Jan Gertheiss, Sara Hogger, Cornelia Oberhauser
& Gerhard Tutz

Selection of Ordinally Scaled Independent Variables

Technical Report Number 062, 2009
Department of Statistics
University of Munich

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

Selection of Ordinally Scaled Independent Variables

Jan Gertheiss[*], [†] Sara Hogger[†], Cornelia Oberhauser[‡] & Gerhard Tutz[†]

July 16, 2009

Abstract

Ordinal categorial variables are a common case in regression modeling. Although the case of ordinal response variables has been well investigated, less work has been done concerning ordinal predictors. This article deals with the selection of ordinally scaled independent variables in the classical linear model, where the ordinal structure is taken into account by use of a difference penalty on adjacent dummy coefficients. It is shown how the Group Lasso can be used for the selection of ordinal predictors, and an alternative blockwise Boosting procedure is proposed. Emphasis is placed on the application of the presented methods to the (Comprehensive) ICF Core Set for chronic widespread pain.

Keywords: Boosting, ICF Core Sets, Lasso, Ordinal Predictors, Ridge, Variable Selection

1 Introduction

Categorial variables which have more than two categories are often measured on ordinal scale level, so that the events described by the category numbers or class labels, lets say $0, \dots, K$, can be considered as ordered but not as equally-spaced. The case of ordinal response variables has been well investigated. Starting with McCullagh's (1980) seminal paper various modeling approaches have been suggested, see for example Armstrong and Sloan (1989), Peterson and Harrell (1990), Cox (1995) for frequentist approaches, or Albert and Chib (2001) for a Bayesian

[*] To whom correspondence should be addressed: jan.gertheiss@stat.uni-muenchen.de.

[†] Department of Statistics, Ludwig-Maximilians-Universität Munich, Germany.

[‡] Institute for Health and Rehabilitation Sciences, Ludwig-Maximilians-Universität Munich, Germany.

modeling approach. A more recent overview on ordered categorical response models has been given by Liu and Agresti (2005).

Less work has been done concerning ordinal predictors, although ordinal independent variables are often found in regression modeling. In this article a subset of the ICF - the *International Classification of Functioning, Disability and Health* (WHO, 2001) is considered as a set of potential regressors in a standard regression model. The ICF consists of about 1400 ordinally scaled factors, also called *categories*, which should not be confused with the categories of a categorial variable. In WHO-terminology *category* denotes the whole factor. For example, category *"walking"* from the component *"activities and participation"* has levels 0 (no difficulty), 1 (mild difficulty), ... , 4 (complete difficulty), and environmental factor *"social norms, practices and ideologies"* is coded by -4 (complete barrier), ... , 0 (no barrier/facilitator), ... , +4 (complete facilitator).

If ordinal variables serve as predictors in regression models it is often seen that factor labels are directly treated as metric covariates, or scores are assigned. Alternatively, simple dummy coding is used as for unordered factors, possibly with monotonicity constraints. The latter is also known under the name *isotonic regression*, see Barlow (1978) for an overview, or Bacchetti (1989) for generalizations to non-normal outcomes. We will focus on the *selection* of ordinal predictors while incorporating the ordinal scale level via a difference penalty approach.

For the representation of ordinal predictors x_j we use the well known dummy coding. That means, with $K_j + 1$ denoting the number of factor levels of x_j, for each x_j we have dummy variables $x_{j0}, \ldots, x_{jK_j}$; i.e.

$$x_{jk} = \begin{cases} 1 & x_j = k, \\ 0 & \text{otherwise.} \end{cases}$$

Given a normal response y, we assume the classical linear model

$$y = \alpha + \sum_{j=1}^{p} \sum_{k=0}^{K_j} \beta_{jk} x_{jk} + \epsilon,$$

with $\epsilon \sim N(0, \sigma^2)$. For means of identifiability, we specify reference category $k = 0$, so that $\beta_{j0} = 0$ for all j.

In matrix notation, $y = (y_1, \ldots, y_n)^T$ denotes the vector of response values, $X = (1|X_1|\ldots|X_p)$ is the design matrix with X_j containing observed (non-redundant) dummies $x_{j1}, \ldots, x_{jK_j}$. With $\beta = (\alpha, \beta_1^T, \ldots, \beta_p^T)^T$ and $\beta_j = (\beta_{j1}, \ldots \ldots, \beta_{jK_j})^T$, as usual, the model is

$$y = X\beta + \epsilon.$$

Variable selection now refers to the selection of whole groups of dummy variables $x_{j1}, \ldots, x_{jK_j}$, or to groupwise exclusion. The latter means that coefficient subvectors β_j are set to zero. Such groupwise selection/exclusion is for example performed by the so-called *Group Lasso* (Yuan and Lin, 2006) and *blockwise Boosting* (Luan and Li, 2008; Tutz and Gertheiss, 2009), which however do not take the ordinal scale level into account. Therefore we present new versions of Group Lasso and blockwise Boosting which are especially suited for ordinal predictors. The paper is organized as follows: In Section 2 and 3 the modified Group Lasso and blockwise Boosting approaches are introduced. Then it is shown what can be done when the focus is on the identification of relevant differences between two adjacent dummy coefficients. In Section 5 the different methods are applied to the (Comprehensive) ICF Core Set for chronic widespread pain. For all computations we used the statistical program R (R Development Core Team, 2009).

2 The Group Lasso

The Group Lasso (Yuan and Lin, 2006) is a modification of the original Lasso (Tibshirani, 1996) which is designed for the selection of grouped variables, as dummy coded factors. One uses a Lasso penalty at the factor level, and a Ridge type penalty within groups of (dummy) coefficients, i.e. the estimated coefficient vector is

$$\hat{\beta}^* = \mathrm{argmin}_\beta\{Q_p(\beta)\}, \tag{1}$$

for the penalized least squares criterion

$$Q_p(\beta) = (y - X\beta)^T(y - X\beta) + \lambda J(\beta), \tag{2}$$

with penalty

$$J(\beta) = \sum_{j=1}^{p} \sqrt{\beta_j^T \Omega_j \beta_j}. \tag{3}$$

Via the L_1-penalty imposed by the square root, the Group Lasso encourages sparsity at the factor level; see Yuan and Lin (2006) for details. Typically, the identity matrix is used for the penalty matrices Ω_j, possibly multiplied by a factor (Yuan and Lin, 2006; Similä and Tikka, 2007; Xie et al., 2008; Wang and Leng, 2008; Meier et al., 2008).

