CentER

Discussion Paper

No. 2004–72

ASYMPTOTICS OF LEAST TRIMMED SQUARES REGRESSION

By P. Čížek

August 2004

ISSN 0924-7815

Tilburg University

Asymptotics of least trimmed squares regression

Pavel Čížek

Department of Econometrics and Operation Research
Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

Abstract

High breakdown-point regression estimators protect against large errors both in explanatory and dependent variables. The least trimmed squares (LTS) estimator is one of frequently used, easily understandable, and thoroughly studied (from the robustness point of view) high breakdown-point estimators. In spite of its increasing popularity and number of applications, there are only conjectures and hints about its asymptotic behavior in regression after two decades of its existence. We derive here all important asymptotic properties of LTS, including the asymptotic normality and variance, under mild β-mixing conditions.

Keywords: nonlinear regression, robust estimation, least trimmed squares
JEL codes: C13, C20

1 Introduction

In statistics and econometrics, a more attention is paid to techniques that can deal with data contamination, which can arise from miscoding or heterogeneity not captured or presumed in a model. This can occur, for instance, if some data points come from a different data-generating process that the majority of observations. Sakata and White (1998) evidence data contamination in financial time series and its adverse effects on estimators such as quasi-maximum likelihood. The sensitivity or robustness of an estimator against large errors and data contamination is typically characterized by the breakdown point, which measures the smallest fraction of a sample that can arbitrarily change the estimator under contamination; see Rousseeuw and Leroy (1987) and Rousseeuw (1997) for an overview, Stromberg and Ruppert (1992) for a breakdown point in nonlinear regression, and Sakata and White (1995) for some finite-sample alternative definitions. In this paper, we study a classical high breakdown-point estimator, the least trimmed squares (LTS), proposed by Rousseeuw (1985) and derive asymptotic results allowing for nonlinear-regression and time-series applications.

The LTS estimator belongs to the class of affine-equivariant estimators that achieve asymptotically the highest breakpoint 1/2 and it is generally preferred to the similar, but slowly converging least median of squares (LMS; Rousseeuw, 1984). Thus, it has been receiving a lot of attention from the theoretical, computational, and application points of view. First, let us mention its extensions to nonlinear regression (Stromberg, 1993) and regression with categorical dependent variables (Christmann, 1998). Results are also available regarding strong consistency (Chen, Stromberg, and Zhou, 1997), sensitivity analysis (Tableman, 1994), small-sample corrections for LTS (Pison, Van Aelst, and Willems, 2002), and bootstrap (Willems and Van Aelst, 2004). Further, there has been a significant development in computational methods (Agulló, 2001; Bai, 2003; Gilloni and Padberg, 2002; Rousseeuw and Van Driessen, 1999). Last, but not least, there are many application areas where LTS has been used: in economics (Beňáček, Jarolím, and Víšek, 1998; Temple, 1998; Zaman, Rousseeuw, and Orhan, 2001), finance (Knez and Ready, 1997; Kelly, 1997), but also in clustering (Ye and Haralick, 2000) and pattern recognition (Wang and Sutter, 2003). Further applications could stem from areas, where LMS is suitable and applicable (see Zinde-Walsh, 2002, for details). In spite of its many extensions and uses, rigorously proved results are limited only to the i.i.d. setting and location model (see Hawkins and Olive, 1999, for an overview) and the knowledge concerning the asymptotic distribution of LTS in regression models consist of a vague conjecture on deriving asymptotic variance made by Stromberg, Hössjer, and Hawkins (2000).

The aim of this work is to address this deficiency and derive the asymptotic distribution of LTS, and as a side effect, to prove the consistency of LTS under weaker conditions than Chen, Stromberg, and Zhou (1997). The main difficulty in deriving such a result stems from the LTS objective function: being a sum of h smallest residuals at any given parameter estimate, it is not differentiable at many points. Thus, the standard tools such as the Taylor expansion of the objective function are not applicable. On the other hand, the standard results of the empirical process theory (see for example Pollard, 1984, van der Vaart and Wellner, 1996, and Andrews, 1993) cannot be readily employed either as noticed by Stromberg, Hössjer, and Hawkins (2000). For this reason, we study first behavior of ordered residual statistics and prove the asymptotic linearity of the LTS normal equations. Next, combining the first set of results with the (uniform) law of large numbers (Andrews, 1987 and 1992) and the stochastic equicontinuity results (Arcones and Yu, 1994, and Yu, 1994) for mixing processes allows us to derive the consistency and the rate of convergence of the LTS estimates. Finally, the consistency of LTS and the asymptotic linearity of the LTS normal equations leads to the asymptotic normality of the LTS estimator.

In the rest of the paper, LTS and its existing extensions to nonlinear regression are introduced in more details in Section 2, where we also extensively discuss assumptions needed for the asymptotic normality of LTS. Asymptotic results are summarized and discussed in

Sections 3 and 4. The proofs are provided in Appendix.

2 Least trimmed squares in nonlinear regression

Let us consider the nonlinear regression model ($i = 1, \ldots, n$)

$$y_i = h(x_i, \beta^0) + \varepsilon_i, \tag{1}$$

where y_i represents the dependent variable, $h(x_i, \beta)$ is a regression function, and β^0 represents the underlying parameter value. The vector $x_i \in \mathbb{R}^k$ represents explanatory variables and the error term ε_i is assumed to form a sequence of independent and identically distributed random variables with an absolutely continuous distribution function.[1] The vector β of unknown parameters is assumed to belong to a parametric space $B \subseteq \mathbb{R}^p$.

The nonlinear least trimmed squares estimator $\hat{\beta}_n^{(LTS,h)}$ is then defined by

$$\hat{\beta}_n^{(LTS,h)} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{h} r_{[i]}^2(\beta), \tag{2}$$

where $r_{[i]}^2(\beta)$ represents the ith order statistics of squared residuals $r_i^2(\beta) = \{y_i - h(x_i, \beta)\}^2$ and $\beta \in B$. The trimming constant h must satisfy $\frac{n}{2} < h \leq n$ and determines the breakdown point of the (nonlinear) LTS estimator since definition (2) implies that $n - h$ observations with the largest residuals do not affect the estimator (except for the fact that the squared residuals of excluded points have to be larger than the hth order statistics of the squared residuals). For $h(x, \beta) = g(x^T\beta)$, where $g(t)$ is unbounded for $t \to \pm\infty$, Stromberg and Ruppert (1992) showed that the breakdown point equals asymptotically 1/2 for $h = [n/2] + 1$ (most robust choice) and 0 for $h = n$ (nonlinear least squares). For other cases, only upper and lower bounds for the breakdown point can be established. For an overview of the properties of LTS in linear and nonlinear regression, see Čížek and Víšek (2000), Víšek (2000), and Čížek (2001), Stromberg (1993), respectively.

Naturally, the choice of the trimming constant h should vary with the sample size n. Because the asymptotic properties of LTS are studied here, that is $n \to \infty$, we have to work with a sequence of trimming constants h_n (for every sample size n, there has to be a corresponding choice of h). As h_n/n determines the fraction of sample included in the LTS objective function, and consequently, the robustness properties of LTS, we want to asymptotically fix this fraction at λ, $\frac{1}{2} \leq \lambda \leq 1$.[2] The trimming constant for a given

[1] Although I assume throughout the work that all variables are of stochastic nature, all presented results hold even in the presence of nonstochastic variables (e.g., seasonal dummies).

[2] The case of $\lambda = 1$ will be excluded for the sake of simplicity from some proofs. This case corresponds to the usual nonlinear least squares estimator, which is extensively studied in the literature anyway. All

sample size n can be then defined by $h_n = [\lambda n]$, where $[x]$ represents the integer part of x; in general, one can also consider any sequence $\{h_n\}_{n\in\mathbb{N}}$ such that $h_n/n \to \lambda$.

In the rest of this section, we discuss assumptions (Section 2.1) and an alternative definition of LTS (Section 2.2) used throughout the paper.

2.1 Assumptions

Let us now complement the model and LTS estimator definition first by some notation and definitions and later by assumptions on the regression function and random variables needed for further analysis.

First, we refer to the distribution functions of ε_i and ε_i^2 as $F(z)$ and $G(z)$ and to the corresponding probability density functions, if they exist, as $f(z)$ and $g(z)$, respectively. Note that since G describes the distribution of the square of the random variable $\varepsilon_i \sim F$, it follows that $G(z) = F(\sqrt{z}) - F(-\sqrt{z})$ for $z > 0$ and $G(z) = 0$ otherwise. Hence, if F is absolutely continuous, G is absolutely continuous too and the corresponding probability density function is $g(z) = \frac{1}{2\sqrt{z}}\{f(\sqrt{z}) + f(-\sqrt{z})\}$ for $z > 0$. Last, but not least, whenever I need to refer to the quantile functions corresponding to F and G, I use F^{-1} and G^{-1}, respectively. Two purely mathematical symbols we need are indicator $I(A)$, which equals 1 for $x \in A$ and 0 elsewhere, and an open δ-neighborhood of a point x in a Euclidian space $\mathbb{R}^l$: $U(x,\delta) = \left\{ z \in \mathbb{R}^l \middle| \, \|z - x\| < \delta \right\}$.

Second, let us introduce the concept of β-mixing, which is central to the distributional assumptions made here. A sequence of random variables $\{X_i\}_{i\in\mathbb{N}}$ is said to be absolutely regular (or β-mixing) if

$$\beta_m = \sup_{t\in\mathbb{N}} \mathsf{E} \sup_{B\in\sigma_{t+m}^f} |P(B|\sigma_t^p) - P(B)| \to 0$$

as $m \to \infty$, where the σ-algebras $\sigma_t^p = \sigma(X_t, X_{t-1}, \ldots)$ and $\sigma_t^f = \sigma(X_t, X_{t+1}, \ldots)$; see Davidson, 1994, or Arcones and Yu, 1994, for details. Numbers $\beta_m, m \in \mathbb{N}$, are called mixing coefficients.

Another concept crucial to this paper are the Vapnik-Cervonenkis (VC) classes of functions, which are rigorously defined and studied in monographs Pollard (1984) and van der Vaart and Wellner (1996), for instance. Very closely related are also the Euclidian classes of functions (Pollard, 1989). To avoid rather technical definitions, let us say that VC classes cover many common functions including any set of functions forming a finite vector space (e.g., polynomial, logarithmic, and exponential functions), functions for which $|f(x,t) - f(x,t')| \le \xi(x)\,\|t - t'\|^\alpha$ for some $\alpha > 0$ and a nonnegative function $\xi(x)$, their

the propositions given later are valid for $\lambda = 1$ too, but their proofs are slightly different or trivial in this case.

sums, products, maxima and minima, monotonic transformations, composed functions, and so on.

Now, I specify all the assumptions necessary to derive the asymptotic linearity of LTS. They form three groups: distributional Assumptions D for random variables in model (1), Assumptions H concerning properties of the regression function $h(x, \beta)$, and finally, the identification Assumptions I.

Assumptions D

D1 Explanatory variables $\{x_i\}_{i\in\mathbb{N}}$ form an absolutely regular sequence with finite second moments and mixing coefficients satisfying

$$m^{r_\beta/(r_\beta-2)} (\log m)^{2(r_\beta-1)/(r_\beta-2)} \beta_m \to 0$$

as $m \to \infty$ for some $r_\beta > 2$.

D2 Let $\{\varepsilon_i\}_{i\in\mathbb{N}}$ be a sequence of independent symmetrically and identically distributed variables with finite second moments, and additionally, let ε_i and x_i are mutually independent. Further, the distribution function F of ε_i is absolutely continuous and its probability density f is assumed to be positive, bounded from above by $M_f > 0$, and continuously differentiable in a neighborhood of $-\sqrt{G^{-1}(\lambda)}$ and $\sqrt{G^{-1}(\lambda)}$.

D3 Assume that $m_G = \inf_{\beta\in B} G_\beta^{-1}(\lambda) > 0$,

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

for some $\delta_g > 0$, and

$$M_{gg} = \sup_{\beta\in B} \sup_{z\in(m_G,+\infty)} g_\beta(z) < \infty,$$

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

Having a general regression function $h(x, \beta)$, Assumption D1 is a necessary condition for the uniform central limit theorem, see Andrews (1993) and Arcones and Yu (1994), for instance. The first part of Assumption D2 is standard and is mainly made for the ease of presentation. The mutual independence of ε_i and x_i can be relaxed, although we need at least conditional symmetry of ε_i given x_i in the later case. The second part of Assumption D2 on distribution function F, especially its twice differentiability around the points corresponding to the λ-quantiles of ε_i^2, is a standard condition needed for the analysis of rank statistics (see

Zinde-Walsh, 2002, for instance). Most importantly, it bounds F and f away from zero in a neighborhood of the mentioned quantiles: $\inf_{z\in U(F^{-1}(\alpha),\varepsilon)} \min\{F(z), f(z)\} > 0$ for $\alpha = (1-\lambda)/2$ and $\alpha = (1+\lambda)/2$. Note that this property together with the absolute continuity of F transfer to G due to the relation $G(z) = F(\sqrt{z}) - F(-\sqrt{z})$. Assumption D3 formalizes this property for the distribution G_β of squared residuals across the whole parameter space B. Although unfamiliar, this assumption excludes above all convergence of G_β to a discountinuous distribution function for some $\beta \in B$ and should not restrict us in common regression models since B is assumed to be compact, see Assumption I below. Finally, although Assumption D implies stochastic nature of all explanatory variables, the presented results are valid also in the presence of nonstochastic variables, such as seasonal dummies.

Next, several conditions on the regression function $h(x, \beta)$ have to be specified. Most of them are just regularity conditions that are employed in almost any work concerning nonlinear regression models. For example, the regression function of a nonlinear regression model is almost always assumed to be twice differentiable; see Amemiya (1983) and White (1980), for example. Further, since some assumptions stated below rely on the value of β and I do not have to require their validity over the whole parametric space, I restrict β to a neighborhood $U(\beta^0, \delta)$ in these cases.

Assumptions H

Let us assume that there are a positive constant $\delta > 0$ and a neighborhood $U(\beta^0, \delta)$ such that the following assumptions hold.

H1 Let $h(x_i, \beta)$ be a continuous (uniformly over any compact subset of the support of x) in $\beta \in B$ and twice differentiable function in β on $U(\beta^0, \delta)$ almost surely. The first derivative is continuous in $\beta \in U(\beta^0, \delta)$.

H2 Furthermore, let us assume that the second derivatives $h''_{\beta_j\beta_k}(x, \beta)$ satisfy locally the Lipschitz property, that is, for any compact subset of $\operatorname{supp} x$ there exists a constant $L_p > 0$ such that for all $\beta, \beta' \in U(\beta^0, \delta)$, and $j, k = 1, \ldots, p$

$$\left|h''_{\beta_j\beta_k}(x, \beta) - h''_{\beta_j\beta_k}(x, \beta')\right| \leq L_p \cdot \|\beta - \beta'\| .$$

H3 Let $\{h(x_i, \beta)|\beta \in B\}$ and $\{h'_\beta(x_i, \beta)|\beta \in U(\beta^0, \delta)\}$ form VC classes of functions such that their envelopes $E_1(x) = \sup_{\beta\in B} |h(x, \beta)|$ and $E_2(x) = \sup_{\beta\in U(\beta^0,\delta)} \left|h'_\beta(x, \beta)\right|$ have finite r_β-th moments.

H4 Let

$$n^{-1/4} \max_{1\leq i\leq n} \max_{1\leq j\leq p} \left|h'_{\beta_j}(x_i, \beta)\right| = \mathcal{O}_p(1) \tag{3}$$

and

$$n^{-1/2} \max_{1\le i\le n} \max_{1\le j,k\le p} \left| h''_{\beta_j\beta_k}(x_i,\beta) \right| = \mathcal{O}_p(1) \tag{4}$$

as $n \to +\infty$ uniformly over $\beta \in U(\beta^0, \delta)$.

H5 Apart from the existence of moments implied by Assumption H3, we also have to postulate the existence of the following expectations:

- Integrals $\mathsf{E}\left[r_i^2(\beta)\right]^m$ and $\mathsf{E}\left[h(x,\beta)\right]^m$ exist and are finite for $m = 1, 2$ and $\beta \in B$.
- Let $\mathsf{E}\left[h''_{\beta_j\beta_k}(x_i,\beta)\right]^m$, $\mathsf{E}\left[h'_{\beta_j}(x_i,\beta^0)h'_{\beta_k}(x_i,\beta^0)\right]^m$, and $\mathsf{E}\left[h'_{\beta_l}(x_i,\beta^0)h''_{\beta_j\beta_k}(x_i,\beta^0)\right]$ exist and are finite for $m = 1, 2$, all $j, k, l = 1, \ldots, p$, and $\beta \in U(\beta^0, \delta)$.

Moreover, assume that $\mathsf{E}\left[h'_\beta(x_i,\beta^0)h'_\beta(x_i,\beta^0)^T\right] = Q_h$, where Q_h is a nonsingular positive definite matrix.

Whereas the differentiability of the regression function and the existence of some moments are standard assumptions (Assumption H5 corresponds to the assumption of finite fourth moments of x_i in the linear case), Assumptions H3 and H4 deserve further comments. First, Assumption H3 limits the class of regression functions $h(x,\beta)$ to a VC class. Even though this assumption does not seem to be very restrictive, it can be omitted as long as we impose stronger distributional assumptions. More specifically, if x_i and ε_i are independent and the distribution function F of ε_i has everywhere differentiable density, it is possible to prove the L^{r_β}-continuity of $I\big(r_i^2(\beta) \le G_\beta^{-1}(\lambda)\big)$ and to limit the braketing cover numbers following results of Andrews (1993). Consequently, the results of Doukhan, Massart, and Rio (1995) could be employed instead of Arcones and Yu (1994) and Yu (1994) that are used in the current paper.

Second, Assumption H4 is a nonlinear equivalent of

$$n^{-1/4} \max_{1\le i,j\le n} |x_{ij}| = \mathcal{O}_p(1), \tag{5}$$

and actually, it is the direct consequence of (5) if $h(x,\beta) = h(x_i^T\beta)$ with bounded derivatives, which implies $h'_{\beta_j}(x,\beta) = h'(x_i^T\beta)x_{ij}$ and $h''_{\beta_j\beta_k}(x,\beta) = h''(x_i^T\beta)x_{ij}x_{ik}$. The restriction (5), in a nonrandom setup, was first introduced by Jurečková (1984) to be able to cope with the discontinuous objective function (this discontinuity has to be understood from the inclusion-of-observations point of view: every observation either fully enters the objective function or does not enter it at all). Nevertheless, it should not pose a considerable restriction on the explanatory variables: for example in the i.i.d. case, it follows from Proposition 2.1 below that equation (5) holds even for some distribution functions with polynomial tails, namely for those that have finite second moments. Additionally, one can notice that random variables with a finite support are not restrained by this assumption in any way.

Proposition 2.1 *Let $x_1, x_2, \ldots$ be a sequence of independent identically distributed random variables with a distribution function $F(z)$. Let $b(z)$ be a lower bound for $F(z)$ in a neighborhood U_1 of $+\infty$. If $b(z)$ can be chosen as $1 - \frac{1}{P_4(z)}$, where $P_4(z)$ is a polynomial of the fourth order, then it holds that $n^{-\frac{1}{4}} \max_{i=1,\ldots,n} x_i = \mathcal{O}_p(1)$ as $n \to +\infty$. Analogously, let $c(z)$ be an upper bound for $F(z)$ in a neighborhood U_2 of $-\infty$. If $c(z)$ can be chosen as $\frac{1}{P_4(z)}$, where $P_4(z)$ is a polynomial of the fourth order, then it holds that $n^{-\frac{1}{4}} \min_{i=1,\ldots,n} x_i = \mathcal{O}_p(1)$ as $n \to +\infty$.*

Proof: See Appendix A. □

Finally, we introduce two standard identification conditions.

Assumptions I

I1 B is a compact space.

I2 For any $\varepsilon > 0$ and $U(\beta^0, \varepsilon)$ such that $B \backslash U(\beta^0, \varepsilon)$ is compact, there exists $\alpha(\varepsilon) > 0$ such that it holds

$$\min_{\beta \in B \backslash U(\beta^0, \varepsilon)} \mathsf{E}\left[r_i^2(\beta) \cdot I\big(r_i^2(\beta) \leq G_\beta^{-1}(\lambda)\big)\right] - \mathsf{E}\left[r_i^2(\beta^0) \cdot I\Big(r_i^2(\beta^0) \leq G_{\beta^0}^{-1}(\lambda)\Big)\right] \;>\; \alpha(\varepsilon).$$

To close this section, let us note that Assumptions D, H, and I are sufficient to prove the asymptotic normality of LTS. If only consistency is needed, one can omit all assumptions on differentiability of the regression function $h(x, \beta)$ and the VC-class Assumption H3. To prove the $\sqrt{n}$-rate of convergence, Assumptions H4 and H5 are superfluous.

2.2 Alternative definition

Before proving the main results of the paper, some basic properties of the LTS objective function $S_n(\beta) = \sum_{i=1}^{h_n} r_{[i]}^2(\beta)$ and its alternative formulation, which is more suitable for deriving asymptotic linearity, are introduced.

Lemma 2.2 *Under Assumptions D2 and H1, $S_n(\beta)$ is continuous on B, twice differentiable at $\hat{\beta}_n^{(LTS,h_n)}$ as long as $\hat{\beta}_n^{(LTS,h_n)} \in U(\beta^0, \delta)$, and almost surely twice differentiable at any fixed point $\beta \in U(\beta^0, \delta)$. Furthermore,*

$$S_n(\beta) = \sum_{i=1}^{n} r_i^2(\beta) \cdot I\big(r_i^2(\beta) \leq r_{[h_n]}^2(\beta)\big), \tag{6}$$

$$S_n'(\beta) = \frac{\partial S_n(\beta)}{\partial \beta} = -2 \sum_{i=1}^{n} r_i(\beta) h_\beta'(x_i, \beta) \cdot I\big(r_i^2(\beta) \leq r_{[h_n]}^2(\beta)\big) \tag{7}$$

$$S_n''(\beta) = \frac{\partial^2 S_n(\beta)}{\partial \beta \partial \beta^T} = 2 \sum_{i=1}^{n} \Big\{ h_\beta'(x_i, \beta) h_\beta'(x_i, \beta)^T - r_i(\beta) h_{\beta\beta}''(x_i, \beta) \Big\} I\big(r_i^2(\beta) \leq r_{[h_n]}^2(\beta)\big) \tag{8}$$

almost surely at any $\beta \in B$ *and* $\beta \in U(\beta^0, \delta)$*, respectively.*

Proof: See Appendix A. □

In general, this definition is not equivalent to the one used in (2) unless all the residuals are different from each other. However, Assumption D2 guarantees this with probability one. Hence, we will use this notation and definition of $S_n(\beta)$ in the rest of the paper.

3 Asymptotic linearity

Although the consistency of LTS can be proved directly using standard tools such as the uniform law of large numbers (see Section 4), this is not the case of the asymptotic normality of LTS. Hence, assuming $\sqrt{n}$-consistency of the LTS estimator, we have to analyze the behavior of the normal equations $\partial S_n(\beta)/\partial\beta = 0$ around β^0 as a function of $\beta - \beta^0$. More specifically, we shall investigate the difference $D_n^1(t) = S_n'(\beta^0 - n^{-\frac{1}{2}}t) - S_n'(\beta^0)$, that is,

$$\begin{aligned} D_n^1(t) \;=\; \sum_{i=1}^{n} \Big[\Big\{ y_i - h\Big(x_i, \beta^0 - n^{-\frac{1}{2}}t\Big) \Big\} \cdot h'_\beta\Big(x_i, \beta^0 - n^{-\frac{1}{2}}t\Big) \times \\ \times I\Big(r_i^2\left(\beta^0 - n^{-\frac{1}{2}}t\right) \le r^2_{[h_n]}\left(\beta^0 - n^{-\frac{1}{2}}t\right)\Big) \\ - \big\{ y_i - h\big(x_i, \beta^0\big) \big\} \cdot h'_\beta\big(x_i, \beta^0\big) \cdot I\big(r_i^2(\beta^0) \le r^2_{[h_n]}(\beta^0)\big) \Big] \end{aligned} \tag{9}$$

for $t \in \mathcal{T}_M = \{t \in \mathbb{R}^p | \, \|t\| \le M\}$, where $0 < M < \infty$ is an arbitrary, but fixed constant. Intuitively, $D_n^1(t)$ describes the change in normal equations when some $\beta = \beta^0 - n^{-\frac{1}{2}}t$ (e.g., an estimate that converges at the $\sqrt{n}$-rate to β^0) is used instead of the true value β^0. We show now that $D_n^1(t)$ behaves asymptotically as a linear function of $n^{\frac{1}{2}}t$ over the whole set $\mathcal{T}_M$, which allows us later to explicitly express the first order approximation of the difference between an estimate $\hat{\beta}_n^{(LTS)}$ and the true value β^0.

