International Statistical Review
Revue Internationale de Statistique

International Statistical Review (2013), 81, 3, 388–406 doi:10.1111/insr.12032

Proportional Odds Models with High-Dimensional Data Structure

Faisal Maqbool Zahid[1] and Gerhard Tutz[2]

[1]*Department of Statistics, Government College University Faisalabad, Pakistan*
E-mail: drfaisal@gcuf.edu.pk
[2]*Department of Statistics, Ludwig-Maximilians-University Munich, Akademiestrasse 1, D-80799 Munich, Germany*
E-mail: gerhard.tutz@stat.uni-muenchen.de

Summary

The proportional odds model 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 to fit 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) that 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 response categories have been proposed; see, for example, McCullagh (1980), Agresti (1999) and Ananth & Kleinbaum (1997). The most widely used model is the proportional odds model (POM), also known as cumulative logit model. Although the parameterisation is sparser than in the multinomial logit model, with increasing number of covariates, the usual maximum likelihood approach may fail. But 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 possible alternative to the usual likelihood approach is penalized likelihood. With high-dimensional settings, ridge regression solves the problem of non-existence of estimates by keeping all the predictors in the model. But it does not reduce the dimension by identifying the relevant/significant predictors to obtain 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 & Tutz (2013) introduced a variable selection procedure based on likelihood-based boosting, which makes variable selection rather than parameter selection as performed by Friedman *et al.* (2010). But for ordered response categories, methods for variable selection seem to be scarce.

In the following, a componentwise boosting technique for ordinal regression models called *pomBoost* is proposed that is able to select predictors. Boosting was initially introduced in the machine learning community to improve classification (see Schapire, 1990, and Freund & Schapire, 1996). Friedman *et al.* (2000) showed that boosting can be seen as an approximation to additive modelling with appropriate likelihood function. In the context of linear models, instead of using the LogitBoost cost function, Bühlmann & Yu (2003) used the L2 loss function. A relation between boosting and lasso was developed by Bühlmann (2006). Bühlmann & Hothorn (2007) provided an overview of boosting. Tutz & Binder (2006) proposed a general likelihood-based boosting procedure for variable selection in generalised additive models.

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 performs 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, which allows categorical predictors with a large number of categories, is used.

The predictor space for POMs may contain different types of predictors, for example, metric, binary, nominal and/or ordinal predictors. Section 2.4 explains how the regularization is implemented for these different types of predictors. The algorithm *pomBoost* for the selection of relevant predictors in the POM is discussed in Section 2.5. 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 3. In Section 4, the boosting technique is used on some real data sets. Some concluding comments are given in Section 5.

2 Ordinal Regression Models

2.1 Basic Models

Let the response variable Y have k ordered categories such that $Y \in \{1, \ldots, k\}$. The response vector of dummy variables for observation i is of the form $\mathbf{y}_i^T = (y_{i1}, \ldots, y_{i,k-1})$, with components

$$y_{ij} = \begin{cases} 1 & \text{if } Y_i = j, \quad j = 1, \ldots, k-1, \\ 0 & \text{if } Y_i = k. \end{cases}$$

The most widely used models for ordinal responses are cumulative type models, in which the ordered response Y may be seen as a coarser version of an unobservable latent variable Z. Let the latent variable be given by $Z = \mathbf{x}^T\boldsymbol{\gamma} + \epsilon$, with the noise variable having distribution function on $F(.)$, and the link between observable and latent variable be determined by $Y = r \Leftrightarrow \gamma_{0,r-1} < Z \leq \gamma_{0r}$ for $r = 1, \ldots, k$, where $-\infty = \gamma_{00} < \gamma_{01} < \ldots < \gamma_{0k} = \infty$ define the category boundaries on the unobservable latent continuum. The resulting model is the cumulative model

$$\phi_r(\mathbf{x}) = P(Y \leq r|\mathbf{x}) = F(\gamma_{0r} - \mathbf{x}^T\boldsymbol{\gamma}). \; r = 1, \ldots, k-1, \tag{1}$$

where $\phi_r(\mathbf{x})$ denote the cumulative probability for the occurrence of response categories up to and including the r-th category for a covariate vector $\mathbf{x}$. The most prominent member of this family of models is the POM, which uses the logistic distribution function $F(.)$. It is given by

$$\phi_r(\mathbf{x}) = P(Y \leq r|\mathbf{x}) = \frac{\exp\left(\gamma_{0r} - \mathbf{x}^T\boldsymbol{\gamma}\right)}{1 + \exp\left(\gamma_{0r} - \mathbf{x}^T\boldsymbol{\gamma}\right)} \qquad r = 1, \ldots, k-1. \tag{2}$$

An alternative form is

$$\log\left[\frac{\phi_r(\mathbf{x})}{1 - \phi_r(\mathbf{x})}\right] = \gamma_{0r} - \mathbf{x}^T\boldsymbol{\gamma} \qquad r = 1, \ldots, k-1. \tag{3}$$

The cumulative model was propagated in particular by McCullagh (1980); extensions were considered, for example, by Genter & Farewell (1985) and Armstrong & Sloan (1989). Alternative models are the sequential model $P(Y = r|Y \geq r, \boldsymbol{\gamma}) = F(\mathbf{x}^T\boldsymbol{\gamma}), r = 1, \ldots, k-1$ and the adjacent categories model $P(Y = r|Y \in \{r, r+1\}, \boldsymbol{\gamma}) = F(\mathbf{x}^T\boldsymbol{\gamma}), r = 1, \ldots, k-1$, where again $F(.)$ is a fixed distribution function. For an overview, see Agresti (2013) or Tutz (2012).

2.2 Sparsity and High-Dimensional Data

Categorical data contain less information than metric data. This is most severe for binary data but is also an issue with ordered categories. One consequence is that for sparse data, which may be due to high-dimensional predictors, maximum likelihood estimates may not exist. In the cumulative model, an additional problem is that intercepts have to be ordered; that is, $\gamma_{01} < \ldots < \gamma_{0,k-1}$ has to hold. So, in high dimensions, the usual Fisher scoring has to be modified. In the following, we sketch the handling of sparse data in the modelling of categorical data referring to categorical predictors as well as categorical responses.

Explicit treatment of sparse data was considered by Simonoff (1983, 1987, 1995) and Simonoff & Tutz (2000) but was restricted to contingency tables. In these approaches, one has large contingency tables, possibly with ordered categories, but small sample size. Therefore, sparsity is due to sample size, not to high-dimensional predictors. The fitting approaches give explicit conditions on the sample size to obtain asymptotic results. Because of the restriction to contingency tables, they seem not to be widely used.

