



Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Another look at Huber's estimator: A new minimax estimator in regression with stochastically bounded noise[☆]

Xuelei Sherry Ni^{a,*}, Xiaoming Huo^b^aDepartment of Mathematics and Statistics, Kennesaw State University, Kennesaw, GA 30144, USA^bSchool of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

ARTICLE INFO

Article history:

Received 16 October 2005

Received in revised form

18 February 2008

Accepted 19 March 2008

Available online 22 May 2008

Keywords:

Huber's estimator

Regression

Asymptotic minimax estimator

ABSTRACT

Huber's estimator has had a long lasting impact, particularly on robust statistics. It is well known that under certain conditions, Huber's estimator is asymptotically minimax. A moderate generalization in rederiving Huber's estimator shows that Huber's estimator is not the only choice. We develop an alternative asymptotic minimax estimator and name it *regression with stochastically bounded noise* (RSBN). Simulations demonstrate that RSBN is slightly better in performance, although it is unclear how to justify such an improvement theoretically. We propose two numerical solutions: an iterative numerical solution, which is extremely easy to implement and is based on the *proximal point method*; and a solution by applying state-of-the-art nonlinear optimization software packages, e.g., SNOPT. Contribution: the generalization of the variational approach is interesting and should be useful in deriving other asymptotic minimax estimators in other problems.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

Huber's estimator has had a long lasting impact on robust statistics. Classical ways of introducing the minimaxity of Huber's estimator seem to imply that it is the only choice (Lehmann, 1991, Section 5.6). In this paper, we rederive minimax estimators in a regression setting, by following Huber's variational approach. We show that other minimax estimators are available by considering the general solutions to the differential equation that is considered in the variational approach. The result is a new estimator which we call regression with stochastically bounded noise (RSBN). To the best of our knowledge, this is the first appearance of such an estimator. The consideration of the general solutions in the variational approach is inspiring to us and has potential applications in other problems where minimax estimators are to be derived.

It is well known that Huber's estimator can be solved via iteratively reweighted least squares. We present two computational approaches for RSBN. First, using the proximal point method in optimization, we develop an iterative approach that is extremely simple to implement—it only takes a few lines in MATLAB. However, its numerical performance is not stable: it can converge extremely quickly in some situations, and extremely slowly in some pathological cases. Second, we present an alternative by using state-of-the-art optimization software packages, such as SNOPT.

RSBN performs well in simulations, although it is not clear whether/why the different choices of minimax estimators make a difference in applications.

The rest of this paper is organized as follows. In Section 2, the formulation is presented. In Section 3, we establish RSBN, as well as its asymptotic minimaxity. In Section 4, we discuss numerical issues. Simulation results are reported in Section 5. Some discussions are presented in Section 6.

[☆] This work was supported in part by NSF grants 0604736 and 0700152.

* Corresponding author.

E-mail address: xni2@kennesaw.edu (X.S. Ni).

2. Formulation

We adopt the standard regression model

$$\mathbf{y} = A\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\mathbf{y} = (y_1, y_2, \dots, y_n)^T \in \mathbb{R}^n$ is the response vector, $\boldsymbol{\beta} \in \mathbb{R}^m$ is the vector of coefficients, $A = [a_1, a_2, \dots, a_n]^T \in \mathbb{R}^{n \times m}$ is the model matrix, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)^T$ represents the random error vector. Without loss of generality, we assume that A is of full column rank (equivalently, $(A^T A)^{-1}$ exists). Furthermore, we assume that the random errors $\varepsilon_i, i = 1, 2, \dots, n$, are i.i.d. with a common density function f .

One can estimate the set of coefficients by choosing the one that solves the following optimization problem:

$$\text{minimize } \sum_{i=1}^n \rho(r_i) \quad \text{subject to } r_i = y_i - a_i^T \boldsymbol{\beta}, \quad i = 1, 2, \dots, n. \tag{2.2}$$

Here, we normally require ρ to be a convex function, because a convex optimization problem is more amenable to a solution than other optimization problems (e.g., combinatorial optimization problems). If we define a vector $\mathbf{r} = (r_1, r_2, \dots, r_n)^T \in \mathbb{R}^n$, the restriction of the above optimization problem can be rewritten as $\mathbf{r} = \mathbf{y} - A\boldsymbol{\beta}$. So, we can express the optimization problem (2.2) in matrix terms as

$$\text{minimize } \rho(\mathbf{r}) = \sum_{i=1}^n \rho(r_i) \quad \text{subject to } \mathbf{r} = \mathbf{y} - A\boldsymbol{\beta}.$$

A key feature of the above formulation is that the objective function is an additive function. The criterion depicted in (2.2) covers many known approaches. For example, when $\rho(x) = x^2$, we have the ordinary least squares estimator. When $\rho(x) = x^2$ for $|x| \leq 1$, and $\rho(x) = 2|x| - 1$, for $|x| > 1$, we have Huber's estimator.

3. Regression achieving asymptotic minimaxity and RSBN

In this paper, we propose the following function for $\rho(x)$:

$$\rho(x) = \begin{cases} -\log \cos \lambda_1(x/\delta) & \text{if } |x/\delta| < 1, \\ \lambda_1 \tan \lambda_1 \cdot |x/\delta| - \lambda_1 \tan \lambda_1 - \log \cos \lambda_1 & \text{if } |x/\delta| \geq 1, \end{cases} \tag{3.1}$$

where $0 < \lambda_1 < \pi/2$. More results on λ_1 will be established when we derive the asymptotic minimaxity of the above estimator. Fig. 1(a) gives a graphical comparison between the above ρ and the objective functions that are used in the least squares estimator and Huber's estimator. The derivatives of these functions are shown in Fig. 1(b).

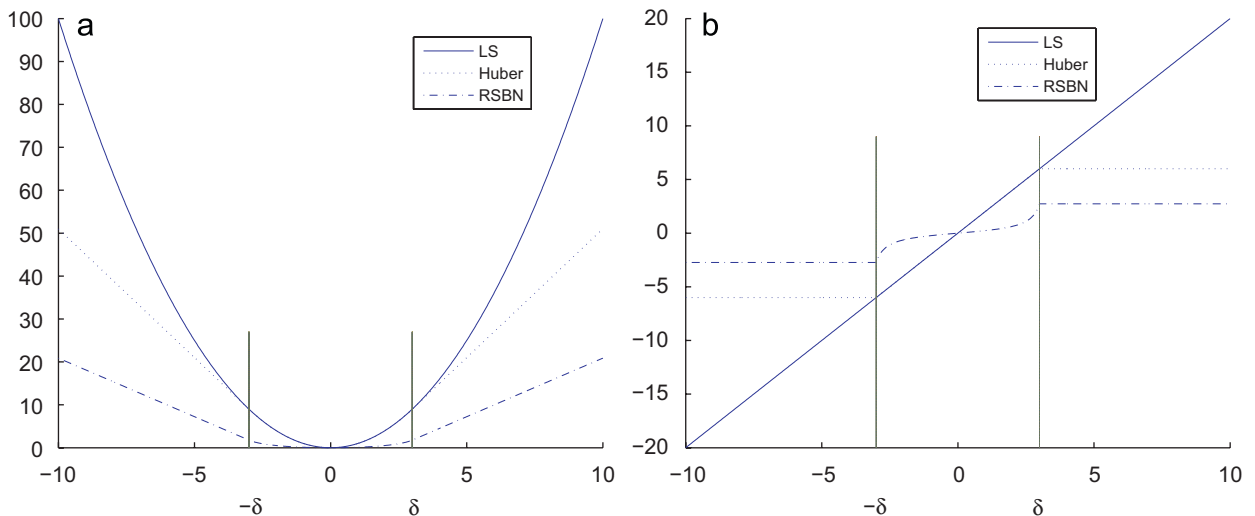


Fig. 1. The objective functions $\rho(x)$ and the first derivatives in ordinary least squares, Huber's estimator, and RSBN: (a) objective functions, (b) first derivatives.

