THE UNIVERSITY OF BRITISH COLUMBIA

DEPARTMENT OF STATISTICS

TECHNICAL REPORT # 243

Asymptotics of Bayesian Median Loss Estimation

By Chi Wai Yu and Bertrand Clarke

September 2008

Asymptotics of Bayesian Median Loss Estimation

By Chi Wai Yu and Bertrand Clarke

Abstract

Here we define estimators based on minimizing the median of a loss function in the Bayesian context. For these estimators, we establish $\sqrt{n}$-consistency, asymptotic normality and efficiency. We contrast the asymptotic behavior of these *medloss* estimators with the $n^{1/3}$ asymptotic behavior of the least median of squares (LMS) estimators and the least trimmed squares (LTS) estimators which are $\sqrt{n}$-consistent. The performance of Bayesian *medloss* estimators is thus intermediate between the LMS and LTS estimators since they use an actual median but still get $\sqrt{n}$-asymptotics.

1 Introduction

The main ideas of Statistical Decision Theory were proposed by Wald [24]. Later, in their book "Theory of Games and Economic Behavior", von Neumann and Morgenstern [23], hereafter vNM, developed axiomatic decision theory for choice behavior in a Frequentist context. Their theory influenced the final shape of Statistical Decision Theory, Wald [25]. In the Bayesian context, Savage [18] extended vNM's reasoning by providing other axioms for the maximization of expected utility to be the criterion for decision making in a subjective probability, i.e. Bayes setting.

However, there are well-known criticisms of the axioms for the existence of vNM's and Savage's expected utility representations. For instance, the Allais paradox, Allais [1], and Ellsberg paradox, Ellsberg [9], show that vNM's Independence axiom and Savage's Sure-Thing principle contradict real life decision making. Consequently, various alternatives to the expected utility models have been proposed.

Manski [12] constructed quantile utility models in a Frequentist context and proposed that the quantile of the utility function should be maximized. However, Manski's approach was not axiomatic. This led Machina and Schmeidler [11], in the Bayesian context, to consider axiomatic models for decision making that did not rest on expected utility. However, their approach does not cover quantile utility models. Most recently Rostek [15] has proposed an axiomatic foundation for Quantile Maximization in the Bayesian context. Her axiomatization means that the best decision should maximize the p^{th} quantile of the utility function, or equivalently to minimize the $(1-p)^{th}$ quantile of the loss.

In a statistical context, Rostek's result justifies using quantiles of the loss, and in this paper, we choose the median of the loss, hereafter called the *medloss*, i.e. we take $p = 0.5$. The median is appropriate because the non-negativity of the loss means that if the loss function itself is regarded as a random variable, it has a right skewed distribution, often strongly right skewed. For such distributions, the median is a more reasonable measure of location than the mean is. In addition, in terms of prediction, using the median helps avoid overprediction and underprediction in terms of the loss.

Parallel to the Bayes estimate or posterior expected-loss estimate in classical decision theory, we define a posterior *medloss* estimate by

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

where $x^n = (x_1, \ldots, x_n)$ are the realizations of the n random variables $X^n = (X_1, \ldots, X_n)$, $\mathcal{L}(d, \theta)$ is the loss function, $d(x^n)$ is the estimate for θ, $\mathcal{D}$ 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 π of Θ given x^n. Our main result is that these estimators are $\sqrt{n}$-consistent, asymptotically normal and efficient.

To the best of our knowledge, no one has provided an axiomatization in the frequentist context which implies that minimizing the *medloss* is the appropriate criterion for choosing an estimator. We conjecture that this can be done, although we do not do so here.

Note that the least median of squares (LMS) estimate is the frequentist version of our median-loss estimate for regression problems. The LMS estimate was introduced by Rousseeuw [16] to estimate regression parameters because of its high robustness to outliers. The consistency of the LMS estimate in nonlinear regression models was established by Stromberg [20]. However, it was also shown to have a slow rate of convergence in the linear regression setting by Andrews *et al.* [4], see also Kim and Pollard [10].

In addition to the Bayesian version in (1.1), we can define the Frequentist *medloss* estimator for θ by

$$\delta(X^n) = \arg\min_{d\in\mathcal{D}} \min_{\theta} \operatorname*{med}_{X^n} \mathcal{L}(d(X^n), \theta), \tag{2}$$

where $\operatorname*{med}_{X^n} \mathcal{L}$ is the median of the loss $\mathcal{L}$ with respect to X^n under the distribution P_θ. We make use of this definition in a nonlinear regression setting. Indeed, consider the nonlinear regression model

$$y_i = h(x_i, \beta_0) + u_i, \quad i = 1, \ldots, n.$$

The LMS estimator and the two-sided least trimmed squares (LTS) estimator are defined by

$$\beta_n = arg\min_{\beta} \operatorname*{median}_{1\le i\le n} [y_i - h(x_i, \beta)]^2,$$

and

$$\beta_n^{(LTS,h)} = \arg\min_{\beta} \sum_{n-h+1}^{h} r_{[i]}^2(\beta),$$

respectively, 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 h must satisfy $\frac{n}{2} < h \le n$. Here we extend the existing asymptotic results for the LMS estimators to the nonlinear regression setting and for the one-sided LTS estimators to two-sided case. We find the LTS estimators are $\sqrt{n}$-consistent and asymptotically normal but not efficient, and the LMS estimators exhibit $n^{1/3}$-asymptotics. The Bayesian *medloss* estimators represent a good tradeoff between

the rate of the LTS estimators and the use of the median in the LMS estimators. This suggests that the Bayesian *medloss* estimators are to be preferred over the LTS and LMS estimators.

In general, the *medloss* estimators in the Bayesian and Frequentist contexts have nice properties such as high robustness to outliers and to the choice of the loss function, and good prediction. We also suggest that the *medloss* estimators should be appropriate when the underlying distribution is asymmetric or heavy-tailed.

The rest of this paper is organized as follows. In Section 2, we present the consistency and asymptotic normality of posterior *medloss* estimators. For comparison, we also state the asymptotic results for LMS estimators in Section 3 and for two-sided LTS estimators in Section 4. In Section 5, we summarize the implications of our work.

2 Consistency and Asymptotic normality for Bayesian *medloss* estimators

Let $X_0^n = (X_0, X_1, \cdots, X_n)$ and define the posterior *medloss* estimator $\delta_n = \delta_n(X_0^n)$ to be the one which minimizes the *medloss*

$$M_n(a) = \underset{\pi(\Theta|X^n)}{med} \mathcal{L}(a, \theta).$$

Consider the setting of Borwanker, et al. [5], in which the consistency and asymptotic normality are established for Markov processes thereby implying the analogous results for IID cases. Suppose that $X_0, X_1, X_2, \cdots$ are random variables forming a strictly stationary ergodic Markov process and taking values in a measurable space $(S, \mathcal{B}_S)$. The stationary initial probability distribution and the transition probability function of the process will be denoted by $P_\theta(A)$ and $P_\theta(y|A)$ for $y \in S$ and $A \in \mathcal{B}_S$ respectively, where $\theta \in \Theta \subset R$. Suppose that there exists a $\sigma-$finite measure μ on $(S, \mathcal{B}_S)$ such that $P_\theta(A)$ and $P_\theta(y|A)$ are both absolutely continuous with respect to μ with densities $f(z|\theta)$ and $f(y, z|\theta)$ respectively. For $\theta \in \Theta$, denote by P_θ the measure on the product measurable space determined by the initial probability distribution and the transition probability function. Given the observations $x_0^n = \{x_0, x_1, \cdots, x_n\}$, the log likelihood function of the process is defined by

$$\ln L_n(\theta, x_0^n) = \ln f(x_0|\theta) + \textstyle\sum_{i=0}^{n-1} f(x_i, x_{i+1}|\theta).$$

Moreover, let θ_0 be the true parameter and $P_0 = P_{\theta_0}$. Borwanker, et al. [5] suggested that $\ln f(x_0|\theta)$ in the above expression may be neglected in the large sample theory. Consider the following assumptions with the observations x_0^n.

Assumption 1.1: The parameter space Θ is an open interval in $\mathcal{R}$. Π is a prior probability measure on $(\Theta, \mathcal{F})$, where $\mathcal{F}$ is the $\sigma-$algebra of Borel subsets of Θ and Π is absolutely continuous and has a density π with respect to the Lebesgue measure on $\mathcal{R}$.

Assumption 1.2: Suppose that $\frac{\partial}{\partial\theta}\ln f(x_0, x_1|\theta)$ and $\frac{\partial^2}{\partial\theta^2}\ln f(x_0, x_1|\theta)$ exist and are continuous in θ for almost all pairs $(x_0, x_1)(\mu \times \mu)$.

Assumption 1.3: For every $\theta \in \Theta$, there exists $\eta(\theta) > 0$ such that

$$E_\theta\Big[sup\Big\{\Big|\tfrac{\partial^2}{\partial\theta^2}\ln f(X_0, X_1|\theta')\Big| : |\theta - \theta'| < \eta(\theta), \theta' \in \Theta\Big\}\Big] < \infty.$$

Assumption 1.4: For every $\theta \in \Theta$ and any $\epsilon > 0$,

$$-\infty < E_\theta[sup\{\ln \tfrac{f(X_0,X_1|\theta')}{f(X_0,X_1|\theta)} : |\theta - \theta'| \geq \epsilon, \theta' \in \Theta\}] < 0.$$

Assumption 1.5: Let

$$i(\theta) = -E_\theta\Big[\tfrac{\partial^2}{\partial\theta^2}\ln f(X_0, X_1|\theta)\Big], \text{ for } \theta \in \Theta.$$

It is clear that $i(\theta) < \infty$ for all $\theta \in \Theta$. Suppose that $i(\theta) > 0$ and $i(\theta)$ is continuous in θ.

Assumption 1.6: The proper prior density π is continuous and positive in an open neighborhood of the true parameter θ_0.

Before showing our main result, we need the following lemma for the asymptotic normality of the MLE in the setting of Markov process.

Lemma 1 (Theorem 2.4 of Borwanker, et al. [5]). *Under Assumptions 1.1-1.5, there exists a compact neighborhood U_{θ_0} of θ_0 such that*

$$\textit{(i) } \hat{\theta}_n \to \theta_0 \textit{ a.s. and (ii) } n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} N(0, i_0^{-1}),$$

where $\hat{\theta}_n = \hat{\theta}_n(x_0^n) = \arg\sup_{\theta \in U_{\theta_0}} \ln L_n(\theta, x_0^n)$.

Under Assumptions 1.1-1.6, now we show the asymptotic normality of the posterior *medloss* estimator, which is an median-loss analog of the asymptotic result for the posterior risk estimator in [5].

Theorem 1. *Let a Markov chain $\{X_n, n \geq 0\}$ satisfying all of the above assumptions. Let $\delta_n = \delta_n(x_0, \cdots, x_n)$ be the posterior medloss estimator of θ for all $(x_0, x_1, \cdots, x_n)$ and all n with respect to a loss function $\mathcal{L}(\theta, a)$ satisfying the following conditions:*

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

(ii) $l(t_1) \geq l(t_2)$ *if* $|t_1| \geq |t_2|$.

Moreover, suppose that there exist a non-negative sequence $\{a_n\}$ *and monotone increasing continuous function* $K(\cdot)$ *such that*

(iii) For any real number c,

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

where $T = \sqrt{n}(\Theta - \hat{\theta}_n)$, $\hat{\theta}_n$ *is the MLE of* θ_0 *and* $\underset{T|X^n}{med}$ *is the median with respect to* T *given* X^n,

(iv) 1/2 is the continuous point of the distribution of $K(Z)$, *and*

(v) $med_Z K(Z+m)$ *has a unique minimum at* $m = 0$, *where* med_Z *is the median with respect to* Z *having a normal distribution* $N(0, i_0^{-1})$.

Then we have

$$\delta_n \to \theta_0 \quad a.s.P_0 \text{ and } n^{1/2}(\theta_0 - \delta_n) \xrightarrow{\mathcal{L}} N(0, i_0^{-1}).$$

To prove Theorem 1, we need the notion of the convergence in quantile, Shorack (2000), and Shorack's Proposition 1.

Definition 1. *For any distribution function* $F(\cdot)$, *the quantile function is*

$$K(t) \overset{\text{def}}{=} F^{-1}(t) = \inf\{x : F(x) \geq t\}, \text{ for } 0 < t < 1.$$

Now denote by K_n *the quantile function associated with the distribution function* F_n *for each* $n \geq 0$. *Then* K_n *converges in quantile to* K_0, *denoted by* $K_n \xrightarrow{\mathcal{Q}} K_0$, *if* $K_n(t) \to K_0(t)$ *at each continuity point* t *of* $K_0(t)$ *in (0, 1).*

Lemma 2 (Shorack [19])**.** *Using the same notation as in Definition 1,*

$$F_n \xrightarrow{\mathcal{L}} F_0 \iff K_n \xrightarrow{\mathcal{Q}} K_0.$$

Now we can prove Theorem 1.

Proof. We prove Theorem 1 in three steps. The first shows that $W_n = n^{1/2}(\hat{\theta}_n - \delta_n)$ is finite a.s. and the second step shows it goes to 0 $a.s.P_0$. Then we complete the proof by using the Slutsky's theorem and the asymptotic normality of $\hat{\theta}_n$.

1. First, for $T = n^{1/2}(\Theta - \hat{\theta}_n)$,

$$\begin{aligned} \limsup_n a_n M_n(\delta_n) &\leq \limsup_n a_n M_n(\hat{\theta}_n) \\ &= \limsup_n a_n \underset{\pi(\Theta|X^n)}{\text{med}}\, l(\Theta - \hat{\theta}_n) \\ &= \limsup_n \underset{T|X^n}{\text{med}}[a_n l(T/n^{1/2})]. \end{aligned}$$

Moreover, $\Big|\text{med}[a_n l(T/n^{1/2})] - \text{med}_Z[K(Z)]\Big| \leq \Big|\text{med}[a_n l(T/n^{1/2})] - \text{med}[K(T)]\Big| + \Big|\text{med}[K(T)] - \text{med}_Z[K(Z)]\Big| \to 0$. The first term goes to zero based on the condition (iii) of the loss function. By Theorem 3.2 of Borwanker, et al. [5] that the density of T converges to that of Z in total variation, we have the convergence of T to Z in distribution because

$$\begin{aligned}|F_T(x) - F_Z(x)| &\leq \int_{-\infty}^{x} |f_T(u) - f_Z(u)|du \\ &\leq \int_{-\infty}^{\infty} |f_T(u) - f_Z(u)|du \to 0, \forall x,\end{aligned}$$

where $F_T(\cdot)$ and $F_Z(\cdot)$ are the cdf's of T and Z, respectively, and $f_T(\cdot)$ and $f_Z(\cdot)$ are the corresponding pdf's. Further, by the continuity of K and the Continuous Mapping Theorem, $K(T)$ converges in distribution to $K(Z)$. Thus, by Lemma 2, $\text{med}K(T) \to \text{med}_Z K(Z)$, which implies that the second term converges to zero. So,

$$\limsup_n a_n M_n(\delta_n) \leq \limsup_n a_n M_n(\hat{\theta}_n) \leq \text{med}_Z K(Z). \tag{3}$$

2. Now we will show $n^{1/2}(\hat{\theta}_n - \delta_n) = W_n < \infty \, a.s.$ by using the argument of Borwanker, et al. [5], but here we consider the posterior *medloss* instead of the posterior risk.

   First, suppose that the statement $W_n < \infty \, a.s.$ is false, then for every $M > 0$, there exists a set A_M with $P_\theta(A_M) > 0$ such that $|W_n(x)| > M \, i.o.$ for $x \in A_M$. Without loss of generality, we can assume that $W_n(x) > M \, i.o.$ Then, for the subsequence $\{n_i\}$ where the inequality holds, we have

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

   by condition (iv). The inequality holds because $XI_{\{A\}} \leq X$ for any non-negative random variable X and an indicator function I with any set A. Note that $K(Z + M)I_{\{Z+M \geq 0\}}$ is a non-decreasing function of M for each fixed Z. So, by Tomkins' corollary in [21], for the median version of the Lebesgue dominated convergence theorem, we have

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

Therefore, for a set of positive probability,

$$\liminf_n a_n M_n(\delta_n) > \mathrm{med}_Z K(Z) \geq \limsup_n a_n M_n(\hat{\theta}_n),$$

which contradicts the definition of δ_n. Thus, $\limsup_n |W_n| < \infty \, a.s. \, P_0$.

