LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Sebastian Petry & Gerhard Tutz

The OSCAR for Generalized Linear Models

Technical Report Number 112, 2011
Department of Statistics
University of Munich

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

The OSCAR for Generalized Linear Models

Sebastian Petry & Gerhard Tutz
Ludwig-Maximilians-Universität München
Akademiestraße 1, 80799 München
{petry, tutz}@stat.uni-muenchen.de

August 8, 2011

Abstract

The Octagonal Selection and Clustering Algorithm in Regression (OSCAR) proposed by Bondell and Reich (2008) has the attractive feature that highly correlated predictors can obtain exactly the same coefficient yielding clustering of predictors. Estimation methods are available for linear regression models. It is shown how the OSCAR penalty can be used within the framework of generalized linear models. An algorithm that solves the corresponding maximization problem is given. The estimation method is investigated in a simulation study and the usefulness is demonstrated by an example from water engineering.

Keywords: Variable Selection, Clustering, OSCAR, LASSO, Generalized Linear Models.

1 Introduction

Within the last decades various regularization techniques for generalized linear models (GLMs) have been developed. Most methods aim at stabilizing estimates and finding simpler models. In particular variable selection has been a major topic. One of the oldest methods is ridge regression, which has been proposed by Hoerl and Kennard (1970). In ridge regression the parameter space is restricted to a p-sphere around the origin $\sum_{j=1}^{p} \beta_j^2 \leq t$, $t \geq 0$. Another popular shrinkage methods is the LASSO for **L**east **A**bsolute **S**hrinkage and **S**election **O**perator (Tibshirani, 1996), where the parameter space is restricted to a p-crosspolytope $\sum_{j=1}^{p} |\beta_j| \leq t$, $t \geq 0$. The restriction induces shrinkage and variables selection. In general, restricted parameter spaces are called penalty regions. For many penalty regions the problem can be transformed into a penalized likelihood problem by adding a penalty term to the log-likelihood. For ridge regression the penalty term is $\lambda \sum_{j=1}^{p} \beta_j^2$ and for the LASSO it is $\lambda \sum_{j=1}^{p} |\beta_j|$, with $\lambda \geq 0$

in both cases. A combination of the ridge and the LASSO uses $\lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j|$. It is well known as the elastic net (Zou and Hastie, 2005).

Zou and Hastie (2005) showed that variable selection leads to unsatisfying results in the case of multicollinearity, that is, if some of the covariates are highly correlated. Then procedures like the LASSO tend to include only a few covariates from a group of the highly correlated covariates. They show that for the elastic net a grouping property holds, which means that the estimated parameters of highly correlated covariates are similar up to sign. An alternative penalty region that enforces grouping of variables was proposed by Bondell and Reich (2008) under the name OSCAR for **O**ctagonal **S**election and **C**lustering **A**lgorithm in **R**egression. For LASSO and the elastic net (EN) several methods have been proposed to solve the penalized log-likelihood problem in generalized linear models (GLMs); (see Park and Hastie, 2007b; Goeman, 2010a; Friedman et al., 2010). For OSCAR it seems that algorithms are available only for the linear model. In the following estimation methods for OSCAR are proposed that work within the more general GLM framework.

In Section 2 we give a short overview on GLMs. In Section 3 the OSCAR penalty region is discussed. In Section 4 we use the results of Section 3 and present an algorithm for estimating the corresponding restricted regression problem based on the active set method. A simulation study is presented in Section 5, which uses settings that are similar to the settings used by Bondell and Reich (2008). A real data example with water engineering background is given in Section 6.

2 Generalized Linear Models

We consider data $(\boldsymbol{y}, \boldsymbol{X})$ where $\boldsymbol{y} = (y_1, ..., y_n)^T$ is the response and $\boldsymbol{X}$ is the $(n \times p)$ matrix of explanatory variables that contains n observations $\boldsymbol{x}_i^T = (x_{i1}, ..., x_{ip})$, $i = 1, ..., n$. In GLMs (McCullagh and Nelder, 1983) it is assumed that the distribution of $y_i|\boldsymbol{x}_i$ is from a simple exponential family

$$f(y_i|\theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right\}, \tag{1}$$

where θ_i is the natural parameter and ϕ is a dispersion parameter; $b(.)$ and $c(.)$ are specific functions corresponding to the type of the family. In addition, it is assumed that the observations are (conditionally) independent. For given data the conditional expectation of $y_i|\boldsymbol{x}_i$, $\mu_i = E(y_i|\boldsymbol{x}_i)$, is modeled by

$$g(\mu_i) = \eta_i \quad \text{or} \quad \mu_i = h(\eta_i),$$

where $\eta_i = \beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}$ is the linear predictor, $g(.)$ is the link function and $h(.) = g^{-1}(.)$ is the response function. Let $\boldsymbol{\beta}_0 = (\beta_0, \boldsymbol{\beta}^T)^T$ denote the parameter vector that includes the intercept. Then the corresponding design matrix is $\boldsymbol{Z} = (\boldsymbol{1}_n, \boldsymbol{X})$ and the linear predictor is $\boldsymbol{\eta} = \boldsymbol{Z}\boldsymbol{\beta}_0$. The maximum likelihood estimate (MLE) is given by

$$\widehat{\boldsymbol{\beta}_0} = \text{argmax}_{\boldsymbol{\beta}_0}\left\{\sum_{i=1}^{n} l_i(\boldsymbol{\beta}_0)\right\}$$

where $l_i(\boldsymbol{\beta}_0)$ is the likelihood function of the ith observation. The maximum likelihood problem can be iteratively solved by

$$\widehat{\boldsymbol{\beta}}_0^{(l+1)} = \text{argmin}_{\boldsymbol{\beta}_0} \left\{ \boldsymbol{\beta}_0^T \boldsymbol{Z}^T \widehat{\boldsymbol{W}}^{(l)} \boldsymbol{Z} \boldsymbol{\beta}_0 - 2\boldsymbol{\beta}_0^T \boldsymbol{Z}^T \widehat{\boldsymbol{W}}^{(l)} \widetilde{\boldsymbol{y}}^{(l)} \right\}, \tag{2}$$

where

$$\widetilde{\boldsymbol{y}}^{(l)} = \boldsymbol{Z}\widehat{\boldsymbol{\beta}}_0^{(l)} + (\widehat{\boldsymbol{D}}^{(l)})^{-1} \left(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}^{(l)} \right)$$

is the working response vector,

$$\widehat{\boldsymbol{W}}^{(l)} = (\widehat{\boldsymbol{D}}^{(l)})^T (\widehat{\boldsymbol{\Sigma}}^{(l)})^{-1} \widehat{\boldsymbol{D}}^{(l)}$$

is the weight matrix with the derivative matrix of the response function,

$$\widehat{\boldsymbol{D}}^{(l)} = \text{diag} \left\{ \frac{\partial h(\widehat{\eta}_i^{(l)})}{\partial \eta} \right\}_{i=1}^{n},$$

and the matrix of variances

$$\widehat{\boldsymbol{\Sigma}}^{(l)} = \text{diag} \left\{ \phi V(h(\widehat{\eta}_i^{(l)})) \right\}_{i=1}^{n},$$

all of them evaluated at the previous step. $V(.)$ is the variance function, which is determined by the distributional assumption and $\widehat{\boldsymbol{\mu}}^{(l)}$ is the estimated prediction of the previous step. The update is repeated until $\|\widehat{\boldsymbol{\beta}}_0^{(l+1)} - \widehat{\boldsymbol{\beta}}_0^{(l)}\| / \|\widehat{\boldsymbol{\beta}}_0^{(l)}\| < \varepsilon$ for small ε. The re-weighted least square estimates

$$\widehat{\boldsymbol{\beta}}_0^{(l+1)} = \left(\boldsymbol{Z}^T \widehat{\boldsymbol{W}}^{(l)} \boldsymbol{Z} \right)^{-1} \boldsymbol{Z}^T \widehat{\boldsymbol{W}}^{(l)} \widetilde{\boldsymbol{y}}^{(l)}$$

is also known as Fisher scoring. The algorithm we will present uses a constrained Fisher scoring combined with the active set method that uses the specific structure of the OSCAR penalty.

