Robust Inference about Conditional Tail Features: A Panel Data Approach*

Yuya Sasaki† and Yulong Wang‡

August 2019

Abstract

We develop a new extreme value theory for panel data and use it to construct asymptotically valid confidence intervals (CIs) for conditional tail features such as conditional extreme quantile and conditional tail index. As a by-product, we also construct CIs for tail features of the coefficients in the random coefficient regression model. The new CIs are robustly valid without parametric assumptions and have excellent small sample coverage and length properties. Applying the proposed method, we study the tail risk of the monthly U.S. stock returns and find that (i) the left tail features of stock returns and those of the Fama-French regression residuals heavily depend on other stock characteristics such as stock size; and (ii) the alpha's and beta's are strongly heterogeneous across stocks in the Fama-French regression. These findings suggest that the Fama-French model is insufficient to characterize the tail behavior of stock returns.

Keywords: panel data, conditional tail, extreme value theory, nonparametric confidence interval

*We thank Federico Bugni, Yanqin Fan, Yoonseok Lee, Zhijie Xiao, Yichong Zhang, and participants at the seminar/conference at Boston College, PSU, SMU, and Econometric Society North American Summer Meeting 2019 for very helpful comments and advice. Wang gratefully acknowledges the financial support by the Applyby-Mosher fund.

†Associate professor of economics, Vanderbilt University. Email: yuya.sasaki@vanderbilt.edu

‡Assistant professor of economics, Syracuse University. Email: ywang402@maxwell.syr.edu.

1 Introduction

Tail risk and extreme events are important research topics in economics and finance. In many applications, the features of interest are conditional tail properties such as conditional tail index and conditional extreme quantile. This article provides a new method to construct confidence intervals for these features. The main advantage of the new method is its robustness against flexible distributional assumptions. In particular, it allows all of the location, scale, and shape to nonparametrically depend on the covariates.

Compared with unconditional tail features, the conditional tail counterparts are much more difficult to study. This is because conditional tails depend on both marginal distributions and their joint behavior. Although the marginal ones can be generally assumed to be Pareto near the tails, the joint has to be fully nonparametric and is hard to study given very limited tail observations. To model a covariate-dependent yet tractable tail, the seminal paper by Chernozhukov (2005) extends the quantile regression (QR) estimator of Koenker and Bassett (1978) from mid-sample to tails, called the extremal quantile regression (EQR). Chernozhukov and Fernández-Val (2011) further investigate the EQR to construct confidence intervals (CIs) based on subsampling.

The EQR approach assumes that the conditional extreme quantile can be well approximated by a parametric location-scale shift model. More specifically, suppose we have independently and identically distributed (i.i.d.) observations $\{Y_i, X_i\}$ for $i = 1, \ldots, n$ and are interested in the τ-th quantile of Y given $X = x$, denoted as $Q_{Y|X=x}(\tau)$. The EQR approach assumes that when τ is close to 1,

$$Q_{Y|X=x}(\tau) \sim \mu(x) + \sigma(x)(1-\tau)^{-\xi} \tag{1}$$

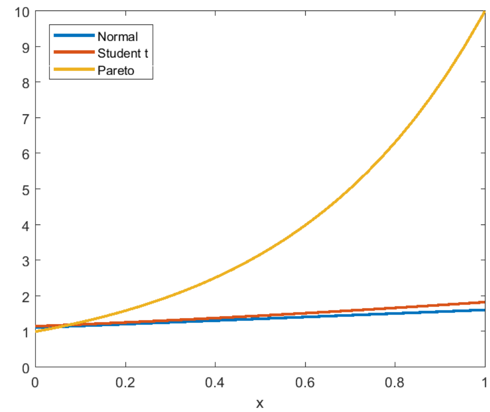
for some *parametric* functions $\mu(x)$ and $\sigma(x)$, which respectively capture the location and the scale. The element $(1-\tau)^{-\xi}$ can be treated as the quantile function of a standard Pareto distribution, that is,

$$\mathbb{P}(Y > y) \sim y^{-1/\xi}, \tag{2}$$

where $1/\xi$ is the Pareto exponent and ξ is called the tail index. This single parameter captures the tail *shape* in the way that a larger ξ implies a heavier tail. The assumption of the model (1) simplifies the conditional tail distribution so that the covariate X only affects the location and scale, but not the shape.[1] This is satisfied if X and Y are jointly normal

[1] Wang and Li (2013) formally establish that the location-shift model assumption is equivalent to assuming ξ remains constant across x.

Figure 1: Plot of the 0.9 Quantile of Y Conditional on X = x from Three Distributions



Note: This figure plots $Q_{Y|X=x}(0.9)$ for $x \in [0, 1]$ where Y and X are distributed as follows: (i) joint normal with zero means, unit variances, and 0.5 correlation; (ii), joint student's t with degree of freedom 3, zero mean, unit variances, and correlation 0.5; and (iii) X is standard normal and $Y|X \sim$Pa$(\xi(x))$, that is, $\mathbb{P}(Y > y|X = x) = 1 - y^{-1/\xi(x)}$ with $\xi(x) = x$.

but violated by many other joint distributions. In contrast to mid-sample features, such a violation may lead to a substantial misspecification bias in studying tail ones.

