LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Faisal Maqbool Zahid & Gerhard Tutz

Proportional Odds Models with High-dimensional Data Structure

Technical Report Number 100, 2011
Department of Statistics
University of Munich

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

Proportional Odds Models with High-dimensional Data Structure

Faisal Maqbool Zahid[a,*], Gerhard Tutz[b]

[a]*Ludwig-Maximilians-University Munich, Ludwigstrasse 33, D-80539 Munich, Germany*

[b]*Ludwig-Maximilians-University Munich, Akademiestraße 1, D-80799 Munich, Germany*

Abstract

The proportional odds model (POM) is the most widely used model when the response has ordered categories. In the case of high-dimensional predictor structure the common maximum likelihood approach typically fails when all predictors are included. A boosting technique *pomBoost* is proposed that fits the model by implicitly selecting the influential predictors. The approach distinguishes between metric and categorical predictors. In the case of categorical predictors, where each predictor relates to a set of parameters, the objective is to select simultaneously all the associated parameters. In addition the approach distinguishes between nominal and ordinal predictors. In the case of ordinal predictors, the proposed technique uses the ordering of the ordinal predictors by penalizing the difference between the parameters of adjacent categories. The technique has also a provision to consider some mandatory predictors (if any) which must be part of the final sparse model. The performance of the proposed boosting algorithm is evaluated in a simulation study and applications with respect to mean squared error and prediction error. Hit rates and false alarm rates are used to judge the performance of *pomBoost* for selection of the relevant predictors.

Key words: Logistic regression, Proportional odds model, Variable selection, Likelihood-based boosting, Penalization, Hit rate, False alarm rate.

1. Introduction

Various regression models for ordered responses categories have been proposed, see for example McCullagh (1980) and Agresti (1999), Ananth and Kleinbaum (1997). The most widely used model is the proportional odds model (POM), also known as cumulative logit model. Although the parameterization is sparser than in the multinomial logit model, with increasing number of covariates the usual maximum likelihood approach may fail. However, in many applications, the number of covariates is much larger than the sample size. In addition the covariates may be categorical with large number of categories. If the number of parameters to be estimated is larger than the sample size, one

*Corresponding author. Tel.: ++49 89 2180 6408; fax.: ++49 89 2180 5040.

Email addresses: faisal-maqbool.zahid@stat.uni-muenchen.de (Faisal Maqbool Zahid), tutz@stat.uni-muenchen.de (Gerhard Tutz)

possible alternative to the usual likelihood approach is penalized likelihood. Ridge regression is one of the oldest penalization methods considered by Zahid (2011) to address the problems in likelihood estimation for ordinal response models with a special focus on proportional odds models. With high dimensional settings ridge regression solves the problem of non-existence of estimates by keeping all the predictors in the model. But it is does not reduce the dimension by identifying the relevant/significant predictors to get a sparse model with an enhanced predictability. For unordered response categories several methods have been proposed. For example, Friedman et al. (2010) used the L1 penalty for parameter selection in the multinomial logit models, Zahid and Tutz (2010) introduced a variable selection procedure based on likelihood-based boosting which makes variable selection rather than parameter selection as done by Friedman et al. (2010). But for ordered response categories, methods for variable selection seem not to be available. In the following a componentwise boosting technique called *pomBoost* is proposed for the fitting of the proportional odds models with implicit selection of the relevant predictors. Boosting was initially introduced in the machine learning community to improve classification (see Schapire (1990) and Freund and Schapire (1996)). Friedman et al. (2000) showed that boosting can be seen as an approximation to additive modeling with appropriate likelihood function. In the context of linear models, instead of using the LogitBoost cost function, Bühlmann and Yu (2003) used the L2 loss function. A relation between boosting and Lasso was developed by Bühlmann (2006). Bühlmann and Hothorn (2007) provided an overview of boosting. Tutz and Binder (2006) proposed a general likelihood-based boosting procedure for variable selection in generalized additive models (GAM).

In this paper we are using the likelihood-based boosting with one step of Fisher scoring. In many application areas, sometimes the experimenter is interested to see the effect of some predictor(s), and wants them to be a necessary part of the final sparse model. The *pomBoost* technique has the provision to declare some predictor(s) as mandatory which will always be the part of model during fitting/selection process. One advantage of the proposed method is that categorical predictors are treated properly by regularization. Our aim is to select predictors not parameters. Therefore a predictor is selected (or omitted) with all of its categories. Our technique also performs the variable selection instead of parameter selection. Also in the case of ordinal predictors the order of the categories is taken into account by regularization. For regularization, the L2 penalty is used which allows categorical predictors with a large number of categories.

The predictor space for proportional odds models may contain different types of predictors e.g., metric, binary, nominal and/or ordinal predictors. Section 2 explains how the regularization is implemented for these different types of predictors. The algorithm *pomBoost* for the selection of relevant predictors in the proportional odds model is discussed in Section 3. The effectiveness of algorithm is evaluated with respect to the mean squared error (MSE) and selection of relevant predictors using a simulation study in Section 4. In Section 5, the boosting technique is used on some real data sets. Some concluding comments are given in Section 6.

2. Design Space and Regularization for Proportional Odds Models

Let the response variable Y have k ordered categories such that $Y \in \{1, \ldots, k\}$. The ordered response Y may be seen as a coarser version of an unobservable latent variable Z as $Y = r \quad \Leftrightarrow \quad \gamma_{0,r-1} < Z \leq \gamma_{0r}$ for $r = 1, \ldots, k$, where $-\infty = \gamma_{00} < \gamma_{01} < \ldots < \gamma_{0k} = \infty$ define the category boundaries on the unobservable latent continuum. Let $\phi_r(\mathbf{x})$ denote the cumulative probability for the occurrence of response categories up to and including the rth category for a covariate vector $\mathbf{x}$. The proportional odds models has the form

$$\phi_r(\mathbf{x}) = P(Y \leq r|\mathbf{x}) = \frac{\exp(\gamma_{0r} - \mathbf{x}^T\boldsymbol{\gamma})}{1 + \exp(\gamma_{0r} - \mathbf{x}^T\boldsymbol{\gamma})} \qquad r = 1, \ldots, q = k - 1, \tag{1}$$

or equivalently

$$\log\left[\frac{\phi_r(\mathbf{x})}{1 - \phi_r(\mathbf{x})}\right] = \gamma_{0r} - \mathbf{x}^T\boldsymbol{\gamma} \qquad r = 1, \ldots, k - 1. \tag{2}$$

