Robust Regression for Highly Corrupted Response by Shifting Outliers

Yoonsuh Jung [1], Seung Pil Lee [2], Jianhua Hu [3]

[1] Department of Statistics, University of Waikato, Hamilton, New Zealand

[2] Division of International Sport and Leisure, Hankuk University of Foreign Studies, Yongin, South Korea

[3] Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX, USA

Address for correspondence: Yoonsuh Jung, Department of Statistics, University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand.

E-mail: yoonsuh@waikato.ac.nz.

Phone: (+64) 7 838 4773.

Fax: (+64) 7 838 4155.

Abstract: Outlying observations are often disregarded at the sacrifice of degrees of freedom or downsized via robust loss functions (for example, Huber's loss) to reduce the undesirable impact on data analysis. In this paper, we treat the outlying status of each observation as a parameter and propose a penalization method to automatically adjust the outliers. The proposed method shifts the outliers towards the fitted values, while preserve the non-outlying observations. We also develop a generally applicable algorithm in the iterative fashion to estimate model parameters and demonstrate the

connection with the maximum likelihood based estimation procedure in the case of least squares estimation. We establish asymptotic property of the resulting parameter estimators under the condition that the proportion of outliers do not vanish as sample size increases. We apply the proposed outlier adjustment method to ordinary least squares and lasso-type of penalization procedure and demonstrate its empirical value via numeric studies. Furthermore, we study applicability of the proposed method to two robust estimators, Huber's robust estimator and Huberized lasso, and demonstrate its noticeable improvement of model fit in the presence of extremely large outliers.

Key words: Case-specific Parameter; Extreme outliers; Huber's estimator; Robust lasso; Robust Linear Model

1 Introduction

Extensive research has been conducted to deal with outliers that can impede proper summarization of the overall tendency in a data set. The existing methods range from two-step approaches of first identifying the outliers and then reducing its effect Pena and Yohai (1999)(Kosinski, 1999) to robust regression procedures (Bloomfield and Steiger, 1983; Huber, 1973; Beaton and Tukey, 1974; Huber and Ronchetti, 2009; Chambers and Heathcote, 1981). Along the direction of adopting an explicit error distribution to account for the outlier data, Lange et al. (1989) used a heavy-tailed t-distribution to accommodate the outliers. Hadi and Simonoff (1993) provides a good literature review on the topic of outlier detection and robust regressions. More

recently, Lee et al. (2012) introduced a penalization method with the parameters indicating the outlying status of all the observations, called "case-specific parameters", and a lasso penalty (Tibshirani, 1996) for the case-specific parameters to identify and downsize the outliers in an automated fashion. Case-specific parameters are devised to deal with each observation and should be penalized to avoid over-saturation. The method of Lee et al. (2012) with case-specific parameter turned out to be identical to Huber's estimator (Huber, 1973) and Huberized lasso (Rosset and Zhu, 2007) depending on the incorporation of lasso penalty. (See Section 3.1 of Lee et al. (2012) for the details.)

In this work, we are particularly interested in robust modeling of a data set containing massive outliers. Specifically, we propose a new penalty function for the case-specific parameters which are tailored to identify the influential outliers and to effectively reduce their impact on the fitted values. For the asymptotic behavior of the proposed method, we consider the nonstandard situation where the proportion of outlying samples remains nonzero as sample size increases.

We start with a classical linear regression model,

$$y_i = x_i'\beta + \epsilon_i, \tag{1.1}$$

where ϵ_is are independently and identically distributed (iid) with mean 0 and a finite variance σ^2, and x_i is a length-p vector for the ith observation. The ϵ_i is typically assumed to be normally distributed, which can be easily violated with the presence of large outliers (or leverage points). To address this, we use an additional parameter γ_i to capture the outlying observations and make the normality assumption more

reasonable. Accordingly, the model is modified as

$$y_i = x_i'\beta + \gamma_i + \epsilon_i, \quad i = 1, ..., n, \tag{1.2}$$

where γ_is are the case-specific parameters (Lee et al., 2012). It takes non-zero values for the outlying observations, and zero for the non-outlying observations. Note that $\gamma_i + \epsilon_i$ represents the deviation from the true regression line. To make γ_i and ϵ_i identifiable, we impose a constraint of setting ϵ_i to zero when its corresponding γ_i is non-zero. It means that γ_i is defined to be $y_i - x_i'\beta$ if $|y_i - x_i'\beta|$ is greater than a pre-specified threshold λ_γ. When $|y_i - x_i'\beta| < \lambda_\gamma$, we set γ_i to zero, which indicates the ith observation is not an outlier.

When there are massive outliers, diminishing the effect of outliers is often not sufficient. In contrast, the suggested methods have an effect of shifting the detected outliers onto the fitted regression line. Thus, most of the undesirable effects can be removed effectively without deleting the observations, which is our major contribution to the literature. Another advantage could be the broad application of this idea to general modelling procedures beyond ordinary least squares.

In Section 2.1, the estimation procedure is cast into the convex function optimization (Rockafellar, 1997). The estimation of β and $\gamma = (\gamma_1, \ldots, \gamma_n)'$ which satisfy the aforementioned conditions is also conducted through an iterative maximum likelihood-type approach, and we show that the two approaches are equivalent. The choice of λ_γ is discussed in Section 2.2. The form of penalty for γ_is appears under the convex minimization approach only. The case-specific parameters and the corresponding penalty are applied to the ordinary least squares (OLS) and Huber's robust regression (Huber and Ronchetti, 2009) in Section 3.1.

As lasso (Tibshirani, 1996) can do variable selection, which is an crucial part of data analysis, we extend the proposed method to lasso and Huberized lasso (Rosset and Zhu, 2007) in Section 3. In Section 4, simulation studies are conducted to illustrate the suggested methods and to show the improvement compared to the methods by (Lee et al., 2012) and other robust methods. Further simulation studies are contained in the supplementary materials. The applications of the proposed methods to real data sets are given in Section 5 where the first data set is a well-known small data set and the second data set is a larger data set with 3,000 samples.

2 Robust regression by shifting outliers

2.1 Model estimation

We set X to be a design matrix of size n by p with full column rank, and Y to be a response vector of size n. For the purpose of estimating (1.2), we minimize the following objective function

$$L(\beta,\gamma)=\frac{1}{2}\sum_{i=1}^{n}(y_i-x_i'\beta-\gamma_i)^2+\sum_{i=1}^{n}\gamma_i^2 I(|y_i-x_i'\beta|<\lambda_\gamma) \text{ subject to } \sum_{i=1}^{n}\gamma_i\epsilon_i=0. \quad (2.1)$$

The first term on the right side of (2.1) is the measure of goodness of fits, and the second term penalizes γ_i only when $|y_i - x_i'\beta|$ is smaller than λ_γ. Therefore, this new penalty is devised to capture the outliers by shrinking the case-specific parameters γ_is towards zero only for non-outlying observations. On the other hand, γ_i can be nonzero only if the ith observation is a potential outlier. Since the first and second term of (2.1) are quadratic, it is clear that $L(\beta,\gamma)$ is a convex function, which will provide convergency of the suggested iterative algorithm.

Note that the form of the penalty for γ_is is similar to hard thresholding rule. When the hard threshold is used for β, it is identical to best subset selection under the orthonormal design matrix. The proposed penalty is somewhat similar to the best subset selection regarding its format. The difference is that we are selecting the observations instead of variables.

The aforementioned model identifiability constraint for model (1.2) implies that $\gamma_i\epsilon_i = 0$, and γ_i and ϵ_i cannot be zero simultaneously. For implementation, we include an addition term of $\sum_{i=1}^{n} \gamma_i\epsilon_i I(|y_i - x_i'\beta| < \lambda_\gamma)$ to (2.1). However, it does not affect the value of the original objective function in (2.1) because $\gamma_i\epsilon_i = 0$. Now, as $\gamma_i^2 + \gamma_i\epsilon_i = \gamma_i(y_i - x_i'\beta)$ from the model (1.2), the objective function in (2.1) is equivalent to

$$L(\beta, \gamma) = \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i'\beta - \gamma_i)^2 + \sum_{i=1}^{n}\gamma_i(y_i - x_i'\beta)I(|y_i - x_i'\beta| < \lambda_\gamma). \tag{2.2}$$

Now, the minimization of $L(\beta, \gamma)$ becomes more apparent. We then adopt an iterative algorithm to estimate β and γ alternately. Due to the iterative fashion, it is noteworthy that the degrees of freedom for estimating regression coefficients in each step are the same as those of OLS without considering variable selection procedure. That is, we adjust the y_is with the estimates of γ, then find the estimates of β with the adjusted y_is. Therefore, incorporating the case-specific parameters γ_is does not increase model complexity, but requires additional computational cost of an iterative algorithm. However, the number of iteration is minimal (< 10) in practice. For the details of algorithm and the adjustment of y_is, we first initialize $\hat{\gamma}^{(0)} = 0$, and find $\hat{\beta}^{(0)} = \text{argmin}_\beta\, L(\beta, 0)$, which is just the ordinary least squares estimate. We denoted it as $\hat{\beta}^{(0)} = \hat{\beta}^{OLS}$.

