# Asymptotics of Bayesian Median Loss Estimation

Chi Wai Yu[a], Bertrand Clarke[b]

[a] *Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong*
[b] *Department of Medicine, Department of Epidemiology and Public Health, and the Center for Computational Sciences, University of Miami, 1120 NW 14th Street, Miami, FL, 33136.*

**Abstract**

We establish the consistency, asymptotic normality, and efficiency for estimators derived by minimizing the median of a loss function in a Bayesian context. We contrast this procedure with the behavior of two Frequentist procedures, the least median of squares (LMS) and the least trimmed squares (LTS) estimators, in regression problems. The LMS estimator is the Frequentist version of our estimator, and the LTS estimator approaches a median-based estimator as the trimming approaches 50% on each side. We argue that the Bayesian median-based method is a good tradeoff between the two Frequentist estimators.

*Key words:* Asymptotics, least median of squares estimator, least trimmed squares estimator, loss function, median, posterior, regression
*2000 MSC:* 62F12, 62F15, 62J02

## 1. Introduction

Conventional statistical techniques like estimation and hypothesis testing can be embedded in the expected loss framework of Wald's Statistical Decision Theory (see Wald [26, 27]). However, Yu and Clarke [29] observe that the loss function as a random variable often has a distribution that is strongly right skewed. It is well known that, for such distributions, the mean is not a good summary statistic because it can be too sensitive to the long right tail. That is, the expected loss, i.e., the risk, will in general not be representative of the location of the distribution of the loss. Consequently, the risk minimizing action will typically permit larger deviations than necessary in prediction problems.

One way around this is to minimize a different feature of the loss function, namely its median, which is well known to be more representative of the location of a skewed distribution than the mean is. Thus, here, we systematically replace the expectation of the loss with the median of the loss (hereafter *medloss*). In terms of prediction, this helps avoid overprediction and underprediction, see [28]. Moreover, it is straightforward to identify a median analog of the Bayes estimator, here called the *posterior medloss estimator*, which minimizes the

*Email addresses:* macwyu@yahoo.com.hk (Chi Wai Yu), bclarke2@med.miami.edu (Bertrand Clarke)

median of the loss with respect to the posterior. That is, the posterior *medloss* estimator is

$$\delta(\mathbf{x}^n) = \arg\min_{\mathbf{a}\in\mathcal{D}} \operatorname*{med}_{\pi(\Theta|x^n)} \mathcal{L}(\mathbf{a}(\mathbf{x}^n), \Theta), \tag{1}$$

where $\mathbf{x}^n = \{\mathbf{x}_i : i = 1, \ldots, n\}$ are the realizations of $n$ $d$-dimensional random vectors $\mathbf{X}^n = \{\mathbf{X}_i \in \mathcal{R}^d : i = 1, \ldots, n\}$, $\mathcal{L}(\mathbf{a}, \theta)$ is a nonnegative loss function, $\mathbf{a}(\mathbf{x}^n)$ is an estimate of $\theta$, $\mathcal{D} \subset \mathcal{R}^p$ is the decision space, and $\operatorname*{med}_{\pi(\Theta|x^n)} \mathcal{L}$ is the median of the loss $\mathcal{L}$ under the posterior density $\pi(\cdot|x^n)$ of $\Theta \in \mathcal{R}^p$ given $\mathbf{x}^n$. The posterior is formed from the prior $\pi(\cdot)$ on $\theta$ varying over the interior of $\mathcal{B}$ and the likelihood function is denoted $f_{\mathbf{X}|\Theta}(\cdot|\theta)$.

One benefit of using (1) is that it is defined more generally than risk based estimators. This is so because the distribution of the loss always has a median but need not always have a mean. (Consider the location Cauchy family for instance.) The insensitivity of the median to the tail behavior will therefore make our method applicable in heavy tailed distributions. Indeed, it will be seen in Theorem 1 below that the main moment-like conditions are expected local suprema in Lemma 1 and are only required for the asymptotic efficiency of the MLE.

In regression problems, $\mathbf{x}$ is regarded as an explanatory variable for the outcomes

$$y_i = h(\mathbf{x}_i, \beta) + u_i, \quad i = 1, \ldots, n, \tag{2}$$

where $y_i, \mathbf{x}_i$ and $u_i$ are the realizations of random elements $Y_i \in \mathcal{R}, \mathbf{X}_i \in \mathcal{R}^d$ and $U_i \in \mathcal{R}$, respectively, and $h$ is a known function in a class of functions $\mathcal{H}$. It is conventional to index regression functions by $\beta$ rather than $\theta$ and we follow this convention here . As above, we suppose that the true value $\beta_0$ of $\beta$ is an element of $\mathcal{B}$, an open subset of $\mathcal{R}^p$, and that $\beta \in \mathcal{B}$ is a random vector from a prior density $\pi$. Furthermore, we suppose the $u_i$'s are independently sampled from the distribution $\mathcal{P}$ on $\mathcal{R}$. Then, to find the posterior *medloss* estimator (1) for $\beta$, we have to derive the posterior density of $\beta$ given $y$ and $\mathbf{x}$ and find the action minimizing its median. The main contribution of this paper is to establish the $\sqrt{n}$-consistency, asymptotic normality, and efficiency of the posterior *medloss* estimator.

In the Frequentist context, one of the most common methods to estimate the regression coefficients $\beta_0$ is the least squares (LS) approach, which minimizes the sum of squares of the residuals. It is well known that the LS estimator is $\sqrt{n}$-consistent and asymptotically normal. However, it is highly sensitive to outliers or other influential observations.

To overcome the excessive sensitivity of the LSE, there are numerous alternative robust approaches. One of them is the least median of squares (LMS) estimator first introduced by Hampel [11](page 380) and then developed by Rousseeuw [20]. Like the LS estimator, the LMS estimator minimizes the median of squares of the residuals, i.e.

$$\hat{\beta}_n^{LMS} = \arg\min_{\beta} \operatorname*{median}_{i=1\le i\le n} [y_i - h(\mathbf{x}_i, \beta)]^2. \tag{3}$$

Because it is based on the median, the LMS estimator has 50% breakdown point. That is, 50 % is the smallest portion of the data that must be contaminated to force the LMS estimator to move an arbitrarily large amount. Asymptotically, Rousseeuw [20] provides a heuristic proof that the LMS estimator has a $\sqrt[3]{n}$ rate of convergence in linear models by

using arguments similar to those Andrews et al. [1] for the shorth estimator. A rigorous proof is given by Kim and Pollard [14] . In nonlinear regression models, Stromberg [24] gives conditions under which the LMS estimator is consistent.

As a compromise of sorts between the LS and LMS estimators, the least trimmed squares (LTS) estimator (see Rousseeuw and Leroy [21]) is sometimes proposed. The LTS improves the the $\sqrt[3]{n}$ rate of convergence of the LMS estimator to a $\sqrt{n}$ rate but the LTS can be less efficient than the LSE. The one-sided LTS estimator is defined by

$$\hat{\beta}_{n,1}^{(LTS,\tau)} = \arg\min_{\beta} \sum_{i=1}^{\tau} r_{[i]}^2(\beta), \tag{4}$$

where $r_{[i]}^2(\beta)$ represents the $i^{th}$ order statistics of squared residuals $r_i^2(\beta) = \{y_i - h(\mathbf{x}_i, \beta)\}^2$, and the trimming constant $\tau$ satisfies $\frac{n}{2} < \tau \leq n$. Its consistency and asymptotic normality for nonlinear regression can be found in Čížek [3, 4]. In Sec. 3, we define the analogous two-sided LTS estimator in (4) and argue it is more reasonable than the one-sided version.

