Submitted to the Annals of Applied Statistics

PROTEOMIC BIOMARKER STUDY USING NOVEL ROBUST PENALIZED ELASTIC NET ESTIMATORS

By Gabriela V. Cohen Freue*, David Kepplinger, Matías Salibián-Barrera*, and Ezequiel Smucler

Department of Statistics, University of British Columbia

Abstract: In large-scale quantitative proteomic studies, scientists measure the abundance of hundreds or thousands of proteins from the human proteome in search of novel biomarkers for a given disease. Despite current innovations in biomedical technologies, advanced statistical and computational methods are still required to harness the rich information contained in these large and complex datasets. While penalized regression estimators can be used to identify potential biomarkers among a large set of molecular features, it is well-known that the performance and statistical properties of the selected model depend on the loss and penalty functions used to construct the regularized estimator. For example, the presence of outlying observations in the data can seriously affect classical estimators that penalize the square error loss function. Similarly, the choice of the penalty function in these estimators is important to be able to preserve groups of correlated proteins in the selected model. Thus, in this paper we propose a new class of penalized robust estimators based on the elastic net penalty, which can be tuned to keep groups of correlated variables together as they enter or leave the model, while protecting the resulting estimator against possibly aberrant observations in the dataset. Our robust penalized estimators have very good robustness properties and are also consistent under relatively weak assumptions. In this paper we also propose an efficient algorithm to compute our robust penalized estimators and we derive a data-driven method to select the penalty term, which is a critical part of any application with real data. Our numerical experiments show that our proposals compare favorably to other robust penalized estimators. Noteworthy, our robust estimators identify new potentially relevant biomarkers of cardiac allograft vasculopathy that are not found with non-robust alternatives. Importantly, the selected model is validated in a new set of 52 test samples, achieving an area under the receiver operating characteristic curve (AUC) of 0.85.

*Supported by NSERC Discovery Grant

MSC 2010 subject classifications: Primary 62J; secondary 62J05, 62J07, 62J07; Primary 62P; secondary 62P10; Primary 62F; secondary 62F35

Keywords and phrases: robust estimation, regularized estimation, penalized estimation, elastic net penalty, proteomics biomarkers

1. Introduction. Biomarkers are indicators of 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 associated 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 linear regression to model the association between hundreds of plasma protein levels and the obstruction of the left anterior descending artery measured in heart transplant patients to identify proteomic biomarkers of cardiac allograft vasculopathy (CAV). CAV is a major complication suffered by 50% of cardiac transplant recipients beyond the first year after transplantation, which is currently diagnosed with highly invasive technics, including cardiac angiography and intravascular ultrasound (IVUS). Identifying these plasma proteomic biomarkers can result in the development of minimally invasive and clinically useful blood tests to diagnose CAV and improve patient care options. 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).

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) (Tibshirani, 1996; Zou and Hastie, 2005). However, most of these estimators penalize the square error loss function and are thus 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 estimators is essential to effectively interrogate the rich information contained in the human proteome.

Although many robust regression methods have been proposed in the literature (see Maronna, Martin and Yohai (2006) 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 (Fan and Peng, 2004; Fan, Li and Wang, 2017), thus not resistant to high leverage outliers commonly observed in large datasets. The first highly robust penalized estimators is the RLARS estimator (Khan, Aelst and Zamar, 2007), a modification of the Least Angle Regression method (Efron et al., 2004) where sample correlations are replaced with robust counterparts. A more recent proposal, SparseLTS (Alfons, Croux and Gelper, 2013), is an L_1-regularized version of the Least Trimmed Squares regression estimator Rousseeuw (1984), 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 (Yohai, 1987).

To overcome these limitations, Maronna (2011) 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 (2017) have recently proposed a penalized MM-LASSO estimator. However, as previously shown for the classical LASSO (Efron et al., 2004), their MM-LASSO estimator cannot select more variables than the number of available observations. In addition, if the data contain groups of highly correlated explanatory variables, LASSO tends to randomly select only one variable within each group ignoring the relevance of other covariates.

In -omics datasets the number of measured features is usually 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 combine robust loss functions with 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. The resulting penalized robust regression estimators are not limited by the number of available samples and can select groups of correlated proteins with common functional objectives or cellular mechanisms, while being protected against possible outliers in the dataset.

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 squared residuals. Second, to get an estimator that is highly robust and at the same time efficient, we use PENSE to initialize a penal-

ized 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 (Zou and Hastie, 2005). In particular in this paper, we use our robust estimators to identify potentially relevant proteomic biomarkers of cardiac allograft vasculopathy from a set of 37 plasma samples from heart transplant patients, collected at 1 year after transplantation. However, 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 our elastic net regularized robust estimators, along with efficient algorithms to compute them. In Section 4, we show that our estimator is robust, in the sense of not being unduly influenced by a small proportion of potentially atypical patients. Before we discuss our findings in the cardiac allograft vasculopathy study in Section 6, we explore the properties of our estimator with a simulation study, reported in Section 5. Final remarks and conclusions can be found in Section 7. The on-line Supplementary Material contains many technical details and proofs.

2. PENSE: a new robust penalized regression estimator. As mentioned before, the relationship 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 $\mu \in \mathbb{R}$, and $\boldsymbol{\beta} \in \mathbb{R}^p$ 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 patient 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). In particular, in the proteomic case study analyzed in this paper, the number of patients (n) is 37 and the number of measured proteins (p) is 81. We assume that the response is centered and the covariates are standardized. Given the potential presence of outliers in our dataset, we center the data using column-wise medians and standardize each variable to have a median absolute deviation (from the median) equal to 1.

Although thousands of molecular features may be measured and analyzed in -omics studies, only a few are usually expected to be associated with a given disease. In other words, we expect this model to be sparse with many coefficients equal to zero, which leads us to consider regularized regression

estimators. In particular, since proteomic biomarkers usually form groups of correlated predictors, we focus on elastic net penalties that tend to keep such groups of variables together as they enter or leave the model.

Furthermore, to protect the resulting estimator against atypical observations commonly present in -omics studies (e.g., due to technical issues in the sample preparation steps, or the presence of patients with unusual molecular profiles), instead of penalizing the classical variance of the residuals (square error loss function), we penalize the square of a *robust* residual scale estimator. More specifically, we propose to use a regularized version of an S-estimator (Rousseeuw and Yohai, 1984) that has good robustness properties to select the most relevant variables in the model and predict the response of interest.

