The Annals of Applied Statistics
2019, Vol. 13, No. 4, 2065–2090
https://doi.org/10.1214/19-AOAS1269

ROBUST ELASTIC NET ESTIMATORS FOR VARIABLE SELECTION AND IDENTIFICATION OF PROTEOMIC BIOMARKERS

BY GABRIELA V. COHEN FREUE[*,1], DAVID KEPPLINGER[*], MATÍAS SALIBIÁN-BARRERA[*,1] AND EZEQUIEL SMUCLER[*,†]

University of British Columbia[*] and Universidad Torcuato Di Tella[†]

In large-scale quantitative proteomic studies, scientists measure the abundance of thousands of proteins from the human proteome in search of novel biomarkers for a given disease. Penalized regression estimators can be used to identify potential biomarkers among a large set of molecular features measured. Yet, the performance and statistical properties of these estimators depend on the loss and penalty functions used to define them. Motivated by a real plasma proteomic biomarkers study, 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 in the selected model and maintain robustness against possible outliers. We also propose an efficient algorithm to compute our robust penalized estimators and derive a data-driven method to select the penalty term. Our robust penalized estimators have very good robustness properties and are also consistent under certain regularity conditions. Numerical results show that our robust estimators compare favorably to other robust penalized estimators. Using our proposed methodology for the analysis of the proteomics data, we identify new potentially relevant biomarkers of cardiac allograft vasculopathy that are not found with nonrobust alternatives. The selected model is validated in a new set of 52 test samples and achieves an area under the receiver operating characteristic (AUC) of 0.85.

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, 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 a certain disease. The innovation of technical resources available for -omic biomarker studies is well recognized, and the development of statistical and computational methods to analyze the resulting datasets is important for validation and clinical implementation of biomarker discoveries.

Received March 2018; revised February 2019.
[1]Supported by NSERC Discovery Grants.
Key words and phrases. Robust estimation, regularized estimation, penalized estimation, elastic net penalty, proteomics biomarkers.

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 in heart transplant patients. The goal is to identify proteomic biomarkers of cardiac allograft vasculopathy (CAV) which is a major complication suffered by 50% of cardiac transplant recipients beyond the first year after transplantation. Identifying these plasma proteomic biomarkers can result in the development of minimally invasive and clinically useful blood tests to diagnose CAV. Although hundreds of proteins were measured and analyzed in these patients, only a few of them are expected to be associated with artery obstruction, resulting in a sparse regression model.

Penalized regression estimators have been proposed to identify a relatively small subset of explanatory variables among a large number of available covariates (even larger than the number of observations) to obtain good prediction (Tibshirani (1996), Zou and Hastie (2005)). However, most of these estimators use the squared error loss function and are thus extremely sensitive to outliers. Since -omics studies usually contain outlying data points associated, for example, with technical problems in sample preparation or patients with rare molecular profiles, robust penalized estimators are needed to effectively interrogate the rich information contained in the data.

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 the penalization of convex M-estimators (Fan, Li and Wang (2017), Fan and Peng (2004)), which are not resistant to high leverage outliers commonly observed in large datasets. Khan, Van Aelst and Zamar (2007) proposed a robust version of the Least Angle Regression method (Efron et al. (2004)) replacing sample correlations with robust counterparts. Since this method does not solve any optimization problem, it is difficult to understand its robustness and asymptotic properties. More recently, Alfons, Croux and Gelper (2013) proposed SparseLTS (Alfons, Croux and Gelper (2013)), an L_1-regularized version of the Least Trimmed Squares regression estimator (Rousseeuw (1984)), which can only be tuned to be either highly robust or highly efficient under the central model.

To obtain simultaneous robustness and efficiency in a regularized context, Maronna (2011) proposed an MM-estimator with a ridge penalty. Although the proposed MM-Ridge regression estimator has good prediction performance, it does not produce sparse solutions and hence cannot be used for variable selection. To address this issue, Smucler and Yohai (2017) recently proposed a penalized MM-LASSO estimator. However, as shown for the classical LASSO (Tibshirani (2013)), if the design matrix is in general position, the 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 from each group ignoring the relevance of the others.

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 propose to combine robust loss functions with the elastic net penalty (Zou and Hastie (2005)), a linear combination between the L_2-penalty of Ridge and the L_1-penalty of LASSO. The resulting penalized robust regression estimators are not limited by the number of available samples and can select groups of correlated proteins 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. Next, to improve the efficiency of PENSE while maintaining its robustness, we use it to initialize an elastic net penalized M-regression estimator. We call this refined estimator PENSEM. We use our robust estimators to identify potentially relevant proteomic biomarkers of cardiac allograft vasculopathy from a set of $n = 37$ plasma samples from heart transplant patients, collected at one year after transplantation.

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 estimators are 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 Supplementary Material contains more technical details and all proofs (Cohen Freue et al. (2019)).

2. PENSE: A new robust penalized regression estimator. The relationship between molecular features and a disease of interest can be modelled by a linear regression model,

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

where $\mu \in \mathbb{R}$ and $\beta \in \mathbb{R}^p$ are the regression coefficients. In biomarkers discovery studies the response variable, $y_i \in \mathbb{R}$, measures the status of a disease (e.g., stenosis of a coronary artery) for the ith patient, and the set of covariates, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\mathsf{T} \in \mathbb{R}^p$, 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, usually only a few are expected to be associated with a given disease. Thus, we focus on regularized regression estimators to select relevant variables. Since proteomics biomarkers usually function in groups, here we consider the elastic net penalty that tends to keep groups of correlated variables together as they enter or leave the model. Furthermore, -omics studies usually contain outlying observations, typically arising from technical issues in the sample preparation steps or the presence of patients with unusual molecular profiles. To protect the resulting estimator against atypical observations, instead of penalizing the classical squared error loss function, we penalize the square of a robust residual scale estimator previously used to define the S-estimators (Rousseeuw and Yohai (1984)).

