Statistical Modelling 2016; **16(3)**: 161–200

Regularized regression for categorical data

Gerhard Tutz[1] and Jan Gertheiss[2]
[1]Department of Statistics, Ludwig-Maximilians-Universität Munich, Germany
[2]Institute of Applied Stochastics and Operations Research, Clausthal University of Technology, Germany

Abstract: In the last two decades, regularization techniques, in particular penalty-based methods, have become very popular in statistical modelling. Driven by technological developments, most approaches have been designed for high-dimensional problems with metric variables, whereas categorical data has largely been neglected. In recent years, however, it has become clear that regularization is also very promising when modelling categorical data. A specific trait of categorical data is that many parameters are typically needed to model the underlying structure. This results in complex estimation problems that call for structured penalties which are tailored to the categorical nature of the data. This article gives a systematic overview of penalty-based methods for categorical data developed so far and highlights some issues where further research is needed. We deal with categorical predictors as well as models for categorical response variables. The primary interest of this article is to give insight into basic properties of and differences between methods that are important with respect to statistical modelling in practice, without going into technical details or extensive discussion of asymptotic properties.

Key words: boosting; categorical data; fused lasso; group lasso; multinomial model; proportional odds model; regression trees

1 Introduction

In recent decades, regularization methods for regression and classification have become a topic of intensive research. Regularization methods typically aim at a sparse representation of the link between predictors and responses, particularly in high-dimensional settings; see, for example, Hastie et al. (2009) and Bühlmann and van de Geer (2011). Only those components that are really needed to model the effect of explanatory variables on an outcome variable should be included in the model. Categorical variables are often a challenge to sparsity, even in seemingly low-dimensional models, because typically at least one parameter is needed for

Address for correspondence: Gerhard Tutz, Department of Statistics, Ludwig-Maximilians-Universität Munich, Germany.
E-mail: gerhard.tutz@stat.uni-muenchen.de

10.1177/1471082X16642560

each category. As a consequence, if the number of categories in a predictor is large, common maximum likelihood estimates tend to fail because they are not unique or deteriorate. If the outcome variable is multi categorical, then similar problems arise because for each level of the outcome variable, a different set of regression parameters is needed to link the response to the covariates.

Another feature of categorical (or more generally speaking 'discrete') data is that other structures beyond those for metric data are of interest. While regularization for metric predictors often means shrinkage and/or variable selection and therefore identification of parameters that should be set to zero, for a categorical predictor we also want to know which categories have to be distinguished when modelling the effect on an outcome variable. In other words, one wants to identify clusters of categories that share the same effect. Furthermore, clustering is not restricted to predictor variables, it is also challenging when modelling categorical outcomes or subject-specific effects.

In this article, we will systematically review regularization methods for categorical data. A very popular approach we will focus on is regularization or constrained estimation by use of penalty terms in the tradition of the lasso, which was introduced by Tibshirani (1996). The primary interest of this article is to give insight into basic properties of and differences between methods that are important with respect to statistical modelling in practice. For instance, which penalty is the right one for which kind of model and which research question? Which penalty serves the purpose of clustering categories? Should the fusion of categories be done individually or groupwise? Which methods can be used for smoothing levels of an ordinal covariate? etc. In order to provide such a broad overview, we will not go into technical details or extensively discuss asymptotic properties. The article is organized as follows: In Section 2, we will present a motivating data example on households' food expenses and the corresponding modelling framework. In Section 3, we will consider penalty methods for categorical covariates. Section 4 deals with models for categorical outcomes, Section 5 with subject-specific models where individuals are to be clustered and Section 6 with pairwise comparisons. In Section 7, we briefly describe some alternatives to penalty methods and Section 8 concludes the article.

2 Data example and modelling framework

2.1 Data example: Spending for food

The data we consider comes from a study about marketing for food products, in particular luxury food (Hartmann, 2015; Hartmann et al., 2016a,b). The primary aim of this study was the segmentation of German consumers based on the perceived dimensions of luxury food. Here, however, we will primarily focus on aspects regarding more general behaviour when buying and consuming food products. Our response of interest is the (approximate) amount of money a household spends each week on groceries in stores (i.e., not in restaurants, for delivery, etc.). Table 1 gives a description of the ordinal covariates. This data set is a typical example of a study using a large number of Likert-type items to investigate habits, attitudes, etc. (variables $X_3, \ldots, X_{20}$). In many studies, the number of items will be even larger than in our illustrative example here. However, not all of the items will be relevant for the response. Therefore, variable selection is important. Furthermore, the relationship between the items, coded by $-2, -1, \cdots, 2$, or $1, 2, \cdots$, and the response will not necessarily be linear (even though many applied papers may assume that). With dummy coding, however, the number of parameters to be fit becomes very large and estimated coefficients will be erratic and thus hard to interpret, even with a relatively large sample size of around 800 as in our case. Therefore, penalized estimation is very attractive in studies of this type.

Table 1 Description of the ordinal covariates

Variable		Coding	
X_1	number of persons in the household	1, 2, 3, 4, 5, 6 or more	

X_2	monthly household net income (in euro)	1: less than 900; 2: 900–1 300;
		3: 1 300–1 500; 4: 1 500–2 000;
		5: 2 000–2 600; 6: 2 600–3 600;
		7: 2 600–5 000; 8: 5 000–7 600;
		9: 7 600–9 000; 10: 9 000–12 600;
		11: 12 600–18 000; 12: 18 000 or more	

	Eating habits:		
X_3	'I like to cook'.	−2:	not true at all
X_4	'I frequently go to *cheap* restaurants'.	−1:	not true
X_5	'I frequently go to *expensive* restaurants'.	0:	partly true
X_6	'I frequently buy convenience foods'.	1:	true
X_7	'I frequently use a delivery service'.	2:	absolutely true
X_8	'I prefer eating at home'.		
	Where are you buying your groceries?		
X_9	at a discount supermarket	1:	very often
X_{10}	at the farmer's market	2:	often
X_{11}	at a wholefood shop	3:	sometimes
X_{12}	directly at the farm	4:	rarely
X_{13}	at the delicatessen store	5:	never
X_{14}	at a specialist shop		
	Time, price, etc.:		
X_{15}	'I like to take my time when buying groceries'.	−2:	I don't agree at all
X_{16}	'I want to be fast when buying groceries'.	−1:	I don't agree
X_{17}	'With food, a high price stands for high quality'.	0:	I partly agree
X_{18}	'I'd rather buy a certain food product	1:	I agree
	if the price is rather high'.	2:	I totally agree
X_{19}	'When buying food, I don't		
	care about the price'.		
X_{20}	'Generally I like luxury food'.		

In addition to the ordinal covariates in Table 1, we consider the two nominal predictors 'family status' (C_1) and the professional category of the main earner (C_2), see Table 2. Here, it is particularly interesting which categories differ from each other with respect to spending behaviour when controlling for the number of persons in the household (X_1) and household income (X_2).

Table 2 Description of the nominal covariates

Variable		Coding	
C_1	family status	1	single
		2	single with child(ren)
		3	in a serious relationship
		4	in a serious relationship with child(ren)
		5	married
		6	married with child(ren)
		7	widowed
C_2	professional category of main earner	1	white-collar worker
		2	executive employee
		3	self-employed without employees
		4	self-employed with employees
		5	freelancer (with university degree)
		6	official
		7	senior official
		8	blue-collar worker
		9	housewife/husband
		10	unemployed
		11	pensioner
		12	student
		13	other

2.2 Structuring categorical predictors

When categorical predictors are included in a regression model, the number of parameters that are needed to specify the impact on the response is typically large. With several categorical predictors and large numbers of categories, the commonly used estimates, such as maximum likelihood or least squares, tend to become unstable. Cases like that particularly call for a sparse representation of the effects on the response by including only the relevant terms or imposing some constraints that facilitate interpretation.

The framework we use here is generalized linear models (GLMs) for which the conditional response $\mu = E(y|\boldsymbol{x})$ is specified by

$$\mu = h(\eta) \quad \text{or} \quad g(\mu) = \eta$$

where $h(\cdot)$ denotes the (known) response function and $g(\cdot) = h(\cdot)^{-1}$ the link function. The linear predictor is determined by the predictors collected in the vector $\boldsymbol{x}$ in the form $\eta = \boldsymbol{x}^\top \boldsymbol{\beta}$. In addition, the conditional distribution $y|\boldsymbol{x}$ is from a simple exponential family (McCullagh and Nelder, 1989).

When predictors are categorical, $\boldsymbol{x}$ usually consists of dummy variables. So let categorical predictors $C_j, j = 1, \ldots, p$ have values $C_j \in \{0, \ldots, k_j\}$. These predictors can be included into a GLM by using dummy variables defined by $x_{jr} = 1$ if $C_j = r$

and $x_{jr} = 0$ otherwise, yielding the linear predictor

$$\eta = \alpha + \sum_{j=1}^{p} \sum_{r=0}^{k_j} x_{jr}\beta_{jr} = \alpha + \sum_{j=1}^{p} \boldsymbol{x}_j^\top \boldsymbol{\beta}_j, \tag{2.1}$$

where $\boldsymbol{\beta}_j$ collects all parameters linked to predictor C_j. For means of identifiability, some constraints need to be placed on $\boldsymbol{\beta}_j$. Typically some reference category is chosen with the corresponding parameter being fixed to zero, most often the first or last category. Here, we will use the first one yielding $\beta_{j0} = 0$ for all j. So $\boldsymbol{x}_j$ and $\boldsymbol{\beta}_j$ from Equation (2.1) can be reduced to $\boldsymbol{x}_j^\top = (x_{j1}, \ldots, x_{jk_j})$ and $\boldsymbol{\beta}_j^\top = (\beta_{j1}, \ldots, \beta_{jk_j})$, respectively. Consequently, predictor C_j adds k_j parameters to the model, and the total number of parameters contributed by the categorical predictors is $k_1 + \cdots + k_p$, which can be very large, in particular if several categorical predictors are included.

In matrix notation, the linear predictor has the form

$$\boldsymbol{\eta} = \boldsymbol{\alpha} + \sum_{j=1}^{p} \boldsymbol{X}_j \boldsymbol{\beta}_j = \boldsymbol{X}\boldsymbol{\beta}, \tag{2.2}$$

where $\boldsymbol{X}_j$ are $(n \times k_j)$ design matrices containing the explanatory variables observed at n individuals or sampling units; $\boldsymbol{\alpha}$ is a vector containing n times the constant α. Alternatively, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_j$ can be collected in a single parameter vector $\boldsymbol{\beta}$ with corresponding design matrix $\boldsymbol{X}$.

Of course, some additional metric covariates $z_1, \ldots, z_q$ may also be included in η. In this case, we have $\eta = \alpha + \sum_{j=1}^{p} \boldsymbol{x}_j^\top \boldsymbol{\beta}_j + \boldsymbol{z}^\top \boldsymbol{\gamma}$, where vector $\boldsymbol{z}$ collects the additional explanatory variables and $\boldsymbol{\gamma}$ is the associated vector of regression coefficients.

The most popular (G)LM, which we will also be using for analyzing our data from Section 2.1, is the linear model

$$y = \alpha + \sum_{j=1}^{p} \boldsymbol{x}_j^\top \boldsymbol{\beta}_j + \epsilon, \tag{2.3}$$

where the link function is just the identity, and ϵ is a normal random variable with mean 0 and variance σ^2. Sometimes, however, model (2.3) is specified without the assumption of normality.

When selecting a model with categorical predictors, it should be distinguished between two problems:

(A) Which categorical predictors should be included in the model?
(B) Which categories within one categorical predictor should be distinguished?

In case (A), it has to be decided whether for some variable j 'all' dummy coefficients β_{jr} are to be set to zero. If so, predictor C_j is excluded from the model. In case (B), the question is whether some β_{jr} and β_{js} are to be set equal. If so, categories r and s of predictor C_j are fused, as one cannot distinguish their effects on the response. In Section 3, we will discuss several approaches that can be used for answering (A), (B) or both.

Sometimes, a categorical covariate can also act as a so-called effect modifier. To distinguish the effect modifier from categorical predictors c_j above, it is denoted by u here. In general, 'effect modifier' u means that the effect of the remaining predictors may be different for different levels of u. For metric or binary predictors $z_1, \ldots, z_q$, the linear predictor then takes the form $\eta = \eta(u) = \alpha(u) + z_1\gamma_1(u) + \ldots z_q\gamma_q(u)$, a so-called 'varying coefficient' model. Whereas varying coefficient models have been studied extensively for continuous effect modifiers (see, e.g., Hastie and Tibshirani, 1993; Hoover et al., 1998; Kauermann and Tutz, 2000; Fan et al., 2003; Lu et al., 2008; Wang et al., 2008; Wang and Xia, 2009; Liu et al., 2014; Klopp and Pensky, 2015), literature on discrete u is relatively rare (Gertheiss and Tutz, 2012; Oelker et al., 2014; Zhao et al., 2014; Ollier and Viallon, 2015, to name a few). With categorical effect modifiers, issues of model selection are closely related to (B) from above. Primarily, for which predictors z_j is the effect actually varying? And for which levels of u is the predictors' effect on the response varying? In addition, we may ask which predictors should be included in the model at all.

An alternative interpretation of the varying coefficient model is in terms of interactions between two (or more) explanatory variables. A very interesting case concerns the interaction between categorical predictors, like in a two-way Analysis of Variance (ANOVA) model. If interactions are present, (B) becomes more difficult because only levels whose interaction effects are all identical should be collapsed. A penalty for solving this problem has been proposed by Post and Bondell (2013), which will also be discussed in Section 3.

3 Regularization for categorical covariates

For answering the questions of model selection stated in Section 2.2, penalties can be very useful tools because they can be tailored to recover, in a data-driven way, exactly those kinds of structures that the researcher is interested in. In what follows, we will give an overview of various penalties that can be used with categorical predictors and discuss what those penalties are designed for from a model-building and fitting perspective.

3.1 Penalty-based methods

Regularization methods that use penalty terms are obtained by maximizing the penalized log-likelihood

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - J_\lambda(\boldsymbol{\beta}),$$

where $l(\boldsymbol{\beta})$ is the usual log-likelihood of the GLM and $J_\lambda(\boldsymbol{\beta})$ is a function that penalizes the size and structure of the parameters β_1, β_2, etc., collected in vector $\boldsymbol{\beta}$. The strength of the penalty is typically determined by a tuning parameter λ, and the penalty has the form $J_\lambda(\boldsymbol{\beta}) = \lambda J(\boldsymbol{\beta})$. A classical penalty is the ridge penalty $J_\lambda(\boldsymbol{\beta}) = \lambda \sum_j \beta_j^2$, which goes back to Hoerl and Kennard (1970). It shrinks estimates towards zero and is able to stabilize estimates but is unable to detect structures within or between the predictors. For the detection of interesting structures in discrete data, other penalties are much more useful and will be discussed in the following sections.

