# PENSE: A Penalized Elastic Net S-Estimator

Gabriela V. Cohen Freue, David Kepplinger
and Matías Salibián-Barrera, Ezequiel Smucler
Department of Statistics, University of British Columbia, Vancouver, BC, Canada

October 26, 2017

## Abstract

Penalized regression estimators have been widely used in recent years to improve the prediction properties of linear models, particularly when the number of explanatory variables is large. It is well-known that different penalties result in regularized estimators with varying statistical properties. Motivated by the analysis of plasma proteomic biomarkers that tend to form groups of correlated predictors, we focus here on estimators with an Elastic Net penalty, in order to keep these groups of variables together as they enter or leave the model. Given the presence of potential outliers in our data, we propose a class of penalized S-estimators which have very good robustness properties. Furthermore, these penalized S-estimators can be used as initial values to compute more efficient penalized M-estimators. In this paper we derive an algorithm to compute our proposed estimators, and also a data-driven method to select the penalty term, which is a critical part of any application with real data. Our robust penalized estimators have very good robustness properties and are also consistent under relatively weak assumptions. Our numerical experiments show that our proposals compare favourably to other robust penalized estimators. When applied to our motivating example, the robust estimators identify new potentially relevant biomarkers that are not found with non-robust alternatives. Moreover, the robust estimators identify two patients with a suspected low obstruction in the artery examined. Further measurements by a more accurate technique validated our predictions.

# 1 Introduction

Biomarkers are indicators used to measure pathogenic processes or responses to therapies. Recent advances in various -omics technologies allow for the simultaneous quantification of thousands of molecules (e.g., genes and proteins) revolutionizing the way that scientists search for molecular biomarkers. For example, in the search of biomarkers of a given disease, mass spectrometry shotgun proteomic techniques can be used to measure the abundance of hundreds of proteins that have not been previously hypothesized to be related with that disease, which can result in the discovery of novel biomarkers. To date, the innovation of technical resources available for -omic biomarker studies is well recognized. Nevertheless, the development of statistical and computational methods to analyze large and complex -omics datasets is of fundamental importance to succeeding in the validation and clinical implementation of biomarker discoveries.

In particular in this paper, we use a linear regression to model the association between hundreds of plasma protein levels and the obstruction of the left anterior descending artery measured in patients who developed cardiac allograft vasculopathy (CAV), a major complication suffered by 50% cardiac transplant recipients beyond the first year after transplant. Although hundreds of proteins were measured and analyzed in these patients, only a few proteins are expected to be associated with the observed artery obstruction, resulting in a sparse regression model (i.e., most regression coefficients equal to zero). Identifying these plasma proteomic biomarkers can result in the development of clinically useful blood tests to diagnose CAV and improve patient care options.

Penalized regression estimators have been proposed to identify a relatively small subset of explanatory variables to obtain good predictions for a response when the number of covariates is large (even larger than the number of observations) [1, 2]. However, most of these estimators are extremely sensitive to outliers. Since -omics datasets usually contain outlying observations associated, for example, with technical problems in sample preparation or patients with rare molecular profiles, the use of a robust penalized estimator is essential to effectively interrogate the rich information contained in the human genome.

Although many robust regression methods have been proposed in the literature (see Maronna et al. [3] for a review), the development of *penalized* robust estimation methods is still in its early stages. Most of the existing work is focused on different penalized versions of convex M-estimators [4, 5], thus not resistant to high leverage outliers commonly observed in large datasets. One of the

first highly robust penalized estimators is the RLARS estimator [6], a modification of the Least Angle Regression method [7] where sample correlations are replaced with robust counterparts. A more recent proposal, SparseLTS [8], is an $L_1$-regularized version of the Least Trimmed Squares regression estimator Rousseeuw [9], which can be shown to have good robustness properties. Both of these estimators are useful for variable selection, but can only be tuned to be either highly robust or highly efficient under the normal model [10].

To overcome these limitations, Maronna [11] has recently proposed an MM-estimator with a ridge penalty to ensure robustness to outliers and leverage points, as well as high efficiency under the normal model. Although the proposed MM-Ridge regression estimator has good prediction performance even in contaminated samples, it does not produce sparse solutions and hence cannot be used for variable selection. To address this issue, Smucler and Yohai [12] have recently proposed a penalized MM-LASSO estimator. However, as previously shown for the classical LASSO [13], their MM-LASSO estimator cannot select more variables than the number of available observations, and if the data contain groups of highly correlated explanatory variables, it tends to randomly select only one variable within each group ignoring the relevance of other covariates.

In usual -omics datasets, the number of measured features is much larger than the number of samples and genes belonging to the same pathway or biological process form groups of correlated variables. Thus, the limitations of Ridge and LASSO methods can jeopardize the discovery of clinically useful biomarkers. In this study, we propose two penalized robust regression estimators using the elastic net penalty, a linear combination between the $L_2$ penalty of Ridge and the $L_1$ penalty of LASSO, which can be tuned to estimate models with different levels of sparsity and a complex correlation structure among covariates. First, we derive the Penalized Elastic Net S-Estimator (PENSE) by penalizing a *robust* (squared) scale function of the residuals, instead of the usual sum of square of the residuals. Second, to get an estimator that is highly robust and at the same time efficient, we use PENSE to initialize a penalized M-regression estimator with the same penalty as that used by PENSE. We call the resulting estimator PENSEM. The proposed estimators can be seen as robust versions of the classical elastic net (EN) class of estimators that contains Ridge and LASSO as special cases [2]. Our estimators are applicable to a wide range of complex high-dimensional datasets commonly found in data science to select explanatory variables while shrinking their estimated coefficients to improve the prediction of the response of interest.

In Sections 2 and 3 we present the PENSE and PENSEM estimators, respectively, along

with efficient algorithms to compute them. In Section 4, we study some important theoretical properties of our estimators. In Section 5, we assess their performance and compare them with other available estimators in a simulation study. In Section 6 we use PENSE to identify potential proteomic biomarkers of CAV and to predict the left anterior descending artery obstruction using the fit model in patients who received a cardiac transplant. In Section 7 we conclude.

# 2 PENSE: a new robust penalized regression estimator

As we mentioned before, the relation between molecular features and a disease of interest can be modelled by a linear regression model:

$$y_i = \mu + \mathbf{x}_i^\intercal \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\mu \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^p$ are the regression coefficients to be estimated from the observed data. In a biomarkers discovery study, the response variable, $y_i \in \mathbb{R}$, measures the status of a disease (e.g., stenosis of a coronary artery) for the $i$-th subject and the set of covariates, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\intercal \in \mathbb{R}^p$, are the measurements of all features (e.g., protein levels). We assume that the response is centered and the covariates are standardized. In Section 5 we explain how to make this transformation robustly. Although thousands of molecular features may be measured and analyzed in -omics studies, only a few are usually expected to be related with a given disease. In other words, we expect this model to be sparse with most coefficients equal to zero.

When the number of observations is smaller than the number of covariates, ordinary least squares (OLS) is a widely used method to estimate the unknown coefficients of a regression model. However, OLS is extremely sensitive to outlying observations in the data. Noting that OLS is the minimizer of a non-robust scale measure of the residuals (the sum of their squares), Rousseeuw and Yohai [14] introduced a family of highly robust regression estimators by replacing the sum of squared residuals with a robust residual scale estimator (e.g., an M-scale estimator [15]). The resulting S-estimators have very good robustness and asymptotic properties [3].

In this Section, we propose to penalize the loss function of S-estimators using an elastic net penalty to estimate sparse regression models and thus select the most relevant variables in the model. More specifically, this new robust penalized estimator, which we call PENSE, is defined as

the minimizer $\hat{\boldsymbol{\theta}}^{PS} = \left(\hat{\mu}^{PS}, \hat{\boldsymbol{\beta}}^{PS}\right)$ of the penalized loss function

$$\mathcal{L}_{\text{PS}}(\mu, \boldsymbol{\beta}) = \hat{\sigma}\left(\mu, \boldsymbol{\beta}\right)^2 + \lambda_S \left(\frac{1}{2}(1-\alpha)||\boldsymbol{\beta}||_2^2 + \alpha||\boldsymbol{\beta}||_1\right) \tag{2}$$

for a level of penalty $\lambda_S \geq 0$, and a combination of the $L_1$ and $L_2$ penalties given by $\alpha \in [0, 1]$. In particular, if $\alpha = 1$, the estimator becomes a LASSO S-estimator, and if $\alpha = 0$, it becomes a Ridge S-estimator.

The robust scale M-estimate, $\hat{\sigma}\left(\mu, \boldsymbol{\beta}\right)$ is defined as the solution of

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{y_i - \mu - \mathbf{x}_i^{\intercal}\boldsymbol{\beta}}{\sigma\left(\mu, \boldsymbol{\beta}\right)}\right) = \delta, \tag{3}$$

for an even and bounded function $\rho$ and tuning constant $\delta \in (0, 1)$. Both $\rho$ and $\delta$ need to be chosen jointly in order to obtain consistent estimators under the normal model. For more details we refer to Maronna et al. [3]. In this paper we use $\rho$ functions in the Tukey's bisquare family, which is given by

$$\rho_c(t) = \min\left\{1, 1 - (1 - (t/c)^2)^3\right\},$$

where $c > 0$ is a parameter that determines the estimator's breakdown point [3].

Minimizing the objective function (2) is challenging due to its non-convexity and the lack of differentiability of the EN penalty at $\boldsymbol{\beta} = \mathbf{0}$. However, since the unpenalized S loss is continuously differentiable and the EN penalty is locally Lipschitz, the penalized S loss (2) is locally Lipschitz and its generalized gradient as defined in Clarke [16] is given by the derivative of the unpenalized S loss, shifted by the generalized gradient of the EN penalty. Moreover, the following Lemma gives a necessary condition for the minimizer of (2).

**Lemma 2.1.** *Let $f(\mathbf{x}) : \mathbb{R}^p \to \mathbb{R}$ be locally Lipschitz around $\mathbf{x}_0$ with generalized gradient $\nabla_{\mathbf{x}} f(\mathbf{x}_0)$. If $f$ attains a local minimum at the point $\mathbf{x}_0$ then $\mathbf{0} \in \nabla_{\mathbf{x}} f(\mathbf{x}_0)$ [16].*

To characterize the local minima of the penalized S loss (2), the gradient of the scale M-estimator of the residuals is required. This can be easily derived by taking the gradient with

respect to $(\mu, \boldsymbol{\beta})$ on both sides of equation (3). Let

$$r_i(\mu, \boldsymbol{\beta}) = y_i - \mu - \mathbf{x}_i^\intercal \boldsymbol{\beta} \text{ and } \tilde{r}_i(\mu, \boldsymbol{\beta}) = \frac{r_i(\mu, \boldsymbol{\beta})}{\sigma(\mu, \boldsymbol{\beta})}$$

denote the residuals and the standardized residuals, respectively. Then,

$$\nabla_{(\mu,\boldsymbol{\beta})} \sigma(\mu, \boldsymbol{\beta})^2 = -\frac{2}{n} \sum_{i=1}^{n} r_i(\mu, \boldsymbol{\beta}) w_i(\mu, \boldsymbol{\beta}) \begin{pmatrix} 1 \\ \mathbf{x}_i \end{pmatrix},$$

with weights

$$w_i(\mu, \boldsymbol{\beta}) = \frac{\rho_c'(\tilde{r}_i(\mu, \boldsymbol{\beta}))/\tilde{r}_i(\mu, \boldsymbol{\beta})}{\frac{1}{n}\sum_{j=1}^{n} \rho_c'(\tilde{r}_j(\mu, \boldsymbol{\beta}))\tilde{r}_j(\mu, \boldsymbol{\beta})}.$$

The generalized gradient of the penalized S loss is thus given by

$$\nabla_{(\mu,\boldsymbol{\beta})} \mathcal{L}_{\text{PS}}(\mu, \boldsymbol{\beta}) = 2 \left[ -\frac{1}{n} \sum_{i=1}^{n} r_i(\mu, \boldsymbol{\beta}) w_i(\mu, \boldsymbol{\beta}) \begin{pmatrix} 1 \\ \mathbf{x}_i \end{pmatrix} + \frac{\lambda_S}{2} \begin{pmatrix} 0 \\ \nabla_{\boldsymbol{\beta}} P_\alpha(\boldsymbol{\beta}) \end{pmatrix} \right], \tag{4}$$

where $P_\alpha(\boldsymbol{\beta}) = \frac{1}{2}(1-\alpha)||\boldsymbol{\beta}||_2^2 + \alpha||\boldsymbol{\beta}||_1$ is the EN penalty. Care must be taken to ensure that the weights do not change the size of the loss function relative to the penalty. Therefore, similar to the unweighted case, the weights are standardized to sum to $n$, i.e., the weights $w_i(\mu, \boldsymbol{\beta})$ in (4) are replaced by their standardized version:

$$w_i^*(\mu, \boldsymbol{\beta}) = n \frac{w_i(\mu, \boldsymbol{\beta})}{\sum_{j=1}^{n} w_j(\mu, \boldsymbol{\beta})}$$

Note that the generalized gradient of the penalized S loss coincides with the subgradient of the classical weighted Elastic Net loss, except that the weights of the former depend on the unknown coefficients $(\mu, \boldsymbol{\beta})$, thus suggesting an iteratively reweighted EN (IRWEN) algorithm to find local optima of the penalized S loss.

## 2.1 Algorithm

Given a fixed penalty parameter $\lambda_S$, an initial estimate $\boldsymbol{\theta}^{\text{init}}$ and its corresponding M-scale estimate $\hat{\sigma}(\boldsymbol{\theta}^{\text{init}})$, we compute PENSE using an IRWEN algorithm. More specifically, we use the initial

estimates to compute weights $w_i^{(0)} = w_i(\boldsymbol{\theta}^{\text{init}})$, with which we obtain a new estimate $\boldsymbol{\theta}^{(1)}$ optimizing the classical weighted EN objective function. This updated estimate is in turn used to compute the new weights $w_i^{(1)} = w_i(\boldsymbol{\theta}^{(1)})$ that define an updated weighted EN objective function. These steps are iterated until the relative change in the coefficient estimate after step $k \geq 1$ is negligible:

$$\frac{\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^{(k-1)}\|_2}{\|\boldsymbol{\theta}^{(k)}\|_2 + \|\boldsymbol{\theta}^{(k-1)}\|_2} < \tau,$$

for a chosen tolerance level $\tau$.

Given an efficient implementation of the IRWEN algorithm, finding local optima of (2) is computationally relatively cheap. However, in order for the above iterations to converge to a good local optimum (or the global minimum), it is necessary to find a "good" starting point for the IRWEN algorithm. The next Section discusses a strategy to find such an initial value for IRWEN.