Our penalized elastic net S-estimator, PENSE, is defined as the minimizer $(\hat{\mu}^{\mathrm{PS}}, \hat{\beta}^{\mathrm{PS}})$ of

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

where $\sigma(\mu, \beta)$ 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 criteria. In our application, we generate a moderate level of sparsity, aiming to select relevant proteins 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 $\sigma(\mu, \beta)$, given implicitly by the solution of

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

for an even and bounded function ρ and tuning constant $\delta \in (0, 1)$. Both ρ and δ need to be chosen jointly 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 nonconvexity and the lack of differentiability of the elastic net penalty at $\beta = \mathbf{0}$. 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 Clarke (1990) we can derive its generalized gradient,

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

where $P_\alpha(\beta) = \frac{1}{2}(1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1$ is the elastic net penalty, $r_i(\mu, \beta) = y_i - \mu - \mathbf{x}_i^{\mathsf{T}}\beta$ are the residuals and the weights $w_i(\mu, \beta)$ are proportional to $\rho'(\tilde{r}_i(\mu, \beta))/\tilde{r}_i(\mu, \beta)$ where $\tilde{r}_i(\mu, \beta) = r_i(\mu, \beta)/\sigma(\mu, \beta)$.

To find a root of (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 (μ, β). This suggests the following iterative procedure. Given an initial estimate $(\mu^{\text{init}}, \beta^{\text{init}})$ and its corresponding M-scale estimate $\sigma(\mu^{\text{init}}, \beta^{\text{init}})$, obtain an improved set of parameter estimates by computing a weighted elastic net with weights $w_i(\mu^{\text{init}}, \beta^{\text{init}})$ as above. Next, update the weights and iterate. We refer to this algorithm as the iteratively reweighted elastic net (IRWEN) (see the Supplementary Material (Cohen Freue et al. (2019)) 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 it is necessary to find a good starting point for IRWEN.

Initial estimators for the nonpenalized S-estimator have been extensively studied in the literature (e.g., Salibian-Barrera and Yohai (2006), Koller and Stahel (2017)). A commonly used strategy is to construct data-driven random starts by fitting the regression model on randomly chosen subsets of the data. The idea is that subsets without outliers will provide good starting points. To maximize the chance of obtaining a clean starting point, subsets are taken with as few points as possible. However, when the number of variables exceeds the sample size (e.g., $p = 81$ proteins and $n = 37$ patients in our case study), it is not clear how to define the size of the random subsets.

Alfons, Croux and Gelper (2013) proposed to compute a LASSO estimator on random subsets of size 3 to initialize the algorithm for SparseLTS. However, the size of the subset limits the number of variables that LASSO can select for the initial estimator. In our application, using an initial estimator based on only three proteins may result in an undesirably sparse final model with the potential loss of relevant biomarkers.

Instead, we adapt the approaches of Peña and Yohai (1999) and Maronna (2011) and construct clean subsets of our proteomics data by removing outlying observations. These 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 the cleaned subset and used as candidate initial estimators for IRWEN. More details can be found in the Supplementary Material (Cohen Freue et al. (2019)).

The optimal level of penalization for PENSE, λ_S^* in (2), is generally unknown in advance and is chosen from a grid of K possible values based on the prediction performance of the penalized estimator. In our problem, this parameter limits the number of potential biomarkers that we migrate to the validation stage. 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 λ_S in the grid, we use "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.

This strategy is commonly used to compute other penalized estimators (Friedman, Hastie and Tibshirani (2010), Tomioka, Suzuki and Sugiyama (2011)). Ideally, starting the "warm" algorithm with a very large penalty value that shrinks all regression coefficients to zero saves the computation of any other initial estimator. However, since the objective function (2) is not convex, this strategy is no longer guaranteed to find good solutions along 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 Material for more details (Cohen Freue et al. (2019)).

3. PENSEM: A refined estimator. Many applications where regularized estimators are used have relatively few observations. 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 models like (1).

Following Yohai (1987), we refine PENSE to obtain a penalized M-estimator, PENSEM, with higher efficiency (lower variance) and same robustness strength. PENSEM is defined as the minimizer $(\hat{\mu}^{\mathrm{PM}}, \hat{\beta}^{\mathrm{PM}})$ of the penalized loss

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

where the residual scale estimate $\hat{\sigma}_0$ is fixed and ρ_2 is an even and bounded function. As with PENSE, we can use an IRWEN algorithm to find local minima of (5), initialized using the PENSE estimate $(\hat{\mu}^{\mathrm{PS}}, \hat{\beta}^{\mathrm{PS}})$. However, the estimation of the initial residual scale, $\hat{\sigma}_0$, requires 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 $\hat{\sigma}_0 = q\hat{\sigma}(\mu^{\mathrm{PS}}, \beta^{\mathrm{PS}})$. Further details on the correction factor q and other adjustments suggested in the literature are given in the Supplementary Material (Cohen Freue et al. (2019)).

Finally, we need to determine the level of penalization of PENSEM which controls the number of variables selected in the final model. Although both PENSE and PENSEM are defined using the same penalty function, it is impossible to determine what values of λ_S and λ_M give the same level of penalization due to the vastly different scale and shape of the loss functions. Thus, the 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^*. Irrespective of the level of penalization induced by each λ_M in the grid, the scale estimate $\hat{\sigma}_0$ is always based on the PENSE estimate obtained with the optimal penalty level λ_S^* since this is the best estimate of the M-scale of the true residuals available. Similarly, for each λ_M in the grid, we start the numerical optimization of the penalized M-loss function (5) at the optimal PENSE estimate $(\hat{\mu}^{\text{PS}}, \hat{\beta}^{\text{PS}})$. Finding a local optimum of the non-convex PENSEM objective function that is close to the optimal PENSE estimate is in line with our goal of refining the optimal PENSE estimate.

4. Properties. In this Section, we study the 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 against outliers is the (finite-sample) breakdown point (Donoho and Huber (1983)), which is the largest proportion of samples in the data set that could be contaminated arbitrarily and still result in a bounded estimator. The larger this proportion, the "safer" the estimator is, in the sense of not being completely determined by a small number of atypical patients in the training set.

One goal in the cardiac allograft vasculopathy study is the detection of atypical samples in the data. Outliers in regression models can be flagged by considering the residuals from the fit. This approach is expected to work well when the estimated parameters are not affected by the outliers; one is trying to detect, so high breakdown point estimators also provide reliable outlier detection methods. We illustrate this successfully in Section 6 below.

