_p(n^{-1})$.

Using these results combined with the redefined objective and the limit behavior from above, one obtains with Slutsky's Theorem that $V_n(\boldsymbol{\beta}) \xrightarrow{d} V(\boldsymbol{\beta})$ for every $\boldsymbol{\beta}$, where

$$V(\boldsymbol{\beta}) = \begin{cases} \frac{1}{2n}\boldsymbol{s}_n(\boldsymbol{\beta}_{\mathcal{D}})^\top \boldsymbol{F}^{-1}(\boldsymbol{\beta}_{\mathcal{D}})\boldsymbol{s}_n(\boldsymbol{\beta}_{\mathcal{D}}) & \text{if } \boldsymbol{\beta}_{\mathcal{B}} = \boldsymbol{\alpha} \otimes \mathbf{1}_q \text{ for some } \boldsymbol{\alpha} \in \mathbb{R}^{|\mathcal{B}|} \text{ and } \boldsymbol{\beta}_{\mathcal{C}} = \mathbf{0} \\ \infty & \text{otherwise} \end{cases}$$

and where $\boldsymbol{s}_n(\boldsymbol{\beta}_{\mathcal{D}})$ denotes the score vector of the oracle model. Since $V_n(\boldsymbol{\beta})$ is convex and the unique minimizer of $V(\boldsymbol{\beta})$ is $(\hat{\boldsymbol{\beta}}_{\mathcal{D}}^{\text{ML}}, \mathbf{0})^\top = (\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{ML}}, \hat{\boldsymbol{\alpha}}^{\text{ML}} \otimes \mathbf{1}_q, \mathbf{0})^\top$, it follows with similar arguments as in Zou (2006) that $\hat{\boldsymbol{\beta}}_{\mathcal{D}} \xrightarrow{d} \hat{\boldsymbol{\beta}}_{\mathcal{D}}^{\text{ML}}$. From the normality of the oracle ML estimator that was shown above, one eventually obtains $\sqrt{n}\big(\hat{\boldsymbol{\beta}}_{\mathcal{D}} - \boldsymbol{\beta}_{\mathcal{D}}^*\big) \xrightarrow{\text{d}} N\big(\mathbf{0}, \boldsymbol{F}^{-1}(\boldsymbol{\beta}_{\mathcal{D}}^*)\big)$.

• **Selection Consistency:**

– *Selection of the relevant effects* follows trivially: From the limit behavior, we know that the penalization on variables in $\mathcal{A}$ goes to zero. Moreover, if a variable x_j with $j \in \mathcal{B}$ is correctly estimated to have a global effect, this global effect is also unpenalized. It therefore remains to show that variables in $\mathcal{B}$ are assigned a global effect and that variables from $\mathcal{C}$ will be set to zero, both with probability 1.

– *Exclusion of zero effects:* We have to show $\lim_{n\to\infty} \mathbb{P}(j \in \mathcal{C}_n) = 1$ for all $j \in \mathcal{C}$, which is proven by contradiction. Assume there exists a $j \in \mathcal{C}$ with $j \notin \mathcal{C}_n$. Then, $\hat{\boldsymbol{\beta}}_{\bullet j} \neq \mathbf{0}$ and $||\hat{\boldsymbol{\beta}}_{\bullet j}|| > 0$, but $\boldsymbol{\beta}_{\bullet j}^* = \mathbf{0}$. Since $\hat{\boldsymbol{\beta}}$ is defined as the minimizer of $V_n(\boldsymbol{\beta})$, it must also minimize $V_n(\boldsymbol{\beta})/\sqrt{n}$. Hence, with $\partial J_n(\hat{\boldsymbol{\beta}}_{\bullet j})$ denoting the subdifferential of the penalty with respect to $\hat{\boldsymbol{\beta}}_{\bullet j}$, the following must be satisfied by $\hat{\boldsymbol{\beta}}_{\bullet j}$ and its score function $\boldsymbol{s}_n(\hat{\boldsymbol{\beta}}_{\bullet j})$:

$$\frac{\boldsymbol{s}_n(\hat{\boldsymbol{\beta}}_{\bullet j})}{\sqrt{n}} \in \frac{1}{\sqrt{n}}\partial J_n(\hat{\boldsymbol{\beta}}_{\bullet j}) := \boldsymbol{\delta}_{nj}, \tag{24}$$

