# NONPARAMETRIC EXTREME REGRESSION QUANTILES

Victor Chernozhukov*

Ver: December 15, 1998

**Abstract**

Let $(x_t, y_t)$ be a random vector in $\mathbb{R}^d \times \mathbb{R}$, and $y_t = g(x_t) + u_t$, $u_t \geq 0$, where $g$ is an unknown smooth function from $\mathbb{R}^d$ to $\mathbb{R}$, s.t. $g(x) = \inf\{z : P(y_t \geq z | x_t = x) > 0\}$. We construct the locally weighted polynomial extreme regression quantile estimator of $g$ and its derivatives. Under certain regularity conditions, the estimator is shown to be a consistent pointwise estimator; the rates of convergence are characterized under various settings; and the asymptotic and approximate distributions are derived. Furthermore, under similar global regularity conditions, the global estimator of $g$ and its derivatives is offered, and is shown to be a consistent estimator in the $L_q, q \in [1, \infty]$ norms. The rates of convergence in these norms are characterized. We then generalize the extreme quantile estimator to a large class of locally weighted polynomial estimators of $g$ and its derivatives and show that the asymptotic behavior of this class is essentially that of the locally weighted extreme quantiles.

The second part of the paper considers the sequential procedures based on the extreme regression quantile estimates, such as the general optimization estimators, where the preliminary estimate of $g$ or its derivatives is required. The conditions under which the preliminary estimation has no effect on either consistency or asymptotic distribution are given. This analysis is extended to the boundary dependent maximum likelihood estimation problem, generalizing the results in Smith (1994) to the case of a nonparametric boundary. This provides the estimates of nuisance parameters entering the asymptotic distribution. Certain truncated quantile regression and censored quantile regression estimators are discussed in which the generalized type-I sample selection settings require the preliminary estimate of the extreme quantile function. The sufficient conditions when the preliminary estimation has no effect on the asymptotic behavior of these estimators are established.

**Acknowledgment.** I am most grateful to Takeshi Amemiya, Roger Koenker, Thomas MaCurdy, and Stephen Portnoy for encouragement, support, and advice. I also thank Gregory Chow, Bo Honore, Kenneth Judd, Robert McMillan, Leonardo Rezende, Mark Watson and participants of econometric seminars at Stanford and Princeton for helpful discussions and questions. I am indebted to Sergei Morozov, Han Hong, Slavi Slavov for giving me extensive comments on the drafts of this paper. The author should be blamed for any errors and typeos encountered.

* Address email correspondence to: vchern@leland.stanford.edu. Address mail correspondence to: Victor Chernozhukov, Dept. of Economics, Stanford, CA 94305. Versions: April, June, October, December 1998. Note: Please do not circulate without consent of the author. Please notify me of any errors found.

# 1 Introduction

*Regression minima and maxima*

Conditional quantile estimation or quantile regression was first introduced by Koenker and Bassett in 1974, 1978, and since then it has emerged as a powerful method that has helped to answer many important economic questions. Examples include those in labor economics (providing wealth and income inequality measures), macroeconomic dynamics (providing robust checks of unit roots), political economy (in median voter models) and the analysis of business performance (see Buchinsky (1995), Powell (1986a),Koenker and Portnoy (1999)). Regression quantiles describe most of the relevant characteristics of the data since they are the natural analogues of order statistics in regression settings.

Currently several notable papers address different aspects of the issue: perhaps the best summary of all work to date is Koenker and Portnoy (1999). The literature has generally addressed the issue of nonextreme conditional $\tau$-th ($0 < \tau < 1$) quantile estimation. The case of $\tau$ equal zero or one, called the *extreme regression quantiles*, is introduced in Portnoy (1987), Bassett (1988), Smith (1994), Koenker and Portnoy (1996).

Intuitively, extreme regression quantiles can be thought of as the regression analogues of extreme order statistics in homogeneous samples. That is, extreme regression quantiles are models of the conditional extremes, such as *conditional minimum* or *maximum*.

To date, Smith has addressed the issues of parametric (linear) extreme quantile estimation, and although his results are not general, they generalize readily to a variety of cases. The important work that strengthened the results of Smith (1994) is by Portnoy and Jureckova (1998). Chernozhukov and Hong (1998a) study the asymptotic behavior of nonlinear extreme regression quantiles.

Our aim here is to investigate most relevant theoretical econometric questions concerning nonparametric extreme regression quantiles, namely the regression extremum in which no assumption about the functional form is made. In particular, we formulate nonparametric extreme quantile estimators. The results on consistency, asymptotic and approximate distributions, as well as rates of convergence are developed. This extends the analysis in Smith (1994) to the estimation of the unknown nonparametric extreme conditional quantile function and its derivatives.

Our motivation in pursuing this line of research is derived from many useful and profound economic and statistical models, which call for investigation of the conditional extremes. Some examples include the economic models of $(S(x), s(x))$ adjustment (such as capital stock adjustment problem, technology update problem, sticky pricing models, the cash-balance problem), minimum wage determination in models of job search, value-at-risk and production frontier estimation. The statistical applications include the use of extreme quantiles in the boundary dependent maximum likelihood estimation, estimation of the fixed sample selection equations in generalized type-I Tobit models and many others.

*Formulation and the Main Results*

Before proceeding, the following is a brief and less formal review of the definitions of the extreme and the nonextreme regression quantiles and the main results of this paper. In the discussion below most of the regularity conditions are omitted for brevity. Consider the model

$$y_t = g_\tau(x_t, \beta_0^\tau) + u_t,$$

where $g_\tau(x, \beta^\tau)$ is the conditional $\tau$-th quantile function, of $y_t$, the dependent variable, given the independent variable $x_t$, $\forall t$. It is parameterized by a finite dimensional parameter $\beta^\tau$, and $\beta_0^\tau$ is usually the unique parameter value that satisfies

$$g_\tau(x, \beta_0^\tau) = F_{y_t}^{-1}(\tau | x_t = x), \forall x, t$$

where $F_{y_t}$ is the distribution function of $y_t$ and $F_{y_t}^{-1}$ is its left-continuous inverse:

$$F_{y_t}^{-1}(v | x_t = x) \equiv \inf(z : F_{y_t}(z | x_t = x) \geq v)$$

The $\tau$-th$(0 < \tau < 1)$ quantile regression estimator of $\beta_0^\tau$ (and hence of $g(x, \beta_0^\tau)$ ) is usually defined as a solution, $\hat{\beta}_0^\tau$, to the optimization problem:

$$\hat{\beta}_0^\tau \in \underset{\beta}{\operatorname{argmin}} \sum_{t=1}^{T} \tau (y_t - g_\tau(x_t, \beta))^+ + (1 - \tau)(y_t - g_\tau(x_t, \beta))^- \tag{1.1}$$

where $\{y_t\}$ are the sample realizations of the dependent variable, and $\{x_t\}$ are the sample realizations of the covariates. The above estimator, under certain regularity conditions, gives the sample analog, $\hat{\beta}_0^\tau$, of the population conditional quantile parameter, $\beta_0^\tau$, which minimizes the infinite sample analog objective function:

$$\beta_0 = \operatorname{argmin}_\beta \lim T^{-1} \sum_{t=1}^{T} E\left[\tau (y_t - g_\tau(x_t, \beta))^+ + (1 - \tau)(y_t - g_\tau(x_t, \beta))^-\right]$$

To understand well (1.1) note that the case when $\tau = 1/2$ is the case of regression median; $\tau = 1/4$ is the case of lower regression quartile.

The above can be extended to the case of $\tau = 0$ or $1$ in the following way. Naturally, we would like to view the extreme quantiles as a limit of the nonextreme quantiles, and hence (for $\tau = 0$, w.l.o.g., and $\beta_0 \equiv \beta_0^0$):

$$g_0(x, \beta_0) = \lim_{\tau \searrow 0} g_\tau(x, \beta_0^\tau) = \lim_{\tau \searrow 0} F_{y_t}^{-1}(\tau | x_t = x) = \inf(z : F_{y_t}(z | x_t = x) > 0), \forall (x, t)$$

This justifies another name for the extreme quantiles in these settings: the *boundary of conditional support of* $y$, or, simply, the *conditional* or *regression boundary*.

Correspondingly, the sample extreme regression quantile estimator, $\hat{\beta}_0$, may be defined as follows[1]:

$$\hat{\beta}_0 \in \lim_{\tau \searrow 0} \underset{\beta}{\operatorname{argmin}} \sum_{t=1}^{T} \tau (y_t - g_\tau(x_t, \beta))^+ + (1 - \tau)(y_t - g_\tau(x_t, \beta))^- \tag{1.2}$$

The above estimator is equivalent to the following one (letting $g_0 \equiv g$ ):

$$\hat{\beta}_0 \in \underset{\beta}{\operatorname{argmax}} \sum_{t=1}^{T} g(x_t, \beta) \quad s.t. \quad g(x_t, \beta) \leq y_t \quad \forall t \tag{1.3}$$

The population, infinite sample, analog can be given by:

$$\lim_{\tau \searrow 0} \underset{\beta}{\operatorname{argmin}} \lim T^{-1} \sum_{t=1}^{T} E\left[\tau (y_t - g(x_t, \beta))^+ + (1 - \tau)(y_t - g(x_t, \beta))^-\right]$$

[1] Setting $\tau < C_T^p$ suffices, where $C_p^T$ is the number of distinct combinations of $p$ points drawn from sample of size $T$ with replacement, where $p = dim(\beta)$

Parameterizing the behavior of conditional quantiles as well-behaved parametric functions of $x$ and some unknown parameters leads to parametric analysis (see Portnoy and Jureckova (1998), Bassett (1988), Smith (1994), Koenker and Portnoy (1996), Chernozhukov and Hong (1998a)). The nonparametric analysis when the conditional quantile function is an unknown function of $x$ has been considered in Chaudhuri (1991), Welsh (1996) for the nonextreme cases. The present work addresses the *nonparametric* analysis of the extreme case.

Thus, the model under consideration here is

$$y_t = g(x_t) + u_t, u_t \geq 0, g(x) = \inf\{z : P(y_t > x | x_t = x) > 0\}, \forall t \tag{1.4}$$

where $g$ is an unknown smooth function. The task is to estimate $g$ and its derivatives at an arbitrary point, $x \in \mathbb{R}^d$.

The idea is to substitute the specification of $g$ in the above objective function, (1.2), by its approximation around $x$, based on the smoothness properties. One such choice is the (local) Taylor approximation:

$$g(x_t) = g(x) + \sum_m (m!)^{-1} D^m g(x)(x_t - x)^m + R_t, \text{ or}$$

$$g(x_t) = \beta_x' z_{x_t} + R_t, \text{ where } R_t = O(|x_t - x|^{k+\gamma}),$$

that is, the above is the polynomial expansion of $g$ up to some fixed order, $k$[2],

$$z_{x_t}' \equiv [1 \ (x_t - x)', [\text{vech}\,(x_t - x)(x_t - x)']', \ldots] \tag{1.5}$$

$\beta_x$ is the vector that contains $\{g(x), (m!)^{-1} D^m g(x), \text{ for } [m] = 1, \ldots k\}$, and $D^m$ is the differential operator (precise formalizations are given in the next section). The task, thus, is to estimate $\beta_x$ at any given point $x$, and hence recover $D^m g(x)$.

Since the above approximation is only good locally, a natural restriction of the objective function (1.2) to the neighborhoods of $x$ yields the objective:

$$\lim_{\tau \searrow 0} \underset{\beta_x}{\text{argmin}} \sum_{t=1}^{T} W_{Tt} \left( \left[ \tau \left( y_t - z_{x_t}' \beta_x \right)^+ + (1 - \tau) \left( y_t - z_{x_t}' \beta_x \right)^- \right] \right) \tag{1.6}$$

where $\{W_{Tt}\}$ are weights that are large for the observations $\{x_t\}$ that are near $x$, and that are small or zero for the observations that are far from $x$. Thus the defined estimator is the locally polynomial estimator.

In this paper, we concentrate the discussion on the case where the weights $\{W_{Tt}, t = 1, ..., T\}$ are the uniform kernel weights:

$$W_{Tt} = 1\left( x_t - x \in [-h_T/2, h_T/2\,]^d \right), \quad (h_T > 0) \tag{1.7}$$

where $[-h_T/2, h_T/2]^d$ is the $d$-dimensional closed box. $\{h_T\}$ is the sequence of positive bandwidth parameters approaching zero as $T \nearrow \infty$. The generalization to other weights (and objective functions) is simple and none of conclusions change for a wide class of general weights, as explained in section 7.

The motivation in defining the estimator in this way is based on the previous literature on a locally polynomial regression. The locally polynomial regression estimators have been proposed to tackle nonparametric estimation of conditional mean and conditional quantiles (nonextreme) (by Stone (1977), Cleveland

[2] In what follows, we introduce the notation vech and diag. For a vector $x \in \mathbb{R}^d$ diag$(x)$ is the diagonal matrix with $x_1, \ldots x_d$ as its diagonal matrix. For a square matrix $X$, diag $(X)$ is the vector that contains diagonal elements of this matrix. For any collection of elements $z$, operation vech $(z)$ simply stacks the elements of $z$ in a vector (in any consistent way).

(1979), Welsh (1996), Chaudhuri (1991)). In these kinds of problems, estimators seem to enjoy optimality in terms of the rate of convergence, as established by Stone, good minimax properties, as explained by Fan (1993), and also tend to have good performance near the boundaries of support of covariates, compared to other estimators (see discussion in Cheng, Fan, and Marron (1997), Wand and Jones (1995)).

In a similar manner, an estimator of the conditional boundaries (conditional extreme population quantiles) is constructed here. In essence, the idea involves considering a locally polynomial specification of conditional quantiles and setting $\tau$ sufficiently small.[3] Of course, the optimality properties of the estimator generally depend on the regularity conditions at hand. In general, optimality does not hold; yet for the class of problems, where the extreme local order statistics are efficient estimators either in terms of rates or relative efficiency, the optimality of the pointwise or global estimates holds. These issues are addressed elsewhere (Chernozhukov 1999b).

The contribution of this work and its organization is as follows. We develop a weighted locally polynomial of the conditional boundary and, importantly, its $m$-th partial derivatives up to a certain fixed order. This estimator is formulated within the framework of regression quantiles, it is highly tractable and easily computable by the available regression quantile routines. The conditional boundary and its derivatives are estimated simultaneously. Conditional boundary estimation has direct relation with the edge and support set estimation literature and thus by itself the problem of boundary or edge estimation may not be new (the recent statistical literature witnessed an outburst of edge estimation research, culminating in such works as Mammen and Tsybakov (1995), Härdle, Park, and Tsybakov (1995), Cuevas and Fraiman (1998), and others). Our procedure offers a tractable and simple estimator of not only the boundaries but the derivatives as well. The need for derivative estimation is paramount in econometrics where one is often interested in studying the marginal effects and not the boundary itself. Moreover, the settings in which our estimator of the boundary and its derivatives is shown to be consistent and has well behaved properties are substantially more general than those in the sources cited. Also, we derive the asymptotic and approximate distributions of our estimator. This result is not available for edge estimators in the sources cited.

The foundation and a main proposition of our analysis is in the inequality that describes the behavior of extreme regression quantiles in terms of local extreme order statistics, dimension of covariates, order of derivative estimated, smoothness of the underlying function, and a bandwidth parameter (section 3.2). The settings in which this inequality applies are ubiquitous and highly general, allowing for general sampling assumptions, including dependence, nonstationarity, and any other non-i.i.d cases, provided that for a given $x$: (i) sufficient accumulation of realizations $\{x_t\}$ in small neighborhoods of $x$ occurs; (ii) the local mean, over such $\{x_t\}$, stays strictly in the interior of the neighborhoods and is not far away from the asymptotic local mean. This is true e.g. when $\{x_t\}$ are sampled from the strictly positive well behaved densities and a LLN holds.

The developed inequality is central to most results in this work in that it allows to link the behavior of extreme quantiles to that of the local extreme order statistics, thus allowing us to apply the immense apparatus of theoretical results on the asymptotic behavior of extrema that has been developed by Gnedenko (1943), Smirnov, Galombos (1978), de Haan (1970), Leadbetter et al. (1983), Resnick (1987), and many others. Relying on the asymptotic theory of extreme order statistics is the distinctive feature of our analysis.

[3]e.g., any $\tau < 1/C_T^p$, where $C_T^p$ is the number of combinations of p-points drawn out of T points, suffices. (Here $p$ equals the dimension of the polynomial that is locally fit, i.e., $p$ is the number of regressors (and their powers), including the constant term, that are used in the regression, and $T$ is the sample size.)

Sections 3.3 and 4 investigate the pointwise consistency and convergence rates under the assumption of independence on the errors (general time series processes are considered in Chernozhukov (1999b)). The treated models are:

1. the model of regular variation, or, equivalently, the model of asymptotic non-degeneracy of local extreme order statistics,
2. the Weibull tail equivalence model, when the conditional distribution functions behave as Weibull distributions near the boundary, as a useful subclass of (1),
3. certain generalizations of (1) and (2), allowing for characterizations of convergence rates in many cases that are not in (1) or (2), yet providing cruder bounds than in (1) or (2),
4. the model of degeneracy of the local extreme order statistics when a positive probability mass is on the boundary.

These models span essentially the entire relevant range of behavior of extremes and, consequently, that of the estimates of the conditional boundary and its derivatives. All the earlier works deal with the special cases in (2) only. On the other hand, the results obtained in the present work also allow for general sampling assumptions on $\{x_t\}$ as well as for data-dependent bandwidth sequences.

In section 5, we derive the joint asymptotic distribution of the boundary estimates and its derivatives in model (2), by studying the convergence of density. This is the first such result in the literature. Our results rely on the works of Smith (1994) and Bassett (1988) that allowed to characterize the distribution of extreme quantiles in the case of the parametric linear model.

We develop the global estimator of conditional extreme quantiles and its derivatives by interpolation procedures in section 6. Rates of convergence of the estimator, in $L_q, q \in [1, \infty]$ norms, are given for the models (1)-(4). These results are required for studying the asymptotic behavior of sequential estimators where the extreme quantiles are estimated in the first stage. In section 7, we generalize the extreme quantile estimator to a large class of locally weighted polynomial estimators of $g$ and its derivatives and show that the asymptotic behavior of this class is essentially that of the locally weighted extreme quantiles.

The second part of this paper considers exactly this issue. Given any two-stage general extremum estimator, we study the conditions on the primitives that eliminate the second-stage effect of the first-stage extreme quantile estimation. The analysis is then extended to the boundary dependent ML estimation of Smith (1994), Smith (1985) to the case of the unknown smooth boundary (section 8.1). In section 8.3, the models of truncation and censoring are discussed where the truncating and censoring rules (sample selection equations) are the fixed unknown functions of covariates.

## 2 Construction of the Local Extreme Regression Quantiles

In this section precise definitions underlying the earlier discussion are given. Let $m = (m_1, \ldots, m_d)$, $m \in \mathbb{Z}_+^d$, the set of $d$-dimensional nonnegative integers. Define $D^m$ to be the differential operator

$$\frac{\partial^{[m]}}{\partial x_1^{m_1} \ldots \partial x_d^{m_d}}, \quad \text{and } [m] = \sum_{i=1}^{d} m_i$$

Construct the estimator of $g$ and its derivatives as follows:

1. Construct the following polynomial, in $x_t \in \mathbb{R}^d$, for all $t \in \{1, ...T\}$:

$$\mathsf{P}_{x_t}(\beta_x, x) = \sum_{m \in \mathbb{Z}_+^d : [m] \leq k} \beta_{xm} (x_t - x)^m$$

where

$$(\bar{x} - x)^m = \prod_{i=1}^{d} (\bar{x}_i - x_i)^{m_i}, \text{ where } x_i \text{ denotes the } i\text{-th sub-element of vector } x,$$

2. Rewrite $\mathsf{P}_{x_t}(\beta_x, x)$ as $\beta_x' z_{x_t}$, where $\beta_x$ is the vector that contains all the coefficients $\beta_{xm}$ (indexed by $m \in \mathbb{Z}_+^d : [m] \leq k$), and $z_{x_t}$ is the vector that contains all elements of set $\{(x_t - x)^m, m \in \mathbb{Z}_+^d : [m] \leq k\}$,

3. Solve the optimization problem (1.6), to obtain $\hat{\beta}_x$,

4. Construct the estimates of derivatives as follows:

$$\widehat{D^m g(x)} = m! \widehat{\beta}_{xm}$$

where $m! \equiv \prod_{i=1}^{d} (m_i!)$, (with the convention $0! = 1$).

Note that because $\{W_{Tt}\}$ are uniform weights over the box neighborhood, the problem in (1.6) reduces to finding an optimum of the following problem:

$$\max_{\beta_x} \sum_{t \in N_{Tx}} \beta_x' z_{x_t}, \quad \text{s.t. } \beta_x' z_{x_t} \leq y_t, \quad \forall t : t \in N_{Tx} \tag{2.1}$$

where $N_{Tx} = \{t : W_{Tt} > 0\}$. Results for general weights do not differ under reasonable conditions on the weights, hence a brief treatment is confined to section 7, which also contains other generalizations.

# 3 Asymptotic Properties: Local Extreme Order Statistics and Rates of Convergence

## 3.1 Main Assumptions and Preliminaries

**Assumption 1** *For $g : \mathbb{R}^d \longrightarrow \mathbb{R}$, fixed $k \in \mathbb{Z}_+$, $C \in \mathbb{R}_{++}$, $\gamma \in (0, 1]$, $m \in \mathbb{Z}^d$*

1. *$D^m g(x)$ is bounded and continuous in $\mathcal{C}$, a compact set in $\mathbb{R}^d$, with minimally smooth boundary, $\forall m : 0 \leq [m] \leq k$.*

2. *k-th order partial derivatives of $g$ are uniformly Holder continuous on $\mathcal{C}$, with some power $\gamma$:*

$$\left| D^u g(x') - D^u g(x'') \right| \leq C \left| x' - x'' \right|^\gamma, \quad \gamma \in (0, 1], \quad \forall u : [u] = k, \quad \forall x', x'' \in \mathcal{C}$$

Let $r = k + \gamma$. Let us refer to $r$ as the *smoothness order* of $g$ on $\mathcal{C}$. Let $\mathcal{G} = \{g : \text{A1 holds for fixed } C, \gamma\}$. Consider probability spaces $(\Omega, \mathcal{F}, P_g)$, where index $g$ denotes possible dependence on $g$. Of course, $\{x_t, y_t\} \in m_{\mathcal{F}}$. In what follows, $P \equiv \sup_{g \in \mathcal{G}} P_g$, $P_* \equiv \inf_{g \in \mathcal{G}} P_g$, that is for any $A \in \mathcal{F}$, $P(A) = \sup_{g \in \mathcal{G}} P_g(A)$, $P_*(A) = \inf_{g \in \mathcal{G}} P_g(A)$.[4]

[4] We say here that $\nu_T \in m_{\mathcal{F}}$ converges to $\nu_T' \in m_{\mathcal{F}}$ in probability, if $\lim_T P(|\nu_T - \nu_T'| > \epsilon) = 0$, and $\nu_T$ converges to $\nu_T'$, a.s. if $P(\limsup\{|\nu_T - \nu_T'| > \epsilon\}) = 0, \forall \epsilon > 0$. Also, $\nu_T = O_p(\nu_T')$, if for any $\epsilon_\epsilon > 0, \exists \delta > 0$: $\limsup_T P(|\nu_T| > \delta|\nu_T'|) < \epsilon$, $\nu_T = o_p(\nu_T')$ if $\nu_T/\nu_T' \to 0$ in probability, $\nu_T = o_{a.s.}(\nu_T')$, if $\nu_T/\nu_T' \to 0$ a.s., and $\nu_T \sim_p \nu_T'$, if $\nu_T = O_p(\nu_T')$ and $\nu_T' = O_p(\nu_T)$.

