Stat Papers (2014) 55:187–207
DOI 10.1007/s00362-013-0516-z

REGULAR ARTICLE

Ultrahigh dimensional variable selection through the penalized maximum trimmed likelihood estimator

N. M. Neykov · P. Filzmoser · P. N. Neytchev

Received: 15 April 2012 / Revised: 9 December 2012 / Published online: 17 April 2013

Abstract The penalized maximum likelihood estimator (PMLE) has been widely used for variable selection in high-dimensional data. Various penalty functions have been employed for this purpose, e.g., Lasso, weighted Lasso, or smoothly clipped absolute deviations. However, the PMLE can be very sensitive to outliers in the data, especially to outliers in the covariates (leverage points). In order to overcome this disadvantage, the usage of the penalized maximum trimmed likelihood estimator (PMTLE) is proposed to estimate the unknown parameters in a robust way. The computation of the PMTLE takes advantage of the same technology as used for PMLE but here the estimation is based on subsamples only. The breakdown point properties of the PMTLE are discussed using the notion of d-fullness. The performance of the proposed estimator is evaluated in a simulation study for the classical multiple linear and Poisson linear regression models.

Keywords Multiple linear regression · Poisson regression · Robust variable screening · Breakdown point · Outlier detection · Maximum penalized trimmed likelihood estimator

N. M. Neykov (✉) · P. N. Neytchev
National Institute of Meteorology and Hydrology,
Bulgarian Academy of Sciences, Sofia, Bulgaria
e-mail: neyko.neykov@meteo.bg

P. N. Neytchev
e-mail: Plamen.Neytchev@meteo.bg

P. Filzmoser
Department of Statistics and Probability Theory,
Vienna University of Technology, Vienna, Austria
e-mail: P.Filzmoser@tuwien.ac.at

1 Introduction

Let (y_i, x_i^T), for $i = 1, \ldots, n$, be identically and independently distributed observations, where y_i is the ith observation of the response variable Y and $x_i^T = (x_{11}, x_{12}, \ldots, x_{1p})$ is the ith row of the covariates matrix X. Assume that y_i depends on x_i through a linear predictor $\eta_i(\theta) = x_i^T\theta$ via the objective function $L(\eta_i(\theta), y_i)$. For instance, $L(\eta_i(\theta), y_i)$ might be a probabilistic model such as likelihood, quasi-likelihood or another discrepancy function related with the ith observation. Without loss of generality, we shall assume that $L(\eta_i(\theta), y_i)$ is the log-likelihood. The Maximum Likelihood Estimator (MLE) is defined as

$$\hat{\theta}_{n,MLE} := \arg\max_{\theta} \left\{ \ell_n(\theta) = \sum_{i=1}^{n} L(\eta_i(\theta), y_i) \right\}. \tag{1}$$

The Penalized MLE (PMLE) is defined as

$$\hat{\theta}_{n,PMLE} := \arg\max_{\theta} \left\{ \ell_n(\theta) - n \sum_{j=1}^{p} p_\lambda(|\theta_j|) \right\}. \tag{2}$$

Here, $p_\lambda(.)$ is a penalty function indexed by the regularization parameter $\lambda \geq 0$. Due to the penalty function, some of the components of θ are shrunk to zero automatically and thus variables selection is performed. A large value of λ tends to choose a simple model whereas a small value of λ favors a complex model. In real applications the parameter λ is not known. It may be chosen by cross-validation or using an information criterion like the Bayesian Information Criterion (BIC), see Bühlmann and van der Geer (2011).

Commonly used penalty functions are the L_1 penalty $p_\lambda(|\theta_j|) = \lambda |\theta_j|$ called LASSO (least absolute shrinkage and selection operator) by Tibshirani (1996), the L_q-norm penalty $p_\lambda(|\theta|) = \lambda |\theta_j|^q$ for $0 < q \leq 2$, (Frank and Friedman 1993), the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li 2001), which is a quadratic spline

$$p_\lambda(|\theta|) = \begin{cases} \lambda|\theta| & \text{if } |\theta| < \lambda, \\ \frac{(a^2-1)\lambda^2 - (|\theta| - a\lambda)^2}{2(a-1)} & \text{if } \lambda \leq |\theta| < a\lambda, \\ \frac{(a+1)\lambda^2}{2} & \text{if } |\theta| \geq a\lambda, \end{cases} \tag{3}$$

where $a = 3.7$, or the minimum concavity penalty (MCP) $p'_\lambda(|\theta_j|) = (\lambda - |\theta|/a)_+$ considered by Zhang (2008). The SCAD and MCP are non-convex penalty functions which possess the oracle property. This means that the important variables can be correctly selected with a high probability whereas the remaining variables will be dropped from the model. Antoniadis et al. (2011) discussed about many other penalty functions and selection criteria for the regularization parameter λ for the generalized linear models (GLMs) framework.

The problem (2) is a convex optimization problem if the $\ell_n(\theta)$ is concave and the L_1 penalty is used. In general, for fixed parameter λ, the penalized likelihood with SCAD penalty function is non-convex and thus special algorithms have been developed to obtain a solution. For instance, Zou and Li (2008) propose an effective locally linear approximation algorithm (LLA) for optimization of (2) with the SCAD penalty function. The idea is to approximate (majorize) the SCAD function by a linear function at the mth iteration

$$p_\lambda(|\theta|) \approx p_\lambda(|\theta^{(m)}|) + p'_\lambda(|\theta^{(m)}|)(|\theta| - |\theta^{(m)}|). \tag{4}$$

As a consequence the penalized maximum likelihood (2) reduces to

$$\ell_n(\theta) - n\sum_{j=1}^{p} w_j^{(m)} \left|\theta_j\right|, \tag{5}$$

where $w_j^{(m)} = p'_\lambda(|\theta_j^{(m)}|)$. By the quadratic approximation of $\ell_n(\theta)$ at $\theta^{(m)}$ this optimization problem becomes weighted L_1 penalized least squares closely related with the adaptive LASSO estimation procedure (Zou 2006) that produce sparse fits and performs variable selection automatically. The LLA algorithm is implemented as a function in the R (R Development Core Team 2012) package SIS of Fan et al. (2009). Zou and Li (2008) discussed also other iterative approaches for solving the corresponding weighted L_1 penalized least squares problem efficiently by the least angle regression (LARS) algorithm (Efron et al. 2004).

For further consideration we use a well known result of Green (1984) concerning computational aspects of the MLE in fitting probabilistic regression models. Because the log-likelihood $\ell_n(\theta)$ is a composite function of the linear predictors $\eta_i(\theta)$, the Fisher scoring algorithm for maximization of $\ell_n(\theta)$ reduces to an iteratively reweighted least squares (IRLS) algorithm. Thus the optimization problem (5) at the $(m+1)$th iteration can be replaced by the following weighted least squares problem with weighted L_1 penalty

$$(z^{(m)} - X\theta)^T A^{(m)}(z^{(m)} - X\theta) + n\sum_{j=1}^{p} w_j^{(m)} \left|\theta_j\right|, \tag{6}$$

where $z = A^{-1}u + \eta$ is an adjusted dependent variable, $u = (\partial\ell_n(\theta)/\partial\eta)$, $\eta = X\theta$ and $A = (uu^T)$, and all these elements are evaluated at the current value $\theta^{(m)}$. The working weight matrix A is diagonal as the observations are independent.

Hence an efficient standard regression procedure with L_1 penalty, e.g., based on the LARS algorithm of Efron et al. (2004) or the coordinate descent algorithm (Friedman et al. 2007, 2010), can be adapted to calculate $\hat{\theta}_{n,PMLE}$ via an IRLS algorithm. A discussion about the applicability and implementation of these two approaches for the penalized logistic regression model with the LLA majorant (surrogate) of the SCAD penalty function is presented by Breheny and Huang (2011). Computational

algorithms within high-dimensional settings are discussed also in Bühlmann and van der Geer (2011), and also in the review paper of Fan and Lv (2010).

It is well known that the least squares estimator, the MLE and quasi-likelihood estimators can be highly sensitive to a small proportion of observations that departs from the model (Huber 1981; Hampel et al. 1986; Maronna et al. 2006). Therefore the penalized least squares estimator and MLE are non-robust against outliers in the data too. To overcome this problem, the penalized M-estimator has been employed (Fan and Li 2001; Fan and Lv 2010). However, within regression models, M-estimators are not robust against outlying observations in the covariates, the so called leverage points, and therefore penalized M-estimators are not robust in such settings as well. We remind that only some redescending M-estimators are robust in linear regression settings with fixed designs (Mizera and Müller 1999).