3.1.1 Smoothing ordered categorical predictors

If the categories of the predictors $C_j \in \{0, \ldots, k_j\}$ are ordered, it is often sensible to assume that the corresponding parameters $\beta_{j1}, \ldots, \beta_{jk_j}$ vary smoothly over the categories. A penalty which enforces that estimates of coefficients for adjacent categories are not too far apart is

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{p} \sum_{r=1}^{k_j} (\beta_{jr} - \beta_{j,r-1})^2, \tag{3.1}$$

where $\beta_{j0} = 0$ refers to the reference category (Gertheiss and Tutz, 2009). The penalty is a generalized ridge type penalty which can be given as a quadratic form. With $\boldsymbol{D}_j$ denoting the matrix that generates differences of fixed order, one obtains

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{p} \boldsymbol{\beta}_j^\top \boldsymbol{D}_j^\top \boldsymbol{D}_j \boldsymbol{\beta}_j = \sum_{j=1}^{p} \boldsymbol{\beta}_j^\top \boldsymbol{\Omega}_j \boldsymbol{\beta}_j,$$

where $\boldsymbol{\Omega}_j = \boldsymbol{D}_j^\top \boldsymbol{D}_j$. It is straightforward to use differences of higher order by building differences of differences. It has been shown that ridge type penalties for differences strongly reduce the mean squared error of estimates if the effect of the categorical predictor on the dependent variable is smooth (Gertheiss and Tutz, 2009). A typical application for this kind of penalty is rating scales (Tutz and Gertheiss, 2014). Also in the example from Section 2.1, many covariates belong to this category of predictors (see Table 1). With larger numbers of predictors, however, estimating of covariate effects should be combined with variable selection, which will be discussed in the next section.

3.1.2 Groupwise selection

Selection of categorical predictors can be obtained by the 'group lasso' (Yuan and Lin, 2006), which is an extension of Tibshirani's lasso (Tibshirani, 1996) that is able to select groups of parameters simultaneously. It uses the penalty term

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{p} \sqrt{k_j}||\boldsymbol{\beta}_j||_2, \tag{3.2}$$

where $||\boldsymbol{\beta}_j||_2 = (\beta_{j1}^2 + \cdots + \beta_{jk_j}^2)^{1/2}$ is the L_2-norm of the parameters of the jth group, which refers to one categorical predictor. The penalty encourages sparsity such that either $\hat{\boldsymbol{\beta}}_j = \mathbf{0}$ or $\beta_{jr} \neq 0$ for all $r = 1, \ldots, k_j$. By encouraging that whole vectors $\boldsymbol{\beta}_j$ are set to zero/non-zero, it aims at the selection of entire (categorical) variables in contrast to simple parameter selection. Instead of L_2-norm $||\boldsymbol{\beta}_j||_2$, any other L_q-norm with $q > 1$ could be used for groupwise selection (Zhao et al., 2009), but we will focus on the group lasso ($q = 2$) here.

When using the group lasso penalty, it is typically assumed that the design matrices for the groups $\boldsymbol{X}_j$ are orthonormal. This is rarely fulfilled. However, one can obtain an orthonormal problem by using the more general penalty

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{p} \sqrt{k_j}||\boldsymbol{\beta}_j||_{\boldsymbol{M}_j}, \tag{3.3}$$

where $||\boldsymbol{\beta}_j||_{\boldsymbol{M}_j} = (\boldsymbol{\beta}_j^{\top}\boldsymbol{M}_j\boldsymbol{\beta}_j)^{1/2}$. For the model with linear predictor $\boldsymbol{\eta} = \sum_{j=1}^{p}\boldsymbol{X}_j\boldsymbol{\beta}_j$, one uses the Gram matrix $\boldsymbol{M}_j = \boldsymbol{X}_j^{\top}\boldsymbol{X}_j/n$ (for centred covariates). With $\boldsymbol{M}_j = \boldsymbol{M}_j^{T/2}\boldsymbol{M}_j^{1/2}$ denoting the Cholesky decomposition and $\tilde{\boldsymbol{X}}_j = \boldsymbol{X}_j\boldsymbol{M}_j^{-1/2}$, $\tilde{\boldsymbol{\beta}}_j = \boldsymbol{M}_j^{1/2}\boldsymbol{\beta}_j$ the predictor is transformed to $\boldsymbol{\eta} = \sum_{j=1}^{p}\tilde{\boldsymbol{X}}_j\tilde{\boldsymbol{\beta}}_j$. Then the penalty (3.3) has the form $(\boldsymbol{\beta}_j^{\top}\boldsymbol{M}_j\boldsymbol{\beta}_j)^{1/2} = (\boldsymbol{\beta}_j^{\top}\boldsymbol{M}_j^{T/2}\boldsymbol{M}_j^{1/2}\boldsymbol{\beta}_j)^{1/2} = (\tilde{\boldsymbol{\beta}}_j^{\top}\tilde{\boldsymbol{\beta}}_j)^{1/2}$. Therefore, maximization of the log-likelihood for the model with design matrices $\boldsymbol{X}_j$, parameters $\boldsymbol{\beta}_j$ and penalty components $||\boldsymbol{\beta}_j||_{\boldsymbol{M}_j}$ is equivalent to maximization of the usual penalized log-likelihood for the model with predictors $\tilde{\boldsymbol{X}}_j = \boldsymbol{X}_j\tilde{\boldsymbol{M}}_j^{-1/2}$ and parameters $\tilde{\boldsymbol{\beta}}_j = \boldsymbol{M}_j^{1/2}\boldsymbol{\beta}_j$. Use of the Gram matrix means that the predictors are groupwise standardized. It allows considering the simpler Euclidean norm because the problem can always be transformed into this simpler penalty problem.

Meier et al. (2008) showed that under sparsity, the resulting estimates are consistent even when the number of predictors is larger than the sample size. Selection consistency under various conditions has been investigated by Nardi and Rinaldo (2008) and Bach (2008).

Extensions of the group lasso are the 'sparse' and the 'adaptive' group lasso. The sparse group lasso proposed by Simon et al. (2013) uses the penalty

$$J_{\lambda}(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} \sqrt{k_j}||\boldsymbol{\beta}_j||_2 + \lambda_2||\boldsymbol{\beta}||_1,$$

where $\boldsymbol{\beta}^\top = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_p^\top)$ collects all the parameters, and $||\boldsymbol{\beta}||_1$ denotes the L_1-norm. The penalty aims at enforcing sparsity of groups 'and' within each group. It was developed with a focus on high-dimensional metric predictors, such as gene expressions. With categorical covariates, however, direct application of this penalty often does not make sense because the parameters collected in $\boldsymbol{\beta}_j$ are typically dummy coefficients specifying differences to the reference category. With the L_1-norm penalty, some of those differences are set to zero, whereas others may not. This is only sensible, however, if the reference category is not arbitrarily chosen but special in some sense, for example, a control group. A more appropriate penalty enforcing sparseness within categorical predictors is the fusion penalty discussed in Section 3.1.3.

The adaptive lasso has been proposed by Zou (2006); a group version has been considered by Wang and Leng (2008) and Wei and Huang (2010). It uses the penalty

$$J_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} w_j ||\boldsymbol{\beta}_j||_2,$$

with adaptive weight $w_j = \sqrt{k_j}/\|\breve{\boldsymbol{\beta}}_j\|$, where $\breve{\boldsymbol{\beta}}_j$ is a preliminary estimate of $\boldsymbol{\beta}_j$, for example, an maximum likelihood (ML) estimate or a ridge estimate. The oracle properties that Zou (2006) derived for the adaptive lasso use that for growing sample size the weights for zero coefficients get inflated, whereas the weights for non-zero coefficients converge to a finite constant.

Besides the group lasso, groupwise selection can be obtained by various other penalties as Huang et al. (2012) showed in their overview on available models. A general form of penalty uses for the jth group of variables $\rho(t; v_j\lambda, \gamma)$, where $\rho(t; v_j\lambda, \gamma)$ is a concave function in t, and γ is an additional tuning parameter. The corresponding penalty is

$$J_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^{p} \rho(||\boldsymbol{\beta}_j||_{\mathbf{M}_j}; v_j\lambda, \gamma).$$

A special case is the function $\rho(t; v_j\lambda) = v_j\lambda|t|$, which yields the (group) lasso, where $v_j = \sqrt{k_j}$. Alternatives are grouped versions of the smoothly clipped absolute deviation (SCAD; Fan and Li, 2001; Wang et al., 2007; Wang et al., 2008) or the minimax concave penalty (MCP; Zhang, 2010; Huang et al., 2012), given by $\rho(t; \lambda, \gamma) = \lambda \int_0^{|t|} \min\{1, (\gamma - s/\lambda)_+/(\gamma - 1)\}ds$, $\gamma > 2$ and $\rho(t; \lambda, \gamma) = \lambda \int_0^{|t|} (1 - s/(\gamma\lambda))_+ ds$, $\gamma > 1$, respectively.

The penalties considered so far can be used for categorical predictors in general. However, they are recommended for 'nominal predictors' only. Given 'ordinal categorical predictors', they ignore the information contained in the ordering of categories. When dealing with ordinal covariates, and when the focus is on selection, the group lasso penalty above should be replaced by

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{p} \sqrt{k_j}||\boldsymbol{D}_j\boldsymbol{\beta}_j||_2,$$

where $\boldsymbol{D}_j$ is a matrix that generates differences of fixed order from the parameters linked to the *j*th predictor as discussed in Section 2.2. In the simplest case of first-order differences, we obtain $||\boldsymbol{D}_j\boldsymbol{\beta}_j||_2 = \sqrt{\sum_r(\beta_{jr} - \beta_{j,r-1})^2}$ as proposed by Gertheiss et al. (2011). This penalty enforces selection of the whole group of parameters that belong to the same categorical predictor and simultaneously smoothes over the ordered categories. It has already been used in several applications (see, e.g., Cieza et al., 2014; Leissner et al., 2014; Oberhauser et al., 2013).

For illustration, we consider the example from Section 2.1 with covariates from Table 1. Figure 1 shows the estimated coefficients for a subset of nine predictors for different values of $\lambda \in \{10^{-1}, 10^{-0.5}, \ldots, 10^5\}$. With increasing λ, coefficients are smoothed and shrunk, and at some point, the entire group of dummy coefficients belonging to one variable is set to zero, which means that the corresponding variable is excluded from the model. It is seen, for example, that for X_5 (expensive restaurants), estimates with small λ are wiggly and thus hardly interpretable, but with larger λ, the coefficients indicate that people who often go to expensive restaurants also tend to spend more on food in stores. However, the most influential variables are, not surprisingly, the number of persons in the household (note, the response is the household's spending on food) and the household's net income. For one to five persons, the increase in spending is virtually linear, but for the last category, increase seems to be stronger. This makes perfect sense since in the last category, households with six 'or more' people are collected. With increasing income, spending for food typically increases too, but not linearly across levels. Furthermore, for the top categories, the effect seems to be slightly reversed.

To obtain one final model, an adequate λ needs to be chosen, which can be done via cross-validation. The corresponding coefficients are drawn in solid black in Figure 1. In this case, 10 variables are removed from the complete set of 20 covariates, and 10 are selected (X_1, X_2, X_3, X_5, X_9, X_{10}, X_{11}, X_{13}, X_{15}, X_{16}). Moreover, coefficients are much easier to interpret than with low or no penalty. Here, this is mainly because most of the solid black coefficients appear to be monotone within factors. A penalty which explicitly favours monotonicity within groups of coefficients is the so-called 'cooperative' lasso (Chiquet et al., 2013). This penalty is designed to select sign-coherent groups of coefficients and thus, if applied to differences of adjacent dummy coefficients, it favours monotonicity. In general, selection penalties applied to differences of dummy coefficients can be used to fuse categories, which will be discussed in more detail below.

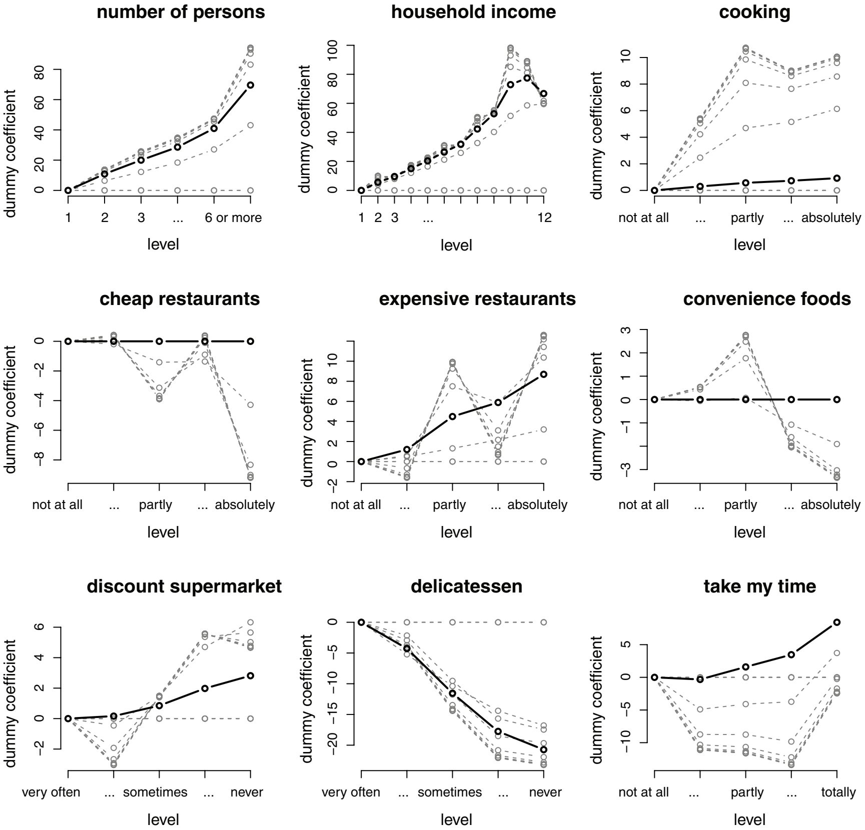

Figure 1 Estimated coefficients for some (ordinal) covariates for different values of λ when applying the smoothing and selection penalty; results for final λ as chosen via 5-fold cross-validation are given in solid black.