Theorem 3.1 *Let Assumptions D, H, and I hold. Given constants* $\lambda \in \left\langle \frac{1}{2}, 1 \right\rangle$ *and* $M > 0$*, it holds that*

$$n^{-\frac{1}{2}} \sup_{t \in \mathcal{T}_M} \left\| D_n^1(t) + n^{\frac{1}{2}} Q_h t \cdot C_\lambda \right\| = o_p(1)$$

as $n \to +\infty$*, where*

$$C_\lambda = \lambda - \sqrt{G^{-1}(\lambda)} \cdot \left\{ f\left(-\sqrt{G^{-1}(\lambda)}\right) + f\left(\sqrt{G^{-1}(\lambda)}\right) \right\} = \lambda - 2G^{-1}(\lambda) g\left(G^{-1}(\lambda)\right).$$

Proof: See Appendix B. □

4 Consistency and asymptotic normality

Let us now present the main asymptotic results concerning LTS: its consistency, rate of convergence, and asymptotic normality. In all cases, we split the LTS objective function to two parts:

$$\begin{aligned} S_n(\beta) &= \sum_{i=1}^{n} r_i^2(\beta) \cdot I\big(r_i^2(\beta) \le r_{[h_n]}^2(\beta)\big) \\ &= \sum_{i=1}^{n} r_i^2(\beta) \cdot \big[I\big(r_i^2(\beta) \le r_{[h_n]}^2(\beta)\big) - I\big(r_i^2(\beta) \le G_\beta^{-1}(\lambda)\big)\big] \qquad (10) \\ &+ \sum_{i=1}^{n} r_i^2(\beta) \cdot I\big(r_i^2(\beta) \le G_\beta^{-1}(\lambda)\big). \qquad (11) \end{aligned}$$

Whereas the first part (10) will be shown to be small because of the convergence of order statistics to quantiles in mean, $r_{[h_n]}^2(\beta) \to G_\beta^{-1}(\lambda)$, the second part (11) will be dealt with by standard asymptotic tools and shown to converge to

$$S(\beta) = \mathsf{E}\left\{r_1^2(\beta) \cdot I\big(r_1^2(\beta) \le G_\beta^{-1}(\lambda)\big)\right\}.$$

First, using the uniform law of large numbers, we prove the consistency of the LTS estimator $\hat{\beta}_n^{(LTS)}$ minimizing $S_n(\beta)$ on the parametric space B. It is worth noticing that we do not have to limit the regression function $h(x, \beta)$ to be from a VC class of functions. Therefore, this consistency result is stronger than the one by Chen, Stromberg, and Zhou (1997) both from the distributional and regression-function points of view.

Theorem 4.1 *Let Assumptions D, H1, H5, and I hold. Then the least trimmed squares estimator $\hat{\beta}_n^{(LTS,h_n)}$ minimizing (6) is weakly consistent, that is, $\hat{\beta}_n^{(LTS,h_n)} \to \beta^0$ in probability as $n \to +\infty$.*

Proof: See Appendix C. □

Next, we will derive the rate of convergence of $\hat{\beta}_n^{(LTS,h_n)}$ to β^0, which should later allow us to employ the asymptotic linearity of LTS. Although the auxiliary results necessary to establish $\sqrt{n}$-consistency are non-trivial, the basic idea of the proof is simple. The second-order differentiability of $S(\beta)$ at β^0 together with Assumption H5, $Q_h > 0$, implies that $\|\partial S(\beta)/\partial\beta\| \ge C\,\|\beta - \beta^0\|$ in a neighborhood $U(\beta^0, \rho)$ for some $C > 0$ and $\rho > 0$. Since the consistency of LTS guarantees that $\hat{\beta}_n^{(LTS,h_n)} \in U(\beta^0, \rho)$ with probability approaching 1 as $n \to +\infty$, we just have to prove that $\left\|\partial S(\hat{\beta}_n^{(LTS,h_n)})/\partial\beta\right\| = \mathcal{O}_p\left(n^{-\frac{1}{2}}\right)$. This can be done again by using decomposition (10)–(11).

Theorem 4.2 *Let Assumptions D, H, and I hold. Then* $\hat{\beta}_n^{(LTS,h_n)}$ *is* $\sqrt{n}$*-consistent, that is,*

$$\sqrt{n}\left(\hat{\beta}_n^{(LTS,h_n)} - \beta^0\right) = \mathcal{O}_p(1)$$

as $n \to +\infty$.

Proof: See Appendix C. □

Finally, the asymptotic distribution of LTS can be derived by combining the $\sqrt{n}$-consistency of the estimator, Theorem 4.2, and its asymptotic linearity, Theorem 3.1. We discuss its main consequences in Section 4.1.

Theorem 4.3 *Let Assumptions D, H, and I are fulfilled and* $C_\lambda = \lambda - 2G^{-1}(\lambda)g\left(G^{-1}(\lambda)\right) \neq 0$. *Then*

$$\sqrt{n}\left(\hat{\beta}_n^{(LTS,h_n)} - \beta^0\right) = n^{-\frac{1}{2}} Q_h^{-1} C_\lambda^{-1} \cdot \sum_{i=1}^{n} \left\{y_i - h\left(x_i, \beta^0\right)\right\} \cdot h'_\beta\left(x_i, \beta^0\right) \cdot I\left(\varepsilon_i^2 \leq G^{-1}(\lambda)\right) + o_p(1)$$

and $\hat{\beta}_n^{(LTS,h_n)}$ *is asymptotically normal*

$$\sqrt{n}\left(\hat{\beta}_n^{(LTS,h_n)} - \beta^0\right) \xrightarrow{F} N(0, V_\lambda),$$

where $V_\lambda = C_\lambda^{-2} \cdot Q_h^{-1}\, \mathsf{var}\left[\varepsilon_i \cdot h'_\beta(x_i, \beta^0) \cdot I(\varepsilon_i^2 \leq G^{-1}(\lambda))\right] Q_h^{-1} = C_\lambda^{-2}\sigma_\lambda^2 \cdot Q_h^{-1}$, *where* $\sigma_\lambda^2 = \mathsf{E}\left[\varepsilon_i^2 \cdot I(\varepsilon_i^2 \leq G^{-1}(\lambda))\right]$.

Proof: See Appendix C. □

Let us note that the symmetry of the distribution function F implies that $\sqrt{G^{-1}(\lambda)} = F^{-1}((1+\lambda)/2)$, and consequently, we can write

$$C_\lambda = \lambda - 2F^{-1}\left(\frac{1+\lambda}{2}\right) f\left(F^{-1}\left(\frac{1+\lambda}{2}\right)\right).$$

Therefore in the case of a location model, the asymptotic variance V derived in Theorem 4.3 corresponds to the results of Tableman (1994) and Hawkins and Olive (1999). The later study also examines the convergence of the finite-sample LTS variance to the asymptotic variance $V_{0.5}$ and documents that the speed of convergence depend on the residual distribution F to a great extent. For example, whereas the asymptotic variance $V_{0.5}$ provides us with a good variance approximation for $n \geq 30$ in the case of the double exponential distribution, one needs several hundreds of observation to claim that $V_{0.5}$ approximates well finite-sample variance in the case of the standard normal distribution.

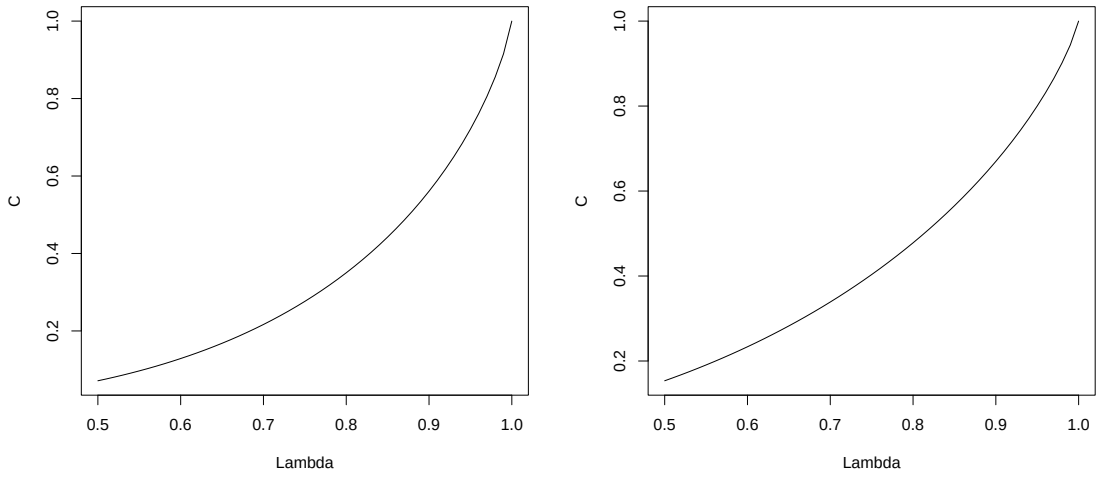

Figure 1: The dependence of C_λ on $\lambda \in (0.5, 1)$ for the Gaussian (left panel) and double exponential (right panel) distributions.

4.1 Implications

The asymptotic normality and variance derived in Theorem 4.3 have several interesting implications that concern the constant C_λ and the variance V_λ as a function of the trimming proportion λ.

Although one can construct a distribution such that $C_\lambda = 0$ at some specific $\lambda > 1/2$, this is not the case of usual unimodal distributions (e.g., the normal, Student, exponential, uniform distributions, see Figure 1). Nevertheless, one can imagine, for example, a mixture of two distributions like $F = 0.80N(0, 1) + 0.10N(c, 1) + 0.10N(-c, 1), c > 0$; the "smaller" parts of the mixture, $N(c, 1)$ and $N(-c, 1)$, can for a large c represent a contamination. In this case, C_λ could be equal or very close to zero for a sufficiently large c and $\lambda \approx 0.80$ and the LTS variance V_λ would extremely increase. This indicates and confirms a common wisdom that even if one has an idea about the maximal contamination α in data, the trimming constant λ should not be set to a value just below $1 - \alpha$ (to keep as much as data points within the objective function), but rather to a significantly smaller value.

Therefore, the choice of the trimming constant λ is very important because it influences both the robustness and variance of LTS. Theorem 4.3 can be used to determine whether there is a trade-off between the high breakdown point (i.e., λ close to 0.5) and the variance of LTS (usually, larger λ reduces variance) and how pronounced it is. To demonstrate, let us compare the behavior of σ_λ/C_λ (the λ-dependent part of asymptotic variance V_λ) under two specific distributions: the Gaussian and double exponential distributions (Figure 2). In the case of the normal distribution, the trade-off is very significant since using the maximal

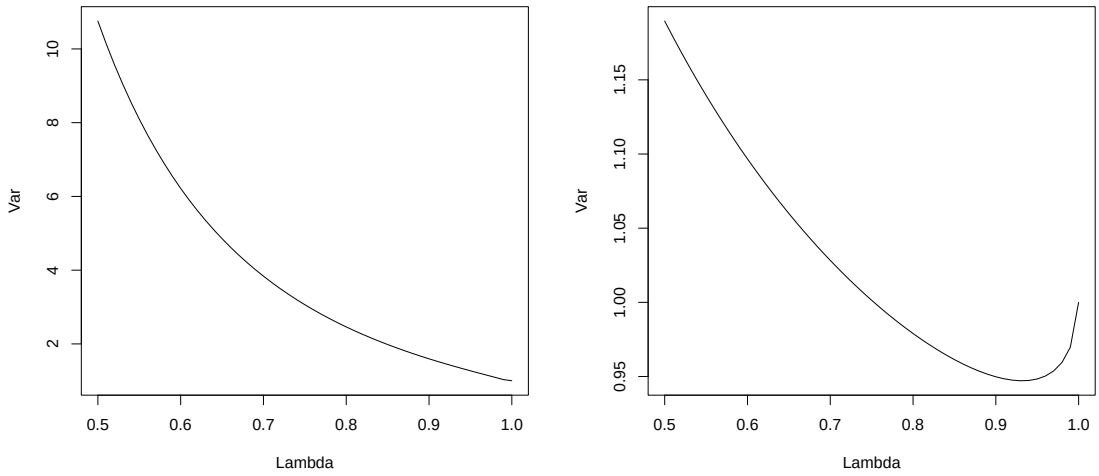

Figure 2: The dependence of σ_λ/C_λ on $\lambda \in (0.5, 1)$ for the Gaussian (left panel) and double exponential (right panel) distributions relative to the least squares, $\lambda = 1$.

trimming, $\lambda = 0.5$, increases the LTS variance more than 10 times compared to the least squares ($\lambda = 1$). Moreover, even a relatively minor increase to $\lambda = 0.6$ reduces V_λ by 42%. Thus, it may be preferable to keep λ above its most robust choice unless there is a strong reason to set $\lambda = 0.5$. On the other hand, in the case of the double exponential distribution, the trade-off between the robustness and variance of LTS is almost negligible. Even the maximal trimming at $\lambda = 0.5$ results only in a 19% increase in variance relative to the least squares.

Although the results mentioned here are just distribution-specific examples, one can often have an approximate idea about the error distribution in applications; for example, from previous evidence, distribution tests, residual analysis and so on. The demonstrated analysis of the trade-off between the breakdown point and variance of LTS can then provide an additional guidance in selecting λ.

5 Conclusion

We consider the least trimmed squares estimator and study its behavior in a nonlinear regression model under mild β-mixing conditions on the explanatory variables. First, we prove its consistency under weaker conditions than Chen, Stromberg, and Zhou (1997). Second, the main result concerns the asymptotic distribution of LTS in regression, which is derived under under conditions allowing for time series applications. Finally, the asymptotic variance of LTS is analyzed with respect to the trade-off between the robustness and variance of LTS. Although the results are distribution-specific, they point out that while

the trade-off is very significant under the normal model, it can be close to non-existent under other distributional laws.

Appendix

Here we present the proofs of important lemmas on the order statistics of squared residuals and the LTS objective function (Appendix A), on the asymptotic linearity of the LTS normal equations (Appendix B), and finally, on the consistency and asymptotic normality of LTS (Appendix C). Note that the alternative definition (6) of LTS is employed in all proofs, and additionally, notation $S_{nn}(\beta) = S_n(\beta)/n$ is used.

We introduce now the notation used in proofs, which extends the notation used in the body of the paper and introduced in Section 2. The dependent variable is denoted $y_i : \Omega_y \to \mathbb{R}$, the vector of explanatory variables is $x_i : \Omega_x \to \mathbb{R}^k$, whereby x_{ij} refers to the value of the jth variable ($1 \leq j \leq k$), and $\varepsilon_i : \Omega_e \to \mathbb{R}$ represents the error term; symbols Ω_y, Ω_x, and Ω_ε refer to the probability spaces that y_i, x_i, and ε_i, respectively, are defined on (thus, $\Omega = \Omega_x \times \Omega_\varepsilon$ is the probability space of the random vector (x_i, ε_i)). The true underlying value of the vector β in model (1) will be referred to by β^0.

Further, the order statistics $r^2_{[i]}(\beta)$ used to define the LTS estimator $\hat{\beta}_n^{(LTS,h_n)}$ in definition (2) stands for the ith order statistics of squared residuals $r_i^2(\beta) = \{y_i - h(x_i, \beta)\}^2$. In other words, it holds that $0 \leq r^2_{[1]}(\beta, \omega) \leq \cdots \leq r^2_{[n]}(\beta, \omega)$ for any $\beta \in B$ and $\omega \in \Omega$.[3] Given an $\omega \in \Omega$, we understand by symbol $r_{[i]}(\beta, \omega)$ the value of residual $r_k(\beta, \omega)$ such that $r_k^2(\beta, \omega) = r^2_{[i]}(\beta, \omega)$; hence, $\left|r_{[i]}(\beta)\right| = \sqrt{r^2_{[i]}(\beta)}$. If it is necessary to refer to the order statistics of sample $r_1(\beta), \ldots, r_n(\beta)$, symbol $r_{(i)}(\beta)$ is used.

To complete notation, I discuss some purely mathematical notation. As observations and parameters considered here always belong to an Euclidean space $\mathbb{R}^l$, we shall need to define a neighborhood of a point $x \in \mathbb{R}^l$: an open neighborhood (open ball) $U(x, \delta) = \left\{ z \in \mathbb{R}^l \middle| \, \|z - x\| < \delta \right\}$ and analogously a closed neighborhood (closed ball) $\bar{U}(x, \delta) = \{z \in \mathbb{R}^l | \, \|z - x\| \leq \delta\}$. Moreover, let us denote a convex span of $x_1, \ldots, x_m \in \mathbb{R}^l$ by $[x_1, \ldots, x_m]_\varkappa$. Finally, several symbols from linear algebra are introduced: 1_n represents n-dimensional vector of ones, $\mathcal{I}_n$ is the identity matrix of dimension n, and $b_1, \ldots, b_n$ are standard basis vectors of $\mathbb{R}^n$, that is, $b_k = (0, \ldots, 0, 1, 0, \ldots, 0)$.

[3] Since $y_i = h(x_i, \beta) + \varepsilon_i$ and $r_i = y_i - h(x_i, \beta) = h\big(x_i, \beta^0\big) - h(x_i, \beta) + \varepsilon_i$, regression residuals can be written as a function of β and $\omega \in \Omega = \Omega_x \times \Omega_\varepsilon$.

A Lemmas on order statistics and LTS objective function

Proof of Proposition 2.1: We prove the proposition just for the case of the lower bound, $b(z)$, the other case can be derived similarly. The cumulative distribution function of $x_{\max} = \max_{i=1,\dots,n} x_i$ is $F_n(z) = F^n(z)$. We want to show that for any $\varepsilon > 0$ there is $K > 0$ such that $P(x_{\max} > K\sqrt[4]{n}) = 1 - F_n(K\sqrt[4]{n}) < \varepsilon$. This is equivalent to the assertion that $F_n(K\sqrt[4]{n}) \to 1$ as $K \to +\infty$ uniformly for $n > n_0$ and some n_0. Because $b(z) < F(z)$, it also holds $b^n(z) < F^n(z) = F_n(z)$, and thus, it is enough to verify that $b^n(K\sqrt[4]{n}) \to 1$ as $K \to +\infty$ uniformly for $n > n_0$. In general, $P_4(z) = a_1z^4 + a_2z^3 + a_3z^2 + a_4z + a_5$ and its leading coefficient a_1 has to be positive—otherwise, $b(z) > 1$ for z large enough and it could not be a lower bound to a distribution function, which is at most equal to one. So, let us assume without loss of generality that $P_4(z) = z^4$ and $b(z) = 1 - \frac{1}{z^4}$. Hence,

$$b^n\left(K\sqrt[4]{n}\right) = \left(1 - \frac{1}{Kn}\right)^n = \left[\left(1 - \frac{1}{Kn}\right)^{Kn}\right]^{\frac{1}{K}} \to \left(\frac{1}{e}\right)^{\frac{1}{K}} = \sqrt[K]{\frac{1}{e}},$$

that is, $b^n\left(K\sqrt[4]{n}\right)$ converges monotonically to a positive number smaller than one for a fixed $K > 0$. Moreover, this number $\frac{1}{\sqrt[K]{e}}$ as well as $b^n\left(K\sqrt[4]{n}\right)$ increase with K. Therefore, we can find $n_0 > 0$ such that $b^n\left(K\sqrt[4]{n}\right) > \sqrt[K]{\frac{1}{3}}$ for all $n > n_0$ and $K > 1$. Since $\sqrt[K]{\frac{1}{3}} \to 1$ for $K \to +\infty$, also $b^n(K\sqrt[4]{n}) \to 1$ as $K \to +\infty$ uniformly for $n > n_0$. This closes the proof. □

Proof of Lemma 2.2: For a given sample size n, let us consider a fixed realization $\omega \in \Omega^n$. The objective function $S_n(\beta)$ at a particular point $\beta \in B$ equals to one of functions $T_1(\beta), \dots, T_l(\beta)$, where $T_j(\beta) = \sum_{i=1}^{h_n} r^2_{k_{ji}}(\beta)$, $j = 1, \dots, l = \binom{n}{h_n}$, and $\{k_{j1}, \dots, k_{jh_n}\} \in \{1, \dots, n\}^{h_n}$ are sets of h_n indices selecting observations from the sample. Each function $T_j(\beta)$ is uniformly continuous on B and twice differentiable in a neighborhood $U(\beta^0, \delta)$. There are two cases to discuss:

1. If one can find an index j and a neighborhood $U(\beta, \varepsilon)$ such that $S_n(\beta) = T_j(\beta)$ for all $\beta \in U(\beta, \varepsilon)$, $S_n(\beta)$ is continuous at β. Additionally, if $\beta \in U(\beta^0, \delta)$ there is a neighborhood $U(\beta, \varepsilon) \subset U(\beta^0, \delta)$ and $S_n(\beta) = T_j(\beta)$ is even twice differentiable at β.

2. In all other cases, β lies on a boundary in the sense that there are some $j_1, \dots, j_m$ such that $S_n(\beta) = T_{j_1}(\beta) = \dots = T_{j_m}(\beta)$ (that is, some residuals being present in the LTS objective function $S_n(\beta)$ are "switching" their place with those that are not present in the objective function and are all equal at this particular β). Since $S_n(\beta) = T_{j_1}(\beta) = \dots = T_{j_m}(\beta)$ and all functions $T_{j_i}, i = 1, \dots, m$ are continuous at

β, $S_n(\beta)$ is continuous at β as well.
Furthermore, $S_n(\beta)$ is also differentiable provided that $T'_{j_1}(\beta) = \ldots = T'_{j_m}(\beta)$ and $\beta \in U(\beta^0, \delta)$. This condition is always satisfied at $\hat{\beta}_n^{(LTS,h_n)}$ as $T'_{j_1}(\beta) = \ldots = T'_{j_m}(\beta) = 0$; otherwise, $\hat{\beta}_n^{(LTS,h_n)}$ would not minimize $S_n(\beta)$.

Now, consider a fixed $\beta \in U(\beta, \delta)$ (n is still fixed). Assumption D2 implies that $r_i^2(\beta) = \{\varepsilon_i + h(x_i, \beta^0) - h(x_i, \beta)\}^2$ is continuously distributed. Therefore, the probability that any two residuals at a given β are equal is zero:

$$P\left(\Omega_0 = \left\{\omega \in \Omega^n \,\middle|\, \exists i, j \in \{1, \ldots, n\}, i \neq j, \text{ such that } r_i^2(\beta^0, \omega) = r_j^2(\beta^0, \omega)\right\}\right) = 0.$$

Moreover, there is a $\delta' > 0$ such that $r_i(\beta)$ is continuous on $\bar{U}(\beta, \delta')$, and therefore, it is also uniformly continuous on $\bar{U}(\beta, \delta')$, $i = 1, \ldots, n$. Therefore, for any given $\omega \notin \Omega_0$ and $\kappa(\omega) = \frac{1}{2} \min_{i,j=1,\ldots,n; i \neq j} \left|r_i^2(\beta) - r_j^2(\beta)\right| > 0$ we can find an $\varepsilon(\omega) > 0$ such that it holds that $\sup_{\beta' \in U(\beta, \delta')} |r_i^2(\beta') - r_i^2(\beta)| < \kappa(\omega)$ for all $i = 1, \ldots, n$. Consequently, the ordering of residuals $r_1^2(\beta'), \ldots, r_n^2(\beta')$ is constant for all $\beta' \in U(\beta, \delta')$ and there exist j such $S_n(\beta) = T_j(\beta)$ on $U(\beta, \delta')$ almost surely as stated in point 1 ($P(\Omega \backslash \Omega_0) = 1$). Thus, $S_n(\beta)$ is twice differentiable at β almost surely .

Finally, since we just derived that there are almost surely no i and j such that $r_i^2(\beta) = r_j^2(\beta)$ at any $\beta \in B$ and any fixed $n \in \mathbb{N}$ and that $S_n(\beta)$ is almost surely twice differentiable at any $\beta \in U(\beta^0, \delta)$, we can write

$$\begin{aligned}
S_n(\beta) &= \sum_{i=1}^{n} r_i^2(\beta) \cdot I\left(r_i^2(\beta) \leq r_{[h_n]}^2(\beta)\right) \\
S'_n(\beta) = \frac{\partial S_n(\beta)}{\partial \beta} &= -2 \sum_{i=1}^{n} r_i(\beta) h'_\beta(x_i, \beta) \cdot I\left(r_i^2(\beta) \leq r_{[h_n]}^2(\beta)\right) \\
S''_n(\beta) = \frac{\partial^2 S_n(\beta)}{\partial \beta \partial \beta^T} &= 2 \sum_{i=1}^{n} \left\{h'_\beta(x_i, \beta) h'_\beta(x_i, \beta)^T - r_i(\beta) h''_{\beta\beta}(x_i, \beta)\right\} \cdot I\left(r_i^2(\beta) \leq r_{[h_n]}^2(\beta)\right)
\end{aligned}$$

almost surely. □

The next lemma just verifies that the uniform law of large numbers is applicable for LTS-like functions.