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

$$\begin{aligned}
\liminf_{n_i} a_{n_i} M_{n_i}(\delta_{n_i}) &= \liminf_{n_i} \operatorname*{med}_{T|X^{n_i}} \Big[a_{n_i} l\Big(\frac{T + W_{n_i}}{n_i^{1/2}} \Big) \Big] \\
&\geq \liminf_{n_i} \operatorname*{med}_{T|X^{n_i}} \Big[a_{n_i} l\Big(\frac{T + W_{n_i}}{n_i^{1/2}} \Big) \Big] - \epsilon \\
&= \mathrm{med}_Z K(Z + m) - \epsilon \\
&> \mathrm{med}_Z K(Z) - \epsilon.
\end{aligned}$$

Since ϵ is arbitrary, we get $\liminf_{n_i} a_{n_i} M_{n_i}(\delta_{n_i}) > \mathrm{med}_Z K(Z)$, which is impossible by (3). Thus, m=0 and $n^{1/2}(\delta_n - \hat{\theta}_n) \to 0 \, a.s. P_0$.

3. Finally, the proof is completed by observing $n^{1/2}(\delta_n - \theta_0) = n^{1/2}(\delta_n - \hat{\theta}_n) + n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} N(0, i_0^{-1})$.

□

Note that conditions (i), (ii) and (iii) are true for L^1 loss with $a_n = n^{1/2}$ and $K(t) = |t|$. Also, since Z has a normal distribution with median 0, conditions (iv) and (v) are satisfied. Therefore, we have the following result.

Corollary 1. *Consider any continuous posterior density of Θ given $X^n = x^n$ with L^p loss, i.e. $\mathcal{L}(\theta, a) = |\theta - a|^p$. Assume that the median of the loss $\mathcal{L}$ is unique. Then for any strictly increasing functions $\mathcal{L}$ of $|\Theta - d(x^n)|$, we have*

$$\delta_n \to \theta_0 \quad a.s. P_0 \text{ and } n^{1/2}(\theta_0 - \delta_n) \xrightarrow{\mathcal{L}} N(0, i_0^{-1}).$$

2.1 For IID random variables

For the corresponding results in an IID setting, we can follow Prakasa Rao [14]. Basically, what we need to do is to change the setting for Markov process to IID random variables. The proofs for the results of IID random variables are similar to those for Markov process. Therefore, we only provide the required settings and assumptions for the IID case; and the proofs are omitted.

Consider IID random variables $X_i, 1 \leq i \leq n$, which are defined on a measurable space $(\Omega, \mathcal{B})$ with probability measure P_θ, $\theta \in \Theta \subset R$. Then similar to Assumption 1.2, we have

Assumptions 1.2*: Suppose $P_\theta << \mu$, μ $\sigma-$finite on $(\Omega, \mathcal{B})$. Let

$$f(x|\theta) = \frac{dP_\theta}{d\mu}(x).$$

Suppose that $\frac{\partial}{\partial\theta} \ln f(x|\theta)$ and $\frac{\partial^2}{\partial\theta^2} \ln f(x|\theta)$ exist and are continuous in θ for x a.e. $[\mu]$.

Assumptions 1.3* to 1.5* are the same as Assumptions 1.3 to 1.5 but using $\ln f(x|\theta)$, instead of $\ln f(x_0, x_1|\theta)$.

Theorem 3 for i.i.d. random variables holds under Assumptions 1.1, 1.2*-1.5* and 1.6. This result is verified by the following theorem.

Theorem 2. *Let $\{X_i, 1 \leq i \leq n\}$ be IID random variables satisfying Assumptions 1.1, 1.2*-1.5* and 1.6. Let $\delta_n = \delta_n(x_1, \ldots, x_n)$ be the posterior medloss estimator of θ with respect to a loss function $\mathcal{L}(a, \theta)$ satisfying the conditions (i)-(iv) defined in Theorem 1, then we have*

$$\delta_n \to \theta_0 \quad a.s. P_0 \text{ and } n^{1/2}(\theta_0 - \delta_n) \xrightarrow{\mathcal{L}} N(0, i_0^{-1}).$$

3 Asymptotic results for the LMS estimator in nonlinear regression models

Next we turn to the asymptotic results for LMS estimators in the regression context. In linear regression models, Kim and Pollard (1990) deduced a limiting Gaussian process for LMS estimators. Here, we extend their result to nonlinear cases.

First, let $\mathcal{H}$ be a vector space of real-valued functions h. Consider the nonlinear regression model

$$y_i = h(x_i, \beta_0) + u_i, \quad i = 1, \ldots, n, \tag{4}$$

where y_i, x_i and u_i are the realizations of random variables $Y_i \in \mathcal{R}, X_i \in \mathcal{R}^p$ and $U_i \in \mathcal{R}$, respectively, and $\beta_0 \in \mathcal{B} \subset \mathcal{R}^d$ is an unknown true parameter for the known function $h \in \mathcal{H}$. Assume that the parameter space $\mathcal{B}$ is compact and β_0 is its interior point, and that (x_i, u_i) are independently sampled from a probability distribution $\mathcal{P}$ on $\mathcal{R}^p \times \mathcal{R}$. So, the LMS estimator is defined by

$$\beta_n = \arg \min_{\beta} \underset{1 \leq i \leq n}{\text{median}} [y_i - h(x_i, \beta)]^2. \tag{5}$$

The asymptotic results of the LMS estimator β_n in the non-linear regression models (4) rely heavily on Kim and Pollard's main theorem [10], so we state this theorem before giving our main result. The notion of manageability used below is discussed in Appendix A.1.

Theorem 3 (Kim and Pollard, [10]). *Consider the empirical processes*

$$E_n g(\cdot, \theta) = \frac{1}{n} \sum_{i \leq n} g(\eta_i, \theta),$$

where $\{\eta_i = (x_i, u_i)\}$ *is a sequence of independent observations taken from a distribution* $\mathcal{P}$ *on* $\mathcal{R}^p \times \mathcal{R}$ *and* $G = \{g(\cdot, \theta) : \theta \in \Theta\}$ *is a class of functions indexed by a subset* Θ *of* $\mathcal{R}^d$.

Define the envelope $G_R(\cdot)$ *as the supremum of* $|g(\cdot, \theta)|$ *over the class* $\mathcal{G}_R = \{g(\cdot, \theta) : \|\theta - \theta_0\| \leq R\}$, *i.e.*

$$G_R(x_i, u_i) = \sup_{g \in \mathcal{G}_R} |g(x_i, u_i, \theta)|.$$

Also make the following assumptions:

1. *Choose a sequence of estimators* $\{\theta_n\}$ *for which* $E_n g(\cdot, \theta_n) \geq \sup_{\theta \in \Theta} E_n g(\cdot, \theta) - o_p(n^{-2/3})$.
2. *The sequence* $\{\theta_n\}$ *converges in probability to the unique* θ_0 *that maximizes* $Eg(\cdot, \theta)$, *the expectation of* $g(\cdot, \theta)$ *with respect to the distribution* P.
3. *The true value* θ_0 *is an interior point of* Θ.

   Let the functions $g(\cdot, \theta_0)$ *be standardized so that* $g(\cdot, \theta_0) = 0$ *and suppose that the class* $\mathcal{G}_R$, *for* R *near 0, is uniformly manageable for the envelopes* G_R. *Then we also require :*

4. $Eg(\cdot, \theta)$ *is twice differentiable with second derivative matrix* $-V$ *at* θ_0.
5. $H(s,t) \equiv \lim_{\alpha \to \infty} \alpha E g(\cdot, \theta_0 + s/\alpha) g(\cdot, \theta_0 + t/\alpha)$ *exists for each* s, t *in* $\mathcal{R}^d$ *and*

   $$\lim_{\alpha \to \infty} \alpha E g(\cdot, \theta_0 + t/\alpha)^2 \{|g(\cdot, \theta_0 + t/\alpha)| > \epsilon \alpha\} = 0$$

   for each $\epsilon > 0$ *and* $t \in \mathcal{R}^d$.
6. $EG_R^2 = O(R)$ *as* $R \to 0$ *and for each* $\epsilon > 0$ *there is a constant* K *such that* $EG_R^2 I_{\{G_R > K\}} < \epsilon R$ *for* R *near 0.*
7. $E|g(\cdot, \theta_1) - g(\cdot, \theta_2)| = O(|\theta_1 - \theta_2|)$ *near* θ_0.

Now, under the above assumptions 1 - 7, we have that the process $n^{2/3} E_n g(\cdot, \theta_0 + t n^{-1/3})$ *converges in distribution to a Gaussian process* $Z(t)$ *with continuous sample paths, expected value* $\frac{-1}{2} t' V t$ *and covariance kernel* H, *as* $n \to \infty$.

Finally, if V *is positive definite and if* Z *has nondegenerate increments, then* $n^{1/3}(\theta_n - \theta_0)$ *converges in distribution to the (almost surely unique) random vector that maximizes* Z, *as* $n \to \infty$.

Now we can state our generalization to nonlinear models. By verifying the assumptions of Theorem 3, we have the following.

Theorem 4. *Suppose*

1. *dim($\mathcal{H}$) is finite.*
2. *$Q_h = E_X[h'(X,\beta_0)h'(X,\beta_0)^T]$ is positive definite.*
3. *u_i has a bounded, symmetric density γ that decreases away from its mode at zero, and it has a strictly negative derivative at r_0, the unique median of $|u|$.*
4. *For any $h \in \mathcal{H}$, h satisfies the Lipschitz condition, i.e.*

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

   and $E_X(L_X) < \infty$.
5. *$E_X\|h'(X,\xi)\| < \infty$ for $\xi \in U(\beta_0, R)$, where $U(a,b)$ means an open ball at center a with radius b, and R is defined for the envelope G_R.*
6. *$E_X|h'(X,\beta_0)^T w| \neq 0$ for any $w \neq 0$.*

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

$$Z(\theta) = \gamma'(1)\theta^T Q_h \theta + W(\theta),$$

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

In the following, we just outline the proof of Theorem 4; the full proof is in Appendix A. First, we recast (5) as a problem of constrained optimization by reparametrizing β by $\beta_0 + \theta$, and taking a first-order Taylor expansion of $h(x,\beta)$ at β_0. Thus,

$$y - h(x,\beta) = u - h'(x,\xi)^T\theta, \tag{6}$$

where $\xi \in (\beta_0, \beta)$ and $\xi \to \beta_0$ as $\theta \to 0$. Then define

$$f_{h,x,u}(\theta, r, \xi) = I_{\{|u - h'(x,\xi)^T\theta| \leq r\}}(x,u),$$

and

$$r_n = \inf\{r : \sup_\theta E_n f_{h,x,u}(\theta, r, \xi) \geq 1/2\}. \tag{7}$$

Let $\theta_n = \beta_n - \beta_0$ be a value at which $\sup_\theta E_n f_{h,x,u}(\theta, r_n, \xi)$ is achieved, where E_n corresponds to the empirical version of the expectation under $\mathcal{P}$.

Assume that the corresponding constrained maximization (7) for the expectation under $\mathcal{P}$ has a unique solution θ_0 and r_0. Without loss of generality, let $\theta_0 = 0$ and $r_0 = 1$. Since $f_{h,x,u}(\theta, r, \xi)$ can be rewritten as

$$I_{\{|y-h(x,\theta+\beta_0)|\le r\}}(x,y) = I_{\{h(x,\theta+\beta_0)-y+r\ge 0 \text{ and } y+r-h(x,\theta+\beta_0)\ge 0\}}(x,y), \tag{8}$$

we let $f_{h,x,y}(\theta, r) = f_{h,x,u}(\theta, r, \xi)$ and define

$$g_{h,x,u}(\theta, \delta, \xi) = f_{h,x,u}(\theta, 1+\delta, \xi) - f_{h,x,u}(0, 1+\delta, \xi).$$

Applying Kim and Pollard's main theorem in [10], here stated as Theorem 3, in the present setting will establish our result Theorem 4. So it suffices to check whether all the required conditions of Kim and Pollard's theorem can be satisfied in the nonlinear case. The verifications are shown in Appendix A.

4 Limiting results for two-sided LTS in nonlinear regression models

Since it is based on a median, the LMS estimator can be viewed as a trimmed mean estimator with a trimming proportion of 50% on both sides. 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 verify this intuition, we see that relaxing the trimming proportion gives the $n^{1/2}$ rate of convergence and asymptotic normality. In this subsection, we propose the two-sided LTS estimator in nonlinear models and establish its limiting behavior. Our work is based on the $n^{1/2}$-convergence and asymptotic normality of the one-sided LTS estimators that were shown by Čížek [6, 7].

Consider the nonlinear regression model (4) and a sequence of the variables $\{x_t\}_{t\in\mathcal{N}}$ satisfying

$$\sup_{t\in\mathcal{N}} E\Big\{ \sup_{B\in\sigma^f_{t+m}} |P(B|\sigma^p_t) - P(B)| \Big\} \to 0,$$

as $m \to \infty$, where $\sigma^p_t = \sigma(x_t, x_{t-1}, \ldots)$ and $\sigma^f_t = \sigma(x_t, x_{t+1}, \ldots)$ are σ-algebras. The two-sided LTS estimator is defined by

$$\beta_n^{(LTS,h)} = \arg\min_{\beta} S_n(\beta), \tag{9}$$

where $S_n(\beta) \overset{def}{=} \sum_{n-h+1}^{h} r^2_{[i]}(\beta)$, $r^2_{[i]}(\beta)$ represents the i^{th} order statistics of squared residuals $r_i^2(\beta) = \{y_i - h(x_i, \beta)\}^2$, and the trimming constant h satisfies $\frac{n}{2} < h \le n$. 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.

The choice of the trimming constant h depends on the sample size n, so consider a sequence of trimming constants h_n. Since h_n/n determines the fraction of sample included in the LTS objective function, we

choose a sequence for which $h_n = [\lambda n]$, where $[z]$ represents the integer part of z so that $h_n/n \to \lambda$ for some $1/2 < \lambda \leq 1$.

Čížek made assumptions for the asymptotic results of the one-sided LTS estimator. They can be classified into three groups: Assumptions D, H and I, where assumptions D are for the distributional assumptions for the random variables, assumptions H for the regression functions h and assumptions I for the identification setting.

Our main results for the two-sided case also follow these assumptions, except for Čížek's assumptions $D3$ and $I2$. We make two alternative assumptions $TD3$ and $TI2$ for the two-sided case in place of his $D3$ and $I2$. Specifically, we have

Assumption $TD3$: assume that for $\lambda \in (0,1)$,

$$m_{gg} \stackrel{def}{=} \inf_{\beta \in \mathcal{B}} \inf_{z \in (-\delta_g, \delta_g)} g_\beta(G_\beta^{-1}(\lambda) + z) > 0$$

for some $\delta_g > 0$. Additionally, when $1/2 < \lambda \leq 1$, suppose that

$$m_G^* \stackrel{def}{=} \sup_{\beta \in \mathcal{B}} G_\beta^{-1}(1-\lambda) > 0 \text{ and } m_G^{**} \stackrel{def}{=} \inf_{\beta \in \mathcal{B}} G_\beta^{-1}(\lambda) > 0,$$

and

$$M_{gg}^* \stackrel{def}{=} \sup_{\beta \in \mathcal{B}} \sup_{z \in (-\infty, m_G^*)} g_\beta(z) < \infty \text{ and } M_{gg}^{**} \stackrel{def}{=} \sup_{\beta \in \mathcal{B}} \sup_{z \in (m_G^{**}, \infty)} g_\beta(z) < \infty,$$

where G_β and g_β are the distribution function and probability density function of $r_i^2(\beta)$.

Assumption $TI2$: For any $\epsilon > 0$ and an open ball $U(\beta_0, \epsilon)$ such that $\mathcal{B} \cap U^c(\beta, \epsilon)$ is compact, there exist $\alpha(\epsilon) > 0$ such that it holds, for $1/2 < \lambda \leq 1$, that

$$\min_{\|\beta - \beta_0\| > \epsilon} E\left[r_i^2(\beta) I_{\{G_\beta^{-1}(1-\lambda) \leq r_i^2(\beta) \leq G_\beta^{-1}(\lambda)\}}\right] > E\left[r_i^2(\beta_0) I_{\{G_{\beta_0}^{-1}(1-\lambda) \leq r_i^2(\beta_0) \leq G_{\beta_0}^{-1}(\lambda)\}}\right] + \alpha(\epsilon).$$

To indicate the modifications of Čížek's assumptions D, H and I, we denote our assumptions by TD, TH and TI, respectively. Our theorems for the two-sided LTS estimator on nonlinear models rely on [6, 7], so our main results on consistency and asymptotic normality of $\beta_n^{(LTS,h)}$ stated in Theorems 5 and 6 rely on numerous preliminary results. Figure ? shows that these preliminary results lead to the desired theorems.