To have a better sense of the misspecification, Figure 1 plots the 90% conditional quantiles of Y given $X = x \in [0, 1]$, for three commonly used joint distributions. First, the blue curve depicts the standard joint normal distribution with correlation 0.5, so that $Q_{Y|X=x}(\tau)$ is always linear in x. Condition (1) is satisfied in this case. Second, the red curve depicts the joint student's t distribution with 3 degrees of freedom (d.f.), zero means, unit variances, and 0.5 covariance. Conditional on $X = x$, the distribution of Y is still student's t but with the mean and the variance depending on x in a highly nonlinear way. The upward slope in the tail reflects this feature. Condition (1) then leads to some bias if $\mu(x)$ and $\sigma(x)$ are misspecified. Third, the yellow curve depicts the distribution where the conditional distribution of Y given $X = x$ is Pareto with exponent $1/x$. The conditional quantile $(1 - \tau)^{-x}$ is highly nonlinear in x, and hence approximating such a nonlinear tail feature with the linear location-scale shift model (1) induces a large misspecification bias.

The misspecification error is not only a theoretical but also an empirical concern in important situations. First, conditional value-at-risk (VaR) is a risk measure commonly used in financial management, insurance, and actuarial science. Estimation and inference are studied by Chernozhukov and Umantsev (2001) and Engle and Manganelli (2004), among many others. Recently Adrian and Brunnermeier (2016) propose a new measure for systemic risk, Δ-CoVar, defined as the difference between two conditional VaRs. The tail shape governs the third- and higher-order moments of the portfolio return, which typically depend on other economic factors, say business cycle. But this is excluded by the location-scale model (1). Second, Kelly and Jiang (2014) find that extreme event risk affects asset pricing in the U.S. stock market. The distribution of stock returns is approximately Pareto in the tail with a time-varying and stock-specific shape parameter. In particular, the shape parameter measures tail risk and varies with other stock characteristics such as stock size. We empirically examine this point in Section 5. Third, top wealth inequality is an active research question in macro finance literature (see, for example, Piketty and Saez (2003), Gabaix, Lasry, Lions, and Moll (2016), and Jones and Kim (2018)). Tail of the wealth distribution is well documented as Pareto, and the exponent is in general a function of fundamentals in general equilibrium models. For example, Beare and Toda (2017) derive a formula for the Pareto exponent and comparative statics results, and Toda (2019) applies that formula in a general equilibrium context. Finally, how infant's birthweight depends on mother's demographics and maternal behavior is an important question in health economics.[2] See Abrevaya (2001), Koenker and Hallock (2001), and Chernozhukov and Fernández-Val (2011). Other economic problems about conditional tail features can be found in the comprehensive review by Chernozhukov, Fernández-Val, and Kaji (2017).

To solve the misspecification issue, there have been some suggestions in the literature on relaxing the local-scale specification (1). To our best knowledge, they all focus on estimation (as opposed to inference) and can be roughly categorized into two classes. The first class maintains some parametric form but relaxes the location-shift model to allow for some non-linearity. They are more flexible but still suffer from misspecification biases. In particular, Wang and Tsai (2009) assume $\xi(x)$ equals to $\exp(x^\intercal\theta_0)$ for some unknown parameter θ_0, and Wang and Li (2013) assume that the Box-Cox transformed Y has a linear conditional quantile in X. The second class is fully nonparametric and constructs some local smooth estimators, including, for example Beirlant, Joossens, and Segers (2004), Gardes, Girard, and Lekina (2010), Gardes, Guillou, and Schorgen (2012), Daouia, Gardes, and Girard (2013), and Martins-Filho, Yao, and Torero (2018). These approaches depend heavily on the richness of the data in the target neighborhood and hence require very large samples.

[2] We also find that mother's net birthweight affects the tail shape of the baby's birthweight in a nonlinear way, which is not reported because of space limit.

In this article, we focus on statistical inference instead of estimation, and provide CIs of conditional tail features that have good coverage and length properties in relatively small samples. In addition, we consider repeated cross-sectional or panel data instead of cross-sectional random samples. The main idea is very intuitive: take one particular observation from each time series and collect them into a cross-sectional sample. Suppose we have panel data of Y and X for many individuals and many time periods, and are interested in some tail feature of the conditional distribution of Y given $X = x_0$, denoted $F_{Y|X=x_0}$. If for every individual, there exists some time period in which X takes x_0, we can simply collect the associated Y's and form a cross-sectional sample from $F_{Y|X=x_0}$. Since this is infeasible when X is continuous, we instead collect from each individual's time series, the induced Y associated with the X that is the nearest neighbor (NN) to x_0. These induced Y's are now *approximately* stemming from $F_{Y|X=x_0}$, and the large (respectively, small) order statistics from them can be used for inference about the right (respectively, left) tail of $F_{Y|X=x_0}$. For multi-dimensional covariates, this is done by defining the NN measured by a certain choice of metric, such as the L^2-norm. If a linear regression model is appropriate, the NN can also be defined using the linear index.

The above approximation is formalized by establishing a new extreme value (EV) theory. The proof is based on the large n and large T asymptotics, where n and T denote the sample sizes in cross-sectional and time dimensions, respectively. A large T guarantees the NN is close enough to the query point x_0, and a large n provides enough observations from the tail. Given the new EV theory, we show how existing suggestions on inference about *unconditional* tail properties can be applied using the induced Y's as input. In particular, we consider both the fixed-k asymptotic inference proposed by Müller and Wang (2017) and the methods developed by Hill (1975) and Smith (1987), which are two leading examples in the numerous increasing-k asymptotic methods. The number k denotes how many largest (smallest) observations are used to approximate the tail. The fixed-k approach is more suitable for a moderate n, say 200, while the increasing-k ones have computational advantage when n is much larger.

In summary, the main idea is a combination of the NN in the time dimension and the EV theory in the cross-sectional dimension. This approach only requires some smoothness condition on the joint distribution and hence is much more flexible than existing methods. A natural question is how much efficiency we lose by using only one out of T observations in each time series. It turns out that if the tail shape depends on the covariate highly nonlinearly, the new NN method dominates existing methods in both coverage and length when T is only moderately large, say 50. When T is very large, say 500, the new CIs also deliver comparable lengths to the kernel regression method with the optimal bandwidth. See the Monte Carlo results in Section 4 for more details.

