ASYMPTOTIC ANALYSIS OF THE HUBERIZED LASSO ESTIMATOR

Xiaohui Chen¹, Z.Jane Wang¹, Martin J. McKeown²

1. Department of Electrical and Computer Engineering, 2. Department of Medicine (Neurology), University of British Columbia, Canada.

ABSTRACT

The Huberized LASSO model, a robust version of the popular LASSO, yields robust model selection in sparse linear regression. Though its superior performance was empirically demonstrated for large variance noise, currently no theoretical asymptotic analysis has been derived for the Huberized LASSO estimator. Here we prove that the Huberized LASSO estimator is consistent and asymptotically normal distributed under a proper shrinkage rate. Our derivation shows that, unlike the LASSO estimator, its asymptotic variance is stabilized in the presence of noise with large variance. We also propose the adaptive Huberized LASSO estimator by allowing unequal penalty weights for the regression coefficients, and prove its model selection consistency. Simulations confirm our theoretical results.

Index Terms— Sparse linear regression, Huberized LASSO, robustness, asymptotic normality, model selection consistency.

1. INTRODUCTION

We consider the problem of estimating the coefficient vector in a linear regression model, defined as

\[ y = X\beta + e \]  

(1)

where the random measurement error vector \( e = (e_1, \ldots, e_n)^T \) is assumed to be independent, identically distributed (i.i.d.) with zero mean and possesses a finite second moment \( \sigma^2 \) for each component. \( X \) is the \( n \times p \) design matrix which can either be non-stochastic or random. As usual, rows of \( X \) represent the \( p \)-dimensional observations and columns of \( X \) represent the predictors/signals. \( y \) is the response vector and \( \beta \) is the coefficient vector to be estimated. In this paper, we are interested in the problem of sparse linear regression. By sparse we mean that \( p \) is large but the number of true predictors with non-zero coefficients is relatively small.

Extracting signals from a sparse regression model is by no means an easy task. The Least Absolute Shrinkage and Selection Operator (LASSO) [1] is probably the most popular and powerful technique for simultaneously performing model selection and parameter estimation. Since the LASSO employs the squared \( \ell_2 \) loss as a goodness-of-fit measure criterion, its estimate is sensitive to large variance noise and outliers. Empirical examples show that the LASSO estimator can behave poorly if data are contaminated [2]. Employing robust losses is critical to overcome this problem. Since the Huber loss is robust to heavy-tailed error distributions and outliers, using the Huber loss coupled with the \( \ell_1 \) penalty yields a robust model selection approach, namely the Huberized LASSO. Specifically, the Huberized LASSO estimator is defined as

\[ \hat{\beta}_n^H = \arg \min_{u \in \mathbb{R}} \left( \frac{1}{n} \sum_{i=1}^{p} L(u; y_i, x_i) + \frac{\lambda_n}{n} \|u\|_1 \right) \]  

(2)

where \( L(\cdot) \) is the Huber loss function

\[ L(u; y_i, x_i) = \begin{cases} 
(y_i - u^*x_i)^2 & \text{if } |y_i - u^*x_i| \leq \delta, \\
2\delta|y_i - u^*x_i| - \delta^2 & \text{if } |y_i - u^*x_i| > \delta.
\end{cases} \]

There are two tuning parameters in (2). \( \delta \geq 0 \) is a tuning parameter for robustness and \( \lambda_n \geq 0 \) is a shrinkage tuning parameter. The Huberized LASSO with a larger \( \delta \) behaves more like the ordinary LASSO, while the Huber loss with a smaller \( \delta \) is more robust to outliers. A larger \( \lambda_n \) yields a sparser linear model (i.e. with less non-zero coefficients) whereas a smaller \( \lambda_n \) corresponds to a less-sparse one. In extreme cases, \( \lambda_n = 0 \) gives the unregularized regression model while \( \lambda_n = \infty \) produces the null model containing no predictor. [3] empirically studied its performance and illustrated the robustness of the Huberized LASSO estimator when contamination is added. However, to our best knowledge, currently no asymptotic analysis of the Huberized LASSO has been presented in the literature. Therefore, in this paper we provide an asymptotic analysis of the Huberized LASSO estimator and show that it has more attractive statistic properties compared with the LASSO estimator when the noise variance is large or when the response is corrupted by outliers.