Note that the degrees of freedom (df) for OLS model has been traditionally defined as the number of independent residuals. That is, when we use p variables with sample

size n, *df* is $(n-p)$ without intercept. When we remove m outliers for better fit, the *df* will be decreased to $(n-m-p)$ for OLS model. In contrast, *df* of OLS^S is maintained as $(n-p)$ with possibly removing major effect of m outliers. Of course, this statement holds only when we do not consider variable selection procedure (or, only when we employ the same p covariates for both OLS and OLS^S). When we adopt variable selection as a part of modelling, then the selected variables for OLS and OLS^S can be different, therefore their *df* may not be equal. Since the variable selection is embedded when incorporating a lasso penalty, the *df* of lasso regression can be different from that of the suggested methods with lasso regression.

Next, $\hat{\gamma}$ can be updated conditional on $\hat{\beta}^{(0)}$ via minimizing

$$L(\hat{\beta}^{(0)}, \gamma) = \frac{1}{2}\sum_{i=1}^{n}(r_i - \gamma_i)^2 + \sum_{i=1}^{n}\gamma_i r_i I(|r_i| < \lambda_\gamma),$$

where $r_i = y_i - x_i'\hat{\beta}^{(0)}$. By solving $\frac{\partial}{\partial \gamma_i}L(\hat{\beta}^{(0)}, \gamma) = 0$ for each i, we have

$$\hat{\gamma}_i^{(1)} = \begin{cases} r_i & \text{if} \quad |r_i| \geq \lambda_\gamma, \\ 0 & \text{if} \quad |r_i| < \lambda_\gamma. \end{cases} \tag{2.3}$$

Now, given $\hat{\gamma}^{(1)}$, we have

$$L(\beta, \hat{\gamma}^{(1)}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i'\beta - \hat{\gamma}_i^{(1)})^2 + \sum_{i=1}^{n}\hat{\gamma}_i^{(1)}(y_i - x_i'\beta)I(|y_i - x_i'\beta| < \lambda_\gamma). \tag{2.4}$$

Note that the second term of right hand side in (2.4) will be zero when $|y_i - x_i'\beta| \geq \lambda_\gamma$. When $|y_i - x_i'\beta| < \lambda_\gamma$, we have $\hat{\gamma}_i^{(1)} = 0$ by (2.3). Thus, $L(\beta, \hat{\gamma}^{(1)})$ in (2.4) is rewritten as $\frac{1}{2}\sum_{i=1}^{n}(y_i - x_i'\beta - \hat{\gamma}_i^{(1)})^2$. We can easily get its minimizer $\hat{\beta}^{(1)} = (X'X)^{-1}X'(Y - \hat{\gamma}^{(1)})$. Again, note that this is a convex minimization with given estimates of γ. We call this estimating method "*outlier shifting* least squares". The tuning parameter λ_γ needs to be pre-specified and we defer its estimation to Section 2.2.

The convergency of the iterative algorithm for convex function minimization is widely studied by Rockafellar (1997). Specifically, the convergency with case-specific parameters is discussed in Lee et al. (2012). Proposition 3 in Lee et al. (2007) can be directly applied to show the convergence of Algorithm 1 to unique minimizer (β^*, γ^*) under the objective function in (2.1).

Next, we summarize the algorithm that can provide more general application beyond the discussed squared error loss function.

Algorithm 1:

1. Initialize $\hat{\gamma}^{(0)} = 0$, and obtain $\hat{\beta}^{(0)} = \arg\min_\beta L(\beta, 0)$, and $Y^{(0)}$.
2. Update $\hat{\gamma}^{(m+1)}$ by finding the minimizer of $L(\hat{\beta}^{(m)}, \gamma)$ with $Y^{(m)}$.
3. Update $Y^{(m+1)}$ with $Y^{(m)} - \hat{\gamma}^{(m+1)}$.
4. Update $\hat{\beta}^{(m+1)}$ by finding the minimizer of $L(\beta, \gamma^{(m+1)})$ with $Y^{(m+1)}$.
5. Iterate between step 2 and step 4 until $||\beta^{(m+1)} - \beta^{(m)}||^2$ is small.

But, there are certain restrictions of the application such as binomial loss where Y is not quantitative. In this case, it is hard to understand and/or interpret the adjustment in step 3. We set $Y^{(0)}$ to be the original observations of the response variable. Note that the estimates of γ are used to modify the original observations in step 3. Thus, when we estimate β in step 4, the degrees of freedom for model is not affected by the number of nonzero estimate of γ, but the outlying observations are modified only. Once y_i is judged as an outlier, then the selected y_is are modified as $y_i' = y_i - \hat{\gamma}_i = y_i - r_i$. As we subtract the residual from the original observation, it shifts the outlier onto the fitted model. If it is not selected, that is, if $\hat{\gamma}_i = 0$,

then there is no adjustment in y_i. As we employ an iterative algorithm which finds the minimizer of β and γ alternately, this adjustment can happen multiple times as shown in Figure 1. This reflects certain advantage of introducing the case-specific parameters to control outliers rather than merely removing them.

A toy example in Figure 1 illustrates how the fitted line and $y_i^{(m)}$ are changed as we iterate the estimation procedure described in algorithm 1.

In Figure 1, only y_1 is detected as an outlier, and is adjusted as $y_1 - \hat{\gamma}_1^{(1)} - \hat{\gamma}_1^{(2)}$ after two iterations of algorithm 1. When the algorithm is converged after two iterations, the final fitted line (solid line) is very close to the underlying truth (red solid line). We also show the fitted line from OLS after removing y_1 (dashed line). In this simple example, removing an observation improves model fit at the sacrifice of one degrees of freedom. In contrast, the proposed method illustrates how the huge outlier can be adjusted effectively without being disregarded. In Figure 1, we observe that the obtained $\hat{\gamma}_1^{(1)}$ and $\hat{\gamma}_1^{(2)}$ have the effect of shifting the potential outlier y_1 towards the fitted line. Thus, we call the suggested penalty in (2.1) *outlier-shifting penalty*.

Remark 1: When we replace the penalty for γ_i in (2.1) with $\lambda_\gamma \sum_{i=1}^n |\gamma_i|$, it will result in the estimator in Lee et al. (2012). In this case, we have $\hat{\gamma}_i = sgn(r_i)(|r_i| - \lambda_\gamma)_+$. The idea of Lee et al. (2012) is good in the sense that outliers are detected and diminished in an automatic fashion. However, large residuals are (even if they are all large) in various size, and the universal shrinkage by λ_γ does not seem to be the most effective way. Extremely large outliers often have high impact on the fitted model after being reduced by λ_γ. Using extremely large value of λ_γ could be a solution, but then, we cannot treat moderately large outliers. This is the main motivation of the suggested penalty in this paper.

