LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK

Sebastian Petry, Claudia Flexeder & Gerhard Tutz

Pairwise Fused Lasso

Technical Report Number 102, 2011
Department of Statistics
University of Munich

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

Pairwise Fused Lasso

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

March 3, 2011

Abstract

In the last decade several estimators have been proposed that enforce the grouping property. A regularized estimate exhibits the grouping property if it selects groups of highly correlated predictor rather than selecting one representative. The pairwise fused lasso is related to fusion methods but does not assume that predictors have to be ordered. By penalizing parameters and differences between pairs of coefficients it selects predictors and enforces the grouping property. Two methods how to obtain estimates are given. The first is based on LARS and works for the linear model, the second is based on quadratic approximations and works in the more general case of generalized linear models. The method is evaluated in simulation studies and applied to real data sets.

Keywords: Regularization, Fused lasso, Fusion estimates, Lasso, Elastic net

1 Introduction

Regularized estimation of regression parameters has been investigated thoroughly within the last decade. With the introduction of the lasso, proposed by Tibshirani (1996), methods for sparse modeling in the high-predictor case became available. In the following many alternative regularized estimators that include variable selection were proposed, among them the elastic net (Zou and Hastie, 2005), SCAD (Fan and Li, 2001), the Dantzig selector (Candes and Tao, 2007) and boosting approaches (for example Bühlmann and Yu, 2003; Bühlmann and Hothorn, 2007). Meanwhile most procedures are also available for generalized linear models (GLMs). Since we will also work within the GLM framework in the following some notation is introduced.

Let the generalized linear model (GLM) with response function $h(.)$ be given by

$$\boldsymbol{\mu} = E(\boldsymbol{y}|\boldsymbol{X}) = h(\mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta}),$$

where $\boldsymbol{y} = (y_1, ..., y_n)^T$ is the response vector and $\boldsymbol{X}$ is the design matrix. It is assumed that the predictors are standardized, $\sum_{i=1}^n x_{ij} = 0$ and $(n-1)^{-1}\sum_{i=1}^n x_{ij}^2 = 1, \forall j \in \{1, ..., p\}$. In the linear predictor $\boldsymbol{\eta} = \mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta}$ the intercept β_0 is separated because usually it is not penalized. With $\boldsymbol{\beta}_0 = (\beta_0, \boldsymbol{\beta}^T)$ we denote the parameter vector including the intercept β_0. Given the ith observation $\boldsymbol{X}_i$ the y_i are (conditionally) independent observations from a simple exponential family

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

where θ_i is the natural parameter of the family, ϕ is a scale or dispersion parameter and $b(.), c(.)$ are specific functions corresponding to the type of the family.

Penalized likelihood estimates of coefficients have the general form

$$\widehat{\boldsymbol{\beta}}_0 = \underset{\boldsymbol{\beta}_0}{\text{argmin}} \; \{l(\boldsymbol{\beta}_0) + P_{\boldsymbol{\lambda}}(\boldsymbol{\beta})\},$$

where $P_{\boldsymbol{\lambda}}(\boldsymbol{\beta})$ is the penalty term that regularizes the estimates and $l(\boldsymbol{\beta}_0)$ is the negative log-likelihood function which corresponds to (1). Ridge regression (Hoerl and Kennard, 1970), which uses $P_\lambda^R(\boldsymbol{\beta}) = \lambda\sum_{j=1}^p \beta_j^2$, frequently has smaller prediction error than ordinary maximum likelihood (ML) estimates but does not select predictors. The lasso penalty $P_\lambda^L(\boldsymbol{\beta}) = \lambda\sum_{j=1}^p |\beta_j|$ proposed by Tibshirani (1996), has the advantage that coefficients whose corresponding predictors have vanishing or low influence on the response are shrunk to zero. As discussed by Zou and Hastie (2005) the lasso does not group predictors and estimates maximal n predictors unequal to 0. In terms of Zou and Hastie (2005) an estimator exhibits the grouping property if it tends to estimate the absolute value of coefficients (nearly) equal if the corresponding predictors are highly correlated. In the case of highly correlated influential covariates the lasso procedure tends to select only few of these. As an alternative Zou and Hastie (2005) presented the elastic net (EN). Its penalty term is the sum of lasso and ridge penalty, $P_{\lambda_1}^L(\boldsymbol{\beta}) + P_{\lambda_2}^R(\boldsymbol{\beta})$. It is a strongly convex penalty which can also perform variable selection. Nowadays `R` packages for solving the lasso- or the EN-penalized likelihood problems for GLMs are available. For example Goemann (2010) and Friedman et al. (2010) proposed algorithms to solve elastic net penalized regression problems. Both algorithms are available as `R`-packages `penalized` and `glmnet`. Lokhorst et al. (2007) and

Park and Hastie (2007) provided the R-packages the `lasso2` and `glmpath` for solving lasso penalized regression problem.

More recently, several alternative methods that also show grouping have been proposed. Bondell and Reich (2008) proposed OSCAR for Octagonal Shrinkage and Clustering Algorithm for Regression. An attractive feature of OSCAR is that it can group very strictly. For specific choice of the tuning parameters the estimates of coefficients are equal. Therefore one obtains clustered predictors where one cluster shares the same coefficient. Typically one big cluster has estimates zero representing the predictors that have not been selected. Tutz and Ulbricht (2009) considered correlation based regularization terms that explicitly take the correlation of predictors into account. In order to obtain variable selection the correlation-based penalty has to be used within a boosting algorithm or an additional lasso term has to be used. For the combination of lasso and correlation-based terms see Anbari and Mkhadri (2008).

In the present paper an alternative method that enforces the grouping effect is proposed. It uses penalty terms that are similar to the fused lasso (FL) proposed by Tibshirani et al. (2005) and shows good performance in terms of variable selection and prediction.

2 Pairwise Fused Lasso (PFL)

The original fused lasso (Tibshirani et al., 2005) was developed for ordered predictors or signals as predictors and metrical response. For such predictors it is possible to use the distances between predictors to obtain sparsity. Thus the fused lasso penalty

$$P^{FL}_{\lambda_1,\lambda_2}(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2} |\beta_j - \beta_{j-1}|, \tag{2}$$

penalizes the difference between the coefficients of adjacent predictors β_j and β_{j-1}. With proper selection of tuning parameters adjacent predictors are fused or grouped. The first summand (the lasso term) of the fused lasso penalty enforces variable selection, the second enforces fusion.

The pairwise fused lasso (PFL), which is proposed here, extends the fused lasso (Tibshirani et al., 2005) to situations where the predictors have no natural ordering. Fusion refers to all possible pairs of predictors and not only to adjacent ones. Thus, the pairwise fused lasso penalty is defined by

$$P^{PFL}_{\lambda,\,\alpha}(\boldsymbol{\beta}) = \lambda \left[\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=2}^{p} \sum_{k=1}^{j-1} |\beta_j - \beta_k| \right], \tag{3}$$

where $\lambda > 0$ and α with $\alpha \in [0,\,1]$ are the tuning parameters. The first term of the pairwise fused lasso penalty is the lasso penalty and accounts for variable selection, the second term represents the sum of the absolute values of all pairwise differences of regression coefficients. This part of the penalty induces clustering.

By using all pairwise differences the pairwise fused lasso assumes no ordering of the predictors. For categorical predictors a similar penalty has been used for factor selection in ANOVA by Bondell and Reich (2009), and for categorical variable selection by Gertheiss and Tutz (2010).

Soil Data - An Illustrating Example

In the soil data, which were used by Bondell and Reich (2008), the response is rich-cove forest diversity (measured by the number of different plants species) in the Appalachian Mountains of North Carolina and the explaining covariates are 15 characteristics. Twenty areas of the same size were surveyed. The number of observations was 20 which is close to the number of predictors which was 15. The data can be partitioned into two blocks. On the one hand there is a group of 7 highly correlated predictors. This group contains cationic covariates, 4 cations (calcium, magnesium, potassium, and sodium) and 3 measurements that are very close to them. The other group of covariates contains 4 other chemical elements and 4 other soil characteristics, for example pH-value. The correlations within this group is not very high. It is remarkable that the design matrix has not full rank.

For illustration we use four different methods, lasso and three PFL methods. The first segments of the coefficient paths given in Figure 1 demonstrate the selecting and grouping property. It is seen that there is a strong similarity between the lasso and the PFL method for $\alpha = 0.98$. For large values of the tuning parameter λ the lasso selects only few covariates. This effect is also seen in the group of the highly correlated cationic covariates. It can bring instability in the estimates as discussed by Zou and Hastie (2005) or Breiman (1996). For smaller value of α the selection part becomes weaker and the fusion part stronger. It is seen that for $\alpha = 0.9$ and more distinctly for $\alpha = 0.1$ the highly correlated variables are fused, but there is hardly any effect beside selection for the weaker correlated variables in the second column of Figure 1.