The paper is organized as follows. Section 2 presents the asymptotic normality and estimation consistency. An adaptive Huberized LASSO is defined and its model selection consistency is shown. The theoretical results are supported by the simulations in Section 3. It is worth emphasizing that, due to

This work was supported by the Pacific Alzheimer Research Foundation and the Canadian Natural Sciences and Engineering Research Council (NSERC).
limited space, proofs of the theorems are not given here. For details, we refer the interested readers to the submitted journal version [4].

We list some major notations as follows. For a vector \( \mathbf{u} \), we interchangeably use \( u_j \) and \( u[j] \) to denote the \( j \)th-element of \( \mathbf{u} \). Define \( \text{supp}(\mathbf{u}) = \{ j \in \{1, \ldots, p\}; |u_j| \neq 0 \} \) and define \( |\mathbf{u}| \) to be the cardinality of \( \mathbf{u} \). We regard \( \mathbf{u} \) as a column vector, and use \( \mathbf{u}^* \) to denote its conjugate transpose. \( X_n \overset{p}{\rightarrow} X \) refers to the convergence in probability and \( X_n \Rightarrow X \) in distribution. For a fixed \( \delta \geq 0 \), define \( K_{0\delta} = P(|e_i| \leq \delta) \) and \( K_{r\delta} = E(|e_i|^r; |e_i| \leq \delta) \) for \( r > 0 \). Similarly, define \( M_{0\delta} = 1 - K_{0\delta} \) and \( M_{r\delta} = E(|e_i|^r; |e_i| > \delta) \).

2. LIMITING BEHAVIOR OF THE HUBERIZED LASSO ESTIMATOR

2.1. Asymptotic normality and estimation consistency

In order to make sure that there is no strong dependency among the predictors (columns of \( X \)), namely model identifiability, we need a regularity assumption on the design matrix. Here, we assume the following conditions:

1. (Design matrix assumption) The Gram matrix \( C_n = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^* \) converges to a positive definite matrix \( C \) as \( n \rightarrow \infty \).

2. (Error assumption) The error \( e_i \) has a common symmetric distribution with respect to (w.r.t.) the origin. So \( E e_i = 0 \) and the median of \( e_i \) is 0.

Remark 1. The design matrix assumption ensures that for all sufficiently large \( n \)’s the observed covariance matrix \( C_n \) is non-degenerate, i.e. the columns of \( X \) are linearly independent, so that the linear model (1) is uniquely represented. Thus, this hypothesis is stronger than the sparse representation condition which only assumes a unique representation of restricted columns of \( X \) [5]. Nevertheless, the classical design matrix condition assumed here (and also by many other authors such as [6] and [7] studying the asymptotics for the LASSO estimator) turns out to greatly simplify the proofs of asymptotic behavior of the Huberized LASSO estimator.

With these two assumptions, we first establish the asymptotic normality of the Huberized LASSO.

Theorem 2.1. Under the assumptions 1 and 2, if \( \lambda_n/\sqrt{n} \rightarrow \lambda_0 \geq 0 \) and \( e_i \) has a continuous probability density function (p.d.f.) w.r.t. the Lebesgue measure on \( \mathbb{R} \), then \( \sqrt{n}(\hat{\beta}_n^H - \beta) \Rightarrow \text{arg min}\{V\} \) where

\[
V(\mathbf{u}) = K_{0\delta} \mathbf{u}^* C \mathbf{u} + 2 \mathbf{u}^* \mathbf{W}
\]

\[+
\lambda_0 \sum_{j=1}^{p} (|u_j\text{sgn}(\beta_j) 1(\beta_j \neq 0) + |u_j|1(\beta_j = 0)]
\]

and \( \mathbf{W} \sim N(0, (\beta_0^2 M_{0\delta} + K_{2\delta}) C) \).

An immediate consequence of Theorem 2.1 is the \( \sqrt{n} \) estimation consistency of the Huberized LASSO estimator \( \hat{\beta}_n^H \).

Corollary 2.2. If \( \lambda_n = o(\sqrt{n}) \), then \( \hat{\beta}_n^H \) is \( \sqrt{n} \)-consistent.

Remark 2. By Corollary 2.2, the asymptotic variance of the Huberized LASSO estimator is \( \frac{\lambda_0^2 M_{0\delta} + K_{2\delta} C^{-1}}{K_{0\delta}} \) and thus it is possible to optimize over \( \delta \in [0, \infty) \). The minimum asymptotic variance indexed by the family of \( \delta \in [0, \infty) \) depends on the specific error distribution. For a small \( \delta \), the asymptotic variance of the Huberized LASSO estimator is in general less than that of the LASSO estimator, see Table 1 for an example.