To implement Figure 1, we start with Lemma 3. Let

$$S_n(\beta) = \sum_{n-h_n+1}^{h_n} r_{[i]}^2(\beta).$$

Lemma 3. *Under assumptions $TD2$ and $TH1$, $S_n(\beta)$ is continuous on B, twice differentiable at $\beta_n^{(LTS,h_n)}$ if $\beta_n^{(LTS,h_n)} \in U(\beta_0, \delta)$, and almost surely twice differentiable at any fixed point $\beta \in U(\beta_0, \delta)$.*

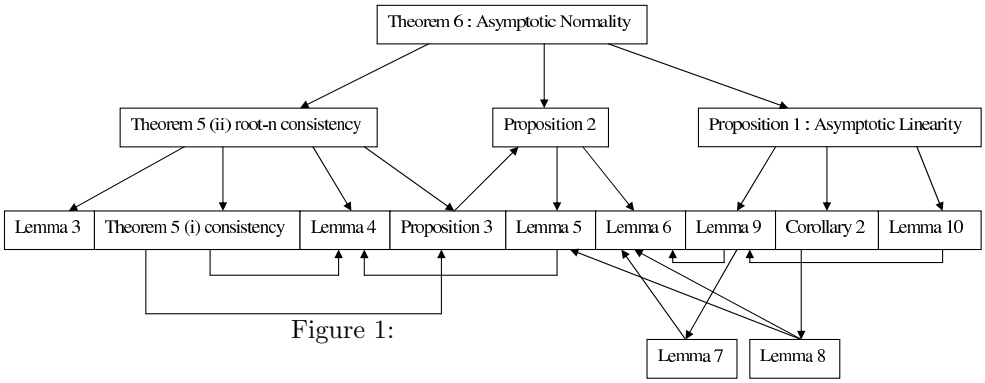

Figure 1:

Denote $I_{[r^2_{[n-h_n+1]}(\beta), r^2_{[h_n]}(\beta)]}(r_i^2(\beta))$ *by* $I_2(r_i^2(\beta))$, $h'_\beta(x_i,\beta)$ *by* h'_β *and* $h''_{\beta\beta}(x_i,\beta)$ *by* $h''_{\beta\beta}$. *Then we have*

$$
\begin{aligned}
S_n(\beta) &= \sum_{i=1}^{n} r_i^2(\beta) I_2(r_i^2(\beta)), \qquad (10)\\
S'_n(\beta) &= -2\sum_{i=1}^{n} r_i^2(\beta) h'_\beta I_2(r_i^2(\beta)),\\
S''_n(\beta) &= 2\sum_{i=1}^{n} \{h'_\beta (h'_\beta)^T - r_i(\beta) h''_{\beta\beta}\} I_2(r_i^2(\beta))
\end{aligned}
$$

almost surely at any $\beta \in \mathcal{B}$ *and* $\beta \in U(\beta_0, \delta)$, *respectively.*

The proof of this lemma is substantially the same as Čížek's, so we omit it here.

Next we establish the consistency and asymptotic normality of $\beta_n^{(LTS,h_n)}$ in three stages using Čížek's idea of asymptotic linearity, which we establish first in Proposition 1.

To investigate the behavior of the normal equations $S'_n(\beta) = 0$ around β_0 as a function of $\beta - \beta_0$, consider the difference

$$
\begin{aligned}
D_n^1(t) &= S'_n(\beta_0 - n^{-1/2}t) - S'_n(\beta_0)\\
&= -2\sum_{i=1}^{n} \Big[\{y_i - h(x_i, \beta_0 - n^{-1/2}t)\} h'_\beta(x_i, \beta_0 - n^{-1/2}t) I_2(r_i^2(\beta_0 - n^{-1/2}t))\\
&\qquad - \{y_i - h(x_i,\beta_0)\} h'_\beta(x_i,\beta_0) I_2(r_i^2(\beta_0)) \Big].
\end{aligned}
$$

Here, $t \in \mathcal{T}_M = \{t \in \mathcal{R}^p | \|t\| \le M\}$, where $0 < M < \infty$ is an arbitrary but fixed constant.

Proposition 1 (Asymptotic Linearity)**.** *Under assumptions TD, TH and TI, and for* $\lambda \in (1/2, 1]$ *and* $M > 0$, *we have*

$$
n^{-1/2} \sup_{t \in \mathcal{T}_M} \| \tfrac{D_n^1(t)}{-2} - n^{1/2} Q_h t C_\lambda \| = o_p(1), \text{ as } n \to \infty,
$$

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

Now we can state our two results on consistency and asymptotic normality.

Theorem 5 (Consistency)**.** *Under assumptions TD, TH*1*, TH*5 *and TI, the two-sided LTS estimator* $\beta_n^{(LTS,h_n)}$ *minimizing (9) is weakly consistent, i.e.*

$$\beta_n^{(LTS,h_n)} \xrightarrow{p} \beta_0, \text{ as } n \to \infty.$$

In addition, if all conditions of H are satisfied. Then $\beta_n^{(LTS,h_n)}$ *is* $\sqrt{n}$*-consistent, i.e.*

$$\sqrt{n}(\beta_n^{(LTS,h_n)} - \beta_0) = O_p(1), \text{ as } n \to \infty.$$

Theorem 6 (Asymptotic Normality)**.** *Suppose that assumptions TD, TH and TI are satisfied and* $C_\lambda \neq 0$*, then we have*

$$\sqrt{n}(\beta_n^{(LTS,h_n)} - \beta_0) \xrightarrow{\mathcal{L}} N(0, V_{2\lambda}),$$

where $V_{2\lambda} = (C_\lambda)^{-2}\sigma_{2\lambda}^2 Q_h^{-1}$, C_λ *and* Q_h^{-1} *are defined in Proposition 1 and* $\sigma_{2\lambda}^2 = Eu_i^2 I_{[G^{-1}(1-\lambda),\leq G^{-1}(\lambda)]}(u_i^2)$.

The proofs of these results are substantially the same as the proofs in [6, 7] for one-sided LTS estimators, so we omit the details. We only extend the required lemmas and propositions for Čížek's one-sided LTS estimator to our two-sided situation. Since the objective function giving the two-sided LTS estimator is not differentiable, we consider the behavior of the ordered residual statistics (Lemmas 5 and 6). Given this, the proof of the asymptotic linearity of the corresponding LTS normal equations as stated in Proposition 1 can be given. Then combining these results with the uniform law of large numbers (Lemma 4) and stochastic equicontinuity for mixing processes, we can prove the consistency and rate of convergence of the two-sided LTS estimates (Theorem 5). Finally, using Proposition 2 below, the proof of the asymptotic normality of the two-sided LTS estimate (Theorem 6) will follow from the consistency and asymptotic linearity of the LTS normal equations.

Now we can begin giving the formal proofs of Proposition 1, and Theorems 5 and 6. We start with Lemma 4.

Lemma 4 (Uniform weak law of large numbers)**.** *Let assumptions TD, TH and TI*1 *hold, and assume that* $t(x,u;\beta)$ *is a real function continuous in* β *uniformly in x and u over any compact subset of the support of* (x,u)*. Also, we suppose that* $E\sup_{\beta\in\mathcal{B}}|t(x,u;\beta)|^{1+\delta} < \infty$*, for some* $\delta > 0$*. Then, letting* $I_3(\beta;K_1,K_2) = I_{[G_\beta^{-1}(1-\lambda)-K_1, G_\beta^{-1}(\lambda)+K_2]}(r_i^2(\beta))$*, we have*

$$\sup_{\beta\in\mathcal{B},K_1,K_2\in\mathcal{R}}\left|\frac{1}{n}\sum_{i=1}^{n}[t(x_i,u_i;\beta)I_3(\beta;K_1,K_2)] - E[t(x_i,u_i;\beta)I_3(\beta;K_1,K_2)]\right| \to 0,$$

as $n \to \infty$ *in probability.*

The proof is in Appendix B.1.

Next, note that Čížek's Lemmas A.2 - A.5 are valid for $\lambda \in (0,1)$ with our assumption TD in place of his assumption D. So we only state these lemmas here without proofs. Denote the i^{th} order statistics of the squared residuals $r_i^2(\beta) = (y_i - h(x_i,\beta))^2$ by $r_{[i]}^2(\beta)$ used to define the two-sided LTS estimator in (9). Thus, we have the following.

Lemma 5. *For $\lambda \in (0,1)$ and $h_n = [\lambda n]$ for $n \in \mathcal{N}$, under assumptions TD, $TH1$ and $TI1$, we have*

$$\sup_{\beta\in\mathcal{B}}\left|r_{[h_n]}^2(\beta) - G_\beta^{-1}(\lambda)\right| \to 0, \tag{11}$$

as $n \to \infty$ in probability. Moreover,

$$E_{G_n} = E\sup_{\beta\in\mathcal{B}}\left|r_{[h_n]}^2(\beta) - G_\beta^{-1}(\lambda)\right| \to 0, \tag{12}$$

as $n \to \infty$.

Lemma 6. *For $\lambda \in (0,1)$ and $h_n = [\lambda n]$ for $n \in \mathcal{N}$, under assumptions TD, $TH1$ and $TI1$, there exist some $\epsilon > 0$ such that*

$$\sqrt{n}\sup_{\beta\in U(\beta_0,\epsilon)}\left|r_{[h_n]}^2(\beta) - G_\beta^{-1}(\lambda)\right| = O_p(1)$$

and

$$E_{L_n} = E\Big\{\sqrt{n}\sup_{\beta\in U(\beta_0,\epsilon)}\left|r_{[h_n]}^2(\beta) - G_\beta^{-1}(\lambda)\right|\Big\} = O_p(1),$$

as $n \to \infty$.

Lemma 7. *Let assumptions TD, TH and $TI1$ hold, and suppose that $\lambda \in (0,1), \tau \in (1/2,1)$, and $h_n = [\lambda n]$ for $n \in \mathcal{N}$. Then, we have*

$$\left|r_{[h_n]}^2(\beta_0 - n^{-1/2}t) - r_{[h_n]}^2(\beta_0)\right| = O_p(n^{-\tau})$$

uniformly in $t \in \mathcal{T}_M = \{t \in \mathcal{R}^k : \|t\| \le M\}$ as $n \to \infty$.

Lemma 8. *Under assumptions $TD, TH1$ and $TI1$, we have that for any $i \le n$ and $\lambda \in (0,1)$,*

$$P_G^0 = P\Big(\sup_{\beta\in\mathcal{B}}\left|I_{\{r_i^2(\beta)\le r_{[h_n](\beta)}^2\}} - I_{\{r_i^2(\beta)\le G_\beta^{-1}(\lambda)\}}\right| \ne 0\Big) = o(1).$$

In addition, under assumptions TD, TH and $TI1$, there exists $\epsilon > 0$ such that

$$P_L^0 = P\Big(\sup_{\beta\in U(\beta_0,\epsilon)}\left|I_{\{r_i^2(\beta)\le r_{[h_n](\beta)}^2\}} - I_{\{r_i^2(\beta)\le G_\beta^{-1}(\lambda)\}}\right| \ne 0\Big) = O(n^{-1/2}).$$

as $n \to \infty$.

By Lemma 8, we have the following result.

Corollary 2. *Under assumptions $TD, TH1$ and $TI1$. For $\lambda \in (1/2, 1)$ and for any $i \leq n$, we have*

$$\begin{aligned} P_G &= P\Big(\sup_{\beta\in\mathcal{B}}\Big|I_{\{r^2_{[n-h_n+1](\beta)}\leq r_i^2(\beta)\leq r^2_{[h_n](\beta)}\}} - I_{\{G_\beta^{-1}(1-\lambda)\leq r_i^2(\beta)\leq G_\beta^{-1}(\lambda)\}}\Big| \neq 0\Big) \\ &= o(1). \end{aligned}$$

In addition, under assumptions TD, TH and $TI1$, there exists $\epsilon > 0$ such that

$$\begin{aligned} P_L &= P\Big(\sup_{\beta\in U(\beta_0,\epsilon)}\Big|I_{\{r^2_{[n-h_n+1](\beta)}\leq r_i^2(\beta)\leq r^2_{[h_n](\beta)}\}} - I_{\{G_\beta^{-1}(1-\lambda)\leq r_i^2(\beta)\leq G_\beta^{-1}(\lambda)\}}\Big| \neq 0\Big) \\ &= O(n^{-1/2}), \ \textit{as } n \to \infty. \end{aligned}$$

Proof of Corollary 2. Denote $A_1 = \{r_i^2(\beta) \leq r^2_{[h_n]}(\beta)\}$, $B_1 = \{r_i^2(\beta) \geq r^2_{[n-h_n+1]}(\beta)\}$, $A_2 = \{r_i^2(\beta) \leq G_\beta^{-1}(\lambda)\}$ and $B_2 = \{r_i^2(\beta) \geq G_\beta^{-1}(1-\lambda)\}$. Let $v_{in}(\beta) = I_{\{r^2_{[n-h_n+1]}(\beta)\leq r_i^2(\beta)\leq r^2_{[h_n]}(\beta)\}} - I_{\{G_\beta^{-1}(1-\lambda)\leq r_i^2(\beta)\leq G_\beta^{-1}(\lambda)\}}$. Thus,

$$v_{in}(\beta) = I_{A_1}I_{B_1} - I_{A_2}I_{B_2}.$$

So we have

$$\begin{aligned} 0 \leq \sup_{\beta\in\mathcal{B}}\Big|v_{in}(\beta)\Big| &= \sup_{\beta\in\mathcal{B}}\Big|I_{A_1}I_{B_1} - I_{A_2}I_{B_1} + I_{A_2}I_{B_1} - I_{A_2}I_{B_2}\Big| \\ &\leq \sup_{\beta\in\mathcal{B}}\Big|I_{A_1} - I_{A_2}\Big|\Big|I_{B_1}\Big| + \sup_{\beta\in\mathcal{B}}\Big|I_{A_2}\Big|\Big|I_{B_1} - I_{B_2}\Big| \\ &\leq \sup_{\beta\in\mathcal{B}}\Big|I_{A_1} - I_{A_2}\Big| + \sup_{\beta\in\mathcal{B}}\Big|I_{B_1} - I_{B_2}\Big|. \end{aligned}$$

Notice that $\sup_{\beta\in\mathcal{B}}\Big|v_{in}(\beta)\Big| \neq 0$ implies that either $\sup_{\beta\in\mathcal{B}}\Big|I_{A_1} - I_{A_2}\Big| \neq 0$ or $\sup_{\beta\in\mathcal{B}}\Big|I_{B_1} - I_{B_2}\Big| \neq 0$. Thus, we have

$$0 \leq P(\sup_{\beta\in\mathcal{B}}\Big|v_{in}(\beta)\Big| \neq 0) \leq P(\sup_{\beta\in\mathcal{B}}\Big|I_{A_1} - I_{A_2}\Big| \neq 0) + P(\sup_{\beta\in\mathcal{B}}\Big|I_{B_1} - I_{B_2}\Big| \neq 0) = o(1).$$

The second last equality holds by the first result of Lemma 8. Similarly, using the above arguments with the second result of Lemma 8, we can prove that there exists $\epsilon > 0$ such that

$$P(\sup_{\beta\in U(\beta_0,\epsilon)}\Big|v_{in}(\beta)\Big| \neq 0) = O(n^{-1/2}).$$

□

Using the same technique as in the proof of Corollary 2, we have the following which is parallel to Čížek's Corollary A.6.

Proposition 2. *Let assumptions $TD, TH1$ and $TI1$ hold and assume that $t(x, u; \beta)$ is a real-valued function continuous in β uniformly in x and u over any compact subset of the support of (x, u). Moreover, assume*

that $E\sup_{\beta\in B}\Big|t(x,u;\beta)\Big| < \infty$. *Then we have that for* $\lambda \in (1/2,1)$,

$$\begin{aligned}
&E\Big\{\sup_{\beta\in\mathcal{B}}\Big|t(x_i,u_i;\beta)\Big[I_{\{r^2_{[n-h_n+1](\beta)}\le r_i^2(\beta)\le r^2_{[h_n](\beta)}\}} - I_{\{G_\beta^{-1}(1-\lambda)\le r_i^2(\beta)\le G_\beta^{-1}(\lambda)\}}\Big]\Big|\Big\}\\
=&\,o(1).
\end{aligned}$$

In addition, under assumptions TD, TH *and* $TI1$, *there exists* $\epsilon > 0$ *such that*

$$\begin{aligned}
&E\Big\{\sup_{\beta\in U(\beta_0,\epsilon)}\Big|t(x_i,u_i;\beta)\Big[I_{\{r^2_{[n-h_n+1](\beta)}\le r_i^2(\beta)\le r^2_{[h_n](\beta)}\}} - I_{\{G_\beta^{-1}(1-\lambda)\le r_i^2(\beta)\le G_\beta^{-1}(\lambda)\}}\Big]\Big|\Big\}\\
=&\,O(n^{-1/2}),\ \textit{as } n\to\infty.
\end{aligned}$$

Proposition 2 controls the upper bound arising from applying Chebyshev's inequality to a weighted sum of differences of indicator functions. This sum of differences expresses the distance between residuals and their limiting quantiles. It is stated in the following.