Alternative approaches that focus on categorical predictors are the smoothing methods derived by Aitchison & Aitken (1976), Aitken (1983), Bowman (1980) and Tutz (1988). The basic idea is to borrow strength from the neighbourhood when predictors are ordinal by using kernel functions that smooth over the neighbourhood categories. The approaches aim at using localised estimates for categorical data but had not much impact on the literature and are rarely applied, which is partly due to their limited usefulness when many predictors are available. With increasing dimensions, kernels as well as other localisation methods suffer quickly from the curse of dimensionality; see, for example, Hastie *et al.* (2009). More timely methods that are able to select categorical predictors and also to find the clusters of categories that have the same effect are based on regularization techniques. They have been proposed for metric responses by Gertheiss & Tutz (2009, 2010). Extensions to more general responses within the framework of generalized linear models were given by Gertheiss *et al.* (2011). Methods for ordinal categorical predictors that allow to select predictors and fuse adjacent categories that are not to be distinguished are investigated in Tutz & Gertheiss (2013). For further regularization techniques for categorical predictors, see also Gertheiss & Oehrlein (2011). Although it is straightforward to derive, the methods seem to have not yet been extended to the case of

ordered responses. The step is not big, but ordered response model is not common generalised linear models but multivariate generalised linear models and therefore makes extended fitting procedures necessary.

Regularized estimators for ordinal categorical responses have been proposed more recently, but available literature is still scarce. Zahid & Ramzan (2012) extended the ridge estimator, which was proposed by Hoerl & Kennard (1970) and adapted to generalised linear models by Nyquist (1991), to the ordinal POM. Archer & Williams (2012) considered L_1-penalization for continuation ratio models with high-dimensional predictors when modelling an ordinal response using gene expression microarray data.

2.3 Boosting for Variable Selection and Model Fitting

The ordinal regression model (2) contains the so-called global parameter $\boldsymbol{\gamma}$, which does not vary across categories, and the intercepts $\{\gamma_{0r}\}$, which vary across categories and must satisfy the condition $\gamma_{01} < \ldots < \gamma_{0,k-1}$ 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 $g(\phi_{i1}, \ldots, \phi_{i,k-1}) = \mathbf{X}_i\boldsymbol{\beta}$, where $g(.)$ is a $(k-1)$-dimensional link function and $\mathbf{X}_i$ a design matrix. For the POM, which is the main model considered here, one has $\log[\phi_{ij}(\mathbf{x}_i)/1-\phi_{ij}(\mathbf{x}_i)] = \mathbf{X}_i\boldsymbol{\beta}$ with $(k-1)\times p^*$ matrix $\mathbf{X}_i = \left[\mathbf{I}_{(k-1)\times(k-1)}, \mathbf{1}_{(k-1)\times 1} \otimes \mathbf{x}_i^T\right]$ and $\boldsymbol{\beta}^T = \left(\boldsymbol{\gamma}_0^T, \boldsymbol{\gamma}^T\right) = (\gamma_{01}, \ldots, \gamma_{0j}, \ldots, \gamma_{0,k-1}, \gamma_1, \ldots, \gamma_p)$ is a vector of length $p^* = p + (k-1)$. The complete design matrix of order $n(k-1)\times p^*$ is given as $\mathbf{X}^T = [\mathbf{X}_1, \ldots, \mathbf{X}_n]$. For the usual maximum likelihood estimator and further details, see McCullagh (1980) or Tutz (2012).

2.4 Regularized Estimators for Categorical Responses

In this section, we are considering the basic estimators (learners) that are used as a building block in the boosting algorithm discussed in the next section. In boosting, 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. One concept in boosting is that of a weak learner, which means that one should not try to maximally improve the fit within one step because selection and resistance to over-fitting would suffer. Originally, weak learners have been considered in classification problems as learners that are slightly better than guessing (Freund & Schapire, 1997). Here, 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 POM in each boosting iteration. For simplicity, in this section, we assume that we have a model with only one predictor that 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 2.5 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

$$l_p(\boldsymbol{\gamma}) = \sum_{i=1}^{n} l_i(\boldsymbol{\gamma}) - \frac{\lambda}{2} J(\boldsymbol{\gamma}), \tag{4}$$

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

$$J(\boldsymbol{\gamma}) = \sum_{j=2}^{K+1} \gamma_j^2 = \boldsymbol{\gamma}^T \mathbf{I}_{K\times K}\, \boldsymbol{\gamma}, \tag{5}$$

in order to obtain regularized estimates. The matrix $\mathbf{I}_{K\times K}$ is an identity matrix that serves for the penalization of K parameter estimates. But if the predictor is ordinal, then differences of 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 & 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

$$J(\boldsymbol{\gamma}) = \sum_{j=2}^{K+1} (\gamma_j - \gamma_{j-1})^2 = \boldsymbol{\gamma}^T\, \boldsymbol{\Omega}\, \boldsymbol{\gamma}, \tag{6}$$

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 (6) for penalization instead of the identity matrix as in (5) 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 (6) is used; otherwise, the penalty term given in (5) will be our choice.

2.5 Boosting for Categorical Response Models

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 that 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 be g candidate predictors as $V_1, \ldots, V_g$, and let K_j denote 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 intercept terms are considered to be as obligatory 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 $\boldsymbol{\beta}^T = \left(\boldsymbol{\beta}_o^T \;\; \boldsymbol{\beta}_c^T\right)$,

with $\boldsymbol{\beta}_c^T = \left(\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_j^T, \ldots, \boldsymbol{\beta}_g^T\right)$ where $\boldsymbol{\beta}_j^T = (\gamma_{2j}, \ldots, \gamma_{K+1,j})$ the parameter vector associated with dummies of the j-th candidate predictor V_j. 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 be considered for refitting in a boosting iteration. The penalized log-likelihood is then given as

