Computational Statistics and Data Analysis 183 (2023) 107730

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Robust variable selection and estimation via adaptive elastic net S-estimators for linear regression ☆

David Kepplinger

Department of Statistics, George Mason University, 4400 University Drive, MSN 4A7, 22030, Fairfax, VA, USA

ARTICLE INFO

Article history:
Received 25 July 2022
Received in revised form 29 January 2023
Accepted 19 February 2023
Available online 1 March 2023

Keywords:
Robust procedures
Outlier resistance
Algorithms
Regularized regression
High-dimensional statistics

ABSTRACT

Heavy-tailed error distributions and predictors with anomalous values are ubiquitous in high-dimensional regression problems and can seriously jeopardize the validity of statistical analyses if not properly addressed. For more reliable variable selection and prediction under these adverse conditions, adaptive PENSE, a new robust regularized regression estimator, is proposed. Adaptive PENSE yields reliable variable selection and coefficient estimates even under aberrant contamination in the predictors or residuals. It is shown that the adaptive penalty leads to more robust and reliable variable selection than other penalties, particularly in the presence of gross outliers in the predictor space. It is further demonstrated that adaptive PENSE has strong variable selection properties and that it possesses the oracle property even under heavy-tailed errors and without the need to estimate the error scale. Numerical studies on simulated and real data sets highlight the superior finite-sample performance in a vast range of settings compared to other robust regularized estimators in the case of contaminated samples. An R package implementing a fast algorithm for computing the proposed method and additional simulation results are provided in the supplementary materials.

1. Introduction

This paper considers robust estimation, prediction and variable selection in the linear regression model

$$\mathcal{Y} = \mu^0 + \boldsymbol{\mathcal{X}}^\intercal \boldsymbol{\beta}^0 + \mathcal{U}, \tag{1}$$

with random predictors $\boldsymbol{\mathcal{X}}$ of fixed dimension p independent of the error term $\mathcal{U}$ and fixed parameters $\mu^0 \in \mathbb{R}$, $\boldsymbol{\beta}^0 \in \mathbb{R}^p$. Based on a sample of n independent realizations of $\mathcal{Y}$ and $\boldsymbol{\mathcal{X}}$, denoted as pairs $(y_i, \mathbf{x}_i)$, $i = 1, \dots, n$, the goal is to estimate the intercept $\mu^0 \in \mathbb{R}$ and slope $\boldsymbol{\beta}^0 \in \mathbb{R}^p$. Emphasis is on prediction accuracy and identification of relevant predictors, i.e., those with non-zero entries in $\boldsymbol{\beta}^0$, when some realizations may be contaminated with aberrant outliers.

With the growing abundance of data, often combined from various sources, it becomes increasingly challenging to make assumptions on the distribution of the error term $\mathcal{U}$ or assume that the data, particularly the predictors, are free of anomalous values or gross outliers. In proteomics studies, for instance, undetected equipment failure, problems with sample preparation, or patients with rare phenotypic profiles, are just a few sources of contamination that can severely impede the accuracy of commonly used statistical methods. If potential contamination or heavy tailed errors are not prop-

☆ Additional information and software are provided as online supplementary material.
E-mail address: dkepplin@gmu.edu.

https://doi.org/10.1016/j.csda.2023.107730

erly addressed, they can jeopardize the validity of statistical analyses and render the results unreliable. The main goal of this work is to develop a practical method for reliable identification of the relevant predictors and estimation of the corresponding non-zero regression coefficients for accurate prediction in the linear model (1), if the errors are heavy-tailed and observations are potentially contaminated.

We formalize this estimation problem as the minimization of a regularized objective:

$$\underset{\mu \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \; \mathcal{O}(\mathbf{y}, \mu + \mathbf{X}\boldsymbol{\beta}) + \Phi(\boldsymbol{\beta}; \lambda), \tag{2}$$

where $\mathbf{y} \in \mathbb{R}^n$ is the vector of response values and $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the matrix of predictor values, $\mathcal{O}(\mathbf{y}, \mu + \mathbf{X}\boldsymbol{\beta})$ is a regression loss (e.g., the sum of squared residuals) and $\Phi(\boldsymbol{\beta}; \lambda)$ is a penalty function (e.g., the L_1 norm). The hyper-parameter λ governs the strength of the penalization. The LASSO (Tibshirani, 1996) and the elastic net (EN) estimator (Zou and Hastie, 2005) are classical examples of regularized regression estimators based on the least-squares (LS) loss. The regularized LS-loss is extensively studied and well understood under numerous settings and penalty functions. Less has been done to ensure simultaneous robust and efficient variable selection and coefficient estimation for heavy-tailed error distributions and under the potential presence of contamination in both the predictors and the response.

The major contributions of this paper are two-fold. First, we introduce the adaptive PENSE estimator, a highly robust method for estimation and variable selection in the linear regression model. The high robustness arises from the particular combination of the S-loss function and the adaptive elastic net penalty. The ensuing estimator exhibits accurate prediction and variable selection as well as favorable asymptotic guarantees without requiring prior knowledge about the scale or distribution of the residuals, or the number of relevant predictors. The S-loss function proposed in Rousseeuw and Yohai (1984) robustly measures the scale of the fitted residuals, thereby circumventing the need for an auxiliary estimate of the residual scale. Second, we demonstrate that adaptive PENSE leads to substantially more robust variable selection than estimators with non-adaptive penalties. We show in the following sections that the adaptive penalty loadings are a fundamental ingredient for robust variable selection in high-dimensions, which has not been established in the literature to our knowledge. Adaptive PENSE therefore possesses superior empirical performance in the presence of aberrant contamination and heavy-tailed errors compared to other regularized estimators for the linear regression model.

We also develop a scalable and reliable algorithm to compute adaptive PENSE estimates. Computation is challenging due to the highly non-convex objective function and many hyper-parameters. Compared to the previous work in Cohen Freue et al. (2019), the algorithms developed here are computationally more tractable while at the same time expand the exploration of the parameter space. This leads to potentially more reliable estimates and makes a wider range of problems amenable to adaptive PENSE. To ensure practitioners can leverage these advances the optimized algorithms are made readily available as an R package in the supplementary materials and on the official R package repository CRAN.

In the supplementary materials we also prove that the adaptive PENSE estimator possesses the oracle property and asymptotic Normality without any moment conditions on the error distribution. Therefore, our theoretical results apply equally to light- and heavy-tailed distributions, including the Cauchy distribution and other symmetric stable distributions. These results stand out from the existing literature as we achieve these guarantees without the need for tuning the estimator to an unknown error distribution and without sacrificing finite-sample robustness towards unusual values in the response or the predictors.

1.1. Related work

The unregularized S-estimator is well-known to be robust towards heavy-tailed errors and arbitrary contamination in the predictors (Rousseeuw and Yohai, 1984). While some regularized S-estimators have been proposed (Maronna, 2011; Gijbels and Vrinssen, 2015; Cohen Freue et al., 2019) their variable selection performance is affected by contamination and the theory is not yet well established. For example, Cohen Freue et al. (2019) show that their PENSE estimator, the S-loss combined with the EN penalty, leads to consistent parameter estimation but it does not possess the oracle property. Following classical robust theory Smucler and Yohai (2017) proposes the adaptive MM-LASSO, a regularized M-estimator with bounded loss function, based on the robust M-scale of the residuals estimated by the S-Ridge (Maronna, 2011). The theory promises the same robustness as the S-Ridge but higher efficiency under the central model. Estimating the residual scale with small finite-sample bias, however, is a difficult task in high-dimensions, even in classical non-robust settings without contamination (Fan et al., 2012; Reid et al., 2016; Tibshirani and Rosset, 2019). Therefore, the performance of adaptive MM-LASSO is in practice often impeded by the finite-sample bias of the M-scale estimate.

Many other M-estimators for the high-dimensional linear regression model have been proposed in the past to shield against heavy-tailed errors, but without protection from contamination in the predictors (e.g., Wang et al., 2007; Zou and Yuan, 2008; Fan et al., 2014; Lambert-Lacroix and Zwald, 2011; Fan et al., 2018; Loh, 2021; Sun et al., 2019; Pan et al., 2021). These estimators replace the LS-loss with convex loss functions which grow slower than the LS-loss for large residuals, so-called unbounded M-loss functions. The convexity of unbounded M-loss functions allows derivations of strong theoretical guarantees and high efficiency under the central model, but these estimators are vulnerable to the potentially devastating effects of contamination in the predictors. Commonly suggested remedies, e.g., down-weighting observations with unusual values in the predictors or univariate winsorizing (Loh, 2017; Sun et al., 2019), are not well-suited for high-dimensional and sparse problems. Applying these ad-hoc remedies and down-weighting observations because of outlying values in irrelevant predictors, for example, sacrifices precious information. To add to the difficulty in practice, these M-estimators with unbounded loss function also require a robust estimate of the residual scale.