Proposition 3. *Let assumptions* $TD, TH1$ *and* $TI1$ *hold and assume that* $t(x,u;\beta)$ *is a real-valued function continuous in* β *uniformly in* x *and* u *over any compact subset of the support of* (x,u). *Moreover, assume that* $E\sup_{\beta\in\mathcal{B}}\Big|t(x,u;\beta)\Big| < \infty$. *Then we have that for* $\lambda\in(1/2,1)$,

$$\begin{aligned}
&\sup_{\beta\in\mathcal{B}}\Big|\frac{1}{n}\sum_{i=1}^n\Big\{t(x_i,u_i;\beta)\Big[I_{\{r^2_{[n-h_n+1](\beta)}\le r_i^2(\beta)\le r^2_{[h_n](\beta)}\}} - I_{\{G_\beta^{-1}(1-\lambda)\le r_i^2(\beta)\le G_\beta^{-1}(\lambda)\}}\Big]\Big\}\Big|\\
=&\,o_p(1).
\end{aligned}$$

In addition, under assumptions TD, TH *and* $TI1$, *there exists* $\epsilon > 0$ *such that*

$$\begin{aligned}
&\sup_{\beta\in U(\beta_0,\epsilon)}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n\Big\{t(x_i,u_i;\beta)\Big[I_{\{r^2_{[n-h_n+1](\beta)}\le r_i^2(\beta)\le r^2_{[h_n](\beta)}\}} - I_{\{G_\beta^{-1}(1-\lambda)\le r_i^2(\beta)\le G_\beta^{-1}(\lambda)\}}\Big]\Big\}\Big|\\
=&\,O_p(1),\ \textit{as } n\to\infty.
\end{aligned}$$

Proof of Proposition 3. Recall that $A_1 = \{r_i^2(\beta)\le r^2_{[h_n]}(\beta)\}$, $B_1 = \{r_i^2(\beta)\ge r^2_{[n-h_n+1]}(\beta)\}$, $A_2 = \{r_i^2(\beta)\le G_\beta^{-1}(\lambda)\}$ and $B_2 = \{r_i^2(\beta)\ge G_\beta^{-1}(1-\lambda)\}$. By the first result of Proposition 2, for any $\epsilon^* > 0$, we have

$$\begin{aligned}
&P\Big(\sup_{\beta\in\mathcal{B}}\Big|\frac{1}{n}\sum_{i=1}^n\Big\{t(x_i,u_i;\beta)\Big[I_{A_1}I_{B_1} - I_{A_2}I_{B_2}\Big]\Big\}\Big| > \epsilon^*\Big)\\
\le&\frac{1}{\epsilon^*}E\Big(\sup_{\beta\in\mathcal{B}}\Big|\frac{1}{n}\sum_{i=1}^n\Big\{t(x_i,u_i;\beta)\Big[I_{A_1}I_{B_1} - I_{A_2}I_{B_2}\Big]\Big\}\Big|\Big)\\
\le&\frac{1}{\epsilon^*}E\Big(\sup_{\beta\in\mathcal{B}}\frac{1}{n}\sum_{i=1}^n\Big|t(x_i,u_i;\beta)\Big[I_{A_1}I_{B_1} - I_{A_2}I_{B_2}\Big]\Big|\Big)\\
=&\frac{1}{\epsilon^*}E\Big(\sup_{\beta\in\mathcal{B}}\Big|t(x_i,u_i;\beta)\Big[I_{A_1}I_{B_1} - I_{A_2}I_{B_2}\Big]\Big|\Big)\to 0.
\end{aligned}$$

Moreover, by the second result of Proposition 2, there exists $\epsilon > 0$ such that

$$E\Big(\sup_{\beta\in U(\beta_0,\epsilon)}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{t(x_i,u_i;\beta)\Big[I_{A_1}I_{B_1}-I_{A_2}I_{B_2}\Big]\Big\}\Big|\Big)$$
$$=\sqrt{n}E\Big(\sup_{\beta\in U(\beta_0,\epsilon)}\Big|\frac{1}{n}\sum_{i=1}^{n}\Big\{t(x_i,u_i;\beta)\Big[I_{A_1}I_{B_1}-I_{A_2}I_{B_2}\Big]\Big\}\Big|\Big)\le O(1).$$

Therefore, using the Chebyshev's inequality again gives the second result. □

In what follows, we study in more detail the differences of probabilities that $I_{\{r^2_{[n-h_n+1](\beta)}\le r_i^2(\beta)\le r^2_{[h_n](\beta)}\}}$ at $\beta=\beta_0$ and β_n for sequences β_n converging to β_0 at $\sqrt{n}$-rate. Our next result gives bounds for how closely residuals at the true parameter β_0 approximate residuals at β in a neighborhood of β_0.

Lemma 9. *Recall that $A_1=\{r_i^2(\beta)\le r^2_{[h_n]}(\beta)\}$ and $B_1=\{r_i^2(\beta)\ge r^2_{[n-h_n+1]}(\beta)\}$. Denote $A_1^0=\{r_i^2(\beta_0)\le r^2_{[h_n]}(\beta_0)\}$ and $B_1^0=\{r_i^2(\beta_0)\ge r^2_{[n-h_n+1]}(\beta_0)\}$. Let assumptions D^* and H hold and $\beta\in U(\beta_0,n^{-1/2}M)$ for some $M>0$. Then for $\lambda\in(1/2,1)$, we have, as $n\to\infty$,*

1. *For the conditional probability*

   (a) $P\Big(I_{A_1^0}I_{B_1^0}\ne I_{A_1}I_{B_1}\Big|x_i\Big)=\Big|(h'_\beta(x_i,\beta_0))^T(\beta-\beta_0)\Big|[H(\lambda)+H(1-\lambda)]+O_p(n^{-1/2})=O_p(n^{-1/4})$, *and*

   (b) $E\Big\{sgn\, r_i(\beta_0)\Big(I_{A_1^0}I_{B_1^0}-I_{A_1}I_{B_1}\Big)\Big|x_i\Big\}=(h'_\beta(x_i,\beta_0))^T(\beta-\beta_0)[H(\lambda)-H(1-\lambda)]+O_p(n^{-1/2})$.

2. *For the corresponding unconditional probability*

$$P\Big(I_{A_1^0}I_{B_1^0}\ne I_{A_1}I_{B_1}\Big)$$
$$=E_X\Big|(h'_\beta(x_i,\beta_0))^T(\beta-\beta_0)\Big|[H(\lambda)+H(1-\lambda)]+O(n^{-1/2})=O(n^{-1/2}).$$

3. *For the conditional probability taken over all $\beta\in U(\beta_0,n^{-1/2}M)$*

$$P\Big(\exists\beta\in U(\beta_0,n^{-1/2}M):I_{A_1^0}I_{B_1^0}\ne I_{A_1}I_{B_1}\Big|x_i\Big)$$
$$=n^{-1/2}M\sum_{j=1}^{p}\Big|h'_{\beta_j}(x_i,\beta_0)\Big|[H(\lambda)+H(1-\lambda)]+O_p(n^{-1/2})$$
$$=O_p(n^{-1/4})$$

4. *For the corresponding unconditional probability taken over all $\beta\in U(\beta_0,n^{-1/2}M)$,*

$$P\Big(\exists\beta\in U(\beta_0,n^{-1/2}M):I_{A_1^0}I_{B_1^0}\ne I_{A_1}I_{B_1}\Big)$$
$$=n^{-1/2}M\sum_{j=1}^{p}E_X\Big|h'_{\beta_j}(x_i,\beta_0)\Big|[H(\lambda)+H(1-\lambda)]+O_p(n^{-1/2})$$
$$=O_p(n^{-1/2})$$

where $H(\lambda) = f(q_\lambda) + f(-q_\lambda)$ *and* $q_\lambda = \sqrt{G^{-1}(\lambda)}$.

Proof of Lemma 9. Note that Čížek's Lemmas A.8 holds for $\lambda \in (0,1)$. First, the result of 1(a) holds because

$$\begin{aligned}
&P(I_{A_1^0} I_{B_1^0} \neq I_{A_1} I_{B_1} | x_i) \\
\leq &P(I_{A_1^0} \neq I_{A_1} | x_i) + P(I_{B_1^0} \neq I_{B_1} | x_i) \\
= &\Big|(h'_\beta(x_i, \beta_0))^T (\beta - \beta_0)\Big| [f(q_\lambda) + f(-q_\lambda)] + O_p(n^{-1/2}) \\
&+ \Big|(h'_\beta(x_i, \beta_0))^T (\beta - \beta_0)\Big| [f(q_{1-\lambda}) + f(-q_{1-\lambda})] + O_p(n^{-1/2}) \\
= &\Big|(h'_\beta(x_i, \beta_0))^T (\beta - \beta_0)\Big| [H(\lambda) + H(1-\lambda)] + O_p(n^{-1/2}) = O_p(n^{-1/4}).
\end{aligned}$$

In addition, 1(b) can be obtained by using Čížek's result in his lemma A.8 with our 1(a), so we omit the proof here.

Second, for the corresponding unconditional probability,

$$\begin{aligned}
&P(I_{A_1^0} I_{B_1^0} \neq I_{A_1} I_{B_1}) \\
= &E_X P(I_{A_1^0} I_{B_1^0} \neq I_{A_1} I_{B_1} | x_i) \leq E_X \Big|(h'_\beta(x_i, \beta_0))^T (\beta - \beta_0)\Big| [H(\lambda) + H(1-\lambda)] \\
&+ O_p(n^{-1/2}).
\end{aligned}$$

Again the proof of the result is completed by using Čížek's result in his lemma A.8.

Third, for the conditional probability taken over all $\beta \in U(\beta_0, n^{-1/2}M)$,

$$\begin{aligned}
&P\Big(\exists \beta \in U(\beta_0, n^{-1/2}M) : I_{A_1^0} I_{B_1^0} \neq I_{A_1} I_{B_1} \Big| x_i\Big) \\
\leq &P\Big(\exists \beta \in U(\beta_0, n^{-1/2}M) : I_{A_1^0} \neq I_{A_1} \Big| x_i\Big) \\
&+ P\Big(\exists \beta \in U(\beta_0, n^{-1/2}M) : I_{B_1^0} \neq I_{B_1} \Big| x_i\Big) \\
\leq &n^{-1/2} M \sum_{j=1}^{p} \Big|h'_{\beta_j}(x_i, \beta_0)\Big| [H(\lambda) + H(1-\lambda)] + O_p(n^{-1/2}) \\
= &O_p(n^{-1/4}).
\end{aligned}$$

The fourth result can be obtained by using the same techniques as in our second and third results, so we omit the proof. □

Čížek's corollary A.9 controls the deviation of residuals in one tail from the Taylor approximation to h. Here, both tails must be controlled, as in the following.

Lemma 10. *Under the assumptions of Lemma 9, suppose that there exists some* $\beta \in U(\beta_0, n^{-1/2}M)$ *such that* $I_{A_1^0} \neq I_{A_1}$ *and* $I_{B_1^0} \neq I_{B_1}$. *Then*

$$\begin{aligned}
&\max\Big\{\Big||r_i(\beta)| - q_\lambda\Big|, \Big||r_i(\beta)| - q_{1-\lambda}\Big|\Big\} \\
\leq &\Big|(h'_\beta(x_i, \xi))^T (\beta - \beta_0)\Big| + O_p(n^{-1/2}) = O_p(n^{-1/4}).
\end{aligned}$$

and

$$\begin{aligned}&\max\Big\{E\Big\{\Big||r_i(\beta)|-q_\lambda\Big|\Big|x_i\Big\},E\Big\{\Big||r_i(\beta)|-q_{1-\lambda}\Big|\Big|x_i\Big\}\Big\}\\ \le&\Big|(h'_\beta(x_i,\xi))^T(\beta-\beta_0)\Big|+O_p(n^{-1/2}),\end{aligned}$$

where $\xi\in(\beta_0,\beta)$.

This lemma is a direct consequence of Čížek's Corollary A.9.

The proofs of Theorems 5 and 6 can be obtained using Čížek's proofs of his theorems 4.1, 4.2 and 4.3 but with our lemmas and propositions for the two-sided LTS estimators. For the sake of completeness, we prove Theorem 6 using Theorem 5 and Proposition 1, because Theorem 6 is the most directly useful in practice.

Proof of Theorem 6 : From Theorem 5, we have $t_n=\sqrt{n}(\beta_0-\beta_n^{(LTS,h_n)})=O_p(1)$, as $n\to\infty$. Then using Proposition 1, with probability approaching to 1, we have

$$\begin{aligned}&n^{-1/2}\Big(\frac{D'_n(t_n)}{-2}-n^{1/2}Q_ht_nC_\lambda\Big)\\ =&n^{-1/2}\Big(\frac{D'_n(\sqrt{n}(\beta_0-\beta_n^{(LTS,h_n)}))}{-2}+n^{1/2}Q_hC_\lambda\sqrt{n}(\beta_n^{(LTS,h_n)}-\beta_0)\Big)\\ =&o_p(1),\end{aligned}$$

where $C_\lambda=(2\lambda-1)+(\frac{q_\lambda+q_{1-\lambda}}{2})[H(\lambda)-H(1-\lambda)]$, $H(\lambda)=f(q_\lambda)+f(-q_\lambda)$ and $q_\lambda=\sqrt{G^{-1}(\lambda)}$.

Then by simple algebra with the definition of $\beta_n^{(LTS,h_n)}$, we have

$$\begin{aligned}&\sqrt{n}(\beta_n^{(LTS,h_n)}-\beta_0)\\ =\;& n^{-1/2}Q_h^{-1}C_\lambda^{-1}\sum_{i=1}^n\{r_i(\beta_0)\}h'_\beta(x_i,\beta_0)I_2^G(\beta_0)+o_p(1) &(13)\\ &+n^{-1/2}Q_h^{-1}C_\lambda^{-1}\sum_{i=1}^n\{r_i(\beta_0)\}h'_\beta(x_i,\beta_0)[I_2(\beta_0)-I_2^G(\beta_0)]. &(14)\end{aligned}$$

First, we show that (14) is negligible in probability. Recall that $r_i(\beta_0)\overset{def}{=}u_i$. Thus, (14) can be rewritten as

$$n^{-1/2}Q_h^{-1}C_\lambda^{-1}\sum_{i=1}^n u_ih'_\beta(x_i,\beta_0)[I_{\{u^2_{[n-h_n+1]}\le u_i^2\le u^2_{[h_n]}\}}-I_{\{G^{-1}(1-\lambda)\le u_i^2\le G^{-1}(\lambda)\}}].$$

Then our Proposition 2 and assumption $TD2$ imply that, for $k=1$ and 2,

$$E\Big|u_i[I_{\{u^2_{[n-h_n+1]}\le u_i^2\le u^2_{[h_n]}\}}-I_{\{G^{-1}(1-\lambda)\le u_i^2\le G^{-1}(\lambda)\}}]\Big|^k=O(n^{-1/2}), \tag{15}$$

as $n\to\infty$. Therefore, the summands in (14) multiplied by $n^{1/4}$ have a finite expectation

$$E\Big|n^{1/4}u_ih'_\beta(x_i,\beta_0)[I_{\{u^2_{[n-h_n+1]}\le u_i^2\le u^2_{[h_n]}\}}-I_{\{G^{-1}(1-\lambda)\le u_i^2\le G^{-1}(\lambda)\}}]\Big|=o(1),$$

and variance

$$
\begin{aligned}
& var\Big\{n^{1/4}u_i h'_\beta(x_i,\beta_0)[I_{\{u^2_{[n-h_n+1]}\le u_i^2\le u^2_{[h_n]}\}} - I_{\{G^{-1}(1-\lambda)\le u_i^2\le G^{-1}(\lambda)\}}]\Big\} \\
\le\ & n^{1/2}E_{X_i}\Big\{h'_\beta(x_i,\beta_0)\cdot var(u_i\cdot\Big|I_{\{u^2_{[n-h_n+1]}\le u_i^2\le u^2_{[h_n]}\}} \\
& - I_{\{G^{-1}(1-\lambda)\le u_i^2\le G^{-1}(\lambda)\}}\Big||X_i)\cdot h'_\beta(x_i,\beta_0)^T\Big\} \\
& + n^{1/2}var_{X_i}\Big\{h'_\beta(x_i,\beta_0)\cdot E(u_i\cdot\Big|I_{\{u^2_{[n-h_n+1]}\le u_i^2\le u^2_{[h_n]}\}} \\
& - I_{\{G^{-1}(1-\lambda)\le u_i^2\le G^{-1}(\lambda)\}}\Big||X_i)\Big\} \\
\le\ & O(1)\Big\{E_{X_i}\{h'_\beta(x_i,\beta_0)h'_\beta(x_i,\beta_0)^T\} + var_{X_i}(h'_\beta(x_i,\beta_0))\Big\} \\
=\ & O(1).
\end{aligned}
$$

by assumption TH5 and the independence of x_i and u_i.