Extended Versions of Fused Lasso

The pairwise fused lasso penalty (3) can be modified by adding different weights to achieve an improvement of the prediction accuracy or of the mean squared error of the estimated parameter vector. Accordingly, a modification of the penalty term is

$$P^{PFL}_{\lambda,\alpha,\boldsymbol{w}}(\boldsymbol{\beta}) = \lambda \left[\alpha \sum_{j=1}^{p} w_j |\beta_j| + (1-\alpha) \sum_{j=2}^{p} \sum_{k=1}^{j-1} w_{jk} |\beta_j - \beta_k| \right], \tag{4}$$

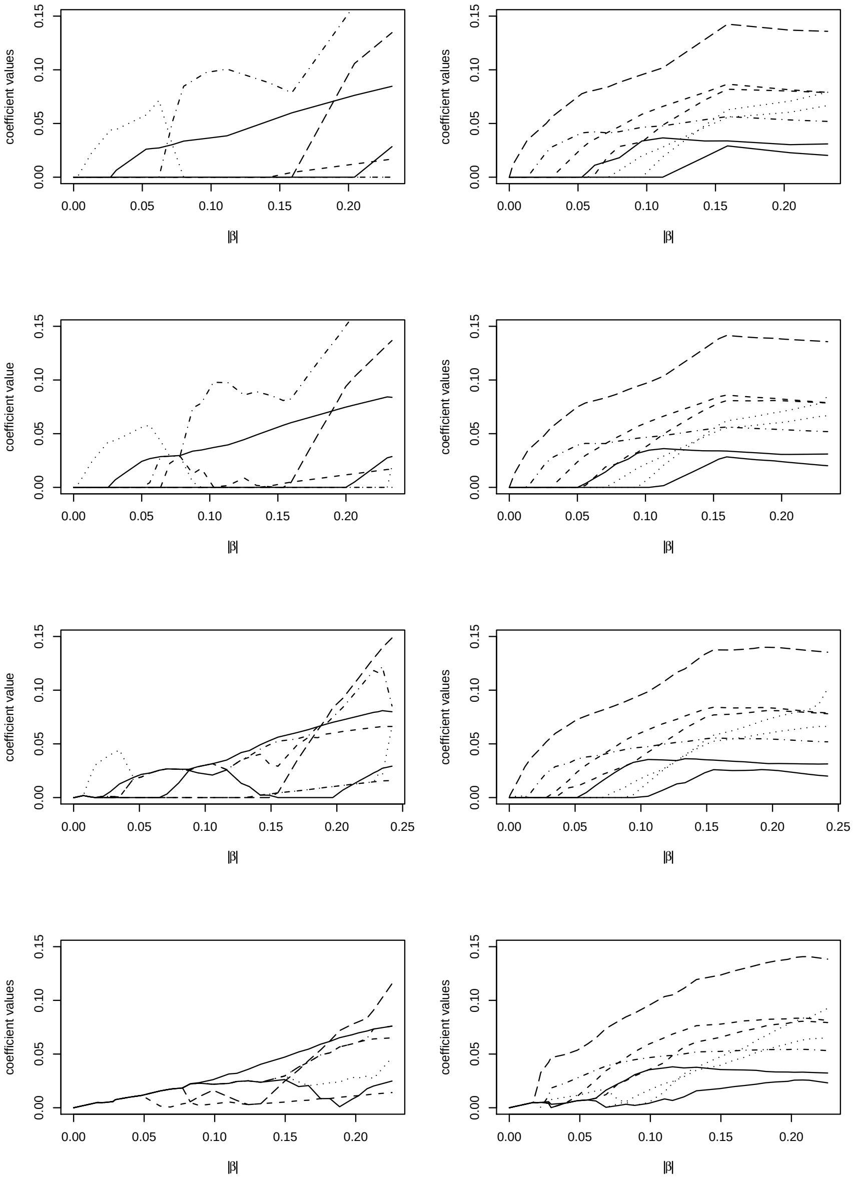

Figure 1: *First segments of the solution paths for standardized coefficients on the whole soil data set for decreasing tuning parameter λ. Left column: paths of the cationic covariates. Right column: paths of the non cationic covariates. First row: coefficient path of the lasso. Second row: coefficient path of PFL model with small clustering part ($\alpha = 0.98$). Third row: coefficient path of PFL model with $\alpha = 0.9$. Fourth row: coefficient path of PFL model with dominating fusion part ($\alpha = 0.02$).*

where w_j and w_{jk} are additional weights. One possibility is to choose $w_j = |\beta_j^{ML}|^{-1}$ and $w_{jk} = |\beta_j^{ML} - \beta_k^{ML}|^{-1}$, where β_i^{ML} denotes the ith component of maximum likelihood estimate. This choice is motivated by the adaptive lasso (Zou, 2006) and its oracle properties. These data-dependent weights can yield better prediction error if the maximum likelihood is well conditioned. In contrast to the simple pairwise fused lasso where all parameters have the same amount of shrinkage strength the penalty varies across coefficients. Large values of $|\beta_i^{ML}|$ yield small weights w_i and consequently weaker shrinkage of the corresponding parameters. If the maximum likelihood estimates of the jth and the kth predictor have nearly the same value, the weight w_{jk} causes large influence of the difference penalty term.

Another possibility is to include the correlation among predictors into the penalty. Zou and Hastie (2005) showed a relationship between correlation and grouping such that strongly correlated covariates tend to be in or out of the model together, but the correlation structure was not used explicitly in the penalty term. A regularization method, which is based on the idea that highly correlated covariates should have (nearly) the same influence on the response except to their sign, is the correlation based penalty considered by Tutz and Ulbricht (2009). Coefficients of two predictors are weighted according to their marginal correlation. As a result, the intensity of penalization depends on the correlation structure. In the same spirit the penalty term of the pairwise fused lasso can be extended to

$$P_{\lambda,\alpha,\widehat{\boldsymbol{\rho}}}^{PFL}(\boldsymbol{\beta}) = \lambda \left[\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=2}^{p} \sum_{k=1}^{j-1} \frac{1}{1 - |\widehat{\rho}_{jk}|} |\beta_j - \text{sign}(\widehat{\rho}_{jk}) \beta_k| \right], \quad (5)$$

where $\widehat{\rho}_{jk}$ denotes the estimated marginal correlation between the jth and the kth predictor. The factor $\text{sign}(\widehat{\rho}_{jk})$ is caused by the fact that two negatively correlated predictors have the same magnitude of influence but different signs. That is, for $\widehat{\rho}_{jk} \to 1$, the coefficients $\widehat{\beta}_j$ and $\widehat{\beta}_k$ are nearly the same and for $\rho_{jk} \to -1$, $\widehat{\beta}_j$ will be close to $-\widehat{\beta}_k$, respectively. In the case of uncorrelated predictors ($\widehat{\rho}_{jk} = 0$) we obtain the usual, unweighted pairwise fused lasso penalty.

Since the marginal correlation measures the interaction between the predictors $\boldsymbol{x}_j$ and $\boldsymbol{x}_k$ without taking further covariates into account, we also investigate the correlation based penalty in Equation (5) with partial correlations instead of the marginal ones. The partial correlation determines to what extent the correlation between two variables depends on the linear effect of the other covariates (Whittaker, 1990). Thereby, the aim is to eliminate this linear effect. We compute the partial correlation matrix with the `R` package `corpcor` (Schäfer et al., 2009). In this package a method for the regularization of (partial) correlation matrix is implemented which makes sense in ill conditioned problems. In general the correlation based weights can be substituted by dependency measurement which are normed on $[-1, 1]$. A combination of correlation and ML weights is possible. But this quite complicate penalty term did not show better performance.

2.1 Solving the Penalized ML Problem

In this section we discuss two procedures for solving the PFL problem

$$\widehat{\boldsymbol{\beta}}_0^{PFL} = \underset{\boldsymbol{\beta}}{\text{argmin}} \; \{l(\boldsymbol{\beta}_0) + P_{\lambda,\alpha}^{PFL}(\boldsymbol{\beta})\},$$

where $P_{\lambda,\alpha}^{PFL}(\boldsymbol{\beta})$ can be modified to include weights or correlation terms. The first approach works only for normally distributed response. It is based on the LARS algorithm from Efron et al. (2004). The second procedure is a generic algorithm based on local quadratical approximation (LQA). The basic principles of this algorithm were given by Osborne et al. (2000) and Fan and Li (2001). The general LQA algorithm can solve a very wide class of penalized likelihood problems (see Ulbricht, 2010b) and is available in an R-package (Ulbricht, 2010a). We will give a short introduction to the algorithm in the second part of this section.