## 2.2 Initial estimator

The challenge of finding initial estimators for the unpenalized S-estimator of regression has been extensively studied in the literature and many reliable and fast procedures have been proposed [e.g., 17, 18]. This is not true, however, for penalized robust regression.

Similar to the unpenalized problem, Alfons et al. [8] initialized their algorithm to calculate SparseLTS using random subsampling, which selects an initial estimator from a set of classical LASSO estimators calculated from subsamples of size $m$, randomly chosen from the $n$ observations. To increase the chance of obtaining an outlier-free subsample among a moderate set examined, they considered subsamples of size $m = 3$. In case of the LASSO penalty, however, this implies that any initial estimator will have at most 3 non-zero estimated coefficients. Although their simulation study shows promising results in very sparse settings, the chance of converging to a good local optimum using initial estimators with only 3 non-zero estimated coefficients decreases with the level of sparsity of the model (i.e., more true non-zero coefficients). While the number of subsamples and their sizes should be guided by a prior assumption on the sparsity of the final solution, the computational time of the algorithm considerably increases as more and larger subsamples are considered.

Maronna [11] used a different strategy to find a good initial estimator to compute the S-Ridge

estimator. Instead of randomly choosing $m$ observation to create a random subsample, he removes potentially outlying observations to build "clean" subsamples using the principal sensitivity components (PSCs) introduced by Peña and Yohai [19] for the OLS estimator. The PSCs are defined as the principal components of the sensitivity observations $\mathbf{r}_i$ given by

$$\mathbf{r}_i = \left(\hat{y}_i - \hat{y}_{i(1)}, \ldots, \hat{y}_i - \hat{y}_{i(n)}\right)^T, \; i = 1, \ldots, n, \tag{5}$$

to find the directions of maximum change in the prediction of the $i$th observation when each point (in parenthesis) is removed from the sample. Expressing the Ridge regression optimization problem as an extended OLS one, Maronna [11] computed the PSCs for Ridge from those of the corresponding extended least squares one. Potentially clean subsamples are obtained by removing the observations with the most extreme values of each PSC.

Since the Ridge penalty does not typically set any coefficient to zero, an initial estimator based on this penalty may result in a poor starting point for reweighted algorithms of other penalized estimators like PENSE. In this paper we compute the PSCs of the EN estimator from its definition, calculating the predictions of the response in (5) with a classical EN estimator.

Specifically, for each of the $\tilde{q} = \text{rank}(\sum_{i=1}^{n} \mathbf{r}_i \mathbf{r}_i^\intercal)$ PSCs, three subsamples are created by removing: i) half of the observations corresponding to its smallest elements, ii) half of the observations corresponding to its largest elements, and iii) half of the observations corresponding to its largest absolute values, respectively. A classical EN estimator is computed on each subsample as well as on the full sample, resulting in a set of $3\tilde{q}+1$ candidate estimators to minimize the penalized scale in (2). Since the PSCs may not detect low-leverage outliers, following a refinement suggested by Peña and Yohai [19], we use the best candidate in this set to compute the residuals of all observations and remove the observations with the largest absolute residuals to recompute the PSCs. This process can be iterated several times or until convergence, resulting in $3\tilde{q}+1$ candidate EN estimators computed on potentially "clean" samples that can be used to initialized the IRWEN algorithm described in Section 2.1.

To save computational time, we run only a small number of iterations of the IRWEN algorithm on all $3\tilde{q}+1$ candidate estimators and select a smaller set of $J$ candidates with lowest values of the objective function (2) to fully iterate the IRWEN algorithm until convergence (or until the maximum number of iterations is reached). We then return the solution with the lowest objective

function (2) found after fully iterating the above set of $J$ candidates.

## 2.3 Computing PENSE on a grid of penalty values

In most applications the optimal level of penalization is not known and is selected based on the performance of the penalized estimator over a grid of possible penalty values, $\lambda_S^{(1)} < \lambda_S^{(2)} < \ldots < \lambda_S^{(K)}$. To ease the burden of computing an initial estimator (or several candidates) based on the EN PSCs to fully iterate the IRWEN algorithm for every $\lambda_S^{(k)}$ in the grid, $k = 1, \ldots, K$, we designed a strategy using "warm" starts, in which a local optimum of (2) at a penalty value in the grid can be used to initiate the algorithm at adjacent penalty levels. Although this "domino effect" is commonly exploited to compute other penalized estimators based on iterative solvers [20, 21], its effectiveness relies on the assumption that estimators at contiguous penalty levels are similar and thus good initial estimators of each other.

The prevalent method of warm starts for penalized estimators is to start with a very large penalty value that shrinks all regression coefficients to zero, thus avoiding the computation of any other initial estimator. However, since the objective function (2) is not convex, this strategy is no longer guaranteed to find a good solution for all $\lambda_S^{(k)}$ in the grid.

Thus, we combine "warm" initial estimates with "cold" initial estimates obtained from EN PSCs to initiate our algorithm, harnessing the benefits of both strategies. The cold initial estimate is computed for only a few values throughout the grid, including the smallest and the largest value, i.e., at $\lambda_S^{(k^*)}$ with $k^* \in \mathcal{K}_{\text{cold}} \subseteq \{1 \ldots, K\}$. The grid is traversed first from the largest value $\lambda_S^{(K)}$ to the smallest value $\lambda_S^{(1)}$, using at each step $k$ the warm estimate from the previous (larger) penalty value $\lambda_S^{k+1}$ and, if available, the cold initial estimates for the current penalty value $\lambda_S^{(k)}$ as potential initial estimators. At nodes with two potential initial estimators (cold and warm, i.e., at $k \in \mathcal{K}_{\text{cold}}$), both candidates are fully iterated at $\lambda_S^{(k)}$ and the solution with lowest objective value is retained. At the end of this process, we obtain one fully iterated estimate at every $\lambda_S$ value in the grid. The process is then repeated in the reverse direction, from smallest to largest $\lambda_S$, obtaining a second estimate at every $\lambda_S^{(k)}$, $k = 1, \ldots, K$, in the grid. At each penalty value, we select the estimate that minimizes the penalized S-loss in (2) resulting in a full regularization path of PENSE across the grid.

# 3 PENSEM: a new penalized MM-estimator

Since there is a trade-off between the efficiency of the S-estimators and their robustness to outlying points (see Hössjer [22]), Yohai [10] proposed the class of MM-estimator, which can maintain both robustness and high-efficiency. It is then natural to also refine the robust PENSE estimator to obtain a penalized elastic net MM-estimator, which we call PENSEM, defined as the minimizer $\hat{\boldsymbol{\theta}}^{PM} = \left(\hat{\mu}^{PM}, \hat{\boldsymbol{\beta}}^{PM}\right)$, of the penalized loss function

$$\mathcal{L}_{\text{PM}}(\mu, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \rho_{c'}\left(\frac{y_i - \mu - \mathbf{x}_i^\intercal \boldsymbol{\beta}}{\hat{\sigma}_0}\right) + \lambda_M \left(\frac{1}{2}(1-\alpha)||\boldsymbol{\beta}||_2^2 + \alpha||\boldsymbol{\beta}||_1\right), \tag{6}$$

where $\rho_{c'}$ is a loss function satisfying $\rho_{c'}(x) \leq \rho_c(x)$ for all $x$ and the scale $\hat{\sigma}_0$ used to standardize the residuals is fixed. The generalized gradient of this objective function is given by

$$\nabla_{(\mu,\boldsymbol{\beta})}\mathcal{L}_{\text{PM}}(\mu, \boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n} r_i(\mu, \boldsymbol{\beta})\tilde{w}_i(\mu, \boldsymbol{\beta}) \begin{pmatrix} 1 \\ \mathbf{x}_i \end{pmatrix} + \lambda_M \begin{pmatrix} 0 \\ \nabla_{\boldsymbol{\beta}} P_\alpha(\boldsymbol{\beta}) \end{pmatrix},$$

with weights

$$\tilde{w}_i(\mu, \boldsymbol{\beta}) = \frac{\rho'_{c'}\left(\frac{r_i(\mu,\boldsymbol{\beta})}{\hat{\sigma}_0}\right)}{\hat{\sigma}_0 r_i(\mu, \boldsymbol{\beta})}.$$

As previously discussed for PENSE, we can use an IRWEN algorithm to find the local minima of the penalized M loss function (6) using normalized weights. Although both PENSE and PENSEM are defined using the same penalty function, the levels of penalization implied by specific values of $\lambda_M$ and $\lambda_S$ are not equivalent since the loss functions are generally different. Thus, the optimum penalty parameter $\lambda_M$ for PENSEM is also chosen from a grid of candidate values which might be different from the grid used to determine $\lambda_S$. To ease the computational burden, the initial estimator to optimize the penalized M loss function (6) for any $\lambda_M$ is fixed at the PENSE estimate $\left(\hat{\mu}^{PS}, \hat{\boldsymbol{\beta}}^{PS}\right)$ at the single penalty level $\lambda_S$ chosen for PENSE.

Ideally, one would use an M-scale based on the residuals from PENSE to compute PENSEM, i.e., $\hat{\sigma}_0 = \hat{\sigma}(\hat{\mu}^{PS}, \hat{\boldsymbol{\beta}}^{PS})$. However, as noted by Maronna and Yohai [23] the scale based on the residuals from an S-estimator underestimates the true error scale if the ratio $p/n$ is high. Since we observed a similar phenomenon for the scale of PENSE, we adjust this initial scale using the

corrections suggested for the scale of the unpenalized S-estimators replacing the actual number of parameters by the estimated effective degrees of freedom. More specifically, $\hat{\sigma}_0 = q\hat{\sigma}(\mu^{PS}, \boldsymbol{\beta}^{PS})$, with $q = q_E$ derived from empirical simulations if $0.1 < \widehat{\text{edf}}/n \leq 0.5$, $q = q_T$ derived from the Taylor expansion of the S-estimation equation if $\widehat{\text{edf}}/n > 0.5$, and $q = 1$ (no adjustment) if $\widehat{\text{edf}}/n < 0.1$ [11, 12].

We note that Maronna and Yohai [23] proposed to (re)estimate the M-scale of the S-Ridge using $\delta = \max\{0.25, 0.5(1 - \hat{q}/n)\}$ in (3), where $\hat{q}$ are the effective degrees of freedom of the initial S-Ridge estimator. Although Maronna's correction for $\delta$ addresses a potential underestimation of the scale from robust penalized estimators, it also decreases the robustness of the resulting estimator. Moreover, for $p \gg n$ and estimations with large effective degrees of freedom this correction usually results in the final scale being computed with $\delta = 0.25$. Thus, in this paper we don't adjust the scale (or equivalently its breakdown point) but notice that, in particular for sparse penalties, the estimation of the scale can be improved by choosing a lower breakdown point.

# 4 Properties

In this Section we study some important robustness and statistical properties of the proposed estimators.

## 4.1 Breakdown Point

One measure of robustness of an estimator against potential outliers is its (finite-sample) breakdown point, which measures the largest proportion of observations that when arbitrarily replaced still result in a bounded estimate. Formally, for a given dataset $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^\intercal$, where $\mathbf{z}_i = (y_i, \mathbf{x}_i^\intercal)$, the replacement finite-sample breakdown point (FBP), $\epsilon^*(\hat{\boldsymbol{\theta}}; \mathbf{Z})$, of a regression estimator $\hat{\boldsymbol{\theta}}$ is defined as

$$\epsilon^*(\hat{\boldsymbol{\theta}}; \mathbf{Z}) = \max\left\{\frac{m}{n} : \sup_{\mathbf{Z}_m \in \mathcal{Z}_m} \|\hat{\boldsymbol{\theta}}(\mathbf{Z}_m)\| < \infty\right\}, \tag{7}$$

where the set $\mathcal{Z}_m$ contains all datasets $\mathbf{Z}_m$ with $0 < m < n$ of the original $n$ observations replaced by arbitrary values [24].

Note that regularized estimators are typically defined as the solution of a penalized optimization problem, which is formally equivalent to that of a constrained one. For example, a LASSO estimator can be obtained minimizing the sum of squares of the residuals subject to $\|\boldsymbol{\beta}\|_1 \leq C$, for some $C \geq 0$. Hence, one may suspect that regularized estimators are "automatically" robust, in the sense that they are necessarily constrained and thus bounded. Unfortunately, this is generally not true, since the bound on the equivalent constrained optimization problem depends on the sample, and thus may grow to infinity when outliers are present. To see this in the case of the LASSO estimator, let $\boldsymbol{\beta}^*$ be a minimizer of the penalized sum of squared residuals objective function: $\sum_{i=1}^n (y_i - \mathbf{x}_i^\intercal \boldsymbol{\beta})^2 + \lambda_0 \|\boldsymbol{\beta}\|_1$, for a fixed $\lambda_0 > 0$ (to simplify the presentation we assume that the data are standardized so that no intercept is present in the model). Following the results in Osborne et al. [25], we have $\|\boldsymbol{\beta}^*\|_1 = C_0 = (\mathbf{r}^*)^\intercal \mathbf{X}\boldsymbol{\beta}^*/\lambda_0$, where $\mathbf{r}^* = (r_1^*, \ldots, r_n^*)^\intercal$ is the vector of residuals for $\boldsymbol{\beta}^*$. If $\boldsymbol{\beta}^*$ is different from the usual least squares estimator, it follows that $\boldsymbol{\beta}^*$ also minimizes $\sum_{i=1}^n (y_i - \mathbf{x}_i^\intercal \boldsymbol{\beta})^2$ subject to $\|\boldsymbol{\beta}\|_1 \leq C_0$ (see Osborne et al. [25]). It is easy to see that, since $C_0$ depends on the sample, it can become arbitrarily large when outliers are present in the data.

The following theorem shows that the PENSE estimator retains the high-breakdown point of the parent unpenalized S-estimator. In other words, setting $\delta = 0.5$ in (3) results in a finite sample breakdown point of 50% for PENSE.

**Theorem 4.1.** *For a dataset of size $n$ and $\delta$ defined in* (3)*, let $m(\delta) \in \mathbb{N}$ be the largest integer smaller than $n \min(\delta, 1-\delta)$. Then the finite-sample breakdown point of PENSE is*

$$\frac{m(\delta)}{n} \leq \epsilon^* \left( \hat{\boldsymbol{\theta}}^{PS}; \mathbf{Z} \right) \leq \delta.$$

A proof of the theorem is given in the Appendix (Section 8.1).

Moreover, the proof in Smucler and Yohai [12] can be used to show that the breakdown point of the elastic net penalized MM-estimator is at least as high as the breakdown point of the initial scale estimator. Therefore, PENSEM retains the high breakdown point of PENSE.