As a by-product, we also study the tail features of the coefficients in a random coefficient regression model. In particular, suppose Y_{it} and X_{it} are generated from the model $Y_{it} = \alpha_i + X_{it}^{\intercal}\beta_i + u_{it}$, where $(\alpha_i, \beta_i^{\intercal})^{\intercal}$ is a random vector drawn from some unknown distribution. We first construct the least squares estimators of α_i and β_i using the i-th time series for all i and collect the largest order statistics from these estimates. Then we show that the estimation error is negligible under the large n and large T framework, and hence the largest (smallest) order statistics among these estimates again satisfy the desired EV theory, which further supports the application of the fixed-k CIs for the tail features of α_i and β_i. This complements the existing literature focusing on the mid-sample properties of heterogeneous effects (e.g., Hsiao and Pesaran (2004) and Wooldridge (2005)).

Applying the proposed methods, we study the tail features of the U.S. monthly stock returns and find strong tail dependence and heterogeneity. In particular, we first construct robust CIs of the conditional extreme quantiles given different stock sizes. The CIs exclude the quantile regression estimators, suggesting that the location-scale model (1) might suffer from some misspecification error. Second, we implement the Fama-French three factor regression model and construct the robust CIs of the extreme quantiles of the regression coefficients, namely, the alpha's and beta's. The CIs for the left and right extreme quantiles are far from overlapping, which implies that the alpha's and beta's are strongly heterogeneous across stocks.

The rest of the paper is organized as follows. Section 2 reviews the classic EV theory for unconditional features, establishes a new one for conditional features, which formalizes the NN approximation, and extends the study to the random coefficient regression model. In light of the main results from Section 2, Section 3 illustrates how to construct new CIs based on the NN and existing approaches for unconditional tail problems. Section 4 implements an extensive Monte Carlo study, which shows that the new CIs have excellent small sample coverage and length properties for moderately large sample sizes. Section 5 applies the new method to the U.S. monthly stock returns. Section 6 concludes with proofs and omitted details collected in the Appendix.

Notation Let $\xrightarrow{p}$ denote convergence in probability and $\xrightarrow{d}$ denote convergence in distribution as $n, T \to \infty$. Let $\mathbf{1}[A]$ denote the indicator function of a generic event A. Let $||B||$ denote the Euclidean norm of a vector or matrix B, and let C denote a generic constant whose value may change across lines. Let $B_\delta(x)$ denote a generic open ball centered at x with radius δ.

2 Main result

We start with reviewing the classic unconditional EV theory in Section 2.1, and then establish a conditional one in the Section 2.2. Finally in Section 2.3, we extend the analysis to the random coefficient regression model.

2.1 Review of the unconditional EV theory

Consider a random sample $Y_1, Y_2, ..., Y_n$ from some population with cumulative distribution function (CDF) F_Y. Denote the marginal quantile function as $Q_Y(\tau) = \inf\{y : F_Y(y) \geq \tau\}$ for some $\tau \in [0,1]$. Let $Y_{(1:n)} \geq Y_{(2:n)} \geq \ldots \geq Y_{(n:n)}$ denote the order statistics, so that $Y_{(1:n)}$ is the sample maximum. From now on, we consider the right tail and suppress ":n" in the subscript for notational ease.

The fundamental result in EV theory is developed by Fisher and Tippett (1928) and Gnedenko (1943), stating that if there exist sequences a_n and b_n such that

$$\frac{Y_{(1)} - b_n}{a_n} \xrightarrow{d} V_1 \text{ as } n \to \infty \tag{3}$$

for some nondegenerate random variable V_1, then the distribution of V_1 is, up to location and scale normalization, the generalized EV distribution with the CDF

$$G_\xi(v) = \begin{cases} \exp(-(1+\xi v)^{-1/\xi}), & 1 + \xi v \geq 0, \text{ for } \xi \neq 0 \\ \exp(-e^{-v}), & v \in \mathbb{R}, \, \xi = 0 \end{cases} \tag{4}$$

where $\xi \in \mathbb{R}$ is the tail index.

EV theory holds if and only if F_Y is within the domain of attraction (DOA) of G_ξ, denoted as $F_Y \in \mathcal{D}(G_\xi)$. This is a very mild assumption as it is satisfied by most commonly used distributions. In particular, the positive ξ case covers the distributions with a Pareto-type tail such as Pareto, student's t, and F. The case with $\xi = 0$ covers the distributions with finite moments of any order. Leading examples are normal and log-normal. The case with a negative ξ covers the distributions with a bounded right end-point, that is, $Q_Y(1) < \infty$. See Chapter 1 in de Haan and Ferreira (2007) for a complete review.

Without loss of generality, assume that any location and scale normalization of V_1 is subsumed in a_n and b_n, so that the CDF of V_1 is equal to G_ξ. It is well known (see, for instance, Theorem 3.5 of Coles (2001)) that if (3) holds, then EV theory also holds jointly for the first k order statistics:

$$\begin{pmatrix} \frac{Y_{(1)}-b_n}{a_n} \\ \vdots \\ \frac{Y_{(k)}-b_n}{a_n} \end{pmatrix} \xrightarrow{d} \mathbf{V} = \begin{pmatrix} V_1 \\ \vdots \\ V_k \end{pmatrix} \tag{5}$$

for any fixed k, where the joint probability density function (PDF) of $\mathbf{V}$ is given by

$$f_{\mathbf{V}|\xi}(v_1, \ldots, v_k) = G_\xi(v_k) \prod_{i=1}^{k} g_\xi(v_i)/G_\xi(v_i) \tag{6}$$

for $v_k \leq v_{k-1} \leq \ldots \leq v_1$ with $g_\xi(v) = \partial G_\xi(v)/\partial v$, and zero otherwise. Note that the constants a_n and b_n depend on ξ and are difficult to estimate. For example, a_n is n^ξ if F_Y is standard Pareto. Since a small estimation error in ξ is amplified by the n-power, inference relying on a good estimate of ξ and the scale usually requires a large k and a even larger sample size n.