2.2. Model selection consistency

In the previous paragraph, we have shown the estimation consistency of the Huberized LASSO estimator. However, in many scenarios, it is also desirable to have the model selection consistency, defined as

\[
P \left( \text{supp}(\hat{\beta}_n^H) = \text{supp}(\beta) \right) \rightarrow 1
\]

as \( n \rightarrow \infty \).

In terms of choosing a sequence of shrinkage tuning parameters \( \{\lambda_n\}_{n \in \mathbb{N}} \), it is often determined by criteria related to the minimization of the prediction error, such as the cross-validation (CV). However, [8] and [9] showed that the ordinary LASSO has a conflict between the consistency for model selection and optimal prediction. As a solution to achieve both estimation and model selection consistency, the adaptive LASSO [10] was proposed and its model selection consistency and asymptotic normality under certain rate of shrinkage were proved. To extend the idea of the adaptive LASSO [10] to the Huberized LASSO, we define the adaptive robust LASSO as

\[
\hat{\beta}_n^H = \arg \min_{\mathbf{u} \in \mathbb{R}^p} \left( \frac{1}{n} \sum_{i=1}^{n} L(\mathbf{u}; y_i, \mathbf{x}_i) + \frac{\lambda_n}{n} \sum_{j=1}^{p} \tilde{w}_j |u_j| \right)
\]

(4)

where \( \mathbf{w} = (\tilde{w}_1, \cdots, \tilde{w}_p)^* \) is a vector of adaptive weights, which allow unequal penalties for the coefficients. For example, we can take \( \mathbf{w} = 1/|\beta_{\text{LS}}|^\gamma \) for some \( \gamma > 0 \) and \( \hat{\beta}_{\text{LS}} \) is the least squares estimator for the full model. Let \( A = \{j : \beta_j \neq 0\} \). Theorem 2.3 shows that the adaptive Huberized LASSO estimator is model selection consistent.

Theorem 2.3. Suppose assumptions 1 and 2 are satisfied. Let \( \lambda_n = o(\sqrt{n}) \) and \( \lambda_n n^{(\gamma-1)/2} \rightarrow \infty \) for some \( \gamma > 0 \). Then the adaptive Huberized LASSO defined in (4) with \( \mathbf{w} = 1/|\hat{\beta}_{\text{LS}}|^\gamma \) has the following two properties:
Table 1. Theoretical asymptotic variances of \( \sqrt{n}(\hat{\beta}_n - \beta) \) for the mixture of Gaussian and student-\( t \) error distributions, respectively. HLASSO means the Huberized LASSO. The adaLASSO and adaHLASSO represent the corresponding adaptive versions. 1,2,5 are the indices of the true non-zero coefficients and 3,4,6,7,8 are the indices of the zero coefficients.

<table>
<thead>
<tr>
<th>Model</th>
<th>Gaussian Mixture</th>
<th>Student-( t )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \mu = 5, \sigma = 1 )</td>
<td>( \nu = 3, \sigma = 3 )</td>
</tr>
<tr>
<td>LASSO</td>
<td>13.5</td>
<td>13.5</td>
</tr>
<tr>
<td>HLASSO</td>
<td>5.65</td>
<td>5.65</td>
</tr>
<tr>
<td>adaLASSO</td>
<td>13.5</td>
<td>0</td>
</tr>
<tr>
<td>adaHLASSO</td>
<td>5.65</td>
<td>0</td>
</tr>
</tbody>
</table>

1. Asymptotic normality for the true non-zero coefficients, i.e.
\[
\sqrt{n}(\hat{\beta}_n^H [A] - \beta[A]) \Rightarrow N \left( 0, \frac{\delta^2 M_{08} + K_{2\delta} C_{11}^{-1}}{K_{2\delta}} \right).
\]