## 4.2 Statistical Consistency

We study the consistency of EN penalized S and M-estimators in linear models with fixed predictor variables in the $p \ll n$ regime. In practical problems, the number of predictors used is often large and might depend on the sample size. This motivates the modelling of $p = p_n$, where $p_n$ may diverge to infinity. See for example [26] and [4]. Hence, we will consider a sequence of regression models

$$y_{i,n} = \mathbf{x}_{i,n}^{\intercal}\boldsymbol{\beta}_{0,n} + u_{i,n}, \quad 1 \leq i \leq n \tag{8}$$

where $y_{i,n} \in \mathbb{R}$, $\mathbf{x}_{i,n} \in \mathbb{R}^{p_n}$, $\boldsymbol{\beta}_{0,n} \in \mathbb{R}^{p_n}$ is to be estimated and $u_{i,n}$ are i.i.d. random variables defined in a common probability space with distribution function $F_0$. When the model contains an intercept, we will assume that the first coordinate of $\mathbf{x}_{i,n}$ is equal to one for all $i$ and $n$, and hence that the first coordinate of $\boldsymbol{\beta}_0$ is the intercept, which is not penalized. From now on, we will drop the $n$ subscript from $y_{i,n}, \mathbf{x}_{i,n}$, $\boldsymbol{\beta}_{0,n}$, $p_n$ and $u_{i,n}$.

We will need the following assumptions:

R0. For $k \in \{c, c'\}$, $\rho_k$ is a bounded $\rho$-function in the sense of [3]. Moreover, $\rho_k$ is continuously differentiable and, for some $m > 0$, $\rho_k(t) = 1$ if $|t| \geq m$.

F0. $F_0$ has an absolutely continuous density, $f_0$. $f_0(t)$ is a monotone decreasing function of $|t|$ and a strictly decreasing function of $|t|$ in a neighbourhood of 0.

B0. $\lambda_S P_\alpha(\boldsymbol{\beta}_0)/n \to 0$.

B1. $\lambda_M P_\alpha(\boldsymbol{\beta}_0)/n \to 0$.

X0. For $0 < \gamma < 1$, let

$$\eta_n(\gamma) = \min_{\mathcal{A} \subset \{1,\ldots,n\}, \#\mathcal{A} = [n\gamma]} \min_{\|\theta\|=1} \max_{i \in \mathcal{A}} |\mathbf{x}_i^{\intercal}\theta|.$$

Then $\liminf_n \eta_n(\gamma) > 0$ for some $0 < \gamma < 1$.

Condition [R0] is satisfied by, for example, Tukey's Bisquare loss function. Condition [F0] allows for extremely heavy tailed errors, making no moment assumptions. Conditions [B0], [B1] are typical in the asymptotic analysis of penalized regression estimators and are satisfied if, for example, the coordinates of $\boldsymbol{\beta}_0$ are bounded and $p\,\lambda_S/n \to 0$ and $p\,\lambda_M/n \to 0$. To prove the

consistency of the regression estimators we will need $p/n \to 0$. The function $\eta_n(\gamma)$ appearing in [X0] was introduced in Davies [27] and used extensively in Smucler [28]. In some sense, it measures the worst possible conditioning of any subset of size $[n\gamma]$ of the predictors. A discussion of this assumption, and conditions under which it holds can be found in the Appendix (Section 8.2).

**Theorem 4.2** (Consistency)**.** *Assume [R0], [F0] and [B0] hold and that $p/n \to 0$.*

*i)* $\hat{\sigma}^{PS} \xrightarrow{P} s(F_0)$*, where $s(F_0)$ is the solution of $\mathbb{E}_{F_0}\rho_c(u/s) = \delta$. Moreover, for any $0 < \gamma < 1$,* $\eta_n(\gamma)\|\hat{\boldsymbol{\beta}}^{PS} - \boldsymbol{\beta}_0\| \xrightarrow{P} 0$.

*ii) If [B1] holds, for any $0 < \gamma < 1$,* $\eta_n(\gamma)\|\hat{\boldsymbol{\beta}}^{PM} - \boldsymbol{\beta}_0\| \xrightarrow{P} 0$.

*In particular, if [X0] holds, $\hat{\boldsymbol{\beta}}^{PS}$ and $\hat{\boldsymbol{\beta}}^{PM}$ are consistent.*

# 5 Simulation Studies

In this Section, we assess the performance of the PENSE and PENSEM estimators in four different simulation settings and compare it to other published robust and/or penalized estimators. In all settings, the response $y_i$ is generated from the linear model

$$y_i = \mathbf{x}_i^\intercal \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad 1, \dots, n.$$

The number of observations ($n$), the number of predictors ($p$), the correlation structure of $\mathbf{x}$, and the true regression coefficients ($\boldsymbol{\beta}$) varied among simulation settings.

Before computing PENSE(M), the training data is standardized using robust measures of univariate location and scale. More specifically, the response and the predictors are centered around their univariate median and each predictor is further scaled to have a median absolute deviation of 1.

We compare our PENSE and PENSEM estimators[1] against the classical LASSO and EN, as well as the robust regularized estimators SparseLTS [8] and the recently published MMLASSO [12], which are robust versions of LASSO. To the best of our knowledge, there are no other robust EN estimators available. Whenever possible, we also include the oracle OLS and MM estimators,

[1] available at `https://cran.r-project.org/web/packages/pense/index.html`

which only estimate the coefficients of the true active set of predictors. For SparseLTS, we use the implementation available in the R package robustHD [29], while for MMLASSO we use the functions available in the authors' github repository[2]. Where possible, the robust estimators are tuned to achieve a 25% breakdown point.

## 5.1 The Penalty Parameters

The level of penalization $\lambda_S$ is chosen from a grid of 100 logarithmically equispaced values to optimize PENSE's prediction performance estimated via 10-fold cross-validation (CV). The training sample might contain contaminated observations, thus we use the robust $\tau$-scale [30] of all $n$ out-of-sample predictions to measure prediction performance instead of the usual root mean squared prediction error. Similarly, we compute PENSEM on a grid of 100 logarithmically equispaced values for $\lambda_M$, always starting from the optimum $\lambda_S^*$ chosen from the previous grid. The optimal $\lambda_M^*$ is again chosen by 10-fold CV using the robust $\tau$-scale.

The balance between the $L_1$ and the $L_2$ penalties as controlled by the parameter $\alpha \in [0, 1]$ is being fixed throughout the selection of $\lambda_S$ and $\lambda_M$. Different strategies can be used to select the appropriate $\alpha$ parameter to compute PENSE(M). In many applications, the user selects this value based on the desired level of sparsity of the resulting model. For example, in the proteomics study analyzed in this paper, the identified potential biomarkers were validated by an independent and more precise technology. Thus, the budget available was used to decide on the number of markers that would be selected and migrated to the validation phase. In other contexts, one can compute the estimators for several different values of $\alpha$ and choose the value $\alpha^*$ that yields the best CV prediction performance. For a comprehensive discussion on this topic we refer to Zou and Hastie [2].

As it was noted for the classical naïve EN estimator [2], PENSE and PENSEM suffer from a "double" penalization due to the combination of the $L_1$ and the $L_2$ penalties in the EN penalty. To achieve better prediction performance while preserving the variable selection properties of the EN penalty, we use a corrected version of both PENSE and PENSEM given by $\sqrt{1 + (1 - \alpha^*)\lambda_S^*}\hat{\boldsymbol{\beta}}$. The intercept is corrected accordingly to maintain centered weighted residuals.

[2] `https://github.com/esmucler/mmlasso`

## 5.2 Simulation Settings

To demonstrate the benefits of the EN over the $L_1$ penalty, we include two simulation settings from Zou and Hastie [2] and Zou and Zhang [31]. In these settings the correlation among the predictors with non-zero regression coefficient is moderate to high. We further extend the settings in Zou and Hastie [2] to a more challenging setting with higher dimensions, having more active predictors than observations. We also assess the performance of the proposed estimators in a very sparse simulation setting with no correlation among the active predictors, which may benefit $L_1$-penalized estimators. The details are as follows.

(a) The first simulation setting is the same as example (d) in Zou and Hastie [2], in which $n = 50$, $p = 40$, $\sigma = 15$, and the vector of true regression coefficients is given by

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

The first 15 predictors are generated from a latent variable model with three latent variables

$$x_j = z_{\lceil j/5 \rceil} + \delta_j \text{ where } z_l \sim N(0,1) \text{ and } \delta_j \sim N(0, 0.01^2) \quad j = 1, \ldots, 15;\ l = 1, 2, 3$$

and the remaining 25 predictors are i.i.d. from a standard Normal distribution, $x_j \sim N(0,1)$, $j = 16, \ldots, 40$.

(b) In the second simulation setting we increase the number of predictors compared to setting (a) and maintain the number of observations $n = 50$. The error term is generated as in setting (a). Here, we use $p = 400$ predictors and the vector of true regression coefficients is given by

$$\boldsymbol{\beta} = (\underbrace{3, \ldots, 3}_{60}, \underbrace{0, \ldots, 0}_{340})^{\mathsf{T}}$$

The latent variable model is still based on three factors, but each factor is associated with 20 predictors, i.e.,

$$x_j = z_{\lceil j/20 \rceil} + \delta_j \text{ where } z_l \sim N(0,1) \text{ and } \delta_j \sim N(0, 0.01^2) \quad j = 1, \ldots, 60;\ l = 1, 2, 3.$$

The other 340 predictors are i.i.d. from a standard Normal distribution, $x_j \sim N(0,1)$, $j = 61, \ldots, 400$.

(c) The third simulation setting is from example 2 in Zou and Zhang [31], in which $n = 100$, $p = 81$, $\sigma = 6$, and the vector of true regression coefficients is given by

$$\boldsymbol{\beta} = (\underbrace{3, \ldots, 3}_{27}, \underbrace{0, \ldots, 0}_{54})^\intercal.$$

The predictors are generated from a multivariate Normal distribution $\mathbf{x} \sim N_p(\mathbf{0}, \ )$ with covariance structure

$$_{jk} = 0.75^{|j-k|} \quad j, k = 1, \ldots, 81.$$

(d) The final setting has a large number of predictors with $p = 995$, moderate sample size of $n = 100$, and a lower standard deviation of the error term $\sigma = 1$. Of these 995 predictors, 15 are active and their raw coefficients, $\gamma_l$, $l = 1, \ldots, 15$, are sampled randomly from a Uniform distribution on the 15-dimensional unit sphere. The indices of the active coefficient are equally spaced at $j = 1, 72, \ldots, 995$:

$$\boldsymbol{\beta} = \sqrt{4}(\gamma_1, \underbrace{0, \ldots, 0}_{71}, \gamma_2, 0, \ldots, 0, \gamma_{14}, \underbrace{0, \ldots, 0}_{71}, \gamma_{15})^\intercal.$$

The predictors are generated from a multivariate Normal distribution $\mathbf{x} \sim N_p(\mathbf{0}, \ )$ with covariance structure

$$_{jk} = 0.5^{|j-k|} \quad j, k = 1, \ldots, 1000$$

and the scaling of the coefficient vector gives a signal-to-noise ratio of 4.

Resistance of the estimators to contaminated observations is assessed by introducing contamination in the first $m = \lfloor \epsilon n \rfloor$ observations $(\mathbf{x}_i, y_i)$ according to the model used in Maronna [11]. Leverage points are introduced by changing the predictors $\mathbf{x}_i$ of the contaminated observations to

$$\mathbf{x}_i \leftarrow \tilde{\mathbf{x}}_i = \boldsymbol{\eta}_i + \frac{k_{lev}}{\sqrt{\mathbf{a}^\intercal \ ^{-1}\mathbf{a}}}\mathbf{a}, \quad i = 1, \ldots, m,$$

where $\boldsymbol{\eta}_i \sim N_p(\mathbf{0}, 0.1^2\mathbf{I}_p)$ and $\mathbf{a} = \tilde{\mathbf{a}} - \frac{1}{p}\tilde{\mathbf{a}}^\intercal \mathbf{1}_p$ with elements of $\tilde{\mathbf{a}}$ uniformly distributed between -1 and 1, $\tilde{a}_j \sim U(-1, 1)$, $j = 1, \ldots, p$. The distance in the direction most influential on the estimator is thus controlled by parameter $k_{lev}$.

In addition to changing the predictors to generate leverage points, we also contaminate the observations in the response by altering the regression coefficient

$$y_i = \tilde{\mathbf{x}}_i\tilde{\boldsymbol{\beta}} \quad \text{with } \tilde{\beta}_j = \begin{cases} \beta_j(1 + k_{slo}) & \text{if } \beta_j \neq 0 \\ k_{slo}\|\boldsymbol{\beta}\|_\infty & \text{o.w.} \end{cases}, \quad i = 1, \ldots, m.$$

If the parameter $k_{slo}$ is 0, no vertical outliers are introduced.

The parameters $k_{lev}$ and $k_{slo}$ control the position of the contaminated observations. To fully evaluate the robustness of the estimator different values for these parameters are considered. Preliminary analysis showed that the effect on all considered estimators was almost the same for any $k_{lev} > 1$, hence we fixed the distance of leverage points at $k_{lev} = 2$. The position of the vertical outliers has a more varying influence on the estimators. Therefore, in each simulation setting we consider a grid of 15 logarithmically spaced values for $k_{slo}$ between 1 and 500.