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

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

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

where $\boldsymbol{\beta}_j$ is a vector of length K and $\mathbf{X}_j^T = [\mathbf{X}_{j1}, \ldots, \mathbf{X}_{jn}]$ with $\mathbf{X}_{ji} = \left[\mathbf{1}_{(k-1)\times 1} \otimes \mathbf{x}_{ji}^T\right]$. 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 $\boldsymbol{\eta}_i = \mathbf{X}_{ji}\boldsymbol{\beta}_j$, $\boldsymbol{\Sigma}_i(\boldsymbol{\beta}_j) = \mathrm{cov}(\mathbf{y}_i)$ is the covariance matrix of i-th observation of $\mathbf{y}$ given parameter vector $\boldsymbol{\beta}_j$ and $\mathbf{W}_i(\boldsymbol{\beta}_j) = \mathbf{D}_i(\boldsymbol{\beta}_j)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\beta}_j)\mathbf{D}_i^T(\boldsymbol{\beta}_j)$. For the full design matrix in matrix notation, $\mathbf{y}$ and $h(\boldsymbol{\eta})$ are given by $\mathbf{y}^T = \left(\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T\right)$ and $h(\boldsymbol{\eta})^T = (h(\boldsymbol{\eta}_1)^T, \ldots, h(\boldsymbol{\eta}_n)^T)$, respectively. The vector function $h(\boldsymbol{\eta}_i)$, which is the inverse of g, $h = g^{-1}$, yields the vector of fitted response values $\boldsymbol{\mu}$ (which is here a vector of probabilities $\boldsymbol{\pi}$) for the i-th observation with the elements computed by use of (2). The matrices have block diagonal form $\boldsymbol{\Sigma}(\boldsymbol{\beta}_j) = \mathrm{diag}\left(\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\beta}_j)\right)$, $\mathbf{D}(\boldsymbol{\beta}_j) = \mathrm{diag}(\mathbf{D}_i(\boldsymbol{\beta}_j))$ and $\mathbf{W}(\boldsymbol{\beta}_j) = \mathrm{diag}(\mathbf{W}_i(\boldsymbol{\beta}_j))$.
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 $\mathbf{X}_{oi} = [\mathbf{I}_{(k-1)\times(k-1)}, \mathbf{1}_{(k-1)\times 1} \otimes \mathbf{x}_{oi}^T]$ 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 $\boldsymbol{\beta}_{o(m)} = \boldsymbol{\beta}_{o(m-1)} + \boldsymbol{\beta}_o^{F1}$ (vector of parameters associated with intercepts, and obligatory predictors if any).

Step 2B: For candidate predictors

(i) For $j = 1, \ldots, g$, fit the model $\boldsymbol{\mu} = h(\hat{\boldsymbol{\eta}}_m + \mathbf{X}_j \boldsymbol{\beta}_j^{F1})$, 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, $\boldsymbol{\beta}_j^{F1}$ is computed for each candidate predictor as

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

where $\nu = \sqrt{df_j}.\lambda$, with df_j = number of parameters associated with jth candidate predictor, λ is ridge penalty, 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 populate the parameter vector associated with all candidate predictors by setting

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

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

(iv) set $\boldsymbol{\beta}_{c(m)} = \boldsymbol{\beta}_{c(m-1)} + \boldsymbol{\beta}_c^{F1}$ (vector of parameters associated with candidate predictors).

In the aforementioned algorithm, in step 2A, we are dealing with obligatory/mandatory predictors (which are always part of the model), and o (for obligatory) is used as index with parameter vector. In step 2B, we are dealing with candidate predictors, each of which is a candidate to be a part of the model, so we are using c (for candidate) as index with parameter vector $\boldsymbol{\beta}$. The boosting algorithm uses ridge penalty to obtain the weak learners. A weak learner means slow learning, which is obtained in this algorithm by ridging. Because different candidate predictors may have different parameters associated with them, so the ridge penalty is adjusted for degrees of freedom by multiplying λ with $\sqrt{\mathrm{df}_j}$. The penalty term is rescaled using $\sqrt{\mathrm{df}_j}$ with respect to the dimensionality of the parameter vector associated with the j-th predictor. Such rescaling ensures that the penalty term is of the order of the number of parameters df_j. The intuition of such rescaling is taken from Meier *et al.* (2008). 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 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

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

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 that 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

$$\frac{\frac{1}{K_i}\sum_{j=1}^{K_i}\left|\hat{\gamma}_{ij}\right|}{\sum_{i=1}^{p}\frac{1}{K_i}\sum_{j=1}^{K_i}\left|\hat{\gamma}_{ij}\right|} < \frac{1}{p}. \tag{7}$$

The use of criterion (7) is heuristic, and the motivation comes from our experience. It avoids that a variable with a tiny parameter is included. It would not collect the tiny parameters linked to one predictor, which is the main advantage of the proposed method. The degrees of freedom df_m used in the criterion AIC or BIC are 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 m-th boosting iteration, an approximate hat matrix for which* $\hat{\boldsymbol{\mu}}_m \approx \mathbf{H}_m\mathbf{y}$ *is given by*

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

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

Proof. We are using the asymmetric hat matrix, which is also discussed by Goeman & Le Cessie (2006). Let the predictor variable $V_j = V_{\text{best}}$ be selected after m boosting iterations, and for POM, we have $\mathbf{D}_m = \mathbf{D}\,(\hat{\boldsymbol{\eta}}_m) = \frac{\partial h(\boldsymbol{\eta}_m)}{\partial \boldsymbol{\eta}_m}$ and $\mathbf{W}_m = \mathbf{W}_m\,(\hat{\boldsymbol{\eta}}_m) = \mathbf{D}_m\,(\hat{\boldsymbol{\eta}}_m)\,\boldsymbol{\Sigma}_m^{-1}\mathbf{D}_m^T\,(\hat{\boldsymbol{\eta}}_m)$, with $\boldsymbol{\Sigma}_m$ as the covariance matrix at m-th boosting iteration. By using the Taylor approximation of first order, that is, $h\,(\hat{\boldsymbol{\eta}}) \approx h(\boldsymbol{\eta}) + (\partial h(\boldsymbol{\eta})/\partial\boldsymbol{\eta}^T)\,(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})$, we obtain

$$\begin{aligned}\hat{\boldsymbol{\mu}}_m - \hat{\boldsymbol{\mu}}_{m-1} &\approx \mathbf{D}_m\,(\hat{\boldsymbol{\eta}}_m - \hat{\boldsymbol{\eta}}_{m-1})\\ &= \mathbf{D}_m\mathbf{X}_j\,\hat{\boldsymbol{\beta}}^{F1}\\ \mathbf{D}_m^{-1}\,(\hat{\boldsymbol{\mu}}_m - \hat{\boldsymbol{\mu}}_{m-1}) &\approx \mathbf{X}_j\left(\mathbf{X}_j^T\mathbf{W}_m\mathbf{X}_j + \nu\,\mathbf{P}\right)^{-1}\mathbf{X}_j^T\,\mathbf{W}_m\,\mathbf{D}_m^{-1}\,(\mathbf{y} - \hat{\boldsymbol{\mu}}_{m-1})\,.\end{aligned}$$