3.1.3 Fusion penalties: the clustering of categories

The group lasso penalty selects predictors, but typically, if a predictor is in the model, all the parameter estimates differ and no clustering is obtained. A penalty that enforces the building of clusters of categories that share the same effect is (cf. Bondell and Reich, 2009; Gertheiss and Tutz, 2010)

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{p} \sum_{r<s} w_{rs}^{(j)} |\beta_{jr} - \beta_{js}|, \tag{3.4}$$

where the sum is over all categories $r, s \geq 0$ and implicitly the reference category zero has been chosen by setting $\beta_{j0} = 0 \; \forall j$. The $w_{rs}^{(j)}$ are additional, appropriately chosen weights (Bondell and Reich, 2009; Chiquet et al., 2015). By using the

L_1-penalized differences between all pairs of parameters that are linked to one categorical predictor, the penalty tends to form clusters of categories that have the same effect. Since the parameter for the reference category ($\beta_{j0} = 0$) is included in the sum, the penalty also enforces variable selection. In the extreme case, for $\lambda \to \infty$, all parameter estimates become zero and the categorical predictors are excluded.

A possible choice for the weights in Equation (3.4) is $w_{rs}^{(j)} = (k_j + 1)^{-1} \sqrt{(n_r^{(j)} + n_s^{(j)})/n}$ (Bondell and Reich, 2009), where $n_r^{(j)}$ and $n_s^{(j)}$ are the numbers of observations in category r and s of predictor c_j, respectively. When multiplying $w_{rs}^{(j)}$ by the additional term $|\breve{\beta}_{jr} - \breve{\beta}_{js}|^{-1}$, with initial/unpenalized estimates $\breve{\beta}_{jr}$ (compare Section 3.1.2), an adaptive version in the lines of Zou (2006) is obtained, for which oracle properties like selection and fusion consistency can be derived (Bondell and Reich, 2009; Gertheiss and Tutz, 2010). On the other hand, the adaptive lasso behaves poorly if true parameters (or differences thereof) are close to zero (Pötscher and Schneider, 2009). It has been our experience though that the adaptive version often yields good results in practice if n is large compared to the number of regression parameters. If not, however, the standard approach omitting $|\breve{\beta}_{jr} - \breve{\beta}_{js}|^{-1}$ is typically superior.

When searching for clusters of categories for ordered predictors, it is natural to assume that clusters of categories refer to adjacent categories. Thus, the penalty should enforce the fusion of adjacent categories, which is obtained by using (cf. Gertheiss and Tutz, 2010; Tutz and Gertheiss, 2014)

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{p} \sum_{r=1}^{k_j} w_r^{(j)} |\beta_{jr} - \beta_{j,r-1}|. \tag{3.5}$$

The effect of the penalty is that one obtains step functions for the ordered predictor: categories that have the same effect are fused. For the weights $w_r^{(j)}$, we can choose the same as above, but divided by 2 and omitting factor $(k_j + 1)^{-1}$ (Gertheiss and Tutz, 2010).

For illustration, Figure 2 shows the estimated coefficients for the same variables as shown in Figure 1 for different values of fraction $s/s_{\max} \in \{0.1, 0.15, 0.2, \ldots, 0.9\}$, with s denoting the actual value of the penalty $J(\hat{\boldsymbol{\beta}})$ and $s_{\max}$ denoting the s value for the (unpenalized) ordinary least squares estimates. The solid black line corresponds to the optimal fraction as chosen via 5-fold cross-validation. With this s, as before, 10 predictors are completely removed from the model, while 10 are selected (X_1, X_2, X_5, X_9, X_{11}, X_{13}, X_{14}, X_{15}, X_{16}, X_{17}). Compared to the smoothing and selection penalty from Section 3.1.2, this yields an overlap of eight predictors (X_1, X_2, X_5, X_9, X_{11}, X_{13}, X_{15}, X_{16}). Besides variable selection, however, with the fusion penalty, it is also possible to fuse adjacent categories, which means that the corresponding dummy coefficients are equal. When looking at Figure 2, one sees that with small s coefficients are not only shrunk towards each other, but also more and more categories are fused.

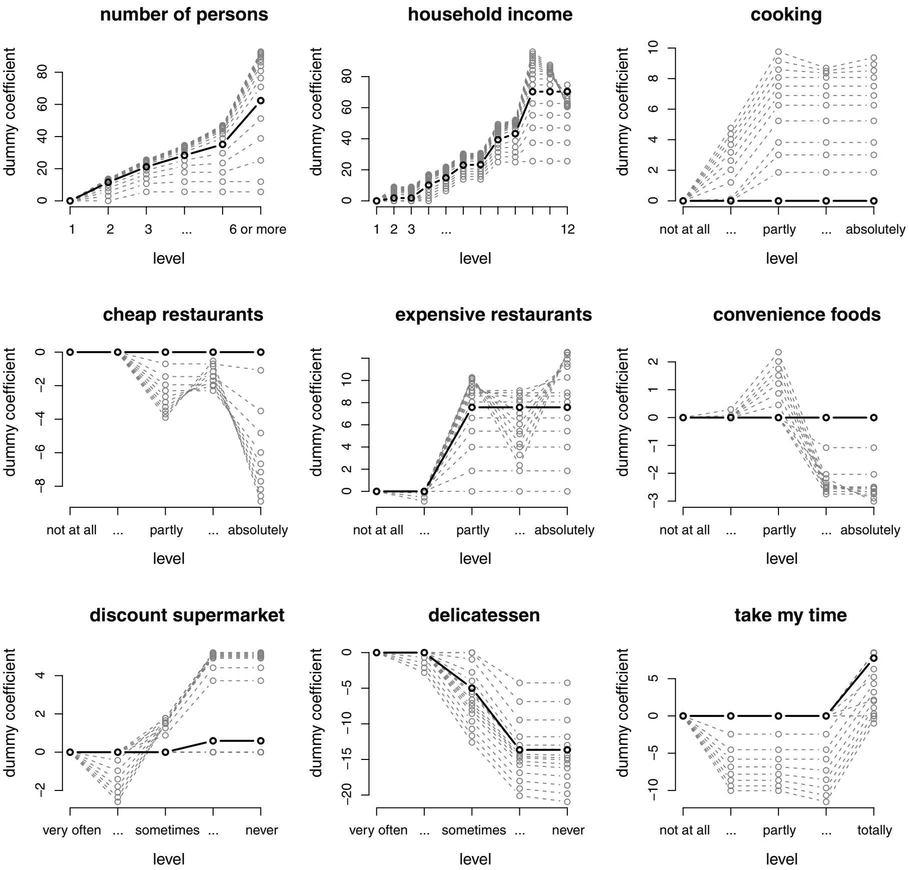

Figure 2 Estimated coefficients for some (ordinal) covariates for different values of fraction $s/s_{\max}$ when applying the fusion penalty; results for final $s/s_{\max}$ as chosen via 5-fold cross-validation are given in solid black.

At some point, all categories are fused, which means that the predictor is implicitly removed from the model, because no distinction is made between the categories. Due to the restriction that the coefficient for the reference category is zero, all coefficients are zero when all levels are fused. For variables that were selected, one clearly sees differences in the results for the smoothing penalty from Figure 1, see in particular the solid black lines in Figure 2. Within variable X_5 (expensive restaurants), for instance, coefficients are not strictly monotone but a step function with only one jump between the negative and the, at least partly, consenting categories. In the case of the household net income, we do not find the inverse u-effect for the top categories but all those levels are fused.

If nominal predictors are also considered, not only differences between adjacent categories are penalized but all pairwise differences of dummy coefficients belonging

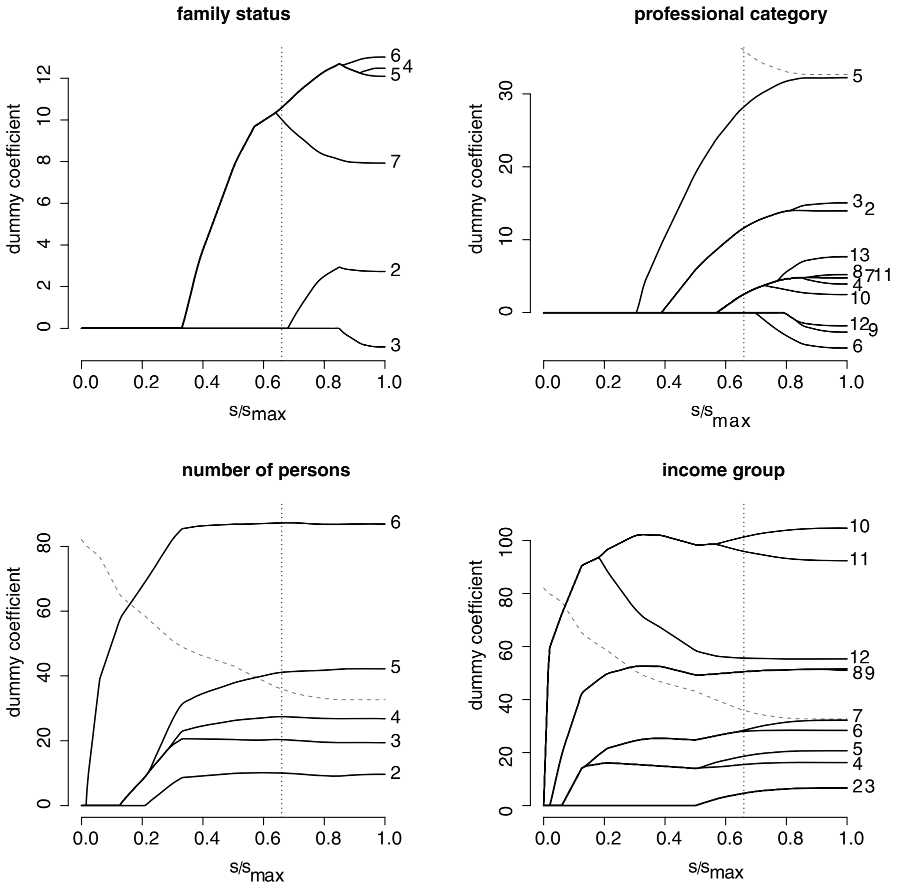

Figure 3 Coefficient path for nominal predictors 'family status' and 'professional category' as well as ordinal covariates 'number of persons' and 'income' (labels, see Tables 1 and 2), dashed gray line refers to the intercept; results for final $s/s_{\max}$ as chosen via 5-fold cross-validation are marked by the dotted vertical line.

to the same covariate, compare Equation (3.4). For illustration, we consider variables C_1 and C_2 from Table 2 while controlling for the ordinal variables number of persons in the household (X_1) and the household's net income (X_2). Since now only four covariates are in the model and hence the number of dummy variables is much smaller than before, we can use the adaptive penalty taking the ordinary least squares estimates into account. Figure 3 shows the corresponding coefficient paths, that is, the estimated dummy coefficients as a function of $s/s_{\max}$. The dashed gray line refers to the intercept; results for final s as chosen via 5-fold cross-validation are marked by the dotted vertical line. We see that, for instance, singles, singles with kid(s) and people in a serious relationship show comparable behaviour, as well as couples with children (married or not) and married couples without kids. With respect to the professional category of the main earner, we see three clusters of different size, plus freelancers (category 5), who tend to spend much more on food than other people.

When looking at Figure 3, we see that categories that are fused at some point remain fused for smaller $s/s_{\max}$ as well, that is, the path contains no splits. This is what Chiquet et al. (2015) call a 'tree structure'. With weights $w_{rs}^{(j)}$ as used so far, however, this is not always the case. Therefore Chiquet et al. (2015) propose alternative weights, such as $w_{rs} = n_r \cdot n_s$ (with $p = 1$), to make sure that the path contains no splits.

With penalty (3.5), the focus is on fusing categories, but as a by-product, variables may also be completely removed from the model. If the focus is rather on variable selection, but some categories of ordinal predictors may also be collapsed, we can apply a sparse group lasso (Simon et al., 2013) to differences of adjacent dummy coefficients. Also, with the cooperative lasso (Chiquet et al., 2013), some (adjacent) levels may be fused. In general, any selection penalty placed on differences of dummy coefficients can be used for level fusion, see also Oelker et al. (2015).

As already mentioned in the Introduction, a model with more than one categorical predictor becomes challenging if interactions of predictors are allowed, like in a multi-factorial ANOVA model. An appropriate penalty for handling such settings has been proposed by Post and Bondell (2013), with the idea being 'that only levels whose interaction effects are all identical should have the possibility of their main effects also declared identical'. This is similar to the concept of heredity, which says that in a regression model with continuous predictors z_1 and z_2, for instance, the interaction term $z_1 z_2$ should only appear in the model if both z_1 and z_2 are present as well. That is why Post and Bondell (2013) call their approach GASH-ANOVA, for 'Grouping And Selection using Heredity in ANOVA'. To ensure heredity, they form groups of parameters where each group contains a main effect difference between two levels of a categorical predictor along with all interaction differences that involve the same two levels, and place an infinity norm penalty on those groups. To simplify notation, let us consider a model with two categorical predictors A and B only, and let the dummy coefficients belonging to A and B be denoted by $\alpha_1, \ldots, \alpha_{k_1}$ and $\beta_1, \ldots, \beta_{k_2}$, respectively. As before, category 0 is taken as the reference category for both factors. Interaction effects are denoted by $(\alpha\beta)_{rs}$, $r = 1, \ldots, k_1$, $s = 1, \ldots, k_2$. Then groups are built in terms of $\phi_{\alpha,rs} = (|\alpha_s - \alpha_r|, |(\alpha\beta)_{s1} - (\alpha\beta)_{r1}|, \ldots, |(\alpha\beta)_{sk_2} - (\alpha\beta)_{rk_2}|)^\top$, $1 \leq r < s \leq k_1$; to penalize differences to the reference category, one additionally has $\phi_{\alpha,0s} = (|\alpha_s|, |(\alpha\beta)_{s1}|, \ldots, |(\alpha\beta)_{sk_2}|)^\top$, $1 \leq s \leq k_1$. Groups $\phi_{\beta,rs}$ are defined analogously. Using those groups, the GASH-ANOVA penalty is

$$J(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{0 \leq r < s \leq k_1} w_{rs}^{(A)} \max\{\phi_{\alpha,rs}\} + \sum_{0 \leq r < s \leq k_2} w_{rs}^{(B)} \max\{\phi_{\beta,rs}\},$$

where $w_{rs}^{(A)}$ and $w_{rs}^{(B)}$ are adequately chosen weights (Post and Bondell, 2013). If an entire group $\phi_{\alpha,rs}$ or $\phi_{\beta,rs}$ is set to zero, corresponding levels of A or B are collapsed. Instead of the max/infinity norm, any other groupwise penalty as discussed in

Section 3.1.2 could be used. If factor levels are ordered, the penalty can be modified to only include differences between adjacent levels. For instance, if factor A is ordinal, one would have $\phi_{\alpha,r-1,r}$ only. A situation like this is often found in practice, for example, in crop sciences when the yield for different genotypes or locations and (ordered) fertilizer levels are analyzed. If more than two factors are considered, groups ϕ can be augmented to include differences of higher order interactions as well. Another advantage of GASH-ANOVA is that it can also be used when only one observation is available for each level combination; see Post and Bondell (2013) for details.