3 The OSCAR Penalty Region

In the following we consider standardized covariates, that is, $\sum_{i=1}^n x_{ij} = 0$ and $(n-1)^{-1} \sum_{i=1}^n x_{ij}^2 = 1$. When Bondell and Reich (2008) introduced the OSCAR for the normal linear regression they also centered the responses by using $\sum_{i=1}^n y_i = 0$. If all covariates and the response are centered no intercept has to be estimated. Then the OSCAR can be given as the constrained least-squares problem

$$\widehat{\boldsymbol{\beta}} = \text{argmax} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2, \text{ s.t. } \boldsymbol{\beta} \in \mathcal{O}_{c,t}(\boldsymbol{\beta}) \right\}, \tag{3}$$

with OSCAR penalty region given by

$$\mathcal{O}_{c,t}(\boldsymbol{\beta}) = \left\{ \boldsymbol{\beta} : \sum_{j=1}^p |\beta_j| + c \sum_{1 \le j < k \le p} \max \{|\beta_j|, |\beta_k|\} \le t \right\}. \tag{4}$$

The first sum $\sum_{j=1}^{p} |\beta_j|$ is the LASSO penalty which induces variable selection. The second sum $c \sum_{1 \leq j < k \leq p} \max\{|\beta_j|, |\beta_k|\}$ accounts for clustering of similar variables. With $c \geq 0$ and $t > 0$ an equivalent form of the OSCAR penalty (4) is

$$\mathcal{O}_{c,t}(\boldsymbol{\beta}) = \left\{ \boldsymbol{\beta} : \sum_{j=1}^{p} \{c(j-1)+1\} |\beta_{(j)}| \leq t \right\}, \tag{5}$$

where $|\beta_{(1)}| \leq |\beta_{(2)}| \leq ... \leq |\beta_{(p)}|$ and $|\beta_{(j)}|$ denotes the jth largest component of $|\boldsymbol{\beta}| = (|\beta_1|, ..., |\beta_p|)^T$. The parameter c controls the clustering and t the amount of shrinkage. Bondell and Reich (2008) gave a `MatLab`-code at `http://www4.stat.ncsu.edu/~bondell/software.html` which solves the least square problem under constraints

$$\mathcal{O}_{\alpha,t}(\boldsymbol{\beta}) = \left\{ \boldsymbol{\beta} : (1-\alpha) \sum_{j=1}^{p} |\beta_j| + \alpha \sum_{1 \leq j < k \leq p} \max\{|\beta_j|, |\beta_k|\} \leq t \right\} \tag{6}$$

$$= \left\{ \boldsymbol{\beta} : \sum_{j=1}^{p} \{\alpha(j-1) + (1-\alpha)\} |\beta_{(j)}| \leq t \right\} \tag{7}$$

where $\alpha \in [0, 1]$ and $t > 0$. If $\alpha = 0$, respectively $c = 0$, the OSCAR is equivalent to the LASSO. For appropriate values of c, α and t the penalty regions (4) and (6) are equivalent. In the following we use $\mathcal{O}_{\alpha,t}(\boldsymbol{\beta})$ from (6) and (7).

In contrast to the Elastic Net penalty the OSCAR enforces that parameters obtain the same value. Bondell and Reich (2008) derived a relationship between the clustering of covariates (which obtain the same value) and their correlation. The word octagonal in OSCAR is motivated by the geometry of the penalty region. The projection of the penalty region into each β_i-β_j-plane is an octagon. The octagonal shape accounts for the estimation of identical parameters as well as variable selection because the coordinates of the vertices have a very specific structure. In particular the absolute values of the coordinates of a vertex on the surface are equal or zero. So each convex combination of vertices on the surface describes an area with specific properties. If less than p vertices are convexly combined one obtains variable selection and/or clustering. For illustration, Figure 1 shows an OSCAR penalty region in $\mathbb{R}^3$.

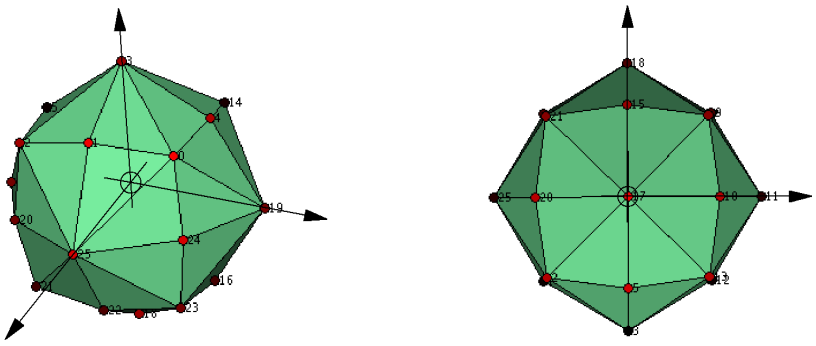

FIGURE 1: *OSCAR penalty region from two different perspectives. On the right it is the projection on a β_i-β_j-plane.*

In Petry and Tutz (2011) it is shown that the OSCAR penalty is the intersection of $2^p \cdot p!$ halfspaces. So $\mathcal{O}_{\alpha,t}(\boldsymbol{\beta})$ can be rewritten into a system of inequations $\boldsymbol{A}\boldsymbol{\beta} \leq \boldsymbol{t}$ where $\boldsymbol{A}$ is the $(2^p \cdot p!) \times p$-dimensional matrix $\boldsymbol{A} = (\boldsymbol{a}_1, ..., \boldsymbol{a}_{2^p \cdot p!})^T$ that contains the normal vectors $\boldsymbol{a}_q$ of each generating hyperplane. Each normal vector $\boldsymbol{a}_q$ is characterized by two attributes:

1. The vector of signs of the components of the normal vector,

$$\text{sign}(\boldsymbol{a}_q) = (\text{sign}(a_{q1}), ..., \text{sign}(a_{qp}))^T$$

   This attribute is induced by the absolute value of the components of $\boldsymbol{\beta}$ (see (6) or (7)).

2. The vector of ranks of the absolute value of the components of the normal vector

$$\text{rank}(|\boldsymbol{a}_q|) = (\text{rank}(|a_{q1}|), .., \text{rank}(|a_{qp}|))^T,$$

   which is a p-dimensional vector. Its jth entry is the position of a_{qj} in the order $|a_{q(1)}| \leq |a_{q(2)}| \leq ... \leq |a_{q(p)}|$ where $|a_{q(j)}|$ denotes the absolute value of the jth largest component of $|(|\boldsymbol{a}_q) = (|a_{q1}|, ..., |a_{qp}|)^T$. This attribute is induced by using the pairwise maximum norm in (6) or the ordered components like in (7) respectively.

Each row of $\boldsymbol{A}$ is given by signs and a permutation of the weights $\boldsymbol{w} = \{(1-\alpha)(j-1) + \alpha : j = 1, ..., p\}$ given in (7). Each half space refers to one constraint of the restricted optimization problem that can be written as

$$\boldsymbol{a}_q = ((1-\alpha) \cdot (\text{rank}(|\boldsymbol{a}_q|) - 1) + \alpha)^T \, \text{diag}(\text{sign}(\boldsymbol{a}_q)) \leq t. \tag{8}$$

Already for small dimensional cases the dimension of $\boldsymbol{A}$ becomes very large, for example, if $p = 5$ the matrix $\boldsymbol{A}$ is 3840×5-dimensional.