Lemma A.1 *Let Assumptions D, H1, and I1 hold and assume that $t(x, \varepsilon; \beta)$ is a real-valued function continuous in β uniformly in x and ε over any compact subset of the support of (x, ε). Moreover, assume that $\mathsf{E} \sup_{\beta \in B} |t(x, \varepsilon; \beta)|^{1+\delta} < \infty$ for some $\delta > 0$. Then*

$$\begin{aligned}
\sup_{\beta \in B, K \in \mathbb{R}} \Bigg| \frac{1}{n} \sum_{i=1}^{n} &\left[t(x_i, \varepsilon_i; \beta) \cdot I\left(r_i^2(\beta) \leq G_\beta^{-1}(\lambda) + K\right)\right] \\
&- \mathsf{E}\left[t(x_i, \varepsilon_i; \beta) \cdot I\left(r_i^2(\beta) \leq G_\beta^{-1}(\lambda) + K\right)\right] \Bigg| \;\to\; 0
\end{aligned}$$

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

Proof: This result is an application of the generic uniform law of large numbers and we use here its variant due to Andrews (1992, Theorem 4).[4] Most of the conditions of the uniform law of large numbers are satisfied trivially or by assumption: (i) the parameter space B is compact by Assumption I1; (ii) differences

$$t(x_i, \varepsilon_i; \beta) \cdot I\big(r_i^2(\beta) \leq G_\beta^{-1}(\lambda) + K\big) - \mathsf{E}\left[t(x_i, \varepsilon_i; \beta) \cdot I\big(r_i^2(\beta) \leq G_\beta^{-1}(\lambda) + K\big)\right]$$

are identically distributed (Assumption D1 and D2) and uniformly integrable because $\mathsf{E} \sup_{\beta \in B} |t(x, \varepsilon; \beta)|^{1+\delta}$ is finite for some $\delta > 0$ (see Davidson, 1994, Theorem 12.10); and (iii) finally, the pointwise convergence of

$$\frac{1}{n} \sum_{i=1}^{n} \left[t(x_i, \varepsilon_i; \beta) \cdot I\big(r_i^2(\beta) \leq G_\beta^{-1}(\lambda) + K\big)\right] - \mathsf{E}\left[t(x_i, \varepsilon_i; \beta) \cdot I\big(r_i^2(\beta) \leq G_\beta^{-1}(\lambda) + K\big)\right] \xrightarrow{P} 0$$

at any $\beta \in B$ and $K \in \mathbb{R}$ follows from the weak law of large numbers for mixingales due to Andrews (1988) (any mixing sequence forms a mixingale, and moreover, the differences $d(x_i, \varepsilon_i; \beta, K)$ are $L^{1+\delta}$-bounded, see Andrews (1988) for more details).

Therefore, the only assumption of Andrews (1992, Theorem 4) which remains to be verified is assumption TSE:

$$\lim_{\rho \to 0} P\left(\sup_{\beta \in B, K \in \mathbb{R}} \sup_{\beta' \in U(\beta, \rho), K' \in U(K, \rho)} |t_I(x_i, \varepsilon_i; \beta', K') - t_I(x_i, \varepsilon_i; \beta, K)| > \kappa \right) = 0 \tag{12}$$

for any $\kappa > 0$, where $t_I(x_i, \varepsilon_i; \beta, K) = t(x_i, \varepsilon_i, \beta) \cdot I\big(r_i^2(\beta) \leq G_\beta^{-1}(\lambda) + K\big)$. To simplify the notation, we write only suprema only with the respective variables β, K, β', K' without the corresponding sets $B, \mathbb{R}, U(\beta, \rho), U(K, \rho)$, respectively, which are fixed throughout the proof. First, note that it holds for all $\beta \in B$ and $K \in \mathbb{R}$

$$\begin{aligned}
& \sup_{\beta, K} \sup_{\beta', K'} |t_I(x_1, \varepsilon_1; \beta', K') - t_I(x_1, \varepsilon_1; \beta, K)| \\
\leq\ & \sup_{\beta, K} \sup_{\beta', K'} \left| t(x_1, \varepsilon_1; \beta') \cdot \left[I\big(r_1^2(\beta') \leq G_{\beta'}^{-1}(\lambda) + K'\big) - I\big(r_1^2(\beta) \leq G_\beta^{-1}(\lambda) + K\big) \right] \right| && (13) \\
+\ & \sup_{\beta, K} \sup_{\beta', K'} \left| [t(x_1, \varepsilon_1; \beta') - t(x_1, \varepsilon_1; \beta)] \cdot I\big(r_1^2(\beta) \leq G_\beta^{-1}(\lambda) + K\big) \right| && (14)
\end{aligned}$$

Hence, we can verify assertion (12) by proving it for expressions (13) and (14). For a given $\varepsilon > 0$, we find $\rho_0 > 0$ such that the probabilities of these two expression exceeding given $\kappa > 0$ are smaller than ε for all $\rho < \rho_0$.

[4] For some function we apply this lemma to, namely to those forming a VC class, the result directly follows from Yu (1994).

1. Let us start with (13). First, observe that

$$\begin{aligned} & \sup_{\beta,K}\sup_{\beta',K'}\left|t(x_1,\varepsilon_1;\beta')\cdot\left[I\left(r_1^2(\beta')\le G_{\beta'}^{-1}(\lambda)+K'\right)-I\left(r_1^2(\beta)\le G_{\beta}^{-1}(\lambda)+K\right)\right]\right| \\ \le\ & \sup_{\beta\in B}\left|t(x_1,\varepsilon_1;\beta)\right|\cdot\sup_{\beta,K}\sup_{\beta',K'}\left|I\left(r_1^2(\beta')\le G_{\beta'}^{-1}(\lambda)+K'\right)-I\left(r_1^2(\beta)\le G_{\beta}^{-1}(\lambda)+K\right)\right| \quad (15) \end{aligned}$$

where $\sup_{\beta\in B}|t(x_1,\varepsilon_1;\beta)|$ is a function independent of β possessing a finite expectation. Because the difference $\left|I\left(r_1^2(\beta')\le G_{\beta'}^{-1}(\lambda)+K'\right)-I\left(r_1^2(\beta)\le G_{\beta}^{-1}(\lambda)+K\right)\right|$ is always lower or equal to one, (13) has an integrable majorant independent of β. Therefore, if we show that the probability

$$\lim_{\rho\to 0}P\left(\sup_{\beta,K}\sup_{\beta',K'}\left|I\left(r_1^2(\beta')\le G_{\beta'}^{-1}(\lambda)+K'\right)-I\left(r_1^2(\beta)\le G_{\beta}^{-1}(\lambda)+K\right)\right|=1\right)=0, \quad (16)$$

it implies, that (15) converges in probability to zero for $\rho\to 0$ and $n\to\infty$ as well.
Second, let us derive an intermediate result regarding the convergence of distribution function $G_{\beta'}$ to G_β. Assumption H1 states that $r_1^2(\beta')\to r_1^2(\beta)$ for $\beta'\to\beta$ uniformly over any compact subset of the support of x, that is, $r_1^2(\beta')\to r_1^2(\beta)$ for $\beta'\to\beta$ in probability uniformly on B. Recalling that $G_\beta(x)$ is the cumulative distribution function of $r_1^2(\beta)$, it follows that $G_{\beta'}(x)\to G_\beta(x)$ for all $x\in\mathbb{R}$ (convergence in distribution) uniformly on B because $G_\beta(x)$ is an absolutely continuous distribution function. The absolute continuity of G_β also implies that $G_{\beta'}^{-1}(\lambda)$ to $G_\beta^{-1}(\lambda)$ uniformly on B.
Third, given the uniform convergence result of the previous paragraph, we can find some $\rho_1>0$ such that $\left|G_{\beta'}^{-1}(\lambda)+K'-G_\beta^{-1}(\lambda)-K\right|<\frac{\varepsilon}{8M_{gg}}$ for any $\beta\in B$, $\beta'\in U(\beta,\rho_1)$, and $K'\in U(K,\rho_1)$, where M_{gg} is the uniform upper bound for the probability density functions of $r_1^2(\beta)$ (Assumption D3). Further, we can find a compact subset $\Omega_1\subset\Omega, P(\Omega_1)>1-\frac{\varepsilon}{2}$, and corresponding $\rho_2>0$ such that $\sup_{\beta,\beta'}|r_1^2(\beta',\omega)-r_1^2(\beta,\omega)|<\frac{\varepsilon}{8M_{gg}}$ for all $\omega\in\Omega_0$ and $\rho<\rho_2$ (Assumption H1). Hence, setting $\rho_0=\min\{\rho_1,\rho_2\}$, it follows that

$$\begin{aligned} & P\left(\sup_{\beta,K}\sup_{\beta',K'}\left|I\left(r_1^2(\beta')\le G_{\beta'}^{-1}(\lambda)+K'\right)-I\left(r_1^2(\beta)\le G_{\beta}^{-1}(\lambda)+K\right)\right|=1\right) \\ \le\ & \frac{\varepsilon}{2}+P\left(\exists\beta\in B: r_1^2(\beta)\in\left(G_\beta^{-1}(\lambda)-\frac{\varepsilon}{4M_{gg}},G_\beta^{-1}(\lambda)+\frac{\varepsilon}{4M_{gg}}\right)\right) \\ \le\ & \frac{\varepsilon}{2}+\frac{2\varepsilon}{4M_{gg}}\cdot M_{gg}=\varepsilon \end{aligned}$$

for any $\rho<\rho_0$ because M_{gg} is the uniform upper bound for the probability density functions of $r_1^2(\beta)$ over all $\beta\in B$. Thus, we have proved (16), and consequently, we have verified that the expectation of (13) converges to zero for $\rho\to 0$ in probability.

2. We should deal now with (14) and prove that for any given $\kappa > 0$

$$\lim_{\rho\to 0} P\left(\sup_{\beta,K}\sup_{\beta',K'}\left|[t(x_1,\varepsilon_1;\beta') - t(x_1,\varepsilon_1;\beta)]\cdot I\big(r_1^2(\beta)\le G_\beta^{-1}(\lambda)+K\big)\right| > \kappa\right) = 0. \qquad (17)$$

First, note that the difference

$$|t(x_1,\varepsilon_1;\beta') - t(x_1,\varepsilon_1;\beta)| \le |t(x_1,\varepsilon_1;\beta')| + |t(x_1,\varepsilon_1;\beta)| \le 2\sup_{\beta\in B}|t(x_1,\varepsilon_1;\beta)|$$

can be bounded from above by a function that is independent of β and has a finite expectation, as follows from the assumptions of this lemma. Let $2\,\mathsf{E}\sup_{\beta\in B}|t(x_1,\varepsilon_1;\beta)| = U_E$. Second, for an arbitrary fixed $\varepsilon > 0$, we can find a compact subset A_ε of the support of (x_1,ε_1) (and its complement $\overline{A_\varepsilon}$) such that $P((x_1,\varepsilon_1)\in A_\varepsilon) > 1 - \frac{\kappa\varepsilon}{2U_E}$ (both x_1 and ε_1 are random variables with finite second moments) and $2\int_{\overline{A_\varepsilon}}\sup_{\beta\in B}|t(x_1,\varepsilon_1;\beta)| < \frac{\kappa\varepsilon}{2}$. Given this set A_ε and $\beta\in B$, we can employ continuity of $t(x_1,\varepsilon_1;\beta)$ in β (uniform over all $(x_1,\varepsilon_1)\in A_\varepsilon$) and find a $\rho_0 > 0$ such that

$$\sup_{(x_1,\varepsilon_1)\in A_\varepsilon}\sup_{\beta,\beta'}|t(x_1,\varepsilon_1;\beta') - t(x_1,\varepsilon_1;\beta)| < \frac{\kappa\varepsilon}{2}.$$

Hence,

$$\begin{aligned}
\mathsf{E}\left\{\sup_{\beta,\beta'}|t(x_1,\varepsilon_1;\beta') - t(x_1,\varepsilon_1;\beta)|\right\} &\le \int_{\overline{A_\varepsilon}} 2\sup_{\beta\in B}|t(x_1,\varepsilon_1;\beta)|\,\mathrm{d}F_x(x_1)\mathrm{d}F_\varepsilon(\varepsilon_1)\\
&+ \int_{A_\varepsilon}\frac{\kappa\varepsilon}{2}\mathrm{d}F_x(x_1)\mathrm{d}F_\varepsilon(\varepsilon_1)\\
&\le \frac{\kappa\varepsilon}{2}+\frac{\kappa\varepsilon}{2} = \kappa\varepsilon,
\end{aligned}$$

and consequently,

$$\begin{aligned}
& P\left(\sup_{\beta,K}\sup_{\beta',K'}\left|[t(x_1,\varepsilon_1;\beta') - t(x_1,\varepsilon_1;\beta)]\cdot I\big(r_1^2(\beta)\le G_\beta^{-1}(\lambda)+K\big)\right| > \kappa\right)\\
\le\ & \frac{1}{\kappa}\,\mathsf{E}\left[\sup_{\beta,K}\sup_{\beta',K'}\left|[t(x_1,\varepsilon_1;\beta') - t(x_1,\varepsilon_1;\beta)]\cdot I\big(r_1^2(\beta)\le G_\beta^{-1}(\lambda)+K\big)\right|\right]\\
\le\ & \kappa\varepsilon/\kappa = \varepsilon
\end{aligned}$$

for any $\rho < \rho_0$. Hence, we have verified that (17).

Thus, the assumption TSE of Andrews (1992), is valid as well and the claim of this lemma follows from the uniform weak law of large numbers. □

The following assertions present some fundamental properties of order statistics of regression residuals.

Lemma A.2 *Let $\lambda \in \left\langle \frac{1}{2}, 1 \right\rangle$ and put $h_n = [\lambda n]$ for $n \in \mathbb{N}$. Under Assumptions D, H1, and I1, it holds that*

$$\sup_{\beta \in B} \left| r^2_{[h_n]}(\beta) - G^{-1}_\beta(\lambda) \right| \to 0 \tag{18}$$

as $n \to +\infty$ in probability, and consequently,

$$E_{Gn} = \mathsf{E} \sup_{\beta \in B} \left| r^2_{[h_n]}(\beta) - G^{-1}_\beta(\lambda) \right| \to 0 \tag{19}$$

as $n \to +\infty$.

Proof: Let us recall that $r^2_i(\beta) \equiv [\varepsilon_i + h(x_i, \beta^0) - h(x_i, \beta)]^2 \sim G_\beta$. Further, let us take an arbitrary $K_1 > 0$, set $K_\varepsilon = K_1 \cdot m_{gg}$ (see Assumption D3 for definition of m_{gg}), and consider some $\varepsilon \in (0, 1)$. For any choice of ε, we find $n_0 \in \mathbb{N}$ such that for all $n > n_0$

$$P\left(\sup_{\beta \in B} \left| r^2_{[h_n]}(\beta) - G^{-1}_\beta(\lambda) \right| > K_1 \right) < \varepsilon, \tag{20}$$

which proves the lemma. Without loss of generality, we can assume that $K_1 < \delta_g$, where δ_g comes from Assumption D3.

First, denote

$$v_{1i}(\beta, K_1) = I\big(r^2_i(\beta) \le G^{-1}_\beta(\lambda) + K_1\big).$$

As it holds for all $\beta \in B$ and $i = 1, \ldots, n$

$$\mathsf{E}\, v_{1i}(\beta, K_1) = P(v_{1i}(\beta, K_1) = 1) = P\big(r^2_i(\beta) \le G^{-1}_\beta(\lambda) + K_1\big) \ge \lambda,$$

it follows that $\mathsf{E}\, v_{1i}(\beta, K_1) \in (\lambda, 1\rangle$. Further, Lemma A.1 for choice $t(x, e, \beta) = 1$ guarantees that we can use the weak law of large numbers for $v_{1i}(\beta, K_1)$ uniformly on $B \times \mathbb{R}_+$. Hence,

$$\sup_{\beta \in B, K_1 \in \mathbb{R}_+} \left| \frac{1}{n} \sum_{i=1}^n \{\nu_{1i}(\beta, K_1) - \mathsf{E}\, \nu_{1i}(\beta, K_1)\} \right| \to 0$$

in probability. Consequently, we can find some n_0 such that it holds for all $n > n_0$

$$P\left(\sup_{\beta \in B, K_1 \in \mathbb{R}_+} \left| \frac{1}{n} \sum_{i=1}^n \{v_{1i}(\beta, K_1) - \mathsf{E}\, v_{1i}(\beta, K_1)\} \right| \le \frac{1}{2} K_\varepsilon \right) > 1 - \frac{\varepsilon}{2}.$$

Thus, it holds uniformly in β and K_1 with probability greater or equal to $1 - \varepsilon/2$

$$-\frac{1}{2} K_\varepsilon + \sum_{i=1}^n \mathsf{E}\, v_{1i}(\beta, K_1) \le \sum_{i=1}^n v_{1i}(\beta, K_1). \tag{21}$$

Second, because $K_1 < \delta_g$, Assumption D3 implies $\mathsf{E}\, v_{1i}(\beta, K_1) > \lambda + K_1 \cdot m_{gg} = \lambda + K_\varepsilon$

for all $\beta \in B$ and $K_1 < \delta_g$. This result together with equation (21) implies that

$$n\lambda + (n - \frac{1}{2})K_\varepsilon = -\frac{1}{2}K_\varepsilon + n(\lambda + K_\varepsilon) < -\frac{1}{2}K_\varepsilon + \sum_{i=1}^{n} \mathsf{E}\, v_{1i}(\beta, K_1) \leq \sum_{i=1}^{n} v_{1i}(\beta, K_1).$$

But this means for all $\beta \in B$ that at least $n\lambda \geq h_n$ of residuals $r_i^2(\beta)$ are smaller than $G_\beta^{-1}(\lambda) + K_1$. In other words, $r_{[h_n]}^2(\beta) \leq G_\beta^{-1}(\lambda) + K_1$ with probability at least $1 - \varepsilon/2$.

The corresponding lower inequality, holding also with probability at least $1 - \varepsilon/2$, can be found by repeating these steps for

$$v_{2i}(\beta, K_1) = I\big(r_i^2(\beta) \geq G_\beta^{-1}(\lambda) - K_1\big).$$

Finally, combining these two inequalities results in (18). Since $r_i^2(\beta)$ is uniformly integrable due to Assumption H5 and Davidson (1994, Theorem 12.10), $r_{[h_n]}^2(\beta)$ is uniformly integrable as well and the second claim follows directly from the (18) by Davidson (1994, Theorem 18.14), which shows that the convergence in probability of uniformly integrable random variables implies the convergence in L^p-norm. □

Lemma A.3 *Let $\lambda \in \left\langle \frac{1}{2}, 1 \right\rangle$ and put $h_n = [\lambda n]$ for $n \in \mathbb{N}$. Under Assumptions D, H, and I1, there is some $\varepsilon > 0$ such that*

$$\sqrt{n} \sup_{\beta \in U(\beta^0, \varepsilon)} \left| r_{[h_n]}^2(\beta) - G_\beta^{-1}(\lambda) \right| = \mathcal{O}_p(1)$$

and

$$E_{Ln} = \mathsf{E}\left\{ \sqrt{n} \sup_{\beta \in U(\beta^0, \varepsilon)} \left| r_{[h_n]}^2(\beta) - G_\beta^{-1}(\lambda) \right| \right\} = \mathcal{O}(1)$$

for $n \to +\infty$.

Proof: The proof has a rather similar structure to the proof of Lemma A.2. First, let us take a fixed $\varepsilon \in (0, 1)$, an arbitrary $K_1 > 0$, and set $K_\varepsilon = K_1 \cdot m_g$. Further, denote

$$v_{1i}(\beta, K_1) = I\Big(r_i^2(\beta) \leq G_\beta^{-1}(\lambda) + n^{-\frac{1}{2}} K_1\Big).$$

As it holds for all $\beta \in B$ and $i = 1, \ldots, n$

$$\mathsf{E}\, v_{1i}(\beta, K_1) = P(v_{1i}(\beta, K_1) = 1) = P\Big(r_i^2(\beta) \leq G_\beta^{-1}(\lambda) + n^{-\frac{1}{2}} K_1\Big) \geq \lambda,$$

it follows that $\mathsf{E}\, v_{1i}(\beta, K_1) \in (\lambda, 1\rangle$.

Now, Assumption H3 and van der Vaart and Wellner (1996, Lemmas 2.6.15 and 2.6.18) imply that $\{v_{1i}(\beta, K_1); \beta \in U(\beta^0, \delta), K_1 \in \mathbb{R}\}$ form a VC class, which is uniformly bounded

by 1. Because of Assumption D1 on the mixing coefficients, we can apply the uniform central limit theorem of Arcones and Yu (1994) to see that

$$\left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\nu_{1i}(\beta, K_1) - \mathsf{E}\,\nu_{1i}(\beta, K_1)\} : \beta \in U(\beta^0, \delta), K_1 > 0 \right\}$$

converges in distribution to a Gaussian processes with uniformly bounded and uniformly continuous pathes. Consequently, we can find some $\varepsilon > 0$ and a constant $U > 0$

$$\sup_{n\in\mathbb{N}} \mathsf{E} \sup_{\beta\in U(\beta^0,\varepsilon), K_1>0} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (v_{1i}(\beta, K_1) - \mathsf{E}\, v_{1i}(\beta, K_1)) \right|^2 < U$$

(functions $v_{1i}(\beta, K_1)$ are bounded). By the Chebyshev inequality $P(|X| > K) \leq \mathsf{E}\,|X|^p / K^p$, it finally follows that

$$P\left(\sup_{\beta\in U(\beta^0,\varepsilon), K_1>0} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (v_{1i}(\beta, K_1) - \mathsf{E}\, v_{1i}(\beta, K_1)) \right| > \frac{1}{2} K_\varepsilon \right) < \frac{4U}{K_\varepsilon^2}.$$

Thus, it holds uniformly in $\beta \in U(\beta^0, \varepsilon)$ with probability greater or equal to $1 - 4U/K_\varepsilon^2$

$$-\frac{1}{2}\sqrt{n} \cdot K_\varepsilon + \sum_{i=1}^{n} \mathsf{E}\, v_{1i}(\beta, K_1) \leq \sum_{i=1}^{n} v_{1i}(\beta, K_1). \tag{22}$$

Further, we can find n_0 such that $n^{-\frac{1}{2}} K_1 < \delta_g$ for all $n > n_0$ (δ_g comes from Assumption D3), and thus, $\mathsf{E}\, v_{1i}(\beta, K_1) > \lambda + n^{-\frac{1}{2}} K_1 \cdot m_g = \lambda + n^{-\frac{1}{2}} K_\varepsilon$ for all $\beta \in U(\beta^0, \varepsilon)$ and $n > n_0$. This result together with equation (22) imply that

$$n\lambda + \frac{1}{2}\sqrt{n} K_\varepsilon = -\frac{1}{2}\sqrt{n} K_\varepsilon + n\lambda + \sqrt{n} K_\varepsilon < -\frac{1}{2}\sqrt{n} K_\varepsilon + \sum_{i=1}^{n} \mathsf{E}\, v_{1i}(\beta) \leq \sum_{i=1}^{n} v_{1i}(\beta).$$

But this means for all $\beta \in U(\beta^0, \varepsilon)$ that at least $n\lambda \geq h_n$ of residuals $r_i^2(\beta)$ are smaller than $G_\beta^{-1}(\lambda) + n^{-\frac{1}{2}} K_\varepsilon$. In other words, $r_{[h_n]}^2(\beta) \leq G_\beta^{-1}(\lambda) + n^{-\frac{1}{2}} K_\varepsilon$ on $U(\beta^0, \varepsilon)$ with probability at least $1 - 4U/K_\varepsilon^2$. The corresponding lower inequality can be found by repeating these steps for

$$v_{2i}(\beta, K_1) = I\Big(r_i^2(\beta) \geq G_\beta^{-1}(\lambda) - n^{-\frac{1}{2}} K_1 \Big).$$

These inequalities can be rewritten as $Z_n = \sup_{\beta\in U(\beta^0,\varepsilon)} n^{-\frac{1}{2}} \left| r_{[h_n]}^2(\beta) - G_\beta^{-1}(\lambda) \right| \leq K_\varepsilon$, which holds with probability $1 - 4U/K_\varepsilon^2$. Thus, for any $\varepsilon > 0$ we find $K_\varepsilon = 1 + \sqrt{4U/\varepsilon}$ such that $P(Z_n(\beta) \leq K_\varepsilon) > 1 - \varepsilon$, so $Z_n = \mathcal{O}_p(1)$. Furthermore, denoting the cumulative

distribution function of Z_n by $F_{z,n}$, the expectation

$$\mathsf{E}\, Z_n = \int_0^\infty [1 - F_{z,n}(x)]dx \leq 1 + \int_1^\infty \frac{4U}{x^2}dx = 1 + 4U$$

is finite. □

Lemma A.4 *Let Assumptions D, H, and I1 be satisfied. Moreover, let* $\lambda \in \left(\frac{1}{2}, 1\right\rangle$, $\tau \in \left(\frac{1}{2}, 1\right)$, *and put* $h_n = [\lambda n]$ *for* $n \in \mathbb{N}$. *Then* $\left| r^2_{[h_n]}\left(\beta^0 - n^{-\frac{1}{2}}t\right) - r^2_{[h_n]}\left(\beta^0\right)\right| = \mathcal{O}_p(n^{-\tau})$ *uniformly in* $t \in \mathcal{T}_M = \{t \in \mathbb{R}^k | \, \|t\| \leq M\}$ *as* $n \to +\infty$.

