No. 2004–130

GENERAL TRIMMED ESTIMATION: ROBUST APPROACH TO NONLINEAR AND LIMITED DEPENDENT VARIABLE MODELS

By P. Čížek

December 2004

ISSN 0924-7815

General trimmed estimation: robust approach to nonlinear and limited dependent variable models

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 and data contamination. Motivated by some – the least trimmed squares and maximum trimmed likelihood estimators – we propose a general trimmed estimator, which unifies and extends many existing robust procedures. We derive here the consistency and rate of convergence of the proposed general trimmed estimator under mild β-mixing conditions and demonstrate its applicability in nonlinear regression, time series, limited dependent variable models, and panel data.

Keywords: consistency, regression, robust estimation, trimming

JEL codes: C13, C20, C24, C25

1 Introduction

In statistics and econometrics, more and 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. Evidence about contamination of data and its adverse effects on estimators such as (quasi-) maximum likelihood is provided, for example, by Gerfin (1996) in labor market data, by Sakata and White (1998) in financial time series, and by Čížek (2004a) in the prices of financial derivates. The global 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). One way to construct a high breakdown-point method is to employ a standard (parametric) estimator and to trim some "unlikely" observations from its objective function. This is, for example, the case of the least trimmed squares (LTS) by Rousseeuw (1985), the least trimmed absolute deviations (LTA) by Bassett (1991), and the maximum trimmed likelihood (MTLE) by Neykov and Neytchev (1990) and Hadi and Luceno (1997). Here we generalize the concept of trimming, prove its consistency, and demonstrate its applicability in many econometric models including nonlinear regression, time series, and limited dependent variable models. Additionally, we mention possible combinations of the "trimming principle" and semiparametric estimation.

First, let us briefly review existing results concerning the LTS, LTA, and MTLE estimators. 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).[1] Thus, LTS has been receiving a lot of attention from the theoretical, computational, and application points of view. There are extensions involving nonlinear regression (Stromberg, 1993), weighted LTS (Víšek, 2002), and adaptive smooth trimming (Čížek, 2002), and in most of these cases, the asymptotic and breakdown behavior is known in the standard regression with i.i.d. errors. Simultaneously, there has been a significant development in computational methods (Agulló, 2001; Gilloni and Padberg, 2002; Rousseeuw and van Driessen, 1999). Last, but not least, there are also first applications of LTS in economics (Beňáček, Jarolím, and Víšek, 1998; Temple, 1998; Zaman, Rousseeuw, and Orhan, 2001) and finance (Knez and Ready, 1997; Kelly, 1997).

Next, the LTA estimator has not attracted much attention yet despite its favorable computational and robustness properties (see Hawkins and Olive, 1999, for an overview and extensions of LTA). The asymptotic properties are known only in the univariate location model (Tableman, 1994). Finally, the MTLE estimator, which can produce the LMS, LTS, maximum likelihood, and some other estimators in special cases (Hadi and Luceno, 1997), has been studied from the robustness point of view (Vandev and Neykov, 1998; Müller and

[1] See also a recent proposal of smoothed LMS by Zinde-Walsh (2002).

Neykov, 2003) and applied in the context of (generalized) linear models (e.g., Neykov et al., 2004). Despite of the appealing concept of the trimmed likelihood, the asymptotic results are known only the case of linear regression with Gaussian errors (Vandev and Neykov, 1993).

The aim of this work is to generalize the principle of LTS, LTA, and MTLE, that is trimming "unlikely" observations from a model point of view. The proposed general trimmed estimator (GTE) does not only include LTS, LTA, and MTLE as special cases, but also allows for application of the trimming principle to many existing parametric and semiparametric estimators. Moreover, we prove its consistency and derive its rate of convergence under rather general conditions, which permit using trimmed estimators in a wide range of econometric applications including time series, panel data, and limited dependent variable models (additional conditions leading to the asymptotic normality of GTE are discussed as well). Thus, the application area of robust trimmed estimators is extended substantially. Another important consequence of the derived results is the consistency of LTA and MTLE in a general multivariate location and regression models, which was not available up to now. The main tools in achieving this are the (uniform) law of large numbers (Andrews, 1988 and 1992) and the uniform central limit theorem (Arcones and Yu, 1994, and Yu, 1994) for mixing processes. On the other hand, computational issues and robustness properties of GTE, which are analogous to LTS, LTA, and MTLE and motivate the use of trimmed estimators also as tools for regression diagnostics, are not discussed here to a larger extent because of a large number of existing studies that address the computation and breakdown behavior of trimmed estimator.

In the rest of the paper, we first propose the general trimmed estimator in Section 2, where we also extensively discuss assumptions needed for studying asymptotic properties of GTE. Asymptotic results are summarized in Section 3. A number of specific trimmed estimators in various econometric models is presented in Section 4. The proofs are provided in Appendix.

2 Generalized trimmed estimator

For the purpose of motivation, let us first present the LTS and MTLE estimators (Section 2.1 and 2.2). Later, the general trimmed estimator and the assumptions used in the paper are discussed (Sections 2.3 and 2.4) as well as an alternative definition of GTE (Section 2.5).

2.1 Least trimmed squares

Let us consider a 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 of explanatory variables x_i and unknown parameters β, and ε_i is a continuously distributed error term. The least trimmed squares estimator $\hat{\beta}_n^{(LTS,h)}$ is then defined by

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

where $r_{[j]}^2(\beta)$ represents the jth order statistics of squared residuals $r_i^2(\beta) = \{y_i - h(x_i, \beta)\}^2$ and B is a parameter space. The trimming constant h must satisfy $\frac{n}{2} < h \leq n$ and determines the breakdown point of the (nonlinear) LTS estimator since definition (4) implies that $n - h$ observations with the largest residuals do not directly affect the estimator. Thus, the observations that are unlikely, that is, observations that have very large residuals in a given parametric model, are dropped from the objective function. For $h(x, \beta) = g(x^\top \beta)$, 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 an overview of the properties of LTS in linear and nonlinear regression, see Čížek and Víšek (2000), Víšek (2000), and Čížek (2004b), Stromberg (1993), respectively.

2.2 Maximum trimmed likelihood

In the same way the LTS estimator is derived from the least squares, the maximum trimmed likelihood estimator follows from the maximum likelihood estimator (MLE). For a sample $(x_i, y_i)_{i=1}^n$, MTLE is defined by

$$\hat{\beta}_n^{(MTLE,h)} = \arg\max_{\beta \in B} \sum_{j=n-h+1}^{n} \ln l_{[j]}(x_i, y_i; \beta), \tag{3}$$

where $l_{[j]}(x_i, y_i; \beta)$ represents the jth order statistics of likelihood contributions $l(x_i, y_i; \beta), i = 1, \ldots, n$, and h is again the trimming constant. Compared to MLE, the $n - h$ observations with smallest likelihood values, that is least probable observations in a given model, are left out of the likelihood function. The robustness properties of MTLE are similar to those of LTS and they were studied in the linear and generalized linear regression models by Vandev and Neykov (1998) and Müller and Neykov (2003), respectively.

2.3 General trimmed estimator

Let us consider a random sample $(x_i, y_i)_{i=1}^n$, where $x_i \in \mathbb{R}^k$ represents a vector of explanatory variables and $y_i \in \mathbb{R}$ denotes the dependent variable.[2] Furthermore, assume that $s(x_i, y_i; \beta)$ represents a loss function identifying the true value β^0 of parameter vector $\beta \in B$, where $B \subseteq \mathbb{R}^p$ is a compact parametric space, and that large values of $s(x_i, y_i; \beta)$ represents unlikely observations for a given model ("bad fit") and small values of $s(x_i, y_i; \beta)$ correpond to likely values ("good fit"). For example, $s(x_i, y_i; \beta) = \{y_i - h(x_i, \beta)\}^2$ in the case of the least squares loss and $s(x_i, y_i; \beta) = -\ln l(x_i, y_i; \beta)$ in the case of the likelihood criterion.[3] The general trimmed estimator $\hat{\beta}_n^{(GTE,h)}$ can be then defined by

$$\hat{\beta}_n^{(GTE,h)} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{j=1}^{h} s_{[j]}(x_i, y_i; \beta), \tag{4}$$

where $s_{[j]}(x_i, y_i; \beta)$ represents the jth order statistics of $s(x_i, y_i; \beta), i = 1, \ldots, n$.[4] Apparently, this definition includes the LTS, LTA, and MTLE estimators as special cases.

Nevertheless, an even more general form of trimming is necessary to make trimmed estimation operational in some models (e.g., binary-choice or panel data models). Let us introduce an auxiliary trimming function $r(x_i, y_i; \beta)$, which also indicates likely and unlikely observations in a given model by small and large values, respectively. Further, let $s_{r:[j]}(x_i, y_i; \beta)$ be the value of $s(x, y; \beta)$ at observation (x_i, y_i) corresponding to the jth order statistics $r_{[j]}(x_i, y_i; \beta)$ of $r(x_i, y_i; \beta), i = 1, \ldots, n$. Then the general trimmed estimator is defined by

$$\hat{\beta}_n^{(GTE,h)} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{j=1}^{h} s_{r:[j]}(x_i, y_i; \beta). \tag{5}$$

[2] The assumption $x_i \in \mathbb{R}^k$ and $y_i \in \mathbb{R}$ correponds to most traditional use in regression models, but the presented results are valid also for $y_i \in \mathbb{R}^l$ and general multivariate models.

[3] For the sake of simplicity, we refer to $s(x_i, y_i; \beta)$ for a given $i \in \mathbb{N}$ as residuals or losses.

[4] For the jth order statistics of $s(x_i, y_i; \beta)$, I use symbol $s_{[j]}(x_i, y_i; \beta)$. In this case, index i inside the order statistics is just a formal notation and does not have any relationship to summation or other indices. It is to be understood so that x_i, y_i inside $s_{[j]}(x_i, y_i; \beta)$ just indicate the sample on which this order statistics is based.

In other words, the ordering of observations and their inclusion in the objective function is not given by ordering values $s(x_i, y_i; \beta)$ of the loss function $s(x, y; \beta)$, but by ordering values $r(x_i, y_i; \beta)$ of the auxiliary trimming function $r(x, y; \beta)$. Although the existing trimmed estimators are based on $r(x, y; \beta) = s(x, y; \beta)$, using GTE in binary-choice models, for instance, requires $r(x, y; \beta) = \mathsf{E}_y\, s(x, y; \beta)$ or $r(x, y; \beta) = \max_y s(x, y; \beta)$ (symbols E_y and $\max_y$ refer to the expectation and maximum taken only with respect to dependent variable y). See Section 4 for more details.