Our new robust penalized estimator, which we call PENSE, is defined as the minimizer $\left(\hat{\mu}^{PS}, \hat{\boldsymbol{\beta}}^{PS}\right)$ of the penalized loss function

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

where $\sigma\left(\mu, \boldsymbol{\beta}\right)$ is a robust residual scale estimator, $\lambda_S \geq 0$ is the penalty level, and $\alpha \in [0, 1]$ determines the desired combination of the L_1- and L_2-penalties. In particular, if $\alpha = 1$, the estimator becomes a LASSO S-estimator, and if $\alpha = 0$, it becomes a Ridge S-estimator. The parameters λ and α determine the size of the identified model and can be chosen using different optimization criteria. In our application, we generate a moderate level of sparsity, aiming to select potentially good proteomic biomarkers while at the same time controlling the number of false biomarkers identified from the data.

In what follows we use a robust M-estimate for $\sigma\left(\mu, \boldsymbol{\beta}\right)$, given implicitly by 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 ρ and tuning constant $\delta \in (0, 1)$. Both ρ and δ need to be chosen jointly in order to obtain robust and consistent estimators. For more details we refer to Maronna, Martin and Yohai (2006).

Given a fixed penalty parameter λ_S, minimizing the objective function (2) is challenging due to its non-convexity and the lack of differentiability of the elastic net penalty at $\boldsymbol{\beta} = \mathbf{0}$. However, since the unpenalized S loss is continuously differentiable and the elastic net penalty is locally Lipschitz, the penalized S loss (2) is locally Lipschitz. Thus, following the results in

Clarke (1990), we can derive its generalized gradient and search for a root that defines a good local optimum of (2). In particular, the generalized gradient of the penalized S loss is given by

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

where $P_\alpha(\boldsymbol{\beta}) = \frac{1}{2}(1-\alpha)||\boldsymbol{\beta}||_2^2 + \alpha||\boldsymbol{\beta}||_1$ is the elastic net penalty, $r_i(\mu,\boldsymbol{\beta}) = y_i - \mu - \mathbf{x}_i^\intercal\boldsymbol{\beta}$ are the residulas, and the weights $w_i(\mu,\boldsymbol{\beta})$ are proportional to $\frac{\sigma(\mu,\boldsymbol{\beta})}{r_i(\mu,\boldsymbol{\beta})}\rho'\left(\frac{r_i(\mu,\boldsymbol{\beta})}{\sigma(\mu,\boldsymbol{\beta})}\right)$.

In order to find a root of the generalized gradient in (4) above, note that it coincides with the subgradient of the classical weighted elastic net loss, except that the weights depend on the unknown coefficients $(\mu,\boldsymbol{\beta})$. This suggests the following iterative procedure: given an initial estimate $(\mu^{\text{init}},\boldsymbol{\beta}^{\text{init}})$ and its corresponding M-scale estimate $\sigma(\mu^{\text{init}},\boldsymbol{\beta}^{\text{init}})$, obtain an improved set of parameter estimates by computing a weighted elastic net with weights $w_i(\mu^{\text{init}},\boldsymbol{\beta}^{\text{init}})$ as above. Next, compute the corresponding updated weights and iterate. We refer to this algorithm as iteratively reweighted elastic net (IRWEN) (see the Supplementary Materials for more details).

2.1. *Initial estimator.* Ideally, we want to find the global minimum of the objective function (2) that defines PENSE. However, because of the lack of convexity of this function, 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 IRWEN.

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., Salibián-Barrera and Yohai, 2006; Koller and Stahel, 2017). These strategies rely on constructing data-driven random starts by fitting the regression model on randomly chosen subsamples. The idea is that subsamples that do not contain outliers will provide regression estimators that are good starting points. To maximize the chance of obtaining a clean starting point, the subsamples are taken with as few points as possible. However, it is not clear how many points to include in the random subsamples when the number of explanatory variables exceeds the sample size. In our application, as well as in many proteomics studies, the number of patients ($n = 37$) is already smaller than the number of measured proteins ($p = 81$), thus unfortunately previous results on subsampling methodologies do not directly generalized to our case.

Alfons, Croux and Gelper (2013) proposed to compute a LASSO estimator

on subsamples of size 3 to initialize the algorithm of their SparseLTS estimator. However, the size of the subsample limits the number of variables that LASSO can select for the initial estimator. In our problem, starting the iterations with only 3 proteins in the model may result in an undesirably sparse final model, with the potential loss of relevant biomarkers. Since increasing the size of the subsamples can considerably increase the computational time of the algorithm, this strategy may not be feasible in many modern statistical applications.

Adapting the approaches of Peña and Yohai (1999) and Maronna (2011), we construct clean subsamples of our proteomics data by removing different sets of outlying observations. These sets of potential outliers are flagged using the principal sensitivity components (PSCs), which measure the effect of each data point on the estimated model. The classical EN estimator is then computed on each resulting subsample and used as candidate initial estimators for IRWEN. More details can be found in the Supplementary Materials.

In most applications the optimal level of penalization is not known in advance, and λ_S in (2) is chosen from a grid of K possible penalty values, $\lambda_S^{(1)} < \lambda_S^{(2)} < \ldots < \lambda_S^{(K)}$, based on the predictive performance of the penalized estimator. Since the number of selected variables (proteins in our case) can vary greatly among different levels of penalization, fine grids with large K are usually preferred. In our case, we examine our estimator at $K = 100$ penalty values to evaluate the contribution of small sets of proteins gradually incorporated in (or removed from) the selected model. To ease the burden of computing an initial estimator (or several candidates) for every $\lambda_S^{(k)}$ in the grid, we use a strategy of "warm" starts, in which a local optimum of (2) at a penalty value in the grid can be used to initiate the iterative algorithm at adjacent penalty levels.

The "domino effect" exploited by "warm" start algorithms is commonly used to compute other penalized estimators based on iterative solvers (Friedman, Hastie and Tibshirani, 2010; Tomioka, Suzuki and Sugiyama, 2011). 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 IRWEN, harnessing the benefits of both strategies. We refer to the Supplementary Materials for more details.

3. PENSEM: a new penalized MM-estimator. Many applications where regularized estimators are utilized have relatively few samples. In particular, proteomics data sets tend to be relatively small due to the costs associated with their collection. Hence, reducing the sampling variability of the regression estimators may help lower the threshold over which protein effects can be detected in linear models like (1).

Following Yohai (1987) we propose to refine the robust PENSE estimator to obtain a penalized elastic net MM-estimator with higher efficiency (lower variance), which we call PENSEM. This estimator is defined as the minimizer $(\hat{\mu}^{PM}, \hat{\boldsymbol{\beta}}^{PM})$ of the penalized loss function

(5)
$$\mathcal{L}_{\text{PM}}(\mu, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \rho_2\left(\frac{y_i - \mu - \mathbf{x}_i^\intercal \boldsymbol{\beta}}{\hat{\sigma}_0}\right) + \lambda_M \left(\frac{1}{2}(1-\alpha)||\boldsymbol{\beta}||_2^2 + \alpha||\boldsymbol{\beta}||_1\right),$$

where the residual scale estimate $\hat{\sigma}_0$ is fixed, and $\rho_2 \leq \rho$, the function used to compute $\hat{\sigma}_0$. It is easy to see that, as previously discussed for PENSE, we can use an IRWEN algorithm to find local minima of the penalized M loss function (5), initialized using the PENSE estimate $\left(\hat{\mu}^{PS}, \hat{\boldsymbol{\beta}}^{PS}\right)$. However, the estimation of the initial residual scale, $\hat{\sigma}_0$, requires some special attention.