Having now considered fully five estimators – posterior *medloss*, LS, LMS, one-sided LTS and two-sided LTS – it is worthwhile to see what they are in a simple example. Consider a flat prior on a unidimensional $\beta$ in a regression through the origin. That is, write

$$y_i = x_i\beta + u_i, \quad i = 1, \ldots, n, \tag{5}$$

where the outcomes $u_i$ are IID $N(0, \sigma^2)$. Then the posterior distribution of $\beta$ given $\{(y_i, x_i) : i = 1, \ldots, n\}$ is also normal with mean $s_{xy}/s_{xx}$, where $s_{xy} = \sum_{i=1}^n x_i y_i$ and $s_{xx} = \sum_{i=1}^n x_i^2$. Then, under squared error loss, the posterior *medloss* estimator is the posterior median; this follows from [29] because the posterior is normal and hence symmetric. That is, the posterior *medloss* estimator is the posterior mean given by $s_{xy}/s_{xx}$ which is the same as the usual LS estimator. In this example, the LMS estimator is

$$\hat{\theta}_n^{LMS} = \arg\min_{\theta} \underset{1 \leq i \leq n}{\text{median}} [y_i - x_i\beta]^2,$$

clearly different from either the LS or *medloss* estimators because medians are rarely numerically equal to means. (This is separate from the fact that the LS and *medloss* estimators are $\sqrt{n}$ convergent whereas the LMS estimator is only $\sqrt[3]{n}$ convergent.)

The one-sided LTS estimator (4) reduces to the usual LS estimator when $\tau = n$ in which case it coincides with the posterior *medloss* estimator. Otherwise, it is different from both the *medloss* and LS and in none of those cases is it the same as the LMS estimator (apart from sets of measure zero). Similarly, the two-sided LTS estimator reduces to the usual LS or posterior *medloss* estimator in the absence of trimming. However, when the trimming is nontrivial, the two-sided LTS numerically differs from all the foregoing estimators (off a set of measure zero) because the trimming is two sided and a mean is taken of the remaining terms. We comment that trimming $n-1$ data points in the one-sided LTS reduces to finding a single point that fits the model perfectly while trimming $\lfloor n/2 \rfloor$ on each side in the two-sided LTS reduces to finding a single point that represents the fit of the model to the whole

data set. (This is exact when $n$ is odd but approximate when $n$ is even.) Between the two extremes of zero trimming and full trimming, the LTS estimators have $\sqrt{n}$ convergence. However, in the limit of full trimming the two-sided LTS reduces to the LMS (with rate $\sqrt[3]{n}$ rate) and the behavior of the one-sided LTS is unclear. Note also that the LTS is $\sqrt{n}$ for any fixed trimming proportion $\rho$ in $[0, 1/2)$ on both sides. The $\sqrt{n}$ rate holds even if $\rho$ is allowed to approach $1/2$ slowly, but at $\rho = 1/2$ the asymptotic rate drops suddenly to $\sqrt[3]{n}$.

If a $N(\mu, \sigma^2)$ prior is used on $\beta$ instead of a flat prior, the above statements remain the same apart from the fact that the posterior *medloss* estimator will be a combination of the prior mean and the sample mean and so will not coincide with the LSE, except in an asymptotic sense.

The rest of this paper is organized as follows. In Section 2, we present our main result giving the asymptotic behavior of the posterior *medloss* estimator for parameter estimation in the absence of covariates. In Section 3, we state results giving the asymptotic behavior for the LMS and LTS estimators. These results are given for the more general case of non-linear models, however, they reduce to the parametric case and so can be compared with our main result in Section 2. Section 4 discusses the comparison of the posterior *medloss* estimator to the LMS and LTS estimators more generally.

## 2. Main results

We establish the asymptotic behavior of the posterior *medloss* estimator $\delta_n$ for finite dimensions in four steps. First, we use the asymptotic normality of the maximum likelihood estimator (MLE) $\hat{\theta}_n$ to identify the limiting distribution. Second, the convergence of posterior density to the normal in total variation is used to show the convergence of their spatial medians. Third, we prove that $\delta_n$ can be approximated by $\hat{\theta}_n$ up to an error of $o_p(n^{-1/2})$. Finally, Slutsky's theorem gives the result we want for $\delta_n$.

Since there are numerous results for the asymptotic normality of the finite-dimensional MLE, it is enough here to quote them without proof. For instance, the following lemma from Schervish [22] gives conditions under which the MLE is asymptotically multivariate normal and efficient in general parametric families.

**Lemma 1.** *Let $\Omega$ be a subset of $\mathcal{R}^p$, and let $\{\boldsymbol{X}_i \in \mathcal{R}^d : i = 1, 2, \ldots\}$ be conditionally IID given $\Theta = \theta \in \mathcal{R}^p$ each with density $f_{\boldsymbol{X}_1|\Theta}(\cdot|\theta)$. Let $\hat{\theta}_n$ be the MLE and assume that it converges to $\theta$ in $P_\theta$ for all $\theta$. Assume that $f_{\boldsymbol{X}_1|\Theta}(\cdot|\theta)$ has continuous second partial derivatives with respect to $\theta$ and that differentiation can be done under the integral sign. Suppose that there exists $M_{r_1}(\boldsymbol{x}, \theta)$ such that, for each interior point $\theta_0$ of $\Omega$ and each $k$, $j$, we have*

$$\sup_{\|\theta-\theta_0\|\le r_1} \left| \frac{\partial^2}{\partial\theta_k\partial\theta_j} \log f_{\boldsymbol{X}_1|\Theta}(\boldsymbol{x}|\theta_0) - \frac{\partial^2}{\partial\theta_k\partial\theta_j} \log f_{\boldsymbol{X}_1|\Theta}(\boldsymbol{x}|\theta) \right| \le M_{r_1}(\boldsymbol{x}, \theta_0),$$

*with $\lim_{r_1\to 0} E_{\theta_0} M_{r_1}(\boldsymbol{X}, \theta_0) = 0$ and that the Fisher information matrix $\mathcal{I}_{\boldsymbol{X}_1}(\theta_0)$ is finite and nonsingular. Then, under $P_{\theta_0}$,*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} N(0, \mathcal{I}^{-1}_{\boldsymbol{X}_1}(\theta_0)). \tag{6}$$

Next we turn to the convergence of posterior density to the normal for finite-dimensional parameters. As noted in Lehmann [15], it is not enough to impose the conditions in Lemma 1 on $\log f_{\mathbf{X}_1|\Theta}(\mathbf{x}|\theta)$ in the neighborhood of $\theta_0$ as is typically the case in asymptotic results. The behavior of $\log f_{\mathbf{X}_1|\Theta}(\mathbf{x}|\theta)$ must be controlled even when $\theta$ is far from $\theta_0$. This is so because the normalizing constant in the posterior density is the marginal for the data that is an integral over the whole parameter space. Again, there are numerous results, see Borwanker et al. [2], Lehmann [15] Ghosh and Ramamoorthi [7] and Prakasa Rao [19]. Here we use the following from Schervish [22]. Note that the limiting distribution has variance given by the Fisher information so the posterior is efficient.