The $G_\xi(v_k)$ term in (6) suggests that the largest k order statistics are not asymptotically independent, given any fixed k. In contrast, this term is negligible if k increases with n, and then $\mathbf{V}$ can be considered as independent draws from the generalized Pareto distribution (GPD) (see Section 3.2 for details).

Based on EV theory, tail features such as extreme quantile can be expressed as known functions of ξ. The inference problem is asymptotically equivalent to the parametric one in which we have k observations drawn from the EV distribution or the GPD and aim for CIs of a function of the single parameter ξ. There have been numerous suggestions on estimation and inference along this line. Depending on whether the asymptotic embedding assumes a fixed or increasing k, we refer to them as the fixed-k or increasing-k approaches, respectively. We discuss them in Sections 3.1 and 3.2. Now, we proceed to study conditional tails and establish a new EV theory with panel data.

2.2 The conditional EV theory

We now present the main result of this paper in this subsection. Let X denote a $\dim(X) \times 1$ vector of continuous random variables with uniformly positive joint PDF.[3] The objects of interest are the tail features of the conditional distribution $F_{Y|X=x}(\cdot) = \mathbb{P}(Y \leq \cdot|X = x)$. To fix idea, we focus on the conditional extreme quantile of Y given X taking the value x_0, that is, $Q_{Y|X=x_0}(\tau)$ for some pre-specified $x_0 \in \mathbb{R}^{\dim(X)}$ and some τ close to 1.

In contrast to the unconditional case, observations from the conditional CDF are not available in a cross-sectional dataset. We overcome this issue by using panel data. Consider a balanced[4] panel dataset $\{Y_{it}, X_{it}\}_{i=1:n,t=1:T}$ that is i.i.d. across i and strictly stationary and weakly dependent across t. Our approach is implemented by the following three steps.

Step 1 Collect, for each i, the induced Y associated with the NN of $\{X_{it}\}_{t=1}^{T}$ to x_0, where the NN is measured by the Euclidean distance $||X_{it} - x_0||$. Denote them as $\{Y_{i,[x_0]}\}_{i=1}^{n}$.

Step 2 Take the largest k order statistics from $\{Y_{i,[x]}\}_{i=1}^{n}$ and denote them as

$$\mathbf{Y} = (Y_{(1),[x_0]}, Y_{(2),[x_0]}, ..., Y_{(k),[x_0]})^\intercal, \tag{7}$$

where $Y_{(1),[x_0]} \geq Y_{(2),[x_0]} \geq \ldots \geq Y_{(n),[x_0]}$ are the order statistics of $\{Y_{i,[x_0]}\}_{i=1}^{n}$.

Step 3 Use $\mathbf{Y}$ as input to apply either the fixed-k or the increasing-k approaches to inference reviewed in Section 3.

The main result of this article is summarized in Theorem 1 below, which states that $\mathbf{Y}$ defined (7) satisfies a similar convergence as in (5). The key idea is heuristically illustrated by the following derivation. For each i, denote the NN among $\{X_{it}\}_{t=1}^{T}$ to x_0 as $X_{i,(x_0)}$. Then for any $y \in \mathbb{R}$,

$$\begin{aligned}
& \mathbb{P}\left(Y_{i,[x_0]} \leq y\right) \\
= \; & \mathbb{E}_{X_{i,(x_0)}}\left[\mathbb{P}\left(Y_{i,[x_0]} \leq y | X_{i,(x_0)}\right)\right] \\
= \; & \mathbb{E}_{X_{i,(x_0)}}\left[F_{Y|X=X_{i,(x_0)}}(y)\right] \text{ (by strict stationarity)} \\
= \; & F_{Y|X=x_0}(y) + \mathbb{E}_{X_{i,(x_0)}}\left[\left.\frac{\partial F_{Y|X=x}(y)}{\partial x^\intercal}\right|_{x=\dot{x}_i}(X_{i,(x_0)} - x_0)\right] \text{ (by mean value expansion)}
\end{aligned}$$

[3]If X contains discrete components, our method is readily applicable by considering the subsample that these discrete variables take their discrete query values.

[4]This is only for notational ease. The new approach is valid as long as T is large for all i.

$$\rightarrow \quad F_{Y|X=x_0}(y) \text{ as } T \rightarrow \infty,$$

where $\dot{x}_i$ lies on the segment connecting $X_{i,(x_0)}$ and x_0. The first equation is by the definition of conditional expectation. The second one is established as Theorem 2.1 in Yang (1977) in the i.i.d. case. It also holds under much more general dependence conditions as long as strict stationarity is maintained. The third equation is valid if the conditional CDF is smooth. The last convergence holds if the NN converges to its query point x_0 and if the CDF is smooth again with bounded derivatives.