For datasets with few explanatory variables (i.e, small p relative to the sample size n), the scale based on the residuals from an S-estimator has been used to compute MM-estimators. However, Maronna and Yohai (2010) have noted that this scale estimator usually underestimates the true error scale if the ratio p/n is high. This problem becomes even more serious in applications like ours where the sample size ($n = 37$) is smaller than the number of explanatory variables ($p = 81$). Following this observation, Maronna (2011) adjusts the residual scale estimator of the Ridge-S if its effective degrees of freedom is larger than 10% of the sample size. Based on the results of our numerical studies and considering the sparsity of our model, we also compute PENSEM using an adjusted residual scale estimator $\hat{\sigma}_0 = q\,\hat{\sigma}(\mu^{PS}, \boldsymbol{\beta}^{PS})$. Further details on the correction factor q and other adjustments suggested in the literature are given in the Supplementary Materials.

Finally, we need to determine the level of penalization of PENSEM, which controls the number of variables selected in the final model. In our problem, this parameter limits the number of potential biomarkers that we migrate to the validation stage. Although both PENSE and PENSEM are defined using the same penalty function, the levels of penalization implied by specific values of λ_M and λ_S are not equivalent since the loss functions are generally different. Thus, the optimum penalty parameter λ_M for PENSEM is also

chosen from a grid of candidate values which might be different from the grid used to determine λ_S. To ease the computational burden, the initial estimator to optimize the penalized M loss function (5) for any λ_M is fixed at the PENSE estimate $\left(\hat{\mu}^{PS}, \hat{\boldsymbol{\beta}}^{PS}\right)$ obtained with the penalty level λ_S selected when computing PENSE.

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

4.1. *Robustness.* Technical challenges with sample preparation, and patients with atypical molecular profiles mean that the potential presence of outliers is an important concern when working with proteomics datasets. One measure of robustness of an estimator against potential outliers is its (finite-sample) breakdown point (Donoho and Huber, 1982), which in our case corresponds to the largest proportion of samples in the data set that could have been contaminated arbitrarily and still result in a bounded regression estimator. The larger this proportion, the "safer" a robust regression estimator is, in the sense of not being completely determined by a small number of atypical patients in the training set.

Another interest in the cardiac allograft vasculopathy study is the detection of potentially atypical samples in the data. Outliers in regression models like the one we use in our study can be flagged by considering the residuals from the fitted model. This approach is expected to work well when the estimated parameters have not been affected by the outliers one is trying to detect. Estimators with high breakdown point thus provide reliable outlier detection methods. We illustrate this successfully in Section 6 below.

Since penalized optimization problems are equivalent to constrained ones, one may conjecture that regularized estimators are "automatically" robust, in the sense that they are necessarily constrained and thus bounded. However, 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, Presnell and Turlach (2000), 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, Presnell and Turlach (2000)). It is easy to see

that, since the bound C_0 depends on the sample, it can become arbitrarily large when outliers are present in the data.

The formal definition of the finite-sample breakdown point of an estimator is as follows. Let $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^\intercal$ be a fixed dataset, where $\mathbf{z}_i = (y_i, \mathbf{x}_i^\intercal)^\intercal$. The replacement finite-sample breakdown point (FBP), $\epsilon^*(\hat{\boldsymbol{\theta}}; \mathbf{Z})$, of an estimator $\hat{\boldsymbol{\theta}}$ is defined as

$$\epsilon^*(\hat{\boldsymbol{\theta}}; \mathbf{Z}) = \max\left\{\frac{m}{n} : \sup_{\mathbf{Z}_m \in \mathcal{Z}_m} \|\hat{\boldsymbol{\theta}}(\mathbf{Z}_m)\| < \infty\right\}, \tag{6}$$

where the set $\mathcal{Z}_m$ contains all possible datasets $\mathbf{Z}_m$ with $0 < m < n$ of the original n observations replaced by arbitrary values (Donoho and Huber, 1982). In the proteomic case study analyzed in this paper, $n = 37$ corresponds to the number of independent plasma samples from cardiac transplant recipients.

The following theorem shows that the PENSE estimator retains the high-breakdown point of the parent unpenalized S-estimator. More specifically, the breakdown point of PENSE is at least $\min(\delta, 1-\delta)$, where δ is the right-hand side of the equation defining the residual scale M-estimator (3). In other words, if we compute PENSE with $\delta = 0.5$, as long as less than half of the patients in our study are representative of the target population, our robust estimator will not be unduly affected by potential outliers in the data.

THEOREM 4.1. *For a dataset of size n, let $m(\delta) \in \mathbb{N}$ be the largest integer smaller than $n\min(\delta, 1-\delta)$, where δ is the right-hand side of (3). Then, the finite-sample breakdown point of the PENSE estimator $(\hat{\mu}^{PS}, \hat{\boldsymbol{\beta}}^{PS})$ satisfies*

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

A proof of the theorem is given in the Supplementary Materials. Moreover, the proof in Smucler and Yohai (2017) 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. *Consistency.* Consistency is a desired statistical property of any estimator that in a sense ensures better estimates of the true model parameters as more data is collected. In addition to being robust, we prove that the coefficients estimated by PENSE and PENSEM converge to the true values when both the number of observations n and the number of predictors p

grow to infinity (Theorem 3.2 of the Supplementary Material). Importantly, our result does not make any moment assumptions on the distribution of the errors, and hence guarantees high-quality estimations in linear models with large sample sizes, even in cases with extremely heavy-tailed error distributions that could cause serious outlying observations. However, our proof of consistency requires that $p < n$, which may not be the case for many available datasets. In particular, in our proteomic biomarkers study the number of patients (n) is 37 and the number of measured proteins (p) is 81. Similarly, the number samples available may remain limited in many applications. Thus, it is important to complement these theoretical asymptotic results with extensive simulation studies.

5. Simulation Studies. Before we discuss our findings on the cardiac allograft vasculopathy study using PENSE and PENSEM in Section 6, we report here the results of a simulation study to further examine the properties of our estimators and compare their performance against that of other published robust and/or penalized estimators in various settings.

We consider data following a linear regression model of the form

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

and four different combinations of the number of observations (n), the number of predictors (p), the correlation structure of the explanatory variables $\mathbf{x}$, and the true regression coefficients ($\boldsymbol{\beta}$) (see Section 5.2 below).