Pre-multiplying with $\mathbf{W}_m$, we obtain

$$\mathbf{W}_m\,\mathbf{D}_m^{-1}\,(\hat{\boldsymbol{\mu}}_m - \hat{\boldsymbol{\mu}}_{m-1}) \approx \mathbf{W}_m\,\mathbf{X}_j\left(\mathbf{X}_j^T\mathbf{W}_m\mathbf{X}_j + \nu\,\mathbf{P}\right)^{-1}\mathbf{X}_j^T\,\mathbf{W}_m\,\mathbf{D}_m^{-1}\,(\mathbf{y} - \hat{\boldsymbol{\mu}}_{m-1})$$

where $\mathbf{W}_m \mathbf{X}_j \left(\mathbf{X}_j^T \mathbf{W}_m \mathbf{X}_j + \nu \mathbf{P}\right)^{-1} \mathbf{X}_j^T$ is the usual hat matrix for ridge regression for multicategory response models. We now have

$$\hat{\boldsymbol{\mu}}_m - \hat{\boldsymbol{\mu}}_{m-1} \approx \mathbf{M}_m \left(\mathbf{y} - \hat{\boldsymbol{\mu}}_{m-1}\right),$$

where

$$\mathbf{M}_m = \mathbf{D}_m \mathbf{W}_m^{-1} \tilde{\mathbf{H}}_m \mathbf{W} \mathbf{D}^{\mathbf{T}^{-1}}$$

with

$$\tilde{\mathbf{H}}_m = \mathbf{W}_m \mathbf{X}_j \left(\mathbf{X}_j^T \mathbf{W}_m \mathbf{X}_j + \nu \mathbf{P}\right)^{-1} \mathbf{X}_j^T.$$

The matrix $\mathbf{M}_m$ yields approximation

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_m &\approx \hat{\boldsymbol{\mu}}_{m-1} + \mathbf{M}_m \left(\mathbf{y} - \hat{\boldsymbol{\mu}}_{m-1}\right) \\
&= \hat{\boldsymbol{\mu}}_{m-1} + \mathbf{M}_m \left[\left(\mathbf{y} - \hat{\boldsymbol{\mu}}_{m-2}\right) - \left(\hat{\boldsymbol{\mu}}_{m-1} - \hat{\boldsymbol{\mu}}_{m-2}\right)\right] \\
&\approx \hat{\boldsymbol{\mu}}_{m-1} + \mathbf{M}_m \left[\left(\mathbf{y} - \hat{\boldsymbol{\mu}}_{m-2}\right) - \mathbf{M}_{m-1} \left(\mathbf{y} - \hat{\boldsymbol{\mu}}_{m-2}\right)\right].
\end{aligned}$$

Because for the intercept model, approximation $\hat{\boldsymbol{\mu}}_0 = \mathbf{M}_0 \mathbf{y}$ holds, one obtains

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_1 &\approx \hat{\boldsymbol{\mu}}_0 + \mathbf{M}_1 \left(\mathbf{y} - \hat{\boldsymbol{\mu}}_0\right) \\
&\approx \mathbf{M}_0 \mathbf{y} + \mathbf{M}_1 (\mathbf{I} - \mathbf{M}_0) \mathbf{y}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_2 &\approx \hat{\boldsymbol{\mu}}_1 + \mathbf{M}_2 \left(\mathbf{y} - \hat{\boldsymbol{\mu}}_1\right) \\
&\approx \left\{\mathbf{M}_0 + \mathbf{M}_1 (\mathbf{I} - \mathbf{M}_0)\right\} \mathbf{y} + \mathbf{M}_2 \left\{\mathbf{I} - \mathbf{M}_0 - \mathbf{M}_1 (\mathbf{I} - \mathbf{M}_0)\right\} \mathbf{y} \\
&= \left\{\mathbf{M}_0 + \mathbf{M}_1 (\mathbf{I} - \mathbf{M}_0) + \mathbf{M}_2 (\mathbf{I} - \mathbf{M}_0)(\mathbf{I} - \mathbf{M}_1)\right\} \mathbf{y} \\
&= \left\{\sum_{j=0}^{2} \mathbf{M}_j \prod_{i=0}^{j-1} (\mathbf{I} - \mathbf{M}_i)\right\} \mathbf{y}.
\end{aligned}$$

Further, in recursive manner, we obtain

$$\hat{\boldsymbol{\mu}}_m \approx \mathbf{H}_m \mathbf{y},$$

with

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

For more details on score function and Fisher matrix, Taylor approximation and derivation of approximate hat matrix in likelihood-based boosting, see Tutz & Groll (2010).

3 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 having non-zero parameter values. For example, in the following simulation study,

we are generating situations, for example, where maximum 10% (out of 50 predictors) and, in some settings, even less than 2% (out of 500 predictors) predictors are relevant and should stay in the final model selected through our proposed variable selection procedure. 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 $\rho^{|j-l|}$ instead of having same correlation pattern among all predictors. Here, rather than taking independent predictors (which is hardly the case in real life), we are using some correlation between them. For example, in our case, if we have 20 continuous predictors and let $\rho = 0.8$ be used, then highest correlation between any two adjacent predictors is 0.8, and minimum correlation is between first predictor and last predictor, which is $0.8^{|20-1|} = 0.014$. In our simulation study, we generate the data with k response categories for 10 different settings with different sample sizes and values of ρ. The description of these 10 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 represent the number of informative continuous/binary predictors having non-zero parameters. 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. For an illustration, consider the abovementioned setting 1, where we have 50 predictors in the predictor space, but in the true model (POM), only five are set as informative with non-zero parameter values, and the rest of the 45 predictors are non-informative having zero parameter values. In simulation study, we investigate the effectiveness of *pomBoost* algorithm in identifying these five informative predictors from a predictor space having 50 predictors. To investigate the performance of the algorithm with categorical predictors, in the 11th setting, for a three-category ordered response model with $n = 100$, nominal and ordinal predictors are generated. We use 80(8) categorical predictors as follows: 20(2) nominal predictors each with three categories, 20(2) nominal predictors each with four categories, 20(2) ordinal predictors each with three categories and 20(2) ordinal predictors each with four categories. The convention 20(2) indicates that two predictors out of 20 are declared informative by setting their associated parameters non-zero and other 18 are non-informative having zero value for all associated parameters. 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 POMs with three and five response categories, respectively. The true parameter vector $\boldsymbol{\beta}^T = \left(\boldsymbol{\gamma}_0^T, \boldsymbol{\gamma}\right)$ is then multiplied with a constant c_{snr}, which is chosen so that the signal-to-noise ratio (SNR) is 3.0. To compute the SNR, the following formula is used.