2.1.1 Metric Regression and the LARS approach

We assume that $\boldsymbol{y}$ is centered and the response is normally distributed. Then one has to solve the penalized least square problem

$$\widehat{\boldsymbol{\beta}}^{PFL} = \underset{\boldsymbol{\beta}}{\text{argmin}} \; \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + P_{\lambda,\alpha}^{PFL}(\boldsymbol{\beta}). \tag{6}$$

It is helpful to reparameterize the problem as follows. Let new parameters be defined by

$$\begin{aligned} \theta_{jk} &= \beta_j - \beta_k, \; 1 \le k < j \le p, \\ \theta_{j0} &= \beta_j, \; 1 \le j \le p, \end{aligned} \tag{7}$$

with the restriction

$$\theta_{jk} = \theta_{j0} - \theta_{k0}, \; 1 \le k < j \le p. \tag{8}$$

With $\mathbf{0}_{p\times\binom{p}{2}}$ denoting a $p \times \binom{p}{2}$-matrix zero matrix an expanded design matrix is $(\boldsymbol{X}|\mathbf{0}_{p\times\binom{p}{2}})$. The corresponding parameter vector is

$$\boldsymbol{\theta} = (\theta_{10}, \ldots, \theta_{p0}, \theta_{21}, \ldots, \theta_{p(p-1)})^T. \tag{9}$$

With the PFL penalty having the form

$$P_{\lambda,\alpha}^{PFL}(\boldsymbol{\theta}) = \lambda \left[\alpha \sum_{j=1}^{p} w_{j0}|\theta_{j0}| + (1-\alpha) \sum_{j=1}^{p-1} \sum_{k=j+1}^{p} w_{jk}|\theta_{jk}| \right]$$

the restiction (8) is incorporated by using an additional quadratic penalty term $\sum_{j=1}^{p-1}\sum_{k=j+1}^{p}(\theta_{j0} - \theta_{k0} - \theta_{jk})^2$ weighted by a large tuning parameter γ. This yields

$$\begin{aligned}\widehat{\boldsymbol{\theta}}^{PFL} \quad = \quad & \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\ \|\boldsymbol{y} - (\boldsymbol{X}|\boldsymbol{0}_{p\times\binom{p}{2}})\|^2 \\ & +\gamma \sum_{j=1}^{p-1}\sum_{k=j+1}^{p} (\theta_{j0} - \theta_{k0} - \theta_{jk})^2 \\ & +\lambda \left[\alpha \sum_{j=1}^{p} w_{j0}|\theta_{j0}| + (1-\alpha)\sum_{j=1}^{p-1}\sum_{k=j+1}^{p} w_{jk}|\theta_{jk}|\right].\end{aligned} \tag{10}$$

For $\gamma \to \infty$ the restriction (8) is fulfilled. The reparameterization (7) allows to formulate the approximate estimator (10) as a lasso type problem. Similar reparameterizations were used by Zou and Hastie (2005) to represent the elastic net problem as a lasso type problem. In the present problem one uses

$$\begin{aligned}\widehat{\boldsymbol{\theta}}^{PFL} \quad = \quad & \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\ \|\boldsymbol{y}_0 - \widetilde{\boldsymbol{D}}\boldsymbol{\theta}\|^2 \\ & +\lambda \left[\alpha \textstyle\sum_{j=1}^{p} w_{j0}|\theta_{j0}| + (1-\alpha)\sum_{j=1}^{p-1}\sum_{k=j+1}^{p} w_{jk}|\theta_{jk}|\right],\end{aligned} \tag{11}$$

where $\boldsymbol{y}_0 = (\boldsymbol{y}^T,\, \boldsymbol{0}^T_{\binom{p}{2}})^T$ and $\boldsymbol{0}$ denotes a zero vector of length $\binom{p}{2}$. $\widetilde{\boldsymbol{D}}$ is the design matrix

$$\widetilde{\boldsymbol{D}} = \begin{pmatrix} \boldsymbol{X}|\boldsymbol{0}_{p\times\binom{p}{2}} \\ \sqrt{\gamma}\boldsymbol{C} \end{pmatrix},$$

where the matrix $\boldsymbol{C}$ is the $p \times \left(\binom{p}{2} + p\right)$-matrix which accounts for the restriction (8) which is equivalent to

$$\theta_{j0} - \theta_{k0} - \theta_{jk} = 0,\ 1 \le k < j \le p. \tag{12}$$

So the restriction (8) is fulfilled if $\boldsymbol{C\theta} = \boldsymbol{0}_{\binom{p}{2}}$ and $\boldsymbol{C}$ has the following form. Let $\boldsymbol{\delta}_{jk}$, $1 \le k < j \le p$, denote a p-dimensional row vector with -1 at the kth and $+1$ at the jth component and zero otherwise. Let $\boldsymbol{\tau}_m$ denote a $\binom{p}{2}$-dimensional row vector whose mth component is -1 and zero otherwise. Then all constrains given by (8) resp. (12) can be summarized in matrix notation

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{\delta}_{21} & \boldsymbol{\tau}_1 \\ \boldsymbol{\delta}_{31} & \boldsymbol{\tau}_2 \\ \vdots & \vdots \\ \boldsymbol{\delta}_{p1} & \boldsymbol{\tau}_{p-1} \\ \boldsymbol{\delta}_{32} & \boldsymbol{\tau}_p \\ \boldsymbol{\delta}_{42} & \boldsymbol{\tau}_{p+1} \\ \vdots & \vdots \\ \boldsymbol{\delta}_{p(p-1)} & \tau_{\binom{p}{2}} \end{pmatrix}. \tag{13}$$

Let $\Theta = \{(i, j)|0 \leq j < i < p\}$ denote the index set of the components of $\boldsymbol{\theta}$ given by (9) one obtains

$$\begin{aligned}\widehat{\boldsymbol{\theta}}^{PFL} &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \ \|\boldsymbol{y}_0 - \widetilde{\boldsymbol{D}}\boldsymbol{\theta}\|^2 + \lambda(\sum_{j=1}^{p} |\alpha\theta_{j0}| + \sum_{j=1}^{p-1}\sum_{k=j+1}^{p} |(1-\alpha)\theta_{jk}|) \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \ \|\boldsymbol{y}_0 - \widetilde{\boldsymbol{D}}\boldsymbol{\theta}\|^2 + \lambda(\sum_{t\in\Theta} |\alpha \cdot \theta_t| + |(1-\alpha) \cdot \theta_t|). \qquad (14)\end{aligned}$$

Equation (14) is a lasso problem on the expanded design matrix $\widetilde{\boldsymbol{D}}$ weighted by α and $(1-\alpha)$. The weights can be included by multiplying $\widetilde{\boldsymbol{D}}$ with the reciprocals of weights

$$\boldsymbol{D} = \widetilde{\boldsymbol{D}} \operatorname{diag}(\alpha w_{10}, ..., \alpha w_{p0}, (1-\alpha)w_{21}, ..., (1-\alpha)w_{p(p-1)})^{-1}. \qquad (15)$$

to obtain

$$\widehat{\boldsymbol{\theta}}^{PFL} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \ \|\boldsymbol{y}_0 - \boldsymbol{D}\boldsymbol{\theta}\|^2 + \lambda(\sum_{t\in\Theta} |\theta_t|).$$

So finally to get $\widehat{\boldsymbol{\beta}}^{PFL}$ we have to multiply the first p components of $\widehat{\boldsymbol{\theta}}^{PFL}$ with $\alpha^{-1} \operatorname{diag}(\alpha w_{10}, ..., \alpha w_{p0})$. For the correlation based pairwise fused lasso we have to modify the submatrix $\boldsymbol{C}$ of $\widetilde{\boldsymbol{D}}$. If $\operatorname{sign}(\widehat{\rho}_{jk}) = -1$ then δ_{jk}, $1 \leq k < j \leq p$, is a p-dimensional row vector where the kth and the jth component are $+1$ and all remaining are zero (see equation (5)). It is remarkable that for $w_{jk} = 1$, $0 \leq 1 < k \leq p$, in (15) we get the unweighted PFL3.