We compare our PENSE and PENSEM estimators[1] against the classical LASSO and EN, as well as the robust regularized estimators SparseLTS (Alfons, Croux and Gelper, 2013) and the recently published MMLASSO (Smucler and Yohai, 2017), which are robust versions of LASSO. PENSE and PENSEM methods are computed using Tukey's Bisquare loss, given by

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

Whenever possible, we also include the oracle OLS and MM estimators, which only estimate the coefficients of the true active set of predictors. For SparseLTS, we use the implementation available in the R package robustHD (Alfons, 2016), 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.

[1] available at https://cran.r-project.org/package=pense

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

5.1. *The Penalty Parameters.* The level of penalization λ_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 τ-scale (Yohai and Zamar, 1988) 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 λ_M, always starting from the optimum λ_S^* chosen from the previous grid. The optimal λ_M^* is again chosen by 10-fold CV using the robust τ-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 λ_S and λ_M. Different strategies can be used to select the appropriate α 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, we chose a moderate level of sparsity to control the risk of missing promising markers and the cost of migrating irrelevant ones to the validation phase. In other contexts, one can compute the estimators for several different values of α and choose the value α^* that yields the best CV prediction performance. For a comprehensive discussion on this topic we refer to Zou and Hastie (2005).

As it was noted for the classical naïve EN estimator (Zou and Hastie, 2005), 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 correct both PENSE and PENSEM as $\sqrt{1 + 1/2(1 - \alpha^*)\lambda^*}\hat{\boldsymbol{\beta}}$. The intercept is corrected accordingly to maintain centered weighted residuals.

5.2. *Simulation Settings.* To demonstrate the benefits of the elastic net over the L_1 penalty, we include two simulation settings from Zou and Hastie (2005) and Zou and Zhang (2009). 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 (2005) 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 (2005), in which the number of predictors $p = 40$ is smaller than the sample size $n = 50$; $\sigma = 15$ and the vector of true regression coefficients is given by

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

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

$$x_j = z_{\lceil j/5 \rceil} + \delta_j \text{ where } z_l \sim N(0,1)\,,\ \delta_j \sim N(0, 0.01^2)\,,$$

for $j = 1, \ldots, 15$, $l = 1, 2, 3$, 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 to $p = 400$, which now exceeds the number of observations $n = 50$. The error term is generated as in setting (a) and the vector of true regression coefficients is given by

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

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

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

for $j = 1, \ldots, 60$, $l = 1, 2, 3$. 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 (2009), 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 $\mathbf{x} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ with covariance structure

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

(d) The final setting has a large number of predictors with $p = 995$, a 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, γ_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 $\mathbf{x} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ with covariance structure

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

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

The 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 (2011). Leverage points are introduced by changing the predictors $\mathbf{x}_i$ of the contaminated observations to

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

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 estimators 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.,

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

The model selection performance is assessed by the sensitivity (SENS) and specificity (SPEC) of the estimated coefficient vector $\hat{\boldsymbol{\beta}}$ defined by

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

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

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

As an 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.

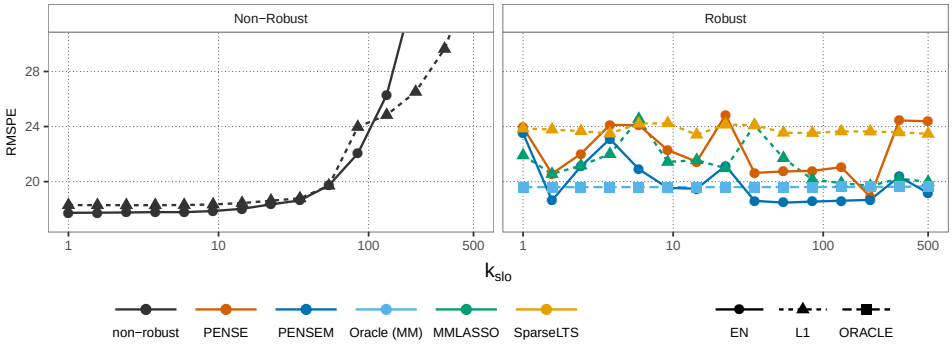

Figure 1: Root mean squared prediction error of different estimators over a grid of k_{slo} values ranging from 1 to 500 with 10% contamination under simulation setting (a).

For each simulation setting, we compute PENSE(M) as well as the classical EN for several values of α. In the results however, we only present the PENSE(M) estimators corresponding to the α^* 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 $\alpha^* = 0.9$, which were both chosen based on the CV performance of each estimator.

Setting (a) is tailored to favor the elastic net penalty 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-penalty of LASSO is that only a single predictor is selected within each group. However, if the penalty parameter λ is small enough, this single predictor can almost fully capture the effect of the entire group. Thus, the benefit of the elastic net penalty is only marginally visible in the prediction performance.

Among the robust estimators, 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 classi-

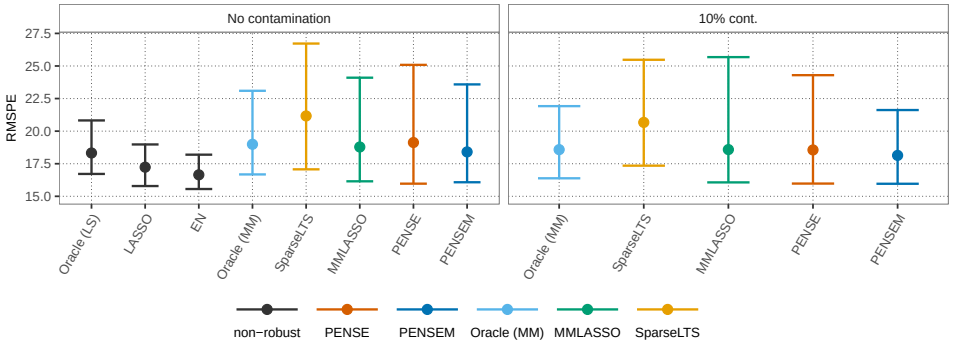

Figure 2: Average prediction performance of different estimators in simulation setting (a). The error bars extend from the 5% to the 95% quantile. For the uncontaminated case we report the RMSPE. For training data with 10% contamination we show the overall measure $\text{RMSPE}_{\text{cont}}$ over a grid of k_{slo} from 1 to 500. Classical EN uses $\alpha^* = 0.7$, while PENSE(M) is using $\alpha^* = 0.9$.

cal estimators, the difference between the robust regularized EN estimators (PENSE and PENSEM) and the MMLASSO is small.

The strength of the elastic net 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 whole groups. Thus, the sensitivity of the LASSO methods is weak compared to that of elastic net methods. For the classical EN and PENSE(M) estimators, the selection of relevant variables brings also some of the irrelevant ones shown by a slight drop in specificity.