Now since all indicators depend only on the squares of the residual u_i^2 and the error terms u_i are symmetrically distributed by assumption TD2, we have that, for any $i = 1, 2, \ldots, n$ and any $n \in \mathcal{N}$,

$$
E\Big\{n^{1/4}u_i h'_\beta(x_i,\beta_0)[I_{\{u^2_{[n-h_n+1]}\le u_i^2\le u^2_{[h_n]}\}} - I_{\{G^{-1}(1-\lambda)\le u_i^2\le G^{-1}(\lambda)\}}]\Big\} = 0.
$$

In the condition case, we get

$$
\begin{aligned}
& E\Big\{n^{1/4}u_i h'_\beta(x_i,\beta_0)[I_{\{u^2_{[n-h_n+1]}\le u_i^2\le u^2_{[h_n]}\}} \\
& - I_{\{G^{-1}(1-\lambda)\le u_i^2\le G^{-1}(\lambda)\}}]\Big|u_1,\ldots,u_{i-1},x_1,\ldots,x_{i-1}\Big\} = 0.
\end{aligned}
$$

Therefore, similar to Čížek's one-sided case,

$$
n^{1/4}u_i h'_\beta(x_i,\beta_0)[I_{\{u^2_{[n-h_n+1]}\le u_i^2\le u^2_{[h_n]}\}} - I_{\{G^{-1}(1-\lambda)\le u_i^2\le G^{-1}(\lambda)\}}]
$$

forms a sequence of martingale differences with finite variances. Applying the law of large numbers for the sum of martingale differences (14), we have

$$
\begin{aligned}
& n^{-1/2}Q_h^{-1}C_\lambda^{-1}\sum_{i=1}^{n}u_i h'_\beta(x_i,\beta_0)\Big[I_{\{u^2_{[n-h_n+1]}\le u_i^2\le u^2_{[h_n]}\}} \\
& - I_{\{G^{-1}(1-\lambda)\le u_i^2\le G^{-1}(\lambda)\}}\Big] \xrightarrow{p} 0,
\end{aligned} \tag{16}
$$

as $n \to \infty$. Thus, (14) is negligible in probability $o_p(1)$. Based on this result, (14) gives

$$
\begin{aligned}
& \sqrt{n}(\beta_n^{(LTS,h_n)} - \beta_0) \\
=\ & n^{-1/2}Q_h^{-1}C_\lambda^{-1}\sum_{i=1}^{n}r_i(\beta_0)h'_\beta(x_i,\beta_0)I_{\{G^{-1}(1-\lambda)\le u_i^2\le G^{-1}(\lambda)\}} + o_p(1).
\end{aligned} \tag{17}
$$

Additionally, using the same arguments as for (14), the summands in (17) form a sequence of identically distributed martingale differences with finite second moments by the assumptions $TD2$ and $TH5$. Then, by the law of large numbers for L^1-mixingales in [2], we have

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n} u_i^2 h'_\beta(x_i,\beta_0)h'_\beta(x_i,\beta_0)^T \cdot I_{\{G^{-1}(1-\lambda)\le u_i^2\le G^{-1}(\lambda)\}} \\ \xrightarrow{p} \quad & var(u_i h'_\beta(x_i,\beta_0)) \cdot I_{\{G^{-1}(1-\lambda)\le u_i^2\le G^{-1}(\lambda)\}}, \end{aligned}$$

as $n \to \infty$. Therefore, the proof of Theorem 8 for the asymptotic normality of the two-sided LTS estimator $\beta_n^{(LTS,h_n)}$ is completed by the central limit theorem for the martingale differences in (17) with the asymptotic variance

$$\begin{aligned} V_{2\lambda} &= C_\lambda^{-2} \cdot Q_h^{-1} \cdot var[u_i h'_\beta(x_i,\beta_0) I_{\{G^{-1}(1-\lambda)\le u_i^2\le G^{-1}(\lambda)\}}] \cdot Q_h^{-1} \\ &= C_\lambda^{-2} \cdot Q_h^{-1} \cdot E\Big\{ [h'_\beta(x_i,\beta_0)u_i I_{\{G^{-1}(1-\lambda)\le u_i^2\le G^{-1}(\lambda)\}}] \\ &\qquad \times [h'_\beta(x_i,\beta_0)u_i I_{\{G^{-1}(1-\lambda)\le u_i^2\le G^{-1}(\lambda)\}}]^T \Big\} \cdot Q_h^{-1} \\ &= C_\lambda^{-2} \cdot Q_h^{-1} \cdot E[h'_\beta(x_i,\beta_0)h'_\beta(x_i,\beta_0)^T] \cdot E[u_i^2 I_{\{G^{-1}(1-\lambda)\le u_i^2\le G^{-1}(\lambda)\}}] \cdot Q_h^{-1} \\ &= C_\lambda^{-2} \cdot Q_h^{-1} \cdot Q_h \cdot \sigma_{2\lambda}^2 \cdot Q_h^{-1} \\ &= C_\lambda^{-2} \cdot \sigma_{2\lambda}^2 \cdot Q_h^{-1}, \end{aligned}$$

where $\sigma_{2\lambda}^2 = E[u_i^2 I_{\{G^{-1}(1-\lambda)\le u_i^2\le G^{-1}(\lambda)\}}]$ and $\lambda \in (1/2, 1)$. □

5 Summary

In place of the conventional expected-loss-based estimators, we used the median of the loss to define a new estimator in the Bayesian and Frequentist contexts. In this paper, the Bayesian *medloss* estimator is shown to have an optimal rate of convergence and asymptotic normality, as in the conventional expected loss case. However, using the median has permitted weaker assumptions. For example, we do not require any moment conditions.

In the Frequentist context, we have also established asymptotic results for the LMS and the two-sided LTS estimators in nonlinear regression models. The former is the Frequentist version of our *medloss* estimator. However, like the linear situation, our LMS estimator only has a cube-root convergence rate. On the other hand, the LMS estimators can be regarded as a limiting case of the LTS estimators with 50% trimming on each side. If any fixed amount of trimming strictly less than 50% on both sides is used, the asymptotic rate increases from $n^{1/3}$ to $\sqrt{n}$ in which case the usual consistency and asymptotic normality can be proved, although efficiency fails.

Taken together these three results demonstrate that, in effect, the Bayesian approach averages over a small region around the LTS estimator to give an estimator close enough to the LMS estimator that the

$\sqrt{n}$-rate and efficiency are obtained. That is, the Bayesian *medloss* estimator is a good tradeoff between using the actual median and using an arbitrary trimming proportion below 50%.

APPENDIX

A Detailed proof for the LMS estimator

In this appendix, we are going to verify that the LMS estimator in nonlinear situations satisfies the conditions of our Theorem 3. Before verifying these conditions in A.4, we need results from A.1-A.3. Finally, we prove the asymptotic results for LMS estimators in A.5. Our method here requires that we first obtain Lemma 4.1 in [10] since it is required for the detailed verification of Theorem 3 here.

A.1 Manageability

Manageability, proposed by Pollard [13], is a notion used to establish an $n^{-1/3}$ rate of convergence for the LMS estimators, and to verify the stochastic equicontinuity conditions for showing the limiting behavior of the LMS estimators in linear models [10].

As explained in [13], the concept of manageability formalizes the idea that maximal inequalities for the maximum deviation of a sum of independent stochastic processes from its expected value can be derived from uniform bounds on the random packing numbers.

Formally, let $\mathcal{F}_{n\omega} = \{(f_1(\omega,t),\ldots,f_n(\omega,t)) : t \in T\}$, and define the packing number $D(\epsilon,\mathcal{F})$ for a subset $\mathcal{F}$ of a metric space with metric d as the largest m for which there exist points $t_1,\ldots,t_m$ in $\mathcal{F}$ with $d(t_i,t_j) > \epsilon$ for $i \neq j$. Also, for each $\alpha = (\alpha_1,\ldots,\alpha_n)$ of nonnegative constants, and each $f = (f_1,\ldots,f_n) \in \mathcal{R}^n$, the pointwise product $\alpha \odot f$ is the vector in $\mathcal{R}^n$ with i^{th} coordinate $\alpha_i f_i$, and $\alpha \odot \mathcal{F}$ is the set of all vectors $\alpha \odot f$ with $f \in \mathcal{F}$.

Then following Pollard [13], a triangular array of random processes $\{f_{ni}(\omega,t) : t \in T, 1 \leq i \leq k_n\}$ is *manageable*, with respect to the envelopes $F_n(\omega)$, for $n = 1,2,\ldots$, if there exists a deterministic function λ, for which

- $\int_0^1 \sqrt{\ln \lambda(x)}dx < \infty$, and
- the random packing number $D(x|\alpha \odot F_n(\omega)|, \alpha \odot \mathcal{F}_{n\omega}) \leq \lambda(x)$ for $0 < x \leq 1$, all ω, all vectors α of nonnegative weights, and all n.

A sequence of processes $\{f_i\}$ is *manageable* if the array defined by $f_{ni} = f_i$ for $i \leq n$ is manageable.

The concept of manageability extends to a definition of uniform manageability based on the maximal inequality. Among those classes of functions which are manageable, those that are also uniformly manageable

satisfy the extra condition that the bound in the maximal inequality is independent of R used in the envelope G_R. See Kim and Pollard [10] for details.

A.1.1 Manageability of the class of functions $f_{h,x,y}(\theta, r)$ and $g_{h,x,u}(\theta, \delta, \xi)$

By the sufficient conditions for manageability [10], we can easily show that the classes of functions $f_{h,x,y}(\theta, r)$ and $g_{h,x,u}(\theta, \delta, \xi)$ for nonlinear models are also manageable.

Lemma 11 (Dudley, [8]). *If G is an m-dimension vector space of real functions on a set, then*

$$VC(\mathcal{C}_g) = dim(G) + 1,$$

where $\mathcal{C}_g = \{x \in X : g(x) \geq 0,\ g \in G\}$ and $VC(\mathcal{C}_g)$ means the VC dimension of $\mathcal{C}_g$.

TO use this result, suppose $\mathcal{G}_1$ and $\mathcal{G}_2$ are the classes $g_1(\theta, r) = h(x, \theta + \beta_0) - y + r$ and $g_2(\theta, r) = y + r - h(x, \theta + \beta_0)$ for any $h \in \mathcal{H}$, respectively. Consider

$$\mathcal{C}_1 = \{(\theta, r) \in \mathcal{R}^{d+1} : 0 \leq g_1(\theta, r),\ g_1 \in \mathcal{G}_1\}$$

and

$$\mathcal{C}_2 = \{(\theta, r) \in \mathcal{R}^{d+1} : 0 \leq g_2(\theta, r),\ g_2 \in \mathcal{G}_2\}.$$

Therefore, by Dudley's lemma and our assumption 1 in Theorem 6, the VC dimensions of $\mathcal{C}_1$ and $\mathcal{C}_2$ are bounded above by $\dim(\mathcal{H})+3 < \infty$. So, $\mathcal{C}_1$ and $\mathcal{C}_2$ form VC-classes, which implies that $\mathcal{C}_1 \cap \mathcal{C}_2$ is also a VC class. Now, the class of functions $f_{h,x,u}(\theta, r, \xi)$ or $f_{h,x,y}(\theta, r)$ forms a VC-subgraph, and hence is manageable.

Recall that

$$g_{h,x,u}(\theta, \delta, \xi) = f_{h,x,u}(\theta, 1 + \delta, \xi) - f_{h,x,u}(0, 1 + \delta, \xi).$$

Since the classes $\mathcal{F}$ and $\mathcal{F}_0$ of $f_{h,x,u}(\theta, r, \xi)$ and $f_{h,x,u}(0, r, \xi)$, respectively, are VC-subgraphs, the class

$$\mathcal{G} = \{f_1 - f_0 : f_1 \in \mathcal{F} \text{ and } f_0 \in \mathcal{F}_0\},$$

is also a VC-subgraph by Lemma 2.6.18 (vander Vaart and Wellner [22]). Thus, subclasses $\mathcal{G}_R$ of G as defined in Kim and Pollard [10] are uniformly manageable with the envelope

$$G_R^h = \sup_{\mathcal{G}_R} |g_{h,x,u}(\theta, r, \xi)|.$$

By the manageability of the class of $f_{h,x,y}(\theta, r)$ and Kim and Pollard's Corollary 3.2 in [10], we have

$$\sup_{\theta, r} |E_n f_{h,x,u}(\theta, r, \xi) - E f_{h,x,u}(\theta, r, \xi)| = O_p(n^{-1/2}). \tag{18}$$

A.2 $O_p(n^{-1/2})$ rate of convergence of r_n in (3.4)

Denote the distribution function of u by Γ. We have

$$\begin{aligned} Ef_{h,x,u}(\theta, r, \xi) &= E_x E_u^x[f_{h,x,u}(\theta, r, \xi)] \\ &= E_x[\Gamma(h'(X,\xi)^T\theta + r) - \Gamma(h'(X,\xi)^T\theta - r)], \end{aligned} \tag{19}$$

where E is the expectation with respect to the product probability measure $\mathcal{P}$ of (x, u), E_u^x means the condition expectation with respect to u given X and E_x is the unconditional expectation taken over X.

Clearly, (19) is a continuous function of θ and r, which is maximized by $\theta = 0$ for each fixed r because of the symmetry of u at 0. In other words, we have

$$\sup_\theta Ef_{h,x,u}(\theta, r, \xi) = \Gamma(r) - \Gamma(-r).$$

Thus, it follows that there exist positive constants k and λ for which

$$\sup_\theta Ef_{h,x,u}(\theta, 1-\delta, \xi) < 1/2 - k\delta \tag{20}$$

$$\text{and} \qquad Ef_{h,x,u}(\theta, 1+\delta, \xi) \geq 1/2 + \lambda\delta, \tag{21}$$

for any $\delta > 0$ small enough. Let $P[A, B]$ and $P_n[A, B]$ represent $EI_{[A,B]}$ and $E_n I_{[A,B]}$, which have probability measures P and P_n, respectively.

By (18), we have

$$\begin{aligned} \triangle_n &\stackrel{\text{def}}{=} \sup_{\theta,r} \left| P_n[h'(x,\xi)^T\theta - r, h'(x,\xi)^T\theta + r] - P[h'(x,\xi)^T\theta - r, h'(x,\xi)^T\theta + r] \right| \\ &= O_p(n^{-1/2}). \end{aligned}$$

Putting $r = 1 - \frac{\triangle_n}{k}$, we get

$$\begin{aligned} & P_n[h'(x,\xi)^T\theta - 1 + \frac{\triangle_n}{k}, h'(x,\xi)^T\theta + 1 - \frac{\triangle_n}{k}] \\ \leq & \triangle_n + P[h'(x,\xi)^T\theta - 1 + \frac{\triangle_n}{k}, h'(x,\xi)^T\theta + 1 - \frac{\triangle_n}{k}]. \end{aligned}$$

Thus by (20), we have

$$\sup_\theta P_n[h'(x,\xi)^T\theta - 1 + \tfrac{\triangle_n}{k}, h'(x,\xi)^T\theta + 1 - \tfrac{\triangle_n}{k}] < \triangle_n + 1/2 - k(\triangle_n/k) = 1/2,$$

which implies that

$$r_n \geq 1 - \triangle_n/k. \tag{22}$$

Similarly, by (21), there exists $\lambda > 0$ such that

$$P[-1-\delta, 1+\delta] \geq 1/2 + \lambda\delta,$$

for all $\delta > 0$ small enough. Therefore,

$$P_n[-1 - \tfrac{\triangle_n}{\lambda}, 1 + \tfrac{\triangle_n}{\lambda}] \geq -\triangle_n + P[-1 - \tfrac{\triangle_n}{\lambda}, 1 + \tfrac{\triangle_n}{\lambda}] \geq -\triangle_n + 1/2 + \lambda(\tfrac{\triangle_n}{\lambda}) = 1/2,$$

which implies

$$r_n \leq 1 + \frac{\triangle_n}{\lambda}. \tag{23}$$

Combining the results in (22) and (23), we get $r_n = 1 + O_p(n^{-1/2})$.