3.1.4 Combined penalties

It should be noted that the given penalty terms can be seen as basic components to obtain sparsity in terms of variables and clusters. In applications, they can apply to main effects but also to interaction terms. Moreover, it is often useful to combine several penalties including simple smoothing penalties as the ridge or extended ridge penalties. For example, if the model contains continuous predictors that are modelled as additive but not necessarily linear effects, one may want to include penalties that ensure that the effects are sufficiently smooth by using P-splines (Eilers and Marx, 1996), see also Section 3.2.1. Another exemplary complex modelling problem that calls for combinations of penalties is discrete survival. As shown in Section 4.2, one typically needs a generalized ridge penalty that smoothes the baseline hazard and a penalty that accounts for the time-varying effects of covariates.

3.1.5 Categorical effect modifier

So far we have considered categorical predictors directly influencing a response variable in a (generalized) linear model. However, the influence can also be indirect by modifying the effect of other covariates on the response. That means the effect γ_j of covariate z_j (potentially) varies across values of categorical 'effect modifier' $u \in \{1, \ldots, k\}$ in terms of

$$\eta = \eta(u) = \gamma_0(u) + z_1\gamma_1(u) + \ldots + z_q\gamma_q(u). \tag{3.6}$$

In general, a model like (3.6) is called a 'varying-coefficient model' (Hastie and Tibshirani, 1993). Here, we will focus on models with metric or binary z_j; interactions of categorical covariates have already been discussed in Section 3.1.3. For categorical u, the varying functions then have the form $\gamma_j(u) = \sum_{r=1}^{k} \gamma_{jr} I(u = r)$, which means that k parameters have to be estimated for each z_j (plus the intercept). Another interpretation of model (3.6) is that for each level of u a regression model with $q + 1$ parameters needs to be fit, with q being the number of covariates being considered in model (3.6). Even with moderate q and k, the number of regression parameters thus may become large and some sort of regularization might be necessary when fitting the model. Depending on the background and objective of the data analysis, the presumed structure of the model and the type of effect modifiers, different penalties have been proposed.

Gertheiss and Tutz (2012) and Oelker et al. (2014) distinguished nominal and ordinal u and proposed fusion penalties in the tradition of the fused lasso (Tibshirani et al., 2005) and the penalties discussed in Section 3.1.3. For nominal u, one can use

$$J(\boldsymbol{\gamma};\psi) = \psi \sum_{j=0}^{q} \sum_{r>s} |\gamma_{jr} - \gamma_{js}| + (1-\psi) \sum_{j=1}^{q} \sum_{r=1}^{k} |\gamma_{jr}|. \tag{3.7}$$

The main difference to penalty (3.4) is that $J(\boldsymbol{\gamma};\psi)$ now consists of two parts, the 'fusion' and the 'selection' parts, which might be differentially weighted using additional tuning parameter ψ. The first one, the fusion part, enforces collapsing categories of the effect modifier. The second term steers selection/exclusion of covariates z_j. In Equation (3.4) above, the selection part was implicitly contained in the fusion part because if the expected value of the response does not vary across the levels of the categorical predictor (given the other covariates), the latter is implicitly excluded from the model. With effect modifying u, by contrast, effect γ_j being constant across u-levels does not mean that z_j is irrelevant. If u is ordinal, the fusion part is modified such that only differences of adjacent coefficients are penalized (as done in Equation [3.5]):

$$J(\boldsymbol{\gamma};\psi) = \psi \sum_{j=0}^{p} \sum_{r=2}^{k} |\gamma_{jr} - \gamma_{j,r-1}| + (1-\psi) \sum_{j=1}^{p} \sum_{r=1}^{k} |\gamma_{jr}|. \tag{3.8}$$

The idea of penalties (3.7) and (3.8) is (a) to identify those levels of u where the effect of z_j on the response is the same/different and (b) to select those z_j that are relevant. It may also be the case that a predictor z_j is important on some levels of u only. Furthermore, when interpreting the results, it should be kept in mind that intercept γ_0 is contained (only) in the fusion term at penalties (3.7) and (3.8), analogously to categorical predictors at penalties (3.4) and (3.5), respectively. The intercept tells us how the mean response changes across levels of u if $z_j = 0$ for all (relevant) j. So if $z_j = 0$ is not plausible/of interest, centring/standardizing z_j before fitting the model is highly recommended. Alternatively, if the direct influence of u on the response is not of (primary) interest and k is not very large compared to the sample size, the intercept might be excluded from the penalty.

Penalties (3.7) and (3.8), however, are not the only way to handle categorical effect modifiers. Ollier and Viallon (2015), for instance, consider the situation where for most categories of u the effect of covariate z_j is the same. They decompose γ_{jr} into a 'global' effect $\bar{\gamma}_j$ and 'effect variations' δ_{jr} around $\bar{\gamma}_j$ and place (differentially weighted) L_1-penalties on both types of parameters. This yields a group of u-categories that are fused, those with global effect only, and some categories with effects differing from the global effect. The latter categories, however, cannot be fused among each other. If for a variable z_j both the global effect and each effect variation is set to zero, z_j is removed from the model.

If primary interest lies in excluding irrelevant predictors from the model, that is, identifying those z_j with no effect on any level of u, a groupwise penalty as described in Section 3.1.2 can be put on vectors $\boldsymbol{\gamma}_j = (\gamma_{j1}, \dots, \gamma_{jk})^\top$ (Negahban and Wainwright, 2011). If it can be assumed that relevant covariates have an effect on some levels of u only, a sparse group lasso (Simon et al., 2013; see Section 3.1.2) can be used. As a generalization of this, Jalali et al. (2013) considered a high-dimensional setting, that is, with large q, where—translated in our framework of varying coefficients models—many z_j have no effect at all, some have an effect on all u-levels and some have an effect on some u-levels only. Here, the matrix $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_q)^\top$ is split into a matrix $\boldsymbol{B}$ of 'shared rows' and a matrix $\boldsymbol{S}$ of 'non-shared rows'. That means for $\boldsymbol{B}$, it is assumed that rows are entirely zero or non-zero, whereas $\boldsymbol{S}$ is sparse in terms of having a few non-zero entries only, but without any further structure. Now a groupwise penalty (compare Section 3.1.2) is put on $\boldsymbol{B}$ for selecting rows, and a lasso-type penalty on $\boldsymbol{S}$ for selecting single elements. With those penalties, however, no fusion of u-categories is possible.

If it is only important to distinguish varying coefficients $\gamma_j(u)$ from constant ones (cf. Leng, 2009), but not to select variables, the appropriate choice is a groupwise selection penalty (like the group lasso) on pairwise differences $\gamma_{jr} - \gamma_{js}$, $r > s$. If levels of u are ordered, we may consider differences $\delta_{jr} = \gamma_{j,r+1} - \gamma_{jr}$ of adjacent coefficients only. If primary interest is in collapsing u-levels, that is, to find clusters of u-levels where regression models (not single coefficients) are identical, a variant of GASH-ANOVA by Post and Bondell (2013) could be used (compare Section 3.1.3). Then groups ϕ_{rs} would contain all pairwise differences $\gamma_{jr} - \gamma_{js}$, $j = 0, \dots, q$ of regression parameters on levels r and s.

As a generalization of model (3.6), one may consider the case of multiple (categorical) effect modifiers $u_1, \dots, u_m$, that is, $\boldsymbol{\gamma} = \boldsymbol{\gamma}(\boldsymbol{u}) = (\gamma_1(\boldsymbol{u}), \dots, \gamma_q(\boldsymbol{u}))^\top$ is a function of $\boldsymbol{u} = (u_1, \dots, u_m)^\top$. For estimating $\boldsymbol{\gamma}(\boldsymbol{u})$, Li et al. (2013) proposed a kernel approach using a variant of the kernel function of Aitchison and Aitken (1976). In a very recent manuscript, Peng et al. (2015) used a group lasso-type penalty for selecting relevant predictors z_j in this setting. Also, the fusion and selection penalty (3.7) can be extended to multiple effect modifiers (Gertheiss and Tutz, 2012), but if $m > 2$ the penalty becomes very complicated.

3.2 Another perspective on smoothing ordinal predictors

One may suggest treating ordinal predictors like metric ones and fitting flexible regression functions to circumvent the problem of (non)linearity in the group labels. In what follows, we will consider ordinal predictors in generalized additive models (GAMs) and show how the quadratic smoothing penalties considered so far can be interpreted as a basis function approach suited to the discrete nature of categorical covariates. Furthermore, we will sketch how mixed models methodology can be used to estimate penalty parameters and to develop an ANOVA procedure tailored to factors with ordered levels.

3.2.1 Ordinal predictors in generalized additive models

Suppose we have ordinal predictors $C_1, \ldots, C_p$ and (potentially) some continuous covariates $z_1, \ldots, z_q$. A GAM has the form (Hastie and Tibshirani, 1990)

$$\eta = \alpha + f_1(C_1) + \ldots + f_p(C_p) + s_1(z_1) + \ldots + s_q(z_q), \tag{3.9}$$

where the conditional mean of the response is given (as before) by $\mu = h(\eta)$, with known response function h. Functions $f_1(.), \ldots, f_p(.), s_1(.), \ldots, s_q(.)$, however, are unknown and need to be estimated. With continuous predictors, a popular approach is to expand $s_1, \ldots, s_q$ in basis functions $B_r(x)$, such as B-splines, to obtain $s_j(x) = \sum_r \theta_{rj} B_r(x)$, with basis coefficients θ_{rj}. When estimating those coefficients, a smoothing penalty is typically applied; see, for example, Eilers and Marx (1996). With ordinal predictor $C_j \in \{0, \ldots, k_j\}$, we could proceed in the same way. However, usual bases for continuous variables, such as B-splines, may not be the best choice for discrete covariates. Given ordinal C_j, we know all possible values. Hence, there is no need to fit a function outside of $\{0, \ldots, k_j\}$. So, a suitable basis for $f_j(.)$ is $A_r(x) = 1$, if $x = r$, and zero otherwise, $r = 0, \ldots, k_j$. If C_j is an ordinal variable, one can use a quadratic difference penalty, in analogy to Eilers and Marx (1996). Since basis functions $A_r(x)$ correspond exactly to the dummy coding as used so far, one ends up with the same approach as discussed in Section 3.1.1.

A very useful tool when fitting GAMs is the close link between quadratic penalties and mixed models. With an adequately chosen covariance matrix, penalized estimates of basis coefficients can typically be interpreted as predictions of random effects in a mixed models framework; see, for example, Ruppert et al. (2003) for details. The appropriate choice to obtain the smoothing penalty (3.1), for instance, is to specify differences $\delta_{jr} = \beta_{j,r} - \beta_{j,r-1}$ of adjacent dummy/basis coefficients as independent identically distributed (iid) normal random effects with mean zero and variance τ_j^2. For higher-order differences, the procedure is similar; see Gertheiss and Oehrlein (2011) for details. In the mixed models framework, the penalty parameter λ_j used for predictor C_j is proportional to $1/\tau_j^2$, and variance parameters τ_j^2 can be estimated by (restricted) maximum likelihood (see, e.g., Ruppert et al., 2003; Jiang, 2007; Wood, 2011, and references therein). This offers a very convenient way to determine penalty parameters, in particular when each ordinal predictor may have its own λ_j, resp. τ_j^2, and cross-validation over a multi-dimensional grid becomes expensive. In high-dimensional settings with a large number of ordinal covariates, however, the standard approach with a single λ is recommended. Nevertheless, generalized additive and mixed models offer a great framework for fitting regression models with some covariates being ordinal. In the R package `mgcv` (Wood, 2006), for instance, only the appropriate basis functions for ordinal predictors need to be defined to have the entire GAM machinery available.

3.2.2 ANOVA with ordinal factors

Mixed models also provide tools for further statistical inference, such as testing for constancy or linearity of a function (Claeskens, 2004; Crainiceanu et al., 2005; Scheipl et al., 2008). In the case of ordinal predictors, this can be used to develop an ANOVA procedure that takes the factor levels' ordering into account. Ordinal factors are often found in practice, for example, fertilizer levels in agricultural sciences, dose levels in pharmaceutical research or ordinal phenotypes (such as tumor stages) in medicine. Standard ANOVA as known from statistical textbooks, however, does not use this additional information provided by the levels' ordering.

Let us first consider one-way ANOVA with one ordinal factor having levels $r = 1, \ldots, k$. Then the usual model is $y_{ri} = \mu_r + \epsilon_{ri}$, with y_{ri} denoting the ith observation of the response on factor level r, $r = 1, \ldots, k$, $i = 1, \ldots, n_r$. The error terms ϵ_{ri} are assumed to be iid normal with mean zero and variance σ^2. The parameters of interest are the level-specific means μ_r, and the null hypothesis to be tested is H_0: $\mu_1 = \ldots = \mu_k$.

To use the factor's ordinal scale level, we proceed by imposing a discrete smoothing penalty within a mixed models framework. Instead of level-specific means μ_r, we consider differences $\delta_r = \mu_{r+1} - \mu_r$ between adjacent means, in analogy to dummy/basis coefficients above. As above, δ_r are specified as iid normal random effects with mean zero and variance τ^2 (Gertheiss, 2014). By using this specification, as sketched above, a penalty is implicitly imposed on the squared differences of adjacent means. Of course, such a penalty shrinking adjacent means towards each other is only reasonable when factor levels are ordered, as in this case we may assume that jumps in the means μ_r and μ_{r+1} between adjacent levels r and $r + 1$ are not extremely large. With the δ-parameterization, testing the null hypothesis, H_0: $\mu_1 = \ldots = \mu_k$ corresponds to testing H_0: $\delta_1 = \ldots = \delta_{k-1} = 0$. Since δ_r are specified as random effects with mean zero and variance τ^2, the latter null hypothesis is also equivalent to H_0: $\tau^2 = 0$.

This test problem, however, is nonstandard because under the null hypothesis, the parameter of interest (τ^2) is on the boundary of the parameter space. Nevertheless, a likelihood ratio test (LRT) or restricted likelihood ratio test (RLRT) can be used for testing, with the finite sample null distribution derived by Crainiceanu and Ruppert (2004); see Gertheiss (2014) for details. Unfortunately, the (R)LRT null distribution cannot be used directly to obtain p-values. However, Crainiceanu and Ruppert (2004) also give a fast algorithm for simulating from it. This algorithm is implemented in R package `RLRsim` (Scheipl et al., 2008; Scheipl and Bolker, 2013). A convenient wrapper for ordinal ANOVA as presented here is included in the R package `ordPens` (Gertheiss, 2015). Extensive simulation studies and real data analyses (Gertheiss, 2014; Sweeney et al., 2016) indicate that RLRT is preferable to LRT, and that an appropriate RLRT as discussed here is often superior to standard ANOVA when factor levels are ordered. If more than one ordinal factor is present (see Gertheiss, 2014), the distributions given by Crainiceanu and Ruppert (2004) cannot be used anymore, but approximate solutions exist (Greven et al., 2008).