Another line of research deals with the mean-shift outlier model (MSOM), in which an additive fixed effect is introduced for each observation to quantify its outlyingness. In She and Owen (2011) the authors propose non-convex thresholding rules, for instance via an L_0 constraint on the mean-shift effects, as a robust and fast way to identify outliers. She et al. (2022) continue in this line of work and develop a theoretical foundation for a broad class of robust estimators defined algorithmically, including the SparseLTS estimator (Alfons et al., 2013), establish minimax bounds for the estimation error under the MSOM, and propose an efficient algorithm for estimation. Similarly, Insolia et al. (2022) propose a mixed integer program (MIP) to constrain both the number of outliers and the number of relevant predictors using the L_0 constraint. The authors develop guarantees for the algorithmic complexity and the estimation error for their MIP. For example, with normally distributed errors and when the true number of relevant predictors and the true number of outliers in the response are known, the MIP possess the oracle property under the MSOM. The MSOM framework and the proposed methods building upon the L_0 constraint provide a promising avenue, particularly in regimes with high signal strength. For lower signal strengths, however, the L_0 constraint for either variable selection or outlier detection tends to suffer from high variability (Hastie et al., 2020; Insolia et al., 2022). For computational tractability, it is also necessary to restrict the optimization to a tight neighborhood around the true parameters and to have good upper bounds for the number of relevant predictors and the number of outliers. Insolia et al. (2022) suggest to use an ensemble of other robust methods to obtain these bounds. The method we propose does not require prior knowledge of the number of relevant predictors and only a rough upper bound on the number of outliers and is thus a good candidate to initialize these methods for the MSOM.

1.2. Notation

The concatenated parameter vector of intercept and slope in the linear regression model (1) is denoted by $\boldsymbol{\theta} = (\mu, \boldsymbol{\beta}^\intercal)^\intercal$. The non-zero elements of a slope parameter $\boldsymbol{\beta}$ are referenced as $\boldsymbol{\beta}_{\mathrm{I}}$, while the zero elements are written as $\boldsymbol{\beta}_{\mathrm{II}}$. Adaptive PENSE estimates are always denoted by a hat, $\hat{\boldsymbol{\theta}}$, and PENSE estimates are marked by a tilde, $\tilde{\boldsymbol{\theta}}$. The subscript $i \in \{1, \dots, n\}$ is exclusively used to denote the i-th observation from the sample, while $j \in \{1, \dots, p\}$ is reserved for indexing predictors. Without loss of generality, it is assumed that the true slope parameter equals the concatenated vector $\boldsymbol{\beta}^0 = (\boldsymbol{\beta}_{\mathrm{I}}^{0\intercal}, \boldsymbol{\beta}_{\mathrm{II}}^{0\intercal})^\intercal$ where the first s elements, $\boldsymbol{\beta}_{\mathrm{I}}^0$, are non-zero and the trailing $p - s$ elements are zero, i.e., $\boldsymbol{\beta}_{\mathrm{II}}^0 = \mathbf{0}_{\text{p-s}}$.

2. Adaptive PENSE

We propose estimating the sparse regression parameter $\boldsymbol{\theta}^0$ in the linear regression model (1) by penalizing the robust S-loss with an adaptive elastic net penalty. In the presence of outliers in the response and unusual values in the predictors, the S-loss is a better alternative to the LS-loss, and is given by

$$\mathcal{O}_{\mathrm{S}}(\mathbf{y}, \hat{\mathbf{y}}) = \hat{\sigma}_{\mathrm{M}}^2(\mathbf{y} - \hat{\mathbf{y}}) = \inf \left\{ s^2 : \frac{1}{n} \sum_{i=1}^{n} \rho\left(\frac{y_i - \hat{y}_i}{|s|}\right) \leq \delta \right\}. \tag{3}$$

Here ρ is a bounded and hence non-convex function and $\delta \in (0, 0.5]$ is a fixed parameter governing robustness as will be shown later. To ease notation, we define the M-scale of the residuals of an estimate $\hat{\boldsymbol{\theta}}$ by $\hat{\sigma}_{\mathrm{M}}(\hat{\boldsymbol{\theta}}) = \hat{\sigma}_{\mathrm{M}}(\mathbf{y} - \hat{\mu} - \mathbf{X}\hat{\boldsymbol{\beta}})$.

Instead of the sample variance of the residuals, the S-loss minimizes the robust M-scale of the fitted residuals, $\hat{\sigma}_{\mathrm{M}}(\mathbf{y} - \hat{\mathbf{y}})$. The robustness of the M-scale depends on two components: (i) the choice of the ρ function and (ii) the fixed quantity δ. The ρ function measures the "size" of the standardized residuals. To get a robust estimate of the scale, ρ must be bounded (Yohai, 1987), i.e., assign the same size to all standardized residuals larger than a threshold. Since the objective is to get a robust estimate, we assume from here on that the ρ function satisfies the condition

[A1] $\rho \colon \mathbb{R} \to [0, 1]$ is even and twice continuously differentiable, with $\rho(0) = 0$, is bounded, $\rho(t) = 1$ for all $|t| \geq c > 0$, and non-decreasing in $|t|$.

The boundedness implies that the derivative of the ρ function is 0 for $|t| \geq c$ and hence residuals greater than c have no effect on the minimization of (3). As evident from the asymptotic theory presented in Section S1 in the supplementary material, the choice of the ρ function directly affects the variance of adaptive PENSE. In practice we use Tukey's bisquare ρ function

$$\rho(t; c) = \min\left(1, 1 - \left(1 - (t/c)^2\right)^3\right), \tag{4}$$

as it is simple and the loss in efficiency compared to an optimized ρ function for Normal errors is negligible for the unregularized S-estimator (Hössjer, 1992). The cutoff c in [A1] and (4) does not affect the S-estimator of the regression parameters or the variance of the M-scale estimator; it is merely a multiplicative factor for the scale estimate. We are therefore fixing $c = 1$ for the reminder of this paper when referring to the S-loss. If an M-scale estimate of the residuals is needed, we use a cutoff which leads to a consistent estimate under Normal errors; this cutoff will depend on δ.

The second component that determines the robustness of an S-estimator is the constant δ which must be in $(0, 0.5]$ for ρ functions of the form [A1]. The M-scale estimate can tolerate up to $\lfloor n \min(\delta, 1-\delta) \rfloor$ gross outliers without exploding to infinity or imploding to 0 (Maronna et al., 2019). Similarly, we show in Theorem 1 in the supplementary material that adaptive PENSE can also tolerate up to $\lfloor n \min(\delta, 1-\delta) \rfloor$ adversely contaminated observations without giving aberrant results. For robustness considerations, an optimal choice is therefore $\delta = 0.5$, which allows the estimator to tolerate gross outliers in the residuals of almost 50% of observations. On the other hand, the variance of the estimator increases with δ and adaptive PENSE with $\delta = 0.5$ achieves only $\sim 30\%$ efficiency under the Normal model while for $\delta = 0.25$ the efficiency is close to 80%. Therefore, a good sense of the expected proportion of contamination helps to get as much efficiency as possible.

The unregularized S-estimator cannot be computed if $p > n(1-\delta) - 1$ and it cannot recover the set of relevant predictors. In Cohen Freue et al. (2019), the S-loss is combined with the elastic net penalty, a convex combination of the L_1 and the squared L_2 norm, given by $\Phi_{\text{EN}}(\boldsymbol{\beta}; \tilde{\lambda}, \tilde{\alpha}) = \tilde{\lambda} \sum_{j=1}^{p} \frac{1-\tilde{\alpha}}{2} \beta_j^2 + \tilde{\alpha} \left|\beta_j\right|$. The hyper-parameter $\tilde{\alpha} \in [0, 1]$ controls the balance between the L_1 and the L_2 penalty and $\tilde{\lambda}$ controls the overall strength of the penalization. With $\tilde{\alpha} = 0$, the EN penalty coincides with the Ridge penalty and for $\tilde{\alpha} = 1$ it recovers the LASSO. If $\tilde{\alpha} < 1$, the elastic net results in a more stable variable selection than the LASSO penalty when predictors are correlated (Zou and Hastie, 2005).