with

$$\boldsymbol{\delta}_{nj} = \frac{\lambda_n}{\sqrt{n}} \frac{1}{||\hat{\boldsymbol{\beta}}_{\bullet j}^{\text{ML}}||} \frac{\hat{\boldsymbol{\beta}}_{\bullet j}}{||\hat{\boldsymbol{\beta}}_{\bullet j}||} + \frac{\zeta_n}{\sqrt{n}} \frac{\min\big(c, ||\hat{\boldsymbol{\beta}}_{\bullet j}^{\text{ML}}||\big)}{c||\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}^{\text{ML}}||} \boldsymbol{D}^\top \mathbf{v}_{2j} = \boldsymbol{\delta}_{1nj} + \boldsymbol{\delta}_{2nj}. \tag{25}$$

Here, $\boldsymbol{D}^\top \mathbf{v}_{2j} = \partial||\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}||$ is the subdifferential of the grouped difference penalty and $\mathbf{v}_{2j}$ has the form given in (23). In particular, one has $||\mathbf{v}_{2j}|| \leq 1$ for any $\hat{\boldsymbol{\beta}}_{\bullet j}$. Since $\frac{\min\left(c,\ ||\hat{\boldsymbol{\beta}}_{\bullet j}^{\text{ML}}||\right)}{c||\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}^{\text{ML}}||} = \mathcal{O}(1)$ and $\zeta_n/\sqrt{n} \to 0$ by assumption, one has $\boldsymbol{\delta}_{2nj} \xrightarrow{\mathbb{P}} 0$, so that only the first term $\boldsymbol{\delta}_{1nj}$ in (25) has to be considered.

From the proof of normality, we know that $n^{-1/2}\boldsymbol{s}_n(\hat{\boldsymbol{\beta}}_{\bullet j}) \xrightarrow{d} N\big(\mathbf{0}, \boldsymbol{F}(\hat{\boldsymbol{\beta}}_{\bullet j})\big)$. Since $\lambda_n \to \infty$ and

$\zeta_n \to \infty$ by assumption, one therefore obtains for any $\epsilon > 0$ that for $t = 1, \dots, q$

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{|s_n(\hat{\beta}_{tj})|}{\sqrt{n}} \leq \lambda_n^{1/4} - \epsilon \right) = 1 \qquad \text{and} \qquad \lim_{n\to\infty} \mathbb{P}\left(\frac{|s_n(\hat{\beta}_{tj})|}{\sqrt{n}} \leq \zeta_n^{1/4} - \epsilon \right) = 1. \tag{26}$$

Since $\hat{\boldsymbol{\beta}}_{\bullet j} \neq \mathbf{0}$, there must exist at least one $t \in \{1, \dots, q\}$ such that $\hat{\beta}_{tj} \neq 0$ and $|\hat{\beta}_{tj}| = \max_{l \in \{1,\dots,q\}} |\hat{\boldsymbol{\beta}}_{lj}|$. Hence, for this t, one has $0 < \frac{1}{q} \leq \left| \frac{|\hat{\beta}_{tj}|}{||\hat{\boldsymbol{\beta}}_{\bullet j}||} \right| \leq 1$. From the considerations about the limit behavior, we also know $\lim_{n\to\infty} \mathbb{P}\left(\sqrt{n} ||\hat{\boldsymbol{\beta}}_{\bullet j}^{\text{ML}}|| \leq \lambda^{1/2} \right) = 1$. Thus, $\lim_{n\to\infty} \mathbb{P}(|\delta_{ntj}| \geq \lambda^{1/4}) = 1$ and with (26), one obtains

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{s_n(\hat{\beta}_{tj})}{\sqrt{n}} = \delta_{ntj} \right) = 0,$$

which contradicts $\hat{\boldsymbol{\beta}}_{\bullet j} \neq \mathbf{0}$. Together with the selection of nonzero coefficients, this implies $\lim_{n\to\infty} \mathbb{P}(\mathcal{C}_n = \mathcal{C}) = 1$.