$$\text{SNR} = \frac{\text{Signal}}{\text{Noise}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k}\left(\pi_{ij} - \bar{y}_{.j}\right)^2}{\sum \text{diag}(\boldsymbol{\Sigma})}.$$

The probabilities π_{ij} are computed for the true parameter vector $\boldsymbol{\beta}$, and $\bar{y}_{.j}$ is the mean of j-th response category ($j = 1, \ldots, k$) in the observed data. The probabilities involved in the covariance matrix $\boldsymbol{\Sigma}$, which is used to compute the noise, are also based on the true parameter vector $\boldsymbol{\beta}$. 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.

3.1 Identification of Informative Predictors

The algorithm *pomBoost* fits the POM 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. During the variable selection procedure, we may encounter two situations as follows: (i) a variable is identified as informative (with non-zero parameter estimates) when its true parameter value is also non-zero, or (ii) a variable is identified as informative (with non-zero parameter estimate) when the variable was actually set as non-informative with zero value for the corresponding parameter. To check the effectiveness of a variable selection technique, the 'hit rates' and 'false alarm rates' are used to measure the proportion of variables in both said situations. The hit rate, which is the proportion of correctly identified informative predictors, also known as sensitivity, is given as

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

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

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

One minus the false alarm rate is also known as specificity. 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. We are using deviance, AIC and BIC for predictor selection in a particular boosting iteration. With all these three criteria for predictor selection, hit rates and false alarm rates are computed and are given in Table 1. If the value of hit rate is 1, it means that algorithm *pomBoost* is selecting (with non-zero estimates) all those predictors that were declared as informative with non-zero parameter value and all informative variables are then part of the finally selected model. The zero value for the false alarm rate is an indication that no irrelevant variable is

Table 1. *Hit rates and false alarm rates 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

HR, hit rates; FAR, false alarm rates; AIC, Akaike information criterion; BIC, Bayesian information criterion.

becoming a part of the model and this is ideally desired for the final selected model. As the value of false alarm rate moves away from zero, it indicates that model contains more non-informative predictors. The hit rates with *pomBoost* given in Table 1 are close to 1 and, in some cases, even exactly 1 (means that none of the informative predictors is left), indicating that the algorithm is performing very good regarding the identification of relevant predictors. However, the value of hit rate less than 1 is an indication that along with informative predictors, algorithm is selecting some non-informative predictors also. Because our focus is on predictor selection, not the parameter selection, setting 11 is an interesting case where we have only categorical predictors and 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 results given in Table 1 suggest that deviance may be our choice if main concern is hit rates during the predictor selection, and if false alarm rate is the main focus, then BIC seems to be a good choice with AIC as its strong competitor. But AIC seems to be more appropriate choice as a predictor selection criterion while considering both of the factors, that is, selection of all relevant predictors with minimum possible irrelevant predictors.

3.2 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 & Ramzan, 2012). The usual maximum-likelihood estimates are not existing in all considered high-dimensional settings. For comparison, we are using three different measures: MSE of the parameter estimates $\hat{\boldsymbol{\beta}}$, mean deviance for the fit (deviance $(\hat{\boldsymbol{\pi}})$) and mean prediction error (MPE). The MSE $\left(\hat{\boldsymbol{\beta}}\right)$ is computed using the formula $\frac{1}{S}\sum_s \left\|\hat{\boldsymbol{\beta}}_s^{\text{method}} - \boldsymbol{\beta}^{\text{true}}\right\|^2$, 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 MPE, we generate a new test data set of size $n = 1000$ observations with the same parameters as in

Table 2. *Comparison of boosting approach and ridge regression in terms of MSE$\left(\hat{\boldsymbol{\beta}}\right)$, deviance$(\hat{\boldsymbol{\pi}})$ and mean prediction error.*

	MSE$\left(\hat{\boldsymbol{\beta}}\right)$				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

MSE, mean squared error; MPE, mean prediction error; AIC, Akaike information criterion; BIC, Bayesian information criterion.

International Statistical Review (2013), **81**, 3, 388–406

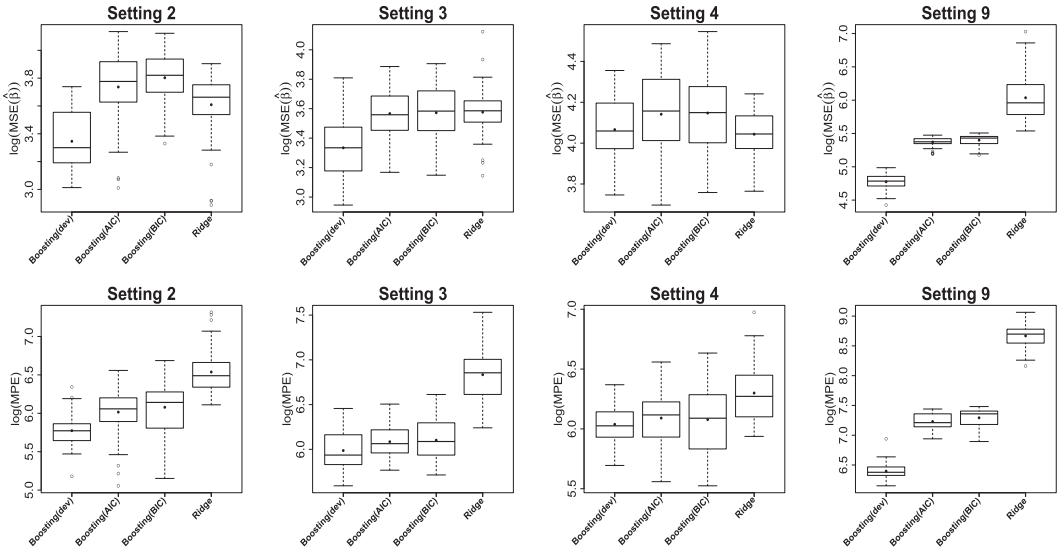

Figure 1. *Illustration of the simulation study: box plots for comparing boosting (with deviance, Akaike information criterion (AIC) and Bayesian information criterion (BIC) as predictor selection criteria) with ridge estimates in terms of* $\log\left(MSE\left(\hat{\boldsymbol{\beta}}\right)\right)$ *(top panel) and mean prediction error, that is, log(MPE) (bottom panel). The solid circles within the boxes represent the mean of the observations for which box plots are drawn.*