2.1 Specifications for Ordinal Predictors

The identity matrix, which has been used for the Group Lasso so far, is applicable to categorial predictors in general. Ordinal covariates, however, provide more information than nominal ones, since the labels' ordering is meaningful. In Gertheiss and Tutz (2009a) a difference penalty for ordinal predictors is proposed. Since categories of covariates are ordered, the response is assumed to change slowly between two adjacent categories of predictor x_j (if all other predictors are held constant). In other words, we try to avoid high jumps and prefer smoother coefficient sub-vectors β_j. So differences between coefficients of adjacent levels are penalized. The corresponding penalty is $J(\beta) = \sum_{j=1}^{p} J_j(\beta_j)$, with

$$J_j(\beta_j) \propto \sqrt{\sum_{k=1}^{K_j} (\beta_{jk} - \beta_{j,k-1})^2},$$

and $\beta_{j0} = 0$ for all j. That means for Ω_j in (3) we choose

$$\Omega_j = K_j(U_j^T U_j), \tag{4}$$

with $K_j \times K_j$ matrix

$$U_j = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 1 & \cdots & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & \cdots & -1 & 1 \end{pmatrix}. \tag{5}$$

The factor K_j ensures that the penalty is of the same order as the number of (free) parameters of β_j. The analogue scaling was also used in Yuan and Lin (2006) and Meier et al. (2008).

2.2 Computation of Estimates

For the computation of Group Lasso estimates there is the R add-on package `grplasso` (Meier et al., 2008) available, which however only allows for (scaled) identity matrices as penalty. For the incorporation of within-group difference penalties, some data transformations are necessary. We use the transformed criterion

$$Q_p(\beta) = (y - X\beta)^T(y - X\beta) + \lambda J(\beta) = (y - \widetilde{X}\widetilde{\beta})^T(y - \widetilde{X}\widetilde{\beta}) + \lambda \widetilde{J}(\widetilde{\beta}) = \widetilde{Q}_p(\widetilde{\beta}),$$

with $\widetilde{X} = (1|\widetilde{X}_1|\ldots|\widetilde{X}_p)$, $\widetilde{\beta} = (\alpha, \widetilde{\beta}_1^T, \ldots, \widetilde{\beta}_p^T)^T$, and

$$\widetilde{X}_j = X_j U_j^{-1},\ \widetilde{\beta}_j = U_j \beta_j,$$

$$\widetilde{J}(\widetilde{\beta}) = \sum_{j=1}^{p} \sqrt{\widetilde{\beta}_j^T I_j \widetilde{\beta}_j},\ I_j = K_j I.$$

Sub-matrix X_j contains dummies corresponding to factor x_j; transformation matrix U_j is taken from (5). Simple matrix multiplication shows that the inverse is given by

$$U_j^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ 1 & \cdots & \cdots & 1 \end{pmatrix}.$$

By multiplying X_j and U_j^{-1} split-coding (Walter et al., 1987) of predictor x_j is obtained. Split-coding means that dummies $\widetilde{x}_{jk}$ are defined by splits at categories $k = 1, \ldots, K_j$, i.e.

$$\widetilde{x}_{jk} = \begin{cases} 1 & \text{if } x_j \geq k, \\ 0 & \text{otherwise.} \end{cases}$$

Now the model is parameterized by coefficients $\widetilde{\beta}_{jk} = \beta_{jk} - \beta_{j,k-1}$, $k = 1, \ldots, K_j$. Thus transitions between category k and $k-1$ of predictor x_j are expressed by coefficient $\widetilde{\beta}_{jk}$. Original dummy coefficients are obtained by back-transformation $\beta_{jk} = \sum_{s=1}^{k} \widetilde{\beta}_{js}$, resp. $\beta_j = U_j^{-1}\widetilde{\beta}_j$. Note that for correct computation of smoothed estimates, in the R `grplasso` function argument `standardize = F` must be chosen.

3 Blockwise Boosting

L_1-penalization as used so far is not the only possibility to obtain variable selection. In order to check reliability and performance of the proposed Group Lasso modification, and for having alternatives available, we develop an alternative selection procedure which is able to specifically select ordinally scaled explanatory variables, too. As before, the ordinal scale level is taken into account via the presented difference penalty. For variable selection, however, Boosting techniques are applied.

3.1 Basic Concept

The Boosting concept has been developed in the machine-learning community with a focus on classification problems (Schapire, 1990; or Freund and Schapire, 1996). More recently, based on work by Breiman (1998) or Breiman (1999), it has been extended to regression problems in articles by Friedman et al. (2000), Bühlmann and Yu (2003), Bühlmann (2006). In the following, consideration is restricted to a version of so-called L_2-Boosting which is essentially repeated least squares fitting of residuals. In common *componentwise* L_2-Boosting in every iteration just one regression coefficient is updated. Therefore, in addition to the fitting of the coefficient, a selection step is included which selects the predictor that produces minimum quadratic loss based on the current residuals. Since all coefficients start with value zero, predictors that are never selected are implicitly excluded from the model. See for example Bühlmann (2006) for details.

Luan and Li (2008) and Tutz and Gertheiss (2009) proposed to simultaneously update whole groups - or *blocks* - of coefficients, instead of single ones. Therefore the procedure is called *blockwise* Boosting in the following. In the paper by Luan and Li (2008) groups are defined by genes which are linked for biological reasons, whereas in Tutz and Gertheiss (2009) blocks of adjacent measurement points in signal regression are considered. In the regression problems considered in this article groups are naturally defined by groups of dummy coefficients belonging to the same categorial predictor. To incorporate the ordinal structure, in every Boosting iteration actual residuals are regressed on each of the corresponding groups in a penalized way, while employing penalty matrices Ω_j from (4). But only the block producing minimum loss is updated. In summary, for fixed tuning parameter λ, the proposed boosting algorithm is as follows.