Proof: As the first and main step, we show that for any $\varepsilon \in (0, 1)$ there exist K_ε and n_ε such that uniformly in $t \in \mathcal{T}_M$ for all $n > n_\varepsilon$

$$P\left(\left| r_{[h_n]}\left(\beta^0 - n^{-\frac{1}{2}}t\right) - r_{[h_n]}\left(\beta^0\right)\right| < n^{-\tau} \cdot K_\varepsilon\right) > 1 - \varepsilon \tag{23}$$

(please, remember the convention introduced in the introduction of Appendix that $r_{[h]}(\beta) = \operatorname{sgn} r_{[h]}(\beta) \cdot \sqrt{r^2_{[h]}(\beta)}$, whereas the order statistics of residuals $r_i(\beta)$ is referred by $r_{(h)}(\beta)$). Additionally, note that assuming $n_\varepsilon^{-\frac{1}{2}} K_\varepsilon < \delta$ (δ comes from Assumptions H), the Taylor expansion leads to

$$r_i(\beta^0 - n^{-\frac{1}{2}}t) = r_i(\beta^0) + h'_\beta(x_i, \xi)^T \cdot n^{-\frac{1}{2}}t, \tag{24}$$

where $\xi \in \left[\beta^0, \beta^0 - n^{-\frac{1}{2}}t\right]$.

Now, all assertions in the following part of the proof are meant conditionally on values of x_i. Let us suppose that $h'_\beta(x_i, \xi)^T t \geq 0$ for a given i (the other case can be analyzed analogously). Then $r_i(\beta^0) + h'_\beta(x_i, \xi)^T n^{-\frac{1}{2}}t \geq r_i(\beta^0)$ which means that all such residuals $r_i(\beta^0 - n^{-\frac{1}{2}}t)$ are larger then residuals $r_i(\beta^0) \equiv \varepsilon_i$. In other words, some residuals evaluated at point $\beta^0 - n^{-\frac{1}{2}}t$ compared to β^0 are shifted out of interval $\left\langle -r_{[h_n]}\left(\beta^0\right), r_{[h_n]}\left(\beta^0\right)\right\rangle$ on its right hand side and some are shifted into it on its left hand side. The assertion (23) can be proved in the following way: considering a bit larger interval $\left\langle -r_{[h_n]}\left(\beta^0\right) - n^{-\tau}K_1,\right.$ $\left. r_{[h_n]}\left(\beta^0\right) + n^{-\tau}K_1\right\rangle$, it is to be shown that such an interval contains at least h_n residuals $r_i(\beta^0 - n^{-\frac{1}{2}}t)$ for some sufficiently large constant K_1. To do so, we shall try to find a number m_1 of indices $i = 1, \ldots, n$ for which (with a probability close to 1)

$$r_i^2(\beta^0) \leq r^2_{[h_n]}(\beta^0) \quad \text{and} \quad r_i(\beta^0 - n^{-\frac{1}{2}}t) \geq r_{[h_n]}\left(\beta^0\right) + n^{-\tau}K_1. \tag{25}$$

Such indices represent the observations that decrease the number of residuals inside the interval $\left\langle -r_{[h_n]}\left(\beta^0\right) - n^{-\tau}K_1, r_{[h_n]}\left(\beta^0\right) + n^{-\tau}K_1\right\rangle$. Similarly, we try to find a number m_2 of indices $i = 1, \ldots, n$ for which (with a probability close to 1)

$$r_i(\beta^0) \leq -r_{[h_n]}(\beta^0) \quad \text{and} \quad r_i(\beta^0 - n^{-\frac{1}{2}}t) \geq -r_{[h_n]}\left(\beta^0\right) - n^{-\tau}K_1. \tag{26}$$

These indices correspond to the observations that were not in the interval $\left\langle -r_{[h_n]}\left(\beta^0\right), r_{[h_n]}\left(\beta^0\right)\right\rangle$ before but they move inside the interval $\left\langle -r_{[h_n]}\left(\beta^0\right) - n^{-\tau}K_1, r_{[h_n]}\left(\beta^0\right) + n^{-\tau}K_1\right\rangle$, and thus, increase the number of residuals contained in it. Since there are just h_n indices among all $i = 1, \ldots, n$ satisfying $r_i^2(\beta^0) \leq r_{[h_n]}^2(\beta^0)$, the number of indices such that $r_i^2(\beta^0 - n^{-\frac{1}{2}}t) \leq r_{[h_n]}^2\left(\beta^0\right) + n^{-\tau}K_1$ equals $h_n - m_1 + m_2$. Therefore, all we have to do is to verify that the difference $m_2 - m_1$ is positive with probability close to 1.

Using (24), case (25) is equivalent to

$$r_{[h_n]}(\beta^0) + n^{-\tau}K_1 - h_\beta^{'}(x_i, \xi)^T \cdot n^{-\frac{1}{2}}t \leq e_i \leq r_{[h_n]}(\beta^0).$$

Similarly, (26) is valid if and only if

$$-r_{[h_n]}(\beta^0) - n^{-\tau}K_1 - h_\beta^{'}(x_i, \xi)^T \cdot n^{-\frac{1}{2}}t \leq e_i \leq -r_{[h_n]}(\beta^0).$$

Thus, it seems to be helpful to study the probability of the events $z \pm n^{-\tau}K_1 - h_\beta^{'}(x_i, \xi)^T \cdot n^{-\frac{1}{2}}t \leq e_i \leq z$ for some $z \in \mathbb{R}$. This probability can be expressed by means of the distribution function $F(x)$ (remember, everything till now is conditional on x_i):

$$F(z) - F\left(z \pm n^{-\tau}K_1 - h_\beta^{'}(x_i, \xi)^T \cdot n^{-\frac{1}{2}}t\right) = \int_{z \pm n^{-\tau}K_1 - h_\beta^{'}(x_i,\xi)^T \cdot n^{-\frac{1}{2}}t}^{z} f(t)dt.$$

Expanding the density in the integral, $f(t) = f(z) + f'(\zeta_t)t$, we get

$$\begin{aligned}
\int_{z \pm n^{-\tau}K_1 - h_\beta^{'}(x_i,\xi)^T \cdot n^{-\frac{1}{2}}t}^{z} f(t)dt &\geq f(z)\left[\pm n^{-\tau}K_1 - h_\beta^{'}(x_i, \xi)^T \cdot n^{-\frac{1}{2}}t\right] + \\
&\quad + L_f\left[\pm n^{-\tau}K_1 - h_\beta^{'}(x_i, \xi)^T \cdot n^{-\frac{1}{2}}t\right]^2 \\
\int_{z \pm n^{-\tau}K_1 - h_\beta^{'}(x_i,\xi)^T \cdot n^{-\frac{1}{2}}t}^{z} f(t)dt &\leq f(z)\left[\pm n^{-\tau}K_1 - h_\beta^{'}(x_i, \xi)^T \cdot n^{-\frac{1}{2}}t\right] + \\
&\quad + U_f\left[\pm n^{-\tau}K_1 - h_\beta^{'}(x_i, \xi)^T \cdot n^{-\frac{1}{2}}t\right]^2,
\end{aligned}$$

which results in

$$F(z) - F\left(z \pm n^{-\tau}K_1 - h_\beta^{'}(x_i, \xi)^T \cdot n^{-\frac{1}{2}}t\right) = f(z)\left[\pm n^{-\tau}K_1 - h_\beta^{'}(x_i, \xi)^T \cdot n^{-\frac{1}{2}}t\right] + \mathcal{O}\left(n^{-1}\right). \tag{27}$$

Having these results in hand, the same idea as in the previous Lemmas A.2 and A.3 can be used. Let us consider for a fixed $z \in \mathbb{R}$

$$w_{1i}(z) = I\left(z + n^{-\tau}K_1 - h_\beta^{'}(x_i, \xi)^T \cdot n^{-\frac{1}{2}}t \leq e_i \leq z\right)$$

and

$$w_{2i}(z) = I\Big(-z - n^{-\tau}K_1 - h'_\beta(x_i,\xi)^T \cdot n^{-\frac{1}{2}}t \le e_i \le -z\Big).$$

Apparently, $m_2 - m_1 = \sum_{i=1}^n (w_{2i}(z) - w_{1i}(z))$ for $z = r_{[h_n]}(\beta^0)$. Let us denote $s_i(z) = w_{2i}(z) - w_{1i}(z)$. Employing (27), we obtain $\mathsf{E}\, s_i(z) = n^{-\tau}K_1 \cdot (f(z) + f(-z)) + \mathcal{O}(n^{-1})$ and hence also $\mathsf{var}\, s_i(z) = 2n^{-\tau}K_1 \cdot (f(z) + f(-z)) + \mathcal{O}(n^{-1})$. Note that both moments do not depend on x_i apart from term $\mathcal{O}(n^{-1})$. The Feller-Lindeberg conditions for the central limit theorem can be easily verified, and thus, two constants K_ε and n_ε such that for all $n \ge n_\varepsilon$

$$P\left(\left|\frac{\sum_{i=1}^n (s_i(z) - \mathsf{E}\, s_i(z))}{C_n}\right| < K_\varepsilon\right) > 1 - \varepsilon$$

can be found, where $C_n^2 = n^{1-\tau}K_1 \cdot (f(z) + f(-z)) + \mathcal{O}(1)$. Further, it follows that with probability greater than $1 - \varepsilon$

$$\sum_{i=1}^n s_i(z) \ge -n^{\frac{1}{2}(1-\tau)}\sqrt{K_1 \cdot (f(z) + f(-z))}K_\varepsilon + n^{1-\tau}K_1 \cdot (f(z) + f(-z)) + \mathcal{O}(1). \quad (28)$$

As $\tau \in \left(\frac{1}{2}, 1\right)$, the last expression increases in n above all limits for a given K_1 because $f(z)$ is bounded from above and away from zero as well in a neighborhood of $G^{-1}(\lambda)$. Thus, we can find n_ε such that for all $n > n_\varepsilon$ the right hand side of (28) is positive, and consequently, the number of the residuals $r_i^2\left(\beta^0 - n^{-\frac{1}{2}}t\right)$ that fall to the interval $\left\langle -r_{[h_n]}(\beta^0) - n^{-\tau}K_1, r_{[h_n]}(\beta^0) + n^{-\tau}K_1\right\rangle$ is at least h_n with probability greater than $1 - \varepsilon$. We can conclude that for some $\infty > K_2 > 0$

$$r_{[h_n]}(\beta^0 - n^{-\frac{1}{2}}t) \le r_{[h_n]}(\beta^0) + n^{-\tau} \cdot K_2.$$

This result was derived conditionally on x_i, but the upper bound $r_{[h_n]}(\beta^0) + n^{-\tau} \cdot K_2$ is independent of x_i realizations, which means that it holds not only conditionally on x_i but without conditioning as well. Analogously, the corresponding lower inequality can be derived.

Finally, Lemma A.3 implies that both $r_{[h_n]}(\beta^0 - n^{-\frac{1}{2}}t)$ and $r_{[h_n]}(\beta^0)$ are bounded in probability. Thus, utilizing equality $a^2 - b^2 = (a+b)(a-b)$, we obtain immediately the assertion of this lemma. □

The following lemma and corollaries translate the results on the convergence of the order statistics of residuals to the convergence of the indicators $I\Big(r_i^2(\beta) \le r_{[h_n]}^2(\beta)\Big)$ to $I\big(r_i^2(\beta) \le G_\beta^{-1}(\lambda)\big)$ and their expectations.

Lemma A.5 *Under Assumptions D, H1, and I1, it holds for any $i \leq n$*

$$P_G = P\left(\sup_{\beta\in B}\left|I\big(r_i^2(\beta) \leq r^2_{[h_n]}(\beta)\big) - I\big(r_i^2(\beta) \leq G_\beta^{-1}(\lambda)\big)\right| \neq 0\right) = o(1).$$

Additionally, under Assumptions D, H, and I1, there exists $\varepsilon > 0$ such that

$$P_L = P\left(\sup_{\beta\in U(\beta^0,\varepsilon)}\left|I\big(r_i^2(\beta) \leq r^2_{[h_n]}(\beta)\big) - I\big(r_i^2(\beta) \leq G_\beta^{-1}(\lambda)\big)\right| \neq 0\right) = \mathcal{O}\Big(n^{-\frac{1}{2}}\Big)$$

as $n \to +\infty$.

Proof: To facilitate easier understanding, let us define the difference between indicators

$$\nu_{in}(\beta) = I\big(r_i^2(\beta) \leq r^2_{[h_n]}(\beta)\big) - I\big(r_i^2(\beta) \leq G_\beta^{-1}(\lambda)\big).$$

Without loss of generality, we discuss only the case $v_{in}(\beta) = -1$, which corresponds to $r^2_{[h_n]}(\beta) < r_i^2(\beta) \leq G_\beta^{-1}(\lambda)$. The other case $v_{in}(\beta) = 1$ can be derived analogously. Also notice that $P\big(\sup_{\beta\in B}|\nu_{in}(\beta)|\big) = P(\exists\beta \in B : |\nu_{in}(\beta)| \neq 0)$ because $|\nu_{in}(\beta)| \in \{0, 1\}$.

So, let us consider an event $\omega = (\omega_1, \ldots, \omega_n) \in \Omega^n$ and assume without loss of generality that $i = n$. Given $\omega' = (\omega_1, \ldots, \omega_{n-1}) \in \Omega^{n-1}$ and $\big(r_1^2(\beta, \omega_1), \ldots, r^2_{n-1}(\beta, \omega_{n-1})\big)$

$$r^2_{[h_n]}(\beta, \omega) = \begin{cases} r^2_{[h_n-1]}(\beta, \omega') & \text{if } r_n^2(\beta, \omega_n) < r^2_{[h_n-1]}(\beta, \omega') \\ r_n^2(\beta, \omega_n) & \text{if } r^2_{[h_n-1]}(\beta, \omega') \leq r_n^2(\beta, \omega_n) \leq r^2_{[h_n]}(\beta, \omega') \\ r^2_{[h_n]}(\beta, \omega') & \text{if } r^2_{[h_n]}(\beta, \omega') < r_n^2(\beta, \omega_n) \end{cases} \tag{29}$$

Denoting Ω_1, Ω_2, and Ω_3 subsets of Ω^n corresponding to the three (disjoint) cases in (29), we can write

$$\begin{aligned} P(\{\omega \in \Omega^n | \exists\beta \in B : \nu_{nn}(\beta) = -1\}) &= P(\{\omega \in \Omega_1 | \exists\beta \in B : \nu_{nn}(\beta) = -1\}) \\ &+ P(\{\omega \in \Omega_2 | \exists\beta \in B : \nu_{nn}(\beta) = -1\}) \\ &+ P(\{\omega \in \Omega_3 | \exists\beta \in B : \nu_{nn}(\beta) = -1\}) \end{aligned}$$

and analyze this sum one by one.

1. $P_1 = P(\{\omega \in \Omega_1 | \exists\beta \in B : \nu_{nn}(\beta) = -1\}) \leq$
   $P\Big(\exists\beta \in B : r^2_{[h_n]}(\beta, \omega) < r_1^2(\beta, \omega_1) < r^2_{[h_n]}(\beta, \omega)\Big) = 0.$
2. $P_2 = P(\{\omega \in \Omega_2 | \exists\beta \in B : \nu_{nn}(\beta) = -1\}) =$
   $P\Big(\exists\beta \in B : r^2_{[h_n-1]}(\beta, \omega') \leq r_n^2(\beta, \omega_n) = r^2_{[h_n]}(\beta, \omega) \leq G_\beta^{-1}(\lambda)\Big)$ can be analyzed in exactly the same way as $P(\{\omega \in \Omega_3 | \exists\beta \in B : \nu_{nn}(\beta) = -1\})$, see point 3.

3. $P_3 = P(\{\omega \in \Omega_3 | \exists \beta \in B : \nu_{nn}(\beta) = -1\}) =$
$P\Big(\exists \beta \in B : r^2_{[h_n]}(\beta, \omega') = r^2_{[h_n]}(\beta, \omega) < r^2_n(\beta, \omega_n) \leq G^{-1}_\beta(\lambda)\Big)$. We can structure this last term in the following way (Assumption D3):

$$P\big(\exists \beta \in B : r^2_{[h_n]}(\beta, \omega') < r^2_n(\beta, \omega_n) \leq G^{-1}_\beta(\lambda)\big) = \tag{30}$$

$$= \int_{\omega' \in \Omega^{n-1}} \int_{\omega_n \in \Omega} \sup_{\beta \in B} I\big(r^2_{[h_n]}(\beta, \omega') < r^2_n(\beta, \omega_n) \leq G^{-1}_\beta(\lambda)\big) dP(\omega_1) dP(\omega') \tag{31}$$

$$= \int_{\omega' \in \Omega^{n-1}} M_{gg} \cdot \sup_{\beta \in B} \left| r^2_{[h_n]}(\beta, \omega') - G^{-1}_\beta(\lambda) \right| dP(\omega') \tag{32}$$

$$= M_{gg} \cdot \mathsf{E} \left\{ \sup_{\beta \in B} \left| r^2_{[h_n]}(\beta, \omega') - G^{-1}_\beta(\lambda) \right| \right\}. \tag{33}$$

The first claim of the lemma, $P_G = o(1)$, is then a direct consequence of Lemma A.2. The second result, $P_L = \mathcal{O}\Big(n^{-\frac{1}{2}}\Big)$, can be derived analogously, if we consider only a neighborhood $U(\beta^0, \varepsilon)$ instead of B, write last expectation as

$$n^{-\frac{1}{2}} M_{gg} \cdot \mathsf{E} \left\{ \sqrt{n} \sup_{\beta \in B} \left| r^2_{[h_n]}(\beta, \omega') - G^{-1}_\beta(\lambda) \right| \right\},$$

and employ Lemma A.3. □

Corollary A.6 *Let Assumptions D, H1, and I1 hold and assume that $t(x, \varepsilon; \beta)$ is a real-valued function continuous in β uniformly in x and ε over any compact subset of the support of (x, ε). Moreover, assume that $\mathsf{E} \sup_{\beta \in B} |t(x, \varepsilon; \beta)| < \infty$. Then it holds that*

$$\mathsf{E} \left\{ \sup_{\beta \in B} \left| t(x_i, \varepsilon_i, \beta) \cdot \big[I\big(r^2_i(\beta) \leq r^2_{[h_n]}(\beta)\big) - I\big(r^2_i(\beta) \leq G^{-1}_\beta(\lambda)\big) \big] \right| \right\} = o(1).$$

Additionally, under Assumptions D, H, and I1, there exists $\varepsilon > 0$ such that

$$\mathsf{E} \left\{ \sup_{\beta \in U(\beta^0, \varepsilon)} \left| t(x_i, \varepsilon_i, \beta) \cdot \big[I\big(r^2_i(\beta) \leq r^2_{[h_n]}(\beta)\big) - I\big(r^2_i(\beta) \leq G^{-1}_\beta(\lambda)\big) \big] \right| \right\} = \mathcal{O}\Big(n^{-\frac{1}{2}}\Big)$$

as $n \to +\infty$.

Proof: This can verified along the same lines are Lemma A.5. Defining functions $\nu_{in}(\beta)$ and sets Ω_1, Ω_2, and Ω_3 exactly the same way as in Lemma A.5, we can express the expectation of any random variable $\mathsf{E}\, X$ as $\left\{ \int_{\Omega_1} + \int_{\Omega_2} + \int_{\Omega_3} \right\} x dF(x)$. By the same argument as in Lemma A.5, we will treat only part concerning $\int_{\Omega_3}$ and assume without loss of generality that $i = n$. Analogously to (30)–(33), we can write

$$\mathsf{E} \left\{ \sup_{\beta \in B} |t(x_n, \varepsilon_n, \beta) \cdot \nu_{in}(\beta)| \right\}$$

$$
\begin{aligned}
&\leq \int_{\Omega_3} \left\{ \sup_{\beta\in B} |t(x_n,\varepsilon_n,\beta)| \cdot \sup_{\beta\in B} |\nu_{in}(\beta)| \right\} dP(\omega) \\
&= \int_{\omega'\in\Omega^{n-1}} \int_{\omega_n\in\Omega} \sup_{\beta\in B} |t(x_n,\varepsilon_n,\beta)| \cdot \sup_{\beta\in B} I\big(r^2_{[h_n]}(\beta,\omega') < r^2_n(\beta,\omega_n) \leq G^{-1}_\beta(\lambda)\big) dP(\omega_n) dP(\omega') \\
&\leq \int_{\omega_n\in\Omega} \sup_{\beta\in B} |t(x_n,\varepsilon_n,\beta)| \cdot \int_{\omega'\in\Omega^{n-1}} \sup_{\beta\in B} I\big(r^2_{[h_n]}(\beta,\omega') < r^2_n(\beta,\omega_n) \leq G^{-1}_\beta(\lambda)\big) dP(\omega') dP(\omega_n) \\
&\leq M_{gg} \cdot \int_{\omega_n\in\Omega} \sup_{\beta\in B} |t(x_n,\varepsilon_n,\beta)| \cdot \int_{\omega'\in\Omega^{n-1}} \sup_{\beta\in B} \left| r^2_{[h_n]}(\beta,\omega') - G^{-1}_\beta(\lambda) \right| dP(\omega') dP(\omega_n).
\end{aligned}
$$

Thus, we obtain from Lemma A.2

$$
\mathsf{E}\left\{ \sup_{\beta\in B} |t(x_n,\varepsilon_n,\beta)\cdot\nu_{in}(\beta)| \right\} \leq M_{gg} E_{Gn} \cdot \int_{\omega_n\in\Omega} \sup_{\beta\in B} |t(x_n,\varepsilon_n,\beta)|\, dP(\omega_n) = o(1).
$$

Similarly, repeating the same steps only over some neighborhood $U(\beta^0,\varepsilon)$ and using Lemma A.3 leads to

$$
\begin{aligned}
&\mathsf{E}\left\{ \sup_{\beta\in B} |t(x_n,\varepsilon_n,\beta)\cdot\nu_{in}(\beta)| \right\} \\
&\leq n^{-\frac{1}{2}} M_{gg} \cdot \int_{\omega_n\in\Omega} \sup_{\beta\in B} |t(x_n,\varepsilon_n,\beta)| \cdot \int_{\omega'\in\Omega^{n-1}} \sup_{\beta\in B} \sqrt{n} \left| r^2_{[h_n]}(\beta,\omega') - G^{-1}_\beta(\lambda) \right| dP(\omega') dP(\omega_n) \\
&\leq n^{-\frac{1}{2}} M_{gg} E_{Ln} \cdot \int_{\omega_n\in\Omega} \sup_{\beta\in B} |t(x_n,\varepsilon_n,\beta)|\, dP(\omega_n) = \mathcal{O}\Big(n^{-\frac{1}{2}}\Big),
\end{aligned}
$$

which closes the proof. □

Corollary A.7 *Let Assumptions D, H1, and I1 hold and assume that $t(x,\varepsilon;\beta)$ is a real-valued function continuous in β uniformly in x and ε over any compact subset of the support of (x,ε). Moreover, assume that $\mathsf{E}\sup_{\beta\in B}|t(x,\varepsilon;\beta)| < \infty$. Then*

$$
\sup_{\beta\in B} \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ t(x_i,\varepsilon_i;\beta) \cdot \left[I\big(r^2_i(\beta) \leq r^2_{[h_n]}(\beta)\big) - I\big(r^2_i(\beta) \leq G^{-1}_\beta(\lambda)\big) \right] \right\} \right| = o_p(1)
$$

as $n \to +\infty$. Additionally, under Assumptions D, H, and I1, there exists $\varepsilon > 0$ such that

$$
\sup_{\beta\in U(\beta^0,\varepsilon)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ t(x_i,\varepsilon_i;\beta) \cdot \left[I\big(r^2_i(\beta) \leq r^2_{[h_n]}(\beta)\big) - I\big(r^2_i(\beta) \leq G^{-1}_\beta(\lambda)\big) \right] \right\} \right| = \mathcal{O}_p(1)
$$

as $n \to +\infty$.