To measure prediction performance of the estimators we generate a validation set of observations $(\mathbf{x}_i^*, y_i^*)$, $i = 1, \ldots, n^*$, with $n^* = 1000$ and without contamination according to the respective simulation settings. Using this independent validation set, we compute the root mean squared prediction error (RMSPE) for an estimate $(\hat{\mu}, \hat{\boldsymbol{\beta}})$, i.e.,

$$\text{RMSPE} = \sqrt{\frac{1}{n^*}\sum_{i=1}^{n^*}(y_i^* - \mathbf{x}_i^{*\intercal}\hat{\boldsymbol{\beta}} - \hat{\mu})^2}.$$

Model selection performance is assessed by the sensitivity (SENS) and specificity (SPEC) of the estimated coefficient vector $\hat{\boldsymbol{\beta}}$, i.e.,

$$\text{SENS} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\#\{j : \beta_j \neq 0 \wedge \hat{\beta}_j \neq 0\}}{\#\{j : \beta_j \neq 0\}}$$
$$\text{SPEC} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{\#\{j : \beta_j = 0 \wedge \hat{\beta}_j = 0\}}{\#\{j : \beta_j = 0\}}.$$

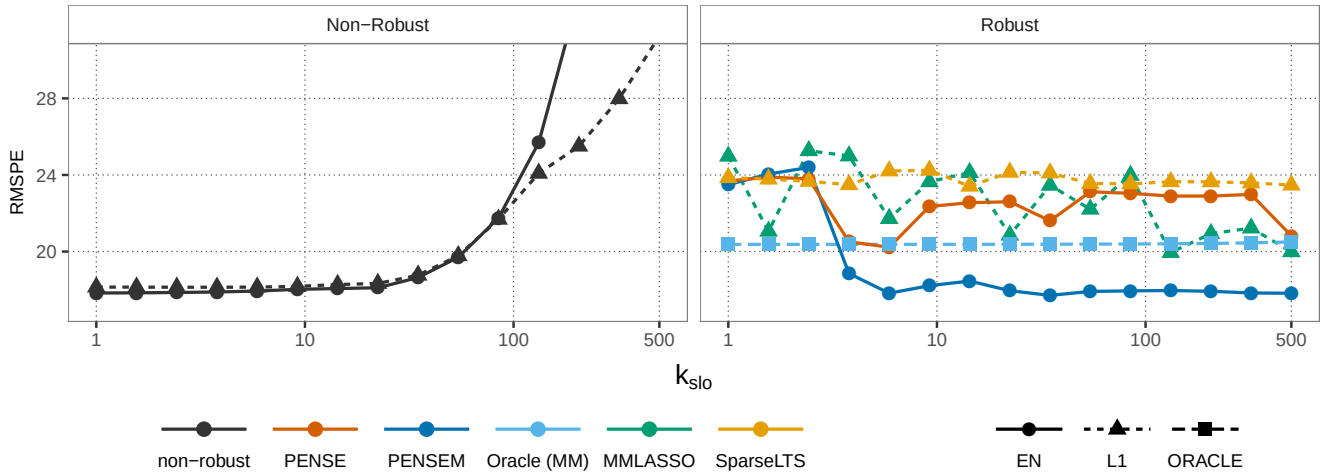


**Figure 1:** Root mean squared prediction error of different estimators over a grid of $k_{slo}$ values ranging from 1 to 500 with 10% contamination under simulation setting (a).

where TP, FP, TN, and FN stand for true and false positive and true and false negative, respectively.

For the uncontaminated cases, these measures provide a good picture of the overall performance of the estimators. When contamination is introduced in the training set, we summarize the performance over the entire grid of vertical outlier positions, $k_{slo}^{(l)}$, $l = 1, \ldots, 15$, by the area under the curve of RMSPE values, $\text{RMSPE}_{\text{cont}}$. Let's denote the estimate at $k_{slo}^{(l)}$ by $(\hat{\mu}^{(l)}, \hat{\boldsymbol{\beta}}^{(l)})$, then the overall RMSPE under contamination is

$$\text{RMSPE}_{\text{cont}} = \frac{1}{k_{slo}^{(15)} - k_{slo}^{(1)}} \sum_{l=2,\ldots,15} \frac{k_{slo}^{(l)} - k_{slo}^{(l-1)}}{2} \left( \text{RMSPE}(\hat{\mu}^{(l-1)}, \hat{\boldsymbol{\beta}}^{(l-1)}) + \text{RMSPE}(\hat{\mu}^{(l)}, \hat{\boldsymbol{\beta}}^{(l)}) \right).$$

For example, Figure 1 shows the curve of RMSPE over $k_{slo}$ from one replication of setting (a) and 10% contamination. It can be seen that the worst case performance might be at a different $k_{slo}$ value for each estimator and the area under the curve reflects the overall performance of the estimator under the different contamination settings examined. We use the same method to summarize the sensitivity and specificity under contamination, denoted by $\text{SENS}_{\text{cont}}$ and $\text{SPEC}_{\text{cont}}$, respectively. Each contamination setting is replicated 200 times, creating 200 of these curves, and corresponding areas, for each simulation setting.

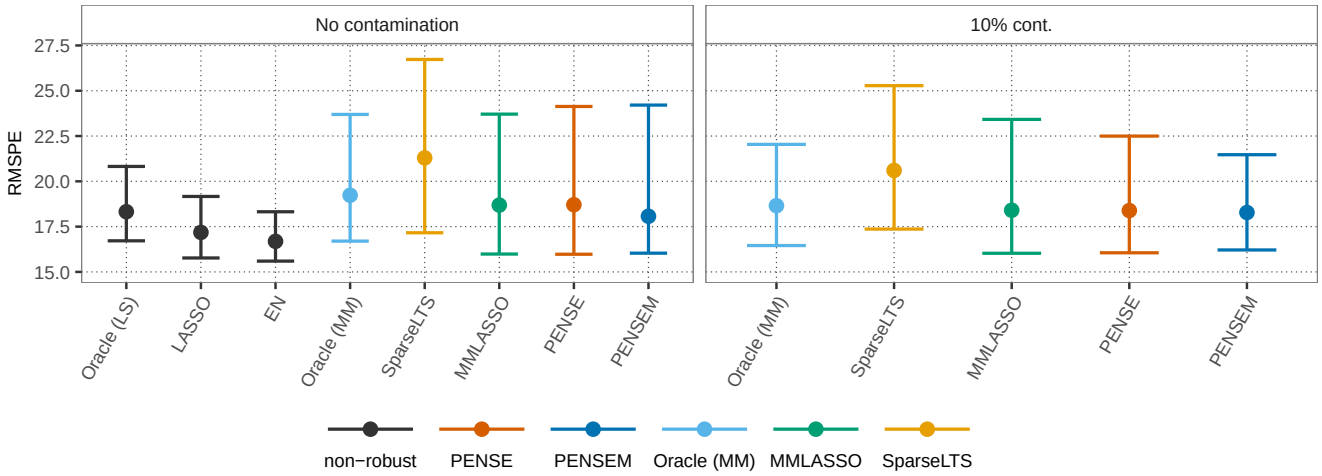


**Figure 2:** Average prediction performance of different estimators in simulation setting (a). The error bars extend from the 5% to the 95% quantile. For the uncontaminated case we report the RMSPE. For training data with 10% contamination we show the overall measure $\text{RMSPE}_{\text{cont}}$ over a grid of $k_{slo}$ from 1 to 500. Classical EN uses $\alpha^* = 0.7$, while PENSE(M) is using $\alpha^* = 0.5$.

For each simulation setting, we compute PENSE(M) as well as the classical EN for several values of $\alpha$. In the results however, we only present the PENSE(M) estimators corresponding to the $\alpha^*$ with the smallest average cross-validated $\text{RMSPE}_{\text{cont}}$. Similarly, we only show the classical EN with smallest average cross-validated RMSPE on the uncontaminated training data.

## 5.3 Simulation Results

*Setting (a)*: The prediction performance measures of PENSE(M) and those of the competing estimators over 200 replications for simulation setting (a) are shown in Figure 2. The solid dots in the plot represent the average values and the error bars mark the 5% and 95% quantiles of the RMSPE (no contamination, left plot) and the $\text{RMSPE}_{\text{cont}}$ (10% contamination in the training set, right plot). In this simulation setting we show the classical EN for $\alpha^* = 0.7$ and PENSE(M) for $\alpha^* = 0.5$, which were both chosen based on the CV performance of each estimator.

Setting (a) is tailored to favor the EN over the $L_1$ penalty due to the extreme grouping of the predictors. Without contamination, the classical EN estimator yields, on average, better prediction performance than LASSO and the oracle OLS estimator. The problem with the $L_1$

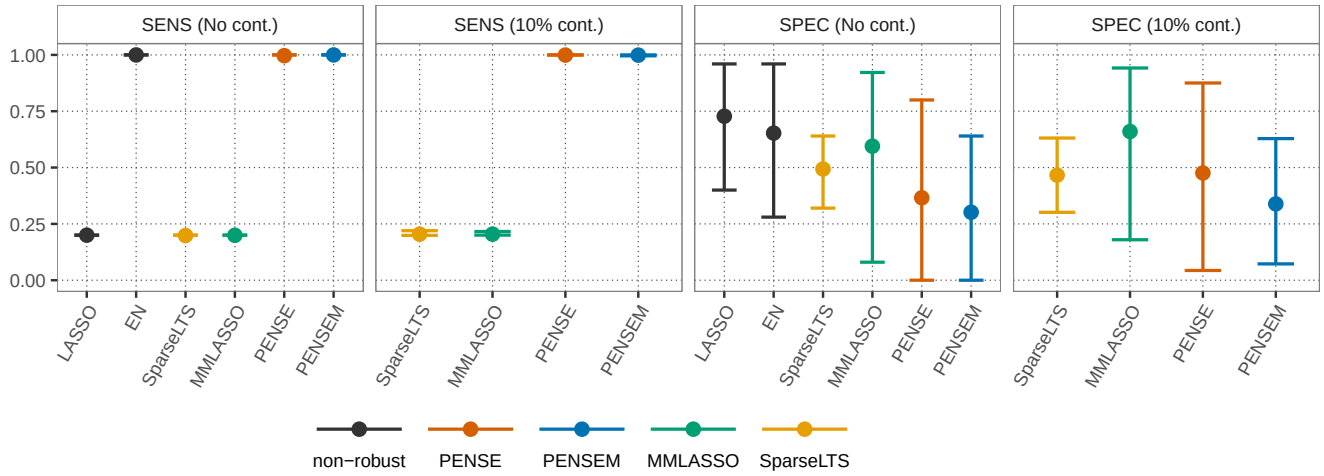


**Figure 3:** Average specificity and sensitivity of different estimators in simulation setting (a). The error bars extend from the 5% to the 95% quantile. For training data with 10% contamination we show the area under the curve ($\text{SENS}_{\text{cont}}$ and $\text{SPEC}_{\text{cont}}$) over a grid of $k_{slo}$ from 1 to 500. Classical EN uses $\alpha^* = 0.7$, while PENSE(M) is using $\alpha^* = 0.5$.

penalty of LASSO is that only a single predictor is selected within each group. However, if the penalty parameter $\lambda$ is small enough, this single predictor can almost fully capture the effect of the entire group. Thus, the benefit of the EN penalty is only marginally visible in the prediction performance.

Of the robust counterparts, PENSEM achieves the smallest RMSPE under no contamination as well as overall under contamination, even outperforming the robust oracle estimators. However, as observed for the classical estimators, the difference between the robust regularized EN estimators (PENSE and PENSEM) and the MMLASSO, is small.

The strength of the EN penalty in this setting becomes more noticeable in the model selection performance in Figure 3. Regardless if the data is contaminated, all of the LASSO-based estimators only pick a single coefficient per group, while the EN estimators consistently select the whole groups. Thus, the sensitivity of the LASSO methods is weak compared to that of EN methods. For the classical EN and PENSEM estimators, the selection of relevant variables brings also some of the irrelevant ones shown in a slight drop in specificity. However, PENSE achieves both high sensitivity and specificity in the uncontaminated and contaminated cases.

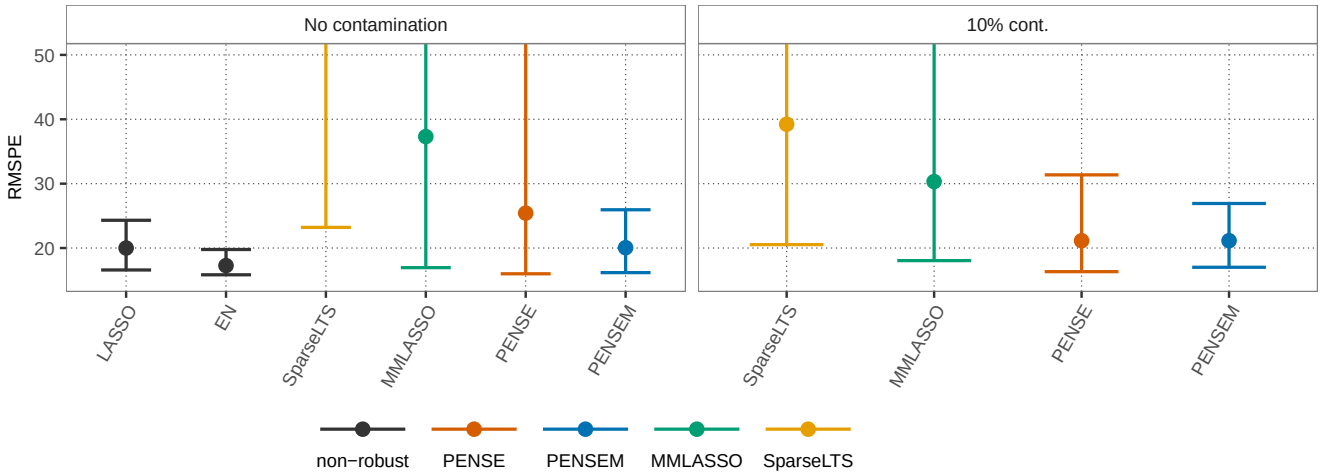


**Figure 4:** Average prediction performance of different estimators in simulation setting (b). The error bars extend from the 5% to the 95% quantile. For the uncontaminated case we report the RMSPE. For training data with 10% contamination we show the overall measure $\text{RMSPE}_{\text{cont}}$ over a grid of $k_{slo}$ from 1 to 500. Classical EN and PENSE(M) are both using $\alpha^* = 0.9$. The oracle estimates cannot be computed in this setting because there are more active predictors than observations.

*Setting (b)*: In this setting, the difference between LASSO and EN estimators is even more pronounced as shown in Figure 4. In addition, the oracle estimates cannot be computed since the number of active predictors is larger than the number of observations. The classical EN as well as PENSE both achieve the best cross-validated prediction performance for $\alpha^* = 0.9$, reflecting the sparsity of this setting. PENSEM has an average RMSPE close to that of the classical EN when no contamination is present. Under contamination, PENSEM performs almost the same as without contamination and PENSE has now a similar prediction performance. As for model selection (Figure 5), we observe again large differences between the sensitivities of LASSO-type and EN-type estimators. The former only selects a single predictor from each group. In contrast to the previous setting, however, PENSE(M) and classical EN have a higher specificity in this setting than in setting (a) due to the large number of noise predictors. Under contamination, PENSE selects all 60 active predictors 99.5% of the time and on average selects only 37 of the 340 noise predictors. In the uncontaminated case the model selection of PENSE is on average even outperforming the classical EN.