A.3 Conditions for Kim and Pollard's Lemma 4.1 are satisfied in nonlinear case

Denote G_R^h at fixed x by $G_R^h(x)$. Note that

$$|g_{h,x,u}(\theta, r, \xi)| \leq I^*_{(-1-\delta, h'(x,\xi)^T\theta - 1 - \delta)}(u) + I^*_{(h'(x,\xi)^T\theta + 1 + \delta, 1+\delta)}(u),$$

for fixed x. Here the asterisk of the indicator function means that the interval may be reversed, that is,

$$I^*_{(a,b)}(u) = I_{(\min(a,b), \max(a,b))}(u).$$

By the boundedness of the density of u, let $M < \infty$ be the supremum of the density of u. Thus,

$$E_u^x G_R^h(x) \leq \sup_{\|\theta - \theta_0\| \leq R} \{2Mh'(x,\xi)^T\theta\}. \tag{24}$$

Recall that we set $\theta_0 = 0$. By the Cauchy-Schwarz inequality, we get that

$$E_u^x G_R^h(x) \leq 2M\|h'(x,\xi)\|R,$$

which implies that

$$EG_R^h = E_x E_u^x G_R^h(x) \leq 2ME_x(\|h'(X,\xi)\|)R.$$

Therefore, by assumption 6 in our Theorem 4, it follows that $EG_R^h = O(R)$, which is required for Lemma 4.1 in [10] to establish the convergence of θ_n or β_n.

A.4 Check the conditions of Kim and Pollard's main theorem/ our Theorem 3

In what follows, we verify that Kim and Pollard's main theorem holds for LMS estimators in nonlinear models, i.e. we check the conditions of our Theorem 3.

A.4.1 Conditions 2, 3 and 4 are satisfied

First, we use Lemma 4.1 of Kim and Pollard [10] for the pair (θ, δ) to show that there exists $M_n = O_p(1)$ such that

$$|E_n g_{h,x,u}(\theta, \delta, \xi) - E g_{h,x,u}(\theta, \delta, \xi)| \leq \epsilon[\|\theta\|^2 + \delta^2] + n^{-2/3} M_n^2, \tag{25}$$

for each $\epsilon > 0$.

To do this, consider

$$E g_{h,x,u}(\theta, \delta, \xi) = E_x[E_u^x[g_{h,x,u}(\theta, \delta, \xi)]],$$

where $E_u^x[g_{h,x,u}(\theta, \delta, \xi)] = E_u[g_{h,x,u}(\theta, \delta, \xi)|x]$ is a continuous function of θ and δ for fixed x.

By Taylor's expansion of $E_u^x[g_{h,x,u}(\theta, \delta, \xi)]$ about $\theta = 0$ and $\delta = 0$, we have $E_u^x[g_{h,x,u}(\theta, \delta, \xi)] = \gamma'(1)\theta^T Q_h^x \theta + o(\|\theta\|^2) + o(\delta^2)$ and

$$E g_{h,x,u}(\theta, \delta, \xi) = \gamma'(1)\theta^T Q_h \theta + o(\|\theta\|^2) + o(\delta^2), \tag{26}$$

where $Q_h = E_x Q_h^x = E_x[h'(x, \beta_0)h'(x, \beta_0)^T]$. (26) is used to verify conditions 2 and 4. However, its derivation is long so it is differed to the end of this subsection.

By (25) and (26), we have

$$E_n g_{h,x,u}(\theta, \delta, \xi) \leq \gamma'(1)\theta^T Q_h \theta + o(1)\|\theta\|^2 + o(1)\delta^2 + \epsilon[\|\theta\|^2 + \delta^2] + O_p(n^{-2/3}).$$

Since θ_n maximizes $E_n g_{h,x,u}(\theta, r_n - 1, \xi)$, we have

$$\begin{aligned} 0 = E_n g_{h,x,u}(0, r_n - 1, \beta_0) &\leq E_n g_{h,x,u}(\theta_n, r_n - 1, \xi) \\ &\leq \gamma'(1)\theta_n^T Q_h \theta_n + (\epsilon + o(1))\|\theta_n\|^2 \\ &\quad + (\epsilon + o(1))(r_n - 1)^2 + O_p(n^{-2/3}). \end{aligned}$$

Since we proved that $r_n = 1 + O_p(n^{-1/2})$, we now obtain

$$0 \leq \gamma'(1)\theta_n^T Q_h \theta_n + (\epsilon + o(1))\|\theta_n\|^2 + (\epsilon + o(1))O_p(n^{-1}) + O_p(n^{-2/3}). \tag{27}$$

Note that Q_h is a symmetric matrix, so we have

$$\lambda_d \leq \tfrac{\theta^T Q_h \theta}{\theta^T \theta} \leq \lambda_1,$$

where λ_1 and λ_d are the largest and smallest eigenvalues of Q_h. In other words, we have

$$\theta^T Q_h \theta \geq \lambda_d \|\theta\|^2,$$

which implies that $\gamma'(1)\theta_n^T Q_h \theta_n \leq \gamma'(1)\lambda_d \|\theta_n\|^2$ ($\because \gamma'(1) < 0$). Thus, by (27),

$$[-\lambda_d\gamma'(1) - (\epsilon + o(1))]\|\theta_n\|^2 \leq (\epsilon + o(1))O_p(n^{-1}) + O_p(n^{-2/3}).$$

Since Q_h is positive-definite, $\lambda_d > 0$. Taking $\epsilon = \frac{-\gamma'(1)}{2}\lambda_d > 0$, we have

$$[\tfrac{-\gamma'(1)}{2}\lambda_d - o(1)]\|\theta_n\|^2 \leq O_p(n^{-2/3}),$$

which implies $\|\theta_n\| = O_p(n^{-1/3})$ or $\|\beta_n - \beta_0\| = O_p(n^{-1/3})$. So, condition 2 holds.

Second, condition 3 is satisfied by the assumption on (4).

Third, to verify condition 4, observe that (26) implies that $Eg_{h,x,u}(\theta, \delta, \xi)$ is twice differentiable in θ and the second derivative matrix with respect to θ at $(0, 0, \beta_0)$ is

$$\begin{aligned}\frac{\partial^2}{\partial\theta\partial\theta^T}Eg_{h,x,u}(0,0,\beta_0) &= E_X\frac{\partial^2}{\partial\theta\partial\theta^T}E_u^x[g_{h,x,u}(0,0,\beta_0)] \\ &= E_X 2\gamma'(1)Q_h^X \\ &= 2\gamma'(1)Q_h.\end{aligned}$$

Finally, we derive the expression (26). Recall that

$$Eg_{h,x,u}(\theta,\delta,\xi) = E_x[E_u^x[g_{h,x,u}(\theta,\delta,\xi)]],$$

where $E_u^x[g_{h,x,u}(\theta,\delta,\xi)] = E_u[g_{h,x,u}(\theta,\delta,\xi)|x]$ is a continuous function of θ and δ for fixed x.

By Taylor's expansion of $E_u^x[g_{h,x,u}(\theta,\delta,\xi)]$ about $\theta = 0$ and $\delta = 0$, we have

$$\begin{aligned}E_u^x[g(\theta,\delta,\xi)] =& E_u^x[g(0,0,\beta_0)] + \theta^T\frac{\partial}{\partial\theta}E_u^x[g(0,0,\beta_0)] + \delta\frac{\partial}{\partial\delta}E_u^x[g(0,0,\beta_0)] \\ &+ \frac{1}{2}\Big[\theta^T\frac{\partial^2}{\partial\theta\partial\theta^T}E_u^x[g(0,0,\beta_0)]\theta + \delta^2\frac{\partial^2}{\partial\delta^2}E_u^x[g(0,0,\beta_0)] \\ &+ 2\delta\theta^T\frac{\partial^2}{\partial\delta\partial\theta}E_u^x[g(0,0,\beta_0)]\Big] + o(\|\theta\|^2) + o(\delta^2),\end{aligned}$$

where we use $g = g_{h,x,u}$ above, so $g_\theta = \frac{\partial}{\partial\theta}g_{h,x,u}$.

By the definition of g_h, we have $E_u^x[g(0,0,\beta_0)] = 0$. Moreover, since

$$\begin{aligned}\frac{\partial}{\partial\theta}E_u^x[g(\theta,\delta,\xi)] &= \frac{\partial}{\partial\theta}\int_{h'(x,\xi)^T\theta-1-\delta}^{h'(x,\xi)^T\theta+1+\delta}\gamma(u)du - 0 \\ &= h'(x,\xi)\gamma(h'(x,\xi)^T\theta + 1 + \delta) - h'(x,\xi)\gamma(h'(x,\xi)^T\theta - 1 - \delta),\end{aligned}$$

we have $\frac{\partial}{\partial\theta}E_u^x[g(0,0,\beta_0)] = h'(x,\beta_0)\gamma(1) - h'(x,\beta_0)\gamma(-1) = \mathbf{0}$, because $\gamma(1) = \gamma(-1)$).

Similarly, we have

$$\begin{aligned}\text{(i)}\quad \frac{\partial}{\partial\delta}E_u^x[g(\theta,\delta,\xi)] &= \frac{\partial}{\partial\delta}\int_{h'(x,\xi)^T\theta-1-\delta}^{h'(x,\xi)^T\theta+1+\delta}\gamma(u)du - \frac{\partial}{\partial\delta}\int_{-1-\delta}^{1+\delta}\gamma(u)du \\ &= [\gamma(h'(x,\xi)^T\theta + 1 + \delta) + \gamma(h'(x,\xi)^T\theta - 1 - \delta)] \\ &\quad - [\gamma(1+\delta) + \gamma(-1-\delta)],\end{aligned}$$

$\Rightarrow \frac{\partial}{\partial\delta}E_u^x[g(0,0,\beta_0)] = 0$.

(ii) $\frac{\partial^2}{\partial\theta\partial\theta^T}E_u^x[g(\theta,\delta,\xi)] = \gamma'(h'(x,\xi)^T\theta + 1 + \delta)h'(x,\xi)h'(x,\xi)^T - \gamma'(h'(x,\xi)^T\theta - 1 - \delta)h'(x,\xi)h'(x,\xi)^T$, thus we have

$$\begin{aligned}\frac{\partial^2}{\partial\theta\partial\theta^T}E_u^x[g(0,0,\beta_0)] &= [\gamma'(1)h'(x,\beta_0)h'(x,\beta_0)^T] - [\gamma'(-1)h'(x,\beta_0)h'(x,\beta_0)^T] \\ &= 2\gamma'(1)Q_h^x,\end{aligned}$$

where $Q_h^x = h'(x,\beta_0)h'(x,\beta_0)^T$ for fixed x and $\gamma'(1) = -\gamma'(-1)$.

(iii) $\frac{\partial^2}{\partial\delta^2}E_u^x[g(\theta,\delta,\xi)] = \frac{\partial}{\partial\delta}[\gamma(h'(x,\xi)^T\theta + 1 + \delta) + \gamma(h'(x,\xi)^T\theta - 1 - \delta)] - \frac{\partial}{\partial\delta}[\gamma(1+\delta) + \gamma(-1-\delta)]$, which implies that $\frac{\partial^2}{\partial\delta^2}E_u^x[g(0,0,\beta_0)] = 0$.

(iv) $\frac{\partial^2}{\partial\delta\partial\theta}E_u^x[g(\theta,\delta,\xi)] = \frac{\partial}{\partial\delta}[h'(x,\xi)\gamma(h'(x,\xi)^T\theta + 1 + \delta) - h'(x,\xi)\gamma(h'(x,\xi)^T\theta - 1 - \delta)]$. Thus, we have $\frac{\partial^2}{\partial\delta\partial\theta}E_u^x[g(0,0,\beta_0)] = \mathbf{0}$.

Thus, we have $E_u^x[g_{h,x,u}(\theta,\delta,\xi)] = \gamma'(1)\theta^T Q_h^x\theta + o(\|\theta\|^2) + o(\delta^2)$ and (26).

A.4.2 Conditions 6 and 7 are satisfied

For condition 6, since u has a bounded density and $E\|h'(X,\xi)\| < \infty$ by our assumption 6, it follows that $E(G_R^h)^2 = O(R)$ by the same technique we used for showing $EG_R^h = O(R)$ in Appendix A.3.

For condition 7, recall that $g_{h,x,u}$ is the difference of two indicator functions $f_{h,x,u}(\theta, 1+\delta, \xi)$ and $f_{h,x,u}(0, 1+\delta, \xi)$, where $f_{h,x,u}(\theta, r, \xi) = I_{\{|u - h'(x,\xi)^T\theta| \le r\}}$. So, for (θ_1, δ_1) and (θ_2, δ_2) near $(0,0)$,

$$\left|g_{h,x,u}(\theta_1,\delta_1,\xi_1) - g_{h,x,u}(\theta_2,\delta_2,\xi_2)\right| \le I^*_{A_1}(u) + I^*_{A_2}(u) + I^*_{A_3}(u) + I^*_{A_4}(u).$$

There are many combinations of intervals of the form A_1, A_2, A_3 and A_4. For example, $A_1 = (-1-\delta_1, -1-\delta_2)$, $A_2 = (1+\delta_2, 1+\delta_1)$, $A_3 = (h'(x,\xi_2)^T\theta_2 - 1 - \delta_2, h'(x,\xi_1)^T\theta_1 - 1 - \delta_1)$ and $A_4 = (h'(x,\xi_2)^T\theta_2 + 1 + \delta_2, h'(x,\xi_1)^T\theta_1 + 1 + \delta_1)$. In all cases the total length of the intervals A_1, A_2, A_3 and A_4 on the right is bounded by $2|h(x,\beta_1) - h(x,\beta_2)| + 4|\delta_2 - \delta_1|$ for fixed x, where $\beta_i = \beta_0 + \theta_i$, and $\xi_i \in (\beta_0, \beta_i)$ for i=1 and 2.

Moreover, $\xi_i \to \beta_0$ as $\theta_i \to 0$. Thus,

$$\begin{aligned}&E_u^x\left|g_{h,x,u}(\theta_1,\delta_1,\xi_1) - g_{h,x,u}(\theta_2,\delta_2,\xi_2)\right| \\ &\le M[2|h(x,\beta_1) - h(x,\beta_2)| + 4|\delta_2 - \delta_1|].\end{aligned}$$

By assumption 4 in our Theorem 4, we have

$$|h(x,\beta_1) - h(x,\beta_2)| \leq L_x\|\beta_1 - \beta_2\| = L_x\|\theta_1 - \theta_2\| \ ,$$

where $L_x > 0$ depends on x. Therefore,

$$E_u^x\big|g_{h,x,u}(\theta_1,\delta_1,\xi_1) - g_{h,x,u}(\theta_2,\delta_2,\xi_2)\big| \leq 2ML_x\|\theta_1 - \theta_2\| + 4M|\delta_2 - \delta_1|,$$

and

$$\begin{aligned}
&E\big|g_{h,x,u}(\theta_1,\delta_1,\xi_1) - g_{h,x,u}(\theta_2,\delta_2,\xi_2)\big| \\
=&E_xE_u^x\big|g_{h,x,u}(\theta_1,\delta_1,\xi_1) - g_{h,x,u}(\theta_2,\delta_2,\xi_2)\big| \\
\leq&2ME_x(L_x)\|\theta_1 - \theta_2\| + 4M|\delta_2 - \delta_1| \\
\leq&2M\big[max\{E_x(L_x),2\}\big]\big[\|\theta_1 - \theta_2\| + |\delta_2 - \delta_1|\big],
\end{aligned}$$

which implies that

$$E\big|g_{h,x,u}(\theta_1,\delta_1,\xi_1) - g_{h,x,u}(\theta_2,\delta_2,\xi_2)\big| = O(\|\theta_1 - \theta_2\| + |\delta_2 - \delta_1|). \tag{28}$$

So Kim and Pollard's condition 7 is satisfied.