The elastic net penalty, just as the LASSO penalty, introduces non-negligible bias and thus cannot lead to a variable selection consistent estimator. We are therefore proposing to combine the robust S-loss with the adaptive EN penalty:

$$\Phi_{\text{AE}}(\boldsymbol{\beta}; \lambda, \alpha, \zeta, \tilde{\boldsymbol{\beta}}) = \lambda \sum_{j=1}^{p} \left|\tilde{\beta}_j\right|^{-\zeta} \left(\frac{1-\alpha}{2} \beta_j^2 + \alpha \left|\beta_j\right| \right), \qquad \zeta \geq 1. \tag{5}$$

The adaptive EN penalty combines the advantages of the adaptive LASSO penalty (Zou, 2006) and the EN penalty (Zou and Zhang, 2009). Contrary to the original definition in Zou and Zhang (2009), (5) applies the penalty loadings $|\tilde{\beta}_j|^{-\zeta}$ to both the L_1 and L_2 penalties. The adaptive EN penalty leverages information from a preliminary regression estimate, $\tilde{\boldsymbol{\beta}}$, to penalize predictors with initially small coefficient values more heavily than predictors with initially large coefficients. This has two advantages over the non-adaptive EN penalty: (i) the bias for large coefficients is reduced and (ii) variable selection is improved by reducing the false-positive rate.

The adaptive EN penalty improves the stability of the estimator in the presence of multicollinearity as compared to the adaptive LASSO (Zou and Zhang, 2009). In applications where there is no multicollinearity, adaptive EN gives the option of setting $\alpha = 1$, i.e., computing an adaptive LASSO estimate. In the presence of multicollinearity, however, the contribution of the L_2 penalty is important for improving the stability of the variable selection and the selection of an appropriate penalization level. The adaptive EN penalty therefore gives the necessary flexibility to cover a wide range of scenarios, particularly scenarios where the researcher has limited control over the predictor values.

Adaptive PENSE is a two-step procedure leveraging a PENSE estimate with $\tilde{\alpha} = 0$. In the first step, a PENSE-Ridge estimate is computed as

$$\tilde{\boldsymbol{\theta}} = \underset{\mu, \boldsymbol{\beta}}{\arg\min}\, \mathcal{O}_{\text{S}}(\mathbf{y}, \mu + \mathbf{X}\boldsymbol{\beta}) + \Phi_{\text{EN}}(\boldsymbol{\beta}; \tilde{\lambda}, 0), \tag{6}$$

and in the second step, using the PENSE-Ridge $\tilde{\boldsymbol{\theta}}$ as the preliminary estimate, adaptive PENSE is computed by

$$\hat{\boldsymbol{\theta}} = \underset{\mu, \boldsymbol{\beta}}{\arg\min}\, \mathcal{O}_{\text{S}}(\mathbf{y}, \mu + \mathbf{X}\boldsymbol{\beta}) + \Phi_{\text{AE}}(\boldsymbol{\beta}; \lambda, \alpha, \zeta, \tilde{\boldsymbol{\beta}}). \tag{7}$$

Fixing the preliminary estimate to a PENSE-Ridge has two important advantages: (i) computation is fast because the Ridge penalty is smooth (hence amenable to more efficient algorithms) and because we do not need to choose from several $\tilde{\alpha}$ values, and (ii) no predictors are discarded prematurely. In high-dimensional problems discarding some predictors in the preliminary stage may be beneficial for computation speed and variable selection performance. Empirical studies, however, suggest that the gains are usually small and only noticeable in very high-dimensional settings. In most other cases adaptive PENSE performs better and is faster if the preliminary stage does not screen out variables.

Even with the first-stage penalty fixed at $\tilde{\alpha} = 0$, computing adaptive PENSE estimates involves choosing four hyper-parameters: (i) $\tilde{\lambda}$, the level of penalization for PENSE-Ridge, (ii) α, the balance of L_1/L_2 regularization for adaptive PENSE, (iii) λ, the level of penalization for adaptive PENSE, and (iv) ζ, the exponent governing the predictor-specific regularization. In general, the larger ζ the stricter the differentiation between small and large coefficient values. If ζ is large, all but a few predictors with largest coefficient values will be heavily penalized and thus likely not selected.

In addition to the large number of hyper-parameters that need to be selected, both optimization problems (6) and (7) are highly non-convex in μ and $\boldsymbol{\beta}$. Finding global minima through numeric optimization is therefore contingent on a starting value that is close to a global minimum. We describe an efficient strategy for deriving potentially good starting values for adaptive PENSE in Section 4.

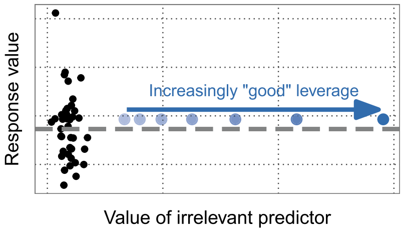

Fig. 1. Example of a good leverage point in a truly irrelevant predictor. The leverage is higher as the value in the irrelevant predictor increases and the residual in the intercept-only model (depicted as dashed line) is neither 0 nor too large.

2.1. Statistical properties

In Theorem 1 in the supplementary materials we show that adaptive PENSE is highly robust, with a finite-sample replacement breakdown point (FBP) of up to 50%. The main message from this theorem is that the adaptive PENSE estimator is bounded away from the boundary of the parameter space if contamination is restrained to fewer than $\lfloor n \min(\delta, 1-\delta) \rfloor$ observations. It is important to stress that the data-driven strategy for choosing the hyper-parameters must itself provision for contamination to not break the robustness of adaptive PENSE. The strategy proposed in Section 4.2 tries to shield against the detrimental effect of contamination by using the robust τ-scale for measuring the prediction accuracy and by repeating cross-validation several times. Without specific assumptions on the type of contamination, the actual performance of the estimator can only be assessed with numerical experiments and we show the favorable empirical performance of adaptive PENSE in Section 5.1.

We now briefly turn to the asymptotic properties of adaptive PENSE as $n \to \infty$ and the dimensionality p remains fixed. Under mild assumptions given in Section S1 in the supplementary materials, we establish the oracle property of the adaptive PENSE estimator and its limiting Normal distribution. This implies that adaptive PENSE has the same asymptotic properties as if the true model would be known in advance. Therefore, adaptive PENSE has the same asymptotic efficiency as an unpenalized S-estimator applied to only the truly relevant predictors. These properties hold in particular if using the PENSE-Ridge as preliminary estimate.

It is noteworthy that the assumption on the residuals does not impose any moment conditions, making our results applicable to heavy tailed errors. Unlike many results concerning regularized M-estimators, we only require a finite second moment of the predictors. Similarly, while many existing regularized estimators possess the oracle property with higher asymptotic efficiency, none achieve these results for heavy tailed errors without prior knowledge or consistent estimate of the residual scale, and simultaneously finite-sample robustness towards contamination in the predictors. Adaptive PENSE fills this gap which is particularly critical for reliable estimation in finite samples.

The theoretical results depend on an appropriate choice for the hyper-parameters $\tilde{\lambda}$ and λ, for PENSE and adaptive PENSE, respectively. For a data-driven procedure, the required conditions are difficult if not impossible to verify. To substantiate the theoretical properties and verify that they translate beneficially in practical problems, we carefully investigate adaptive PENSE's properties in finite samples with data-driven hyper-parameter selection in Section 5.

3. Reliable variable selection

A key benefit of adaptive PENSE over estimators with other convex or folded-concave penalties is that it possesses more robust variable selection properties. Good leverage points, i.e., observations where the response follows the regression model but one or more predictors have unusually large values carry important information for estimating the model parameters and should not be considered outliers. If the leverage, however, is due to large values in truly irrelevant predictors, robust estimators have difficulty discerning irrelevance from infinitesimal effects. Such good leverage points caused by irrelevant predictors, as illustrated in Fig. 1, can lead to a large number of false positives in robust estimates using non-adaptive penalties.