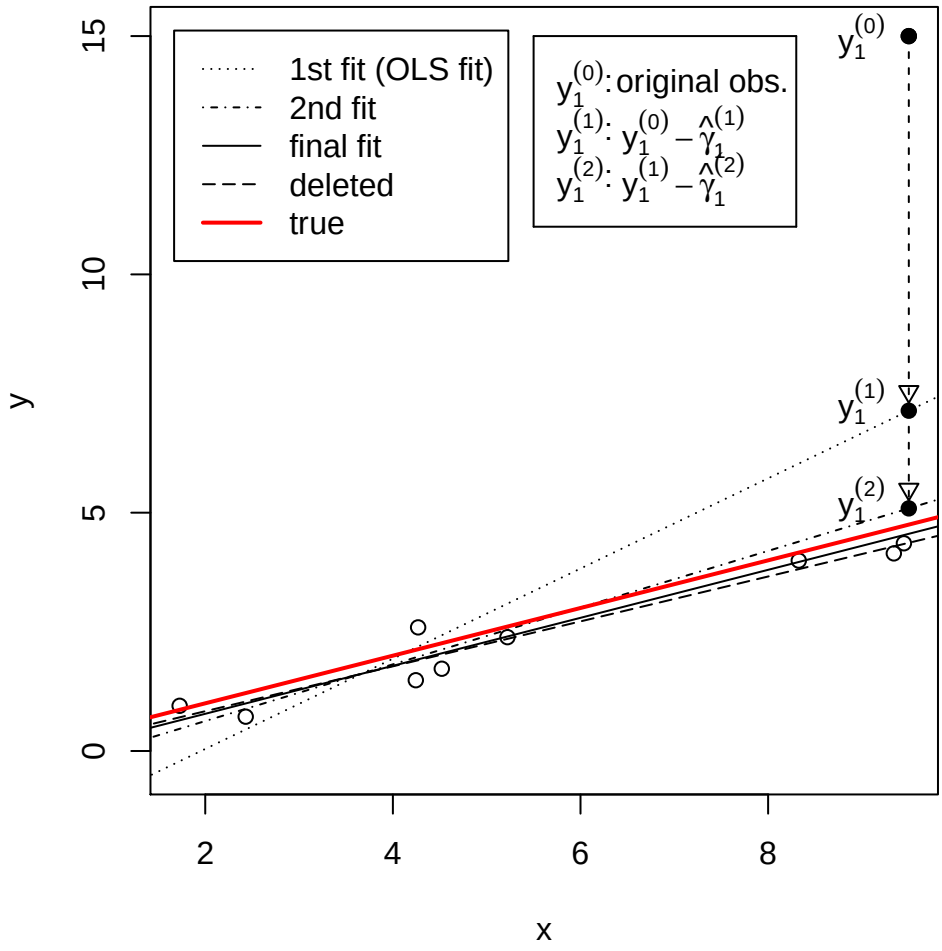

Figure 1: A toy example for the illustration of Algorithm 1 with 10 observations where the first observation y_1 is highly corrupted. '1st fit' is the initial fit with unadjusted observations, and '2nd fit' is the fit after 1 iteration. The 'final fit' is from the estimates when the algorithm 1 is converged after 2 iterations. 'deleted' stands for the fitted line by *OLS* after removing y_1. The thick solid line (in red) represents the underlying truth.

2.2 A default value of λ_γ

Because we target detecting highly contaminated data by non-zero γ_is, it is reasonable to assume that ϵ_is are not contaminated and normally distributed. Thus, we propose to select λ_γ such that

$$P(|\epsilon_i| \geq \lambda_\gamma) \leq n_o/n, \tag{2.5}$$

where n_o is the size of highly contaminated y_is. As we are interested in highly corrupted data, we assume the proportion of outliers do not disappear. That is, $n_o/n > 0$ as $n \to \infty$.

Under the assumption of $\epsilon_i \sim N(0, \sigma^2)$, the inequality (2.5) is equivalent to

$$\lambda_\gamma \geq \sigma \Psi^{-1}\left(\frac{2n - n_o}{2n}\right),$$

where $\Psi^{-1}(\cdot)$ is the inverse C.D.F. of the standard normal distribution. Herein, σ should be estimated and so does the number of outliers n_o. Since λ_γ is required to be specified prior to the model estimation procedure, it is desirable to use a robust estimate for σ, such as interquartile range, median absolute deviation (MAD), or robust scale estimators suggested by Rousseeuw and Croux (1993). Investigation of the residuals from OLS often suggests reasonable number of outliers n_o. Then we set λ_γ to be the following value,

$$\hat{\lambda}_\gamma = \hat{\sigma} \Psi^{-1}\left(\frac{2n - \hat{n}_0}{2n}\right). \tag{2.6}$$

The above default value, which is devised to catch highly contaminated observations, differs from a conventional rule of defining outliers as a fixed percentage (typically, 1% or 5%) of the observations. Instead, $\hat{\lambda}_\gamma$ is adaptive to the data through the proportion of outliers n_o/n.

2.3 Asymptotic properties

We investigate the asymptotic properties of $\hat{\beta}^{(m)}$ under the outlier-shifting least squares and show its limiting distribution. Note that the classical linear model in (1.1) has *iid* errors, which is far from the reality when outliers exist. To incorporate the outliers in the error term, errors should be a mixture of *iid* errors and outliers. For this purpose, we may use an error of $\epsilon_i^* = (1-\alpha)\epsilon_i + \alpha\xi_i$ where ϵ_i is *iid* errors with mean 0 and finite variance, and ξ_is are independent and follow some heavy tail distribution with finite mean and variance. Finally, $\alpha(> 0)$ is the proportion of outliers. Note that ϵ_i^*s are not necessarily identical nor symmetric. Then, under the model of

$$y_i = x_i'\beta + \epsilon_i^*, \tag{2.7}$$

we have $E(\hat{\beta}^{OLS}) = \beta + (X'X)^{-1}E(X'\epsilon^*)$. For $\hat{\beta}^{OLS}$ to be consistent, $E(X'\epsilon^*)$ needs to approach to zero with probability 1. However, this is not the case when the distribution of ξ_i does not have mean zero. So, we impose a restriction on ϵ_i^* that the mean of ϵ_i^* tends to zero with probability 1. This implies that asymptotically equal amount of outliers exist in each tail. Then, under some mild conditions on X, $\hat{\beta}^{OLS}$ is consistent by Linderberg-Feller type of central limit theorem under independent but not identical errors. Here are the list of conditions we consider.

CONDITIONS

1. $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} x_i x_i' = Q$, where Q is a non-negative definite matrix.
2. $E(x_i) < \infty$ for all i.
3. $\lambda_\gamma = O(1)$
4. $E(\xi_i) \to 0$ w.p. 1.

Condition 1 and 2 are the standard moment conditions for X. Condition 3 arises from the assumption that the proportion of outliers n_o/n approaches to a non-zero constant discussed in Section 2.2. Condition 4 implies the amount of outliers in each tail is asymptotically equal. In practice, the proposed method works well when all the outliers are in one tail as they will be effectively shifted. For the asymptotic results, we need further clarification. When the number of outliers are ignorable ($n_o/n \to 0$), then we do not need the condition 4. Since $n_o = \sqrt{n}$ is also ignorable, we may not need condition 4 in many cases except for the case of $n_o/n \to c$, where $c > 0$. Then, the proposed estimator is $\sqrt{n}$-consistent as shown in Theorem 1.

Theorem 1 *Assuming conditions 1-4 hold, we have, for any m,*

$$\sqrt{n}(\hat{\beta}^{(m)} - \beta) \xrightarrow{d} N(0, \sigma^2 Q^{-1}).$$

Proof. See Appendix.

3 Other applications of shifting outliers

3.1 Outlier-shifting in robust loss function

A well-known robust approach for improving the least squares method is Huber's loss function:

$$\rho^H(u) = \frac{u^2}{2} I(|u| \le c) + c(|u| - \frac{c}{2}) I(|u| > c). \tag{3.1}$$

Even for this robust loss function, we observe from our empirical investigations that it is still possible to improve the performance in the presence of massively contaminated

observations. Therefore, we add the case-specific parameters with the *outlier-shifting penalty* to Huber's loss function, resulting in the following outlier-shifting Huber's loss function;

$$L(\beta, \gamma) = \sum_{i=1}^{n} \rho^H(y_i - x_i'\beta - \gamma_i) + \sum_{i=1}^{n} \gamma_i^2 I(|y_i - x_i'\beta| < \lambda_\gamma).$$

Algorithm 1 can be implemented to estimate the parameters. For the threshold value c in (3.1), we set it scale invariant as $c = c_0\sigma$, and also use a robust scale estimator for σ as discussed in Section 2.2. We conduct extensive numeric comparisons among the estimates from ordinary least squares, outlier-shifting least squares, Huber's loss, and outlier-shifting Huber's loss in Section 4.

The Huber's approach shares the idea of M-quantile (Breckling and Chambers, 1988) for robust regression. The two methods has the same form of loss function when M-quantile targets conditional median. Thus, we naturally see that M-quantile regression might get benefit by adopting the proposed *outlier-shifting penalty* for the quantile regression. We may need to adjust the form of the proposed penalty due to the asymmetric loss function in quantile regression except for the median. The details of adopting the proposed penalty in the light of quantile regression will not be considered here, here is the sketch of the idea. Under the linear model of $y_i = x_i'\beta + \epsilon$, quantile regression estimator for qth quantile is defined as the minimizer of the check loss function;

$$L_q(\beta) = \sum_{i=1}^{n} (y_i - x_i'\beta)\{q - 1 + I(y_i - x_i'\beta \geq 0)\}.$$

Then, we can incorporate an asymmetric outlier shifting penalty with case-specific parameter γ_i as

$$\begin{aligned} L_q(\beta, \gamma) \quad &= \textstyle\sum_{i=1}^{n} (y_i - x_i'\beta - \gamma_i)\{q - 1 + I(y_i - x_i'\beta - \gamma_i \geq 0)\} \\ &+ \textstyle\sum_{i=1}^{n} \gamma_i q I(0 < y_i - x_i'\beta < \lambda_\gamma) + \sum_{i=1}^{n} \gamma_i (1-q) I(-\lambda_\gamma < y_i - x_i'\beta < 0), \end{aligned}$$

which may result in robustified quantile regression.