4 Categorical response variables

In what follows, we will discuss penalty methods for categorical responses. For a categorical response variable, regularization to obtain sparsity has to be adapted to the multivariate nature of the response.

4.1 Multinomial models

The classical model for response categories $Y \in \{1, \dots, K\}$ is the multinomial model, which can be seen as a multivariate GLM. In its generic form, it specifies

$$\pi_r = P(Y = r|\boldsymbol{x}) = \frac{\exp(\beta_{r0} + \boldsymbol{x}^\top \boldsymbol{\beta}_r)}{\sum_{s=1}^{K} \exp(\beta_{s0} + \boldsymbol{x}^\top \boldsymbol{\beta}_s)} = \frac{\exp(\eta_r)}{\sum_{s=1}^{k} \exp(\eta_s)}, \tag{4.1}$$

where $\boldsymbol{\beta}_r^\top = (\beta_{r1}, \dots, \beta_{rp})$. Since parameters $\beta_{10}, \dots, \beta_{K0}$, $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K$ are not identifiable, additional constraints are needed. Typically, one of the response categories is chosen as reference category, for example, by setting $\beta_{K0} = 0$, $\boldsymbol{\beta}_K = \mathbf{0}$.

In the multinomial logit model, the effect of covariates is specified by the linear predictors $\eta_r,\ r = 1, \dots, K-1$, which correspond to the log odds between category r and the reference category K. Here, we will consider a more general version of the model that allows for category-specific variables. For example, when the response is the choice of transportation mode, the attributes could be price and duration, which vary across the alternatives and therefore are category-specific. Then, in addition to the global predictors $\boldsymbol{x}$, a set of category-specific predictors $\boldsymbol{w}_1, \dots, \boldsymbol{w}_K$ is available, where $\boldsymbol{w}_r$ contains the attributes of category r. The set of linear predictors generalizes to

$$\eta_{ir} = \beta_{r0} + \boldsymbol{x}^\top \boldsymbol{\beta}_r + (\boldsymbol{w}_{ir} - \boldsymbol{w}_{iK})^\top \boldsymbol{\alpha}, \qquad r = 1, \dots, K-1. \tag{4.2}$$

The second term specifies the effect of the global variables, and the third term specifies the effect of the difference $\boldsymbol{w}_{ir} - \boldsymbol{w}_{iK}$ on the choice between category r and the reference category. When choosing a transport mode, it can be the difference in price that has an effect on the choice.

If the simple lasso is used by penalizing all parameters by a sum over $|\beta_{rj}|$ (Friedman et al., 2010), one does not necessarily obtain variable selection but parameter selection because if one of the parameters β_{rj} is not deleted, the whole variable x_j is still in the model. To select variables, one has to group all the parameters that correspond to one variable and penalize them simultaneously. This is obtained by the categorically structured (CATS) penalty

$$J(\boldsymbol{\beta}) = \psi \sum_{j=1}^{p} \phi_j \, ||\boldsymbol{\beta}_{.j}|| \ + \ (1-\psi) \sum_{l=1}^{L} \varphi_l \, |\alpha_l|, \tag{4.3}$$

where $\boldsymbol{\beta}_{.j}^{\top} = (\beta_{1j}, \ldots, \beta_{K-1,j})$ collects all parameters linked to predictor x_j; ψ is an additional tuning parameter that balances the penalty on the global and the category-specific variables. The parameters ϕ_j and φ_l are weights that assign different amounts of penalization to different parameter groups. Typically, they are chosen by $\phi_j = \sqrt{K-1}$ and $\varphi_l = 1$.

The penalty enforces variable selection, that is, all the parameters in $\boldsymbol{\beta}_{.j}$ are simultaneously shrunk towards zero. It is strongly related to the classical group lasso (Yuan and Lin, 2006; see Section 3.1.2). However, in the group lasso, the grouping refers to the parameters that are linked to a categorical predictor within a univariate regression model, whereas in the present model, grouping arises from the multivariate response structure.

By representing the logit model as a multivariate GLM (Fahrmeir and Tutz, 1997; Tutz, 2012), the score function and the Fisher matrix have similar forms as in univariate GLMs. However, specific software is needed when penalties are included. The R package `MRSP` (Pössnecker, 2014) for the fitting of multinomial regression models with a structured penalization uses the fast iterative shrinkage thresholding algorithm (FISTA; Beck and Teboulle, 2009).

Early suggestions for regularization in multinomial logit models (Krishnapuram et al., 2005; Friedman et al., 2010) used L_1-type penalties that shrink all the parameters individually. They do not use the natural grouping of coefficients and cannot directly promote variable selection (see above). Grouped variable selection was used in general multivariate regression, among others, by Turlach et al. (2005) and Argyriou et al. (2007). Preliminary versions of the group lasso within the multinomial regression framework have been considered by Tutz (2012), Vincent and Hansen (2014), Chen and Li (2013) and Simon et al. (2013). The version with category-specific variables was proposed by Tutz et al. (2015). They also consider the case with categorical predictors, in which the penalty also has to account for grouping of the parameters linked to covariates. Simple lasso-type penalties for multinomial regression with category-specific variables were considered by Mauerer et al. (2015).

As an example, we consider the choice of travel mode of $n=210$ passengers in Australia. The data has been used by Greene (2003) and is available from the R package `Ecdat` (Croissant, 2006). The alternatives of travel mode were air, train, bus and car, which have frequencies of 0.276, 0.300, 0.142 and 0.280, respectively. Here, car serves as the reference category. As the global variables, which do not vary over categories, we consider household income (income) and size of the travel group (size); as category-specific variables, we consider terminal waiting time (wait), the vehicle cost component (vcost), the total cost (tcost) and the travel time in the vehicle (timevcl). All variables have been standardized to have variance one. Figure 4 shows the coefficient build-ups for the coefficients as a function of the tuning parameter with the weight ψ fixed at $\psi = 0.5$. For the global variables, one has three coefficient paths, one for each response category; for the category-specific variables, only one path is given. The category-specific parameter estimates of the global variables obtained by using the CATS penalty are given in Table 3; the estimates for the category-specific

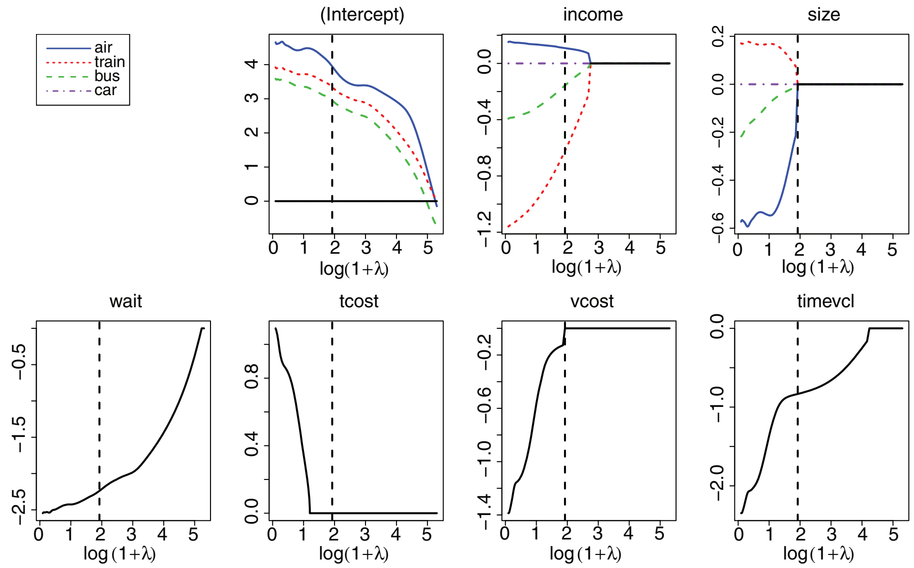

Figure 4 Coefficient build-ups for travel mode data; estimates chosen via BIC are marked by the vertical dashed line.

Table 3 Estimated category-specific coefficients for travel mode data

	intercept	income	size
air	3.954	0.109	0
train	3.351	−0.621	0
bus	2.963	−0.156	0

variables are given in Table 4. It is seen that the size as well the costs are excluded from the model. With income in the predictor, the travel costs seem to be negligible.

In the modelling of choice data, sometimes also for category-specific predictors, category-specific effects are assumed. If one assumes that latent utilities that characterize alternatives are given by $U_r = u_r + \varepsilon_r$, $u_r = \boldsymbol{x}^\top \boldsymbol{\gamma}_r + \boldsymbol{w}_r^\top \boldsymbol{\alpha}_r$ and alternatives are determined by the principle of maximum random utility, that is, $Y = r \Leftrightarrow U_r = \max_{j=1,\dots,K} U_j$, one obtains the logistic model if the $\varepsilon_1, \dots, \varepsilon_K$ are iid variables with distribution function $F(x) = \exp(-\exp(-x))$ (e.g., McFadden, 1973; Yellott, 1977). The predictors are obtained as differences of utilities and have the form

$$\eta_r = u_r - u_K = \boldsymbol{x}^\top \boldsymbol{\beta}_r + \boldsymbol{w}_r^\top \boldsymbol{\alpha}_r - \boldsymbol{w}_K^\top \boldsymbol{\alpha}_K,$$

Table 4 Estimated global coefficients for travel mode data

wait	vcost	tcost	timevcl
−2.242	0	0	−0.844

where $\boldsymbol{\beta}_r = (\boldsymbol{\gamma}_r - \boldsymbol{\gamma}_K)$. The total set of parameters that defines the total vector now contains the $K-1$ $\boldsymbol{\beta}$-parameters $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_{K-1}$ and the K $\boldsymbol{\alpha}$-vectors $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_K$. If $\boldsymbol{\alpha}_1 = \dots = \boldsymbol{\alpha}_K = \boldsymbol{\alpha}$, one obtains the special model (4.2). As far as the coefficients $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_K$ are concerned, one has to distinguish between the cases of category-specific parameters $\boldsymbol{\alpha}_r$, global effects, that is, $\boldsymbol{\alpha}_1 = \dots = \boldsymbol{\alpha}_K = \boldsymbol{\alpha}$ and no effect, that is, $\boldsymbol{\alpha}_1 = \dots = \boldsymbol{\alpha}_K = 0$. Appropriate penalties for this effect type selection are the same as considered in the next section for ordinal response data. For multinomial responses, however, the corresponding penalties seem not to have been used before and no software seems yet available.

4.2 Ordinal regression and discrete survival models

As before with predictors, categorical responses can be distinguished into nominal and ordinal variables. Typically, different types of models are used for those two, with model (4.1) from above being the typical one for nominal variables. In what follows, we will consider models for ordinal response and discrete survival models, which are closely related.

4.2.1 Basic models

If the categorical response has ordered levels, the multinomial model (4.1) is often replaced by the cumulative model

$$P(Y \leq r|\boldsymbol{x}) = F(\beta_{r0} + x_1\beta_{r1} + \dots x_p\beta_{rp}), \tag{4.4}$$

where $F(\cdot)$ is a fixed distribution function. The most widespread model uses the logistic function $F(\eta) = \exp(\eta)/(1+\exp(\eta))$ yielding

$$\log\left(\frac{P(Y \leq r|\boldsymbol{x})}{P(Y > r|\boldsymbol{x})}\right) = \beta_{r0} + x_1\beta_{r1} + \dots x_p\beta_{rp}. \tag{4.5}$$

Although the left-hand side only makes sense if levels are ordered, the model still has the same complexity as the multinomial model. So to actually take the ordinal nature of the response into account, it is typically assumed that β_{rj} is constant over r, that is, $\beta_{rj} = \beta_j$ for all j, r. This leads to the so-called proportional odds model. While the general model uses the predictor $\eta_r = \beta_{r0} + \boldsymbol{x}^\top\boldsymbol{\beta}_r$ with $\boldsymbol{\beta}_r^\top = (\beta_{r1}, \dots, \beta_{rp})$, in the proportional odds model, the predictor is simplified to $\eta_r = \beta_{r0} + \boldsymbol{x}^\top\boldsymbol{\beta}$, where $\boldsymbol{\beta}^\top = (\beta_1, \dots, \beta_p)$. The strength of the proportional odds model is that two populations characterized by the covariate values $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ can be compared by

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

Since the proportion of the 'cumulative odds' $P(Y \leq r|\boldsymbol{x})/P(Y > r|\boldsymbol{x})$ does not depend on the category, interpretation of parameters is much easier than in the general model, see also, for example, Agresti (2010).

Alternative ordinal regression models are the adjacent categories model

$$P(Y = r + 1|\boldsymbol{x})/P(Y = r|\boldsymbol{x})] = F(\beta_{0r} + \boldsymbol{x}^\top \boldsymbol{\beta}_r),\ r = 1, \ldots, K - 1,$$

and the sequential model

$$P(Y = r|Y \geq r, \boldsymbol{x}) = F(\beta_{0r} + \boldsymbol{x}^\top \boldsymbol{\beta}_r),\ r = 1, \ldots, K - 1,$$

where $F(\cdot)$ is again a fixed distribution function, see Agresti (2010) and Tutz (2012). In particular, the sequential model is interesting because it is equivalent to models used in the modelling of discrete survival data. Let the response Y refer to discrete time, for example, to months of unemployment or days spent in the hospital. Then the conditional probability $\lambda(r|\boldsymbol{x}) = P(Y = r|Y \geq r, \boldsymbol{x})$ is called the 'discrete hazard' and corresponds to the probability that a transition takes place, for example, from unemployment to employment given the individual is still unemployed at time r. In the corresponding discrete hazards model $\lambda(r|\boldsymbol{x}) = F(\beta_{0r} + \sum_j x_j \beta_{jr})$ the β_{jr} are time-varying coefficients. That means, for example, that the effect of covariates like gender or age on the conditional probability of finding a job varies over the time of unemployment, which is often a realistic assumption. In the case of survival models, the parameters β_{0r} represent the baseline hazard, which is shared by all individuals. Estimation of discrete survival (including censored data) can be obtained by considering the conditional survival given fixed time, that is, one considers the binary events $Y = r|Y \geq r$ and $Y > r|Y \geq r$. Then discrete time is an ordered categorical predictor and the time-varying effects can be seen as an interaction between the x-predictors and time. For details on estimation including censored data, see, for example, Fahrmeir and Tutz (1997), Chapter 9, and Tutz and Schmid (2016).