2.1.2 Generalized Linear Models and the LQA Approach

A general class of penalized generalized linear models can be fitted by using the local quadratic approximation (LQA) approach (Ulbricht, 2010b). The LQA algorithm solves penalized minimization problems

$$\widehat{\boldsymbol{\beta}}_0 = \underset{\boldsymbol{\beta}_0}{\operatorname{argmin}} \ \left\{l(\boldsymbol{\beta}_0) + P_{\boldsymbol{\lambda}}^{\delta}(\boldsymbol{\beta})\right\}, \qquad (16)$$

where $l(\boldsymbol{\beta}_0)$ is the negative log-likelihood of the underlying generalized linear model and the penalty term is a sum of J penalty functions having the form

$$P_{\boldsymbol{\lambda}}^{\delta}(\boldsymbol{\beta}) = \sum_{j=1}^{J} p_{\lambda_j,j}(|\boldsymbol{a}_j^T\boldsymbol{\beta}|), \qquad (17)$$

where the $\boldsymbol{a}_j$ are known vector of constants.Let the superscript δ denote the specific penalty family, e.g. $P_{\boldsymbol{\lambda},\alpha}^{PFL}(\boldsymbol{\beta})$ denotes the pairwise fused lasso penalty. The penalty proposed by Fan and Li (2001) has the special structure $P_{\lambda}^{\delta}(\boldsymbol{\beta}) = \sum_{j=1}^{p} p_{\lambda}(|\beta_j|)$. Since for that structure the vectors $\boldsymbol{a}_j$ have only one non-zero element it cannot be used to include interactions between the predictors. Hence, the approach of Fan and Li (2001) can be applied only to penalty families such as ridge and lasso, but not to the fused lasso or pairwise fused lasso.

In 17 the sum of all J penalty functions $p_{\lambda_j,j}(|\boldsymbol{a}_j^T\boldsymbol{\beta}|)$ determines the penalty region, the number J of penalty functions in in general not equal to the number of regressors p. Furthermore, the type of the penalty function and the tuning parameter λ_j do not have to be the same for all J penalty functions. It is easily seen that the pairwise fused lasso penalty can be described by

$$P_{\lambda,\alpha}^{PFL}(\boldsymbol{\beta}) = \sum_{j=1}^{p+\binom{p}{2}} p_{\lambda,\alpha,j}(|\boldsymbol{a}_j^T\boldsymbol{\beta}|).$$

The first p penalty functions are

$$p_{\lambda,\alpha,j}(\cdot) = \lambda \cdot \alpha|\boldsymbol{a}_j^T\boldsymbol{\beta}|, \quad j = 1, \ldots, p,$$

where $\boldsymbol{a}_j = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ with a one at the jth position. The next $\binom{p}{2}$ penalty functions for the difference penalty term are

$$p_{\lambda,\alpha,j}(\cdot) = \lambda\,(1-\alpha)\,|\boldsymbol{a}_j^T\boldsymbol{\beta}|, \quad j = p+1, \ldots, \tilde{p}+p$$

with the p-dimensional vectors having the form $\boldsymbol{a}_j = (0, \ldots, -1, 0, \ldots, 1, 0, \ldots, 0)$ which describes the differences between two parameters..

An often applied principle in solving a convex optimization problem is to use a quadratic approximation of the objective function. If the latter is twice continuously differentiable iterative procedures of the Newton type apply. Therefore, we need the gradient and the Hessian of the objective function. Since the first term of (16) is the negative log-likelihood, we can use the corresponding score function and expected Fisher information matrix. For the second term, one cannot proceed the same way because it includes L_1-norm terms. Therefore, Ulbricht (2010b) developed a quadratic approximation of the penalty term (17) which is shortly sketched in the following. Based on this approximation, Newton-type algorithms can be applied.

Let the following properties hold for all J penalty functions:

1. $p_{\lambda,j} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with $p_{\lambda,j}(0) = 0$,
2. $p_{\lambda,j}$ is continuous and monotone in $|\boldsymbol{a}_j^T\boldsymbol{\beta}|$,
3. $p_{\lambda,j}$ is continuously differentiable for all $\boldsymbol{a}_j^T\boldsymbol{\beta} \neq 0$, i.e. $dp_{\lambda,j}\left(|\boldsymbol{a}_j^T\boldsymbol{\beta}|\right)/d|\boldsymbol{a}_j^T\boldsymbol{\beta}| \geq 0$ for all $\boldsymbol{a}_j^T\boldsymbol{\beta} \geq 0$.

Let $\boldsymbol{\beta}_{(k)}$ denote the approximation of the estimate $\widehat{\boldsymbol{\beta}}$ at the kth iteration of the LQA algorithm. Then the first order Taylor expansion of the jth penalty function in the neighborhood of $\boldsymbol{\beta}_{(k)}$ can be written as

$$p_{\lambda,j}\left(|\boldsymbol{a}_j^T\boldsymbol{\beta}|\right) \approx p_{\lambda,j}\left(|\boldsymbol{a}_j^T\boldsymbol{\beta}_{(k)}|\right) + \frac{1}{2}\frac{p'_{\lambda,j}\left(|\boldsymbol{a}_j^T\boldsymbol{\beta}_{(k)}|\right)}{\sqrt{\left(\boldsymbol{a}_j^T\boldsymbol{\beta}_{(k)}\right)^2 + c}}\left(\boldsymbol{\beta}^T\boldsymbol{a}_j\boldsymbol{a}_j^T\boldsymbol{\beta} - \boldsymbol{\beta}_{(k)}^T\boldsymbol{a}_j\boldsymbol{a}_j^T\boldsymbol{\beta}_{(k)}\right) \tag{18}$$

which is a quadratic function of $\boldsymbol{\beta}$. Thereby, $p'_{\lambda,j}\left(|\boldsymbol{a}_j^T\boldsymbol{\beta}_{(k)}|\right) = dp_{\lambda,j}\left(|\boldsymbol{a}_j^T\boldsymbol{\beta}|\right)/d|\boldsymbol{a}_j^T\boldsymbol{\beta}| \geq 0$ denotes the first derivative and c is a small positive integer (for our computations we choose $c = 10^{-8}$). Using matrix notation and summation over all J penalty functions the Taylor expansion is equivalent to

$$\sum_{j=1}^{J} p_{\lambda,j}\left(|\boldsymbol{a}_j^T\boldsymbol{\beta}|\right) \approx \sum_{j=1}^{J} p_{\lambda,j}\left(|\boldsymbol{a}_j^T\boldsymbol{\beta}_{(k)}|\right) + \frac{1}{2}\left(\boldsymbol{\beta}^T\boldsymbol{a}_\lambda\boldsymbol{\beta} - \boldsymbol{\beta}_{(k)}^T\boldsymbol{a}_\lambda\boldsymbol{\beta}_{(k)}\right), \tag{19}$$

with

$$\boldsymbol{a}_\lambda = \sum_{j=1}^{J}\frac{p'_{\lambda,j}\left(|\boldsymbol{a}_j^T\boldsymbol{\beta}_{(k)}|\right)}{\sqrt{\left(\boldsymbol{a}_j^T\boldsymbol{\beta}_{(k)}\right)^2 + c}}\boldsymbol{a}_j\boldsymbol{a}_j^T \tag{20}$$

which does not depend on the parameter vector $\boldsymbol{\beta}$. Since an intercept is included in the model, the penalty matrix is extended to

$$\boldsymbol{a}_\lambda^* = \begin{bmatrix} 0 & \boldsymbol{0}^T \\ \boldsymbol{0} & \boldsymbol{a}_\lambda \end{bmatrix}, \tag{21}$$

where $\boldsymbol{0}$ is the p-dimensional zero vector. Then, starting with the initial value $\boldsymbol{b}_{(0)}$, the update step of this Newton-type algorithm based on local quadratic approximations of the penalty term is

$$\boldsymbol{b}_{(k+1)} = \boldsymbol{b}_{(k)} - \left(\boldsymbol{F}(\boldsymbol{b}_{(k)}) + \boldsymbol{a}_\lambda^*\right)^{-1}\left\{-_{(}\boldsymbol{b}_{(k)}) + \boldsymbol{a}_\lambda^*\boldsymbol{b}_{(k)}\right\}. \tag{22}$$