Several robust alternatives of the MLE that are robust simultaneously against outliers in the response and covariates have been developed, e.g., the weighted MLE of Markatou et al. (1997) and the maximum trimmed likelihood estimator (TLE) of Neykov and Neytchev (1990). To our knowledge, none of these estimators have been used in high dimensional data modeling. Thus, the goal of this paper is to develop an alternative of the penalized MLE for variable selection based on the penalized maximum TLE (PMTLE) in order to reduce the influence of the outliers in the covariates. The TLE is looking for that subsample of $k > n/2$ observations out of n with the optimal likelihood. The trimming number of observations can be chosen by the user in appropriate bounds to get a high breakdown point (BDP) and optimal efficiency. Because the TLE accommodates the classical MLE, the variable selection methodology, which is based mainly on the penalized MLE, can be adapted and further developed. In this paper the superiority of this approach in comparison with the penalized MLE is illustrated.

The paper is organized as follows. In Sect. 2 we define the generalized trimmed estimator (GTE), consider its penalized version and characterize its finite sample BDP. The applicability of the PMTLE is considered for the iterative sure independence screening (ISIS) framework of Fan et al. (2009) in Sect. 3. In Sect. 4 a simulation study is performed to illustrate the effectiveness of the proposed estimator in comparison with the ISIS procedure for the classical multiple and Poisson linear regression models. Finally, conclusions are given in Sect. 5.

2 Penalized maximum trimmed likelihood estimator

For introducing the penalized maximum trimmed likelihood estimator (PMTLE), we first need to review the definition and some properties of the GTE introduced by Vandev and Neykov (1998). Let $f_i : \Theta^p \to \mathbb{R}^+$, where $\Theta^p \subseteq \mathbb{R}^p$ is an open set.

Definition 1 The GTE $\hat{\theta}^k_{n,GTE}$ of θ is defined as the solution of the optimization problem

$$\hat{\theta}^{\mathrm{k}}_{\mathrm{n}}, \mathrm{GTE} := \arg\min_{\theta\in\Theta^p} \left\{ S_{n,k}(\theta) = \min_{I\in I_k} \sum_{i\in I} f_i(\theta) \right\}, \tag{7}$$

where I_k is the set of all k-subsets of the index set $\{1, \ldots, n\}$ and k is the trimming constant.

The trimming parameter k determines the robustness properties of the GTE as $n-k$ functions with the largest values of $f_i(\theta)$ are excluded from the loss function. The BDP of the GTE is not less than $\frac{1}{n}\min\{n-k, k-d\}$ if the set $F = \{f_i(\theta) : i = 1, \ldots, n\}$ is d-full. F is called d–full if for any subset of cardinality d of F, the supremum of this subset is a subcompact function. A real valued function $\varphi(\theta)$ is called subcompact if the sets $L_{\varphi(\theta)}(C) = \{\theta : \varphi(\theta) \leq C\}$ are contained in a compact set for every real constant C. Details can be found in Vandev and Neykov (1998), Müller and Neykov (2003), and Dimova and Neykov (2004). Thus, if one wants to study the BDP of the GTE, one has to find the fullness parameter d of F and then the BDP can be exemplified by the range of values of k. The BDP is maximized for $k = \lfloor (n+d+1)/2 \rfloor$, when it approximately equals $1/2$ for large n. Therefore, by selecting the value of k properly one can control the level of robustness of the GTE. Further, the asymptotic properties of the GTE estimator (7) were studied by Čížek (2008) for the case of twice differentiable functions $f_i(\theta)$.

The optimization problem (7) defining the GTE is of combinatorial nature,

$$\min_{\theta \in \Theta^p} S_{n,k}(\theta) = \min_{\theta \in \Theta^p} \min_{I \in I_k} \sum_{i \in I} f_i(\theta) = \min_{I \in I_k} \min_{\theta \in \Theta^p} \sum_{i \in I} f_i(\theta). \tag{8}$$

Therefore, it follows that all possible $\binom{n}{k}$ partitions of the set $\{f_1, \ldots, f_n\}$ have to be considered and $\hat{\theta}^{\mathrm{k}}_{\mathrm{n,GTE}}$ is defined by the partition with the minimal value of $S_{n,k}(\theta)$. Hence, an exact computation of the GTE is not feasible for large samples. To get an approximative GTE solution, an algorithm was developed by Neykov et al. (2012). It repeatedly (i) sets $s = 0$, selects a small subset $\{f_{i_1}, \ldots, f_{i_{k^*}}\}$ of k^* functions from F and forms $I_s = \{i_1, \ldots, i_{k^*}\}$, (ii) minimizes the objective function $\sum_{i \in I_s} f_i(\theta)$ with respect to θ, and uses the obtained estimate $\hat{\theta}_s$, (iii) sets $s = s + 1$, orders the functions of F in ascending order, $f_{\nu(1)}(\hat{\theta}_s) \leq f_{\nu(2)}(\hat{\theta}_s) \leq \ldots \leq f_{\nu(k)}(\hat{\theta}_s) \leq \ldots \leq f_{\nu(n)}(\hat{\theta}_s)$, where $\nu(.)$ is the permutation of the indices $\{1, 2, \ldots, n\}$, and forms $I_s = \{\nu(1), \ldots, \nu(k)\}$; the steps (ii) and (iii) are repeated as long as the newly obtained estimates $\hat{\theta}_s$ produce smaller values of the objective function $\sum_{i \in I_s} f_i(\theta)$.

The trial subsample size k^* should be greater than or equal to d, which is necessary for the existence of (7). However, the chance to get at least one good subsample of data points is larger if $k^* = d$. Obviously, only for very small samples all possible subsets of size $k^* = k$ can be considered to obtain the precise instead of an approximative solution. The algorithm could be further accelerated for large data sets by applying the partitioning and nesting techniques as discussed by Neykov et al. (2012).

Particular cases of the GTE are the least trimmed squares (LTS) estimator (Rousseeuw 1984) if $f(\theta)$ in (7) is replaced by the squared regression residuals, the least median of squares (Rousseeuw 1984), the maximum trimmed likelihood estimator (TLE) (Neykov and Neytchev 1990) if $f(\theta) = -L(\eta_i(\theta); y_i)$, the least trimmed quantile regression (Neykov et al. 2012), the extended trimmed quasi-likelihood estimator (Neykov et al. 2012), to name a few.

For high dimensional statistical optimization problems, where p is large in comparison with the sample size n, we need to consider a penalized version of the GTE.

Definition 2 The penalized GTE is defined as

$$\min_{\theta\in\Theta^p} S^P_{n,k}(\theta) = \min_{\theta\in\Theta^p}\left\{\min_{I\in I_k}\sum_{i\in I} f_i(\theta) + k\sum_{j=1}^{p} p_\lambda\left(\left|\theta_j\right|\right)\right\} \tag{9}$$

$$= \min_{I\in I_k}\left\{\min_{\theta\in\Theta^p}\left[\sum_{i\in I} f_i(\theta) + k\sum_{j=1}^{p} p_\lambda\left(\left|\theta_j\right|\right)\right]\right\} \tag{10}$$

$$= \min_{I\in I_k}\left\{\min_{\theta\in\Theta^p}\sum_{i\in I}\left[f_i(\theta) + \sum_{j=1}^{p} p_\lambda\left(\left|\theta_j\right|\right)\right]\right\}. \tag{11}$$

One can see that the penalized GTE refers to a penalized optimization problem, however, defined over all k-subsets. Thus the aforementioned algorithm can be used to obtain an approximate solution. For fixed λ, the BDP of the penalized GTE can be characterized via the d-fullness index of the set of functions $F_\lambda = \{f_i(\theta) + \sum_{j=1}^{p} p_\lambda(|\theta_j|),\ i = 1,\ldots,n\}$. Let F be d-full. Due to the inclusion

$$\left\{\theta\in R^p : \max_{j\in J}\left(f_j(\theta) + \sum_{l=1}^{p} p_\lambda(|\theta_l|)\right) \le C\right\} \subseteq \left\{\theta\in R^p : \max_{j\in J} f_j(\theta) \le C\right\}$$

it follows that F_λ is d-full because F is d-full. We see that the set F_λ is even 1-full provided the set $\{\theta\in R^p : \sum_{l=1}^{p} p_\lambda(|\theta_l|) \le C\}$ is contained in a compact set. For the convex penalty functions such as L_1 this is obvious whereas for the non-convex function such as SCAD the generalized d-fullness technique (Dimova and Neykov 2004) can be employed. From a computational point of view, the LLA defined by (4) can be used to get an approximate solution of the penalized GTE with SCAD penalty function. As the LLA is a convex majorant of the SCAD function, this ensures d-fullness of the corresponding set of functions F_λ. Therefore we conclude that a solution of the penalized GTE always exists if the set of functions F_λ is d-full. We note that this solution may not be unique, and thus additional conditions are required to achieve this.