The proportional odds model contains the so-called global parameter $\boldsymbol{\gamma}$, which does not vary across categories, and the intercepts $\{\gamma_{0r}\}$ while vary across categories and must satisfy the condition $\gamma_{01} < \ldots < \gamma_{0q}$ in order to obtain positive probabilities. For the estimation purpose, with k ordered response categories and p predictor parameters, the model can be written as $\log\left[\frac{\phi_{ir}(\mathbf{x})}{1-\phi_{ir}(\mathbf{x})}\right] = \mathbf{X}_i\boldsymbol{\beta}$ with $q \times p^*$ matrix $\mathbf{X}_i = [\mathbf{I}_{q\times q}, \mathbf{1}_{q\times 1} \otimes \mathbf{x}_i^T]$ and $\boldsymbol{\beta}^T = (\boldsymbol{\gamma}_0^T, \boldsymbol{\gamma}^T) = (\gamma_{01}, \ldots, \gamma_{0q}, \gamma_1, \ldots, \gamma_p)$ is a vector of length $p^* = p + q$. The complete design matrix of order $nq \times p^*$ is given as $\mathbf{X}^T = [\mathbf{X}_1, \ldots, \mathbf{X}_n]$. For further details see McCullagh and Nelder (1989) and Fahrmeir and Tutz (2001).

2.1. Regularization

In the boosting algorithm discussed in the next section, each of the predictors (with all of its corresponding parameters) is considered individually for its possible inclusion in the model at a particular boosting iteration. To obtain weak learner in a boosting iteration L2 regularization is used. The intercept terms $\boldsymbol{\gamma}_0$ and mandatory predictors (if any) are considered a necessary part of proportional odds model in each boosting iteration. For simplicity in this section we assume that we have a model with only one predictor which has K parameters associated with it. It means that if the predictor is metric or binary then $K = 1$, otherwise we have a categorical predictor with $K + 1$ categories. The *pomBoost* algorithm discussed in Section 3 updates the regularized estimates of the parameters associated with one variable at a time on the basis of one step Fisher scoring. The regularization using the L2 penalty is implemented differently according to the nature of predictor. Assume that K dummies for the $K + 1$ categories labeled $1, \ldots, K + 1$ are associated to the only predictor in the model. The penalized log-likelihood with ridge penalty is given as

$$l_p(\boldsymbol{\gamma}) = \sum_{i=1}^{n} l_i(\boldsymbol{\gamma}) - \frac{\lambda}{2} J(\boldsymbol{\gamma}), \tag{3}$$

where $l_i(\boldsymbol{\gamma})$ is the log-likelihood contribution of the ith observation and λ is a tuning parameter. If the predictor is nominal then we use the penalty term

$$J(\boldsymbol{\gamma}) = \sum_{j=2}^{K+1} \gamma_j^2 = \boldsymbol{\gamma}^T \mathbf{I}_{K\times K}\, \boldsymbol{\gamma}, \tag{4}$$

in order to obtain regularized estimates. The matrix $\mathbf{I}_{K\times K}$ is an identity matrix which serves for the penalization of K parameter estimates. But if the predictor is ordinal, then parameter estimates of adjacent categories are penalized. Penalizing such differences leads to avoid large differences among the parameter estimates of adjacent categories and provides a smoother coefficient vector. With penalization, the order of the ordinal predictors is not so much focused in the literature. Gertheiss and Tutz (2009) used these differences for penalization rather than using the parameter estimates themselves. In the case of ordinal predictor, the first category is treated as reference category such that $\gamma_1 = 0$ and the penalty term $J(\boldsymbol{\gamma})$ is given by

$$J(\boldsymbol{\gamma}) = \sum_{j=2}^{K+1}(\gamma_j - \gamma_{j-1})^2 = \boldsymbol{\gamma}^T\, \boldsymbol{\Omega}\, \boldsymbol{\gamma}, \tag{5}$$

with $\boldsymbol{\Omega} = \mathbf{U}^T\mathbf{U}$, for a $K \times K$ matrix $\mathbf{U}$ given by

$$\mathbf{U} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -1 & 1 & \ddots & & \vdots \\ 0 & -1 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -1 & 1 \end{bmatrix}.$$

The use of square matrix $\boldsymbol{\Omega}$ in (5) for penalization instead of the identity matrix as in (4) causes the penalization of differences between the parameter estimates of adjacent categories of ordinal predictor. In the next section for having weak learners in our boosting algorithm, two types of penalty terms $J(\boldsymbol{\gamma})$ will be used. If the predictor is ordinal then the penalty term given in (5) is used otherwise the penalty term given in (4) will be our choice.

3. Boosting for Variable Selection and Model Fitting

The likelihood-based componentwise boosting algorithm *pomBoost* proposed in this section uses one step Fisher scoring with ridge penalty in order to obtain a weak learner. The intercept terms $\{\gamma_{0r}\}$, as well as predictor variables that are declared as obligatory will not be penalized. Along with intercepts and mandatory predictors the predictor which improves the fit maximally will be used for updating within a boosting iteration. In order to obtain a weak learner regularization is applied to the candidate predictors. All the predictors are divided into two groups: obligatory predictors (along with the intercept terms) and candidate predictors which are possible candidates to be a part of the final sparse model. Let there are g candidate predictors as $V_1, \ldots, V_g$ and let K_j denotes the number of parameters/dummies associated with the candidate predictor $V_j,\ j = 1, \ldots, g$. So the predictor variable indices $V = \{1, \ldots, p\}$ are partitioned into two mutually exclusive sets as $V = V_o \cup V_1 \cup \ldots \cup V_g$, where V_o represents the obligatory predictors (each predictor may have one or more parameters associated with it) and $V_1, \ldots, V_g$ are g candidate predictors. The total predictor space which is divided into two groups as $V = V_0 \cup V_c$ with $V_c = V_1 \cup \ldots \cup V_g$ has the parameter vector

$\boldsymbol{\beta}^T = (\boldsymbol{\beta}_o^T \ \boldsymbol{\beta}_c^T)$.