When $\rho(x)$ has the form in (3.1), the resulting estimate is named the RSNB estimate. We will prove that the RSNB estimator achieves asymptotic minimaxity, which is our main reason for choosing (3.1) as ρ . This result is summarized in the following theorem and will be proved later in this section.

Theorem 3.1. *The RSNB estimate is asymptotically locally minimax, given that $\varepsilon_i, 1 \leq i \leq n$, satisfy $\text{Prob}(|\varepsilon_i| > \delta) \leq \alpha$, where $\delta > 0$ and $0 < \alpha < 1$ are predetermined.*

We call the errors that satisfy the above condition ‘stochastically bounded noise’.

Theorem 3.1 is proved through Sections 3.1–3.6.

3.1. Asymptotic normality

The solution to (2.2) is an M-estimator. The asymptotic normality of M-estimators is well known in statistics, cf. Hampel et al. (1986) and Huber (1981, Chapter 7.6). This property is summarized in the following.

Lemma 3.2. *Suppose $\rho(\cdot)$ has a monotone first derivative $\rho' = \psi$, and a second derivative $\rho'' = \psi'$. Also suppose $\int \psi'(x - c) dF(x)$ exists, where $F(x)$ is the cumulative distribution functions (c.d.f.) of f and c is a constant. Then, the estimate given by (2.2) has the asymptotic distribution*

$$N\left(\beta_0, \frac{1}{n} \frac{\int \psi^2 dF}{(\int \psi' dF)^2} (A^T A)^{-1}\right),$$

where β_0 is the vector of the true values of the coefficients.

We take the quantity $\int \psi^2 dF / (\int \psi' dF)^2$ as a natural measure of the performance of an M-estimator and call it the asymptotic variance. Obviously, the smaller the asymptotic variance is, the closer the M-estimator is to the true value of the parameter.

3.2. Minimum asymptotic variance estimate

Based on the above analysis, we hope to achieve the minimum asymptotic variance. The following lemma summarizes some related results, which has been shown in some statistics books, cf. Lehmann (1991, Chapter 5.6). The detailed proof of this lemma can also be found in Ni (2005) and Huo and Ni (2005).

Lemma 3.3. *The asymptotic variance of the estimator from (2.2) is bounded below by $1/I(f)$, where $I(f)$ is the Fisher information of the distribution f . The lower bound is achieved when $\rho \propto (-\log f)$, i.e., when the estimate is the maximum likelihood estimate.*

When $\rho = -\log f$, we call the solution to (2.2) the minimum asymptotic variance estimate.

3.3. Least informative distribution

We restate the variational approach, which is one way to derive Huber's estimator. Obviously, the smaller the Fisher information $I(f)$, the larger the lower bound of the asymptotic variance. So we are interested in the least informative distribution, which is the solution to the following optimization problem: (Note the variable, f , is a function.)

$$\text{minimize } I(f) \quad \text{subject to } \int v(x)f(x) dx \leq 0, \quad \int f(x) dx = 1. \tag{3.2}$$

Notice that in our framework, f is assumed to have a nonzero information $I(f)$. For example, a piecewise constant function f may lead to $I(f) = 0$, which leads to an infinite asymptotic variance. Such a case is excluded by demanding the smoothness of f : f' exists everywhere and does not equal to 0 for all x .

The first constraint in (3.2) is a general form of many types of restrictions on noise distributions. For example, if

$$v(x) = -\alpha \quad \text{when } |x| < \delta \quad \text{and} \quad v(x) = 1 - \alpha \quad \text{when } |x| \geq \delta, \tag{3.3}$$

we have $\int_{-\delta}^{\delta} f \geq 1 - \alpha$. This implies that the errors are stochastically bounded. This condition is meaningful when there are outliers. If $v(x) = x^2 - B$, we have $\int x^2 f(x) dx \leq B$, which is the second moment constraint. Similarly, we can have some other moment constraints. The second constraint in (3.2) is equivalent to requiring f to be a p.d.f.

To find the solution to (3.2), we consider the following function:

$$\mu(f) = I(f) + \eta_1 \left[\int v(x)f(x) dx + \gamma^2 \right] + \eta_2 \left[\int f(x) dx - 1 \right],$$

where η_1 and η_2 are the Lagrange multipliers, and $\gamma \in \mathbb{R}$ is a pseudo-variable: $\int v(x)f(x) dx + \gamma^2 = 0$. We consider a variational approach. Assume f_0 is a minimizer in (3.2). For any other p.d.f. f_1 , consider $f_t = (1 - t)f_0 + tf_1$, $0 \leq t \leq 1$. Because f_0 is a minimizer, we must have $(d/dt)\mu(f_t)|_{t=0} \geq 0$ for any f_1 , which is equivalent to

$$-4 \int \frac{(\sqrt{f_0})''}{\sqrt{f_0}} (f_1 - f_0) dx + \eta_1 \int v \cdot (f_1 - f_0) dx + \eta_2 \int (f_1 - f_0) dx \geq 0.$$

The above holds if and only if

$$4 \frac{(\sqrt{f_0})''}{\sqrt{f_0}} - \eta_1 \cdot v - \eta_2 = 0. \tag{3.4}$$

Note the above is a necessary condition for f_0 to be the solution to (3.2), and it also holds if f_0 is a local minimizer. I.e., if f_0 solves (3.2) within a neighborhood, then $(d/dt)\mu(f_t)|_{t=0} \geq 0$ is still true. We summarize this result in the following lemma.

Lemma 3.4. *If a function f_0 has second derivative and achieves a local minimum in (3.2), then it satisfies (3.4).*

3.4. Regression with stochastically bounded noise (RSBN)

We construct a function f_0 that satisfies (3.4). The constructed function f_0 leads to the objective function that is used in RSBN. Our objective is to find an appropriate ρ in (2.2), so that the solution to (2.2) is both easy to compute and optimal within a family of distributions for random errors. We list the conditions that are to be satisfied.

- [Probability density function] Function f is a probability density function: $f : \mathbb{R} \rightarrow \mathbb{R}^+$ ($f \geq 0$) and $\int f = 1$. In the previous discussion, we assumed that f has a finite Fisher information, $I(f) < \infty$. We also assume that f is symmetric about 0.
- [Stochastically bounded noise] We have $\int_{-\delta}^{\delta} f \geq 1 - \alpha$. This indicates that the probability of errors having absolute values no larger than δ is at least $1 - \alpha$. Usually α is small. As mentioned earlier, an expression equivalent to this condition is $\int v(x)f(x) dx \leq 0$, where v is defined in (3.3).
- [Convexity] Function $\rho(x) = -\log f(x)$ must be convex, and ρ' and ρ'' both exist. Under these conditions, (2.2) becomes a nonlinear convex optimization problem.
- [Minimaxity] When $\rho(x) = -\log f(x)$, according to Lemma 3.3, the minimum asymptotic variance is achieved. If f also minimizes the objective in (3.2), the minimum variance is achieved in the worst scenario. Such an estimator is called an asymptotic minimax estimator. From Lemma 3.4, the above mentioned minimizer f should satisfy Eq. (3.4).

The reader can verify that the following function is a solution to Eq. (3.4).

$$f_0(x) = \begin{cases} c \left[\cos \lambda_1 \frac{x}{\delta} \right]^2, & |x| < \delta, \\ c \cdot \exp\left(-2\lambda_2 \frac{|x|}{\delta}\right) \cdot \cos^2 \lambda_1 \cdot \exp(2\lambda_2), & |x| \geq \delta, \end{cases} \tag{3.5}$$

where $0 < \lambda_1 < \pi/2$, $\lambda_2 > 0$. The above is constructed by considering the general solutions to the differential equation (3.4). One of the simplest forms that satisfies all the aforementioned conditions is chosen. Special care is given to ensure that $\log(f_0)$ has second derivative. More discussion regarding our choice of f_0 , especially how it differs from Huber's estimator, will be provided in Section 3.5.