– *Shrinkage of category-specific to global effects:* We must show $\lim_{n\to\infty} \mathbb{P}(j \in \mathcal{B}_n) = 1$ for all $j \in \mathcal{B}$. Assume there exists a $j \in \mathcal{B}$ with $j \notin \mathcal{B}_n$. Then, $\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j} \neq \mathbf{0}$, $||\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}|| > 0$ and $||\hat{\boldsymbol{\beta}}_{\bullet j}|| > 0$, but $\boldsymbol{D}\boldsymbol{\beta}_{\bullet j}^{*} = \mathbf{0}$ and $\boldsymbol{\beta}_{\bullet j}^{*} \neq \mathbf{0}$. Recycling the arguments from the previous case, (24) must be satisfied. This time, $||\hat{\boldsymbol{\beta}}_{\bullet j}^{\text{ML}}|| \to ||\boldsymbol{\beta}_{\bullet j}^{*}|| > 0$, so that $\boldsymbol{\delta}_{1nj} \xrightarrow{\mathbb{P}} 0$ due to $\lambda_n / \sqrt{n} \to 0$ and only the second term $\boldsymbol{\delta}_{2nj}$ from (25) has to be considered, which this time is guaranteed to be differentiable:

$$\boldsymbol{\delta}_{2nj} = \frac{\zeta_n}{\sqrt{n}} \frac{\min\left(c, ||\hat{\boldsymbol{\beta}}_{\bullet j}^{\text{ML}}||\right)}{c||\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}^{\text{ML}}||} \frac{\boldsymbol{D}^{\mathsf{T}}\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}}{||\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}||}.$$

Note that the term $\min\left(c, ||\hat{\boldsymbol{\beta}}_{\bullet j}^{\text{ML}}||\right)/c \to const$ and can therefore be ignored. Since $\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j} \neq \mathbf{0}$, it follows that $\hat{\boldsymbol{\beta}}_{\bullet j} \notin \text{Null}(\boldsymbol{D}^{\mathsf{T}}\boldsymbol{D})$. Hence, there must exist at least one $t \in \{1, \dots, q\}$ such that $\left[\boldsymbol{D}^{\mathsf{T}}\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}\right]_t \neq 0$ and $\left|\left[\boldsymbol{D}^{\mathsf{T}}\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}\right]_t\right| = \max_{l \in \{1,\dots,q\}} \left|\left[\boldsymbol{D}^{\mathsf{T}}\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}\right]_l\right|$. Hence, one has $0 < \left| \frac{\left[\boldsymbol{D}^{\mathsf{T}}\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}\right]_t}{||\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}||} \right| < 2$. We also know that $\lim_{n\to\infty} \mathbb{P}\left(\sqrt{n} ||\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}^{\text{ML}}|| \leq \zeta^{1/2} \right) = 1$ and thus $\lim_{n\to\infty} \mathbb{P}(|\delta_{ntj}| \geq \zeta^{1/4}) = 1$. Together with (26), this yields

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{s_n(\hat{\beta}_{tj})}{\sqrt{n}} = \delta_{ntj} \right) = 0,$$

which contradicts $\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j} \neq \mathbf{0}$. Together with nonzero global coefficients being unpenalized, this implies $\lim_{n\to\infty} \mathbb{P}(\mathcal{B}_n = \mathcal{B}) = 1$ and therefore completes the proof of Theorem 3. □

Remark: The above proof of selection consistency relies on the fact that the scaled penalty slope $\boldsymbol{\delta}_{nj}$ diverges at a rate that cannot be reached by $\boldsymbol{s}_n(\hat{\boldsymbol{\beta}}_{\bullet j})/\sqrt{n}$ whenever incorrect structures are present in $\hat{\boldsymbol{\beta}}$. Since both $||\hat{\boldsymbol{\beta}}_{\bullet j}^{\text{ML}}||$ and $||\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}^{\text{ML}}||$ go to zero at the same rate for $\boldsymbol{\beta}_{\bullet j}^{*} = \mathbf{0}$, usage of the "naive" adaptive weight $w_{2j} = 1/||\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\bullet j}^{\text{ML}}||$ could not rule out the possiblity that, for various values

of $\hat{\boldsymbol{\beta}}_{\bullet j}$, $\boldsymbol{\delta}_{1nj} \to \infty$ and $\boldsymbol{\delta}_{2nj} \to -\infty$ (or vice versa) at a similar rate, so that $|\delta_{ntj}| \to const$ for all $t = 1, \ldots, q$ in (25).

A.4. Proof of Theorem 4

It is immediately seen that large parts of the proof of Theorem 3 do not depend on the specific choice of the penalty function ρ. There are three critical parts:

Since ρ in (19) is assumed to be continuously differentiable for nonzero arguments, its gradient must be finite for every finite, nonzero input. Hence, when $\boldsymbol{\beta}^*_{\bullet j}$ or $\boldsymbol{D\beta}^*_{\bullet j}$ are nonzero, the gradient vectors $\nabla\rho(\boldsymbol{\beta}^*_{\bullet j})$ or $\nabla\rho(\boldsymbol{D\beta}^*_{\bullet j})$ are constant and finite. Thus, the derivation of the limit behavior for the general penalty yields the same results as in the proof of Theorem 3.

