

Regularization of Case-Specific Parameters for Robustness and Efficiency

Yoonkyung Lee, *The Ohio State University*
Steven N. MacEachern, *The Ohio State University*
Yoonsuh Jung, *The Ohio State University*

Technical Report No. 799

July, 2007

**Department of Statistics
The Ohio State University
1958 Neil Avenue
Columbus, OH 43210-1247**

Regularization of Case-Specific Parameters for Robustness and Efficiency

Yoonkyung Lee, Steven N. MacEachern, and Yoonsuh Jung

Department of Statistics, The Ohio State University

Columbus, Ohio 43210

yklee@stat.osu.edu, snm@stat.osu.edu, and yoons@stat.osu.edu

Abstract

Regularization methods allow one to handle a variety of inferential problems where there are more covariates than cases. This allows one to consider a potentially enormous number of covariates for a problem. We exploit the power of these techniques, supersaturating models by augmenting the “natural” covariates in the problem with an additional indicator for each case in the data set. We attach a penalty term for these case-specific indicators which is designed to produce a desired effect. For regression methods with squared error loss, an ℓ_1 penalty produces a regression which is robust to outliers and high leverage cases; for quantile regression methods, an ℓ_2 penalty decreases the variance of the fit enough to overcome an increase in bias. The paradigm thus allows us to robustify procedures which lack robustness and to increase the efficiency of procedures which are robust.

We provide a general framework for the inclusion of case-specific parameters in regularization problems, describing the impact on the effective loss for a variety of regression and classification problems. We outline a computational strategy by which existing software can be modified to solve the augmented regularization problem, providing conditions under which such modification will converge to the optimum solution. We illustrate the benefits of including case-specific parameters in the context of mean regression and median regression through simulation and analysis of a linguistic data set.

KEYWORDS: Case indicator; Large margin classifier; LASSO; Leverage point; Outlier; Penalized method; Quantile regression

1 Introduction

A core part of regression analysis involves the examination and handling of individual cases (Weisberg; 2005). Traditionally, cases have been removed or downweighted as outliers or because they exert an overly large influence on the fitted regression surface.

The mechanism by which they are downweighted or removed is through inclusion of case-specific indicator variables. For a least-squares fit, inclusion of a case-specific indicator in the model is equivalent to removing the case from the data set; for a normal-theory, Bayesian regression analysis, inclusion of a case-specific indicator with an appropriate prior distribution is equivalent to inflating the variance of the case and hence downweighting it. The tradition in robust regression is to handle the case-specific decisions automatically, most often by downweighting outliers according to an iterative procedure (Huber; 1981).

This idea of introducing case-specific indicators also applies naturally in the context of regularized regression. Model selection criteria such as AIC or BIC take aim at choosing a model by attaching a penalty for each additional parameter in the model. These criteria can be applied directly to a larger space of models—namely those in which the covariates are augmented by a set of case indicators, one for each case in the data set. When considering inclusion of a case indicator for a large outlier, the criterion will judge the trade-off between the empirical risk (here, negative log-likelihood) and model complexity (here, number of parameters) as favoring the more complex model. It will include the case indicator in the model, and, with a least-squares fit, effectively remove the case from the data set. A more considered approach would allow differential penalties for case-specific indicators and “real” covariates. With adjustment, one can essentially recover the familiar t-tests for outliers (e.g. Weisberg (2005)), either controlling the error rate at the level of the individual test or controlling the Bonferroni bound on the familywise error rate.

Case-specific indicators can also be used in conjunction with more recent regularization methods such as the LASSO (Tibshirani; 1996). Again, care must be taken with details of their inclusion. If these new covariates are treated in the same fashion as the other covariates in the problem, one is making an implicit judgement that they should be penalized in the same fashion. Alternatively, one can allow a second parameter that governs the severity of the penalty for the indicators. This penalty can be set with a view of achieving robustness in the analysis, and it allows one to tap into a large, extant body of knowledge about robustness (Huber; 1981).

With regression often serving as a motivating theme, a host of methods for regularized model selection and estimation problems have been developed. These methods range broadly across the field of statistics. In addition to traditional normal-theory linear regression, we find many methods motivated by a loss which is composed of a negative log-likelihood and a penalty for model complexity. Among these regularization methods are penalized linear regression methods (e.g. ridge regression (Hoerl and Kennard; 1970) and the LASSO), regression with a nonparametric mean function, (e.g. smoothing splines (Wahba; 1990) and generalized additive models (Hastie and Tibshirani; 1990)), and extension to regression with non-normal error distributions, namely, generalized linear models (McCullagh and Nelder; 1989). In all of these cases, we envision adding case-specific indicators along with an appropriate penalty in order to yield an automated, robust analysis. It should be noted that, in addition

to a different severity for the penalty term, the case-specific indicators sometimes require a different form for their penalty term. One can also add these indicators for inference procedures which are not regularized.

A second class of procedures open to regularization are those motivated by minimization of an empirical risk function. While the risk function may be a negative log-likelihood, it may also be of very different form. Quantile regression (whether linear or nonlinear) falls into this category, as do modern classification techniques such as the support vector machine (Vapnik; 1998) and the psi-learner (Shen et al.; 2003). Many of these procedures are designed with the robustness of the analysis in mind, often operating on an estimand defined to be the population-level minimizer of the empirical risk. The procedures are consistent across a wide variety of data-generating mechanisms and hence are asymptotically robust. They have little need of further robustification. Instead, scope for bettering these procedures lies in improving their finite sample properties. Interestingly, the finite sample performance of many procedures in this class can be improved by including case-specific indicators in the problem, along with an appropriate penalty term for them.

In this paper, we investigate the use of case-specific indicators in regularized problems. Section 2 provides a more concrete description of the problem than has been given here, along with a computational algorithm and conditions that ensure the algorithm will obtain the global solution to the regularized problem. Section 3 explains the methodology for a selection of regression methods, motivating particular forms for the penalty terms. Section 4 describes how the methodology applies to several classification schemes. Section 5 gives details of computational implementation for a robust version of the LASSO. Section 6 contains a simulation study; Section 7 a worked example. We discuss implications of the work and potential extensions in Section 8.

2 Robust and Efficient Modeling Procedures

Suppose that we have n pairs of observations denoted by (x_i, y_i) , $i = 1, \dots, n$, for statistical modeling and prediction. Here $x_i = (x_{i1}, \dots, x_{ip})^\top$ with p covariates and the y_i 's are responses. As in the standard setting of regression and classification, the y_i 's are assumed to be conditionally independent given the x_i 's. In this paper, we take modeling of the data as a procedure of finding a functional relationship between x_i and y_i , $f(x; \beta)$ with unknown parameters $\beta \in \mathbb{R}^p$ that is consistent with the data. The discrepancy or lack of fit of f is measured by a loss function $\mathcal{L}(y, f(x; \beta))$. Consider a modeling procedure, say, \mathcal{M} of finding f which minimizes the empirical risk

$$R_n(f) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}(y_i, f(x_i; \beta))$$

or its penalized version

$$R_n(f) + \lambda J(f) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}(y_i, f(x_i; \beta)) + \lambda J(f),$$

where λ is a positive penalty parameter for balancing the data fit and the model complexity of f measured by $J(f)$. A variety of common modeling procedures are subsumed under this formulation, including ordinary linear regression, generalized linear models, nonparametric regression, and supervised learning techniques. For brevity of exposition, we identify f with β and view $J(f)$ as a functional depending on β .

Motivated by the LASSO type ℓ_1 norm regularization, we propose a general scheme to modify the modeling procedure \mathcal{M} . First, we introduce case-specific parameters, $\gamma = (\gamma_1, \dots, \gamma_n)^\top$, for the n observations and modify \mathcal{M} to be the procedure of finding the original model parameters, β , together with the case-specific parameters, γ , that minimize

$$L(\beta, \gamma) = \sum_{i=1}^n \mathcal{L}(y_i, f(x_i; \beta) + \gamma_i) + \lambda_\beta J(f) + \lambda_\gamma J_2(\gamma). \quad (1)$$

If λ_β is zero, \mathcal{M} is empirical risk minimization, otherwise it is penalized risk minimization. In general, $J_2(\gamma)$ measures the size of γ . When concerned with robustness, we often take $J_2(\gamma) = \|\gamma\|_1 = \sum_{i=1}^n |\gamma_i|$. A rationale for this choice is that with added flexibility, the case-specific parameters can curb the undesirable influence of individual cases on the fitted model. Such a case specific adjustment of the model would be necessary only for a small number of potential outliers, and the ℓ_1 norm which yields sparsity works to that effect. When concerned with efficiency, we often take $J_2(\gamma) = \|\gamma\|_2^2 = \sum_{i=1}^n \gamma_i^2$. This choice has the effect of increasing the impact of selected, non-outlying cases on the analysis.

In subsequent sections, we will take a few standard statistical methods for regression and classification and illustrate how this general scheme applies. For each method, particular attention will be paid to the form of adjustment to the loss function for the penalized case-specific parameters.