3.2 Outlier-shifting in Lasso-type models

In this section, we switch to the often encountered high dimensional problem with a large number of covariates. To address this problem, the lasso type of penalties is widely used. The idea of case-specific parameters and shifting outlier penalty is directly applicable to lasso (Tibshirani, 1996) and Huberized lasso (Rosset and Zhu, 2007) to improve robustness. lasso is defined as finding a minimizer of

$$L(\beta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda_\beta \sum_{j=1}^{p}|\beta_j|. \tag{3.2}$$

We denote $\hat{\beta}^{lasso}$ as the minimizer of (3.2). By introducing case-specific parameters γ and the outlier-shifting penalty used in (2.1), the outlier-shifting lasso procedure is defined as finding the minimizer of

$$L(\beta,\gamma) = \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i'\beta - \gamma_i)^2 + \lambda_\beta \sum_{j=1}^{p}|\beta_j| + \sum_{i=1}^{n}\gamma_i^2 I(|y_i - x_i'\beta| < \lambda_\gamma). \tag{3.3}$$

To estimate β and γ, algorithm 1 in Section 2.1 can be directly implemented. However, the L1-type penalty on β is non-differentiable. Thus, we resort to the coordinate decent type of methods (Wu and Lange, 2008).

With initial values of $\hat{\gamma}^{(0)} = 0$ and $\hat{\beta}^{(0)} = \hat{\beta}^{lasso}$, we proceed to iterate updating $\hat{\gamma}$ and $\hat{\beta}$ alternately as in the algorithm 2. At the $(m+1)$th iteration, the outlier-shifting lasso estimates are

$$\begin{aligned}
\hat{\gamma}_i^{(m+1)} &= (y_i - x_i'\hat{\beta}^{(m)})I(|y_i - x_i'\hat{\beta}^{(m)}| \geq \lambda_\gamma), \text{ for } i = 1,\ldots,n, \\
\hat{\beta}^{(m+1)} &= \arg\min_{\beta\in\mathbb{R}^p}\{\frac{1}{2}\sum_{i=1}^{n}(y_i - x_i'\beta - \hat{\gamma}_i^{(m+1)})^2 + \lambda_\beta\sum_{j=1}^{p}|\beta_j|\}.
\end{aligned}$$

We use the same estimated value of λ_γ provided in Section 2.2.

Rosset and Zhu (2007) considered Huberized lasso where the squared error loss in (3.2) is replaced by Huber's loss function. In fact, Huberized lasso is coincidentally the same as what proposed by Lee et al. (2012) with case-specific parameters. Analogous to Section 3.1, we can add the outlier-shifting penalty to Huberized lasso to intensify the robustness. Then, the *outlier-shifting* Huberized lasso is defined as the minimizer of

$$L(\beta, \gamma) = \frac{1}{2}\sum_{i=1}^{n} \rho^H(y_i - x_i'\beta - \gamma_i) + \lambda_\beta \sum_{j=1}^{p} |\beta_j| + \sum_{i=1}^{n} \gamma_i^2 I(|y_i - x_i'\beta| < \lambda_\gamma), \quad (3.4)$$

where $\rho^H(\cdot)$ is given in (3.1). The performance of lasso, outlier-shifting lasso, Huberized lasso, and outlier-shifting Huberized lasso is investigated through the simulation studies in the next section.

Remark: Although we choose lasso penalty for variable selection, the choice of penalty for β is not a crucial part. It is obvious that other penalties for β can be combined with outlier-shifting penalty. For example, smoothly clipped absolute deviation penalty (SCAD) proposed by Fan and Li (2001) can replace lasso penalty.

3.3 Connection to maximum likelihood approach

Now, treating the negative log likelihood as a loss function, we consider a specific form of Algorithm 1 with normal likelihood. Because γ_is are employed to account for highly corrupted outliers, we may assume normally distributed ϵ_is in model (1.2). Since γ_is explain extreme outliers, the normality assumption is reasonable. And by the constraint given in Section 1, ϵ_is are zero when γ_is are non-zero. Then, the

likelihood is simply

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - x_i'\beta - \gamma_i)^2}{2\sigma^2}\right\}.$$

The maximum likelihood estimate of β, σ^2, and γ can be written as

$$\begin{aligned}
\hat{\beta} &= (X'X)^{-1}X'(Y - \hat{\gamma}), \\
\hat{\sigma}^2 &= ||Y - X\hat{\beta} - \hat{\gamma})||^2/n, \\
\hat{\gamma} &= Y - X\hat{\beta}.
\end{aligned}$$

Note that $\hat{\gamma} = (\gamma_1, \ldots, \gamma_n)'$ are in fact the residuals, which take non-zero values for both the outliers and non-outlying observations. To be coherent with the description of γ_i in Section 1, $\hat{\gamma}_i$ is modified to be $\hat{\gamma}_i = (y_i - x_i'\hat{\beta})I(|y_i - x_i'\hat{\beta}| \geq \lambda_\gamma)$, for i=1, ..., n, where λ_γ is a pre-specified threshold. With this modified $\hat{\gamma}$, we define an updated response variable $y_i^* = y_i - \hat{\gamma}_i$. It implies that the observations of $\mathbf{y} = (y_1, \ldots, y_n)'$ are treated as outliers and accordingly adjusted only if $|y_i - x_i'\hat{\beta}|$ is larger than λ_γ. Or equivalently, $y_i^* = x_i'\hat{\beta}$ if $|r_i| \geq \lambda_\gamma$; and $y_i^* = y_i$, otherwise. This adjustment is to shift the outlying observations onto the fitted values, while preserving the other observations. Treating the modified y_i^* as a new observation, we repeat the estimation procedure to update $\hat{\beta}$, $\hat{\sigma}^2$, and $\hat{\gamma}$. Here is the detailed algorithm of the estimation procedure.

Algorithm 2:

1. Initialize $\beta^{(0)}$ and $\sigma^{2(0)}$ with their least square estimates, set $\hat{\gamma}^{(0)} = 0$, and use the original observations of $\mathbf{y}$ as $Y^{(0)}$.

2. Update $\hat{\gamma}_i^{(m+1)}$ with $(y_i^{(m)} - x_i'\hat{\beta}^{(m)})I(|y_i^{(m)} - x_i'\hat{\beta}^{(m)}| \geq \lambda_\gamma)$, for $i = 1, \ldots, n$.

3. Update $Y^{(m+1)}$ with $(Y^{(m)} - \hat{\gamma}^{(m+1)})$, where $\hat{\gamma}^{(m+1)} = (\hat{\gamma}_1^{(m+1)}, \ldots, \hat{\gamma}_n^{(m+1)})'$.

4. Update $\hat{\beta}^{(m+1)}$ with $(X'X)^{-1}X'Y^{(m+1)}$.

5. Update $\hat{\sigma}^{2(m+1)}$ with $||Y^{(m+1)} - X\hat{\beta}^{(m+1)}||^2/n$.

6. Iterate steps 2 to 5 until $||\beta^{(m+1)} - \beta^{(m)}||^2$ is small.

Remark 2: Notice that the obtained $\hat{\gamma}$ is identical to (2.3), which is estimated with the outlier-shifting penalty described in the previous section. Therefore, this provides an intuitive interpretation of the outlier-shifting penalty and as well as an alternative algorithm to optimize the objective function (2.1).

Interestingly, this outlier adjustment method also shares some similarity with the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) in terms of shrinkage towards the fitted value in an iterative manner. A clear distinction between the two is that the the outlier adjustment method automatically determines the *unidentified outliers* via the tuning parameter λ_γ, while the missing observations are *identified* in advance in the EM algorithm.

4 Simulation Studies

We conduct two sets of simulations, investigating various sizes and proportions of contaminated observations. The second set is in the supplementary materials. In the simulations, we compare the suggested estimators to existing methods. They are (i) ordinary least squares (OLS); (ii) outlier-shifting least squares (OLS^S); (iii) Huber's estimator (H); and (iv) outlier-shifting Huber's estimator (H^S), (v) median regression (med), (vi) $lasso$, (vii) outlier-shifting lasso ($lasso^S$); (viii) Huberized lasso

($Hlasso$); and (ix) outlier-shifting Huberized lasso ($Hlasso^S$). The proposed outlier-shifting penalty based approaches are notated with superscript 'S'.