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 shows again the best prediction performance of the robust estimators. It is clearly visible that the M-step reduces variability in the 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 select 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 irrelevant predictors. Under

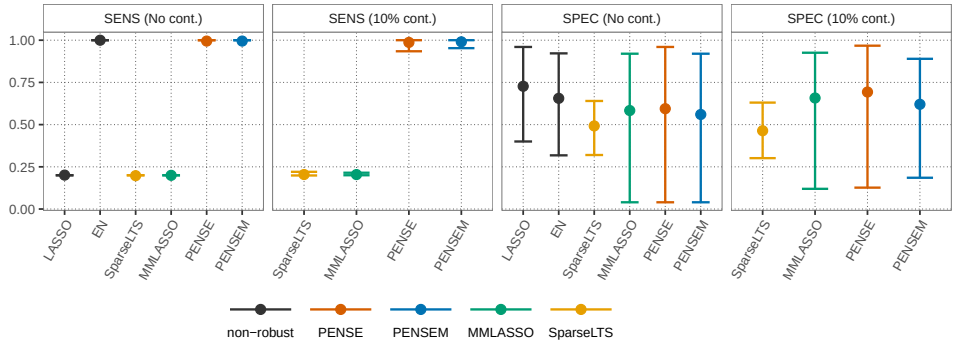

Figure 3: Average specificity and sensitivity of different estimators in simulation setting (a). The error bars extend from the 5% to the 95% quantile. For training data with 10% contamination we show the area under the curve ($\text{SENS}_{\text{cont}}$ and $\text{SPEC}_{\text{cont}}$) over a grid of k_{slo} from 1 to 500. Classical EN uses $\alpha^* = 0.7$, while PENSE(M) is using $\alpha^* = 0.9$.

contamination, PENSE selects all 60 active predictors 88% of the time and on average selects only 23 of the 340 irrelevant predictors. In the uncontaminated case the model selection of PENSE is on average even outperforming the classical EN.

Setting (c): This is the last setting where the elastic net penalty should have an advantage over the L_1 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 based on an L_1 penalty. It is clearly visible that the $L1$-based estimators have difficulty addressing the moderate to high correlation among active predictors in this setting. For model selection, as shown in Figure 7, the classic EN and PENSE(M) again outperform $L1$-based methods, which not surprisingly comes at the cost of a drop in their specificity. PENSE selects around 17 of the 54 irrelevant 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): The 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 the $L1$ to the $L2$ penalties, thus being adjustable to different degrees of sparsity. As expected, PENSEM results are very sim-

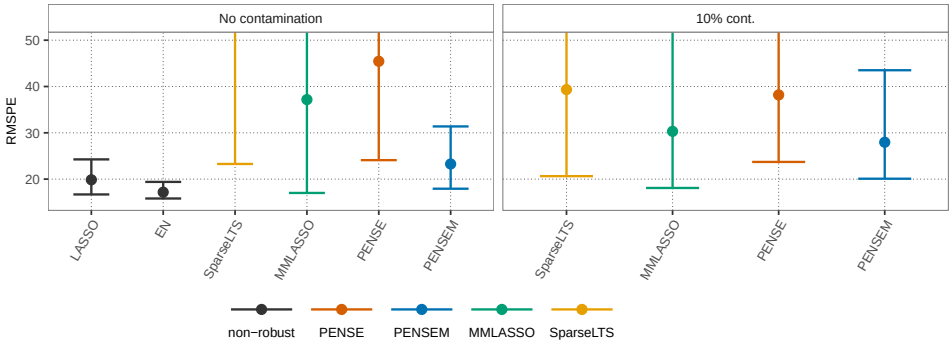

Figure 4: Average prediction performance of different estimators in simulation setting (b). The error bars extend from the 5% to the 95% quantile. For the uncontaminated case we report the RMSPE. For training data with 10% contamination we show the overall measure $\text{RMSPE}_{\text{cont}}$ over a 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.

ilar to MMLASSO, with observed differences coming from the initial estimators used to initialize the 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, these simulation results 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 $L1$-based estimates if important predictors are correlated. Especially in settings with large number of relevant correlated covariates relative to the sample size, the elastic net penalty is beneficial for both prediction performance and identification of important predictors.

6. Biomarkers of Cardiac Allograft Vasculopathy. In this Section, we use PENSEM to select potential plasma biomarkers of cardiac al-

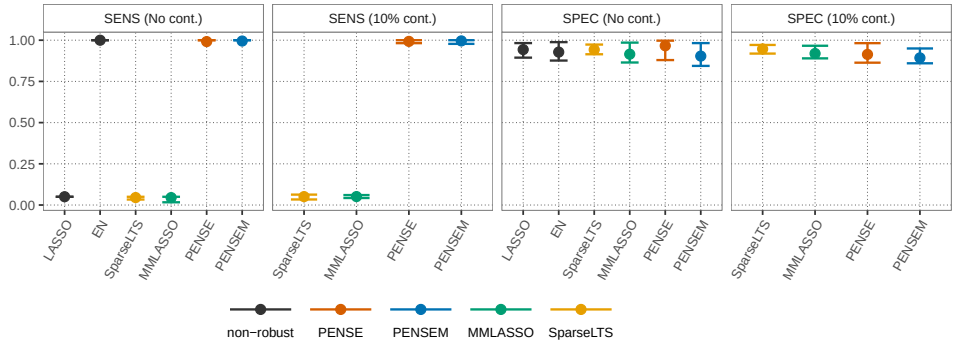

Figure 5: Average specificity and sensitivity of different estimators in simulation setting (b). The error bars extend from the 5% to the 95% quantile. For training data with 10% contamination we show the area under the curve ($\text{SENS}_{\text{cont}}$ and $\text{SPEC}_{\text{cont}}$) over a grid of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

lograft vasculopathy (CAV), a major complication suffered by 50% of cardiac transplant recipients beyond the first year after transplantation. 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 (Schmauss and Weis, 2008). 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 (BiT) 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, 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. (2013), which developed a proteomic classifier of CAV using a preliminary univariate robust screening of proteins and a classical EN classification method. PENSE and

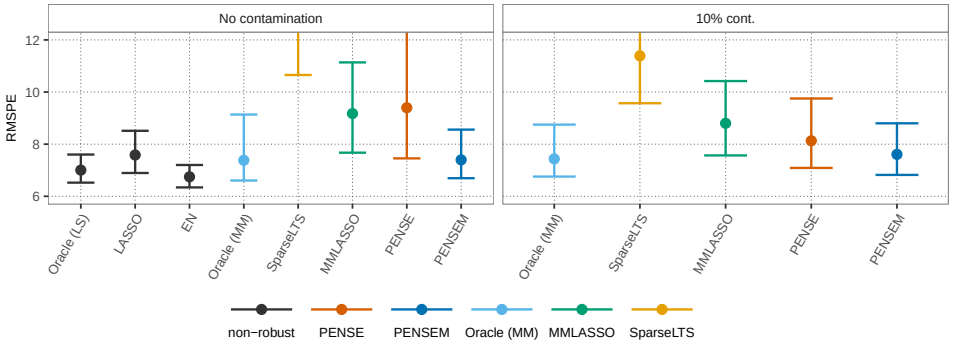

Figure 6: Average prediction performance of different estimators in simulation setting (c). The error bars extend from the 5% to the 95% quantile. For the uncontaminated case we report the RMSPE. For training data with 10% contamination we show the overall measure $\text{RMSPE}_{\text{cont}}$ over a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

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.