From the penalized GTE definition it follows that when $k = n$, and for suitable choices of $f_i(\theta)$ and $p_\lambda(.)$, we can derive different penalized estimators such as the LASSO of Tibshirani (1996), the penalized L_1-likelihood of Tibshirani (1997), the penalized likelihood with the SCAD function of Fan and Li (2001), the LAD-LASSO of Wang et al. (2007), or the penalized M-estimator (Fan and Li 2001). The lack of robustness with respect to outlying leverage points in the regression framework is the main weakness of these estimators. Exceptions are the high BDP penalized MCD estimator (Croux and Haesbroeck 2010) and the penalized LTS estimator (Alfons et al. 2013) which are defined over subsamples. The last two estimators can also be

derived from the penalized GTE by substituting $f_i(.)$ with the Mahalanobis distances and squared regression residuals, respectively.

Definition 3 The PMTLE is defined as a particular case of the penalized GTE when the function $f_i(\theta)$ in (9) is replaced by the negative log-likelihood of the ith observation.

The PMTLE can attain the highest BDP provided the set F_λ of penalized negative log-likelihoods is d–full. As the set F_λ inherits the index of fullness of F, it is sufficient to derive the index of fullness of the set F comprised by the negative log-likelihoods.

We remind that in the classical settings, when $p < n$, the d–fullness indices of various sets of functions have been characterized. For instance, Vandev and Neykov (1993) determined the index of fullness $d = p$ for the set of p-variate normal distributions. Müller and Neykov (2003) related the index of fullness of the negative log-likelihoods sets of the linear logistic, Poisson and r-th power exponential distribution regression models with the quantity $\mathcal{N}(X)+1$, where $\mathcal{N}(X) = \max_{0\neq\theta\in\mathbb{R}^p} \text{card}\{i \in \{1,\ldots,n\}; x_i^T\theta = 0\}$ provides the maximum number of covariates $x_i \in R^p$ lying in a subspace, Müller (1995). If the observations x_i^T are linearly independent then $\mathcal{N}(X) = p-1$, and this is the minimal value for $\mathcal{N}(X)$. If the covariates are qualitative variables such as factors with several levels, then $\mathcal{N}(X)$ is much larger. Neykov et al. (2012) derived the index of fullness $d = \max(\mathcal{N}(X), \mathcal{N}(Z)) + 1$ of the set of extended quasi-log-likelihoods where X and Z are the mean and dispersion models covariates data matrices. Neykov et al. (2012) characterized the index of fullness $d = \mathcal{N}(X)+1$ of the quantile linear regression residuals set. Hence the indices of fullness of the corresponding F_λ sets with convex penalty functions are available for direct use. As consequence of this, the BDP of the PMTLE for the above probabilistic models equals $\frac{1}{n}\min\{n-k, k-\mathcal{N}(X)-1\}$. If the parameter of trimming k satisfies the inequalities $\lfloor\{n+\mathcal{N}(X)+1\}/2\rfloor \le k \le \lfloor\{n+\mathcal{N}(X)+2\}/2\rfloor$ the BDP is maximized and equals $\frac{1}{n}\lfloor\{n-\mathcal{N}(X)-1\}/2\rfloor$. Obviously, the BDP of the PMTLE can be small in modeling experimental data with qualitative (categorical) covariates. Thus the PMTLE is more suitable for data with continuous covariates.

Now the question is how to proceed with the characterization of the BDP in high-dimensional data when $p >> n$. As Bühlmann and van der Geer (2011) pointed out: "The philosophy that will generally rescue us, is to 'believe' that in fact only a few coordinates of the θ are non-zero". Armed with this 'belief' we postpone the BDP discussion of the PMTLE to the next section.

In order to reduce the outlier's influence on the selection of the penalization parameter λ we recommend the usage of the penalized BIC based on trimming, defined by $\text{PTBIC}(\lambda) = -2\log(S^P_{n,k}(\hat{\theta})) + df(\lambda)\log(k)$ where $S^P_{n,k}(\hat{\theta})$ is the PMTLE and $df(\lambda)$ are the model degrees of freedom given by the non-zero estimated components of $\hat{\theta}$. Obviously, PTBIC reduces to BIC if $k = n$ and $\lambda = 0$.

In the next section, the applicability of the PMTLE is investigated, and its BDP properties for the ultrahigh dimensional multiple linear regression and Poisson regression model are considered.

3 Robust SIS and ISIS based on trimming

The usage of penalization is limited in ultrahigh dimensional settings. According to Fan and Lv (2008), the ultrahigh dimensionality concerns with variable selection in the cases when p is much larger than n, i.e., $\log(p) = O(n^{\alpha})$ for some $0 < \alpha < 1$. In this section we focus on the so called sure independence screening (SIS) technique and its variations for variable selection developed by Fan et al. (2009). SIS is a preprocessing technique which aims at a drastic reduction of the number of covariates to a dimension less than the sample size by conventional marginal utility methods, with the hope to catch the most informative covariates, and then to use a penalization technique to select the carriers, see Fan and Lv (2010). Such a two-stage procedure is acceptable because the penalty based variable selection techniques work reasonably well with a moderate number of covariates. Fan and Lv (2008), Fan et al. (2009), and Fan and Lv (2010) have provided theoretical results that all important covariates can be selected by such a procedure with high probability. Unfortunately, the SIS techniques that rely on MLE, quasi-likelihood and robust M-estimators of Huber (1981), are not resistant against outliers in the covariates, and so their applicability is of limited use. This can be overcome by replacing these estimators by their high BDP counterparts based on trimming. The usage of the PMTLE for the classical multiple linear and Poisson regression models will be demonstrated in the following. In order to aid the presentation, we briefly review the SIS formulation, following closely (Fan et al. 2009).

3.1 Variable ranking by marginal utility

Without loss of generality, we shall assume that $L(.)$ is the negative log-likelihood, although other functions such as the quasi-likelihood, the least squares estimator can be used. Let us define the marginal utility of the jth covariate X_j, for $j = 1, \ldots, p$, by

$$L_0 = \min_{\theta_0} \frac{1}{n} \sum_{i=1}^{n} L(y_i, \theta_0), \tag{12}$$

$$L_j = \min_{\theta_0, \theta_j} \frac{1}{n} \sum_{i=1}^{n} L(y_i, \theta_0 + x_{ij}\theta_j), \tag{13}$$

where L_j is the loss function of using $\theta_0 + x_{ij}\theta_j$ to predict y_i.

The idea behind SIS is to compute the marginal utilities $L_1, \ldots, L_p$, rank them in ascending order, $L_{\nu(1)} \leq \ldots \leq L_{\nu(q)} \leq \ldots \leq L_{\nu(p)}$, where $(\nu(1), \ldots, \nu(p))$ is the permutation of the indices $(1, \ldots, p)$, and select the q-vector of covariates $(X_{\nu(1)}, X_{\nu(2)}, \ldots, X_{\nu(q)})$ for further consideration. In this way the covariate X_j is selected by SIS according to the magnitude of its marginal utility. Computing the L_j is fast as the fitting model has two parameters only, and so even for ultrahigh dimensional data this is not an intensive computational problem. Fan and Lv (2008) recommend to take $q = \lfloor n/\log n \rfloor$ for multiple regression and $q = \lfloor n/(2\log n) \rfloor$ for Poisson regression. The parameter q is usually chosen large enough but $q < n$ to

ensure the sure screening property. As q is specified in advance, only the q smallest marginal utilities have to be ordered, and an ordering of the remaining values is not required, hereby saving computation time. We note that the influence of θ_0 can be excluded by the marginal utility via the marginal likelihood ratio $LR_j = L_0 - L_j$ that assesses the increments of the log-likelihood and equals the deviance differences for GLMs. Obviously this will not change the ordering of L_j. For the multiple regression model this is equivalent to centering the dependent variable by its mean. On the other hand, the covariates have to be standardized to reduce the influence of their magnitude.