For OLS^S, we take the pre-specified value in (2.6) as λ_γ, wherein $MAD/0.6745$ of residuals from fitting the median regression is used as the robust estimator of σ. The median absolute deviation (MAD) should be scaled to be a scale estimator. Thus, we divided it by 0.6745, which is known to be a robust estimator of σ when the errors does not follow normal distribution (Huber and Ronchetti, 2009). For Huber's function defined in (3.1), we report the result at c=1.5 that is observed to have the best empirical performance among various examined values of c. To implement H, we use `rlm` function in `MASS` package. For the implementation of *lasso*, `glmnet` function in `glmnet` package is utilized. For the selection of penalty parameter in *lasso*, we use 10-fold cross-validation with 100 candidate values of the penalty parameter, λ_β.

In all the simulation studies, we generate $K = 500$ data sets with sample size $n = 100$ from the model,

$$Y = X\beta + \sigma\epsilon,$$

where the components of X and ϵ follow standard normal distribution. We use $\sigma = 1$ for the base model. We consider contaminating 10%, 20%, and 30% of the response in the base model with $\sigma = 3, 6$, and 10, respectively, which produces 9 different combinations of the scenarios. Thus, the generated errors follow mixture normal distribution of $(1-\pi)N(0,1) + \pi N(0,\sigma)$ where $\pi = 0.1$, 0.2, and 0.3. For each case, we report the mean squared error (MSE) between the true and the fitted regression lines over K simulated data set to assess the performance of each estimator, $MSE = \sum_{k=1}^{K}(\hat{\beta}^k - \beta)'E(X'X)(\hat{\beta}^k - \beta)/K$, where $\hat{\beta}^k$ is the parameter estimate from the kth simulated data set.

Simulation set 1: We consider $p = 8$ covariates and set $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)'$ and take the correlation between x_i and x_j to be $\varrho^{|i-j|}$ with ϱ=0.5, which are also used in Tibshirani (1996). (The results based on orthogonal covariates with ϱ=0 is in supplementary material.) The MSE values obtained by the nine estimators are provided in Table 1.

As expected, we see that OLS is the most vulnerable to outliers. MSE values obtained by the methods of OLS^S, H, and H^S are similar when 10% of the data is contaminated with $\sigma = 3$. As the coverage and the magnitude of contamination increases, the advantage of the proposed H^S and $Hlasso^S$ over the other methods becomes much more prominent; it is not surprising that H performs well. For the base model without outliers, OLS, OLS^S, H, and H^S yield MSE values at the similar levels of 0.0927 0.0924 0.0937, and 0.1012, respectively. In summary, H^S and $Hlasso^S$, which are doubly robustified by Huber's loss function and the proposed outlier-shifting penalty, manifest the most benefit for a severely contaminated data set. We also consider a nonparametric robust approach-median regression (Koenker, 2005)-for comparison in this simulation study. It pursues the conditional median regression model rather than the mean models considered thus far. However, the mean and median are equivalent in the considered scenario of symmetric error distributions. The median regression approach is always inferior to H^S and superior to $lasso^S$, and fall between OLS^S and H most of the time. Because the median is known to be less efficient than the mean under normal distributions, we can expect that the median regression obtains larger variances of parameter estimates than the mean regression. The phenomenon is demonstrated by Figure 2, which shows the distributions of the regression coefficient estimates obtained by the investigated five methods. We see that the estimates of the median regression approach have slightly larger variance than that

Table 1: Point estimate of MSE, and its standard error (in parentheses) multiplied by 1000 based on 500 simulated data sets obtained by OLS, OLS^S, H, H^S, med, $lasso$, $lasso^S$, $Hlasso$, and $Hlasso^S$ under $\sigma = 3$, 6, and 10 with the contaminated proportions of 10%, 20%, and 30%.

	10%	20%	30%
		$\sigma = 3$	
OLS	161.6 (4.3)	234.7 (6.1)	312.0 (8.1)
OLS^S	122.8 (3.5)	165.1 (4.7)	220.8 (6.4)
H	121.3 (3.0)	159.5 (4.1)	212.2 (5.7)
H^S	126.3 (3.2)	156.0 (4.2)	197.5 (5.3)
med	168.1 (4.1)	200.3 (4.9)	243.2 (6.3)
$lasso$	131.1 (4.2)	191.9 (5.9)	250.3 (7.7)
$lasso^S$	103.1 (3.2)	136.2 (4.2)	177.3 (5.7)
$Hlasso$	101.7 (3.2)	136.9 (4.3)	181.7 (5.8)
$Hlasso^S$	102.3 (3.1)	129.3 (4.0)	167.8 (5.3)
		$\sigma = 6$	
OLS	393.7 (11.6)	712.3 (19.2)	1045 (27.0)
OLS^S	198.5 (6.3)	341.7 (10.8)	536.8 (16.5)
H	134.4 (3.4)	205.9 (5.5)	340.5 (9.9)
H^S	128.3 (3.2)	159.4 (4.0)	223.3 (6.2)
med	175.0 (4.3)	221.5 (5.5)	296.4 (7.9)
$lasso$	320.5 (11.3)	585.8 (19.1)	845.7 (26.2)
$lasso^S$	127.8 (4.1)	194.6 (6.8)	288.5 (10.4)
$Hlasso$	127.6 (4.1)	218.0 (7.1)	358.6 (12.2)
$Hlasso^S$	107.6 (3.3)	143.7 (4.5)	212.6 (7.6)
		$\sigma = 10$	
OLS	943.7 (29.2)	1843 (51.0)	2776 (71.4)
OLS^S	394.2 (13.8)	776.7 (26.1)	1286 (40.5)
H	141.3 (3.6)	234.7 (6.3)	450.7 (13.7)
H^S	128.1 (3.2)	162.3 (4.1)	248.3 (7.3)
med	178.2 (4.5)	230.6 (5.8)	321.6 (8.8)
$lasso$	774.8 (28.7)	1532 (51.9)	2276 (70.0)
$lasso^S$	204.5 (8.8)	356.4 (15.3)	553.3 (21.7)
$Hlasso$	171.9 (5.8)	368.3 (13.7)	724.3 (26.3)
$Hlasso^S$	111.1 (3.6)	161.5 (5.5)	271.4 (10.3)

Table 2: Average false negative rate of three nonzero coefficients and average false positive rate of five zero coefficients (in parenthesis) based on 500 simulated data sets obtained by *lasso*, $lasso^S$, *Hlasso*, and $Hlasso^S$ with $\sigma = 6$, and 10 with contaminated proportions of 10%, 20%, and 30%.

	$\sigma = 6$			$\sigma = 10$		
	10%	20%	30%	10%	20%	30%
lasso	0 (.445)	0 (.443)	0 (.466)	0.001 (.447)	0.005 (.435)	0.024 (.457)
$lasso^S$	0 (.519)	0 (.548)	0 (.581)	0 (.570)	0 (.572)	0 (.583)
Hlasso	0 (.471)	0 (.501)	0 (.508)	0 (.517)	0 (.522)	0 (.531)
$Hlasso^S$	0 (.492)	0 (.548)	0 (.579)	0 (.517)	0 (.568)	0 (.626)

of H^S. Overall, the difference in the variances of the five estimators enlarges as the level of contamination increases. And H^S produces the smallest variance among the five methods as shown in Figure 2.

As the four regression parameter penalized methods (*lasso*, $lasso^S$, *Hlasso*, and $Hlasso^S$) can produce sparse solutions, we examine the false negative rates and false positive rates. Note that there are three nonzero coefficients in true β. We record the average probability of identifying the nonzero coefficient from 500 simulated data sets in Table 2. We also present the average false positive rate of the five truly zero coefficients in the parentheses in the same table. As the standard errors of the average false positive rates all around 0.01, they are not presented. From table 2, the four methods select more variables than they should as the false positive rates are considerably high. As a consequence, the true positive rates are all close to 1. The false discovery rates from the suggested methods are higher than their competitors.