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), $\epsilon^*(\hat{\theta}; \mathbf{Z})$, of an estimator $\hat{\theta}$ is defined as

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

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 (1983)). In the proteomic case study analyzed in this paper, $n = 37$ corresponds to the number of independent plasma samples from cardiac transplant recipients.

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. For example, Alfons, Croux and Gelper (2013) show that the breakdown point of the LASSO is $1/n$. In general, 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 β^* be a minimizer of the penalized sum of squared residuals objective function, $\sum_{i=1}^n (y_i - \mathbf{x}_i^\intercal \beta)^2 + \lambda_0 \|\beta\|_1$, 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 $\|\beta^*\|_1 = C_0 = (\mathbf{r}^*)^\intercal \mathbf{X}\beta^*/\lambda_0$, where $\mathbf{r}^* = (r_1^*, \ldots, r_n^*)^\intercal$ is the vector of residuals for β^*. If β^* is different from the usual least squares estimator, it follows that β^* also minimizes $\sum_{i=1}^n (y_i - \mathbf{x}_i^\intercal \beta)^2$ subject to $\|\beta\|_1 \leq C_0$. Since the bound 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. More specifically, the breakdown point of PENSE is at least $\min(\delta, 1 - \delta)$, where δ is the tuning constant in (3) defining the residual scale M-estimator. Thus, 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 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 $(\hat{\mu}^{\mathrm{PS}}, \hat{\beta}^{\mathrm{PS}})$ satisfies*

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

A proof of the theorem is given in the Supplementary Material (Cohen Freue et al. (2019)). Moreover, the proof in Smucler and Yohai (2017) can be used to show that the breakdown point of PENSEM is at least as high as the breakdown point of the initial scale estimator 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 in Cohen Freue et al. (2019)). Importantly, our result does not require any moment assumptions on the distribution of the errors, and

hence guarantees high-quality estimations with large sample sizes, even in cases with extremely heavy-tailed error distributions. 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. Results of an extensive simulation study to complement the above asymptotic theory in the case $n < p$ are discussed in the next section.

5. Simulation studies. We report here the results of our numerical experiments conducted to further study the properties of our estimators and compare them with other robust and/or penalized estimators.

We consider data following a linear regression model of the form

$$y_i = \mathbf{x}_i^\intercal \beta_0 + \varepsilon_i, \quad \varepsilon_i \sim N\big(0, \sigma^2\big), 1, \ldots, n, \tag{7}$$

with 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 (β_0) (see Section 5.2 below).

We compare our PENSE and PENSEM estimators with the following classical and robust sparsity-inducing estimators: the classical LASSO, the classical EN, SparseLTS and MMLASSO. Ridge estimators are not included since they can not be used for variable selection. Whenever possible, we also include the oracle OLS and MM estimators, which estimate only the coefficients of the true active set of predictors. PENSE and PENSEM are computed using Tukey's Bisquare loss $\rho_c(t) = \min\{1, 1 - (1 - (t/c)^2)^3\}$. Computer code for PENSE and PENSEM is publicly available at https://cran.r-project.org/package=pense. For SparseLTS we use the implementation available in the R package robustHD (Alfons (2016)). The penalty parameter was chosen to optimize its prediction performance estimated by cross-validation. For MMLASSO we use the functions available in the authors' github repository https://github.com/esmucler/mmlasso. The implementation of the MM-LASSO chooses the breakdown point adaptively between 25% and 50%, depending on the estimated degrees of freedom of the initial S-Ridge estimate. Other robust estimators are tuned to achieve a 25% breakdown point.

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). Since the training sample might contain contaminated observations, instead of the usual root mean squared prediction error we use a robust τ-scale (Yohai and Zamar (1988)) estimate of the out-of-sample prediction errors.

Similarly, we compute PENSEM on a grid of 100 logarithmically equispaced values for λ_M, starting from the optimum λ_S^*. The optimal λ_M^* is again chosen by 10-fold CV minimizing the τ-scale of the prediction errors.

The balance between the L_1- and the L_2-penalties as controlled by the parameter $\alpha \in [0, 1]$ is fixed throughout the selection of λ_S and λ_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 with the classical EN estimator (Zou and Hastie (2005)), PENSE and PENSEM suffer from "double" penalization due to the combination of the L_1- and the L_2- penalties and we correct them as $\hat{\beta}\sqrt{1 + 1/2(1 - \alpha^*)\lambda^*}$. The intercept is also adjusted to maintain centered weighted residuals.

5.2. *Simulation scenarios.* To demonstrate the benefits of the EN over the L_1 penalty, we include two scenarios from Zou and Hastie (2005) and Zou and Zhang (2009). In these scenarios the correlation among the predictors with nonzero regression coefficient is moderate to high. We also modify the scenario in Zou and Hastie (2005) to a more challenging one with more active predictors than observations. Finally, we consider a very sparse scenario with no correlation among the active predictors, which may favor L_1-penalized estimators. Using the notation in (7), the four scenarios are:

(1) Example (d) in Zou and Hastie (2005): $p = 40$, $n = 50$; $\sigma = 15$ and

$$\beta_0 = (\underbrace{3, \ldots, 3}_{15}, \underbrace{0, \ldots, 0}_{25})^\intercal.$$

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

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

for $j = 1, \ldots, 15$, $l = 1, 2, 3$. The remaining 25 predictors are i.i.d. standard Normal.

(2) Same as (1) except $p = 400 \gg n = 50$, and

$$\beta_0 = (\underbrace{3, \ldots, 3}_{60}, \underbrace{0, \ldots, 0}_{340})^\intercal.$$

The latent variable model is based on three factors, and each factor is associated with 20 predictors, that is,

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

for $j = 1, \ldots, 60$, $l = 1, 2, 3$. The other 340 predictors are i.i.d. $N(0, 1)$.