Recalling that $\rho = -\log f_0$, we have

$$\rho(x) = \begin{cases} -\log c - 2 \log \cos \lambda_1 \frac{x}{\delta}, & |x| < \delta, \\ -\log c + 2\lambda_2 \frac{|x|}{\delta} - 2\lambda_2 - 2 \log \cos \lambda_1, & |x| \geq \delta. \end{cases} \tag{3.6}$$

Note that $\rho(x)$ can be simplified without changing the optimization problem in (2.2): i.e., replacing $\rho(x)$ with $a\rho(x) + b$, $a > 0$ in (2.2) gives an equivalent optimization problem. Note that $\rho(x)$ is linear outside the interval $[-\delta, \delta]$.

3.5. Our choice of objective function versus Huber's estimator

Our choice of the objective function $\rho(\cdot)$ is rooted in (3.4). We now present the justification for choosing the functional solution as in (3.5). Because $-\eta_1 \cdot v(x) - \eta_2$ in (3.4) is piecewise constant with discontinuity points $-\delta$ and δ , we consider a generic differential equation:

$$\frac{g''}{g} + C = 0, \tag{3.7}$$

where $C \in \mathbb{R}$ is a constant and $g = \sqrt{f_0}$. The general solution to the above equation, up to a constant, is:

- if $C = 0$, $g = x + c_1$,
- if $C > 0$, $g = \cos(x + c_2)$, and
- if $C < 0$, $g = \exp\{-\sqrt{-C}|x|\}$,

where c_1 and c_2 are constants. Since we want $g(\pm\infty) = 0$, we must assume $-\eta_1 \cdot v(x) - \eta_2 < 0$, i.e., $C < 0$, outside the interval $[-\delta, \delta]$, which leads to the only functional form that vanishes at infinity. Inside the interval $[-\delta, \delta]$, we assumed $-\eta_1 \cdot v(x) - \eta_2 > 0$, which leads to the objective function in RSBN. If we assume $-\eta_1 \cdot v(x) - \eta_2 = 0$, then we have $g(x) = x$, which will eventually lead to Huber's estimator. Here is where we deviate from Huber's derivation.

As mentioned earlier, some historical description seems to imply that Huber's estimator is a unique minimax estimator, e.g., see Lehmann (1991, Section 5.6). They consider an asymptotic minimax estimator among all the c.d.f. $F(x) = (1 - \varepsilon)G(x) + \varepsilon H(x)$, where constant ε and c.d.f. $G(x)$ are known, and c.d.f. $H(x)$ is unknown but satisfies some general conditions. When $G(x) = \Phi(x)$, the c.d.f. of the standard normal distribution, the minimax estimator is Huber's estimator. We solve the minimax problem in a more general sense.

3.6. Parameters in RSBN

The parameters $c, \delta, \alpha, \lambda_1, \lambda_2$ satisfy the following conditions:

$$\int_{-\delta}^{\delta} f_0(x) dx = 1 - \alpha \quad \text{and} \quad \lim_{x \rightarrow \delta^+} f'(x) = \lim_{x \rightarrow \delta^-} f'(x). \tag{3.8}$$

So, we have

$$\lambda_2 = \lambda_1 \tan \lambda_1. \tag{3.9}$$

The relationship between λ_1 and α is revealed by the following lemma.

Lemma 3.5. *The proportion α defined in the stochastically bounded noise and the parameter λ_1 in RSBN have the relation:*

$$\alpha = \frac{\cos^3 \lambda_1}{\lambda_1 \cdot \sin \lambda_1 + \cos \lambda_1}. \tag{3.10}$$

The details of the derivation of the above relation and a figure that illustrates the relationship between α and λ_1 can be found in an online technical report (Huo and Ni, 2005).

Now we consider a simplified version of (3.6). As an objective function in (2.2), the following ρ is equivalent to the one in (3.6):

$$\rho(x) = \begin{cases} -\log \cos \lambda_1 \frac{x}{\delta}, & |x| < \delta, \\ \lambda_2 \frac{|x|}{\delta} - \lambda_2 - \log \cos \lambda_1, & |x| \geq \delta. \end{cases} \tag{3.11}$$

Bringing in (3.9), we get exactly the expression in (3.1). From all the above, we have established Theorem 3.1.

The following flow chart summarizes the procedure for constructing ρ :

$$\alpha, \delta \xrightarrow{(3.12)} \lambda_1 \xrightarrow{(3.11)} \lambda_2 \xrightarrow{(3.13)} \rho.$$

3.7. Other properties

- *Fisher information of the least informative distribution:* We consider the Fisher information $I(f_0)$. Let $f_\theta = f_0(x - \theta)$, where f_0 is the least informative distribution in Section 3.4. We have

$$I(f_0) = 4 \frac{\lambda_1^2}{\delta^2} \frac{\lambda_1 \cdot \sin \lambda_1}{\lambda_1 \cdot \sin \lambda_1 + \cos \lambda_1}. \tag{3.12}$$

The details of validating the above equation are posted in Huo and Ni (2005). Taking $\delta = 1.0$ and combining (3.10) and (3.12), we have the relationship between the Fisher information $I(f_0)$ and α . Since $\lambda_1 \in [0, \pi/2]$, the range of the Fisher information $I(f_0)$ is from 0 to π^2/δ^2 . A figure in Huo and Ni (2005) shows the relationship between α and $I(f_0)$. We find that a small α leads to a large Fisher information.

- *Asymptotic variance of RSBN:* As for the asymptotic variance, in Section 3.1, we have already shown that

$$\text{asymptotic variance} = \frac{\int \psi^2 dF}{(\int \psi' dF)^2}. \tag{3.13}$$

Since $\psi = \rho'$, the formulae for $\psi(\cdot)$ and $\psi'(\cdot)$ can be easily derived. Expressions for these two functions can be found in [Huo and Ni \(2005\)](#).

4. Numerical approaches

Two distinct numerical approaches are presented in Sections 4.1 and 4.2, respectively.

4.1. Utilizing the proximal point method

In this subsection, we describe a proximal point algorithm which solves RSBN. The purpose is to give the readers who do not have access to a sophisticated software package an extremely easy-to-use algorithm. [Huo and Ni \(2005\)](#) give a detailed description of the proximal point method.

Considering the first order necessary condition for (2.2), we have

$$0 = A^T \psi(A\beta - y), \quad (4.1)$$

where $\psi = \rho'$, $\psi(y - A\beta) = [\psi((y - A\beta)_1), \psi((y - A\beta)_2), \dots, \psi((y - A\beta)_n)]^T$, and $(y - A\beta)_i$ denotes the i th component of the vector $y - A\beta$. Eq. (4.1) is equivalent to

$$\text{find } u, v : \begin{cases} u = A\beta \\ v = \psi(u - y) \\ 0 = A^T v \end{cases} \quad \text{i.e., finding } u, v : \begin{cases} u \in \text{Range}(A), \\ v \in \text{Kernel}(A), \\ v = \psi(u - y). \end{cases} \quad (4.2)$$

Applying the proximal point method to RSBN, we have the following algorithm for RSBN.