Before discussing assumptions concerning GTE, let us shortly return to the trimming constant h. Naturally, the choice of the trimming constant h should vary with the sample size n, and therefore, we have to work with a sequence of trimming constants h_n. As h_n/n determines the fraction of sample included in the GTE objective function, and consequently, the robustness properties of GTE, we want to asymptotically fix this fraction at λ, $\frac{1}{2} \leq \lambda \leq 1$. The trimming constant for a given 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 $h_n/n \to \lambda$.

2.4 Assumptions

Let us now complement the GTE definition first by some notation and definitions and later by assumptions on the loss and trimming functions and random variables needed for further analysis.

First, we refer to the distribution functions of $s(x_i, y_i; \beta)$ and $r(x_i, y_i; \beta)$ as $F_\beta(z)$ and $G_\beta(z)$ and to the corresponding probability density functions, if they exist, as $f_\beta(z)$ and $g_\beta(z)$, respectively. At the true parameter value β^0, we also use a simpler notation $F \equiv F_{\beta^0}$ and $G \equiv G_{\beta^0}$, and similarly for density functions, $f \equiv f_{\beta^0}$ and $g \equiv g_{\beta^0}$. Further, whenever we need to refer to the quantile functions corresponding to F_β and G_β, notation F_β^{-1} and G_β^{-1} is used, respectively. Next, because the derivatives of functions $s(x, y; \beta)$ and $r(x, y; \beta)$ are taken only with respect to β here, we denote tham simply by $s'(x, y; \beta)$, $r'(x, y; \beta)$, and so on. Two purely mathematical symbols we need are the indicator function $I(A)$, 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

$$\beta_m = \sup_{t\in\mathbb{N}} \mathsf{E} \sup_{B\in\sigma^f_{t+m}} |P(B|\sigma^p_t) - P(B)| \to 0$$

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

Now, I specify all the assumptions necessary to derive the $\sqrt{n}$ consistence of GTE (a smaller subset of assumptions sufficient for the consistency of GTE is discussed at the end of the section). They form three groups: distributional Assumptions D for random variables (x_i, y_i), Assumptions F concerning properties of the loss function $s(x, y; \beta)$ and auxiliary trimming function $r(x, y; \beta)$, and finally, identification Assumptions I.

Assumptions D

D1 Random variables $\{y_i, x_i\}_{i\in\mathbb{N}}$ form an identically distributed absolutely regular sequence of random vectors 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 The distribution function G_β of $r(x_i, y_i; \beta)$ is absolutely continuous for any $\beta \in B$.

D3 Assume that for $m_G = \inf_{\beta\in B} G^{-1}_\beta(\lambda)$ and $M_G = \sup_{\beta\in B} G^{-1}_\beta(\lambda)$, it holds that

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

and

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

for some $\delta_g > 0$.

Having a general objective function $s(x, y; \beta)$, Assumption D1 is a necessary condition for the uniform central limit theorem, see Andrews (1993) and Arcones and Yu (1994), for instance. Assumption D2 indicates that at least one random variable have to be continuously distributed. Assumption D3 formalizes two things: first, the density function g_β has to be bounded uniformly in $\beta \in B$, which actually prevents distribution G_β to become or be arbitrarily close to a discrete one for some $\beta \in B$. Second, the density function has to be positive in a neighborhood of the λ-quantile of G_β, that is, around the chosen "trimming" point of $r(x_i, y_i; \beta)$ distribution. In a less general setting, when structure of a model is known, Assumption D3 is usually implied by $G \equiv G_{\beta^0}$ being absolutely continuous with a density function $g \equiv g_{\beta^0}$ positive, bounded, and differentiable around $G^{-1}(\lambda)$; see Čížek (2004b) for the case of nonlinear regression model. Differentiability of density function g around the point corresponding to the λ-quantile of the $r(x_i, y_i; \beta^0)$ distribution is a standard condition needed for the analysis of rank statistics (see Víšek, 1999, and Zinde-Walsh, 2002, for instance).

Next, several conditions on the loss function $s(x_i, y_i; \beta)$ and auxiliary trimming function $r(x_i, y_i; \beta)$ 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 objective function of an estimator is almost always assumed to be twice differentiable; see Pakes and Pollard (1989). 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 F

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

F1 Let $s(x_i, y_i; \beta)$ and $r(x_i, y_i; \beta)$ be a continuous (uniformly over any compact subset of the support of x) in $\beta \in B$ and $s(x_i, y_i; \beta)$ be twice differentiable in β on $U(\beta^0, \delta)$ almost surely.

F2 Let $\{r(x_i, y_i; \beta) | \beta \in U(\beta^0, \delta)\}$ and $\{s'(x_i, y_i; \beta) | \beta \in U(\beta^0, \delta)\}$ form VC classes of functions such that their envelopes $E_1(x) = \sup_{\beta \in B} |r(x_i, y_i; \beta)|$ and $E_2(x) = \sup_{\beta \in U(\beta^0, \delta)}$ $|s'(x_i, y_i; \beta)|$ have finite r_β-th moments.

F3 Expectations $\mathsf{E} \sup_{\beta \in B} |r(x_i, y_i; \beta)|^{1+\delta}$, $\mathsf{E} \sup_{\beta \in B} |s(x_i, y_i; \beta)|^{1+\delta}$, and finally, $\mathsf{E} \sup_{\beta \in U(\beta^0, \delta)}$

$|s(x_i, y_i; \beta)|^{1+\delta}$ exist and are finite for $l = 1, 2$. Moreover, assume that $\mathsf{E}\, s''(x_i, y_i; \beta^0) = Q_s > 0$, where Q_s is a nonsingular positive definite matrix.

F4 Conditional expectation $\mathsf{E}\left\{\sup_{\beta \in U(\beta^0,\delta)} \left[s'(x_i, y_i; \beta) | \, r(x_i, y_i; \beta) \in C\right]\right\}$ is uniformly bounded over all intervals $C \in \mathbb{R}$ such that $G^{-1}(\lambda) \in C$.

Whereas the differentiability of the objective function and the existence of some moments are standard assumptions, Assumption F2 deserves further comments, because it limits the class of functions $s'(x, y; \beta)$ and $r(x, y; \beta)$ to VC classes (see Pollard, 1984, and van der Vaart, 1996, for a definition). Although limited, they cover many common functions including polynomial, logarithmic, and exponential functions, functions such that $|f(x, t) - f(x, t')| \leq \xi(x) \, \|t - t'\|^{\alpha}$ for some $\alpha > 0$ and nonnegative $\xi(x)$, their sums, products, maxima and minima, composed function and so on. Even though this assumption is not necessarily restrictive in many contexts and it is not needed for the proof of consistency, it can be omitted as long as we impose stronger distributional assumptions. For example, assume that function $r(x, y; \beta)$ is continuously differentiable in $\beta \in U(\beta^0, \delta)$, its derivative $r'(x_i, y_i; \beta)$ can be bounded by $M(z_i)$ on $\beta \in U(\beta^0, \delta)$, where M is an integrable function of a subset z_i of variables (y_i, x_i), and the distribution of $r(x_i, y_i; \beta)$ conditional on z_i is absolutely continuous (this is satisfied in linear regression for $r(x, y; \beta) = (y - x^{\top}\beta)^2$ and $z_i = x_i$, for instance). Then it is possible to prove the L^{r_β}-continuity of $I\Big(r(x_i, y_i; \beta) \leq G_{\beta}^{-1}(\lambda)\Big)$ in $U(\beta^0, \delta)$ 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.

Additionally, the proof of $\sqrt{n}$ consistency requires an unusual regularity assumption Assumption F4, which is the only and rather weak link between the loss function $s(x, y; \beta)$ and auxiliary trimming function $r(x, y; \beta)$. First notice that conditioning by large intervals C is not important here since the conditional expectation converges to the unconditional one for $C \to \mathbb{R}$ (Assumption F3). Considering small intervals around $G^{-1}(\lambda)$, Assumption F4 just expresses the idea that the loss function should not behave wildly "around" the trimming point (i.e., for x_i, y_i, and β such that $r(x_i, y_i; \beta)$ is close to $G_{\beta}^{-1}(\lambda)$). To exemplify, let us use once again a linear regression model with $s(x, y; \beta) = r(x, y; \beta) = (y - x^{\top}\beta)^2$. Then $s'(x_i, y_i; \beta) = (y_i - x_i^{\top}\beta)x_i$ and conditioning has a form $\{(y_i - x_i^{\top}\beta)^2 - G^{-1}(\lambda)\} \in (-a, b)$, where $a, b > 0$. For $a \to 0, b \to 0$, the conditional expectation becomes

$$\mathsf{E}\left\{(y_i - x_i^\top \beta)x_i | (y_i - x_i^\top \beta)^2 = G^{-1}(\lambda)\right\} = G^{-1}(\lambda)\,\mathsf{E}\left\{x_i \operatorname{sgn}(y_i - x_i^\top \beta)\right\},$$

which is guaranteed by the existence of the second moments of x_i.

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

$$\begin{aligned} \min_{\beta \in B \backslash U(\beta^0,\varepsilon)} & \mathsf{E}\left[s(x_i, y_i; \beta) \cdot I\Big(r(x_i, y_i; \beta) \le G_\beta^{-1}(\lambda)\Big)\right] \\ & - \mathsf{E}\left[s(x_i, y_i; \beta^0) \cdot I\Big(r(x_i, y_i; \beta^0) \le G_{\beta^0}^{-1}(\lambda)\Big)\right] \quad > \quad \alpha(\varepsilon). \end{aligned}$$

To close this section, let us note that Assumptions D, F, and I are sufficient to prove the $\sqrt{n}$ consistency of GTE. If only consistency is needed, one can omit all assumptions concerning differentiability and derivatives of the regression function $s(x_i, y_i; \beta)$ (Assumptions F), Assumption F2 on VC classes, and also weaken Assumption D1, since centered $s(x_i, y_i; \beta)$ can form a $L^{1+\delta}$-mixingale in the most general case (Andrews, 1988).

2.5 Alternative definition

Before proving the main results of the paper, some basic properties of the GTE objective function $S_n(\beta) = \sum_{j=1}^{h_n} s_{r:[j]}(x_i, y_i; \beta)$ and its alternative formulation, which is more suitable for deriving asymptotic results, are introduced.