2.1 Algorithm for Finding Solutions

Although the computational details for obtaining the solution to (1) are specific to each modeling procedure \mathcal{M} , it is feasible to describe a common computational strategy which is effective for a wide range of procedures. For fixed λ_β and λ_γ , the solution pair of $\hat{\beta}$ and $\hat{\gamma}$ to the modified \mathcal{M} can be found with little extra computational cost. A generic algorithm below alternates estimation of β and γ . Given $\hat{\gamma}$, minimization of $L(\beta, \hat{\gamma})$ is done via the original modeling procedure \mathcal{M} . Take $J_2(\gamma) = \|\gamma\|_1$ as an example here. Fixing $\hat{\beta}$, we seek to minimize $L(\hat{\beta}, \gamma)$, which decouples to a minimization of $\mathcal{L}(y_i, f(x_i; \hat{\beta}) + \gamma_i) + \lambda_\gamma |\gamma_i|$ for each γ_i . In most cases, an explicit form of the

minimizer $\hat{\gamma}$ of $L(\hat{\beta}, \gamma)$ can be obtained. This adjustment is equivalent to changing the loss from $\mathcal{L}(y, f(x; \beta))$ to $\mathcal{L}(y, f(x; \beta) + \hat{\gamma})$, which we call the γ -adjusted loss of \mathcal{L} . Alternatively, one may view

$$\mathcal{L}_{\lambda_\gamma}(y, f(x; \beta)) := \min_{\gamma \in \mathbb{R}} \{\mathcal{L}(y, f(x; \beta) + \gamma) + \lambda_\gamma |\gamma|\} = \mathcal{L}(y, f(x; \beta) + \hat{\gamma}) + \lambda_\gamma |\hat{\gamma}|$$

as an “effective loss”. Concrete examples of the adjustments will be given in the next sections.

These considerations lead to the following iterative algorithm for finding $\hat{\beta}$ and $\hat{\gamma}$.

1. Initialize $\hat{\gamma}^{(0)} = 0$ and $\hat{\beta}^{(0)} = \arg \min_{\beta} L(\beta, 0)$ (the ordinary \mathcal{M} solution).
2. Iteratively alternate the following two steps, $m = 0, 1, \dots$
 - $\hat{\gamma}^{(m+1)} = \arg \min_{\gamma} L(\hat{\beta}^{(m)}, \gamma)$ modifies “residuals”.
 - $\hat{\beta}^{(m+1)} = \arg \min_{\beta} L(\beta, \hat{\gamma}^{(m+1)})$. This step amounts to reapplying the \mathcal{M} procedure to $\hat{\gamma}^{(m+1)}$ -adjusted data although the nature of the data adjustment would largely depend on L .
3. Terminate the iteration when $\|\hat{\beta}^{(m+1)} - \hat{\beta}^{(m)}\|^2 < \epsilon$, where ϵ is a prespecified convergence tolerance.

In a nutshell, the algorithm attempts to find the joint minimizer (β, γ) by combining the minimizers β and γ resulting from the projected subspaces. Convergence of the iterative updates can be established under appropriate conditions. Before we state the conditions and results for convergence, we briefly describe implicit assumptions on the loss function and the complexity or penalty terms, $J(f)$ and $J_2(\gamma)$. $\mathcal{L}(y, f(x; \beta))$ is assumed to be non-negative. For simplicity, we assume that $J(f)$ of $f(x; \beta)$ depends on β only, and that it is of the form $J(f) = \|\beta\|_p^p$ and $J_2(\gamma) = \|\gamma\|_p^p$ for $p \geq 1$. The LASSO penalty has $p = 1$ while a ridge regression type penalty sets $p = 2$. Many other penalties of this format for $J(f)$ can be adopted as well to achieve better model selection properties or certain desirable performance of \mathcal{M} . Examples include those for the elastic net (Zou and Hastie; 2005), the grouped LASSO (Yuan and Lin; 2006), and the hierarchical LASSO.

For certain combinations of the loss \mathcal{L} and the penalty functionals, $J(f)$ and $J_2(\gamma)$, more efficient computational algorithms can be devised, as in Hastie et al. (2004); Efron et al. (2004); Rosset and Zhu (2007). However, in an attempt to provide a general computational recipe applicable to a variety of modeling procedures which can be implemented with simple modification of existing routines, we do not pursue the optimal implementation tailored to a specific procedure in this paper.

2.2 Convergence

Convexity of the loss and penalty terms plays a primary role in characterizing the solutions of the iterative algorithm. For a general reference to properties of convex functions and convex optimization, see Rockafellar (1997). First, we ensure that the minimizer pair (β, γ) in each step is properly defined.

Lemma 1. *Suppose that $L(\beta, \gamma)$ in (1) is continuous and strictly convex in β and γ for fixed λ_β and λ_γ . Given γ , there exists a unique minimizer $\beta(\gamma) = \arg \min_\beta L(\beta, \gamma)$, and vice versa.*

Proof. Given γ , choose an arbitrary $\beta^0 \in \mathbb{R}^p$ and consider $A_\beta := \{\beta \in \mathbb{R}^p \mid L(\beta, \gamma) \leq L(\beta^0, \gamma)\}$. By the continuity of L , A_β is closed. For a fixed $\lambda_\beta > 0$, it is bounded because it is contained in the ℓ_p ball of $\{\beta \in \mathbb{R}^p \mid \|\beta\|_p \leq L(\beta^0, \gamma)/\lambda_\beta\}$. Thus, there exists β in the compact set A_β attaining the minimum. Uniqueness follows from the strict convexity of L . Similarly, given β , there exists a unique minimizer $\gamma(\beta) = \arg \min_\gamma L(\beta, \gamma)$. \square

Remark The assumption that $L(\beta, \gamma)$ is strictly convex holds if the loss $\mathcal{L}(y, f(x; \beta))$ itself is strictly convex. Also, it is satisfied when a convex $\mathcal{L}(y, f(x; \beta))$ is combined with $J(f)$ and $J_2(\gamma)$ strictly convex in β and γ , respectively.

Lemma 2. *Under the same condition as Lemma 1, let $h : \mathbb{R}^{p+n} \rightarrow \mathbb{R}^{p+n}$ be the mapping of (β, γ) to its one-step update (β^1, γ^1) by the iterative algorithm. Then, h is continuous.*

Proof. By the iterative strategy, (β^1, γ^1) depends on β only, and β^1 is a composite mapping of β to β^1 . So, it is sufficient to show that the mappings of β to γ^1 and γ^1 to β^1 are continuous. Although the former, finding γ^1 given β , is a much simpler optimization than the latter in general for an array of $L(\beta, \gamma)$'s of practical interest, by the symmetry in the problem, we will show only that the mapping of γ^1 to β^1 is continuous. Dropping the superscript for notational simplicity, consider $g(\gamma) = \inf_\beta L(\beta, \gamma)$. Since $L(\beta, \gamma)$ is convex, g is convex in γ (see, for example, Rockafellar (1997), p.38) and thus continuous. For any sequence, $\{\gamma_n\}_{n=1}^\infty$ converging to γ , we want to show that the corresponding sequence of β minimizers, $\{\beta(\gamma_n)\}_{n=1}^\infty$, converges to $\bar{\beta} := \beta(\gamma)$.

As g is continuous, for $\epsilon > 0$, there is N such that $n \geq N$ implies $g(\gamma_n) \leq g(\gamma) + \epsilon$, that is, $L(\beta(\gamma_n), \gamma_n) \leq L(\bar{\beta}, \gamma) + \epsilon$. Let $M := \max\{\max_{n=1, \dots, N} L(\beta(\gamma_n), \gamma_n), L(\bar{\beta}, \gamma) + \epsilon\}$ and $A := \{(\beta, \gamma) \mid L(\beta, \gamma) \leq M\}$. Note that A is closed and bounded for fixed λ_β and λ_γ , and it contains the sequence of $\{(\beta(\gamma_n), \gamma_n)\}_{n=1}^\infty$. Therefore $\{(\beta(\gamma_n))_{n=1}^\infty\}$ has a convergent subsequence $\{(\beta(\gamma_{n_k}))_{k=1}^\infty\}$ with a limit, say, β^* . Then by the continuity of g and L ,

$$L(\bar{\beta}, \gamma) = g(\gamma) = \lim_{k \rightarrow \infty} g(\gamma_{n_k}) = \lim_{k \rightarrow \infty} L(\beta(\gamma_{n_k}), \gamma_{n_k}) = L(\beta^*, \gamma). \quad (2)$$

The uniqueness of the minimizer $\bar{\beta}$ at γ and (2) imply that $\beta^* = \bar{\beta}$. Consequently, this proves that every convergent subsequence of the bounded sequence $\{(\beta(\gamma_n))\}_{n=1}^{\infty}$ has the same limit $\bar{\beta}$. Hence the limit of $\{(\beta(\gamma_n))\}_{n=1}^{\infty}$ is $\bar{\beta}$, which completes the proof. \square

Proposition 3. *Suppose that $L(\beta, \gamma)$ is strictly convex in β and γ with a unique minimizer (β^*, γ^*) for fixed λ_β and λ_γ . Then, the iterative algorithm gives a sequence of $(\hat{\beta}^{(m)}, \hat{\gamma}^{(m)})$ with strictly decreasing $L(\hat{\beta}^{(m)}, \hat{\gamma}^{(m)})$. Moreover, $(\hat{\beta}^{(m)}, \hat{\gamma}^{(m)})$ converges to (β^*, γ^*) .*

Proof. For fixed λ_β and λ_γ , by the definition of $\hat{\gamma}^{(m+1)}$ and $\hat{\beta}^{(m+1)}$ in the algorithm, we have

$$L(\hat{\beta}^{(m)}, \hat{\gamma}^{(m)}) \geq L(\hat{\beta}^{(m)}, \hat{\gamma}^{(m+1)}) \geq L(\hat{\beta}^{(m+1)}, \hat{\gamma}^{(m+1)}).$$

Unless $(\hat{\beta}^{(m)}, \hat{\gamma}^{(m)}) = (\beta^*, \gamma^*)$, at least one of the inequalities must be strict by the strict convexity of L . Thus, $\{L(\hat{\beta}^{(m)}, \hat{\gamma}^{(m)})\}_{m=0}^{\infty}$ is a monotonically decreasing sequence. Since it is bounded below by zero, the sequence has a limit, say, L^* . Now consider $A := \{(\beta, \gamma) \mid L(\beta, \gamma) \leq L(\hat{\beta}^{(0)}, \hat{\gamma}^{(0)})\}$, which is closed and bounded. The sequence of the minimizers, $(\hat{\beta}^{(m)}, \hat{\gamma}^{(m)})$, is contained in A , and therefore it has a convergent subsequence $\{(\hat{\beta}^{(m_k)}, \hat{\gamma}^{(m_k)})\}_{k=1}^{\infty}$. Let $(\bar{\beta}, \bar{\gamma})$ denote the limit of the subsequence. By the continuity of L , $\lim_{k \rightarrow \infty} L(\hat{\beta}^{(m_k)}, \hat{\gamma}^{(m_k)}) = L(\bar{\beta}, \bar{\gamma})$.