**Lemma 2.** *In addition of the assumptions in Lemma 1, suppose that for any $r_3 > 0$, there exists an $\epsilon > 0$ such that*

$$P_{\theta_0}\Big\{ \sup_{\|\theta-\theta_0\|>r_3} \frac{1}{n}(L_n(\theta) - L_n(\theta_0)) \le -\epsilon \Big\} \to 1,$$

*where $L_n(\theta) = \sum_{i=1}^n \log f_{\boldsymbol{X}_1|\Theta}(\boldsymbol{x}_i|\theta)$. Assume also that the prior has a density $\pi(\theta)$ with respect to Lebesgue measure and that $\pi(\cdot)$ is continuous and positive at $\theta_0$. Then, we have that, as $n \to \infty$,*

$$\int_{\mathcal{R}^p} \Big| \pi^*(\boldsymbol{t}|\boldsymbol{x}^n) - (2\pi)^{-p/2}|\mathcal{I}_{\boldsymbol{X}_1}(\theta_0)|^{1/2} exp\{-\boldsymbol{t}^T\mathcal{I}_{\boldsymbol{X}_1}(\theta_0)\boldsymbol{t}/2\} \Big| \, d\boldsymbol{t} \xrightarrow{P_{\theta_0}} 0, \tag{7}$$

*where $\boldsymbol{x}^n = \{\boldsymbol{x}_i : i = 1, \ldots, n\}$ and $\pi^*(\cdot|\boldsymbol{x}^n)$ is the posterior density of $\boldsymbol{T} = \sqrt{n}(\Theta - \hat{\theta}_n(\boldsymbol{x}^n))$.*

To state Lemma 3, we make the following definitions. For any distribution function $F(\cdot)$, let

$$Q(t) \stackrel{\text{def}}{=} F^{-1}(t) = \inf\{x : F(x) \ge t\}, \text{ for } 0 < t < 1$$

be its quantile function. Now denote by $Q_n$ the quantile function associated with the distribution function $F_n$ for each $n \ge 0$. If $Q_n(t) \to Q_0(t)$ at each continuity point $t$ of $Q_0(t)$ in (0, 1), $Q_n$ is said to converge in quantile to $Q_0$, denoted by $Q_n \xrightarrow{\mathcal{Q}} Q_0$. We have the following from Proposition 3.1 in Chapter 7 in Shorack [23].

**Lemma 3.** *Convergence in distribution is equivalent to convergence in quantile, i.e.,*

$$F_n \xrightarrow{\mathcal{L}} F_0 \iff Q_n \xrightarrow{\mathcal{Q}} Q_0.$$

Now we can establish our asymptotic results for the posterior *medloss* estimator.

**Theorem 1.** *Suppose that the assumptions of Lemma 2 hold and that the convergence of the MLE, $\hat{\theta}_n$, to $\theta_0$ is a.s., i.e. $\hat{\theta}_n \to \theta_0$ a.s. $P_0 = P_{\theta_0}$. Further, let $\delta_n = \delta_n(\boldsymbol{x}^n)$ be the posterior medloss estimator of $\theta \in \mathcal{R}^p$ for all realizations $\{\boldsymbol{x}_i : i = 1, \ldots, n\}$ of $\{\boldsymbol{X}_i \in \mathcal{R}^d : i = 1, \ldots, n\}$ and all $n$ with respect to a nonnegative loss function $\mathcal{L}(\theta, \boldsymbol{a})$ satisfying the following conditions:*

*(i)* $\mathcal{L}(\theta, \boldsymbol{a}) = l(\theta - \boldsymbol{a}) \geq 0$,

*(ii)* $l(\boldsymbol{t}_1) \geq l(\boldsymbol{t}_2)$ *if* $\|\boldsymbol{t}_1\| \geq \|\boldsymbol{t}_2\|$.

*Moreover, suppose that there exist a non-negative sequence* $\{a_n\}$ *and continuous function* $K : \mathcal{R}^p \to \mathcal{R}$ *such that*

*(iii) For any real-valued vector* $\boldsymbol{c}_n$ *depending on* $n$,

$$\lim_{n\to\infty} \left| \underset{\boldsymbol{T}|\boldsymbol{X}^n}{med}[a_n l((\boldsymbol{T} + \boldsymbol{c}_n)/n^{1/2})] - \underset{\boldsymbol{T}|\boldsymbol{X}^n}{med}[K(\boldsymbol{T} + \boldsymbol{c}_n)] \right| = 0,$$

*where* $\boldsymbol{T} = \sqrt{n}(\Theta - \hat{\theta}_n)$,

*If* $\boldsymbol{Z}$ *has the normal distribution* $N(\boldsymbol{0}, \mathcal{I}^{-1}_{\boldsymbol{X}_1}(\theta_0))$, *i.e. the limiting distribution of the posterior density in Lemma 2, suppose that*

*(iv) 1/2 is a continuous point of the distribution of* $K(\boldsymbol{Z})$, *and*

*(v)* $med_{\boldsymbol{Z}} K(\boldsymbol{Z} + \boldsymbol{m})$ *has a unique minimum at* $\boldsymbol{m=0}$, *where* $med_{\boldsymbol{Z}}$ *is the median with respect to* $\boldsymbol{Z}$.

*Then we have*

$$\delta_n \to \theta_0 \ a.s. \ P_0 \ and \ n^{1/2}(\theta_0 - \delta_n) \xrightarrow{\mathcal{L}} N(0, \mathcal{I}^{-1}_{\boldsymbol{X}_1}(\theta_0)).$$

Before giving the proof, we note that it is the asymptotic normality of the MLE and posterior that is central to the proof of Theorem 1. That is, the assumptions in Lemma 1 and Lemma 2 only constitute readily verifiable conditions for asymptotically normal MLE's and posteriors. We do not use the formal assumptions again in the proof of the asymptotics of the posterior *medloss* estimator below.

**Remark 1:** Note that conditions (i), (ii) and (iii) are true for $L^1$ loss with $a_n = n^{1/2}$ and $K(\mathbf{t}) = \|\mathbf{t}\|$. Furthermore, in this case, $\mathbf{Z}$ is multivariate normal with median $\mathbf{0}$, so conditions (iv) and (v) can be satisfied, verifying the conditions are not vacuous.