**Assumption 2** *$\forall\varepsilon > 0$ fixed, $\inf_{x\in X, g\in\mathcal{G}} P_g\left(y_t - g(x) < \varepsilon|x\right) > \delta_\varepsilon > 0$, $\forall t$, $X$ is a subset of $\mathbb{R}^d$ that contains $C$. (in addition to $y_t \geq g(x_t)$, given $x_t$)*

**Assumption 3** *$\{h_T\}: h_T \in m_{\mathcal{F}}, h_T \searrow 0$, $Th_T^d T^{-a'} \to \infty$, $Th_T^d T^{-a} \to 0$, for some $a', a > 0$, as $T \to \infty$, either*
*(3a) in probability, or,*
*(3b) almost surely.*

The following is a general sampling assumption on $\{x_t\}$. Let for $\epsilon$ fixed and small:

$$A(x,T,h_T) \equiv \left\{t : 1 \leq t \leq T \text{ and } \left|x_t - x\right| \leq h_T/2\right\}$$

$$N(x,T,h_T) \equiv \text{card}\left(A(x,T,h_T)\right)$$

$$\bar{x}(x,T,h_T,u) \equiv N(x,T,h_T)^{-1} \sum_{t\in A(x,T,h_T)} (x_t - x)^u, \text{ for } u \in \mathbb{Z}^d$$

$$\mathcal{S}_{1T} \equiv \left\{w \in \Omega : N(x,T,h_T) \geq K_1 T h_T^d, \quad \forall x \in C\right\},$$

$$\mathcal{S}_{2T} \equiv \left\{w \in \Omega : h_T^{-[u]}\left|\bar{x}(x,T,h_T,u)\right| \leq K_2 2^{-[u]}, \quad \forall x \in C, u : 0 < [u] \leq k\right\}$$

**Assumption 4** *For any sequence $\{h_T\}$ satisfying 3a, 3b, there are $K_1 > 0, 0 < K_2 < 1$, s.t.*
*(4a) $\lim_T P_*\left(\mathcal{S}_{1T} \cap \mathcal{S}_{2T}\right) = 1$ under assumption 3a, or,*
*(4b) $P_*\left(\mathcal{S}_{1T} \cap \mathcal{S}_{2T}, ev.\right) = 1$ under assumption 3b.*[5]

REMARKS:

(1) Assumption 1 is the smoothness assumption. It is typically made in the nonparametric regression literature (see Stone (1982) and others). It covers not only smooth functions, but also non-differentiable functions that are Holder continuous ( e.g., any continuous piecewise smooth function is Holder continuous).

(2) Assumption 2 is the identifiability condition that simply says that the unknown function $g$ is the boundary of support of $y_t$, given $x_t = x$.

(3) Assumption 4a is the sampling assumption which requires that there should be enough observations in any shrinking box centered around any point in a relevant domain and in addition, that the observations (on $\{x_t\}$ in any of these shrinking neighborhoods be more or less evenly distributed so that the sample mean, based on the observations in these neighborhoods stays away from their boundaries. This assumption (in the form that $\lim_T P_*(\mathcal{S}_{1T}) = 1$) is, in a sense, a classical assumption, as suggested by Stone (1982). All other conditions used in the nonparametric literature imply this one (see the sources cited below ). The additional requirement that $\lim_T P_*(\mathcal{S}_{2T}) = 1$ is somewhat artificial in the sense that it is hard to envision a situation where event $\mathcal{S}_{1T}$ would not imply $\mathcal{S}_{2T}$. However, it seems that indeed this may not hold. The assumption allows $x_t$ to be dependent and non-stationary as long as the above conditions are satisfied.

(4) Assumption 4b is somewhat stronger than 4a in that it requires events $\mathcal{S}_{iT}, i = 1,2$ to hold eventually, a.s. Assumption 4b implies 4a. Assumption 4a is required for establishing the asymptotic results for convergence and bounds in probability and assumption 4b is needed for convergence a.s.

[5] $\{A, ev.\} \equiv \{\liminf A\}$

(5) Assumptions 3a and 3b allow for data dependent bandwidth sequences. These are typical of the non-parametric literature. Assumption 3b implies 3a but assumption 3a is required for establishing the asymptotic results for convergence and bounds in probability, whereas assumption 3b – for convergence and bounds a.s..

To demonstrate the simple ( yet general case ) that fall under the above sampling assumptions, consider

**Assumption 5** *Define $F_T(\chi) \equiv T^{-1}\sum_t 1(x_t < \chi)$ and $F_x$ to be a distribution function that has uniformly continuous and strictly positive density[6], $w_x$*; $\{X_t\}$ are i.i.d. $F_X$.

**Lemma 1** [7] *Assumptions 5 and 3b imply assumption 4b, while 5 and 3a imply 4a.*

## 3.2 Relation to the Local Extreme Order Statistics

In this section, a relation with the local (minimum) order statistics is made to obtain a rather general characterization of the rates of convergence of the estimator proposed here. Once this representation is obtained, a reference may be made to the rich literature on the asymptotic and finite sample behavior of extreme order statistics in order to infer the behavior of the estimator considered here. In particular, the results on the rates of convergence of extreme order statistics can be used to derive the rates of convergence of the locally polynomial extreme regression quantile. Some rather general results along these lines are considered in the next section. The result given below may be used to characterize the rates of convergence and other properties in all conceivable cases.

Let

$$\tilde{N}_h(x) \equiv [x - (h/2)\mathbf{1}; x + (h/2)\mathbf{1}] \subset \mathbb{R}^d,$$

$$N_h(x) \equiv \{z_{\tilde{x}} : \tilde{x} \in \tilde{N}_h(x)\} \subset \mathbb{R}^p, \quad \text{where} \quad p \equiv dim(z_x).$$

The following construction is valid w.p. $\to 1$ or ev. a.s. under assumptions 3a, 4a and 3b, 4b, correspondingly:

split $N_h(x)$ in $p+1$ subsets: $\{N_h^1(x), ..., N_h^{p+1}(x)\}$ s.t.

(1) $N_h^j(x) \cap N_h^i(x) = \varnothing$ if $p \geq 2$, $\forall i, j \leq p+1 : i \neq j$, and $N_h^j(x)$ are closed for $\forall i \leq p$,

(2) For the projection mapping $\pi : \mathbb{R}^p \to \mathbb{R}^{p-1}$ defined by $\pi x = (x_2, ..., x_p)$, $\Lambda(\pi N_h^j(x)) \propto \Lambda(\pi N_h^i(x)), \forall h > 0$, $\forall i, j \leq p+1$, where $\Lambda$ is the Lebesgue measure on $\mathbb{R}^{p-1}$,

(3) For $\mathcal{X}_T \equiv \text{vech}\{h_T^{-[m]} : 0 \leq [m] \leq k\}$ so that $z'_{x_t}\text{diag}\{\mathcal{X}_T\} = O(\mathbf{1}'), \forall T > T_o$, $\bar{u}(x, h_T, T) \equiv \text{card}(N_{Tx})^{-1}\sum_{t \in N_{Tx}} \text{diag}\{\mathcal{X}_T\} z_{x_t} = \text{diag}\{\mathcal{X}_T\}\tilde{z}\lambda_{\tilde{z}}$, uniformly in $\tilde{z} \in \times_{j=1}^p N_{h_T}^j(x)$, where $\lambda_{\tilde{z}} : \lambda'_{\tilde{z}}\mathbf{1} = 1, \lambda_{\tilde{z}} > \epsilon\mathbf{1}$, for an appropriate unit vector $\mathbf{1}$, and fixed $\epsilon \in (0,1)$ ,

(4) for $\forall$ matrices $\tilde{z} \in \times_{j=1}^p N_{h_T}^j(x)$, $\delta_T$, the smallest characteristic root of the matrix $\text{diag}\{\mathcal{X}_T\}\tilde{z}$, is bounded away from zero, $\forall T > T_o$,

[6] Density, here, is the Radon-Nyckodym derivative of $F_T$ with respect to the Lebesgue measure. Uniformly positive means that $w_x > 0$ on $X$ uniformly in $g \in \mathcal{G}, x \in X$. Uniform continuity means here that $\sup_{x \in X, g \in \mathcal{G}} \limsup_{x' \to x} |w_x(x) - w_x(x')| = 0$.

[7] Apart from Lemmas 1-4, other Lemmas and all proofs appear in the appendix

The $N_h(x)$ is a subspace of $\mathbb{R}^p$ that linearly spans $\mathbb{R}^p$, implying that this construction is valid (w.p. $\to 1$ or ev. a.s.). The appendix B explains this and shows explicitly the construction for the cases when $p = 1$ and $p = 2$.

Consider $p$ *local first order statistics*, $\{U_{(1)jh_T}(x), \quad j \in \{1, ...p\}\}$ :

$$U_{(1)jh_T}(x) \equiv \min\{u_t : z_t \in N^j_{h_T}(x)\} \tag{3.1}$$

By construction of $\{N^j_{h_T}(x)\}$, these are well-defined:

$$U_{(1)jh_T}(x) \neq \varnothing, \forall x \in C,$$

under assumptions 4a, 3a, w.p. $\to 1$, and under assumptions 4b, 3b, eventually a.s. Hence, in the discussion below, ignore either finitely many $T$ or all events s.t. the local extreme order statistics above are not defined. We relate now the behavior of estimates $\widehat{D^m g(x)}$ with that of $\{U_{(1)jh_T}(x), \quad j \in \{1, ...p\}\}$ .

**Theorem 1** RELATION TO THE LOCAL EXTREME ORDER STATISTICS. *There are fixed independent of $T$ and $x$ constants $K_4, K_5 > 0$ s.t. the inequality*

$$\left|\widehat{D^m g(x)} - D^m g(x)\right| \leq K_4 h_T^{-[m]} \left(\max_j U_{(1)jh_T}(x) + K_5 h_T^r\right), \forall x \in C \tag{3.2}$$

*holds w.p. $\to 1$ with assumptions 1, 2, 4a, 3a, and, eventually, a.s with assumptions 1, 2, 4b, 3b .*

REMARKS:

(1) The above establishes a link needed to find the pointwise and global rates of convergence. This is a fairly general result, reducing the study of extreme regression quantiles to that of the local extreme order statistics when it concerns the rates of convergence. The result permits substantial generality, including general forms of time dependence of $\{y_t, x_t\}$ (see Robinson (1997), Andrews (1995) for time series settings in which the main assumptions hold).

(2) The following intuition is in place here: (a) $K_5 h^r$ is the "cost" (in measuring $D^m g$) of using approximation rather than parametric function, (b) $\max_j U_{(1)jh_T}(x)$ is the cost of randomness, and (c) $K_4 h_T^{-[m]}$ is the cost multiplier arising when marginal effects are measured; it equals 1, when $[m] = 0$. Thus lower $m$ (marginal effects) and higher $r$ (smoothness of $g$) increase the rate of convergence. The effects of bandwidth choice and dimensionality, $d$, are not transparent here, as the effect is realized through $\max_j U_{(1)jh_T}(x)$ (see later section).

(3) The subsequent section gives some results of different generality regarding the rates of convergence when $\{y_t\}$ are independent. In all of the analysis to follow the tail behavior of the error distribution functions is to be parameterized to infer the rates of convergence. This is not very restrictive since it is possible to properly estimate the appropriate normalizing constants (and hence the rates of convergence) .

## 3.3 Pointwise Consistency

In this section a general result on pointwise consistency is presented. The assumptions are mild and the result may be useful for some applications. This result, due to its generality, does not allow to characterize the rates of convergence. We further need an additional assumption here.

**Assumption 6** *$\{u_t\}$ are independent.*

Generalizations to the dependent sequence of errors are considered in Chernozhukov (1999b).

**Lemma 2** *Assume 2 - 4, 6. Then in pr., under assumptions 3a, 4a, and a.s. under 3b, 4b.*

$$\max_j U_{(1)jh_T}(x) = o(1)$$

Theorem 1 and this Lemma yield:

**Theorem 2** POINTWISE CONSISTENCY. *Assume 1 - 4, 6 and that $\{h_T\}$ in addition satisfies $h_T^{-[m]} U_{(1)jh_T}(x) = o(1)$ in pr. (or a.s.). Then*

$$\left|\widehat{D^m g(x)} - D^m g(x)\right| = o(1)$$

*in probability under assumptions 4a, 3a, and a.s. with assumptions 4b, 3b.*

REMARKS:

(1) For $m = 0$, $D^m g(x) = g(x)$, the estimator is always consistent under the assumptions stated.

(2) All of the general settings to be discussed in subsequent section allow for the sequences satisfying assumption 3 and yielding at the same time $h_T^{-[m]} U_{(1)jh_T}(x) = o(1)$ in pr. or a.s. Thus the consistency of estimators of all partial derivatives $\widehat{D^m g(x)}$ is a generic result.

# 4 On the Rates of Convergence

This is the section where the pointwise *uniform* rates of convergence of the proposed estimator are studied. Uniformity here is in the sense that bounds are defined in relation to $P$, defined in section 2 as $P(A) \equiv \sup_{g \in \mathcal{G}} P_g(A)$, for any $A \in \mathcal{F}$. The reader may recall that we say $\nu_T = O_p(\nu_T')$, if $\forall \epsilon > 0\ \exists \delta > 0$, s.t. $\limsup_T \sup_{g \in \mathcal{G}} P_g(|\nu_T| > \delta |\nu_T') \leq \epsilon$. The first subsection addresses general issues (model 1). The second subsection considers useful special cases (model 2). The third and fourth subsections deal with the cases that are not considered in the first two (models 3 and 4). In particular, model 3 is a simple generalization of 1 and 2, yet it does not necessarily provide sharp bounds as in case of 1 and 2.

## 4.1 Some General Results on the Rates of Convergence

The developments here are based on the asymptotic distributions of extrema. These paralleled the normal Central Limit Theorems for averages and have been extensively developed in this century by Gnedenko (1943), de Haan (1970), Smirnov, Von-Misses, and others. There is a large number of general references available on the topic: see, for example, Reiss (1989), Galombos (1978), (1984), Leadbetter, Lindgren, and Rootzén (1983), and others. Here we focus on the (local) distributions that imply that the (local) extreme order statistics belong to the domain of attraction of a max-stable distribution. Such a distribution, if non-degenerate, in the settings described here, can only belong to the family of Weibull distributions. Consider the following Lemma due to Gnedenko (1943), which is conveniently restated here with the extra index $x$ denoting the dependence on $x$ of the conditional distribution.

**Lemma 3** *Let $\{u_1, ..., u_{\mathcal{N}}\}$, $u_t > 0$, $\forall t$, be i.i.d. with regularly varying law (distribution function) $F(\cdot|x)$. If there exists a sequence of constants $\{\gamma_{\mathcal{N}}(x)\}$ s.t. $\gamma_{\mathcal{N}}(x)U_{(1)|x} \Rightarrow \mathcal{W}$, where $\mathcal{W}$ is nondegenerate random variable, $\Rightarrow$ denotes weak convergence, and $U_{(1)|x} \equiv min\{u_1, ..., u_{\mathcal{N}}\}$, then $\mathcal{W}$ is a random variable with a Weibull distribution function:*

$$W(s) = 1(s > 0)\left(1 - \exp\{-s^{\alpha(x)}\}\right), \quad \alpha(x) > 0.$$

*Furthermore, the above weak convergence occurs if and only if $F(\cdot|x)$ is regularly varying, i.e., it satisfies for $\forall \varsigma \in \mathbb{R}_{++}$:*

$$\lim_{z \searrow 0} \frac{F(\varsigma z|x)}{F(z|x)} = \varsigma^{\alpha(x)} \tag{4.1}$$

*And, furthermore, a choice of $\{\gamma_{\mathcal{N}}(x)\}$ is given by $\gamma_{\mathcal{N}} = \left(F^{-1}(\mathcal{N}^{-1}|x)\right)^{-1}$.*

If $U_{(1)|x}$, which we term here as the *conditional order statistic*, appropriately normalized, is in the domain of attraction of a Weibull random variable, then write $U_{(1)|x} \in \mathcal{D}\left(\mathcal{W}_{\alpha(x)}\right)$. It is easy to check that if $F(\cdot|x)$ is regularly varying in $u$ with power $\alpha(x)$ in the above sense, then $F(\cdot|x)$ can be written as:

$$F(u|x) = u^{\alpha(x)}C(u, x), u > 0, \alpha(x) > 0, C(\cdot, \cdot) > 0. \tag{4.2}$$

where $C(u, x)$ is regularly varying with power 0 (slowly varying). It is also convenient to define for $\mho > 0$,

$$\tilde{\gamma}_{\mathcal{N}}(x, \mho) \equiv \left(\text{esssup}_{x' \in \tilde{N}_{\mho}(x)} F^{-1}\left(\mathcal{N}^{-1}|x'\right)\right)^{-1}.$$

In the model to be introduced we require that:

$$\limsup_{y \to 0+} \sup_{x \in \mathcal{C}, g \in \mathcal{G}} \left|F(F^{-1}(y|x)|x) - y\right| = 0 \tag{4.3}$$

Note that since $F^{-1}$ is just left-continuous inverse of $F$, generally $F(F^{-1}(y|x)|x) \neq y$. The above is not a restriction if, of course, $F$ is continuous and strictly monotone. Furthermore, it is not difficult to show that for any fixed $g$ the above is not a restriction given other assumptions, stated next (see discussion in Resnick (1987), p.15 ).

**Assumption 7** (MODEL 1) *$P_g\left(u_t < \varsigma | x_t = x'\right) = F_g\left(\varsigma|x'\right), \forall t$, and $F_g(\cdot|x')$ satisfies (4.3) and (4.2), uniformly in $x' \in \mathcal{C}, g \in \mathcal{G}$ in the sense that the condition on $C(u, x)$ holds unifromly and $\sup_{(x,g) \in (\mathcal{C}, \mathcal{G})} \alpha(x) \leq \bar{\alpha}$, $\inf_{(x,g) \in (\mathcal{C}, \mathcal{G})} \alpha(x) \geq \underline{\alpha}$, $(\bar{\alpha}, \underline{\alpha}) > 0$.*

We suppress the possible dependence of $F$ on $g$ for brevity of notation.

**Lemma 4** *For any $x \in \mathcal{C}$, given assumptions 3, 4, 6, 7,*

$$\max_j U_{(1)jh_T}(x) = O_p\left(\left(\tilde{\gamma}_{Th_T^d}(x, h_T)\right)^{-1}\right).$$

**Theorem 3** *With assumptions 1, 3, 4, 6, 7, for any $x \in \mathcal{C}$,*

$$\left|\widehat{D^m g(x)} - D^m g(x)\right| = O_p\left(h_T^{-[m]}\left(\tilde{\gamma}_{Th_T^d}(x, h_T)^{-1} + h_T^r\right)\right). \tag{4.4}$$

REMARKS:

(1) Having obtained this result, one can characterize the rates of convergence and find the optimal sequence of $\{h_T(x)\}$ if it exists. Clearly, the above result shows that should this sequence exist, it generally depends on $x \in C$. This is quite untypical for the nonparametric literature on conditional mean and median modeling where bandwidth sequences are, except for normalization constants, the functions of the sample size only. (One exception, due to Hengartner and Linton (1996), offers cases of non-standard sampling of $x_t$).

(2) If $\tilde{\gamma}_{Th_T^d}(x, h_T)$ is increasing in $h_T$[8], then optimal sequences of $\{h_T(x)\}$ maximizing the rate of convergence, can be determined up to the normalization constants, $K_6 \sim_p 1$, by solving for $h_T, \forall T$, the equation:

$$\left(\tilde{\gamma}^{-1}_{[Th_T^d]}(x, h_T)\right) = (K_6 h_T)^r \tag{4.5}$$

Clearly, any approximate solution works and the $[\cdot]$ above can be replaced by $(\cdot)$. Then, for that sequence $\{h_T^*(x)\}$,

$$\left|\widehat{D^m g(x)} - D^m g(x)\right| = O_p\left(\left(\tilde{\gamma}_{Th_T^{*d}}(x, h_T^*(x))^{\frac{[m]-r}{r}}\right)\right) = O_p\left(h_T^*(x)^{r-[m]}\right)$$

In practice, if $F$ is unknown, one needs to estimate $F$ in order to construct $\{h_T^*\}$. These issues are further addressed in Chernozhukov (1999c). The next section offers the cases that are more transparent in that an easy characterization of tails is employed, which enables us to find optimal or near-optimal sequences $\{h_T\}$.

## 4.2 Special Cases and Practicalities: Rates of Convergence Under Weibull Tail Parameterization

Two points are addressed here. First, we would like to consider some simplifications that yield us the rates of convergence that do not depend on any particular point $x \in C$. That should increase the practicality of the methods offered as far as their asymptotic characterization is concerned. Second, we would like to consider the case of the Weibull tail-equivalent distribution functions as a useful and simple class (see Fig.1). As argued, this class approximates well many regularly varying distribution functions. Restriction to that class involves somewhat more stringent tail parameterization than that considered so far. It has been demonstrated as quite useful in applied analysis (see Smith (1994)) and may be used for all practical purposes, unless there is a compelling evidence to deviate from such model.

### 4.2.1 Rates of Convergence: Specialization and Bounds

Here we further specialize the obtained results to the cases when the rates of convergence do not depend on $x$ in $C$.

**Assumption 8** *In addition to assumption 7, uniformly in $x, x' \in C, g \in C$, for $\forall \{T, h_T\}$ satisfying assumption* 3

$$\frac{\gamma_{Th_T^d}(x)}{\gamma_{Th_T^d}(x')} \to C_{gxx'} \text{ in pr}, C_{gxx'} \in [K_7, K_8], K_7, K_8 \in \mathbb{R}$$

[8] more generally, the optimal $\{h_T^*\}$ is found by minimizing $h_T^{-[m]}\left(\tilde{\gamma}_{Th_T^d}(x, h_T)^{-1} + h_T^r\right)$ s.t. the constraints in assumption 3

Since this implies that the rates of convergence are invariant with respect to the location of $x$ in $\mathcal{C}$, it is useful to economize and write $\gamma_{Th_T^d}$ instead of $\gamma_{Th_T^d}(x)$. This gives a convenient corollary:

**Corollary 1** *With assumptions 1 - 4, 6, 8, for any $x \in \mathcal{C}$,*

$$\left|\widehat{D^m g(x)} - D^m g(x)\right| = O_p\left(h_T^{-[m]}\left(\bar{\gamma}_{Th_T^d}^{-1} + h_T^r\right)\right)$$

*then for an optimal $\{h_T\}$ : $F^{-1}\left((Th_T^d)^{-1}\right) \sim_p h_T^r$,*

$$\left|\widehat{D^m g(x)} - D^m g(x)\right| = O_p\left(h_T^{r-m}\right) = O_p\left(\left(\bar{\gamma}_{Th_T^d}\right)^{([m]-r)/r)}\right)$$

This simplification also allows to construct the bounds on the rates of convergence that are more transparent then the ones obtained so far. The following corollary uses Lemma 6.

**Corollary 2** *With assumptions 1-4, 6, 8 for any $x \in \mathcal{C}$, for $\alpha$ defined by equation (4.2), for any fixed $\alpha' > \alpha$,*

$$\left|\widehat{D^m g(x)} - D^m g(x)\right| = O_p\left(h_T^{-[m]}\left((Th_T^d)^{-1/\alpha'} + h_T^r\right)\right)$$

*and if $\{h_T\}$ is s.t. $h_T \sim_p T^{-\frac{1}{r\alpha'+d}}$*

$$\left|\widehat{D^m g(x)} - D^m g(x)\right| = O_p\left(T^{\frac{r-[m]}{r\alpha'+d}}\right)$$

REMARKS:

(1) This result has an interesting relationship with the rates of convergence in the Weibull tail-equivalent cases, as shown next.

(2) The result is remarkably simpler than the one obtained in Theorem 3. It also allows transparent interpretation of the determinants of the speed of convergence, namely that:

(a) the curse of dimensionality is present – higher dimensionality causes slower rate of convergence,

(b) the smoothness order matters: the smoother the underlying function is, the higher rate of convergence can be achieved asymptotically, provided that the order of polynomial fit corresponds to this smoothness, i.e., the highest polynomial power should be equal $k = [r]$,

(c) parameter $\alpha$ controls how much probability mass is located near the boundary; low values of $\alpha$ correspond to higher masses, and high values – to the lower masses; the higher the mass, the higher rate of convergence can be achieved.

Now let us parameterize the tail behavior of the distribution function of the positive error $u$, conditional on $x$, as follows:

**Assumption 9** (MODEL 2) *Assumption 7 holds, and, in addition, $C(u,x)/C(x) \to 1$, uniformly in $\mathcal{C}, \mathcal{G}$, so that*

$$P_g(u_t \leq \varsigma|x) = F(\varsigma|x) \sim C(x)\varsigma^{\alpha(x)}, \quad as \quad \varsigma \longrightarrow 0^+,$$

(where $C(x), \alpha(x)$ is possibly dependent on $g \in \mathcal{G}$). From the Gnedenko Theorem (Lemma 3), it follows that $\mathcal{N}^{(1/\alpha(x))} U_{(1)|x} \Rightarrow W_{\alpha(x)}$, since, for $\varsigma \in \mathbb{R}_{++}$ fixed

$$\lim_{z \searrow 0} \frac{F(\varsigma z|x)}{F(z|x)} = \lim_{z \searrow 0} \frac{C(x)(\varsigma z)^{\alpha(x)}}{C(x)(z)^{\alpha(x)}} = \varsigma^{\alpha(x)}$$

and a choice of $\{\gamma_{\mathcal{N}}(x)\}$, as follows from Lemma 5, is given by:

$$\gamma_{\mathcal{N}}(x) = \left(\frac{\mathcal{N}}{C(x)}\right)^{(1/\alpha(x))} \quad \text{and} \quad \tilde{\gamma}_{\mathcal{N}}(x, \mho) = \text{essinf}_{x' \in N_{\mho}(x)} \left(\frac{\mathcal{N}}{C(x')}\right)^{(1/\alpha(x'))}.$$

**Theorem 4** *Assume 1 - 4, 6, 9. 1. Then, for any $x$ in $\mathcal{C}$ fixed, equation (4.4) holds. 2. If, in addition, assumption 8 is true, then*

$$\left|\widehat{D^m g(x)} - D^m g(x)\right| = O_p\left(h_T^{-[m]}\left(\left(Th_T^d\right)^{\left(\frac{-1}{\alpha}\right)} + h_T^r\right)\right)$$

*so that when* $h_T \sim_p T^{\frac{-1}{r\alpha+d}}$, $\left|\widehat{D^m g(x)} - D^m g(x)\right| = O_p\left(T^{\frac{([m]-r)}{r\alpha+d}}\right)$

REMARKS:

(1) Figure 1 demonstrates the cases of densities corresponding to the above parameterization. The density exhibits a nice range of behavior near the boundary. Higher values of $\alpha$ correspond to the lower mass near boundary.