We validate our results on an independent set of 52 patients collected by BiT in the second phase of their study. For the validation phase, the plasma samples collected around one year after transplantation were analyzed with a much more sensitive proteomics technology, called Multiple Reaction Monitoring (MRM), which allows the quantification of targeted proteins (Cohen Freue and Borchers, 2012; Domanski et al., 2012). Since the use of MRM requires the development of stable isotope-labeled standard peptides to measure the targeted proteins, only a subset of candidate proteins is usually available in this validation phase. The stenosis of the LAD artery was measured equally in all patients from the discovery and test cohorts, using cardiac angiography.

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, both tuned to achieve a 25% breakdown point. In this application we induce a moderate level of sparsity using $\alpha^* = 0.6$, aiming to control the number of potential false biomarkers identified and potential good biomarkers missed in this study. As explained

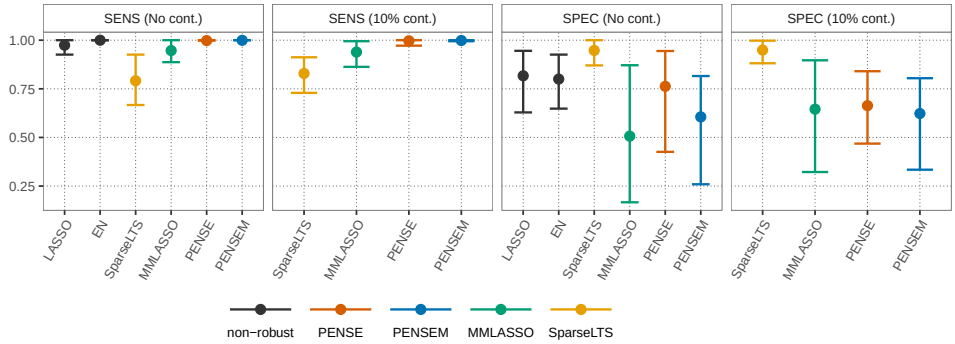

Figure 7: Average specificity and sensitivity of different estimators in simulation setting (c). The error bars extend from the 5% to the 95% quantile. For training data with 10% contamination we show the area under the curve ($\text{SENS}_{\text{cont}}$ and $\text{SPEC}_{\text{cont}}$) over a grid of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

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 CV. To make this selection more stable, we repeat this estimation 200 times over the full grid of penalty values and select λ_S^* as the maximum λ 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 35 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 (λ_M^*) with the same criteria used to select λ_S^*. PENSEM selects 15 out of the 35 potential markers selected by PENSE to predict the diameter of the LAD artery. Analogously, using the "one standard error" (1SE) rule such that the CV error is within one standard error of that of the minimum, the classical EN estimator (using the same α parameter) does not select any variable (i.e., the intercept-only model is selected). Figure 10 illustrates PENSEM's estimates of the regression coefficients for different values of λ_M (i.e., PENSEM's regularization path), highlighting in blue the coefficients selected at the optimal level of penalization chosen (i.e., λ_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 C4B/C4A, APOE, AMBP, and SHBG (Lin et al., 2013). However, further analysis of this dataset using our estimators allows the identification of new potential markers, including some

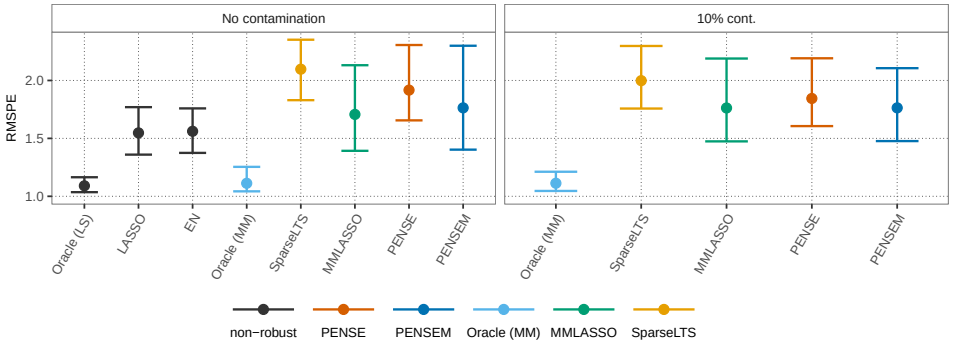

Figure 8: Average prediction performance of different estimators in simulation setting (d). The error bars extend from the 5% to the 95% quantile. For the uncontaminated case we report the RMSPE. For training data with 10% contamination we show the overall measure $\text{RMSPE}_{\text{cont}}$ over a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

additional proteins of the coagulation and complement cascades (F10 and CFB, respectively), another apolipoprotein (APOC2), and new homoglobin subunits (HBD, HBA, HBZ), among other biologically relevant proteins. 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 by 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 four patients in the lower part of the figure (B-514, B-584, B-527 and B-561 measured in weeks 51 and 52 after transplant as indicated by the sample labels). However, a second measurement of the LAD of the last three patients using a more accurate technique (IVUS) indicates that their arteries present a mild stenosis with about 16% area reduction, as suggested by PENSEM's predictions (negative residuals). Similarly, the stenosis of B-381 might have been overestimated by the angiography performed at week 51 (91% area reduction) compared to the results of the IVUS test (79% area reduction). Other outlying measurements may be present in the iTRAQ protein measurements of these patient samples highlighted by PENSEM.