Proof: The corollary follows directly from the Chebyshev inequality for non-negative ran-

dom variables, $P(X \geq K) \leq \mathsf{E}\, X/K$, since by Corollary A.6

$$\begin{aligned} & \mathsf{E}\left\{\sup_{\beta\in B}\left|\frac{1}{n}\sum_{i=1}^{n} t(x_i,\varepsilon_i;\beta)\cdot\left[I\big(r_i^2(\beta)\leq r^2_{[h_n]}(\beta)\big)-I\big(r_i^2(\beta)\leq G_\beta^{-1}(\lambda)\big)\right]\right|\right\} \\ \leq\ & \mathsf{E}\left\{\sup_{\beta\in B}\left|t(x_i,\varepsilon_i;\beta)\cdot\left[I\big(r_i^2(\beta)\leq r^2_{[h_n]}(\beta)\big)-I\big(r_i^2(\beta)\leq G_\beta^{-1}(\lambda)\big)\right]\right|\right\} \\ =\ & o(1) \end{aligned}$$

and

$$\begin{aligned} & \mathsf{E}\left\{\sup_{\beta\in U(\beta^0,\varepsilon)}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} t(x_i,\varepsilon_i;\beta)\cdot\left[I\big(r_i^2(\beta)\leq r^2_{[h_n]}(\beta)\big)-I\big(r_i^2(\beta)\leq G_\beta^{-1}(\lambda)\big)\right]\right|\right\} \\ \leq\ & n^{1/2}\,\mathsf{E}\left\{\sup_{\beta\in U(\beta^0,\varepsilon)}\left|t(x_i,\varepsilon_i;\beta)\cdot\left[I\big(r_i^2(\beta)\leq r^2_{[h_n]}(\beta)\big)-I\big(r_i^2(\beta)\leq G_\beta^{-1}(\lambda)\big)\right]\right|\right\} \\ =\ & \mathcal{O}(1) \end{aligned}$$

as $n \to +\infty$ and the expectation is thus uniformly bounded in $n \in \mathbb{N}$. □

Finally, the last two lemmas of this section study in more details differences of probabilities that $I\Big(r_i^2(\beta^0)\leq r^2_{[h_n]}(\beta^0)\Big)$ and $I\Big(r_i^2(\beta_n)\leq r^2_{[h_n]}(\beta_n)\Big)$ for sequences β_n converging to β^0 at $\sqrt{n}$ rate.

Lemma A.8 *Let Assumptions D and H hold and $\beta \in U(\beta^0, n^{-\frac{1}{2}}M)$ for some $M > 0$. Then it holds as $n \to +\infty$*

1. *For the conditional probability*

$$\begin{aligned} & P\big(I\big(r_i^2\big(\beta^0\big)\leq r^2_{[h_n]}(\beta^0)\big)\neq I\big(r_i^2(\beta)\leq r^2_{[h_n]}(\beta)\big)\big|\, x_i\big) \\ =\ & \Big|h'_\beta\big(x_i,\beta^0\big)^T(\beta-\beta^0)\Big|\cdot\Big\{f\left(-\sqrt{G^{-1}(\lambda)}\right)+f\left(\sqrt{G^{-1}(\lambda)}\right)\Big\}+\mathcal{O}_p\Big(n^{-\frac{1}{2}}\Big) \\ =\ & \mathcal{O}_p\Big(n^{-\frac{1}{4}}\Big) \end{aligned}$$

and

$$\begin{aligned} & \mathsf{E}\left\{\operatorname{sgn} r_i(\beta^0)\cdot\big(I\big(r_i^2\big(\beta^0\big)\leq r^2_{[h_n]}(\beta^0)\big)-I\big(r_i^2(\beta)\leq r^2_{[h_n]}(\beta)\big)\big)\big|\, x_i\right\} = \\ & = h'_\beta\big(x_i,\beta^0\big)^T(\beta-\beta^0)\cdot\Big\{f\left(-\sqrt{G^{-1}(\lambda)}\right)+f\left(\sqrt{G^{-1}(\lambda)}\right)\Big\}+\mathcal{O}_p\Big(n^{-\frac{1}{2}}\Big). \end{aligned}$$

2. *For the corresponding unconditional probability*

$$\begin{aligned} & P\big(I\big(r_i^2\big(\beta^0\big)\leq r^2_{[h_n]}(\beta^0)\big)\neq I\big(r_i^2(\beta)\leq r^2_{[h_n]}(\beta)\big)\big) \\ =\ & \mathsf{E}_x\Big|h'_\beta\big(x_i,\beta^0\big)^T(\beta-\beta^0)\Big|\cdot\Big\{f\left(-\sqrt{G^{-1}(\lambda)}\right)+f\left(\sqrt{G^{-1}(\lambda)}\right)\Big\}+\mathcal{O}\Big(n^{-\frac{1}{2}}\Big) \end{aligned}$$

$$= \mathcal{O}\Big(n^{-\frac{1}{2}}\Big).$$

3. For the conditional probability taken over all $\beta \in U(\beta^0, n^{-\frac{1}{2}}M)$

$$\begin{aligned} & P\Big(\exists\beta \in U\left(\beta^0, n^{-\frac{1}{2}}M\right) : I\big(r_i^2(\beta^0) \le r_{[h_n]}^2(\beta^0)\big) \neq I\big(r_i^2(\beta) \le r_{[h_n]}^2(\beta)\big)\Big|\, x_i\Big) \\ = \; & n^{-\frac{1}{2}}M \cdot \sum_{j=1}^{p} \left|h'_{\beta_j}(x_i, \beta^0)\right| \cdot \left\{f\left(-\sqrt{G^{-1}(\lambda)}\right) + f\left(\sqrt{G^{-1}(\lambda)}\right)\right\} + \mathcal{O}_p\Big(n^{-\frac{1}{2}}\Big) \\ = \; & \mathcal{O}_p\Big(n^{-\frac{1}{4}}\Big). \end{aligned}$$

4. For the corresponding unconditional probability taken over all $\beta \in U(\beta^0, n^{-\frac{1}{2}}M)$

$$\begin{aligned} & P\Big(\exists\beta \in U\left(\beta^0, n^{-\frac{1}{2}}M\right) : I\big(r_i^2(\beta^0) \le r_{[h_n]}^2(\beta^0)\big) \neq I\big(r_i^2(\beta) \le r_{[h_n]}^2(\beta)\big)\Big) \\ = \; & n^{-\frac{1}{2}}M \cdot \sum_{j=1}^{p} \mathsf{E}_x\left|h'_{\beta_j}(x_i, \beta^0)\right| \cdot \left\{f\left(-\sqrt{G^{-1}(\lambda)}\right) + f\left(\sqrt{G^{-1}(\lambda)}\right)\right\} + \mathcal{O}\Big(n^{-\frac{1}{2}}\Big) \\ = \; & \mathcal{O}\Big(n^{-\frac{1}{2}}\Big) \end{aligned}$$

as $n \to +\infty$.

Proof: To facilitate easier understanding, let us first define the constant $q_\lambda = \sqrt{G^{-1}(\lambda)}$, difference between residuals $\Delta_h(x_i, \beta) = r_i(\beta^0) - r_i(\beta)$ at β^0 and β, and difference between indicators

$$\nu_{in}(\beta) = I\big(r_i^2(\beta) \le r_{[h_n]}^2(\beta)\big) - I\big(r_i^2(\beta^0) \le r_{[h_n]}^2(\beta^0)\big)$$

at β and β^0. Then we have to compute

$$P\big(\big|I\big(r_i^2(\beta^0) \le r_{[h_n]}^2(\beta^0)\big) - I\big(r_i^2(\beta) \le r_{[h_n]}^2(\beta)\big)\big| = 1\big|\, x_i\big) = P(\,|\nu_{in}(\beta)| = 1|\, x_i)$$

and to prove that the corresponding unconditional probability is (asymptotically) linear in $\beta - \beta^0$. In addition to that, we shall estimate

$$P\left(\sup_{\beta \in U(\beta^0, n^{-\frac{1}{2}}M)} |\nu_{in}(\beta)| = 1 \middle| x_i\right) = P\Big(\exists\beta \in U(\beta^0, n^{-\frac{1}{2}}M) : |\nu_{in}(\beta)| = 1\Big|\, x_i\Big)$$

(these two probabilities are equivalent because the supremum is always attained—$|\nu_{in}(\beta)|$ can be only zero or one). Note that randomness, that is the dependence on events $\omega \in \Omega$, is represented just by the index i here: $r_i(\beta) = h(x_i, \beta) - h(x_i, \beta^0) + \varepsilon_i$ is a function of the ith realization (x_i, ε_i) and the same applies to $\nu_{in}(\beta)$ as a function of $r_i^2(\beta), i = 1, \ldots, n$. Finally, let us assume without the loss of generality that $n > [M^2/\min\{\delta, \varepsilon\}^2]$, where δ and

ε come from Assumption H1 and Lemma A.3, respectively.

1. First, let us compute $P(\nu_{in}(\beta) = -1|\, x_i)$. In the following derivations, it is necessary to keep in mind that we consider all $\beta \in U(\beta^0, n^{-\frac{1}{2}}M)$ so that most of the results can be reused later when $P\Big(\exists\beta \in U(\beta^0, n^{-\frac{1}{2}}M) : \nu_{in}(\beta) = -1\Big|\, x_i\Big)$ is estimated. Apparently, $\nu_{in}(\beta) = -1$ if and only if

$$r_i^2(\beta) > r_{[h_n]}^2(\beta) \quad \text{and} \quad r_i^2(\beta^0) \leq r_{[h_n]}^2(\beta^0).$$

It holds that

$$r_i^2\big(\beta^0\big) \leq r_{[h_n]}^2(\beta^0) \Rightarrow r_i(\beta^0) \in \big(-r_{[h_n]}(\beta^0), r_{[h_n]}(\beta^0)\big) \tag{34}$$

and

$$r_i^2(\beta) > r_{[h_n]}^2(\beta) \Rightarrow r_i(\beta) \in \big(-\infty, -r_{[h_n]}(\beta)\big) \cup \big(r_{[h_n]}(\beta), +\infty\big)\,. \tag{35}$$

By means of the Taylor expansion we can write (for a given $\omega \in \Omega$)

$$\begin{aligned} r_i(\beta) &= \{y_i - h(x_i, \beta)\} \\ &= \big\{y_i - h\big(x_i, \beta^0\big)\big\} - h'_\beta(x_i, \xi)^T(\beta - \beta^0) \\ &= r_i(\beta^0) - h'_\beta(x_i, \xi)^T(\beta - \beta^0) \\ &= r_i(\beta^0) - \Delta_h(x_i, \beta) \end{aligned}$$

where $\xi \in [\beta^0, \beta]_{\varkappa}$ and difference $\Delta_h(x_i, \beta) = h'_\beta(x_i, \xi)^T\,(\beta - \beta^0)$ ($[\cdot,\cdot]_{\varkappa}$ denotes a convex span, see the introduction to Appendix). Taking this result into account, assertions (34) and (35) imply that

$$r_i(\beta^0) \in \big(-r_{[h_n]}(\beta^0), -r_{[h_n]}(\beta) + \Delta_h(x_i, \beta)\big) \cup \big(r_{[h_n]}(\beta) + \Delta_h(x_i, \beta), r_{[h_n]}(\beta^0)\big)\,, \tag{36}$$

where the convention $(a, b) = \emptyset$ if $b < a$ is used. For $\nu_i(\beta) = 1$, it is possible to derive analogously

$$r_i(\beta^0) \in \big(-r_{[h_n]}(\beta) + \Delta_h(x_i, \beta), -r_{[h_n]}(\beta^0)\big) \cup \big(r_{[h_n]}(\beta^0), r_{[h_n]}(\beta) + \Delta_h(x_i, \beta)\big)\,. \tag{37}$$

Given results (36) and (37), we can write $P(\,|\nu_i(\beta)| = 1|\, x_i)$ as

$$P\big(r_i(\beta^0) \in \big[-r_{[h_n]}(\beta^0), -r_{[h_n]}(\beta) + \Delta_h(x_i, \beta)\big]_{\varkappa} \cup \big[r_{[h_n]}(\beta^0), r_{[h_n]}(\beta) + \Delta_h(x_i, \beta)\big]_{\varkappa}\big|\, x_i\big). \tag{38}$$

Lemma A.4 allows us to simplify this expression even further:

$$P\big(r_i(\beta^0) \in \big[-r_{[h_n]}(\beta^0), -r_{[h_n]}(\beta^0) + \Delta_h(x_i, \beta)\big]_{\varkappa} \cup \big[r_{[h_n]}(\beta^0), r_{[h_n]}(\beta^0) + \Delta_h(x_i, \beta)\big]_{\varkappa}\big|\, x_i\big) + o_p\Big(n^{-\frac{1}{2}}\Big) \tag{39}$$

as $n \to +\infty$. Please, notice that, conditionally on x_i, $\nu_{in}(\beta) \neq 0$ implies $\operatorname{sgn} r_i(\beta^0) \cdot \nu_{in}(\beta) = \operatorname{sgn} \Delta_h(x_i, \beta)$ with probability approaching 1 as $1 - \mathcal{O}\Big(n^{-\frac{1}{2}}\Big)$ with $n \to \infty$. First, $\Delta_h(x_i, \beta)$ is given by x_i and β, and for a fixed x_i, it is bounded by $\Delta_h(x_i, \beta) = h'_\beta(x_i, \xi)^T(\beta - \beta^0) \leq \mathcal{O}(1)(\beta - \beta^0)$ and converges to zero for $\beta \to \beta^0$. So we can choose $n_0 \in \mathbb{N}$ such that $|\Delta_h(x_i, \beta)| < \sqrt{\frac{1}{2}G^{-1}(\lambda)}$ for all $n > n_0$ (remember, $\beta \in U(\beta^0, n^{-\frac{1}{2}}M)$). Second, Lemma A.3 implies that $P\Big(r^2_{[h_n]}(\beta^0) < \frac{1}{2}G^{-1}(\lambda)\Big) = \mathcal{O}\Big(n^{-\frac{1}{2}}\Big)$, and consequently, $P\Big(\big|r_{[h_n]}(\beta^0)\big| < \sqrt{\frac{1}{2}G^{-1}(\lambda)}\Big) = \mathcal{O}\Big(n^{-\frac{1}{2}}\Big)$ as $n \to \infty$. Therefore, we can write with probability higher than $1 - \mathcal{O}\Big(n^{-\frac{1}{2}}\Big)$ that for $\Delta_h(x_i, \beta) > 0$ and $n > n_0$ (see (36) and (37))

- $\nu_i(\beta) = 1$ corresponds to $r_i(\beta^0) \in \big(r_{[h_n]}(\beta^0), r_{[h_n]}(\beta^0) + \Delta_h(x_i, \beta)\big) \subset (0, +\infty)$, thus $\nu_i(\beta) > 0$ if $r_i(\beta^0) > 0$.
- $\nu_i(\beta) = -1$ corresponds to $r_i(\beta^0) \in \big(-r_{[h_n]}(\beta^0), -r_{[h_n]}(\beta^0) + \Delta_h(x_i, \beta)\big) \subset (-\infty, 0)$, thus $\nu_i(\beta) < 0$ if $r_i(\beta^0) < 0$.

Similarly for the case of $\Delta_h(x_i, \beta) < 0$.

Let us now analyze probability (39). Keeping in mind that residual $r_i(\beta^0) \equiv \varepsilon_i$, its probability density function $f(x)$ is bounded from above by a positive constant M_f and is differentiable in a neighborhood of $\sqrt{G^{-1}(\lambda)}$ due to Assumption D2, we can write using Lemma A.3 (remember that q_λ denotes $\sqrt{G^{-1}(\lambda)}$):

$$\begin{aligned} &P\big(r_i(\beta^0) \in \big[-r_{[h_n]}(\beta^0), -r_{[h_n]}(\beta^0) + \Delta_h(x_i, \beta)\big]_{\varkappa} \cup \big[r_{[h_n]}(\beta^0), r_{[h_n]}(\beta^0) + \Delta_h(x_i, \beta)\big]_{\varkappa} \big|\, x_i\big) \\ &\quad = P\big(r_i(\beta^0) \in [-q_\lambda - \xi_1, -q_\lambda - \xi_1 + \Delta_h(x_i, \beta)]_{\varkappa} \cup [q_\lambda + \xi_1, q_\lambda + \xi_1 + \Delta_h(x_i, \beta)]_{\varkappa} \big|\, x_i\big), \end{aligned}$$

where ξ_1 and ξ_2 are random variables behaving like $\mathcal{O}_p\Big(n^{-\frac{1}{2}}\Big)$. Taylor's expansion for the distribution function of ε_i further implies

$$\begin{aligned} &P\big(r_i(\beta^0) \in \big[-r_{[h_n]}(\beta^0), -r_{[h_n]}(\beta^0) + \Delta_h(x_i, \beta)\big]_{\varkappa} \cup \big[r_{[h_n]}(\beta^0), r_{[h_n]}(\beta^0) + \Delta_h(x_i, \beta)\big]_{\varkappa} \big|\, x_i\big) \\ = \;& |\Delta_h(x_i, \beta)| \cdot \{f(-q_\lambda) + f(q_\lambda) + f'(\xi_3) \cdot (\Delta_h(x_i, \beta) + \xi_1) + f'(\xi_4) \cdot (\Delta_h(x_i, \beta) + \xi_2)\} \\ = \;& \Big|h'_\beta(x_i, \xi)^T\big(\beta - \beta^0\big)\Big| \cdot \Big\{f\Big(-\sqrt{G^{-1}(\lambda)}\Big) + f\Big(\sqrt{G^{-1}(\lambda)}\Big)\Big\} + \mathcal{O}_p\Big(n^{-\frac{1}{2}}\Big), \end{aligned} \tag{40}$$

(remember, $\beta \in U(\beta^0, n^{-\frac{1}{2}}M)$, so $\Delta_h(x_i, \beta) = h'_\beta(x_i, \xi)^T(\beta - \beta^0) = \mathcal{O}_p\Big(n^{-\frac{1}{4}}\Big)$), where the last step uses Taylor's expansion of the first derivative of $h(x, \beta)$ at point β^0:

$$h'_\beta(x_i, \xi) = h'_\beta\big(x_i, \beta^0\big) + h''_{\beta\beta}(x_i, \zeta)(\xi - \beta^0) = h'_\beta\big(x_i, \beta^0\big) + \mathcal{O}_p(1)$$

(see Assumption H4). Hence, the first assertion of part 1 is proved—the inequality

$$P\big(I\big(r_i^2\big(\beta^0\big) \leq r^2_{[h_n]}(\beta^0)\big) \neq I\big(r_i^2(\beta) \leq r^2_{[h_n]}(\beta)\big) \big|\, x_i\big) \leq \mathcal{O}_p\Big(n^{-\frac{1}{4}}\Big)$$

follows from Assumptions D2, H3, and the fact that $\beta \in U(\beta^0, n^{-\frac{1}{2}}M)$. The second assertion follows immediately from the note explaining that $\operatorname{sgn} r_i(\beta^0) \cdot \nu_i(\beta) = \operatorname{sgn} \Delta_h(x_i, \beta)$ with probability higher than $1 - \mathcal{O}\left(n^{-\frac{1}{2}}\right)$.

2. Next, we shall evaluate the corresponding unconditional probability, that is the expectation of $P(\,|\nu_i(\beta)| = 1|\,x_i)$ over x_i, and check its asymptotic linearity in $\beta - \beta^0$. As

$$\begin{aligned} & P\big(I\big(r_i^2\big(\beta^0\big) \le r_{[h_n]}^2(\beta^0)\big) \neq I\big(r_i^2(\beta) \le r_{[h_n]}^2(\beta)\big)\big) \\ = \;& \mathsf{E}_x\, P(\,|\nu_i(\beta)| = 1|\,x_i) \\ = \;& \mathsf{E}_x \left| h_\beta'\big(x_i, \beta^0\big)^T(\beta - \beta^0)\right| \cdot \left\{ f\left(-\sqrt{G^{-1}(\lambda)}\right) + f\left(\sqrt{G^{-1}(\lambda)}\right)\right\} + \mathcal{O}\left(n^{-\frac{1}{2}}\right), \end{aligned}$$

the result is apparent once we take into account $\beta \in U(\beta^0, n^{-\frac{1}{2}}M)$ and the fact that the random variable denoted $\mathcal{O}_p\left(n^{-\frac{1}{2}}\right)$ in (40) can be expressed as a product of a random variable and difference $\beta - \beta^0$.

3. We have derived in part 1 of this proof that

$$\begin{aligned} P(\,|\nu_i(\beta)| = 1|\,x_i) \;=\; & P\big(r_i(\beta^0) \in \big[-r_{[h_n]}(\beta^0), -r_{[h_n]}(\beta^0) + \Delta_h(x_i, \beta)\big]_{\varkappa} \cup \\ & \cup \big[r_{[h_n]}(\beta^0), r_{[h_n]}(\beta^0) + \Delta_h(x_i, \beta)\big]_{\varkappa}\big|\, x_i\big) + o_p\left(n^{-\frac{1}{2}}\right) \end{aligned} \tag{41}$$

as $n \to \infty$, where $o_p\left(n^{-\frac{1}{2}}\right)$ holds uniformly over all $\beta \in U(\beta^0, n^{-\frac{1}{2}}M)$ due to Lemma A.4. The length of the intervals in (41) is a function of $\beta - \beta^0$. Further, notice that the lower bound of the interval and $r_i(\beta^0)$ itself does not depend on β, only the length of the interval is β-dependent, and this length converges to zero as $\beta \to \beta^0$ with increasing n. Now, the crucial point here is that the set of events $\omega \in \Omega$ such that a continuously distributed random variable $r_i(\beta^0) \equiv \varepsilon_i$ belongs to intervals specified in (41) depends purely on the lower and upper bounds of the intervals, and consequently, only on their lengths $\Delta_h(x_i, \beta) = h_\beta'(x_i, \xi)^T(\beta - \beta^0)$ in our case. Therefore, the set of events $\omega \in \Omega$ such that there exists $\beta \in U(\beta^0, n^{-\frac{1}{2}}M)$ for which the continuously distributed random variable ε_i belongs to the intervals specified in (41) and the probability of this set reduce to finding the supremum of the length of the interval over all $\beta \in U(\beta^0, n^{-\frac{1}{2}}M)$.

Hence, using the argument employed to derive (40), we can write

$$\begin{aligned} & P\left(\exists \beta \in U\left(\beta^0, n^{-\frac{1}{2}}M\right) : r_i(\beta^0) \in \big[-r_{[h_n]}(\beta^0), -r_{[h_n]}(\beta^0) + \Delta_h(x_i, \beta)\big]_{\varkappa} \cup \right. \\ & \qquad \left. \cup \big[r_{[h_n]}(\beta^0), r_{[h_n]}(\beta^0) + \Delta_h(x_i, \beta)\big]_{\varkappa}\big|\, x_i\right) \\ = \;& \sup_{\beta \in U\left(\beta^0, n^{-\frac{1}{2}}M\right)} \left| h_\beta'\big(x_i, \beta^0\big)^T\left(\beta - \beta^0\right)\right| \cdot \left\{ f\left(-\sqrt{G^{-1}(\lambda)}\right) + f\left(\sqrt{G^{-1}(\lambda)}\right)\right\} + \mathcal{O}_p\left(n^{-\frac{1}{2}}\right) \\ \le \;& n^{-\frac{1}{2}}M \cdot \sum_{j=1}^{p} \left| h_{\beta_j}'\big(x_i, \beta^0\big)\right| \cdot \left\{ f\left(-\sqrt{G^{-1}(\lambda)}\right) + f\left(\sqrt{G^{-1}(\lambda)}\right)\right\} + \mathcal{O}_p\left(n^{-\frac{1}{2}}\right) \end{aligned}$$

as $n \to +\infty$. Thus, the third assertion is verified using the same argument as in part 1 of the proof.