For the set of obligatory predictors, log-likelihood function is given as $l(\boldsymbol{\beta}_o) = \sum_{i=1}^n l_i(\boldsymbol{\beta_o})$ with score function $s(\boldsymbol{\beta}_o) = \sum_{i=1}^n \mathbf{X}_{oi}^T \mathbf{D}_i(\boldsymbol{\beta}_o)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\beta}_o)[\mathbf{y}_i - h(\boldsymbol{\eta}_i)] = \mathbf{X}_o^T \mathbf{D}(\boldsymbol{\beta}_o)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}_o)[\mathbf{y} - h(\boldsymbol{\eta})]$. For the set of candidate predictors, let the predictor V_j is considered for refitting in a boosting iteration, and let $\gamma_2, \ldots, \gamma_{K+1}$ are the global parameters associated with K dummies of the predictor V_j. The penalized log-likelihood is then given as

$$l_p(\boldsymbol{\gamma}) = \sum_{i=1}^n l_i(\boldsymbol{\gamma}) - \frac{\lambda}{2} J(\boldsymbol{\gamma}) = \sum_{i=1}^n l_i(\boldsymbol{\gamma}) - \frac{\lambda}{2} \boldsymbol{\gamma}^T \mathbf{P}\, \boldsymbol{\gamma}$$

The penalty matrix $\mathbf{P}$ assumes the value $\boldsymbol{\Omega}_{K_j \times K_j}$ if the predictor variable V_j is ordinal otherwise it is replaced by $\mathbf{I}_{K_j \times K_j}$. The score function for this penalized log-likelihood is given as

$$\begin{aligned} s_p(\boldsymbol{\gamma}) &= \sum_{i=1}^n \mathbf{X}_{ji}^T \mathbf{D}_i(\boldsymbol{\gamma})\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\gamma})[\mathbf{y}_i - h(\boldsymbol{\eta}_i)] - \lambda \mathbf{P}\boldsymbol{\gamma} \\ &= \mathbf{X}_j^T \mathbf{D}(\boldsymbol{\gamma})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\gamma})[\mathbf{y} - h(\boldsymbol{\eta})] - \lambda \mathbf{P}\boldsymbol{\gamma}, \end{aligned}$$

where $\boldsymbol{\gamma}$ is a vector of length K and $\mathbf{X}_j^T = [\mathbf{X}_{j1}, \ldots, \mathbf{X}_{jn}]$ with $\mathbf{X}_{ji} = [\mathbf{1}_{q\times 1} \otimes \mathbf{x}_{ji}^T]$. The matrix $\mathbf{D}_i(\boldsymbol{\gamma}) = \frac{\partial h(\boldsymbol{\eta}_i)}{\partial \boldsymbol{\eta}}$ is the derivative of $h(\boldsymbol{\eta})$ evaluated at $\boldsymbol{\eta}_i = \mathbf{X}_{ji}\boldsymbol{\gamma}$, $\boldsymbol{\Sigma}_i(\boldsymbol{\gamma}) = \text{cov}(\mathbf{y}_i)$ is the covariance matrix of ith observation of $\mathbf{y}$ given parameter vector $\boldsymbol{\gamma}$ and $\mathbf{W}_i(\boldsymbol{\gamma}) = \mathbf{D}_i(\boldsymbol{\gamma})\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\gamma})\mathbf{D}_i^T(\boldsymbol{\gamma})$. For the full design matrix, in matrix notation $\mathbf{y}$ and $h(\boldsymbol{\eta})$ are given by $\mathbf{y}^T = (\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T)$ and $h(\boldsymbol{\eta})^T = (h(\boldsymbol{\eta}_1)^T, \ldots, h(\boldsymbol{\eta}_n)^T)$ respectively. The matrices have block diagonal form $\boldsymbol{\Sigma}(\boldsymbol{\gamma}) = \text{diag}(\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\gamma}))$, $\mathbf{D}(\boldsymbol{\gamma}) = \text{diag}(\mathbf{D}_i(\boldsymbol{\gamma}))$ and $\mathbf{W}(\boldsymbol{\gamma}) = \text{diag}(\mathbf{W}_i(\boldsymbol{\gamma}))$.

The *pomBoost* algorithm can be described as follows:

Algorithm: *pomBoost*

Step 1: (Initialization)

Fit the intercept model $\boldsymbol{\mu}_0 = h(\boldsymbol{\eta}_0)$ by maximizing the likelihood function to obtain $\hat{\boldsymbol{\eta}}_0$ and $h(\hat{\boldsymbol{\eta}}_0)$.

Step 2: Boosting iterations

For $m = 1, 2, \ldots$

Step 2A: For obligatory/mandatory predictors

(i) Fit the model $\boldsymbol{\mu} = h(\hat{\boldsymbol{\eta}}_{m-1} + \mathbf{X}_o\, \boldsymbol{\beta}_o^{F1})$, where $\hat{\boldsymbol{\eta}}_{m-1}$ is treated as an offset and $\mathbf{X}_o^T = [\mathbf{X}_{o1}, \ldots, \mathbf{X}_{on}]$ for $\mathbf{X}_{oi} = [\mathbf{I}_{q\times q}, \mathbf{1}_{q\times 1} \otimes \mathbf{x}_{oi}^T]$ is the design matrix based on the parameters/columns corresponding to V_o. $\boldsymbol{\beta}_o^{F1}$ is computed with one-step Fisher scoring as

$$\boldsymbol{\beta}_o^{F1} = (\mathbf{X}_o^T \mathbf{W}(\hat{\boldsymbol{\eta}}_{m-1})\mathbf{X}_o)^{-1}\, \mathbf{X}_o^T\, W(\hat{\boldsymbol{\eta}}_{m-1})\, \mathbf{D}^{-1}\, (\mathbf{y} - \hat{\boldsymbol{\mu}}_{m-1}).$$

(ii) set $\hat{\boldsymbol{\eta}}_m = \hat{\boldsymbol{\eta}}_{m-1} + \mathbf{X}_o\, \boldsymbol{\beta}_o^{F1}$

(iii) set $\boldsymbol{\beta}_{o(m)} = \boldsymbol{\beta}_{o(m-1)} + \boldsymbol{\beta}_o^{F1}$

Step 2B: For candidate predictors

(i) For $j = 1, \ldots, g$, fit the model $\boldsymbol{\mu} = h(\hat{\boldsymbol{\eta}}_m + \mathbf{X}_j \boldsymbol{\gamma}_j^{F1})$, with offset $\hat{\boldsymbol{\eta}}_m$ and $\mathbf{X}_j$ is the design matrix corresponding to V_j. With one-step Fisher scoring by maximizing penalized log-likelihood, $\boldsymbol{\gamma}_j^{F1}$ is computed as

$$\boldsymbol{\gamma}_j^{F1} = (\mathbf{X}_j^T \mathbf{W}(\hat{\boldsymbol{\eta}}_m)\mathbf{X}_j + \nu\, \mathbf{P})^{-1}\, \mathbf{X}_j^T\, W(\hat{\boldsymbol{\eta}}_m)\, \mathbf{D}^{-1}\, (\mathbf{y} - \hat{\boldsymbol{\mu}}_m),$$

where $\nu = \sqrt{df}_j.\lambda$, with ridge penalty λ and $\mathbf{P} = \boldsymbol{\Omega}_{K_j \times K_j}$, if V_j is ordinal otherwise $\mathbf{P} = \mathbf{I}_{K_j \times K_j}$.