BlockBoost

Step 1 (Initialization)

For $j = 1, \dots, p$ fit a linear model to data (y, X_j) by using generalized Ridge regression. From the resulting estimates $\hat{b}_j = (X_j^T X_j + \lambda\Omega_j)^{-1} X_j^T y$, select the best by minimizing the residual sum of squares, i.e. $\hat{j}_0 = \arg\min_{1 \le j \le p} \|y - X_j \hat{b}_j\|^2$. Let $\hat{\beta}^{(0)} = (\hat{\beta}_1^{(0)}, \dots, \hat{\beta}_p^{(0)})^T$ be defined by components:

$$\hat{\beta}_j^{(0)} = \begin{cases} \hat{b}_j & j = \hat{j}_0 \\ 0 & \text{otherwise} \end{cases}$$

Step 2 (Residual Fit)

For $r = 1, 2, \dots, M$ compute residuals $u_i = y_i - x_i^T \hat{\beta}^{(r-1)}, i = 1, \dots, n$ and fit for $j = 1, \dots, p$ a linear model to data (u, X_j) where $u^T = (u_1, \dots, u_n)$. From the resulting Ridge type estimates $\hat{b}_j = (X_j^T X_j + \lambda\Omega_j)^{-1} X_j^T u$, choose $\hat{j}_r$ such that the residual sum of squares is minimized, i.e.

$$\hat{j}_r = \arg\min_{1 \le j \le p} \|u - X_j \hat{b}_j\|^2.$$

Let $\hat{\beta}^{(r)}$ be defined by components:

$$\hat{\beta}_j^{(r)} = \begin{cases} \hat{\beta}_j^{(r-1)} + \hat{b}_j & j = \hat{j}_r \\ \hat{\beta}_j^{(r-1)} & \text{otherwise} \end{cases}$$

If every predictor is just binary we have $K_j = 1$ for all j, and common componentwise Boosting is obtained. A common approach for selecting penalty parameter λ and the number of Boosting iterations M is K-fold cross validation. Alternatively, Boosting iterations may be stopped via a corrected version of the AIC (Hurvich et al., 1998) as proposed by Bühlmann (2006). This criterion is based on the boosting hat matrix, which maps the response vector y into the space of fitted values. With convention "$\prod_{m=1}^{0}(\cdot) = I$", in the rth iteration the boosting hat matrix B_r is defined by (see Bühlmann and Yu, 2003)

$$B_r = \sum_{l=0}^{r} H_{\hat{j}_l} \prod_{m=1}^{l} (I - H_{\hat{j}_{l-m}}),$$

with Ridge type hat matrix $H_j = X_j(X_j^T X_j + \lambda\Omega_j)^{-1} X_j^T$ and j_l denoting the variable block / factor that has been selected in the lth boosting iteration. It is simple to show that

$$B_r = I - \prod_{m=0}^{r} (I - H_{\hat{j}_{r-m}}) = I - (I - H_{\hat{j}_r})(I - H_{\hat{j}_{r-1}}) \cdots (I - H_{\hat{j}_0}).$$

The (corrected) AIC in the rth iteration is defined by (Hurvich et al., 1998)

$$AIC_c(r) = \log\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - (B_r y)_i)^2\right) + \frac{1 + \text{trace}(B_r)/n}{1 - (\text{trace}(B_r) + 2)/n}.$$

Given an upper bound R^* for the candidate number of boosting iterations, the optimum iteration number M can be estimated by (Bühlmann, 2006) $\hat{M} = \text{argmin}_{0 \le r \le R^*} AIC_c(r)$. For selecting penalty parameter λ the AIC can be minimized, too. Of course, also the simple least squares based selection criterion applied within the Boosting algorithm may be replaced by the AIC or a cross validation criterion.

3.2 Relations to Isotonic Regression

Sometimes a monotonic relationship between one or more explanatory variables and the response can be assumed, for example in dose-response-analysis. One advantage of (blockwise) Boosting over the Group Lasso is that monotonicity constraints can be easily incorporated. Leitenstorfer and Tutz (2007), for example, dealt with monotonicity in generalized additive models based on B-splines and estimation by Boosting techniques.

Without loss of generality we assume that a nondecreasing relationship between (ordinal) predictor x_j and response y can be supposed. To ensure that corresponding restrictions on the dummy coefficients are satisfied, in every Boosting iteration optimization needs to be carried out with restrictions. In the first iteration (Step 1 in the algorithm on page 6) we have

$$\hat{b}_j = \arg\min_{b \in \mathcal{B}_j} \left\{(y - X_j b)^T (y - X_j b) + \lambda b^T \Omega_j b\right\},$$

with

$$\mathcal{B}_j = \{b : b_1 \le \ldots \le b_{K_j}\}.$$

In the following iterations (Step 2 on page 6) the restrictions become

$$\mathcal{B}_j = \left\{ b : b_2 - b_1 \geq \hat{\beta}_{j1}^{(r-1)} - \hat{\beta}_{j2}^{(r-1)}, \ldots, b_{K_j} - b_{K_j - 1} \geq \hat{\beta}_{j,K_j-1}^{(r-1)} - \hat{\beta}_{jK_j}^{(r-1)} \right\}.$$

For practical optimization quadratic programming methods can be used. In R, for example, there are the add-on packages `quadprog` (Turlach, 2007) or `kernlab` (Karatzoglou et al., 2004) available.