(2) As noted earlier, this result has an interesting relationship with the result in the previous corollary, namely, that some good bounds could be read off the bounds of the "nearby" Weibull tail-equivalent densities. This is briefly generalized in the next section.

(3) It should also be pointed that the tail parameterization given includes many well known densities, such as exponential, gamma, Weibull, all truncated distributions with strictly positive density at the truncation point. The last case, corresponding to tail parameterization with exponent $\alpha = 1$ is useful in sample selection models, e.g., censored and truncated quantile regressions. One of the concluding sections of this paper is entirely devoted to discussion and examples of application of extreme quantile regression to these kinds of models. See also Chernozhukov and Hong (1998b), (1998c) for other applications. The other cases, where $\alpha \neq 1$, are also quite general. For applications see Smith (1994) and section 9.

### 4.3 Simple Bounds on the Pointwise Rates of Convergence

The results developed in the previous section are useful in leading us to consider the cases where no good characterization of the tail behavior is known but the probability and a.s. bounds can be stated with only rough information. These contemplations route us to

**Theorem 5** (MODEL 3) *Fixing any $x \in \mathcal{C}$, suppose that, for some small fixed $\delta > 0$ and $T^* > 0$:*

$$\liminf_{\varsigma \searrow 0} \inf_{t > T^*, g \in \mathcal{G}} \left(\frac{\inf_{x' \in \bar{N}_\delta(x)} P_g\left(u_t \leq \varsigma | x_t = x\right)}{\varsigma^{\alpha(x)}}\right) > 0, \tag{4.6}$$

*then with assumptions 1 - 4, 6, for any $x \in \mathcal{C}$, probability bounds in Theorem 4 apply here (with $\alpha(x)$ replacing $\alpha$).*

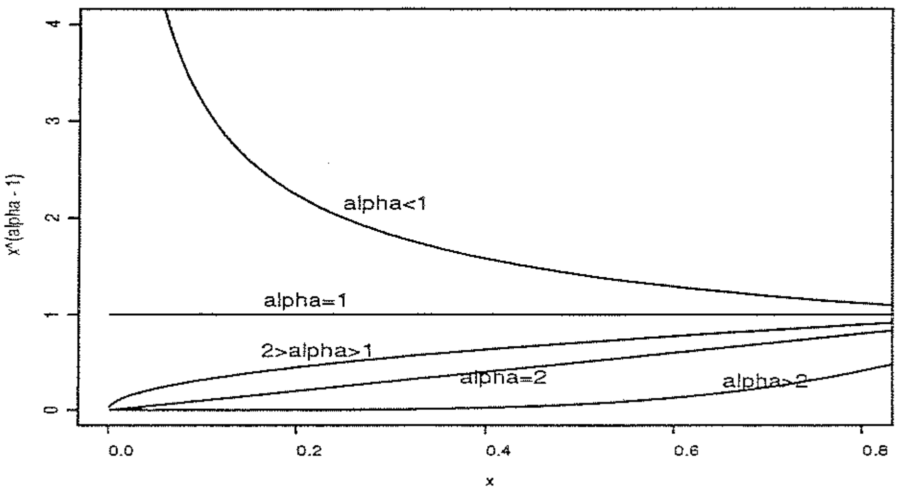


Figure 1: Tail Behavior of Weibull Tail Equivalent Densities

### 4.4 Rates of Convergence in Degenerate Cases

The term degeneracy refers to the asymptotic degeneracy of the conditional extreme order statistics. Namely, the case considered here is: for some fixed $T_*, \zeta > 0$, and $x \in \mathcal{C}$

$$\inf_{t > T_*} \inf_{x' \in N_\zeta(x), g \in \mathcal{G}} P_g\left(u_t = 0 \middle| x_t = x'\right) > 0, \tag{4.7}$$

that is, errors $\{u_t\}$ follow laws that assign strictly positive probability to the boundary. This is our MODEL 4. There are other cases which can not be subjected to the Gnedenko Limit Theorem or its generalizations (see Leadbetter et al and Galombos for examples). These can be treated on a case by case basis using the result of the previous section.

**Theorem 6** *Given assumptions 1 - 4, 6 and if (4.7) holds, for* $x \in \mathcal{C}$, $\limsup_T P\left(U_{(1)xjT} > 0\right) = 0$, *and hence if* $\{h_T\}$ *is s.t.* $Th_T^{\frac{d}{2}}/T^{\epsilon} \longrightarrow \infty$, $Th_T^{\frac{d}{2}}/T^{\epsilon'} \longrightarrow 0$, *in pr., for some small fixed* $\epsilon, \epsilon' > 0$:

$$\left|\widehat{D^m g}(x) - D^m g(x)\right| = o_p\left(T^{-(1-\epsilon')\frac{r-m}{d}}\right).$$

REMARKS:

(1) The basic result here is that $\epsilon'$ can be set equal to a very small fixed number. Thus the rate of convergence is roughly $O_p\left(T^{\frac{(|m|-r)}{d}}\right)$. All situations where there is positive probability mass at the boundary fall in this class of problems, and the rates of convergence, as shown, do not depend on other characteristics of the error distribution.

# 5 Asymptotic Distributions under Weibull Tail Equivalence

## 5.1 Approximate and Asymptotic Distributions

Assume $u_t$, conditionally on $x_t$, are continuously distributed near zero s.t. the distribution function has density near zero that satisfies $(\forall x \in \mathcal{C}, \forall g \in \mathcal{G})$:

$$f(z|x) \sim C(x)\alpha z^{\alpha-1}, \quad \text{as} \quad z \searrow 0, C(x) \in [K_{10}, K_{11}], K_{10}, K_{11} > 0 \tag{5.1}$$

**Assumption 10** *Assume(5.1), and that $C(x)$ is s.t. $\forall x, x' \in \mathcal{C} : |x - x'| \to 0, \forall g \in \mathcal{G}$, and some $\rho > 0$,*

$$|C(x') - C(x)| \leq O\|x' - x\|_\infty^\rho$$

Define $f_{x_t} \equiv f(\cdot|x_t), F_{x_t}(\cdot) \equiv F(\cdot|x_t)$, $I(x,T)$ be the collection of all sets $J$ of cardinality $p$ whose elements are distinct indices $\{t\}$ s.t. $t \in N_{Tx}$, for $p = dim(z_{x_t})$.

$$\Psi_T \equiv \text{diag}\left\{\text{vech}\left\{h_T^{[m]}, \quad m \in \mathbb{Z}_+^d : 0 \leq [m] \leq k\right\}\right\},$$

so that the order of elements in the diagonal of matrix $\Psi_T$ is chosen in the way that $\Psi_T^{-1} z_{x_t} = O_p(1)$; $\tilde{u}_{x_t} \equiv \Psi_T^{-1} z_{x_t}$, $\Xi \equiv \sum_{m \in \mathbb{Z}_+^d : 0 \leq [m] \leq k} [m]$ . Let $\tilde{W} \equiv \mathcal{H}_T\left(\beta_x - \hat{\beta}_x\right)$, where

$$\mathcal{H}_T = \text{diag}\left\{\text{vech}\left\{\left(Th_T^d\right)^{1/\alpha}\left(h_T^{[m]}\right), \quad m \in \mathbb{Z}_+^d : 0 \leq [m] \leq k\right\}\right\}$$

where $\mathbb{Z}_+^d$ is the set of all nonnegative $d$-dimensional integers and the order of elements in the diagonal of matrix $\mathcal{H}_T$ is chosen so that $\tilde{W} = O_p(1)$, which is possible by Theorem 4 and noting that assumption 9 is satisfied. Also $\{h_T\}$ is taken to satisfy an additional assumption, for some small $\epsilon > 0$:

$$\left(Th_T^d\right)^{(1/\alpha)} h_T^{\tau-\epsilon} \longrightarrow 0, \text{ a.s. or in pr.} \tag{5.2}$$

Here we fix $g \in \mathcal{G}$, and let $P = P_g$. Let $f_T$ denote the exact conditional on $\{x_t\}$ density of $\tilde{W}$ . The density $f_T^*$ is said here to be *approximate* if $f_T - f_T^*$ converges in $L_1$ norm to 0, w.p.1 or in pr. (implying thus the weak convergence of random measures associated with each density w.p.1). Denote such convergence as $f_T \approx f_T^*$. Precisely, $f_T \approx f_T^*$ if $f_T, f_T^*$ exist *ev.* P$-a.s.$[9] and for $F_T(z) \equiv \int_{v \leq z} f_T(v)dv$ and $F_T^*(z) \equiv \int_{v \leq z} f_T^*(v)dv$ [10] for any bounded continuous function $k : \mathbb{R}^p \to \mathbb{R}$,

$$\int_{\mathbb{R}^p} k(z)(dF_T(z) - dF_T^*(z))dz \to 0, \text{ P-a.s. as } T \to \infty$$

and $f_T \approx_p f_T^*$ if $f_T$ exists w. p. $\to 1$ and the above convergence occurs in probability (P).

We also say that $f$ is asymptotic density of $\tilde{W}$ if $f \approx f_T$ (or $f \approx_p f_T^*$), if $f$ is independent of $T$ and integrates to 1 (denote such relationship as $f_T \Rightarrow f$ or $f_T \Rightarrow_p f$).

**Theorem 7** *Fix $g \in \mathcal{G}, x \in \mathcal{C}$, assume 1, 3b, 5, 6, 10, and that $\{h_T\}$ satisfies (5.2) a.s. Then $f_T^* \approx f_T$, where $f_T$ is the exact density of $\tilde{W}$ and*

$$f_T^*(w) = \left(Th_T^d\right)^{-p}\left(\sum_{J \in I(x,T)} \bar{D}(J) \prod_{t \in J}\left(C(x)\left(\tilde{u}_{x_t}' w\right)_+^{\alpha-1}\right)\right) \times \prod_{t \in N_{Tx}\}}\left(1 - C(x)\left(\tilde{u}_{x_t}' w / Th_T^d\right)_+^\alpha\right) \tag{5.3}$$

[9] Existence is discussed in the appendix, see also Smith (1994) and Portnoy and Jureckova (1998).

[10] When the density does not exist we take it to be a continuous density of an arbitrary fixed probability measure on a compact set in $\mathbb{R}^d$, so that the statements below are well defined.

*where*

$$\tilde{D}(J)=\begin{cases} det\left[\tilde{u}_{x_{t_1}},...,\tilde{u}_{x_{t_p}}\right], & for\quad \{t_1,...,t_p\}=J,\quad if\quad \bar{u}(x)\in co\left(\tilde{u}_{x_t},t\in J\right), \\ 0, & otherwise \end{cases}$$

*and given assumptions 3a and if* $\{h_T\}$ *satisfies (5.2) in pr.,* $f_T \approx_p f_T^*$.

Next, the asymptotic density is stated as:

**Theorem 8** *Fix* $g\in\mathcal{G}, x\in\mathcal{C}$*, assume 1, 3b, 5, 6, 10 and that* $\{h_T\}$ *is s.t. (5.2) holds a.s. Then the asymptotic density of* $\tilde{W}$*,* $f^*$ $(f_T \Rightarrow f)$ *is given by (for* $s\in\mathbb{R}^p$*) :*

$$f^*(s)=\frac{(w_x(x))^p}{p!}\left(\int_{\Delta^p}\left(D(\{u^{\{m\}}(j),j\in P\})\prod_{j=1}^{p}\left(\alpha C(x)\left(\left(u^{\{m\}}(j)\right)'s\right)_+^{\alpha-1}\right)\right)\prod_{j=1}^{p}dF_*(u^{\{m\}}(j))\right)$$
$$\times\left(exp\quad\left\{-w_x(x)C(x)\int_{\Delta}\left(\left(u^{\{m\}}\right)'s\right)_+^{\alpha}dF_*(u^{\{m\}})\quad\right\}\right)$$

*where for* $u\in\mathbb{R}^d$ *and* $m\in\mathbb{Z}_+^d$*,*

$$u^{\{m\}}\equiv vech\{u^m,m:[m]\le k\}\in\Delta$$
$$\Delta\equiv\{1\}\underset{m:[m]=2^j-1,j\in J_*}{\times}\left[-1/2^j,1/2^j\right]\underset{m:[m]=2^j,j\in J_*}{\times}\left[0,1/2^j\right]\subset\mathbb{R}^p,\ where$$
$$J_*\equiv\{j\in\mathbb{Z}_{++}:(2^j-1)\vee 2^j\le k\}$$

$F_*(\cdot)$ *is the distribution function of the p-dimensional random vector, s.t. for* $P_T$*, the empirical measure of* $\{x_t\}$*:* $lim_{T\to\infty}P_T(\tilde{u}_{x_t}\le y)=F_*(y)$*, for* $\forall y\in\mathbb{R}^p$*, where* $F_*$ *does not depend on the choice of* $\{h_T\}$[11]*; and for* $\{u^{\{m\}}(i), i\in P\equiv\{1,...p\}\}\in\mathbb{R}^{p\times p}$

$$D\left(\{u^{\{m\}(j)},j\in P\}\right)=\begin{cases} det\left[u^{\{m\}}(1),...,u^{\{m\}}(p)\right], & if\quad \bar{u}^{\{m\}}\in co\left(u^{\{m\}}(i),i\in P\right), \\ 0, & otherwise, \end{cases}$$

*where* $\bar{u}^{\{m\}}=\int_{\Delta}u^{\{m\}}dF_*(u^{\{m\}})$*. Further, under assumption 3a and that (5.2) holds in pr.,* $f_T\Rightarrow_p f$.

The formula is not cumbersome: it is not at all hard to compute it. For example, consider the locally constant regression, i.e., $z_{x_t}=1,\forall t$, then the asymptotic formula above reduces to a simple Weibull density.

**Corollary 3** *Assume 1 with k=0, 3a, 5, 6, 10 and that* $\{h_T\}$ *is s.t.* $(Th_T^d)^{(1/\alpha)}h_T^{\gamma-\epsilon}\to 0$*, for small* $\epsilon>0$*. Then an approximate density of* $\left((Th_T^d)^{(1/\alpha)}(g(x)-\widehat{g(x)})\right)$ *is given by:*

$$f_T^*(z)=\frac{1}{(Th_T^d)}\{card\{N_{Tx}\}\alpha C(x)z_+^{\alpha-1}\}\{1-C(x)z_+^{\alpha}/Th_T^d\}^{card\{N_{Tx}\}}$$

*(where* $card\{N_{Tx}\}$ *can be replaced by* $w_x(x)Th_T^d$*) and the asymptotic density is given by:*

$$f^*(z)=\alpha w_x(x)C(x)z_+^{\alpha-1}\exp\{-w_x(x)C(x)z_+^{\alpha}\}$$

REMARKS:

(1) $F_*$ above is the distribution function of vector $u^{\{m\}}\equiv$ vech$\{u^m,m:0\le[m]\le k\}$, where $u$ is the vector of $d$ independent variables $u_1,...u_d$, that are uniformly distributed on $[-1/2,1/2]$.

[11] $F_*$ is further characterized in the remarks

(2) Consistency is not required to obtain the above theorems. However, the theorems imply consistency of $\hat{g}(x)$. In order to have consistency of $\widehat{D^m g(x)}$ implied from these theorems, we must in addition have: $(Th_T^d)^{-1/\alpha} h_T^{-[m]} \to 0$, in pr.

(3) The choice of a bandwidth sequence above is not optimal and hence involves undersmoothing. This is a typical feature of nonparametric estimators: undersmoothing is needed to obtain an (well-behaved) asymptotic distribution. If the optimal bandwidth choice is used instead, an additional "bias" term, that depends on $\{D^m g(x), m : [m] = k\}$, appears. For practical reasons, we have ignored such a case.

(4) To present an example, take locally linear regression, and $dim(x) = 1$, i.e. $z_{x_t} = (1, x_t - x)$, then $dF_*(u)$ is $\delta_{u_1=1} 1(u_2 \in [-1/2, 1/2])$ ($\delta_{u_1=1}$ is Dirac delta); this is probability measure that assigns mass 1 to all cases when $u_1 = 1$ and uniform on $[-1/2, 1/2]$ in its second coordinate: i.e. $f_*(u_2|u_1 = 1) = 1(u_2 \in [-1/2, 1/2])$. This implies that the joint asymptotic density of $\bar{W} \equiv \Big((Th_T^d)^{1/\alpha}(g(x) - \hat{g}(x)), (Th_T^d)^{1/\alpha} h_T^{-1}(D^1 g(x) - \widehat{D^1}$ is given by (note that $\Delta = 1 \times [-1/2, 1/2], \bar{u}^{\{m\}} = (1, 0)$):

$$f^*(s_1, s_2) = \frac{(w_x)^2}{2}\Big(\int_{[-1/2,1/2]^2} (D(\{(1,u_1)', (1,u_2)'\}) \prod_{j=1}^{2} \alpha C(x)((s_1 + s_2 u_j)_+^{\alpha-1}) \prod_{j=1}^{2} du_j\Big)$$
$$\times\Big(exp \quad \{-w_x C(x) \int_{[-1/2,1/2]} (s_1 + s_2 u)_+^{\alpha} du \quad \}\Big)$$

(5) From the proof of Theorem 9, it follows that moments of the local extreme quantile estimates converge to the moments of the asymptotic distribution, provided $\{u_t\}$ has finite higher moments.

(6) The result allows for random, i.e. data-driven, bandwidth sequences, provided they satisfy the general conditions ( assumption 3a or b ).

(7) Appendix C provides a brief Monte-Carlo experiment, plotting the simulated finite sample-densities vs. asymptotic approximations.

## 5.2 Asymptotic and Bootstrap Methods of Inference

Conventional bootstrap methods are not consistent for extreme quantile regression models, either parametric or non-parametric. An example can be constructed along the lines of the well known example of Bickel and Friedman. Subsampling methods work in our settings.

# 6 Global Estimation of the Extreme Quantile Function and Its Derivatives

Here we present results on the rates of convergence in the case where the whole function $g(x)$ (or $D^m g(x)$) is estimated over some compact domain with a minimally smooth boundary, $\mathcal{C}$. The results are useful for the problems where extreme quantiles are used sequentially. See later sections on the sequential uses of extreme quantiles.

## 6.1 Construction of the Global Estimator

It is not feasible to compute the point estimate at every point in the regressors space. Hence some kind of interpolation procedure is needed in order to construct the global estimator. Also, we define the $L_q$ norm of a measurable function $y : \mathbb{R}^d \to \mathbb{R}$ on set $\mathcal{C}$ as:

$$\left\|y\right\|_q \equiv \left[\int_{\mathcal{C}} |y(x)|^q dx\right]^{1/q}, \left\|y\right\|_\infty \equiv \left[\sup_{x\in\mathcal{C}} |y(x)|\right]$$

GENERAL INTERPOLATION. Following a suggestion by Stone (1982), define the estimator only for the grid of points $\{x\} \in C_T$, and then define the global estimator $\overline{D^m g}$ by an interpolation. Let $\mathcal{Z} \equiv \{\widehat{D^m g(x)}, x \in C_T\}$, $\mathcal{V} \equiv \{D^m g(x), x \in C_T\}$ then for some map $\Gamma$ (that may depend on T) $\overline{D^m g(x)} \equiv \Gamma(\mathcal{Z})$, $\widetilde{D^m g(x)} \equiv \Gamma(\mathcal{V})$. Then

$$\left\|\overline{D^m g} - D^m g\right\|_q = O\left(\left\|\overline{D^m g} - \widetilde{D^m g}\right\|_q + \left\|D^m g - \widetilde{D^m g}\right\|_q\right) \equiv \mathcal{E}_1 + \mathcal{E}_2$$

If $\mathcal{E}_1 = o_p(\mathcal{E}_2)$, the rate of convergence depends only on the quality of approximation, and not on any of the statistical properties. This is the subject of approximation theory and not ours. Thus we require $\mathcal{E}_2 = O_p(\mathcal{E}_1)$. Further, we note that $C_T$ is the collection of $O(L_T^d)$ points in $\mathcal{C}$, where $1/L_T$ is the maximum distance ($\|\cdot\|_\infty$ ) between the two closest points $x, x' \in C_T : x \neq x'$. We require that card$(C_T) = O_p(T^b), b > 0$ and of $\{\Gamma\}$ to have the property:

$$\left\|\overline{D^m g} - \widetilde{D^m g}\right\|_q = O\left(\left\|\overline{D^m g} - D^m g\right\|_{q,C_T}\right)$$

where for $y : \mathbb{R}^d \to \mathbb{R}$, $\|y\|_{q,C_T} = |\text{card}(C_T)^{-1} \sum_{x\in C_T} |y(x)|^q|^{1/q}$, for $1 \le q < \infty$ and $\|y\|_{\infty,C_T} \equiv \sup_{x\in C_T} |y(x)|$.
EXAMPLE: A LINEAR INTERPOLATION. For every $x \in \mathcal{C}$ there are points $x(l) \in C_T \cap \mathcal{C}$, for some index $l$, that are within $1/L_T$ from $x$. Then define the estimate at point $x$ as:

$$\overline{D^m g(x)} \equiv \Gamma\left(z \equiv \text{vech}\{\widehat{D^m g(x)}, x \in x(l)\}, x\right) = \Gamma(z, x) \tag{6.1}$$

Here pick $\Gamma(z, x) = \lambda_x' z$, where $\lambda_x > 0, \lambda_x' 1 = 1$ is chosen so that $x$ is on the plane adjoining points in $z$. It is straightforward to check that this interpolation satisfies the earlier requirements, and that $\mathcal{E}_2 = O_p(\mathcal{E}_1)$ in the context of the next theorems 9 and 10, if, for instance, for any fixed $m \in \mathbb{Z}_+^d : [m] \le k$, we pick

$$1/L_T \sim_p h_T^{(r-[m])K}, \text{ for } K \ge \frac{1}{(r-[m]) \wedge 1} \tag{6.2}$$

To have one rule that works for any $m \in \mathbb{Z}_+^d : [m] \le k$, set

$$1/L_T \sim_p h_T^{(r)K}, \text{ for } K \ge \frac{1}{r \wedge 1} \tag{6.3}$$

Alternatively, one may use the following rules (these rules work for both Theorem 9 and 10 as a consequence of Lemma 6, although it is perhaps more obvious then to replace the terms $(Th_T^d)^{-1}$ with $\ln T(Th_T^d)^{-1}$ in Theorem 10):
under assumption 7 (model 1), set

$$1/L_T \sim_p \left(\text{esssup}_{x\in\mathcal{C}} \left[F^{-1}\left((Th_T^d)^{-1}|x\right)\right]^{\left(\frac{\varsigma}{(r-[m])\wedge 1}\right)}\right) \text{ for fixed } \varsigma \ge 1; \tag{6.4}$$

for the special case of assumption 9 (model 2) and for the case described in (4.6), model 3, choose

$$1/L_T \sim_p \left([T^K]\right) \text{ for } K \ge \frac{-r + [m]}{((r-[m]) \wedge 1)(r\underline{\alpha} + d)}; \tag{6.5}$$

for the degenerate case (eq. 4.7), model 4, set

$$1/L_T \sim_p \left(K'[T^K]\right) \text{ for } K > \frac{-r+[m]}{((r-[m])\wedge 1)d} \tag{6.6}$$

And to make the later set of rules applicable to all $m \in \mathbb{Z}_+^d : [m] \leq k$, set $[m] = 0$ in the above formulas. We (subjectively) find that the rules in equations (6.2), (6.3) are easier to use than the rules in (6.4), (6.5), since the latter would typically require the estimates of tail indices, $\alpha(x)$ (see section 8 ).