Lemma 1 *Under Assumptions D2 and F1, $S_n(\beta)$ is continuous on B, twice differentiable at $\hat\beta_n^{(GTE,h_n)}$ as long as $\hat\beta_n^{(GTE,h_n)} \in U(\beta^0, \delta)$, and almost surely twice differentiable at any fixed point $\beta \in U(\beta^0, \delta)$. Furthermore,*

$$S_n(\beta) = \sum_{i=1}^{n} s(x_i, y_i; \beta) \cdot I\big(r(x_i, y_i; \beta) \le r_{[h_n]}(x_i, y_i; \beta)\big), \tag{6}$$

$$\begin{aligned} S_n^{'}(\beta) = \frac{\partial S_n(\beta)}{\partial \beta} &= \sum_{i=1}^{n} s'(x_i, y_i; \beta) \cdot I\big(r(x_i, y_i; \beta) \leq r_{[h_n]}(x_i, y_i; \beta)\big) && (7) \\ S_n^{''}(\beta) = \frac{\partial^2 S_n(\beta)}{\partial \beta \partial \beta^T} &= \sum_{i=1}^{n} s''(x_i, y_i; \beta) \cdot I\big(r(x_i, y_i; \beta) \leq r_{[h_n]}(x_i, y_i; \beta)\big) && (8) \end{aligned}$$

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 (4) 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 Consistency

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

$$\begin{aligned} S_n(\beta) &= \sum_{i=1}^{n} s(x_i, y_i; \beta) \cdot I\big(r(x_i, y_i; \beta) \leq r_{[h_n]}(x_i, y_i; \beta)\big) \\ &= \sum_{i=1}^{n} s(x_i, y_i; \beta) \Big[I\big(r(x_i, y_i; \beta) \leq r_{[h_n]}(x_i, y_i; \beta)\big) - I\Big(r(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda)\Big)\Big] && (9) \\ &+ \sum_{i=1}^{n} s(x_i, y_i; \beta) \cdot I\Big(r(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda)\Big). && (10) \end{aligned}$$

Whereas the first part (9) will be shown to be small because of the convergence of order statistics to quantiles in mean, $r_{[h_n]}(x_i, y_i; \beta) \to G_\beta^{-1}(\lambda)$, the second part (10) will be dealt with by standard asymptotic tools and shown to converge to

$$S(\beta) = \mathsf{E}\left\{ s(x, y; \beta) \cdot I\Big(r(x, y; \beta) \leq G_\beta^{-1}(\lambda)\Big)\right\}.$$

First, using the uniform law of large numbers, we prove the consistency of the GTE estimator $\hat{\beta}_n^{(GTE, h_n)}$ minimizing $S_n(\beta)$ on the parameter space B.

Theorem 2 *Let $s(x_i, y_i; \beta)$ and $r(x_i, y_i; \beta)$ be continuous functions on B as specified in Assumption F1 and let Assumptions D, F3, and I hold. Then the general trimmed estimator*

$\hat{\beta}_n^{(GTE,h_n)}$ *minimizing (6) is weakly consistent, that is,* $\hat{\beta}_n^{(GTE,h_n)} \to \beta^0$ *in probability as* $n \to +\infty$.

Proof: See Appendix B. □

Next, we will derive the rate of convergence of $\hat{\beta}_n^{(GTE,h_n)}$ to β^0. 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 F3, $Q_s > 0$, implies that $\|\partial S(\beta)/\partial\beta\| \geq C\left\|\beta - \beta^0\right\|$ in a neighborhood $U(\beta^0, \rho)$ for some $C > 0$ and $\rho > 0$. Since the consistency of GTE guarantees that $\hat{\beta}_n^{(GTE,h_n)} \in U(\beta^0, \rho)$ with probability approaching 1 as $n \to +\infty$, we just have to prove that $\left\|\partial S(\hat{\beta}_n^{(GTE,h_n)})/\partial\beta\right\| = \mathcal{O}_p\left(n^{-\frac{1}{2}}\right)$. This can be again done by using decomposition (9)–(10).

Theorem 3 *Let Assumptions D, F, and I hold. Then* $\hat{\beta}_n^{(GTE,h_n)}$ *is* $\sqrt{n}$*-consistent, that is,*

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

as $n \to +\infty$.

Proof: See Appendix B. □

Finally, the asymptotic distribution of GTE would be of interest, but we are not able to derive it in this general setting. Let us note however that the asymptotic normality was proved in the case of nonlinear regression for LTS (Čížek, 2004b) and the same idea and steps can be used in practically any regression model with reduced form (1) under (slightly extended) conditions for L^p-continuity of $I\left(r(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda)\right)$ mentioned in Section 2.4. More precisely, function $r(x, y; \beta)$ should be continuously differentiable in $\beta \in U(\beta^0, \delta)$, its derivative $r'(x_i, y_i; \beta)$ has to be bounded by $M(z_i)$ on $\beta \in U(\beta^0, \delta)$, where M is an integrable function of a subset z_i of variables (y_i, x_i), and the distribution of $r(x_i, y_i; \beta)$ conditional on z_i is absolutely continuous with a density function, which is positive, bounded, and differentiable around $G^{-1}(\lambda)$.

4 Examples of trimmed estimators

In this section, we discuss various trimmed estimators and models where they can be applied. To verify their feasibility, we check the identification Assumption I2, at least locally, as discussed in Section 4.1. Later, we present examples of trimmed estimators based on the least-squares loss in nonlinear, times series, truncated, and censored regression (Section 4.2), on the likelihood function in nonlinear and binary-response regression (Section 4.3), and their use in panel data context (Section 4.4). Finally, we shortly treat possible combinations of the GTE approach and semiparametric estimator (Section 4.5).

4.1 Identification condition

A crucial ingredient of the consistency of GTE is the identification Assumption I2, which differs from a usual least squares or maximum likelihood identification condition by inclusion of trimming. Plainly, the identification Assumption I2 can also be formulated such that

$$IC(\beta) = \mathsf{E}\left[s(x_i, y_i; \beta) \cdot I\Big(r(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda)\Big)\right] \tag{11}$$

as a function of β has a unique minimum at β^0. Since it is rather difficult to verify that β^0 is a global minimum without having a specific model in hand, we concentrate only on local behavior of $IC(\beta)$: we try to justify that β^0 is a local minimum of $IC(\beta)$ by checking

$$\frac{\partial IC(\beta^0)}{\partial \beta} = 0 \quad \text{and} \quad \frac{\partial^2 IC(\beta^0)}{\partial \beta^2} > 0 \tag{12}$$

(twice differentiability of $s(x_i, y_i; \beta)$ is guaranteed by Assumption F1). Additionally, using Lemmas 1 and 4 (Appendix A), we can write

$$\frac{\partial IC(\beta^0)}{\partial \beta} = \mathsf{E}\left[s'(x_i, y_i; \beta^0) \cdot I\big(r(x_i, y_i; \beta^0) \leq G^{-1}(\lambda)\big)\right] = 0 \tag{13}$$

$$\frac{\partial^2 IC(\beta^0)}{\partial \beta^2} = \mathsf{E}\left[s''(x_i, y_i; \beta^0) \cdot I\big(r(x_i, y_i; \beta^0) \leq G^{-1}(\lambda)\big)\right] > 0, \tag{14}$$

which are limits of (7) and (8); see the proof of Theorem 3 in Appendix B for details.

In the rest of this section, we try to verify conditions (13) and (14) for various models and estimators. In all cases, we assume that a given "normal" estimator, which corresponds to no trimming, $\lambda = 1$, is locally identified and we discuss additional assumption necessary for "trimmed" identification conditions, $\lambda < 1$.

4.2 Least trimmed squares

Let us now discuss GTE based on the least-squares loss, which in (non)linear regression coincides with well-known LTS. After dealing with identification Assumption I2 and the use of GTE in time series, an example of least-squares based GTE in truncated and censored regression is given to demonstrate wider applicability of GTE compared to LTS.

The LTS estimator, considered here for nonlinear regression model (1), is a special case of GTE for $r(x, y; \beta) = s(x, y; \beta) = \{y - h(x, \beta)\}^2$. To apply GTE in the context of nonlinear regression, the standard identification assumptions for least squares estimator – the orthogonality $\mathsf{E}(\varepsilon_i | x_i) = 0$ and spheriality $\mathsf{E}\left[h'(x_i, \beta^0) h'(x_i, \beta^0)^\top\right] = Q_h > 0$ conditions – have to be augmented by the symmetry of the conditional distribution of ε_i given x_i, which guarantees that

$$\mathsf{E}\left[\varepsilon_i \cdot I\big(\varepsilon_i^2 \leq G^{-1}(\lambda)\big) \big| \, x_i\right] = 0. \tag{15}$$

First, let us verify condition (13):

$$\begin{aligned}
\frac{\partial IC(\beta^0)}{\partial \beta} &= \mathsf{E}\left[s'(x_i, y_i; \beta^0) \cdot I\big(r(x_i, y_i; \beta^0) \leq G^{-1}(\lambda)\big)\right] \\
&= \mathsf{E}\left[-2\left\{y_i - h(x_i, \beta^0)\right\} h'_\beta(x_i, \beta^0) \cdot I\Big(\left\{y_i - h(x_i, \beta^0)\right\}^2 \leq G^{-1}(\lambda)\Big)\right] \\
&= \mathsf{E}\Big\{-2h'_\beta(x_i, \beta^0)\, \mathsf{E}\left[\varepsilon_i \cdot I\big(\varepsilon_i^2 \leq G^{-1}(\lambda)\big) \big| \, x_i\right]\Big\} = 0.
\end{aligned}$$

Second, condition (14) can be verified similarly:[5]

$$\begin{aligned}
\frac{\partial^2 IC(\beta^0)}{\partial \beta^2} &= \mathsf{E}\left[s''(x_i, y_i; \beta^0) \cdot I\big(r(x_i, y_i; \beta^0) \leq G^{-1}(\lambda)\big)\right] \\
&= \mathsf{E}\left[2h'_\beta(x_i, \beta^0) h'_\beta(x_i, \beta^0)^\top \cdot I\Big(\left\{y_i - h(x_i, \beta^0)\right\}^2 \leq G^{-1}(\lambda)\Big)\right] \\
&\quad - \mathsf{E}\left[2\left\{y_i - h(x_i, \beta^0)\right\} h''_\beta(x_i, \beta^0) \cdot I\big(\left\{y_i - h(x_i, \beta^0)\right\} \leq G^{-1}(\lambda)\big)\right] \\
&= \mathsf{E}\left[2h'_\beta(x_i, \beta^0) h'_\beta(x_i, \beta^0)^T \cdot \mathsf{E}\left\{I\big(\varepsilon_i^2 \leq G^{-1}(\lambda)\big) \big| \, x_i\right\}\right] \\
&\quad - \mathsf{E}\Big\{-2h''_\beta(x_i, \beta^0)\, \mathsf{E}\left[\varepsilon_i \cdot I\big(\varepsilon_i^2 \leq G^{-1}(\lambda)\big) \big| \, x_i\right]\Big\} \\
&= \lambda Q_{hh} > 0.
\end{aligned}$$

Let us note that given Assumptions D for the consistency of GTE, its application in nonlinear regression models is not limited only to a classical cross-sectional regression. Linear

[5] For the sake of simplicity, we assume homoscedasticity here.

and nonlinear regression models are used also in time series estimation, for example, in the smooth threshold autoregressive model (STAR), which allows for a smooth transition between states by means of a general function $h(y_{t-d}; c, \delta) : \mathbb{R} \to \langle 0, 1\rangle$:

$$y_t = \alpha_0 + \sum_{i=1}^{p} y_{t-i}\alpha_i + \left(\delta_0 + \sum_{i=1}^{p} y_{t-i}\delta_i\right) \cdot h(y_{t-d}; c, \delta) + \varepsilon_t$$

(see Dijk, Terasvirta, and Franses, 2000, for a survey). For $h \equiv 0$, we obtain a standard autoregressive process of order p. In this context, a use of a robust method such as GTE is very advisable because, contrary to cross-sectional estimation, a single observation influences not only its own residual, but also regression residuals of $p - 1$ following observations.