**Remark 2:** The role of asymptotic normality of the MLE is so essential that virtually any time we have sufficient conditions for the MLE to be asymptotically normal, we have a corresponding result for the posterior *medloss* estimator. This means that the substance of Theorem 1 holds, in particular, for many regression problems in linear and non-linear cases. For instance, in generalized linear models (GLM), if the first and second conditional moments of the response variable (given the explanatory variables) exist then, as in Fahrmeir and Kaufmann [6], we get asymptotic normality of the MLE and hence, by our Theorem 1, an asymptotic normality result for posterior *medloss* estimators in generalized linear models. More generally, we can use quasi-likelihood to obtain a version of Theorem 1 for GLM's. Under various regularity conditions, the quasi-maximum likelihood estimator (QMLE) has a $\sqrt{n}$ rate of convergence and is asymptotically normal. Then, as used in Step 4 in the proof of Theorem 1 below, $\delta_n$ is asymptotically equivalent to the QMLE and has a $\sqrt{n}$ rate of convergence and is asymptotically normal in a regression settings. This contrasts sharply with the LMS which does not have these properties, as will be seen in Theorem 2 below.

Similarly, when asymptotic normality holds for nonlinear models, see Gallant [8], we obtain, via Theorem 1, an asymptotic normality result for posterior *medloss* estimators in nonlinear models. Finally, the results of Koenker and Bassett [17], Bassett and Koenker [9] can be used to obtain asymptotic normality of the posterior *medloss* estimator, via Theorem 1, for quantile regression models.

*Proof.* We prove Theorem 1 by way of contradiction in four steps. The first step obtains an inequality we will need for the second step which shows $n^{1/2}(\hat{\theta}_n - \delta_n)$ is finite a.s. The third step shows $n^{1/2}(\hat{\theta}_n - \delta_n)$ goes to $\mathbf{0}$ a.s. $P_0$. Then we complete the proof by using Slutsky's theorem and the asymptotic normality of $\hat{\theta}_n$. Denote the posterior *medloss* with respect to $\mathcal{L}(\theta, \mathbf{a})$ by $M_n(\mathbf{a}) = \underset{\pi(\Theta|\mathbf{X}^n)}{\text{med}} \mathcal{L}(\theta, \mathbf{a})$.

1. First, $\limsup_n a_n M_n(\delta_n) \leq \limsup_n a_n M_n(\hat{\theta}_n) = \limsup_n \underset{\mathbf{T}|\mathbf{X}^n}{\text{med}}[a_n l(\mathbf{T}/n^{1/2})]$. Moreover,

$$\begin{aligned}
&\Big|\underset{\mathbf{T}|\mathbf{X}^n}{\text{med}}[a_n l(\mathbf{T}/n^{1/2})] - \text{med}_{\mathbf{Z}}[K(\mathbf{Z})]\Big| \\
&\leq \Big|\underset{\mathbf{T}|\mathbf{X}^n}{\text{med}}[a_n l(\mathbf{T}/n^{1/2})] - \underset{\mathbf{T}|\mathbf{X}^n}{\text{med}}[K(\mathbf{T})]\Big| + \Big|\underset{\mathbf{T}|\mathbf{X}^n}{\text{med}}[K(\mathbf{T})] - \text{med}_{\mathbf{Z}}[K(\mathbf{Z})]\Big|. \qquad (8)
\end{aligned}$$

   The first term in (8) goes to zero by condition (iii) of the loss function. By (7), we can show that $\mathbf{T}$ converges in distribution to $\mathbf{Z}$, which implies that $K(\mathbf{T})$ also converges to $K(\mathbf{Z})$ in distribution by the Continuous Mapping Theorem. So, using Lemma 3 with condition (iv), we have $\underset{\mathbf{T}|\mathbf{X}^n}{\text{med}} K(\mathbf{T}) \to \text{med}_{\mathbf{Z}} K(\mathbf{Z})$ and therefore the second term in (8) converges to zero. Thus,

$$\limsup_n a_n M_n(\delta_n) \leq \limsup_n a_n M_n(\hat{\theta}_n) \leq \text{med}_{\mathbf{Z}} K(\mathbf{Z}). \qquad (9)$$

2. Let $\mathbf{W}_n = n^{1/2}(\hat{\theta}_n - \delta_n)$. Now we show $\limsup_n |\mathbf{W}_n| < \infty$ a.s.
   First, suppose that the statement $\limsup_n |\mathbf{W}_n| < \infty$ a.s. is false. Then, for every positive vector $\mathbf{M}$, there exists a set $A_M$ with $P_\theta(A_M) > 0$ such that $\mathbf{W}_n(\mathbf{x}) > \mathbf{M}$ or $\mathbf{W}_n(\mathbf{x}) < -\mathbf{M}$ *i.o.* for $\mathbf{x} \in A_M$. Without loss of generality, we can assume that $\mathbf{W}_n(\mathbf{x}) > M$ *i.o.* Then, for the subsequence $\{n_i\}$ where the inequality holds, we have

$$\begin{aligned}
a_{n_i} M_{n_i}(\delta_{n_i}) = a_{n_i} \underset{\pi(\Theta|\mathbf{X}^{n_i})}{\text{med}} l(\Theta - \delta_{n_i}) &= \underset{\mathbf{T}|\mathbf{X}^{n_i}}{\text{med}} \Big[a_{n_i} l\Big(\frac{\mathbf{T}+\mathbf{W}_{n_i}}{n_i^{1/2}}\Big)\Big] \\
&\geq \underset{\mathbf{T}|\mathbf{X}^{n_i}}{\text{med}} \Big[a_{n_i} l\Big(\frac{\mathbf{T}+\mathbf{W}_{n_i}}{n_i^{1/2}}\Big) I_{\{\mathbf{T}\,+\,\mathbf{M}\geq\mathbf{0}\}}\Big] \\
&\geq \underset{\mathbf{T}|\mathbf{X}^{n_i}}{\text{med}} \Big[a_{n_i} l\Big(\frac{\mathbf{T}+\mathbf{M}}{n_i^{1/2}}\Big) I_{\{\mathbf{T}\,+\,\mathbf{M}\geq\mathbf{0}\}}\Big] \\
&\to \text{med}_{\mathbf{Z}}\Big[K(\mathbf{Z}+\mathbf{M}) I_{\{\mathbf{Z}\,+\,\mathbf{M}\geq\mathbf{0}\}}\Big]. \qquad (10)
\end{aligned}$$

The first inequality holds because $l(\mathbf{X})I_{\{\mathbf{X}\in\mathcal{B}\}} \le l(\mathbf{X})$ for any non-negative random vector $\mathbf{X}$ and an indicator function $I$ with any set $\mathcal{B}$. The second inequality holds by the assumption that $\mathbf{W}_n(\mathbf{x}) > M$ *i.o.* and by condition (ii) with $\mathbf{T} + \mathbf{W}_{n_i} > \mathbf{T} + \mathbf{M} \ge 0$. Then we use arguments similar to those for the convergence of $\underset{\mathbf{T}|\mathbf{X}^n}{\text{med}}[a_n l(\mathbf{T}/n^{1/2})]$ to $\text{med}_{\mathbf{Z}}[K(\mathbf{Z})]$ in Step 1 to get (10).
According to Tomkins' median version of the Lebesgue dominated convergence theorem in [25] and condition (v) in our Theorem 1, we have

$$\begin{aligned}\lim_{\mathbf{M}\to+\infty} \text{med}_{\mathbf{Z}}\Big[K(\mathbf{Z}+\mathbf{M})I_{\{\mathbf{Z}\,+\,\mathbf{M}\ge\mathbf{0}\}}\Big] &= \text{med}_{\mathbf{Z}} \lim_{\mathbf{M}\to+\infty}\Big[K(\mathbf{Z}+\mathbf{M})I_{\{\mathbf{Z}\,+\,\mathbf{M}\ge\mathbf{0}\}}\Big] \\ &= K(+\infty) > \text{med}_{\mathbf{Z}} K(\mathbf{Z}).\end{aligned}$$