Corresponding to the log-likelihood $l(\boldsymbol{b})$, $_{(}\boldsymbol{b})$ and $\boldsymbol{F}(\boldsymbol{b})$ denote the score function and Fisher information matrix, respectively. Iterations are carried out until the relative distance moved during the kth step is less or equal to a specified threshold ϵ, i.e. the termination condition is

$$\frac{||\boldsymbol{b}_{(k+1)} - \boldsymbol{b}_{(k)}||}{||\boldsymbol{b}_{(k)}||} \leq \epsilon, \quad \epsilon > 0. \tag{23}$$

3 Simulation Study

In this section we investigate the performance of the pairwise fused lasso and compare it to established procedures. All simulations are based on the generalized linear model

$$E(\boldsymbol{y}|\boldsymbol{X}) = h(\boldsymbol{X}\boldsymbol{\beta}_{true})$$

where $h(.)$ is the canonical response function. 50 replications are performed for every simulation scenario and in each replication we generate a training, a validation and a test data set. The observation numbers of the corresponding data sets are denoted by $n_{train}/n_{vali}/n_{test}$. The training set is used to fit the models defined by the different tuning parameter(s). The optimal tuning parameter(s) are determined by the minimizing the deviance on the validation data set. Finally we use the test data set to evaluate the prediction by the predictive deviance on the test dataset, $DEV = -2(l(\widehat{\boldsymbol{y}}_{test}) - l(\boldsymbol{y}_{test}))$. Further we use $MSE = \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|^2$ to measure the accuracy of the estimate of $\boldsymbol{\beta}$. The result are illustrated by boxplot where the outliers are not shown. As abbreviation for the differently weighted PFLs we will use the following:

- PFL denotes PFL penalty with all weights set to 1.
- PFL.ml denotes PFL penalty with ML-weights.
- PFL.cor denotes PFL penalty with correlation driven weights.
- PFL.pcor denotes PFL penalty with partial correlation driven weights.

We give the lasso, EN, and the ML estimates for comparison. The lasso and the EN estimates are calculated by the `lqa` routine. Since we investigate a regularization method with both variable selection and grouping property, we use the following simulation scenarios.

Normal Regression

Setting 1: The first setting is specified by the parameter vector $\boldsymbol{\beta}_{true} = (3, 1.5, 0, 0, 0, 2, 0, 0)^T$ and standard error $\sigma = 3$. The correlation between the i-th and the j-th predictor is

$$\operatorname{corr}(i, j) = 0.9^{|i-j|}, \ \forall i, j \in \{1, \ldots, 8\}. \tag{24}$$

The observation numbers are 20/20/200.

Setting 2: In this setting we have $p = 20$ predictors. The parameter vector is structured into blocks:

$$\boldsymbol{\beta}_{true} = \big(\underbrace{0, \ldots, 0}_{5}, \underbrace{2, \ldots, 2}_{5}, \underbrace{0, \ldots, 0}_{5}, \underbrace{2, \ldots, 2}_{5}\big)^{\mathrm{T}}.$$

The standard error σ is 15 and the correlation between two predictors $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ is given by $\operatorname{corr}(i, j) = 0.5$. The observation numbers are 50/50/400.

Setting 3: This setting consists of $p = 20$ predictors. The parameter vector is given by

$$\boldsymbol{\beta}_{true} = \big(5, 5, 5, 2, 2, 2, 10, 10, 10, \underbrace{0, \ldots, 0}_{11}\big)^T.$$

and $\sigma = 15$. The design matrix $\boldsymbol{X}$ is specified by the following procedure. First we generate three auxiliary predictors $Z_j \sim N_n(\mathbf{0}, \boldsymbol{I})$, $j \in \{1, 2, 3\}$. With these predictors we generate

$$\begin{aligned} \boldsymbol{X}_i &= Z_1 + \tilde{\boldsymbol{\epsilon}}_i,\ i \in \{1, 2, 3\}, \\ \boldsymbol{X}_i &= Z_2 + \tilde{\boldsymbol{\epsilon}}_i,\ i \in \{4, 5, 6\}, \\ \boldsymbol{X}_i &= Z_3 + \tilde{\boldsymbol{\epsilon}}_i,\ i \in \{7, 8, 9\}, \end{aligned}$$

with $\tilde{\boldsymbol{\epsilon}}_i \sim N_n(\mathbf{0}, 0.01\boldsymbol{I})$, $i \in \{1, \ldots, 9\}$. The predictors $\boldsymbol{X}_i$, $i \in \{10, \ldots, 20\}$, are white noise, i.e. $\boldsymbol{X}_i \sim N_n(\mathbf{0}, \boldsymbol{I})$. Thus, within the first three blocks of 3 variables there is a quite high correlation, but there is no correlation between these blocks. The observation numbers are 50/50/400.

Binary Regression

In each simulation scenario the observation numbers $n_{train}/n_{vali}/n_{test}$ correspond to 100/100/400. Furthermore, the predictor $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}_{true}$ from the Normal case is multiplied by a factor a in order to realize a appropriate domain for the logistic response function. The value range of the predictor should be approximately the interval $[-4, 4]$. Thus, for each setting we determine a factor a and multiply the true parameter vector from the normal case by this factor. The corresponding factor a and the modified parameter vector for each simulation setting are given by:

Setting 1:

$$a = 0.40 \quad \rightarrow \quad \boldsymbol{\beta}_{true} = (1.2, 0.6, 0, 0, 0, 0.8, 0, 0)^T$$

Setting 2:

$$a = 0.15 \quad \rightarrow \quad \boldsymbol{\beta}_{true} = \big(\underbrace{0, \ldots, 0}_{5}, \underbrace{0.3, \ldots, 0.3}_{5}, \underbrace{0, \ldots, 0}_{5}, \underbrace{0.3, \ldots, 0.3}_{5}\big)^T$$

Setting 3:

$$a = 0.10 \quad \rightarrow \quad \boldsymbol{\beta}_{true} = \big(0.75, 0.75, 0.75, 0.3, 0.3, 0.3, 1.5, 1.5, 1.5, \underbrace{0, \ldots, 0}_{11}\big)^T$$

The response is finally modeled by $y_i = Bin(1,\ (1 + \exp(-\eta_i))^{-1})$. In Figure 3 the result is illustrated by boxplots

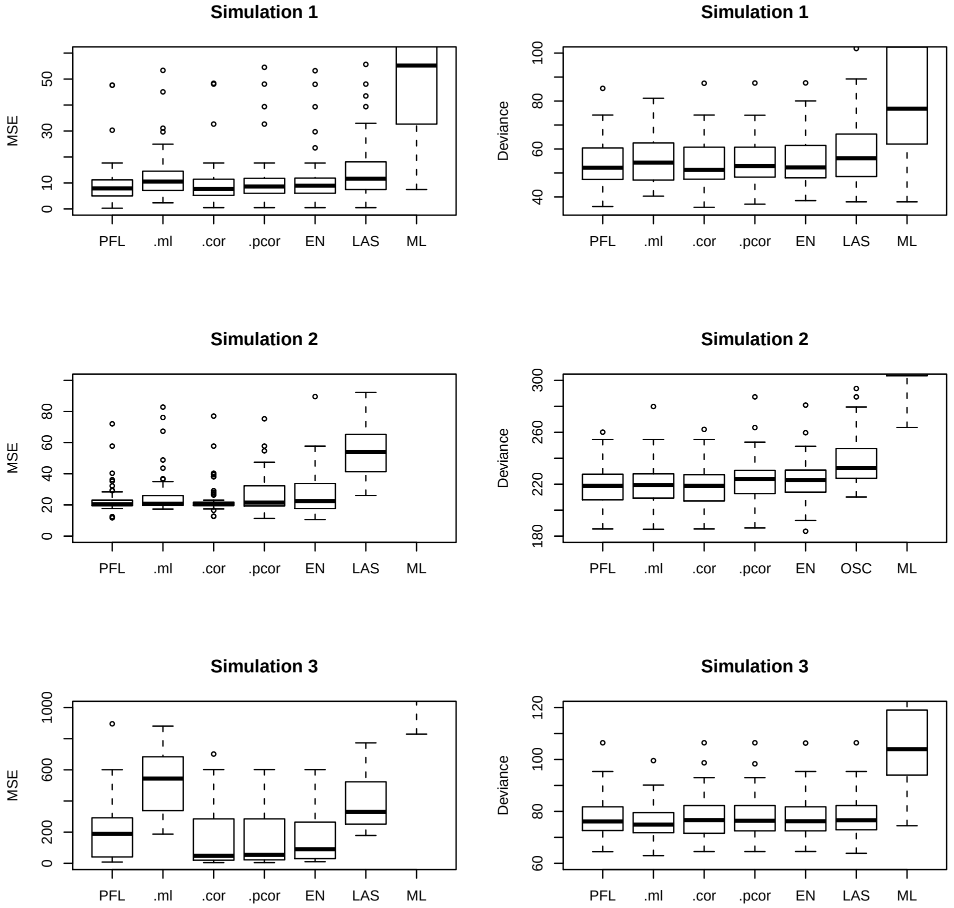

FIGURE 2: *Boxplots of the MSE and Deviance for simulations with normal distributed response.*