The effected predictors are often the first to enter the model. This is due to a combination of how large values in predictors affect the sub-gradient of the objective function and robust scaling of predictors. This phenomenon is shared among all robust estimators which are based on down-weighting outlying observations and can be best seen in the sub-gradient of the objective function at $\boldsymbol{\beta} = \mathbf{0}_p$. For simplicity, we focus here only on the PENSE estimator with sub-gradient at $\boldsymbol{\beta} = \mathbf{0}_p$ given by

$$\boldsymbol{\nabla}_{\boldsymbol{\beta}} \left\{ \mathcal{O}_{\mathrm{S}}(\mathbf{y}, \mu + \mathbf{X}\boldsymbol{\beta}) + \Phi_{\mathrm{EN}}(\boldsymbol{\beta}; \tilde{\lambda}, \tilde{\alpha}) \right\} \Big|_{\boldsymbol{\beta}=\mathbf{0}_p} = -\frac{1}{n} \sum_{i=1}^{n} w_i^2 \,(y_i - \mu)\, \mathbf{x}_i + \tilde{\lambda}[-\tilde{\alpha}; \tilde{\alpha}],$$

where w_i are derived from the S-loss evaluated at the intercept-only model. The weight w_i is >0 if and only if the residual from the intercept-only model is not too large (relative to all other residuals) and different from 0 (not fit exactly). Consider now that predictor j is truly irrelevant and contains a very large value in observation i, but the residual for observation i

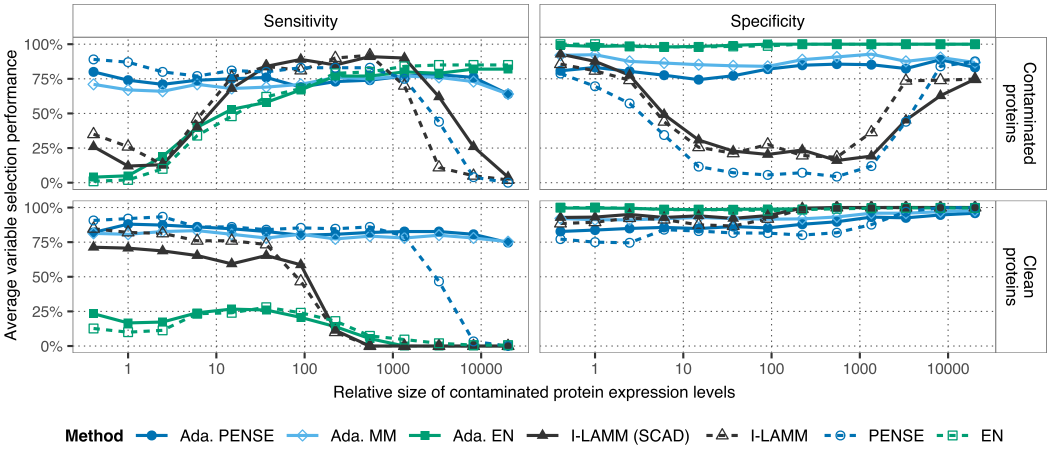

Fig. 2. Effect of high-leverage points on the sensitivity and specificity of various regularized regression estimators. Sensitivity (left column) is the number of selected proteins relative to the number of truly relevant proteins; specificity (right column) is the number of not-selected proteins relative to the number of truly irrelevant proteins. Average performance over 50 replications is reported for proteins with contaminated values (top row) and proteins free from any contamination (bottom row). Computational details for the displayed estimators are provided in Section 5.1.

in the intercept-only model is small such that $w_i \gg 0$, like the observation shown in Fig. 1. Robust scaling of the predictor likely does not shrink this large value substantially and hence the j-th predictor dominates the sub-gradient at $\boldsymbol{\beta} = \mathbf{0}_p$; therefore, it enters the model first. This single aberrant value leads to the false impression that the j-th predictor is relevant. However, because the leverage is caused by an extreme value in an irrelevant predictor, the robustly estimated coefficient for this predictor is likely very small, compared to coefficients of truly relevant predictors. Importantly, the higher the leverage of this observation, the smaller the estimated coefficient. This allows adaptive PENSE to screen out this predictor, making it more robust against this form of contamination.

This form of good leverage points may occur in many practical problems, particularly in sparse high-dimensional settings. In protein expression data, for example, a group of proteins could be highly expressed in a small fraction of subjects while only trace amounts of the protein are detected in the vast majority of subjects. Even if the group of proteins is not relevant for the outcome of interest, non-robust methods or robust methods with non-adaptive penalties are prone to selecting these proteins. An example of this behavior using synthetic data is shown in Fig. 2 with $n = 100$ and $p = 32$. Here, five out of 28 irrelevant proteins have higher expression levels in 20 observations. In another five observations, the response variable is outlying and two out of the five relevant proteins exert high-leverage. Details on how the shown estimators are computed are given in Section 5.1, and additional information on the simulation setting is provided in Section S5.1 in the supplementary materials.

Clearly, adaptive PENSE is the only estimator for which the variable selection is not breaking down. Overall, the effect of such leverage points is more pronounced the higher the leverage of the contaminated observations, but the estimators are affected differently. Non-robust EN estimators tend to select only the two contaminated truly relevant predictors, but the actual parameter estimates are highly biased due to the contamination. I-LAMM with either the LASSO or SCAD penalty doesn't select any proteins if the leverage caused by the affected proteins is too high. The PENSE estimator, on the other hand, selects almost all of the affected irrelevant proteins but tends towards not selecting any proteins at all as leverage increases. Only adaptive PENSE is mostly unaffected by the contamination, identifying on average four out of five truly relevant predictors, while screening out 24 of the 28 irrelevant predictors.

A notion of finite sample robustness of variable selection, comparable to the finite sample replacement breakdown point for the estimation bias, is unfortunately still missing from the literature. We thus resort to empirical studies to analyze how various forms of contamination affect finite sample variable selection performance. The results presented in Section 5 underscore the superior robustness and reliability of adaptive PENSE in terms of variable selection.

4. Computing adaptive PENSE estimates

The non-convex objective function paired with the need to select hyper-parameters requires optimized algorithms for computing adaptive PENSE estimates. The problem is split into two stages: (1) computing a preliminary PENSE-Ridge estimate and (2) computing an adaptive PENSE estimate based on this (fixed) preliminary estimate. These two stages are done sequentially and hyper-parameters are selected independently in each. The only information passed down is the preliminary estimate.

The biggest challenge for computing adaptive PENSE estimates is the potentially severe non-convexity of the objective function. Undesirable local minima of the objective function, however, are artifacts of contamination and should thus be avoided. This insight is used in Cohen Freue et al. (2019) to find approximate initial solutions which are likely closer to good local minima than to local minima caused by contamination. They introduce the elastic net Peña-Yohai (EN-PY) estimator to obtain initial estimates for their non-convex optimization routine at a few penalization levels. They use these initial estimates as starting points to compute PENSE estimates at these penalization levels. For all other penalization levels, a warm-start strategy is employed, where the solution obtained for the next largest penalization level acts as starting point. The issue with this approach is that a given penalization level does not lead to the same penalization in the initial EN-PY estimator and PENSE, and the starting point can thus be too far away from the local optima of interest. We improve their EN-PY procedure to be computationally less demanding and adapt it for the adaptive EN penalty.

4.1. Algorithm for adaptive PENSE

We develop an algorithm which is optimized for computing adaptive PENSE estimates over a fine grid of penalization levels, with hyper-parameters α and ζ fixed. Using our refined adaptive EN-PY procedure, we compute initial solutions for adaptive PENSE at a coarse grid of penalization levels (e.g., at every fifth penalization level). We then employ the following novel warm-start strategy to maximize exploration of the search space while maintaining computational feasibility.

Even for a fixed penalty level, the effect of penalization on the initial solutions obtained by adaptive EN-PY is vastly different than on adaptive PENSE; therefore, we combine the initial solutions from all penalization levels into one large set of initial estimates. Beginning at the largest value in the range of penalization levels, we use every solution in the set of initial estimates to start an iterative algorithm following the minimization by majorization (M-M) paradigm (Lange, 2016). Our iterative M-M algorithm locates a local optimum by solving a sequence of weighted adaptive LS-EN problems, each with updated observation weights derived from the robust S-loss evaluated at the current iterate. Instead of fully iterating until convergence for every initial solution, the M-M algorithm is stopped prematurely, leading to a set of candidates solutions. Of those, only the most promising candidates are fully iterated. The fully iterated solution with smallest value of the objective function is finally taken as the adaptive PENSE estimate at the largest penalization level. This two-step approach – exploration and improvement – is successfully applied for many other types of robust estimators (e.g., Salibián-Barrera and Yohai, 2006; Rousseeuw and Van Driessen, 2006; Alfons et al., 2013) and works very well for adaptive PENSE, too. Section S3 in the supplementary materials give additional details about our M-M algorithm.