Therefore, for large $\mathbf{M}$, on a set of positive probability,

$$\limsup_n a_n M_n(\delta_n) > \text{med}_{\mathbf{Z}} K(\mathbf{Z}) \ge \limsup_n a_n M_n(\hat{\theta}_n),$$

which contradicts the definition of $\delta_n$.
Thus, $\limsup_n \mathbf{W}_n < \infty$ a.s. $P_0$. Similarly, we have $\liminf_n \mathbf{W}_n > -\infty$ a.s. $P_0$.

3. Next for any arbitrary $\epsilon > 0$, we denote by $B_M$ the set such that for $\mathbf{x} \in B_M$, $-\mathbf{M} \le \mathbf{W}_n \le \mathbf{M}$ for every $n$ and $P_\theta(B_M) > 1 - \epsilon$. For a fixed $\mathbf{x} \in B_M$, $\mathbf{W}_n(\mathbf{x})$ is a bounded sequence, so it has a limit point $\mathbf{m}$. Assume that $\mathbf{m} \ne \mathbf{0}$. Then, for the subsequence $\{n_i\}$ where $\mathbf{W}_{n_i}(\mathbf{x}) \to \mathbf{m}$, we have

$$\begin{aligned}&\liminf_{n_i} a_{n_i} M_{n_i}(\delta_{n_i}) \\ =\ &\liminf_{n_i} \underset{\mathbf{T}|\mathbf{X}^{n_i}}{\text{med}}\Big[a_{n_i} l\Big(\frac{\mathbf{T}+\mathbf{W}_{n_i}}{n_i^{1/2}}\Big)\Big] \\ \ge\ &\text{med}_{\mathbf{Z}} K(\mathbf{Z}+\mathbf{m}) - \limsup_{n_i}\left|\underset{\mathbf{T}|\mathbf{X}^{n_i}}{\text{med}}\Big[a_{n_i} l\Big(\frac{\mathbf{T}+\mathbf{W}_{n_i}}{n_i^{1/2}}\Big)\Big] - \text{med}_{\mathbf{Z}} K(\mathbf{Z}+\mathbf{m})\right|.\end{aligned}$$

Note that

$$\begin{aligned}&\left|\underset{\mathbf{T}|\mathbf{X}^{n_i}}{\text{med}}\Big[a_{n_i} l\Big(\frac{\mathbf{T}+\mathbf{W}_{n_i}}{n_i^{1/2}}\Big)\Big] - \text{med}_{\mathbf{Z}} K(\mathbf{Z}+\mathbf{m})\right| \\ \le\ &\left|\underset{\mathbf{T}|\mathbf{X}^{n_i}}{\text{med}}\Big[a_{n_i} l\Big(\frac{\mathbf{T}+\mathbf{W}_{n_i}}{n_i^{1/2}}\Big)\Big] - \underset{\mathbf{T}|\mathbf{X}^{n_i}}{\text{med}} K(\mathbf{T}+\mathbf{W}_{n_i})\right| \\ &+\ \left|\underset{\mathbf{T}|\mathbf{X}^{n_i}}{\text{med}} K(\mathbf{T}+\mathbf{W}_{n_i}) - \text{med}_{\mathbf{Z}} K(\mathbf{Z}+\mathbf{m})\right|.\end{aligned}$$

Then, by condition (iii) and arguments similar to those for the convergence of $\underset{\mathbf{T}|\mathbf{X}^n}{\text{med}}[K(\mathbf{T})]$ to $\text{med}_{\mathbf{Z}}[K(\mathbf{Z})]$ in Step 1, we have

$$\left|\underset{\mathbf{T}|\mathbf{X}^{n_i}}{\text{med}}\Big[a_{n_i} l\Big(\frac{\mathbf{T}+\mathbf{W}_{n_i}}{n_i^{1/2}}\Big)\Big] - \text{med}_{\mathbf{Z}} K(\mathbf{Z}+\mathbf{m})\right| \le \epsilon.$$

Thus, by condition (v), we have

$$\begin{aligned}\liminf_{n_i} a_{n_i} M_{n_i}(\delta_{n_i}) &\geq \text{med}_{\mathbf{Z}} K(\mathbf{Z}+\mathbf{m}) - \epsilon \\ &> \text{med}_{\mathbf{Z}} K(\mathbf{Z}) - \epsilon.\end{aligned}$$

Since $\epsilon$ is arbitrary, we get $\liminf_{n_i} a_{n_i} M_{n_i}(\delta_{n_i}) > \text{med}_{\mathbf{Z}} K(\mathbf{Z})$, which is impossible by (9). Therefore, $\mathbf{m}=\mathbf{0}$ and $n^{1/2}(\delta_n - \hat{\theta}_n) \to \mathbf{0}$ a.s. $P_0$.

4. Finally, the proof is completed by observing

$$n^{1/2}(\delta_n - \theta_0) = n^{1/2}(\delta_n - \hat{\theta}_n) + n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} N(0, \mathcal{I}_{\mathbf{X}_1}^{-1}(\theta_0)).$$

□

To summarize the key conceptual point of our result we state the following.

**Corollary 1.** *Suppose asymptotic normality of the MLE and of the posterior density hold and consider any continuous posterior density of $\Theta$ given $\boldsymbol{X}^n = \boldsymbol{x}^n$ under $L^1$ loss, i.e. $\mathcal{L}(\theta, \boldsymbol{a}) = \|\theta - \boldsymbol{a}\|$. If the median of the $L^1$ loss is unique, then we have*

$$\delta_n \to \theta_0 \quad a.s.\ P_0 \quad \textit{and} \quad n^{1/2}(\theta_0 - \delta_n) \xrightarrow{\mathcal{L}} N(\boldsymbol{0}, \mathcal{I}_{\boldsymbol{X}_1}^{-1}(\theta_0)). \tag{11}$$

*More generally, (11) holds when the $L^1$ loss is replaced by any strictly increasing function of $\|\Theta - \boldsymbol{d}(\boldsymbol{x}^n)\|$, provided the median of the function is unique.*

We remark that the results in Theorem 1 and Corollary 1 can be easily extended to the case of Markov chain settings by using arguments similar to those of Borwanker et al. [2] for the asymptotic behavior of Bayesian estimators.

## 3. Asymptotics for two related estimators

In Section 1, the posterior *medloss* was contrasted with four other estimators in the context of the normal example. In this section, we focus on two of these, the LMS and the two-sided LTS. Although Theorem 1 is shown only for the purely parametric case, i.e., no covariates, the discussion after Theorem 1 shows that it holds for a variety of model classes. Consequently, for generality, we state results for the LMS and two-sided LTS for the context of non-linear models.