4 The g|mOSCAR A|gorithm

For GLMs the least-squares problem (3) turns into the restricted maximum likelihood problem

$$\widehat{\boldsymbol{\beta}_0} = \text{argmax} \left\{ \sum_{i=1}^{n} l_i(\boldsymbol{\beta}_0), \text{ s.t. } \boldsymbol{\beta}_0 \in \mathbb{R} \times \mathcal{O}_{\alpha,t}(\boldsymbol{\beta}) \right\},$$

where $l_i(.)$ is the log-likelihood of a GLM. In contrast to the linear normal regression, where responses are easily centered, now an unrestricted intercept has to be included. The new penalty region is $\mathbb{R} \times \mathcal{O}_{\alpha,t}$, which can be rewritten as an system of inequations

$$(\boldsymbol{0}, \boldsymbol{A})\boldsymbol{\beta}_0 \leq \boldsymbol{t}, \tag{9}$$

where $\boldsymbol{\beta}_0 = (\beta_0, \boldsymbol{\beta}^T)^T$. The region (9) is an unbounded intersection of subspaces called polyhedron. Each row of (9) refers to one constraint. In general a constraint of a system of inequations is called active if the equal sign holds in the corresponding row of the system of inequations. If the equal sign holds the solution lies on the corresponding face

of the polyhedron. Only the active constraints have an influence on the solution. The remaining constraints are fulfilled but have no influence on the solution, and are called inactive constraints. Removing inactive constraints has no influence on the solution of the constrained log-likelihood problem. The solution is unique and each point in $\mathbb{R}^p$ can be represented by the intersection of p hyperplanes of dimension $p-1$. Therefore, the number of constraints can be reduced from $2^p \cdot p!$ to p. Only by numerical reasons sometimes in the algorithm more than p constraints are set active. Because of (8) and (9) for an given parameter vector $\boldsymbol{\beta}_0$ an active constraint from (9) has the following form

$$\boldsymbol{a}(\boldsymbol{\beta}_0)\boldsymbol{\beta}_0 = (0,\, ((1-\alpha)\cdot(\text{rank}(|\boldsymbol{\beta}|)-1)+\alpha)^T \,\text{diag}(\text{sign}(\boldsymbol{\beta})))\boldsymbol{\beta}_0 = t. \tag{10}$$

It is important that rank($|\boldsymbol{\beta}|$) is a p-dimensional vector where all elements of $\{1,\, 2,\, ...,\, p\}$ are used as entries. If some elements of $|\boldsymbol{\beta}|$ are equal the assembly of their ranks is arbitrary.

The following algorithm is an active set method combined with Fisher scoring. There are two parts.

AS (Active Set): This step accounts for the creation of the active set and is indexed by $^{(k)}$.

FS (Fisher Scoring): This step solves the restricted ML problem. It is indexed by $^{(l)}$ in analogy to (2). The constraints are given by the active set that is determined by the AS-step.

First we initialize $k = 0$ and choose an initial value $\widehat{\boldsymbol{\beta}}_0^{(0)}$, for instance, the MLE.

AS-step

We set k to $k+1$. With $\widehat{\boldsymbol{\beta}}_0^{(k-1)}$ we determine $\boldsymbol{a}(\widehat{\boldsymbol{\beta}}_0^{(k-1)}) = \boldsymbol{a}^{(k)}$ as given in (10). The new active constraint $\boldsymbol{a}^{(k)}$ is added as a new row to $(\boldsymbol{0},\, \boldsymbol{A})^{(k-1)}\boldsymbol{\beta}_0 \leq \boldsymbol{t}$ if $\boldsymbol{a}^{(k)}$ is not a row of $(\boldsymbol{0},\, \boldsymbol{A})^{(k-1)}$

$$\begin{pmatrix} (\boldsymbol{0},\, \boldsymbol{A})^{(k-1)} \\ \boldsymbol{a}^{(k)} \end{pmatrix} \boldsymbol{\beta}_0 = (\boldsymbol{0},\, \boldsymbol{A})^{(k)}\boldsymbol{\beta}_0 \leq \boldsymbol{t}. \tag{11}$$

Finally we remove all inactive constraints from $(\boldsymbol{0},\, \boldsymbol{A})^{(k)}\boldsymbol{\beta}_0 \leq \boldsymbol{t}$.

FS-step

We have to solve the constrained ML problem

$$\widehat{\boldsymbol{\beta}}_0^{(k)} = \text{argmin}_{\boldsymbol{\beta}_0} \left\{ -\sum_{i=1}^{n} l_i(\boldsymbol{\beta}_0),\ \text{s.t. } (\boldsymbol{0},\, \boldsymbol{A})^{(k)}\boldsymbol{\beta}_0 \leq \boldsymbol{t} \right\}, \tag{12}$$

which is a combination of the unconstrained least square problem (2) and the penalty region $(\boldsymbol{0},\, \boldsymbol{A})^{(k)}\boldsymbol{\beta}_0 \leq \boldsymbol{t}$ from the AS. For clarity we do not use double indexing. For solving (12) we use the following constrained Fisher scoring

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_0^{(l+1)} \quad = \quad \text{argmin}_{\boldsymbol{\beta}_0} \Big\{ & \boldsymbol{\beta}_0^T \boldsymbol{Z}^T \widehat{\boldsymbol{W}}^{(l)} \boldsymbol{Z}\boldsymbol{\beta}_0 - 2\boldsymbol{\beta}_0^T \boldsymbol{Z}^T \widehat{\boldsymbol{W}}^{(l)} \widetilde{\boldsymbol{y}}^{(l)}, \\ & \text{s.t. } (\boldsymbol{0},\, \boldsymbol{A})^{(k)}\boldsymbol{\beta}_0 \leq \boldsymbol{t} \Big\}. \end{aligned} \tag{13}$$

It is solved iteratively with the `quadprog` package from `R` (see Turlach, 2009). The constrained update (13) is repeated up to convergence $\delta_2 = \|\hat{\boldsymbol{\beta}}_0^{(l)} - \hat{\boldsymbol{\beta}}_0^{(l+1)}\|/\|\hat{\boldsymbol{\beta}}_0^{(l)}\| \leq \varepsilon$, for small ε. After convergence $\widehat{\boldsymbol{\beta}}_0^{(l+1)}$ is the solution of (12). With $\widehat{\boldsymbol{\beta}}_0^{(k)}$ we start the AS-step again.

The AS-step envelops the FS-step. Both loops are repeated until $\delta_1 = \|\hat{\boldsymbol{\beta}}_0^{(k)} - \hat{\boldsymbol{\beta}}_0^{(k+1)}\|/\|\hat{\boldsymbol{\beta}}_0^{(k)}\| \leq \varepsilon$, for small ε.

Algorithm: glmOSCAR

Step 1 (Initialization) Choose $\widehat{\boldsymbol{\beta}}_0^{(0)}$ and set $\delta_1 = \infty$.

Step 2 (Iteration)

AS: While $\delta_1 > \varepsilon$.

- Determine $\boldsymbol{a}^{(k)}$ as described in (10).
- Determine $(\mathbf{0},\, \boldsymbol{A})^{(k)}$ as described in (11) and remove the inactive constraints.

FS: Set $\delta_2 = \infty$.