(ii) From the candidate predictors $V_1, \ldots, V_g$, select the predictor say V_{best}, which improves the fit maximally and set

$$\boldsymbol{\beta}_c^{F1} = \begin{cases} \boldsymbol{\gamma}_j^{F1} & \text{if } j \in V_{\text{best}} \\ 0 & \text{if } j \notin V_{\text{best}} \end{cases}$$

(iii) set $\hat{\boldsymbol{\eta}}_m \leftarrow \hat{\boldsymbol{\eta}}_m + \mathbf{X}_c \boldsymbol{\beta}_c^{F1}$

(iv) set $\boldsymbol{\beta}_{c(m)} = \boldsymbol{\beta}_{c(m-1)} + \boldsymbol{\beta}_c^{F1}$

The boosting algorithm uses ridge penalty to obtain the weak learners. But different candidate predictors may have different parameters associated with them, so the ridge penalty is adjusted for degrees of freedom by multiplying λ with $\sqrt{\text{df}_j}$. As a result, for a fixed value of λ with increasing number of parameters for a candidate predictor, the learner becomes more weak than the other candidate predictors with less degrees of freedom. For selecting a predictor for refit in a boosting iteration, different criteria can be used. One possible choice can be the deviance and the predictor with minimum value of deviance $\text{Dev}(\hat{\boldsymbol{\eta}}_m)$ among all candidate predictors is considered for refit. The other choices which should be more appropriate with varying number of parameters for different predictors are Akaike information criterion (AIC) and Bayesian information criterion (BIC), because they also involve the degrees of freedom. Both of these measures are given by

$$\text{AIC} = \text{Dev}(\hat{\boldsymbol{\eta}}_m) + 2\ \text{df}_m,$$

and

$$\text{BIC} = \text{Dev}(\hat{\boldsymbol{\eta}}_m) + \log(n)\ \text{df}_m,$$

where df_m is the effective degrees of freedom given by the trace of the approximate hat matrix $\mathbf{H}_m$ obtained after m boosting iterations. The use of AIC or BIC for predictor selection in a boosting iteration seems to be better choices than the deviance because they involve the effective degrees of freedom. But using these measures can slow the computational process significantly for large sample size and increasing number of candidate predictors. In case of large samples with high-dimensional structure using deviance for predictor selection can reduce the computational burden and makes the algorithm more efficient regarding processing time. In the boosting it is possible that some of the predictors are considered for updating only for a very few number of times. In such case those predictors which are not contributing in the model in a real sense can become a part of the final sparse model. The *pomBoost* algorithm

avoids such predictors (with too small estimates) to be a part of the final model after m boosting iterations. The estimates $\hat{\boldsymbol{\gamma}}_j$ associated with the candidate predictor V_j are set to zero after m boosting iterations if

$$\frac{\frac{1}{K_i}\sum_{j=1}^{K_i}|\hat{\gamma}_{ij}|}{\sum_{i=1}^{p}\frac{1}{K_i}\sum_{j=1}^{K_i}|\hat{\gamma}_{ij}|} < \frac{1}{p}. \tag{6}$$

The degrees of freedom df_m used in the criterion AIC or BIC is computed from the approximate hat matrix $\mathbf{H}_m$ after m boosting iterations. The approximate hat matrix $\mathbf{H}_m$ is defined in the following proposition.

Proposition: In the mth boosting iteration, an approximate hat matrix for which $\hat{\boldsymbol{\mu}}_m \approx \mathbf{H}_m\mathbf{y}$ is given by

$$\mathbf{H}_m = \sum_{j=0}^{m}\mathbf{M}_j\prod_{i=0}^{j-1}(\mathbf{I}-\mathbf{M}_0),$$

where $\mathbf{M}_m = \mathbf{W}_m(\mathbf{X}_m^T\mathbf{W}_m\mathbf{X}_m + \nu\boldsymbol{\Lambda})^{-1}\mathbf{X}_m$ for $\mathbf{D}_m = \mathbf{D}(\hat{\boldsymbol{\eta}}_m))$ and $\mathbf{W}_m = \mathbf{D}_m\boldsymbol{\Sigma}_m^{-1}\mathbf{D}_m^T$.

Proof: Let the predictor variable $V_j = V_{\text{best}}$ is selected after m boosting iterations and for proportional odds model we have $\mathbf{D}_m = \mathbf{D}(\hat{\boldsymbol{\eta}}_m)$ and $\mathbf{W}_m = \mathbf{W}_m(\hat{\boldsymbol{\eta}}_m) = \mathbf{D}_m\hat{\boldsymbol{\eta}}_m\boldsymbol{\Sigma}_m^{-1}\mathbf{D}_m^T\hat{\boldsymbol{\eta}}_m$. By using the Taylor approximation of first order i.e., $h(\hat{\boldsymbol{\eta}}) \approx h(\boldsymbol{\eta}) + (\partial h(\boldsymbol{\eta})/\partial\boldsymbol{\eta}^T)(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})$, we obtain $\hat{\boldsymbol{\mu}}_m \approx \hat{\boldsymbol{\mu}}_{m-1} + \mathbf{D}_m(\hat{\boldsymbol{\eta}}_m - \hat{\boldsymbol{\eta}}_{m-1}) = \hat{\boldsymbol{\mu}}_{m-1} + \mathbf{D}_m\mathbf{X}_j\,\hat{\boldsymbol{\beta}}^{F1} = \hat{\boldsymbol{\mu}}_{m-1} + \mathbf{D}_m\mathbf{X}_j(\mathbf{X}_j^T\mathbf{W}_m\mathbf{X}_j + \nu\,\mathbf{P})^{-1}\,\mathbf{X}_j^T\,(\mathbf{y} - \hat{\boldsymbol{\mu}}_{m-1})$. So we have $\hat{\boldsymbol{\mu}}_m \approx \hat{\boldsymbol{\mu}}_{m-1} + \mathbf{M}_m(\mathbf{y} - \hat{\boldsymbol{\mu}}_{m-1})$ with $\mathbf{M}_m = \mathbf{D}_m\mathbf{W}_m^{-1}\tilde{\mathbf{H}}_m\mathbf{W}\mathbf{D}^{\mathbf{T}^{-1}}$ where $\tilde{\mathbf{H}}_m = \mathbf{W}_m\mathbf{X}_j(\mathbf{X}_j^T\mathbf{W}_m\mathbf{X}_j + \nu\,\mathbf{P})^{-1}\mathbf{X}_j^T$. The expression for $\hat{\boldsymbol{\mu}}_m$ can be written as $\hat{\boldsymbol{\mu}}_m \approx \hat{\boldsymbol{\mu}}_{m-1} + \mathbf{M}_m(\mathbf{y} - \hat{\boldsymbol{\mu}}_{m-1}) = \mathbf{H}_{m-1}\mathbf{y} + \mathbf{M}_m(\mathbf{I} - \mathbf{H}_{m-1})\mathbf{y}$. Expanding in the same way, for mth boosting iteration, the general form of the approximate hat matrix is $\mathbf{H}_m = \sum_{j=0}^{m}\mathbf{M}_j\prod_{i=0}^{j-1}(\mathbf{I} - \mathbf{M}_0)$, with $\hat{\boldsymbol{\mu}}_m \approx \mathbf{H}_m\mathbf{y}$ and the starting value $\hat{\boldsymbol{\mu}}_0 = \mathbf{M}_0\mathbf{y}$.