3.2 Penalized pseudo-likelihood

The subset of variables selected by SIS may still include many unimportant covariates. To improve performance, Fan et al. (2009), and Fan and Song (2010) recommend the usage of the penalized likelihood to further delete unimportant variables. By reordering the covariates if necessary, we may assume without loss of generality that $X_1, \ldots, X_q$ are the covariates recruited by SIS. Let $x_{i,q} = (x_{i1}, \ldots, x_{iq})^T$ and redefine $\theta = (\theta_1, \ldots, \theta_q)^T$. Minimization of the penalized log-likelihood

$$\frac{1}{n}\sum_{i=1}^{n} L(y_i, \theta_0 + x_{i,q}^T\theta) + \sum_{j=1}^{q} p_\lambda(|\theta_j|), \tag{14}$$

will yield a sparse regression parameter estimate θ, where the regularization parameter may be chosen by cross-validation. Let us denote the nonzero components of θ by $\widehat{\mathcal{M}}$.

Fan et al. (2009) refer to this two-stage procedure as SIS-Lasso or SIS-SCAD, depending on the choice of the penalty function. The screening stage solves only bivariate optimization, see (13), whereas the fitting part solves only the optimization problem (14) with moderate size q. This is an attractive feature in ultrahigh dimensional statistical learning.

3.3 Robust SIS-SCAD based on trimming

The two-stage SIS-SCAD estimation procedure are based on the MLE and penalized MLE which are not robust against outlying observations in the covariates in probabilistic regression models. A naive approach would be to replace the optimization problems (12), (13) and (14) by their counterparts based on trimming and to solve them separately to get the corresponding extremes keeping the trimming parameter k at the lowest possible levels to guarantee maximal BDP. This means that the GTE algorithm needs to be used in $p + 2$ separate optimization problems.

However, the GTE combinatorial optimization principle dictates that the two-stage SIS-SCAD estimation procedure has to be applied to all k-subsets in order to get that k-subset with the optimal value of the objective function (14). In this way we formally define the two-stage Trimmed SIS-SCAD (TSIS-SCAD) estimation procedure as follows:

$$\min_{I\in I_k}\begin{cases} \text{ISIS procedure} \\ \begin{cases} L^{trim}_{0,\widehat{\mathcal{M}}_1} = \min\limits_{\theta_0,\theta_{\widehat{\mathcal{M}}_1}} k^{-1}\sum\limits_{i\in I} L(y_i,\theta_0 + x^T_{i,\widehat{\mathcal{M}}_1}\theta_{\widehat{\mathcal{M}}_1}) \\ L^{(2,trim)}_j = \min\limits_{\theta_0,\theta_{\widehat{\mathcal{M}}_1},\theta_j} k^{-1}\sum\limits_{i\in I} L(y_i,\theta_0 + x^T_{i,\widehat{\mathcal{M}}_1}\theta_{\widehat{\mathcal{M}}_1} + x_{ij}\theta_j) \end{cases} \\ \text{ISIS}-\text{SCAD procedure} \\ \tilde{S}^{P,trim}_{k,n} = \min\limits_{\theta_0,\theta_{\widehat{\mathcal{M}}_1},\theta_{\widehat{\mathcal{A}}_2}} \left(k^{-1}\sum\limits_{i\in I} L(y_i,\theta_0 + x^T_{i,\widehat{\mathcal{M}}_1}\theta_{\widehat{\mathcal{M}}_1} + x^T_{i,\widehat{\mathcal{A}}_2}\theta_{\widehat{\mathcal{A}}_2}) \right. \\ \left. + \sum\limits_{j\in\widehat{\mathcal{M}}_1\cup\widehat{\mathcal{A}}_2} p_\lambda(|\theta_j|) \right) \end{cases} \tag{15}$$

Therefore, for all k-subsets the linked optimization problems (19) have to be solved subsequently and the penalized TSIS-SCAD estimate is defined by that k-subset with the minimal value of $S^{P,trim}_{k,n}$. Because this is not feasible for large data sets, an approximate estimate can be obtained by the use of the GTE algorithm. Obviously, the covariates have to be standardized using the means and standard deviations computed from each subset in order to reduce the influence of their magnitude as the GTE algorithm consists of optimization problems over data subsets.

An important choice for the algorithm is the trimming parameter k which controls the identifiability of the model parameters and determines the finite sample BDP of the estimator. Since here we will only consider simple linear and Poisson regression models, the d-fullness index of $F_j = \{L(y_i,\theta_0 + x_{ij}\theta_j)$ for $i = 1,\dots,n\}$ is $\mathcal{N}(X_j)+1$ for $j = 1,\dots,p$, according to Müller and Neykov (2003). Thus, the finite sample BDP of the TLE utility estimator equals $\frac{1}{n}\min\left\{n-k, k-\mathcal{N}(X_j)-1\right\}$, whereas for the penalized TSIS-SCAD estimator the finite sample BDP is $\frac{1}{n}\min\left\{n-k, k-\mathcal{N}(X_{n\times q})-1\right\}$, see Müller and Neykov (2003). Therefore, the finite sample BDP of the two-stage TSIS-SCAD estimation procedure equals $\frac{1}{n}\min\{n-k, k-D-1\}$ where $D = \max[\max_j \mathcal{N}(X_j), \mathcal{N}(X_{n\times q})]$. This BDP is maximized for $\lfloor\{n+D+1\}/2\rfloor \le k \le \lfloor\{n+D+2\}/2\rfloor$ and equals $\frac{1}{n}\lfloor\{n-D-1\}/2\rfloor$.

Instead of assigning a minimal value of the trimming parameter k to gain maximal BDP we prefer to take a subset of data of size $k = \lfloor\alpha n\rfloor$ for $\alpha \in (0.5, 1]$, provided all covariates are continuous. For instance, the choice $\alpha = 0.80$ ensures simultaneously a resistance against 20 % outliers in the data and leads to a higher efficiency of the estimator.

3.4 Iterative feature selection

Independent variable screening as it is done in the SIS procedure may have poor performance if variables are marginally weakly correlated with the response variable but jointly related with the response, or if a variable is jointly uncorrelated with the response but its marginal correlation with the response is higher than for some other

important variable. These problems are addressed by iterative SIS (ISIS) proposed by Fan and Lv (2008), Fan et al. (2009), and Fan and Song (2010) which incorporates the joint covariance information.

In the first step of ISIS, the two stage SIS-SCAD procedure is performed to select the subset $\widehat{\mathcal{M}}_1$ of covariates. Then Fan et al. (2009) propose to compute the following loss function in order to assess the importance of the covariate X_j which has not been included by the SIS-SCAD procedure:

$$L_j^{(2)} = \min_{\theta_0, \theta_{\widehat{\mathcal{M}}_1}, \theta_j} n^{-1} \sum_{i=1}^{n} L\left(y_i, \theta_0 + x_{i,\widehat{\mathcal{M}}_1}^T \theta_{\widehat{\mathcal{M}}_1} + x_{ij}\theta_j\right), \tag{16}$$

for $j \in \widehat{\mathcal{M}}_1^c = \{1, \ldots, p\} \setminus \widehat{\mathcal{M}}_1$, where $x_{i,\widehat{\mathcal{M}}_1}$ is the sub-vector of x_i consisting of those elements in $\widehat{\mathcal{M}}_1$.

The optimization problem (16) is low-dimensional and thus easy to solve. The additional contribution of variable X_j given the existence of variables in $\widehat{\mathcal{M}}_1$ can be assessed by the marginal likelihood ratio test (difference by the two deviance functions for the GLM setting):

$$L_j^{LR} = \min_{\theta_0, \theta_{\widehat{\mathcal{M}}_1}} n^{-1} \sum_{i=1}^{n} L(y_i, \theta_0 + x_{i,\widehat{\mathcal{M}}_1}^T \theta_{\widehat{\mathcal{M}}_1}) - L_j^{(2)}. \tag{17}$$

After ordering of L_j^{LR} in ascending order for $j \in \widehat{\mathcal{M}}_1^c$ we take the indices corresponding to the smallest m_2 elements and form the set $\widehat{\mathcal{A}}_2$.