A.4.3 Condition 1 is satisfied

Now we show that θ_n comes close to maximizing $E_nf_{h,x,u}(\theta,1,\xi)$, which is equivalent to saying that β_n maximizes $P_n(|y - h(x,\beta)| \leq 1)$. Kim and Pollard's technique needs to check whether or not the two-parameter centered process

$$\begin{aligned}
X_n(a,b) = &n^{2/3}E_ng_{h,x,u}(an^{-1/3},bn^{-1/3},\xi(a)) \\
&- n^{2/3}Eg_{h,x,u}(an^{-1/3},bn^{-1/3},\xi(a))
\end{aligned}$$

satisfies the uniform tightness (i.e. stochastic equicontinuity) condition used for the weak convergence of the process. In their lemma 4.6, Kim and Pollard [10] show that the process X_n satisfies the uniform tightness. The main hypotheses of lemma 4.6 are uniform manageability and conditions 6 and 7. In Appendix A.1.1 we have shown the classes of $f_{h,x,u}$ and $g_{h,x,u}$ are manageable. Also, in Appendix A.4.2 we establish conditions 6 and 7. Now X_n is uniformly tight. Given this, we must show that β_n comes close to maximizing $P_n(|y - h(x,\beta)| \leq 1)$. So, using $n^{1/3}(r_n - 1) = o_p(1)$, we have

$$X_n(n^{1/3}\theta, n^{1/3}(r_n - 1)) - X_n(n^{1/3}\theta, 0) = o_p(1)$$

uniformly over θ in an $O_p(n^{-1/3})$ neighborhood of zero. That is,

$$\begin{aligned}
&E_ng_{h,x,u}(\theta, r_n - 1, \xi) \\
=&Eg_{h,x,u}(\theta, r_n - 1, \xi) + E_ng_{h,x,u}(\theta,0,\xi) - Eg_{h,x,u}(\theta,0,\xi) + o_p(n^{-2/3}),
\end{aligned}$$

uniformly over an $O_p(n^{-1/3})$ neighborhood. Within such a neighborhood, by (26) we have

$$Eg_{h,x,u}(\theta, r_n - 1, \xi) - Eg_{h,x,u}(\theta, 0, \xi) = o((r_n - 1)^2) = o_p(n^{-2/3}).$$

Then if m_n maximizes $E_n g_{h,x,u}(\theta, 0, \xi)$ just as θ_n maximizes $E_n g_{h,x,u}(\theta, r_n - 1, \xi)$, we have $m_n = O_p(n^{-1/3})$. Therefore,

$$\begin{aligned} E_n g_{h,x,u}(\theta_n, 0, \xi) &= E_n g_{h,x,u}(\theta_n, r_n - 1, \xi) - o_p(n^{-2/3}) \\ &\geq E_n g_{h,x,u}(m_n, r_n - 1, \xi) - o_p(n^{-2/3}) \\ &= E_n g_{h,x,u}(m_n, 0, \xi) - o_p(n^{-2/3}). \end{aligned}$$

In other words, we have

$$E_n g_{h,x,u}(\theta_n, 0, \xi) \geq \sup_{\theta} E_n g_{h,x,u}(\theta_n, 0, \xi) - o_p(n^{-2/3}),$$

which means that θ_n comes close to maximizing $E_n f_{h,x,u}(\theta, 1, \xi)$.

A.4.4 Condition 5 is satisfied

Consider the one-parameter class of functions $\{g_{h,x,u}(\theta, 0, \xi) : \theta \in \mathcal{R}^d, \xi \in (\beta_0, \beta)\}$ with $\theta = \beta - \beta_0$. Using the same techniques as in the verification of conditions 6 and 7, we have for fixed s and t,

$$\begin{aligned} & E_u^x |g_{h,x,u}(\frac{s}{\alpha}, 0, \xi_s) - g_{h,x,u}(\frac{t}{\alpha}, 0, \xi_t)|^2 \\ =& |\Gamma(1 + h'(x, \xi_s)^T \frac{s}{\alpha}) - \Gamma(1 + h'(x, \xi_t)^T \frac{t}{\alpha})| \\ & + |\Gamma(-1 + h'(x, \xi_s)^T \frac{s}{\alpha}) - \Gamma(-1 + h'(x, \xi_t)^T \frac{t}{\alpha})|. \end{aligned}$$

By Taylor's expansion of the first two terms at 1 and the last two at -1 with $\gamma(1) = \gamma(-1)$, we have

$$\begin{aligned} & E_u^x |g_{h,x,u}(\frac{s}{\alpha}, 0, \xi_s) - g_{h,x,u}(\frac{t}{\alpha}, 0, \xi_t)|^2 \\ =& 2\big|\gamma(1)[h'(x, \xi_s)^T \frac{s}{\alpha} - h'(x, \xi_t)^T \frac{t}{\alpha}] + o(1/\alpha)\big|, \end{aligned} \tag{29}$$

where $\xi_s \in (\beta_0, \beta_s)$ and $\xi_t \in (\beta_0, \beta_t)$, $\beta_s = \beta_0 + \frac{s}{\alpha}$ and $\beta_t = \beta_0 + \frac{t}{\alpha}$. In fact, $\|\xi_s - \beta_0\| \leq \|s/\alpha\|$ and $\|\xi_t - \beta_0\| \leq \|t/\alpha\|$. As $\alpha \to \infty$, ξ_s and ξ_t will tend to β_0. Thus we have

$$\begin{aligned} L(s - t) \equiv & \lim_{\alpha \to \infty} \alpha E |g_{h,x,u}(\frac{s}{\alpha}, 0, \xi_s) - g_{h,x,u}(\frac{t}{\alpha}, 0, \xi_t)|^2 \\ =& 2 \lim_{\alpha \to \infty} E_x \big|\gamma(1)[h'(x, \xi_s)^T s - h'(x, \xi_t)^T t] + \alpha o(1/\alpha)\big| \\ =& 2\gamma(1) E_x |h'(x, \beta_0)^T (s - t)|. \end{aligned}$$

Similarly, we can also prove that

$$L(s) \equiv \lim_{\alpha \to \infty} \alpha E |g_{h,x,u}(\tfrac{s}{\alpha}, 0, \xi_s)|^2 = 2\gamma(1) E_x |h'(x, \beta_0)^T s|$$

and

$$L(t) \equiv \lim_{\alpha\to\infty} \alpha E|g_{h,x,u}(\tfrac{t}{\alpha}, 0, \xi_t)|^2 = 2\gamma(1)E_x|h'(x,\beta_0)^T t|.$$

Thus, the limiting covariance function is

$$\begin{aligned} H(s,t) &\equiv \lim_{\alpha\to\infty} \alpha E[g_{h,x,u}(\frac{s}{\alpha}, 0, \xi_s) g_{h,x,u}(\frac{t}{\alpha}, 0, \xi_t)] \\ &= \frac{1}{2}[L(s) + L(t) - L(s-t)], \end{aligned}$$

by the identity $2xy = x^2 + y^2 - (x-y)^2$.

A.5 Proof of asymptotic results for the LMS estimators in nonlinear models

Conditions 1-7 are satisfied, it is enough to complete the proof of Theorem 4 by verifying that the limiting Gaussian process has nondegenerate increments.

Note that in Appendix A.4.4., since $L(0) = 0$, $H(s,s) = L(s)$ and $H(t,t) = L(s)$. Thus, by our assumption 6, we have

$$H(s,s) - 2H(s,t) + H(t,t) = L(s-t) \neq 0, \text{ for any } s \neq t. \tag{30}$$

Under (30), Kim and Pollard's lemma 2.6 in [10] can be applied to give that the limiting Gaussian process has nondegenerate increments. Consequently, applying Kim and Pollard's main theorem with our assumption 3 on the positive definiteness of Q_h, we can identify the limit distribution of $n^{1/3}\theta_n$, i.e. $n^{1/3}(\beta_n - \beta_0)$, with the $\arg\max_\theta$ of the Gaussian process

$$Z(\theta) = \gamma'(1)\theta^T Q_h \theta + W(\theta),$$

where W has zero means, covariance kernel H and continuous sample paths.

B LTS

The following results are used for the proof of the asymptotic behavior of the two-sided LTS estimator.

B.1 Proof of Lemma 4 for the uniform law of large numbers

Proof. We prove the uniform weak law of large numbers in lemma 4 by verifying the four conditions of Andrews' theorem 4 in [3]. First, (i) The condition of total boundedness (BD) is ensured by assumption $TI1$ for the compactness of the parameter space $\mathcal{B}$.

(ii) Note that, since $E\sup_{\beta\in\mathcal{B}}|t(x,u;\beta)|^{1+\delta} < \infty$, for some $\delta > 0$,

$$t(x_i,u_i;\beta)I_3(\beta;K_1,K_2) - E[t(x_i,u_i;\beta)I_3(\beta;K_1,K_2)]$$

are identically distributed by assumptions $TD1$ and $TD2$; they are also uniformly integrable. Thus, Andrews' domination condition (DM) is satisfied.

(iii) Additionally, the pointwise convergence of

$$\frac{1}{n}\sum_{i=1}^{n}[t(x_i,u_i;\beta)I_3(\beta;K_1,K_2)] - E[t(x_i,u_i;\beta)I_3(\beta;K_1,K_2)] \xrightarrow{p} 0$$

at any $\beta \in \mathcal{B}$ and $K_1,\ K_2 \in \mathcal{R}$ follows from the weak law of large numbers for mixingales in [2].

(iv) The last condition of termwise stochastic equicontinuity (TSE) in Andrews' Theorem 4 [3] that

$$\lim_{\rho\to 0} P\Big(\sup_{\beta,K_1,K_2\beta',K_1',K_2'}\Big|t_I(x_i,u_i;\beta',K_1',K_2') - t_I(x_i,u_i;\beta,K_1,K_2)\Big| > k\Big) = 0 \tag{31}$$

is satisfied for any $k>0$, where $t_I(x_i,u_i;\beta,K_1,K_2) = t(x_i,u_i;\beta)I_3(\beta;K_1,K_2)$ and the suprema $\beta, K_1, K_2, \beta', K_1'$ and K_2' are taken over the sets $\mathcal{B},\mathcal{R},\mathcal{R},U(\beta,\rho),U(K_1,\rho)$ and $U(K_2,\rho)$.

To see that (31) holds, first notice that for all $\beta\in\mathcal{B}$ and $K_1,K_2\in\mathcal{R}$, we have

$$\begin{aligned}
&\sup_{\beta,K_1,K_2\beta',K_1',K_2'}\Big|t_I(x_i,u_i;\beta',K_1',K_2') - t_I(x_i,u_i;\beta,K_1,K_2)\Big| \\
\le &\sup_{\beta,K_1,K_2\beta',K_1',K_2'}\Big|t(x_i,u_i;\beta')[I_3(\beta';K_1',K_2') - I_3(\beta;K_1,K_2)]\Big| \qquad (32)\\
+ &\sup_{\beta,K_1,K_2\beta',K_1',K_2'}\Big|[t(x_i,u_i;\beta') - t(x_i,u_i;\beta)]I_3(\beta;K_1,K_2)\Big|. \qquad (33)
\end{aligned}$$

Now it is enough to show that given $\epsilon>0$, we can find $\rho_0>0$ such that the probabilities of the expression (32) and (33) exceeding given $k>0$ are smaller than ϵ for all $\rho<\rho_0$.

1. Consider the expression (32). First note that

$$\begin{aligned}
&\sup_{\beta,K_1,K_2\beta',K_1',K_2'}\Big|t(x_i,u_i;\beta')[I_3(\beta';K_1',K_2') - I_3(\beta;K_1,K_2)]\Big| \\
\le &\sup_{\beta\in\mathcal{B}}\Big|t(x_i,u_i;\beta)\Big| \sup_{\beta,K_1,K_2\beta',K_1',K_2'}\Big|I_3(\beta';K_1',K_2') - I_3(\beta;K_1,K_2)\Big|, \qquad (34)
\end{aligned}$$

where $\sup_{\beta\in\mathcal{B}}\Big|t(x_i,u_i;\beta)\Big|$ is a function independent of β with a finite expectation. In addition, $\Big|I_3(\beta';K_1',K_2') - I_3(\beta;K_1,K_2)\Big|$ is always less than or equal to 1, so (32) has an integrable upper bound independent of β. Thus, if we can show that the probability

$$\lim_{\rho\to 0} P(\sup_{\beta,K_1,K_2\beta',K_1',K_2'}\Big|I_3(\beta';K_1',K_2') - I_3(\beta;K_1,K_2)\Big| = 1) = 0, \tag{35}$$

then we get that (34) converges in probability to zero for $\rho\to 0$ and $n\to\infty$ as well. So to prove (34) it is enough to prove (35).

Our strategy for proving (34) has three steps. (1) We use Čížek's argument that $G_{\beta'}^{-1}(\lambda)$ to $G_{\beta}^{-1}(\lambda)$ uniformly on $\mathcal{B}$ for all λ by the absolute continuity of G_β. (2) By the result of the uniform convergence of G_β^{-1}, we can find some $\rho_1>0$ such that for $1/2<\lambda\le 1$,

$$\left|(G_{\beta'}^{-1}(1-\lambda)+K_1')-(G_{\beta}^{-1}(1-\lambda)+K_1)\right|<\epsilon(16M_{gg}^{*})^{-1}$$

and

$$\left|(G_{\beta'}^{-1}(\lambda)+K_2')-(G_{\beta}^{-1}(\lambda)+K_2)\right|<\epsilon(16M_{gg}^{**})^{-1},$$

for any $\beta\in\mathcal{B},\beta'\in U(\beta,\rho_1)$ and $K_j'\in U(K_j,\rho_1)$ for $j=1,2$, where M_{gg}^{*} and M_{gg}^{**}, defined in assumption $TD3$, are the uniform upper bounds in both sides for the probability density functions of $r_i^2(\beta)$. (3) If we denote the product probability space of (x_i,u_i) by Ω and consider a compact subset $\Omega_1\subset\Omega$, such that $P(\Omega_1)>1-\epsilon/2$, and choose $\rho_2>0$ such that

$$\sup_{\beta\in\mathcal{B}}\sup_{\beta'\in U(\beta,\rho_2)}\left|r_i^2(\beta',\omega)-r_i^2(\beta,\omega)\right|<\epsilon(16\max\{M_{gg}^{*},M_{gg}^{**}\})^{-1},\tag{36}$$

for all $\omega\in\Omega_1$ and $\rho<\rho_2$ by assumption $TH1$.

Therefore, letting $\rho_0=\min\{\rho_1,\rho_2\}$ and $\rho<\rho_0$, we can apply steps (1), (2) and (3) to get the following sequence of inequalities. We have that

$$\begin{aligned}
& P(\sup_{\beta,K_1,K_2\beta',K_1',K_2'}\left|I_3(\beta';K_1',K_2')-I_3(\beta;K_1,K_2)\right|=1)\\
= \;& P(\sup_{\beta,K_1,K_2\beta',K_1',K_2'}\left|I_3(\beta';K_1',K_2')-I_3(\beta;K_1,K_2)\right|=1,\Omega_1)\\
& +P(\sup_{\beta,K_1,K_2\beta',K_1',K_2'}\left|I_3(\beta';K_1',K_2')-I_3(\beta;K_1,K_2)\right|=1,\Omega|\Omega_1)\\
\leq \;& P(\sup_{\beta,K_1,K_2\beta',K_1',K_2'}\left|I_3(\beta';K_1',K_2')-I_3(\beta;K_1,K_2)\right|=1,\Omega_1)+P(\Omega|\Omega_1)\\
\leq \;& P(\sup_{\beta,K_1,K_2\beta',K_1',K_2'}\left|I_3(\beta';K_1',K_2')-I_3(\beta;K_1,K_2)\right|=1,\Omega_1)+\epsilon/2\\
= \;& P(\exists\beta\in\mathcal{B}:r_i^2(\beta)\in[G_{\beta}^{-1}(\lambda)+K_2-\epsilon(8M_{gg}^{**})^{-1},G_{\beta}^{-1}(\lambda)+K_2+\epsilon(8M_{gg}^{**})^{-1}]\\
& \qquad\cup[G_{\beta}^{-1}(1-\lambda)-K_1-\epsilon(8M_{gg}^{*})^{-1},G_{\beta}^{-1}(1-\lambda)-K_1+\epsilon(8M_{gg}^{*})^{-1}])+\epsilon/2\\
\leq \;& M_{gg}^{**}(\frac{\epsilon}{4M_{gg}^{**}})+M_{gg}^{*}(\frac{\epsilon}{4M_{gg}^{*}})\\
= \;& \frac{\epsilon}{2}.
\end{aligned}$$

Thus, (35) is proved, and finally, the expectation of (32) converges to zero for $\rho\to 0$ in probability.