- Solve $\widehat{\boldsymbol{\beta}}_0^{(k+1)} = \text{argmin}\left\{-\sum_{i=1}^n l_i(\boldsymbol{\beta}_0),\ \text{s.t. } (\mathbf{0},\, \boldsymbol{A})^{(k)}\boldsymbol{\beta}_0 \leq \boldsymbol{t}\right\}$ using a constrained Fisher scoring from (13) up to convergence $\delta_2 < \varepsilon$.
- After converging the constrained Fisher scoring (13) compute $\delta_1 = \frac{\|\hat{\boldsymbol{\beta}}_0^{(k)} - \hat{\boldsymbol{\beta}}_0^{(k+1)}\|}{\|\hat{\boldsymbol{\beta}}_0^{(k)}\|}$ and go to AS.

This algorithm can be generalized to a wide class of linearly restricted GLMs if the restricting halfspaces are defined by sign and rank.

5 Simulation Study

The settings of the simulation study are similar to the settings of Bondell and Reich (2008). However, we adapt the true parameter vectors to GLMs with canonical link function by scaling and changed the number of observations for some settings. We compare the OSCAR penalty with the MLE and two established methods:

LASSO: The LASSO penalty, which uses the penalty $\lambda \sum_{j=1}^p |\beta_j|$,

Elastic Net (EN): The EN, which uses a combination of the LASSO penalty term and the ridge term $\lambda \left[\alpha \sum_{i=1}^p |\beta_j| + (1-\alpha) \sum_{i=1}^p \beta_j^2\right]$.

Several program packages in `R` that fit the EN and the LASSO for GLMs are available (for example Lokhorst et al., 2007; Park and Hastie, 2007a; Friedman et al., 2008; Goeman, 2010b). We use the `R`-package `glmnet` (see Friedman et al., 2008, 2010; Simon et al., 2011).

The predictive performance is measured by the predictive deviance

$$\text{Dev}(\boldsymbol{y},\, \widehat{\boldsymbol{\mu}}, \phi) = -2\phi \sum_i (l(y_i,\, \mu_i) - l(y_i,\, y_i)),$$

where $\widehat{\boldsymbol{\mu}}$ is the estimated prediction based on data $(\boldsymbol{y},\, \boldsymbol{X})$. First we fit the models for different tuning parameters on a training data set with n_{train} observations to get a set of parameter vector $\boldsymbol{B} = \{\widehat{\boldsymbol{\beta}}_0^{[1]},\, ...,\, \widehat{\boldsymbol{\beta}}_0^{[q]}\}$ where the superscript $^{[q]}$ indicates the tuning parameter constellation. Then a validation data set with n_{vali} observations is used to determine the optimal tuning parameter constellation that minimizes the predictive deviance on the validation data set

$$\widehat{\boldsymbol{\beta}}_0^{[opt]} = \text{argmin}_{\widehat{\boldsymbol{\beta}}_0 \in \boldsymbol{B}} \left\{ \text{Dev}(\boldsymbol{y}_{vali},\, h(\boldsymbol{Z}_{vali}\widehat{\boldsymbol{\beta}}_0), \phi) \right\}.$$

The test data is used to measure the predictive deviance

$$\text{Dev}(\boldsymbol{y}_{test},\, h(\boldsymbol{Z}_{test}\widehat{\boldsymbol{\beta}}_0^{[opt]}), \phi).$$

In addition we give the mean square error of $\boldsymbol{\beta}$ $\text{MSE} = p^{-1}\|\boldsymbol{\beta}_{true} - \widehat{\boldsymbol{\beta}}^{[opt]}\|^2$. We will consider the following settings.

Normal Case

For completeness we repeat the simulation study from Bondell and Reich (2008) with small modifications as described above. The generationg model for all data sets is $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_{true} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma\boldsymbol{I})$.

Norm1 The true parameter vector is $\boldsymbol{\beta}_1 = (3,\, 2,\, 1.5,\, 0,\, 0,\, 0,\, 0,\, 0)^T$ and the covariates are from $N(\boldsymbol{0},\, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = \{\sigma_{ij}\}_{i,j}$ with $\sigma_{ij} = 0.7^{|i-j|}$, $i,\, j = 1,\, ...,\, 8$. The number of observations are $n_{train} = 20$, $n_{vali} = 20$, and $n_{test} = 100$. As Bondell and Reich (2008) we choose $\sigma = 3$ for the standard deviation of the error term.

Norm2 This setting is the same as Norm1 but the true parameter vector is $\boldsymbol{\beta}_2 = (3,\, 0,\, 0,\, 1.5,\, 0,\, 0,\, 0,\, 2)^T$.

Norm3 This setting is the same as Norm1 and Norm2 but the true parameter vector is $\boldsymbol{\beta}_3 = 0.85 \cdot \boldsymbol{1}_8$.

Norm4 The true parameter vector is

$$\boldsymbol{\beta}_4 = (\underbrace{0,\, ...,\, 0}_{10},\, \underbrace{2,\, ...,\, 2}_{10},\, \underbrace{0,\, ...,\, 0}_{10},\, \underbrace{2,\, ...,\, 2}_{10})^T.$$

In each block of ten the covariates are from a $N(\boldsymbol{0},\, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = \{\sigma_{ij}\}_{i,j}$ with $\sigma_{ij} = 0.5$ if $i \neq j$ and $\sigma_{ii} = 1$, $i,\, j = 1,\, ...,\, 10$. Between the four blocks there is no correlation. The number of observations are $n_{train} = 100$, $n_{vali} = 100$, and $n_{test} = 500$. The standard devation of the error term is $\sigma = 15$ (compare Bondell and Reich, 2008).

Norm5 The true parameter vector is

$$\boldsymbol{\beta}_5 = (\underbrace{3, ..., 3}_{15}, \underbrace{0, ..., 0}_{25})^T.$$

and the number of observations are $n_{train} = 50$, $n_{vali} = 50$, and $n_{test} = 250$. The covariates are generated as follows. V_1, V_2, and V_3 are iid from a univariate $N(0, 1)$ with $X_i = V_1 + \varepsilon_i,\ i = 1, ..., 5$, $X_i = V_2 + \varepsilon_i,\ i = 6, ..., 10$, $X_i = V_3 + \varepsilon_i,\ i = 11, ..., 15$, $X_i \sim N(0, 1)$, $i = 16, ..., 40$. where $\varepsilon_i \sim N(0, 0.16)$. So only the influencial covariates are parted in three blocks of five. Inner each block the covariates are correlated. Between these blocks there is no correlation. The non influential covariates are uncorrelated and the standard deviation of the error term is $\sigma = 15$ (compare Bondell and Reich, 2008).

The results of this part of the simulation study is shown in Figure 2.

Poisson case

In the first three settings we divide the true parameter vector of the first three setting from Bondell and Reich (2008) by 4. The generating model of the Poisson setting has the form $y_i \sim Pois(\boldsymbol{x}_i^T \boldsymbol{\beta}_{true})$. The covariates are generated in the same way as in the normal case NormX.

Pois1 The true parameter vector is $\boldsymbol{\beta}_1 = (0.75, 0.5, 0.375, 0, 0, 0, 0, 0)^T$. The number of observations are $n_{train} = 20$, $n_{vali} = 20$, and $n_{test} = 100$.

Pois2 This setting is the same as Pois1 apart from the true parameter vector which is $\boldsymbol{\beta}_2 = (0.75, 0, 0, 0.375, 0, 0, 0, 0.5)^T$.

Pois3 This setting is the same as Pois1 and Pois2 apart from the true parameter vector $\boldsymbol{\beta}_3 = 0.2125 \cdot \mathbf{1}_8$.

Pois4 For this setting we divide the true parameter vector from Bondell and Reich (2008) by 20

$$\boldsymbol{\beta}_4 = (\underbrace{0, ..., 0}_{10}, \underbrace{0.1, ..., 0.1}_{10}, \underbrace{0, ..., 0}_{10}, \underbrace{0.1, ..., 0.1}_{10})^T.$$

The number of observations are $n_{train} = 100$, $n_{vali} = 100$, and $n_{test} = 500$.