2. Model selection consistency.

3. NUMERIC EXAMPLES

In this section, two simulations are used to validate the theoretical results we have derived. The underlying model is assumed as follows:
\[
y_i = x_i^T \beta + e_i
\]
where \( \beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^T \). \( X \) is realized from a Gaussian random matrix with zero mean and unit variance. So we have \( C = I_{8 \times 8} \). The errors are generated based on two different mechanisms, with more details given shortly. The intercept term is not considered since it can always be estimated by the mean of \( y \). Therefore, the response \( y \) is centered before applying any shrinkage model. The shrinkage tuning parameter \( \lambda_n \) is chosen to be \( n^{1/3} \) for the LASSO and Huberized LASSO models such that they have both parameter estimation and model selection consistency for adaptive weight \( \tilde{w} = 1/|\hat{\beta}_n^{LS}|^\gamma \) with \( \gamma = 1 \). We set \( \delta = 1 \) for the Huber loss. Simulations are averaged and reported over 100 simulated data sets, each of which contains \( n = 1,000 \) data points. The following two error distributions are considered:

1. Symmetric Gaussian mixture with three components.
The errors are simulated from a Gaussian mixture distribution with symmetric two-side outliers, i.e. the error is assumed to have the following p.d.f.
\[
f(e_i) = \frac{1}{4} N(e_i; -\mu, \sigma^2) + \frac{1}{2} N(e_i; 0, \sigma^2) + \frac{1}{4} N(e_i; \mu, \sigma^2)
\]
where \( N(e_i; \mu, \sigma^2) \) denotes the p.d.f. of a normal random variable with mean \( \mu \) and variance \( \sigma^2 \). It is clear that \( f \) satisfies the error assumption. The results are reported for \( \mu = 5 \) and \( \sigma = 1 \). Fig.1. shows the histograms of \( \sqrt{n}(\hat{\beta}_n^H - \beta) \) for the non-adaptive (the first two rows) and adaptive (the last two rows) models. The theoretical variances of the limiting distribution of \( \sqrt{n}(\hat{\beta}_n^H - \beta) \) are shown in Table 1. From Table 1 and Fig.1, we can report several observations here:

(a) The variances of \( \sqrt{n}(\hat{\beta}_n^H - \beta) \) calculated based on the simulations closely match the theoretical asymptotic variances (see the first two rows of Table 1).

(b) The asymptotic variances of the Huberized LASSO estimates are smaller than that of the LASSO estimates, as expected (see Table 1).

(c) Although the adaptive LASSO has been proved to be model selection consistent, the simulation study shows that the adaptive LASSO performs poorly when the noise variance \( \sigma^2 \) is large, even for a relatively large sample size \( n = 1,000 \). In contrast, the adaptive Huberized LASSO provides significant performance improvements over the ordinary LASSO (see the last two rows of Fig.1. for the zero coefficients).

2. Student-\( t \) errors with heavy tails. Here the setup is the same as in the Gaussian mixture case, except that the errors are generated from a student-\( t \) distribution with the degree of freedom \( \nu = 3 \) and \( \sigma = 3 \). The theoretical values of the variances of the asymptotic distribution of \( \sqrt{n}(\hat{\beta}_n^H - \beta) \) are given in Table 1. Based on the histogram results, same observations can be noted as in the Gaussian mixture case. Therefore, we do not report the figures here due to the space concern.

4. CONCLUSIONS AND DISCUSSIONS

Although the Huberized LASSO has been used in literature as a robust variable selection tool in linear regression, we derive in this paper the limiting distribution of the Huberized LASSO estimator. Both estimation consistency and model selection consistency are established. The simulations confirm the theoretical analysis and demonstrate the superiority of the Huberized LASSO over the LASSO in the presence of noise with large variance or outliers.

The asymptotic results derived in this paper are based on a fixed design matrix \( X \). In practice, measurement errors exist not only in the response, but also in the predictors. Thus, the relaxation of the current fixed-design framework to the random design is desirable for practical applications, such as graphical and autoregression models. To this end, additional mild measurability assumptions on the random errors are required and the results presented in this paper still hold for the random design setup [4].
Fig. 1. Histograms of $\sqrt{n}(\hat{\beta}_n - \beta)$ for the Gaussian mixture distributed errors with $\mu = 5$. Green curve represents the fitted normal distribution to the estimated values of $\sqrt{n}(\hat{\beta}_n - \beta)$ based on over 100 simulations. Red curve represents the theoretical asymptotic normal distribution of $\text{arg min}(V)$. From top to the bottom, the four rows represent the LASSO, Huberized LASSO, adaptive LASSO, and adaptive Huberized LASSO models, respectively. The columns represent the eight predictors in the order of $(3, 1.5, 0, 0, 2, 0, 0, 0)$.

5. REFERENCES