4 Clustering Categories

If many differences of adjacent dummy coefficients β_{jk} and $\beta_{j,k-1}$ are assumed to be zero, coefficient (sub-)vectors β_j should not be modeled as a smooth function as done so far, but as a step function. The effect is that some (adjacent) categories are clustered, since the predictor's influence on the response is modeled as piecewise constant over categories. Such clustering - or *fusion* - of class levels can be easily obtained by changing the penalty in (2), as described in detail in Gertheiss and Tutz (2009b). One employs

$$Q_p(\beta) = (y - X\beta)^T (y - X\beta) + \lambda J(\beta),$$

with penalty

$$J(\beta) = \sum_{j=1}^{p} \sum_{k=1}^{K_j} |\beta_{jk} - \beta_{j,k-1}|.$$

The used penalty is a version of so-called Variable Fusion (Land and Friedman, 1997) for ordinal predictors. Given ordered metric predictors, the same type of penalty has also been used for the Fused Lasso (Tibshirani et al., 2005). In the latter case, beside the given difference penalty, an additional L_1-penalty is put on the regression coefficients to ensure variable selection. Since in the situation considered here $\beta_{j0} = 0$ for all j by definition, selection on the factor level is implicitly included with a penalty as given above. If for a certain predictor x_j estimated coefficients $\hat{\beta}_{jk} = \hat{\beta}_{j,k-1}$ for all $k = 1, \ldots, K_j$, then x_j is excluded from the model. In general, due to Lasso (Tibshirani, 1996) typical selection characteristics, only for some j and k, $\hat{\beta}_{jk} \neq \hat{\beta}_{j,k-1}$ will hold. That means, not the whole ordinal predictor x_j is selected, but only relevant transitions between adjacent categories. Practical estimation can be done via split-coding of the ordinal predictors (see Section 2), and applying standard Lasso methodology, for example the famous `lars`-algorithm (Efron et al., 2004), also available as a R add-on package.

It is clear that after split-coding not only the Lasso but every estimation procedure with implicit variable selection may be applied in order to select jumps between levels of x_j; for example L_2-Boosting with componentwise least squares (Bühlmann, 2006). Walter et al. (1987) intended the use of classical tests for such identification of substantial *"between-strata differences"*.

5 Application to ICF Core Sets for Chronic Widespread Pain

ICF Core Sets constitute an approach to make the International Classification of Functioning, Disability and Health (ICF) (WHO, 2001) applicable in clinical practice. The ICF was officially endorsed by the Fifty-fourth World Health Assembly in 2001. Its overall aim is to provide a unified and standard language and framework for the description of functioning and disability (WHO, 2001). The ICF consists of about 1400 so-called ICF categories, each of which refers to a health or a health-related domain. ICF categories can be used by health professionals, i.e. physicians, nurses, physiotherapists, occupational therapists, etc., to document the health and functioning of patients by using an ordinal scale, ranging from no problem/limitation to complete problem/limitation. However, since the ICF is a very comprehensive classification, its application in clinical practice is a major challenge. Therefore, ICF Core Sets, which represent a selection of ICF categories relevant to persons with specific health conditions or treated in specific settings, are being developed.

ICF Core Sets are defined within the scope of international consensus conferences, in which evidence from preliminary studies is presented and serves as basis for a decision-making and consensus process. The preliminary studies include an expert survey, a systematic review, an empirical and a qualitative study (Cieza, Ewert et al., 2004). For each health condition two different Core Sets are developed: A Comprehensive and a Brief ICF Core Set. The *Comprehensive ICF Core Set* is intended to serve as standard for multi-professional assessment, while the *Brief ICF Core Set* is intended to serve as minimal standard for the assessment and reporting of functioning and health for clinical studies (Stucki and Grimby, 2004). One of the health conditions for which ICF Core Sets have been developed is chronic widespread pain (CWP). Though there is no universally accepted definition of CWP, it may be characterized by pain involving several regions of the body, which causes problems in functioning, psychological distress, poor sleep quality, or difficulties in daily life (Cieza, Stucki et al., 2004).

The Comprehensive ICF Core Set for CWP consists of 67 and the Brief ICF Core Set for CWP of 26 ICF categories; see Cieza, Stucki et al. (2004) for details. These first versions of ICF Core Sets are undergoing extensive validation from different perspectives and in different settings. Within these validation studies a multicenter, international study has been performed in which health professionals rated the level of impairment/limitation of persons with CWP in each of the ICF categories contained in the Comprehensive ICF Core Set.

In the following we deal with the Comprehensive ICF Core Set for CWP and the data collected in the multicenter, international study, including the data of the well established SF-36 questionnaire (Ware and Sherbourne, 1992), which

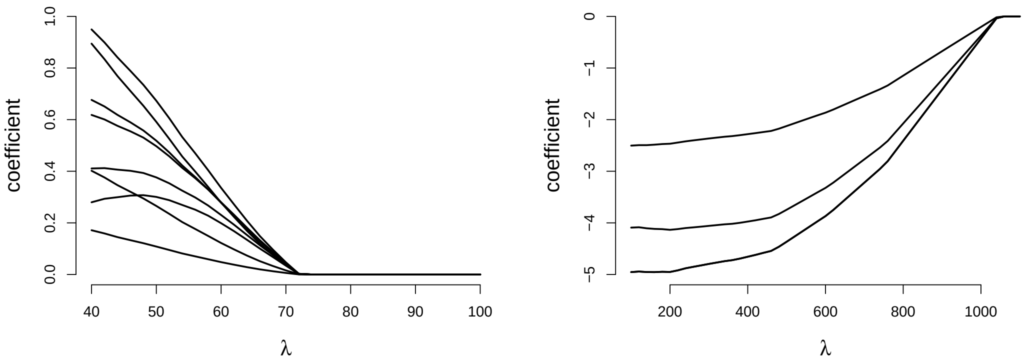

Figure 1: Paths of Group Lasso estimates of dummy coefficients as functions of penalty parameter λ; considered are environmental factor "social norms, practices and ideologies" (left) as well as ICF category "walking" (right).

is answered by the patients. Based on the SF-36 a physical health component summary (PCS) can be computed (McHorney et al., 1993); the higher this score the better the patient's subjective physical health condition.