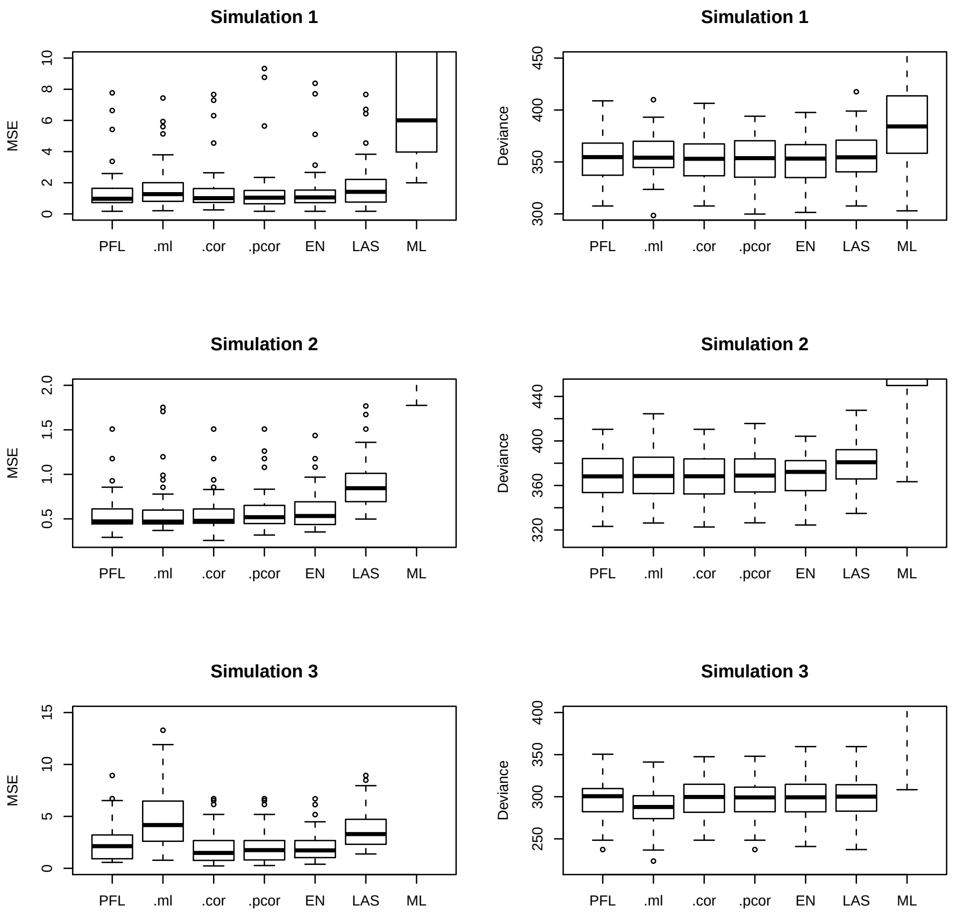

FIGURE 3: *Boxplots of the MSE and Deviance for simulations with binomial distributed response*

Poisson Regression

Analogously to the simulation study on binary responses, the predictor $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}_{true}$ is multiplied by a factor a. Since the value range of the mean $\boldsymbol{\mu} = \exp(\boldsymbol{\eta})$ should be approximately in the interval $[0, 8]$, we again determine for each setting the corresponding factor a. We model the response by $y_i = Pois(\exp(\eta_i))$. The modified parameter vectors and the factor a determine the settings:

Setting 1:

$$a = 0.15 \quad \rightarrow \quad \boldsymbol{\beta}_{true} = (0.45, 0.225, 0, 0, 0, 0.3, 0, 0)^T$$

Setting 2:

$$a = 0.05 \quad \rightarrow \quad \boldsymbol{\beta}_{true} = \big(\underbrace{0, \ldots, 0}_{5}, \underbrace{0.1, \ldots, 0.1}_{5}, \underbrace{0, \ldots, 0}_{5}, \underbrace{0.3, \ldots, 0.3}_{5}\big)^T$$

Setting 3:

$$a = 0.03 \ \rightarrow \ \boldsymbol{\beta}_{true} = \big(0.15, 0.15, 0.15, 0.06, 0.06, 0.06, 0.3, 0.3, 0.3, \underbrace{0, \ldots, 0}_{11}\big)^T$$

We model the response by $y_i = Pois(1, \exp(\eta_i))$. Figure 4 sums up the result by boxplots

Summing Up the Result

The results of the simulation studies are summarized in Table 1. It is seen that the PFL is competitive in terms of the predictive deviance and the MSE. The simulation study gives no clear indication which weights are best. The performance of both correlation based weights is quite similar. The correlation based weights seem to perform quite well across all settings. In general, apart from the ML based estimate, the PFL penalties distinctly outperform the lasso and are strong competitors for the elastic net. The pairwise penalization seems to be an appropriate way for improving the performance of estimates. The exception are methods based on ML weights which suffer from the instability of the ML estimate. In ill-conditioned cases one should replace the MLE by a regularized estimate which does not select variables like the ridge estimator. It should be noted that in contrast to the elastic net the PFL penalty enforces identical coefficients for “similar” variables where the meaning of “similar” is specified by the chosen weights.

4 Data Example

In this section we give two real data examples. One for the Binomial case and one for the Gaussian. In both cases we split the data set 50 times in two parts. One training data set with n_{train} observations and a test data set with n_{test} observations. We use the training data set to learn the model by a 5-fold cross

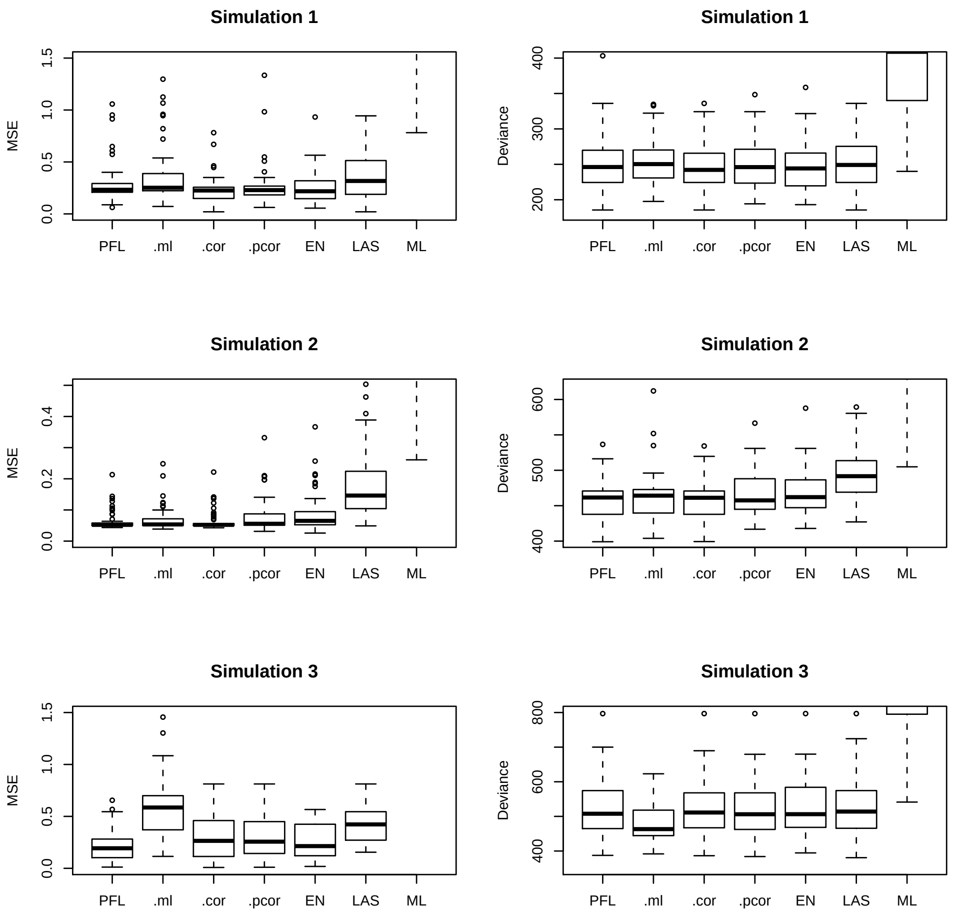

FIGURE 4: *Boxplots of the MSE and Deviance for simulations with Poisson distributed response*