Additionally, GTE can be used in a wider range of models than LTS. For example, least squares and LTS are not consistent in a truncated regression model, where a linear regression $y_i^* = x_i^\top \beta + \varepsilon_i$ with symmetrically distributed ε_i is presumed, but $(y_i = y_i^*, x_i)$ can be observed only if $y_i^* > 0$. On the other hand, Powell (1986) proposed symmetrically truncated least squares (STLS) estimator, which restores the symmetry of distribution Φ_x of ε conditional on x by truncating its tail and employes least squares afterwards. Specifically in our example, Φ_x is truncated from below at $-x^\top\beta$, and therefore, it can be symmetrized by truncating from above at $+x^\top\beta$. Powell (1986) shows that this can be achieved by minimizing $\sum_{i=1}^{n}(y_i - x_i^\top\beta)^2 \cdot I(y_i - x_i^\top\beta < x_i^\top\beta)$ with respect to β. Since the objective function is continuous and differentiable in β almost everywhere, it is possible to define the corresponding trimmed STLS estimator $\hat{\beta}_n^{(GTE-STLS,h)}$ by setting

$$s(x, y; \beta) = r(x, y; \beta) = (y_i - x_i^\top\beta)^2 \cdot I(y_i < 2x_i^\top\beta).$$

Note that this also applies in censored regression models, where STLS would be replaced by symmetrically censored least squares (SCLS) of Powell (1986).

To conclude this example, let us verify identification conditions (13) and (14) for GTE–SLTS under the previously mentioned assumptions: orthogonality $\mathsf{E}(\varepsilon_i|x_i) = 0$, spheriality $\mathsf{E}(x_i x_i^\top) = Q_x > 0$, and conditional symmetry of Φ_x distribution. The first derivative of $IC(\beta)$ equals

$$\frac{\partial IC(\beta^0)}{\partial\beta} \quad = \quad \mathsf{E}\left[-2(y_i - x_i^\top\beta^0)x_i \cdot I\Big(y_i \leq 2x_i^\top\beta^0\Big) \cdot I\big(s(x_i, y_i; \beta^0) \leq G^{-1}(\lambda)\big)\Big|\, x_i\right]$$

$$= \mathsf{E}\left[x_i \, \mathsf{E}\left\{-2\varepsilon_i \cdot I\Big(\varepsilon_i \le x_i^\top \beta^0\Big) \cdot I\big(\varepsilon_i^2 \le G^{-1}(\lambda)\big)\Big|\, x_i\right\}\right] = 0.$$

Similarly, the second derivative is

$$\begin{aligned}\frac{\partial^2 IC(\beta^0)}{\partial \beta^2} &= \mathsf{E}\left[2x_i x_i^\top \cdot I\Big(y_i \le 2x_i^\top \beta^0\Big) \cdot I\big(s(x_i, y_i; \beta^0) \le G^{-1}(\lambda)\big)\Big|\, x_i\right] \\ &= \mathsf{E}\left\{2x_i x_i^\top \, \mathsf{E}\left[I\Big(\varepsilon_i \le x_i^\top \beta^0\Big) \cdot I\big(\varepsilon_i^2 \le G^{-1}(\lambda)\big)\Big|\, x_i\right]\right\} > 0\end{aligned}$$

as long as $\mathsf{E}\left[I\big(\varepsilon_i \le x_i^\top \beta^0\big)\big|\, x_i\right] > 0$.

4.3 Maximum trimmed likelihood

Our next examples concern GTE based on the likelihood function, which in (non)linear regression coincides with MTLE. After mentioning briefly identification of MTLE in nonlinear regression, we again focus on examples, where standard MTLE does not apply, but it is possible to construct a likelihood-based GTE: binary-choice and truncated regression.[6]

The MTLE estimator in nonlinear regression model (1) is also a special case of GTE for $r(x, y; \beta) = s(x, y; \beta) = \ln \phi\{y - h(x, \beta)\}$, where ϕ denotes the density function of ε_i. Conditions (13) and (14) can be verified in the same way as for LTS in Section 4.2. The most important additional assumption is again the (conditional) symmetry of the ε_i distribution, which implies that introducing "trimming" into the identification conditions does not invalidate them. For example under conditional symmetry of ϕ given x_i, $\mathsf{E}[\phi'(\varepsilon_i)/\phi(\varepsilon_i)|x_i] = 0$ implies

$$\mathsf{E}\left[\frac{\phi'(\varepsilon_i)}{\phi(\varepsilon_i)} I\big(-\ln \phi(\varepsilon_i) \le G^{-1}(\lambda)\big)\Big|\, x_i\right] = 0.$$

Applying the GTE concept to maximum likelihood estimation becomes less trivial once we consider less "continuous" models, such as binary-choice models. In this case, the dependent variable takes on only two values, $y_i \in \{0, 1\}$, and its conditional expectation is described by $\mathsf{E}(y_i|x_i) = P(y_i = 1|x_i) = \Phi(x_i^\top \beta)$, where Φ is a symmetric absolutely continuous distribution function (e.g., standard normal distribution function in the case of probit). The log-likelihood contribution is then described by

$$s(x_i, y_i; \beta) = -\ln l(x_i, y_i; \beta) = y_i \ln \Phi(x_i^\top \beta) + (1 - y_i) \ln\Big\{1 - \Phi(x_i^\top \beta)\Big\}.$$

[6] Even though the results are applicable in censored regression as well, we opt for truncated regression for the sake of easier and more concise presentation.

The MTLE estimator, which uses $r(x_i, y_i; \beta) = s(x_i, y_i; \beta)$, cannot be applied because the identification condition (13) is not satisfied for $\lambda < 1$ (ϕ denotes the density function corresponding to Φ):

$$\begin{aligned}
\frac{\partial IC(\beta^0)}{\partial \beta} &= \mathsf{E}\left[\left\{-\frac{y_i \phi(x_i^\top \beta^0)}{\Phi(x_i^\top \beta^0)} x_i + \frac{(1-y_i)\phi(x_i^\top \beta^0)}{1-\Phi(x_i^\top \beta^0)} x_i\right\} \cdot I\big(-\ln l(x_i, y_i; \beta^0) \le G^{-1}(\lambda)\big)\right] \\
&= \mathsf{E}\left[-P(y_i = 1|x_i)\frac{\phi(x_i^\top \beta^0)}{\Phi(x_i^\top \beta^0)} x_i \cdot I\big(-\ln l(x_i, 1; \beta^0) \le G^{-1}(\lambda)\big)\right] \qquad (16) \\
&+ \mathsf{E}\left[+P(y_i = 0|x_i)\frac{\phi(x_i^\top \beta^0)}{1-\Phi(x_i^\top \beta^0)} x_i \cdot I\big(-\ln l(x_i, 0; \beta^0) \le G^{-1}(\lambda)\big)\right] \qquad (17) \\
&= \mathsf{E}\Big\{\phi(x_i^\top \beta^0) x_i \times \\
&\quad \times \Big[I\Big(-\ln\{1-\Phi(x_i^\top \beta^0)\} \le G^{-1}(\lambda)\Big) - I\Big(-\ln \Phi(x_i^\top \beta^0) \le G^{-1}(\lambda)\Big)\Big]\Big\}.
\end{aligned}$$

equals in general zero only if for all possible values of the random vector x,

$$I\Big(-\ln \Phi(x^\top \beta^0) \le G^{-1}(\lambda)\Big) = I\Big(\ln\{1-\Phi(x^\top \beta^0)\} \le G^{-1}(\lambda)\Big), \qquad (18)$$

that is, if $G^{-1}(\lambda) = +\infty$ and $\lambda = 1$.[7]

On the other hand, this derivation hints that the identification condition would be satisfied if the trimming function $r(x, y; \beta)$ satisfies $r(x, 0; \beta) = r(x, 1; \beta)$; see (16)–(18). Therefore, we propose to set $r(x_i, y_i; \beta) = \max\big\{-\ln \Phi(x^\top \beta^0), -\ln[1-\Phi(x^\top \beta^0)]\big\}$ and use GTE minimizing

$$\sum_{j=1}^{h} -\ln l(x_i, y_i; \beta) \cdot I\Big(-\max_{y \in \{0,1\}} \ln l(x_i, y; \beta) \le G_\beta^{-1}(\lambda)\Big).$$

The conditions (13) and (14) can be then verified analogously to (16)–(17).

Finally, let us recall the truncated regression model mentioned in Section 4.2, which are usually estimated by a maximum likelihood estimator. As we learned, a crucial condition for applying the trimming principle is the symmetry of the error distribution. Therefore, MTLE cannot be used in such cases because even if the underlying error distribution is symmetric, limited observability (truncation or censoring) destroys the symmetry. On the other hand, it is possible to construct a likelihood-based GTE estimator using the idea of Powell (1986)'s STLS: we can symmetrically truncate the conditional distribution Φ_x of ε given x so that symmetry is restored. For example, if Φ_x and its density ϕ_x are truncated from below at

[7] We neglect the other "solution," $\lambda = 0$, which results in objective function constantly equal to zero.