At the next smallest penalization level, the M-M algorithm is started from all initial estimates and all fully iterated, most promising candidates from the previous penalization level. Again, only a few iterations of the M-M algorithm are performed for these starting points and only the most promising solutions are iterated until convergence. This cycle is repeated for every value in the range of penalization levels, from largest to smallest. Carrying forward the most promising solution from previous penalization levels combined with initial estimates from adaptive EN-PY leads to efficient and extensive exploration of the parameter space. These computational solutions are implemented in the R package `pense` in the supplementary materials.

4.2. Selecting hyper-parameters

For both PENSE-Ridge and adaptive PENSE we select hyper-parameters via repeated K-fold cross-validation (CV) over a grid of possible values. In the first stage, only the penalization level is selected as $\tilde{\alpha}$ is fixed at 0. In the second stage, the two additional hyper-parameters α and ζ are selected from a small set of pairs. For every pair, the penalization level is chosen via separate CVs (using the same splits), and the combination resulting in the best prediction performance is selected. The range of penalization levels is different in both stages, as well as for every pair (α, ζ) considered in the second stage.

Estimating the prediction error of robust estimators via cross-validation suffers from high variability due to the non-convexity of the objective function, even more so if contamination is present. The prediction error estimated from a single CV is highly dependent on the random split and hence unreliable for selecting the hyper-parameter. We reduce this volatility by repeating CV several times to get a more reliable assessment of the prediction performance for given values of the hyper-parameters. In addition to repeating CV, the measure of the prediction error needs to be stable in the presence of gross errors in the observed response values. For PENSE and adaptive PENSE we use the robust τ-scale of the uncentered prediction errors (Maronna and Zamar, 2002) to estimate the prediction accuracy in each individual CV run:

$$\hat{\tau} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \min\left(c_\tau, \frac{|y_i - \hat{y}_i|}{\underset{i'=1,\dots,n}{\operatorname{Median}} |y_{i'} - \hat{y}_{i'}|}\right)^2}.$$

Here, $\hat{y}_i$ is the predicted response value from the CV split where the i-th observation is in the test set. The cutoff $c_\tau > 0$ governs the tradeoff between efficiency and robustness of the estimate. Following previous works (Cohen Freue et al., 2019; Smucler and Yohai, 2017), all analyses in this manuscript use $c_\tau = 3$. Since the presence of gross errors in the predictions is handled by the robust τ-scale, we average the prediction errors from the replicated CVs using the sample mean. This gives an overall measure of prediction performance for each set of hyper-parameters. Repeating CV additionally gives insights into the variability of the prediction performance and affords more sensible selection of the penalization level, e.g., using the "1-standard-error-rule" (Hastie et al., 2009).

Another important remedy to reduce the variability incurred by CV is robust scaling of the data to make penalization levels more comparable across CV splits. Specifically, the predictors in the full data set are scaled to have unit M-scale such that the penalization level λ leads to comparable penalization of all predictors. Furthermore, in each CV split we standardize each predictor to also have unit M-scale. Therefore, a fixed penalization level λ induces a level of sparsity to the parameter estimate computed on the training data comparable to the sparsity when computed on the standardized input data.

5. Numerical studies

We conduct numerical studies both on simulated and real data. The simulation study covers scenarios where $p < n$ and $p > n$, multiple sparsity settings and various heavy-tailed error distributions. The application showcases the potential of adaptive PENSE in a challenging setting where other estimators are highly affected by contamination.

5.1. Simulation study

In the simulation study we compare the prediction and model selection performance of adaptive PENSE and the proposed hyper-parameter selection procedure with other commonly used robust and non-robust estimators. We assess the reliability of the estimators by considering many different contamination structures and heavy-tailed error distributions. Below we present the results from a scenario with $n = 200$ observations and $p = 32,\ 128,\ 512$ predictors, of which $s = \lfloor 0.9\sqrt{p} \rfloor$ are relevant. Additional simulation results are presented in the supplementary materials. Section S5.3 shows results for $n = 100$ and more severe contamination. Section S5.4 shows empirical results supporting the theoretical property that the estimation error of adaptive PENSE is decreasing as n increases.

Here we consider multiple error distributions and either no adverse contamination or contamination in 10% of observations. The outliers are placed such that they are severe enough to bias the estimators, but not as far away from the true signal as to be easily detectable by robust estimators. Since the outliers tend to be difficult to detect they are most damaging to adaptive PENSE. On the other hand, it would be easy to find contamination settings that are more detrimental to the non-robust estimators, as their coefficient estimates can get infinitely biased. This unfortunately also holds for MM-type estimators because a severely biased preliminary scale estimate can also lead to the break down of MM-type estimators. However even under these adverse conditions adaptive PENSE performs better or similar to other robust methods, highlighting the reliability of adaptive PENSE.

For each combination of error distribution and presence/absence of contamination, we randomly generate 50 data sets according to the recipe outlined below. The predictors are drawn from a multivariate t-distribution with four degrees of freedom and a grouped correlation structure. Predictors within each of the $\lfloor 1 + \sqrt{p}/2 \rfloor$ groups are highly correlated (correlation of 0.96), and mildly correlated with predictors in other groups. Elastic net penalties provide the required stability of variable selection for this grouping structure among predictors. The true regression parameter is $\boldsymbol{\beta}^0 = (1, \dots, 1, 0, \dots 0)$ with s 1's and $p - s$ 0's. The residuals follow a symmetric stable distribution with stability parameter values $\nu = 1.33$ ("Stable(1.33) residuals") and $\nu = 1$ ("Cauchy residuals"). Section S5.2 in the supplementary materials also shows results for light-tailed Normal errors ($\nu = 2$). The residuals are scaled such that the true model explains 25% of the variation in the response, corresponding to a signal-to-noise ratio of 1/3.

In settings with 10% contamination, bad leverage points are introduced to $\log_2(p)$ randomly chosen irrelevant predictors by changing the values in these predictors to create a moderate amount of leverage. Preliminary studies suggest adaptive PENSE is most affected by the chosen leverage, while results for other estimators are substantially worse when introducing higher leverage. This is expected as edge-cases are the most difficult to handle for robust methods, while extremely aberrant values are easier to detect and hence their influence can be reduced. The response variable is contaminated with gross outliers by using a different regression parameter with 1's for the affected predictors and 0's otherwise. The residuals for the outlying observations are scaled such that the contaminated model explains about 91% of the variation in the contaminated response. In addition to bad leverage points and outliers, all settings have 20% of observations as good leverage points by changing the values in $(p - s)/2$ irrelevant predictors. We introduce a moderately amount of good leverage, leading to the largest effect on adaptive PENSE. More details about the data generation are provided in Section S5.2 in the supplementary materials.

The hyper-parameters for all estimators are selected according to the schema outlined in Section 4.2 using 20 replications of five-fold CV. For I-LAMM with the LASSO and SCAD penalties and (adaptive) EN we use the mean absolute error (MAE) to measure prediction accuracy. For the other estimators, we use the uncentered τ-scale estimate with $c_\tau = 3$. The hyper-parameter α in EN-type penalties is chosen from $\{0.5, 0.75, 1\}$ and for adaptive EN-type penalties we consider ζ values in $\{1, 2\}$. For each estimator we consider 50 values of the penalization level λ, covering the full range of models, from the intercept-only model to the almost saturated model. The I-LAMM estimates are computed with a modified version of the package by Fan et al. (2018), provided in the supplementary materials. The only modification to the package developed by the paper's authors is the change to the MAE for hyper-parameter selection. Except for the number of penalization levels, the penalty type, and the number of CV folds, all arguments for I-LAMM use the package's default values. Non-robust EN-type estimators are computed with the R package `glmnet` (Friedman et al., 2010), again using default argument values except for those controlling the penalization strength and number of CV folds. Adaptive PENSE and MM estimates are computed with the R package `pense` also provided in the supplementary materials. The adaptive MM estimator considered in this study is similar to the adaptive MM-LASSO (Smucler and Yohai, 2017), but replacing the LASSO by the EN penalty and using the scale of the residuals from PENSE-Ridge. The hyper-parameters are selected using the "1-SE-rule", i.e., the