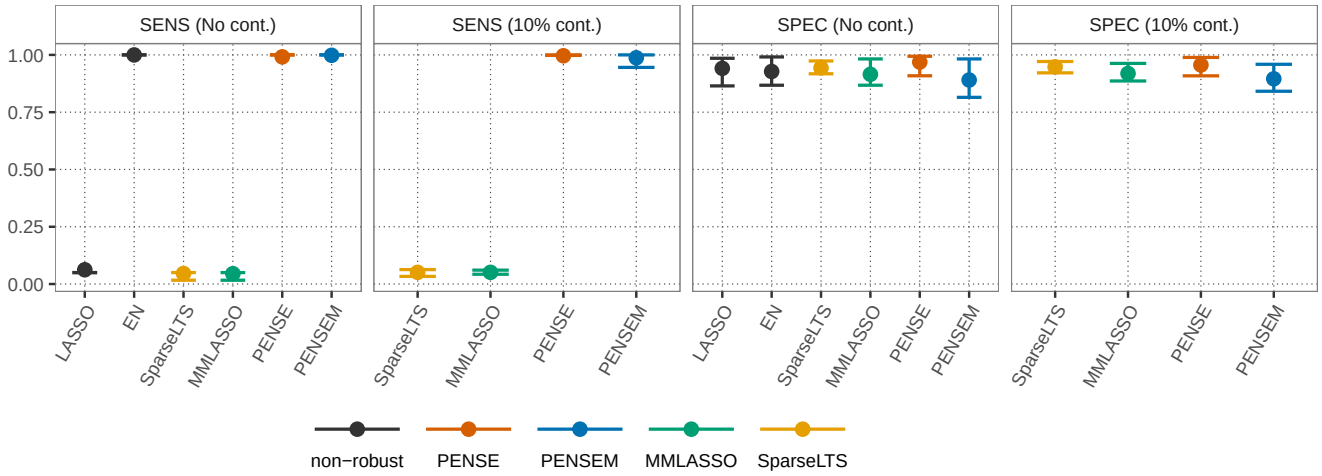


**Figure 5:** Average specificity and sensitivity of different estimators in simulation setting (b). The error bars extend from the 5% to the 95% quantile. For training data with 10% contamination we show the area under the curve ($\text{SENS}_{\text{cont}}$ and $\text{SPEC}_{\text{cont}}$) over a grid of $k_{slo}$ from 1 to 500. Classical EN and PENSE(M) are both using $\alpha^* = 0.9$.

*Setting (c)*: This is the last setting where the elastic net penalty should have an advantage over the LASSO penalty. In terms of prediction performance (Figure 6), PENSE and PENSEM (with $\alpha^* = 0.7$) perform, on average, almost as well as the robust oracle estimate and notably better than the other robust estimators which are using a LASSO penalty. It is clearly visible that the LASSO-based estimators have difficulty addressing the moderate to high correlation among active predictors in this setting. For model selection, as shown in Figure 7, classic EN and PENSE(M) again outperform LASSO-based methods, which not surprisingly comes at the cost of a drop in their specificity. PENSE selects around 15 of the 54 noise predictors on average under contamination, while PENSEM selects roughly 21. SparseLTS seems to generally select smaller models with decent accuracy, while MMLASSO chooses as many noise predictors as PENSE, but is less sensitive.

*Setting (d)*: Results of this very sparse setting are shown in Figure 8. Not surprisingly, the best CV performance for PENSE(M) is achieved with an $L_1$ penalty ($\alpha^* = 1$). This example illustrates the flexibility of the EN penalty, which ranges from a LASSO to a Ridge penalty, thus being adjustable to different degrees of sparsity. As expected, PENSEM results are very similar to MMLASSO, with observed differences coming from the initial estimators used to initialized the

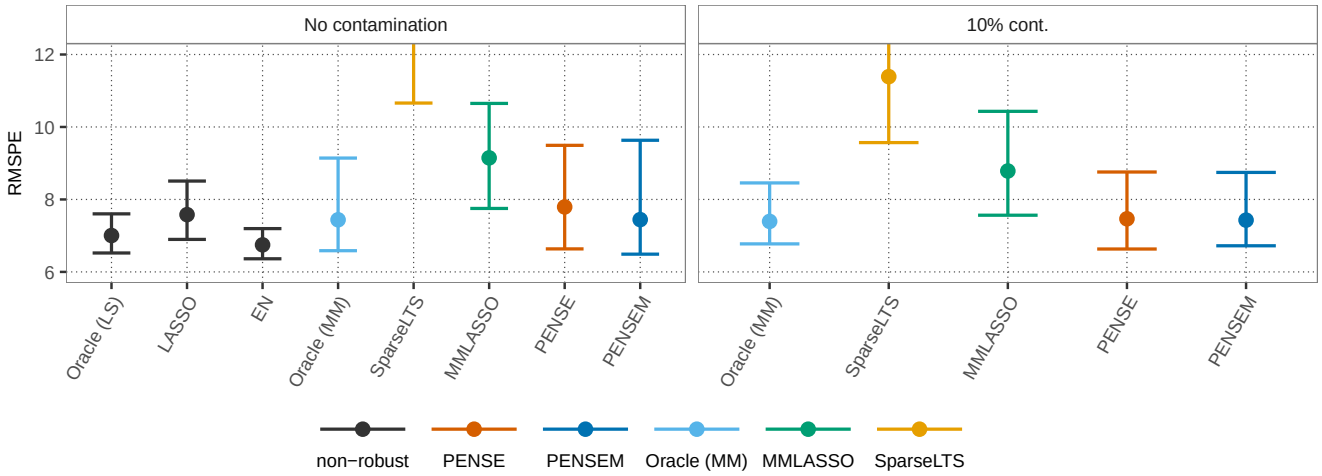


**Figure 6:** Average prediction performance of different estimators in simulation setting (c). The error bars extend from the 5% to the 95% quantile. For the uncontaminated case we report the RMSPE. For training data with 10% contamination we show the overall measure $\text{RMSPE}_{\text{cont}}$ over a grid of $k_{slo}$ from 1 to 500. Classical EN and PENSE(M) are both using $\alpha^* = 0.7$.

M-steps and the algorithms used to optimize the associated objective functions. MMLASSO has a slightly smaller average RMSPE than PENSEM in the uncontaminated case. However, under contamination, PENSEM shows a little better average performance and less variation. When examining model selection as presented in Figure 9 we can observe that all methods struggle to identify all 15 active covariates. This can be mainly attributed to the fact that coefficients are sampled on the unit sphere which results in some coefficients being very small compared to others. PENSEM generally exhibits less variation in sensitivity and has a very similar average as MMLASSO in both measures under contamination.

In summary, the simulations show that PENSE and PENSEM are performing competitively compared to other robust regularized estimators of regression. The flexible elastic net penalty makes PENSE(M) applicable to a broad range of settings and clearly outperforms LASSO-based estimates if important predictors are correlated. Especially in settings (b) and (c) the elastic net penalty is beneficial for both prediction performance and identification of important predictors.

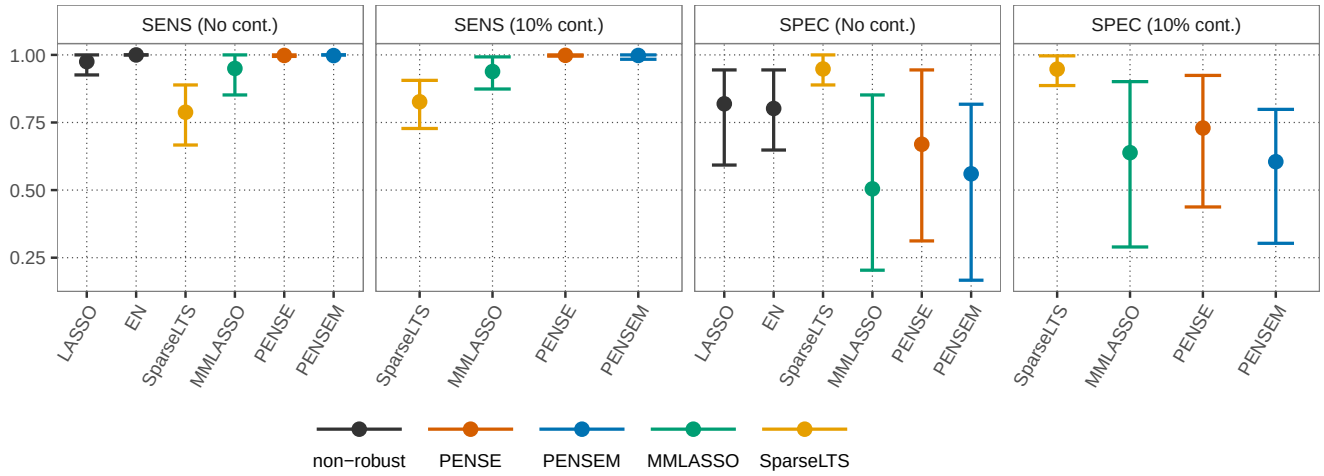


**Figure 7:** Average specificity and sensitivity of different estimators in simulation setting (c). The error bars extend from the 5% to the 95% quantile. For training data with 10% contamination we show the area under the curve ($\text{SENS}_{\text{cont}}$ and $\text{SPEC}_{\text{cont}}$) over a grid of $k_{slo}$ from 1 to 500. Classical EN and PENSE(M) are both using $\alpha^* = 0.7$.

# 6 Biomarkers of Cardiac Allograft Vasculopathy

In this Section, we use PENSE to select potential plasma biomarkers of cardiac allograft vasculopathy (CAV), a major complication suffered by 50% cardiac transplant recipients beyond the first year after transplant. The most typical screening and diagnosis of CAV requires the examination of the coronary arteries that supply oxygenated blood to the heart. Despite its invasiveness, cost, and associated risks of complications, to date, coronary angiography remains the most widely used tool to assess the narrowing and stenosis of the coronary arteries [32]. The identification of plasma biomarkers of CAV can result in the development of a simple blood test to diagnose and monitor this condition significantly improving current patient care options.

The Biomarkers in Transplantation initiative has collected plasma samples from a cohort of patients who received a heart transplant at St. Paul's Hospital, Vancouver, British Columbia, and consented to be enrolled in the study. Around one year after transplantation, some of these patients presented signs of coronary artery narrowing, measured by the stenosis of the left anterior descending (LAD) artery, as an indicator of CAV development. To identify potential biomarkers of this condition, protein levels from 37 plasma samples, collected at 1 year after transplantation,

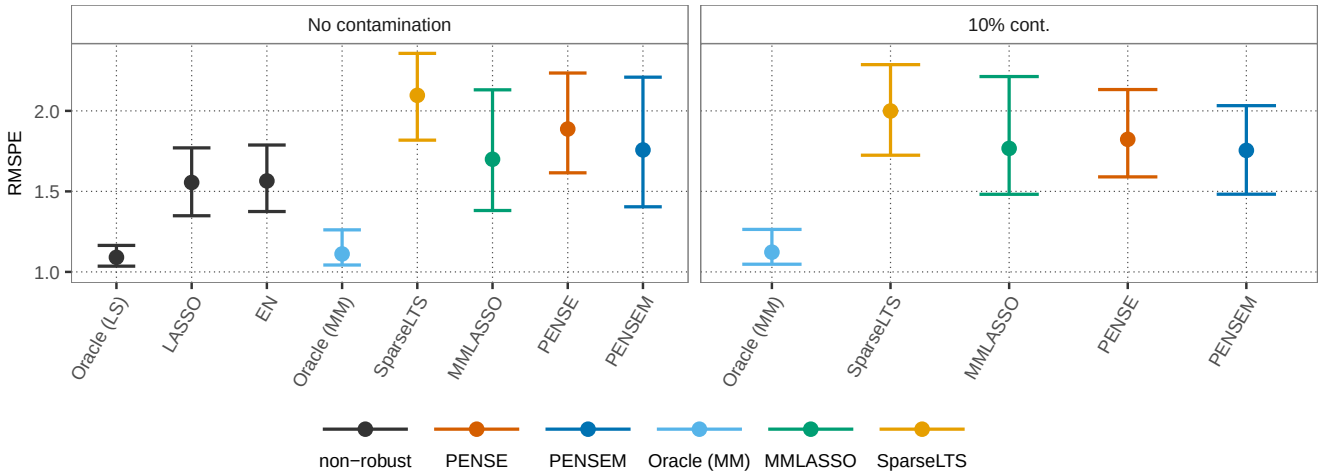


**Figure 8:** Average prediction performance of different estimators in simulation setting (d). The error bars extend from the 5% to the 95% quantile. For the uncontaminated case we report the RMSPE. For training data with 10% contamination we show the overall measure $\text{RMSPE}_{\text{cont}}$ over a grid of $k_{slo}$ from 1 to 500. Classical EN uses $\alpha^* = 0.9$ while PENSE(M) is fitted with $\alpha^* = 1$.

were measured using isobaric tags for relative and absolute quantitation (iTRAQ) technology. This mass spectrometry technique enabled the simultaneous identification and quantification of multiple proteins present in the samples. A full description of this proteomics study is given by Lin et al. [33], which developed a proteomic classifier of CAV using a preliminary univariate robust screening of proteins and a classical EN classification method. PENSE and PENSEM combine robustness, variable selection and modelling in a single step, taking full advantage of the multivariate nature of the data that can result in the identification of new potential markers of CAV and a better prediction.

Although hundreds of proteins were detected and measured by iTRAQ in most patient samples, only a few proteins are expected to be associated with the observed artery obstruction, resulting in a sparse regression model (i.e., most regression coefficients equal to zero). Thus, we use PENSE to select a candidate set of relevant proteins among the 81 proteins that were detected in all samples and PENSEM to refine this set. In this application we induce a moderate level of sparsity using $\alpha^* = 0.75$, aiming to control the number of potential false biomarkers identified and potential good biomarkers missed in this study. As explained in Section 5.1, the selection of the level of penalization is based on a robust measure of the size of the prediction errors estimated by 10-fold

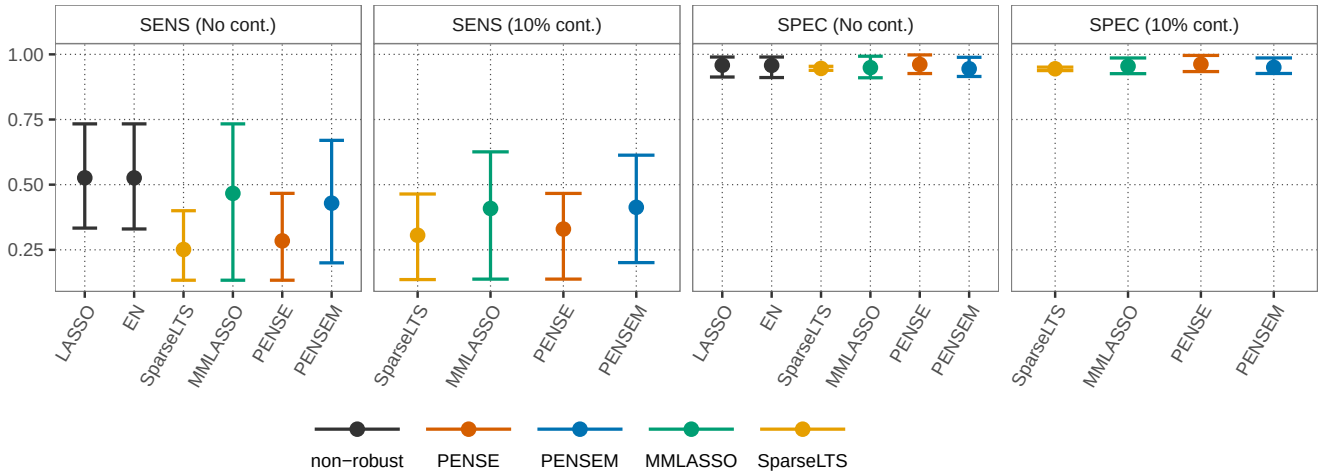


**Figure 9:** Average specificity and sensitivity of different estimators in simulation setting (d). The error bars extend from the 5% to the 95% quantile. For training data with 10% contamination we show the area under the curve ($\text{SENS}_{\text{cont}}$ and $\text{SPEC}_{\text{cont}}$) over a grid of $k_{slo}$ from 1 to 500. Classical EN uses $\alpha^* = 0.9$ while PENSE(M) is fitted with $\alpha^* = 1$.