The above pre-screening step is followed by the penalized likelihood for obtaining a sparse estimate

$$\theta_2 = \underset{\theta_0, \theta_{\widehat{\mathcal{M}}_1}, \theta_{\widehat{\mathcal{A}}_2}}{\arg\min} \left(n^{-1} \sum_{i=1}^{n} L(y_i, \theta_0 + x_{i,\widehat{\mathcal{M}}_1}^T \theta_{\widehat{\mathcal{M}}_1} + x_{i,\widehat{\mathcal{A}}_2}^T \theta_{\widehat{\mathcal{A}}_2}) + \sum_{j \in \widehat{\mathcal{M}}_1 \cup \widehat{\mathcal{A}}_2} p_\lambda(|\theta_j|) \right). \tag{18}$$

As a result we obtain a new estimated set $\widehat{\mathcal{M}}_2$ of active indices consisting of those indices of θ_2 that are non-zero. Thus, this procedure allows to delete variables from the previously selected features with indices in $\widehat{\mathcal{M}}_1$. The process, which iteratively recruits and deletes features, can then be repeated until we obtain a set of indices $\widehat{\mathcal{M}}_l$ which either has reached the prescribed size q, or satisfies $\widehat{\mathcal{M}}_l = \widehat{\mathcal{M}}_{l-1}$. In this way a final estimated parameter vector θ_l is obtained.

In their R package *SIS*, Fan et al. (2009) chose $k_1 = \lfloor 2q/3 \rfloor$, and thereafter at the rth iteration, they take $m_r = q - |\widehat{\mathcal{M}}_{r-1}|$. This ensures that the iterated versions of SIS take at least two iterations to terminate.

3.5 Robust iterated variable selection based on trimming

Similar to robustifying the two-stage SIS-SCAD estimation procedure, we could replace the optimization problems (16) and (18) by their counterparts based on trim-

ming and solve them for all k-subsets out of n cases in order to get that k-subset with the optimal objective value of (18). This way we formally define the two-stage Trimmed ISIS-SCAD (TISIS-SCAD) estimation procedure as

$$\min_{I\in I_k}\begin{cases}\text{ISIS procedure}\\ \begin{cases} L^{trim}_{0,\widehat{\mathcal{M}}_1}=\min\limits_{\theta_0,\theta_{\widehat{\mathcal{M}}_1}} k^{-1}\sum\limits_{i\in I} L(y_i,\theta_0+x^T_{i,\widehat{\mathcal{M}}_1}\theta_{\widehat{\mathcal{M}}_1})\\ L^{(2,trim)}_j=\min\limits_{\theta_0,\theta_{\widehat{\mathcal{M}}_1},\theta_j} k^{-1}\sum\limits_{i\in I} L(y_i,\theta_0+x^T_{i,\widehat{\mathcal{M}}_1}\theta_{\widehat{\mathcal{M}}_1}+x_{ij}\theta_j)\end{cases}\\ \text{ISIS}-\text{SCAD procedure}\\ \tilde{S}^{P,trim}_{k,n}=\min\limits_{\theta_0,\theta_{\widehat{\mathcal{M}}_1},\theta_{\widehat{\mathcal{A}}_2}}\left(k^{-1}\sum\limits_{i\in I} L(y_i,\theta_0+x^T_{i,\widehat{\mathcal{M}}_1}\theta_{\widehat{\mathcal{M}}_1}+x^T_{i,\widehat{\mathcal{A}}_2}\theta_{\widehat{\mathcal{A}}_2})+\sum_{j\in\widehat{\mathcal{M}}_1\cup\widehat{\mathcal{A}}_2} p_\lambda(|\theta_j|)\right)\end{cases}\tag{19}$$

Therefore for all k-subsets the linked optimization problems (19) have to be solved subsequently and the penalized TISIS-SCAD estimate is defined by that k-subset with the minimal value of $\tilde{S}^{P,trim}_{k,n}$. This procedure would not be computationally feasible for larger data sets, and therefore an approximate estimate can be obtained by the use of the GTE algorithm. Again the variable standardization has to be done within the subsets.

Let $r=|\widehat{\mathcal{M}}_1\cup\widehat{\mathcal{A}}_2|$ be the cardinality of $\widehat{\mathcal{M}}_1\cup\widehat{\mathcal{A}}_2$ and $\widehat{\mathcal{M}}^*_1=\widehat{\mathcal{M}}_1+1$. Similar to the previous section we can conclude that the corresponding utility sets are $(\mathcal{N}(X_{n\times r})+1)$ and $(\mathcal{N}(X_{n\times\widehat{\mathcal{M}}^*_1})+1)$ full, and these are the minimal numbers of observations that guarantee identifiability of θ (Müller and Neykov 2003). Hence, the finite sample BDP of the TLE utility estimator defined by (19) equals $\frac{1}{n}\min\left\{n-k,k-\mathcal{N}(X_{n\times\widehat{\mathcal{M}}^*_1})-1\right\}$ whereas for the penalized maximum trimmed ISIS-SCAD estimator it is $\frac{1}{n}\min\{n-k,k-\mathcal{N}(X_{n\times r})-1\}$. Using the notation $\tilde{D}=\max[\mathcal{N}(X_{n\times\widehat{\mathcal{M}}^*_1}),\mathcal{N}(X_{n\times r})]$, the BDP of the two-stage TISIS-SCAD estimation procedure (19) equals $\frac{1}{n}\min\left\{n-k,k-\tilde{D}-1\right\}$. This BDP is maximized for $\left\lfloor\left\{n+\tilde{D}+1\right\}/2\right\rfloor\le k\le\left\lfloor\left\{n+\tilde{D}+2\right\}/2\right\rfloor$ and equals $\frac{1}{n}\left\lfloor\left\{n-\tilde{D}-1\right\}/2\right\rfloor$.

As mentioned above, one can select $k=\lfloor\alpha n\rfloor$ with $\alpha=0.80$, for instance.

4 Simulation study

In this section, we study the performance of SIS-SCAD, ISIS-SCAD and their trimmed counterparts on simulated data for the multiple and Poisson linear regression framework. Two different data configurations are presented and discussed.

4.1 Performance measures

According to the simulation designs described in the next sections we generate training data without and with contamination, and estimate the regression parameters θ with the different methods. In addition, n test set observations are generated according to the same scheme but without outliers. We denote the test set covariates by $\tilde{x}_i$ and the response by $\tilde{y}_i$, for $i = 1, \ldots, n$. The predictions $\tilde{\eta}_i = \tilde{x}_i^T \hat{\theta}$ for the linear regression model, and $\log(\tilde{\eta}_i) = \tilde{x}_i^T \hat{\theta}$ for Poisson regression are evaluated by the root mean squared error of prediction (RMSEP),

$$\text{RMSEP}(\hat{\theta}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i - \tilde{\eta}_i)^2}.$$

The RMSEP is computed for each estimator and simulated test data set, and we report averages and medians over all simulations. Further, we compare also with the so called oracle estimator, where the true regression coefficients θ are used for the evaluation.

We evaluate the methods also according to their ability to select the correct variables, using the false positive rate (FPR) and the false negative rate (FNR). False positives refer to variables that are selected by the method, while their coefficients in the simulation design are zero. In contrast, a false negative is a coefficient estimated as zero, while it was generated as non-zero. Formally, FPR and FNR can be defined as

$$FPR(\hat{\theta}) = \frac{|\{j \in \{1, \ldots, p\} : \hat{\theta}_j \neq 0 \wedge \theta_j = 0\}|}{\{j \in \{1, \ldots, p\} : \theta_j = 0\}} \tag{20}$$

$$FNR(\hat{\theta}) = \frac{|\{j \in \{1, \ldots, p\} : \hat{\theta}_j = 0 \wedge \theta_j \neq 0\}|}{\{j \in \{1, \ldots, p\} : \theta_j \neq 0\}} \tag{21}$$

These rates are computed for each simulated data set, and we will report average numbers over all simulations. The better the sparseness structure is identified by the method, the smaller these rates should be.

In order to compare the simulation results with those of Fan et al. (2009) for the Poisson regression model, we also report the median values of the evaluation measures $||\theta - \hat{\theta}||_1 = \sum_{i=0}^{p} |\theta_j - \hat{\theta}_j|$ and $||\theta - \hat{\theta}||_2 = (\sum_{i=0}^{p}(\theta_j - \hat{\theta}_j)^2)^{1/2}$, the AIC - Akaike's information criterion, and the BIC—Bayesian information criterion.