Pois5 The true parameter vector Bondell and Reich (2008) is divided by 30

$$\boldsymbol{\beta}_5 = (\underbrace{0.1, ..., 0.1}_{15}, \underbrace{0, ..., 0}_{25})^T.$$

The number of observations are $n_{train} = 100$, $n_{vali} = 100$, and $n_{test} = 500$.

The result of this part of the simulation study is shown in Figure 3.

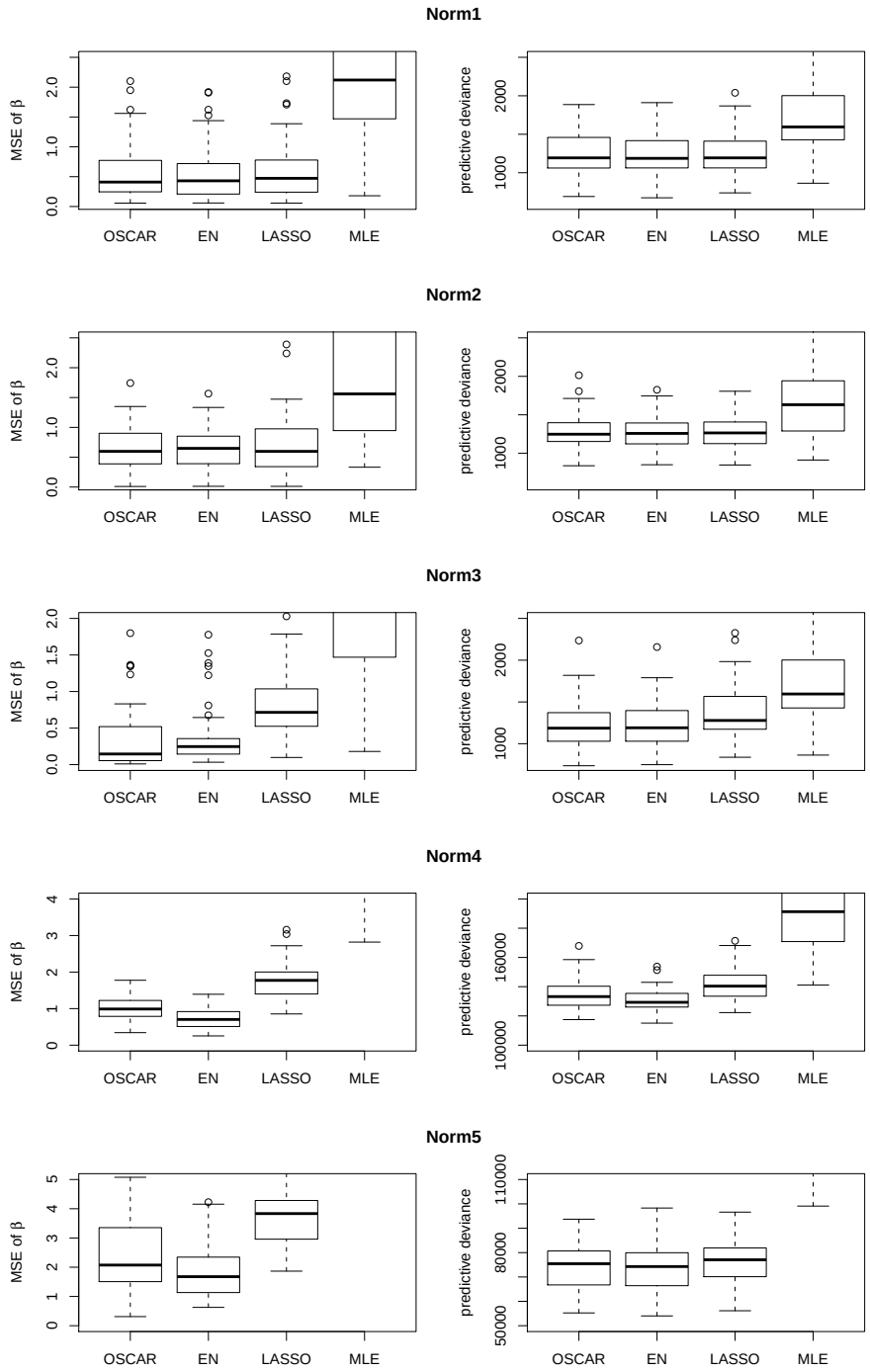

FIGURE 2: *Boxplots of MSE of $\boldsymbol{\beta}$ and the predictive deviance of the 5 settings in the normal case (see Bondell and Reich, 2008).*

Binomial case

In all setting covariates are generated in the same way as in the corresponding Poisson setting PoisX. The generating model is $y_i \sim Bin(\boldsymbol{x}_i^T \boldsymbol{\beta}_{true})$. For the first three settings we divide the true parameter vector of the first three setting from Bondell and Reich (2008) by 2.