2. Now we turn to expression (33) and prove that for any given $k>0$,

$$\lim_{\rho\to 0}P(\sup_{\beta,K_1,K_2\beta',K_1',K_2'}\left|[t(x_i,u_i;\beta')-t(x_i,u_i;\beta)]I_3(\beta;K_1,K_2)\right|>k)=0.\tag{37}$$

By Čížek's result that

$$E\Big\{\sup_{\beta,\beta'}\Big|t(x_i,u_i;\beta')-t(x_i,u_i;\beta)\Big|\Big\}\leq k\epsilon,$$

we have

$$\begin{aligned} & P(\sup_{\beta,K_1,K_2\beta',K_1',K_2'}\Big|[t(x_i,u_i;\beta')-t(x_i,u_i;\beta)]I_3(\beta;K_1,K_2)\Big|>k) \\ \leq\ & \frac{1}{k}E\Big[\sup_{\beta,K_1,K_2\beta',K_1',K_2'}\Big|[t(x_i,u_i;\beta')-t(x_i,u_i;\beta)]I_3(\beta;K_1,K_2)\Big|\Big] \\ \leq\ & k\epsilon/k=\epsilon, \end{aligned}$$

for any $\rho<\rho_0$. Thus, (B.7) is proved.

Consequently, the assumption of TSE in [3] is valid and the proof of this lemma is completed by applying the uniform weak law of large numbers. □

B.2 Proof of Proposition 1 on Asymptotic Linearity

Now we can prove Proposition 1.

Proof. Recall that

$$\begin{aligned} D_n^1(t) &= S_n'(\beta_0-n^{-1/2}t)-S_n'(\beta_0) \\ &= -2\sum_{i=1}^{n}\Big[\{y_i-h(x_i,\beta_0-n^{-1/2}t)\}h_\beta'(x_i,\beta_0-n^{-1/2}t)I_2(\beta_0-n^{-1/2}t) \\ &\qquad -\{y_i-h(x_i,\beta_0)\}h_\beta'(x_i,\beta_0)I_2(\beta_0)\Big], \end{aligned}$$

Here, $I_2(\beta)=I_2(r_i^2(\beta))=I_{[r^2_{[n-h_n+1]}(\beta),r^2_{[h_n]}(\beta)]}(r_i^2(\beta))$ and $t\in\mathcal{T}_M=\{t\in\mathcal{R}^p:\|t\|\leq M\}$. For any $M>0$, there is an $n_0\in\mathcal{N}$ such that $\beta_0-n^{-1/2}t\in U(\beta_0,\delta)$, for all $n\geq n_0$ and $t\in\mathcal{T}_M$. Therefore, using Taylor's expansion for $n>n_0$ and $t\in\mathcal{T}_M$, we have

$$h(x,\beta_0-n^{-1/2}t)=h(x,\beta_0)-h_\beta'(x,\xi)^Tn^{-1/2}t$$

and

$$h_\beta'(x,\beta_0-n^{-1/2}t)=h_\beta'(x,\beta_0)-h_{\beta\beta}''(x,\xi')^Tn^{-1/2}t,$$

where ξ and ξ' are between β_0 and $\beta_0-n^{-1/2}t$. Let $B_1(x)=h(x,\beta_0), B_2(x)=h_\beta'(x,\xi)^Tn^{-1/2}t$, $C_1(x)=$

$h'_\beta(x,\beta_0)$ and $C_2(x) = h''_{\beta\beta}(x,\xi')^T n^{-1/2}t$. Thus, $D_n^1(t)$ can be rewritten as

$$\begin{aligned}
\frac{D_n^1(t)}{-2} &= \sum_{i=1}^{n}\Big[\{y_i - B_1(x_i)\}C_1(x_i)I_2(\beta_0 - n^{-1/2}t) - \{y_i - B_1(x_i)\}C_1(x_i)I_2(\beta_0) \\
&\quad - \{y_i - B_1(x_i)\}C_2(x_i)I_2(\beta_0 - n^{-1/2}) + B_2(x_i)C_1(x_i)I_2(\beta_0 - n^{-1/2}t) \\
&\quad - B_2(x_i)C_2(x_i)I_2(\beta_0 - n^{-1/2}t)\Big] \\
&= \sum_{i=1}^{n}\Big[\{(y_i - B_1(x_i))C_1(x_i)\}[I_2(\beta_0 - n^{-1/2}t) - I_2(\beta_0)]\Big] \qquad (38) \\
&\quad - \sum_{i=1}^{n}\Big[(y_i - B_1(x_i))C_2(x_i)I_2(\beta_0)\Big] \qquad (39) \\
&\quad - \sum_{i=1}^{n}\Big[(y_i - B_1(x_i))C_2(x_i)[I_2(\beta_0 - n^{-1/2}t) - I_2(\beta_0)]\Big] \qquad (40) \\
&\quad + \sum_{i=1}^{n}\Big[B_2(x_i)C_1(x_i)I_2(\beta_0)\Big] \qquad (41) \\
&\quad + \sum_{i=1}^{n}\Big[B_2(x_i)C_1(x_i)[I_2(\beta_0 - n^{-1/2}t) - I_2(\beta_0)]\Big] \qquad (42) \\
&\quad - \sum_{i=1}^{n}\Big[B_2(x_i)C_2(x_i)I_2(\beta_0 - n^{-1/2}t)\Big] \qquad (43)
\end{aligned}$$

Using techniques substantially like those in Čížek's proofs for his (42)-(47), we can show that the sums in (39), (40), (42) and (43) are $O_p(n^{1/4})$ or $o_p(n^{1/2})$, and therefore, are asymptotically negligible in comparison with (38) and (41), which are $O_p(n^{1/2})$. Thus, we omit the proofs here, except for (38) and (41).

To deal with (38), let $v_i(n,t) = I_2(\beta_0 - n^{-1/2}t) - I_2(\beta_0)$. So (38) can be rewritten as

$$\begin{aligned}
&\sum_{i=1}^{n} \Big(\{(y_i - B_1)C_1\}[I_2(\beta_0 - n^{-1/2}t) - I_2(\beta_0)] \Big) \\
=&\sum_{i=1}^{n} \Big(\{(y_i - h(x_i,\beta_0))h'_\beta(x_i,\beta_0)\}[I_2(\beta_0 - n^{-1/2}t) - I_2(\beta_0)] \Big) \\
=&\sum_{i=1}^{n} r_i(\beta_0) \cdot h'_\beta(x_i,\beta_0) \cdot v_i(n,t) \\
=&\frac{1}{2}\sum_{i=1}^{n} \Big[\{r_i(\beta_0) - sgn\, r_i(\beta_0) \cdot q_\lambda\} + \{r_i(\beta_0) - sgn\, r_i(\beta_0) \cdot q_{1-\lambda}\} \\
&\qquad + sgn\, r_i(\beta_0)[q_\lambda + q_{1-\lambda}] \Big] \cdot h'_\beta(x_i,\beta_0) \cdot v_i(n,t) \\
=&\frac{1}{2}\Big[\sum_{i=1}^{n} \{r_i(\beta_0) - sgn\, r_i(\beta_0) q_\lambda\} \cdot h'_\beta(x_i,\beta_0) \cdot v_i(n,t) \qquad (44) \\
&+ \sum_{i=1}^{n} \{r_i(\beta_0) - sgn\, r_i(\beta_0) q_{1-\lambda}\} \cdot h'_\beta(x_i,\beta_0) \cdot v_i(n,t) \qquad (45) \\
&+ \sum_{i=1}^{n} sgn\, r_i(\beta_0)(q_\lambda + q_{1-\lambda}) \cdot h'_\beta(x_i,\beta_0) \cdot v_i(n,t) \Big]. \qquad (46)
\end{aligned}$$

Again, using techniques substantially like those of Čížek with our Lemmas 9 and 10, (44) and (45) multiplied by $n^{-1/4}$ can be shown to be bounded in probability for $\lambda \in (1/2, 1)$. Moreover (46) can be rewritten as

$$\begin{aligned}
&\sum_{i=1}^{n} sgn\, r_i(\beta_0)(q_\lambda + q_{1-\lambda}) \cdot h'_\beta(x_i,\beta_0) \cdot v_i(n,t) \\
=&n^{1/2}(q_\lambda + q_{1-\lambda})[H(\lambda) - H(1-\lambda)]Q_h t + O(1) + o_p(n^{1/2}).
\end{aligned}$$

Therefore, we conclude that

$$\begin{aligned}
\sup_{t \in \mathcal{T}_M} \| &\sum_{i=1}^{n} \{r_i(\beta_0)\} h'_\beta(x_i,\beta_0) v_i(n,t) \\
&- \frac{1}{2} n^{1/2}(q_\lambda + q_{1-\lambda})[H(\lambda) - H(1-\lambda)]Q_h t\| = o_p(1),
\end{aligned}$$

as $n \to \infty$.

Finally we split (41) into two parts :

$$\begin{aligned}
B_2 C_1 I_2(\beta_0) &= \sum_{i=1}^{n} h'_\beta(x_i,\xi)^T n^{-1/2} t \cdot h'_\beta(x_i,\beta_0) I_2(\beta_0) \\
&= \sum_{i=1}^{n} h'_\beta(x_i,\beta_0)^T n^{-1/2} t \cdot h'_\beta(x_i,\beta_0) I_2(\beta_0) \qquad (47) \\
&\quad + \sum_{i=1}^{n} n^{-1/2} t^T h''_{\beta\beta}(x_i,\xi'') \cdot n^{-1/2} t \cdot h'_\beta(x_i,\beta_0) I_2(\beta_0), \qquad (48)
\end{aligned}$$

where ξ'' is between β_0 and $\beta_0 - n^{-1/2}t$.

Note that the supremum of (48) over $t \in \mathcal{T}_M$ is $O_p(1)$. Since

$$\begin{aligned}&\Big|\sum_{i=1}^{n} n^{-1/2}t^T h''_{\beta\beta}(x_i,\xi'')\cdot n^{-1/2}t\cdot h'_\beta(x_i,\beta_0)I_2(\beta_0)\Big|\\ \le&\sum_{i=1}^{n}\|n^{-1/2}t^T\cdot h''_{\beta\beta}(x_i,\xi'')\cdot n^{-1/2}t\cdot h'_\beta(x_i,\beta_0)\|,\end{aligned}$$

by the law of large numbers for mixingales in [2] and the uniform law of large numbers in [3] for the right hand side of the inequality over $\beta'' \in U(\beta,\delta)$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\Big|t^T h''_{\beta\beta}(x_i,\beta'')t\cdot h'_\beta(x_i,\beta_0)\Big| \xrightarrow{p} E\Big|t^T h''_{\beta\beta}(x_i,\beta'')t\cdot h'_\beta(x_i,\beta_0)\Big|,$$

as $n\to\infty$. Moreover, (48) is bounded in probability because the expectation is bounded uniformly over $t\in\mathcal{T}_M$ by assumption $TH5$ and $\|t\|\le M$.

Next we turn to (47). Similarly, split it into three parts :

$$\begin{aligned}&\sum_{i=1}^{n} h'_\beta(x_i,\beta_0)^T n^{-1/2}t\cdot h'_\beta(x_i,\beta_0)I_2(\beta_0)\\ =&\sum_{i=1}^{n} h'_\beta(x_i,\beta_0)^T n^{-1/2}t\cdot h'_\beta(x_i,\beta_0)\Big[I_2(\beta_0)-I_2^G(\beta_0)\Big] &(49)\\ &+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{h'_\beta(x_i,\beta_0)h'_\beta(x_i,\beta_0)^T I_2^G(\beta_0)-E[h'_\beta(x_i,\beta_0)h'_\beta(x_i,\beta_0)^T I_2^G(\beta_0)]\Big\}t &(50)\\ &+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}E[h'_\beta(x_i,\beta_0)h'_\beta(x_i,\beta_0)^T I_2^G(\beta_0)]t, &(51)\end{aligned}$$

where $I_2^G(\beta_0) = I_{\{G^{-1}(1-\lambda)\le r_i^2(\beta_0)\le G^{-1}(\lambda)\}}$ with $\lambda\in(1/2,1)$. The supremum of (49) taken over $t\in\mathcal{T}_M$ is $O_p(n^{1/4})$ as $n\to\infty$, and (50) is bounded in probability by applying the central limit theorem to (50). Again, the proofs are omitted here because they are so similar to Čížek's.

Finally, since

$$\begin{aligned}&E[h'_\beta(x_i,\beta_0)h'_\beta(x_i,\beta_0)^T I_2^G(\beta_0)]\\ =&E_{X_i}[h'_\beta(x_i,\beta_0)h'_\beta(x_i,\beta_0)^T\cdot\{EI_{\{G^{-1}(1-\lambda)\le r_i^2(\beta_0)\le G^{-1}(\lambda)\}}|X_i\}]\\ =&(2\lambda-1)E_{X_i}[h'_\beta(x_i,\beta_0)h'_\beta(x_i,\beta_0)^T]\\ =&(2\lambda-1)Q_h, &(52)\end{aligned}$$

(51) can be rewritten as $n^{1/2}(2\lambda-1)Q_h t$, where $\lambda\in(1/2,1)$. Thus, we can conclude that

$$\sup_{t\in\mathcal{T}_M}\Big\|\sum_{i=1}^{n} h'_\beta(x_i,\beta_0)n^{-1/2}t\cdot h'_\beta(x_i,\beta_0)\cdot I_2(\beta_0)-n^{1/2}(2\lambda-1)Q_h t\Big\| = O_p(1),$$

as $n\to\infty$.

The proof of Proposition 1 is completed by combining all of the above results. □

References

[1] Allais, M. (1953). Le comportement de lhomme rationnel devant le risque: critique des postulats et axiomes de lecole Americaine.

[2] Andrews, D.W.K. (1988). Laws of large numbers for dependent non-identically distributed random variables. *Econometric theory* **4** 458–467.

[3] Andrews, D.W.K. (1992). Generic uniform convergence. *Econometric theory* **8** 241–257.

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

[5] Borwanker, J., Kallianpur, G., and Prakasa Rao, B.L.S. (1971). The Bernstein-Von Mises Theorem for Markov Processes. *The Annals of Mathematical Statistics* **42(4)** 1241–1253.

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

[7] Čížek, P. (2005). Least trimmed squares in nonlinear regression under dependence. *Journal of Statistical Planning and Inference* **136** 3967–3988.

[8] Dudley, R.M. (1979). Balls in $\mathcal{R}^k$ do not cut all subsets of $k+2$ points. *Advances in Mathematics* **31** 306–308.

[9] Ellsberg, D. (1961). Risk, Ambiguity, and the Savage Axioms. *Quarterly Journal of Economics* **75** 643–669.

[10] Kim, J and Pollard, D. (1990). Cube Root Asymptotics. *The Annals of Statistics* **18(1)** 191–219.

[11] Machina, M. and D. Schmeidler. (1992). A More Robust Definition of Subjective Probability. *Econometrica* **60(4)** 745–780.

[12] Manski, C. F. (1988). Ordinal Utility Models of Decision Making Under Uncertainty. *Theory and Decision* **25(1)** 79–104.

[13] Pollard, D. (1990). *Empirical Processes: Theory and Applications.* Conference Board of the Mathematical Sciences, NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 2, Hayward, Calif. : Institute of Mathematical Statistics, and Alexandria, Va. : American Statistical Association.

[14] Prakasa Rao, B.K.S. (1987). *Asymptotic Theory of Statistical Inference.* John Wiley and Sons.

[15] Rostek, M.J. (2006). *Ph.D. Dissertation* Yale University.

[16] Rousseeuw, P.J. (1984). Least Median of Squares Regression. *Journal of the American Statistical Association* **79** 871-880.

[17] Rukhin, A.L. (1978). Universal Bayes Estimators. *The Annals of Statistics* **6(6)** 1345–1351.

[18] Savage, L.J. (1954). *The Foundations of Statistics*. Wiley, New York.

[19] Shorack, G.R. (2000). *Probability for Statisticians*. Springer texts in statistics. Springer-Verlag, New York.

[20] Stromberg, A.J. (1995). Consistency of the least median of squares estimator in nonlinear regression. *Commun. Statist.-Theory Meth.* **24(8)** 1971–1984.

[21] Tomkins, R.J. (1978). Convergence Properties of Conditional Medians. *The Canadian Journal of Statistics* **6(2)** 169–177.

[22] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak convergence and empirical processes: with applications to statistics*. Springer series in statistics. Springer-Verlag, New York.

[23] Von Neumann, J. and Morgenstern, O. (1947). *The Theory of Games and Economic Behaviour*, 2nd ed. (1st edn 1944). Princeton, Princeton University Press.

[24] Wald, A. (1939). Contributions to the Theory of Statistical Estimation and Testing Hypotheses. *The Annals of Mathematical Statistics* **10(4)** 299–326.

[25] Wald, A. (1950). *Statistical Decision Functions*. John Wiley, New York.