The objective is to identify those ICF categories that contribute to the explanation of PCS. Therefore, the PCS is regressed on the Comprehensive ICF Core Set for CWP. That means we are faced with a regression problem with ordinal predictors. Moreover, it is tried to identify a set of relevant ICF categories, i.e. variable selection is intended. Ordinary least squares estimation (without variable selection) is quite unstable since "only" $n = 420$ observations are given for more than 300 dummy coefficients. A detailed summary of all categories from the Comprehensive ICF Core Set is given in the Appendix, Table 1, 2 and 3.

In the considered application it can be assumed that the predictors' influence on the response varies continuously over categories, and that there are not just a few distinct jumps. Hence, the smooth modeling using the Group Lasso or blockwise Boosting (Sections 2 and 3) is preferred over the ordinal Fused Lasso outlined in Section 4.

After some preprocessing of the data, which primarily means imputation of some missing values using R add-on package `Amelia` (Honaker et al., 2009), the Group Lasso and blockwise Boosting are applied as described before. Estimation is done without any monotonicity constraints, and selection within the Boosting algorithm is based on the AIC. Figure 1 shows some paths of Group Lasso estimates of dummy coefficients. It shows the factors "walking" (right panel) and "social norms, practices and ideologies" (left panel). With increasing penalty λ coefficients are shrunken more and more towards zero, until at a certain point the whole group of coefficients is simultaneously set to zero, which means that the corresponding factor is excluded from the model. In the case of "walking" this

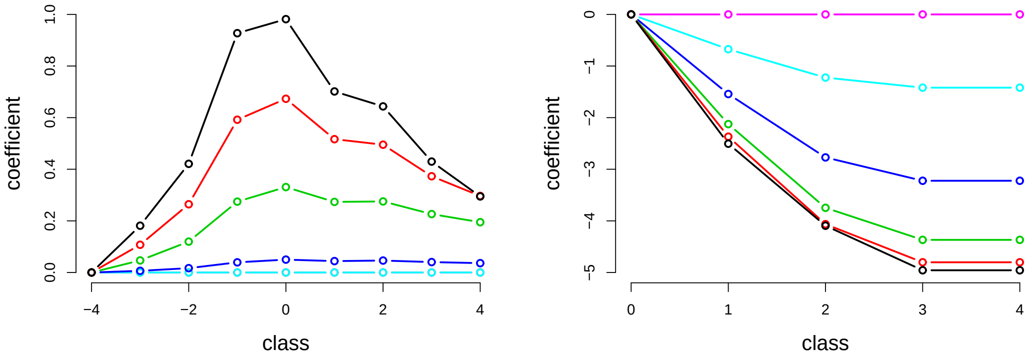

Figure 2: Group Lasso estimates of dummy coefficients as functions of class labels for different choices of λ; considered are environmental factor "social norms, practices and ideologies" ($\lambda \in \{40, 50, \ldots\}$, left) as well as ICF category "walking" ($\lambda \in \{100, 300, \ldots\}$, right).

happens much later than for "social norms,..."; i.e. even for very high penalty parameter λ, factor "walking" is still selected, whereas "social norms,..." is excluded from the model. Figure 2 offers another perspective: dummy coefficients as functions of class labels, for some distinct λ values. In addition to shrinkage and selection also smoothing behavior is clearly seen. The plot of factor "social norms, practices and ideologies" (Figure 2, left) makes clear that monotonicity constraints may be counterproductive in the considered application. Although ICF category "walking" has four free dummy coefficients, in the right panel of Figure 1 only three paths are plotted. The reason is that class 4 is not observed in the data. Nevertheless, a corresponding coefficient is fitted (see Figure 2, right), but due to the difference penalty, coefficients of class 3 and 4 are set equal, with the effect that corresponding paths cannot be distinguished.

The feature that coefficients are fitted even if corresponding levels are not observed in the data has some advantages, in particular if (K-fold) cross validation is carried out. If some classes are rarely observed (in the extreme case they are observed only once) it may happen that all these observations are found in the same fold. The penalty, however, ensures that a coefficient is always fitted, and test set prediction remains possible. Moreover, if the fitted model is to be used for prediction on future data, it may occur that classes which are not observed in the training data, may be observed in future data sets.

For selection of the adequate penalty for the Group Lasso and the adequate smoothing and number of iterations of the BlockBoost we employed 5-fold cross validation. Corresponding scores are shown in Figure 3. In both cases the optimal penalty is $\lambda = 80$. For the BlockBoost, minimization must be done over a two-dimensional grid of λ and M values. It turns out that $M = 29$ Boosting iterations seem enough. The behavior of the cross validation score for the Group Lasso

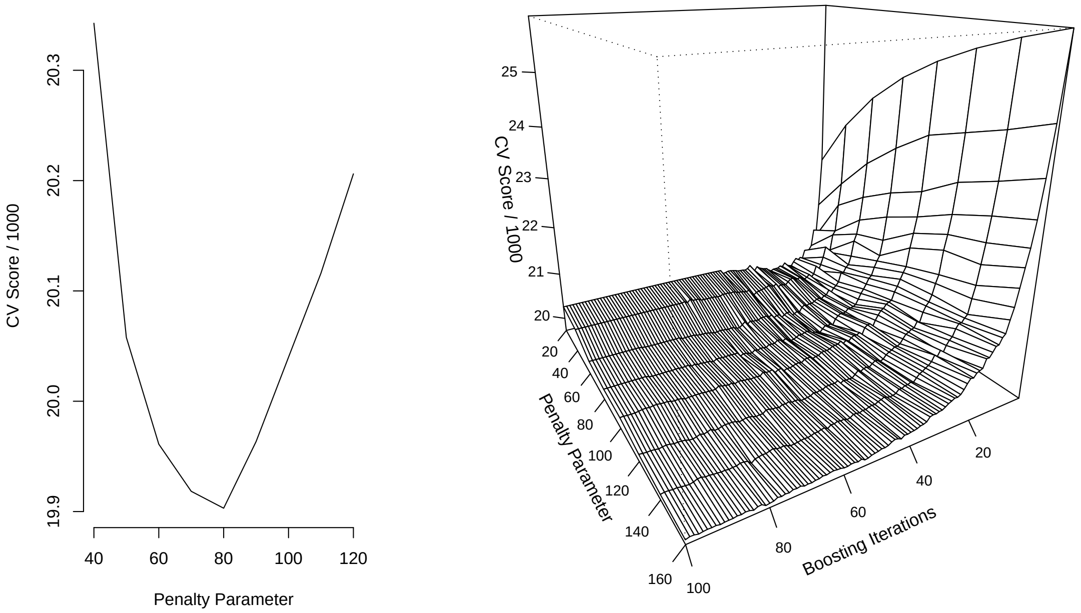

Figure 3: 5-fold cross validation scores for Group Lasso (left) and blockwise Boosting (right) as functions of penalty parameter λ and number of Boosting iterations (if applicable).