For the LMS, we recall that Kim and Pollard [14] established a cube-root rate of convergence to a limiting Gaussian process for linear regression models. Our first result extends this to non-linear regression models of the form (2). To state our result, let $\mathcal{H}$ be a finite dimensional vector space of real-valued regression functions of the form $h = h(\mathbf{x}, \beta)$ for $\beta \in \mathcal{B}$. Let $R > 0$ and define the envelope $H_R(\cdot)$ to be the supremum of $|h(\cdot, \theta)|$ over $\mathcal{H}_R = \{h(\cdot, \beta) | \|\beta - \beta_0\| \leq R\}$, i.e., $H_R(\mathbf{x}) = \sup_{h \in \mathcal{H}_R} |h(\mathbf{x}, \beta)|$. Then, we have the following.

**Theorem 2.** *Suppose*

1. $\boldsymbol{X}_i$ *and* $u_i$ *are independent for* $i = 1, \ldots, n$.
2. $h(\boldsymbol{x}_i, \beta)$ *is continuous in* $\beta \in \mathcal{B}$ *and is differentiable in* $\beta$ *on a neighborhood of* $\beta_0$.
3. $Q_h = E_{\boldsymbol{X}}[h'(\boldsymbol{X}, \beta_0)h'(\boldsymbol{X}, \beta_0)^T]$ *is positive definite.*
4. $u_i$ *comes from a bounded, symmetric density* $\gamma$ *that decreases away from its mode at zero, and has a strictly negative derivative at* $r_0$*, the unique median of* $|u|$.
5. *For any* $h \in \mathcal{H}$*,* $h$ *satisfies the Lipschitz condition, i.e.*

$$|h(\boldsymbol{X}, \beta_1) - h(\boldsymbol{X}, \beta_2)| \leq L_{\boldsymbol{X}} \|\beta_1 - \beta_2\|, \text{ where } L_{\boldsymbol{X}} > 0 \text{ depends on } \boldsymbol{X},$$

   *and* $E_{\boldsymbol{X}}(L_{\boldsymbol{X}}) < \infty$.
6. $E_{\boldsymbol{X}}\|h'(\boldsymbol{X}, \xi)\| < \infty$ *for* $\xi \in U(\beta_0, R)$*, where* $U(a, b)$ *is an open ball at center* $a$ *with radius* $b$*, and* $H_R$ *is well-defined for* $R$.
7. $E_{\boldsymbol{X}}|h'(\boldsymbol{X}, \beta_0)^T \boldsymbol{w}| \neq 0$ *for any* $\boldsymbol{w} \neq \boldsymbol{0}$.

*Then we have that* $n^{1/3}(\hat{\beta}_n^{LMS} - \beta_0)$ *converges in distribution to the* $\arg\max_\theta$ *of the Gaussian process*

$$Z(\theta) = \gamma'(r_0)\theta^T Q_h \theta + W(\theta),$$

*as* $n \to \infty$*, where* $\theta = \beta - \beta_0$ *and the Gaussian process* $W$ *has zero mean and continuous sample paths.*

*Proof.* This theorem is the case $q = 1/2$ in Yu and Clarke [30]. □

This result shows that in nonlinear regression models, the LMS estimator has a slow rate of convergence. Note that since the LMS is based on a median, it can also be viewed as a trimmed mean estimator with a trimming proportion approaching 50% on both sides. Clearly, the more the trimming, the fewer data points that contribute directly to the estimator. Consequently, the rate of convergence slows from root-$n$ to cube root $n$. To reinforce this intuition, we observe that when the trimming proportion is strictly less than 50% on each side the $n^{1/2}$ rate of convergence and asymptotic normality are recovered.

To define the second estimator, the two-sided LTS, it is worth recalling the one-sided LTS from (4). The asymptotic consistency, normality, and variance of the one-sided LTS were established in [3, 4]. By contrast, the two-sided LTS estimator is

$$\hat{\beta}_{n,2}^{(LTS,\tau)} = \arg\min_{\beta} \sum_{n-\tau+1}^{\tau} r_{[i]}^2(\beta), \tag{12}$$

where $r_{[i]}^2(\beta)$ represents the $i^{th}$ order statistics of squared residuals $r_i^2(\beta) = \{y_i - h(x_i, \beta)\}^2$, and the trimming constant $\tau$ satisfies $\frac{n}{2} < \tau \leq n$. The one-sided LTS trims off only the large values of the $r_i$'s whereas the two-sided LTS trims off the small and large $r_i$'s equally.

Parallel to [3, 4], we establish the asymptotic consistency, normality, and variance of the two-sided LTS. Again, consider the nonlinear regression model (2) and assume let $\{X_t\}_{t\in\mathcal{N}}$ be sequence of $\beta$-mixing random variables, i.e., the variables satisfy

$$\beta_m = \sup_{t\in\mathcal{N}} E\Big\{ \sup_{B\in\sigma_{t+m}^f} |P(B|\sigma_t^p) - P(B)| \Big\} \to 0, \tag{13}$$

as $m \to \infty$, where $\sigma_t^p = \sigma(X_t, X_{t-1}, \ldots)$ and $\sigma_t^f = \sigma(X_t, X_{t+1}, \ldots)$ are $\sigma$-algebras; this is a condition roughly ensuring that when $B$ is defined by (future) variables that are sufficiently separated from the (past) variables in the $\sigma$-algebra $\sigma_t^p$ that the two are independent. In particular, if the $X_t$'s are independent then $\beta_m = 0$ for all $m \geq 1$.

To state our result, denote the distribution functions of $U_i$ and $U_i^2$ by $F$ and $G$, the corresponding pdf's by $f$ and $g$, and quantile functions by $F^{-1}$ and $G^{-1}$, respectively. Also, observe that the choice of the trimming constant $\tau$ in (3) may depend on the sample size $n$. So, we assume that a given sequence of trimming constants $\tau_n$ is given with the property that $\tau_n/n$ determines the fraction of sample included in (3) and that $\tau_n = [\lambda n]$, where $[z]$ represents the integer part of $z$, so that $\tau_n/n \to \lambda$ for some $1/2 < \lambda \leq 1$. Now we have the following for $\hat{\beta}_n^{(LTS,\tau_n)}$.

**Theorem 3.** *For $\beta$-mixing explanatory variables and further regularity conditions (see Yu and Clarke [31]), we have, when $\lambda \in (1/21)$, that*

1. $\hat{\beta}_{n,2}^{(LTS,\tau_n)} \xrightarrow{p} \beta_0$, *as* $n \to \infty$; *and*
2. $\sqrt{n}(\hat{\beta}_{n,2}^{(LTS,\tau_n)} - \beta_0) \xrightarrow{\mathcal{L}} N(0, V_{2\lambda})$,

*where* $V_{2\lambda} = (C_\lambda)^{-2}\sigma_{2\lambda}^2 Q_h^{-1}$, $Q_h = E_X[h'(X,\beta_0)h'(X,\beta_0)^T]$, $C_\lambda = (2\lambda - 1) + (\frac{q_\lambda + q_{1-\lambda}}{2})[H(\lambda) - H(1-\lambda)]$, $H(\lambda) = f(q_\lambda) + f(-q_\lambda)$, $q_\lambda = \sqrt{G^{-1}(\lambda)}$ *and* $\sigma_{2\lambda}^2 = EU_i^2 I_{[G^{-1}(1-\lambda), G^{-1}(\lambda)]}(U_i^2)$.