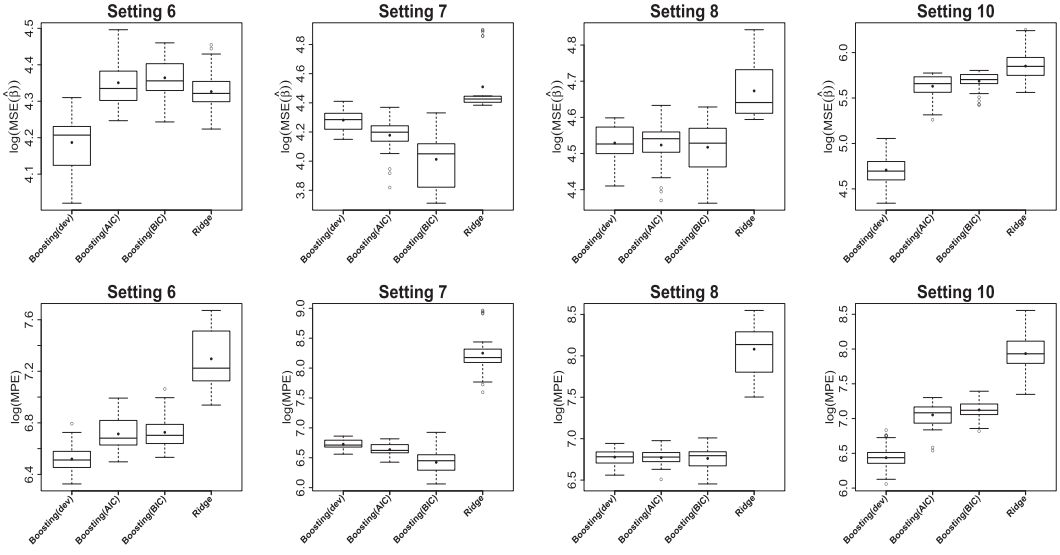

Figure 2. *Illustration of the simulation study for high-dimensional settings (settings 6, 7 and 8) and setting 10 with five response categories model: box plots for comparing boosting (with deviance, Akaike information criterion (AIC) and Bayesian information criterion (BIC) as predictor selection criteria) with ridge estimates in terms of* $\log\left(MSE\left(\hat{\boldsymbol{\beta}}\right)\right)$ *(top panel) and mean prediction error, that is, log(MPE) (bottom panel). The solid circles within the boxes represent the mean of the observations for which box plots are drawn.*

simulation study for each setting. The MPE based on the deviance measure for this test data set is computed as $\text{MPE} = \frac{1}{S}\sum_s D_s = \frac{1}{S}\sum_s 2.\left[\sum_{i=1}^{n}\sum_{j=1}^{k}\pi_{ijs}^{\text{test}}\log\left(\frac{\pi_{ijs}^{\text{test}}}{\hat{\pi}_{ijs}^{\text{test}}}\right)\right]$. 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 of boosting approach are better than those of the ridge regression with an exception for setting 1 where the boosting is showing some weak results in terms of MSE $\left(\hat{\boldsymbol{\beta}}\right)$ 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 settings 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 Figures 1 and 2. In both figures, the box plots associated with five-category response models (settings 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.

4 Application

In this section, we use the *pomBoost* algorithm to analyze gene expression data from a study on prostate cancer (Singh *et al.*, 2002). Based on microarray experiments, the expression levels of 12 600 genes are compared with the Gleason score, which evaluates how effectively cancer cells are able to structure themselves. A Gleason score between 2 and 4 means well-differentiated cells, a score between 5 and 6 describes intermediate differentiation, a score of 7 is intermediate to badly differentiated, and a score between 8 and 10 means badly or undifferentiated tumors. The study included 52 patients, 26 with score 6 and 20 with score 7. The six patients with score larger than 7 were merged to form the top level. In the original analysis of Singh *et al.* (2002), the Gleason score was treated as a metric variable. In Chu *et al.* (2005), it was pointed out that the Gleason score is clearly an ordinal variable, and thus, the problem of predicting the score from gene expression data is a typical problem of ordinal regression. Although the focus of Chu *et al.* (2005) was on prediction, whereas Singh *et al.* (2002) was applying a multiple testing approach, it should be noted that there is not a single gene that was selected both by Singh *et al.* (2002) and by Chu *et al.* (2005). Frommlet (2010) fitted ordinal response models but used univariate models for each single gene.

When using *pomBoost* algorithm, we are considering 244 genes (predictors) out of 12 600 genes. The so-called preselection is based on three criteria: (i) 218 genes with p-value < 0.05 based on simple F-test (which means significant marginals without any correction for multiple testing); (ii) 15 genes reported by Chu *et al.* (2005), which were not already selected in (i); and (iii) 11 genes reported in the original work of Singh *et al.* (2002), which were not already selected in (i). The data of 244 predictors and the response variable being used in this section can be accessed at http://shahla.userweb.mwn.de/pomBoost-GenesData.zip.

The AIC and BIC are used as criteria for the selection of a gene in a boosting iteration. So at each boosting iteration, that gene (predictor), out of total 244 genes, is selected/updated, which has minimum AIC (BIC is also used for selection of a predictor and updating the associated parameters at each boosting iteration). The deviance is then used as stopping criterion; that is, to stop the boosting iterations and to obtain convergence, we use deviance measure with 10-fold cross-validation. To limit the boosting iterations between 200 and 300, the value of ridge

Table 3. *Genes indices and their estimates selected with pomBoost algorithm, when AIC (with $\lambda = 90$) and BIC (with $\lambda = 50$) are used for selecting/updating a predictor in a boosting iteration. Deviance is used as stopping criterion with* 10*-fold cross-validation.*