Suppose that $(\bar{\beta}, \bar{\gamma}) \neq (\beta^*, \gamma^*)$, i.e., $L(\bar{\beta}, \bar{\gamma}) > L(\beta^*, \gamma^*)$. Then we can obtain the one-step update of $(\bar{\beta}, \bar{\gamma})$ denoted by $(\bar{\beta}^1, \bar{\gamma}^1)$ and further reduce the objective value by $\epsilon := L(\bar{\beta}, \bar{\gamma}) - L(\bar{\beta}^1, \bar{\gamma}^1) > 0$. By Lemma 2, the mapping of $(\bar{\beta}, \bar{\gamma})$ to $(\bar{\beta}^1, \bar{\gamma}^1)$ is continuous. So, there exists a $\delta > 0$ such that for any (β, γ) in an open ball centered at $(\bar{\beta}, \bar{\gamma})$ with a radius ϵ , $|L(\beta, \gamma) - L(\bar{\beta}^1, \bar{\gamma}^1)| < \epsilon/2$. This implies that for sufficiently large k , $L(\hat{\beta}^{(m_k+1)}, \hat{\gamma}^{(m_k+1)}) \leq L(\bar{\beta}^1, \bar{\gamma}^1) + \epsilon/2$. However, this leads to a contradiction that $L(\hat{\beta}^{(m_k+1)}, \hat{\gamma}^{(m_k+1)}) \leq \{L(\bar{\beta}, \bar{\gamma}) - \epsilon\} + \epsilon/2 = L(\bar{\beta}, \bar{\gamma}) - \epsilon/2$. Therefore, $(\bar{\beta}, \bar{\gamma}) = (\beta^*, \gamma^*)$. Furthermore, since the limit of any convergent subsequence is the same, we conclude that $(\hat{\beta}^{(m)}, \hat{\gamma}^{(m)})$ converges to (β^*, γ^*) . \square

3 Regression

Consider a linear model of the form $y_i = x_i^\top \beta + \epsilon_i$. Without loss of generality, we assume that each covariate is standardized. We also assume that the y_i 's are centered to zero. Let X be an n by p design matrix with x_i^\top in the i th row and let $Y = (y_1, \dots, y_n)^\top$.

3.1 LASSO

The LASSO (Tibshirani; 1996) estimate of $\beta = (\beta_1, \dots, \beta_p)^\top$ is defined as the solution $\hat{\beta} \in \mathbb{R}^p$ that minimizes

$$L_\lambda(\beta) = \frac{1}{2}(Y - X\beta)^\top(Y - X\beta) + \lambda \sum_{j=1}^p |\beta_j|, \quad (3)$$

where λ is a regularization parameter for balancing the data fit and the amount of shrinkage of β . We take this as a baseline model fitting procedure for illustration.

To reduce the sensitivity of the LASSO solution to influential observations, the given p covariates are augmented by n case indicators. Let z_i be the indicator variable taking 1 for the i th observation and 0 otherwise, and $\gamma = (\gamma_1, \dots, \gamma_n)^\top$ be the coefficients of the case indicators. The proposed modification of the LASSO with $J_2(\gamma) = \|\gamma\|_1$ leads to the robust LASSO. For the robust LASSO, we find $\hat{\beta} \in \mathbb{R}^p$ and $\hat{\gamma} \in \mathbb{R}^n$ that minimize

$$L(\beta, \gamma) = \frac{1}{2}\{Y - (X\beta + \gamma)\}^\top\{Y - (X\beta + \gamma)\} + \lambda_\beta \sum_{j=1}^p |\beta_j| + \lambda_\gamma \sum_{i=1}^n |\gamma_i|, \quad (4)$$

where λ_β and λ_γ are fixed regularization parameters constraining β and γ . Just as the ordinary LASSO in (3) stabilizes the solution by shrinking and selecting β , the additional penalty in the robust LASSO in (4) has the same effect on γ , whose components gauge the extent of case influences.

The minimizer $\hat{\gamma}$ of $L(\hat{\beta}, \gamma)$ for a fixed $\hat{\beta}$ can be found by soft-thresholding the residual vector $r = Y - X\hat{\beta}$. That is, $\hat{\gamma} = \text{sgn}(r)(|r| - \lambda_\gamma)_+$. For observations with small residuals, $|r_i| \leq \lambda_\gamma$, $\hat{\gamma}_i$ is set equal to zero with no effect on the current fit and for those with large residuals, $|r_i| > \lambda_\gamma$, $\hat{\gamma}_i$ is set equal to the residual $r_i = y_i - x_i^\top \hat{\beta}$ offset by λ_γ towards zero. Combining $\hat{\gamma}$ with $\hat{\beta}$, we define the adjusted residuals to be $r_i^* = y_i - x_i^\top \hat{\beta} - \hat{\gamma}_i$. That is, $r_i^* = r_i$ if $|r_i| \leq \lambda_\gamma$, and $r_i^* = \text{sgn}(r_i)\lambda_\gamma$, otherwise. Thus, introduction of the case-specific parameters along with the ℓ_1 penalty on γ amounts to winsorizing the ordinary residuals. The γ -adjusted loss is equivalent to truncated squared error loss which is $(y - x^\top \beta)^2$ if $|y - x^\top \beta| \leq \lambda_\gamma$, and is λ_γ^2 otherwise. Figure 1 shows (a) the relationship between the ordinary residual r and the corresponding γ , (b) the residual and the adjusted residual r^* , and (c) the γ -adjusted loss as a function of r .

The effective loss is $\mathcal{L}_{\lambda_\gamma}(y, x^\top \beta) = (y - x^\top \beta)^2/2$ if $|y - x^\top \beta| \leq \lambda_\gamma$, and $\lambda_\gamma^2/2 + \lambda_\gamma(|y - x^\top \beta| - \lambda_\gamma)$ otherwise. This effective loss matches Huber's loss function for robust regression (Huber; 1981), and yields the Huberized LASSO described by Rosset and Zhu (2004). As in robust regression, we choose a sufficiently large λ_γ so that only a modest fraction of the residuals are adjusted.

3.1.1 Bayesian Interpretation of the Robust LASSO

The LASSO solution has a dual interpretation as a posterior mode of β when the prior distribution of β is independent double exponential. An analogue can be drawn for the robust LASSO. Suppose that the β_j 's and γ_i 's have independent double exponential priors with mean 0 and the scale parameters σ_β and σ_γ , respectively. Consider a normal distribution with mean 0 and standard deviation σ for the errors ϵ_i . Treating the hyperparameters σ_β and σ_γ , and the error variance σ^2 as known constants, we have the posterior density function of β and γ

$$p(\beta, \gamma | X, Y) \propto \exp\left(-\frac{1}{2\sigma^2}\{Y - (X\beta + \gamma)\}^\top\{Y - (X\beta + \gamma)\} - \frac{1}{\sigma_\beta} \sum_{j=1}^p |\beta_j| - \frac{1}{\sigma_\gamma} \sum_{i=1}^n |\gamma_i|\right).$$

Thus, the posterior mode of (β, γ) is the maximizer of $\log p(\beta, \gamma | X, Y)$, or equivalently the minimizer of

$$\frac{1}{2}\{Y - (X\beta + \gamma)\}^\top\{Y - (X\beta + \gamma)\} + \frac{\sigma^2}{\sigma_\beta} \sum_{j=1}^p |\beta_j| + \frac{\sigma^2}{\sigma_\gamma} \sum_{i=1}^n |\gamma_i|.$$

Reparametrization of σ^2/σ_β and σ^2/σ_γ as λ_β and λ_γ yields the objective function of the robust LASSO in (4). With this reformulation, we have another interpretation of λ_β and λ_γ as the so-called noise-to-signal ratios.

3.2 Location Families

More generally, a wide class of regularization problems can be cast in the form of a minimization of

$$L_\lambda(\beta) = \sum_{i=1}^n g(y_i - x_i^\top \beta) + \lambda J(\beta),$$

where $g(\cdot)$ is the negative log-likelihood derived from a location family. The assumption that we have a location family implies that the negative log-likelihood is a function only of $r_i = y_i - x_i^\top \beta$. Dropping the subscript, common choices for the negative log-likelihood, $g(r)$ include r^2 (least squares, normal distributions) and $|r|$ (least absolute deviations, Laplace distributions).

Introducing the case specific parameters γ_i , we wish to minimize

$$L(\beta, \gamma) = \sum_{i=1}^n g(y_i - x_i^\top \beta - \gamma_i) + \lambda_\beta J(\beta) + \lambda_\gamma \|\gamma\|_1.$$

For minimization with a fixed $\hat{\beta}$, the next result applies to a broad class of $g(\cdot)$ (but not to $g(r) = |r|$).