The performance of the estimators is initially evaluated by 200 replications of 10-folds cross-validations and compared to that of the classical EN and some robust estimators (see Table 2). An α value of 0.6 is used for all estimators based on the elastic net penalty. In terms of prediction, all es-

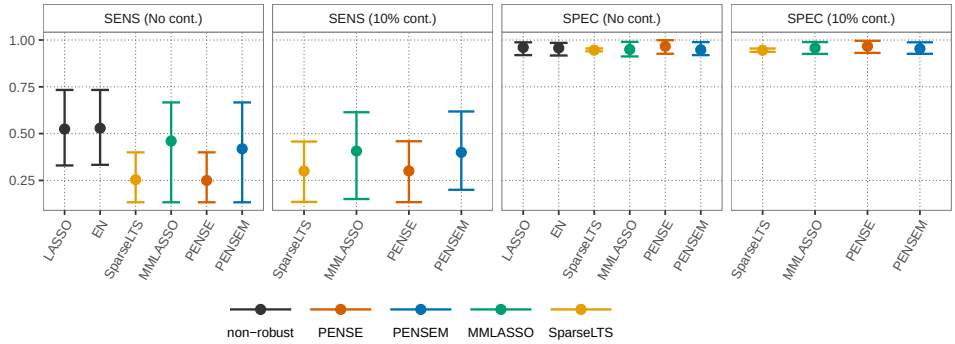

Figure 9: Average specificity and sensitivity of different estimators in simulation setting (d). The error bars extend from the 5% to the 95% quantile. For training data with 10% contamination we show the area under the curve ($\text{SENS}_{\text{cont}}$ and $\text{SPEC}_{\text{cont}}$) over a grid of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

timators perform similarly, with PENSEM showing, on average, a slightly better performance.

A subset of 6 proteins (marked with asterisk in Table 1 and represented with solid lines in Figure 10) out the 15 selected proteins, were successfully developed and measured with MRM on all 37 discovery samples, as well as 52 new test samples. Thus, to validate the results of PENSEM's protein selection, we train and test a model based on these independent and more precise protein measurements. We use an MM-estimator to train the model based on the 6 available proteins since no additional selection is required at this stage. The MM-estimator is conceptually equivalent to PENSEM when the penalty parameters λ_S and λ_M are set to 0.

The model is trained on the same 37 training plasma samples, except that the protein levels were now measured by MRM instead of iTRAQ. Interestingly, the MM-estimator flags the samples B-381W51, B527W51 and B-561W52 as outlying even when proteins are measured by MRM. Some of the other samples flagged by PENSEM as outliers are diagnosed as borderline outliers by the MM-estimator.

The 52 test samples are from new patients, not involved in any phase of the discovery, so they constitute an independent test set to validate our estimated model. Among these test samples, 12 are flagged as outlying. Since robust estimators are not trained to predict the response of outlying samples we exclude these samples to estimate the performance of our robust estimated model. The predicted response of the remaining 40 test samples is used to classify the disease status of the test patients.

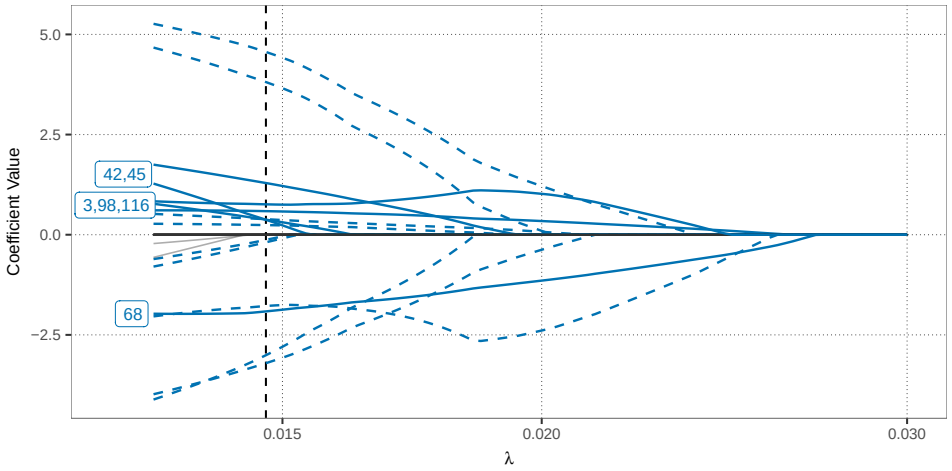

Figure 10: PENSEM's regularization path. The regularization path illustrates how the estimated coefficients shrink at different levels of penalization. The optimal level of penalization λ_M^* is represented by the vertical dashed line. The path of the variables selected at this level of penalization are highlighted in blue. Solid lines are used for the coefficients of the proteins available in the MRM test set. The numbers in the labels correspond to in-house protein IDs.

In clinical practice, a percentage of diameter stenosis below 20 suggests that the patient is not suffering from CAV, and a percentage above 40 is an indication of CAV. To have enough samples in both groups, we use a middle cut-off of 30 to classify patients into CAV and non-CAV based on our predicted percentage of diameter stenosis. Training a model based on 6 out of the 15 proteins selected by PENSEM and using an MM-estimator, we can predict the percentage of diameter stenosis with sufficient accuracy to distinguish CAV from non-CAV test patients achieving an AUC of 0.85.

Overall, results demonstrate the ability of PENSEM to identify promising biomarkers of CAV, some of which could be migrated to a more sensitive and cost-effective platform (MRM) to validate the model in an external cohort of patients, without antibody dependencies. While the migration of proteins is in general a challenging step in a biomarkers pipeline, our model preserves the accuracy in predicting the percentage of diameter stenosis in new test samples. The plasma proteomic biomarkers of CAV selected by PENSEM may offer a relevant post-transplant monitoring tool to effectively guide clinical care. Our robust PENSE and PENSEM estimators provide a reference for a wide range of other biomarker studies and the analysis of other complex datasets.

TABLE 1

Potential biomarkers of CAV identified by PENSEM. A validation Multiple Reaction Monitoring (MRM) assay was developed for the proteins identified with an asterisk. The first column shows an in-house protein ID used to match proteins from different datasets.

Protein ID	Gene Symbol	Protein Name
3	C4B/C4A*	Complement C4-B/C4-A
13	CFB	Complement factor B
30	F2	Prothrombin (Fragment)
42	APOE*	Apolipoprotein E
45	AMBP*	Protein AMBP
46	ECM1	Extracellular matrix protein 1
59	ITIH3	Inter-alpha-trypsin inhibitor heavy chain H3
68	SHBG*	Sex hormone-binding globulin
69	SERPINF1	Pigment epithelium-derived factor
98	PROS1*	Vitamin K-dependent protein S
101	F10	Coagulation factor X
116	APOC2*	Apolipoprotein C-II
139	HBD	Hemoglobin subunit delta
141	LCAT	Phosphatidylcholine-sterol acyltransferase
298	HBA2;HBA1;HBZ	Hemoglobin subunit alpha/zeta

TABLE 2

Mean and standard deviation (SD) of the prediction τ-scales.

	Lasso	EN	PENSE	PENSEM	MMLasso	SparseLTS
Mean	17.20	17.17	17.53	16.99	18.20	18.07
SD	1.58	1.47	1.53	1.30	1.74	1.45

7. Conclusions. In this paper we propose 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 show 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 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 LASSO.