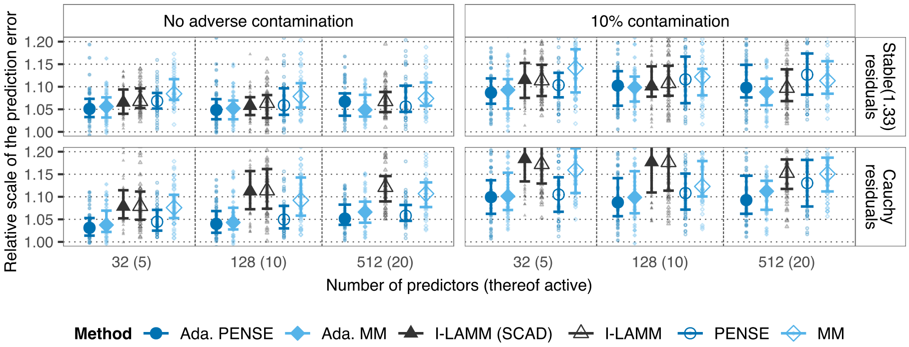

Fig. 3. Prediction accuracy of robust estimators, measured by the uncentered τ-scale of the prediction errors relative to the true scale of the error distribution (lower is better). The median out of 50 replications is shown by the large points, the error bars depict the interquartile range, and small points show the Individual results. Adaptive PENSE and adaptive MM outperform other methods, especially for heavy-tailed error distributions.

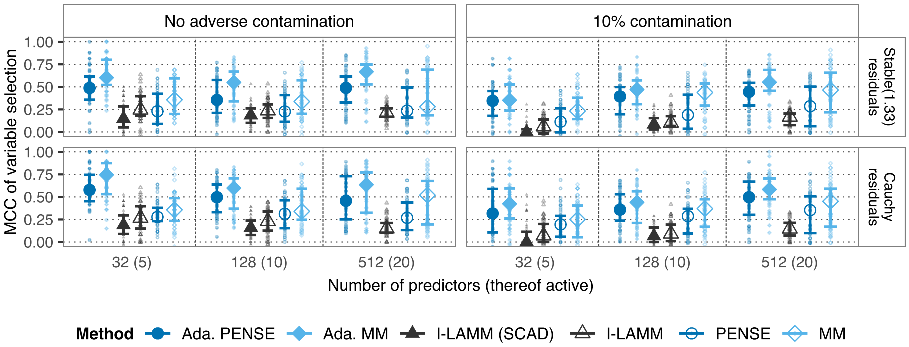

Fig. 4. Overall variable selection performance of robust estimators, measured by the MCC defined in (8) (higher is better). The median out of 50 replications is depicted by the large points, the error bars show the interquartile range, and the small points represent individual results. Adaptive PENSE and adaptive MM are substantially better than the other considered estimators in all settings.

hyper-parameters leading to the sparsest model while staying within one standard error of the best prediction accuracy. The full code to reproduce the simulations is provided in the supplementary materials.

The prediction performance is shown in Fig. 3, where the scale of the prediction errors is estimated by the uncentered τ-scale with $c_\tau = 3$. Adaptive PENSE and adaptive MM are noticeably less affected by contamination or heavy tailed errors than other estimators. Moreover, estimators with adaptive EN penalties are outperforming non-adaptive penalties, likely due to the presence of good leverage points. With moderately heavy-tailed errors, adaptive PENSE has slightly lower prediction accuracy than adaptive MM; likely because the residual scale can often be estimated accurately in these settings. Especially in high-dimensional settings the efficiency gains of adaptive MM are visible. In the presence of contamination and more heavy-tailed errors, however, adaptive PENSE is generally more reliable than adaptive MM. I-LAMM with either penalty performs only slightly worse than adaptive PENSE for Stable(1.33) errors in this simulation scenario, even in the presence of contamination. However, the I-LAMM estimators have similar issues with good leverage points as PENSE and they are more affected by bad leverage points and gross errors in the residuals under Cauchy errors. LS-EN estimators are excluded from the display in Fig. 3 as their prediction performance is badly affected by the heavy-tailed errors. Additional plots and results, including the non-robust EN estimators, are provided in Section S5.2 in the supplementary materials.

Fig. 4 shows the overall variable selection performance of the estimators. The Matthews correlation coefficient (MCC) is calculated from the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN):

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}. \tag{8}$$

Adaptive penalties lead to better variable selection than non-adaptive penalties; even without adverse contamination, good leverage points distort variable selection by non-adaptive penalties. Inspection of the sensitivity and specificity of variable selection (shown in Figure S.5 in the supplementary materials) highlights that non-adaptive penalties tend to select a large proportion of the irrelevant predictors with large values. While adaptive penalties also screen out truly relevant predictors at a higher rate than non-adaptive penalties, adaptive PENSE and adaptive MM substantially improve variable selection overall. Importantly, the robust estimators with adaptive penalty are selecting none or only few of the irrelevant predictors with extreme values.

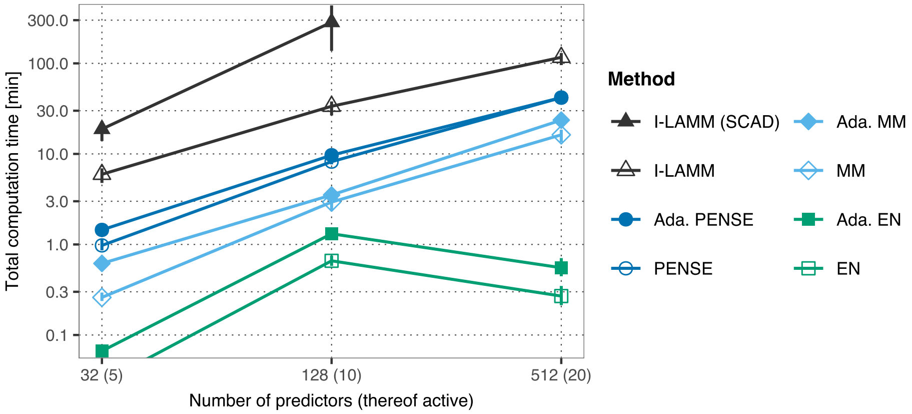

Fig. 5. Total computation time for each method using 4 CPU cores (Intel(R) Xeon(R) Gold 6240R). The median out of 50 replications is depicted by the points and the error bars show the interquartile range. I-LAMM (SCAD) estimates for $p = 512$ take prohibitively long to compute and are thus omitted. The vertical and horizontal axes are shown on the log scale.

Finally, Fig. 5 shows the total computation time for all considered methods. All methods have access to four Intel(R) Xeon(R) Gold 6240R CPUs. Methods implemented in the `pense` package (adaptive PENSE, adaptive MM, PENSE, MM) use multi-threading, while the other methods parallelize the 20 CV replications across the four cores using R's `parallel` package. Unsurprisingly non-robust EN and adaptive EN estimates are substantially faster to compute than the robust estimates. Adaptive PENSE takes on average about 1.5 minutes for the simplest setting with $p = 32$ and up to 45 minutes for $p = 512$. I-LAMM with the LASSO penalty takes about 5 minutes in the smallest setting and more than 1.5 hours in the largest setting. When using the SCAD penalty, I-LAMM requires 15 minutes for $p = 32$ and on average 5 hours for $p = 128$. Because of the prohibitively long computation time for $p = 512$, I-LAMM (SCAD) has been omitted from this high-dimensional setting. Similar results are shown in Figures S.11 and S.13 in the supplementary materials for the other simulation scenarios. These timings demonstrate that the algorithm implemented in the `pense` package make it possible to apply adaptive PENSE and the other robust estimators in a wide range of settings.

The simulation study highlights that adaptive PENSE and the strategy for choosing hyper-parameters are highly resilient towards many different forms of contamination in the predictors, even if they occur in tandem with heavy-tailed errors and gross outliers in the residuals. Both the prediction accuracy and the variable selection performance are on-par and most often better than other robust and non-robust regularized estimators for high-dimensional regression. While adaptive MM often leads to similar results as adaptive PENSE, in some situations it is substantially more affected by contamination. This is particularly noticeable in settings with severe contamination presented in Section S5.3 in the supplementary materials. There are several instances in which adaptive MM has more than 30% higher prediction error than adaptive PENSE. Figure S.10 in the supplementary material shows that the higher prediction error in these instances is connected with poor variable selection by adaptive MM, likely caused by adverse contamination in combination with heavy-tailed errors. This underscores the issue of relying on an auxiliary scale estimate, as it can lead to an unexpected breakdown of the estimator.