4. Finally, we should find the corresponding unconditional probability, that is the expectation of $P\Big(\exists\beta \in U(\beta^0, n^{-\frac{1}{2}}M) : |\nu_i(\beta)| = 1\Big)$. The assertion is a direct consequence of the fact that

$$\begin{aligned} & \mathsf{E}_x\, P\Big(\exists\beta \in U\left(\beta^0, n^{-\frac{1}{2}}M\right) : |\nu_i(\beta)| = 1\Big| x_i\Big) \\ = \;& n^{-\frac{1}{2}}M \cdot \sum_{j=1}^{p} \mathsf{E}_x \left|h'_{\beta_j}(x_i, \beta^0)\right| \cdot \left\{f\left(-\sqrt{G^{-1}(\lambda)}\right) + f\left(\sqrt{G^{-1}(\lambda)}\right)\right\} + \mathcal{O}\Big(n^{-\frac{1}{2}}\Big) \end{aligned}$$

(note again that $\mathsf{E}_x\, \mathcal{O}_p\Big(n^{-\frac{1}{2}}\Big) = \mathcal{O}\Big(n^{-\frac{1}{2}}\Big)$ because we integrate a random variable multiplied by the non-random difference $\beta - \beta^0 \in U(0, n^{-\frac{1}{2}}M)$). $\square$

Corollary A.9 *Under the assumptions of Lemma A.8, suppose that there exists some* $\beta \in U(\beta^0, n^{-\frac{1}{2}}M)$ *such that*

$$I\big(r_i^2\big(\beta^0\big) \le r^2_{[h_n]}\big(\beta^0\big)\big) \ne I\big(r_i^2(\beta) \le r^2_{[h_n]}(\beta)\big).$$

Then

$$\begin{aligned} \Big|\big|r_i(\beta^0)\big| - \sqrt{G^{-1}(\lambda)}\Big| &= \Big|r_i(\beta^0) - \operatorname{sgn} r_i(\beta^0) \cdot \sqrt{G^{-1}(\lambda)}\Big| \\ &\le \Big|h'_\beta(x_i, \xi)^T(\beta - \beta^0)\Big| + \mathcal{O}_p\Big(n^{-\frac{1}{2}}\Big) \\ &= \mathcal{O}_p\Big(n^{-\frac{1}{4}}\Big) \end{aligned}$$

and

$$\mathsf{E}\left\{\left|\big|r_i(\beta^0)\big| - \sqrt{G^{-1}(\lambda)}\right|\Big| x_i\right\} \le \Big|h'_\beta(x_i, \xi)^T(\beta - \beta^0)\Big| + \mathcal{O}_p\Big(n^{-\frac{1}{2}}\Big).$$

Proof: In the proof of Lemma A.8, see (36)–(39), we have shown that

$$\nu_{in}(\beta) = I\big(r_i^2(\beta) \le r^2_{[h_n]}(\beta)\big) - I\big(r_i^2\big(\beta^0\big) \le r^2_{[h_n]}(\beta^0)\big)$$

can be non-zero for a given x_i if and only if

$$\begin{aligned} & r_i(\beta^0) \in \big[-r_{[h_n]}(\beta^0), -r_{[h_n]}(\beta) + \Delta_h(x_i, \beta)\big]_{\varkappa} \cup \big[r_{[h_n]}(\beta^0), r_{[h_n]}(\beta) + \Delta_h(x_i, \beta)\big]_{\varkappa} \Longleftrightarrow \\ & \Longleftrightarrow r_i(\beta^0) \in \Big[-\sqrt{G^{-1}(\lambda)} - \xi_1, -q_\lambda - \xi_1 + \Delta_h(x_i, \beta)\Big]_{\varkappa} \cup \\ & \qquad \cup \Big[\sqrt{G^{-1}(\lambda)} + \xi_1, q_\lambda + \xi_1 + \Delta_h(x_i, \beta)\Big]_{\varkappa}, \end{aligned}$$

where ξ_1 and ξ_2 are random variables behaving like $\mathcal{O}_p\Big(n^{-\frac{1}{2}}\Big)$ and $\Delta_h(x_i, \beta) = h'_\beta(x_i, \xi)^T(\beta -$

β^0). Hence,

$$\Big|\big|r_i(\beta^0)\big| - \sqrt{G^{-1}(\lambda)}\Big| \leq \Big|h'_\beta(x_i,\xi)^T(\beta-\beta^0)\Big| + \mathcal{O}_p\Big(n^{-\frac{1}{2}}\Big),$$

and by the first claim of Lemma A.8, we also obtain

$$\mathsf{E}\Big\{\Big|\big|r_i(\beta^0)\big| - \sqrt{G^{-1}(\lambda)}\Big|\,\Big|\,x_i\Big\} \leq \Big|h'_\beta(x_i,\xi)^T(\beta-\beta^0)\Big| + \mathcal{O}_p\Big(n^{-\frac{1}{2}}\Big).$$

Finally, $\Delta_h(x_i,\beta) = h'_\beta(x_i,\xi)^T(\beta-\beta^0) = \mathcal{O}_p\Big(n^{-\frac{1}{4}}\Big)$ due to Assumption H4, and consequently,

$$\Big|\big|r_i(\beta^0)\big| - \sqrt{G^{-1}(\lambda)}\Big| = \mathcal{O}_p\Big(n^{-\frac{1}{4}}\Big)$$

as $n \to +\infty$. □

B Proof of asymptotic linearity

Proof of Theorem 3.1: We are to analyze the term $D_n^1(t) = S'_n(\beta^0 - n^{-\frac{1}{2}}t) - S'_n(\beta^0)$, that is,

$$\begin{aligned} D_n^1(t) = \sum_{i=1}^n \Big[\Big\{y_i - h\Big(x_i,\beta^0 - n^{-\frac{1}{2}}t\Big)\Big\}\cdot h'_\beta\Big(x_i,\beta^0 - n^{-\frac{1}{2}}t\Big)\times \\ \times I\Big(r_i^2\Big(\beta^0 - n^{-\frac{1}{2}}t\Big) \leq r^2_{[h_n]}\Big(\beta^0 - n^{-\frac{1}{2}}t\Big)\Big) \\ - \big\{y_i - h\big(x_i,\beta^0\big)\big\}\cdot h'_\beta\big(x_i,\beta^0\big)\cdot I\big(r_i^2(\beta^0) \leq r^2_{[h_n]}(\beta^0)\big)\Big] \end{aligned}$$

for $t \in \mathcal{T}_M = \{t \in \mathbb{R}^p |\, \|t\| \leq M\}$. There is apparently an $n_0 \in \mathbb{N}$ such that $\beta^0 - n^{-\frac{1}{2}}t \in U(\beta^0,\delta)$ for all $n \geq n_0$ and $t \in \mathcal{T}_M$ ($M > 0$ is a given constant). Therefore, using Taylor's expansion for all $n \geq n_0$ and $t \in \mathcal{T}_M$, we get

$$h\Big(x,\beta^0 - n^{-\frac{1}{2}}t\Big) = h\big(x,\beta^0\big) - h'_\beta(x,\xi)^T n^{-\frac{1}{2}}t$$

and

$$h'_\beta\Big(x,\beta^0 - n^{-\frac{1}{2}}t\Big) = h'_\beta\big(x,\beta^0\big) - h''_{\beta\beta}(x,\xi')n^{-\frac{1}{2}}t,$$

where $\xi,\xi' \in \Big[\beta^0,\beta^0 - n^{-\frac{1}{2}}t\Big]_{\asymp}$. Consequently, we may write $D_n^1(t)$ in the following form:

$$\begin{aligned} D_n^1(t) \;=\; \sum_{i=1}^n \Big[\Big\{\big(y_i - h\big(x_i,\beta^0\big)\big)\cdot h'_\beta\big(x_i,\beta^0\big)\cdot I\Big(r_i^2\Big(\beta^0 - n^{-\frac{1}{2}}t\Big) \leq r^2_{[h_n]}\Big(\beta^0 - n^{-\frac{1}{2}}t\Big)\Big) \\ - \big(y_i - h\big(x_i,\beta^0\big)\big)\cdot h'_\beta\big(x_i,\beta^0\big)\cdot I\big(r_i^2\big(\beta^0\big) \leq r^2_{[h_n]}\big(\beta^0\big)\big)\Big\} \\ - \big(y_i - h\big(x_i,\beta^0\big)\big)\cdot h''_{\beta\beta}(x_i,\xi')n^{-\frac{1}{2}}t\cdot I\Big(r_i^2\Big(\beta^0 - n^{-\frac{1}{2}}t\Big) \leq r^2_{[h_n]}\Big(\beta^0 - n^{-\frac{1}{2}}t\Big)\Big) \end{aligned}$$

$$-h_{\beta}^{'}(x_i,\xi)^T n^{-\frac{1}{2}}t \cdot h_{\beta}^{'}(x_i,\beta^0) \cdot I\Big(r_i^2\Big(\beta^0 - n^{-\frac{1}{2}}t\Big) \leq r_{[h_n]}^2\Big(\beta^0 - n^{-\frac{1}{2}}t\Big)\Big)$$
$$+ h_{\beta}^{'}(x_i,\xi)^T n^{-\frac{1}{2}}t \cdot h_{\beta\beta}^{''}(x_i,\xi')n^{-\frac{1}{2}}t \cdot I\Big(r_i^2\Big(\beta^0 - n^{-\frac{1}{2}}t\Big) \leq r_{[h_n]}^2\Big(\beta^0 - n^{-\frac{1}{2}}t\Big)\Big)\Big]$$

$$= \sum_{i=1}^{n}\Big[\Big\{\big(y_i - h\big(x_i,\beta^0\big)\big) \cdot h_{\beta}^{'}\big(x_i,\beta^0\big)\times$$
$$\times\Big[I\Big(r_i^2\Big(\beta^0 - n^{-\frac{1}{2}}t\Big) \leq r_{[h_n]}^2\Big(\beta^0 - n^{-\frac{1}{2}}t\Big)\Big) - I\big(r_i^2\big(\beta^0\big) \leq r_{[h_n]}^2\big(\beta^0\big)\big)\Big]\Big\} \quad (42)$$
$$-\big(y_i - h\big(x_i,\beta^0\big)\big) \cdot h_{\beta\beta}^{''}(x_i,\xi')n^{-\frac{1}{2}}t \cdot I\big(r_i^2\big(\beta^0\big) \leq r_{[h_n]}^2\big(\beta^0\big)\big) \quad (43)$$
$$-\big(y_i - h\big(x_i,\beta^0\big)\big) \cdot h_{\beta\beta}^{''}(x_i,\xi')n^{-\frac{1}{2}}t \times \quad (44)$$
$$\times\Big[I\Big(r_i^2\Big(\beta^0 - n^{-\frac{1}{2}}t\Big) \leq r_{[h]}^2\Big(\beta^0 - n^{-\frac{1}{2}}t\Big)\Big) - I\big(r_i^2\big(\beta^0\big) \leq r_{[h_n]}^2\big(\beta^0\big)\big)\Big]$$
$$-h_{\beta}^{'}(x_i,\xi)^T n^{-\frac{1}{2}}t \cdot h_{\beta}^{'}\big(x_i,\beta^0\big) \cdot I\big(r_i^2\big(\beta^0\big) \leq r_{[h_n]}^2\big(\beta^0\big)\big) \quad (45)$$
$$-h_{\beta}^{'}(x_i,\xi)^T n^{-\frac{1}{2}}t \cdot h_{\beta}^{'}\big(x_i,\beta^0\big) \times$$
$$\times\Big[I\Big(r_i^2\Big(\beta^0 - n^{-\frac{1}{2}}t\Big) \leq r_{[h_n]}^2\Big(\beta^0 - n^{-\frac{1}{2}}t\Big)\Big) - I\big(r_i^2\big(\beta^0\big) \leq r_{[h_n]}^2\big(\beta^0\big)\big)\Big] \quad (46)$$
$$+ h_{\beta}^{'}(x_i,\xi)^T n^{-\frac{1}{2}}t \cdot h_{\beta\beta}^{''}(x_i,\xi')n^{-\frac{1}{2}}t \cdot I\Big(r_i^2\Big(\beta^0 - n^{-\frac{1}{2}}t\Big) \leq r_{[h_n]}^2\Big(\beta^0 - n^{-\frac{1}{2}}t\Big)\Big)\Big] \quad (47)$$

Let us now analyze the parts of the previous expression one by one. We will show that sums (43), (44), (46), and (47) behave like $\mathcal{O}_p\Big(n^{\frac{1}{4}}\Big)$ or $o_p\Big(n^{\frac{1}{2}}\Big)$, and therefore, are asymptotically negligible with respect to parts (42) and (45), which behave like $\mathcal{O}_p\Big(n^{\frac{1}{2}}\Big)$. Moreover, we find asymptotic representations of (42) and (45).

First of all, the last part (47) can be bounded from above in the following way (see Assumption H4):

$$\sup_{t\in\mathcal{T}_M}\sum_{i=1}^{n}\Big\|h_{\beta}^{'}(x_i,\xi)^T n^{-\frac{1}{2}}t \cdot h_{\beta\beta}^{''}(x_i,\xi')n^{-\frac{1}{2}}t \cdot I\Big(r_i^2\Big(\beta^0 - n^{-\frac{1}{2}}t\Big) \leq r_{[h_n]}^2\Big(\beta^0 - n^{-\frac{1}{2}}t\Big)\Big)\Big\|$$
$$\leq \mathcal{O}_p\Big(n^{-\frac{3}{4}}\Big)\sup_{t\in\mathcal{T}_M}\sum_{i=1}^{n}\Big\|h_{\beta\beta_k}^{''}(x_i,\xi')\Big\|$$
$$\leq \mathcal{O}_p\Big(n^{-\frac{3}{4}}\Big)\left(\sum_{i=1}^{n}\Big\|h_{\beta\beta_k}^{''}(x_i,\beta^0)\Big\| + \mathcal{O}_p\Big(n^{\frac{1}{2}}\Big)\right),$$

where the last result follows from Assumption H2 (the Lipschitz property for $h_{\beta\beta}^{''}(x_i,\beta)$) and the fact that $\xi' \in \Big[\beta^0, \beta^0 - n^{-\frac{1}{2}}t\Big]_{\varkappa}$. Once we realize that Assumptions D1 and H5 and the law of large numbers (e.g., Andrews, 1988) guarantee $\sum_{i=1}^{n}\big\|h_{\beta\beta}^{''}(x_i,\beta^0)\big\| = \mathcal{O}_p(n)$ as $n \to +\infty$, we get immediately

$$\sup_{t\in\mathcal{T}_M}\sum_{i=1}^{n}\Big\|h_{\beta}^{'}(x_i,\xi)^T n^{-\frac{1}{2}}t \cdot h_{\beta\beta}^{''}(x_i,\xi)n^{-\frac{1}{2}}t \cdot I\Big(r_i^2\Big(\beta^0 - n^{-\frac{1}{2}}t\Big) \leq r_{[h_n]}^2\Big(\beta^0 - n^{-\frac{1}{2}}t\Big)\Big)\Big\|$$

$$= \mathcal{O}_p\Big(n^{\frac{1}{4}}\Big)$$

as $n \to +\infty$.

Next, we are going to analyze part (46), that is,

$$\sup_{t\in\mathcal{T}_M} \sum_{i=1}^{n} \left\| h'_\beta(x_i,\xi)^T n^{-\frac{1}{2}} t \cdot h'_\beta\big(x_i,\beta^0\big) \cdot \nu_i(n,t) \right\|,$$

where $\nu_i(n,t)$ denotes the difference of indicators

$$\nu_i(n,t) = I\Big(r_i^2\Big(\beta^0 - n^{-\frac{1}{2}}t\Big) \le r^2_{[h_n]}\Big(\beta^0 - n^{-\frac{1}{2}}t\Big)\Big) - I\big(r_i^2(\beta^0) \le r^2_{[h_n]}(\beta^0)\big).$$

As

$$\begin{aligned}
&\sup_{t\in\mathcal{T}_M} \sum_{i=1}^{n} \left\| h'_\beta(x_i,\xi)^T n^{-\frac{1}{2}} t \cdot h'_\beta\big(x_i,\beta^0\big) \cdot \nu_i(n,t) \right\| \\
&= \sup_{t\in\mathcal{T}_M} \sum_{i=1}^{n} \left\| \Big(h'_\beta\big(x_i,\beta^0\big)^T n^{-\frac{1}{2}} t \cdot h'_\beta\big(x_i,\beta^0\big) + n^{-\frac{1}{2}} t^T \cdot h''_{\beta\beta}(x_i,\xi) \cdot n^{-\frac{1}{2}} t \cdot h'_\beta\big(x_i,\beta^0\big) \Big) \cdot \nu_i(n,t) \right\| \\
&\le \sup_{t\in\mathcal{T}_M} \sum_{i=1}^{n} \left\| h'_\beta\big(x_i,\beta^0\big)^T n^{-\frac{1}{2}} t \cdot h'_\beta\big(x_i,\beta^0\big) \right\| \cdot |\nu_i(n,t)| \qquad (48) \\
&+ \mathcal{O}_p(1) \cdot \sup_{t\in\mathcal{T}_M} \sum_{i=1}^{n} \left\| n^{-\frac{1}{2}} t \cdot h'_\beta\big(x_i,\beta^0\big) \right\| \cdot |\nu_i(n,t)| \qquad (49)
\end{aligned}$$

(see Assumption H4), we need to analyze these two summands. This can be done in the same way for both of them, so we will do it here just for (48). To do this, we employ the Chebyshev inequality for non-negative random variables: for any non-negative random variable X it holds that $P(X > K) < \mathsf{E}\, X/K$. Therefore,

$$\begin{aligned}
& P\left(\sup_{t\in\mathcal{T}_M} \sum_{i=1}^{n} \left\| h'_\beta\big(x_i,\beta^0\big)^T n^{-\frac{1}{2}} t \cdot h'_{\beta_k}\big(x_i,\beta^0\big) \cdot \nu_i(n,t) \right\| > K n^{\frac{1}{4}} \right) \\
\le\ & \frac{1}{K n^{\frac{1}{4}}}\, \mathsf{E} \left(\sup_{t\in\mathcal{T}_M} \sum_{i=1}^{n} \left\| h'_\beta\big(x_i,\beta^0\big)^T n^{-\frac{1}{2}} t \cdot h'_\beta\big(x_i,\beta^0\big) \cdot \nu_i(n,t) \right\| \right) \\
\le\ & \frac{n^{-\frac{3}{4}}}{K} \sum_{i=1}^{n} \mathsf{E} \left\{ \sup_{t\in\mathcal{T}_M} \left\| h'_\beta\big(x_i,\beta^0\big)^T t \cdot h'_\beta\big(x_i,\beta^0\big) \right\| \cdot \sup_{t\in\mathcal{T}_M} |\nu_i(n,t)| \right\}
\end{aligned}$$

and by the Schwartz inequality and Lemma A.8

$$\frac{n^{-\frac{3}{4}}}{K} \sum_{i=1}^{n} \mathsf{E} \left\{ \sup_{t\in\mathcal{T}_M} \left\| h'_\beta\big(x_i,\beta^0\big)^T t \cdot h'_\beta\big(x_i,\beta^0\big) \right\| \cdot \sup_{t\in\mathcal{T}_M} |\nu_i(n,t)| \right\}$$

$$
\begin{aligned}
&\leq \frac{n^{-\frac{3}{4}}}{K}\sum_{i=1}^{n}\sqrt{\mathsf{E}\left(\sup_{t\in\mathcal{T}_M}\left\|h'_\beta(x_i,\beta^0)^T t\right\|\cdot\left\|h'_\beta(x_i,\beta^0)\right\|\right)^2\cdot\mathsf{E}\sup_{t\in\mathcal{T}_M}|\nu_i(n,t)|}\\
&\leq \frac{n^{\frac{1}{4}}}{K}\sqrt{\mathsf{E}\left(M\cdot\sum_{j=1}^{p}\left|h'_{\beta_j}(x_i,\beta^0)\right|\cdot\left\|h'_\beta(x_i,\beta^0)\right\|\right)^2\mathcal{O}\Big(n^{-\frac{1}{2}}\Big)}\\
&\leq \frac{\mathcal{O}(1)}{K}=\frac{\text{const}}{K}. \qquad (50)
\end{aligned}
$$

Apparently, for any $\varepsilon>0$ there is a $K>0$ such that the constant term (50), which is proportional to $\frac{1}{K}$, is smaller than ε. Thus, we have shown that

$$
\begin{aligned}
&\sup_{t\in\mathcal{T}_M}\sum_{i=1}^{n}\Big\|h'_\beta(x_i,\xi)^T n^{-\frac{1}{2}}t\cdot h'_\beta\big(x_i,\beta^0\big)\times\\
&\quad\times\Big(I\Big(r_i^2\Big(\beta^0-n^{-\frac{1}{2}}t\Big)\leq r^2_{[h_n]}\Big(\beta^0-n^{-\frac{1}{2}}t\Big)\Big)-I\big(r_i^2(\beta^0)\leq r^2_{[h_n]}(\beta^0)\big)\Big)\Big\|=\mathcal{O}_p\Big(n^{\frac{1}{4}}\Big)
\end{aligned}
$$

as $n\to+\infty$. Please, note that (44) can be estimated in the same way, so we have also shown how to prove

$$
\begin{aligned}
&\sup_{t\in\mathcal{T}_M}\sum_{i=1}^{n}\Big\|\big\{y_i-h\big(x_i,\beta^0\big)\big\}\cdot h''_{\beta\beta}(x_i,\xi')n^{-\frac{1}{2}}t\times\\
&\quad\times\Big(I\Big(r_i^2\Big(\beta^0-n^{-\frac{1}{2}}t\Big)\leq r^2_{[h_n]}\Big(\beta^0-n^{-\frac{1}{2}}t\Big)\Big)-I\big(r_i^2(\beta^0)\leq r^2_{[h_n]}(\beta^0)\big)\Big)\Big\|=\mathcal{O}_p\Big(n^{\frac{1}{4}}\Big)
\end{aligned}
$$

as $n\to+\infty$.

The next summand to be analyzed is (43):

$$
\sup_{t\in\mathcal{T}_M}\sum_{i=1}^{n}\big\{y_i-h\big(x_i,\beta^0\big)\big\}\cdot h''_{\beta\beta}(x_i,\xi')n^{-\frac{1}{2}}t\cdot I\big(r_i^2\big(\beta^0\big)\leq r^2_{[h_n]}\big(\beta^0\big)\big).
$$

This can be rewritten as (Assumption H2)

$$
\begin{aligned}
&\sup_{t\in\mathcal{T}_M}\sum_{i=1}^{n}\big\{y_i-h\big(x_i,\beta^0\big)\big\}\cdot h''_{\beta\beta}\big(x_i,\beta^0\big)n^{-\frac{1}{2}}t\cdot I\big(r_i^2\big(\beta^0\big)\leq r^2_{[h_n]}\big(\beta^0\big)\big) \qquad (51)\\
&+\mathcal{O}_p\big(n^{-1}\big)\cdot\sum_{i=1}^{n}\big\{y_i-h\big(x_i,\beta^0\big)\big\}\cdot I\big(r_i^2\big(\beta^0\big)\leq r^2_{[h_n]}\big(\beta^0\big)\big). \qquad (52)
\end{aligned}
$$

Assumption D2 implies that the expectation of (51) conditional on x_i is equal to zero, thus the unconditional expectation is zero as well. Moreover, the variance of a component of (51) given by indices $j,k,l\in\{1,\ldots,p\}$ equals (Assumptions D2 and H5 are used)

$$
\mathsf{var}\left[\big(y_i-h\big(x_i,\beta^0\big)\big)\cdot h''_{\beta_j\beta_k}\big(x_i,\beta^0\big)t_l\cdot I\big(r_i^2\big(\beta^0\big)\leq r^2_{[h_n]}\big(\beta^0\big)\big)\right]
$$

$$
\begin{aligned}
&= \mathsf{var}_x\left\{h''_{\beta_j\beta_k}(x_i,\beta^0)t_l \cdot \mathsf{E}\left[\,(y_i - h(x_i,\beta^0)) \cdot I\big(r_i^2(\beta^0) \le r^2_{[h_n]}(\beta^0)\big)\big|\, x_i\right]\right\} \\
&+ \mathsf{E}_x\left\{\Big(h''_{\beta_j\beta_k}(x_i,\beta^0)t_l\Big)^2 \cdot \mathsf{var}\left[\,(y_i - h(x_i,\beta^0)) \cdot I\big(r_i^2(\beta^0) \le r^2_{[h_n]}(\beta^0)\big)\big|\, x_i\right]\right\} \\
&= \mathsf{var}_x\left\{h''_{\beta_j\beta_k}(x_i,\beta^0)t_l \cdot 0\right\} + \mathsf{E}_x\left\{\Big(h''_{\beta_j\beta_k}(x_i,\beta^0)t_l\Big)^2 \cdot \sigma^2\right\} \\
&= \sigma^2 t_l^2 \cdot \mathsf{E}\left\{h''_{\beta_j\beta_k}(x_i,\beta^0)\right\}^2 \le \sigma^2 M^2 \cdot \mathsf{E}\left\{h''_{\beta_j\beta_k}(x_i,\beta^0)\right\}^2,
\end{aligned}
$$

so it exists and is finite and uniformly bounded over all $t \in \mathcal{T}_M$. Because of Assumption D2, the summands in (51) form a triangular array of martingale differences and we can employ the law of large numbers for martingales (see Davidson, 1994, Theorem 19.7, for instance) to conclude for components of (51) that for every $j,k,l = 1,\ldots,p$,

$$
n^{-\frac{3}{4}} \sum_{i=1}^{n} \big\{y_i - h(x_i,\beta^0)\big\} \cdot h''_{\beta_j\beta_k}(x_i,\beta^0)t_l \cdot I\big(r_i^2(\beta^0) \le r^2_{[h_n]}(\beta^0)\big) \to 0
$$

in probability (uniformly in $t \in \mathcal{T}_M$ since $\|t\| \le M$ is non-random). Because (52) is apparently bounded in probability, it holds that

$$
\sup_{t\in\mathcal{T}_M} \sum_{i=1}^{n} \big\{y_i - h(x_i,\beta^0)\big\} \cdot h''_{\beta\beta}(x_i,\xi')n^{-\frac{1}{2}}t \cdot I\big(r_i^2(\beta^0) \le r^2_{[h_n]}(\beta^0)\big) = o_p\Big(n^{\frac{1}{4}}\Big)
$$

as $n \to \infty$.