$-x^\top\beta$, they can be truncated from above at $x^\top\beta$ to achieve symmetry. Consequently, the original likelihood contribution $l(x, y; \beta) = \phi_x(y - x^\top\beta)/\{1 - \Phi_x(y - x^\top\beta)\} \cdot I(y > 0)$ is replaced by

$$l(x, y; \beta) = \frac{\phi_x(y - x^\top\beta)}{\{\Phi_x(x^\top\beta) - \Phi_x(-x^\top\beta)\}} \cdot I\Big(0 \leq y \leq 2x^\top\beta\Big). \tag{19}$$

4.4 Panel data

Even though regression estimation in panel data is based to a large extent on the same methods as cross-section and time series estimation, and therefore, the application of GTE seems to follow the rules discussed in Sections 4.2 and 4.3, there is one extra feature of GTE worth mentioning. Since we allow that the loss function $s(x, y; \beta)$ is in general different from the auxiliary trimming function $r(x, y; \beta)$, it is possible to apply trimming to something else then just individual observations. For example, panel data typically consist of observations on a large number N of inviduals (cross-sectional units) over T time periods: $(y_{it}, x_{it})_{i=1,t=1}^{N,\ T}$. Especially if the number T of time periods is small, one can consider, instead of trimming single observations, to perform trimming across individuals. In such a case, the trimming function $r(x, y; \beta)$ could be a sum of losses per each individual, $r(x_{it}, y_{it}; \beta) = \sum_{t=1}^{T} s(x_{it}, y_{it}; \beta)$, or the worst loss of each individual, $r(x_{it}, y_{it}; \beta) = \max_{t=1,\ldots,T} s(x_{it}, y_{it}; \beta)$, for all $t = 1, \ldots, T$.

4.5 Semiparametric estimation

Last, but not least, one can ask whether the trimming principle used in GTE can be combined with semi- and nonparametric estimators.[8] Unfortunately, the derived results do not allow in their current form to plug in a nonparametric estimator, for example, to propose a trimmed form of Ichimura (1993)'s semiparametric least squares estimator of (1) with an unknown regression function. Moreover, such an estimator would be probably computationally infeasible. On the other hand, some estimators based on approximating unknown regression or likelihood functions by a series expansion could be suitable candidates for deriving a corresponding trimmed method. For instance, the seminonparametric likelihood approach by Gallant and Nychka (1987) relies on maximum likelihood principle and approximation of an

[8] Note that previously mentioned STLS and SCLS are often considered semiparametric estimators too, but here we have in mind estimators using smoothing or series expansions to approximate an unknown regression or likelihood function.

unknown density function $\phi(x)$ by

$$\phi(x) \approx \phi^a(x) = P_k^2(x-\tau)\phi^2\{x|\tau, \mathrm{diag}(\gamma)\},$$

where P_k^2 is a polynomial of order $k \in \mathbb{N}$. Hence, defining GTE by $s(x, y; \beta) = r(x, y; \beta) = -\ln \phi^a(x, y; \beta)$ leads to a computationally feasible semiparametric estimator provided that $\phi^a(x)$ is a symmetric function, that is, coefficients of polynomial $P_k^2(x-\tau)$ are zero for odd powers of $x - \tau$.

5 Conclusion

Motivated by LTS, LTA, and MTLE, we proposed a general trimmed estimator, which extends the applicability of high breakdown-point methods to a wide range of econometric models, including nonlinear regression, time series, and limited dependent variable models. Thus, GTE allows to employ classical parametric methods, but adds a protection against contamination of data. The following conclusions concerns further asymptotic properties of GTE, its extensions and use in applications.

Although we proved the consistency and the rate of convergence under rather general conditions, it seems that results concerning the asymptotic distribution of GTE can be derived only if the structure of a model and an underlying estimator becomes more specific. Thus, this asymptotic result has to be probably derived on the case-by-case basis, although the arguments are likely to follow similar lines as the proof of asymptotic normality of LTS by Čížek (2004b).

Furthermore, we discussed only the most basic form of trimmed estimation, where observations are either included in or excluded from the GTE objective function. Nevertheless, various weighted trimmed estimators and data-adaptive choice of trimming, only recently introduced for LTS and MTLE, are straightforward to apply.

Finally, we argued that computational, robustness, and finite sample properties of GTE should be analogous to existing results concerning LTS, LTA, and MTLE. On the other hand, most existing robust estimators are studied and applied in the context of location or linear regression models, whereas possible applications of GTE also involve rather complex nonlinear models. Hence, simulation studies have to be employed to learn more about finite sample behavior of GTE under different circumstances. Last, but not least, existing algorithms for evaluating trimmed estimators have to be adapted to many different models and implemented.

Appendix

Here we present the proofs of lemmas and theorems on the order statistics of $\{r(x_i, y_i; \beta)\}_{i=1}^n$ and the GTE objective function (Appendix A) and on the consistency of GTE (Appendix B). Note that the alternative definition (6) of GTE is employed in all proofs. Additionally, notation $S_{nn}(\beta) = S_n(\beta)/n$ and symbol Ω for the probability space, on which $\{x_i, y_i\}$ is defined, are used.

A Lemmas on order statistics and GTE objective function

Proof of Lemma 1: 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 $T_1(\beta), \ldots, T_l(\beta)$, where $T_j(\beta) = \sum_{i=1}^{h_n} s(x_{k_{ji}}, y_{k_{ji}}; \beta)$, $j = 1, \ldots, l = \binom{n}{h_n}$, and $\{k_{j1}, \ldots, k_{jh_n}\} \in \{1, \ldots, n\}^{h_n}$ 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 (almost surely).

2. In all other cases, β lies on a boundary in the sense that there are some $j_1, \ldots, j_m$ such that $S_n(\beta) = T_{j_1}(\beta) = \ldots = T_{j_m}(\beta)$ (that is, some residuals being present in the GTE 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 $S_n(\beta) = T_{j_1}(\beta) = \ldots = T_{j_m}(\beta)$ and all functions $T_{j_i}, i = 1, \ldots, m$, 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 $\hat{\beta}_n^{(GTE,h_n)} \in U(\beta^0, \delta)$ as $T'_{j_1}(\hat{\beta}_n^{(GTE,h_n)}) =$

$$\ldots = T^{'}_{j_m}(\hat{\beta}_n^{(GTE,h_n)}) = 0; \text{ otherwise, } \hat{\beta}_n^{(GTE,h_n)} \text{ would not minimize } S_n(\beta).$$

Now, consider a fixed $\beta \in U(\beta^0, \delta)$ (n is still fixed). Assumption D2 implies that $r(x_i, y_i; \beta)$ is continuously distributed. Therefore, the probability that any two residuals at a given β are equal is zero:

$$P\left(\Omega_0 = \{\omega \in \Omega^n \,|\exists i, j \in \{1,\ldots,n\}, i \neq j, \text{ such that } r(x_i, y_i; \beta, \omega) = r(x_j, y_j; \beta, \omega)\,\}\right) = 0.$$

Moreover, there is a $\delta' > 0$ such that $r(x_i, y_i; \beta)$ 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 $\kappa(\omega) = \frac{1}{2}\min_{i,j=1,\ldots,n; i\neq j} |r(x_i, y_i; \beta, \omega) - r(x_j, y_j; \beta, \omega)| > 0$ we can find an $\varepsilon(\omega) > 0$ such that it holds that $\sup_{\beta' \in U(\beta,\delta')} |r(x_i, y_i; \beta') - r(x_i, y_i; \beta)| < \kappa(\omega)$ for all $i = 1, \ldots, n$. Consequently, the ordering of $r(x_1, y_1; \beta), \ldots, r(x_n, y_n; \beta)$ is constant for all $\beta' \in U(\beta, \delta')$ and there exist j such that $S_n(\beta) = T_j(\beta)$ 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 $r(x_i, y_i; \beta) = r(x_j, y_j; \beta)$ 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

$$\begin{aligned} S_n(\beta) &= \sum_{i=1}^{n} s(x_i, y_i; \beta) \cdot I\big(r(x_i, y_i; \beta) \leq r_{[h_n]}(x_i, y_i; \beta)\big) \\ S_n^{'}(\beta) = \frac{\partial S_n(\beta)}{\partial \beta} &= \sum_{i=1}^{n} s'(x_i, y_i; \beta) \cdot I\big(r(x_i, y_i; \beta) \leq r_{[h_n]}(x_i, y_i; \beta)\big) \\ S_n^{''}(\beta) = \frac{\partial^2 S_n(\beta)}{\partial \beta \partial \beta^T} &= \sum_{i=1}^{n} s''(x_i, y_i; \beta) \cdot I\big(r(x_i, y_i; \beta) \leq r_{[h_n]}(x_i, y_i; \beta)\big) \end{aligned}$$

almost surely for $\beta \in B$ and $\beta \in U(\beta^0, \delta)$, respectively. □

The next lemma just verifies that the uniform law of large numbers is applicable for trimmed sums.

Lemma 4 *Let Assumptions D, F1, and I1 hold and assume that $t(x, y; \beta)$ is a real-valued function continuous in β uniformly in x and y over any compact subset of the support of*

(x, y). Moreover, assume that $\mathsf{E}\sup_{\beta\in B}|t(x_i, y_i; \beta)|^{1+\delta} < \infty$ for some $\delta > 0$. Then

$$\begin{aligned}\sup_{\beta\in B, K\in\mathbb{R}} \Bigg| \frac{1}{n}\sum_{i=1}^{n} &\Big[t(x_i, y_i; \beta)\cdot I\Big(r(x_i, y_i; \beta) \le G_{\beta}^{-1}(\lambda) + K\Big)\Big] \\ &- \mathsf{E}\Big[t(x_i, y_i; \beta)\cdot I\Big(r(x_i, y_i; \beta) \le G_{\beta}^{-1}(\lambda) + K\Big)\Big]\Bigg| \quad \to \quad 0\end{aligned}$$

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).[9] 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

$$\begin{aligned}d(x_i, y_i; \beta, K) &= t(x_i, y_i; \beta)\cdot I\Big(r(x_i, y_i; \beta) \le G_{\beta}^{-1}(\lambda) + K\Big) \\ &- \mathsf{E}\Big[t(x_i, y_i; \beta)\cdot I\Big(r(x_i, y_i; \beta) \le G_{\beta}^{-1}(\lambda) + K\Big)\Big]\end{aligned}$$