Proposition 4. *Suppose that g is strictly convex with the minimum at 0, and $\lim_{r \rightarrow \pm\infty} g'(r) = \pm\infty$, respectively. Then,*

$$\hat{\gamma} = \arg \min_{\gamma} g(r - \gamma) + \lambda_{\gamma} |\gamma| = \begin{cases} r - g'^{-1}(\lambda_{\gamma}) & \text{for } r > g'^{-1}(\lambda_{\gamma}) \\ 0 & \text{for } g'^{-1}(-\lambda_{\gamma}) \leq r \leq g'^{-1}(\lambda_{\gamma}) \\ r - g'^{-1}(-\lambda_{\gamma}) & \text{for } r < g'^{-1}(-\lambda_{\gamma}). \end{cases}$$

The proposition follows from straightforward algebra. Set the first derivative of the decoupled minimization equation equal to 0 and solve for γ . Inserting these values for $\hat{\gamma}_i$ into the equation for $L(\beta, \gamma)$ yields

$$L(\hat{\beta}, \hat{\gamma}) = \sum_{i=1}^n g(r_i - \hat{\gamma}_i) + \lambda_{\beta} J(\hat{\beta}) + \lambda_{\gamma} \|\hat{\gamma}\|_1.$$

The first term in the summation can be decomposed into three parts. Large r_i contribute $g(r_i - r_i + g'^{-1}(\lambda_{\gamma})) = g(g'^{-1}(\lambda_{\gamma}))$. Large, negative r_i contribute $g(g'^{-1}(-\lambda_{\gamma}))$. Those r_i with intermediate values have $\hat{\gamma}_i = 0$ and so contribute $g(r_i)$. Thus a graphical depiction of the γ -adjusted loss is much like that in Figure 1, panel (c), where the loss is truncated above. For asymmetric distributions (and hence asymmetric log-likelihoods), the truncation point may differ for positive and negative residuals. It should be remembered that when $|r_i|$ is large, the corresponding $\hat{\gamma}_i$ is large, implying a large contribution of $\|\gamma\|_1$ to the overall minimization problem. The residuals will tend to be large for vectors β that are at odds with the data. Thus, in a sense, some of the loss which seems to disappear due to the effective truncation of g is shifted into the penalty term for γ . Hence the effective loss $\mathcal{L}_{\lambda_{\gamma}}(y, f(x; \beta)) = g(y - f(x; \beta) - \hat{\gamma}) + \lambda_{\gamma} |\hat{\gamma}|$ is the same as the original loss, $g(y - f(x; \beta))$ when the residual is in $[g'^{-1}(-\lambda_{\gamma}), g'^{-1}(\lambda_{\gamma})]$ and is linear beyond the interval. The linearized part of g is joined with g such that $\mathcal{L}_{\lambda_{\gamma}}$ is differentiable.

Computationally, the minimization of $L(\beta, \hat{\gamma})$ given $\hat{\gamma}$ entails application of the same modeling procedure \mathcal{M} with g to winsorized pseudo responses $y_i^* = y_i - \hat{\gamma}_i$, where $y_i^* = y_i$ for $g'^{-1}(-\lambda_{\gamma}) \leq r_i \leq g'^{-1}(\lambda_{\gamma})$, $y_i^* = g'^{-1}(\lambda_{\gamma})$ for $r > g'^{-1}(\lambda_{\gamma})$, and $y_i^* = g'^{-1}(-\lambda_{\gamma})$ for $r < g'^{-1}(-\lambda_{\gamma})$. So, the $\hat{\gamma}$ -adjusted data in Step 2 of the main algorithm consist of (x_i, y_i^*) pairs in each iteration.

3.3 Quantile Regression

In many applications, the assumption of normality on the distribution of the errors ϵ may not be appropriate. For instance, the error distribution may be skewed or heavy-tailed. More suited to such situations, quantile regression (Koenker and Bassett; 1978; Koenker and Hallock; 2001) aims to estimate conditional quantiles of y given x instead of the mean. For the $100\alpha^{th}$ quantile, the check function ρ_{α} is employed:

$$\rho_{\alpha}(r) = \begin{cases} \alpha r & \text{for } r \geq 0 \\ -(1 - \alpha)r & \text{for } r < 0. \end{cases} \quad (5)$$

Consider median regression with absolute deviation loss $\mathcal{L}(y, x^\top \beta) = |y - x^\top \beta|$. It can be verified easily that the ℓ_1 -adjustment of \mathcal{L} is void due to the piecewise linearity of the loss, reaffirming that the median regression is a robust procedure. For an effectual adjustment, the ℓ_2 norm regularization of the case-specific parameters is considered. In general, the penalized median regression estimate of β is defined to be the minimizer of

$$L_\lambda(\beta) = \sum_{i=1}^n |y_i - x_i^\top \beta| + \lambda_\beta J(\beta). \quad (6)$$

With the case-specific parameters γ_i , we have the following objective function for median regression:

$$L(\beta, \gamma) = \sum_{i=1}^n |y_i - x_i^\top \beta - \gamma_i| + \lambda_\beta J(\beta) + \frac{\lambda_\gamma}{2} \|\gamma\|^2. \quad (7)$$

For a fixed $\hat{\beta}$ and residual $r = y - x^\top \hat{\beta}$, the $\hat{\gamma}$ minimizing $|r - \gamma| + \frac{\lambda_\gamma}{2} \gamma^2$ is given by

$$\text{sgn}(r) \frac{1}{\lambda_\gamma} I(|r| > \frac{1}{\lambda_\gamma}) + r I(|r| \leq \frac{1}{\lambda_\gamma}).$$

The γ -adjusted loss for median regression is

$$\mathcal{L}(y, x^\top \beta + \hat{\gamma}) = |y - x^\top \beta - \frac{1}{\lambda_\gamma}| I(|y - x^\top \beta| > \frac{1}{\lambda_\gamma}),$$

as shown in Figure 2 (a). Interestingly, this ℓ_2 -adjusted absolute deviation loss is the same as the so-called “ ϵ -insensitive linear loss” for support vector regression (Vapnik; 1998) with $\epsilon = 1/\lambda_\gamma$.

With this adjustment, the effective loss is Huberized squared error loss. As illustrated in Section 6, the ℓ_2 adjustment makes quantile regression more efficient by rounding the sharp corner of the loss, and leads to a hybrid procedure which lies between mean and median regression. Note that, to achieve the desired effect for quantile regression, one chooses quite a different value of λ_γ than one would for when adjusting squared error loss for a robust mean regression. Also, for quantiles other than the median, the effective loss differs from Huberized squared error loss.

4 Classification

Now suppose that y_i 's indicate binary outcomes. For modeling and prediction of the binary responses, we mainly consider margin-based procedures such as logistic regression, support vector machines (Vapnik; 1998), and boosting (Freund and Schapire; 1997). These procedures can be modified by the addition of case indicators.

4.1 Penalized Logistic Regression

Although it is customary to label a binary outcome as 0 or 1 in logistic regression, we instead adopt the symmetric labels of $\{-1, 1\}$ for y_i 's. The symmetry facilitates comparison of different classification procedures. Logistic regression takes the negative log likelihood as a loss for estimation of logit $f(x) = \log[p(x)/(1 - p(x))]$. The loss, $\mathcal{L}(y, f(x)) = \log[1 + \exp(-yf(x))]$, can be viewed as a function of the so-called margin, $yf(x)$. This functional margin of $yf(x)$ is a pivotal quantity for defining a family of loss functions in classification similar to the residual in regression.

Penalized logistic regression can be modified with case indicators:

$$L(\beta_0, \beta, \gamma) = \sum_{i=1}^n \log(1 + \exp(-y_i\{f(x_i) + \gamma_i\})) + \lambda_\beta \|\beta\|_1 + \lambda_\gamma \|\gamma\|_1, \quad (8)$$

where $f(x) = \beta_0 + x^\top \beta$. For fixed $\hat{\beta}_0$ and $\hat{\beta}$, γ_i is determined by minimizing

$$\log(1 + \exp(-y_i\{f(x_i) + \gamma_i\})) + \lambda_\gamma |\gamma_i|.$$

First note that the minimizer γ_i must have the same sign as y_i . Letting $\eta = yf$ and assuming that $0 < \lambda_\gamma < 1$, we have

$$\begin{aligned} & \arg \min_{\gamma \geq 0} \log(1 + \exp(-\eta - \gamma)) + \lambda_\gamma |\gamma| \\ &= \begin{cases} \log\{(1 - \lambda_\gamma)/\lambda_\gamma\} - \eta & \text{if } \eta \leq \log\{(1 - \lambda_\gamma)/\lambda_\gamma\}, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

This yields a truncated negative log likelihood given by

$$\mathcal{L}(y, f(x)) = \begin{cases} \log(1 + \lambda_\gamma/(1 - \lambda_\gamma)) & \text{if } yf(x) \leq \log\{(1 - \lambda_\gamma)/\lambda_\gamma\}, \\ \log(1 + \exp(-yf(x))) & \text{otherwise.} \end{cases}$$

See Figure 2 (b) where $\eta_\lambda = \log\{(1 - \lambda_\gamma)/\lambda_\gamma\}$, and it is a decreasing function of λ_γ . λ_γ determines the level of truncation of the loss. As λ_γ tends to 1, there is no truncation.