4.2 Simulation design: multiple linear regression

We use the third simulation design considered in Alfons et al. (2013) where the sparse LTS regression estimator with L_1 penalty (L1-penalized trimmed LTS, trimmed LASSO) was introduced. We compare their estimator with the SIS-SCAD and its trimmed version TSIS-SCAD, because SIS-SCAD exhibits better performance than SIS-LASSO according to the simulation study (without contamination) of Fan et al.

(2009). We note that Fan et al. (2009) denote SIS-SCAD and ISIS-SCAD as Van-SIS and Van-ISIS.

In this setting, we generate $n = 100$ observations from a p-dimensional normal distribution $N_p(0, \Sigma)$, with $p = 1,000$. The covariance matrix $\Sigma = (\Sigma_{ij})_{1 \le i,j \le p}$ is given by $\Sigma_{ij} = 0.5^{|i-j|}$, creating correlated predictor variables. The coefficient vector $\theta = (\theta_1, \ldots, \theta_p)^T$ has components $\theta_1 = \theta_7 = 1.5$, $\theta_2 = 0.5$, $\theta_4 = \theta_{11} = 1$, and $\theta_j = 0$ for $j \in \{1, \ldots, p\} \backslash \{1, 2, 4, 7, 11\}$.

The response variable is generated according to the multiple linear regression model $y_i = x_i^T \theta + \varepsilon_i$, where the error terms ε_i follow a normal distribution with $\mu = 0$ and $\sigma = 0.5$. We apply the same contamination scheme as Alfons et al. (2013), see also Khan et al. (2007), who proposed:

1. No contamination
2. Vertical outliers: 10 % of the errors terms in the regression model follow a normal $N(20, \sigma^2)$, instead of a $N(0, \sigma^2)$.
3. Leverage points: Same as in 2., but the 10 % contaminated observations contain high-leverage values, by drawing the predictor variables from independent $N(50, 1)$ distributions.

The results of the simulation experiment are given in Table 1. The first and second row of this table are taken from Tables 3 of Alfons et al. (2013) in order to make a comparison. *L1-LTSraw* is the result of the L1-penalized trimmed LTS procedure, and *L1-LTS* is a reweighted version of the estimator (see Alfons et al. 2013). The means (*mean*) and medians (*med*), respectively, of the RMSEP, FPR and FNR over 500 simulation runs are reported for every method; ISIS-SCAD is denoted by *ISIS*, and its trimmed version by *TISIS-XX*, where *XX* shows the percentage of trimming - 10, 20, 25.

The results based on the means and medians are almost the same in our simulation experiments. Larger differences could refer to possible problems with the algorithm. We see that the performance of the ISIS-SCAD estimator is excellent for the scenario without contamination, and the RMSEP is close to the oracle estimator. However, ISIS-SCAD breaks down in the presence of vertical outliers or leverage points, whereas the robust methods L1-LTS and TISIS are stable. TISIS shows excellent performance: the RMSEP is close to the oracle estimator, and the false positive and false negative rates are very small. Moreover, the different trimming percentages result in about the same performance.

4.3 Simulation design: Poisson regression

The simulation configurations of this section are the same as in Fan et al. (2009). The following three settings of covariates $X_1, \ldots, X_p$ and regression coefficients $\theta_0, \theta_1, \ldots, \theta_p$, for $p = 1,000$ and sample size $n = 200$ are generated:

1. $X_1, \ldots, X_p$ are independent and identically distributed $N(0, 1)$ random variables; $\theta_0 = 5$, $\theta_1 = -0.5423$, $\theta_2 = -0.5314$, $\theta_3 = -0.5012$, $\theta_4 = -0.4850$, $\theta_5 = -0.4133$, $\theta_6 = -0.5234$, and $\theta_j = 0$ for $j > 6$;

Table 1 Results for the simulation scheme in the multiple linear regression case, where $n = 100$ and $p = 1{,}000$

Method	No contamination			Vertical outliers			Leverage points		
	RMSEP	FPR	FNR	RMSEP	FPR	FNR	RMSEP	FPR	FNR
L1-LTSraw	0.79	0.02	0.00	0.74	0.02	0.00	0.72	0.02	0.00
L1-LTS	0.74	0.01	0.00	0.70	0.01	0.00	0.70	0.02	0.00
ISIS (mean)	0.53	0.00	0.00	4.89	0.01	0.75	2.17	0.01	0.33
ISIS (med)	0.52	0.00	0.00	4.88	0.01	0.79	2.13	0.01	0.40
TISIS-10 (mean)	0.53	0.00	0.00	0.55	0.00	0.00	0.55	0.00	0.00
TISIS-20 (mean)	0.53	0.00	0.00	0.56	0.00	0.01	0.57	0.00	0.03
TISIS-25 (mean)	0.53	0.00	0.00	0.59	0.00	0.02	0.58	0.00	0.04
TISIS-10 (med)	0.52	0.00	0.00	0.53	0.00	0.00	0.53	0.00	0.00
TISIS-20 (med)	0.52	0.00	0.00	0.54	0.00	0.00	0.54	0.00	0.00
TISIS-25 (med)	0.52	0.00	0.00	0.55	0.00	0.00	0.56	0.00	0.00
Oracle	0.50								

The means and medians of RMSEP, FPR and FNR over 500 simulation runs are reported for every method: L1-LTSraw and L1-LTS refer to the raw and weighted penalized LTS regression estimator of Alfons et al. (2013), respectively, ISIS and TISIS (with the percentage of trimming) corresponds to the original and trimmed version of ISIS-SCAD, respectively, and Oracle uses the true regression parameters

Table 2 Poisson regression, Case 1 of the simulation scheme with 0, 10 and 20 % of contamination by vertical outliers (VO), $n = 200$ and $p = 1{,}000$

	0 % Cont.	VO-10 % contamination			VO-20 % contamination		
	ISIS	ISIS	TISIS-10	TISIS-20	ISIS	TISIS-20	TISIS-30
$\|\|\theta - \hat{\theta}\|\|_1$	0.12	3.58	0.13	0.17	4.83	0.14	0.18
$\|\|\theta - \hat{\theta}\|\|_2$	0.03	0.99	0.03	0.05	1.36	0.04	0.05
FPR	0.01	0.01	0.01	0.01	0.01	0.01	0.01
FNR	0	0	0	0	0.17	0	0
AIC	1544.82	*	1393.09	1175.19	*	1232.82	1022.87
AICt	1666.58	26502.49	1675.22	1749.27	*	1685.46	1786.56
BIC	1607.49	*	1453.76	1233.61	*	1291.25	1078.26
BICt	1729.24	26563.51	1737.89	1811.93	*	1748.13	1849.23
RMSPE.t	24.84	385.37	26.22	32.28	493.55	28.4	36.74
RMSEP.o	17.38						

2. $X_1, \ldots, X_p$ are jointly Gaussian, marginally $N(0, 1)$, and with $corr(X_i, X_4) = 1/\sqrt{2}$ for all $i \neq 4$ and $corr(X_i, X_j) = 1/2$ if i and j are distinct elements of $\{1, \ldots, p\}\backslash\{4\}$;
   $\theta_0 = 5, \theta_1 = \theta_2 = \theta_2 = 0.6, \theta_4 = -0.9\sqrt{2}$; and $\theta_j = 0$ for $j > 4$;
3. $X_1, \ldots, X_p$ are jointly Gaussian, marginally $N(0, 1)$, and with $corr(X_i, X_5) = 0$ for all $i \neq 5$, $corr(X_i, X_4) = 1/\sqrt{2}$ for all $i \notin \{4, 5\}$, and $corr(X_i, X_j) = 1/2$ if

Table 3 Poisson regression, Case 1 of the simulation scheme with 0, 10 and 20 % of contamination by leverage points (LP), $n = 200$ and $p = 1{,}000$

	0 % Cont.	LP-10 % contamination			LP-20 % contamination		
	ISIS	ISIS	TISIS-10	TISIS-20	ISIS	TISIS-20	TISIS-30
$\|\|\theta - \hat{\theta}\|\|_1$	0.12	3.84	0.14	0.17	3.92	0.15	0.21
$\|\|\theta - \hat{\theta}\|\|_2$	0.03	1.41	0.04	0.05	1.42	0.04	0.06
FPR	0.01	0.01	0.01	0.01	0.01	0.01	0.01
FNR	0	0.83	0	0	1	0	0
AIC	1544.82	*	1381.66	1179.74	*	1228.92	1033.77
AICt	1666.58	*	1691.73	1727.89	*	1693.31	1770
BIC	1607.49	*	1442.22	1237.93	*	1286.54	1089.67
BICt	1729.24	*	1754.4	1790.56	*	1754.99	1832.67
RMSPE.t	24.84	493.32	27.61	31.63	511.07	29.26	43.89
RMSEP.o	17.38						

i and j are distinct elements of $\{1, \ldots, p\}\backslash\{4, 5\}$;
$\theta_0 = 5, \theta_1 = \theta_2 = \theta_2 = 0.6, \theta_4 = -0.9\sqrt{2}, \theta_5 = 0.15$, and $\theta_j = 0$ for $j > 5$.