The above derivation states that the collection of the induced order statistics Y associated with the NN to x_0 can be treated as approximately stemming from the true conditional CDF $F_{Y|X=x_0}$ asymptotically. Thus the largest (cross-sectional) order statistics $\mathbf{Y}$ can be treated as draws from the tail of $F_{Y|X=x_0}$. Given the assumption about the maximum domain of attraction, the problem reduces back to its unconditional analogue and hence existing suggestions on unconditional tail problems become applicable. Note that the normalizing constants a_n and b_n now depend on $\xi(x)$ evaluated at $x = x_0$.

A formal establishment requires the following conditions.

Condition 1.1 $(Y_{i1}, X_{i1}^{\intercal})^{\intercal}, \ldots, (Y_{iT}, X_{iT}^{\intercal})^{\intercal}$ are i.i.d. across i. $(Y_{it}, X_{it}^{\intercal})^{\intercal}$ for each $t = 1, \ldots, T$ is strictly stationary and β-mixing with the mixing coefficient satisfying $\beta(t) = O(t^{-2-\varepsilon})$ for some $\varepsilon > 0$. In addition, $f_X(x)$ is uniformly continuously differentiable and bounded away from 0 in an open ball centered at x_0.

Condition 1.1 requires the data to be independent across i and weakly dependent across t. In addition, it also requires the density of X to be positive in an open neighborhood around the query point x_0. This condition is sufficient to establish that the NN converges to the query point x_0 almost surely at some power rate. For readability, we formalize this result in Lemma 1 in Appendix A.1. Note that we intentionally choose only one NN to allow for weak dependence across t. If data are independent across both i and t, more than one NNs can potentially be chosen to enlarge the effective sample. We leave this for future research.

Condition 1.2 $F_{Y|X=x_0} \in \mathcal{D}\left(G_{\xi(x_0)}\right)$ with $\xi(x_0) \geq 0$.

This condition requires that the underlying conditional distribution is in the domain of attraction of the generalized EV distribution. This is a mild condition as it is satisfied by many commonly used joint distributions. In particular, it generalizes the conditional location-scale shift model (1) by allowing $\mu(x), \sigma(x)$, and $\xi(x)$ to be all unknown (but smooth) functions of x. The condition on a non-negative $\xi(x_0)$ is for illustrational ease since the Y in most applications involving tail features has an unbounded support.[5] To illustrate the mildness of this condition, we discuss the following three examples. In particular, Condition 1.2 is satisfied in all three of them but the location-scale model assumption (1) is not.

Example 1 (Joint Normal) Suppose (Y, X) are joint normal with zero means, unit variances, and correlation ρ. Then Y given $X = x$ is normal with mean ρx, and variance $1 - \rho^2$. The conditional tail index is $\xi(x) = 0$ for all $x \in \mathbb{R}$. The conditional quantile is $Q_{Y|X=x}(\tau) = \rho x + \sqrt{1-\rho^2}\Phi^{-1}(\tau)$, where $\Phi^{-1}(\cdot)$ is the quantile function of the standard normal distribution. Thus the location-scale model assumption (1) is satisfied.

Example 2 (Joint Student's t) Suppose (Y, X) are jointly student's t distributed with d.f. v, zero means, unit variances, and correlation $\rho \neq 0$. Then Y given $X = x$ is student's t distributed with d.f. $v + 1$, mean ρx, and variance $(1 - \rho^2)(v + x^2)/(v + 1)$. The conditional tail index is $\xi(x) = 1/(v + 1)$ for all $x \in \mathbb{R}$.[6] The conditional quantile is $Q_{Y|X=x}(\tau) = \rho x + \sqrt{(1-\rho^2)(v+x^2)/(v+1)}Q_{t(v)}(\tau)$, where $Q_{t(v)}(\cdot)$ is the quantile function of the standard student's t distribution with d.f. v. This specification satisfies the location-scale shift model (1) but the scale function is highly nonlinear in x.

Example 3 (Conditional Pareto) Suppose X is half-normal with positive support and Y given $X = x$ is the Pareto distribution such that $\mathbb{P}(Y \leq y|X = x) = 1-(y+1)^{-1/x}$ for $y \geq 0$ and any $x > 0$. Then the conditional tail index is $\xi(x) = x$ and the conditional quantile is $Q_{Y|X=x}(\tau) = -1 + (1-\tau)^{-x}$, which violates the location-scale shift model (1).

Let y_0 denote the end-point of the conditional CDF, that is, $y_0 = Q_{Y|X=x_0}(1) \leq \infty$. The next assumption is a high level regularity condition on the tail of the conditional CDF, whose primitive conditions are discussed in Appendix A.1.

Condition 1.3 $f_{Y|X=x}(y)$ is uniformly bounded and continuously differentiable in x and y. In addition, for any fixed $y > 0$ with $u_n = a_n y + b_n \to y_0$, and any open ball $B_{\eta_T}(x_0)$ centered at x_0 with radius $\eta_T \equiv O(T^{-\eta})$

[5] The derivation for the $\xi(x_0) < 0$ case requires additional assumptions on the relative magnitudes of T and n, and is suppressed for illustrational ease.

[6] See Ding (2016) for the exact expression for the PDF.

for some $\eta > 0$, $\lim_{u_n \to y_0} \sup_{x \in B_{\eta_T}(x_0)} T^{-\eta} \left\| \frac{\partial F_{Y|X=x}(u_n)/\partial x}{1-F_{Y|X=x_0}(u_n)} \right\| = 0$ and $\lim_{u_n \to y_0} \sup_{x \in B_{\eta_T}(x_0)} T^{-\eta} \left\| \frac{\partial f_{Y|X=x}(u_n)/\partial x}{f_{Y|X=x_0}(u_n)} \right\| = 0$ as $n \to \infty$ and $T \to \infty$.