*Proof.* The proof follows [3] and [4] closely; see Yu and Clarke [31]. □

Theorem 3 shows that $\sqrt{n}$-convergence and asymptotic normality hold for the two-sided LTS estimator, but that it is inefficient. If the role of $\lambda$ is examined closely, it can be seen that the asymptotic variance increases as the amount of trimming increases. Moreover, comparing Theorem 3 with [3, 4] shows that from an asymptotic standpoint one- and- two-sided trimming are equivalent: Both one- and two-sided LTS estimators have the same rates and asymptotic variances.

Nevertheless, we argue that two-sided trimming makes more sense than one-sided trimming in many contexts. Consider the following four examples. First, suppose we correctly fit a model $Y = \mu + U$ where $U$ is a symmetric error. If we use quantile data from $U$ in which the data at the $n$-th stage consists of $n-1$ points representing the $q/n$ quantiles of $U$ for $q = 1, \ldots, n-1$ then one- and two-sided trimming are essentially equivalent: The largest residuals removed in one-sided trimming will be from both tails of $U$ in equal numbers while in two-sided trimming the largest residuals and the smallest residuals will be equally from the tails of $U$ and from the center of $U$. Also, both sides of 0 will be equally represented.

Next, suppose that $U$ is asymmetric, perhaps shaped like an exponential. The large residuals will be mostly from the side which has the heavier tail. One-sided trimming will remove the large residuals and lead to an estimate which underestimates $\mu$ if the heavier tail is to the right and which overestimates $\mu$ if the heavier tail is to the left. Two-sided trimming will have the same problem, but to a much lesser extent since only half the residuals trimmed will be from the heavier tail and they will be balanced somewhat by the removal of small residuals which will slightly over-represent the side with the lighter tail. Overall, we suggest

that when the model class is good, it is the error term that determines whether two-sided trimming is better than or equivalent to one-sided trimming.

Now consider the same two error terms, but suppose the model space is all linear functions of $\mathbf{x}$ while the true model is $Y = X_1^2 + U$ so that $X_2, \ldots, X_d$ are irrelevant. Suppose also that the values of $X_1$ are in three clusters, say 2/5 are around $X_1 = 1$, another 2/5 are around $X_1 = -1$ and 1/5 are around $X_1 = 0$. When $U$ is symmetric, removing the largest residuals (from the points near $X_1 = 0$) will reinforce the idea that the linear model is true. Removing some of the largest and some of the smallest residuals will still pull the fitted line down probably making it a better fit to future data than the line defined only from the clusters at $X_1 = 1$ and $X_1 = -1$.

When $U$ is asymmetric the situation is even worse. If the tail of $U$ is to the right so that the largest residuals come from points above the curve $Y = X_1^2$, then the largest residuals will come from points with $X_1$ near 0 and with $Y$ values below the curve $Y = X_1^2$. Removing them will again reinforce the idea that a straightline curve is correct. The reinforcement in this case will be stronger because the points with $X_1$ close to 0 that are above the curve $Y = X_1^2$ will genuinely look like they came from a straightline model. Again, this may happen with two-sided trimming too, but to a much smaller extent. However, as in the last case, two-sided trimming will probably yield a straightline which is a better fit to future data than the line from the one-sided trimming. That is, two-sided trimming is likely to give a better wrong model.

More generally, comparing one- and two-sided trimming depends on the symmetry of the error and the adequacy of the model class. The generic case seems to be the following. Fix a model class and suppose we use maximal trimming in the two-sided case, i.e., we let $\tau$ approach $n/2$ and permit $n$ to increase. Then, in the limit, we obtain an estimate for $\beta$ that depends on a small number of pairs $(y_i, \mathbf{x}_i)$ that give the median discrepancy between the model and the data. By contrast, if we use the analogous procedure on the one-sided LTS, i.e., we let $\tau$ approach 1, then we obtain an estimate of $\beta$ that again depends on a small number of points but now these points give the minimal discrepancy between the model and the data. Clearly, in a setting where there is high model uncertainty or high data variability due to the error i.e., the modeling can be easily misled by errant data, an estimator derived from the median discrepancy between the model and a small number of data points will be better than an estimator derived from the minimal discrepancy between the model and a small number of data points. The same sort of difference will hold for smaller amounts of trimming, but be correspondingly less. On the other hand, when the data is sparse, i.e., $n$ is small relative to $d$, the extent to which two-sided trimming gives more representative estimators than one-sided trimming does will tend to be larger. This occurs because there will be more variability among points that can be found to fit the model exceptionally well (one-sided trimming) than among points which give a representative fit of the data to the model. Otherwise put, one-sided trimming gives an estimator that may be a function of inliers – points which may fit the model well but are suspect or overly influential for other reasons – whereas two-sided trimming finds an estimator based on a typical fit.

## 4. A comparison of posterior medloss estimator, LMS, and LTS estimators

In the last section we argued that the two-sided LTS was better than the one sided LTS. Next, we argue that the posterior *medloss* estimator is a better choice than either the LMS or the two-sided LTS asymptotically and in more realistic settings.

First, Theorem 1 shows that the posterior *medloss* estimator is $\sqrt{n}$-convergent and its asymptotic variance is the inverse of the Fisher information matrix, i.e. the posterior *medloss* estimator is consistent, asymptotically normal with rate $\sqrt{n}$ and efficient. By contrast, for a fixed $\tau \in (n/2, n]$, Theorem 3 shows that the 2-sided LTS estimator is also $\sqrt{n}$-convergent, but will in general not be efficient unless $\tau = n$. That is, the posterior *medloss* estimator has a higher asymptotic relative efficiency than the two-sided (or one-sided) LTS. For the LMS estimator, Theorem 2 establishes a $\sqrt[3]{n}$ rate of convergence. Thus, in an asymptotic sense, the LMS estimator is worse than either the posterior *medloss* estimator or the 2-sided LMS estimator. Given these observations, there is no asymptotic reason to prefer either the LMS or the LTS estimators over the *medloss* estimator.

From a finite $n$ perspective, we note that the posterior *medloss* depends directly on all the data while the LMS and LTS only depend on a subset of the data. This means that when the data are good, in the minimal sense that none of them can be thrown out on the grounds that they were collected improperly, the *medloss* retains them even though the LMS and LTS do not. Indeed, all data points are treated symmetrically by the posterior (in the IID case) but the LMS and LTS by definition throw out some data. We expect this will often give the *medloss* estimator more desirable stability properties when the data are good even when there are overly influential observations.