4. Simulation Study

In this section properties of *pomBoost* algorithm are investigated using simulated data. For the response variable with three and five ordered categories, we generate the predictor space with continuous and binary covariates for different samples of size n. Our main focus is on sparse model fitting and we are using the high-dimensional predictor space with few relevant predictors. The continuous covariates are drawn from a $p-$dimensional multivariate normal distribution with variance 1 and correlation between two covariates $\mathbf{x}_j$ and $\mathbf{x}_l$ being $\rho^{|j-l|}$. We generate the data with k response categories for ten different settings with different sample sizes and values of ρ. The description of these ten settings is as follows:

Setting	k	n	Continuous	Binary	ρ	p_{info}	S
1	3	50	40 (4)	10 (1)	0.3	5	50
2	3	100	40 (4)	10 (1)	0.3	5	50
3	3	50	40 (4)	10 (1)	0.8	5	50
4	3	100	40 (4)	10 (1)	0.8	5	50
5	3	100	180 (6)	20 (2)	0.3	8	50
6	3	100	180 (6)	20 (2)	0.8	8	50
7	3	100	400 (6)	100 (2)	0.3	8	25
8	3	100	400 (6)	100 (2)	0.8	8	25
9	5	100	40 (4)	10 (1)	0.3	5	50
10	5	100	40 (4)	10 (1)	0.8	5	50

The numbers within brackets are the number of informative continuous/binary predictors. p_{info} is the total number of informative predictors (informative predictors have non-zero values for true parameters while all other non-informative predictors have zero parameter values) in a particular setting and S is the number of simulations. To investigate the performance of the algorithm with categorical predictors, in the eleventh setting nominal and ordinal predictors with three and four categories are considered with three ordered categories for the response variable. For a sample of size 100, predictor space of total 80 predictors with 20(2) predictors of each of four types is generated. With this setting $S = 50$ samples are generated. For the true parameter values $\sum_{j=1}^{p_{\text{info}}} K_j$ values (where p_{info} is the total number of informative predictors) are obtained by the formula $(-1)^j \exp(-2(j-1)/20)$ for $j = 1, \ldots, \sum_{j=1}^{p_{\text{info}}} K_j$. These values are randomly allotted to the global parameters $\boldsymbol{\gamma}_{\text{info}}$ corresponding to the informative predictors. The true values of the intercepts $\boldsymbol{\gamma}_0^T = (-0.3, 0.8)$ and $\boldsymbol{\gamma}_0^T = (-0.8, -0.3, 0.3, 0.8)$ are used for proportional odds models with three and five response categories respectively. The true parameter vector $\boldsymbol{\beta}^T = (\boldsymbol{\gamma}_0^T, \boldsymbol{\gamma})$ is then multiplied with a constant c_{snr} which is chosen so that the signal-to-noise ratio is 3.0. For the componentwise boosting we set the maximum number of iterations equal to 400. For regularization, we tried to use the same value of ridge penalty λ for all samples in a particular setting for a particular variable selection criterion. The value of λ is chosen so that there are at least 50 boosting iterations in each sample of all settings. Deviance with 10−fold cross-validation is used as a stopping criterion. In some instances the optimal (final) boosting iteration less than 400 is not obtained and the results are then based on maximum number of iterations.

4.1. Identification of Informative Predictors

The algorithm *pomBoost* fits the proportional odds model by implicitly selecting the relevant predictors. In high-dimensional structures, it is important that the final sparse model contains all informative predictors and ideally no irrelevant predictor. The ”hit rates” and ”false alarm rates” are used to evaluate the performance of the algorithm

regarding proper variable selection. The hit rate which is the proportion of correctly identified informative predictors is given as

$$\text{hit rate} = \frac{\sum_{j=1}^{p} I(\boldsymbol{\gamma}_j^{\text{true}} \neq \mathbf{0}).I(\hat{\boldsymbol{\gamma}}_j \neq \mathbf{0})}{\sum_{j=1}^{p} I(\boldsymbol{\gamma}_j^{\text{true}} \neq \mathbf{0})},$$

and false alarm rate which is the proportion of non-informative predictors identified as informative is given by

$$\text{false alarm rate} = \frac{\sum_{j=1}^{p} I(\boldsymbol{\gamma}_j^{\text{true}} = \mathbf{0}).I(\hat{\boldsymbol{\gamma}}_j \neq \mathbf{0})}{\sum_{j=1}^{p} I(\boldsymbol{\gamma}_j^{\text{true}} = \mathbf{0})}.$$

The vector $\boldsymbol{\gamma}_j^{\text{true}}$, $j = 1, \ldots, p$ contains true global parameter values for the predictor V_j and $\hat{\boldsymbol{\gamma}}_j$ is the vector of corresponding estimates. The indicator function I(expression) assumes value 1, if "expression" is true and 0 otherwise. The hit rates and false alarm rates with deviance, AIC and BIC as the predictor selection criteria are given in Table 1. The results show that as far as selection of relevant predictors is concerned, the algorithm is selecting most of the times all relevant predictors in all settings with three or five response categories. In some settings, even no relevant predictor is missed in any sample although some acceptable number of non-informative predictors are also selected with informative predictors. For setting 11 where we have only categorical predictors and the each predictor with all of its associated parameters is to be selected or rejected for updating, the hit rate is good especially with deviance as predictor selection criterion. But with respect to false alarm rate AIC and BIC are performing better than deviance. The general view of Table 1 reflects that results of hit rates and false alarm rates for all predictor selection criteria are very close to each other. More specifically if hit rates are focused then deviance may be our choice for predictor selection and if false alarm rates are considered then BIC seems to be a good choice with AIC as its strong competitor. But AIC seems to be more appropriate choice while considering both of the factors that is selection of all relevant predictors with minimum possible irrelevant predictors.