(3) Example 2 in Zou and Zhang (2009), with $n = 100$, $p = \lfloor 4n^{2/3} \rfloor - 5 = 81$, $\sigma = 6$, and

$$\beta_0 = (\underbrace{3, \ldots, 3}_{27}, \underbrace{0, \ldots, 0}_{54})^\intercal.$$

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

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

(4) In this scenario, $p = 995 \gg n = 100$, and $\sigma = 1$. Of the 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 coefficients are equally spaced at $j = 1, 72, \ldots, 995$:

$$\beta_0 = \sqrt{4}(\gamma_1, \underbrace{0, \ldots, 0}_{71}, \gamma_2, 0, \ldots, 0, \gamma_{14}, \underbrace{0, \ldots, 0}_{71}, \gamma_{15})^\intercal.$$

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

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

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

Scenarios (1) and (2) are closely related to the biomarkers discovery study in Section 6 since the sample size is comparable to the number of patients in the study (37) and the strong grouping structure mirrors the correlation typically found in proteomics studies. Furthermore, the number of available proteins (81) is between the number of predictors considered in the first two scenarios. The performance of the estimators in scenarios (1) and (2) is thus indicative of their performance in our biomarkers discover study.

We study the robustness of the estimators by contaminating the first $m = \lfloor \epsilon n \rfloor$ observations $(\mathbf{x}_i, y_i)$ as in Maronna (2011). Leverage points are introduced by replacing the predictors $\mathbf{x}_i$ with

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

where $\eta_i \sim N_p(\mathbf{0}, 0.1^2 \mathbf{I}_p)$ and $\mathbf{a} = \tilde{\mathbf{a}} - \frac{1}{p} \tilde{\mathbf{a}}^\intercal \mathbf{1}_p$ with the entries $\tilde{a}_j$ of $\tilde{\mathbf{a}}$ following a $U(-1, 1)$ distribution, $j = 1, \ldots, p$. The parameter k_{lev} controls the distance in the direction most influential for the estimator.

We also contaminate the observations in the response by altering the regression coefficient

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

The parameters k_{lev} and k_{slo} control the position of the contaminated observations. Preliminary analysis showed that the effect on all considered estimators was almost the same for any $k_{\text{lev}} > 1$, hence we fixed $k_{\text{lev}} = 2$. The position of the vertical outliers affects more of the estimators, and we consider a grid of 15 logarithmically spaced values for k_{slo} between 1 and 500.

To measure prediction performance of the estimators, we generated a clean validation set for each scenario, $(\mathbf{x}_i^*, y_i^*)$, $i = 1, \ldots, n^*$, $n^* = 1000$, and computed the root mean squared prediction error (RMSPE) of $(\hat{\mu}, \hat{\beta})$ as

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

Model selection performance was assessed with the sensitivity (SENS) and specificity (SPEC) of $\hat{\beta}$:

$$\text{SENS} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\#\{j : \beta_{0j} \neq 0 \wedge \hat{\beta}_j \neq 0\}}{\#\{j : \beta_{0j} \neq 0\}},$$

$$\text{SPEC} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{\#\{j : \beta_{0j} = 0 \wedge \hat{\beta}_j = 0\}}{\#\{j : \beta_{0j} = 0\}},$$

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

Sensitivity measures the proportion of active predictors detected by the estimator, while specificity is the proportion of noise predictors correctly omitted from the final model. Ideally, both measures should be close to 1, but since none of the estimators in our study is variable selection consistent in these scenarios, the exact true model is rarely chosen by any of them. In the context of the biomarkers discovery study, it is important to achieve a high sensitivity to ensure none of the important proteins are missed and to simultaneously keep a reasonably high specificity to control the cost of future validation experiments.

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, $k_{\text{slo}}^{(l)}$, $l = 1, \ldots, 15$, by the area under the curve of RMSPE values, $\text{RMSPE}_{\text{cont}}$. Let's denote the estimate at $k_{\text{slo}}^{(l)}$ by $(\hat{\mu}^{(l)}, \hat{\beta}^{(l)})$, then the overall RMSPE under contamination is

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

As an example, Figure 1 shows the curve of RMSPE over k_{slo} from one replication of scenario (1) 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

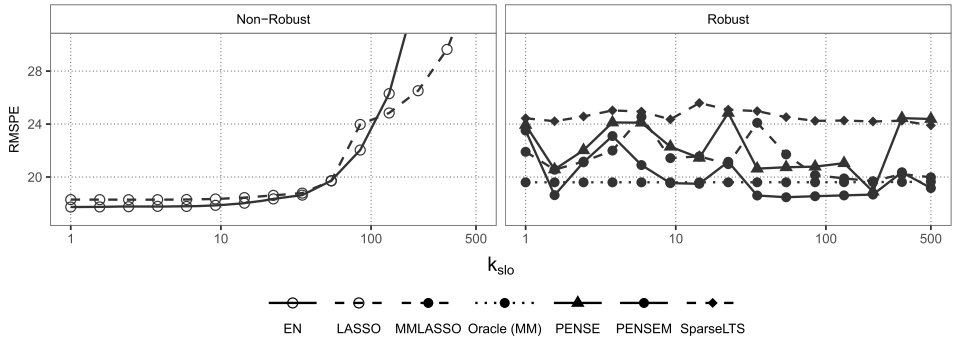

FIG. 1. *Root mean squared prediction error of different estimators over a grid of* k_{slo} *values ranging from* 1 *to* 500 *with* 10% *contamination under scenario* (1).

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 scenario.

For each scenario we compute PENSE(M) and the classical EN for several values of α. To save space, we present results only for the PENSE(M) estimators corresponding to the α^* with the smallest average cross-validated $\text{RMSPE}_{\text{cont}}$ and the classical EN with smallest average cross-validated RMSPE on the uncontaminated training data.

5.3. *Simulation results*. *Scenario (1)*: The prediction performance measures of PENSE(M) and those of the competing estimators over 200 replications for Scenario (1) 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 scenario, 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.