For instance, when the data are noisy because of a large error term and complex because finding a good model is difficult, omitting data might be bad because it's unclear which data points are most reliable. So, better than omitting data would be allowing the seemingly better data to outweigh the seemingly worse data and this is what the *medloss* does. Indeed, the *medloss* does this in two ways: First, by retaining all the data and second, by using the prior as a sort of sanity check. That is, when data lead to a model that is poorly representative there is a good chance that the model will be in a region of low prior probability. This means that the posterior cannot assign it high probability. That is, the effect of the prior in the *medloss* estimator will tend to be to pull the poor model to a region of better models.

For instance, recall the normal example of Section 1 given by (5). The *medloss* estimator under squared error loss with a flat prior comes from the posterior which is located at $s_{xy}/s_{xx}$. Thus in the nicest case, the *medloss* and the LS or LTS effectively coincide when the data is well-behaved; this is not so in general. Note that the location has a breakdown point of zero because moving any one data point can move the location an arbitrary amount. Since the posterior is normal, it too has a breakdown point of zero. However, when a proper prior is used, a deviation in a single $x_1$ must be relatively larger to move the *medloss* estimator a fixed amount than to move $s_{xy}/s_{xx}$ even though the breakdown point is still zero. We suggest that a breakdown point of zero for the *medloss* estimator essentially only occurs when the posterior depends on statistics that have a breakdown point of zero; otherwise we expect a very high breakdown point. Note that the breakdown point of the LTS depends on the level

of trimming. Moreover, when the data are good and $s_{xy}/s_{xx}$ is overly influenced by a small number of data points, it may indicate that the model class is inadequate. Distinguishing between good but overly influential data and good but not overly influential data will be useful whereas throwing out the subset of data that leads to poor fit for a given model class can be misleading.

These differences become more pronounced outside the normal error setting with simple models, good data, and decent fit. For instance, suppose $n$ is small relative to $d$. Then, even when all the data is good, they often clump in dispersed regions with large empty regions between them. In this setting throwing out data means we are left with estimates that depend on a small number of incompletely representative points. The consequence of this is that an LTS estimator exacerbates data sparsity and nonrepresentativity while the posterior *medloss* estimator does the best it can with all the data, making it preferable.

Note that the argument here depends partially on using proper priors both on the parameters within a model and across the model space. The field of prior selection for parameters is well developed and several excellent reviews are available, [13], [10]. The field of prior selection for model spaces and their exploration via the posterior is less well developed but is currently under very active investigation, see [5], [16], and [12], among others. Overall, the implications for the present context seem to be that any 'reasonable' proper prior will give better behavior than a non-proper prior. That is, the propriety of the prior is what lets it serve as a sanity check by ruling out some regions of the model space or parameter spaces that are unrealistic. It can be seen from the statement of Theorem 1 that the choice of prior only affects the finite sample properties of estimators.

**Acknowledgments**
This research was supported by by NSERC operating grant number RGPIN 138122.

**References**

[1] D.F. Andrews, P.J. Bickel, F.R. Hampel, P.J. Huber, W.H. Rogers, J.W. Tukey, Robust Estimates of Location: Survey and Advances, Princeton University Press, Princeton, New Jersey, 1972.

[2] J. Borwanker, G. Kallianpur, B.L.S. Prakasa Rao, The Bernstein-Von Mises Theorem for Markov Processes, Ann. Math. Stat. 42(1971) 1241-1253.

[3] P. Čížek, Asymptotics of least trimmed squares regression, CentER Discussion Paper 2004-72, Tilburg University, The Netherlands, 2004.

[4] P. Čížek, Least trimmed squares in nonlinear regression under dependence, J. Statist. Plann. Inference 136(2005) 3967-3988.

[5] C. Carvalho, J. Scott, Objective Bayesian model selection in Gaussian graphical models, Biometrika 96 (2009) 497-512.

[6] L. Fahrmeir, H. Kaufmann, Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in Generalized Linear Models, Ann. Statist. 12(1985) 342-368.

[7] J.K. Ghosh, R.V. Ramamoorthi, Bayesian Nonparametrics, Springer Series in Statistics, Springer, 2003.

[8] Gallant, A.R, Nonlinear Statistical Models, John Wiley and Sons, 1987.

[9] G.W. Bassett, R.W. Koenker, Strong Consistency of Regression Quantiles and Related Empirical Processes, Econometric Theory 2(1986) 191-201.

[10] M. Ghosh, Objective priors: A selective Review, Technical Report, Dept. of Statistics, Univ. of Florida.

[11] F.R. Hampel, Beyond Location Parameters: Robust Concepts and Methods, Bulletin of the International Statistical Institute. 46 375-382, 1975.

[12] J. Hu, V. Johnson, Bayesian model selection using test statistics, J. R. S. S. Ser. B 71-1 (2009) 143-158.

[13] R. Kass, L. Wasserman, The Selection of Prior Distributions by Formal Rules, J. Amer. Stat. Assoc. 91 (1996) 1343-1370..

[14] J. Kim, D. Pollard, Cube Root Asymptotics, Ann. Statist. 18(1990) 191-219.

[15] E.L. Lehmann, Theory of Point Estimation, John Wily and Sons, 1983.

[16] F. Liang, R. Paulo, G. Molina, M. Clyde, and J. Berger, Mixtures of $g$-priors for Bayesian variable selection. J. Amer. Stat. Assoc. 103 (2008) 410-423.

[17] R.W. Koenker, G.W. Bassett, Regression Quantiles, Econometrica 46(1978) 33-50.

[18] P. Milasevic, G.R. Ducharme, Uniqueness of the Spatial Median, Ann. Statist. 15(1987) 1332-1333.

[19] B.K.S. Prakasa Rao, Asymptotic Theory of Statistical Inference, John Wiley and Sons, 1987.

[20] P.J. Rousseeuw, Least Median of Squares Regression, J. Amer. Statist. Assoc. 79(1984) 871-880.

[21] P.J. Rousseeuw, A.M. Leroy, Robust Regression and Outlier Detection, John Wiley and Sons, 2003.

[22] M.J. Schervish, Theory of Statistics, Springer Series in Statistics, Springer, 1997.

[23] G.R. Shorack, Probability for Statisticians, Springer texts in statistics, Springer-Verlag, New York, 2000.

[24] A.J. Stromberg, Consistency of the least median of squares estimator in nonlinear regression, Comm. Statist. Theory Methods 24(1995) 1971-1984.

[25] R.J. Tomkins, Convergence Properties of Conditional Medians, Canad. J. Statist. 6(1978) 169-177.

[26] A. Wald, Contributions to the Theory of Statistical Estimation and Testing Hypotheses, Ann. Math. Stat. 10(1939) 299-326.

[27] A. Wald, Statistical Decision Functions, John Wiley, New York, 1950.

[28] C. W. Yu, Median Loss Analysis and its Application to Model Selection. PhD thesis, Dept. of Statistics, University of British Columbia.

[29] C.W. Yu, B. Clarke, Median Loss Analysis, Submitted, 2008.

[30] C.W. Yu, B. Clarke, Cube root asymptoics of the Least Quantile of Squares Estimator in Nonlinear Regression Models, Submitted, 2008

[31] C.W. Yu, B. Clarke, Asymptotics of Bayesian Median Loss Estimation, Technical Report # 243, September 2008, Department of Statistics, University of British Columbia.