4.2. Empirical Results

In this section we are comparing the estimates/fit for the sparse model chosen from componentwise boosting procedure with ridge estimates (see Zahid (2011)). The usual MLE is not existing in all considered high-dimensional settings. For comparison we are using three different measures: mean squared error (MSE) of the parameter estimates $\hat{\boldsymbol{\beta}}$, mean deviance for the fit (deviance($\hat{\boldsymbol{\pi}}$)) and mean prediction error (MPE). The MSE($\hat{\boldsymbol{\beta}}$) is computed using the formula $\frac{1}{S}\sum_s \|\hat{\boldsymbol{\beta}}_s^{\text{method}} - \boldsymbol{\beta}^{\text{true}}\|^2$ and the deviance for the fit is computed as $D = 2.\sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij}\log\left(\frac{y_{ij}}{\hat{\pi}_{ij}}\right)$ with $y_{ij}\log\left(\frac{y_{ij}}{\hat{\pi}_{ij}}\right) = 0$ for $y_{ij} = 0$. To compute the mean prediction error (MPE), we generate a new test data set of size $n = 1000$ observations with the same parameters as in simulation study for each setting. The mean prediction error based on the deviance measure for this test data set is computed as $\text{MPE} = \frac{1}{S}\sum_s D_s = \frac{1}{S}\sum_s 2.\Big[\sum_{i=1}^{n}\sum_{j=1}^{k} \pi_{ijs}^{\text{test}}\log\left(\frac{\pi_{ijs}^{\text{test}}}{\hat{\pi}_{ijs}^{\text{test}}}\right)\Big]$. The values of these three measures are given in Table 2 for boosting technique with deviance, AIC and BIC as predictor selection criteria and ridge regression with all predictors (informative and non-informative) in the model. The results with boosting are better than those for the ridge regression with an exception for setting 1 where the boosting is showing some weak

results in terms of MSE($\hat{\boldsymbol{\beta}}$) and the fit but still performing much better in terms of prediction error. If we look at the results of boosting with different predictor selection criteria, the use of deviance as the predictor selection criterion is showing better results than AIC and BIC in all settings with three or five categories response models. But for high-dimensional settings with moderate correlation among the continuous predictors such as in setting 5 and 7, BIC is showing the best results followed by the AIC. The log values of $\hat{\boldsymbol{\beta}}$ and MPE for boosting and ridge regression in some selected settings are shown graphically in terms of box plots in Figure 1 and 2. In both figures, the box plots associated with 5−categories response models (setting 9 and 10) are reflecting more significant improvement for boosting approach (especially with deviance as a predictor selection criterion) over the ridge estimates. The solid circles within each box of the box plots represent the mean of the data for which the box plots are drawn.

TABLE 1: Hit rates (HR) and false alarm rates (FAR) for identifying the informative predictors when deviance, AIC and BIC are used as criteria for selecting a predictor in a boosting iteration. Deviance is used as stopping criterion with 10−fold cross-validation.

	Deviance		AIC		BIC	
	HR	FAR	HR	FAR	HR	FAR
Setting 1	0.9920	0.1164	0.9120	0.0378	0.9160	0.0396
Setting 2	1.0000	0.0787	1.0000	0.0640	1.0000	0.0622
Setting 3	0.9920	0.1040	0.9880	0.0804	0.9920	0.0671
Setting 4	1.0000	0.1164	1.0000	0.1000	1.0000	0.0840
Setting 5	0.8300	0.0924	0.8175	0.0767	0.8000	0.0548
Setting 6	0.9250	0.0874	0.8075	0.0572	0.7975	0.0501
Setting 7	0.9850	0.0487	0.9950	0.0452	1.0000	0.0302
Setting 8	0.8300	0.0550	0.8150	0.0454	0.8000	0.0364
Setting 9	1.0000	0.0676	0.9520	0.0729	0.8560	0.0724
Setting 10	1.0000	0.0342	0.9680	0.1067	0.9280	0.1222
Setting 11	0.9100	0.1542	0.7450	0.0878	0.7250	0.0903

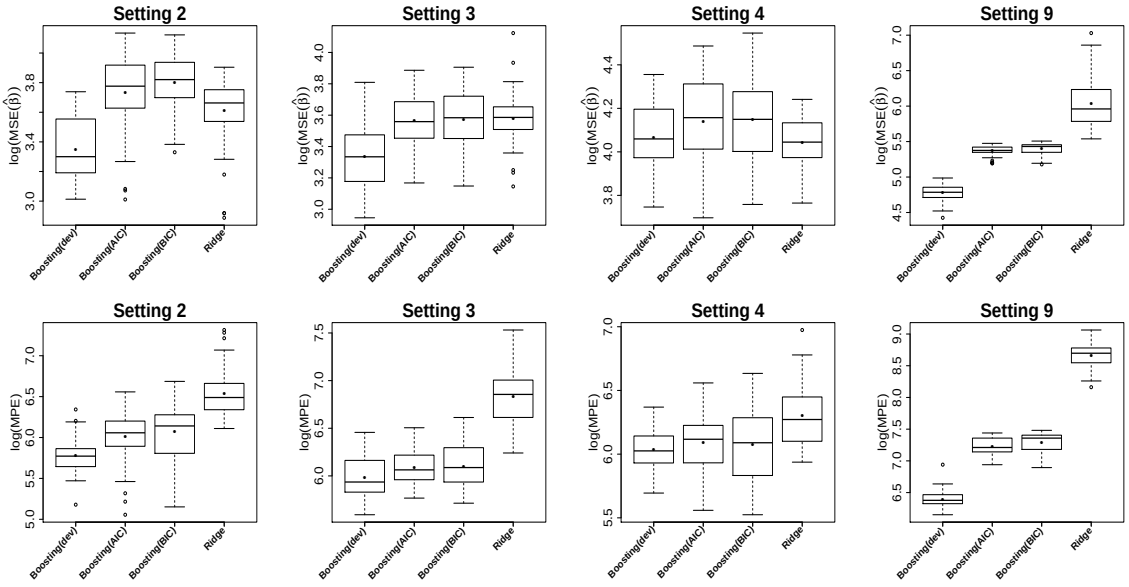

FIGURE 1: Illustration of the simulation study: Box plots for comparing Boosting (with deviance, AIC and BIC as predictor selection criteria) with ridge estimates in terms of log(MSE($\hat{\boldsymbol{\beta}}$)) (top panel) and Mean Prediction Error i.e., log(MPE) (bottom panel). The solid circles within the boxes represent the mean of the observations for which box plots are drawn.