AIC		BIC	
Gene index	Estimate	Gene index	Estimate
15	0.044	15	0.058
19	−0.090	19	−0.131
34	−0.215	22	0.112
45	0.039	34	−0.474
52	−0.279	52	−0.171
58	−0.086	58	−0.086
64	0.209	64	0.263
75	−0.970	75	−1.110
136	0.976	88	−0.406
140	0.079	95	−0.038
167	−0.208	97	0.233
174	0.188	136	1.117
183	−0.858	140	0.289
188	0.086	167	−0.335
215	0.111	174	0.264
223	0.049	176	0.118
225	0.131	183	−0.855
228	0.053	215	0.156
233	0.267	222	−0.040
238	0.050	233	0.333
242	0.084	242	0.096
243	0.060	243	0.177

AIC, Akaike information criterion; BIC, Bayesian information criterion.

penalty for regularization is used as $\lambda = 90$ and $\lambda = 50$ for AIC and BIC, respectively. The parameter estimates of genes selected using *pomBoost* algorithm are given in Table 3. The deviance, as a stopping criterion, suggests 251 and 221 iterations for AIC and BIC as predictor selection criteria, respectively. The resulting parameter estimates for the selected genes are given in Table 3. With AIC and BIC as predictor selection criteria, 22 genes out of 244 are

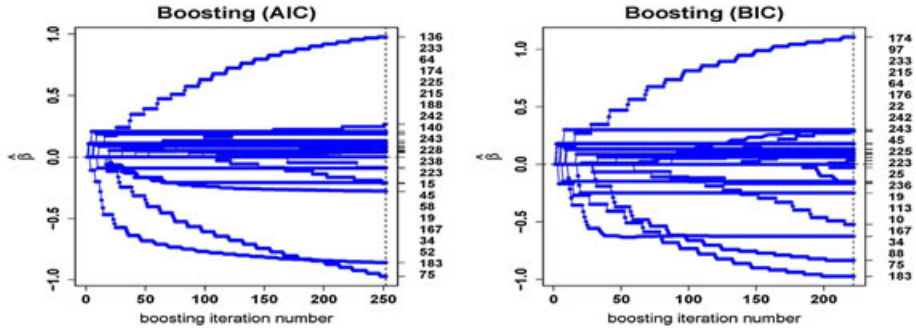

Figure 3. *Coefficients build-up with componentwise boosting for genes data. Deviance with* 10*-fold cross-validation is used as stopping criterion. The optimal results are obtained at* 251 *and* 221 *iterations for Akaike information criterion (AIC; left panel) and Bayesian information criterion (BIC; right panel) as predictor selection criteria, respectively. The indices of selected genes are given on the right side of each graph.*

selected. In both cases, 15 genes (with genes indices 15, 19, 34, 52, 58, 64, 75, 136, 140, 167, 174, 183, 233, 242, 243) are the same while other seven are different (with AIC, they have indices 45, 188, 215, 223, 225, 228, 238, and with BIC, these are with indices 22, 88, 95, 97, 176, 215, 222). The data set is mainly used to illustrate that the method works in high dimensions. Selection results can hardly be compared to the univariate procedures that have been used for the Gleason data before. The coefficients build-up for *pomBoost* are given in Figure 3 where the indices of selected genes are given on the right side of each graph.

5 Concluding Remarks

In regression, POMs 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 modelling bias. But, in particular when many predictors are available, one wants to know which variables are influential. 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 been somewhat neglected in the literature. Although Lu & Zhang (2007) refer to variable selection for POMs, they discussed variable selection in survival analysis. The proposed algorithm is an effort to fill this gap. The proposed boosting technique fits POM by implicitly selecting the relevant predictors. Unlike multinomial logit models that have category specific estimates, POMs have so called global parameters. But in the case of a categorical predictor, more than one parameter is 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 cases, 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 POM, only the procedure is easily extended to any ordered regression model.

It should be noted that not only alternative regularization methods like L_1-type penalization (Tibshirani, 1996; Yuan & Lin, 2006; Meier *et al.*, 2008) but also methods like the Dantzig selector (James & Radchenko, 2009) or smoothly clipped absolute deviation (Fan & Li, 2001) have been successfully used to select variables. But they are designed for univariate response models and seem to have not yet been extended to the case of ordinal responses. The same holds for the methods to handle categorical predictors mentioned in Section 2.2. One advantage of boosting methods over penalization methods is the easy handling of different forms of predictors by using appropriate weak learners. When using L_1-type penalization, one has to construct complex penalty terms that account for the different forms of available predictors. They have to include not only methods for nominal categorical predictors as in the group lasso (Yuan & Lin, 2006) but also terms for ordered categorical variables as in Tutz & Gertheiss (2013) and terms for metric predictors. In boosting approaches, as proposed here for ordinal responses, it is easy to use differently structured learners that account for the scale level of the predictor. On the other hand, lasso type estimators have the advantage that the form of the regularization is given by an explicit penalty term. While boosting is an algorithm-based regularization method, the penalty term in the lasso makes the constraints used in estimation distinct. In special cases,

as in the linear model, specific versions of boosting called forward stage-wise regression with infinitesimally small step sizes produce solutions that are approximately equivalent to the solutions obtained by lasso (Efron *et al.*, 2004). It is to be expected that also for the ordinal model, if tuning parameters are chosen appropriately, solutions will be similar. However, in the general case, with a mixture of metric, categorical nominal and categorical ordinal predictors, specific software would yet have to be developed.

Acknowledgements

The authors thank the three referees and the associate editor for their valuable comments and suggestions that helped to improve the paper.

References

Agresti, A. (1999). Modelling ordered categorical data: Recent advances and future challenges. *Stat. Med.*, **18**, 2191–2207.

Agresti, A. (2013). *Categorical Data Analysis*, 3rd ed. New York: Wiley.

Aitchison, J. & Aitken, C.G.G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*, **63**, 413–420.

Aitken, C.G.G. (1983). Kernel methods for the estimation of discrete distributions. *J. Stat. Comput. Simul.*, **16**, 189–200.

Ananth, C.V. & Kleinbaum, D.G. (1997). Regression models for ordinal responses: A review of methods and applications. *Int. J. Epidemiol.*, **26**(6), 1323–1333.

Archer, K. & Williams, A. (2012). L1 penalized continuation ratio models for ordinal response prediction using high-dimensional datasets. *Stat. Med.*, **31**(14), 1464–1474.

Armstrong, B. & Sloan, M. (1989). Ordinal regression models for epidemiologic data. *Amer. J. Epidemiol.*, **129**, 191–204.

Bowman, A.W. (1980). A note on consistency of the kernel method for the analysis of categorical data. *Biometrika*, **67**, 682–684.