Condition 1.3 requires that the derivatives of the conditional CDF and PDF are smooth and decay quickly. This is a mild condition again, which is satisfied by the above examples by straightforward calculation. We give details in Appendix A.1.

Condition 1.4 $n \to \infty$, $T \to \infty$, and $T/n \to \lambda$ for some $\lambda \in (0, \infty)$.

Condition 1.4 requires both n and T to be large. A large n guarantees that the error due to the EV approximation is negligible, and a large T controls the distance between the NN and the query point. The parameter λ can be any positive constant, and hence T can be much smaller than n.

Given the above conditions, we establish the following theorem that formalizes the idea of treating the induced Y's as draws from the true conditional distribution.

Theorem 1 *Under Conditions 1.1-1.4, there exist sequences of constants $a_n > 0$ and b_n depending on x_0 such that*

$$\frac{\mathbf{Y} - b_n}{a_n} \xrightarrow{d} \mathbf{V} \tag{8}$$

where $\mathbf{Y}$ is defined in (7) and $\mathbf{V}$ is jointly EV distributed with PDF (6) and $\xi = \xi(x_0)$.

Theorem 1 is the main result of this paper. It allows us to apply the existing approaches to inference about unconditional tail features for the purpose of inference about conditional tail features. In light of the convergence (8), $\mathbf{Y}$ can be simply treated as input for the existing approaches. See Section 3 for details. On the one hand, the new method is robust against the misspecification mentioned in the introduction since no functional form is imposed. On the other hand, the limit observation $\mathbf{V}$ has a known PDF, which renders an efficiency gain compared with the fully nonparametric methods. This feature is demonstrated in Section 4 by Monte Carlo experiments. Before that, we apply our proof strategy for the main result to a widely studied case: the random coefficient regression model in the following subsection.

2.3 Extension to the random coefficient regression model

Consider the following panel data model

$$Y_{it} = \alpha_i + X_{it}^{\intercal} \beta_i + u_{it}, \tag{9}$$

where $(\alpha_i, \beta_i^\intercal)^\intercal$ denotes the random coefficients and u_{it} the error term. This setup covers the classic panel regression model with fixed effects in which $\beta_i = \beta_0$ for all i (cf. Chapter 10 of Wooldridge (2002)). As long as Conditions 1.1-1.4 are satisfied, the previously introduced methods naturally apply here. In addition, the parametric assumption (9) allows us to conduct inference about the tail features of α_i and β_i, which we illustrate now.

Let $(\hat{\alpha}_i, \hat{\beta}_i^\intercal)^\intercal$ be the OLS estimator by regressing Y_{it} on $(1, X_{it}^\intercal)^\intercal$ using the time series associated with the i-th individual. Collect $\{(\hat{\alpha}_i, \hat{\beta}_i^\intercal)^\intercal\}_{i=1}^n$ and sort each series of estimates descendingly. Then we define

$$\mathbf{A} = (\hat{\alpha}_{(1)}, ..., \hat{\alpha}_{(k)})^\intercal,$$

that is, the largest k order statistics of $\{\hat{\alpha}_i\}$, and

$$\mathbf{B} = (\hat{\beta}_{j,(1)}, ..., \hat{\beta}_{j,(k)})^\intercal,$$

that is, the largest k order statistics of the j-th components of $\{\hat{\beta}_i\}_{i=1}^n$. Without loss of generality, we focus on the first component of β_i and suppress the subscript j.

The following conditions are imposed for the convergence of $\mathbf{A}$ and $\mathbf{B}$.

Condition 2.1 $(\alpha_i, \beta_i^\intercal, u_{it}, X_{it}^\intercal)^\intercal$ are i.i.d. across i and strictly stationary and weakly dependent across t;

Condition 2.2 $F_\alpha \in \mathcal{D}\left(G_{\xi_\alpha}\right)$ and $F_\beta \in \mathcal{D}\left(G_{\xi_\beta}\right)$ with $\xi_\alpha \geq 0$ and $\xi_\beta \geq 0$;

Condition 2.3 $\sup_i ||\hat{\beta}_i - \beta_i|| = o_p(1)$, $\sup_i |\bar{u}_i| = o_p(1)$, and $\sup_i \left|\left|\bar{X}_i\right|\right| = O_p(1)$, where $\bar{X}_i = T^{-1}\sum_{t=1}^T X_{it}$ and $\bar{u}_i = T^{-1}\sum_{t=1}^T u_{it}$. In addition, if $\xi_\omega = 0$, $\sup_i ||\hat{\beta}_i - \beta_i||/f_\omega\left(Q_\omega(1-1/n)\right) = o_p(1)$ and $\sup_i |\bar{u}_i|/f_\omega\left(Q_\omega(1-1/n)\right) = o_p(1)$ for $\omega = \alpha$ or β, where $Q_\omega(\cdot)$ and $f_\omega(\cdot)$ denote the quantile function and the PDF of ω, respectively.