- (1) **Choose** $\mu^{(0)} \in \mathbb{R}^n$, $k = 0$.
- (2) **Find** u_i , such that $\psi(u_i - y_i) + u_i = \mu_i^{(k)}$, $i = 1, 2, \dots, n$.
- (3) **Let** $v_i = \mu_i^{(k)} - u_i$, $i = 1, 2, \dots, n$.
- (4) **Project** $u = (u_1, \dots, u_n)^T$, $v = (v_1, \dots, v_n)^T$.

$$\begin{aligned} \mu^{(k+1)} &= \mathbf{P}_A(u) + \mathbf{P}_{\text{Ker}(A)}(v) \\ &= v + A(A^T A)^{-1} A^T (u - v). \end{aligned}$$

Here \mathbf{P}_A and $\mathbf{P}_{\text{Ker}(A)}$ are the projection operators projecting the input onto the range of A and the kernel of A , respectively.

- (5) If $\mu^{(k)}$ does not converge, $k = k + 1$, go back to step (2).

In step (2), since ψ is monotone increasing, u_i will have a unique solution. But because there is a tangent function in ψ in RSBN, one needs to implement a line search algorithm to solve it. We can see that if ψ is piecewise polynomial, this method is quite appealing, because a closed-form solution is available to the equation in step (2). This approach has been used to solve for Huber's estimator—see [Michelot and Bougeard \(1994\)](#). After finding u , β can be found via $u = A\beta$. It is possible that our algorithm converges slowly to the solution—we refer to the analysis that has been conducted in [Huo and Ni \(2005\)](#).

4.2. Other implementation: SNOPT with SQP

As an alternative, we use a state-of-the-art optimization software package to solve for the RSBN *directly*. We use a general-purpose optimization package—SNOPT, a software package developed in [Gill et al. \(1998\)](#). It minimizes a linear or nonlinear function subject to bounds on the variables, as well as sparse linear or nonlinear constraints. It is suitable for large-scale linear and quadratic programming and for linearly constrained optimization, as well as for general nonlinear programs. In our case, in (2.2), we have linear constraints and a nonlinear but convex objective function.

SNOPT finds a solution that is *locally optimal*. Ideally, all the nonlinear functions should be smooth and the user should provide the gradients. In our case, since the objective function in (2.2) is convex, the *local optimal* solution will coincide with the global optimal solution.

SNOPT uses a sequential quadratic programming (SQP) algorithm that obtains a search direction from a sequence of quadratic programming subproblems. Each QP subproblem minimizes a quadratic model of a certain Lagrangian function subject to a linearization of the constraints. An augmented Lagrangian merit function is reduced along each search direction to ensure convergence from any starting point.

The source code for SNOPT is written in Fortran. In order to use it, a Fortran compiler is required. The numerical examples in the present paper are a result of combining some MATLAB programming, Unix shell programming, Fortran programming, and SNOPT.

5. Simulation

5.1. An illustrative example: variable star

In this section, we study a well-known data set in time series analysis—the magnitudes of a variable star at midnight on 600 successive nights. Bloomfield (1976) showed that it is a superposition of two ‘dominant’ sinusoid functions. We take a slightly different viewpoint, and assume that the underlying signal (denoted by \mathbf{s}) is a smooth signal residing in a low dimensional linear subspace. The observed magnitudes, denoted by $\mathbf{y} = (y_1, y_2, \dots, y_{600})^T$, are an approximation to \mathbf{s} . In our case, \mathbf{y} is the rounded version of \mathbf{s} : i.e., $y_i = [s_i + 0.5]$, where $[x]$ is the largest integer no larger than x . It is evident that the mapping from \mathbf{s} to \mathbf{y} is completely nonlinear. Let $\mathbf{y} = \mathbf{s} + \mathbf{n}$, where \mathbf{n} is the so-called noise sequence. Considering the source from where the noise sequence is generated, we must agree that the Gaussian assumption on the distribution of \mathbf{n} is not appropriate. In this case, we compare the ordinary least squares estimator, Huber’s estimator, and RSBN.

We consider the discrete cosine transform (DCT) of the original signal. The DCT with signal length n has the k th basis function:

$$c_k(i) = \begin{cases} \sqrt{1/n}, & k = 0, \quad i = 1, 2, \dots, n, \\ \sqrt{2/n} \cos \left[\left(i - \frac{1}{2} \right) k \frac{\pi}{n} \right], & k \neq 0, \quad i = 1, 2, \dots, n. \end{cases}$$

The reasons for choosing DCT are: (a) DCT is an analogous of the Fourier transform, which is widely adopted in representing cyclic signals; (b) there are fast numerical algorithms to implement DCT.

In this simulation, we first find the subspace that contains most of the signal’s energy. This can be done by carrying out a DCT transform, then retaining the coefficients with the largest magnitudes. Later, we intentionally distort the observation. Then, three different projections are compared. We illustrate the optimality of our method.

Fig. 2(a) shows the magnitude vector \mathbf{y} of the variable star. We take the DCT of \mathbf{y} , and keep 10% of the coefficients with the largest magnitudes. The associated 10% basis functions span the subspace that contains the largest possible proportion of the energy. We denote the subspace by A . The dimension of A is 60. Projecting the observation \mathbf{y} onto the subspace spanned by A (i.e., $\text{Range}(A)$) by ordinary least squares regression, we have $\mathbf{P}_{A,LS}(\mathbf{y}) = \hat{\mathbf{s}}_{LS,1}$, where ‘LS’ indicates the least squares estimator, and ‘1’ indicates it is from the original observation \mathbf{y} . (In other words, vector $\hat{\mathbf{s}}_{LS,1}$ minimizes $\|\mathbf{u} - \mathbf{y}\|_{\ell_2}^2$ among all the vectors \mathbf{u} that are in $\text{Range}(A)$.) We then project the observation \mathbf{y} to A by using RSBN. We choose $\delta = 0.5$, $\lambda_1 = 0.46\pi$, and ρ is given in (3.11). We denote $\mathbf{P}_{A,RSBN}(\mathbf{y}) = \hat{\mathbf{s}}_{RSBN,1}$. Since the deviations, $\mathbf{y} - \hat{\mathbf{s}}_{RSBN,1}$ (shown in Fig. 2(c)) or $\mathbf{y} - \hat{\mathbf{s}}_{LS,1}$ (shown in Fig. 2(b)), are supposed to be round-off errors, ideally they should fall within the interval $[-0.5, 0.5]$. For the least squares estimator, there are 70 deviations

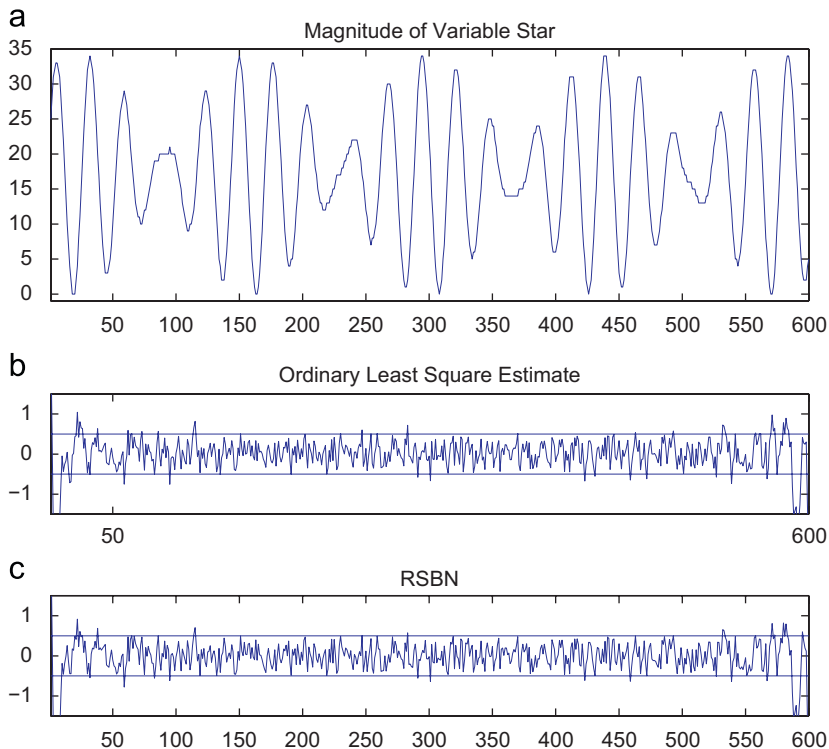


Fig. 2. Variable star data set and the deviations by the least squares regression and RSBN.