CV. To make this selection more stable, we repeat this estimation 100 times over the full grid of penalty values and select $\lambda_S^*$ as the maximum $\lambda$ such that the median estimated prediction error at this value is within 1.5 MAD of the minimum median error across the grid. At this selected level of penalization, PENSE identifies 16 potential markers to predict the diameter of the LAD artery and thus assess the level of obstruction in that artery.

To refine the selection given by PENSE, PENSEM is computed over a grid of lambda values, using the selected PENSE as an initial estimator, and selecting the optimal level of penalization ($\lambda_M^*$) with the same criteria used to select $\lambda_S^*$. PENSEM selects 27 out the 11 potential markers selected by PENSE to predict the diameter of the LAD artery. The classical EN estimator (using the same $\alpha$ parameter) identified only 6 markers of which 5 are also identified by our estimators. Figure 10 illustrates PENSEM's estimations of the regression coefficients for different values of $\lambda_M$ (i.e., PENSEM's regularization path), highlighting in blue the coefficients selected at the optimal level of penalization chosen (i.e., $\lambda_M^*$ represented by the vertical dashed line). The names of the selected markers are given in Table 1. Interestingly, many of these markers were previously related to CAV, including the complements C4B/C4A, and C7, APOE, AMBP, and SHBG [33]. However, further analysis of this dataset using our estimator allowed the identification of new potential

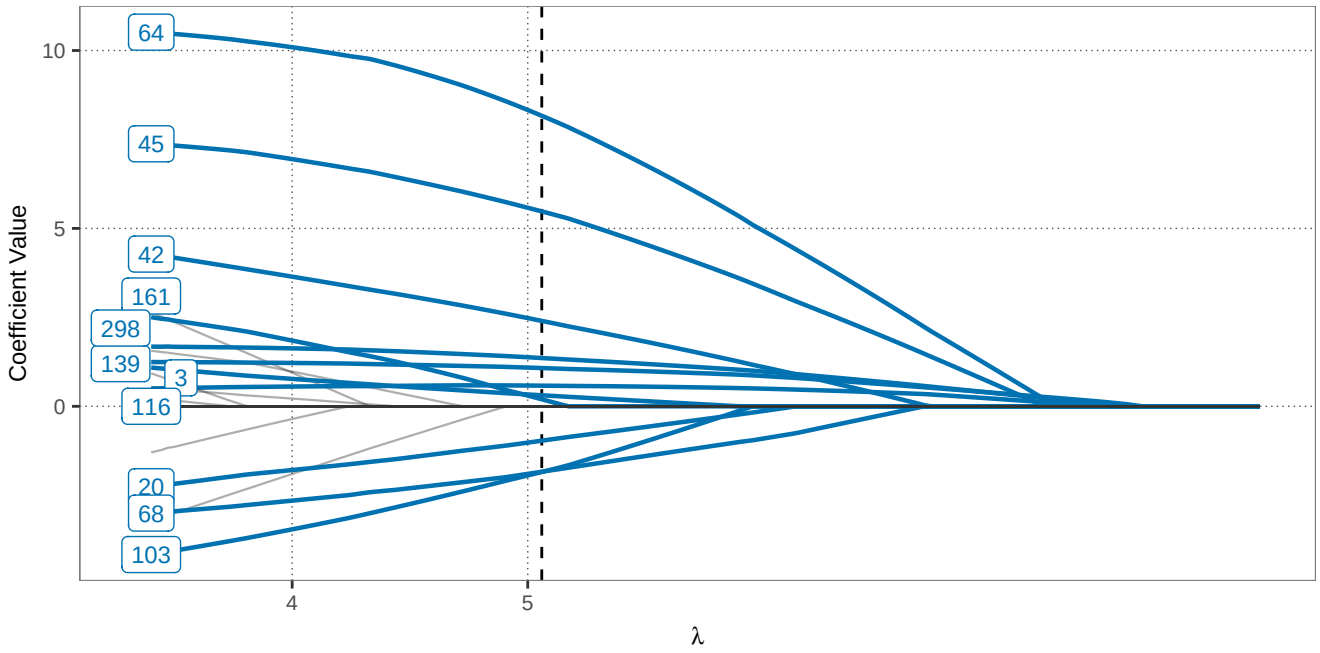


**Figure 10:** PENSEM's regularization path. The regularization path illustrates how the estimated coefficients shrink at different levels of penalization. The optimal level of penalization $\lambda_M^*$ is represented by the vertical dashed line. The path of the variables selected at this level of penalization are highlighted in blue. The numbers in the labels correspond to the protein IDs.

markers, including CFI, APOC2, and three homoglobin subunits HBD, HBA and HBZ. Overall, results illustrate the involvement of complex mechanisms of CAV, such as complement system activation and regulation, immune-recognition, inflammation, and apoptosis related mechanisms among others.

An additional advantage of using a robust estimator to estimate the regression coefficients is that outlying observations can be flagged looking at the residuals of each point versus their fitted values (see Figure 11). Based on the results of the angiography, no obstruction was detected in the LAD artery of the two patients in the lower part of the figure (B-527-W51 and B-561-W52). However, a second measurement of the LAD of these two patients using a more accurate technique (intravascular ultrasound, IVUS) indicates that their arteries present a mild stenosis with about 16% area reduction, as suggested by PENSEM's predictions (negative residuals).

In the absence of an independent test set, the performance of the estimators was evaluated by 200 replications of 10-folds cross-validations and compared to that of the classical EN estimator

**Table 1:** Potential biomarkers of CAV identified by PENSEM.

| Protein ID | Gene Symbol | Protein Name |
|---|---|---|
| 3 | C4B/C4A | Complement C4-B/C4-A |
| 20 | C7 | Complement component C7 |
| 42 | APOE | Apolipoprotein E |
| 45 | AMBP | Protein AMBP |
| 64 | CFI | Complement factor I |
| 68 | SHBG | Sex hormone-binding globulin |
| 116 | APOC2 | Apolipoprotein C-II |
| 139 | HBD | Hemoglobin subunit delta |
| 298 | HBA2;HBA1;HBZ | Hemoglobin subunit alpha/zeta |

(see Table 2). In terms of prediction, all estimators perform similarly, with PENSE showing, on average, a slightly better performance.

**Table 2:** Mean and standard deviation (SD) of the prediction $\tau$-scales.

| | Lasso | EN | PENSE(3/4) | PENSEM(3/4) | MMLasso | Sparse-LTS |
|---|---|---|---|---|---|---|
| **Mean** | 17.50 | 17.41 | 17.07 | 17.55 | 18.35 | 18.97 |
| **SD** | 1.63 | 1.36 | 1.57 | 1.54 | 1.73 | 1.81 |

# 7 Conclusions

In this paper we proposed regularized robust estimators with an Elastic Net penalty, which we call PENSE and PENSEM. The first one is a penalized version of an S-estimator, while the second one corresponds to a penalized high-breakdown M-estimator, which generally results in an increase of efficiency for the parameter estimates.

We showed that these estimators retain the robustness properties of their un-penalized counterparts (high breakdown point and consistency), which makes them very useful when one may have outlying or other atypical observations in the data. At the same time, our numerical experiments

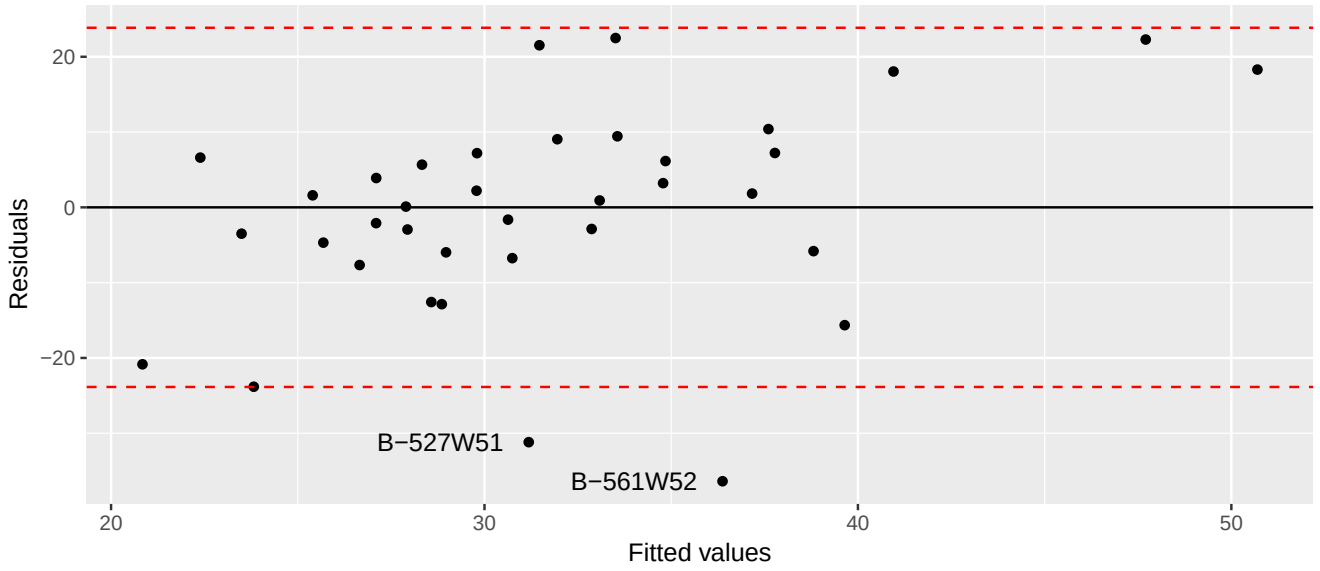


**Figure 11:** Outlying patients identified by PENSEM. The red dashed lines represent $\pm 2$ times the robust $\tau$-scale of the residuals.

show that PENSE and PENSEM also inherit the prediction and model selection properties of the Elastic Net penalty. In particular, highly correlated explanatory variables enter or leave the model in groups, unlike what is observed with the $L_1$ penalty of the LASSO.

In this paper we proposed an algorithm for computing PENSE and PENSEM that works very well in practice. Computing regression estimators with good robustness properties is computationally very costly because the loss functions that need to be optimized to compute high-breakdown point robust estimators are necessarily non-convex. Moreover, the presence of a non-differentiable penalty term for the penalized estimators increases their computational difficulty. Our algorithm relies on an iterative procedure derived from the first-order conditions of the optimization problem that defines the penalized estimators. These iterations are initialized from a relatively small number of robust starting values that are constructed following the ideas of Peña and Yohai [19].

A very important part of any practical use of penalized estimators is choosing an "optimal" value for the penalty term. Although cross-validation is a very popular method to do this, in our case we need to be concerned with the possibility of having outliers or other atypical observations in our data, which may affect the estimated prediction error. Following other proposals in the literature we used a robust scale estimator of the prediction errors obtained via cross-validation

instead of the mean squared prediction error. An implementation in R of our algorithm (including the robust cross-validation step) is publicly available from CRAN in an R-package called 'pense' (https://cran.r-project.org/package=pense).

Finally, we used PENSE and PENSEM to study the association between hundreds of plasma protein levels and a measure of artery obstruction on cardiac transplant recipients. The robust estimators identified new potentially relevant biomarkers that were not found with non-robust alternatives. Moreover, the analysis based on our robust estimators flagged two patients with suspiciously atypical artery obstruction values. Later measurements with more accurate techniques of the artery obstruction of these patients confirmed that the original values were inaccurate.

# Acknowledgement

We thank the NCE CECR PRevention of Organ Failure (PROOF) Centre of Excellence to share data of the heart transplant cohort, collected and processed by the Genome Canada-funded Biomarkers in Transplantation initiative. GCF and MSB were supported by NSERC Discovery grants.

# 8 Technical Appendix

## 8.1 Breakdown Point

To emphasize the sample matrix $\mathbf{Z}$ used in the optimization problem, we modify the notation of the objective function for this section to:

$$\mathcal{L}_{\text{PS}}(\mu, \boldsymbol{\beta}, \mathbf{Z}) = \hat{\sigma}(\mu, \boldsymbol{\beta}, \mathbf{Z})^2 + \lambda_S \left( \frac{1}{2}(1-\alpha)||\boldsymbol{\beta}||_2^2 + \alpha||\boldsymbol{\beta}||_1 \right) \tag{9}$$

To show that the breakdown point of PENSE, $\epsilon^* \left( \hat{\boldsymbol{\theta}}^{PS}; \mathbf{Z} \right)$ is bounded below by $\frac{m(\delta)}{n}$ and above by $\delta$, we need the following lemma from Maronna et al. [3, p. 166]:

**Lemma 8.1.** *Consider any sequence of samples* $\left( \tilde{\mathbf{Z}}^{(N)} \right)_{N \in \mathbb{N}}$ *and corresponding residuals* $r_{N,i} = r_i(\hat{\mu}^{(N)}, \hat{\boldsymbol{\beta}}^{(N)}) = \tilde{y}_{N,i} - \hat{\mu}^{(N)} - \tilde{\mathbf{x}}_{N,i}^\intercal \hat{\boldsymbol{\beta}}^{(N)}$ *for* $(\tilde{y}_{N,i}, \tilde{\mathbf{x}}_{N,i}^\intercal)$ *a row in* $\tilde{\mathbf{Z}}^{(N)}$.

*Then*

*(i) Let* $C = \{i : |r_{N,i}| \to \infty\}$. *If* $\#(C) > n\delta$, *then* $\hat{\sigma}(\hat{\beta}_0^{(N)}, \hat{\boldsymbol{\beta}}^{(N)}, \tilde{\mathbf{Z}}^{(N)}) \to \infty$ *for* $N \to \infty$.

*(ii) Let* $D = \{i : |r_{N,i}| \text{ is bounded}\}$. *If* $\#(D) > n - n\delta$, *then* $\hat{\sigma}(\hat{\beta}_0^{(N)}, \hat{\boldsymbol{\beta}}^{(N)}, \tilde{\mathbf{Z}}^{(N)})$ *is bounded.*

The proof of Theorem 4.1 can be done separately for the boundedness from below and above.