Conditions 2.1-2.3 are again mild. In particular, Condition 2.1 is similar to Condition 1.1. Since the objective of interest is the unconditional tail features of α_i and β_i, we do not need the NN condition on the covariate. The dependence structure is left unspecified as long as it is sufficient for Condition 2.3. Condition 2.2 assumes the distributions of α_i and β_i are respectively in the domain of attraction of G_{ξ_α} and G_{ξ_β} with non-negative tail indices. Condition 2.3 requires the estimator $\hat{\beta}_i$ to be consistent for all i and the moments of sample averages of u_{it} and X_{it} across t are bounded. If the tail index is zero, these bounds need to be stronger to accommodate the fact that $a_n \to 0$.[7]

[7]Straightforward calculation yields that normal distribution satisifies Condition 2.3, if $\sup_i ||\hat{\beta}_i -$

Corollary 1 *Under Conditions 2.1-2.3, there exist sequences of constants* $a_n > 0$ *and* b_n *that depend on* ξ_α *or* ξ_β *such that*

$$\frac{\mathbf{A} - b_n(\xi_\alpha)}{a_n(\xi_\alpha)} \xrightarrow{d} \mathbf{V}(\xi_\alpha)$$

and

$$\frac{\mathbf{B} - b_n(\xi_\beta)}{a_n(\xi_\beta)} \xrightarrow{d} \mathbf{V}(\xi_\beta),$$

where $\mathbf{V}(\cdot)$ *are jointly EV distributed with PDF (6) and tail index* ξ_α *or* ξ_β*.*

Under Conditions 2.1-2.3, Corollary 1 establishes the desired convergence of the largest order statistics among $\{\hat{\alpha}_i\}$ and $\{\hat{\beta}_i\}$. Thus, we can apply the methods reviewed in the following section to conduct inference about the tail features of α_i and β_i.

3 Construction of CIs for conditional tail features

Our main result presented in the previous section formalizes the NN approximation in the panel data setup. This NN approach facilitates application of existing methods of unconditional tail inference for the purpose of inference about conditional tails. In this section, we illustrate how existing inference approaches can be applied for construction of CIs for conditional tail features in light of our NN approximation result. We consider two types of approaches, which are based on either the fixed-k asymptotics (Section 3.1) or the increasing-k asymptotics (Section 3.2).

3.1 The fixed-k approaches

We first consider the fixed-k approaches proposed by Müller and Wang (2017) and Wang and Xiao (2019). The major advantage of these approaches is their excellent performance in coverage probability and length given relatively small sample sizes, say $n = 200$. Literally, the fixed-k approaches rely on the asymptotic embedding that the largest k observations converge in distribution to the joint EV distributed vector $\mathbf{V}$. If the tail feature under investigation is also asymptotically equivalent to some known function of ξ, we essentially

$\beta_i|| = O_p(T^{-\varepsilon})$ and $\sup_i |\bar{u}_i| = O_p(T^{-\varepsilon})$ for some $\varepsilon > 0$ and if $n/T \to \lambda$ for some $\lambda \in (0, \infty)$. This is seen by $1/f_\alpha\left(Q_\alpha\left(1 - 1/n\right)\right) \leq O(\log(n))$ when f_α and Q_α are standard normal PDF and quantile functions, respectively (cf. Example 1.1.7 in de Haan and Ferreira (2007)).

end up with a straightforward parametric problem: construction of CIs for a function of ξ with one draw from $\mathbf{V}$ whose PDF is characterized only by ξ.

Suppose we have the dataset as in Section 2.2 and aim for a $1-\alpha$ CI for the conditional extreme quantile $Q_{Y|X=x_0}(\tau)$ for τ close to 1. To be precise, we rewrite τ as $1-h/n$ for some $h>0$ as similarly considered in Chernozhukov (2005) and Chernozhukov and Fernández-Val (2011). This setup means that the extreme quantile is of the same order of the sample maximum from n random draws from the true conditional CDF $F_{Y|X=x_0}$. Such an extreme quantile is too far in the tail for the normal approximation to perform well.

Following Steps 1-3 in Section 2.2, the effective data becomes $\mathbf{Y}$ as in (7) and the objective is to construct a confidence set $S(\mathbf{Y}) \subset \mathbb{R}$ such that $\mathbb{P}(Q_{Y|X=x_0}(\tau) \in S(\mathbf{Y})) \geq 1-\alpha$, at least as $n \to \infty$ and $T \to \infty$. In particular, EV theory suggests that

$$\frac{Q_{Y|X=x_0}(1-h/n) - b_n}{a_n} \to q(\xi, h) \equiv \begin{cases} \frac{h^{-\xi}-1}{\xi} & \text{if } \xi \neq 0 \\ -\log(h) & \text{if } \xi = 0 \end{cases}$$

where we suppress x_0 in ξ for notational ease. Note that $q(\xi,h)$ is the $\exp(h)$ quantile of V_1. The normalizing constants a_n and b_n also implicitly depend on ξ and hence are unknown. Since they are shared by both $\mathbf{Y}$ and $Q_{Y|X=x_0}(1-h/n)$, we can impose location and scale equivariance on the CI to cancel them out. Specifically, we impose that for any constants $a>0$ and b, $S(a\mathbf{Y}+b) = aS(\mathbf{Y})+b$, where $aS(\mathbf{Y})+b = \{y : (y-b)/a \in S(\mathbf{Y})\}$. Under this equivariance constraint, we can write

$$\begin{aligned} & \mathbb{P}(Q_{Y|X=x_0}(1-h/n) \in S(\mathbf{Y})) \\ = \;& \mathbb{P}\left(\frac{Q_{Y|X=x_0}(1-h/n) - Y_{(k),[x_0]}}{Y_{(1),[x_0]} - Y_{(k),[x_0]}} \in S\left(\frac{\mathbf{Y} - Y_{(k),[x_0]}}{Y_{(1),[x_0]} - Y_{(k),[x_0]}}\right)\right) \\ \to \;& \mathbb{P}_\xi\left(Y^* \in S(\mathbf{V}^*)\right), \end{aligned}$$

where we introduce the self-normalized statistics

$$\begin{aligned} Y^* &= \frac{q(\xi,h) - V_k}{V_1 - V_k} \\ \mathbf{V}^* &= \left(\frac{V_1 - V_k}{V_1 - V_k}, \frac{V_2 - V_k}{V_1 - V_k}, \ldots, \frac{V_k - V_k}{V_1 - V_k}\right), \end{aligned}$$

and highlight with the subscript ξ that the densities of Y^* and $\mathbf{V}^*$ now depend solely on ξ. They can be computed by using (5), (6), and change of variables.