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

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} &\Big[t(x_i, y_i; \beta)\cdot I\Big(r(x_i, y_i; \beta) \le G_{\beta}^{-1}(\lambda) + K\Big)\Big] \\ &- \mathsf{E}\Big[t(x_i, y_i; \beta)\cdot I\Big(r(x_i, y_i; \beta) \le G_{\beta}^{-1}(\lambda) + K\Big)\Big] \quad \xrightarrow{P} \quad 0\end{aligned}$$

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 $d(x_i, y_i; \beta, K)$ 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:

$$\lim_{\rho\to 0} P\left(\sup_{\beta\in B, K\in\mathbb{R}}\ \sup_{\beta'\in U(\beta,\rho), K'\in U(K,\rho)} \left|t_I(x_i, y_i; \beta', K') - t_I(x_i, y_i; \beta, K)\right| > \kappa\right) = 0 \qquad (20)$$

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

for any $\kappa > 0$, where $t_I(x_i, y_i; \beta, K) = t(x_i, y_i, \beta) \cdot I\Big(r(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda) + K\Big)$. To simplify the notation, we write 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 $K \in \mathbb{R}$

$$\begin{aligned}
& \sup_{\beta,K} \sup_{\beta',K'} \left|t_I(x_i, y_i; \beta', K') - t_I(x_i, y_i; \beta, K)\right| \\
\leq & \sup_{\beta,K} \sup_{\beta',K'} \left|t(x_i, y_i; \beta') \left[I\Big(r(x_i, y_i; \beta') \leq G_{\beta'}^{-1}(\lambda) + K'\Big) - I\Big(r(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda) + K\Big)\right]\right| \qquad (21) \\
+ & \sup_{\beta,K} \sup_{\beta',K'} \left|\left[t(x_i, y_i; \beta') - t(x_i, y_i; \beta)\right] I\Big(r(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda) + K\Big)\right| \qquad (22)
\end{aligned}$$

Hence, we can verify assertion (20) by proving it for expressions (21) and (22). 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 $\rho < \rho_0$.

1. Let us start with (21). First, observe that

$$\begin{aligned}
& \sup_{\beta,K} \sup_{\beta',K'} \left|t(x_i, y_i; \beta') \left[I\Big(r(x_i, y_i; \beta') \leq G_{\beta'}^{-1}(\lambda) + K'\Big) - I\Big(r(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda) + K\Big)\right]\right| \\
\leq & \sup_{\beta \in B} |t(x_i, y_i; \beta)| \times \qquad (23) \\
\times & \sup_{\beta,K} \sup_{\beta',K'} \left|I\Big(r(x_i, y_i; \beta') \leq G_{\beta'}^{-1}(\lambda) + K'\Big) - I\Big(r(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda) + K\Big)\right|,
\end{aligned}$$

where $\sup_\beta |t(x_1, y_1; \beta)|$ is a function independent of β possessing a finite expectation. Because the difference $\left|I\Big(r(x_i, y_i; \beta') \leq G_{\beta'}^{-1}(\lambda) + K'\Big) - I\Big(r(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda) + K\Big)\right|$ is always lower or equal to one, (21) has an integrable majorant independent of β. Therefore, if we show that

$$P\left(\sup_{\beta,K} \sup_{\beta',K'} \left|I\Big(r(x_i, y_i; \beta') \leq G_{\beta'}^{-1}(\lambda) + K'\Big) - I\Big(r(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda) + K\Big)\right| = 1\right) \to 0 \qquad (24)$$

as $\rho \to 0$, it implies, that (23) 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 F1 states that $r(x_i, y_i; \beta') \to r(x_i, y_i; \beta)$ for $\beta' \to \beta$ uniformly over any compact subset of the support of x, that is, $r(x_i, y_i; \beta') \to r(x_i, y_i; \beta)$ for $\beta' \to \beta$ in probability uniformly on B. Recalling that $G_\beta(x)$ is the cumulative distribution function of $r(x_i, y_i; \beta)$, 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 (Assumption D2) also implies that $G_{\beta'}^{-1}(\lambda)$ converges 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 $r(x_i, y_i; \beta)$ (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 $\sup_{\beta,\beta'} |r(x_i, y_i; \beta', \omega) - r(x_i, y_i; \beta, \omega)| < \frac{\varepsilon}{8M_{gg}}$ for all $\omega \in \Omega_1$ and $\rho < \rho_2$ (Assumption F1). Hence, setting $\rho_0 = \min\{\rho_1, \rho_2\}$, it follows that

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

for any $\rho < \rho_0$ because M_{gg} is the uniform upper bound for the probability density functions of $r(x_i, y_i; \beta)$ around $G_\beta^{-1}(\lambda)$ over all $\beta \in B$. Thus, we have proved (24), and consequently, we have verified that the expectation of (21) converges to zero for $\rho \to 0$ in probability.

2. We should deal now with (22) and prove that for any given $\kappa > 0$

$$\lim_{\rho\to 0} P\left(\sup_{\beta,K}\sup_{\beta',K'}\left|\left[t(x_i,y_i;\beta') - t(x_i,y_i;\beta)\right] I\left(r(x_i,y_i;\beta) \le G_\beta^{-1}(\lambda)+K\right)\right| > \kappa\right) = 0. \quad (25)$$

First, note that the difference

$$\left|t(x_i,y_i;\beta') - t(x_i,y_i;\beta)\right| \le \left|t(x_i,y_i;\beta')\right| + |t(x_i,y_i;\beta)| \le 2\sup_\beta |t(x_i,y_i;\beta)|$$

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 $2\,\mathsf{E}\sup_\beta |t(x_i, y_i; \beta)| = U_E$.

Second, for an arbitrary fixed $\varepsilon > 0$, we can find a compact subset A_ε of the support of (x_i, y_i) (and its complement $\overline{A_\varepsilon}$) such that $P((x_i, y_i) \in A_\varepsilon) > 1 - \frac{\kappa\varepsilon}{2U_E}$ (both x_i and y_i are random variables with finite second moments) and $2\int_{\overline{A_\varepsilon}} \sup_{\beta\in B} |t(x_i, y_i; \beta)| < \frac{\kappa\varepsilon}{2}$. Given this set A_ε and $\beta \in B$, we can employ continuity of $t(x_i, y_i; \beta)$ in β (uniform over all $(x_1, y_1) \in A_\varepsilon$) and find a $\rho_0 > 0$ such that

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

Hence,

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

and consequently,

$$\begin{aligned}
& P\left(\sup_{\beta,K}\sup_{\beta',K'} \left|\left[t(x_i,y_i;\beta') - t(x_i,y_i;\beta)\right] \cdot I\Big(r(x_i,y_i;\beta) \leq G_\beta^{-1}(\lambda) + K\Big)\right| > \kappa\right) \\
\leq\ & \frac{1}{\kappa}\,\mathsf{E}\left[\sup_{\beta,K}\sup_{\beta',K'} \left|\left[t(x_i,y_i;\beta') - t(x_i,y_i;\beta)\right] \cdot I\Big(r(x_i,y_i;\beta) \leq G_\beta^{-1}(\lambda) + K\Big)\right|\right] \\
\leq\ & \kappa\varepsilon/\kappa = \varepsilon
\end{aligned}$$

for any $\rho < \rho_0$. Hence, we have verified that (25).

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 5 *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, F1, F3, and I1, it holds that*

$$\sup_{\beta\in B} \left|r_{[h_n]}(x_i,y_i;\beta) - G_\beta^{-1}(\lambda)\right| \to 0 \qquad (26)$$

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

$$E_{Gn} = \mathsf{E}\sup_{\beta\in B} \left|r_{[h_n]}(x_i,y_i;\beta) - G_\beta^{-1}(\lambda)\right| \to 0 \qquad (27)$$

as $n \to +\infty$.

Proof: Let us recall that $r(x_i, y_i; \beta) \sim G_\beta$. 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 will find $n_0 \in \mathbb{N}$ such that for all $n > n_0$

$$P\left(\sup_{\beta \in B} \left| r_{[h_n]}(x_i, y_i; \beta) - G_\beta^{-1}(\lambda) \right| > K_1 \right) < \varepsilon, \tag{28}$$

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

$$v_{1i}(\beta, K_1) = I\Big(r(x_i, y_i; \beta) \le G_\beta^{-1}(\lambda) + K_1 \Big).$$

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

$$\mathsf{E}\, v_{1i}(\beta, K_1) = P(v_{1i}(\beta, K_1) = 1) = P\Big(r(x_i, y_i; \beta) \le G_\beta^{-1}(\lambda) + K_1 \Big) \ge \lambda,$$

it follows that $\mathsf{E}\, v_{1i}(\beta, K_1) \in (\lambda, 1\rangle$. Further, Lemma 4 for choice $t(x, y; \beta) = 1$ 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 $1 - \varepsilon/2$

$$-\frac{1}{2} K_\varepsilon + \sum_{i=1}^{n} \mathsf{E}\, v_{1i}(\beta, K_1) \le \sum_{i=1}^{n} v_{1i}(\beta, K_1). \tag{29}$$

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 (29) implies that

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

But this means for all $\beta \in B$ that at least $n\lambda \geq h_n$ of values $r(x_i, y_i; \beta)$ are smaller than $G_\beta^{-1}(\lambda) + K_1$. In other words, $r_{[h_n]}(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda) + K_1$ 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

$$v_{2i}(\beta, K_1) = I\Big(r(x_i, y_i; \beta) \geq G_\beta^{-1}(\lambda) - K_1\Big).$$

Finally, combining these two inequalities results in (26). Since $r(x_i, y_i; \beta)$ is uniformly integrable due to Assumption F3 and Davidson (1994, Theorem 12.10), $r_{[h_n]}(x_i, y_i; \beta)$ is uniformly integrable as well and the second claim follows directly from the (26) 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 6 *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, F, and I1, there is some $\varepsilon > 0$ such that*

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

and

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

for $n \to +\infty$.