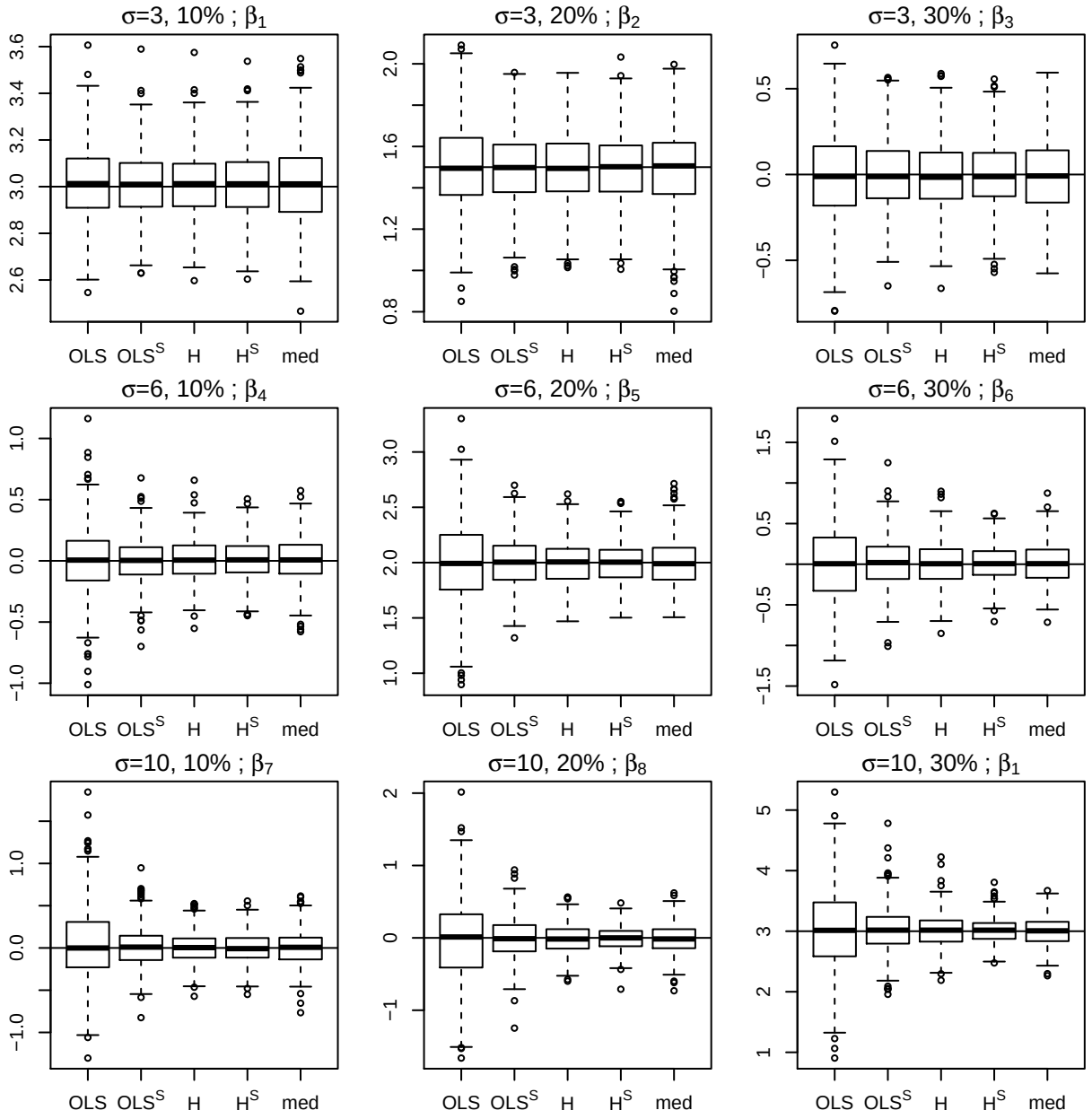

Figure 2: Distribution of estimates from 500 simulated samples under the simulation set 1. See the text for the details of simulation setting. Horizontal lines represent the true regression coefficient.

Finally, we examine the performance of $\hat{\gamma}_i$s. Note that we generate *iid* errors from $N(0,1)$ first. Then, 10% (or, 20%, 30%) of the errors (out of n) are multiplied by 3 (or, 6, 10). Thus, the generated errors follow mixture normal distribution of $(1-\pi)N(0,1)+\pi N(0,\sigma)$ where $\pi = 0.1$, 0.2, and 0.3, and $\sigma = 3$, 6, and 10 are considered. It is very possible that some errors are still quite small after being multiplied by 10. Thus, we regard ϵ_is that satisfy $|\epsilon_i| > 2.5$ as outliers. As the sample size $n = 100$ and $\sigma = 1$ in the data set without contamination, we would expect to observe about one ϵ_i which satisfy the above criterion under the assumed normality. Then, we count the number of non-zero $\hat{\gamma}_i$s that captures the outliers. As the contaminated proportion increases, we expect larger number of outliers are generated. Thus, we report the average percentage of correct detection of the outliers in Table 3. As standard errors are all small (around 0.01, or smaller), we omit them in the table. In general, as the contaminated proportion decreases, and as the contaminated magnitude increases, we see higher rate of identifying the outliers.

Table 3: Averaged probability of identifying the outliers with $|\epsilon_i| > 2.5$ based on 500 simulated data sets obtained by OLS^S, H^S, $lasso^S$, and $Hlasso^S$ with contaminated proportions of 10%, 20%, and 30% and magnitude of $\sigma = 3$, 6, and 10.

	$\sigma = 3$			$\sigma = 6$			$\sigma = 10$		
	10%	20%	30%	10%	20%	30%	10%	20%	30%
OLS^S	0.82	0.78	0.70	0.88	0.83	0.73	0.87	0.81	0.71
H^S	0.88	0.85	0.78	0.94	0.92	0.86	0.96	0.94	0.89
$lasso^S$	0.87	0.85	0.77	0.92	0.88	0.81	0.91	0.87	0.79
$Hlasso^S$	0.59	0.56	0.51	0.74	0.73	0.69	0.88	0.88	0.82

5 Applications

5.1 *Analysis of stack loss data*

The stack loss data set (Brownlee, 1965) has been examined by many researchers (Daniel and Wood, 1980; Chambers and Heathcote, 1981; Carroll and Ruppert, 1985). The details of pre-analysis can be found in Chapter 5.4 of Daniel and Wood (1971), and using non-linear covariates are included in Denby and Mallows (1977).

The data set has 21 observations with 3 explanatory variables. The data describe the operation of a plant. Researchers have sought to explain the percentage of unconverted ammonia that escapes from the plant (Y) by the flow of cooling air (X_1), inlet temperature of cooling water (X_2), and concentration of acid (X_3). Thus, we consider the following standard linear model;

$$y_i = \beta_0 + \sum_{j=1}^{3} \beta_j x_{ij} + \epsilon_i, i = 1, \ldots, n,$$

where ϵ_is are iid random variable with mean 0 and variance σ^2. Most people concluded that the observations 1, 3, 4, and 21 are outliers. Chambers and Heathcote (1981) and Hoeting et al. (1996) suggested deleting the two observations of 4 and 21 can improve the fitted surface. Thus, we compose three different data sets of containing 1) all 21 observations, 2) 19 observations (observations 4 and 21 are deleted), and 3) 17 observations (observations 1, 3, 4 and 21 are deleted), which were examined by Chambers and Heathcote (1981). We fit *outlier-shifting* least squares (OLS^S) model to these three data sets then, we record the change in the estimates of the regression coefficients. To estimate the scale parameter, the median absolute deviation of the residuals from fitting a median regression is utilized. For comparison, the estimates

from OLS, H, and Tukey's robust regression ($Tukey$) by (Beaton and Tukey, 1974) are examined. The loss function for Tukey's robust regression is defined as,

$$\rho^T(u) = \frac{1}{6}[1 - \{1 - (u/(6c))^2\}^3]I(\mid u \mid \leq 6c). \tag{5.1}$$

The values used for the bending constant c in $\rho^H(u)$ and $\rho^T(u)$ are 1.345 and 4.685, respectively. Huber and Ronchetti (2009) argued that $c = 1.345$ in $\rho^H(c)$ is a good choice, and showed that asymptotically, it is 95% as efficient as least squares if the true distribution is normal. Similarly, $c = 4.645$ in $\rho^T(c)$ provides 95% of asymptotic efficiency under the normal distribution. `rlm` function in R package `MASS` is used for implementing H and Tukey's robust regression. The estimates from the above four methods are represented in Table 4.

Table 4: Estimates of β_0, β_1, β_2, β_3 for the stack loss data. data 1, data 2, and data 3 stand for the data set with 21, 19, and 17 observations, respectively.

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
		data 1		
OLS	17.524	6.561	4.094	-0.815
H	17.596	7.604	2.927	-0.685
$Tukey$	17.800	8.504	2.057	-0.602
OLS^S	17.112	7.614	1.781	-0.387
		data 2		
OLS	17.688	8.770	1.756	-0.583
H	17.593	8.600	1.702	-0.581
$Tukey$	17.625	8.673	1.703	-0.603
OLS^S	17.228	7.914	1.431	-0.517
		data 3		
OLS	16.943	7.313	1.825	-0.359
H	17.006	7.503	1.643	-0.397
$Tukey$	17.018	7.508	1.636	-0.390
OLS^S	17.131	7.722	1.441	-0.484

To measure the *sensitivity* of the estimates as we remove two and four observations, the absolute difference between the minimum and the maximum estimates among the three data sets are calculated. For example, *sensitivity* of OLS for β_1 is |6.561 - 8.770| = 2.209. Therefore, estimator with large value of *sensitivity* indicates that the estimator is sensitive to outliers. Three variables are standardized to have scale invariant sensitivity. The *sensitivity* of the estimates from the five methods are demonstrated in Table 5.