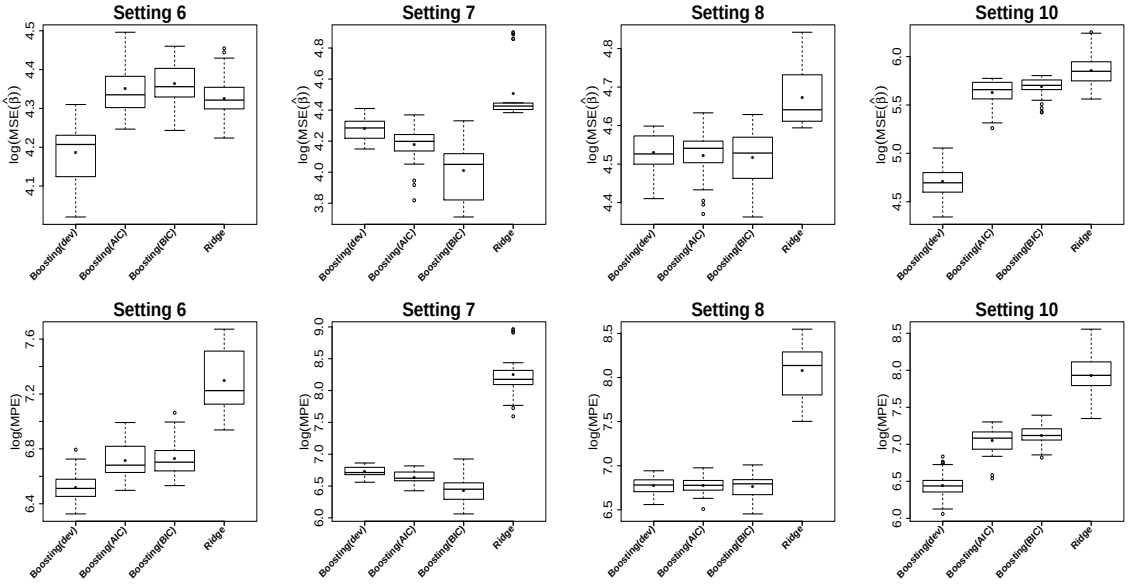

FIGURE 2: Illustration of the simulation study for high dimensional settings: Box plots for comparing Boosting (with deviance, AIC and BIC as predictor selection criteria) with ridge estimates in terms of log(MSE($\hat{\boldsymbol{\beta}}$)) (top panel) and Mean Prediction Error i.e., log(MPE) (bottom panel). The solid circles within the boxes represent the mean of the observations for which box plots are drawn.

TABLE 2: Comparison of Boosting approach with ridge regression in terms of MSE($\hat{\boldsymbol{\beta}}$), deviance($\hat{\boldsymbol{\pi}}$) and Mean Prediction Error (MPE).

	MSE($\hat{\boldsymbol{\beta}}$)				deviance($\hat{\boldsymbol{\pi}}$)				MPE			
	Boosting				Boosting				Boosting			
	deviance	AIC	BIC	Ridge	deviance	AIC	BIC	Ridge	deviance	AIC	BIC	Ridge
Setting 1	70.2648	111.1998	112.1113	56.8419	20.6806	54.3948	55.4751	35.3729	698.0362	1205.8165	1226.7801	2127.6189
Setting 2	29.0005	43.2247	45.4894	37.8050	27.8537	35.0938	36.9378	32.0932	328.3775	428.2221	463.4275	721.3418
Setting 3	28.6156	35.8152	36.1927	36.2654	15.8493	19.5011	19.8625	22.1099	407.9411	446.9234	457.4776	972.9473
Setting 4	58.9571	64.1723	64.4525	57.3699	42.0505	43.7915	43.8009	37.3850	424.0350	454.0069	454.2585	560.1846
Setting 5	58.3875	53.1730	47.5466	65.0128	44.6954	41.7891	39.1264	82.5753	696.7068	668.2237	637.7509	2638.3867
Setting 6	65.9816	77.6853	78.6995	75.7883	41.9446	59.2422	61.3988	49.8337	682.1554	830.3144	839.9394	1515.1373
Setting 7	72.5417	65.7581	56.2641	91.1384	59.5380	53.6569	46.7804	129.5409	836.1576	766.0989	629.7030	4069.3070
Setting 8	92.7841	92.3475	91.8610	107.3350	56.7862	58.9721	60.3961	109.0961	882.7764	876.0458	870.5403	3372.8829
Setting 9	119.3682	215.8329	221.8824	447.5947	43.2587	111.5447	120.1233	174.2679	605.6147	1391.4402	1483.4945	5908.5421
Setting 10	112.1967	280.6495	295.9767	352.5364	64.3926	104.4435	112.5791	92.3607	632.6943	1170.4883	1251.0021	2866.8977