(Figure 3, left) indicates that penalization distinctly improves the ordinary least squares fit which is obtained for $\lambda = 0$.

In Figure 4 some fitted coefficients are shown for ICF categories which were selected by both blockwise Boosting and the Group Lasso. Qualitatively, results are similar. Also fitted constants are almost equal. In case of the Group Lasso $\hat{\alpha} = 36.16$ is obtained, and $\hat{\alpha} = 36.65$ for blockwise Boosting. With respect to fitted regression coefficients, predictor *d450 ("walking")* seems to have the largest effect on the response; *d450* is also the covariate which was selected first. In the case of blockwise Boosting that means it was selected in the first iteration, and with highest penalty λ if the Group Lasso is applied. This finding is not surprising, since in the SF-36 questionnaire (where the response is based on) three items are related to "walking". The shape of the coefficient curve is monotone descreasing, which means that a patient feels better if he/she has less difficulties (with walking). An interesting ICF category is *e450*: If "individual attitudes of health professionals" are seen as a barrier (negative class labels) this has almost no effect on the patient's well-being, if they are seen as a facilitator (positive class labels), however, there is a clear positive relationship. But it should be kept in mind that the questions from the ICF are answered by the doctor, i.e. a health professional. In Table 1, 2 and 3 in the Appendix it is reported in detail which ICF categories are selected by the used methods for the chosen tuning parameters

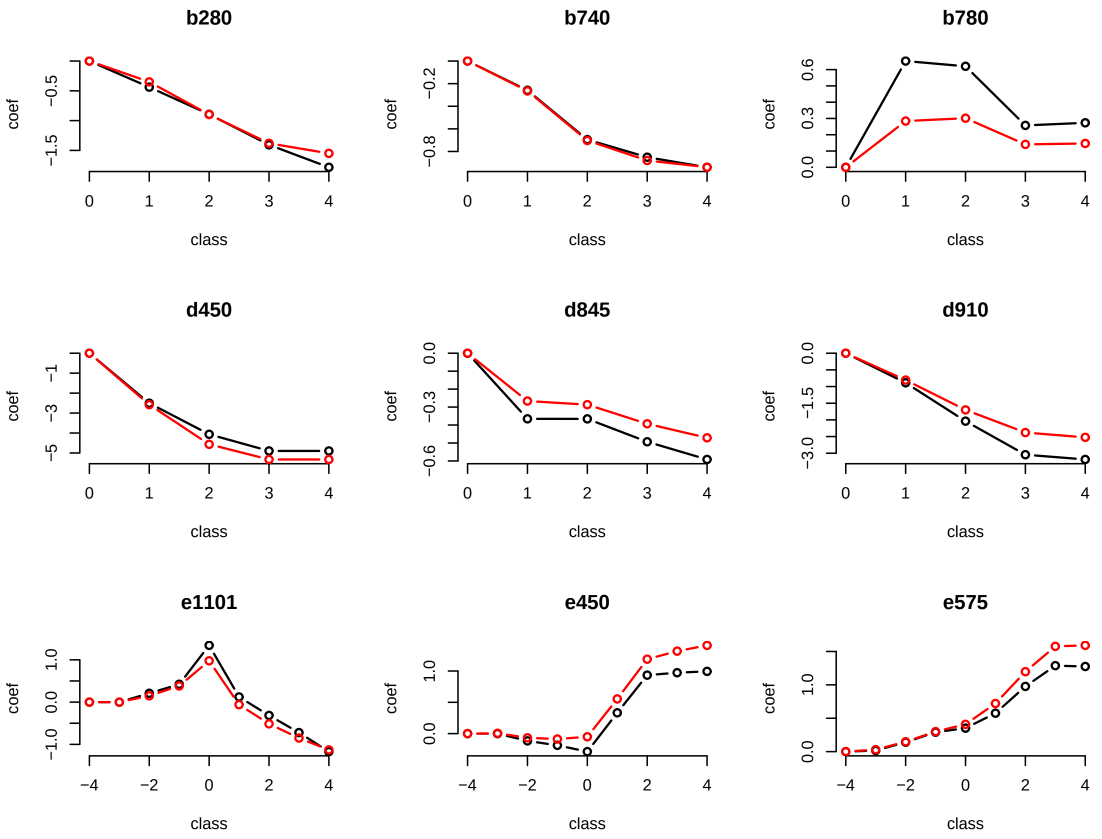

Figure 4: Coefficients of some predictors selected by the Group Lasso (black) and blockwise Boosting (red); for a description of all potential predictors see Table 1, 2 and 3.

$\lambda = 80$ and $M = 29$. The Group Lasso selects much more predictors (30) than blockwise Boosting (18). However, all ICF categories which are selected by the BlockBoost are among those chosen by the Group Lasso. That means the overlap (18) is highly significant. The p-value which results from the hypergeometric distribution is about 10^{-8}.

6 Summary and Discussion

We considered selection of ordinally scaled explanatory variables in the classical linear model. Factors were dummy coded and the ordinal structure was taken into account via a difference penalty on adjacent dummy coefficients. For selection purposes two different approaches were presented: a modification of the Group Lasso where L_1-penalization on the factor level is employed, and a blockwise Boosting procedure. In the latter case a whole block of dummy coefficients is selected and updated in each Boosting iteration - as a generalization of componentwise Boosting, which just selects single coefficients.