Proof: The proof has a structure rather similar to the proof of Lemma 5. 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

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

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

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

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

Now, Assumption F2 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 process with uniformly bounded and uniformly continuous pathes. Consequently, we can find some $\varepsilon>0$ and a constant $U>0$

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

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

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

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

$$-\frac{1}{2}\sqrt{n}\cdot K_\varepsilon+\sum_{i=1}^{n}\mathsf{E}\,v_{1i}(\beta,K_1)\le\sum_{i=1}^{n}v_{1i}(\beta,K_1). \tag{30}$$

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 (30) imply that

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

But this means for all $\beta\in U(\beta^0,\varepsilon)$ that at least $n\lambda\ge h_n$ of values $r(x_i,y_i;\beta)$ are smaller than $G_\beta^{-1}(\lambda)+n^{-\frac{1}{2}}K_\varepsilon$. In other words, $r_{[h_n]}(x_i,y_i;\beta)\le G_\beta^{-1}(\lambda)+n^{-\frac{1}{2}}K_\varepsilon$ 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

$$v_{2i}(\beta,K_1)=I\Big(r(x_i,y_i;\beta)\ge G_\beta^{-1}(\lambda)-n^{-\frac{1}{2}}K_1\Big).$$

These inequalities can be rewritten as $Z_n=\sup_{\beta\in U(\beta^0,\varepsilon)}n^{-\frac{1}{2}}\left|r_{[h_n]}(x_i,y_i;\beta)-G_\beta^{-1}(\lambda)\right|\le K_\varepsilon$, 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. □

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

Lemma 7 *Under Assumptions D, F1, F3, and I1, it holds for any $i \leq n$*

$$P_G = P\Bigg(\sup_{\beta \in B} \Big| I\big(r(x_i, y_i; \beta) \leq r_{[h_n]}(x_i, y_i; \beta)\big) - I\Big(r(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda)\Big)\Big| \neq 0\Bigg) = o(1).$$

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

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

as $n \to +\infty$.

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

$$\nu_{in}(\beta) = I\big(r(x_i, y_i; \beta) \leq r_{[h_n]}(x_i, y_i; \beta)\big) - I\Big(r(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda)\Big).$$

Without loss of generality, we discuss only the case $v_{in}(\beta) = -1$, which corresponds to $r_{[h_n]}(x_i, y_i; \beta) < r(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda)$. 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 $(r(x_1, y_1; \beta, \omega_1), \ldots, r(x_{n-1}, y_{n-1}; \beta, \omega_{n-1}))$

$$r_{[h_n]}(x_i, y_i; \beta, \omega) = \begin{cases} r_{[h_n - 1]}(x_i, y_i; \beta, \omega') & \text{if } r(x_n y_n; \beta, \omega_n) < r_{[h_n-1]}(x_i, y_i; \beta, \omega') \\ r(x_n, y_n; \beta, \omega_n) & \text{if } r_{[h_n-1]}(x_i, y_i; \beta, \omega') \leq r(x_n, y_n; \beta, \omega_n) \\ & \text{and } r(x_n, y_n; \beta, \omega_n) \leq r_{[h_n]}(x_i, y_i; \beta, \omega') \\ r_{[h_n]}(x_i, y_i; \beta, \omega') & \text{if } r_{[h_n]}(x_i, y_i; \beta, \omega') < r(x_n, y_n; \beta, \omega_n) \end{cases} \qquad (31)$$

Denoting Ω_1, Ω_2, and Ω_3 subsets of Ω^n corresponding to the three (disjoint) cases in (31), 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. $P_1 = P(\{\omega \in \Omega_1 | \exists \beta \in B : \nu_{nn}(\beta) = -1\})$
$\leq P\big(\exists \beta \in B : r_{[h_n]}(x_i, y_i; \beta, \omega) < r(x_n, y_n; \beta, \omega_n) < r_{[h_n]}(x_i, y_i; \beta, \omega)\big) = 0.$

2. $P_2 = P(\{\omega \in \Omega_2 | \exists \beta \in B : \nu_{nn}(\beta) = -1\})$
$= P\Big(\exists \beta \in B : r_{[h_n-1]}(x_i, y_i; \beta, \omega') \leq r(x_n, y_n; \beta, \omega_n) = r_{[h_n]}(x_i, y_i; \beta, \omega) \leq G_\beta^{-1}(\lambda)\Big)$ 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. $P_3 = P(\{\omega \in \Omega_3 | \exists \beta \in B : \nu_{nn}(\beta) = -1\})$
$= P\Big(\exists \beta \in B : r_{[h_n]}(x_i, y_i; \beta, \omega') = r_{[h_n]}(x_i, y_i; \beta, \omega) < r(x_n, y_n; \beta, \omega_n) \leq G_\beta^{-1}(\lambda)\Big).$ We can structure this last term in the following way (Assumption D3):

$$\begin{aligned} & P\Big(\exists \beta \in B : r_{[h_n]}(x_i, y_i; \beta, \omega') < r(x_n, y_n; \beta, \omega_n) \leq G_\beta^{-1}(\lambda)\Big) = && (32) \\ & = \int_{\omega' \in \Omega^{n-1}} \int_{\omega_n \in \Omega} \sup_{\beta \in B} I\Big(r_{[h_n]}(x_i, y_i; \beta, \omega') < r(x_n, y_n; \beta, \omega_n) \leq G_\beta^{-1}(\lambda)\Big) dP(\omega_1) dP(\omega') \\ & = \int_{\omega' \in \Omega^{n-1}} M_{gg} \cdot \sup_{\beta \in B} \left| r_{[h_n]}(x_i, y_i; \beta, \omega') - G_\beta^{-1}(\lambda) \right| dP(\omega') \\ & = M_{gg} \cdot \mathsf{E} \left\{ \sup_{\beta \in B} \left| r_{[h_n]}(x_i, y_i; \beta, \omega') - G_\beta^{-1}(\lambda) \right| \right\}. && (33) \end{aligned}$$

The first claim of the lemma, $P_G = o(1)$, is then a direct consequence of Lemma 5. 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

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

and employ Lemma 6. □

Corollary 8 *Let Assumptions D, F1, F3, and I1 hold and assume that $t(x, y; \beta)$ is a real-valued function continuous in β uniformly in x and y over any compact subset of the support of (x, y). Moreover, assume that $\mathsf{E} \sup_{\beta \in B} |t(x_i, y_i; \beta)| < \infty$. Then it holds that*

$$\mathsf{E}\left\{\sup_{\beta\in B}\left|t(x_i,y_i,\beta)\left[I\big(r(x_i,y_i;\beta)\le r_{[h_n]}(x_i,y_i;\beta)\big)-I\Big(r(x_i,y_i;\beta)\le G_\beta^{-1}(\lambda)\Big)\right]\right|\right\}=o(1).$$

Additionally, if Assumptions D, F, and I1 hold and there exists $\varepsilon > 0$ such that

$$\mathsf{E}\left\{\sup_{\beta\in U(\beta^0,\varepsilon)}\left[|t(x_i,y_i,\beta)||\,I\big(r(x_i,y_i;\beta)\le r_{[h_n]}(x_i,y_i;\beta)\big)\ne I\Big(r(x_i,y_i;\beta)\le G_\beta^{-1}(\lambda)\Big)\right]\right\}<M_t$$

is bounded,

$$\mathsf{E}\left\{\sup_{\beta\in U(\beta^0,\varepsilon)}\left|t(x_i,y_i,\beta)\left[I\big(r(x_i,y_i;\beta)\le r_{[h_n]}(x_i,y_i;\beta)\big)-I\Big(r(x_i,y_i;\beta)\le G_\beta^{-1}(\lambda)\Big)\right]\right|\right\}$$
$$=\mathcal{O}\Big(n^{-\frac{1}{2}}\Big)$$

as $n \to +\infty$.

Proof: This can verified along the same lines as Lemma 7. Defining functions $\nu_{in}(\beta)$ and sets Ω_1, Ω_2, and Ω_3 exactly the same way as in Lemma 7, 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 7, we will treat only part concerning $\int_{\Omega_3}$ and assume without loss of generality that $i = n$.

First, since the expectation

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

has an integrable majorant and $P\big(\sup_{\beta \in B} |\nu_{in}(\beta)| = 1\big)$ converges to zero as $n \to +\infty$ (Lemma 7), the whole expectation converges to zero as well, which is the first claim of this corollary.

Second, similarly to (32)–(33), we can write

$$\mathsf{E}\left\{\sup_{\beta\in U(\beta^0,\varepsilon)}|t(x_n,y_n,\beta)\cdot\nu_{in}(\beta)|\right\}\le\int_{\Omega_3}\left\{\sup_{\beta\in U(\beta^0,\varepsilon)}|t(x_n,y_n,\beta)\cdot\nu_{in}(\beta)|\right\}dP(\omega)$$
$$\le\int_{\omega'\in\Omega^{n-1}}\int_{\omega_n\in\Omega}\mathsf{E}\left\{\sup_{\beta\in U(\beta^0,\varepsilon)}[|t(x_n,y_n,\beta)||\,|\nu_{in}(\beta)|=1]\right\}\sup_{\beta\in U(\beta^0,\varepsilon)}|\nu_{in}(\beta)|\,dP(\omega')dP(\omega_n)$$

$$\begin{aligned}
&\leq M_t \int_{\omega' \in \Omega^{n-1}} \int_{\omega_n \in \Omega} \sup_{\beta \in U(\beta^0,\varepsilon)} |\nu_{in}(\beta)| \, dP(\omega') dP(\omega_n) \\
&\leq n^{-\frac{1}{2}} M_t M_{gg} \int_{\omega' \in \Omega^{n-1}} \sqrt{n} \sup_{\beta \in U(\beta^0,\varepsilon)} \left| r_{[h_n]}(x_i, y_i; \beta, \omega') - G_\beta^{-1}(\lambda) \right| dP(\omega')
\end{aligned}$$

Thus, we obtain from Lemma 6

$$\mathsf{E}\left\{ \sup_{\beta \in B} |t(x_n, y_n, \beta) \cdot \nu_{in}(\beta)| \right\} \leq n^{-\frac{1}{2}} M_t M_{gg} E_{Ln} \cdot \int_{\omega_n \in \Omega} \sup_{\beta \in B} |t(x_n, y_n, \beta)| \, dP(\omega_n) = \mathcal{O}\Big(n^{-\frac{1}{2}}\Big),$$

which closes the proof. □

Corollary 9 *Under assumptions of Corollary 8, it holds that*

$$\sup_{\beta \in B} \left| \frac{1}{n} \sum_{i=1}^n \Big\{ t(x_i, y_i, \beta) \left[I\big(r(x_i, y_i; \beta) \leq r_{[h_n]}(x_i, y_i; \beta)\big) - I\Big(r(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda)\Big) \right] \Big\} \right| = o_p(1)$$

and there exists $\varepsilon > 0$ *such that*

$$\sup_{\beta \in U(\beta^0,\varepsilon)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \Big\{ t(x_i, y_i, \beta) \left[I\big(r(x_i, y_i; \beta) \leq r_{[h_n]}(x_i, y_i; \beta)\big) - I\Big(r(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda)\Big) \right] \Big\} \right| = \mathcal{O}_p(1)$$

as $n \to +\infty$.