Scenario (1) 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

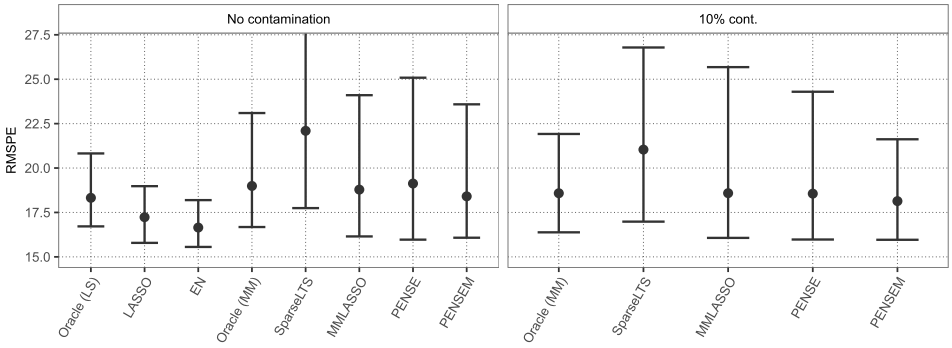

FIG. 2. *Average prediction performance of different estimators in scenario* (1). *The error bars extend from the* 5% *to the* 95% *quantile. For the uncontaminated case we report the RMSPE. For training data with* 10% *contamination, we show the overall measure* $\text{RMSPE}_{\text{cont}}$ *over a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

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 elastic net penalty in this scenario 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. Although this

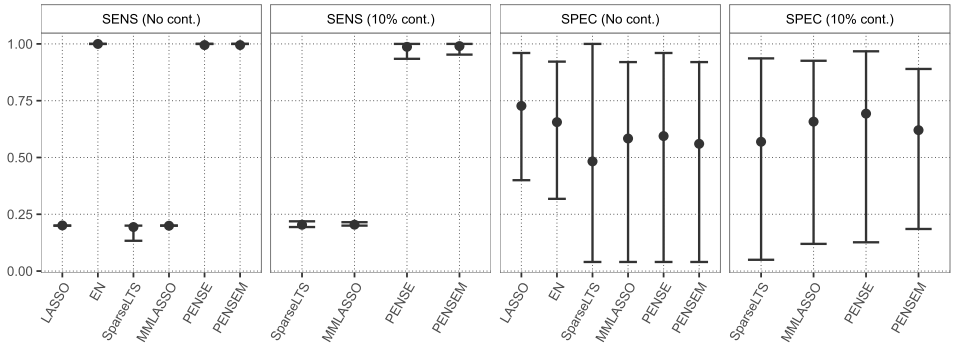

FIG. 3. *Average specificity and sensitivity of different estimators in scenario* (1). *The error bars extend from the* 5% *to the* 95% *quantile. For training data with* 10% *contamination, we show the area under the curve* ($\text{SENS}_{\text{cont}}$ *and* $\text{SPEC}_{\text{cont}}$) *over a grid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

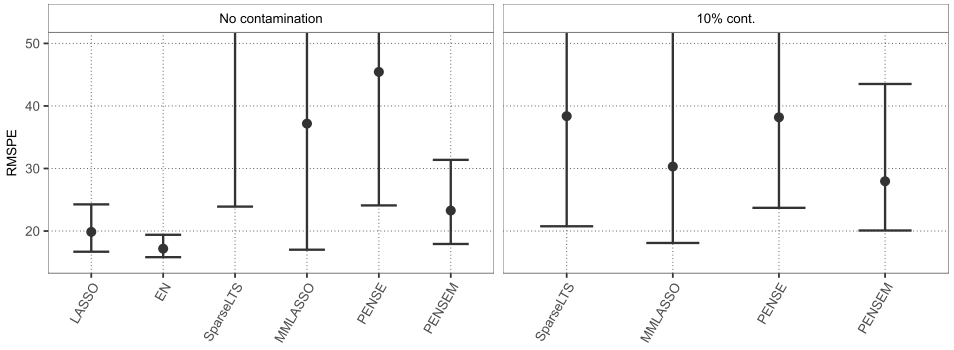

FIG. 4. *Average prediction performance of different estimators in scenario* (2). *The error bars extend from the* 5% *to the* 95% *quantile. For the uncontaminated case we report the RMSPE. For training data with* 10% *contamination, we show the overall measure* $RMSPE_{\text{cont}}$ *over a grid of* 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 scenario because there are more active predictors than observations. Classical LASSO and EN are omitted from the right panel to avoid distortion of the plotting scale and hamper comparisons.*

may result in an incremented cost of the validation phase for a protoemics study, in practice not all proteins can be successfully migrated due to their chemical properties. Thus, identifying a group of correlated proteins also increases the options to build a strong validation assay.

Scenario (2): In this scenario, the difference between LASSO and EN estimators is even more pronounced as shown in Figure 4.

Additionally, this scenario is similar to the biomarker study in terms of sample size, $n < p$, the grouped correlation structure and potential contamination. We expect the performance of PENSE and PENSEM to be indicative of their performance in the next section.

In addition, the oracle estimates cannot be computed since the number of active predictors is larger than the sample size. The classical EN as well as PENSE both achieve the best cross-validated prediction performance for $\alpha^* = 0.9$, reflecting the sparsity of this scenario. PENSEM shows again the best prediction performance of the robust estimators. 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 select only a single predictor from each group. In contrast to the previous scenario, however, PENSE(M) and classical EN have a higher specificity in this scenario than in scenario (1) due to the large number of irrelevant predictors. Under contamination PENSE selects all 60 active predictors 88% of the time and, on average, selects only 23 of the 340 irrelevant predictors. Without contamination the model selection of PENSE is on average even outperforming the classical EN.