5.2. Real-data example

We apply adaptive PENSE and the other methods considered in the simulation study to the analysis of the chemical composition of 180 archaeological glass vessels from the 15–17th century (Janssens et al., 1998). The analysis is performed on electron probe X-ray micro analysis spectra comprising 1920 frequencies. This data set has been analyzed in other papers on robust high-dimensional regression (e.g., Smucler and Yohai, 2017; Loh, 2021) as it is known to contain observations contaminated in the response and the frequency spectrum (Maronna, 2011). The X-ray spectra are available in the R package *cellWise* (Raymaekers and Rousseeuw, 2021), and the chemical composition in the R package *chemometrics* (Filzmoser and Varmuza, 2017). Of the 1920 frequencies available, only 487 frequencies with meaningful variation between vessels are used for the analysis, in line with the analyses conducted in comparable studies. The goal is to predict the concentration of the chemical compound P_2O_5, measured as the total amount of the compound relative to the total weight of the glass fragment [% w/w]. To get a predictive model, we model the log-concentration of P_2O_5 as a linear function of the spectrum. With similar dimensions and potential contamination as analyzed in the above simulation study, we can be confident that both adaptive PENSE and adaptive MM are good candidates for fitting this predictive model.

The breakdown point for the robust estimators is set to 28%, allowing up to 50 observations with contaminated residuals. Section S4 in the supplementary materials demonstrate that adaptive PENSE is not sensitive to the exact choice of the breakdown point, as similar results are achieved for breakdown points between 20%–33%. Hyper-parameters are selected via six-fold CV, repeated 20 times for all considered estimators. We consider 50 penalization levels for all estimators. The α parameter for all EN-type estimators is selected from the set $\{0.5, 0.75, 1\}$, and the ζ parameter for all estimators with

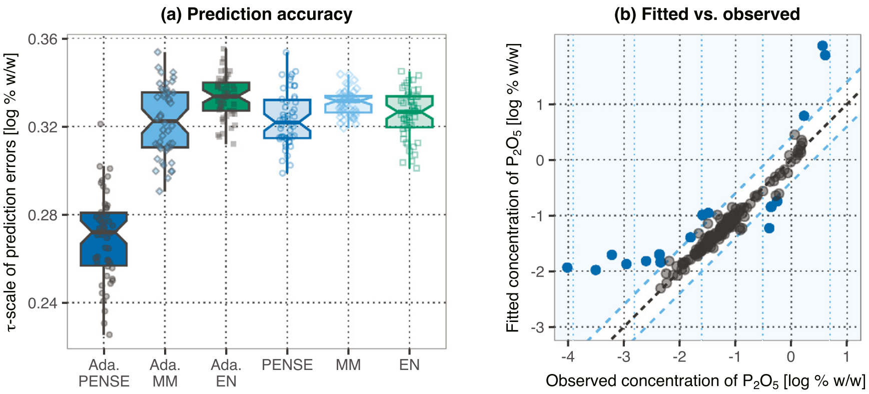

Fig. 6. Accuracy of predicting the concentration of compound P_2O_5 from the glass-vessel data set (a) and observed vs. fitted values from adaptive PENSE (b). The prediction accuracy in (a) is estimated by the uncentered τ-scale of the prediction errors from 50 replications of six-fold cross-validation. Hyper-parameters are selected independently in each CV via an inner six-fold CV with 20 replications. Adaptive PENSE is outperforming the other methods for predicting the concentration. In (b) the 14 data points in the shaded areas have unusually low or high log-concentrations of $\log(P_2O_5)$ ($>$ three times the estimated residual scale).

adaptive EN penalty is chosen from $\{1, 2\}$. The I-LAMM estimator is computed with the LASSO and the SCAD penalty, each considering five different robustness parameters for Huber's loss (the default in the package).

Prediction accuracy is estimated via a nested cross-validation, with the outer loop comprising 50 replications of six-fold CV. In each outer CV fold, the hyper-parameters are chosen via an inner six-fold CV using 20 replications. The prediction accuracy in each outer CV is estimated by the uncentered τ-scale of the prediction errors for all 180 observations.

Fig. 6 (a) shows that adaptive PENSE clearly outperforms the other methods for predicting the concentration of P_2O_5 from the frequency spectrum. Both I-LAMM estimators are omitted from these plots because their prediction accuracies are substantially worse than the other methods (median τ-scale of the prediction error of $0.63 \log \% w/w$), almost identical to the intercept-only model. The non-robust and non-adaptive estimators all yield very similar prediction accuracy. These estimators and adaptive MM select substantially fewer frequencies (Median ≤ 22) than adaptive PENSE (Median $= 44$) in the 50 replications. It seems more frequencies are beneficial for prediction, and the fact that less reliable estimators are missing these frequencies suggests a heavy-tailed error distribution and unusual values in the spectrum have deleterious effects on most methods except for adaptive PENSE.

The adaptive PENSE fit computed on all 180 glass vessels, shown in Fig. 6 (b), suggests 16 vessels have unusually large residuals. The adaptive PENSE fit also suggests a heavy-tailed error distribution, further explaining why it gives better predictions than other estimators in this application and why I-LAMM estimators fail to select any predictors. As demonstrated in the experiments above, in such a setting with heavy-tailed errors the contamination in the frequency spectrum can have detrimental effects on most estimators. Adaptive PENSE has been shown to retain reliable variable selection performance in such a challenging scenario, selecting 43 frequencies belonging to several groups of adjacent and hence highly correlated frequencies.

6. Conclusions

Extreme predictor values, paired with heavy-tailed error distributions or even gross outliers in the residuals can have severe ramifications for statistical analyses if not handled properly. Particularly in high-dimensional problems such extreme values are highly likely. Whether these extreme values have a detrimental effect on the statistical analysis, however, is unknown. Omitting affected observations or predictors is therefore ill-advised and even fallacious; these anomalous values are often well hidden in the multivariate structure and thus difficult if not impossible to detect. We propose the adaptive PENSE estimator which can cope with such unusual values in the predictors and remove their influence if they are paired with aberrant residuals.

We demonstrate that adaptive PENSE leads to estimates with high prediction accuracy and reliable variable selection even under adverse contamination. Unlike other robust estimators, adaptive PENSE is capable of correctly screening out truly irrelevant predictors even if they contain unusual values. The extensive simulation study shows that adaptive PENSE achieves overall better prediction and variable selection performance than competing regularized estimators. While adaptive MM performs better for Normal errors and often similar for heavy-tailed error distributions, in some settings adaptive MM can be substantially affected by the contamination. Overall, adaptive PENSE is more resilient and reliable in challenging scenarios. This is underscored by adaptive PENSE's superior prediction performance for predicting the concentration of the compound P_2O_5 in ancient glass vessels from their spectra. Adaptive PENSE not only achieves comparable or better prediction accuracy in our numerical studies, its variable selection performance is more reliable, as demonstrated by its stability across all considered contamination scenarios. In addition to the strong empirical performance, we have established theoretical guarantees for adaptive PENSE. It is asymptotically able to uncover the true set of relevant predictors and the estimate of the respective coefficients converges to a Normal distribution. Importantly, these guarantees hold regardless of the tails of the error distribution and overall very mild assumptions on the distribution of the predictors or the residuals.

When using adaptive PENSE, a conservative upper limit for the proportion of contaminated response values should be provided for the desired breakdown point. If the breakdown point is set to low, adaptive PENSE can break down and it may be difficult to diagnose this problem if the outliers are masked. Using the maximal breakdown point of 50%, on the other hand, would yield a highly robust estimate, but at the expense of efficiency if substantially fewer response values are outlying. It is thus advisable to draw on subject-level expertise or prior knowledge to select a lower breakdown point if reasonable. As pointed out by an anonymous reviewer, exploring techniques for choosing the breakdown point adaptively based on the data at hand, similar to the method in Xiong and Joseph (2013), could lead to an adaptive PENSE estimator achieving high robustness and efficiency. Important for high-dimensional data sets, however, extreme predictor values are of concern to adaptive PENSE only if they occur in observations with outlying response value. It is therefore sufficient to have an upper limit on the number of contaminated response values, but it is not necessary to know the number of unusual values hidden in the predictors.