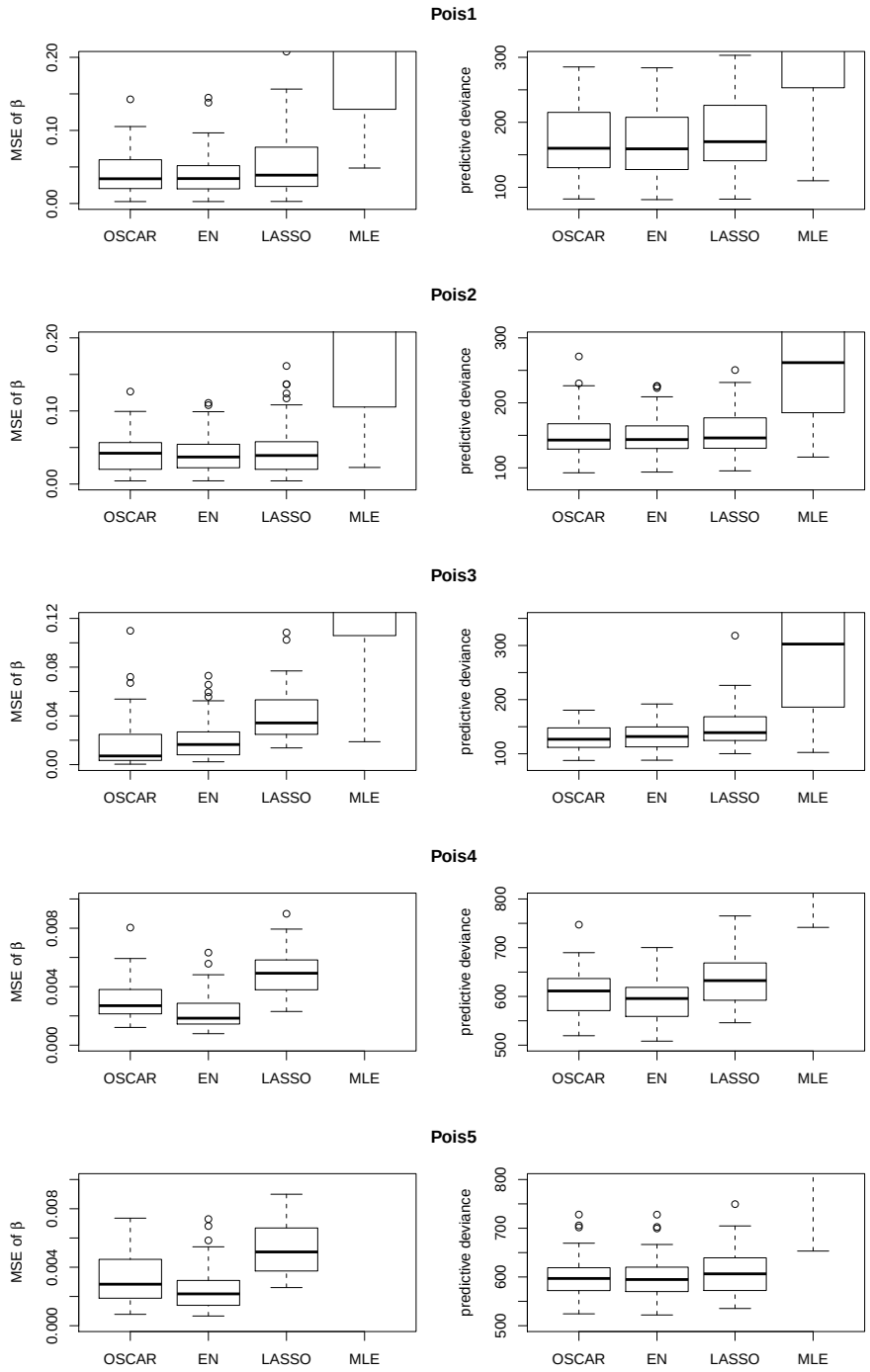

FIGURE 3: *Boxplots of MSE of $\boldsymbol{\beta}$ and the predictive deviance of the 5 Poisson settings.*

Bin1 The true parameter vector is $\boldsymbol{\beta}_1 = (1.5, 1, 0.75, 0, 0, 0, 0, 0)^T$. The number of observations are $n_{train} = 100$, $n_{vali} = 100$, and $n_{test} = 500$.

Bin2 This setting is the same as Bin1 but the true parameter vector is $\boldsymbol{\beta}_2 = (1.5, 0, 0, 0.75, 0, 0, 0, 1)^T$.

Bin3 This setting is the same as Bin1 and Bin2 but the true parameter vector $\boldsymbol{\beta}_3 = 0.425 \cdot \mathbf{1}_8$.

Bin4 We divide the true parameter vector from Bondell and Reich (2008) by 10

$$\boldsymbol{\beta}_4 = (\underbrace{0, ..., 0}_{10}, \underbrace{0.2, ..., 0.2}_{10}, \underbrace{0, ..., 0}_{10}, \underbrace{0.2, ..., 0.2}_{10})^T.$$

and increase the number of observations to $n_{train} = 200$, $n_{vali} = 200$, and $n_{test} = 1000$

Bin5 The true parameter vector Bondell and Reich (2008) is divided by 15

$$\boldsymbol{\beta}_5 = (\underbrace{0.2, ..., 0.2}_{15}, \underbrace{0, ..., 0}_{25})^T.$$

The number of observations is equal to Bin4.

In Figure 4 the results are illustrated by boxplots.

The results are summarized in Table 1. As a general tendency, it is seen that the procedures with clustering or grouping property outperform the LASSO, with the exception of settings Norm2 and Bin2. In the third settings the exact clustering of OSCAR seems to have an advantage over the non-exact grouping of the Elastic Net. Here the OSCAR dominates the other estimates. In the fourth setting OSCAR and EN outperform the LASSO, but the EN is the best for both criteria for all distributions. In the fifth setting the differences of the predictive deviance are quite small. With the exception of setting Bin2 the OSCAR is the best or second best for both criteria. In summary, the OSCAR for GLMs is a strong competitor to the Elastic Net, which outperforms the LASSO.

6 Application

The data were collected in water engineering in Southern California and contain 43 years worth of precipitation measurements. They are available from the `R`-package `alr3` (see Weisberg, 2011, 2005). The response variable is the stream runoff near Bishop (CA) in acre-feet. There are six covariates which are the snowfall in inches at different measurement stations labeled by `APMAM`, `APSAB`, `APSLAKE`, `OPBPC`, `OPRC`, and `OPSLAKE`. The covariates are grouped by its position. The covariates with labels that start with the same letter are quite close to each other and are highly correlated. The correlation structure is shown in Figure 5. We consider two cases: First we fit a linear normal model to predict the stream runoff. Then we split the response variable in two parts by setting the response $y_i = 0$ if $y_i < \text{median}(\boldsymbol{y})$ and $y_i = 1$ if $y_i \geq \text{median}(\boldsymbol{y})$. With this binary response we fit a GLM with binomial distribution and logit link. The tuning parameter are determined by

$$AIC = 2 \sum_{i=1}^{n} l_i(\boldsymbol{\beta}_0) + 2(df + 1).$$

Bondell and Reich (2008) proposed for df the number of coefficients that are absolute unique but non zero, or, in other words, the number of distinct non zero entries of $|\boldsymbol{\beta}|$.

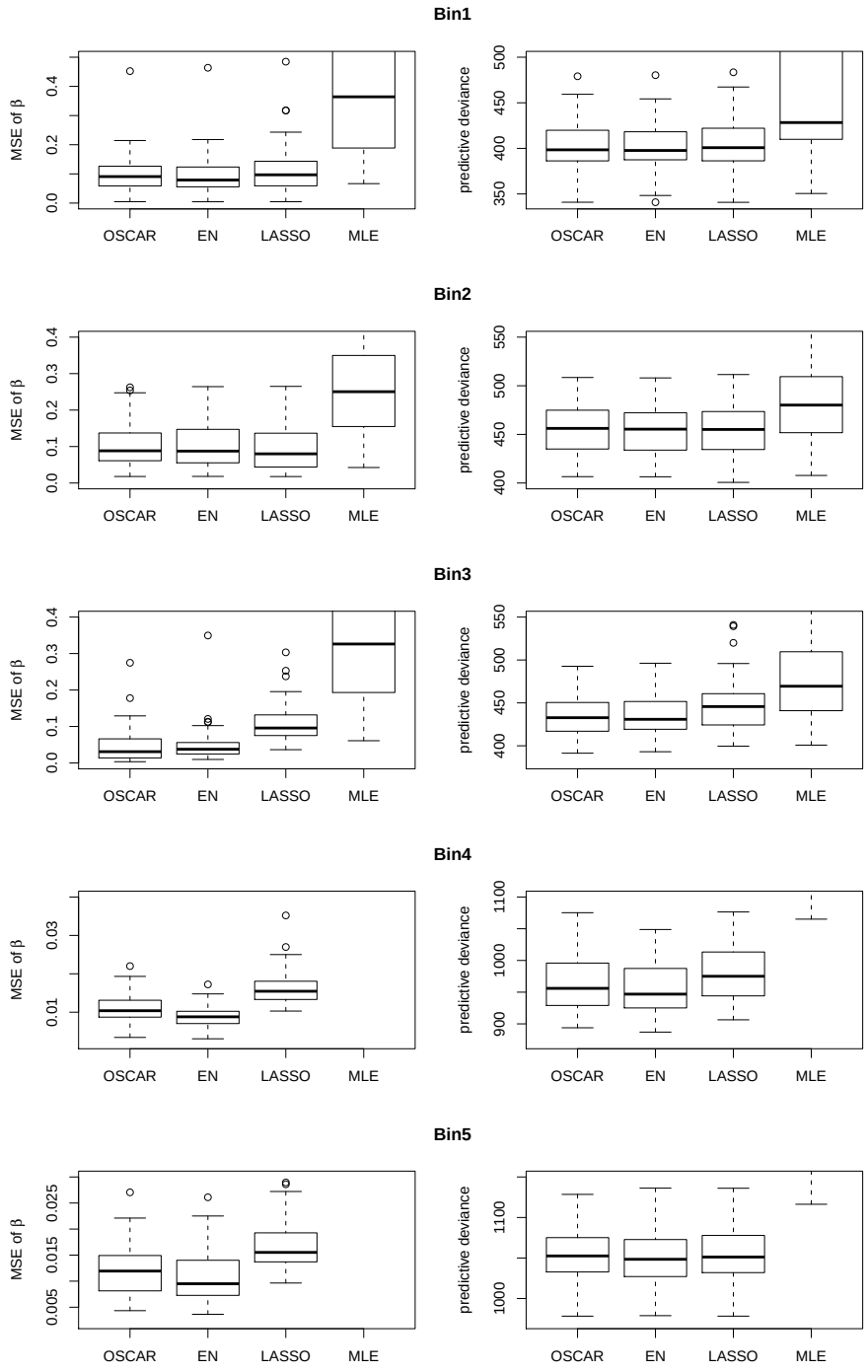

FIGURE 4: *Boxplots of MSE of $\boldsymbol{\beta}$ and the predictive deviance of the 5 binomial settings.*

We use the *AIC* to determine the tuning parameters because the MLE exists, which is necessary for using the `quadprog` procedure (see Turlach, 2009). Cross-validation does not work, especially in the binomial case, because the MLE does not exist for all subsamples. For the binomial case $c = 0.9$ and for the normal case $c = 0.2$ was determined.