The last but one term to be estimated is (45), that is,

$$
\begin{aligned}
&\sum_{i=1}^{n} h'_\beta(x_i,\xi)^T n^{-\frac{1}{2}}t \cdot h'_\beta(x_i,\beta^0) \cdot I\big(r_i^2(\beta^0) \le r^2_{[h_n]}(\beta^0)\big) \\
&= \sum_{i=1}^{n} h'_\beta(x_i,\beta^0)^T n^{-\frac{1}{2}}t \cdot h'_\beta(x_i,\beta^0) \cdot I\big(r_i^2(\beta^0) \le r^2_{[h_n]}(\beta^0)\big) \qquad (53) \\
&+ \sum_{i=1}^{n} n^{-\frac{1}{2}}t^T \cdot h''_{\beta\beta}(x_i,\xi'') \cdot n^{-\frac{1}{2}}t \cdot h'_\beta(x_i,\beta^0) \cdot I\big(r_i^2(\beta^0) \le r^2_{[h_n]}(\beta^0)\big). \qquad (54)
\end{aligned}
$$

The supremum of the second part (54) over $t \in \mathcal{T}_M$ behaves like $\mathcal{O}_p(1)$, as we shall argue now. Since

$$
\begin{aligned}
&\left|\sum_{i=1}^{n} n^{-\frac{1}{2}}t^T \cdot h''_{\beta\beta}(x_i,\xi'') \cdot n^{-\frac{1}{2}}t \cdot h'_\beta(x_i,\beta^0) \cdot I\big(r_i^2(\beta^0) \le r^2_{[h_n]}(\beta^0)\big)\right| \\
&\le \sum_{i=1}^{n} \left|n^{-\frac{1}{2}}t^T \cdot h''_{\beta\beta}(x_i,\xi'') \cdot n^{-\frac{1}{2}}t \cdot h'_\beta(x_i,\beta^0)\right|,
\end{aligned}
$$

we can simply use the law of large numbers for mixingales (Andrews, 1988) and the uniform

law of large number (Andrews, 1992) for the right hand side of the inequality over all $\beta'' \in U(\beta, \delta)$:

$$\frac{1}{n}\sum_{i=1}^{n}\left|t^T \cdot h''_{\beta\beta}(x_i, \beta'') \cdot t \cdot h'_\beta\big(x_i, \beta^0\big)\right| \to \mathsf{E}\left|t^T \cdot h''_{\beta\beta}(x_1, \beta'') \cdot t \cdot h'_\beta\big(x_1, \beta^0\big)\right|$$

in probability as $n \to \infty$ (the conditions BD, TSE-1D, DM, and P-WLLN of Andrews, 1992, are satisfied by means of Assumptions I1, H2, H5, and D1, respectively). Since the expectation is bounded uniformly over $t \in \mathcal{T}_M$ ($\|t\| \leq M$ and Assumption H5), (54) is bounded in probability.

Let us look now at (53):

$$\begin{aligned}
& \sum_{i=1}^{n} h'_\beta\big(x_i, \beta^0\big)^T n^{-\frac{1}{2}} t \cdot h'_\beta\big(x_i, \beta^0\big) \cdot I\big(r_i^2\big(\beta^0\big) \leq r^2_{[h_n]}\big(\beta^0\big)\big) \\
= & \sum_{i=1}^{n} h'_\beta\big(x_i, \beta^0\big)^T n^{-\frac{1}{2}} t \cdot h'_\beta\big(x_i, \beta^0\big) \times & (55) \\
& \qquad \times \left[I\big(r_i^2\big(\beta^0\big) \leq r^2_{[h_n]}\big(\beta^0\big)\big) - I\big(r_i^2\big(\beta^0\big) \leq G^{-1}(\lambda)\big)\right] \\
+ & \; n^{-\frac{1}{2}} \sum_{i=1}^{n} \Big\{h'_\beta\big(x_i, \beta^0\big) \cdot h'_\beta\big(x_i, \beta^0\big)^T \cdot I\big(r_i^2\big(\beta^0\big) \leq G^{-1}(\lambda)\big) & (56) \\
& \qquad - \mathsf{E}\left[h'_\beta\big(x_i, \beta^0\big) \cdot h'_\beta\big(x_i, \beta^0\big)^T \cdot I\big(r_i^2\big(\beta^0\big) \leq G^{-1}(\lambda)\big)\right]\Big\} t \\
+ & \; n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathsf{E}\left[h'_\beta\big(x_i, \beta^0\big) \cdot h'_\beta\big(x_i, \beta^0\big)^T \cdot I\big(r_i^2\big(\beta^0\big) \leq G^{-1}(\lambda)\big)\right] t. & (57)
\end{aligned}$$

The supremum of the first part, that is, sum (55), over $t \in \mathcal{T}_M$ behaves again like $\mathcal{O}_p\Big(n^{\frac{1}{4}}\Big)$ for $n \to \infty$. This can be proved in the same manner as we did for (48), this time utilizing Lemma A.5. Next, using the central limit theorem, each element of matrix (56) converges in distribution to a normally distributed random variable with zero mean and a finite variance uniformly bounded for $t \in \mathcal{T}_M$ (the result of Arcones and Yu, 1994, applies due to Assumptions D1, D2, and H3; alternatively, one can apply standard central limit theorem such as Davidson, 1994, Theorem 24.5). Hence, it is bounded in probability as well. Finally, the last element (57) can be rewritten as $n^{\frac{1}{2}} \cdot \lambda \cdot Q_h t$ since

$$\begin{aligned}
& \mathsf{E}\left[h'_\beta\big(x_i, \beta^0\big) \cdot h'_\beta\big(x_i, \beta^0\big)^T \cdot I\big(r_i^2\big(\beta^0\big) \leq G^{-1}(\lambda)\big)\right] \\
= & \; \mathsf{E}_x\left[h'_\beta\big(x_i, \beta^0\big) \cdot h'_\beta\big(x_i, \beta^0\big)^T \cdot \mathsf{E}\left\{I\big(r_i^2\big(\beta^0\big) \leq G^{-1}(\lambda)\big)\big|\, x_i\right\}\right] \\
= & \; \lambda \cdot \mathsf{E}_x\left[h'_\beta\big(x_i, \beta^0\big) \cdot h'_\beta\big(x_i, \beta^0\big)^T\right] = \lambda \cdot Q_h.
\end{aligned}$$

Therefore, we can conclude that

$$\sup_{t\in\mathcal{T}_M}\left\|\sum_{i=1}^{n} h_\beta^{'}\big(x_i,\beta^0\big)n^{-\frac{1}{2}}t\cdot h_\beta^{'}\big(x_i,\beta^0\big)\cdot I\big(r_i^2\big(\beta^0\big)\le r_{[h_n]}^2\big(\beta^0\big)\big)-n^{\frac{1}{2}}\cdot\lambda\cdot Q_h t\right\|=\mathcal{O}_p(1)$$

as $n\to+\infty$.

Finally, let us move our attention to the term (42). Using once again notation

$$\nu_i(n,t)=I\Big(r_i^2\Big(\beta^0-n^{-\frac{1}{2}}t\Big)\le r_{[h_n]}^2\Big(\beta^0-n^{-\frac{1}{2}}t\Big)\Big)-I\big(r_i^2(\beta^0)\le r_{[h_n]}^2(\beta^0)\big),$$

we can rewrite (42) as

$$\begin{aligned}
&\sum_{i=1}^{n}\big\{y_i-h\big(x_i,\beta^0\big)\big\}\cdot h_\beta^{'}\big(x_i,\beta^0\big)\cdot\nu_i(n,t)\\
=\;&\sum_{i=1}^{n} r_i(\beta^0)\cdot h_\beta^{'}\big(x_i,\beta^0\big)\cdot\nu_i(n,t)\\
=\;&\sum_{i=1}^{n}\Big\{r_i(\beta^0)-\operatorname{sgn} r_i(\beta^0)\cdot\sqrt{G^{-1}(\lambda)}\Big\}\cdot h_\beta^{'}\big(x_i,\beta^0\big)\cdot\nu_i(n,t) \qquad (58)\\
+\;&\sum_{i=1}^{n}\operatorname{sgn} r_i(\beta^0)\cdot\sqrt{G^{-1}(\lambda)}\cdot h_\beta^{'}\big(x_i,\beta^0\big)\cdot\nu_i(n,t). \qquad (59)
\end{aligned}$$

For the simplicity of notation, let us use $q_\lambda=\sqrt{G^{-1}(\lambda)}$. The first part (58) multiplied by $n^{-\frac{1}{4}}$ is bounded in probability. This can be shown as follows: the Chebyshev inequality implies

$$\begin{aligned}
&P\left(n^{-\frac{1}{4}}\sup_{t\in\mathcal{T}_M}\left\|\sum_{i=1}^{n}\big\{r_i(\beta^0)-\operatorname{sgn} r_i(\beta^0)\cdot q_\lambda\big\}\cdot h_\beta^{'}\big(x_i,\beta^0\big)\cdot\nu_i(n,t)\right\|>K\right) \qquad (60)\\
\le\;&\frac{1}{K}\,\mathsf{E}\left(n^{-\frac{1}{4}}\sup_{t\in\mathcal{T}_M}\left\|\sum_{i=1}^{n}\big\{r_i(\beta^0)-\operatorname{sgn} r_i(\beta^0)\cdot q_\lambda\big\}\cdot h_\beta^{'}\big(x_i,\beta^0\big)\cdot\nu_i(n,t)\right\|\right)\\
\le\;&\frac{n^{\frac{3}{4}}}{K}\,\mathsf{E}\left(\big|r_i(\beta^0)-\operatorname{sgn} r_i(\beta^0)\cdot q_\lambda\big|\cdot\left\|h_\beta^{'}\big(x_i,\beta^0\big)\right\|\cdot\sup_{t\in\mathcal{T}_M}|\nu_i(n,t)|\right)
\end{aligned}$$

and by Lemma A.8 together with Corollary A.9 ($r_i(\beta^0)\equiv\varepsilon_i$ and x_i are independent random variables)

$$\begin{aligned}
&\frac{n^{\frac{3}{4}}}{K}\,\mathsf{E}_x\left(\left\|h_\beta^{'}\big(x_i,\beta^0\big)\right\|\cdot\mathsf{E}\left[\big|r_i(\beta^0)-\operatorname{sgn} r_i(\beta^0)\cdot q_\lambda\big|\cdot\sup_{t\in\mathcal{T}_M}|\nu_i(n,t)|\,\middle|\,x_i\right]\right)\\
\le\;&\frac{n^{\frac{3}{4}}}{K}\,\mathsf{E}\left(\mathcal{O}\big(n^{-1}\big)\cdot\left\|h_\beta^{'}\big(x_i,\beta^0\big)\right\|\cdot\Big[\left\|h_\beta^{'}(x_i,\xi)\right\|+\mathcal{O}_p(1)\Big]\cdot\Big[\left\|h_{\beta\beta}^{''}\big(x_i,\beta^0\big)\right\|+\mathcal{O}_p(1)\Big]\right)
\end{aligned}$$

$$\leq \quad \frac{\mathcal{O}(1)}{K} = \frac{\text{const.}}{K}.$$

Therefore, the probability (60) can be made smaller than ε by an appropriate choice of K, and hence, (58) multiplied by $n^{-\frac{1}{4}}$ is bounded in probability. In other words, it holds that

$$\sup_{t\in\mathcal{T}_M} \left\| \sum_{i=1}^{n} \left\{ r_i(\beta^0) - \operatorname{sgn} r_i(\beta^0) \cdot q_\lambda \right\} \cdot h'_\beta\big(x_i, \beta^0\big) \times \right.$$
$$\left. \times \left[I\Big(r_i^2\Big(\beta^0 - n^{-\frac{1}{2}}t\Big) \leq r^2_{[h_n]}\Big(\beta^0 - n^{-\frac{1}{2}}t\Big)\Big) - I\big(r_i^2\big(\beta^0\big) \leq r^2_{[h_n]}\big(\beta^0\big)\big) \right] \right\| = \mathcal{O}_p\Big(n^{\frac{1}{4}}\Big)$$

as $n \to \infty$. All we have to do now is to treat

$$\sum_{i=1}^{n} \operatorname{sgn} r_i(\beta^0) \cdot q_\lambda \cdot h'_\beta\big(x_i, \beta^0\big) \cdot \nu_i(n,t). \tag{61}$$

This is done again in two steps: first, we show that the sum less its expectation is $o_p\Big(n^{\frac{1}{2}}\Big)$, and second, the expectation of the sum is evaluated. For the first part, we have shown in Lemma A.8 that the probability of

$$\nu_i(n,t) = I\Big(r_i^2\Big(\beta^0 - n^{-\frac{1}{2}}t\Big) \leq r^2_{[h_n]}\Big(\beta^0 - n^{-\frac{1}{2}}t\Big)\Big) - I\big(r_i^2\big(\beta^0\big) \leq r^2_{[h_n]}\big(\beta^0\big)\big)$$

being non-zero conditional on x_i (and thus the conditional expectation of this term in absolute value) is equal to

$$\mathsf{E}\left(\nu_i(n,t)\middle|\, x_i\right) = \left| h'_\beta\big(x_i, \beta^0\big)^T(\beta - \beta^0) \right| \cdot \left\{ f\left(-\sqrt{G^{-1}(\lambda)}\right) + f\left(\sqrt{G^{-1}(\lambda)}\right) \right\} + \mathcal{O}_p\Big(n^{-\frac{1}{2}}\Big)$$

as $n \to +\infty$, and that the expectation of this conditional probability behaves like $\mathcal{O}\Big(n^{-\frac{1}{2}}\Big)$. Therefore, the random variable $\nu_i(n,t)$ multiplied by $n^{\frac{1}{2}}$ will have its expectation conditional on x_i behaving like $\left\|h'_\beta(x_i,\beta^0)\right\| \cdot \mathcal{O}(1) + \mathcal{O}_p(1)$ in absolute value. Consequently, Assumption H5 implies for any $j = 1, \ldots, k$,

$$\begin{aligned}
& \mathsf{E}\left| n^{\frac{1}{2p}} \cdot \operatorname{sgn} r_i(\beta^0) \cdot h'_{\beta_j}\big(x_i, \beta^0\big) \cdot \nu_i(n,t) \right|^2 \\
= \quad & \mathsf{E}\left[\left| h'_{\beta_j}\big(x_i, \beta^0\big) \right|^2 \mathsf{E}\left\{ n^{\frac{1}{p}} \left|\nu_i(n,t)\right| \middle| x_i \right\} \right] \\
\leq \quad & \mathsf{E}\left\{ \left| h'_{\beta_j}\big(x_i, \beta^0\big) \right|^2 \cdot n^{\frac{1}{p}-\frac{1}{2}} \left[\left\| h'_\beta\big(x_i, \beta^0\big) \right\| \cdot \mathcal{O}(1) + \mathcal{O}_p(1) \right] \right\} = \mathcal{O}(1).
\end{aligned}$$

Hence, the law of large numbers for L^2-mixingales (Davidson and de Jong, 1997, Corollary

2.1) can be applied to the following sum of random variables:

$$\frac{1}{n^{\frac{1}{2}+\frac{1}{2p}}}\sum_{i=1}^{n} n^{\frac{1}{2p}}\left\{\operatorname{sgn} r_i(\beta^0)\cdot q_\lambda h'_{\beta_j}\big(x_i,\beta^0\big)\cdot\nu_i(n,t)-\mathsf{E}\left[\operatorname{sgn} r_i(\beta^0)\cdot q_\lambda h'_{\beta}\big(x_i,\beta^0\big)\cdot\nu_i(n,t)\right]\right\}.$$

As a direct consequence, it follows that

$$\begin{aligned}&\sum_{i=1}^{n}\operatorname{sgn} r_i(\beta^0)\cdot q_\lambda\cdot h'_\beta\big(x_i,\beta^0\big)\cdot\nu_i(n,t)-\\ -&\sum_{i=1}^{n}\mathsf{E}\left\{\operatorname{sgn} r_i(\beta^0)\cdot q_\lambda\cdot h'_\beta\big(x_i,\beta^0\big)\cdot\nu_i(n,t)\right\} \;=\; o_p\big(n^{1/2}\big)\end{aligned}$$

as $n\to+\infty$.

Finally, the expectation of (61)

$$\mathsf{E}\left\{\sum_{i=1}^{n}\operatorname{sgn} r_i(\beta^0)\cdot\sqrt{G^{-1}(\lambda)}\cdot h'_\beta\big(x_i,\beta^0\big)\cdot\nu_i(n,t)\right\} \tag{62}$$

$$=\;\mathsf{E}_x\left\{\sum_{i=1}^{n} q_\lambda\cdot h'_\beta\big(x_i,\beta^0\big)\cdot\mathsf{E}\left(\operatorname{sgn} r_i(\beta^0)\cdot\nu_i(n,t)\big|\,x_i\right)\right\} \tag{63}$$

can be proved to be a linear function of t by means of Lemma A.8. Since

$$\mathsf{E}\left\{\operatorname{sgn} r_i(\beta^0)\cdot\nu_i(n,t)\big|\,x_i\right\}=h'_\beta\big(x_i,\beta^0\big)^T n^{-\frac{1}{2}}t\cdot\{f(-q_\lambda)+f(q_\lambda)\}+\mathcal{O}_p\Big(n^{-\frac{1}{2}}\Big),$$

(63) can be rewritten as

$$\begin{aligned}&\mathsf{E}_x\left\{\sum_{i=1}^{n} q_\lambda\cdot h'_\beta\big(x_i,\beta^0\big)\cdot\left[h'_\beta\big(x_i,\beta^0\big)^T n^{-\frac{1}{2}}t\cdot\{f(-q_\lambda)+f(q_\lambda)\}+\mathcal{O}_p\Big(n^{-\frac{1}{2}}\Big)\right]\right\}=\\ &=q_\lambda\cdot\{f(-q_\lambda)+f(q_\lambda)\}\cdot\left\{\sum_{i=1}^{n}\mathsf{E}_x\left[h'_\beta\big(x_i,\beta^0\big)\cdot h'_\beta\big(x_i,\beta^0\big)^T\right]+\mathcal{O}\Big(n^{-\frac{1}{2}}\Big)\right\}\cdot n^{-\frac{1}{2}}t\\ &=q_\lambda\cdot\{f(-q_\lambda)+f(q_\lambda)\}\cdot Q_h\cdot n^{\frac{1}{2}}t+\mathcal{O}(1).\end{aligned}$$

Therefore, we can conclude that

$$\begin{aligned}\sup_{t\in\mathcal{T}_M}\Bigg\|&\sum_{i=1}^{n}\big\{y_i-h\big(x_i,\beta^0\big)\big\}\cdot h'_\beta\big(x_i,\beta^0\big)\cdot\nu_i(n,t)-\\ &-n^{\frac{1}{2}}\cdot\sqrt{G^{-1}(\lambda)}\cdot\left\{f(-\sqrt{G^{-1}(\lambda)})+f(\sqrt{G^{-1}(\lambda)})\right\}\cdot Q_h t\Bigg\| \;=\; o_p(1)\end{aligned}$$

as $n\to+\infty$. This closes the proof once we recall that $g(z)=\frac{1}{2\sqrt{z}}\{f(\sqrt{z})+f(-\sqrt{z})\}$. □

C Proof of consistency and asymptotic normality

Proof of Theorem 4.1: This is a standard proof of consistency based on the uniform law of large numbers and the convergence of the order statistics $r^2_{[h_n]}(\beta)$ to the corresponding quantile $G^{-1}_\beta(\lambda)$. Let us denote the LTS objective function and its expectation by

$$\begin{aligned} S_{nn}(\beta) &= \frac{1}{n}\sum_{i=1}^{n} r_i^2(\beta)\cdot I\big(r_i^2(\beta)\le r^2_{[h_n]}(\beta)\big), \\ S(\beta) &= \mathsf{E}\left\{r_1^2(\beta)\cdot I\big(r_1^2(\beta)\le G^{-1}_\beta(\lambda)\big)\right\}. \end{aligned}$$

By definition, $P\Big(S_{nn}\left(\hat\beta_n^{(LTS,h_n)}\right) < S_{nn}\left(\beta^0\right)\Big) = 1$. For any $\delta>0$ and an open neighborhood $U(\beta^0,\delta)$ of β^0

$$\begin{aligned} 1 &= P\Big(S_{nn}\left(\hat\beta_n^{(LTS,h_n)}\right) < S_{nn}\left(\beta^0\right)\Big) \\ &= P\Big(S_{nn}\left(\hat\beta_n^{(LTS,h_n)}\right) < S_{nn}\left(\beta^0\right) \quad \text{and} \quad \hat\beta_n^{(LTS,h_n)}\in U(\beta^0,\delta)\Big) \\ &+ P\Big(S_{nn}\left(\hat\beta_n^{(LTS,h_n)}\right) < S_{nn}\left(\beta^0\right) \quad \text{and} \quad \hat\beta_n^{(LTS,h_n)}\in B\backslash U(\beta^0,\delta)\Big) \\ &\le P\Big(\hat\beta_n^{(LTS,h_n)}\in U(\beta^0,\delta)\Big) + P\Big(\inf_{\beta\in B\backslash U(\beta^0,\delta)} S_{nn}\left(\beta\right) < S_{nn}\left(\beta^0\right)\Big). \end{aligned}$$

Therefore, $P\big(\inf_{\beta\in B\backslash U(\beta^0,\delta)} S_{nn}\left(\beta\right) < S_{nn}\left(\beta^0\right)\big)\to 0$ as $n\to+\infty$ implies

$$P\Big(\hat\beta_n^{(LTS,h_n)}\in U(\beta^0,\delta)\Big)\to 1$$

as $n\to+\infty$, that is, the consistency of $\hat\beta_n^{(LTS,h_n)}$ (δ was an arbitrary positive number). To verify $P\big(\inf_{\beta\in B\backslash U(\beta^0,\delta)} S_{nn}\left(\beta\right) < S_{nn}\left(\beta^0\right)\big)\to 0$ note that

$$\begin{aligned} & P\left(\inf_{\beta\in B\backslash U(\beta^0,\delta)} S_{nn}\left(\beta\right) < S_{nn}\left(\beta^0\right)\right) \\ = \; & P\left(\inf_{\beta\in B\backslash U(\beta^0,\delta)} \left[S_{nn}\left(\beta\right) - S(\beta) + S(\beta)\right] < S_{nn}\left(\beta^0\right)\right) \\ = \; & P\left(\inf_{\beta\in B\backslash U(\beta^0,\delta)} \left[S_{nn}\left(\beta\right) - S(\beta)\right] < S_{nn}(\beta^0) - \inf_{\beta\in B\backslash U(\beta^0,\delta)} S(\beta)\right) \\ \le \; & P\left(\sup_{\beta\in B}\left|S_{nn}\left(\beta\right) - S(\beta)\right| > \inf_{\beta\in B\backslash U(\beta^0,\delta)} S(\beta) - S_{nn}(\beta^0)\right) \\ \le \; & P\left(2\sup_{\beta\in B}\left|S_{nn}\left(\beta\right) - S(\beta)\right| > \inf_{\beta\in B\backslash U(\beta^0,\delta)} S(\beta) - S(\beta^0)\right). \end{aligned}$$

Since the identification Assumption I2 implies

$$(\forall \delta > 0)\,(\exists \alpha > 0)\left(\inf_{\beta \in B - U(\beta^0, \delta)} S(\beta) - S(\beta^0) > \alpha\right),$$

it is enough to show that for all $\alpha > 0$

$$P\left(\sup_{\beta \in B} |S_n(\beta) - S(\beta)| > \alpha\right) \to 0 \text{ as } n \to +\infty.$$

This is a direct consequence of Lemma A.1 and Corollary A.7 for function $t(x_i, \varepsilon_i; \beta) = r_i^2(\beta)$, see Assumptions D, H1, and H5, because

$$\begin{aligned} S_{nn}(\beta) - S(\beta) &= \frac{1}{n}\sum_{i=1}^{n}\left\{r_i^2(\beta)\cdot\left[I\left(r_i^2(\beta) \le r_{[h_n]}^2(\beta)\right) - I\left(r_i^2(\beta) \le G_\beta^{-1}(\lambda)\right)\right]\right\} \\ &+ \frac{1}{n}\sum_{i=1}^{n}\left\{r_i^2(\beta)\cdot I\left(r_i^2(\beta) \le G_\beta^{-1}(\lambda)\right) - \mathsf{E}\left[r_1^2(\beta)\cdot I\left(r_1^2(\beta) \le G_\beta^{-1}(\lambda)\right)\right]\right\}. \end{aligned}$$

□

Proof of Theorem 4.2: We already know that $\hat{\beta}_n^{(LTS,h_n)}$ is consistent. Hence $P\left(\left\|\hat{\beta}_n^{(LTS,h_n)} - \beta^0\right\| > \rho\right) \to 0$ as $n \to \infty$ for any $\rho > 0$ (Theorem 4.1).