EXAMPLE: OTHER INTERPOLATION METHODS. The linear interpolation method defined above is of a simplest kind. Other methods include construction of global polynomials that pass through a given set of points defined by the grid net constructed as in the previous section. These methods are usually the Lagrange, Hermite interpolations (or, more generally, cardinal interpolations), Chebyshev, and other methods. The convergence theory for these methods is poor and often relies on the condition that the underlying function is in $C^\infty$ and on the construction of nonuniform grids (such as by the Chebyshev method)(e.g., Atkinson (1989), Rivlin (1990)). Since $\widehat{D^m g}$ approaches $D^m g$ at the grid points, the interpolation is essentially that of $D^m g$, evaluated at the grid points, hence these methods qualify as unnecessarily restrictive. Another group of interpolation methods that includes linear interpolation as a special case are the piecewise Hermite polynomials, splines, and B-splines. In general, a higher order polynomial spline interpolation decreases the order of cardinality of the grid net required to achieve the necessary quality of approximation (see Prenter (1989), de Boor (1978)). Certain modifications of these methods allow the global estimates to be treated as nice functions in a stochastic entropy sense (see the definitions of type III function in Andrews (1994)).

## 6.2 Convergence Rates in $L_q$ norm, $1 \leq q < \infty$

In this subsection, we consider the uniform rates of convergence of global estimate $\overline{D^m g}$ in $L_q$ norm. The subsequent developments rely on the probability bounds generated by the moments of the extreme order statistics. Thus the analysis relies on studying the asymptotic behavior of the moments.

Theorem 1 and the bounds on the moment behavior (that can be referred to in the appendix, Lemma 8 ) deliver the following theorem. This theorem covers the models 1-4. The section on pointwise rates of convergence presents a detailed discussion of each of the cases, hence it it is not repeated here and the result is stated in a compact form.

**Theorem 9** *Under assumptions 1 − 3, 6,*

$$\left\|\overline{D^m g} - D^m g\right\|_q = O_p\left(h_T^{[-m]}\left(\Upsilon_T(h_T) + h_T^r\right)\right)$$

*(1) where in the model (1) of regular variation (assumption 7),*

$$\Upsilon_T(h_T) = \int_C \text{esssup}_{x' \in \tilde{N}_{h_T}(x)}\left(F^{-1}\left((Th_T^d)^{-1}|x'\right)\right)dx$$

*and if $h_T^*$ is s.t $h_T^* \sim_p \left(\Upsilon(h_T^*)^{1/r}\right)$, then $\left\|\overline{D^m g} - D^m g\right\|_q = O_p\left((\Upsilon_T(h_T^*))^{\frac{r-m}{r}}\right)$,*
*(2) in particular, in the model (2) of Weibull tail equivalence (assumption 9),*

$$\Upsilon_T(h_T) = \int_C \text{esssup}_{x' \in \tilde{N}_{h_T}(x)}\left((Th_T^d)^{-\frac{1}{\alpha(x)}}\right)dx,$$

*(3) in model 3, when* $\forall x \in \mathcal{C}$ *(4.6) holds,*

$$\Upsilon_T(h_T) = \int_{\mathcal{C}} \left( (Th_T^d)^{-\frac{1}{\alpha(x)}} \right) dx,$$

*(4) in the model of degeneracy (model 4), when for some fixed* $T_*, \zeta > 0$,

$$\inf_{t > T_*} \inf_{x' \in \mathcal{C}, g \in \mathcal{G}} P_g \left( u_t = 0 \middle| x_t = x' \right) > 0, \tag{6.7}$$

*for any fixed small* $a' > 0$ *in assumption* 3,

$$\Upsilon_T(h_T) = 0$$

*and hence if* $\{h_T\}$ *is s.t.* $Th_T^d / T^\epsilon \longrightarrow \infty$, $Th_T^d / T^{\epsilon'} \longrightarrow 0$, *in pr. (P), for some small fixed* $\epsilon, \epsilon' > 0$:

$$\left\| \overline{D^m g} - D^m g \right\|_q = o_p \left( T^{-(1-\epsilon')\frac{r-|m|}{d}} \right).$$

The following corollary states the simplifications that result from assuming special cases which were discussed in 4.2.

**Corollary 4** *Under assumptions 1 − 3, 6 and 8:*

$$\left\| \overline{D^m g} - D^m g \right\|_q = O_p \left( h_T^{-|m|} \left( \Upsilon_T(h_T) + h_T^r \right) \right)$$

*(1) where in the model (1) of regular variation (assumption 7),*

$$\Upsilon_T(h_T) = F^{-1} \left( (Th_T^d)^{-1} \right)$$

*and if* $h_T^*$ *is optimal:* $h_T^* \sim_p \left( \left( F^{-1} \left( (Th_T^{*d})^{-1} \right) \right)^{1/r} \right)$, *then*

$$\left\| \overline{D^m g} - D^m g \right\|_q = O_p \left( F^{-1} \left( (Th_T^{*d})^{-1} \right)^{\frac{r-|m|}{r}} \right),$$

*(2) in particular, in the model of Weibull tail equivalence (assumption 9), model 2,*

$$\Upsilon_T(h_T) = (Th_T^d)^{-\frac{1}{\alpha}}$$

*and if* $h_T^*$ *is optimal:* $h_T^* \sim_p \left( T^{-\frac{1}{r\alpha + d}} \right)$, *then*

$$\left\| \overline{D^m g} - D^m g \right\|_q = O_p \left( T^{\frac{|m|-r}{r\alpha + d}} \right),$$

*(3) in model 3, when* $\forall x \in \mathcal{C}$, *equation (4.6) holds, and if* $\alpha(x) = \alpha$, *then the bounds of model (2) apply.*

## 6.3 Rates of Convergence in $L_\infty$ norm

**Theorem 10** *Under assumptions 1 − 4, 6*

$$\left\| \overline{D^m g} - D^m g \right\|_\infty = O_p \left( h_T^{-|m|} \left( \Upsilon_T(h_T) + h_T^r \right) \right)$$

*(1) where in the model (1) of regular variation (assumption 7),*

$$\Upsilon_T(h_T) = esssup_{x' \in \mathcal{C}} F^{-1} \left( lnT (Th_T^d)^{-1} | x' \right)$$

*and if* $h_T^*$ *is s.t.* $h_T^* \sim_p \left(\Upsilon\left(h_T^*\right)^{1/r}\right)$, *then* $\left\|\overline{D^m g} - D^m g\right\|_\infty = O_p\left(\left(\Upsilon_T\left(h_T^*\right)\right)^{\frac{r-[m]}{r}}\right)$
*(2) in particular, in the model (2) of Weibull tail equivalence (assumption 9),*

$$\Upsilon_T\left(h_T\right) = esssup_{x\in C}\left(\left((lnT)^{-1}(Th_T^d)\right)^{-\frac{1}{\alpha(x)}}\right)$$

*(3) in model 3, when* $\forall x \in C$ *(4.6) holds,*

$$\Upsilon_T\left(h_T\right) = esssup_{x\in C}\left(\left((lnT)^{-1}Th_T^d\right)^{-\frac{1}{\alpha(x)}}\right),$$

*(4) in the model of degeneracy (model 4), when for some fixed* $T_*, \zeta > 0$,

$$\inf_{t>T_*}\inf_{x'\in C, g\in G} P_g\left(u_t = 0\middle| x_t = x'\right) > 0, \tag{6.8}$$

*for any fixed* $a' > 0$ *in assumption* 3,

$$\Upsilon_T\left(h_T\right) = 0$$

*and hence if* $\{h_T\}$ *is s.t.* $Th_T^d/T^\epsilon \longrightarrow \infty$, $Th_T^d/T^{\epsilon'} \longrightarrow 0$, *in pr. (P), for some small fixed* $\epsilon, \epsilon' > 0$:

$$\left\|\overline{D^m g} - D^m g\right\|_\infty = o_p\left(T^{-(1-\epsilon')\frac{r-m}{d}}\right)$$

A corollary similar to that stated after Theorem 9 can be stated here as well.
REMARKS:

(1) While we needed to know the behavior of (fixed number of) moments of (truncated) local extreme statistic in Theorem 9, we require here a much stronger result – an exponential inequality on the deviation of $U_{(1)jh_T}(x)$ from $\gamma_{Th_T^d}(x, h_T) = F^{-1}((Th_T^d)^{-1})$ (take $\{u_t\}$ i.i.d. case for discussion).

(2) To develop this inequality, a previous version of this work introduced a concept of **log-regularity** of $F$ in order to use the natural exponential inequality – the limit law of $U_{(1)jh_T(x)}$, so that for large $T$ ( and $\alpha : F \in \mathcal{D}(\mathcal{W}_\alpha)$, $C > 0$)

$$P\left(U_{(1)jh_T}(x) > \ln T F^{-1}((Th_T^d)^{-1})\right) \le \exp\{-(\ln T)^\alpha C\}$$

But this required a large-deviation result (which we called log-regularity). [12] Log-regularity imposes, as $y \to \infty$:

$$yF\left((\ln y)^{1/\alpha}F^{-1}(y^{-1})\right) \sim \ln y$$

in order to obtain the above bound (in addition to assumption 7). This puts restrictions on $F$, but it does not provide bounds, sharper than those in Theorem 10, since by assumption 7, in its first part:

$$yF\left(F^{-1}(\ln y y^{-1})\right) \sim \ln y$$

This implies that given assumption 7, log-regularity is redundant. Using only assumption 7, a somewhat different exponential inequality is developed in appendix (Lemma 9 ). It bounds the probabilities of deviation of $U_{(1)jh_T}(x)$ from $F^{-1}(\ln T(Th_T^d)^{-1})$. We conjecture, however, that the log-regular case is possibly the only case when the exponential bound is sharp (for i.i.d case).

[12] see Anderson (1978), (1984), Goldie and Smith (1987), de Haan and Hordijk (1972) and listed there citations on large deviation analysis in extreme value theory. The concept of log-regularity is a special case of the concept of super-slow variation.

# 7 Sequential Uses of Extreme Quantiles

It is immediate that the global estimates can be used sequentially to help estimate other statistical models. For example, a variety of pseudo-maximum likelihood methods or other semi-parametric general extremum estimators could be considered.

Let $\mathbf{y}_T = ((y_1, x_1)..., (y_T, x_T))'$, $D^m g$ be a function: $\mathbb{R}^d \to \mathbb{R}$, $B$ be a compact subset of $\mathbb{R}$, and $\hat{\beta}$ be an estimator defined as:

$$\hat{\beta} \in \text{argmax}_{\beta' \in B} Q_T \left(\mathbf{y}, D^m g, \beta'\right),$$

and $\tilde{\beta}$ be an estimator defined as:

$$\tilde{\beta} \in \text{argmax}_{\beta' \in B} Q_T \left(\mathbf{y}, \overline{D^m g}, \beta'\right).$$

The above estimators are general extremum estimators, as defined in Amemiya (1985) (see also Newey and McFadden (1994)). The special cases include practically all known estimators: GMM, minimum-distance, m-estimators, MLE, etc. Here we look at conditions that suffice for 1) consistency of $\tilde{\beta}$, 2) asymptotic equivalence of $\tilde{\beta}$ and $\hat{\beta}$. The latter is a fairly important question from a practical standpoint, since if preliminary estimate, $\overline{D^m g}$, affects the asymptotic distribution of $\tilde{\beta}$, the properties of this distribution are largely not known at this time, as no standard functional Delta methods extend to the estimation problems that we consider. To address this concern, we look here at simple sufficient regularity conditions that guarantee no asymptotic second-stage effect. For further and more general discussions of asymptotic behavior of two-stage estimators, in which the first stage is non-parametric and has no effect on the second stage, can be found in Andrews (1994). In what follows, the probability statements are with respect to the outer probability measure, which we denote here as $P$ (that is, notation is not altered) (see van der Vaart and Wellner (1996) for definitions)[13].

CONSISTENCY. In all applications that we study here and elsewhere (Chernozhukov and Hong (1998b), Chernozhukov and Hong (1998c)), uniformly in $\beta'$, $Q_T(\mathbf{y}, D^m g, \beta')$ converges in probability to function $Q(\beta')$, that is uniquely maximized at $\beta$. By Theorem 4.1.1 in Amemiya (1985), $\hat{\beta} \to \beta$ in pr. To assert $\tilde{\beta} \to \beta$, in pr., it suffices to show $\mathcal{P}(\overline{D^m g}, D^m g) \to 0$ in pr., where $\mathcal{P}(f_1, f_2) \equiv \sup_{\tilde{\beta} \in B} \left|Q_T(\mathbf{y}, f_1, \tilde{\beta}) - Q_T(\mathbf{y}, f_2, \tilde{\beta})\right|$. In many cases, simple analysis emerges when the following holds:

$$\mathcal{P}(\overline{D^m g}, D^m g) = O_p \left(M_T k \left(\left\|\overline{D^m g} - D^m g\right\|_q\right)\right)$$

where $1 = O_p(M_T)$, and $k > 0$ is strictly monotone and continuous function s. t. $k(0) = 0$. In this instance, using models and assumptions stated in the Theorems 9 and 10, one is to find the set of $(r, d, [m], F(\cdot|x))$ such that $\left\|\overline{D^m g} - D^m g\right\|_q = k^{-1}\left(o_p\left(M_T^{-1}\right)\right)$. Typically, when $M_T \sim_p 1$, $r - [m] > 0$ suffices (for models 1-4). Such situations arise in the examples of sections 8.1-8.2. In more general settings, one may directly establish the (weak) uniform stochastic equicontinuity (u.s.e.)[14]: Let $\rho$ be a pseudo-metric (e.g. defined by $\|\cdot\|_q$) and $\mathcal{G}$ be a functional space, then $\{Q_T\}$ is $(\mathcal{G}, \rho)$-u.s.e. if for $\forall \epsilon > 0, \exists \delta > 0$ s.t.

$$\limsup_T P \left( \sup_{f_1, f_2 \in \mathcal{G}: \rho(f_1, f_2) < \delta} \mathcal{P}(f_1, f_2) > \epsilon \right) < \epsilon$$

[13] Here we repeatedly use Theorems 4.1.1., 4.1.3 in Amemiya (1985). These theorems are trivially extended to the cases when we use $P$, the outer probability measure. Alternatively, one may refer to van der Vaart and Wellner (1996) for the analogues of these theorems under $P$.

[14] The definition of stochastic equicontinuity slightly differs from that in e.g. Andrews (1994)

$\mathcal{G}$ and $\rho$ have to be picked so that $\mathcal{G}$ is well-behaved and totally bounded under $\rho$ space, $P\left(D^m g \in \mathcal{G}\right) \to 1$, and $\rho\left(\overline{D^m g}, D^m g\right) \to 0$. These and u.s.e. condition yield the necessary uniform convergence: $\mathcal{P}(\overline{D^m g}, D^m g) \to 0$, in pr.[15] In the context of weak convergence, to be discussed, it may be desirable to take $\mathcal{G}$ to be a space of functions with good entropy bounds. The global estimators defined in section 6 can be shown to posses such quality, at least with probability converging to one.

ASYMPTOTIC DISTRIBUTION. For some increasing function $b$, $\tilde{\beta}$ and $\hat{\beta}$, under the regularity conditions of Theorem 4.1.3 in Amemiya (1985), satisfy the estimating equations (w.p. $\to$ 1):

$$b(T)\left(\tilde{\beta}-\beta\right) = R_T\left(\mathrm{y}, \overline{D^m g}, \tilde{\beta}\right) + o_p(1), \quad b(T)\left(\hat{\beta}-\beta\right) = R_T\left(\mathrm{y}, D^m g, \hat{\beta}\right) + o_p(1).$$

where $R_T\left(\mathbf{y}, D^m g, \hat{\beta}\right)$ converges weakly to $Z$ ($\Rightarrow$). For instance, in sections 8.2 and 8.3, $b(T) = \sqrt{T}$ and $R_T(y,\beta,g)$ is of the form $T^{-1/2}\sum_t \eta_t(y_t, x_t, g(x_t), \beta)$. To find the conditions under which $R_T(\mathrm{y}, \overline{D^m g}, \tilde{\beta}) \Rightarrow Z$ (and hence $b(T)(\tilde{\beta}-\beta) \Rightarrow Z$ ), one should repeat the earlier analysis by redefining $\mathcal{P}(f_1, f_2) \equiv \sup_{\tilde{\beta}\in B}\left|R_T(\mathrm{y}, f_1, \tilde{\beta}) - R_T(\mathrm{y}, f_2, \tilde{\beta})\right|$ and proceeding as above. Reader may further refer to Andrews (1994) for further discussion of techniques of establishing stochastic equicontinuity.

## 7.1 Estimation of Auxiliary Parameters by Boundary-Dependent Maximum Likelihood

The following discussion is based on the model of Smith (1994), (1985). Indeed, although Smith (1994) considers parametric estimation of the (linear) regression quantiles, his results are extendible to the first step nonparametric estimation. The model is as before:

$$y_t = g(x_t') + u_t, u_t > 0$$

where $g$ is an unknown smooth function, $u_t$ are random variables. Define $\nabla_z m$ as the partial derivative of $m$ w.r.t. $z$. It is assumed that the errors $\{u_t\}$ have density of the form:

$$f(y, g(x'), \phi, x) = 1(y > g(x'))s\left(y - g(x'), x, \phi\right)\left(y - g(x')\right)^{\alpha(\phi)-1}, \alpha(\phi) > 0, \tag{7.1}$$

Note that $\alpha(\phi)$ is the tail index. Earlier, we have mentioned that in many instances we can estimate $\alpha(\phi)$ and hence deduce the rate of convergence and optimal bandwidth sequences. The presented here MLE is one such method [16]. Of course, other parameters of the likelihood function, $\phi$, are estimated by this method as well. Assume here for $s = s(z, x, \phi)$, $\nabla_z s$, $\nabla_\phi s$, $\nabla_{\phi z} s$, $\nabla_{\phi\phi' z} s$ , $\nabla_\phi \alpha$, $\nabla_{\phi\phi'} \alpha$ are 1) bounded uniformly in $z \in \mathbb{R}_+, \phi \in \Phi, \{x_t\} \in X$ w.p. 1, 2) continuous in all arguments; 3) $s(z, x_t, \phi) \to \alpha(\phi)C(\phi, x_t)$ as $z \to 0$, uniformly in $\{x_t\} \in$X, $\phi \in \Phi$. It is convenient also to define notation: $f(y, g(x'), \phi, x, a)$, which equals $f(y, g(x'), \phi, x)$, except that we replace $1(y > g(x'))$ with $1(y > g(x') + a)$. Define now the log-likelihood function: $L_T(\phi, g, a) \equiv T^{-1}\sum_{t=1}^T \ln f(y_t, g(x_t), x_t', \phi, a)$ ( with $\ln 0 \equiv 0$) and two ML estimators. $\hat{\phi}, \tilde{\phi}$ of the form:

$$\hat{\phi}_T \in \text{argmax}_\Phi L_T(\phi, \hat{g}, a_T), \quad \tilde{\phi}_T \in \text{argmax}_\Phi L_T(\phi, g, 0).$$

[15] Using u.s.e. is of benefit, since it is generally easy to show it using the approaches suggested in Andrews (1994), van der Vaart and Wellner (1996), and others. Usually, one can decompose the functional form of the objective function and so on into well-behaved classes of functions, in the entropy sense, and then apply entropy stability results to assert the s.e.

[16] A more general form of tail indices an their estimation methods may be found in Chernozhukov (1999c)

(where we put $a_T = (T/(\ln T)^{1+\epsilon})^{\frac{-r}{r\alpha+d}})$[17]. In (7.1), we clearly have model 2, with $\alpha(x) = \alpha, \forall x$[18]. Therefore, using Theorem 10 and the optimal bandwidth choice [19], $h_T \sim_p T^{-\frac{1}{r\alpha+d}}$ we have $\|\hat{g}-g\|_\infty = O_p\left((T/lnT)^{-\frac{r}{r\alpha+d}}\right)$. If we do not know $\alpha$, we should first estimate it by the "pilot" MLE, and then use $\hat{\alpha}$ in this expression. This does not affect the results, as explained in the proof.

**Theorem 11** *Let $(x', x, \phi_0, y) \in X' \times X \times interior\Phi \times \mathbb{R} \subset \mathbb{R}^d \times \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \times \mathbb{R}$. $X' \subset X, X', \Phi$ are compact. Fix $g \in \mathcal{G}$, let assumptions 1-4a ($X' = \mathcal{C}$), 6, and (7.1) hold. Then, for $\alpha \equiv \alpha(\phi_0)$*

$$\hat{\phi} - \tilde{\phi} = \begin{cases} O_p\left(\left(T\,(lnT)^{-(1+\epsilon)}\right)^{\frac{-r\alpha}{r\alpha+d}} lnT\right) & \text{if} \quad 0 < \alpha < 1, \\ O_p\left(\left(T\,(lnT)^{-1}\right)^{\frac{-r}{r\alpha+d}}\right) & \text{if} \quad \alpha \geq 1, \end{cases}$$

$$\hat{\phi} - \tilde{\phi} = \begin{cases} o_p\left(T^{-1/2}\right) & \text{if} \quad 0 < \alpha < 1 \quad \text{and} \quad r > \frac{d}{\alpha}, \\ o_p\left(T^{-1/2}\right) & \text{if} \quad 1 \leq \alpha < 2 \quad \text{and} \quad r > \frac{d}{2-\alpha}, \end{cases}$$

*and in the latter case $T^{1/2}\left(\hat{\phi} - \phi_0\right) \Rightarrow N\left(0, I_f^{-1}\right)$, where $I_f$ is the information matrix, provided, given $g$, $L_T(\phi, g, 0)$ satisfies the regularity conditions (Smith (1985), Ch. 4.2.3 in Amemiya (1985)).*

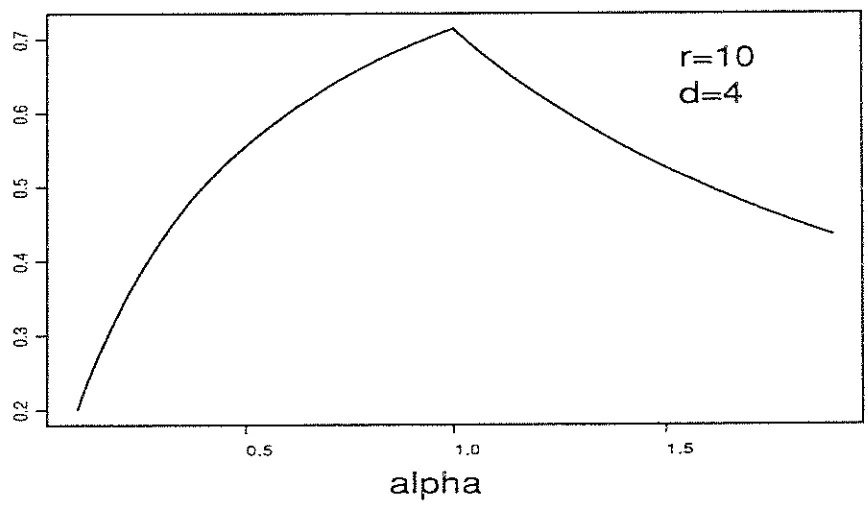


Figure 2: $O_p(T^{-x})$ difference between $\hat{\phi}$ and $\tilde{\phi}$, $x$ is on the vertical axis

The above result can be useful in applications as noted in Smith (1994). It should also be said that the results are quite analogous to those of Smith (1994). The substantial difference is that $g$ is modeled as an

[17]The "adjustment" by $a_T$ is for purely technical reasons. This is because both the MLE and $\hat{g}$ are assumed to be estimated using the same sample. If a sample splitting procedure is used, this ajustment is unnecessary.

[18]The estimation of the index tail function, $\alpha(x)$ is considered in Chernozhukov (1999c)

[19]Note the apparent circularity in the argument. In effect to operationalize this estimator, one need the estimate of $\alpha$. In practice one needs to first obtain a consistent estimate of $\alpha$, by first conduction MLE with the unoptimal bandwidth choice, satisfying assumption 3. For this we need only an assumption of what the range of values of $\alpha$ is. The estimate will be consistent, as follows from the proof, and will have at least a polynomial rate of convergence, thus not affecting the rates of covergence found here. This is discussed and explained in the proof. See Chernozhukov (1999c) for more on this.