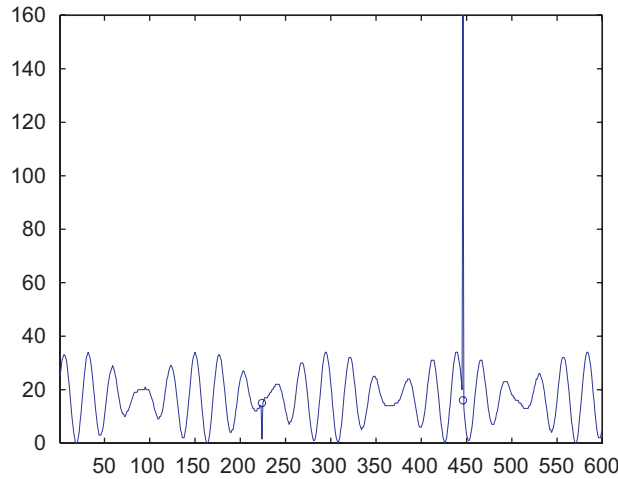


Fig. 3. The distorted variable star signal. Circles indicate the true values.

with magnitudes larger than 0.5, and 16 of them having magnitudes larger than 1.0. For RSBN, there are 44 deviations having magnitudes larger than 0.5, and 15 of them having magnitudes larger than 1.0. In this case, compared to the ordinary least squares estimator, RSBN has fewer deviations falling beyond the ideal interval $[-0.5, 0.5]$. Of course, at the same time, we should observe a loss in the mean square error, which is what the least squares estimator aims to minimize. The sum of squared deviations for the least squares estimator is 10.4337, and for RSBN it is 10.6022.

Now we randomly pick two positions in the variable star sequence. In particular, we choose position 224 and 446. Originally, $y_{224} = 15$ and $y_{446} = 16$. Suppose the decimal points in these numbers were somehow misspecified, and the recorded values became $y'_{224} = 1.5$ and $y'_{446} = 160$. Without loss of generality, let \mathbf{y}' denote the new sequence. Fig. 3 shows \mathbf{y}' .

Recall $\mathbf{P}_{A,LS}$ and $\mathbf{P}_{A,RSBN}$ denote the projection operators onto $\text{Range}(A)$ by the least squares method and RSBN, respectively. Let $\mathbf{P}_{A,H}$ denote a projection operator onto A via Huber's estimator. Recall that a general form of Huber's estimator is, for $\Delta > 0$,

$$\rho(x) = \begin{cases} x^2, & |x| < \Delta, \\ 2\Delta|x| - \Delta^2, & |x| \geq \Delta, \end{cases}$$

in (2.2) (Huber, 1977, 1981). In Huber's estimator, ρ is piecewise linear (outside a neighborhood of the origin) or quadratic (inside the neighborhood).

Consider the projections $\hat{\mathbf{s}}_{LS,2} = \mathbf{P}_{A,LS}(\mathbf{y}')$, $\hat{\mathbf{s}}_{RSBN,2} = \mathbf{P}_{A,RSBN}(\mathbf{y}')$, and $\hat{\mathbf{s}}_{H,2} = \mathbf{P}_{A,H}(\mathbf{y}')$. The deviations $\mathbf{y} - \hat{\mathbf{s}}_{LS,2}$, $\mathbf{y} - \hat{\mathbf{s}}_{RSBN,2}$ and $\mathbf{y} - \hat{\mathbf{s}}_{H,2}$ are plotted in Fig. 4(a)–(c). Note these are the deviations from the estimates of the 'original' signal sequence \mathbf{y} (not \mathbf{y}'). Table 1 shows some statistics on the performance of these three different methods.

There are several phenomena worth noting. First of all, the deviations of the least squares estimate are significantly worse than the other two. This illustrates that the least squares estimator is not robust. Second, the performance of RSBN has almost no difference between the two cases: \mathbf{y} and \mathbf{y}' . In other words, $\hat{\mathbf{s}}_{RSBN,2}$ is as close to \mathbf{y} as $\hat{\mathbf{s}}_{RSBN,1}$ is. Third, RSBN performs slightly better than Huber's estimator. It is not surprising that RSBN and Huber's estimator have similar performance, since the objective functions in (2.2) for these two are very close to each other. One commonality is that are both linear outside an interval: $(-\delta, \delta)$ in RSBN and $(-\Delta, \Delta)$ in Huber's.

5.2. Comparison with the ordinary least squares estimator and Huber's estimator

We compare three different regression methods: the ordinary least squares estimator, RSBN, and Huber's estimator. We demonstrate that for mixed Gaussian noise, RSBN does the best job.

5.2.1. Design of the simulation

We choose $m = 15$ and $n = 600$. In each experiment, for the model in (2.1), A is generated by sampling each entry $(A_{ij}, 1 \leq i \leq n, 1 \leq j \leq m)$ from a standard normal distribution $(N(0, 1))$, with the constraint that A must have a full column rank. If the generated matrix A does not have a full column rank, the process is repeated instead of proceeding to the next step. The vector $\boldsymbol{\beta}$ is generated as a standard normal vector in \mathbb{R}^m , $\boldsymbol{\beta} \sim N(0, \mathbf{I}_m)$. The vector $\boldsymbol{\varepsilon}$ is generated as a standard normal vector in \mathbb{R}^n , $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}_n)$. The observation vector \mathbf{y} is then computed as $\mathbf{y} = A\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

Recall the operator $\mathbf{P}_{A,LS} : \mathbb{R}^n \rightarrow \text{Range}(A)$ is the projection operator from the Euclidean space \mathbb{R}^n to the linear subspace $\text{Range}(A)$. In other words,

$$\mathbf{P}_{A,LS}(\mathbf{y}) = \underset{\mathbf{u} \in \text{Range}(A)}{\text{argmin}} \|\mathbf{u} - \mathbf{y}\|_{\ell_2}^2.$$

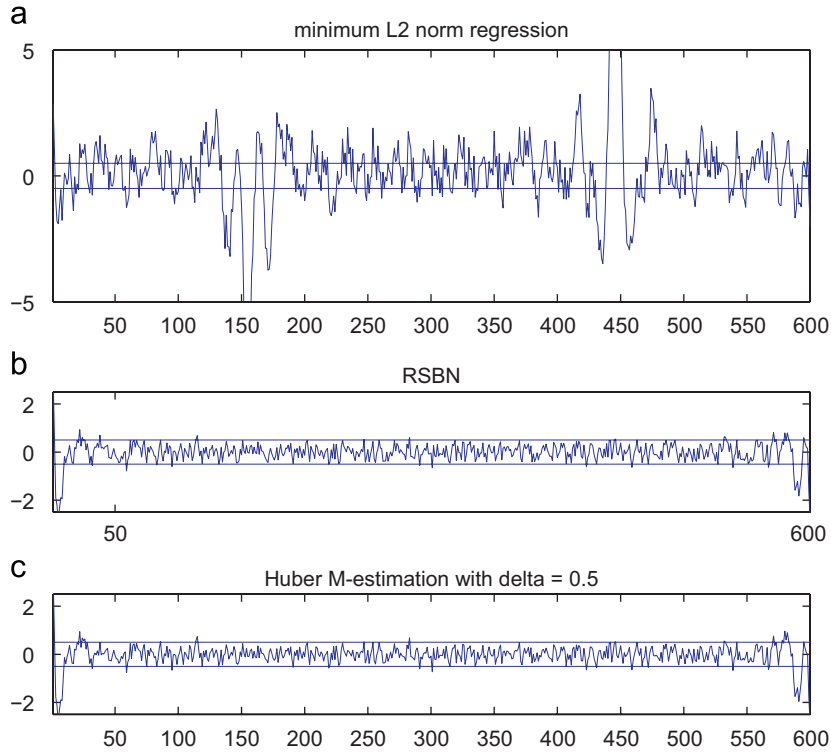


Fig. 4. Differences between the original variable star signal and the estimates from the distorted signal by three methods.