Computing adaptive PENSE estimates is challenging, but the proposed computational solutions ensure adaptive PENSE is applicable to many analyses. With these computational solutions, adaptive PENSE is computable in a reasonable time for data sets of moderate size on modern personal computers. Moreover, the algorithms are designed to efficiently leverage parallel computing resources, making large data sets amenable to adaptive PENSE on high-performance computing clusters. These efficient algorithms and the resilient hyper-parameter search are readily available in the R package `pense` provided in the supplementary materials. The superior reliability even under adverse contamination in predictors and responses alike, combined with the easy-to-use and scalable computational tools, make adaptive PENSE an ideal choice in a wide range of applications.

Acknowledgement

The author would like to thank anonymous referees for thoughtful comments on a previous version of this manuscript, and Gabriela V. Cohen Freue for numerous stimulating discussions on the topic of robust regularized estimators and her feedback on early drafts of this manuscript. Ezequiel Smucler provided important input and proofreading during the development of the proofs. Keith Crank gave valuable writing suggestions to polish the final draft. The numerical studies and the real-world application were enabled by computing resources provided by the Digital Research Alliance of Canada (https://alliancecan.ca), as well as by resources provided by the Office of Research Computing at George Mason University (https://orc.gmu.edu) and funded in part by grants from the National Science Foundation (Awards Number 1625039 and 2018631).

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2023.107730.

References

Alfons, A., Croux, C., Gelper, S., 2013. Sparse least trimmed squares regression for analyzing high-dimensional large data sets. Ann. Appl. Stat. 7, 226–248. https://doi.org/10.1214/12-aoas575.

Cohen Freue, G.V., Kepplinger, D., Salibián-Barrera, M., Smucler, E., 2019. Robust elastic net estimators for variable selection and identification of proteomic biomarkers. Ann. Appl. Stat. 13, 2065–2090. https://doi.org/10.1214/19-AOAS1269.

Fan, J., Guo, S., Hao, N., 2012. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. J. R. Stat. Soc., Ser. B, Stat. Methodol. 74, 37–65. https://doi.org/10.1111/j.1467-9868.2011.01005.x.

Fan, J., Fan, Y., Barut, E., 2014. Adaptive robust variable selection. Ann. Stat. 42, 324–351. https://doi.org/10.1214/13-AOS1191.

Fan, J., Liu, H., Sun, Q., Zhang, T., 2018. I-LAMM for sparse learning: simultaneous control of algorithmic complexity and statistical error. Ann. Stat. 46, 814–841. https://doi.org/10.1214/17-AOS1568.

Filzmoser, P., Varmuza, K., 2017. chemometrics: multivariate statistical analysis in chemometrics. https://CRAN.R-project.org/package=chemometrics. r package version 1.4.2.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33, 1–22. https://doi.org/10.18637/jss.v033.i01.

Gijbels, I., Vrinssen, I., 2015. Robust nonnegative garrote variable selection in linear regression. Comput. Stat. Data Anal. 85, 1–22. https://doi.org/10.1016/j.csda.2014.11.009.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, 2nd ed. Springer, New York, NY.

Hastie, T., Tibshirani, R., Tibshirani, R., 2020. Best subset, forward stepwise or lasso? Analysis and recommendations based on extensive comparisons. Stat. Sci. 35, 579–592. https://doi.org/10.1214/19-STS733.

Hössjer, O., 1992. On the optimality of S-estimators. Stat. Probab. Lett. 14, 413–419. https://doi.org/10.1016/0167-7152(92)90103-C.

Insolia, L., Kenney, A., Chiaromonte, F., Felici, G., 2022. Simultaneous feature selection and outlier detection with optimality guarantees. Biometrics 78, 1592–1603. https://doi.org/10.1111/biom.13553.

Janssens, K.H., Deraedt, I., Schalm, O., Veeckman, J., 1998. Composition of 15–17th century archaeological glass vessels excavated in Antwerp, Belgium. In: Love, G., Nicholson, W.A.P., Armigliato, A. (Eds.), Modern Developments and Applications in Microbeam Analysis. Springer, pp. 253–267.

Lambert-Lacroix, S., Zwald, L., 2011. Robust regression through the Huber's criterion and adaptive lasso penalty. Electron. J. Stat. 5, 1015–1053. https://doi.org/10.1214/11-EJS635.

Lange, K., 2016. MM Optimization Algorithms. Society for Industrial and Applied Mathematics.

Loh, P.L., 2017. Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. Ann. Stat. 45, 866–896. https://doi.org/10.1214/16-AOS1471.

Loh, P.L., 2021. Scale calibration for high-dimensional robust regression. Electron. J. Stat. 15, 5933–5994. https://doi.org/10.1214/21-EJS1936.

Maronna, R., 2011. Robust ridge regression for high-dimensional data. Technometrics 53, 44–53. https://doi.org/10.1198/TECH.2010.09114.

Maronna, R., Martin, D., Yohai, V., Salibián-Barrera, M., 2019. Robust Statistics: Theory and Methods (with R). Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ.

Maronna, R.A., Zamar, R.H., 2002. Robust estimates of location and dispersion for high-dimensional datasets. Technometrics 44, 307–317. https://doi.org/10.1198/004017002188618509.

Pan, X., Sun, Q., Zhou, W.X., 2021. Iteratively reweighted ℓ_1-penalized robust regression. Electron. J. Stat. 15, 3287–3348. https://doi.org/10.1214/21-EJS1862.

Raymaekers, J., Rousseeuw, P., 2021. cellWise: analyzing data with cellwise outliers. https://CRAN.R-project.org/package=cellWise. r package version 2.2.5.

Reid, S., Tibshirani, R., Friedman, J., 2016. A study of error variance estimation in lasso regression. Stat. Sin. 26, 35–67. https://doi.org/10.5705/ss.2014.042.

Rousseeuw, P.J., Van Driessen, K., 2006. Computing LTS regression for large data sets. Data Min. Knowl. Discov. 12, 29–45. https://doi.org/10.1007/s10618-005-0024-4.

Rousseeuw, P.J., Yohai, V.J., 1984. Robust regression by means of S-estimators. In: Franke, J., Härdle, W., Martin, D. (Eds.), Robust and Nonlinear Time Series Analysis. Springer, New York, NY, pp. 256–272.

Salibián-Barrera, M., Yohai, V.J., 2006. A fast algorithm for S-regression estimates. J. Comput. Graph. Stat. 15, 414–427. https://doi.org/10.1198/106186006X113629.

She, Y., Owen, A.B., 2011. Outlier detection using nonconvex penalized regression. J. Am. Stat. Assoc. 106, 626–639. https://doi.org/10.1198/jasa.2011.tm10390.

She, Y., Wang, Z., Shen, J., 2022. Gaining outlier resistance with progressive quantiles: fast algorithms and theoretical studies. J. Am. Stat. Assoc. 117, 1282–1295. https://doi.org/10.1080/01621459.2020.1850460.

Smucler, E., Yohai, V.J., 2017. Robust and sparse estimators for linear regression models. Comput. Stat. Data Anal. 111, 116–130. https://doi.org/10.1016/j.csda.2017.02.002.

Sun, Q., Zhou, W.X., Fan, J., 2019. Adaptive Huber regression. J. Am. Stat. Assoc. 115, 254–265. https://doi.org/10.1080/01621459.2018.1543124.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc., Ser. B, Stat. Methodol. 58, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Tibshirani, R.J., Rosset, S., 2019. Excess optimism: how biased is the apparent error of an estimator tuned by SURE? J. Am. Stat. Assoc. 114, 697–712. https://doi.org/10.1080/01621459.2018.1429276.

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. https://doi.org/10.1198/073500106000000251.

Xiong, S., Joseph, V.R., 2013. Regression with outlier shrinkage. J. Stat. Plan. Inference 143, 1988–2001. https://doi.org/10.1016/j.jspi.2013.06.007.

Yohai, V.J., 1987. High breakdown-point and high efficiency robust estimates for regression. Ann. Stat. 15, 642–656. https://doi.org/10.1214/aos/1176350366.

Zou, H., 2006. The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. 101, 1418–1429. https://doi.org/10.1198/016214506000000735.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc., Ser. B, Stat. Methodol. 67, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00527.x.

Zou, H., Yuan, M., 2008. Composite quantile regression and the oracle model selection theory. Ann. Stat. 36, 1108–1126. https://doi.org/10.1214/07-AOS507.

Zou, H., Zhang, H.H., 2009. On the adaptive elastic-net with a diverging number of parameters. Ann. Stat. 37, 1733–1751. https://doi.org/10.1214/08-AOS625.