For the adjacent categories and sequential models, the same holds as for the cumulative models. If the weight parameters depend on the category, the models have the same complexity as the multinomial model. Therefore, reduction of parameters is desirable to obtain sparser models with parameters that are easier to interpret.

4.2.2 Penalization

Although the assumption that parameters do not vary over response categories (such as the proportional odds assumption in cumulative models) is very popular and reduces model complexity, it is quite restrictive and sometimes not reasonable. On the other hand, the number of parameters β_{jr} to estimate with non-proportional odds models can be large, which makes estimation and interpretation more challenging. As before with categorical predictors, penalties can help here.

Smooth effects. An assumption that is often reasonable is that effects β_{rj} and $\beta_{r+1,j}$ do not change drastically between two adjacent categories r and $r + 1$. A potential penalty to stabilize estimates is $J(\boldsymbol{\beta}) = \sum_{j=1}^{p} \sum_r (\beta_{r+1,j} - \beta_{rj})^2$. In discrete hazard models, it means that the time-varying effect is smooth over time. Here, the same penalty is often used for the baseline hazard. By using $J(\boldsymbol{\beta}) = \sum_r (\beta_{r+1,0} - \beta_{r0})^2$ and appropriate choice of the smoothing parameter, one enforces that the baseline hazard is smooth.

Structural breaks in sequential and discrete survival models. In particular for discrete hazard models, the fusion penalty $J(\boldsymbol{\beta}) = \sum_{j=1}^{p} \sum_{r} |\beta_{r+1,j} - \beta_{rj}|$ is attractive. Then some coefficients would be fused, giving us groups of (time) categories for which the effect of a covariate is stable. It might be used to find structural breaks in the effect of covariates.

Effect type selection. In many applications, one wants to distinguish between three forms of effects. For variable j, the most general is the category-varying effect β_{rj}, the more parameter sparse effect β_j, which does not vary over categories, and the no-effect case in which $\beta_{rj} = \beta_j = 0$. A penalty that is able to distinguish between these hierarchically nested effect types is $J(\boldsymbol{\beta}) = \sum_{j=1}^{p} \lambda_1 \sqrt{\sum_{r=1}^{K-1} \beta_{rj}^2} + \lambda_2 \sqrt{\sum_{r=1}^{K-1} (\beta_{r+1,j} - \beta_{r,j})^2} = \sum_{j=1}^{p} \lambda_1 ||\boldsymbol{\beta}_j|| + \lambda_2 ||\boldsymbol{D}_j \boldsymbol{\beta}_j||$, where $\boldsymbol{\beta}_j^\top = (\beta_{1j}, \ldots, \beta_{K-1,j})$ again collects all parameters linked to predictor x_j, and $\boldsymbol{D}_j$ generates the differences between adjacent categories. The second part of the penalty is of the group-lasso type and enforces that parameters of specific variables are set equal. For large enough λ_2, one obtains $\beta_{1j} = \cdots = \beta_{K-1,j} = \beta_j$ for some variables x_j. The first term enforces that parameters linked to the same variable are set to zero simultaneously and therefore the respective variable can be excluded from the model. If for a specific covariate all coefficients are set equal, the effect of this variable is time-constant in the discrete survival model. For the cumulative logit model, it means that the covariate has proportional odds across all categories, which gives a partial proportional odds model (Peterson and Harrell, 1990). For further properties of these penalties, see Pössnecker and Tutz (2016).

5 Subject-specific models

In this section, models for repeated measurements are considered. For repeated measurements, penalty methods provide an alternative to random effects models with good performance in terms of estimation accuracy. Repeated measurements can be represented by $(y_{ij}, \boldsymbol{x}_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$, where y_{ij} denotes the response of unit i at measurement occasion j, and $\boldsymbol{x}_{ij}$ is a vector of covariates that potentially varies across measurements. A common approach to model heterogeneity across units is to use random effects models. In a 'generalized linear mixed effects model' (GLMM), the structural assumption specifies that the conditional mean, $\mu_{ij} = E(y_{ij}|b_i, \boldsymbol{x}_{ij}, z_{ij})$, has the form

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{b}_i, \tag{5.1}$$

where g is a monotonic and continuously differentiable link function, $\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}$ is a linear parametric term with parameter vector $\boldsymbol{\beta}^\top = (\beta_0, \beta_1, \ldots, \beta_p)$ that includes an intercept and z_{ij} is a covariate vector associated with random effects. The second term contains the random effects that model the heterogeneity of the units. For the random effects, one assumes a distributional form, typically a normal distribution, $\boldsymbol{b}_i \sim N(0, \boldsymbol{Q})$.

The focus of the random effects models is on the fixed effects; the distribution of the random effects is mainly used to account for the heterogeneity of the units. Although it is the most popular model that accounts for heterogeneity, it has several drawbacks. The assumption of a specific distribution for the random effects may affect the inference, see, for example, Agresti et al. (2004) and Litière et al. (2007). In particular, if the distributional assumption is far from the data generating distribution, inference can be strongly biased. Moreover, assuming a continuous distribution means that the effects of units cannot be identical. Therefore, by assumption, no clustering of units is available. One further aspect is that it is assumed that the random effects and the covariates observed per second level unit are independent, a criticism that has a long tradition, particularly in the econometric literature. If random effects and covariates are correlated, the estimates obtained from random effects models can be poor. For an overview on the choice between fixed and random effects models, see, for example, Townsend et al. (2013).

As an alternative, we consider the 'fixed effect or subject-specific model'

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{\beta}_i. \tag{5.2}$$

The model specifies that each unit has its own coefficient $\boldsymbol{\beta}_i$, $i = 1, \ldots, n$.

The problem with these models is that the large number of parameters can render the estimates unstable and encourage overfitting. Typically, there is not enough information available to distinguish among all the units; but under the assumption that observations form clusters with respect to their effect on the response, the number of parameters can be reduced and estimates are available. The tool to obtain sparsity of subject-specific parameters and clusters is the use of the penalty

$$J(\boldsymbol{\beta}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n) = \sum_{r>m} ||\boldsymbol{\beta}_r - \boldsymbol{\beta}_m||. \tag{5.3}$$

If $\lambda = 0$, one obtains the unpenalized estimates of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n$ and each unit has its own parameter. If $\lambda \to \infty$, the penalty enforces that the estimates of all subject-specific parameters are the same. It has been demonstrated in Tutz and Oelker (2016) that the method outperforms the random effects model, particularly if correlation between the random intercept and the random effect, the so-called level 2 endogeneity, is present. The model is also compared to alternative approaches as the discrete mixture model (Aitkin, 1999).

As an example, we consider the modelling of the effect of beta blockers on mortality after myocardial infarction, see also Aitkin (1999). In a 22-centre clinical trial, for each centre, the number of deceased/successfully treated patients in control/treatment groups was observed. The binary response (1 = deceased/0 = not deceased) suggests a mixed logit model of the form

$$\text{logit}\, P(y_{ij} = 1) = \beta_0 + \beta_{i0} + \beta_T \cdot \text{Treatment}_{ij}, \quad i = 1, \ldots, 22 \text{ Centres}, \tag{5.4}$$

where Treatment$_{ij}$ codes the treatment in hospital i for patient j. If β_{i0} is replaced by a random effect b_i with normal distribution, implicitly the hospitals are considered as a random sample and all the effects of hospitals are assumed to differ. In contrast, the fixed effects model with regularization assumes that some of the hospitals have the same treatment effect. Figure 5 shows the coefficient build-ups against regularization, where the vertical line refers to the cross-validated choice of the tuning parameter. Apparently, there are essentially five clusters of hospitals with comparable effect sizes within clusters (but two clusters only consisting of a single hospital).

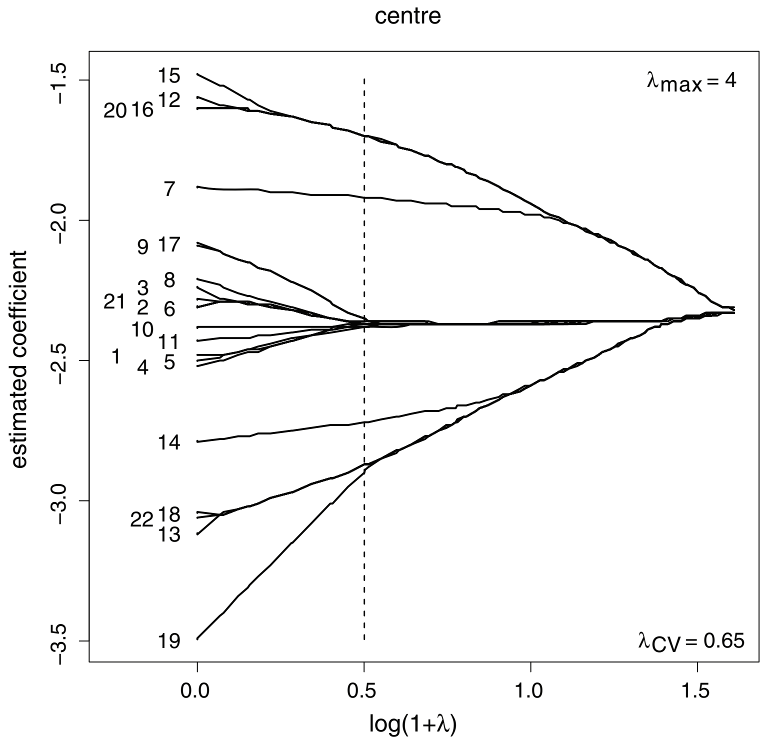

Figure 5 Coefficient path for beta blocker data showing the estimated fixed effects of hospitals; vertical line shows the estimates for cross-validated choice of the tuning parameter.

6 Paired comparison models

In what follows, we briefly consider paired comparisons, which is a well-established method to measure the relative preference or dominance of objects or items. The aim is to find the underlying preference scale by presenting the items or objects in pairs (Bradley, 1976). Paired comparisons have been extensively used in psychometrics and marketing, see, for example, David (1988) and Moore and Lehmann (1989). They are also found in sports whenever two players or teams compete in a tournament (Fahrmeir and Tutz, 1994; Glickman and Stern, 1998). The non-observable scale to be found refers to the strengths of the competitors. Paired comparison data for binary or ordinal responses can be seen as special repeated measurements with a multinomial response, typically modelled as fixed effects models. As in multinomial models, regularization helps to identify relevant structures in the data. Overviews on paired comparison modelling are found in the reviews of David (1988) and Cattelan (2012).

A concrete example of a paired comparison model for ordinal responses is the cumulative Bradley-Terry-Luce (BTL) model. Let $\{a_1, \dots, a_m\}$ denote the set of objects or items to be compared in a paired comparison experiment and $Y_{(r,s)} \in \{1, \dots, K\}$ denote the ordinal response when objects a_r and a_s are compared. The response $Y_{(r,s)} = 1$ corresponds to a strong preference of a_r over a_s and $Y_{(r,s)} = K$ corresponds to a strong preference of a_s over a_r. The cumulative model (Tutz, 1986) has the form

$$P(Y_{(r,s)} \le k) = \frac{\exp(\theta_k + \gamma_r - \gamma_s)}{1 + \exp(\theta_k + \gamma_r - \gamma_s)}, \tag{6.1}$$

with the restriction $\sum_{r=1}^{m} \gamma_r = 0$. The parameters $\gamma_1, \dots, \gamma_m$ contain the attractiveness or strength of the objects. With increasing γ_r, the probability for low response categories, and therefore the strong preference of a_r over a_s, increases while the probability for large response categories denoting dominance of a_s decreases. The threshold parameters determine the preference for specific categories. The threshold for the last category K is $\theta_K = \infty$ so that $P(Y_{(r,s)} \le K) = 1$ holds. Further restrictions on the threshold parameters are useful to ensure equal probabilities for corresponding categories if the order of the paired comparison is reversed (see Tutz, 1986). If an order effect is required, for example, to model the home advantage in sport competitions, an additional parameter can be included. Formally, model (6.1) is a cumulative logit model, also called a proportional odds model (compare Section 4.2). Alternatives with different link functions also exist (see, e.g., Agresti, 1992).

If model (6.1) is used to model the strengths of competitors, one obtains for each competitor, for example, a football team, a strength parameter γ_r. By using a difference penalty that contains all the differences $|\gamma_r - \gamma_s|$, one can identify clusters of competitors that share the same strength. Penalties of this type have been used by Masarotto and Varin (2012) and Tutz and Schauberger (2015) to model the strengths of teams in sports tournaments.

In extended versions of the BTL model, heterogeneity of respondents can be modelled by including explanatory variables. Let $Y_{i(r,s)}$ denote the response of person i for a given pair of items (r, s) and $\boldsymbol{x}_i^\top = (x_{i1}, \dots x_{ip})$ be a person-specific covariate vector. It is assumed that the strength of the preference of item a_r for person i is determined by $\gamma_{ir} = \beta_{r0} + \boldsymbol{x}_i^\top \boldsymbol{\beta}_r$. That means there is a global strength parameter β_{r0} but the effective strength is modified by the covariates. The parameter $\boldsymbol{\beta}_r^\top = (\beta_{r1}, \dots, \beta_{rp})$ contains the effect of the covariates on item a_r. The corresponding model has the form

$$P(Y_{i(r,s)} \leq k \mid \boldsymbol{x}_i) = \frac{\exp(\theta_k + \beta_{r0} - \beta_{s0} + \boldsymbol{x}_i^\top(\boldsymbol{\beta}_r - \boldsymbol{\beta}_s))}{1 + \exp(\theta_k + \beta_{r0} - \beta_{s0} + \boldsymbol{x}_i^\top(\boldsymbol{\beta}_r - \boldsymbol{\beta}_s))}, \tag{6.2}$$

with the sum-to-zero constraints $\sum_{r=1}^{m} \beta_{rj} = 0, j = 0, 1, \ldots, p$ for identifiability. Since each item has its own parameter $\boldsymbol{\beta}_r$, one typically has too many parameters to obtain stable maximum likelihood estimates. A penalty that yields clusters of items which share the same effect of covariates is given by

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{p} \sum_{r<s} w_{rsj} |\beta_{rj} - \beta_{sj}|.$$

The penalty ensures stable estimates and helps to find clusters. Schauberger and Tutz (2015) used it to model the preference of political parties in Germany.

It is much more difficult to include item or team-specific explanatory variables. Let $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m$ denote the vectors of explanatory variables linked to item or team a_r. In sports tournament data, it can be, for example, the budget of a club, which should be influential because the budget determines if a club is able to get the best and most expensive players. In a model that accounts for team-specific variables, the strength of the teams γ_r is replaced by $\gamma_r + \boldsymbol{w}_r^T \boldsymbol{\beta}$ yielding the linear predictor

$$\eta_{rst} = \alpha_r + \theta_t + \gamma_r - \gamma_s + (\boldsymbol{w}_r - \boldsymbol{w}_s)^T \boldsymbol{\beta}.$$

However, in this model, parameters are not identifiable because the parameters γ_r cannot be distinguished from the parameters $\tilde{\gamma}_r = \gamma_r + \boldsymbol{w}_r^T \boldsymbol{\beta}$. Therefore, additional constraints are needed to obtain unique estimates.