The minimum slope condition (20) ensures that the arguments which establish selection consistency also hold using the general penalty (19).

The most critical part is the proof of normality, which requires the objective function $V_n(\boldsymbol{\beta})$ to be convex around its global minimum. Since the definition of the general penalty in (19) allows ρ to be concave, this convexity cannot be guaranteed without severly restricting the possible choices of ρ and must therefore be assumed for Theorem 4. □

A.5. Proof of Lemma 1

Let $\tilde{\boldsymbol{\beta}} = \mathbf{Prox}_{\text{ETL}}(\boldsymbol{u}|\, 0, \zeta)$, let $S := \left(1 - \frac{\lambda}{||\tilde{\boldsymbol{\beta}}||_2}\right)_+ \in [0,1]$ denote the shrinkage factor applied by the proximal map of the L_2-norm penalty and let $\hat{\boldsymbol{\beta}} = S\tilde{\boldsymbol{\beta}}$. Lemma 1 states that $\mathbf{Prox}_{\text{ETL}}(\boldsymbol{u}|\, \lambda, \zeta)$ is given by this $\hat{\boldsymbol{\beta}}$.

If $\boldsymbol{D\tilde{\beta}} = \mathbf{0}$, then $\tilde{\boldsymbol{\beta}} = \alpha \cdot \mathbf{1}_q$ for some $\alpha \in \mathbb{R}$. Hence, $S\tilde{\boldsymbol{\beta}} = S\alpha \cdot \mathbf{1}_q$, so that $\boldsymbol{D\hat{\beta}} = \boldsymbol{D\tilde{\beta}} = \mathbf{0}$. If $\boldsymbol{D\tilde{\beta}} \neq \mathbf{0}$, one obtains $\boldsymbol{D\hat{\beta}} = \boldsymbol{D}(S \cdot \tilde{\boldsymbol{\beta}}) = S \cdot \boldsymbol{D\tilde{\beta}}$ and therefore

$$\frac{\boldsymbol{D\hat{\beta}}}{||\boldsymbol{D\hat{\beta}}||_2} = \frac{S \cdot \boldsymbol{D\tilde{\beta}}}{S \cdot ||\boldsymbol{D\tilde{\beta}}||_2} = \frac{\boldsymbol{D\tilde{\beta}}}{||\boldsymbol{D\tilde{\beta}}||_2}.$$

Hence, one obtains, with the subdifferential of $||\boldsymbol{D\beta}||_2$ which was given in (23), that $\partial||\boldsymbol{D\hat{\beta}}||_2 \supseteq \partial||\boldsymbol{D\tilde{\beta}}||_2$. As shown in Yu (2013), this is a sufficient condition that proves Lemma 1. □

References

Agresti, A. (2013). *Categorical Data Analysis, 3d Edition*, Wiley, New York.

Archer, K. & Williams, A. (2012). L1 penalized continuation ratio models for ordinal response prediction using high-dimensional datasets, *Statistics in Medicine* **31**, 1464–1474.

Archer, K. J., Hou, J., Zhou, Q., Ferber, K., Layne, J. G. & Gentry, A. E. (2014). ordinalgmifs: An R Package for Ordinal Regression in High-dimensional Data Settings, *Cancer Informatics* **13**, 187–195.

Barbero, Á. & Sra, S. (2011). Fast Newton type Methods for Total Variation with Applications, *Proceedings of the 28th International Conference on Machine Learning*, 313–320.

Beck, A. & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM Journal on Imaging Sciences* **2**, 183–202.

Bondell, H. & Reich, B. (2009). Simultaneous factor selection and collapsing levels in ANOVA, *Biometrics* **65**, 169–177.

Boyd, S. & Vandenberghe, L. (2004). *Convex Optimization*, Cambridge University Press, Cambridge.

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression, *Biometrics* **46**, 1171–1178.

Chen, J. & Li, H. (2013). Variable selection for sparse Dirichlet-Multinomial regression with an application to microbiome data analysis, *The Annals of Applied Statistics* **7**, 418–442.