Among the four aforementioned methods, the proposed method is the most stable; That is, the amount of change in the estimates is the smallest for the suggested method as we change the sample size by removing two and four outliers.

Table 5: *Sensitivity* of five estimators in Table 4 at β_1, β_2, β_3 with the stack loss data.

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
OLS	2.2092	2.3381	0.4558
H	1.0967	1.2842	0.2878
$Tukey$	1.1642	0.4205	0.2125
OLS^S	0.3005	0.3498	0.1293

Figure 3 shows the residual plots from the four methods of OLS, H, $Tukey$, and OLS^S with all observations. In the lower right panel, the residuals from the observation 1, 3, 4, and 21 (marked with *) are very close to zero due to shifting. This partly reflects that our method reduces the undesirable impact effectively. Overall, the residuals from OLS^S are smaller than the other methods. Another interesting point is that the residual from observation 2 is very close to zero with our method, even if it is not shifted. Some researchers (Andrews and Pregibon, 1978; Dempster and Gasko-Green, 1981; Rousseeuw and Leroy, 2005) reported that the observation 2 is a moderate outlier, although it is not clearly judged by Figure 3. Our conjecture is that when the injurious effect of observations 1, 3, 4, and 21 are significantly reduced by our method, the observation 2 is no more an outlier. Especially, the fit of the observation 21 is significantly improved by the suggested method as shown in Figure 3, whereas the other methods still suffer from the large residual from the observation 21.

5.2 *Analysis of mid-Atlantic wage data*

For analysis of real data, we numerically compare the eight models (OLS, OLS^S, H, H^S, $lasso$, $lasso^S$, $Hlasso$, and $Hlasso^S$) employed in simulation studies in Section

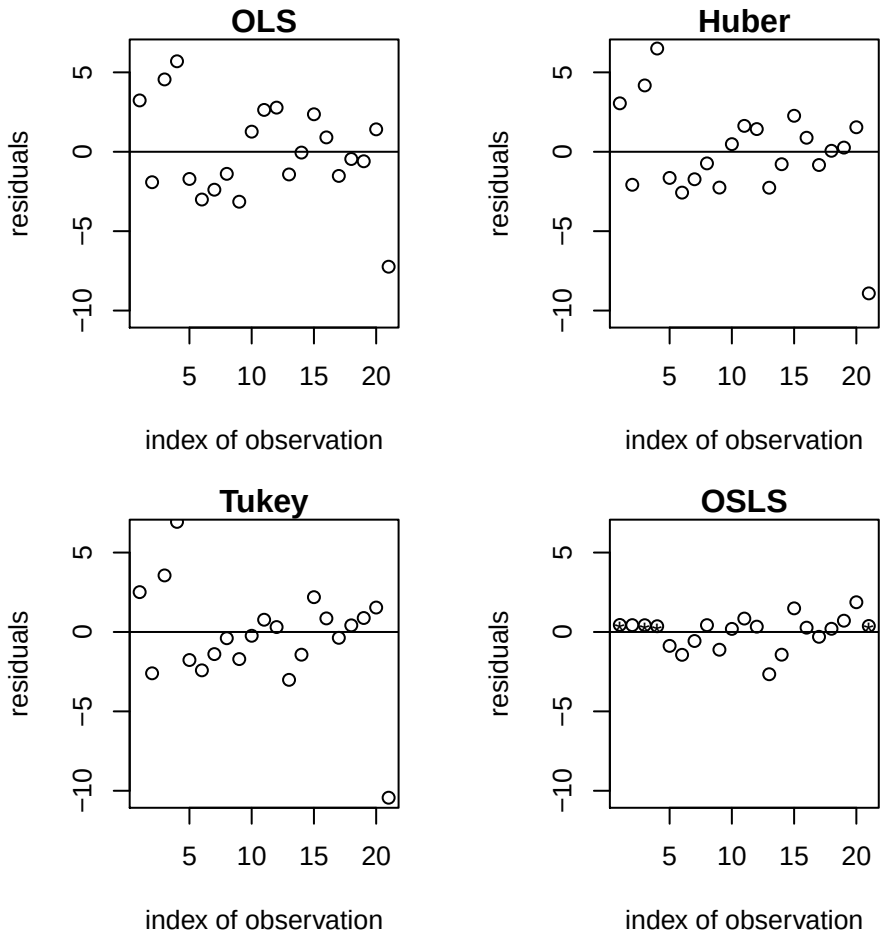

Figure 3: Residual plots from the four methods of OLS, H, $Tukey$, and OLS^S with all observations. In the OLS^S, the four observations with $*$ are detected as outliers (observations 1,3,4, and 21) and shifted by our method.

4. The mid-Atlantic wage data is manually assembled by Steve Miller of Open BI (`www.openbi.com`) from the March 2011 supplement to Current population and is available in R package `ISLR`. The data set is from 3,000 males in the mid-Atlantic region of the United States. There are eight explanatory variables of *year, age, married, race, edu, class, health, and ins*, and one response variable *wage*. Here are the details of each variable.

1. *wage*: Workers raw wage in 1,000 USD
2. *year*: 2003 to 2009
3. *age*: 18 or above
4. *married*: A factor with levels 1. Married and 2. The others (composed of never married, widowed, divorced, separated)
5. *race*: A factor with levels 1. White 2. African American and 3. The others
6. *edu*: A factor with levels 1. Less than high school graduate 2. High school graduate 3. Some college 4. College graduate and 5. Advanced degree
7. *class*: A factor with levels 1. Industrial and 2. Information indicating type of job
8. *health*: A factor indicating health status of 1. Good or less than good and 2. Very good
9. *ins*: A factor for status of having health insurance. 1. Yes and 2. No

We perform pre-analysis and modified the original variables slightly and do log transformation of the response variable due to heteroscedasticity. For the variable *married*, we combine the four categories of never married, widowed, divorced and separated into one category since there seems little difference. Similarly, the others category in *race* includes Asian which was a separate category in the original data in `ISLR` package. We treat *year* and *age* as continuous variables and the others as categorical variables. Since age^2 is a significant variable (details can be found at Chapter 1 of Games et al. (2013)), we add it into the model. To incorporate the categorical variables in the regression model, we make indicator variables. For example, we make two

indicator variables for *race* where the first level is considered as a baseline. Thus, we have $race_{2i}$ and $race_{3i}$, where $race_{2i} = 1$ if ith worker is an African American and 0 otherwise. In each categorical variable, level of 1 is considered as a baseline. Finally, we take log transformation on the response variable. The model we consider is

$$\begin{aligned}\log(wage_i) &= \beta_0 + \beta_1 year_i + \beta_2 age_i + \beta_3 age_i^2 + \beta_4 married_i + \beta_5 race_{2i} + \beta_6 race_{3i} \\ &\quad + \beta_7 edu_{2i} + \beta_8 edu_{3i} + \beta_9 edu_{4i} + \beta_{10} edu_{5i} + \beta_{11} class_i + \beta_{12} health_i + \beta_{13} ins_i + \epsilon_i,\end{aligned}$$

where ϵ_is are independent error with mean 0 and finite variance. To compare the eight candidate models, we perform cross-validation studies. In detail, first, we randomly split the data into two parts where the first part with 2,000 samples is regarded as a training data set, and the other 1,000 samples is test data. Second, we fit the candidate models and predict the value of response variable in the test data. Then, the absolute deviation between predicted value and observed value is measured. We incorporate a robust measure of absolute deviation since the squared distance is very sensitive to outliers. Finally, cross-validated score (CVS) is calculated by $\sum_{i=1}^{1000} |y_i - y_i^{pred}|/1000$, where y_i is the observed value of the response variable ($\log(wage_i)$) in the test data and y_i^{pred} is the predicted value of y_i. To reduce variation raised from random split, we repeat the above procedure for 500 times.

The mean of 500 CVS values is in Table 6. Standard errors of the means are all very close to 0.02, thus omitted in the table. The results of the predictive performance of the eight models illustrate that *lasso* and OLS^S perform better than the others. The value of c in Table 6 is the threshold in Huber's loss function in (3.1). As the improvement by our methods is quite minor, the values of the estimates by the above methods are similar in general. When income is treated as a response variable in economic data, extreme outliers are often detected even if the response variable is

Table 6: Mean of CVS based on 500 random split of mid-Atlantic wage data by models of OLS, OLS^S, H, H^S, $lasso$, $lasso^S$, $Hlasso$, and $Hlasso^S$. All values are multiplied by 100.