One way to constrain estimates is to use a random effects model. By assuming that the strengths are random effects, for example, by assuming $\gamma_r \sim N(0, \sigma^2)$, parameters can be estimated within a random effects model, see Turner and Firth (2012) who used random effects models to account for correlations between responses. An alternative approach is to use penalized estimation procedures within a fixed effects model framework. Assuming that teams are clustered, one can again use the penalty that contains all the differences $|\gamma_r - \gamma_s|$. It penalizes the abilities that are not explained by covariates, γ_r, $r = 1, \ldots, m$, but not the parameter $\boldsymbol{\beta}$. If the tuning parameter gets large, $\lambda \to \infty$, all strength parameters γ_r are estimated as identical and the total strength is determined solely by $\boldsymbol{w}_r^T \boldsymbol{\beta}$. By using a regularization term with positive tuning parameter, the parameters are defined and estimable, compare also Friedman et al. (2010), where this procedure has been used in overparameterized multinomial regression models.

7 Alternative methods

The focus of this article is on penalty methods for categorical variables. Penalty-based methods, however, are not the only way to regularize when fitting and selecting models with categorical variables. In this section, we hence briefly discuss some alternative regularization approaches that can be used to model categorical data.

7.1 Boosting

Boosting is an algorithmic regularization method that also allows selecting predictors. Statistical theory of boosting was developed by Friedman (2001), Friedman et al. (2000) and Bühlmann and Yu (2003), among others. An overview on gradient boosting is found in Bühlmann and Hothorn (2007), likelihood-based boosting was considered, for example, by Tutz and Binder (2006); for a brief introduction, comparison of both approaches and discussion, also see Mayr et al. (2014a), Mayr et al. (2014b) and Bühlmann et al. (2014). The basic principle of boosting to use weak learners to iteratively improve selected regression coefficients together with early stopping yields variable selection and has been successfully applied in high-dimensional settings. If one uses blockwise boosting, which updates blocks of coefficients, for example, all the coefficients that are connected to a categorical predictor, an alternative to the group lasso with similar properties is obtained. For example, when selecting ordinal predictors, the smoothing penalty-discussed here (Gertheiss and Tutz, 2009; see Section 3.1.1) can also be incorporated in a boosting procedure (Gertheiss et al., 2011; Hofner et al., 2011a; Hofner et al., 2011b). The selection of covariates in multinomial regression models by boosting techniques was considered by Zahid and Tutz (2013). For the clustering of categories as considered in Section 3.1.3, however, penalty-based estimation seems preferable since boosting algorithms that identify clusters are hard to obtain.

7.2 Finite mixtures

Another method to identify clusters of subjects with the same tendency to respond, which has been used in particular for repeated measurement, is finite mixture models. In finite mixtures of GLMs, it is assumed that the density or mass function of observation y given $\boldsymbol{x}$ is a mixture

$$f(y|\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k f_k(y|\boldsymbol{x}, \boldsymbol{\beta}_k, \phi_k), \tag{7.1}$$

where $f_k(y|\boldsymbol{x}, \boldsymbol{\beta}_k, \phi_k)$ represents the kth component of the mixture that follows a simple exponential family parameterized by the parameter vector from the model $\mu_k = \mathrm{E}(y|\boldsymbol{x}, k) = h(\boldsymbol{x}^\top \boldsymbol{\beta}_k)$ with response function $h(\cdot)$ and the dispersion parameter ϕ_k. The unknown component weights follow $\sum_{k=1}^{K} \pi_k = 1, \pi_k > 0, k = 1, \dots, K$.

For hierarchical settings like repeated measurements on persons, the components can be linked to the second level units represented by persons. Let $C = \{1, \dots, n\}$ denote the set of units that are observed. Then, one specifies for the mean of the jth measurement of the *i*th unit in the *k*th component

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

where $\beta_{k(i)}$ denotes that the component membership is fixed for each second level unit, that is, $\beta_{k(i)} = \beta_k$ for all $i \in C_k$, where $C_1, \ldots, C_K$ is a disjunct partition of C. Therefore, the units are clustered into subsets with identical intercepts with the total vector of coefficients being given by $(\beta_1, \ldots, \beta_K, \boldsymbol{\beta}^{\top})$.

Mixture models were, for example, considered by Follmann and Lambert (1989), Aitkin (1999) and Fruehwirth-Schnatter (2006). Grün and Leisch (2008) consider identifiability for mixtures of multinomial logit models and provide the R package flexmix (Leisch and Grün, 2012) with various applications.

For the estimation of mixture models with a fixed number of mixture components, typically, the EM-algorithm is employed. In simulation studies (Tutz and Oelker, 2016), it has been shown that mixture models tend to underestimate the number of clusters; regularization methods as considered in Section 5 showed better performance here.

7.3 Tree-based approaches

Penalization is a useful tool to identify relevant categorical predictors and clusters of categories but becomes computationally demanding if the number of categories is very large. In this case, approximations or alternative procedures have to be used. An alternative is recursive partitioning methods, also called trees. The big advantage of classical trees or recursive partitioning procedures as CART (Breiman et al., 1984) and C4.5 (Quinlan, 1993) is that they automatically find interactions. The concept of interactions is at the core of recursive partitioning. However, if one uses just the first splits in nominal or categorical ordered predictors accounting for the other variables, one can use recursive partitioning to identify clusters. First approaches that use this method have been considered by Tutz and Berger (2014) for generalized regression models and Bürgin and Ritschard (2015) in varying coefficient regression.

7.4 Dirichlet processes

An alternative Bayesian approach to model clustered random effects is based on Dirichlet processes. Dirichlet processes were proposed by Ferguson (1973) and studied, for example, by Sethuraman (1994) and Hjort et al. (2010). The main advantage of Dirichlet processes is their cluster property, which allows flexibly modelling discrete distributions. Assuming a Dirichlet process for the distribution of random effects creates ties among the random effects. The resulting Dirichlet process mixture yields clusters of units. Dirichlet process priors have been used within the linear mixed models framework by Bush and MacEachern (1996) and Müller and Rosner (1997). A frequentist approach to linear mixed models with Dirichlet process mixtures was given by Heinzl and Tutz (2013); a combination of Dirichlet processes and fusion penalties was considered in Heinzl and Tutz (2014, 2016). Although an interesting tool for linear models, extensions to generalized mixed models seem not to have been considered so far.

7.5 Bayesian approaches with spike and slab priors

Bayesian approaches to effect fusion may also be based on the spike and slab distribution (George and McCulloch, 1993). A general framework of structured additive regression with spike and slab priors was proposed by Scheipl et al. (2012). They propose a Markov chain Monte Carlo approach with good mixing and convergence properties. Sparse Bayesian modelling of the effects of nominal and ordinal categorical predictors within a regression framework was considered more recently by Pauger and Wagner (2014). Instead of the fusion penalty discussed in Section 3.1.3, a spike and slab prior is placed on appropriate differences of regression coefficients. The approach is attractive and competes directly with the fusion penalties considered here. Evaluation of the method is certainly an interesting topic for future research.

7.6 Kernel methods

In Section 3.1.5, we already mentioned the use of kernel methods in varying coefficient models with categorical effect modifiers. The strategy sketched there can also be useful when having categorical predictors in a more general framework. Racine and Li (2004), for instance, presented a kernel-based approach for nonparametric estimation of regression functions with both continuous and categorical predictors. The use of regression splines for the continuous predictors under the presence of categorical covariates has been considered by Ma and Racine (2013) and Ma et al. (2015). Each of those methods use a variant of the Aitchison and Aitken (1976) kernel for the categorical predictors; more precisely,

$$K(C_j, c_j, \lambda_j) = \begin{cases} 1 & \text{if } C_j = c_j, \\ \lambda_j & \text{otherwise.} \end{cases}$$

Here, $\lambda_j \in [0, 1]$ is the smoothing parameter for predictor C_j, but complexity can be reduced by setting $\lambda_s = \lambda$ for all j. Using those kernels, the weight of each observation in the nonparametric estimation procedure is determined by the product kernel $\prod_j K(C_j, c_j, \lambda_j)$. For ordinal predictors, kernels can be replaced by $K(C_j, c_j, \lambda_j) = \lambda_j^d$, with $d = |C_j - c_j|$. This makes sure that a higher weight is assigned to observations that are 'close' with respect to the levels' ordering. For kernel-based smoothing of categorical data in contingency tables, see, for example, Simonoff (1996) and Simonoff and Tutz (1999). So far, kernel- and penalty-based methods for categorical data have developed rather independently of each other. A thorough comparison of both strategies is still lacking.

8 Concluding remarks

In this article we presented various penalty methods that can be used when fitting statistical models with categorical variables involved. The use of penalty methods for qualitative data is relatively new and gained much less attention than penalty methods for quantitative variables, which are particularly useful and popular in high-dimensional models. With categorical data, the number of parameters often becomes large even for a moderate number of covariates due to the parametrization typically used: dummy variables in the case of categorical predictors and category-specific parameters for categorical responses. So it is not surprising that penalty-based methods can also be very helpful when modelling categorical data; increased use can be expected for the future.

We mainly distinguished between models with categorical predictors and models with categorical response, and in each case, between nominal and ordinal variables. We saw that with the right penalty being chosen, not only variability of estimates reduces, but also structures in the data can be revealed, making interpretation of results much easier. Interestingly, the same type of penalty, such as groupwise or fusion penalties, can often be used with both categorical predictors and response, but in a different way and with different results and interpretation. In this article, we wanted to illustrate the potential of penalty methods in statical modelling with categorical data. Certainly not all penalties that have ever been used or could be used for categorical variables were mentioned—any substantial gap will hopefully be filled by the discussants.

A variety of alternative approaches for regularizing models with categorical variables has been proposed. We sketched only some of them very briefly. Sometimes there is even a very close connection between methods. The choice of a certain penalty that drives estimates into a certain direction, for instance, might also be interpreted as some kind of ‘prior belief’, and in some cases, such as quadratic penalties, there is even a one-to-one connection to Bayesian methods. So we do not want to promote penalization as the one and only approach, but definitely as a very useful one.

Acknowledgements

We thank Laura Hartmann and Armin Spiller from the Department of Agricultural Economics and Rural Development, Georg-August-Universität Göttingen for providing the data on households’ spending for food and related information. Furthermore, we thank Wolfgang Pössnecker and Moritz Berger for contributing to the data analysis given in Sections 4 and 6 and Catherine Lao for proofreading and language editing.

References

Agresti A (1992) Analysis of ordinal paired comparison data. *Applied Statistics*, **41**, 287–97.

Agresti A (2010) *Analysis of Ordinal Categorical Data*, 2nd edn. Hoboken, NJ: Wiley.

Agresti A, Caffo B and Ohman-Strickland P (2004) Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies.

Computational Statistics & Data Analysis, **47**, 639–53.

Aitchison J and Aitken CGG (1976) Multivariate binary discrimination by the kernel method. *Biometrika*, **63**, 413–20.

Aitkin M (1999) A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, **55**, 117–28.

Argyriou A, Evgeniou T and Pontil M (2007) Multi-task feature learning. In Schölkopf B, Platt J and Hoffman T, eds. *Advances in Neural Information Processing Systems, 19*, pages 41–48. Cambridge, MA: MIT Press.

Bach F (2008) Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, **9**, 1179–225.

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

Bondell HD and Reich BJ (2009) Simultaneous factor selection and collapsing levels in anova. *Biometrics*, **65**, 169–77.

Bradley RA (1976) Science, statistics, and paired comparisons. *Biometrics*, **32**, 213–39.

Breiman L, Friedman JH, Olshen RA and Stone JC (1984) *Classification and Regression Trees*. Monterey, CA: Wadsworth.

Bühlmann P, Gertheiss J, Hieke S, Kneib T, Ma S, Schumacher M, Tutz G, Wang C-Y, Wang Z and Ziegler A (2014) Discussin of 'the evolution of boosting algorithms' and 'extending statistical boosting'. *Methods of Information in Medicine*, **53**, 436–45.

Bühlmann P and Hothorn T (2007) Boosting algorithms: Regularization, prediction and model fitting (with discussion). *Statistical Science*, **22**, 477–505.

Bühlmann P and van de Geer S (2011) *Statistics for High-dimensional Data: Methods, Theory and Applications*. Berlin/Heidelberg: Springer.

Bühlmann P and Yu B (2003) Boosting with the L2 loss: Regression and classification. *Journal of the American Statistical Association*, **98**, 324–39.

Bürgin R and Ritschard G (2015) Tree-based varying coefficient regression for longitudinal ordinal responses. *Computational Statistics & Data Analysis*, **86**, 65–80.

Bush CA and MacEachern SN (1996) A semi-parametric bayesian model for randomised block designs. *Biometrika*, **83**, 275–85.

Cattelan M (2012) Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science*, **27**, 412–33.

Chen J and Li H (2013) Variable selection for sparse dirichlet-multinomial regression with an application to microbiome data analysis. *The Annals of Applied Statistics*, **7**, 418–42.

Chiquet J, Grandvalet Y, and Charbonnier C (2013) Sparsity with sign-coherent groups of variables via the cooperative-lasso. *The Annals of Applied Statistics*, **6**, 795–830.

Chiquet J, Gutierrez P and Rigaill G (2015) Fast tree inference with weighted fusion penalties. *Journal of Computational and Graphical Statistics*, accepted for publication.

Cieza A, Oberhauser C, Bickenbach J, Chatterji S and Stucki G (2014) Towards a minimal generic set of domains of functioning and health. *BMC Public Health*, **14**, 218.

Claeskens G (2004) Restricted likelihood ratio lack-of-fit tests using mixed spline models. *Journal of the Royal Statistical Society B*, **66**, 909–26.

Crainiceanu CM and Ruppert D (2004) Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society B*, **66**, 165–85.

Crainiceanu CM, Ruppert D, Claeskens G and Wand MP (2005) Exact likelihood ratio tests for penalised splines. *Biometrika*, **92**, 91–103.

Croissant Y (2006) *Ecdat: Data Sets for Econometrics*. R package version 0.1–5.

David HA (1988) *The Method of Paired Comparisons*. Griffin's Statistical Monographs and Courses 41, 2nd edn. London: Griffin.

Eilers PHC and Marx BD (1996) Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89–121.

Fahrmeir L and Tutz G (1994) Dynamic stochastic models for time-dependent ordered paired comparison systems. *Journal of the American Statistical Association*, **89**, 1438–49.