*Proof of Theorem 4.1, bounded from below.* We consider an arbitrary sequence of contaminated samples $\left( \tilde{\mathbf{Z}}_m^{(N)} \right)_{N \in \mathbb{N}}$ with $m \leq m(\delta)$. We will show that the corresponding sequence of $\left( \hat{\mu}^{(N)}, \hat{\boldsymbol{\beta}}^{(N)} \right)_{N \in \mathbb{N}}$ is bounded.

First, let $\left( \tilde{\mu}, \tilde{\boldsymbol{\beta}} \right)$ such that $|\tilde{\mu}| + \|\tilde{\boldsymbol{\beta}}\|_1 = K_1 < \infty$ which implies also finite L2 norm of $\tilde{\boldsymbol{\beta}}$, $\|\tilde{\boldsymbol{\beta}}\|_2^2 = K_2 < \infty$. For the uncontaminated observations $(y_i, \mathbf{x}_i^\intercal)$ which are also in the contaminated sample $\tilde{\mathbf{Z}}_m^{(N)}$, we know from the triangle inequality that $|r_{N,i}| < \infty$. Hence the number of bounded residuals $\#(D) \geq n - m \geq n - n\delta$ and therefore part (ii) of Lemma 8.1 says that $\hat{\sigma}(\tilde{\mu}, \tilde{\boldsymbol{\beta}}, \tilde{\mathbf{Z}}_m^{(N)})$ is bounded:

$$\sup_{N \in \mathbb{N}} \hat{\sigma}(\tilde{\mu}, \tilde{\boldsymbol{\beta}}, \tilde{\mathbf{Z}}_m^{(N)}) < \infty. \tag{10}$$

Now suppose that $\left( \|\hat{\boldsymbol{\beta}}^{(N)}\|_1 \right)_{N \in \mathbb{N}}$ is unbounded. Due to (10) we know there exists a $N_0 \in \mathbb{N}$ such that $\|\hat{\boldsymbol{\beta}}^{(N_0)}\|_1 > K_1 + \frac{1}{\alpha \lambda_S} \sup_{N \in \mathbb{N}} \hat{\sigma}(\tilde{\mu}, \tilde{\boldsymbol{\beta}}, \tilde{\mathbf{Z}}_m^{(N)})^2$ and $\|\hat{\boldsymbol{\beta}}^{(N_0)}\|_2^2 > K_2$. Thus, for every $N' \geq N_0$,

$$\begin{aligned}\mathcal{L}_{\text{PS}}(\hat{\mu}^{(N')}, \hat{\boldsymbol{\beta}}^{(N')}, \tilde{\mathbf{Z}}_m^{(N')}) &> \hat{\sigma}(\hat{\mu}^{(N')}, \hat{\boldsymbol{\beta}}^{(N')}, \tilde{\mathbf{Z}}_m^{(N')})^2 + \lambda_S \left(\frac{1}{2}(1-\alpha)K_2 + \alpha K_1\right) + \sup_{N\in\mathbb{N}} \hat{\sigma}(\tilde{\mu}, \tilde{\boldsymbol{\beta}}, \tilde{\mathbf{Z}}_m^{(N)})^2 \\ &\geq \mathcal{L}_{\text{PS}}(\tilde{\mu}, \tilde{\boldsymbol{\beta}}, \tilde{\mathbf{Z}}_m^{(N')})\end{aligned} \tag{11}$$

contradicting the assumption that $\left(\hat{\mu}^{(N')}, \hat{\boldsymbol{\beta}}^{(N')}\right)$ minimizes the objective function. This proofs that $\hat{\boldsymbol{\beta}}^{(N)}$ is bounded for $m \leq m(\delta)$. It remains to show that the intercept $\hat{\beta}_0^{(N)}$ is bounded as well.

Since $\left(\|\hat{\boldsymbol{\beta}}^{(N)}\|_1\right)_{N\in\mathbb{N}}$ is bounded, $|y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}^{(N)}|$ is bounded for the $n-m$ uncontaminated observations $(y_i, \mathbf{x}_i)$ in the sample $\tilde{\mathbf{Z}}_m^{(N)}$. Assume now that $|\hat{\mu}^{(N)}| \to \infty$, then the residuals of the uncontaminated observations also tend to infinity and hence $\#(C) > n\delta$. According to part (i) of Lemma 8.1 this implies that $\hat{\sigma}(\hat{\mu}^{(N)}, \hat{\boldsymbol{\beta}}^{(N)}, \tilde{\mathbf{Z}}_m^{(N)}) \to \infty$. Thus, there exists an integer $N_1 \in \mathbb{N}$ such that $\hat{\sigma}(\hat{\mu}^{(N_1)}, \hat{\boldsymbol{\beta}}^{(N_1)}, \tilde{\mathbf{Z}}_m^{(N_1)})^2 > \sup_{N\in\mathbb{N}} \hat{\sigma}(\tilde{\mu}, \tilde{\boldsymbol{\beta}}, \tilde{\mathbf{Z}}_m^{(N)})^2 + \lambda_S \left(\frac{1}{2}(1-\alpha)K_2 + \alpha K_1\right)$. Similar to (11), this can be used to show that for all $N' \geq N_1$, $\mathcal{L}_{\text{PS}}(\hat{\mu}^{(N')}, \hat{\boldsymbol{\beta}}^{(N')}, \tilde{\mathbf{Z}}_m^{(N')}) \geq \mathcal{L}_{\text{PS}}(\tilde{\mu}, \tilde{\boldsymbol{\beta}}, \tilde{\mathbf{Z}}_m^{(N')})$. Therefore, $\hat{\mu}^{(N)}$ and $\hat{\boldsymbol{\beta}}^{(N)}$ are bounded for $m \leq m(\delta)$. □

*Proof of Theorem 4.1, bounded from above.* Take $m > n\delta$, then we can show that the estimator breaks down. Denote by $C \subset \{1, \ldots, n\}$ the indices in the contaminated samples $\tilde{\mathbf{Z}}_m^{(N)}$ that are changed from the original sample $\mathbf{Z}$ with $\#(C) = m$. If we choose an arbitrary $\mathbf{x}_0$ with $\|\mathbf{x}_0\|_2 = 1$ and consider a sequence of samples $\left(\tilde{\mathbf{Z}}_m^{(N)}\right)_{N\in\mathbb{N}}$ defined by

$$(y_{N,i}, \mathbf{x}_{N,i}) = \begin{cases} (N^{\nu+1}, \mathbf{x}_0 N) & i \in C \\ (y_i, \mathbf{x}_i) & i \notin C \end{cases},$$

where $0 < \nu \leq 1$. We will show that the sequence of estimators $\left((\hat{\mu}^{(N)}, \hat{\boldsymbol{\beta}}^{(N)})\right)_{N\in\mathbb{N}}$ for $\tilde{\mathbf{Z}}_m^{(N)}$ can not be bounded.

Assume first that $(\hat{\mu}^{(N)}, \hat{\boldsymbol{\beta}}^{(N)})$ is bounded in norm. Similar as in the proof above we have $|r_{N,i}| < \infty$ for $i \notin C$ and all $N \in \mathbb{N}$. Residuals for contaminated samples on the other hand are bounded below by

$$|r_{N,i}| \geq N \left| N^{\nu} - \|\mathbf{x}_0\|_1 \|\hat{\boldsymbol{\beta}}^{(N)}\|_1 \right| - |\hat{\mu}^{(N)}|.$$

Because the norms of $\hat{\mu}$ and $\hat{\boldsymbol{\beta}}^{(N)}$ are bounded, the right-hand side goes to infinity, and so do the residuals for $i \in C$. According to part (i) of Lemma 8.1, this implies the scale $\hat{\sigma}(\hat{\mu}^{(N)}, \hat{\boldsymbol{\beta}}^{(N)}, \tilde{\mathbf{Z}}_m^{(N)})$ tends to infinity as well. When we decompose the M-estimation equation to

$$\sum_{i \notin C} \rho\left(\frac{r_{N,i}}{\hat{\sigma}(\hat{\mu}^{(N)}, \hat{\boldsymbol{\beta}}^{(N)}, \tilde{\mathbf{Z}}_m^{(N)})}\right) + \sum_{i \in C} \rho\left(\frac{r_{N,i}}{\hat{\sigma}(\hat{\mu}^{(N)}, \hat{\boldsymbol{\beta}}^{(N)}, \tilde{\mathbf{Z}}_m^{(N)})}\right) = n\delta$$

and take the limit for $N \to \infty$, we see that the argument to $\rho$ in the first sum tends to zero because of the bounded residuals, which in turn means the first sum tends to 0. The summands in the second term are all identical, hence the limit is

$$\lim_{N\to\infty} \rho\left(\frac{1 - \hat{\mu}^{(N)}/N^{\nu+1} - \mathbf{x}_0^\intercal \hat{\boldsymbol{\beta}}^{(N)}/N^{\nu}}{\hat{\sigma}(\hat{\mu}^{(N)}, \hat{\boldsymbol{\beta}}^{(N)}, \tilde{\mathbf{Z}}_m^{(N)})/N^{\nu+1}}\right) = \frac{n\delta}{m}. \tag{12}$$

The function $\rho(x)$ is continuous and increasing for $x > 0$ such that $\rho(x) < 1 = \rho(\infty)$. Because $n\delta/m < 1 = \rho(\infty)$ there exists a unique value $\gamma$ such that

$$\rho\left(\frac{1}{\gamma}\right) = \frac{n\delta}{m}. \tag{13}$$

The numerator in the argument in (12) tends to 1 and due to (13) any converging subsequence of $\hat{\sigma}(\hat{\mu}^{(N)}, \hat{\boldsymbol{\beta}}^{(N)}, \tilde{\mathbf{Z}}_m^{(N)})/N^{\nu+1}$ must have limit $\gamma$. Therefore, the boundedness of $(\hat{\mu}^{(N)}, \hat{\boldsymbol{\beta}}^{(N)})$ implies

$$\lim_{N\to\infty} \frac{1}{N^{\nu+1}} \hat{\sigma}(\hat{\mu}^{(N)}, \hat{\boldsymbol{\beta}}^{(N)}, \tilde{\mathbf{Z}}_m^{(N)}) = \gamma.$$

and thus

$$\lim_{N\to\infty} \frac{1}{N^{2\nu+2}} \mathcal{L}_{\mathrm{PS}}(\hat{\mu}^{(N)}, \hat{\boldsymbol{\beta}}^{(N)}, \tilde{\mathbf{Z}}_m^{(N)}) = \gamma^2 \tag{14}$$

due to the bounded norm of $\hat{\boldsymbol{\beta}}^{(N)}$.

Next we define an unbounded sequence of parameters as $\tilde{\mu}^{(N)} \equiv 0$ and $\tilde{\boldsymbol{\beta}}^{(N)} = \frac{N^\alpha}{2}\mathbf{x}_0$. With this

sequence of parameters, the residuals become

$$r_{N,i} = \begin{cases} \frac{N^{\nu+1}}{2} & i \in C \\ y_i - \frac{N^\nu}{2}\mathbf{x}_0^\intercal \mathbf{x}_i & i \notin C \end{cases},$$

which means all residuals tend to infinity for $N \to \infty$. Again, this implies that $\hat{\sigma}(\tilde{\mu}^{(N)}, \tilde{\boldsymbol{\beta}}^{(N)}, \tilde{\mathbf{Z}}_m^{(N)}) \to \infty$. The decomposition of the M-estimation equation yields

$$\sum_{i \notin C} \rho\left(\frac{y_i - \frac{N^\nu}{2}\mathbf{x}_0^\intercal \mathbf{x}_i}{\hat{\sigma}(\tilde{\mu}^{(N)}, \tilde{\boldsymbol{\beta}}^{(N)}, \tilde{\mathbf{Z}}_m^{(N)})}\right) + \sum_{i \in C} \rho\left(\frac{N^{\nu+1}/2}{\hat{\sigma}(\tilde{\mu}^{(N)}, \tilde{\boldsymbol{\beta}}^{(N)}, \tilde{\mathbf{Z}}_m^{(N)})}\right) = n\delta.$$

Taking the limit for $N \to \infty$ on all terms, the first sum goes to 0 and with the same argument as before, we have

$$\lim_{N\to\infty} \frac{1}{N^{2\nu+2}} \mathcal{L}_{\text{PS}}(\tilde{\mu}^{(N)}, \tilde{\boldsymbol{\beta}}^{(N)}, \tilde{\mathbf{Z}}_m^{(N)}) = \frac{\gamma^2}{4} \tag{15}$$

because the L1 norm of $\mathbf{x}_0$ is finite.

Hence, for large enough $N_0$, the limits (14) and (15) give

$$\frac{1}{N^{2\nu+2}} \mathcal{L}_{\text{PS}}(\tilde{\mu}^{(N)}, \tilde{\boldsymbol{\beta}}^{(N)}, \tilde{\mathbf{Z}}_m^{(N)}) < \frac{1}{N^{2\nu+2}} \mathcal{L}_{\text{PS}}(\hat{\mu}^{(N)}, \hat{\boldsymbol{\beta}}^{(N)}, \tilde{\mathbf{Z}}_m^{(N)}) \quad \forall N \geq N_0. \tag{16}$$

This implies a bounded $(\hat{\mu}^{(N)}, \hat{\boldsymbol{\beta}}^{(N)})$ can not be the minimum of the objective function for the contaminated sample. □

## 8.2 Statistical Consistency

We first discuss Condition [X0]. The following two lemmas, first discussed in [28] and included here for completeness sake, give sufficient assumptions for it to hold.