Table 1
Some statistics for the three regression methods on the distorted variable star data

	OLS	Huber's	RSBN
Square root of sum of squares, $\ \mathbf{y} - \hat{\mathbf{s}}_{*,2}\ _2$	47.2771	10.6658	10.6053
Number of magnitudes > 0.5	372	53	45
Number of magnitudes > 1.0	196	16	15

Let $d_{LS,1}$ denote the deviation vector from the least squares projection $\mathbf{P}_{A,LS}(\mathbf{y})$ to the true linear component $A\boldsymbol{\beta}$. Note here the first subscript 'LS' indicates the least squares method, and the second subscript '1' in $d_{LS,1}$ indicates the Gaussian noise vector (ε). We have

$$d_{LS,1} = \mathbf{P}_{A,LS}(\mathbf{y}) - A\boldsymbol{\beta} = \mathbf{P}_{A,LS}(\varepsilon). \tag{5.1}$$

We then distort the Gaussian vector ε . We randomly select 1% of the entries in ε , multiply them by 200 (value 200 is arbitrarily chosen). The new vector is denoted by ε' . Effectively, each entry of ε' follows a mixed normal distribution: $\varepsilon'_i \sim 0.99N(0, 1) + 0.01N(0, 200^2)$, $1 \leq i \leq n$. Denote $\mathbf{y}' = A\boldsymbol{\beta} + \varepsilon'$.

Let $d_{LS,2}$, $d_{RSBN,2}$ and $d_{Huber,2}$ denote the deviation vectors corresponding to the least squares estimator, RSBN, and Huber's estimator, respectively. Here, the second subscript '2' indicates the distorted noise vector ε' . We have $d_{*,2} = \mathbf{P}_{A,*}(\mathbf{y}') - A\boldsymbol{\beta} = \mathbf{P}_{A,*}(\varepsilon')$, where '*' can be LS, RSBN, or Huber.

We repeat the experiments 1000 times. Each time, for the distorted errors, the three methods lead to three deviation vectors: $d_{LS,2}$, $d_{RSBN,2}$ and $d_{Huber,2}$. Letting $d_{LS,2}^{(i)}$, $d_{RSBN,2}^{(i)}$ and $d_{Huber,2}^{(i)}$ denote the deviation vectors we get in the i th experiment, we have in total 3000 n -D vectors: $d_{LS,2}^{(i)}$, $d_{RSBN,2}^{(i)}$, $d_{Huber,2}^{(i)}$, $i = 1, 2, \dots, 1000$.

The smaller the deviations, the better the regression method.

5.2.2. Cut-off value

To measure the robustness of different methods, it is natural to compare the deviation vectors $d_{LS,2}$, $d_{RSBN,2}$, and $d_{Huber,2}$ with the deviation vector $d_{LS,1}$, because $d_{LS,1}$ is the deviation of an ideal method (least squares estimation or MLE) in the ideal situation (with Gaussian errors). We propose to study the number of deviations with magnitudes above a quantity τ : i.e., for $1 \leq i \leq 1000$,

$$\#\{j : |(d_{*,2}^{(i)})_j| > \tau, 1 \leq j \leq n\},$$

where * can be LS, RSBN, or Huber. Here # stands for the cardinality of a finite set. The *j*th component of $d_{*,2}^{(i)}$ is denoted as $(d_{*,2}^{(i)})_j$. Quantity τ , which is called the *cut-off* value, can be viewed as a quantile of $\|d_{LS,1}\|_\infty$.

The following is to derive a reasonable value of τ . We study the distribution of the random variable $\|d_{LS,1}\|_\infty$. Let $\|d_{LS,1}\|_2$ denote the ℓ_2 norm of $d_{LS,1}$. We have

$$\|d_{LS,1}\|_\infty = \|d_{LS,1}\|_2 \cdot \frac{\|d_{LS,1}\|_\infty}{\|d_{LS,1}\|_2}.$$

We list three facts. For details, please refer to the Appendix of [Huo and Ni \(2005\)](#):

- $\|d_{LS,1}\|_2$ and $\|d_{LS,1}\|_\infty/\|d_{LS,1}\|_2$ are independent.
- $\|d_{LS,1}\|_2^2$ has a χ_m^2 distribution with m degrees of freedom, where m is the column rank of A .
- Assume that the projection $\mathbf{P}_{A,LS}$ has the eigenvalue decomposition

$$\mathbf{P}_{A,LS} = U^T \begin{pmatrix} \mathbf{I}_m & \\ & \mathbf{0} \end{pmatrix} U,$$

where U is orthogonal. Let

$$\beta = U^T \begin{pmatrix} \beta_m \\ \mathbf{0}_{(n-m) \times 1} \end{pmatrix},$$

where β_m is uniform on the unit sphere in \mathbb{R}^m , $\|\beta_m\|_2 = 1$, and $\mathbf{0}_{(n-m) \times 1}$ is an all zero vector. Let $\rho_{\max,m} = \|\beta\|_\infty$. The ratio $\|d_{LS,1}\|_\infty/\|d_{LS,1}\|_2$ has the same distribution as $\rho_{\max,m}$. The analytical form of the probability density function of $\rho_{\max,m}$ could be too complicated to be useful, however.

Based on the above three facts, we can find the distribution of $\|d_{LS,1}\|_\infty$ and the cut-off value through simulations. In this paper, we choose the cut-off value $\tau = 1$. The related probability $P\{\|d_{LS,1}\|_\infty > \tau\}$ is approximately 3.1×10^{-4} , which is estimated from 100,000 simulations.

5.2.3. Simulation results

[Fig. 5](#) illustrates the results from all the steps of one simulation.

- [Fig. 5\(a\)](#) shows the Gaussian noise vector ε . Each element of it satisfies the distribution $\text{Normal}(0, 1)$.
- [Fig. 5\(b\)](#) shows the deviation vector $(d_{LS,1})$ of the least squares regression in the Gaussian noise case.
- [Fig. 5\(c\)](#) shows the distorted Gaussian noise vector ε' . The vector ε' is generated by multiplying six randomly chosen elements of ε by 200.
- [Fig. 5\(d\)](#) shows the deviation vector $d_{LS,2}$ from the least squares regression with the distorted Gaussian noise ε' .
- [Fig. 5\(e\)](#) shows the corresponding deviation vector $d_{RSBN,2}$ from RSBN.
- [Fig. 5\(f\)](#) shows the corresponding deviation vector $d_{Huber,2}$ from Huber's estimator.

From [Section 5.2.1](#), we get 3000 deviation vectors out of 1000 simulations: $d_{LS,2}^{(i)}, d_{RSBN,2}^{(i)}, d_{Huber,2}^{(i)}, i = 1, 2, \dots, 1000$.

We choose two ways to compare the three different methods. One is to study the relative ratio of the ℓ_2 norms of a pair of the deviation vectors. The other is to count the number of magnitudes above the cut-off line (determined by the τ value developed in [Section 5.2.2](#)) in each deviation vector.