4.2 Large Margin Classifiers

With the symmetric class labels, the foregoing characterization of the case-specific parameter γ in logistic regression can be easily generalized to various margin based classification procedures. In classification, potential outliers are those cases with large negative margins. Let $g(\tau)$ be a loss function of the margin $\tau = yf(x)$. The following proposition holds for a general family of loss functions. It is analogous to Proposition 4.

Proposition 5. *Suppose that g is convex and monotonically decreasing in τ , and $\lim_{\tau \rightarrow -\infty} g'(\tau) = \infty$. Then,*

$$\hat{\gamma} = \arg \min_{\gamma} g(\tau + \gamma) + \lambda_{\gamma} |\gamma| = \begin{cases} g'^{-1}(-\lambda_{\gamma}) - \tau & \text{for } \tau \leq g'^{-1}(-\lambda_{\gamma}) \\ 0 & \text{for } \tau > g'^{-1}(-\lambda_{\gamma}). \end{cases}$$

The proof is straightforward. Examples of the margin based loss g satisfying the assumption include the exponential loss $g(\tau) = \exp(-\tau)$ in boosting, the squared hinge loss $g(\tau) = \{(1 - \tau)_+\}^2$ in the support vector machine, and the negative log likelihood $g(\tau) = \log(1 + \exp(-\tau))$ in logistic regression. Although their theoretical targets are different, all the loss functions are truncated above for large negative margins when adjusted by γ . Thus, the effective loss $\mathcal{L}_{\lambda_{\gamma}}(yf(x; \beta)) = g(yf(x; \beta) + \hat{\gamma}) + \lambda_{\gamma} |\hat{\gamma}|$ is obtained by linearizing g for $yf(x; \beta) < g'^{-1}(-\lambda_{\gamma})$.

However, the effect of $\hat{\gamma}$ -adjustment depends on the form of g , and hence on the classification method. For boosting,

$$\hat{\gamma} = \begin{cases} -\log \lambda_{\gamma} - yf(x) & \text{if } yf(x) \leq -\log \lambda_{\gamma} \\ 0 & \text{otherwise.} \end{cases}$$

This gives $L(\beta, \hat{\gamma}) = \sum_{i=1}^n \exp(-y_i f(x_i; \beta) - \hat{\gamma}_i) = \sum_{i=1}^n \exp(-\hat{\gamma}_i) \exp(-y_i f(x_i; \beta))$. So, finding β given $\hat{\gamma}$ amounts to weighted boosting, where the positive case-specific parameters $\hat{\gamma}_i$ downweight the corresponding cases by $\exp(-\hat{\gamma}_i)$. For the squared hinge loss in the SVM,

$$\hat{\gamma} = \begin{cases} 1 - yf(x) - \lambda_{\gamma}/2 & \text{if } yf(x) \leq 1 - \lambda_{\gamma}/2 \\ 0 & \text{otherwise.} \end{cases}$$

A positive case-specific parameter $\hat{\gamma}_i$ has the effect of relaxing the margin requirement, that is, lowering the joint of the hinge individually. It allows the associated slack variable to be smaller in the primal formulation. Accordingly, the adjustment affects the coefficient of the linear term in the dual formulation of the quadratic programming problem.

4.3 Support Vector Machines

The linear Support Vector Machine (SVM) looks for the optimal hyperplane $f(x) = \beta_0 + x^{\top} \beta = 0$ minimizing

$$L_{\lambda}(\beta_0, \beta) = \sum_{i=1}^n [1 - y_i f(x_i)]_+ + \frac{\lambda}{2} \|\beta\|^2, \quad (9)$$

where $[t]_+ = \max(t, 0)$ and $\lambda > 0$ is a regularization parameter. Using the case indicators z_i and their coefficients γ_i , we modify (9), arriving at the problem of minimizing

$$L(\beta_0, \beta, \gamma) = \sum_{i=1}^n [1 - y_i \{f(x_i) + \gamma_i\}]_+ + \frac{\lambda_{\beta}}{2} \|\beta\|^2 + \frac{\lambda_{\gamma}}{2} \|\gamma\|^2. \quad (10)$$

For fixed $\hat{\beta}_0$ and $\hat{\beta}$, the minimizer $\hat{\gamma}$ of $L(\hat{\beta}_0, \hat{\beta}, \gamma)$ is obtained by solving the decoupled optimization problem of

$$\min_{\gamma} [1 - y_i f(x_i) - y_i \gamma]_+ + \frac{\lambda_{\gamma}}{2} \gamma^2 \text{ for each } \gamma_i.$$

With an argument similar to that for logistic regression, the minimizer $\hat{\gamma}_i$ should have the same sign as y_i . Let $\xi = 1 - y_i f$. A simple calculation shows that

$$\arg \min_{\gamma \geq 0} [\xi - \gamma]_+ + \frac{\lambda_{\gamma}}{2} \gamma^2 = \begin{cases} 0 & \text{if } \xi \leq 0 \\ \xi & \text{if } 0 < \xi < 1/\lambda_{\gamma} \\ 1/\lambda_{\gamma} & \text{if } \xi \geq 1/\lambda_{\gamma}. \end{cases}$$

Hence, the increase in margin $y_i \hat{\gamma}_i$ due to inclusion of γ is given by

$$\{1 - y_i f(x_i)\} I(0 < 1 - y_i f(x_i) < \frac{1}{\lambda_{\gamma}}) + \frac{1}{\lambda_{\gamma}} I(1 - y_i f(x_i) \geq \frac{1}{\lambda_{\gamma}}).$$

The γ -adjusted hinge loss is $\mathcal{L}(y, f) = [1 - 1/\lambda_{\gamma} - yf]_+$ with the hinge lowered by $1/\lambda_{\gamma}$ as shown in Figure 2 (c).

It can be shown that reapplying the SVM procedure to the γ -adjusted data is equivalent to the ordinary SVM with the regularization parameter λ inflated by $\lambda_{\gamma}/(\lambda_{\gamma} - 1)$ (for $\lambda_{\gamma} > 1$). In other words, the SVM formulation is intrinsically robust to outliers as long as λ is properly controlled. In fact, it is well known that $1/\lambda$ acts as the upper bound of the Lagrange multiplier of each observation in the dual formulation, which reflects the magnitude of the influence of each case.

5 Computation

Taking the LASSO as a primary example, we describe details of computational implementation and illustrate how existing software can be easily altered to fit the robust LASSO.

5.1 Robust LASSO Algorithm

Recall the problem of finding the solution $\hat{\beta}$ and $\hat{\gamma}$ to the robust LASSO in (4) for fixed λ_{β} and λ_{γ} . The iterative algorithm in Section 2 can be restated for the LASSO as follows:

1. Initialize $\hat{\gamma}^{(0)} = 0$ and get the ordinary LASSO solution $\hat{\beta}^{(0)} = \arg \min L(\beta, 0)$.
2. Iteratively alternate the following two steps, $m = 0, 1, \dots$
 - $\hat{\gamma}^{(m+1)} = \arg \min L(\hat{\beta}^{(m)}, \gamma) = \text{sgn}(r^{(m)}) (|r^{(m)}| - \lambda_{\gamma})_+$, where $r^{(m)} = Y - X \hat{\beta}^{(m)}$.

- $\hat{\beta}^{(m+1)} = \arg \min L(\beta, \hat{\gamma}^{(m+1)})$. This step is LASSO optimization with a winsorized response $Y^* = Y - \hat{\gamma}^{(m+1)} = X\hat{\beta}^{(m)} + r^{*(m)}$.

3. Terminate when $\sum_{j=1}^p (\hat{\beta}_j^{(m+1)} - \hat{\beta}_j^{(m)})^2/p < \epsilon$ for a prespecified small value ϵ .

For fitting β , existing algorithms for LASSO such as LARS (Efron et al.; 2004) can be used. Since the LARS algorithm can generate the entire solution path of $\hat{\beta}$ as a function of λ_β , it may be numerically more efficient to incorporate tuning of λ_β with the iterative algorithm than to find $\hat{\beta}$ for a fixed λ_β . So, we may consider joint updating of λ_β and $\hat{\beta}$ as well as λ_γ and $\hat{\gamma}$ as suggested in Gu (1992) for the similar issue of combining an iterative algorithm with tuning. Namely, one can choose the best λ_β at each iteration for fitting β and the best λ_γ for fitting γ with respect to some selection criteria.

5.2 Selection of the Penalty Parameters

As with methods of regularization in general, the choice of the penalty parameters is important for the effectiveness of the robust LASSO. Appropriate selection of λ_β and λ_γ needs to be combined with the iterative algorithm.

The role of λ_γ as a bending constant for winsorizing the residuals makes it sensible to set $\lambda_\gamma = k \cdot \sigma$ with a proper choice of k . The standard robust statistics literature (Huber; 1981) suggests that good choices of k lie in the range from 1 to 2. The effect of the k on estimation error is illustrated in the simulation studies in Section 6. Setting λ_γ in this fashion requires an estimate of σ . One may use a robust estimate by fitting a full robust regression model with an M estimator as implemented in **MASS** package. Alternatively, one can refit σ in each iteration and dynamically update λ_γ . Empirically, little difference has been observed between the two methods.