**Lemma 8.2.** *For $\mathcal{A} \subset \{1, ..., n\}, \#\mathcal{A} = [n\gamma]$ let*

$$\Sigma(\mathcal{A}) = \frac{1}{[n\gamma]} \sum_{i \in \mathcal{A}} \mathbf{x}_i \mathbf{x}_i^\intercal.$$

*Let $\rho_{1,n}(\mathcal{A})$ be the smallest eigenvalue of $\Sigma(\mathcal{A})$. Then*

$$\min_{\mathcal{A}\subset\{1,\ldots,n\},\#\mathcal{A}=[n\gamma]} \rho_{1,n}(\mathcal{A}) \leq \eta_n(\gamma)^2.$$

*Hence, $\liminf_n \eta_n(\gamma) > 0$ holds if the smallest eigenvalues of the covariance matrices formed from any subsample of size $[n\gamma]$ are uniformly bounded away from zero.*

*Proof.* Take $\boldsymbol{\theta}$ with $\|\boldsymbol{\theta}\| = 1$. Then

$$\boldsymbol{\theta}^\intercal \Sigma(\mathcal{A})\boldsymbol{\theta} \leq \max_{i\in\mathcal{A}} |\mathbf{x}_i^\intercal \boldsymbol{\theta}|^2.$$

Hence

$$\rho_{1,n}(\mathcal{A}) \leq \min_{\|\boldsymbol{\theta}\|=1} \max_{i\in\mathcal{A}} |\mathbf{x}_i^\intercal \boldsymbol{\theta}|^2$$

which implies that

$$\min_{\mathcal{A}\subset\{1,\ldots,n\},\#\mathcal{A}=[n\gamma]} \rho_{1,n}(\mathcal{A}) \leq \eta_n(\gamma)^2.$$

□

The following Lemma shows that if $\mathbf{X}_i$, $i = 1, \ldots, n$ are independent and identically distributed random vectors in $\mathbb{R}^p$ sampled from an appropriate distribution, then [X0] holds in probability for $\mathbf{X}_i$, $i = 1, \ldots, n$. In particular, [X0] will hold in probability for $\mathbf{x}_i = \mathbf{X}_i$, $i = 1, \ldots, n$, conditional on $\mathbf{X}_i$, $i = 1, \ldots, n$. Note that the assumption of Lemma 8.3 holds for example if $\mathbf{X}_i \sim N_p(\mathbf{0}, \mathbf{M}_n)$ and there exists some $\kappa > 0$ such that the smallest eigenvalue of $\mathbf{M}_n$ is bounded below by $\kappa$ for all $n$.

**Lemma 8.3.** *Assume $\mathbf{X}_i$, $i = 1, \ldots, n$ are independent and identically distributed random vectors in $\mathbb{R}^p$ such that there exists $\eta_1, \eta_2$ with $0 < \eta_1, \eta_2 < 1$ such that, for all $n$*

$$\sup_{\|\boldsymbol{\theta}\|=1} \mathbb{P}\left(|\mathbf{X}^\intercal \boldsymbol{\theta}| < \eta_1\right) < 1 - \eta_2.$$

*Moreover, assume $p/n \to 0$. Then [X0] holds in probability for $\mathbf{X}_i$, $i = 1, \ldots, n$.*

*Proof.* It can be shown, using maximal inequalities such as those of Theorem 5.1 of Smucler [28], that if $p/n \to 0$

$$\sup_{\|\boldsymbol{\theta}\|=1} \left| \frac{1}{n} \sum_{i=1}^{n} I\left\{|\mathbf{X}_i^\intercal \boldsymbol{\theta}| < \eta_1\right\} - \mathbb{P}\left(|\mathbf{X}^\intercal \boldsymbol{\theta}| < \eta_1\right) \right| \xrightarrow{P} 0.$$

Hence, with arbitrarily high probability, for large enough $n$,

$$\sup_{\|\boldsymbol{\theta}\|=1} \frac{1}{n} \sum_{i=1}^{n} I\left\{|\mathbf{X}_i^\intercal \boldsymbol{\theta}| < \eta_1\right\} < \sup_{\|\boldsymbol{\theta}\|=1} \mathbb{P}\left(|\mathbf{X}^\intercal \boldsymbol{\theta}| < \eta_1\right) + \eta_2/2 < 1 - \eta_2/2.$$

In this case, for any $\gamma$ such that $1 - \eta_2/2 < \gamma < 1$, for large enough $n$ it follows that for all $\boldsymbol{\theta}$ with $\|\boldsymbol{\theta}\| = 1$ and all subsets $\mathcal{A}$ of $\{1, \ldots, n\}$ with $\#\mathcal{A} = [n\gamma]$ there exists $i \in \mathcal{A}$ such that $|\mathbf{X}_i^\intercal \boldsymbol{\theta}| \geq \eta_1$, which implies $\eta_n(\gamma) \geq \eta_1$. □

The following Lemma, proven in Smucler [28], gives necessary conditions for $\liminf_n \eta_n(\gamma) > 0$ to hold. Note that the Lemma implies that the smallest eigenvalue of the sample covariance matrix of the predictors is bounded away from zero.

**Lemma 8.4.** *Assume there exists a constant $M > 0$ such that $1/n \sum_{i=1}^{n} \|\mathbf{x}_i\|^2 \leq pM$ for all $n$. Then if $\liminf_n \eta_n(\gamma) > 0$ for some $0 < \gamma < 1$, there exists positive numbers $\eta_1, \eta_2$ and $n_0$ such that*

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\intercal I\{\|\mathbf{x}_i\| < \eta_1 \sqrt{p}\} - \eta_2 \mathbf{I}_p$$

*is positive definite for all $n \geq n_0$.*

See also Examples 1, 2 and 3 of Davies [27].

Next, we prove Theorem 4.2. Our proofs leverage heavily on the results of Smucler [28].

*Proof of Theorem 4.2.* We will prove (i), the proof of (ii) is very similar and is thus omitted. For $v \in \mathbb{R}$, $s \in \mathbb{R}^+$ let

$$R(v, s) = \mathbb{E}_{F_0} \rho_c \left( \frac{u - v}{s} \right).$$

Let $\hat{\boldsymbol{\beta}}^S$ be the S-estimator corresponding to $\hat{\boldsymbol{\beta}}^{PS}$. Then

$$\hat{\sigma}(\hat{\boldsymbol{\beta}}^S)^2 \leq (\hat{\sigma}^{PS})^2 \leq \hat{\sigma}(\boldsymbol{\beta}_0)^2 + \frac{\lambda_S}{n}(P_\alpha(\boldsymbol{\beta}_0)/n) \tag{17}$$

By Lemma 4.1 of Smucler [28], we have that $\hat{\sigma}(\hat{\boldsymbol{\beta}}^S) \xrightarrow{P} s(F_0)$. Since by [B0], $\lambda_S(P_\alpha(\boldsymbol{\beta}_0)/n) \to 0$, it suffices to show that $\hat{\sigma}(\boldsymbol{\beta}_0) \xrightarrow{P} s(F_0)$. Fix $\varepsilon > 0$. We can find $\zeta > 0$ such that

$$\mathbb{E}_{F_0}\rho_c\left(\frac{u}{s(F_0)-\varepsilon}\right) \geq \delta + \zeta \text{ and } \mathbb{E}_{F_0}\rho_c\left(\frac{u}{s(F_0)+\varepsilon}\right) \leq \delta - \zeta.$$

Then by the Law of Large Numbers

$$\lim \frac{1}{n}\sum_{i=1}^n \rho_c\left(\frac{u_i}{s(F_0)-\varepsilon}\right) \geq \delta + \zeta \; a.s. \text{ and } \lim \frac{1}{n}\sum_{i=1}^n \rho_c\left(\frac{u_i}{s(F_0)+\varepsilon}\right) \leq \delta - \zeta \; a.s.$$

Since $\rho_c$ is monotone in $|t|$, with probability one, for large enough $n$, $s(F_0)-\varepsilon \leq \hat{\sigma}(\boldsymbol{\beta}_0) \leq s(F_0)+\varepsilon$. In particular, $\hat{\sigma}(\boldsymbol{\beta}_0) \xrightarrow{P} s(F_0)$.

We now prove the second part of (i). Fix $\varepsilon > 0$. Note that by the definitions of $\hat{\boldsymbol{\beta}}^{PS}$ and $\hat{\sigma}^{PS}$

$$\frac{1}{n}\sum_{i=1}^n \rho_c\left(\frac{u_i - \mathbf{x}_i^\intercal(\hat{\boldsymbol{\beta}}^{PS} - \boldsymbol{\beta}_0)}{\hat{\sigma}^{PS}}\right) = \delta.$$

By Lemma 5.2 of Smucler [28] we have that

$$\sup_{\mathbf{b}\in\mathbb{R}^p, 0<s<2s(F_0)} \frac{1}{n}\left|\sum_{i=1}^n \left(\rho_c\left(\frac{u_i - \mathbf{x}_i^\intercal \mathbf{b}}{s}\right) - R(\mathbf{x}_i^\intercal \mathbf{b}, s)\right)\right| \xrightarrow{P} 0. \tag{18}$$

Hence

$$\frac{1}{n}\sum_{i=1}^n R(\mathbf{x}_i^\intercal(\hat{\boldsymbol{\beta}}^{PS} - \boldsymbol{\beta}_0), \hat{\sigma}^{PS}) \xrightarrow{P} \delta \tag{19}$$

By Lemma 4.2 of Smucler [28] (i)

$$R(0, \hat{\sigma}^{PS}) \xrightarrow{P} \delta. \tag{20}$$

By (19), for any given $\zeta > 0$, with arbitrarily high probability, for large enough $n$ we have that

$$\frac{1}{n}\sum_{i=1}^{n} R(\mathbf{x}_i^\intercal(\hat{\boldsymbol{\beta}}^{PS} - \boldsymbol{\beta}_0), \hat{\sigma}^{PS}) \leq \delta + \zeta. \tag{21}$$

The rest of the proof now follows along the same lines of the proof of Theorem 4.3 of [28]. □

# References


[1] R. Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), pp. 267–288. ISSN: 00359246.

[2] H. Zou and T. Hastie. "Regularization and Variable Selection via the Elastic Net". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67.2 (2005), pp. 301–320. ISSN: 13697412, 14679868.

[3] R. Maronna, D. Martin, and V. Yohai. *Robust statistics: theory and methods*. Wiley Series in Probability and Statistics. Wiley, 2006. ISBN: 9780470010921.

[4] J. Fan and H. Peng. "Nonconcave penalized likelihood with a diverging number of parameters". In: *The Annals of Statistics* 32.3 (June 2004), pp. 928–961.

[5] J. Fan, Q. Li, and Y. Wang. "Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.1 (2017), pp. 247–265. ISSN: 1467-9868.

[6] J. A. Khan, S. V. Aelst, and R. H. Zamar. "Robust Linear Model Selection Based on Least Angle Regression". In: *Journal of the American Statistical Association* 102.480 (2007), pp. 1289–1299.

[7] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. "Least angle regression". In: *The Annals of Statistics* 32.2 (Apr. 2004), pp. 407–499.

[8] A. Alfons, C. Croux, and S. Gelper. "Sparse least trimmed squares regression for analyzing high-dimensional large data sets". In: *The Annals of Applied Statistics* 7.1 (Mar. 2013), pp. 226–248.

[9] P. J. Rousseeuw. "Least Median of Squares Regression". In: *Journal of the American Statistical Association* 79.388 (1984), pp. 871–880.

[10] V. J. Yohai. "High Breakdown-Point and High Efficiency Robust Estimates for Regression". In: *The Annals of Statistics* 15.2 (1987), pp. 642–656. ISSN: 00905364.

[11] R. A. Maronna. "Robust Ridge Regression for High-Dimensional Data". In: *Technometrics* 53.1 (2011), pp. 44–53. ISSN: 00401706.

[12] E. Smucler and V. J. Yohai. "Robust and sparse estimators for linear regression models". In: *Computational Statistics & Data Analysis* 111.C (2017), pp. 116–130.

[13] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al. "Least angle regression". In: *The Annals of statistics* 32.2 (2004), pp. 407–499.

[14] P. Rousseeuw and V. J. Yohai. "Robust regression by means of S-estimators". In: *Robust and nonlinear time series analysis*. Springer, 1984, pp. 256–272.

[15] P. J. Huber and E. M. Ronchetti. *Robust statistics*. Hoboken, NJ, USA.: John Wiley & Sons, Inc., 2009.

[16] F. Clarke. *Optimization and nonsmooth analysis*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1990. ISBN: 9781611971309.

[17] M. Salibián-Barrera and V. J. Yohai. "A Fast Algorithm for S-Regression Estimates". In: *Journal of Computational and Graphical Statistics* 15.2 (2006), pp. 414–427.

[18] M. Koller and W. A. Stahel. "Nonsingular subsampling for regression S estimators with categorical predictors". In: *Computational Statistics* 32.2 (2017), pp. 631–646.

[19] D. Peña and V. Yohai. "A Fast Procedure for Outlier Diagnostics in Large Regression Problems". In: *Journal of the American Statistical Association* 94.446 (1999), pp. 434–445. ISSN: 01621459.

[20] R. Tomioka, T. Suzuki, and M. Sugiyama. "Super-Linear Convergence of Dual Augmented Lagrangian Algorithm for Sparsity Regularized Estimation". In: *Journal of Machine Learning Research* 12 (July 2011), pp. 1537–1586. ISSN: 1532-4435.

[21] J. Friedman, T. Hastie, and R. Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software, Articles* 33.1 (2010), pp. 1–22. ISSN: 1548-7660.

[22] O. Hössjer. "On the optimality of S-estimators". In: *Statistics & Probability Letters* 14.5 (1992), pp. 413–419. ISSN: 0167-7152.

[23] R. Maronna and V. Yohai. "Correcting MM estimates for "fat" data sets". In: *Computational Statistics & Data Analysis* 54 (Dec. 2010), pp. 3168–3173.

[24] D. L. Donoho and P. J. Huber. "The notion of breakdown point". In: *A Festschrift For Erich L. Lehmann*. Ed. by P. J. Bickel, D. K., and J. Hodges. CRC Press, 1982, pp. 157–184.

[25] M. R. Osborne, B. Presnell, and B. A. Turlach. "On the LASSO and its Dual". In: *Journal of Computational and Graphical Statistics* 9.2 (2000), pp. 319–337.

[26] P. J. Huber. "Robust Regression: Asymptotics, Conjectures and Monte Carlo". In: *Ann. Statist.* 1.5 (Sept. 1973), pp. 799–821.

[27] L. Davies. "The Asymptotics of S-Estimators in the Linear Regression Model". In: *Ann. Statist.* 18.4 (Dec. 1990), pp. 1651–1675.

[28] E. Smucler. "Asymptotic Statistical Properties of Redescending M-estimators in Linear Models with Increasing Dimension". In: *ArXiv e-prints* (Dec. 2016).

[29] A. Alfons. *robustHD: Robust Methods for High-Dimensional Data*. R package version 0.5.1. 2016.

[30] V. J. Yohai and R. H. Zamar. "High Breakdown-Point Estimates of Regression by Means of the Minimization of an Efficient Scale". In: *Journal of the American Statistical Association* 83.402 (1988), pp. 406–413. ISSN: 01621459.

[31] H. Zou and H. H. Zhang. "On the adaptive elastic-net with a diverging number of parameters". In: *The Annals of Statistics* 37.4 (Aug. 2009), pp. 1733–1751.

[32] D. Schmauss and M. Weis. "Cardiac Allograft Vasculopathy". In: *Circulation* 117.16 (2008), pp. 2131–2141. ISSN: 0009-7322.

[33] D. Lin et al. "Plasma protein biosignatures for detection of cardiac allograft vasculopathy". In: *The Journal of Heart and Lung Transplantation* 32.7 (2013), pp. 723–733. ISSN: 1053-2498.