The proposed methods were applied to the Comprehensive ICF Core Set for chronic widespread pain, and showed a significant overlap with respect to factor selection, while blockwise Boosting resulted in a much sparser model. Fitted coefficient curves seemed quite similar, at least if factors were selected by both methods. An advantage of the Group Lasso over blockwise Boosting is that in the first case only one tuning parameter is needed, whereas blockwise Boosting requires specification of two parameters - the extent of smoothing within each iteration, and the total number of iterations. But also the Group Lasso can be used with two parameters. If the method is generalized to a *Group-Lasso-Ridge Hybrid* (Meier et al., 2008) the (first) tuning parameter λ is just used for variable selection. In a second step, the selected model is refitted using another parameter κ. When analyzing the Comprehensive ICF Core Set for CWP, however, this modification improved the fit just slightly.

Beside smooth modeling of predictor-specific coefficient vectors via quadratic penalization, it was also outlined how (ordered) class levels can be clustered using L_1-penalization. A similar technique for fusion of nominal categories is presented in Bondell and Reich (2009) and Gertheiss and Tutz (2009b).

Since there are only a few approaches available which are especially designed for both ordinal predictors and variable selection (but many applications of that type) the proposed methods have high potential for the future. For example, there are ICF Core Sets under development for many kinds of diseases, not just CWP. A possible alternative to the presented approaches within the (generalized) linear model is isotonic regression, where a monotone relationship is assumed between predictors and (transformed mean of the) response. As shown, however, the blockwise Boosting approach can be easily modified, such that certain linear restrictions are satisfied. Generalizations to non-normal outcomes are possible via e.g. likelihood-based Boosting; see for example Tutz and Binder (2006) or Tutz and Leitenstorfer (2007). Also the Group Lasso can be generalized, as shown by Meier et al. (2008).

Within nonparametric methods, regression trees (see e.g. Hastie et al., 2001; Hothorn et al., 2006) are able to exploit the additional information which is provided by the categories' ordinal structure. Since trees suffer from instability, however, Random Forests (Breiman, 2001) are often used instead. Though some measures of variable importance are provided, Forests are like "black boxes". Hence, results are hard to interpret, and common (linear) regression models are preferred by many analysts.

Acknowledgements

This work was partially supported by DFG project TU62/4-1 (AOBJ: 548166).

Appendix: Composition of the Comprehensive ICF Core Set for CWP

In the following a short summary of the Comprehensive ICF Core Set for chronic widespread pain is given, see Cieza, Stucki et al. (2004) for details. For a more detailed description of the ICF categories in general, such as information concerning inclusion and exclusion criteria for the single categories, see WHO (2001).

ICF code	ICF category title	Selected by	
		GL	BB
b122	Global psychosocial functions		
b126	Temperament and personality functions		
b130	Energy and drive functions		
b134	Sleep functions		
b140	Attention functions	✓	✓
b147	Psychomotor functions		
b152	Emotional functions	✓	
b1602	Content of thought		
b164	Higher-level cognitive functions		
b180	Experience of self and time functions		
b260	Proprioceptive function		
b265	Touch function		
b270	Sensory functions related to temperature and other stimuli		
b280	Sensation of pain	✓	✓
b430	Haematological system functions		
b455	Exercise tolerance functions	✓	✓
b640	Sexual functions		
b710	Mobility of joint functions	✓	
b730	Muscle power functions	✓	
b735	Muscle tone functions	✓	
b740	Muscle endurance functions	✓	✓
b760	Control of voluntary movement functions		
b780	Sensations related to muscles and movement functions	✓	✓
s770	Additional musculoskeletal structures related to movement		

Table 1: ICF categories of the components "body functions" and "body structures" included in the Comprehensive ICF Core Set for CWP (cf. Cieza, Stucki et al., 2004), and selection results of Group Lasso (GL) and BlockBoost (BB); possible levels are 0 (no impairment), 1 (mild impairment), . . . , 4 (complete impairment), see WHO (2001).

ICF code	ICF category title	Selected by	
		GL	BB
d160	Focusing attention	✓	
d175	Solving problems		
d220	Undertaking multiple tasks		
d230	Carrying out daily routine		
d240	Handling stress and other psychological demands		
d410	Changing basic body position	✓	✓
d415	Maintaining a body position		
d430	Lifting and carrying objects	✓	
d450	Walking	✓	✓
d455	Moving around	✓	✓
d470	Using transportation		
d475	Driving		
d510	Washing oneself		
d540	Dressing	✓	
d570	Looking after one's health		
d620	Acquisition of goods and services		
d640	Doing housework	✓	✓
d650	Caring for household objects	✓	
d660	Assisting others	✓	
d720	Complex interpersonal interactions	✓	✓
d760	Family relationships		
d770	Intimate relationships		
d845	Acquiring, keeping and terminating a job	✓	✓
d850	Remunerative employment	✓	
d855	Non-remunerative employment	✓	✓
d910	Community life	✓	✓
d920	Recreation and leisure		

Table 2: ICF categories of the component "activities and participation" included in the Comprehensive ICF Core Set for CWP (cf. Cieza, Stucki et al., 2004), and selection results of Group Lasso (GL) and BlockBoost (BB); possible levels are 0 (no difficulty), 1 (mild difficulty), . . . , 4 (complete difficulty), see WHO (2001).