For the choice of λ_β , C_p -type risk estimates can be used. Note that the C_p criterion requires an accurate estimate of σ , and its implementation in the LARS algorithm uses $\hat{\sigma}$ from the full OLS fit. As potential outliers may significantly influence this assessment of fit through C_p , a robust estimate of σ is recommended for C_p as well. Ronchetti and Staudte (1994) obtain a robust version of C_p by replacing the squared error loss with its truncated version and p with $c \cdot p$ in C_p , where c is a constant, slightly smaller than 1, that depends on the the bending constant k . For instance, $c \approx 0.9817$ for $k = 1.345$ and 0.9985 for $k = 2$. The ordinary sum of squared residuals with the updated Y^* in the C_p evaluation is effectively the same as the empirical risk with respect to truncated squared error loss. With k as large as 2, the C_p in the LARS iterations is close to the robust C_p . This justifies selection of λ_β using C_p in each iteration.

An alternative method for selection of λ_β is generalized cross validation (GCV). Unlike C_p , its evaluation does not depend on $\hat{\sigma}^2$ although the leave-one-out cross validation identity, the theoretical basis for the GCV has not been established for the robust LASSO.

With C_p modified by the robust estimate of σ , numerical experiments indicate that the robust LASSO algorithm typically converges in several iterations, with a noticeable change in coefficients, if any, occurring in the first update only. For large k , fewer iterations are needed.

6 Simulations

We conducted a simulation study to investigate the sensitivity of the LASSO (or LARS) and the robust LASSO (robust LARS, $k = 2$) to contamination of the data. For brevity, we report only that portion of the results pertaining to accuracy of the fitted regression surface and inclusion of variates in the model. Similar results were obtained for k near 2. The results differ for extreme values of k . Throughout the simulation, the standard linear model $y = x^\top \beta + \epsilon$ was assumed. Following the simulation setting in Tibshirani (1996), we generated $x = (x_1, \dots, x_8)^\top$ from a multivariate normal distribution with mean zero and standard deviation 1. The correlation between x_i and x_j was set to $\rho^{|i-j|}$ with $\rho = 0.5$. Three scenarios were considered with a varying degree of sparsity in terms of the number of non-zero true coefficients: i) sparse: $\beta = (5, 0, 0, 0, 0, 0, 0, 0)$, ii) intermediate: $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$, and iii) dense: $\beta_j = 0.85$ for all $j = 1, \dots, 8$. In all cases, the sample size was 100. For the base case, ϵ_i was assumed to follow $N(0, \sigma^2)$ with $\sigma = 3$. For potential outliers in ϵ , the first 5% of the ϵ_i 's were tripled, yielding a data set with more outliers. We also investigated sensitivity to high leverage cases. For this setting, we tripled the first 5% of the values of x_1 . Thus the replicates were blocked across the three settings. There were 100 replicates in the simulation. The C_p criterion was used to select the model.

Figure 3 shows mean square error (MSE) between the fitted and true regression surfaces, omitting intercepts. MSE is integrated across the distribution of a future X , taken to be that for the base case of the simulation. Over the n replicates in the simulation, $MSE = n^{-1} \sum_{i=1}^n (\hat{\beta}^i - \beta)^\top \Sigma (\hat{\beta}^i - \beta)$, where $\hat{\beta}^i$ is the estimate of β for the i^{th} replicate. LARS and robust LARS perform comparably in the base case, with the MSE for robust LARS being greater by 1 to 6 percent. For both LARS and robust LARS, MSE in the base case increases as one moves from the sparse to the dense scenario. MSE increases noticeably when ϵ is contaminated, by a factor of 1.31 to 1.41 for LARS. For robust LARS, the factor for increase over the base case with LARS is 1.12 to 1.22. For contamination in X , results under LARS and robust LARS are similar in the intermediate and dense cases, with increases in MSE over the base case. For the sparse case, the coefficient of the contaminated covariate, x_1 , is large relative to the other covariates. Here, robust LARS performs noticeably better than LARS, with a smaller increase in MSE .

Table 1 presents results on the difference in number of selected variables for pairs of models. In each pair, a contaminated model is contrasted with the corresponding uncontaminated model. The top half of the table presents results for contamination of ϵ . The distribution of the differences in the number of selected variables for the

pairs of fitted models has a mode at 0 in each scenario for both LARS and robust LARS. There is, however, substantial spread around 0. The fitted models for the data with contaminated errors tend to have fewer variables than those for the original data, especially in the dense scenario. This may well be attributed to inflated estimates of σ^2 used in C_p for the contaminated data, favoring relatively smaller models. The effect is stronger for LARS than for robust LARS, in keeping with the lessened impact of outliers on the robust estimate of σ^2 .

The bottom half of Table 1 presents results for contamination of X . Again, the distributions of differences in model size have modes at 0 in all scenarios. The distributions have substantial spread around 0. Under the sparse scenario in which the contamination has a substantial impact on MSE , the distribution under robust LARS is more concentrated than under LARS.

The simulation demonstrates that the proposed robustification is successful in dealing with both contaminated errors and contaminated covariates. As expected, in contrast to LARS, robust LARS is effective in identifying observations with large measurement errors and lessening their influence. It is also effective at reducing the impact of high leverage cases, especially when the high leverage arises from a covariate with a large regression coefficient. The combined benefits of robustness to outliers and high leverage cases render robust LARS effective at dealing with influential cases in an automated fashion.

We also examined quantile regression, finding that regular quantile regression can be made more efficient by inclusion and ℓ_2 regularization of case-specific parameters. With Huberized loss being the effective loss in this case, we examined the effect of k on regression of conditional medians, where $k \cdot \sigma$ is used as the actual bending constant. In addition to the normal distribution, we considered a linear transformed log normal distribution with median zero and $\sigma \approx 3$ as a skewed distribution for errors. The same simulation settings were used for X and β , and 100 replicates with normally distributed errors and log normally distributed errors were generated. Note that for both error distributions, the conditional mean functions in the previous simulation give the conditional median functions in this simulation. Huber’s robust regression was applied to the simulated data. Figure 4 presents approximate 95% confidence intervals for MSE for this “efficient” median regression as a function of k . $k = 0$ corresponds to the ordinary median regression while $k = 3$ approximates least squares regression. When the error distribution is normal, mean regression has smaller MSE than median regression. When the error distribution is skewed, there is a virtue of moving away from the regular median regression for efficient estimation of conditional medians. Huber’s regression with k around 0.5 provides more accurate fits to the median regression coefficients than does median regression.

7 Analysis of Language Data

Balota et al. (2004) conducted an extensive lexical decision experiment in which

subjects were asked to identify whether a string of letters was an English word or a non-word. The words were monosyllabic, and the non-words were constructed to closely resemble words on a number of linguistic dimensions. Two groups were studied – college students and older adults. The data consist of response times by word, averaged over the thirty subjects in each group. For each of word, a number of covariates was recorded. Goals of the experiment include determining which features of a word (i.e., covariates) affect response time, and whether the active features affect response time in the same fashion for college students and older adults. The authors make a case for the need to conduct and analyze studies with regression techniques in mind, rather than simpler ANOVA techniques.

Baayen (2007) conducts an extensive analysis of a slightly modified data set which is available in his `languageR` package. In his analysis, he creates and selects variables to include in a regression model, addresses issues of nonlinearity, collinearity and interaction, and removes selected cases as being influential and/or outlying. He trims a total of 87 of the 4568 cases. The resulting model, based on “typical” words, is used to address issues of linguistic importance. It includes seventeen basic covariates which enter the model as linear terms, a non-linear term for the written frequency of a word (fit as a restricted cubic spline with five knots), and an interaction term between the age group and the (nonlinear) written frequency of the word. We take his final model as an expert fit, and use it as a target to compare the performance of the robust LASSO to the LASSO.

We consider two sets of potential covariates for the model. The small set consists of Baayen’s 17 basic covariates and three additional covariates representing a squared term for written frequency and the interaction between age group and the linear and squared terms for written frequency. Age group has been coded as ± 1 for the interactions. The large set augments these covariates with nine additional covariates that were not included in Baayen’s final model. Baayen excluded some of these covariates for a lack of significance, others because of collinearity. The LASSO and robust LASSO were fit to all 4568 cases with the small and large sets of covariates. For the robust LASSO, a variety of values were considered for k . In all cases, the model was selected via the minimum C_p criterion. The estimated error standard deviations were approximately 0.0775 and 0.0789 for both of the LASSO and the robust LASSO models with the small and large sets of covariates, respectively. A comparison of the models in terms of sum of squared deviations (SSD) from Baayen’s fitted values, with the sum excluding those cases that he removed as outliers and/or influential cases, shows mixed results, with the robust LASSO having a smaller value of SSD for some values of k and larger SSD for other values. The comparison highlights the discontinuity of the fitted surface in k whenever a different model is selected by minimum C_p . In order to assess the relative performance of the two methods, we must average over a number of data sets.

To obtain the needed replicates and to examine the performance of the robust LASSO with a smaller sample size, we conducted a simulation study. For one repli-

cate of the simulation, we sampled 400 cases from the data set and fit the LASSO and robust LASSO. Models were selected with the minimum C_p criterion, and SSD computed on the full data set, removing the cases that Baayen did. A summary of SSD is presented in Figure 5. The figure also presents confidence intervals, based on 5,000 replicates, for the robust LASSO. We see that the robust LASSO outperforms the LASSO for both the small and large covariate sets over a wide range of k .