The first case with independent predictors is the simplest situation for variable selection. Here, the coefficients $\theta_1, \ldots, \theta_6$ were generated as $\left(\frac{\log n}{\sqrt{n}} + |Z|/8\right) U$ with $Z \sim N(0, 1)$ and $U = 1$ with probability 0.5 and $U = -1$ with probability 0.5, independently of Z. The last two cases are more complicated because of serial correlations. Even more, although $\theta_4 \neq 0$, the choices of the other regression coefficients in Cases 2 and 3 ensure that $\text{corr}(X_4, Y) = 0$, which makes variable selection more difficult. The coefficient $\theta_0 = 5$ is used to control an appropriate signal-to-noise ratio.

The data (x_i^T, y_i) for $i = 1, \ldots, 200$ are independent copies of a pair where y_i is conditionally on x_i distributed as Poisson$(\mu(x))$, where $\log(\mu(x)) = \theta_0 + x_i^T \theta$.

We apply the following contamination scheme:

1. No contamination
2. Vertical outliers: 10 and 20 % data contamination is introduced by changing respectively the first 20 and 40 observations to $y_i := y_i + \exp(7)$, for $i = 1, \ldots, 20$, respectively 40.
3. Leverage points: 10 and 20 % data contamination is introduced by modifying respectively the first 20 and 40 rows of the covariates matrix according to $x_{ij} := -3B_j \text{sign}(x_{ij})$ for $i = 1, \ldots, 20$, where $B_j = \max_{1 \leq i \leq n} (|x_{ij}|)$ for $j = 1, \ldots, p$.

Following the suggestion of Fan et al. (2009), we perform the computation for ISIS-SCAD and TISIS-SCAD with $q = \left\lfloor \frac{n}{2 \log n} \right\rfloor = 18$ as a sensible choice based on asymptotic results. The final regularization parameter for the SCAD penalty was chosen via 10-fold cross-validation as recommended by Fan et al. (2009). However, the BIC is used to choose the SCAD regularization parameter at each intermediate stage of the ISIS procedures in the three cases.

Table 4 Poisson regression, Case 2 of the simulation scheme with 0, 10 and 20 % of contamination by vertical outliers (VO), $n = 200$ and $p = 1{,}000$

	0 % Cont.	VO-10 % contamination			VO-20 % contamination		
	ISIS	ISIS	TISIS-10	TISIS-20	ISIS	TISIS-20	TISIS-30
$\|\|\theta - \hat{\theta}\|\|_1$	0.26	5.54	0.28	0.32	6.54	0.29	0.33
$\|\|\theta - \hat{\theta}\|\|_2$	0.07	1.66	0.08	0.09	1.88	0.08	0.1
FPR	0.01	0.02	0.01	0.01	0.02	0.01	0.01
FNR	0	0.25	0	0	0.5	0	0
AIC	1535.93	*	1381.11	1174.58	*	1226.36	1024.64
AICt	1674.54	26274.8	1683.06	1703.42	38396.37	1686.06	1732.16
BIC	1598.6	*	1441.77	1233	*	1284.79	1080.53
BICt	1737.21	26334.17	1745.73	1766.09	38459.04	1748.73	1792.69
RMSPE.t	17.52	212.38	17.59	18.9	291.46	18.29	19.83
RMSPE.o	13.79						

The estimators were applied to the training data and evaluated on the test data with $n = 200$ observations, which were generated according to the same schemes without contamination. For the TISIS-SCAD procedure we report the result for different trimming percentages. In the tables 2–7, we report several performance measures, all of which are based on 100 Monte Carlo repetitions. The tables contain the medians of these measures. The first two rows give the estimation errors $||\theta - \hat{\theta}||_1$ and $||\theta - \hat{\theta}||_2$, respectively, evaluated for the training data. In the 3^{rd} and 4^{th} row we report the FPR and FNR, respectively, for the training data. The fifth, sixth, seventh and eighth rows give Akaike's information criterion (Akaike 1974), AIC, and the Bayesian information criterion (Schwartz 1978), BIC, computed over the training and test (indicated by the additional "t") data. The last two rows give the RMSEP for the test data (RMSEP.t) and the true regression parameter (RMSEP.o). The symbols "*" in the tables refer to very big values greater than 250,000. Two consecutive tables are used for one simulation setting, where the first table contains the results for the vertical outliers, and the second table is for the leverage points.

For the simulation experiments without contamination, our results for ISIS-SCAD closely follow those based on Van-ISIS presented at Tables 5, 6 and 7 of Fan et al. (2009). In case of contamination (vertical outliers or leverage points) we see that the ISIS-SCAD estimator fails; all error measures are (much) worse, independent of the simulation scheme. An exception is the FPR, which means that in case of contamination the correct zero-coefficients are indeed set to zero. However, since FNR increases considerably, many non-zero coefficients are also set to zero. The robust version TISIS-SCAD shows excellent behavior for all simulation schemes, and for uncontaminated and contaminated data. Generally, the results are close to the ISIS-SCAD estimator when no contamination is present. Remarkable are the results for FPR and FNR of TISIS-SCAD, which are not higher than 1 % in all scenarios.

Table 5 Poisson regression, Case 2 of the simulation scheme with 0, 10 and 20 % of contamination by leverage points (LP), $n = 200$ and $p = 1{,}000$

	0 % Cont.	LP-10 % contamination			LP-20 % contamination		
	ISIS	ISIS	TISIS-10	TISIS-20	ISIS	TISIS-20	TISIS-30
$\|\|\theta - \hat{\theta}\|\|_1$	0.26	3.41	0.28	0.32	3.46	0.3	0.32
$\|\|\theta - \hat{\theta}\|\|_2$	0.07	1.66	0.08	0.09	1.66	0.09	0.09
FPR	0.01	0.01	0.01	0.01	0.01	0.01	0.01
FNR	0	0.75	0	0	0.75	0	0
AIC	1535.93	17359.63	1379.97	1174.5	16466.75	1226.01	1027.9
AICt	1674.54	18796.81	1680.45	1704.82	19419.59	1697.08	1716.28
BIC	1598.6	17389.32	1440.63	1232.54	16521.18	1284.44	1083.79
BICt	1737.21	18834.75	1743.11	1767.49	19468.78	1759.75	1778.95
RMSPE.t	17.52	157.28	17.66	18.41	159.36	18.22	19.34
RMSEP.o	13.79						

Table 6 Poisson regression, Case 3 of the simulation scheme with 0, 10 and 20 % of contamination by vertical outliers (VO), $n = 200$ and $p = 1{,}000$

	0 % Cont.	VO-10 % contamination			VO-20 % contamination		
	ISIS	ISIS	TISIS-10	TISIS-20	ISIS	TISIS-20	TISIS-30
$\|\|\theta - \hat{\theta}\|\|_1$	0.26	5.63	0.27	0.31	6.6	0.29	0.33
$\|\|\theta - \hat{\theta}\|\|_2$	0.07	1.65	0.08	0.09	1.9	0.09	0.1
FPR	0.01	0.02	0.01	0.01	0.02	0.01	0.01
FNR	0	0.4	0	0	0.6	0	0
AIC	1539.62	*	1384.54	1177.33	*	1231.26	1031.1
AICt	1674.91	26836.25	1683.5	1705.71	47284.46	1689.55	1729.97
BIC	1602.29	*	1445.2	1235.75	*	1289.36	1086.99
BICt	1737.57	26898.92	1746.17	1768.38	47345.48	1752.22	1792.64
RMSPE.t	17.58	214.84	17.95	18.94	288.5	18.47	19.8
RMSEP.o	13.91						