From these results we are confident that PENSE and PENSEM are very well suited to unmask potential biomarkers for cardiac allograft vasculopathy in our

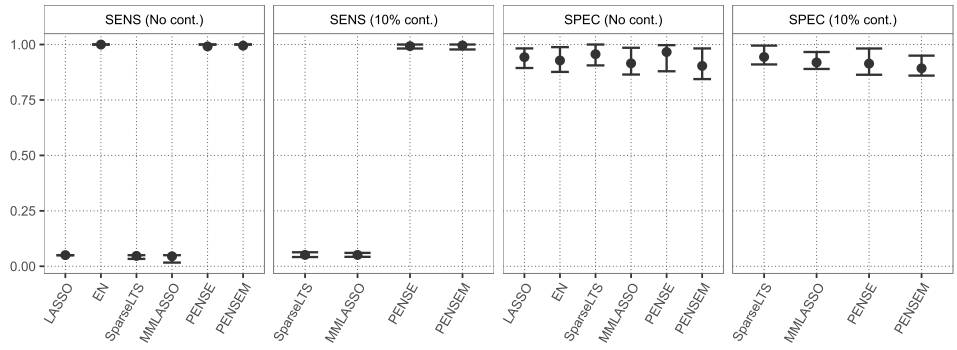

FIG. 5. *Average specificity and sensitivity of different estimators in scenario* (2). *The error bars extend from the* 5% *to the* 95% *quantile. For training data with* 10% *contamination, we show the area under the curve* ($\text{SENS}_{\text{cont}}$ *and* $\text{SPEC}_{\text{cont}}$) *over a grid of* k_{slo} *from* 1 *to* 500. *Classical EN and PENSE*(*M*) *are both using* $\alpha^* = 0.9$. *Classical LASSO and EN are omitted from the panels with contamination to avoid distortion of the plotting scale and hamper comparisons.*

application. Especially, the variable selection performance shows that our methodology keeps the risk of missing important proteins very low while simultaneously maintaining the cost of migrating unnecessary proteins low.

Scenario (3): This is the last scenario 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 sce-

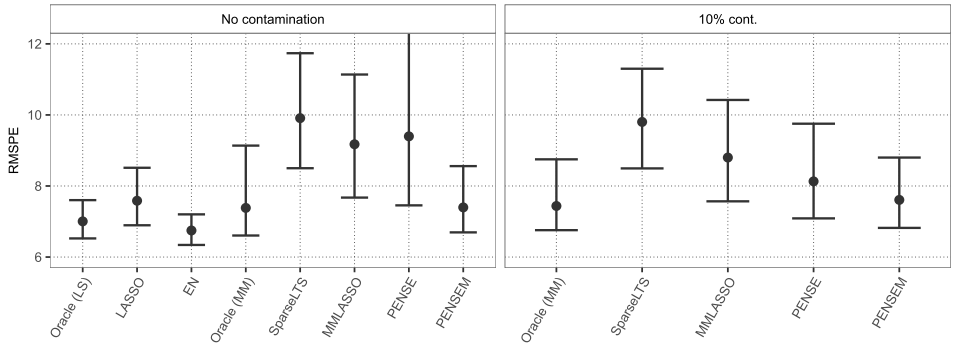

FIG. 6. *Average prediction performance of different estimators in scenario* (3). *The error bars extend from the* 5% *to the* 95% *quantile. For the uncontaminated case we report the RMSPE. For training data with* 10% *contamination, we show the overall measure* $\text{RMSPE}_{\text{cont}}$ *over a grid of* k_{slo} *from* 1 *to* 500. *Classical EN and PENSE*(*M*) *are both using* $\alpha^* = 0.7$. *Classical LASSO and EN are omitted from the right panel to avoid distortion of the plotting scale and hamper comparisons.*

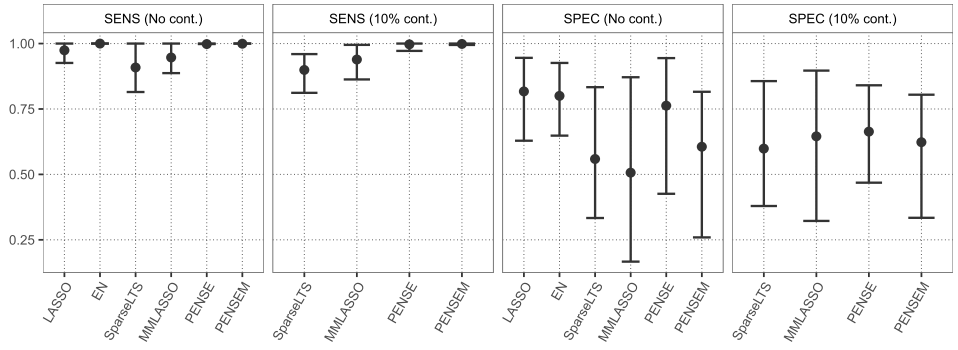

FIG. 7. *Average specificity and sensitivity of different estimators in scenario* (3). *The error bars extend from the* 5% *to the* 95% *quantile. For training data with* 10% *contamination, we show the area under the curve* ($\text{SENS}_{\text{cont}}$ *and* $\text{SPEC}_{\text{cont}}$) *over a grid of* k_{slo} *from* 1 *to* 500. *Classical EN and PENSE*(*M*) *are both using* $\alpha^* = 0.7$. *Classical LASSO and EN are omitted from the panels with contamination to avoid distortion of the plotting scale and hamper comparisons.*

nario. 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.

Scenario (4): The results of this very sparse scenario are shown in Figure 8. Not surprisingly, the best CV performance for PENSE(M) is achieved with an L_1-

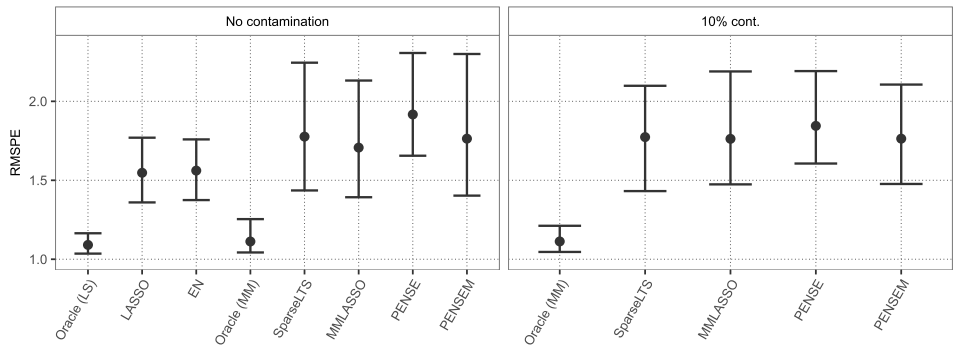

FIG. 8. *Average prediction performance of different estimators in scenario* (4). *The error bars extend from the* 5% *to the* 95% *quantile. For the uncontaminated case we report the RMSPE. For training data with* 10% *contamination, we show the overall measure* $\text{RMSPE}_{\text{cont}}$ *over a grid of* k_{slo} *from* 1 *to* 500. *Classical EN uses* $\alpha^* = 0.9$ *while PENSE*(*M*) *is fitted with* $\alpha^* = 1$. *Classical LASSO and EN are omitted from the right panel to avoid distortion of the plotting scale and hamper comparisons.*