In the simulation study, we also tracked the covariates in the selected model, with a summary displayed for $k = 1.6$ in Figure 6. Two key features of the coefficients are whether or not they are zero and how large they are. Several interesting features appear. First, there is little apparent difference between the 20 covariates in the small set (covariates 1 through 20) and the additional nine covariates in the large set (covariates 21 through 29). Both sets of covariates often have coefficients near 0, as indicated by the left-hand panel. The vertical bars, extending from the 10th to the 90th percentile for a given coefficient, show many effects that are, on the whole, small. The right-hand panel presents the fraction of replicates in which a given coefficient is non-zero in the model selected by the minimum C_p criterion. Again, we see little difference between the two sets of covariates. Overall, covariates in the small set had non-zero coefficients 0.73 of the time under the LASSO and 0.75 of the time under the robust LASSO. The remaining covariates had non-zero coefficients 0.68 and 0.70 of the time, respectively. The models selected by the robust LASSO averaged 21.4 covariates; those selected by the LASSO averaged 20.7 covariates. Second, in spite of the relatively high frequency with which covariates were “included” in the model, they often had very small coefficients, suggesting that there is substantial uncertainty about the “correct” model. Third, the LASSO and the robust LASSO provide coefficients of similar magnitude and with similar non-zero frequency. For this data set and simulation, we expect this behavior, as the fraction of outlying data is small and the magnitude of the outliers is modest. Even in this difficult situation for a robust procedure, as Figure 5 shows, the robust LASSO can better match the expert fit. Fourth, examination of particular covariates demonstrates the appeal of regularization methods. Covariates 20 (`WrittenFrequency`) and 29 (`Familiarity`) address the same issue. See Balota et al. (2004) for a more complete description of the covariates. Both covariates appear in nearly all of the models for both the LASSO and the robust LASSO. Subjects are able to decide that a familiar word is a word more quickly (and more accurately) than an unfamiliar word, and so we see negative coefficients for both covariates. Although there seems to be no debate on whether this conceptual effect of similarity exists, there are a variety of viewpoints on how to best capture this effect. Regularization methods facilitate inclusion of a suite of covariates that address single conceptual effect. The robust LASSO retains this ability.

8 Discussion

In the preceding sections, we have laid out an approach to modifying regularized estimation problems. The approach is based on the creation of case-specific covariates that are included as part of the regularization problem. With appropriate choices of penalty terms, the addition of these covariates allows us to robustify those procedures which lack robustness and also allows us to improve the efficiency of procedures which are very robust, but not particularly efficient. The techniques are easy to implement, as they often require little modification of existing software. In some cases, there is no need for modification of software, as one merely feeds a modified data set into existing routines.

The motivation behind this work is a desire to move regularized estimation in the direction of traditional data analysis (e.g., Weisberg (2004)). An important component of this type of analysis is the ability to take different looks at a data set. These different looks may suggest creation of new variates and differential handling of individual cases or groups of cases. Robust methods allow us to take such a look, even when data sets are large. Coupling robust regression techniques with the ability to examine an entire solution path provides a sharper view of the impact of unusual cases on the analysis. Another important component of a traditional analysis is mapping observed patterns in data into established modeling concepts, such as the fixed effect/random effect distinction, overdispersion, and case-specific uncertainty. The approach that we have taken suggests how these issues can be handled in the context of regularized regression. The next several paragraphs provide capsule descriptions.

There are two versions of a model which incorporates random effects. In one version, the effects are marginalized, never explicitly appearing in the model; in a second version, the effects appear, and their distribution—often normal with mean 0—is described. Pursuing the second approach, the covariate set can be expanded by including the random effects. In keeping with an assumption of normality of the effects, an ℓ_2 penalty can be imposed upon them. In the context of the generalized linear model, the random effects can account for overdispersion. In the context of random effects regression (Laird and Ware; 1982), the covariates for the random effects would be more than indicator variables for sets of cases.

For continuous response variates, differential case-uncertainty translates into differential case-dispersion. This can be accomplished by including case-specific indicators to represent random effects and adjusting the penalty for the individual case to match the case-uncertainty. The computational strategy described in Section 2 applies, as the case-specific terms decouple for the minimization. Thus, inclusion of the additional terms provides a direct means of extending the LASSO to heteroscedastic regression, and from there, to heteroscedastic robust regression.

The simulation on quantile regression shows the potential for improvement in accuracy due to inclusion of case-specific covariates. For large enough samples, the bias of the enhanced estimator will typically outweigh its benefits. The natural approach is to adjust the penalty attached to the case-specific covariates as the sample size

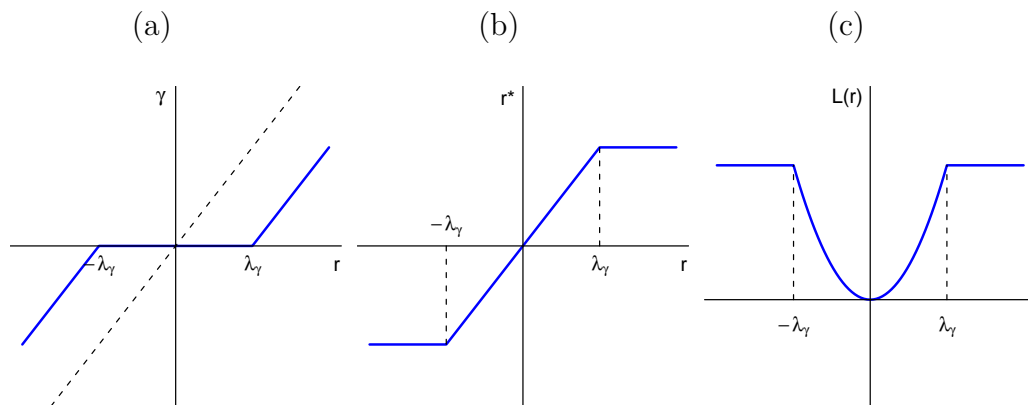


Figure 1: Modification of the squared error loss with a case-specific parameter. (a) γ versus the residual r . (b) the adjusted residual r^* versus the ordinary residual r . (c) a truncated squared error loss.

increases. This can be accomplished in two different ways. First, the parameter λ_γ can be increased as the sample size grows; second, the norm for the penalty can be decreased, moving toward the ℓ_1 norm.

References

- Baayen, R. (2007). *Analyzing Linguistic Data: a practical introduction to statistics*, Cambridge University Press, Cambridge, England.
- Balota, D., Cortese, M., Sergent-Marshall, S., Spieler, D. and Yap, M. (2004). Visual word recognition of single-syllable words, *Journal of Experimental Psychology* **133**: 283–316.
- Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression, *Annals of Statistics* **32**(2): 407–499.
- Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences* **55**(1): 119–139.
- Gu, C. (1992). Cross-validating non-Gaussian data, *Journal of Computational and Graphical Statistics* **1**: 169–179.
- Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, Chapman & Hall/CRC.

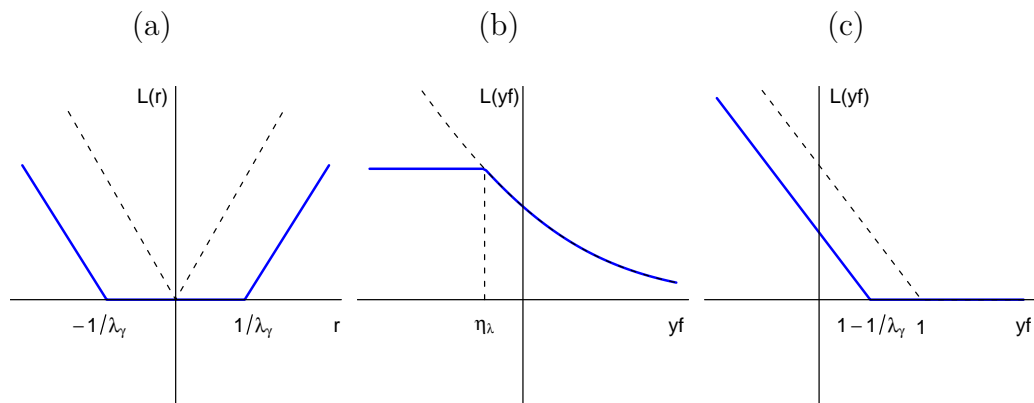


Figure 2: Robust versions of (a) absolute deviation loss for median regression, (b) negative log likelihood for logistic regression, and (c) hinge loss for the support vector machine.

Hastie, T., Rosset, S., Tibshirani, R. and Zhu, J. (2004). The entire regularization path for the support vector machine, *Journal of Machine Learning Research* **5**: 1391–1415.

Hoerl, A. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics* **12**(3): 55–67.

Huber, P. J. (1981). *Robust statistics*, John Wiley & Sons, New York.

Koenker, R. and Bassett, G. (1978). Regression quantiles, *Econometrica* **1**: 33–50.

Koenker, R. and Hallock, K. (2001). Quantile regression, *Journal of Economic Perspectives* **15**: 143–156.

Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data, *Biometrics* **38**: 963–974.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, Second Edition*, Chapman & Hall/CRC.

Rockafellar, R. T. (1997). *Convex Analysis (Princeton Landmarks in Mathematics and Physics)*, Princeton University Press.

Ronchetti, E. and Staudte, R. G. (1994). A robust version of Mallows’s C_p , *Journal of the American Statistical Association* **89**(426): 550–559.

Rosset, S. and Zhu, J. (2004). Discussion of “Least angle regression” by Efron, Hastie, Johnstone and Tibshirani, *Annals of Statistics* **32**(2): 469–475.