		OSCAR	Elastic Net	LASSO	MLE
Normal Case (Results of predictive deviances are divided by 100)					
β_1	MSE	0.4095 (0.0754)	0.4303 (0.0667)	0.4724 (0.0781)	2.1195 (0.1283)
	DEV	11.93 (0.270)	11.87 (0.321)	11.93 (0.328)	15.95 (0.531)
β_2	MSE	0.5985 (0.0587)	0.6484 (0.0555)	0.5981 (0.0750)	1.5609 (0.3545)
	DEV	12.48 (0.265)	12.58 (0.226)	12.64 (0.400)	15.44 (0.950)
β_3	MSE	0.1212 (0.0610)	0.2272 (0.0333)	0.6321 (0.0733)	1.9654 (0.1888)
	DEV	11.27 (0.371)	11.72 (0.495)	12.54 (0.268)	15.45 (0.936)
β_4	MSE	0.9893 (0.0449)	0.7034 (0.0609)	1.7730 (0.0753)	6.7606 (0.4305)
	DEV	1332.13 (15.919)	1293.09 (12.569)	1403.87 (23.670)	1912.99 (63.918)
β_5	MSE	2.0738 (0.2089)	1.6770 (0.1335)	3.8346 (0.2317)	64.4542 (3.9849)
	DEV	754.37 (26.295)	742.73 (21.561)	770.61 (18.180)	3053.32 (192.65)
Poisson Case					
β_1	MSE	0.0339 (0.0053)	0.0341 (0.0045)	0.0388 (0.0071)	0.2710 (0.0459)
	DEV	160.01 (12.845)	159.17 (8.859)	170.15 (13.045)	354.85 (58.980)
β_2	MSE	0.0422 (0.0055)	0.0370 (0.0050)	0.0391 (0.0049)	0.2116 (0.0534)
	DEV	142.75 (4.031)	143.50 (4.108)	145.98 (4.727)	261.91 (42.965)
β_3	MSE	0.0071 (0.0027)	0.0165 (0.0031)	0.0342 (0.0046)	0.3171 (0.0590)
	DEV	126.84 (5.399)	131.85 (5.928)	139.03 (6.500)	302.63 (51.125)
β_4	MSE	0.0027 (0.0003)	0.0018 (0.0002)	0.0049 (0.0002)	0.0333 (0.0032)
	DEV	611.19 (8.774)	595.71 (9.000)	632.49 (7.975)	1295.12 (43.155)
β_5	MSE	0.0028 (0.0004)	0.0022 (0.0002)	0.0050 (0.0002)	0.0515 (0.0035)
	DEV	596.93 (7.479)	594.78 (6.857)	606.44 (6.359)	1192.28 (119.52)
Binomial Case					
β_1	MSE	0.0908 (0.0140)	0.0790 (0.0115)	0.0968 (0.0188)	0.3642 (0.0809)
	DEV	398.41 (3.693)	397.71 (4.283)	400.76 (4.194)	428.32 (10.937)
β_2	MSE	0.0883 (0.0134)	0.0875 (0.0152)	0.0800 (0.0096)	0.2504 (0.0194)
	DEV	456.17 (5.394)	455.39 (6.554)	455.07 (6.657)	480.13 (7.296)
β_3	MSE	0.0309 (0.0053)	0.0376 (0.0040)	0.0958 (0.0089)	0.3262 (0.0475)
	DEV	432.73 (4.484)	430.86 (5.779)	445.72 (5.523)	469.44 (7.657)
β_4	MSE	0.0104 (0.0005)	0.0095 (0.0005)	0.0155 (0.0009)	0.1921 (0.0136)
	DEV	956.02 (7.059)	946.90 (7.069)	975.07 (6.156)	1354.75 (33.367)
β_5	MSE	0.0119 (0.0009)	0.0093 (0.0008)	0.0155 (0.0010)	0.1806 (0.0221)
	DEV	1052.56 (6.903)	1048.43 (6.005)	1051.21 (7.999)	1366.61 (25.889)

TABLE 1: *Summary of the results of the Simulation study*

The EN and the LASSO paths were calculated with the `glmnet` (see Friedman et al., 2008, 2010; Simon et al., 2011). For the EN we determine $\alpha = 0.9$ in the normal case and $\alpha = 0.5$ in the binomial case. The coefficient buildups of standardized coefficients for the different procedures are shown in Figure 6, OSCAR is in the first row, LASSO in the second row, and the Elastic Net is in the third row of Figure 6. On the left the solution paths of the normal distribution case and on the right of the binomial distribution case are given. The dotted horizontal line show the optimal tuning parameter t.

The coefficient buildups of the OSCAR show a strong influence of the measurement

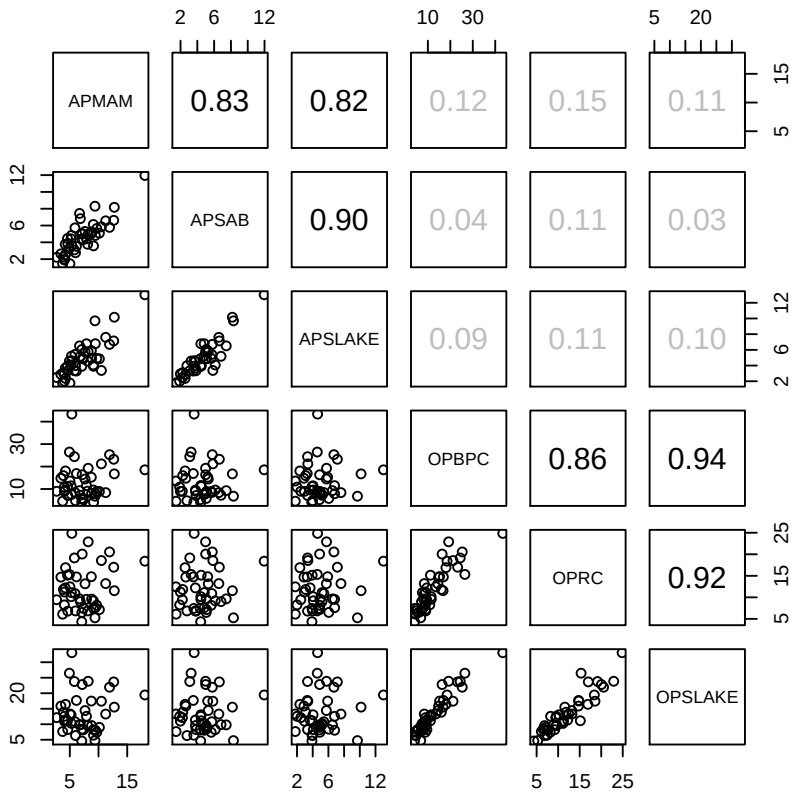

FIGURE 5: *Correlation structure of the covariates of the water data.*

stations that have names starting with "O". Especially in the normal case the clustering and the variable selection of OSCAR is quite impressive. All variables of the group starting with "O" are estimated equal and the second group is shrunken to zero for AIC optimal t. In the binary case clustering and variable selection is somewhat weaker, but still impressive, in particular when compared to to the elastic net. For optimal t `OPBPC` and `OPSLAKE` are clustered as well as two weaker correlated covariates (`APMAM` and `OPRC`). Only the variable `APSAB` is shrunken to zero. In the normal case the clustering coefficient buildups of EN and LASSO are quite similar. In the binomial case the EN has at least a tendency to cluster the covariates starting with "O". The exact clustering of covariates is easy to interpret, especially in the normal case. The snowfall at adjacent measurement stations has the same influence on the stream runoff. But only the influence of the snowfall at the measurement stations that have names starting with "O" have non-zero influence. The remaining (starting with an "A") are shrunk to zero.