Bühlmann, P. (2006). Boosting for high-dimensional linear models. *Ann. Statist.*, **34**, 559–583.

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

Bühlmann, P. & Yu, B. (2003). Boosting with l2 loss: Regression and classification. *J. Amer. Statist. Assoc.*, **98**, 324–339.

Chu, W., Ghahramani, Z., Falciani, F. & Wild, D.L. (2005). Biomarker discovery in microarray gene expression data with Gaussian processes. *Bioinformatics*, **21**, 3385–3393.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least angle regression. *Ann. Statist.*, **32**, 407–499.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, **96**, 1348–1360.

Freund, Y. & Schapire, R.E. (1996). Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference*, Bari, Italy, pp. 148–156.

Freund, Y. & Schapire, R.E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.*, **55**, 119–139.

Friedman, J., Hastie, T. & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1–22.

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

Frommlet, F. (2010). Some properties of a recently introduced approach to ordinal regression. *Austral. J. Statist.*, **39**, 182–202.

Genter, F.C. & Farewell, V.T. (1985). Goodness-of-link testing in ordinal regression models. *Canad. J. Statist.*, **13**, 37–44.

Gertheiss, J., Hogger, S., Oberhauser, C. & Tutz, G. (2011). Selection of ordinally scaled independent variables with applications to international classification of functioning core sets. *J. R. Stat. Soc. Ser. C.*, **60**(3), 377–395.

Gertheiss, J. & Oehrlein, F. (2011). Testing linearity and relevance of ordinal predictors. *Electron. J. Stat.*, **5**, 1935–1959.

Gertheiss, J. & Tutz, G. (2009). Penalized regression with ordinal predictors. *Int. Stat. Rev.*, **77**, 345–365.
Gertheiss, J. & Tutz, G. (2010). Sparse modeling of categorial explanatory variables. *Ann. Appl. Stat.*, **4**, 2150–2180.
Goeman, J.J. & Le Cessie, S. (2006). A goodness-of-fit test for multinomial logistic regression. *Biometrics*, **62**, 980–985.
Hastie, T., Tibshirani, R. & Friedman, J.H. (2009). *The Elements of Statistical Learning*, 2nd ed. New York: Springer-Verlag.
Hoerl, A.E. & Kennard, R.W. (1970). Ridge regression: Bias estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.
James, G. & Radchenko, P. (2009). A generalized Dantzig selector with shrinkage tuning. *Biometrika*, **96**, 323–337.
Lu, W. & Zhang, H.H. (2007). Variable selection for proportional odds model. *Stat. Med.*, **26**, 3771–3781.
McCullagh, P. (1980). Regression model for ordinal data (with discussion). *J. R. Stat. Soc. Ser. B*, **42**, 109–142.
Meier, L., van de Geer, S. & Bühlmann, P. (2008). The group lasso for logistic regression. *J. R. Stat. Soc.*, **70**, 53–71.
Nyquist, H. (1991). Restricted estimation of generalized linear models. *Appl. Stat.*, **40**, 133–141.
Schapire, R.E. (1990). The strength of weak learnability. *Mach. Learn.*, **5**, 197–227.
Simonoff, J.S. (1983). A penalty function approach to smoothing large sparse contingency tables. *Ann. Statist.*, **11**, 208–218.
Simonoff, J.S. (1987). Probability estimation via smoothing in sparse contingency tables with ordered categories. *Statist. Probab. Lett.*, **5**, 55–63.
Simonoff, J.S. (1995). Smoothing categorical data. *J. Statist. Plann. Inf.*, **47**, 41–69.
Simonoff, J.S. & Tutz, G. (2000). Smoothing methods for discrete data. In *Smoothing and Regression. Approaches, Computation and Application*, Ed. M. Schimek, pp. 193–228. New York: Wiley.
Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P., Reshaw, A.D., Amico, A., Richie, J., Lander, E., Loda, M., Kantoff, P., Golub, T. & Sellers, W. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203–209.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B*, **58**, 267–288.
Tutz, G. (1988). Smoothing for discrete kernels in discrimination. *Biom. J.*, **6**, 729–739.
Tutz, G. (2012). *Regression for Categorical Data.* New York: Cambridge University Press.
Tutz, G. & Binder, H. (2006). Generalized additive modelling with implicit variable selection by likelihood based boosting. *Biometrics*, **62**, 961–971.
Tutz, G. & Gertheiss, J. (2013). Rating scales as predictors – the old question of scale level and some answers. *Psychometrika*, DOI: 10.1007/S11336-013-9343-3.
Tutz, G. & Groll, A. (2010). Generalized linear mixed models based on boosting. In *Statistical Modelling and Regression Structure*, pp. 197–215. Berlin, Heidelberg: Physica-Verlag.
Yuan, M. & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. B*, **68**, 49–67.
Zahid, F.M. & Ramzan, S. (2012). Ordinal ridge regression with categorical predictors. *J. Appl. Stat.*, **39**, 161–171.
Zahid, F.M. & Tutz, G. (2013). Multinomial logit models with implicit variable selection. *Adv. Data Anal. Classif.,* DOI 10.1007/s11634-013-0136-4.

Résumé

Le modèle des *odds proportionnels* (*rapports des chances* proportionnels) est le modèle le plus couramment utilisé dans l'analyse de réponses de type ordinal. En présence d'un grand nombre de covariables possibles, l'approche par maximum de vraisemblance usuelle est typiquement mise en échec si toutes les covariables sont prises en compte. Une méthode de type *boosting*, *pomBoost*, est proposée, par laquelle le modèle est estimé via une sélection implicite des prédicteurs les plus pertinentes. Cette approche fait la distinction entre variables *métriques* et *catégorielles*. Dans le cas de variables catégorielles, l'objectif est une sélection simultanée d'un ensemble de prédicteurs. La méthode fait la distinction, de surcroît, entre variables nominales et ordinales. Dans ce dernier cas, la relation d'ordre intervient dans le calcul de la pénalisation. La méthode permet également d'imposer la présence de certaines covariables dans le modèle final. Les performances de l'algorithme de *boosting* sont évaluées, du point de vue de l'erreur quadratique moyenne et de l'erreur de prédiction, au moyen d'une étude de simulation et d'applications à des données empiriques. Les taux de succès et de fausse alarme sont considérés pour l'évaluation des performances de *pomBoost* dans la sélection des prédicteurs.

[Received April 2012, accepted July 2013]