In addition, we propose an efficient algorithm to compute both PENSE and PENSEM that works very well in practice. Computing regression es-

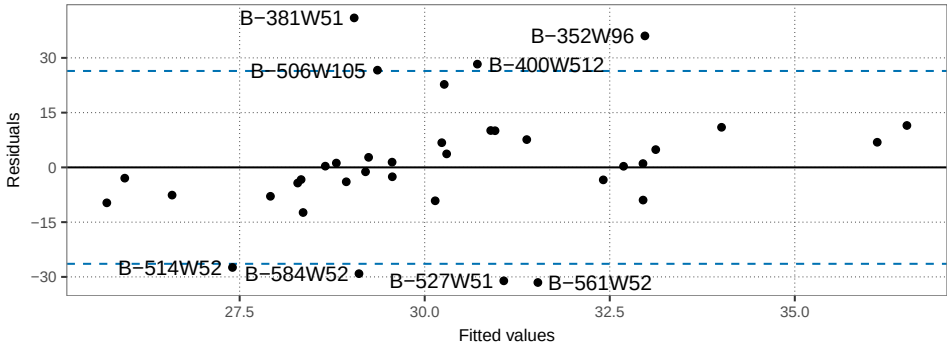

Figure 11: Patients flagged by PENSEM as outlying based on 15 proteins selected using iTRAQ data in the discovery phase. The blue dashed lines represent ± 2 times the robust τ-scale of the residuals.

timators 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 (Peña and Yohai, 1999).

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 use 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 use PENSE and PENSEM to study the association between hundreds of plasma protein levels and a measure of artery obstruction on cardiac transplant recipients. Our robust estimators identify new potentially relevant biomarkers that are not found with non-robust alternatives. Moreover, the analysis based on our robust penalized estimators flags eight patients with suspiciously atypical artery obstruction values. Later mea-

surements with more accurate techniques of the artery obstruction of four of these patients confirm that the original values were inaccurate. Importantly, a model based on most of the proteins selected by PENSEM is validated in a new set of 52 test samples, achieving an AUC of 0.85 when classifying 40 non-outlying samples.

Overall, our robust PENSE and PENSEM estimators, as well as the computational methodologies proposed in this study advance the current knowledge of robust regularized regression estimators and provide flexible and computationally feasible robust estimation for complex and large datasets.

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. Most of the numerical results were generated using GCF's computational infrastructure funded by the Canada Foundation for Innovation (CFI). GCF and MSB were supported by NSERC Discovery grants.

References.

Alfons, A. (2016). robustHD: Robust Methods for High-Dimensional Data R package version 0.5.1.

Alfons, A., Croux, C. and Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics* **7** 226–248.

Clarke, F. H. (1990). *Optimization and Nonsmooth Analysis. Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Cohen Freue, G. V. and Borchers, C. H. (2012). Multiple Reaction Monitoring (MRM). *Circulation: Cardiovascular Genetics* **5** 378.

Domanski, D., Percy, A. J., Yang, J., Chambers, A. G., Hill, J. S., Freue, G. V. C. and Borchers, C. H. (2012). MRM-based multiplexed quantitation of 67 putative cardiovascular disease biomarkers in human plasma. *PROTEOMICS* **12** 1222–1243.

Donoho, D. L. and Huber, P. J. (1982). The notion of breakdown point. In *A Festschrift For Erich L. Lehmann*, (P. J. Bickel, D. K. and J. L. Hodges, eds.). *Wadsworth international statistics* 157–184. CRC Press, Belmont, CA.

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

Fan, J., Li, Q. and Wang, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79** 247–265.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* **32** 928–961.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software, Articles* **33** 1–22.

Khan, J. A., Aelst, S. V. and Zamar, R. H. (2007). Robust Linear Model Selection Based on Least Angle Regression. *Journal of the American Statistical Association* **102** 1289-1299.

Koller, M. and Stahel, W. A. (2017). Nonsingular subsampling for regression S estimators with categorical predictors. *Computational Statistics* **32** 631–646.

Lin, D., Freue, G. C., Hollander, Z., Mancini, G. B. J., Sasaki, M., Mui, A., Wilson-McManus, J., Ignaszewski, A., Imai, C., Meredith, A., Balshaw, R., Ng, R. T., Keown, P. A., McMaster, W. R., Carere, R., Webb, J. G. and McManus, B. M. (2013). Plasma protein biosignatures for detection of cardiac allograft vasculopathy. *The Journal of Heart and Lung Transplantation* **32** 723 - 733.

Maronna, R. A. (2011). Robust Ridge Regression for High-Dimensional Data. *Technometrics* **53** 44-53.

Maronna, R. A., Martin, D. R. and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods. Wiley Series in Probability and Statistics*. Wiley, West Sussex, England.

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

Osborne, M. R., Presnell, B. and Turlach, B. A. (2000). On the LASSO and its Dual. *Journal of Computational and Graphical Statistics* **9** 319-337.

Peña, D. and Yohai, V. (1999). A Fast Procedure for Outlier Diagnostics in Large Regression Problems. *Journal of the American Statistical Association* **94** 434-445.

Rousseeuw, P. J. (1984). Least Median of Squares Regression. *Journal of the American Statistical Association* **79** 871-880.

Rousseeuw, P. and Yohai, V. J. (1984). Robust regression by means of S-estimators. In *Robust and nonlinear time series analysis* 256–272. Springer, Berlin Heidelberg.

Salibián-Barrera, M. and Yohai, V. J. (2006). A Fast Algorithm for S-Regression Estimates. *Journal of Computational and Graphical Statistics* **15** 414-427.

Schmauss, D. and Weis, M. (2008). Cardiac Allograft Vasculopathy. *Circulation* **117** 2131–2141.

Smucler, E. and Yohai, V. J. (2017). Robust and sparse estimators for linear regression models. *Computational Statistics & Data Analysis* **111** 116-130.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58** 267-288.

Tomioka, R., Suzuki, T. and Sugiyama, M. (2011). Super-Linear Convergence of Dual Augmented Lagrangian Algorithm for Sparsity Regularized Estimation. *Journal of Machine Learning Research* **12** 1537–1586.

Yohai, V. J. (1987). High Breakdown-Point and High Efficiency Robust Estimates for Regression. *The Annals of Statistics* **15** 642-656.

Yohai, V. J. and Zamar, R. H. (1988). High Breakdown-Point Estimates of Regression by Means of the Minimization of an Efficient Scale. *Journal of the American Statistical Association* **83** 406-413.

Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **67** 301-320.

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics* **37** 1733–1751.

Department of Statistics
University of British Columbia
3182-2207 Main Mall
Vancouver, British Columbia, V6T 1Z4
Canada
E-mail: gcohen@stat.ubc.ca; d.kepplinger@stat.ubc.ca; matias@stat.ubc.ca; ezequiels.90@gmail.com