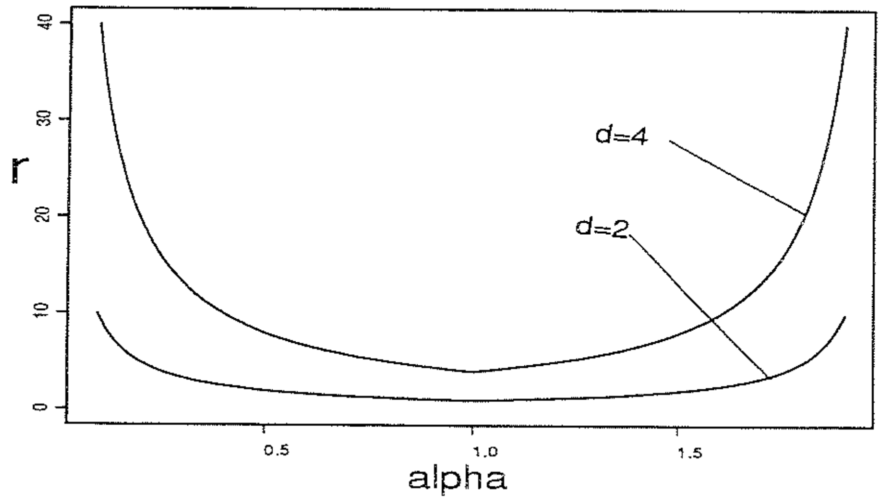


Figure 3: Order of smoothness needed to achieve $\sqrt{T}$ equivalence

unknown smooth function here while Smith models $g$ as a parametric function. Because of this, the results obtained specifically state what smoothness assumptions are required in order to assert the necessary rates of convergence. The intuition behind the results is clear: the smoother $g$ is, that is, the higher $r$ is, the higher rate of convergence can be achieved. In the limit, the rate of convergence approaches the parametric rate and one essentially can recover the results of Smith for the parametric case (the difference is by a $\ln T$ factor). Figures 2,3 demostrate intuition: 1) when $\alpha$ is smaller, rate of convergence of $\hat{g}$ to $g$ is higher, but since smaller $\alpha$ involves higher mass (of $\{u_t\}$) near zero, this causes larger difference between the likelihoods $L_T(\phi, \hat{g}, \cdot)$ and $L_T(\phi, g, 0)$, hence two forces operate in different directions; 2) on the other hand, when $\alpha$ is larger, the rate of convergence of $\hat{g}$ to $g$ is smaller, and at the same time the difference between the two likelihoods is smaller, hence these two factors again operate in different directions. A kind of balance is observed when $\alpha = 1$.

## 7.2 Truncated and Censored Data: Truncated Mean Estimation under Symmetry

These examples are an extension of the analysis in Powell (1986b). Consider the following latent variable model:

$$y_t^* = m(x_t, \alpha) + \epsilon_t$$

where $m$ is a regression function (can be a quantile, median, or mean function).
The general Type-I *censored model* is the case when the observed data $\{x_t, y_t\}$ consists of:

$$(x_t, y_t) = (x_t, y_t^*) \quad \text{if} \quad y_t^* > g(x_t)$$
$$(x_t, y_t) = (x_t, \text{"NA"}) \quad \text{if} \quad y_t^* \le g(x_t)$$

The general Type-I *truncated model* is when the observed data $\{x_t, y_t\}$ consists of:

$$(x_t, y_t) = (x_t, y_t^*) \quad \text{if} \quad y_t^* > g(x_t)$$

Assume that $\epsilon_t$ is symmetric around zero (for now). A subsequent example will mention the estimator where this restriction is relaxed for the case of the censored model (see Chernozhukov and Hong (1998b), Powell (1986a), Buchinsky (1995)). Analysis here is clearly an extension of Powell (1986b) in the sense that the censoring (truncating) line or function is no longer identically zero; it is modeled as an unknown smooth function $g(x)$. See Buchinsky and Hahn (1995) for an analysis of the censored model in such settings, yet from a different prospective.

The estimators are defined, by Powell, for these models as follows :
a) for the truncated model, $\hat{\beta} \in \operatorname{argmin} R_T(\beta, g)$, where:

$$R_T(\beta, g) \equiv \sum_{t=1}^{T} \left( y_t - \max\left\{ \frac{y_t + g(x_t)}{2}, m(x_t, \beta) \right\} \right)^2, \tag{7.2}$$

b) for the censored model: $\hat{\beta} \in \operatorname{argmin} S_T(\beta, g)$, where

$$\begin{aligned} S_T(\beta, g) \equiv & \sum_{t=1}^{T} \left( \tilde{y}_t - \max\left\{ \frac{\tilde{y}_t + g(x_t)}{2}, m(x_t, \beta) \right\} \right)^2 \\ & + \sum_{t=1}^{T} 1(\tilde{y}_t > 2m(x_t, \beta) - g(x_t)) \left[ \left( \frac{\tilde{y}_t + g(x_t)}{2} \right)^2 - max\{g(x_t), m(x_t, \beta)\}^2 \right]. \end{aligned} \tag{7.3}$$

where $\tilde{y}_t \equiv y_t$ if uncensored , $g(x_t)$ if censored . The motivation behind such objective functions is found by differentiation (e.g., left differentiation) and setting the first derivative equal to zero. For further discussion the reader may refer to Powell (1986b). The basic idea is to symmetrically trim (censor or truncate) observations on $y_t$ to consistently restore symmetry and then, using this sample, estimate the mean (median) function $m$. Since in (7.2), (7.3), $g$ is not known, we suggest to replace $\{g(x_t)\}$ by $\{\hat{g}(x_t)\}$, including in $\{\tilde{y}_t\}$. It is clear here that global estimate $\hat{g}$ is constructed on the basis of sample $\{y_t^*, x_t\}$. Clearly, data $\{y_t^*, x_t\}$ would conform to model 2 with assumptions 2, 9 (with $\alpha(x)$ equal 1) provided that the density of $y_t^*$ is uniformly positive at or near the boundary $g(x_t)$. Sampling assumptions on $\{x_t\}$ further guarantee consistency and smoothness assumptions on $g$ provides the necessary fast rate of convergence. See the tecnical report for this paper(Chernozhukov 1998) for proofs, but the idea is very simple: once $\hat{g}$ converges uniformly to $g$ at a sufficiently fast rate, the resulting asymptotics is that obtained by Powell. Let $\nabla m(x_t, \beta) \equiv \frac{\partial m(x_t, \beta)}{\partial \beta}(\beta)$.

**Assumption 11** (C1) *For $x' \in x$: $g(x') = g(x)$, $(x', x, \beta_0, y) \in X' \times X \times interior\Theta \times \mathbb{R} \subset \mathbb{R}^d \times \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \times \mathbb{R}$. $X' \subset X$, $X, \Theta$ are compact and convex. Fix $g \in \mathcal{G}$.*

(C2) *Assumptions 1 - 4a hold ($X' = C$), $g$ has order of smoothness $r > d = dim(x_t')$ on $X'$.*

(C3) *$\nu_T(\beta)$, the minimum characteristic root of*

$$N_T(\beta) \equiv T^{-1} \sum_{t=1}^{T} E1(m(x_t, \beta_0) > g(x_t) + \epsilon_0) \nabla m(x_t, \beta) \nabla m(x_t, \beta)'$$

*is s.t. uniformly in $\Theta$, $\nu_T(\beta) > \nu_0$, for $T > T_0$, for some $\epsilon_0, \nu_0, T_0 > 0$*

(C4) *$\{\epsilon_t\}$ are independent, and, conditionally on $\{x_t\}$, are continuously and symmetrically distributed around zero, with conditional (on $\{x_t\}$) densities $\{f_t\}$, which are bounded above and positive in a small neighborhood of zero and of $m(x_t, \beta_0) - g(x_t)$, uniformly in $t$.*

(C5) *Conditional densities $\{f_t\}$ are strictly unimodal, uniformly in $t$, in some neighborhood of zero (See Powell (1986b) for definition).*

(C6) *$m(x,\beta)$ is three times differentiable in $\beta, x$ with uniformly bounded over $X$ and $\Theta$ derivatives, and the specification of $m$ includes an intercept.*

**Theorem 12** *For the censored regression model, under assumptions (C1-C6), $\hat{\beta} \in argminS_T(\beta,\hat{g})$ is consistent and asymptotically normal, and $\sqrt{T}$-equivalent to $\bar{\beta} \in argminS_T(\beta, g)$ $\hat{\beta} - \bar{\beta} = o_p(1/\sqrt{T})$*

**Theorem 13** *For the truncated regression model, under assumptions (C1-C6) $\hat{\beta} \in argminR_T(\beta,\hat{g})$ is consistent, asymptotically normal, and $\sqrt{T}$ equivalent to $\bar{\beta} \in argminR_T(\beta, g)$: $\hat{\beta} - \bar{\beta} = o_p(1/\sqrt{T})$.*

## 7.3 Truncated Data: Truncated Regression Quantiles under Symmetry

Having obtained the estimates of the median as in the example on the truncated regression, it is possible to consistently estimate all truncated regression quantiles by "untruncation" of the data, using the symmetry restriction. Let $q(x,\beta_0)$ be the quantile function of $\{y_t^*\}$, and let $m(x_t,\alpha_0)$ be the median function of $\{y_t^*\}$. Let $\hat{\beta} \in$ argmin

$$V_T(\beta,\hat{g}) = \sum_{t=1}^{T} 1(m(x_t,\hat{\alpha}) > \hat{g}(x_t') + c)\{(\mathcal{D}_t(\hat{\alpha},\hat{g})\rho_\tau(2m(x_t,\hat{\alpha}) - y_t - q(x_t,\beta)) + \rho_\tau(y_t - q(x_t,\beta))\}$$

where $m(x_t,\hat{\alpha})$ is a preliminary estimate of conditional median, based e.g. on (7.2), and

$$\mathcal{D}_t(\hat{g}) = \begin{cases} 1 & \text{if} \quad y_t \geq 2m(x_t,\hat{\alpha}) - \hat{g}(x_t') \\ 0 & \text{if} \quad y_t < 2m(x_t,\hat{\alpha}) - \hat{g}(x_t') \end{cases}$$

literature. The idea is extremely simple – that is, to symmetrically "flip," around $m(x_t,\alpha_0)$, observations exceeding $2m(x_t,\alpha_0) - g(x_t)$, to consistently "untruncate" the data. Figure 5 demonstrates the intuition. The analysis of this model and the estimation techniques is found in Chernozhukov and Hong (1998c), where the details are also worked out for any $\sqrt{T}$ consistent first-stage estimator of parametric conditional median and alternative estimators are also considered.

Other alternative estimators and generalizations of this model are discussed in Chernozhukov and Hong (1998c). In this work, preliminary estimation of $g$ does not affect the first order asymptotics of the estimators due to faster than $\sqrt{T}$ rate of convergence (when $r > d$), however preliminary estimate of $m$ does affect it.

## 7.4 Conditional Quantile Estimators for Censored Model

Chernozhukov and Hong (1998c) analyzed the censored model and several different estimators of the censored regression quantiles, applying the results of the present paper. The asymptotic distribution of the estimators is not affected by the preliminary estimation of $\hat{g}$. Again, the proof uses the conditions guaranteeing sufficiently fast rate of convergence of $\hat{g}$ to $g$, in particular assumptions 9 ($\alpha(x) = 1$), 6 , 3a, 4a, and the smoothness assumption on $g$: $r > dim(x_t') = d$.

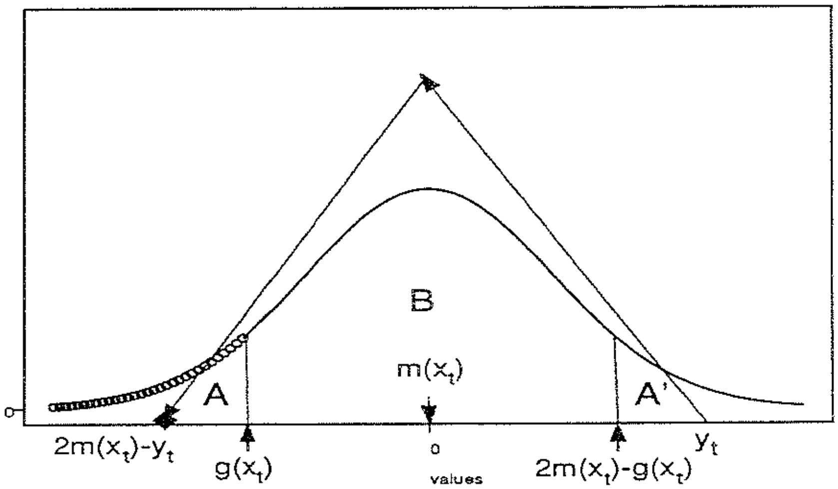


Figure 4: When $m(x_t) > g(x_t)$ the whole shape of conditional distribution is identified

# 8 Economic Applications

There are many theoretical and practical examples which call for modeling of conditional boundaries of support, that is, the extreme quantiles. Most notable examples are production frontier analysis, any economic sample selection models, and job serach models. See (Chernozhukov 1999a) for more discussion.

# 9 Conclusion

Nonparametric extreme regression quantiles are considered. Global and pointwise rates of convergence of the extreme quantile functions and their derivatives are characterized in a variety of settings. Some related sequential estimation procedures are considered, and applications are discussed. In our view, further interesting directions include the extension of the results to time series settings, specification tests of functional form and semi-parametric estimation of extreme quantile functions.

APPENDIX A

Proofs of Theorems and Lemmas

NOTATION. In what follows, $P \equiv \sup_{g\in\mathcal{G}} P_g$, $P_* \equiv \inf_{g\in\mathcal{G}} P_g$, that is for any $A \in \mathcal{F}$, $P(A) = \sup_{g\in\mathcal{G}} P_g(A)$, $P_*(A) = \inf_{g\in\mathcal{G}} P_g(A)$. $E_g$ is the expectation operator, indexed by $g$. Distribution functions, denoted by $F$, also depend on $g$, but this dependence is suppressed for brevity. $M_1, M_2, ...$ will denote some fixed positive constants, enumeration of which is renewed in every part of the appendix– thus enumeration in any part is independent of that in any other part. Constants $K_1, K_2, ...$ are those used in the statements of Theorems, and hence they are referred in this way throughout, whenever such reference is relevant. Term density means the usual Radon-Nykodim derivative with respect to the Lebesgue measure.

PROOF OF LEMMA 1. (omitted for brevity– see technical report)

PROOF OF THEOREM 1. To stress the dependence of $z_{x_t}$ on $x$ via its definition, write, in this proof only, $z_{xx_t}$, instead of $z_{x_t}$. Let $\bar{z}_{xT} \equiv \text{card}(N_{xT})^{-1} \sum_{t\in N_{Tx}} z_{xx_t}$,

$$\mathcal{L}_T \equiv \{w \in \Omega : \text{diag}\{\mathcal{X}_T\}\bar{z}_{xT} = \text{diag}\{\mathcal{X}_T\}\bar{z}\lambda_{\bar{z}},\ \text{for } \forall \bar{z} \in \times_{j=1}^{p} N_{h_T}^{j}(x), \lambda_{\bar{z}} \in \mathbb{R}^{p}_{++} : \lambda_{\bar{z}}'\mathbf{1} = 1, \lambda_{\bar{z}} > \epsilon\mathbf{1},\ \text{for fixed } \epsilon \in (0,1), \forall x \in \mathcal{C}\}$$

$$\mathcal{K}_T \equiv \{w \in \Omega : \text{card}\{t : z_{x_t} \in N_{h_T}^{j}(x)\} > M_1 T h_T^{d}, \forall x \in \mathcal{C}, \forall j \in \{\overline{1,p}\}\},\ \text{for some fixed } M_1 > 0,$$

Under assumption 4a, 3a, and by construction of $\{N_h^j(x), x \in \mathcal{C}\}$, $\lim_{T\to\infty} P_*\left(\mathcal{L}_T \cap \mathcal{K}_T\right) = 1$. With 4b, 3b, $P_*\left(\mathcal{L}_T \cap \mathcal{K}_T,\ \text{ev.}\right) = 1$.

Therefore, what is considered below is conditional on the event $\mathcal{L}_T \cap \mathcal{K}_T$. Consider the $p$ minimum order statistics, $\{U_{(1)jh_T}(x)\}$ defined as before. Let $l_x$ be the index:

$$l_x \equiv \{t : u_t = U_{(1)jh_T} = \min\{u_t : z_{xx_t} \in N_{h_T}^{j}(x), \quad j \in \{\overline{1,p}\}\}$$

Denote

$$\mathcal{U}_{(1)}(x) \equiv u(l_x) \equiv \text{vech}\{u_t : t \in l_x\},$$

$$z(l_x) \equiv \{z_{xx_t} : t \in l_x\},\ \text{so that } z_{xx_t} \text{ are column vectors in } z_x(l_x)$$

$$\mathcal{R}(l_x) \equiv \text{vech}\{\sum_{m\in\mathbb{Z}^d:[m]=k} m!\left(D^m g(x_t^*) - D^m g(x)\right)(x_t - x)^m : \quad t \in l_x\}$$

where $x_t^*$ is s.t. $\|x_t^* - x\|_\infty \leq h_T$, so that $\forall t \in l_x$, (using assumption 1), $g(x_t) = z_{xx_t}'\beta_x + \mathcal{R}(t)$. Then for the estimator, $\hat{\beta}_x$, defined in section 2, $z(l_x)'\hat{\beta}_x \leq y(l_x) = z(l_x)'\beta_x + \mathcal{R}(l_x) + \mathcal{U}_{(1)}(x)$.

By assumption 1, there exists $M_2', M_2 > 0$, s.t. for $t \in l_x$

$$\sup_{x\in\mathcal{C}} \sup_{t\in l_x} |\mathcal{R}(t)| \leq M_2' \left(\sup_{t\in l_x} \|x_t - x\|_\infty^k h_T^\gamma\right) \leq M_2\left(h_T^r\right) \tag{.1}$$

From the above, $z(l_x)'(\hat{\beta}_x - \beta_x) \leq \mathcal{R}(l_x) + \mathcal{U}_{(1)}(x)$. Denote $\hat{\theta}_x = (\hat{\beta}_x - \beta_x)$, hence

$$z(l_x)'(\hat{\theta}_x) \leq \mathcal{R}(l_x) + \mathcal{U}_{(1)}(x) \tag{.2}$$

This establishes an upper bound. To find a lower bound, recall that by construction of $\{N_{h_T}^j(x), j = \overline{1,p}\}$ (and conditioning on the event $\mathcal{L}_T \cap \mathcal{K}_T$ ), $\forall x \in \mathcal{C}$,

$$\bar{z}_{xT} = \lambda_{z'}' z',\ \text{for } \forall z' \in \times_{j=1}^{p} N_{h_T}^{j}(x), \quad \text{where } \lambda_{z'} \in \mathbb{R}^{p}_{++}, \lambda_{z'}\mathbf{1} = 1, \lambda_{z'} >> \epsilon\mathbf{1},$$

for appropriate unit vector $\mathbf{1}$, and fixed $\epsilon$. By definition of $\hat{\theta}_x$, conditioning on the event $\mathcal{K}_T \cap \mathcal{L}_T$, by Theorem 1 in Bassett (1988) and assumption 1, $\exists M_3 > 0$:

$$\bar{z}'_{xT}\hat{\theta}_x > 0 - M_3 h_T^r \tag{.3}$$

This implies (for $w \in \mathbb{R}^p$, let $w^+ = (w_1 \bigvee 0, ..., w_p \bigvee 0)'$, and define $w^-$, $|w|$ accordingly), $\forall x \in \mathcal{C}$,

$$\lambda'_{z(l_x)} \left(z(l_x)'\theta\right)^+ > \left| \left(\lambda_{z(l_x)}\right)_j \cdot \left(z(l_x)'\theta\right)_j \right| - M_3 \left(h_T^r\right), \quad \forall j \in \overline{1,p} \Longrightarrow$$

$$p \bigvee_{j=1}^{p} \left(\lambda_{z(l_x)}\right)_j \bigvee_{j=1}^{p} \left(z(l_x)'\theta\right)_j^+ > \bigwedge_{j=1}^{p} \left(\lambda_{z(l_x)}\right)_j \bigvee_{j=1}^{p} \left(\left|(z(l_x)'\theta\right|\right)_j - M_3\left(h_T^r\right), \Longrightarrow$$

$$\frac{p(1-\epsilon)}{\epsilon} \bigvee_{j=1}^{p} \left(z(l_x)'\theta\right)_j^+ > \bigvee_{j=1}^{p} \left(\left|z(l_x)'\theta\right|\right)_j - M_3\left(h_T^r\right) \quad (*)$$

But from (.2), $\bigvee_{j=1}^{p} \left(z(l_x)'\theta\right)_j^+ \leq \bigvee_{t \in l_x} \left(u(t) + \left|\mathcal{R}(t)\right|\right)$ Then (*), (.2), (.1), (.3) imply, for $M_4 = p(1-\epsilon)/\epsilon$, $M_5 = M_2 M_4 + M_3$, $\forall x \in \mathcal{C}$,

$$\left\|z(l_x)'\hat{\theta}_x\right\|_\infty \leq \left(M_4 \bigvee_{t \in l_x} (u(t)) + M_5\left(h_T^r\right)\right) \equiv \left(M_4 \max_j U_{(1)jh_T}(x) + M_5 h_T^r\right), \Longrightarrow$$

$$\hat{\theta}'_x z(l_x) z(l_x)'\hat{\theta}_x \leq \left(M_4 \max_j U_{(1)jh_T}(x) + M_5 h_T^r\right)^2 p$$

Define $\tilde{\theta}_x \equiv \left(diag\{\mathcal{X}_T\}\right)^{-1}\hat{\theta}_x$, $\tilde{Q}_T \equiv \text{diag}\{\mathcal{X}_T\}' z(l_x) z(l_x)' \text{diag}\{\mathcal{X}_T\}$, then $\hat{\theta}'_x z(l_x) z(l_x)'\hat{\theta}_x = \tilde{\theta}'_x \tilde{Q}_T \tilde{\theta}_x$. By a matrix inequality (p.460 in Amemiya (1985)): $\tilde{\theta}'_x\tilde{\theta}_x \leq \delta_T^{-1}(\tilde{Q}_T)\left(\tilde{\theta}'_x\tilde{Q}_T\tilde{\theta}_x\right)$, where $\delta_T(\tilde{Q}_T)$ is the smallest characteristic root of $\tilde{Q}_T$. By construction of $\{N_{h_T}^j(x), j = \overline{1,p}\}$, $\delta_T(\tilde{Q}_T)$ is bounded away from zero and from above, uniformly in $T > T_0$, for some $T_0 > 0$. Therefore $\exists M_6, M_7 > 0$, s.t for any $m \in \mathbb{Z}_+^d : [m] \leq k$,

$$\left(\hat{\theta}_{xm}\right)^2 = (m!)^2 \left(\widehat{D^m g(x)} - D^m g(x)\right)^2 \leq \left(\left(\max_j U_{(1)jh_T}(x) + M_7 h_T^r\right)\left(M_6 h_T^{-[m]}\right)\right)^2$$

The above bound is conditional on the event $\mathcal{L}_T \cap \mathcal{K}_T$, thus the conclusion follows.

■

PROOF OF LEMMA 2. Let $\mathcal{A}_T \equiv \{w \in \Omega : U_{(1)jh_T}(x) > \epsilon\}$, $N_{xjT} \equiv \text{card}\{z_{x_t} \in N_{h_T}^j(x)\}$, $\mathcal{B}_T \equiv \{w \in \Omega : N_{xjT} > T^{a'}\}$.

1. To show $\lim_T P(\mathcal{A}_T) = 0$ under assumptions 2, 3a, 4a.

$$\lim_T P(\mathcal{A}_T) \leq \lim_T P(\mathcal{A}_T \cap \mathcal{B}_T) + \lim_T P(\mathcal{B}_T^c) \leq \lim_T P(\mathcal{A}_T \cap \mathcal{B}_T) \leq \lim_T (1-\delta_\epsilon)^{T^{a'}} = 0$$

where the second inequality follows since $\lim_T P(\mathcal{B}_T^c) = 0$ by assumptions 4, 3.

2. To show $P(\limsup_T \mathcal{A}_T) = 0$, given assumptions 2, 3b, 4b.

$$P(\limsup \mathcal{A}_T) \leq P(\limsup\{\mathcal{A}_T \cap \mathcal{B}_T\}) + P(\limsup\{\mathcal{B}_T^c\}) \leq P(\limsup\{\mathcal{A}_T \cap \mathcal{B}_T\})$$

where the second inequality follows since $P(\limsup_T \mathcal{B}_T^c) = 0$ by assumptions 4b, 3b. Then

$$\sum_T P(\mathcal{A}_T \cap \mathcal{B}_T) \leq \sum_T (1-\delta_\epsilon)^{T^{a'}} \sim \ln(1-\delta_\epsilon)^{-\frac{1}{a}}\Gamma\left((a')^{-1}\right) < \infty,$$

implies by the Borel-Cantelli Lemma $P(\limsup\{\mathcal{A}_T \cap \mathcal{B}_T\}) = 0$.

■

PROOF OF THEOREM 2 . The Theorem is a direct consequence of the Slutsky Theorem, Lemma 2, and assumption 3a for convergence in probability and 3b for convergence a.s.