Cox, C. (1995). Location-scale cumulative odds models for ordinal data: A generalized non-linear model approach, *Statistics in Medicine* **14**, 1191–1203.

Fahrmeir, L. & Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models, *The Annals of Statistics* **13**, 342–368.

Fahrmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models (Second Edition)*, Springer-Verlag, New York.

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

Gertheiss, J., Hogger, S., Oberhauser, C. & Tutz, G. (2011). Selection of ordinally scaled independent variables with applications to international classification of functioning core sets, *Journal of the Royal Statistical Society Series C* **60**, 377–395.

Gertheiss, J., Maity, A. & Staicu, A.-M. (2013). Variable selection in generalized functional linear models, *Stat* **2**, 86–101.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, New York.

Heinzl, F. & Tutz, G. (2014). Clustering in linear mixed models with a group fused lasso penalty, *Biometrical Journal* **1**, 44–68.

Huang, J., Breheny, P. & Ma, S. (2012). A Selective Review of Group Selection in High-Dimensional Models, *Statistical Science* **27**, 481–499.

James, M. (1978). The Generalised Inverse, *The Mathematical Gazette* **62**, 109–114.

Masters, G. N. (1982). A Rasch Model for Partial Credit Scoring, *Psychometrika* **47**, 149–174.

McCullagh, P. (1980). Regression Model for Ordinal Data (with Discussion), **B 42**, 109–127.

McKinley, T. J., Morters, M. & Wood, J. L. N. (2015). Bayesian Model Choice in Cumulative Link Ordinal Regression Models, *Bayesian Analysis* **10**, 1–30.

McLachlan, G. & Peel, D. (2000). *Finite Mixture Models*, Wiley & Sons, New York.

Meier, L., van de Geer, S. & Bühlmann, P. (2008). The group lasso for logistic regression, *Journal of the Royal Statistical Society B* **70**, 53–71.

Meier, L., van de Geer, S. & Bühlmann, P. (2009). High-dimensional additive Modeling, *The Annals of Statistics* **37**, 3779–3821.

Oelker, M.-R., Gertheiss, J. & Tutz, G. (2014). Regularization and model selection with categorical predictors and effect modifiers in generalized linear models, *Statistical Modelling* **14**, 157–177.

Parikh, N. & Boyd, S. (2013). Proximal Algorithms, *Foundations and Trends in Optimization* **1**, 123–231.

Peterson, B. & Harrell, F. E. (1990). Partial Proportional Odds Models for Ordinal Response Variables, *Applied Statistics* **39**, 205–217.

Peyhardi, J., Trottier, C. & Guédon, Y. (2015). A new specification of generalized linear models for categorical data, *Biometrika* **102**, 889–906.

Pößnecker, W. (2014). *MRSP: Multinomial Response Models with Structured Penalties*, R package version 0.4.3.

R Development Core Team (2014). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Simon, N., Friedman, J. & Hastie, T. (2013). A blockwise descent algorithm for group-penalized multiresponse and multinomial regression, *arXiv preprint*.

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

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. (2005). Sparsity and smoothness via the fused lasso, *Journal of the Royal Statistical Society B* **67**, 91–108.

Tutz, G. (2012). *Regression for Categorical Data*, Cambridge University Press, Cambridge.

Tutz, G., Pößnecker, W. & Uhlmann, L. (2015). Variable Selection in General Multinomial Logit Models, *Computational Statistics & Data Analysis* **82**, 207–222.

Tutz, G. & Schmid, M. (2016). *Discrete Time to Event Models*, Springer Series in Statistics.

Vincent, M. & Hansen, N. (2014). Sparse group lasso and high dimensional multinomial classification, *Computational Statistics* & *Data Analysis* **71**, 771–786.

Wang, H. & Leng, C. (2008). A note on adaptive group lasso, *Computational Statistics* & *Data Analysis* **52**, 5277–5286.

Wytock, M., Sra, S. & Kolter, J. (2014). Fast Newton methods for the group fused lasso, *Proceedings of the 30th Uncertainty in Artificial Intelligence (UAI) conference*.

Yu, Y.-L. (2013). On Decomposing the Proximal Map, *Advances in Neural Information Processing Systems* **26**, 91–99.

Yuan, M. & Lin, Y. (2006). Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society B* **68**, 49–67.

Zou, H. (2006). The adaptive lasso and its oracle properties, *Journal of the American Statistical Association* **101**, 1418–1429.