ICF code	ICF category title	Selected by	
		GL	BB
e1101	Drugs	✓	✓
e310	Immediate family		
e325	Acquaintances, peers, colleagues, neighbours and community members		
e355	Health professionals	✓	
e410	Individual attitudes of immediate family members		
e420	Individual attitudes of friends	✓	
e425	Individual attitudes of acquaintances, peers, colleagues, neighbours and community members		
e430	Individual attitudes of people in positions of authority		
e450	Individual attitudes of health professionals	✓	✓
e455	Individual attitudes of health-related professionals		
e460	Societal attitudes		
e465	Social norms, practices and ideologies		
e570	Social security services, systems and policies	✓	✓
e575	General social support services, systems and policies	✓	✓
e580	Health services, systems and policies	✓	✓
e590	Labour and employment services, systems and policies		

Table 3: ICF categories of the component "environmental facors" included in the Comprehensive ICF Core Set for CWP (cf. Cieza, Stucki et al., 2004), and selection results of Group Lasso (GL) and BlockBoost (BB); possible levels are -4 (complete barrier), ... , 0 (no barrier/facilitator), ... , +4 (complete facilitator), see WHO (2001).

References

Albert, J. H. and S. Chib (2001). Sequential ordinal modeling with applications to survival data. *Biometrics 57*, 829–836.

Armstrong, B. and M. Sloan (1989). Ordinal regression models for epidemiologic data. *American Journal of Epidemiology 129*, 191–204.

Bacchetti, P. (1989). Additive isotonic models. *Journal of the American Statistical Association 84*, 289–294.

Barlow, R. E. (Ed.) (1978). *Statistical inference under order restrictions: the theory and application of isotonic regression.* Chichester: Wiley.

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

Breiman, L. (1998). Arcing classifiers. *Annals of Statistics 26*, 801–849.

Breiman, L. (1999). Prediction games and arcing algorithms. *Neural Computation 11*, 1493–1517.

Breiman, L. (2001). Random forests. *Machine Learning 45*, 5–32.

Bühlmann, P. (2006). Boosting for high-dimensional linear models. *Annals of Statistics 34*, 559–583.

Bühlmann, P. and B. Yu (2003). Boosting with the L2 loss: Regression and classification. *Journal of the American Statistical Association 98*, 324–339.

Cieza, A., T. Ewert, T. B. Üstün, S. Chatterji, N. Kostanjsek, and G. Stucki (2004). Development of ICF Core Sets for patients with chronic conditions. *Journal of Rehabilitation Medicine Suppl. 44*, 9–11.

Cieza, A., G. Stucki, M. Weigl, L. Kullmann, T. Stoll, L. Kamen, N. Kostanjsek, and N. Walsh (2004). ICF Core Sets for chronic widespread pain. *Journal of Rehabilitation Medicine Suppl. 44*, 63–68.

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

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *Annals of Statistics 32*, 407–499.

Freund, Y. and R. E. Schapire (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pp. 148–156. San Francisco, CA: Morgan Kaufmann.

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

Gertheiss, J. and G. Tutz (2009a). Penalized regression with ordinal predictors. *International Statistical Review*. (accepted for publication).

Gertheiss, J. and G. Tutz (2009b). Sparse modeling of categorial explanatory variables. Technical Report 60, Department of Statistics, Ludwig-Maximilians-Universität München. (submitted).

Hastie, T., R. Tibshirani, and J. H. Friedman (2001). *The Elements of Statistical Learning*. New York: Springer.

Honaker, J., G. King, and M. Blackwell (2009). *Amelia: Amelia II: A Program for Missing Data*. R package version 1.2-2.

Hothorn, T., K. Hornik, and A. Zeileis (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics 15*, 651–674.

Hurvich, C. M., J. S. Simonoff, and C. Tsai (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society B 60*, 271–293.

Karatzoglou, A., A. Smola, K. Hornik, and A. Zeileis (2004). kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software 11*(9), 1–20.

Land, S. R. and J. H. Friedman (1997). Variable fusion: A new adaptive signal regression method. Technical report 656, Department of Statistics, Carnegie Mellon University Pittsburg.

Leitenstorfer, F. and G. Tutz (2007). Generalized monotonic regression based on B-splines with an application to air pollution data. *Biostatistics 8*, 654–673.

Liu, Q. and A. Agresti (2005). The analysis of ordinal categorical data: An overview and a survey of recent developments. *Test 14*, 1–73.

Luan, Y. and H. Li (2008). Group additive regression models for genomic data analysis. *Biostatistics 9*, 100–113.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society B 42*, 109–142.

McHorney, C. A., J. E. Ware, and A. E. Raczek (1993). The MOS 36-item short-form health survey (SF-36): II. psychometric and clinical tests of validity in measuring physical and mental health constructs. *Medical Care 31*, 247–263.

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

Peterson, B. and F. E. Harrell (1990). Partial proportional odds models for ordinal response variables. *Applied Statistics 39*, 205–217.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning 5*, 197–227.

Similä, T. and J. Tikka (2007). Input selection and shrinkage in multiresponse linear regression. *Computational Statistics & Data Analysis 52*, 406–422.

Stucki, G. and G. Grimby (2004). Foreword: Applying the ICF in medicine. *Journal of Rehabilitation Medicine Suppl. 44*, 5–6.

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

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

Turlach, B. A. (2007). *quadprog: Functions to solve Quadratic Programming Problems.* R package version 1.4-11, S original by Berwin A. Turlach, R port by Andreas Weingessel.

Tutz, G. and H. Binder (2006). Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics 62*, 961–971.

Tutz, G. and J. Gertheiss (2009). Feature extraction in signal regression: A boosting technique for functional data regression. *Journal of Computational and Graphical Statistics*. (accepted for publication).

Tutz, G. and F. Leitenstorfer (2007). Generalized smooth monotonic regression in additive modeling. *Journal of Computational and Graphical Statistics 16*, 165–188.

Walter, S. D., A. R. Feinstein, and C. K. Wells (1987). Coding ordinal independent variables in multiple regression analysis. *American Journal of Epidemiology 125*, 319–323.

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

Ware, J. E. and C. Sherbourne (1992). The MOS 36-item short-form health survey (SF-36): I. conceptual framework and item selection. *Medical Care 30*, 473–483.

WHO (2001). *International Classification of Functioning, Disability and Health: ICF.* Geneva: World Health Organization.

Xie, B., W. Pan, and X. Shen (2008). Variable selection in penalized model-based clustering via regularization on grouped parameters. *Biometrics 64*, 921–930.

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