OLS	OLS^S	$H(c=2)$	$H(c=2.5)$
20.13	19.95	20.08	20.08
$H^S(c=2)$	$H^S(c=2.5)$	$lasso$	$lasso^S$
20.07	20.06	19.95	20.07
$Hlasso(c=2)$	$Hlasso(c=2.5)$	$Hlasso^S(c=2.5)$	$Hlasso^S(c=2.5)$
20.07	20.06	20.08	20.10

log transformed. An robust model may improve the prediction as our analysis shows here.

6 Discussion

In this work, we propose a method of automatically adjusting severely outlying observations by using case-specific parameters and shifting outlier penalty to reduce the undesirable impact. We provide a default value for the threshold of the outliers to capture huge outliers and to attain consistency. The outlier shifting trace is demonstrated with the example displayed in Figure 1. In terms of model estimation, we demonstrate the equivalence between the maximum likelihood based and convex minimization approaches in the case of least squares estimation. The convex minimization approach enables applying the proposed method to a wide range of modeling procedures, such as Huber's regression, lasso, and Huberized lasso. Our empirical in-

vestigation shows that the advantage of the proposed method in accurately capturing the true regression surface increases with the extent of data contamination. Despite of our focus on linear models in this work, the new method can be tailored to non-linear models involving convex minimization problems. We will extensively investigate more complicated cases in the future work.

Appendix

Proof of Theorem 1

First consider m=1. For all i, we have,

$$P\left(|\hat{\gamma}_i^{(1)}| > 0\right) = P\left(|y_i - x_i'\hat{\beta}^{(0)}| > \lambda_\gamma\right). \tag{6.1}$$

Since $|y_i - x_i'\hat{\beta}^{(0)}| = O_p(1)$ and $\lambda_\gamma = O(1)$, $P(|\hat{\gamma}_i^{(1)}| > 0)$ does not approach to zero as n increases. Thus, $\hat{\gamma}_i^{(1)} = O_p(1)$. On the other hand, $\hat{\beta}^{(1)} = (X'X)^{-1}X'(Y^{(0)} - \hat{\gamma}^{(1)}) = \hat{\beta}^{OLS} - (X'X)^{-1}X'\hat{\gamma}^{(1)}$. Now, a slight modification shows that

$$(X'X)^{-1}X'\hat{\gamma}^{(1)} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} x_i\hat{\gamma}_i^{(1)}.$$

Then, we have $(\frac{1}{n}\sum_{i=1}^{n} x_i x_i')^{-1} \xrightarrow{p} Q^{-1}$ by condition 1. Since $\hat{\gamma}_i^{(1)} = 0$ or r_i from (2.3), $\frac{1}{n}\sum_{i=1}^{n}\hat{\gamma}_i^{(1)} \xrightarrow{p} E(\xi_i)$. When this result is combined with condition 4, we have $\frac{1}{n}\sum_{i=1}^{n} x_i\hat{\gamma}_i^{(1)} = o_p(1)$. Thus, $\hat{\beta}^{(1)} = \hat{\beta}^{OLS} + o_p(1)$. Now, we consider $\hat{\gamma}_i^{(2)}$. First, we replace $\hat{\gamma}_i^{(1)}$ with $\hat{\gamma}_i^{(2)}$, and $\hat{\beta}^{(0)}$ with $\hat{\beta}^{(1)}$ in (6.1), which readily shows that $\hat{\gamma}_i^{(2)} = O_p(1)$. Next, we consider $\hat{\beta}^{(2)}$.

$$\begin{aligned}
\hat{\beta}^{(2)} &= (X'X)^{-1}X'(Y^{(1)} - \hat{\gamma}^{(2)}) \\
&= (X'X)^{-1}X'(Y^{(0)} - \hat{\gamma}^{(1)} - \hat{\gamma}^{(2)}) \\
&= \hat{\beta}^{OLS} - (X'X)^{-1}X'(\hat{\gamma}^{(1)} + \hat{\gamma}^{(2)}).
\end{aligned}$$

Again, $(X'X)^{-1}X'(\hat{\gamma}^{(1)} + \hat{\gamma}^{(2)})$ is $o_p(1)$ by the similar argument. Since $\sqrt{n}(\hat{\beta}^{OLS} - \beta) \xrightarrow{d} N(0, \sigma^2 Q^{-1})$ from the central limit theorem, we have the same result for $\hat{\beta}^{(2)} + (X'X)^{-1}X'(\hat{\gamma}^{(1)} + \hat{\gamma}^{(2)})$, and this holds for general m with $\hat{\beta}^{(m)}$, which completes the proof. □

Acknowledgements

Jung's work was supported by University of Waikato Research Trust (#103406), Lee's work was supported by Hankuk University of Foreign Studies Research Fund of 2015, and Hu's work was partially supported by the National Institute of Health Grants R01GM080503 and R01CA158113.

References

Andrews, D. F. and Pregibon, D. (1978). Finding the outliers that matter. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **40**(1), 85 – 93.

Beaton, A. E. and Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, **16**(2), 147 – 185.

Bloomfield, P. and Steiger, W. (1983). *Least Absolute Deviations: Theory, Applications and Algorithms*. Birkhauser Boston, Boston, 1 edition. ISBN 0817631577, 9780817631574.

Breckling, J. and Chambers, R. (1988). M-quantiles. *Biometrika*, **75**, 761 – 771.

Brownlee, K. (1965). *Statistical theory and methodolooy in science and engineering.* New York : Wiley, 2nd edition.

Carroll, R. J. and Ruppert, D. (1985). Transformations in regression: A robust analysis. *Technometrics*, **27**(1), 1 – 12.

Chambers, R. L. and Heathcote, C. (1981). On the estimation of slope and the identification of outliers in linear regression. *Biometrika*, **68**(1), 21 – 33.

Daniel, C. and Wood, F. S. (1971). *Fitting equations to data; computer analysis of multifactor data for scientists and engineers.* Wiley, New York.

Daniel, C. and Wood, F. S. (1980). *Fitting Equations to Data: Computer Analysis of Multifactor Data.* John Wiley & Sons, Inc., New York, NY, USA, 2nd edition. ISBN 0471053708.

Dempster, A. P. and Gasko-Green, M. (1981). New tools for residual analysis. *Ann. Stat.*, **9**(5), 945 – 959.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **39** (1), 1–38.

Denby, L. and Mallows, C. (1977). Two diagnostic displays for robust regression analysis. *Technometrics*, **19**(1), 1 – 13.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.*, **96**(456), 1348 –1360.

Games, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R.* Springer-Verlag, New York, www.StatLearning.com.

Hadi, A. S. and Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models. *J. Am. Stat. Assoc.*, **88**(424), 1264 – 1272.

Hoeting, J., Raftery, A. E., and Madigan, D. (1996). A method for simultaneous variable selection and outlier identification in linear regression. *Comput. Statist. Data Anal.*, **22**, 251 – 270.

Huber, P. and Ronchetti, E. M. (2009). *Robust Statistics*. Probability and Statistics. John Wiley & Sons, Inc., Hoboken, New Jersey, 2 edition. ISBN 978-0-470-12990-6.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures, and monte carlo. *Ann. Statist.*, **1**, 799–821.

Koenker, R. (2005). *Quantile Regression*. Cambridge U. Press. ISBN 0521608279.

Kosinski, A. S. (1999). A procedure for the detection of multivariate outliers. *Comput. Statist. Data Anal.*, **29**, 145 – 161.

Lange, K. L., Little, R. J. A., and Taylor, J. M. G. (1989). Robust statistical modeling using the t distribution. *J. Am. Stat. Assoc.*, **84**(408), 881 – 896.

Lee, Y., MacEachern, S. N., and Jung, Y. (2007). Regularization of case-specific parameters for robustness and efficiency. Technical Report 799, The Ohio State University.

Lee, Y., MacEachern, S. N., and Jung, Y. (2012). Regularization of case-specific parameters for robustness and efficiency. *Statist. Sci.*, **27**(3), 350 – 372.

Pena, D. and Yohai, V. (1999). A fast procedure for outlier diagnostics in large regression problems. *J. Am. Stat. Assoc.*, **94**(446), 434 – 445.

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

Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths. *Ann. Stat.*, **35**(3), 1012 – 1030.

Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the median absolute deviation. *J. Am. Stat. Assoc.*, **88**(424), 1273 –1283.

Rousseeuw, P. J. and Leroy, A. M. (2005). *Robust Regression and Outlier Detection*. John Wiley & Sons, Inc., Hoboken, NJ, USA. ISBN 9780471852339.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **58**(1), 267 – 288.

Wu, T. T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat.*, **2**(1), 224–244.