Further, we employ the almost sure second-order differentiability of

$$S_{nn}(\beta) = \frac{1}{n}\sum_{i=1}^{n} r_i^2(\beta)\cdot I\left(r_i^2(\beta) \le r_{[h_n]}^2(\beta)\right)$$

and

$$S(\beta) = \mathsf{E}\left\{r_1^2(\beta)\cdot I\left(r_1^2(\beta) \le G_\beta^{-1}(\lambda)\right)\right\}$$

at β^0 (see Lemma 2.2 and Assumption H1). Since

$$S_{nn}(\beta) = \frac{1}{n}\sum_{i=1}^{n} r_i^2(\beta)\cdot\left[I\left(r_i^2(\beta) \le r_{[h_n]}^2(\beta)\right) - I\left(r_i^2(\beta) \le G_\beta^{-1}(\lambda)\right)\right] \tag{64}$$

$$+ \frac{1}{n}\sum_{i=1}^{n} r_i^2(\beta)\cdot I\left(r_i^2(\beta) \le G_\beta^{-1}(\lambda)\right), \tag{65}$$

Assumptions H, Lemma A.1 and Corollary A.7 imply $S_{nn}(\beta) \to S(\beta)$ as $n \to \infty$ in probability. Using the same argument for the first two derivatives of $S_{nn}(\beta)$, see Lemma 2.2, $S'_{nn}(\beta) \to S'(\beta)$ and $S''_{nn}(\beta) \to S''(\beta)$ as $n \to \infty$ uniformly in $\beta \in U(\beta^0, \delta)$, whereby

$$S''(\beta^0) = 2\,\mathsf{E}\left\{\left[h'_\beta\left(x_i, \beta^0\right)h'_\beta\left(x_i, \beta^0\right)^T + r_1(\beta^0)h''_{\beta\beta}\left(x_i, \beta^0\right)\right]\cdot I\left(r_1^2(\beta) \le G_\beta^{-1}(\lambda)\right)\right\}$$

$$= \quad 2\,\mathsf{E}\left\{\left[h_\beta'(x_i,\beta^0)h_\beta'(x_i,\beta^0)^T\right]\cdot I\big(r_1^2(\beta)\le G_\beta^{-1}(\lambda)\big)\right\} = 2\lambda Q_h > 0$$

by Assumptions D2 and H5. Since Q_h is a positive definite matrix by Assumption H5, there is a constant $\rho, \delta > \rho > 0$, such that $\left\|S'(\beta)\right\| \ge C\left\|\beta-\beta^0\right\|$ for all $\beta \in U(\beta^0,\rho)$ and some $C>0$. Due to the consistency of $\hat\beta_n^{(LTS,h_n)}$, this implies that for any $\varepsilon>0$ there is some $n_0\in\mathbb{N}$ such that $\hat\beta_n^{(LTS,h_n)} \in U(\beta^0,\rho)$ and subsequently $\left\|S(\hat\beta_n^{(LTS,h_n)})\right\| \ge C\left\|\hat\beta_n^{(LTS,h_n)}-\beta^0\right\|$ for all $n>n_0$ with probability at least $1-\varepsilon$. Therefore, it is sufficient to show that $\sqrt{n}\left\|S'(\hat\beta_n^{(LTS,h_n)})\right\| = \mathcal{O}_p(1)$ to prove the theorem.

To analyze $\sqrt{n}S(\hat\beta_n^{(LTS,h_n)})$, let us express it for $n>n_0$ with probability greater than $1-\varepsilon$ as

$$\begin{aligned}
&\sqrt{n}\,\mathsf{E}\left\{r_1(\hat\beta_n^{(LTS)})h_\beta'\Big(x_i,\hat\beta_n^{(LTS)}\Big)\cdot I\Big(r_1^2(\hat\beta_n^{(LTS)})\le G_\beta^{-1}(\lambda)\Big)\right\}\\
\le\ &\sup_{\beta\in U(\beta^0,\rho)}\frac{1}{\sqrt{n}}\sum_{i=1}^n\Big\{-r_1(\beta)h_\beta'(x_i,\beta)\cdot I\big(r_1^2(\beta)\le G_\beta^{-1}(\lambda)\big) \qquad (66)\\
&\qquad\qquad +\mathsf{E}\left[r_1(\beta)h_\beta'(x_i,\beta)\cdot I\big(r_1^2(\beta)\le G_\beta^{-1}(\lambda)\big)\right]\Big\}\\
+\ &\sup_{\beta\in U(\beta^0,\rho)}\frac{1}{\sqrt{n}}\sum_{i=1}^n\Big\{r_1(\beta)h_\beta'(x_i,\beta)\cdot\big[I\big(r_1^2(\beta)\le G_\beta^{-1}(\lambda)\big)-I\big(r_1^2(\beta)\le r_{[h_n]}^2(\beta)\big)\big]\Big\} \qquad (67)
\end{aligned}$$

(recall that $S_{nn}'(\hat\beta_n^{(LTS)})=0$ by Lemma 2.2). We only have to show that both terms are bounded in probability. This result for (67) is a consequence of Lemma A.7 together with Assumptions H1 and H5. The other part (66) can be bounded in probability by the following argument. Assumption H3 together with van der Vaart and Wellner (1996, Lemma 2.6.18) imply that

$$\mathcal{F}_{n,\delta} = \Big\{r_1(\beta)h_\beta'(x_i,\beta)\cdot I\big(r_1^2(\beta)\le G_\beta^{-1}(\lambda)\big) : \beta\in U(\beta^0,\delta)\Big\}$$

form a VC class of functions. Therefore, Assumptions D1 and H3 permit the use of uniform central limit theorem of Arcones and Yu (1994), which implies that $\mathcal{F}_{n,\delta}$ converges in distribution to a Gaussian process with uniformly bounded paths, which confirms that (66) is bounded in probability. □

Proof of Theorem 4.3: Due to Theorem 4.2, $t_n = \sqrt{n}\left(\hat\beta_n^{(NLTS,h_n)}-\beta^0\right) = \mathcal{O}_p(1)$ as $n\to+\infty$. Therefore, using the asymptotic linearity of LTS (Theorem 3.1), we can write with probability arbitrarily close to one

$$\begin{aligned}
&n^{-\frac12}\left(D_n^1(t_n)+n^{\frac12}Q_h t_n\cdot C_\lambda\right)\\
=\ &n^{-\frac12}\left\{D_n^1\left[\sqrt{n}\left(\hat\beta_n^{(LTS,h_n)}-\beta^0\right)\right]+n^{\frac12}Q_hC_\lambda\cdot\sqrt{n}\left(\hat\beta_n^{(LTS,h_n)}-\beta^0\right)\right\}\\
=\ &o_p(1),
\end{aligned}$$

where $C_\lambda = \lambda - 2G^{-1}(\lambda)g\left(G^{-1}(\lambda)\right)$. Substituting for $D_n^1(t)$ yields

$$\begin{aligned} n^{-\frac{1}{2}} & \sum_{i=1}^{n}\Big[\Big\{y_i - h\Big(x_i,\hat\beta_n^{(LTS,h_n)}\Big)\Big\}\,h_\beta'\Big(x_i,\hat\beta_n^{(LTS,h_n)}\Big)\cdot I\Big(r_i^2\left(\hat\beta_n^{(LTS,h_n)}\right)\le r^2_{[h_n]}\left(\hat\beta_n^{(LTS,h_n)}\right)\Big)\\ &\qquad - \big\{y_i - h\big(x_i,\beta^0\big)\big\}\,h_\beta'\big(x_i,\beta^0\big)\cdot I\big(r_i^2(\beta^0)\le r^2_{[h_n]}(\beta^0)\big)\Big] +\\ + \quad & n^{\frac{1}{2}}Q_hC_\lambda\cdot\sqrt{n}\left(\hat\beta_n^{(LTS,h_n)}-\beta^0\right)\\ = \quad & o_p(1), \end{aligned}$$

and since the first summand in the previous equation is by the definition of $\hat\beta_n^{(LTS,h_n)}$ equal to zero, it follows that

$$\begin{aligned} \sqrt{n}\left(\hat\beta_n^{(LTS,h_n)}-\beta^0\right) &= n^{-\frac{1}{2}}Q_h^{-1}C_\lambda^{-1}\cdot\sum_{i=1}^{n}\big\{y_i-h\big(x_i,\beta^0\big)\big\}\,h_\beta'\big(x_i,\beta^0\big)\cdot I\big(r_i^2(\beta^0)\le r^2_{[h_n]}(\beta^0)\big)\\ &\qquad +o_p(1)\\ &= n^{-\frac{1}{2}}Q_h^{-1}C_\lambda^{-1}\cdot\sum_{i=1}^{n}r_i(\beta^0)h_\beta'\big(x_i,\beta^0\big)\cdot I\big(r_i^2(\beta^0)\le G^{-1}(\lambda)\big)+o_p(1)\\ &+ n^{-\frac{1}{2}}Q_h^{-1}C_\lambda^{-1}\cdot\sum_{i=1}^{n}r_i(\beta^0)h_\beta'\big(x_i,\beta^0\big)\times \qquad (68)\\ &\qquad\qquad\times\left[I\big(r_i^2(\beta^0)\le r^2_{[h_n]}(\beta^0)\big)-I\big(r_i^2(\beta^0)\le G^{-1}(\lambda)\big)\right]. \end{aligned}$$

First, we show that term (68) is negligible in probability. Recalling that $r_i(\beta^0)\equiv\varepsilon_i$, we can rewrite (68) as

$$n^{-\frac{1}{2}}Q_h^{-1}C_\lambda^{-1}\cdot\sum_{i=1}^{n}\varepsilon_i\cdot h_\beta'\big(x_i,\beta^0\big)\cdot\left[I\big(\varepsilon_i^2\le\varepsilon^2_{[h_n]}\big)-I\big(\varepsilon_i^2\le G^{-1}(\lambda)\big)\right].$$

Assumption D2 and Corollary A.6 implies for $k=1$ and 2 that

$$\mathsf{E}\left|\varepsilon_i\cdot\left[I\big(\varepsilon_i^2\le\varepsilon^2_{[h_n]}\big)-I\big(\varepsilon_i^2\le G^{-1}(\lambda)\big)\right]\right|^k=\mathcal{O}\Big(n^{-\frac{1}{2}}\Big)$$

as $n\to\infty$. Therefore, the summands in (68) multiplied by $n^{\frac{1}{4}}$ have a finite expectation and variance (ε_i and x_i are independent random variables):

$$\mathsf{E}\left|n^{\frac{1}{4}}\cdot\varepsilon_i\cdot h_\beta'\big(x_i,\beta^0\big)\cdot\left[I\big(\varepsilon_i^2\le\varepsilon^2_{[h_n]}\big)-I\big(\varepsilon_i^2\le G^{-1}(\lambda)\big)\right]\right|=o(1)$$

and by Assumption H5

$$\begin{aligned} & \mathsf{var}\left\{n^{\frac{1}{4}}\cdot\varepsilon_i\cdot h_\beta'\big(x_i,\beta^0\big)\cdot\left[I\big(\varepsilon_i^2\le\varepsilon^2_{[h_n]}\big)-I\big(\varepsilon_i^2\le G^{-1}(\lambda)\big)\right]\right\}\\ \le\quad & n^{\frac{1}{2}}\,\mathsf{E}_x\left\{h_\beta'\big(x_i,\beta^0\big)\cdot\mathsf{var}\left(\varepsilon_i\cdot\left|I\big(\varepsilon_i^2\le\varepsilon^2_{[h_n]}\big)-I\big(\varepsilon_i^2\le G^{-1}(\lambda)\big)\right|\,\middle|\,x_i\right)\cdot h_\beta'\big(x_i,\beta^0\big)^T\right\} \end{aligned}$$

$$
\begin{aligned}
&+ \quad n^{\frac{1}{2}} \, \mathsf{var}_x \left\{ h'_\beta\big(x_i, \beta^0\big) \cdot \mathsf{E}\left(\varepsilon_i \cdot \big| I\big(\varepsilon_i^2 \leq \varepsilon^2_{[h_n]}\big) - I\big(\varepsilon_i^2 \leq G^{-1}(\lambda)\big) \big| \, \big| \, x_i \right) \right\} \\
&\leq \quad \mathcal{O}(1) \left\{ \mathsf{E}\left[h'_\beta\big(x_i, \beta^0\big) \right] + \mathsf{var}\left[h'_\beta\big(x_i, \beta^0\big) \right] \right\} = \mathcal{O}(1).
\end{aligned}
$$

Now, because all indicators depend only on the squares of residuals ε_i^2 and error terms ε_i are symmetrically distributed (Assumption D2), we get for any $i = 1, \ldots, n$ and any $n \in \mathbb{N}$

$$
\mathsf{E}\left\{ n^{\frac{1}{4}} \cdot \varepsilon_i \cdot h'_\beta\big(x_i, \beta^0\big) \cdot \big[I\big(\varepsilon_i^2 \leq \varepsilon^2_{[h_n]}\big) - I\big(\varepsilon_i^2 \leq G^{-1}(\lambda)\big) \big] \right\} = 0,
$$

and even conditionally,

$$
\mathsf{E}\left\{ n^{\frac{1}{4}} \cdot \varepsilon_i \cdot h'_\beta\big(x_i, \beta^0\big) \cdot \big[I\big(\varepsilon_i^2 \leq \varepsilon^2_{[h_n]}\big) - I\big(\varepsilon_i^2 \leq G^{-1}(\lambda)\big) \big] \Big| \, \varepsilon_1, \ldots, \varepsilon_{i-1}, x_1, \ldots, x_{i-1} \right\} = 0.
$$

Therefore, $n^{\frac{1}{4}} \cdot \varepsilon_i \cdot h'_\beta(x_i, \beta^0) \cdot \left[I\left(\varepsilon_i^2 \leq \varepsilon^2_{[h_n]}\right) - I(\varepsilon_i^2 \leq G^{-1}(\lambda)) \right]$ forms a sequence of martingale differences with finite variances and we can apply the law of large number for the sum of martingale differences (68) (see Davidson, 1994, Theorem 20.11, for instance):

$$
n^{-\frac{3}{4}} Q_h^{-1} C_\lambda^{-1} \cdot \sum_{i=1}^{n} n^{\frac{1}{4}} \cdot \varepsilon_i \cdot h'_\beta\big(x_i, \beta^0\big) \cdot \big[I\big(\varepsilon_i^2 \leq \varepsilon^2_{[h_n]}\big) - I\big(\varepsilon_i^2 \leq G^{-1}(\lambda)\big) \big] \to 0
$$

in probability as $n \to \infty$. Thus, (68) is negligible in probability $o_p(1)$. Given this result,

$$
\begin{aligned}
\sqrt{n}\left(\hat{\beta}_n^{(LTS,h_n)} - \beta^0 \right) &= n^{-\frac{1}{2}} Q_h^{-1} C_\lambda^{-1} \sum_{i=1}^{n} \big\{ y_i - h\big(x_i, \beta^0\big) \big\} \cdot h'_\beta\big(x_i, \beta^0\big) \cdot I\big(r_i^2(\beta^0) \leq G^{-1}(\lambda)\big) \\
&\quad + o_p(1) \\
&= n^{-\frac{1}{2}} Q_h^{-1} C_\lambda^{-1} \sum_{i=1}^{n} \varepsilon_i \cdot h'_\beta\big(x_i, \beta^0\big) \cdot I\big(\varepsilon_i^2 \leq G^{-1}(\lambda)\big) + o_p(1), \qquad (69)
\end{aligned}
$$

which is the first assertion of the theorem.

Second, by the same argument as used in the above discussion of (68), the summands in (69) form a sequence of identically distributed martingale differences with finite second moments (Assumptions D2 and H5). Since by the law of large numbers for L^1-mixingales (Andrews, 1988)

$$
\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 \cdot h'_\beta\big(x_i, \beta^0\big) h'_\beta\big(x_i, \beta^0\big)^T \cdot I\big(\varepsilon_i^2 \leq G^{-1}(\lambda)\big) \to \mathsf{var}\left[\varepsilon_i \cdot h'_\beta\big(x_i, \beta^0\big) \cdot I\big(\varepsilon_i^2 \leq G^{-1}(\lambda)\big) \right]
$$

in probability as $n \to \infty$, we can employ the central limit theorem for martingale differences (for example, Davidson, 1994, Theorem 24.3) for (69). This results directly in the

asymptotic normality of $\hat{\beta}_n^{(LTS,h_n)}$. The asymptotic variance can be then expressed as

$$\begin{aligned} V &= C_\lambda^{-2} \cdot Q_h^{-1} \cdot \mathsf{var}\left[h'_\beta(x_i,\beta^0)\varepsilon_i \cdot I\left(\varepsilon_i^2 \le G^{-1}(\lambda)\right)\right] \cdot Q_h^{-1} \\ &= C_\lambda^{-2} \cdot Q_h^{-1} \cdot \mathsf{E}\left[h'_\beta(x_i,\beta^0)\varepsilon_i \cdot I\left(\varepsilon_i^2 \le G^{-1}(\lambda)\right)\right]\left[h'_\beta(x_i,\beta^0)\varepsilon_i \cdot I\left(\varepsilon_i^2 \le G^{-1}(\lambda)\right)\right]^T \cdot Q_h^{-1} \\ &= C_\lambda^{-2} \cdot Q_h^{-1} \cdot \mathsf{E}\left[h'_\beta(x_i,\beta^0)h'_\beta(x_i,\beta^0)^T\right] \cdot \mathsf{E}\left[\varepsilon_i^2 \cdot I\left(\varepsilon_i^2 \le G^{-1}(\lambda)\right)\right] \cdot Q_h^{-1} \\ &= C_\lambda^{-2} \cdot Q_h^{-1} \cdot Q_h\sigma_\lambda^2 \cdot Q_h^{-1} = C_\lambda^{-2}\sigma_\lambda^2 \cdot Q_h^{-1}. \end{aligned}$$

□

References

[1] Agulló J. (2001) New algorithms for computing the least trimmed squares regression estimator, *Computational Statistics & Data Analysis* **36(4)**: 425–439.

[2] Amemiya T. (1983) Non-linear regression models, in Griliches Z. and Intriligator M. D. (eds.) *Handbook of Econometrics Vol. 1*, North Holland, Amsterdam, 333–389.

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

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

[5] Andrews D. W. K. (1993) An introduction to econometric applications of empirical process theory for dependent random variables, *Econometric Reviews* **12(2)**: 183–216.

[6] Arcones M A. and Yu B. (1994) Central limit theorems for empirical and U-processes of stationary mixing sequences, *Journal of Theoretical Probability* **7**: 47–71.

[7] Bai E.-W. (2003) A random least-trimmed-squares identification algorithm, *Automatica* **39**: 1651–1659.

[8] Beňáček V., Jarolím M., and Víšek J. Á. (1998) Supply-side characteristics and the industrial structure of Czech foreign trade, *Proceedings of the conference Business and economic development in central and eastern Europe: Implications for economic integration into wider Europe, ISBN 80-214-1202-X*, Technical university in Brno together with University of Wisconsin, Whitewaters, and the Nottingham Trent university, 51–68.

[9] Chen Y., Stromberg A., and Zhou M. (1997) The least trimmed squares estimate in nonlinear regression. *Technical report*, 1997/365, Department of statistics, University of Kentucky.

[10] Christmann A. (1998) On positive breakdown point estimators in regression models with discrete response variables, *Habilitation thesis*, University of Dortmund, Germany.

[11] Čížek, P. (2001) Robust estimation in nonlinear regression models, *SFB 373 Discussion paper,* 2001/25*,* Humboldt University, Berlin.

[12] Čížek P. and Víšek J. Á. (2000) Least trimmed squares, in Härdle W., Hlávka Z., and Klinke S. (eds.) *XploRe Application Guide*, Springer, Heidelberg.

[13] Davidson, J. (1994) *Stochastic Limit Theory*, Oxford University Press, New York.

[14] Davidson, J. and de Jong, R. (1997) Strong laws of large numbers for dependent and heterogeneous processes: a synthesis of new and recent results, *Econometric Reviews* **16(3)**: 251–79.

[15] Giloni A. and Padberg M. (2002) Least trimmed squares regression, least median squares regression, and mathematical programming, *Mathematical and Computer Modelling* **35(9)**: 1043–1060.

[16] Hawkins D. M. and Olive D. (1999) Applications and algorithms for least trimmed sum of absolute deviations regression, *Computational Statistics & Data Analysis* **32**: 119–134.

[17] Jurečková J. (1984) Regression quantiles and trimmed least squares estimator under a general design, *Kybernetika* **20**: 345–357.

[18] Kelly M. (1997) Do noise traders influence stock prices?, *Journal of Money, Credit and Banking* **29(3)**: 351–363.

[19] Knez P. J. and Ready M. J. (1997) On the robustness of size and book-to-market in cross-sectional regressions, *The Journal of Finance* **52(4)**: 1355–1382.

[20] Pison G., Van Aelst S., and Willems G. (2002) Small sample corrections for LTS and MCD, *Metrika* **55**: 111–123.

[21] Pollard D. (1984) *Convergence of Stochastic Processes*, Springer, New York.

[22] Pollard D. (1989) Asymptotics via empirical processes, *Statistical Science* **4(4)**: 341–366.

[23] Rousseeuw P. J. (1984): Least median of squares regression. *Journal of American Statistical Association* **79**: 871–880.

[24] Rousseeuw P. J. (1985): Multivariate estimation with high breakdown point, in Grossman W., Pflug G., Vincze I., and Wertz W. (eds.) *Mathematical statistics and applications, Vol. B*, Reidel, Dordrecht, Netherlands, 283–297.

[25] Rousseeuw P. J. (1997) Introduction to positive-breakdown methods, in Maddala G. S. and Rao C. R. (eds.) *Handbook of statistics, Vol. 15: Robust inference*, Elsevier, Amsterdam, 101–121.

[26] Rousseeuw P. J. and Leroy A. M. (1987): *Robust regression and outlier detection*, Wiley, New York.

[27] Rousseeuw, P. J., and Van Driessen, K. (1999): Computing LTS regression for large data sets, *Technical report, University of Antwerp*, submitted.

[28] Sakata S. and White H. (1995) An alternative definition of finite-sample breakdown point with application to regression model estimators, *Journal of the American Statistical Association* **90**: 1099–1106.

[29] Sakata S. and White H. (1998) High breakdown point conditional dispersion estimation with application to S&P 500 daily returns volatility, *Econometrica* **66(3)**: 529–567.

[30] Stromberg A. J. (1993): High breakdown estimation of nonlinear regression parameters, *Journal of American Statistical Association* **88**: 237–244.

[31] Stromberg A. J., Hössjer O., and Hawkins D. M. (2000) The least trimmed difference regression estimator and alternatives, *Journal of the American Statistical Association* 95: 853–864.

[32] Stromberg A. J. and Ruppert D. (1992) Breakdown in nonlinear regression, *Journal of American Statistical Association* **87**: 991–997.

[33] Tableman M. (1994) The influence functions for the least trimmed squares and the least trimmed absolute deviations estimators, *Statistics & Probability Letters* **19(4)**: 329–337.

[34] Temple J. R. W. (1998) Robustness tests of the augmented solow model, *Journal of Applied Econometrics* **13(4)**: 361–375.

[35] Van der Vaart A. W. and Wellner J. A. (1996): *Weak convergence and empirical processes: with applications to statistics*, Springer, New York.

[36] Víšek J. Á. (2000): On the diversity of estimates, *Computational Statistics & Data Analysis* **34**: 67–89.

[37] Wang H. and Suter D. (2003) Using symmetry in robust model fitting, *Pattern Recognition Letters* **24(16)**: 2953–2966.

[38] Willems G. and Van Aelst S. (2004) Fast and robust bootstrap for LTS, *Computational Statistics & Data Analysis*, in press.

[39] White H. (1980) Nonlinear regression on cross-section data, *Econometrica* **48(3)**: 721–746.

[40] Ye M. and Haralick R. M. (2000): Optical flow from a least-trimmed squares based adaptive approach, *International conference on pattern recognition ICPR 2000*, Barcelona, Spain.

[41] Yu B. (1994) Rates of convergence for empirical processes of stationary mixing sequences, *The Annals of Probability* **22(1)**: 94–116.

[42] Zaman A., Rousseeuw P. J., and Orhan M. (2001) Econometric applications of high-breakdown robust regression techniques, *Economics Letters* **71**: 1–8.

[43] Zinde-Walsh V. (2002) Asymptotic theory for some high breakdown point estimators, *Econometric Theory* **18**: 1172–1196.