[Fig. 6\(a\)](#) gives a histogram of the ratios of the ℓ_2 norms of deviation vectors from Huber's estimator and RSBN: $\|d_{Huber,2}^{(i)}\|_2^2 / \|d_{RSBN,2}^{(i)}\|_2^2, i = 1, 2, \dots, 1000$. We observe that most of them are above 1. This implies that RSBN tends to give smaller sum square of deviations than Huber's estimator. [Fig. 6\(b\)](#) shows a histogram of the logarithm (base 10) of the ratios corresponding to the least squares estimator and RSBN: $\log_{10}(\|d_{LS,2}^{(i)}\|_2^2 / \|d_{RSBN,2}^{(i)}\|_2^2), i = 1, 2, \dots, 1000$. The reason to take logarithms is that some ratios can be extremely large. Obviously, the least squares regression for non-Gaussian errors leads to much higher sum of squares of deviations than the RSBN does.

Define the numbers of magnitudes above the cut-off in the following way:

$$\Gamma_{*,2}^{(i)} = \#\{j : |(d_{*,2}^{(i)})_j| > 1, 1 \leq j \leq n\},$$

where * can be the subscripts: LS, RSBN, or Huber. We observe that for all $1 \leq i \leq 1000$, $\Gamma_{RSBN,2}^{(i)} = 0$. This means that RSBN is very robust (in the sense that there is no outstanding deviation from the true signal). Huber's estimator performs comparably. [Fig. 6\(c\)](#) gives a histogram of $\Gamma_{Huber,2}^{(i)}, i = 1, 2, \dots, 1000$. We observe that 15 of them have 1 deviation whose magnitude is larger than 1, and only 1 has 2 deviations whose magnitudes are greater than 1. [Fig. 6\(d\)](#) shows a histogram of $\Gamma_{LS,2}^{(i)}, i = 1, 2, \dots, 1000$. We can see that in most simulations, the number of deviations with magnitudes above the cut-off 1 is large. The average number of deviations with magnitudes above the cut-off is 421, which is roughly 70% of the signal.

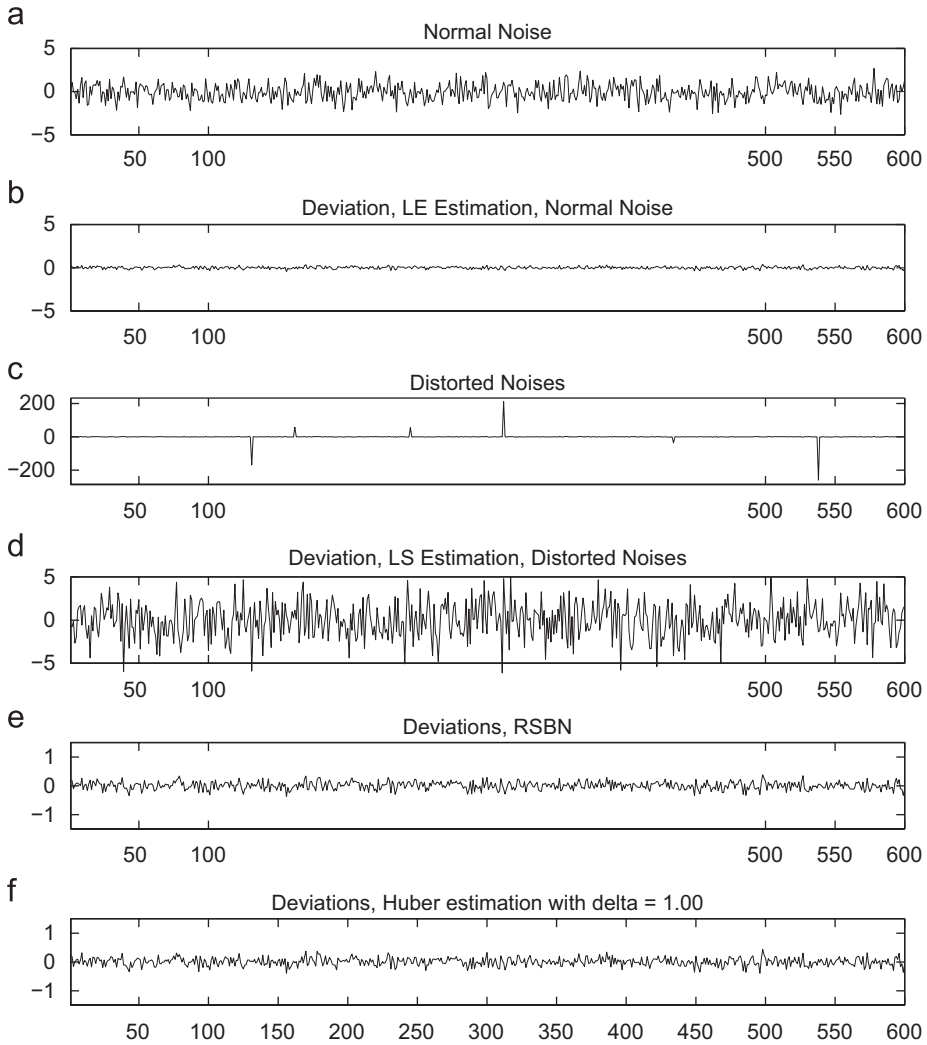


Fig. 5. Quantitative results in one simulation. See text for the details.

In this simulation, RSBN outperforms Huber's estimator, and both significantly outperforms the ordinary least squares estimator.

6. Discussion

6.1. A general regression formulation

Eq. (2.2) is consistent with many approaches that exist in the literature.

- For $\Delta > 0$, we have $\rho(x) = 0$, when $|x| < \Delta$; and $\rho(x) = |x| - \Delta$, when $|x| \geq \Delta$. Formulation (2.2) is an ℓ_1 regression with a 'dead zone'. By adding some slack variables, (2.2) can be formulated as a linear programming problem. The reader can verify that the following linear programming problem is equivalent to the problem in (2.2):

$$\begin{aligned}
 &\text{minimize} && \sum_{i=1}^n t_i \\
 &\text{subject to} && -t_i - \Delta \leq y_i - a_i^T \beta, \quad i = 1, 2, \dots, n, \\
 &&& y_i - a_i^T \beta \leq t_i + \Delta, \quad i = 1, 2, \dots, n, \\
 &&& 0 \leq t_i, \quad i = 1, 2, \dots, n.
 \end{aligned}$$

The idea behind adding a dead zone is to make the large residual relatively more important.