Fahrmeir L and Tutz G (1997) *Multivariate Statistical Modelling based on Generalized Linear Models*. New York: Springer.

Fan J and Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348–60.

Fan J, Yao Q and Cai Z (2003) Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society, Series B*, **65**, 57–80.

Ferguson TS (1973) A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, **1**, 209–30.

Follmann DA and Lambert D (1989) Generalizing logistic regression by non-parametric mixing. *Journal of the American Statistical Association*, **84**, 295–300.

Friedman JH (2001) Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, **29**, 1189–232.

Friedman JH, Hastie T and Tibshirani R (2000) Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, **28**, 337–407.

Friedman J, Hastie T and Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1–22.

Fruehwirth-Schnatter S (2006) *Finite Mixture and Markov Switching Models*. New York: Springer.

George EI and McCulloch RE (1993) Variable selection via gibbs sampling. *Journal of the American Statistical Association*, **88**, 881–9.

Gertheiss J (2014) Anova for factors with ordered levels. *Journal of Agricultural, Biological, and Environmental Statistics*, **19**, 258–77.

Gertheiss J (2015) *ordPens: Selection and/or Smoothing of Ordinal Predictors*. R package version 0.3–1.

Gertheiss J, Hogger S, Oberhauser C and 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 (Applied Statistics)*, **60**, 377–96.

Gertheiss J and Oehrlein F (2011) Testing relevance and linearity of ordinal predictors. *Electronic Journal of Statistics*, **5**, 1935–59.

Gertheiss J and Tutz G (2009) Penalized regression with ordinal predictors. *International Statistical Review*, **77**, 345–65.

Gertheiss J and Tutz G (2010) Sparse modeling of categorical explanatory variables. *Annals of Applied Statistics*, **4**, 2150–80.

Gertheiss J and Tutz G (2012) Regularization and model selection with categorical effect modifiers. *Statistica Sinica*, **22**, 957–82.

Glickman ME and Stern HS (1998) A state-space model for national football league scores. *Journal of the American Statistical Association*, **93**, 25–35.

Greene W (2003) *Econometric Analysis*. Harlow: Pearson Education Limited.

Greven S, Crainiceanu C, Küchenhoff H and Peters A (2008) Restricted likelihood ratio testing for zero variance components in linear mixed models. *Journal of Computational and Graphical Statistics*, **17**, 870–91.

Grün B and Leisch F (2008) Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. *Journal of Classification*, **25**, 225–47.

Hartmann LH (2015) *Consumption Motives in Luxury Marketing—An Analysis of Two Agricultural Markets: Equestrian Sports and Food*. Göttingen: Cuvillier.

Hartmann LH, Nitzko S and Spiller A (2016a) Segmentation of German consumers based on perceived Dimensions of Luxury Food. Journal of Food Products Marketing, accepted for publication.

Hartmann LH, Nitzko S and Spiller A (2016b) The significance of definitional dimensions of luxury food. British Food Journal, accepted for publication

Hastie T and Tibshirani R (1990) *Generalized Additive Models*. London: Chapman & Hall.

Hastie T and Tibshirani R (1993) Varying-coefficient models. *Journal of the Royal Statistical Society, Series B*, **55**, 757–96.

Hastie T, Tibshirani R and Friedman JH (2009) *The Elements of Statistical Learning, 2nd edn*. New York: Springer.

Heinzl F and Tutz G (2013) Clustering in linear mixed models with approximate dirichlet process mixtures using em algorithm. *Statistical Modelling*, **13**, 41–67.

Heinzl F and Tutz G (2014) Clustering in linear-mixed models with a group fused lasso penalty. *Biometrical Journal*, **56**, 44–68.

Heinzl F and Tutz G (2016) Additive mixed models with approximate Dirichlet process mixtures: The EM approach. *Statistics and Computing*, **26**, 73–92.

Hjort NL, Holmes C, Müller P and Walker SG (2010) *Bayesian Nonparametrics*, Vol. 28. Cambridge: Cambridge University Press.

Hoerl AE and Kennard RW (1970) Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Hofner B, Hothorn T, Kneib T and Schmid M (2011a) A framework for unbiased model selection based on boosting. *Journal of Computational and Graphical Statistics*, **20**, 956–71.

Hofner B, Mueller J and Hothorn T (2011b) Monotonicity-constrained species distribution models. *Ecology*, **92**, 1895–901.

Hoover DR, Rice JA, Wu CO and Yang L-P (1998) Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, **85**, 809–22.

Huang J, Breheny P and Ma S (2012) A selective review of group selection in high-dimensional models. *Statistical Science*, **27**, 481–99.

Jalali A, Ravikumar P and Sanghavi S (2013) A dirty model for multiple sparse regression. *IEEE Transactions on Information Theory*, **59**, 7947–68.

Jiang J (2007) *Linear and Generalized Linear Mixed Models and their Applications*. New York: Springer.

Kauermann G and Tutz G (2000) Local likelihood estimation in varying coefficient models including additive bias correction. *Journal of Nonparametric Statistics*, **12**, 343–71.

Klopp O and Pensky M (2015) Sparse high-dimensional varying coefficient model: Nonasymptotic minimax study. *The Annals of Statistics*, **43**, 1273–99.

Krishnapuram B, Carin L, Figueiredo M and Hartemink A (2005) Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**, 957–68.

Leisch F and Grün B (2012) *flexmix: Flexible Mixture Modeling*. R package.

Leissner J, Coenen M, Froehlich S, Loyola D and Cieza A (2014) What explains health in persons with visual impairment? *Health and Quality of Life Outcomes*, **12**, 65.

Leng C (2009) A simple approach for varying-coefficient model selection. *Journal of Statistical Planning and Inference*, **139**, 2138–46.

Li Q, Ouyang D and Racine JS (2013) Categorical semiparametric varying-coefficient models. *Journal of Applied Econometrics*, **28**, 551–79.

Litière S, Alonso A and Molenberghs G (2007) Type I and type II error under random-effects misspecification in generalized linear mixed models. *Biometrics*, **63**, 1038–44.

Liu J, Li R and Wu R (2014) Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *Journal of the American Statistical Association*, **109**, 266–74.

Lu Y, Zhang R and Zhu L (2008) Penalized spline estimation for varying-coefficient models. *Communications in Statistics—Theory and Methods*, **37**, 2249–61.

Ma S and Racine JS (2013) Additive regression splines with irrelevant categorical and continuous regressors. *Statistica Sinica*, **23**, 515–41.

Ma S, Racine JS and Yang L (2015) Spline regression in the presence of categorical predictors. *Journal of Applied Econometrics*, **30**, 705–17.

Masarotto G and Varin C (2012) The ranking lasso and its application to sport tournaments. *The Annals of Applied Statistics*, **6**, 1949–70.

Mauerer I, Pössnecker W, Thurner P and Tutz G (2015) Modeling electoral choices in multiparty systems with high-dimensional data: A regularized selection of parameters using the lasso approach. *Journal of Choice Modelling*, **16**, 23–42

Mayr A, Binder H, Gefeller O and Schmid M (2014a) The evolution of boosting algorithms: From machine learning to statistical modelling. *Methods of Information in Medicine*, **53**, 419–27.

Mayr A, Binder H, Gefeller O and Schmid M (2014b) Extending statistical boosting: An overview of recent methodological developments. *Methods of Information in Medicine*, **53**, 428–35.

McCullagh P and Nelder JA (1989) *Generalized Linear Models*, 2nd edn. New York: Chapman & Hall.

McFadden D (1973) Conditional logit analysis of qualitative choice behavior. In Zarembka, P, ed. *Frontiers in Econometrics*, pages 105–142, 16, 23–42. New York: Academic Press.

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

Moore WL and Lehmann DR (1989) A paired comparison nested logit model of individual preference structures. *Journal of Marketing Research*, **26**, 420–28.

Müller P and Rosner GL (1997) A Bayesian population model with hierarchical mixture priors applied to blood count data. *Journal of the American Statistical Association*, **92**, 1279–92.

Nardi Y and Rinaldo A (2008) On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, **2**, 605–33.

Negahban SN and Wainwright MJ (2011) Simultaneous support recovery in high dimensions: Benefits and perils of block l_1/l_∞-regularization. *IEEE Transactions on Information Theory*, **57**, 3841–63.

Oberhauser C, Escorpizo R, Boonen A, Stucki G and Cieza A (2013) Statistical validation of the brief international classification of functioning, disability and health core set for osteoarthritis based on a large international sample of patients with osteoarthritis. *Arthritis Care & Research*, **65**, 177–86.

Oelker M-R, Gertheiss J and Tutz G (2014) Regularization and model selection with categorical predictors and effect modifiers in generalized linear models. *Statistical Modelling*, **14**, 157–77.

Oelker M-R, Pößnecker W and Tutz G (2015) Selection and fusion of categorical predictors with L_0-type penalties. *Statistical Modelling*, **15**, 389–410.

Ollier E and Viallon V (2015) Regression modeling on stratified data: Automatic and covariate-specific selection of the reference stratum with simple l_1-norm penalties. Retrieved 25 April 2016, from http://arxiv.org/abs/1508.05476.

Pauger D and Wagner H (2014) Bayesian effect fusion for categorical and ordinal predictors. In Kneib T, Sobotka F, Fahrenholz J and Irmer H, eds. *Proceedings of the 29th International Workshop on Statistical Modelling*, pages 261–267. Göttingen.

Peng B, Ren Z and Zhang X (2015) Shrinkage estimation of categorical semiparametric varying-coefficient models.

Peterson B and Harrell FE (1990) Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **39**, 205–17.

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

Pössnecker W and Tutz G (2016) A General Framework for the Selection of Effect Type in Ordinal Regression. *Technical Report 186*, Department of Statistics, LMU, Munich.

Post J and Bondell H (2013) Factor selection and structural identification in the interaction ANOVA model. *Biometrics*, **69**, 70–9.

Pötscher BM and Schneider U (2009) On the distribution of the adaptive lasso estimator. *Journal of Statistical Planning and Inference*, **139**, 2775–90.

Quinlan JR (1993) *Programs for Machine Learning*. San Francisco: Morgan Kaufmann Publisher Inc.

Racine JS and Li Q (2004) Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, **119**, 99–130.

Ruppert D, Wand MP and Carroll RJ (2003) *Semiparametric Regression*. Cambridge: Cambridge University Press.

Schauberger G and Tutz G (2015) Modelling heterogeneity in paired comparison data—An l1 penalty approach with an application to party preference data. *Technical Report 183*, Department of Statistics, LMU, Munich.

Scheipl F and Bolker B (2013) *RLRsim: Exact (Restricted) Likelihood Ratio Tests for Mixed and Additive Models*. R package version 2.0–12.

Scheipl F, Fahrmeir L and Kneib T (2012) Spike-and-slab priors for function selection in structured additive regression models. *Journal of the American Statistical Association*, **107**, 1518–32.

Scheipl F, Greven S and Küchenhoff H (2008) Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics & Data Analysis*, **52**, 3283–99.

Sethuraman J (1994) A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639–50.

Simon N, Friedman J, Hastie T and Tibshirani R (2013) A sparse-group lasso. *Journal of Computational and Graphical Statistics*, **22**, 231–45.

Simonoff J and Tutz G (1999) Smoothing methods for discrete data. In Schimek M, ed. *Smoothing and Regression. Approaches, Computation and Application*, pages 193–228. Wiley.

Simonoff JS (1996) *Smoothing Methods in Statistics*. New York: Springer.

Sweeney E, Crainiceanu C and Gertheiss J (2016) Testing differentially expressed genes in dose-response studies and with ordinal phenotypes. *Statistical Applications in Genetics and Molecular Biology*, accepted for publication.

Tibshirani R (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267–88.

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

Townsend Z, Buckley J, Harada M and Scott MA (2013) The choice between fixed and random effects. In Marc BDM, Scott A and Simonoff JS, eds. *The SAGE Handbook of Multilevel Modeling*. London: SAGE.

Turlach B, Venables W and Wright S (2005) Simultaneous variable selection. *Technometrics*, **47**, 349–63.

Turner H and Firth D (2012) Bradley-Terry models in R: The BradleyTerry2 package. *Journal of Statistical Software*, **48**, 1–21.

Tutz G (1986) Bradley-Terry-Luce models with an ordered response. *Journal of Mathematical Psychology*, **30**, 306–16.

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

Tutz G and Berger M (2014) Tree-structured modelling of categorical predictors in regression. *Technical Report 169*, Department of Statistics, LMU, Munich.

Tutz G and Binder H (2006) Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics*, **62**, 961–71.

Tutz G and Gertheiss J (2014) Rating scales as predictors – The old question of scale level and some answers. *Psychometrika*, **79**, 357–76.

Tutz G and Oelker M (2016) Modeling clustered heterogeneity: Fixed effects, random effects and mixtures. *International Statistical Review*, accepted for publication.

Tutz G, Pössnecker W and Uhlmann L (2015) Variable selection in general multinomial logit models. *Computational Statistics and Data Analysis*, **82**, 207–22.

Tutz G and Schauberger G (2015) Extended ordered paired comparison models with application to football data from German bundesliga. *Advances in Statistical Analysis*, **99**, 209–27.

Tutz G and Schmid M (2016) *Modelling Discrete Time-to-event Data*. New York: Springer.

Vincent M and Hansen N (2014) Sparse group lasso and high dimensional multinomial classification. *Computational Statistics & Data Analysis*, **71**, 771–86.

Wang H and Leng C (2008) A note on adaptive group lasso. *Computational Statistics & Data Analysis*, **52**, 5277–86.

Wang H and Xia Y (2009) Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association*, **104**, 747–57.

Wang L, Chen G and Li H (2007) Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics*, **23**, 1486–94.

Wang L, Li H and Huang JZ (2008) Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association*, **103**, 1556–68.

Wei F and Huang J (2010) Consistent group selection in high-dimensional linear regression. *Bernoulli*, **16**, 1369.

Wood SN (2006) *Generalized Additive Models: An Introduction with R*. London: Chapman and Hall/CRC.

Wood SN (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, **73**, 3–36.

Yellott JI (1977) The relationship between Luce's choice axiom, Thurstone's theory of comparative judgement, and the double exponential distribution. *Journal of Mathematical Psychology*, **15**, 109–44.

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

Zahid FM and Tutz G (2013) Multinomial logit models with implicit variable selection. *Advances in Data Analysis and Classification*, **7**, 393–416.

Zhang C-H (2010) Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, **38**, 894–942.

Zhao P, Rocha G and Yu B (2009) The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, **37**, 3468–97.

Zhao W, Zhang R and Liu J (2014) Regularization and model selection for quantile varying coefficient model with categorical effect modifiers. *Computational Statistics & Data Analysis*, **79**, 44–62.

Zou H (2006) The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418–29.