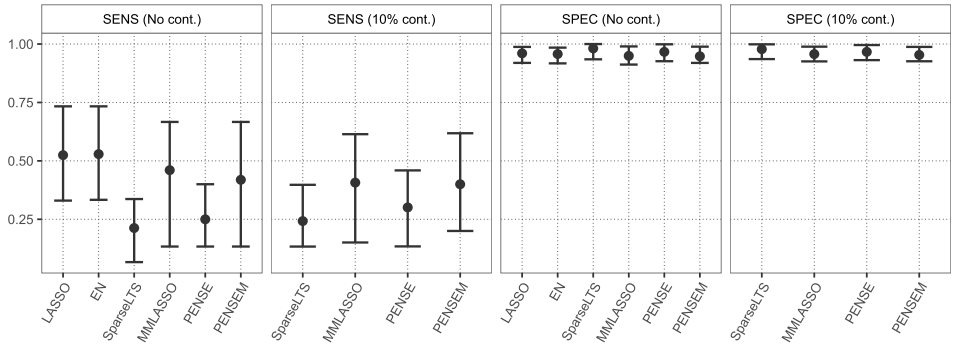

FIG. 9. *Average specificity and sensitivity of different estimators in scenario* (4). *The error bars extend from the* 5% *to the* 95% *quantile. For training data with* 10% *contamination, we show the area under the curve* ($\text{SENS}_{\text{cont}}$ *and* $\text{SPEC}_{\text{cont}}$) *over a grid of* k_{slo} *from* 1 *to* 500. *Classical EN uses* $\alpha^* = 0.9$ *while PENSE*(*M*) *is fitted with* $\alpha^* = 1$. *Classical LASSO and EN are omitted from the panels with contamination to avoid distortion of the plotting scale and hamper comparisons.*

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 similar 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. MM-LASSO 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 scenarios and clearly outperforms $L1$-based estimates if important predictors are correlated. Especially in scenarios with large number of relevant correlated covariates, 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 allograft 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 one 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 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 modeling 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 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 λ_M 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 additional proteins of the coagulation and complement cascades (F10 and CFB, respectively), another apolipoprotein (APOC2) and new homoglobin subunits (HBD, HBA, HBZ),

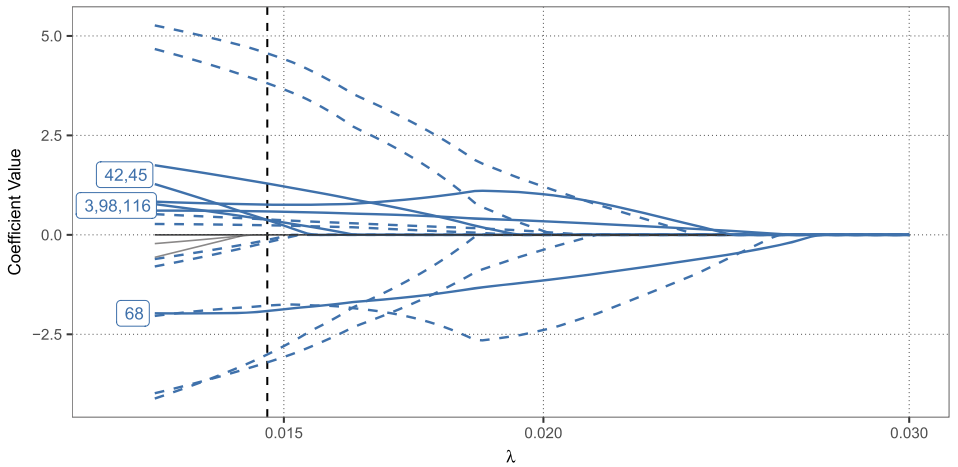

FIG. 10. *PENSEM's regularization path. The regularization path illustrates how the estimated coefficients shrink at different levels of penalization. The optimal level of penalization λ_M^* is represented by the vertical dashed line. The path of the variables selected at this level of penalization are 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.*

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

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

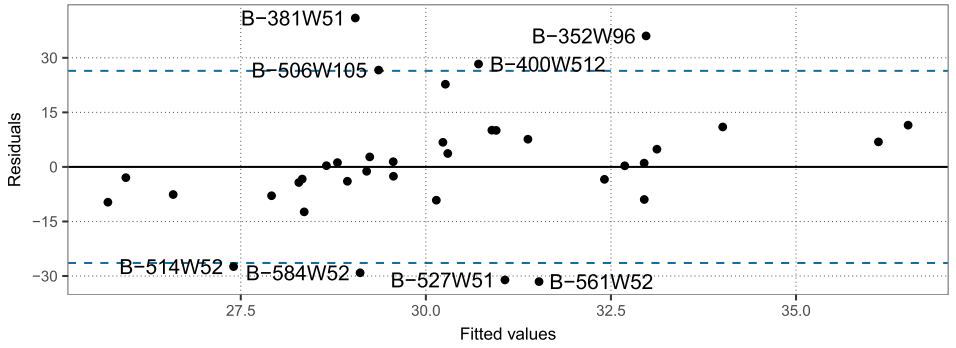

FIG. 11. *Patients flagged by PENSEM as outlying based on* 15 *proteins selected using iTRAQ data in the discovery phase. The blue dashed lines represent* ± 2 *times the robust* τ*-scale of the residuals.*

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

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-fold 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 elastic net estimators. In terms of prediction, all estimators perform similarly, with PENSEM showing, on average, a slightly better performance.

A subset of six proteins (marked with asterisk in Table 1 and represented with solid lines in Figure 10) out of 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 six 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.

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 nonCAV based on our predicted percentage of diameter stenosis. Training a model based on six 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 nonCAV 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 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 protein 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 biomarkers studies and complex datasets.

7. Conclusions. In this paper, we propose regularized robust estimators with an elastic net penalty, which we call PENSE and PENSEM. The former is a penalized S-estimator, while the latter corresponds to a penalized high-breakdown M-estimator to increase the efficiency of the parameter estimates.