5. Application

The data set being analyzed in this section is taken from UCI repository (`http://archive.ics.uci.edu/ ml/ datasets/Housing`). The data is about housing values in suburbs of Boston. The response variable Median value of owner-occupied homes in \$1000's (MEDV) is categorized using the four ordered categories as MEDV< 10, $10 \leq$ MEDV < 25, $25 \leq$ MEDV < 40 and MEDV ≥ 40. There are thirteen predictors as: per capita crime rate by town (CRIM); proportion of residential land zoned for lots over 25,000 sq.ft. (ZN); proportion of non-retail business acres per town (INDUS); Charles River dummy variable (CHAS= 1 if tract bounds river; CHAS= 0 otherwise); nitric oxides concentration (NOX)(parts per 10 million); average number of rooms per dwelling (RM); proportion of owner-occupied units built prior to 1940 (AGE); weighted distances to five Boston employment centres (DIS); index of accessibility to radial highways (RAD); full-value property-tax rate per $\$10,000$ (TAX); pupil-teacher ratio by town (PTRATIO); $1000(\text{Bk} - 0.63)^2$ where Bk is the proportion of blacks by town (B) and % lower status of the population (LSTAT). The algorithm *pomBoost* is used for fitting the proportional odds model with selection of predictors. For predictor selection in a boosting iteration deviance, AIC and BIC are used as criteria and deviance is used as stopping criteria with 10−fold cross-validation. For regularization same value of ridge penalty i.e., $\lambda = 500$ is used for all three predictor selection criteria. The deviance as predictor selection criterion suggests four predictors RM, PTRATIO, B and LSTAT as informative. AIC and BIC as selection criteria are more strict and exclude PTRATIO and B. For boosting we used 400 as the maximum number of iterations. The parameter estimates with boosting algorithm are given in Table 3. For the deviance and AIC the selected number of iterations was 109 and 74, respectively. For BIC no optimal iteration less than 400 was found, so the maximum number of iteration is used for the results. The boosting coefficients build-up with deviance, AIC and BIC as predictor selection criteria are given in Figure 3.

TABLE 3: *Parameter estimates for Housing data with boosting when deviance, AIC and BIC are used as predictor selection criteria and deviance is used as stopping criterion based on* 10−*fold cross-validation.*

Predictor	Deviance	AIC	BIC	Predictor	Deviance	AIC	BIC
Intercept 1	−5.9429	−5.0165	−5.2761	RM	−1.0631	−1.4932	−1.4455
Intercept 2	2.1398	1.7592	1.8516	AGE	0	0	0
Intercept 3	5.2280	4.7416	4.8856	DIS	0	0	0
CRIM	0	0	0	RAD	0	0	0
ZN	0	0	0	TAX	0	0	0
INDUS	0	0	0	PTRATIO	0.4873	0	0
CHAS	0	0	0	B	−0.3670	0	0
NOX	0	0	0	LSTAT	1.6842	0.9941	1.3231

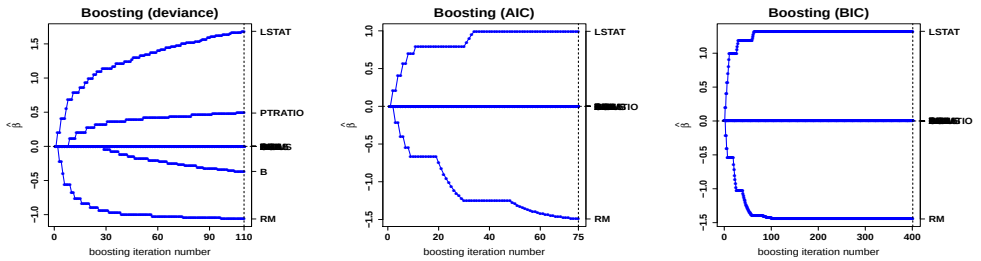

FIGURE 3: Coefficients build-up with component wise boosting for Housing data. Deviance (left panel), AIC (central panel) and BIC (right panel) are used as predictor selection criteria. Deviance with 10−fold cross-validation is used as stopping criterion. The names of non-informative predictors are overlapped against zero value on right side of each graph.

6. Concluding Remarks

In regression, proportional odds models (POM) are commonly used to model response variable with ordered categories. In many application areas it is common to consider a large number of predictors for the initial model to reduce the modeling bias. But to increase the predictive ability of the model and for a better interpretation of the parameters, models should be sparse. Variable selection is an important but challenging part of model building. A judicious predictor selection criterion selects a subset of significant predictors which have potential influence on the response variable. The issue of variable selection in ordinal regression has somewhat neglected in the literature. Although Lu and Zhang (2007) refer to variable selection for proportional odds models they discussed variable selection in survival analysis. The proposed algorithm is an effort to fill this gap. The proposed boosting technique fits proportional odds model by implicitly selecting the relevant predictors. Unlike multinomial logit models which have category specific estimates, proportional odds models have so called global parameters. But in case of a categorical predictor more than one parameters are linked with it. To obtain the weak learner in a boosting iteration, regularization with ridge penalty is used. Regularization allows to include categorical predictors with large number of categories in the model. The predictor selection indicates the selection of all parameters for a predictor. Our componentwise boosting procedure picks potentially influential predictors not the parameters. The algorithm *pomBoost* distinguishes between mandatory predictor(s) and other predictors among which selection is required. For regularization with ordinal predictors, the ordering of the categories should be considered. In such case rather than penalizing the estimates, differences between the parameter estimates of adjacent categories should be penalized. The proposed method differentiates among nominal, ordinal and binary/continuous predictors and performs the regularization for the candidate predictors according to their nature. Although we are considering the proportional odds model only the procedure is easily extended to any ordered regression model.

References

Agresti, A., 1999. Modelling ordered categorical data: Recent advances and future challenges. Statistics in Medicine 18, 2191–2207.

Ananth, C. V., Kleinbaum, D. G., 1997. Regression models for ordinal responses: A review of methods and applications. International Journal of Epidemiology 26, Number 6, 1323–1333.

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

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

Bühlmann, P., Yu, B., 2003. Boosting with l2 loss: Regression and classification. Journal of the American Statistical Association 98, 324–339.

Fahrmeir, L., Tutz, G., 2001. Multivariate Statistical Modelling based on Generalized Linear Models. second ed. Springer-Verlag NewYork, Inc.

Freund, Y., Schapire, R. E., 1996. Experiments with a new boosting algorithm. Machine Learning: Proc. of the Thirteenth International Conference, 148–156.

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

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

Gertheiss, J., Tutz, G., 2009. Penalized regression with ordinal predictors. International Statistical Review 77, 345–365.

Lu, W., Zhang, H. H., 2007. Variable selection for proportional odds model. Statistics in Medicine 26, 3771–3781.

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

McCullagh, P., Nelder, J., 1989. Generalized Linear Models. second ed. Chapman & Hall, NewYork.

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

Tutz, G., Binder, H., 2006. Generalized additive modelling with implicit variable selection by likelihood based boosting. Biometrics 62, 961–971.

Zahid, F. M., 2011. Ordinal ridge regression with categorical predictors. Technical Report No. - -. Institute of Statistics, Ludwig-Maximilians-University Munich, Germany.

Zahid, F. M., Tutz, G., 2010. Multinomial logit models with implicit variable selection. Technical Report No. 89. Institute of Statistics, Ludwig-Maximilians-University Munich, Germany.