■

PROOF OF LEMMA 3. See Gnedenko (1943)

■

PROOF OF LEMMA 4. It suffices to show the result only for any fixed $j$, since the maximum is taken over $p$ (finitely many) elements. To show, for any $\epsilon > 0$, $\exists\delta$ sufficiently large s.t.

$$\limsup_{T} P\left(U_{(1)jh_T}(x) > \delta\tilde{\gamma}_{Th_T^d}(x,h_T)^{-1}\right) < \epsilon \tag{.4}$$

Take $\delta > 1$, for concreteness, and let $N_{Tjx} \equiv \{t : z_{x_t} \in N^j_{h_T}(x)\}$, $N_{xjT} \equiv \text{card}\{N_{Tjx}\}$, $\mathcal{D}_T = \{w \in \Omega : h_T = \bar{h}_T, N_{xjT} = n\} \in \mathcal{F}$, $\mathcal{E}_T \equiv \{w \in \Omega : Th_T^d > T^{a'}, N_{xjT} > M_0 Th_T^d\}$, for $M_0 > 0$ fixed and sufficiently small so that by assumptions 3, 4 $\lim_T P(\mathcal{E}_T^c) = 0$.

$$\begin{aligned}
P\Big(U_{(1)jh_T}(x) &> \delta\tilde{\gamma}_{Th_T^d}(x,h_T)^{-1}|\mathcal{D}_T \cap \mathcal{E}_T\Big) \\
&\le \sup_{g\in\mathcal{G}} \prod_{t\in N_{Tjx}} \left(1 - F\left(\delta F^{-1}\left((T\tilde{h}_T^d)^{-1}\Big|x_t\right)\Big|x_t\right)\right) \\
&\le \sup_{g\in\mathcal{G}} \exp\left\{\sum_{t\in N_{Tjx}} \ln\left(1 - F\left(\delta F^{-1}\left((T\tilde{h}_T^d)^{-1}\Big|x_t\right)\Big|x_t\right)\right)\right\} \\
&\le \sup_{g\in\mathcal{G}} \exp\left\{-M_1 \sum_{t\in N_{Tjx}} F\left(\delta F^{-1}\left((T\tilde{h}_T^d)^{-1}\Big|x_t\right)\Big|x_t\right)\right\} \\
&\le \sup_{g\in\mathcal{G}} \exp\left\{-M_1 \sum_{t\in N_{Tjx}} \left[\frac{F\left(\delta F^{-1}\left((T\tilde{h}_T^d)^{-1}\Big|x_t\right)\Big|x_t\right)}{F\left(F^{-1}\left((T\tilde{h}_T^d)^{-1}\Big|x_t\right)\Big|x_t\right)}\right] F\left(F^{-1}\left((T\tilde{h}_T^d)^{-1}\Big|x_t\right)\Big|x_t\right)\right\} \\
&\le \exp\left\{-M_1 n \inf_{g\in\mathcal{G}} \text{essinf}_{x\in\mathcal{C}} \left(\left[\frac{F\left(\delta F^{-1}\left((T\tilde{h}_T^d)^{-1}\Big|x\right)\Big|x\right)}{F\left(F^{-1}\left((T\tilde{h}_T^d)^{-1}\Big|x\right)\Big|x\right)}\right] F\left(F^{-1}\left((T\tilde{h}_T^d)^{-1}\Big|x\right)\Big|x\right)\right)\right\} \\
&\le \exp\left\{-M_2\left[\delta^{\underline{\alpha}}\right] n \inf_{g\in\mathcal{G}} \text{essinf}_{x\in\mathcal{C}} F\left(F^{-1}\left((T\tilde{h}_T^d)^{-1}\Big|x\right)\Big|x\right)\right\} \\
&\le \exp\left\{-M_3\left[\delta^{\underline{\alpha}}\right] n \inf_{g\in\mathcal{G}} \text{essinf}_{x\in\mathcal{C}} F\left(F^{-1}\left(n^{-1}\Big|x\right)\Big|x\right)\right\} \\
&\le \exp\left\{-M_4\ \delta^{\underline{\alpha}}\right\}
\end{aligned}$$

where all of the inequalities hold for sufficiently large $T\tilde{h}_T^d, n$, for some $M_1, M_2, M_3, M_4 > 0$. The first and second inequalities are clear, the third inequality holds because $\ln(1-z) \sim -z$, as $z \to 0$, and since by assumption 7 and Lemma 6, for some $M_5 > 0$

$$\sup_{g\in\mathcal{G}} \inf_{x\in\mathcal{C}} F^{-1}(y|x) \le y^{M_5} \to 0, \text{ as } y \to 0^+; \tag{.5}$$

the fourth inequality and fifth inequalities are clear as well; the sixth inequality holds since by the regular variation property (assumption 7) and Lemma 6,

$$\liminf_{y\to\infty} \inf_{g\in\mathcal{G}} \text{essinf}_{x\in\mathcal{C}} \left[\frac{F\left(\delta F^{-1}\left(y^{-1}\Big|x\right)\Big|x\right)}{F\left(F^{-1}\left(y^{-1}\Big|x\right)\Big|x\right)}\right] \ge \delta^{\underline{\alpha}} \tag{.6}$$

The seventh inequality is by conditioning on $\mathcal{D} \cap \mathcal{E}_T$, and the last is due to assumption 7. The obtained inequality implies that

$$\limsup_{T} P\left(\left\{U_{(1)jh_T}(x) > \delta\tilde{\gamma}_{Th_T^d}(x,h_T)^{-1}\right\} \cap \mathcal{E}_T\right) \le \exp\left\{-M_4\ \delta^{\underline{\alpha}}\right\}$$

which together with $\limsup_T P(\mathcal{E}_T^c) = 0$ (due to assumptions 3, 4) yields the result.

■

PROOF OF THEOREM 3. The Theorem is a direct consequence of the Slutsky Theorem, Theorem 1, and Lemma 4.

■

LEMMAS 5 AND 6.

Two distribution functions are said here to be tail equivalent if they have the same left endpoint 0 and for some $A > 0$,

$$\lim_{u \searrow 0} (1 - F(u)) / (1 - G(u)) = A$$

**Lemma 5** EQUIVALENCE CLASSES. *Let $F$ and $G$ be the distribution functions and suppose $W$ is a Weibull distribution function. Let $H_1 = W_\alpha$ suppose that $F \in \mathcal{D}(H_1)$, and that*

$$1 - (1 - F(\gamma_{\mathcal{N}} x))^{\mathcal{N}} \Rightarrow H_1(x), \quad \text{as } \mathcal{N} \to \infty$$

*for normalizing constants $\{\gamma_{\mathcal{N}}\}$, then*

$$1 - (1 - G(\gamma_{\mathcal{N}} x))^{\mathcal{N}} \Rightarrow H_2(x), \quad \text{as } \mathcal{N} \to \infty$$

iff for some $a > 0$, $H_2(u) = H_1(au)$ and $F$ and $G$ are tail equivalent in the above sense.

PROOF. This is a corollary of Proposition 1.19 in Resnick (1987).

■

**Lemma 6** *Let $F$ be a regularly varying distribution function, $F \in \mathcal{RV}_\rho$, then for any small fixed $\epsilon > 0$, as $y \to 0^+$, $v \to 0^+$,*

$$F(y)/y^{\rho+\epsilon} \longrightarrow \infty, \; F(y)/y^{\rho-\epsilon} \longrightarrow 0, \; F^{-1}(v)/v^{1/(\rho+\epsilon)} \longrightarrow 0, \; F^{-1}(v)/v^{1/(\rho-\epsilon)} \longrightarrow \infty.$$

PROOF. The first two results follow by noting that $F(u)/u^\rho$ is slowly varying, and the other two – from the first two by noting that for sufficiently small $v$, and $0 < \epsilon' < \epsilon$, $v^{1/(\rho-\epsilon')} \ll F^{-1}(v) \ll v^{1/(\rho+\epsilon')}$, by definition of $F^{-1}$ and monotonicity of $F$.

■

PROOF OF THEOREM 4. This is a corollary of the previous Theorem.

■

PROOF OF THEOREM 5. Proceed as in the proof of Lemma 4, defining the same notation (and letting $\alpha \equiv \alpha(x)$), for large $T\bar{h}_T^d$

$$P\Big(U_{(1)jh_T}(x) > \delta(T\bar{h}_T^d)^{-1/\alpha} | \mathcal{D}_T \cap \mathcal{E}_T\Big) \le \sup_{g \in \mathcal{G}} \prod_{t \in N_{Tjx}} \Big(1 - F\Big(\delta(T\bar{h}_T^d)^{-1/\alpha} \Big| x_t\Big)\Big)$$

$$\le \sup_{g \in \mathcal{G}} \prod_{t \in N_{Tjx}} \Big(1 - \Big(\delta(T\bar{h}_T^d)^{-1/\alpha}\Big)^\alpha\Big) \le \exp\Big\{ - M_4 \; \delta^\alpha \Big\}$$

Conclude as in Lemma 4 and then as in Theorem 3.

■

PROOF OF THEOREM 6. This proof is notationally the same as the proof of Lemma 2, except that we should replace $\delta_\epsilon$ with $\delta$ and redefine $\mathcal{A}_T$ as $\{w \in \Omega : U_{(1)xjT} \neq 0\}$. Conclusion is entirely trivial.

■

**Lemma 7** UNIFORM CONVERGENCE OF LOCAL EMPIRICAL DISTRIBUTION FUNCTION. *Fix $x \in C$. Let $F_T^l$ be the local distribution function of $\{\tilde{u}_{xT}\}$, i.e. $F_T^l(\xi) \equiv card(N_{Tx})^{-1} \sum_{t \in N_{Tx}} 1(\tilde{u}_{xT} \le \xi)$, then under assumption 5, for $B_T = [T^{q'}, T^q]$, for fixed $q, q' < 0$.*

$$\lim_{T \to \infty} \sup_{x \in C, h_T \in B_T, \xi \in \Delta} \left| F_T^l(\xi) - F_x(\xi) \right| = 0, \quad a.s.$$

*where $F_{\cdot}$ is differentiable (has density) in $\xi_2, ...\xi_p$, and furthermore:*

$$\lim_{T\to\infty} \sup_{x\in C, h_T \in B_T} \left| card(N_{Tx})/Th_T^d - w_x(x)\right| = 0, \quad a.s.$$

Proof. Omitted for breivity (all that's omitted is avaliable in the technical report (Chernozhukov 1998)

**Lemma 8** BOUNDS ON THE MOMENT BEHAVIOR *Fix $x \in \mathcal{C}$, let $\bar{U}_{(1)jh_T}(x) \equiv U_{(1)jh_T}(x) \wedge \mathcal{K}$, for $\mathcal{K} > 0$ fixed, $\mathcal{A}_T \equiv \{w \in \Omega : h_T \in [T^{-k}, T^{-k'}\}$, for $0 < k' < k$, s.t. assum. (3), then*

$$E\left(\bar{U}_{(1)jh_T}(x)|\mathcal{A}_T\right) = O\left(\Upsilon_T(h_T, x)\right), \text{ and } \lim_T P\left(\{\sup_{x\in C_T} U_{(1)jh_T}(x) > \mathcal{K}\} \cup \mathcal{A}_T{}^c\right) = 0.$$

*where $\Upsilon_T(h_T, x)$ is defined as $\Upsilon_T(h_T)$, except that $\int_C(\cdot)dx$ is replaced with $\int_{\bar{N}_{h_T}(x)}(\cdot)dx$, and $E(\cdot) \equiv \sup_{g\in G} E_g(\cdot)$*

Proof. The second part is a direct consequence of Theorem 10 (proof of which does not rely on this Theorem) and assumption 3. Consider the first part. Let $\Psi_T \equiv \Upsilon_T(h_T, x)^{-1}\bar{U}_{(1)jh_T}(x)$. Start with models 1-3. To show the above, it suffices to show (see Pickands (1968)) the asymptotic uniform integrability of $\{\Psi_T\}$:

$$\lim_{L\to\infty} \text{limsup}_T E\left[\Psi_T^q 1(\Psi_T > L)|\mathcal{A}_T\right] = 0$$

For brevity, we suppress "$\mathcal{A}_T$" in notation. Then

$$E\Psi_T^q 1(\Psi_T > L) \le E\int_0^{\Psi_T} qs^{q-1}ds1(\Psi_T > L) = E\int_0^L qs^{q-1}ds1(\Psi_T > L) + E\int_L^{\mathcal{K}/\Upsilon_T(h_T,x)} qs^{q-1}1(\Psi_T > s)ds$$
$$\le L^q P(\Psi_T > L) + E\int_L^{\mathcal{K}/\Upsilon_T(h_T,x)} qs^{q-1}P(\Psi_T > s)ds \equiv M(L,T) + N(L,T)$$

$$N(L,T) = E\int_L^{\mathcal{K}/(\Upsilon_T(h_T,x))^\mu} qs^{q-1}P(\Psi_T > s)ds + E\int_{\mathcal{K}/(\Upsilon_T(h_T,x))^\mu}^{\mathcal{K}/\Upsilon_T(h_T,x)} qs^{q-1}P(\Psi_T > s)ds \equiv N_1(L,T) + N_2(L,T)$$

for $\mu < 1$. Then,

$$\lim_{L\to\infty} \limsup_T M(L,T) = \lim_{L\to\infty} \exp\{M_1 \ln L - L^\rho\} = 0$$

for some $\rho > 0$ ($\rho = \underline{\alpha}$ in model 1, $\rho = \alpha$ in model 3), which follows from Lemmas 4 and 5. Next consider $N_1(L,T)$:

$$\lim_{L\to\infty} \limsup_T N_1(L,T) \le \lim_{L\to\infty} \exp\{M_2 \ln L - M_3 L^{\rho'}\} = 0$$

for models 1-3 (some $\rho' > 0$ ). The proof of this inequality is very similar to the proofs of Lemma 4 and 5, and hence we omit it. Then, consider $N_2(L,T)$: by assumptions 3, 4 and 6, for a fixed $0 < \delta < 1$,

$$\lim_{L\to\infty} \limsup_T N_2(L,T) \le \lim_{L\to\infty} \limsup_T EM_4 \left(\mathcal{K}/\Upsilon_T(h_T,x)\right)^{q-1}(1-\delta)^{T^{M_5}}$$
$$\le \lim_{L\to\infty} \limsup_T EM_4 \left(\mathcal{K}/\Upsilon_T(h_T,x)\right)^{q-1} \exp\{-T^{M_5}M_6\} = 0.$$

The last conclusion follows, since by Lemma 6 and conditioning on $\mathcal{A}_T$, $T^{-b} \ll \Upsilon_T(h_T, x) \ll T^{-b'}$, for some fixed $b, b' > 0$. Next consider model 4. By assumption 3, 4,

$$E\bar{U}_{(1)jh_T}(x) = E\bar{U}_{(1)jh_T}(x)1\left(\bar{U}_{(1)jh_T}(x) > 0\right) \le \mathcal{K}(1-\delta)^{T^a} \to 0$$

Lemma is proved.

PROOF OF THEOREM 9 . By construction of interpolation procedure:

$$\int_C \left|\overline{D^m g(x)} - D^m g(x)\right|^q dx = O_p\left(\|\widehat{D^m g(x)} - D^m g(x)\|_{q,C_T}^q\right) = O_p\left(\left(\left(\Upsilon_T(h_T) + h_T^r\right)h_T^{-[m]}\right)^q\right)$$

where the second inequality follows by Chebyshev inequality and Lemma 8.

■

**Lemma 9** ASYMPTOTIC EXPONENTIAL INEQUALITIES FOR EXTREME ORDER STATISTICS. *Under the assumptions of Theorem 10, for the grid $C_T$, for models 1-3:*

$$\limsup_T P\Big(max_{x\in C_T} U_{(1)jh}(x) > \delta\Upsilon_T(h_T)\Big) \leq \limsup_T exp\Big\{-M_0\,(\delta^\rho - b)\,lnT\Big\}$$

*for some constants $\rho, M_0 > 0$, independent of $\delta$, where $\Upsilon_T(h_T)$ is stated in the Theorem 10 for each of models 1-3, $\delta > 1$, $\rho = \underline{\alpha}$ for model 1 (and 2),and $\rho = \alpha$ for model 3. For model 4,*

$$\limsup_T P\Big(max_{x\in C_T} U_{(1)jh}(x) > 0\Big) = 0.$$

Proof. The proof closely follows one of Lemma 4. Take $\delta > 1$, for concreteness, and let $N_{Txj} \equiv \{t : x_t \in N^j_{h_T}(x)\}$, $N_{xjT} \equiv \text{card}\{N_{Txj}\}$, $\mathcal{D}_T = \{w \in \Omega : h_T = \bar{h}_T, N_{xjT} = n_x, \forall x \in C_T\} \in \mathcal{F}$, $\mathcal{E}_T \equiv \{w \in \Omega : T^{m''} > Th^d_T > T^{m'}, N_{xjT} > M_2 Th^d_T, \forall x \in C_T\}$, for $M_2 > 0$ fixed and sufficiently small, $m'', m' > 0$ so that by assumptions 3, 4 $\lim_T P(\mathcal{E}^c_T) = 0$. Consider at first the set of assumptions corresponding to model 1 (and model 2 by implication).

$$\begin{aligned}
&P\Big(\max_{x\in C_T} U_{(1)jh_T}(x) > \delta\Upsilon_T(h_T)|\mathcal{D}_T\cap\mathcal{E}_T\Big)\\
&\leq \sup_{g\in\mathcal{G}} \sum_{x\in C_T}\prod_{t\in N_{Txj}}\Big(1 - F\Big(\delta F^{-1}\Big(\ln T(T\bar{h}^d_T)^{-1}\Big|x_t\Big)\Big|x_t\Big)\Big)\\
&\leq \sup_{x\in C_T, g\in\mathcal{G}} \text{card}(C_T)\exp\Big\{\sum_{t\in N_{Txj}}\ln\Big(1 - F\Big(\delta F^{-1}\Big(\ln T(T\bar{h}^d_T)^{-1}\Big|x_t\Big)\Big|x_t\Big)\Big)\Big\}\\
&\leq \sup_{x\in C_T, g\in\mathcal{G}} \text{card}(C_T)\exp\Big\{-M_1\sum_{t\in N_{Txj}} F\Big(\delta F^{-1}\Big(\ln T(T\bar{h}^d_T)^{-1}\Big|x_t\Big)\Big|x_t\Big)\Big\}\\
&\leq \sup_{x\in C_T, g\in\mathcal{G}} \text{card}(C_T)\exp\Big\{-M_1\sum_{t\in N_{Txj}}\left[\frac{F\Big(\delta F^{-1}\Big(\ln T(T\bar{h}^d_T)^{-1}\Big|x_t\Big)\Big|x_t\Big)}{F\Big(F^{-1}\Big(\ln T(T\bar{h}^d_T)^{-1}\Big|x_t\Big)\Big|x_t\Big)}\right] F\Big(F^{-1}\Big(\ln T(T\bar{h}^d_T)^{-1}\Big|x_t\Big)\Big|x_t\Big)\Big\}\\
&\leq \sup_{x'\in C_T, g\in\mathcal{G}} \text{card}(C_T)\exp\Big\{-M_1 n_{x'}\inf_{g\in\mathcal{G}}\text{essinf}_{x\in C}\left(\left[\frac{F\Big(\delta F^{-1}\Big(\ln T(T\bar{h}^d_T)^{-1}\Big|x\Big)\Big|x\Big)}{F\Big(F^{-1}\Big(\ln T(T\bar{h}^d_T)^{-1}\Big|x\Big)\Big|x\Big)}\right] F\Big(F^{-1}\Big(\ln T(T\bar{h}^d_T)^{-1}\Big|x\Big)\Big|x\Big)\right)\Big\}\\
&\leq \sup_{x'\in C_T, g\in\mathcal{G}} \text{card}(C_T)\exp\Big\{-M_2\Big[\delta^{\underline{\alpha}}\Big]\inf_{g\in\mathcal{G}}\text{essinf}_{x\in C} n_{x'} F\Big(F^{-1}\Big(\ln T(T\bar{h}^d_T)^{-1}\Big|x\Big)\Big|x\Big)\Big\}\\
&\leq \sup_{x'\in C_T, g\in\mathcal{G}} \text{card}(C_T)\exp\Big\{-M_3\Big[\delta^{\underline{\alpha}}\Big]\inf_{g\in\mathcal{G}}\text{essinf}_{x\in C} n_{x'} F\Big(F^{-1}\Big(\ln T(n_{x'})^{-1}\Big|x\Big)\Big|x\Big)\Big\}\\
&\leq \sup_{g\in\mathcal{G}} \text{card}(C_T)\exp\Big\{-M_4\,\delta^{\underline{\alpha}}\ln T\Big\}
\end{aligned}$$

where all of the inequalities hold for sufficiently large $T\bar{h}_T, n$, for some $M_1, M_2, M_3, M_4 > 0$. All of inequalities are justified exactly as in the proof of Lemma 4.

Conditionally on $\mathcal{E}_T$ and from Lemma 6, $\sup_{g\in\mathcal{G}}\text{card}(C_T) = O(T^b)$, for some $b > 0$. This and the above inequality imply:

$$\limsup_T P\Big(\Big\{U_{(1)jh_T}(x) > \delta\Upsilon_T(h_T)\Big\}\cap\mathcal{E}_T\Big) \leq \limsup_T \exp\Big\{-M_5\ \ (\delta^{\underline{\alpha}} - b)\,lnT\Big\}$$

which together with $\limsup_T P(\mathcal{E}^c_T) = 0$ (due to assumptions 3, 4) yields the result.

Consider Model 3 next. Proceed as above, defining the same notation, then for sufficiently large $T$, $P\Big(\sup_{x\in C_T} U_{(1)jh_T}(x) > \delta(T\bar{h}^d_T)^{-1/\alpha}|\mathcal{D}_T\cap\mathcal{E}_T\Big) \leq$

$$\begin{aligned}
&\leq \sup_{g\in\mathcal{G}}\sum_{x\in C_T}\prod_{t\in N_{Txj}}\Big(1 - F\Big(\delta((\ln T)^{-1}T\bar{h}^d_T)^{-1/\alpha}\Big|x_t\Big)\Big) \leq \sup_{g\in\mathcal{G}}\sum_{x\in C_T}\prod_{t\in N_{Txj}}\Big(1 - \Big(\delta((\ln T)^{-1}T\bar{h}^d_T)^{-1/\alpha}\Big)^\alpha\Big)\\
&\leq \sup_{g\in\mathcal{G}}\sum_{x\in C_T}\Big(1 - \Big(\delta((\ln T)^{-1}n_x M_6)^{-1/\alpha}\Big)^\alpha\Big)^{n_x} \leq \sup_{g\in\mathcal{G}} \text{card}(C_T)\exp\Big\{-M_7\ \delta^\alpha\Big\}
\end{aligned}$$

Then conclude as above.

Consider Model 4. $P\Big(\sup_{x\in C_T} U_{(1)jh_T}(x) > 0 | \mathcal{D}_T \cap \mathcal{E}_T\Big) \leq \text{card}(C_T)(1-\delta)^{M_8 T h_T^d} \leq T^{M_8}(1-\delta)^{T^{M_{10}}}$. Hence $\limsup_T P\Big(\{\sup_{x\in C_T} U_{(1)jh_T}(x) > 0\} \cap \mathcal{E}_T\Big) \leq \limsup_T T^{M_7}(1-\delta)^{T^{M_8}} = 0$, and since $\limsup_T P(\mathcal{E}_T^c) = 0$ (due to assumptions 3, 4) the result is obtained.

■

PROOF OF THEOREM 10.

To prove Theorem 10, we use Lemma 9, stated before the Theorem. By construction of interpolation procedure:

$$\left\|\widehat{D^m g(x)} - D^m g(x)\right\|_\infty = O_p\left(\left\|\widehat{D^m g(x)} - D^m g(x)\right\|_{\infty, C_T}\right) = O_p\left(\left(\left(\Upsilon_T(h_T) + h_T^r\right) h_T^{-[m]}\right)\right)$$

where the second equality follows from Lemma 9.

■

DISCUSSION AND PROOF OF THEOREM 11

1. Discussion and definitions. Here we give the definitions of the "pilot" ML estimator and define some notation.

By compactness, $\alpha \in [\underline{\alpha}, \overline{\alpha}]$. Hence, let $h_T^1 \sim_p (T/\ln T)^{\frac{-1}{r\overline{\alpha}+d}}$, $a_T^1 \sim_p \left(T/(\ln T)^{1+\epsilon}\right)^{\frac{-r}{r\overline{\alpha}+d}}$. By Theorem 10, $\bar{g} \equiv g^1$, the global estimator under such $\{h_T^1\}$, is s.t. $\|\bar{g} - g\|_\infty = O_p\left((T/\ln T)^{\frac{-r}{r\overline{\alpha}+d}}\right)$. Then let

$$\bar{\phi} \in \text{argmax}_\phi L_T(\phi, \bar{g}, a_T^1)$$

$\bar{\phi}$ is said to be *pilot* estimator here. Therefore the estimator defined in the statement of the Theorem is:

$$\hat{\phi} \in \text{argmax}_\phi L_T(\phi, \hat{g}, a_T^2)$$

where $\hat{g} \equiv g^2$ is the global estimator defined with the (optimal) bandwidth sequence $h_T^2 \sim_p (T/\ln T)^{\frac{-r}{r\hat{\alpha}+d}}$, where $\hat{\alpha} \equiv \alpha(\bar{\phi})$. As we shall prove, $\hat{\alpha} - \alpha = O_p(T^{-k})$, for some $k > 0$. This implies

$$h_T^2 \sim_p T^{\frac{-r}{r\hat{\alpha}+d}} \sim_p T^{\frac{-r}{r\alpha+d}} O_p(exp\{M_1 \ln T T^{-k}\}) \sim_p T^{\frac{-r}{r\alpha+d}}$$

This implies, by Theorem 10, $\|\hat{g} - g\|_\infty = O_p\left((T/\ln T)^{\frac{-r}{r\alpha+d}}\right)$. Let $a_T^2$ be defined then as $a_T^1$, with $\alpha$ in place of $\overline{\alpha}$.