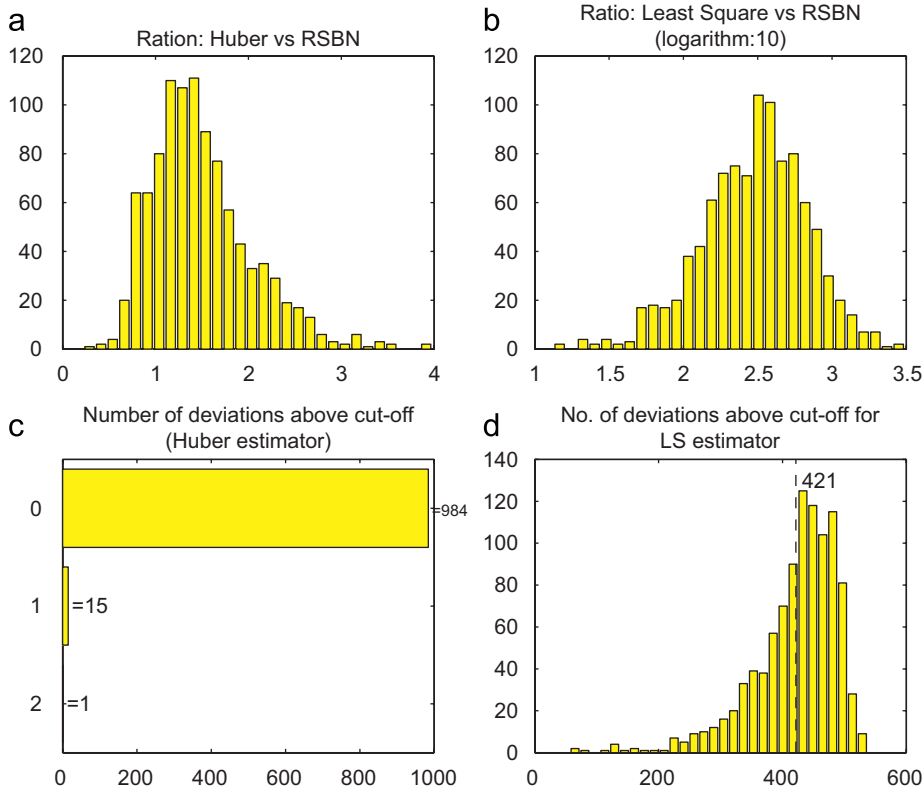


Fig. 6. (a) The histogram of the ratios between Huber's estimator and RSBN: $\|d_{\text{Huber},2}^{(i)}\|_2^2 / \|d_{\text{RSBN},2}^{(i)}\|_2^2$, $i = 1, 2, \dots, 1000$; (b) the histogram of the logarithm ratios $\log_{10}(\|d_{\text{LS},2}^{(i)}\|_2^2 / \|d_{\text{RSBN},2}^{(i)}\|_2^2)$, $i = 1, 2, \dots, 1000$, for the least squares regression and RSBN; (c) for Huber's estimator, the number of deviations whose magnitudes are above the cut-off; (d) for the ordinary least squares regression, the histogram of the number of deviations whose magnitudes are above the cut-off.

- If $\rho(x) = |x|$, (2.2) is the standard least ℓ_1 norm estimation (Dodge, 1987). It can be solved as a linear programming problem (Vanderbei, 1996). This can be viewed as a special case of the last problem: $\Delta = 0$. This formulation is interesting when the errors are Laplacian: i.e., the errors satisfy an exponential distribution. Li and Swetits (1998) established an analytical connection between Huber's estimator (with Δ being a tuning parameter) and the least ℓ_1 norm estimator (which was called the linear ℓ_1 estimator in Li and Swetits, 1998). Their result is based on analyzing the solutions to the dual problem.

6.2. Other theoretical results

Some results that are related to estimators in regression are worth mentioning. Researchers have explored the robustness of some regression approaches. For example, Ellis and Morgenthaler (1992) analyzed the 'leverage' and 'breakdown' in minimum ℓ_1 norm regression. The objective in that paper is different from ours: e.g., they do not consider the asymptotic performance as what we discussed and they do not consider minimaxity. However, their work is inspiring. A citation search of Ellis and Morgenthaler (1992) gives a good sense of what is known about the robustness of some estimators in regression.

In our formulation, we assume the errors are independent. Other conditions regarding the regularity of the probability density function of the errors—e.g., the existence of the second derivative of the density, as well as some integrability conditions—are embedded in the derivation of the asymptotic minimaxity. Researchers have studied the conditions for an M-estimate to be consistent. The introduction of Berlinet et al. (2000) provides a nice overview. Further citation search for the papers cited there gives a full spectrum of the results that are available. In this paper, we did not intend to address some of those issues. However, it will be an interesting future research topic to derive a minimax M-estimator under weaker regularity conditions.

6.3. Misc.

- *Convexity of Fisher information $I(f)$* : In our derivation, we implicitly used the result that Fisher information $I(f)$ is a convex function of f . We give a brief verification of such a convexity in the discussion of Huo and Ni (2005).

- *Local minimaxity*: We can only verify that our RSBN estimator is minimax at a neighborhood of function f_0 —cf. Lemma 3.4. Proving that RSBN is a minimax estimator globally (i.e., for all the functions satisfying the ‘stochastically bounded noise’ condition) seems to be a difficult task. This problem has not been solved in our paper.
- *ℓ_1 function*: There is an interesting similarity between the derived minimax estimator and some criterion functions in model selection. Specifically, the fact that the objective function becomes linear outside a neighborhood of the origin coincides with the ℓ_1 -norm principle that has recently gained popularity via methods such as LASSO (Tibshirani, 1996) and Basis Pursuit (Chen et al., 2001).
- *Unknown scale*: In the ‘stochastically bounded noise’ condition, the parameters δ and α are assumed to be known. In practice, these parameters need to be estimated. In fact, Huber’s estimator assumes a known scale of the errors, and so does RSBN. Neither method is scale-invariant. Moreover, both methods assume that there is no leverage point present. This is the limitation of using the formulation (2.2).
- *Regarding practicality*: In this work, we restrict ourselves to M-estimates. Many other robust statistics have been proposed in the literature, going beyond the framework of (2.2). It is impossible and unnecessary to summarize them here. We would not say that RSBN is by far the best robust estimator in regression. In fact, our simulations seem to indicate that there is no significant difference between Huber’s estimator and RSBN. Huber’s estimator has been outperformed in many publications. The use of the variational approach in deriving an asymptotic minimax estimator is the main contribution of this paper.

References

- Berlinet, A., Liese, F., Vajda, I., 2000. Necessary and sufficient conditions for consistency of m-estimates in regression models with general errors. *J. Statist. Plann. Inference* 89 (1–2), 243–267.
- Bloomfield, P., 1976. *Fourier Analysis of Time Series: An Introduction*. Wiley, New York.
- Chen, S.S., Donoho, D.L., Saunders, M.A., 2001. Atomic decomposition by basis pursuit. *SIAM Rev.* 43 (1), 129–159 reprinted from *SIAM J. Sci. Comput.* 20(1) (1998) 33–61.
- Dodge, Y., 1987. *Statistical Data Analysis: Based on the L_1 -norm and Related Methods*. North-Holland, Amsterdam.
- Ellis, S.P., Morgenthaler, S., 1992. Leverage and breakdown in l_1 regression. *J. Amer. Statist. Assoc.* 87, 143–148.
- Gill, P.E., Murray, W., Saunders, M.A., 1998. User’s guide for SNOPT 5.3: a Fortran package for large-scale nonlinear programming. Draft.
- Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 1986. *Robust Statistics: The Approach Based on Influence Functions*. Wiley Series in Probability and Mathematical Statistics.
- Huber, P.J., 1977. *Robust Statistical Procedures*, vol. 27. CBMS-NSF.
- Huber, P.J., 1981. *Robust Statistics*. Wiley Series in Probability and Mathematical Statistics.
- Huo, X., Ni, X.S., 2005. RSBN: regression with stochastically bounded noises. Technical Report, Georgia Institute of Technology, Atlanta, GA (<http://www2.isye.gatech.edu/statistics/papers/05-22.pdf>), May.
- Lehmann, E.L., 1991. *Theory of Point Estimation*. Wadsworth & Brooks/Cole, Pacific Grove, CA.
- Li, W., Swetits, J.J., 1998. The linear l_1 estimator and the Huber m-estimator. *SIAM J. Optim.* 8 (2), 457–475.
- Michelot, C., Bougeard, M.L., 1994. Duality results and proximal solutions of the Huber m-estimator problem. *Appl. Math. Optim.* 30, 203–221.
- Ni, X.S., 2005. New results in detection, estimation, and model selection. Ph.D. Thesis, Georgia Institute of Technology. Available at: (<http://etd.gatech.edu/theses/available/etd-12042005-190654/>).
- Tibshirani, R.J., 1996. Regression shrinkage and selection via the LASSO. *J. Roy. Statist. Soc. Ser. B* 58, 267–288.
- Vanderbei, R.J., 1996. *Linear Programming*. Kluwer Academic Publishers, Dordrecht.