5 Summary and conclusions

We introduced a robust version of the penalized MLE based on the idea of trimming and characterized its BDP based on the notion of d-fullness. The finite sample properties of the proposed estimator were studied via an extended simulation study within high-dimensional multiple and Poisson linear regression settings. The new estimator generally performs very well, which is confirmed by the simulation experiments and by a comparison to other proposals. To handle the computations, the SIS/ISIS procedure of Fan et al. (2009) was used. However, any other procedure that implements penalization/regularization techniques can be employed instead. The computation of

Table 7 Poisson regression, Case 3 of the simulation scheme with 0, 10 and 20 % of contamination by leverage points (LP), $n = 200$ and $p = 1{,}000$

	0 % Cont.	LP-10 % contamination			LP-20 % contamination		
	ISIS	ISIS	TISIS-10	TISIS-20	ISIS	TISIS-20	TISIS-30
$\|\|\theta - \hat{\theta}\|\|_1$	0.26	3.58	0.27	0.31	3.65	0.31	0.31
$\|\|\theta - \hat{\theta}\|\|_2$	0.07	1.67	0.07	0.09	1.67	0.09	0.09
FPR	0.01	0.01	0.01	0.01	0.01	0.01	0.01
FNR	0	0.8	0	0	1	0	0
AIC	1539.62	17816.74	1385.37	1181.93	16620.12	1229.87	1031
AICt	1674.91	19809.14	1677.79	1704.42	20470.47	1697.85	1724.25
BIC	1602.29	17849.23	1446.04	1240.27	16672.89	1288.3	1086.89
BICt	1737.57	19835.52	1740.46	1767.09	20526.54	1760.52	1786.92
RMSPE.t	17.58	164.91	17.86	18.84	166.76	19.06	19.96
RMSEP.o	13.91						

the estimator is taking advantage of the same technology as used for its classical counterpart, but here the estimation is based on subsamples only. The used algorithm consisting of a trial and a refinement step (Neykov et al. 2012) follows the ideas of the FAST-LTS algorithm of Rousseeuw (1999), Müller and Neykov (2003). An important choice for estimators based on trimming is the trimming percentage. In the numerical experiments, it has been shown that a trimming percentage lower than the contamination level can lead to very poor estimates, but any higher trimming percentage gives very reasonable results. Therefore, a general rule is to work with a conservative choice of the trimming percentage or to estimate the amount of trimming similarly to Čížek (2010), and Gervini and Yohai (2002).

Acknowledgments The authors are thankful to the Vienna University of Technology and the ESF (COST Action IC0702) for supporting the stay of N. Neykov and P. Neytchev in Vienna.

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19:716–723

Alfons A, Croux C, Gelper S (2013) Sparse least trimmed squares regression. Ann Appl Stat 7:226–248. doi:10.1214/12-AOAS575

Antoniadis A, Gijbels I, Nikolova M (2011) Penalized likelihood regression for generalized linear models with non-quadratic penalties. Ann Inst Stat Math 63:585–615. doi:10.1007/s10463-009-0242-4

Breheny P, Huang J (2011) Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. Ann Appl Stat 5:232–253. doi:10.1214/10-AOAS388

Bühlmann P, van der Geer S (2011) Statistics for high dimensional data: methods theory and applications. Springer, New York

Čížek P (2008) General trimmed estimation: robust approach to nonlinear and limited dependent variable models. Econom Theory 24:1500–1529

Čížek P (2010) Reweighted least trimmed squares: an alternative to one-step estimators. Center Discussion Paper 2010/91, Tilburg University, The Netherlands

Croux C, Haesbroeck G (2010) Robust scatter regularization. Compstat, Book of Abstracts, Paris: Conservatoire National des Arts et Metiers (CNAM) and the French National Institute for Research in Computer Science and Control (INRIA)
Dimova R, Neykov NM (2004) Generalized d-fullness technique for breakdown point study of the trimmed likelihood estimator with applications. In: Hubert M, Pison G, Struyf A, Van Aelst S (eds) Theory and applications of recent robust methods. Birkhäuser, Basel, pp 83–91
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 Am Stat Assoc 96:1348–1360
Fan J, Lv J (2008) Sure independence screening for ultrahigh dimensional feature space (with discussion). J Royal Stat Soc B 70:849–911
Fan J, Lv J (2010) A selective overview of variable selection in high dimensional feature space. Statistica Sinica 20:101–148
Fan J, Samworth R, Wu Y (2009) Ultrahigh dimensional variable selection: beyond the linear model. J Mach Learn Res 10:1829–1853
Fan J, Song R (2010) Sure independence screening in generalized linear models with NP-dimensionality. Ann Stat 38:3567–3604
Frank IE, Friedman JH (1993) A statistical view of some chemometrics regression tools (with discussion). Technometrics 35:109–148
Friedman J, Hastie T, Höfling H, Tibshirani R (2007) Pathwise coordinate optimization. Ann Appl Stat 1:302–332
Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33:1–22. http://www.jstatsoft.org/v33/i01/
Gervini D, Yohai VJ (2002) A class of robust and fully efficient regression estimators. Ann Stat 30:583–616
Green PJ (1984) Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. J Royal Stat Soc Ser B 46:149–192
Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (1986) Robust statistics. The approach based on influence functions. Wiley, New York
Huber PJ (1981) Robust statistics. Wiley, New York
Khan JA, Van Aelst S, Zamar RH (2007) Robust linear model selection based on least angle regression. J Am Stat Assoc 102:1289–1299
Markatou M, Basu A, Lindsay B (1997) Weighted likelihood estimating equations: the discrete case with applications to logistic regression. J Stat Plan Inference 57:215–232
Maronna RA, Martin RD, Yohai VJ (2006) Robust statistics: theory and methods. Wiley, New York
Mizera I, Müller CH (1999) Breakdown points and variation exponents of robust M-estimators in linear models. Ann Statist 27:1164–1177
Müller CH (1995) Breakdown points for designed experiments. J Stat Plan Inference 45:413–427
Müller CH, Neykov NM (2003) Breakdown points of the trimmed likelihood and related estimators in generalized linear models. J Stat Plan Inference 116:503–519
Neykov NM, Neytchev P (1990) A robust alternative of the maximum likelihood estimators. COMPSTAT'90—Short Communications, Dubrovnik, Yugoslavia, pp 99–100
Neykov NM, Müller CH (2003) Breakdown point and computation of trimmed likelihood estimators in generalized linear models. In: Dutter R, Filzmoser P, Gather U, Rousseeuw PJ (eds) Developments in robust statistics. Physica-Verlag, Heidelberg, pp 277–286
Neykov NM, Filzmoser P, Neytchev PN (2012a) Robust joint modeling of mean and dispersion through trimming. Comput Stat Data Anal 56:34–48. doi:10.1016/j.csda.2011.07.007
Neykov NM, Cizek P, Filzmoser P, Neytchev PN (2012b) The least trimmed quantile regression. Comput Stat Data Anal 56:1757–1770. doi:10.1016/j.csda.2011.10.023
R Development Core Team (2012) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
Rousseeuw PJ (1984) Least median of squares regression. J Am Stat Assoc 79:851–857
Rousseeuw PJ, Van Driessen K (1999) Computing least trimmed of squares regression for large data sets. Estadistica 54:163–190
Schwartz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464
Tibshirani R (1996) Regression shrinkage and selection via the LASSO. J Royal Stat Soc Ser B 58:267–288
Tibshirani R (1997) The lasso method for variable selection in the Cox model. Stat Med 16:385–395

Vandev DL, Neykov NM (1993) Robust maximum likelihood in the Gaussian case. In: Ronchetti E, Stahel WA (eds) New directions in data analysis and robustness. Birkhäuser Verlag, Basel, pp 259–264
Vandev DL, Neykov NM (1998) About regression estimators with high breakdown point. Statistics 32:111–129
Wang H, Li G, Jiang G (2007) Robust regression shrinkage and consistent variable selection through the LAD-lasso. J Bus & Econ Stat 25:347–355
Zhang CH (2008) Discussion of one-step sparse estimates in nonconcave penalized likelihood models by H. Zou and R. Li. Ann Stat 36:1553–1560
Zou H (2006) The Adaptive lasso and its oracle properties. J Am Stat Assoc 101:1418–1429
Zou H, Li R (2008) One-step sparse estimates in nonconcave penalized likelihood models (with discussion). Ann Stat 36:1509–1533