		PFL	PFL.ml	PFL.cor	PFL.pcor	EN	LASSO	ML
Normal distribution								
Setting 1	MSE	7.90	10.54	**7.64**	8.64	8.95	11.64	55.22
		(0.88)	(0.68)	(0.94)	(0.61)	(0.57)	(1.83)	(8.13)
	DEV	52.17	54.33	**51.25**	52.85	52.32	56.13	76.79
		(2.05)	(3.345)	(1.90)	(2.31)	(2.68)	(2.86)	(4.62)
Setting 2	MSE	20.39	20.82	**20.35**	21.53	22.37	54.01	284.16
		(0.25)	(0.52)	(0.20)	(1.50)	(1.57)	(3.77)	(22.11)
	DEV	**218.82**	219.21	218.82	223.90	223.04	232.50	336.00
		(3.01)	(2.66)	(2.82)	(2.52)	(2.05)	(3.05)	(12.71)
Setting 3	MSE	189.15	543.81	**48.07**	54.40	90.79	330.20	4057.24
		(44.89)	(51.63)	(76.95)	(70.67)	(58.76)	(26.06)	(315.11)
	DEV	76.12	**74.90**	76.66	76.39	76.22	76.60	103.97
		(1.37)	(1.00)	(1.24)	(1.00)	(1.34)	(1.30)	(2.77)
Binomial distribution								
Setting 1	MSE	**0.97**	1.27	1.01	1.0404	1.06	1.42	6.00
		(0.12)	(0.13)	(0.11)	(0.11)	(0.14)	(0.15)	(1.23)
	DEV	354.66	354.11	**353.04**	353.66	353.24	354.49	384.20
		(4.50)	(3.31)	(4.53)	(4.54)	(4.55)	(5.38)	(4.34)
Setting 2	MSE	0.47	**0.47**	0.48	0.52	0.53	0.84	8.46
		(0.015)	(0.01)	(0.01)	(0.02)	(0.03)	(0.05)	(1.39)
	DEV	**368.18**	368.46	368.26	368.91	372.20	380.85	528.39
		(2.22)	(3.05)	(0.99)	(2.60)	(3.51)	(2.97)	(40.19)
Setting 3	MSE	2.13	4.17	**1.48**	1.75	1.73	3.30	399.04
		(0.33)	(0.26)	(0.43)	(0.35)	(0.24)	(0.44)	(100.04)
	DEV	300.64	**287.81**	299.71	299.21	299.34	300.14	544.94
		(2.36)	(4.36)	(3.63)	(2.99)	(3.71)	(3.66)	(51.69)
Poisson distribution								
Setting 1	MSE	0.23	0.25	0.23	0.23	**0.22**	0.32	5.70
		(0.01)	(0.02)	(0.01)	(0.01)	(0.02)	(0.05)	(1.06)
	DEV	246.19	250.30	**242.14**	246.06	244.20	249.11	408.90
		(6.44)	(6.50)	(5.40)	(6.66)	(6.88)	(6.05)	(55.29)
Setting 2	MSE	0.05	0.05	**0.05**	0.06	0.07	0.15	1.54
		(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.02)	(0.14)
	DEV	461.56	464.22	461.23	**457.51**	462.09	491.67	929.31
		(4.53)	(4.15)	(3.08)	(7.00)	(6.04)	(5.59)	(61.26)
Setting 3	MSE	**0.19**	0.59	0.26	0.26	0.21	0.42	20.19
		(0.03)	(0.04)	(0.03)	(0.03)	(0.04)	(0.05)	(2.58)
	DEV	507.66	**463.25**	511.19	506.18	506.36	513.92	1061.44
		(12.33)	(8.10)	(18.07)	(15.28)	(18.65)	(19.69)	(63.48)

Table 1: *Results of the simulation studies.*

validation. The model is determined by a parameter vector $\widehat{\boldsymbol{\beta}}_{train}$. The test data set is used for measuring the predictive deviance $-2(l(\boldsymbol{y}_{test}, \widehat{\boldsymbol{y}}_{test}) - l(\boldsymbol{y}_{test}, \boldsymbol{y}_{test}),$ where $l(., .)$ denotes the log likelihood function and $\widehat{y}_{test} = h((\mathbf{1}, \boldsymbol{X}_{test})\boldsymbol{\beta}_{train})$ is the modeled expectation for the test data set.

Biopsy Data Set

The biopsy dataset is from the R-package MASS Venables and Ripley (2002). It contains 699 observations and 9 covariates. We exclude the 16 observations with missing values. The covariates are whole-number scores between 0 and 10. Their description is given in Table 2. The response contains two classes of breast cancer "benign" or "malignant" and so we fit a logistic regression model. The predictive deviance is given in Figure 5 and in Table 3 In Figure 6 the estimates are shown.

Number	Explanation
1	clump thickness
2	uniformity of cell size
3	uniformity of cell shape
4	marginal adhesion
5	single epithelial cell size
6	bare nuclei
7	bland chromatin
8	normal nucleoli
9	mitoses

TABLE 2: *Covariates of the biopsy data*

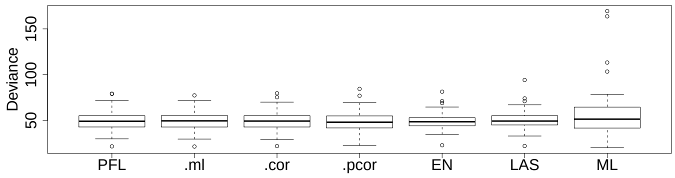

FIGURE 5: *Boxplots of the predictive deviance for the Biopsy Data Set*

In contrast to the Elastic Net estimates the grouping property of the PFL is stronger. Further it is remarkable that different models have similar predictive deviances. The MLE leads some to perfect discrimination of the groups and the procedures gives warning.

PFL	PFL.ml	PFL.cor	PFL.pcor	EN	LASSO	ML
49.2292 (10.8875)	49.6492 (10.9686)	49.4307 (11.3377)	48.18492 (10.7444)	48.6917 (8.8604)	49.4356 (11.0634)	51.5290 (27.7673)

Table 3: *The median of predictive deviance on test data for the Bones Data Set. The bootstrap variance based on 500 bootstrap samples is bracketed.*