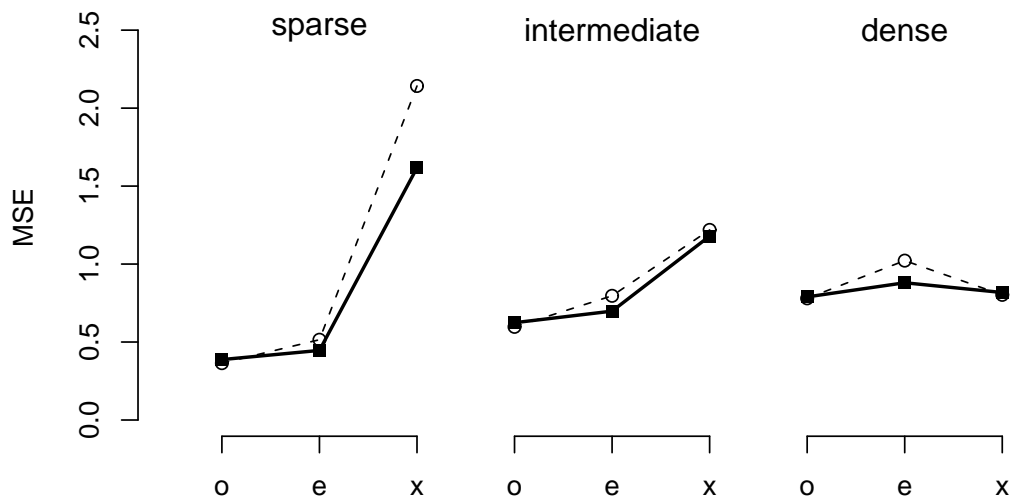


Figure 3: Mean squared error (MSE) of $\hat{\beta}$ for LARS and its robust version under three different scenarios in the simulation study. In each scenario, o, e, and x indicate clean data, data with contaminated measurement errors, and data with mismeasured first covariate. The dotted lines are for LARS while the solid lines are for robust LARS. The points are the average MSE for 100 replicates.

Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths, *The Annals of Statistics*.

Shen, X., Zhang, X., Tseng, G. C. and Wong, W. H. (2003). On ψ -learning, *Journal of the American Statistical Association* **98**(463): 724–734.

Tibshirani, R. (1996). Regression selection and shrinkage via the lasso, *Journal of the Royal Statistical Society, Series B* **58**(1): 267–288.

Vapnik, V. N. (1998). *Statistical Learning Theory*, Wiley-Interscience.

Wahba, G. (1990). *Spline Models for Observational Data*, Series in Applied Mathematics, Vol. 59, SIAM, Philadelphia.

Weisberg, S. (2004). Discussion of “Least angle regression” by Efron, Hastie, Johnstone and Tibshirani, *The Annals of Statistics* **32**(2): 490–494.

Weisberg, S. (2005). *Applied Linear Regression*, 3rd edn, Wiley, Hoboken NJ.

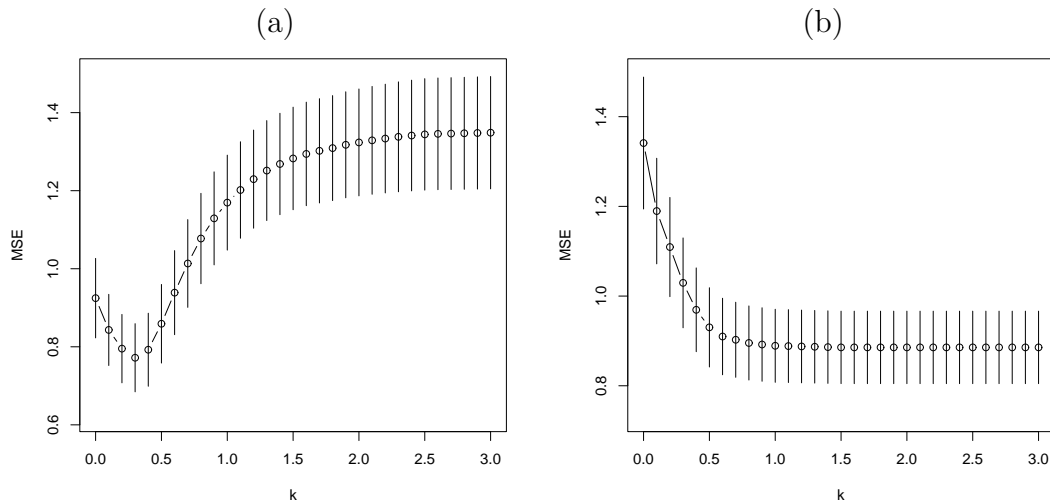


Figure 4: MSE for the “efficient” median regression as a function of k . Panel (a) is for errors with a shifted log normal distribution, and panel (b) is for normally distributed errors. The points represent means of 100 replicates, and the vertical lines give approximate 95% confidence intervals for the MSE

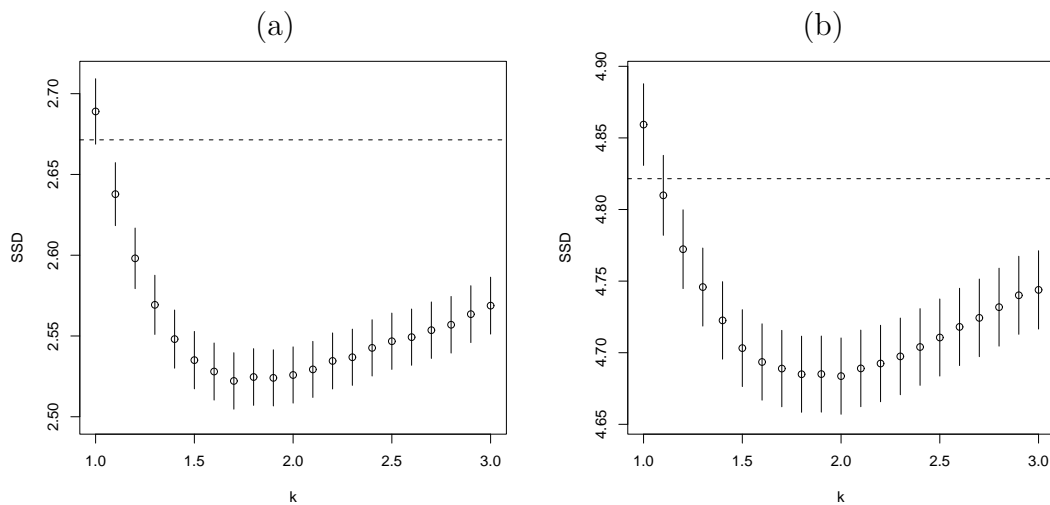


Figure 5: Sum of squared deviations (SSD) from Baayen’s fits in the simulation study. The horizontal line is the mean SSD for the the LASSO while the points represent the mean of SSDs for the robust LASSO. The vertical lines give approximate 95% confidence intervals for the mean SSDs. Panel (a) presents results for the small set of covariates and panel (b) presents results for the large set of covariates.

Table 1: Difference in the number of selected variables for the fitted model to contaminated data from that to clean data

Scenario	LARS							robust LARS						
	-3	-2	-1	0	1	2	3	-3	-2	-1	0	1	2	3
ϵ contamination														
Sparse	5*	6	21	48	13	5	2*	1*	4	12	71	7	5	0
Intermediate	5	10	14	46	21	3	1	1	3	14	64	14	4	0
Dense	2	1	16	80	1	0	0	0	0	8	89	3	0	0
X contamination														
Sparse	7*	5	15	34	20	7	12*	5*	3	16	36	22	12	6
Intermediate	1	5	1	55	21	3	2	1	3	18	50	23	4	1
Dense	0	0	5	93	2	0	0	0	0	4	94	2	0	0

NOTE: The entries with * are the cumulative counts of the specified case and more extreme cases.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(1): 49–67.

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**(2): 301–320.

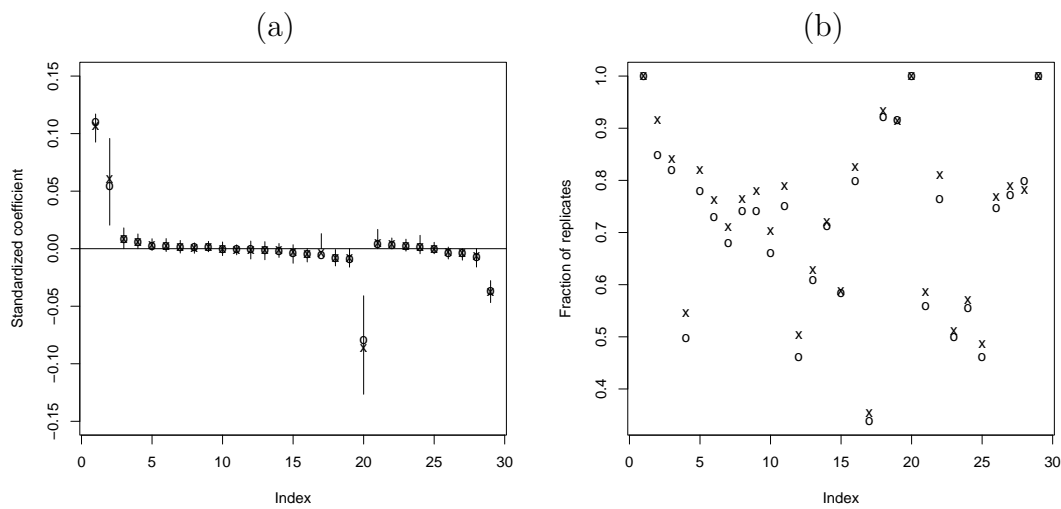


Figure 6: Coefficients in the simulation study. The first 20 variates are the small set of covariates, while the remaining 9 variates are in the large set of covariates. \times is for the robust LASSO and \circ is for the LASSO. Panel (a) contains a plot of mean coefficients for standardized variates in the simulation. The vertical lines extend from the 10th to 90th percentiles of the coefficients under the robust LASSO. Panel (b) shows the fraction of replicates in which coefficients in the selected model are non-zero.