2. Proof. At first, $\bar{\phi}$ is consistent by 4.2.2 and Theorem 4.1.1 in Amemiya (1985). By Theorem 4.1.1, to show $\hat{\phi} - \phi = o_p(1)$, $\bar{\phi} - \phi = o_p(1)$, it suffices to show $\sup_{\phi\in\Phi}\left|\left(L_T(\phi, g^i, a_T^i) - L_T(\phi, g^i, 0)\right|\right. = o_p(1)$, $i = 1, 2$. (see also section 8.1). To show $\hat{\phi} - \bar{\phi} = o_p\left(T^{-k}\right)$ and the same for $\bar{\phi}$, it suffices to show (Theorem 4.1.3 in Amemiya, section 8.1): $\sup_{\phi\in\Phi}\left|\nabla_{\phi^j}^j \left(L_T(\phi, g^i, a_T^i) - L_T(\phi, g^i, 0)\right)\right| = o_p(1)$ for $j = 2$ and $o_p\left(T^{-k}\right)$ for $j = 1$. Also define:

$$\mathcal{P} \equiv \{t : y_t > \hat{g}^i(x_t) + a_T^i\}, \mathcal{P}^c \equiv \{t : y_t \leq \hat{g}^i(x_t) + a_T^i\},$$

$$\mathcal{P}_1^c \equiv \{t : y_t \leq g(x_t) + .5a_T^i\}, \mathcal{P}_1 \equiv \{t : y_t \leq g(x_t) + 1.5a_T^i\} \setminus \mathcal{P}_1^c, \mathcal{P}_2 \equiv \mathcal{P} \setminus \mathcal{P}_1$$

Following Smith (1994), (1985), decompose

$$\begin{aligned} L_T(\phi, g, 0) - L_T(\phi, g^i, a_T^i) &= T^{-1}(\alpha - 1)\Big\{\sum_{t\in\mathcal{P}} \ln(y_t - g(x_t)) - \ln(y_t - g^i(x_t))\Big\} \\ &+ T^{-1}\Big\{\sum_{t\in\mathcal{P}} \ln s(y_t - g(x_t), \phi, x_t) - \ln s(y_t - g^i(x_t), \phi, x_t)\Big\} \\ &+ T^{-1}\Big\{\sum_{t\in\mathcal{P}^c} (\alpha - 1)\ln(y_t - g(x_t)) + \ln s(y_t - g(x_t), \phi, x_t)\Big\} \\ &\equiv S_{1T} + S_{2T} + S_{3T} \end{aligned}$$

Further,

$$\sup_{\phi\in\Phi}\left|S_{2T}\right| = O_p\left(\left[T^{-1}\sum_{t\in\mathcal{P}} \frac{s_1(y_t - g^*(x_t), \phi, x_t)}{s(y_t - g^*(x_t), \phi, x_t)}\right]\left\|g^i - g\right\|_\infty\right)$$

where $s_1$ denotes the derivative of $s$ with respect to its first argument, and $\{g^*(x_t)\}$ are on the line segment adjoining $\{g(x_t)\}$ and $\{g^i(x_t)\}$. By assumptions on $s$,

$$\sup_{\phi\in\Phi}\left|S_{2T}\right| = O_p\left(\left\|g^i - g\right\|_\infty\right) = O_p\left(\left(T\ln T^{-1}\right)^{\frac{-r}{r\alpha^i+d}}\right) \text{ for } \alpha_1 = \overline{\alpha}, \alpha_2 = \alpha \tag{.7}$$

W.p. $\to 1$, $g^i \in \mathcal{G}_T \equiv \{g^i : \|g^i - g\|_\infty \leq a_T^i\}$. Hence w.p. $\to 1$, perturbing $g^i$ in $\mathcal{G}_T$ does not affect $\mathcal{P}_1^c$. Therefore,

$$\begin{aligned}\sup_{\phi\in\Phi}\left|S_{3T}\right| &\leq \sum_{t\in\mathcal{P}_1^c}\left|\ln(y_t - g(x_t))\right| + \sum_{t\in\mathcal{P}_1}\left|\ln(y_t - g(x_t))\right| + \sum_{t\in\mathcal{P}_1\cup\mathcal{P}_1^c} M_1 \\ &\leq O_p\left(\int_0^{M_2 a_T^i} \ln w w^{\alpha-1}dw\right) + O_p\left(\int_0^{M_3 a_T^i} \ln a_T w^{\alpha-1}dw\right) + O_p\left(\int_0^{M_4 a_T^i} w^{\alpha-1}dw\right) \\ &= O_p\left(\left(a_T^i\right)^\alpha \ln T\right)\end{aligned} \tag{.8}$$

Write $S_{1T} = S_{1T1} + S_{1T2}$, where $S_{1T1}$ is summation of elements, as defined above, over subscripts $t$ in $\mathcal{P}_2$ and $S_{1T2}$ is summation of elements, as defined above, over subscripts $t$ in $\mathcal{P}_1$. Expanding in $g^i$, and simple calculations yield:

$$\sup_{\phi}\left|S_{1T1}\right| = O_p\left(\int_{M_5 a_T^i}^{M_6}\left\|g^i - g\right\|_\infty w^{\alpha-1}dw\right) = O_p\left(\left\|g^i - g\right\|_\infty\right) = O_p\left(\left(T\ln T^{-1}\right)^{\frac{-r}{r\alpha^i+d}}\right) \tag{.9}$$

for $\alpha_1 = \overline{\alpha}, \alpha_2 = \alpha$ and

$$\sup_{\phi}\left|S_{1T2}\right| = O_p\left(\int_0^{M_7 a_T^i}\left(a_T^i\right)^{-1}\left\|g^i - g\right\|_\infty w^{\alpha-1}dw\right) = O_p\left(\left(a_T^i\right)^\alpha\right) \tag{.10}$$

So it follows from (.7), (.8), (.9), (.10):

$$\sup_{\phi}\left|L_T(\phi,\hat{g},a_T^2) - L_T(\phi,g,0)\right| = \begin{cases} O_p\left(\left(T\left(\ln T\right)^{-(1+\epsilon)}\right)^{\frac{-r\alpha}{r\alpha+d}}\ln T\right) & \text{if } 0<\alpha<1 \\ O_p\left(\left(T\left(\ln T\right)^{-1}\right)^{\frac{-r}{r\overline{\alpha}+d}}\right) & \text{if } \alpha \geq 1\end{cases} \tag{.11}$$

Further calculations give:

$$\sup_{\phi}\left|L_T(\phi,\hat{g},a_T^2) - L_T(\phi,g,0)\right| = \begin{cases} o_p\left(T^{-1/2}\right) & \text{if } 0<\alpha<1 \text{ and } r > \frac{d}{\alpha} \\ o_p\left(T^{-1/2}\right) & \text{if } 1 \leq \alpha < 2 \text{ and } r > \frac{d}{2-\alpha}\end{cases} \tag{.12}$$

Also note:

$$\sup_{\phi}\left|L_T(\phi,\bar{g},a_T^1) - L_T(\phi,g,0)\right| = O_p\left(\left(a_T^1\right)^\alpha \ln T + \left(T\ln T^{-1}\right)^{\frac{-r}{r\overline{\alpha}+d}} + \left(a_T^1\right)^\alpha\right) = O_p\left(T^{-k}\right) \tag{.13}$$

for any $k$: $0 < k < \min\left(\frac{r\alpha}{r\overline{\alpha}+d}, \frac{r}{r\overline{\alpha}+d}\right)$. This implies by Theorem 4.1.1 in Amemiya (1985) consistency of $\hat{\phi}$ and $\bar{\phi}$ and hence that $\hat{\phi}, \bar{\phi} \in$ interior$\Phi$ w.p. $\to 1$. Further, since derivatives $\nabla^j_{\phi^j} s$, $\nabla^j_{\phi^j}\alpha$ are bounded, observe that the bounds (.11), (.12), (.13) apply to $\sup_\phi\left|\nabla^j_{\phi^j}\left(L_T(\phi,g^i,a_T^i) - L_T(\phi,g,0)\right)\right|$ (for j=1,2), respectively for i=2 (.11, .12 ) and i=1 (.13). Thus we have shown that the probability bounds in (.11), (.12) apply to $\hat{\phi} - \tilde{\phi}$ and those in (.13) apply to $\bar{\phi} - \tilde{\phi}$. Note that, as we claimed, the rate of convergence of the pilot estimator, $\bar{\alpha}$ is $T^{-k'}$, $k' > 0$, since $\alpha_\phi$ is bounded, and from 4.1.3 in Amemiya (1985) or Smith (1985), we have: $\tilde{\phi} - \phi = O_p(T^{-1/2})$. This implies $k' = \min(1/2, k)$. Asymptotic normality follows as in Smith (1994),Smith (1985).

■

APPENDIX B: EXAMPLE: CONSTRUCTION OF THE DEFINING SETS FOR THE LOCAL EXTREME ORDER STATISTICS (SECTION 3.1)

Fix $x$. For $p = 1$, $N_h(x) = \{1\}$. Construct $N_h^1(x) = \{1\}$ and $N_h^2(x) = \{1\}$. Clearly condition 1 is satisfied. Condition 2 is relevant only for case $p \geq 2$. In condition 3, $\mathcal{X}_T = 1$, $\bar{u}(x, T, h_T) = 1$ and $\lambda$ is e.g. $(0.5, 0.5)$. In condition 4, $\delta_T = 1 > 0$.

Let $p = 2$. Take also $d = 1$, for brevity of notation. Fix any $0 < \varepsilon < \varepsilon' \leq 1 - K_2$. $\tilde{N}_h(x) = [x - h/2, x + h/2] \in \mathbb{R}$, $N_h(x) = \{(1, \tilde{x} - x)', \tilde{x} \in \tilde{N}_h(x)\}$, $\mathcal{X}_T = (1, 1/h_T)$. Then let $N_h^1(x) = \{(1, y)' : y \in [-h/2, -h/2(1-\varepsilon)]\}$, $N_h^2(x) = \{(1, y)' : y \in [h/2, h/2(1-\varepsilon)]\}$, $N_h^3(x) = \{(1, y)' : y \in [-h/2(1-\varepsilon), h/2(1-\varepsilon)]\}$ for $\varepsilon$ fixed $\in (0, 1)$ that depends on the constant $K_2$ in assumption 4. Thus conditions 1,2 are satisfied. By assumption 4, $\bar{u}(x, T, h_T) \in co^+((1, 1/2(1-\varepsilon'))', (1, -(1/2)(1-\varepsilon'))')$ ev. a.s. (or w.p. $\to 1$), thus condition 3 is satisfied ( Clearly, all $\lambda_{\tilde{z}} > \epsilon(\varepsilon') > 0$). The smallest value of the determinant of matrix diag$\{\mathcal{X}_T\}\tilde{z}$, for $\tilde{z}$ in $\times_{j=1}^2 N_{h_T}^j(z)$ is $(1 - \varepsilon)$, thus condition 4 holds as well. Case when $p > 2, d > 1$ follows in a way similar to the above, with yet more involved notation.

□

APPENDIX C: EXAMPLE: ASYMPTOTIC APPROXIMATION AND THE EMPIRICAL DISTRIBUTION OF THE NONPARAMETRIC EXTREME QUANTILE ESTIMATOR

Here we consider a small-scale simulation example that compares the asymptotic distribution of extreme quantile estimator with the simulated finite-sample distribution of this estimator. We consider simple setting and the case of locally constant regression. The set up is as follows:

$$y_t = \sin(2\pi x_t) + |u_t|, \quad \text{where } u_t \sim \text{ i.i.d } N(0,1), \quad x_t \sim \text{ i.i.d Uniform } (0,1)$$

This is an example of Model 2 with $\alpha = 1, d = 1$. The locally constant estimator ($p = 1$) is constructed using bandwidth choices computed as:

$$h_T = \frac{1}{T^a \ln T}, \quad a = 1/2 + 10^{-5}$$

This bandwidth satisfies assumption 3: $Th_T \to \infty$ and also the additional condition imposed in (5.2): $Th_T h_T^{r-\epsilon} \to 0$, for sufficiently small $\epsilon$. In this set up $r = 1$. Two sample sizes were considered : $T = 500$ and $T = 100$. Figures 6 and 7 present the asymptotic density of the normalized extreme quantile estimate, $\tilde{W} = Th_T (\hat{g}(x) - g(x))$, at $x = 0.5$, calculated on the basis of Corollary 3 ) and the NW-kernel estimate of the finite-sample distributions, based on 5000 replications.

APPENDIX D: PROOFS TO BE OMITTED

PROOF OF THEOREM 7. In what follows we take assumptions 1, 3b, and so on as the center of discussion. In round brackets, $(\cdot)$ we denote the modifications of the argument for the set of assumptions corresponding to 1, 3a and so on. In the proof, we should accommodate for stochastic nature of $\{h_T\}$. Note, by assumptions 3b,4b and (5.2)

$$h_T \in B_T \equiv \left[ M_1 T^{-\frac{1-(\alpha'+\epsilon)}{d}}, \; M_2 T^{-\frac{1}{d+\alpha(r-\epsilon)}} \right] \text{ ev. a.s.( in pr. )}$$

We, therefore, should establish the weak convergence *uniformly* in $\{h_T\} \in \{B_T\}$. Because of this complication, we employ the general definition of weak convergence, as defined in section 5, and show that it holds directly. Let $D(J) = 1\left(\bar{z}_{xT} \in \text{co}\left(z_{x_t} \in J\right)\right) \det\left[\{z_{x_t}, t \in J\}\right]$. Under assumptions stated the event when the exact conditional density $f_T$ exists (let $\mathcal{D}_T \equiv \{w \in \Omega : \exists f_T\}$) is when $\exists J$ s.t. $D(J) > 0$ (see Smith (1994)). This is true if e.g. $\{z_{x_t}, t \in J\} \in \times_j N_T^j(x)$, which happens ev. a.s. by Lemma 1 and by construction of $\{N_T^j(x)\}$ (in pr.). Given $\mathcal{D}_T$, the exact density of $\left(\beta_x - \hat{\beta}_x\right)$, $f_T^*$, conditionally on $N_{Tx}, \{z_{x_t}\}, \{h_T\}$, is given by:

$$\bar{f}_T(s) = \sum_{J \in I(x,T)} \left( D(J) \prod_{t \in J} \left\{ f_{x_t}\left( \left( z_{x_t}' s - \delta_{x_t} \right)_+ \right) \right\} \prod_{t \in N_{Tx} \setminus J} \left\{ 1 - F_{x_t}\left( \left( z_{x_t}' s - \delta_{x_t} \right)_+ \right) \right\} \right)$$

where the terms are defined in the statement of the Theorem, and

$$\delta_{x_t} \equiv \sum_{m \in \mathbb{Z}_+^d : [m] = k} \left( D^m g(x_t^*) - D^m g(x) \right) (x_t - x)^m \text{ for } x_t^* : \|x_t^* - x\|_\infty \leq h_T$$

The formula directly follows from Bassett results (1988) and the argument of Smith (1994), employing assumption 1. Let $\mathcal{H}_T$ be as defined in the statement of the Theorem, and the order of its diagonal elements is chosen so that $\tilde{W} = \mathcal{H}_T\left(\beta_x - \hat{\beta}_x\right) = O_p(1)$, which is possible by Theorem 4. Therefore the density of $\tilde{W}$ is:

$$f_T(w) = \frac{1}{\det\left[\mathcal{H}_T\right]} \left( \sum_{J \in I(x,T)} D(J) \prod_{t \in J} f_{x_t}\left( \left( z_{x_t}' \mathcal{H}_T^{-1} w - \delta_{x_t} \right)_+ \right) \times \prod_{t \in N_{Tx} \setminus J} \left( 1 - F_{x_t}\left( \left( z_{x_t}' \mathcal{H}_T^{-1} w - \delta_{x_t} \right)_+ \right) \right) \right)$$

Further, let $\tilde{u}_{x_t}$ be defined as in the text. Then $\det\left[\mathcal{H}_T\right] = \left(Th_T^d\right)^{p/\alpha} h_T^{\Xi}$, and $\det\left[\Psi_T\right] = h_T^{\Xi}$, where $\Xi = \sum_{m \in \mathbb{Z}^d : 0 \leq [m] \leq k} [m]$. By assumption 1, $\sup_{x_t \in C} |\delta_{x_t}| \leq M_4 h_T^r$, yielding that for any compact set $\mathcal{D} \in \mathbb{R}^p$:

$$\sup_{h_T \in B_T, u_{x_t} \in \Delta, w \in \mathcal{D}} \left| \left(Th_T^d\right)^{1/\alpha} \tilde{u}_{x_t}' w - \delta_{x_t} \right| \to 0 \tag{.14}$$

($\Delta$ is defined in Theorem 8). Also elementary matrix analysis gives $D(J) = \bar{D}(J) \times \det[\Psi_T]$. This and assumption 10 yield us the conjecture for approximate density (let $C_x \equiv C(x)$)

$$f_T^a(w) = \left(Th_T^d\right)^{-\frac{p}{\alpha}} \left( \sum_{J \in I(x,T)} \bar{D}(J) \prod_{t \in J} C_{x_t} \left( \left(Th_T^d\right)^{-\frac{1}{\alpha}} \bar{u}'_{x_t} w - \delta_{x_t} \right)_+^{\alpha - 1} \right) \times \prod_{t \in N_{Tx}} \left( 1 - C_{x_t} \left( \left(Th_T^d\right)^{-\frac{1}{\alpha}} \bar{u}'_{x_t} w - \delta_{x_t} \right)_+^{\alpha} \right)_+$$

Using assumption 10 again, and further calculations give us another conjecture

$$f_T^b(w) = \left(Th_T^d\right)^{-p} \left( \sum_{J \in I(x,T)} \bar{D}(J) \prod_{t \in J} C_x \left( \bar{u}'_{x_t} w - \left(Th_T^d\right)^{\frac{1}{\alpha}} \delta_{x_t} \right)_+^{\alpha - 1} \right) \times \prod_{t \in N_{Tx}} \left( 1 - C_x \left( \left(Th_T^d\right)^{-\frac{1}{\alpha}} \bar{u}'_{x_t} w - \delta_{x_t} \right)_+^{\alpha} \right)_+ \quad (.15)$$

Let $D_{\mathcal{K}} \equiv \{w \in \mathbb{R}^p : \|w\|_\infty \leq \mathcal{K}\}$ and $f_T^b(w, h_T)$ be defined as $f_T^b$ except that $\sum_{J \in I(x,T)}$ is replaced with $\sum_{J \in \Phi_T'}$ where $\Phi_T \equiv \{J : \forall t \in J : \bar{u}'_{x_t} w > h_T^r (Th_T^d)^{\frac{1}{\alpha}} M_5\}$, for some fixed sufficiently large $M_5 > 0$ (s.t. for sufficiently large T, $h_T^r (Th_T^d)^{\frac{1}{\alpha}} M_5 \gg (Th_T^d)^{\frac{1}{\alpha}} \delta_{x_t}, \forall t \in N_{Tx}$, which is possible by assumptions 1, 3). Let $f_T(w, h_T)$ be defined analogously from $f_T$, and $f_T^a(w, h_T)$ – from $f_T^a$. We start with the claim that for any fixed compact set $\mathcal{D}_{\mathcal{K}}$ and bounded $k : \mathbb{R}^p \to \mathbb{R}$ (take $k > 0$ w.l.o.g),

$$\sup_{h_T \in \mathcal{B}_T} \int_{\mathcal{D}_{\mathcal{K}}} k(w) \left| f_T(w, h_T) - f^b(w, h_T) \right| dw \to 0 \text{ a.s. ( in pr ).} \quad (.16)$$

Indeed by (.14) and assumption 10 $\sup_{h_T \in \mathcal{B}_T} \sup_{w \in \mathcal{D}_{\mathcal{K}}} k(w) \left| f_T(w, h_T) - f^a(w, h_T) \right| \to 0$, and by assumption 10, $\sup_{h_T \in \mathcal{B}_T} \sup_{w \in \mathcal{D}_{\mathcal{K}}} k(w) \left| f_T^a(w, h_T) - f_T^b(w, h_T) \right| = \sup_{h_T \in \mathcal{B}_T} O\left(h_T^{\rho r}\right) \to 0$. Next we have to show that for any fixed compact set $\mathcal{D}_{\mathcal{K}}$,

$$\sup_{h_T \in \mathcal{B}_T} \int_{\mathcal{D}_{\mathcal{K}}} k(w) \left| \bar{f}_T(w, h_T) - \bar{f}_T^b(w, h_T) \right| dw \to 0 \text{ a.s. ( in pr ).} \quad (.17)$$

where $\bar{f}_T(w, h_T) \equiv f_T(w) - f_T(w, h_T)$ and $\bar{f}_T^b$ is defined analogously. Then, the LHS of (.17)

$$\leq M_7 \sup_{h_T \in \mathcal{B}_T} 2 \int_{\mathcal{D}_{\mathcal{K}}} k(w) \left| \bar{f}_T^c(w, h_T) \right| dw = (*) \text{ for suff. large T}$$

where $\bar{f}_T^c(w, h_T)$ is defined as $\bar{f}_T^b(w, h_T)$ except that $\alpha$ is relaced by $\alpha' < \alpha$. This inequality is consequence of assumption 10 and definition of $\mathcal{B}_T$. Continuing,

$$(*) \leq M_8 \sup_{h_T \in \mathcal{B}_T} \int_{\mathcal{D}_{\mathcal{K}}} \left| \bar{f}_T^c(w, h_T) \right| dw \leq M_9 \sup_{h_T \in \mathcal{B}_T} \int_{\mathcal{D}_{\mathcal{K}}} (Th_T^d)^{-p} \sum_{J \in I(x,T) \setminus \Phi_T} \bar{D}(J) \prod_{t \in J} C_x \left( \bar{u}'_{x_t} w - (Th_T^d)^{1/\alpha} \delta_{x_t} \right)_+^{\alpha' - 1} dw$$

$$\leq M_{10} \sup_{h_T \in \mathcal{B}_T} (Th_T^d)^{-p} \sum_{J \in I(x,T)} \int_{\tilde{\mathcal{D}}_{\mathcal{K}}} \prod_{i_j \in J} C_x \left( z_{i_j} - (Th_T^d)^{1/\alpha} \delta_{x_{i_j}} \right)_+^{\alpha' - 1} \times_{i_j \in J} dz_{i_j} = (**)$$

where $\tilde{\mathcal{D}}_k \equiv \{-\infty \leq z_{i_1} \leq M_{11} (Th_T^d)^{\frac{1}{\alpha}} h_T^r, -\infty \leq z_{i_j} \leq p\mathcal{K}, \text{ for } i_j \neq i_1 \text{ and } \{i_j\} = J\}$. The first inequality follows since $k(\cdot)$ is bounded, the second – by observing that the product terms in (.15) ($\prod_{t \in N_{Tjx}}(\cdot)$) are bounded by 1. The third inequality follows by direct calculations and change of variables, setting $M_{11}$ sufficiently large and fixed, and also by observing that by definition of $\bar{u}_{x_t}$, $\mathcal{D}_{\mathcal{K}}$, $p\mathcal{K} \geq p \|\bar{u}_{x_i}\|_\infty \|w\|_\infty \geq \|z_i\|_\infty$, $\forall i \in J$. Continuing,

$$(**) \leq M_{12} \sup_{h_T \in \mathcal{B}_T} C_x \frac{\text{card}(N_{Tx})!}{(\text{card}(N_{Tx}) - p)! p!} \left(Th_T^d\right)^{-p} \left(Th_T^d\right) h_T^{r\alpha} \leq \sup_{h_T \in \mathcal{B}_T} M_{14} w_x \left(Th_T^d\right) h_T^{r\alpha} \to 0.$$

where the first inequality follows by a simple combinatorial calculation, the second is by (next) Lemma 7. Next we claim that

$$\sup_{h_T \in \mathcal{B}_T} \int_{\mathcal{D}_{\mathcal{K}}} k(w) \left| f_T^*(w) - f_T(w) \right| dw \to 0, \text{ a.s. ( in pr. )} \quad (.18)$$

where $f_T^*$ is defined in the statement of the Theorem. Let $f_T^*(\cdot, \gamma)$ be defined as $f_T^*$ where summation $\sum_{J \in I(x,T)}$ is replaced by $\sum_{J \in \Phi_T'}$, where $\Phi_T' \equiv \{J : u'_{x_t} w > \gamma\}$ and let $f_T^e(\cdot, \gamma)$ be defined from $f_T^b$ similarly. By (.15),(.16), (.17) to show (.18) it suffices to establish

$$\sup_{h_T \in \mathcal{B}_T} \int_{\mathcal{D}_{\mathcal{K}}} k(w) \left| f_T^*(w) - f_T^e(w) \right| dw \to 0, \text{ a.s. ( in pr. )} \quad (.19)$$

Then by Taylor expansion,

$$\sup_{h_T\in B_T}\int_{\mathcal{D}_{\mathcal{K}}} k(w)\big|f_T^*(w,\gamma)-f_T^c(w,\gamma)\big|dw \le M_{15}\sup_{h_T\in B_T,w\in\mathcal{D}_{\mathcal{K}},t\in N_{Tx}}\left(Th_T^d\right)^{\frac{1}{\alpha}}\delta_{x_t}\le \sup_{h_T\in B_T} M_{16}\left(Th_T^d\right)^{\frac{1}{\alpha}}h_T^{\tau}\to 0.$$

For $\tilde{f}_T^*(\cdot,\gamma)\equiv f_T^*-f_T^*(\cdot,\gamma)$, $\tilde{f}_T^c(\cdot,\gamma)\equiv f_T^c-f_T^c(\cdot,\gamma)$, proceeding as in the proof of (.17) we find that for any $\epsilon>0\ \exists\gamma_\epsilon$ small, s.t. for sufficiently large T,

$$\sup_{h_T\in B_T}\int_{\mathcal{D}_{\mathcal{K}}} k(w)\big|\tilde{f}_T^*(w,\gamma)-\tilde{f}_T^c(w,\gamma_\epsilon)\big|dw\le\epsilon$$

which shows (.18). Now from the proof of the next Theorem (proof of which does not rely on this paragraph), for any $\varepsilon>0\ \exists\mathcal{K}$ sufficiently large so that $\int_{\mathcal{D}_{\mathcal{K}}^c} f_T^* dw\le\varepsilon/(4\overline{k})]$, ev. a.s. ( w.p. $\to 1$), where $\overline{k}=\sup_w k(w)$. This provides the necessary integrability ("tightness") condition for $f_T$: from (.18) $\exists\, T$ sufficiently large s.t $\int_{\mathcal{D}_{\mathcal{K}}} k(w)(f_T^*-f_T)(w)dw\le\varepsilon/4$, and hence $\int_{\mathbb{R}^p} k(w)(f_T^*-f_T)(w)dw\le\varepsilon$, ev. a.s. ( w.p. $\to 1$). Finally, since $h_T\in B_T$ ev. a.s., (w.p. $\to 1$), we have completed the proof.