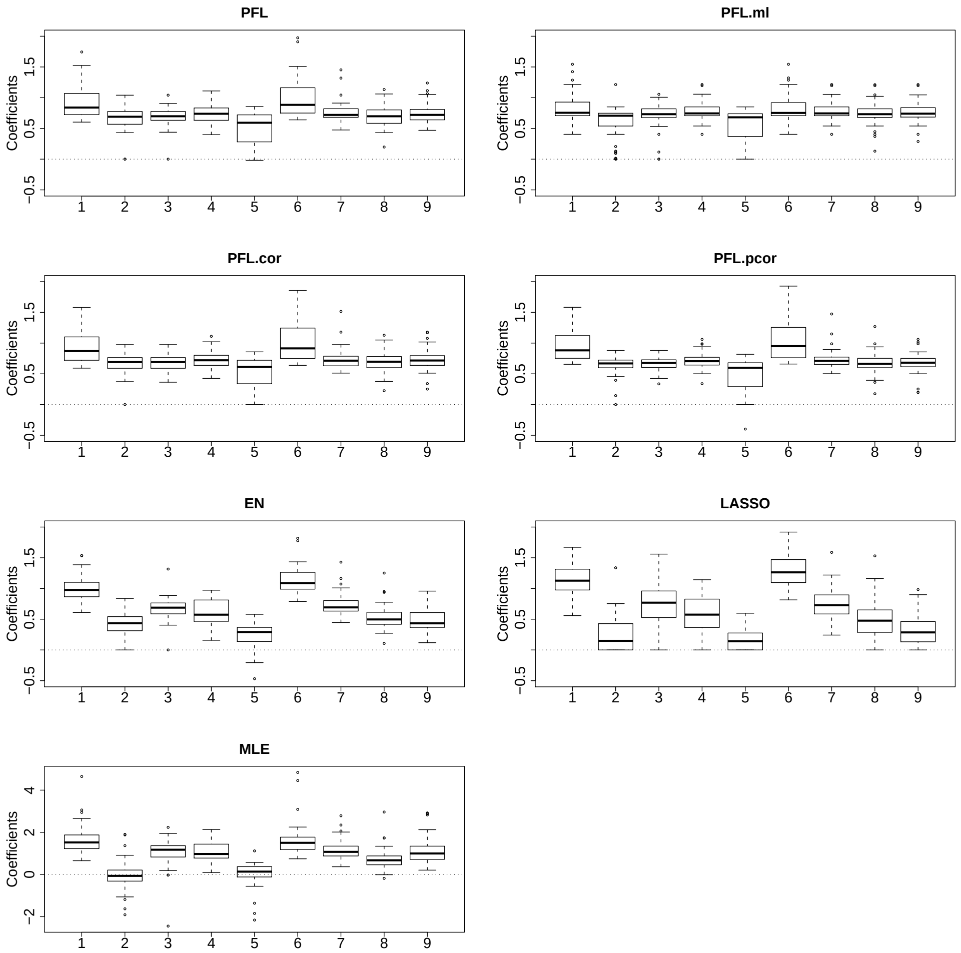

Figure 6: *Boxplots of the coefficient estimates for the Biopsy Data Set*

Bones Data Set

This study aims at estimating the age by various measurements of bones for 87 persons. The underlying data set consists 20 predictors. They are bones characteristics and the gender of the deceased person. The data based on the Basel-Kollektiv and are provided by Stefanie Doppler from the department of anthropology and human genetics of the LMU Munich. The predictors are given in Table 4. Some of the predictors are highly correlated, i.e. $\rho_{ij} \approx 0.9$. We choose the normal model. We randomly split the data set 25-times into a test data set with 60 observations and a test data set with $n = 27$. The predictive deviance on test data and for each method are given in Table 5 and illustrated in Figure 7. We give standardized estimates by boxplots of the coefficient estimates. Here the selecting and grouping effect appears. All regularized estimators select variables. The MLE-weighted PFL tends to group the covariates 12,13, and 14. It has the best predictive deviance. It is remarkable that variable selection dominates clustering in the other cases. Although the MLE is quite ill-conditioned using ML weights improves the prediction.

Number	Explanation
1	gender
2	size of an compact bone
3	femur class
4	type I osteon
5	type II osteon
6	osteon fragments
7	osteon population density
8	Haverssche canals
9	non Haverssche canals
10	Volkmannsche canals
11	resorption lacuna
12	percentage of resorption lacuna
13	percentage of general lamellae
14	percentage of osteonal bones
15	percentage of fragmental bones
16	surface of an osteon
17	surface of a resorption lacuna
18	quotient of the surface of a resorption lacuna and the surface of an osteon
19	activation frequency
20	bone formation rate

Table 4: *Covariates of the bones data*

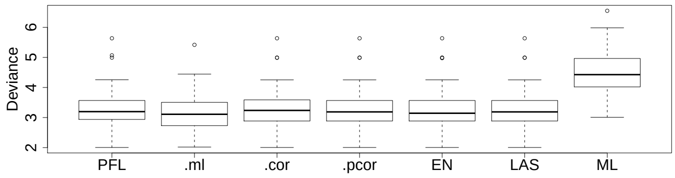

FIGURE 7: *Boxplots of the predictive deviance for the Bones Data Set*

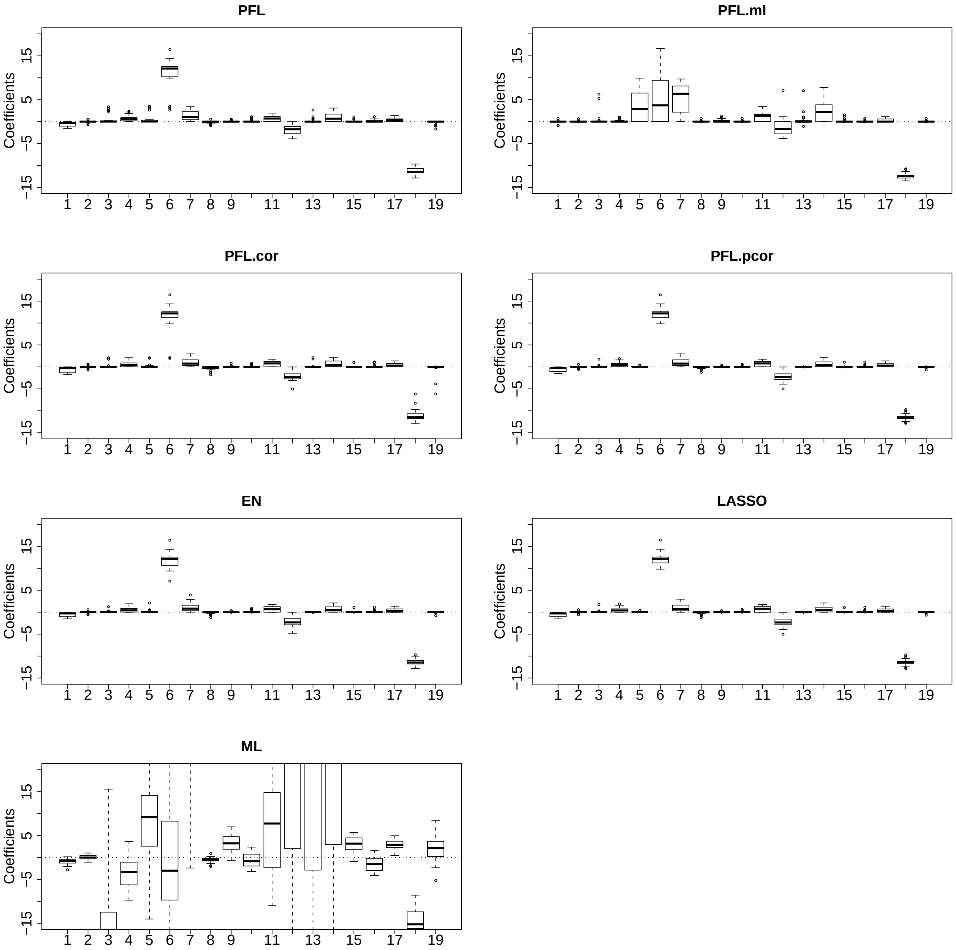

FIGURE 8: *Boxplots of the predictors for the Bones Data Set*

PFL	PFL.ml	PFL.cor	PFL.pcor	EN	LASSO	ML
3.1969 (0.9178)	3.1085 (0.7589)	3.2367 (0.9112)	3.1873 (0.8401)	3.1432 (0.8366)	3.1873 (0.9212)	4.4276 (0.8909)

Table 5: *The median of predictive deviance on test data for the Bones Data Set. The bootstrap variance (B=500) is bracketed.*

5 Concluding Remarks

We proposed a regularization method that enforces the grouping property by including pairwise differences of coefficients in the penalty term. It works for linear as well as generalized linear models and is strong competitor for the lasso and the elastic net. Although it uses fusion methodology it does not assume that a metric on predictors is available. Therefore it can used for common regression problems.

Acknowledgments

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

References

Anbari, M. E. and A. Mkhadri (2008). Penalized regression combining the l1 norm and a correlation based penalty. *INRIA Research Report 6746.*

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

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

Breiman, L. (1996). Heuristics of instability and stabilisation in model selection. *The Annals of Statistics 24*, 2350–2383.

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

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

Candes, E. and T. Tao (2007, DEC). The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics 35*(6), 2313–2351.

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

Fan, J. and R. Li (2001). Variable selection via nonconcave penalize likelihood and its oracle properties. *Journal of the American Statistical Association 96*, 1348–1360.

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

Gertheiss, J. and G. Tutz (2010). Sparse modeling of categorial explanatory variables. *The Annals of Applied Statistics 136*, 100–107.

Goemann, J. (2010). L1 penalized estimation in the cox proportinal hazards model. *Biometrical Journal 1*(52), 70–84.

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.

Osborne, M., B. Presnell, and B. Turlach (2000). On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 319–337.

Park, M. Y. and T. Hastie (2007). An l1 regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society* (69), 659–677.

Schäfer, J., R. Opgen-Rhein, and K. Strimmer (2009). *Efficient estimation of covariance and (partial) correlation.* R package version 1.5.3.

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

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

Tutz, G. and J. Ulbricht (2009). Penalized regression with correlation based penalty. *Statistics and Computing 19*, 239–253.

Ulbricht, J. (2010a). *lqa: Local quadratic approximation.* R package version 1.0-2.

Ulbricht, J. (2010b). *Variable selection in generalized linear models.* Dissertation, Ludwig-Maximilians-Universität, München.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S. Fourth edition*. Springer.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.

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