We show that these estimators retain the robustness properties of their unpenalized counterparts (high breakdown point and consistency) which was essential in our biomarkers study since the data contain some outlying data points. 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 practice, this property enables the identification of potentially correlated, but equally relevant, biomarkers.

In addition, we propose an efficient algorithm to compute both PENSE and PENSEM that works very well in practice. Computing regression estimators with good robustness properties is computationally very costly because their loss functions are necessarily nonconvex. Moreover, the presence of a nondifferentiable 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)). Our algorithm was designed to work effectively in a variety of datasets including those with more variables than observations as that of our proteomics study.

In practice, an important step of the computation of penalized estimators is choosing an "optimal" penalty level. 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 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 nonrobust alternatives. Moreover, our robust penalized estimators flag eight patients with suspiciously atypical artery obstruction values. Measurements with more accurate techniques for four of these patients confirm that the original values of obstruction 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 nonoutlying samples.

Overall, our robust PENSE and PENSEM estimators and the algorithms to compute them 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). Part of this work was conducted while ES was a postdoctoral research fellow in the departments of Statistics and of Computer Science at The University of British Columbia.

SUPPLEMENTARY MATERIAL

Supplementary material for "Robust elastic net estimators for variable selection and identification of proteomic biomarkers". (DOI: 10.1214/19-AOAS1269SUPP; .pdf). We provide additional details on PENSE algorithm, properties and mathematical proofs.

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. *Ann. Appl. Stat.* **7** 226–248. MR3086417

CLARKE, F. H. (1990). *Optimization and Nonsmooth Analysis*, 2nd ed. *Classics in Applied Mathematics* **5**. SIAM, Philadelphia, PA. MR1058436

COHEN FREUE, G. V. and BORCHERS, C. H. (2012). Multiple Reaction Monitoring (MRM). *Circ. Cardiovasc. Genet.* **5** 378.

COHEN FREUE, G. V, KEPPLINGER, D., SALIBIÁN-BARRERA, M. and SMUCLER, E. (2019). Supplement to "Robust elastic net estimators for variable selection and identification of proteomic biomarkers." DOI:10.1214/19-AOAS1269SUPP.

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. and HUBER, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann*. *Wadsworth Statist./Probab. Ser.* 157–184. Wadsworth, Belmont, CA. MR0689745

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166

FAN, J., LI, Q. and WANG, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 247–265. MR3597972

FAN, J. and PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** 928–961. MR2065194

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

KHAN, J. A., VAN AELST, S. and ZAMAR, R. H. (2007). Robust linear model selection based on least angle regression. *J. Amer. Statist. Assoc.* **102** 1289–1299. MR2412550

KOLLER, M. and STAHEL, W. A. (2017). Nonsingular subsampling for regression S estimators with categorical predictors. *Comput. Statist.* **32** 631–646. MR3656977

LIN, D., FREUE, G. C., HOLLANDER, Z., MANCINI, G. B. J., SASAKI, M., MUI, A., WILSON-MCMANUS, J., IGNASZEWSKI, A., IMAI, C. et al. (2013). Plasma protein biosignatures for detection of cardiac allograft vasculopathy. *J. Heart Lung Transplant.* **32** 723–733.

MARONNA, R. A. (2011). Robust ridge regression for high-dimensional data. *Technometrics* **53** 44–53. MR2791951

MARONNA, R. A., MARTIN, R. D. and YOHAI, V. J. (2006). *Robust Statistics*. *Wiley Series in Probability and Statistics*: *Theory and Methods*. Wiley, Chichester. MR2238141

MARONNA, R. A. and YOHAI, V. J. (2010). Correcting MM estimates for "fat" data sets. *Comput. Statist. Data Anal.* **54** 3168–3173. MR2727743

OSBORNE, M. R., PRESNELL, B. and TURLACH, B. A. (2000). On the LASSO and its dual. *J. Comput. Graph. Statist.* **9** 319–337. MR1822089

PEÑA, D. and YOHAI, V. (1999). A fast procedure for outlier diagnostics in large regression problems. *J. Amer. Statist. Assoc.* **94** 434–445. MR1702315

ROUSSEEUW, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.* **79** 871–880. MR0770281

ROUSSEEUW, P. and YOHAI, V. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis* (*Heidelberg*, 1983). *Lect. Notes Stat.* **26** 256–272. Springer, New York. MR0786313

SALIBIAN-BARRERA, M. and YOHAI, V. J. (2006). A fast algorithm for S-regression estimates. *J. Comput. Graph. Statist.* **15** 414–427. MR2246273

SCHMAUSS, D. and WEIS, M. (2008). Cardiac allograft vasculopathy. *Circ.* **117** 2131–2141.

SMUCLER, E. and YOHAI, V. J. (2017). Robust and sparse estimators for linear regression models. *Comput. Statist. Data Anal.* **111** 116–130. MR3630222

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TIBSHIRANI, R. J. (2013). The lasso problem and uniqueness. *Electron. J. Stat.* **7** 1456–1490. MR3066375

TOMIOKA, R., SUZUKI, T. and SUGIYAMA, M. (2011). Super-linear convergence of dual augmented Lagrangian algorithm for sparsity regularized estimation. *J. Mach. Learn. Res.* **12** 1537–1586. MR2813147

YOHAI, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Statist.* **15** 642–656. MR0888431

YOHAI, V. J. and ZAMAR, R. H. (1988). High breakdown-point estimates of regression by means of the minimization of an efficient scale. *J. Amer. Statist. Assoc.* **83** 406–413. MR0971366

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327

ZOU, H. and ZHANG, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Ann. Statist.* **37** 1733–1751. MR2533470

G. V. COHEN FREUE
D. KEPPLINGER
M. SALIBIÁN-BARRERA
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

E. SMUCLER
DEPARTMENT OF MATHEMATICS AND STATISTICS
UNIVERSIDAD TORCUATO DITELLA
AVENIDA FIGUEROA ALCORTA 7350
BUENOS AIRES 1428
ARGENTINA
E-MAIL: esmucler@utdt.edu