■

PROOF OF THEOREM 8 . The proof proceeds based on the set of assumptions stated in the Theorem and when the bandwidth is chosen by assumption 3b. In brackets $(\cdot)$ we indicate changes of the argument when assumption 3a is used instead. Let $\mathcal{D}_{\mathcal{K}}$ de defined as in the proof of Theorem 7. We begin with the claim that, uniformly in $w$ in $\mathcal{D}_{\mathcal{K}}$,

$$\prod_{t\in N_{Tx}}(1-(\tilde{u}_{x_t}'w)_+/(Th_T^d)^{\frac{1}{\alpha}})^\alpha\to\exp\{-w_x(x)E^*(u^{\{m\}'}w)_+^\alpha\}\text{ a.s (in probability)}\tag{.20}$$

where $E^*$ is the expectation operator w.r.t. to the distribution function $F^*$ (of $u^{\{m\}}$) defined in the statement. Indeed uniformly in $(u_{x_t},w)\in\Delta\times\mathcal{D}_{\mathcal{K}}$, $(u_{x_t}'w)_+/(Th_T^d)^{1/\alpha}\to 0$ a.s. (in pr.) , therefore $\left(\prod_{t\in N_{Tx}}(1-(\tilde{u}_{x_t}'w)_+/(Th_T^d)^{\frac{1}{\alpha}})^\alpha\right)/\left(\exp\{-\sum_{t\in N_{Tx}}(\tilde{u}_{x_t}'w)_+^\alpha/Th_T^d\}\right)\to 1$. Since by Lemma 7, card$(N_{Tx})/Th_T^d\to w_x(x)$ a.s. ( in pr.,) $\left(\exp\{-\sum_{t\in N_{Tx}}(\tilde{u}_{x_t}'w)_+^\alpha/Th_T^d\}\right)=\exp\{-w_x(x)E_T^l(\tilde{u}'w)_+^\alpha\}+o(1)$, a.s. ( in pr.) where $E_T^l$ is the expectation operator w.r.t. the local empirical distribution function in Lemma 7 (of $\{\tilde{u}_{x_t}\}$), and $\exp\{-w_x(x)E_T^l(\tilde{u}'w)_+^\alpha\}\to\exp\{-w_x(x)E^*(u^{\{m\}'}w)_+^\alpha\}$ a.s. ( in pr.) by the Lemma 7 and the Helly-Bray Theorem. To show the uniform convergence over $\mathcal{D}_{\mathcal{K}}$, we use the Arzela-Ascoli Theorem (Theorems 21.7, 21.8, 21.9 in Davidson (1994)), by establishing strong stochastic equicontinuity, s.s.e. (weak– in case of assumption 3b, w.s.e.). First, note that $((\mathcal{D}_{\mathcal{K}},\|\cdot\|_\infty))$ is of course totally bounded metric space; then let $B(w,\delta)$ be a box with side length $\delta$ around $w$, then let $\zeta(T,h,\epsilon)\equiv\big|\text{card}(N_{Tx})^{-1}(\sum_{t\in N_{Tx}:\tilde{u}_{x_t}'w\ge(\epsilon/4)^{1/\alpha}}(\tilde{u}_{x_t}'w')_+^\alpha-(\tilde{u}_{x_t}'w)_+^\alpha)\big|$, and let $\zeta'(T,h,\epsilon)$ be defined similarly, but with summation $\sum_{t\in N_{Tx}:\tilde{u}_{x_t}'w<(\epsilon/4)^{1/\alpha}}$. For case when $\alpha\ge 1$, by expansion (recall $p=\dim(u_{x_t})$):

$$\eta\equiv\sup_{w\in\mathcal{D}_{\mathcal{K}},h_T\in B_T}\sup_{w'\in B(w,\delta)}\zeta(T,h,0)\le\sup_{h_T\in B_T,t\in N_{Tx}}\|\tilde{u}_{x_t}\|_\infty p\delta\le p\delta,\tag{.21}$$

noting that, by construction, $\sup_{h_T\in B_T,t\in N_{Tx}}\|\tilde{u}_{x_t}\|_\infty\le 1$. For case: $\alpha<1$,

$$\begin{aligned}\eta\le\sup_{w\in\mathcal{D}_{\mathcal{K}},h_T\in B_T}\sup_{w'\in B(w,\delta)}(\zeta(T,h,\epsilon)+\zeta'(T,h,\epsilon))&\le\left(\left(\frac{\epsilon}{4}\right)^{\frac{1}{\alpha}}-\sup_{h_T\in B_T,t\in N_{Tx}}\|u_{x_t}\|_\infty p\delta\right)^{\alpha-1}\sup_{h_T\in B_T,t\in N_{Tx}}\|u_{x_t}\|_\infty\delta\\&+\left(\left(\frac{\epsilon}{4}\right)^{\frac{1}{\alpha}}+\sup_{h_T\in B_T,t\in N_{Tx}}\|u_{x_t}\|_\infty p\delta\right)^\alpha\le\delta\left(\left(\frac{\epsilon}{4}\right)^{\frac{1}{\alpha}}-p\delta\right)^{\alpha-1}+\left(\left(\frac{\epsilon}{4}\right)^{\frac{1}{\alpha}}+p\delta\right)^\alpha\end{aligned}\tag{.22}$$

Then, for any $\epsilon>0$, set $\delta$ sufficiently small s.t $\eta<\epsilon$, so that $P(\eta>\epsilon,i.o.)=0$, (or $\limsup_T P(\eta>\epsilon)=0$), which shows s.s.e. (w.s.e.). Next, let

$$f_T^1(w)=\left(Th_T^d\right)^{-p}\Big(\sum_{J\in I(x,T)}\tilde{D}(J)\prod_{t\in J}C(x)\,(\tilde{u}_{x_t}'w)_+^{\alpha-1}\Big)\times\exp\{-w_x(x)E^*(u^{\{m\}'}w)_+^\alpha\}$$

then $\sup_{h_T\in B_T}\int_{\mathcal{D}_{\mathcal{K}}}|f_T^1-f_T^*|(w)dw\to 0$ by (.20). Next, let $\tilde{u}(J)\equiv\{\tilde{u}_{x_t},t\in J\}$, $\overline{D}(J)\equiv 1(\bar{u}^{\{m\}}\in co(\tilde{u}(J)))\det[\tilde{u}(J)]$, and let $f_T^2$ be defined as $f_T^1$ where $\tilde{D}$ is replaced with $\overline{D}$, hence

$$\begin{aligned}\sup_{h_T\in B_T}\int_{\mathcal{D}_{\mathcal{K}}}|f_T^2(w)-f_T^*(w)|dw&\le(Th_T^d)^{-p}\sum_{J\in I(x,T)}\big|1(\tilde{u}^{\{m\}}\in co(\tilde{u}(J)))-1(\tilde{u}_{xT}\in co(\tilde{u}(J)))\big|O(1)\\&\le\big(\sup_{h_T\in B_T}\big|P_T^{lJ}(A_{1T})-P_T^{lJ}(A_{2T})\big|\big)O(1)\end{aligned}\tag{.23}$$

where $P_T^{lJ}$ is the empirical probability measure of $\{\text{vech}\{\tilde{u}(J)\}, J \in I(x,T)\}$. It is trivial to check that uniformly in $x \in \mathbb{R}^{p\times p}$, $\sup_{h_T\in B_T} \left|P_T^{lJ}(\text{vech}\{\tilde{u}_{x_t}, t \in J\} \leq x) - \left(P_T^{l}(\tilde{u}_{x_t} \leq x)\right)^p\right| \to 0$, a.s. (in pr.) where $P_T^l$ is the local empirical probability measure defined by the local empirical distribution function in Lemma 7. Above, sets $A_{1T}, A_{2T}$ are hyper semi-quadrands in $\Delta^p$ symmetric around $\bar{u}^{\{m\}}$ and $\bar{u}_{xT}$, respectively. The Lebesgue difference between the two sets projected on $\mathbb{R}^{p\times(p-1)}$ is at most of order $\|\bar{u}^{\{m\}} - \bar{u}_{xT}\|_\infty$, and hence by Lemma 7 the LHS in (.23 ) is o(1) a.s. (in pr.) (see Portnoy and Jureckova for a similar argument).

Let $\lambda\left(u(J)\right) \equiv \|((\lambda-1)^+\|_2 + \|\lambda^-\|_2$, if $\exists\lambda \in \mathbb{R}^p : u(J)'\lambda = \bar{u}^{\{m\}}$. By assumption 5 and Lemma 7, card$(J \in I(x,T) : \nexists\lambda \in \mathbb{R}^p : u(J)'\lambda = \bar{u}^{\{m\}})/$ card$(J \in I(x,T))= o(1)$ a.s. (in pr.). From Lemma 7 and simple combinatorial calculus as in the proof of Theorem 7, card$(J \in I(x,T))/Th_T^d \to (w_x)^p/p!$. Then, let $\chi(\epsilon,\bar{\lambda}) \equiv 1 - G(\bar{\lambda}/\epsilon)$, where $G$ is smooth distribution function s.t. $G(x) = 1$ if $x \geq 1$, $G(x) = 0$, if $x < 1$, then let $f_T^3(\cdot,\epsilon)$ be defined as $f_T^2(\cdot)$, but where $1(\bar{u}^{\{m\}} \in co(\tilde{u}(J)))$ is replaced by $\chi(\epsilon, \lambda(u(J)))$ ( That is we smooth the discontinuity). Analysis, similar to that in the proof of (.23) shows $(Th_T^d)^{-p}\sum_{J\in I(x,T)}\left|\chi(\epsilon,\lambda(u(J))) - 1(\bar{u}^{\{m\}} \in co(\tilde{u}(J)))\right| \leq M_2\epsilon$ ev. a.s. (w.p. $\to 1$), which as in (.23) shows that:

$$\sup_{h_T\in B_T}\int_{\mathcal{D}_K}\left|f_T^3(w,\epsilon) - f_T^2(w)\right|dw \leq M_3\epsilon \text{ ev. as. ( w.p.} \to 1\text{ )} \tag{.24}$$

Finally, by the Helly-Bray Theorem, and by Lemma 7, uniformly in $h_T \in B_T$, (and then by Fubini)

$$\begin{aligned}\int_{\mathcal{D}_K} f_T^3(w,\epsilon)dw \to \int_{\mathcal{D}_K}\frac{w_x^p}{p!}\Bigg(\int_{\Delta^p}\left(\chi\left(\epsilon,\lambda(u^{\{m\}}(P)\right)\right)\det\left[u^{\{m\}}(j), j \in P)\right]\prod_{j=1}^{p}\alpha C(x) \\ \left(\left(u^{\{m\}}(j)\right)'w\right)_+^{\alpha-1}\prod_{j=1}^{p}dF_*(u^{\{m\}}(j))\Bigg)\times exp\;\left\{-w_x(x)C(x)\int_\Delta((u^{\{m\}})'w)_+^\alpha dF_*(u)\right\}dw\end{aligned} \tag{.25}$$

where notation is defined in the statement of this Theorem. Now noting that the RHS (denoting it as $f^*(\cdot,\epsilon)$ of (.26) is continuous in $\epsilon$, and since $\epsilon$ is arbitrary in (.24), (.25), we have that a.s. (in pr. ) uniformly in $h_T \in B_T$,

$$\int_{\mathcal{D}_K} f_T^2(w)dw \to \int_{\mathcal{D}_K} f^*(w,0)dw = \int_{\mathcal{D}_K} f^*(w)dw \tag{.26}$$

where $f^*$ is defined in the statement of this Theorem. By steps (.20)-(.26), the last statement is true for $f_T^*$, thus the conclusion follows by also noting that $\{f_T(w)\}$ (the exact densities of $\{\tilde{\mathcal{W}}_T\}$ , defined in the proof of the previous Theorem) are uniformly integrable, by the argument of Portnoy and Jureckova) and that the limiting density integrates to 1.

■

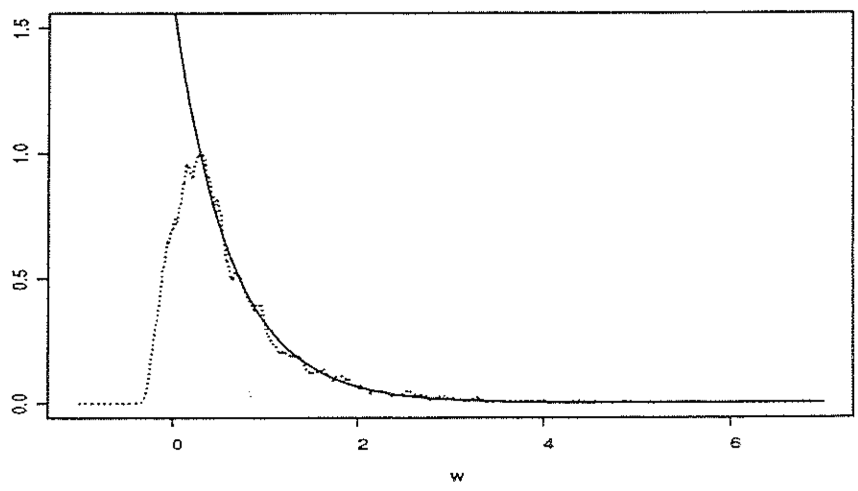


Figure 5: Asymptotic (*solid line* ) and the Kernel-Smoothed Monte-Carlo Density (*dotted line*). $T = 100$

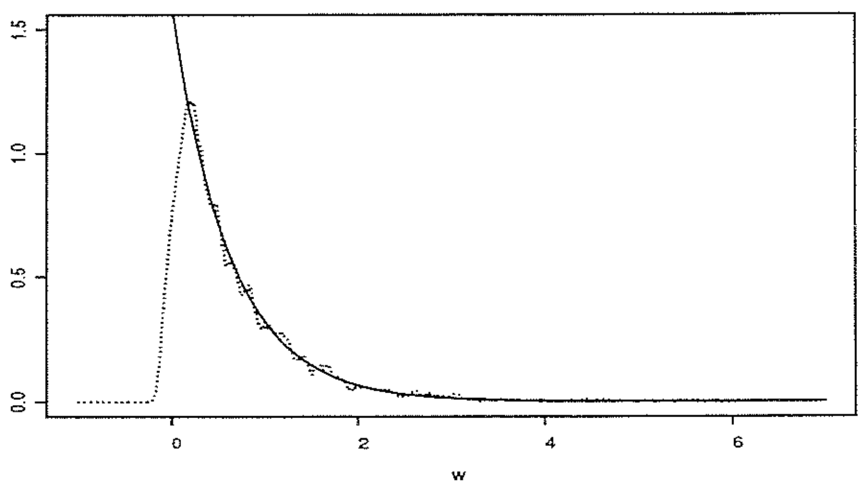


Figure 6: Asymptotic (*solid line* ) and the Kernel-Smoothed Monte-Carlo Density (*dotted line*). $T = 500$

## References


AMEMIYA, T. (1985): *Advanced Econometrics*. Harvard University Press.

ANDERSON, C. W. (1978): "Super-slowly varying functions in extreme value theory," *J. Roy. Statist. Soc. Ser. B*, 40(2), 197–202.

——— (1984): "Large deviations of extremes," in *Statistical extremes and applications (Vimeiro, 1983)*, pp. 325–340. Reidel, Dordrecht-Boston, Mass.

ANDREWS, D. W. K. (1994): "Empirical Process Methods in Econometrics," in Engle, R.F. and D. McFadden (eds), Handbook of Econometrics, Vol. IV.

——— (1995): "Nonparametric kernel estimation for semiparametric models," *Econometric Theory*, 11(3), 560–596.

ATKINSON, K. (1989): *An Introduction to Numerical Analysis*. New York: Wiley.

BASSETT, G. S. (1988): "A Property of the Observations Fit by the Extreme Regression Quantiles," *Computational Statistics and Data Analysis*, 6, 353–359.

BOOR DE, C. (1978): *A Practical Guide to Splines*. New York: Springer.

BUCHINSKY, M. (1995): "Estimating the Asymptotic Covarinace Matrix for Quantile Regression: A Monte-Carlo Study," *J. Econometrics*, 68, 303–338.

CHAUDHURI, P. (1991): "Nonparametric estimates of regression quantiles and their local Bahadur representation.," *Ann. Statist.*, 19, 760–777.

CHENG, M.-Y., J. FAN, AND J. S. MARRON (1997): "On automatic boundary corrections," *Ann. Statist.*, 25(4), 1691–1708.

CHERNOZHUKOV, V. (1998): "Nonparametric Extreme Regression Quantiles: Technical Report," .

——— (1999a): "Conditional Extremes and Near-Extremes," Job Market Paper.

——— (1999b): "Conditional Extremes and Near-Extremes: Estimation, Inference, and Economic Applications," draft of Ph.D. Dissertation, Stanford, Department of Economics.

——— (1999c): "Inference for Extreme and Near-Extreme Regression Quantiles," in progress , Stanford.

CHERNOZHUKOV, V., AND H. HONG (1998a): "Asymptotic Behavior of Nonlinear Extreme Regression Quantiles," October, 1998, mimeo, Stanford.

——— (1998b): "Censored Quantile Regression with Extreme Quantile Estimate of the Unknown Censoring Points," October, Stanford.

——— (1998c): "Quantile Estimation in Truncated Model," October, 1998, Paper presented at CEME seminar on the Recent Developments in Semi-Parametric Econometrics, NY, Oct. 31.

CLEVELAND, W. S. (1979): "Robust locally weighted regression and smoothing scatterplots.," *J. Amer. Statist. Assoc.*, 74, 829–836.

CUEVAS, A., AND R. FRAIMAN (1998): "A plug-in approach to support estimation," *Ann. Statist.*, 25(6), 2300–2312.

DAVIDSON, J. (1994): *Stochastic limit theory*. The Clarendon Press, Oxford University Press, New York.

DE HAAN, L. (1970): *On Regular Variation and Its Applications to the Weak Convergence*. Mathematical Centre Tract 32, Mathematical Centre, Amsterdam, Holland.

DE HAAN, L., AND A. HORDIJK (1972): "The rate of growth of sample maxima," *Ann. Math. Statist.*, 43, 1185–1196.

FAN, J. (1993): "Local linear regression smoothers and their minimax efficiencies.," *Ann. Statist.*, 21, 196–216.

GALOMBOS, J. (1978): *The Asymptotic Theory of Extreme Order Statistics*. Wiley, Ney York.

——— (1984): *Order Statistics. In:Handbook of Statistics. Vol4, Eds. P.R.Krishnaiah and P.K.Sen, pp. 359-382.* Amsterdam:North-Holland.

GNEDENKO, B. (1943): "Sur la distribution limité du terme d' une série alétoire," *Ann. math.*, 44, 423–453.

GOLDIE, C. M., AND R. L. SMITH (1987): "Slow variation with remainder: theory and applications," *Quart. J. Math. Oxford Ser. (2)*, 38(149), 45–71.

HÄRDLE, W., B. U. PARK, AND A. B. TSYBAKOV (1995): "Estimation of non-sharp support boundaries," *J. Multivariate Anal.*, 55(2), 205–218.

HENGARTNER, N. W., AND O. B. LINTON (1996): "Nonparametric regression estimation at design poles and zeros," *Canad. J. Statist.*, 24(4), 583–591.

KOENKER, R., AND G. S. BASSETT (1978): "Regression Quantiles," *Econometrica*, 46, 33–50.

KOENKER, R., AND S. PORTNOY (1996): "Fundamental Properties of Regression Quantiles," preprint, Department of Statistics and Economics, University of Illinois.

——— (1999): "Regression Quantiles," preprint.

LEADBETTER, M. R., G. LINDGREN, AND H. ROOTZÉN (1983): *Extremes and related properties of random sequences and processes.* Springer-Verlag, New York-Berlin.

MAMMEN, E., AND A. B. TSYBAKOV (1995): "Asymptotical minimax recovery of sets with smooth boundaries," *Ann. Statist.*, 23(2), 502–524.

NEWEY, W., AND D. MCFADDEN (1994): "Large Sample Estimation and Hypothesis Testing," in Engle, R.F. and D. McFadden (eds), Handbook of Econometrics, Vol. IV.

PICKANDS, J. (1968): "Moment convergence of sample extremes," *Ann. Math. Statist.*, 39, 881–889.

PORTNOY, S. (1987): "Using Regression Fractiles to Identify Outliers," *Y. Dodge (Ed) Statistical Data Analysis Based on $L_1$ Norm and Related Methods*, North Holland, Amsterdam.

PORTNOY, S., AND J. JURECKOVA (1998): "On Extreme Regression Quantiles," preprint.

POWELL, J. L. (1986a): "Censored Regression Quantiles," *Journal of Econometrics*, 32, 143–155.

——— (1986b): "Symmetrically Trimmed Least Squares Estimation For Tobit Models," *Econometrica*, 54, 1435–1460.

PRENTER, P. M. (1989): *Splines and Variational Methods.* New York: Wiley.

REISS, R. (1989): *Approximate Distributions of Order Statistics.* Springer.

RESNICK, S. (1987): *Extreme Values, Regular Variation, and Point Processes.* Springer.

RIVLIN, T. (1990): *Chebyshev Polynomials : From Approximation Theory to Algebra and Number theory.* New York: Wiley.

ROBINSON, P. M. (1997): "Large-sample inference for nonparametric regression with dependent errors," *Ann. Statist.*, 25(5), 2054–2083.

SMITH, R. L. (1985): "Maximum Likelihood Estimation in a Class of Nonregular Cases," *Biometrika*, 72, 67–90.

——— (1994): "Nonregular Regression," *Biometrika*, 81, 173–183.

STONE, C. J. (1977): "Consistent nonparametric regression.," *Ann. Statist.*, 5, 595–645.

——— (1982): "Optimal Global Rates of Convergence for Nonparametric Regression.," *Ann. Statist.*, 10, 1040–1053.

VAN DER VAART, A. W., AND J. A. WELLNER (1996): *Weak convergence and empirical processes.* Springer-Verlag, New York.

WAND, M. P., AND M. C. JONES (1995): *Kernel smoothing.* Chapman and Hall, Ltd., London.

WELSH, A. H. (1996): "Robust estimation of smooth regression and spread functions and their derivatives," *Statist. Sinica*, 6, 347–366.