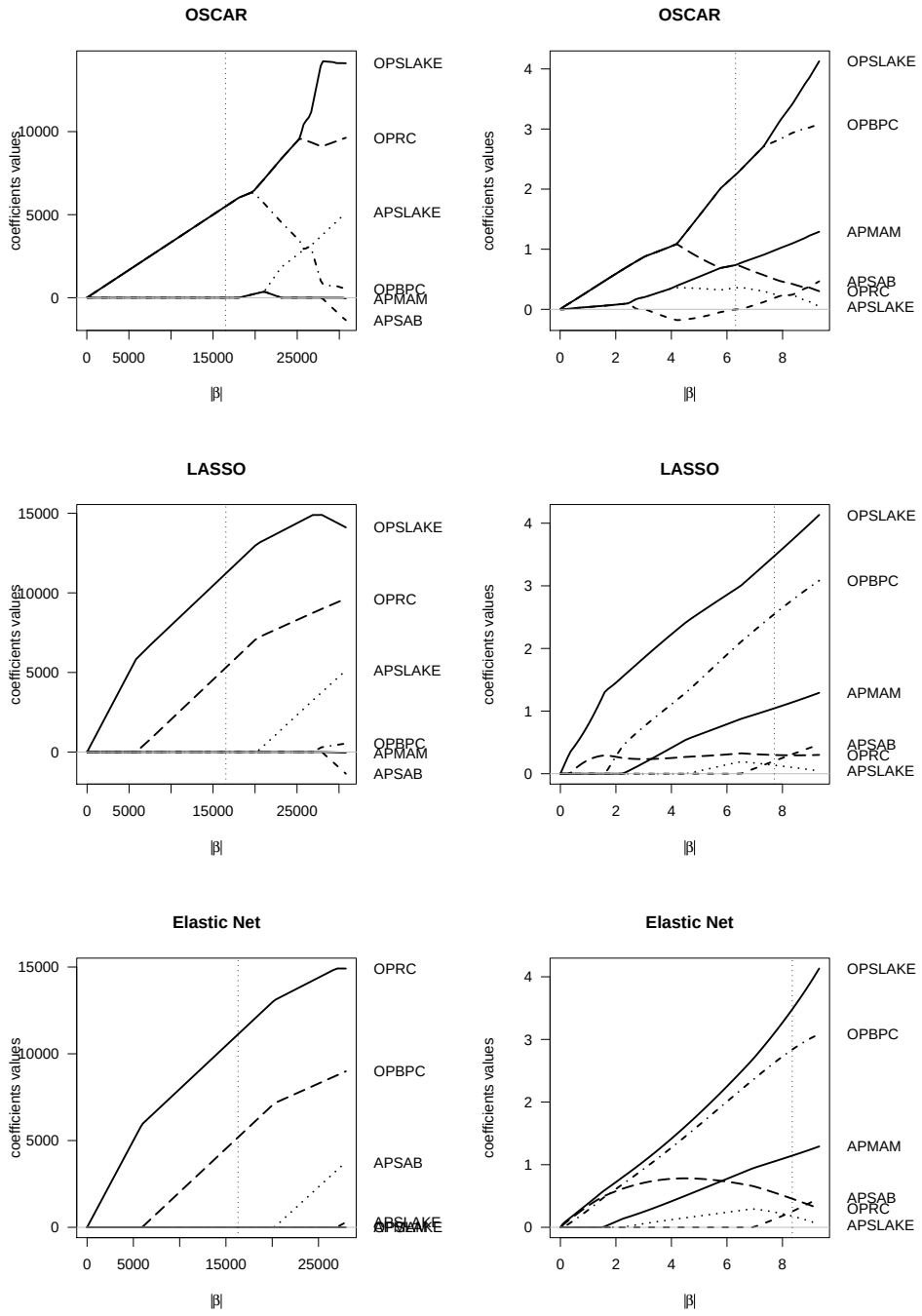

FIGURE 6: *Coefficient buildups for the water data. The left column shows the normal case and the right column shows the binary distribution case. In the first row the solution paths of the OSCAR are given, the second row shows the LASSO- and the third row the EN-paths.*

7 Conclusion and Remarks

We adapt the OSCAR penalty to GLMs. For solving the constrained log-likelihood problem we present an algorithm which combines the active set method and Fisher

scoring. It turns out that the OSCAR is quite competitive. In the simulation study it is the best or second best (with the exception of one setting) in terms of the MSE and the predictive deviance. Especially in the normal case the result of the data example is good to interpret. The snowfall at closed measurement stations is quite similar and so it can be assumed that their influence on the stream runoff is nearly equal. The data example also illustrates that the LASSO picks only two highly correlated covariates out of the group of three.

Acknowledgements

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

References

Bondell, H. D. and B. J. Reich (2008). Simultaneous regression shrinkage, variable selction and clustering of predictors with oscar. *Biometrics 64*, 115–123.

Friedman, J., T. Hastie, and R. Tibshirani (2008). *glmnet: Lasso and elastic-net regularized generalized linear models*. R package version 1.1.

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

Goeman, J. (2010a). L1 penalized estimation in the cox proportional hazards model. *Biometrical Journal 52*, 70–84.

Goeman, J. (2010b). *penalized: L1 (lasso) and L2 (ridge) penalized estimation in GLMs and in the Cox model*. R package version 0.9-32.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Bias estimation for nonorthogonal problems. *Technometrics 12*, 55–67.

Lokhorst, J., B. Venables, B. Turlach, and M. Maechler (2007). *lasso2: L1 constrained estimation aka 'lasso'*. R package version 1.2-6.

McCullagh, P. and J. A. Nelder (1983). *Generalized Linear Models*. New York: Chapman & Hall.

Park, M. Y. and T. Hastie (2007a). *glmpath: L1 Regularization Path for Generalized Linear Models and Cox Proportional Hazards Model*. R package version 0.94.

Park, M. Y. and T. Hastie (2007b). L1 regularization-path algorithm for generalized linear models. *Journal of the Royal Statistical Society B 69*, 659–677.

Petry, S. and G. Tutz (2011). Shrinkage and variable selection by polytopes. *Journal of Statistical Planning and Inference (to appear)*.

Simon, N., J. Friedman, T. Hastie, and R. Tibshirani (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software 39*(5), 1–13.

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

Turlach, B. A. (2009). *quadprog: Functions to solve Quadratic ProgrammingProblems*. R package version 1.4-11, S original by Berwin A. Turlach, R portby Andreas Weingessel.

Weisberg, S. (2005). *Applied Linear Regression* (Third ed.). Hoboken NJ: Wiley.

Weisberg, S. (2011). *alr3: Data to accompany Applied Linear Regression 3rd edition*. R package version 2.0.3.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B 67*, 301–320.