Proof: The corollary follows directly from the Chebyshev inequality for non-negative random variables, $P(X \geq K) \leq \mathsf{E}\, X/K$, since by Corollary 8

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

and

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

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

B Proof of consistency and convergence rate

Proof of Theorem 2: This is a standard proof of consistency based on the uniform law of large numbers and the convergence of the order statistics $r_{[h_n]}(x_i, y_i; \beta)$ to the corresponding quantile $G_\beta^{-1}(\lambda)$. Let us recall the GTE objective function $S_{nn}(\beta)$ and denote

$$S(\beta) \quad = \quad \mathsf{E}\left\{ s(x_i, y_i; \beta) \cdot I\Big(r(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda) \Big) \right\}.$$

By definition, $P\Big(S_{nn}\left(\hat{\beta}_n^{(GTE,h_n)}\right) < S_{nn}\left(\beta^0\right) \Big) = 1$. For any $\delta > 0$,

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

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

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

as $n \to +\infty$, that is, the consistency of $\hat{\beta}_n^{(GTE,h_n)}$ (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

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

Since the identification Assumption I2 implies

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

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 Corollary 9 and Lemma 4 for function $t(x_i, y_i; \beta) = s(x_i, y_i; \beta)$, see Assumptions D, F1, and F3, because

$$\begin{aligned} & S_{nn}(\beta) - S(\beta) \\ = \; & \frac{1}{n}\sum_{i=1}^{n}\left\{s(x_i, y_i; \beta)\left[I\Big(r(x_i, y_i; \beta) \le r_{[h_n]}(x_i, y_i; \beta)\Big) - I\Big(r(x_i, y_i; \beta) \le G_\beta^{-1}(\lambda)\Big)\right]\right\} \\ + \; & \frac{1}{n}\sum_{i=1}^{n}\left\{s(x_i, y_i; \beta)I\Big(r(x_i, y_i; \beta) \le G_\beta^{-1}(\lambda)\Big) - \mathsf{E}\left[s(x_i, y_i; \beta)I\Big(r(x_i, y_i; \beta) \le G_\beta^{-1}(\lambda)\Big)\right]\right\}. \end{aligned}$$

□

Proof of Theorem 3: We already know that $\hat\beta_n^{(GTE,h_n)}$ is consistent. Hence $P\Big(\Big\|\hat\beta_n^{(GTE,h_n)} - \beta^0\Big\| > \rho\Big) \to 0$ as $n \to \infty$ for any $\rho > 0$ (Theorem 2).

Further, we employ the almost sure second-order differentiability of $S_{nn}(\beta)$ and

$$S(\beta) = \mathsf{E}\left\{s(x_i, y_i; \beta)\cdot I\Big(r(x_i, y_i; \beta) \le G_\beta^{-1}(\lambda)\Big)\right\}$$

at β^0 (see Lemma 1 and Assumption F1). Since

$$S_{nn}(\beta) = \frac{1}{n}\sum_{i=1}^{n} s(x_i, y_i; \beta)\cdot\left[I\Big(r(x_i, y_i; \beta) \le r_{[h_n]}(x_i, y_i; \beta)\Big) - I\Big(r(x_i, y_i; \beta) \le G_\beta^{-1}(\lambda)\Big)\right] \tag{34}$$

$$+ \frac{1}{n}\sum_{i=1}^{n} s(x_i, y_i; \beta)\cdot I\Big(r(x_i, y_i; \beta) \le G_\beta^{-1}(\lambda)\Big), \tag{35}$$

Assumptions F, Lemma 4, and Corollary 9 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 1, $S'_{nn}(\beta) \to S'(\beta)$

and $S''_{nn}(\beta) \to S''(\beta)$ as $n \to \infty$ uniformly in $\beta \in U(\beta^0, \delta)$, whereby

$$S''(\beta^0) \;=\; \mathsf{E}\left\{s''(x_i, y_i; \beta^0) \cdot I\Big(r(x_i, y_i; \beta) \le G_\beta^{-1}(\lambda)\Big)\right\} = Q_s > 0$$

by Assumptions D2 and F3. Since Q_s is a positive definite matrix by Assumption F3, 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 $\hat{\beta}_n^{(GTE,h_n)}$, this implies that for any $\varepsilon > 0$ there is some $n_0 \in \mathbb{N}$ such that $\hat{\beta}_n^{(GTE,h_n)} \in U(\beta^0, \rho)$ and subsequently $\left\|S(\hat{\beta}_n^{(GTE,h_n)})\right\| \ge C\left\|\hat{\beta}_n^{(GTE,h_n)} - \beta^0\right\|$ for all $n > n_0$ with probability at least $1-\varepsilon$. Therefore, it is sufficient to show that $\sqrt{n}\left\|S'(\hat{\beta}_n^{(GTE,h_n)})\right\| = \mathcal{O}_p(1)$ to prove the theorem.

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

$$\begin{aligned}
&\sqrt{n}\,\mathsf{E}\left\{s'(x_i, y_i; \hat{\beta}_n^{(GTE,h_n)}) I\Big(r(x_i, y_i; \hat{\beta}_n^{(GTE,h_n)}) \le G_\beta^{-1}(\lambda)\Big)\right\} \\
\le\; & \sup_{\beta \in U(\beta^0,\rho)} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Bigg\{ -s'(x_i, y_i; \beta) I\Big(r(x_i, y_i; \beta) \le G_\beta^{-1}(\lambda)\Big) \qquad (36) \\
& \qquad\qquad + \mathsf{E}\left[s'(x_i, y_i; \beta) I\Big(r(x_i, y_i; \beta) \le G_\beta^{-1}(\lambda)\Big)\right] \Bigg\} \\
+\; & \sup_{\beta \in U(\beta^0,\rho)} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ s'(x_i, y_i; \beta) \left[I\Big(r(x_i, y_i; \beta) \le G_\beta^{-1}(\lambda)\Big) - I\big(r(x_i, y_i; \beta) \le r_{[h_n]}(x_i, y_i; \beta)\big) \right] \right\} \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (37)
\end{aligned}$$

(recall that $S'_{nn}(\hat{\beta}_n^{(GTE,h_n)}) = 0$ by Lemma 1). We only have to show that both terms are bounded in probability. This result for (37) is a consequence of Lemma 9 together with Assumptions F1, F3, and F4. The other part (36) can be bounded in probability by the following argument. Assumption F2 together with van der Vaart and Wellner (1996, Lemma 2.6.18) imply that

$$\mathcal{F}_{n,\delta} = \left\{s'(x, y; \beta) \cdot I\Big(r(x, y; \beta) \le G_\beta^{-1}(\lambda)\Big) : \beta \in U(\beta^0, \delta)\right\}$$

form a VC class of functions. Therefore, Assumptions D1 and F2 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 (36) is bounded in probability. □

References

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

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

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

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

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

[6] Bassett, G. W. (1991) Equivariant, monotonic, 50% breakdown estimators. *American Statistician* **45**, 135–137.

[7] Beňáček V., M. Jarolím & J. Á. Víšek (1998) Supply-side characteristics and the industrial structure of Czech foreign trade. In *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, pp. 51–68.

[8] Čížek, P. (2002) Robust estimation with discrete explanatory variables. In W. Härdle and B. Rönz (eds.) *COMPSTAT 2002 – Proceedings in Computational Statistics.* Berlin: Springer, pp. 509–514.

[9] Čížek, P. (2004a) Smoothed local L-estimation with an application, in Hubert M., Pison G., Struyf A. and Aelst S. (eds.), Theory and Applications of Recent Robust Methods, Birhaeuser Verlag, Basel, 59–70.

[10] Čížek, P. (2004b) Least trimmed squares in nonlinear regression under dependence, submitted.

[11] Čížek, P. & J. Á. Víšek (2000) Least trimmed squares. In W. Härdle, Z. Hlávka, and S. Klinke (eds.) *XploRe Application Guide*. Heidelberg: Springer, pp. 49–64.

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

[13] Dijk, D., T. Terasvirta & P. H. Franses (2000) Smooth transition autoregressive models—a survey of recent developments. *SSE/EFI Working paper series in Economics and Finance* 380.

[14] Gallant, A. R. & D. W. Nychka (1987) Semi-nonparametric maximum likelihood estimation. *Econometrica* **55(2)**, 363–390.

[15] Gerfin, M. (1996) Parametric and semi-parametric estimation of the binary response model of labour market participation. *Journal of Applied Econometrics* **11(3)**, 321–339.

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

[17] Hadi, A. & A. Luceno (1997): Maximum trimmed likelihood estimators: a unified approach, examples and algorithms. *Computational Statistics and Data Analysis* **25**, 251–272.

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

[19] Ichimura, H. (1993) Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* **58**, 71–120.

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

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

[22] Müller, C. H. & N. K. Neykov (2003) Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models. *Journal of Statistical Planning and Inference* **116**, 503–519.

[23] Neykov, N. M. & P. N. Neytchev (1990) A robust alternative of the maximum likelihood estimator. *Short communications of COMPSTAT, Dubrovnik 1990*, 99–100.

[24] Neykov, N., P. Filzmoser, R. Dimova & P. Neytchev (2004) Mixture of GLMs and the trimmed likelihood methodology. In J. Antoch (ed.) *COMPSTAT 2004 – Proceedings in Computational Statistics.* Heidelberg: Springer, pp. 1585–1592.

[25] Pakes, A. & D. Pollard (1989) Simulation and the asymptotics of optimization estimators. *Econometrica* **57(5)**, 1027–1057.

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

[27] Powell, J. L. (1986) Symmetrically trimmed least squares estimation for Tobit models. *Econometrica* **54**, 1435–1460.

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

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

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

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

[32] Rousseeuw, P. J. & K. van Driessen (1999) Computing LTS regression for large data sets. *Technical report, University of Antwerp*, submitted.

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

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

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

[36] Tableman, M. (1994) The asymptotics of the least trimmed absolute deviations (LTAD) estimator. *Statistics & Probability Letters* **19(4)**, 329–337.

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

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

[39] Vandev, D. L. & N. M. Neykov (1993) Robust maximum likelihood in the Gaussian case. In S. Morgenthaler, E. Ronchetti, and W. A. Stahel (eds.), *New directions in data analysis and robustness.* Basel: Birkhäuser Verlag, pp. 259–264.

[40] Vandev, D. L. & N. M. Neykov (1998): About regression estimators with high breakdown point. *Statistics* **32**, 111–129.

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

[42] Víšek, J. Á. (2002) The least weighted squares II. Consistency and asymptotic normality. *Bulletin of the Czech Econometric Society* **16(16)**, 1–28.

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

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

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