Given only a finite number of observations, a consistent estimation of ξ is out of the question. Instead of imposing the correct size control for the true ξ, we impose it for all the values of ξ that are empirically relevant. In this sense the fixed-k approach is conservative but more robust to misspecification, especially when the sample size is not large enough to support a precise estimation of ξ. Let $\Xi \subset \mathbb{R}$ be the set of tail indices for which we impose the asymptotically correct coverage.[8] The asymptotic problem then is to construct a location and scale equivariant S that satisfies

$$\mathbb{P}_{\xi}\left(Y^* \in S(\mathbf{V}^*)\right) \geq 1-\alpha \text{ for all } \xi \in \Xi, \tag{10}$$

since any S that satisfies (10) also satisfies $\liminf_{n\to\infty, T\to\infty} \mathbb{P}(Q_{Y|X=x_0}(1-h/n) \in S(\mathbf{Y})) \geq 1-\alpha$ under (8) and the continuous mapping theorem. Among all solutions to this problem, we choose the optimal one that minimizes the weighted average expected length criterion

$$\int \mathbb{E}_{\xi}[\text{lgth}(S(\mathbf{V}))]dW(\xi), \tag{11}$$

where W is a positive measure with support on Ξ,[9] and $\text{lgth}(A) = \int \mathbf{1}[y \in A]dy$ for any Borel set $A \subset \mathbb{R}$. The equivariance of S further implies $\mathbb{E}_{\xi}[\text{lgth}(S(\mathbf{V}))] = \mathbb{E}_{\xi}[(V_1 - V_k)\,\text{lgth}(S(\mathbf{V}^*))]$. Thus the program of minimizing (11) subject to (10) among all equivariant sets S asymptotically becomes

$$\begin{array}{l} \min_{S(\cdot)} \int_{\Xi} \mathbb{E}_{\xi}[(V_1 - V_k)\,\text{lgth}(S(\mathbf{V}^*))]dW(\xi) \\ \text{s.t. } \mathbb{P}_{\xi}(Y^* \in S(\mathbf{V}^*)) \geq 1-\alpha \text{ for all } \xi \in \Xi. \end{array} \tag{12}$$

Note that the above expectation and probability are w.r.t. the distribution of Y^* and $\mathbf{V}^*$. This distribution depends on $\xi(x)$ evaluated at x_0. Solution to problem (12) is numerically calculated with the corresponding MATLAB program provided on the author's website. The computation cost is only several seconds using a modern PC.

In addition to the conditional extreme quantile, we can also construct a fixed-k CI for the conditional tail index $\xi(x_0)$. Consider the testing problem

$$\begin{aligned} H_0 &: \xi(x_0) = \xi_0 \text{ against} \\ H_1 &: \xi(x_0) \in \Xi \backslash \{\xi_0\}. \end{aligned}$$

With some weighting function W again, the likelihood ratio test is constructed as

$$\varphi(\mathbf{v}^*) = \mathbf{1}\left[\frac{\int_{\Xi} f_{\mathbf{V}^*|\xi}(\mathbf{v}^*)dW(\xi)}{f_{\mathbf{V}^*|\xi_0}(\mathbf{v}^*)} > \text{cv}_{(\alpha;\xi_0)}\right], \tag{13}$$

[8] We use $\Xi = [-1/2, 1/2]$ for inference about conditional extreme quantiles in later applications, which covers all the distributions with finite variance. This range can be easily extended.

[9] We use the uniform weight in later sections.

where $\mathbf{1}[\cdot]$ is the indicator function, and $\text{cv}_{(\alpha;\xi_0)}$ denotes the critical value that depends on the significance level α and the null value ξ_0. The CI for ξ is then obtained by inverting this test. Theorem 1 and the continuous mapping theorem again yield that the test (13) and the corresponding CI are asymptotically valid.

Other tail related quantities such as the conditional tail expectation are also covered as long as they can be expressed as functions of the conditional tail index.

3.2 The increasing-k approaches

If the cross-sectional sample size n is large enough, we can choose a large k and switch to the approaches based on the increasing-k asymptotics (see de Haan and Ferreira (2007) for an overview). In particular, we consider two popular methods developed respectively by Hill (1975) and Smith (1987), and show that they are applicable with $\mathbf{Y}$ being the input if k is large. Application of other methods is also possible but requires a method-specific consideration.

For the unconditional distribution, Pickands (1975) states that if $F_Y \in \mathcal{D}(G_\xi)$, the generalized Pareto distribution is a good approximation of the tail of F_Y in the sense that

$$\lim_{v_n\to\infty} \sup_{0<y<\infty} \left| \frac{F_Y(y+v_n) - F_Y(v_n)}{1-F_Y(v_n)} - F_{\text{GP}}(y;\xi,\sigma) \right| = 0, \tag{14}$$

where v_n denotes the cutoff of tail and

$$F_{\text{GP}}(y;\xi,\sigma) = \begin{cases} 1-(1+\xi y/\sigma)^{-1/\xi} & \text{if } \xi \neq 0 \\ 1-\exp(y/\sigma) & \text{if } \xi = 0. \end{cases}$$

The cutoff v_n determines the number of tail observations k in the way that $Y_{(k),[x_0]} \geq v_n$ and $Y_{(k+1),[x_0]} < v_n$. Thus, there is little difference in choosing v_n to determine k or the other way.

Given the choice of v_n (and accordingly k), Hill's estimator can be constructed in the panel framework as

$$\hat{\xi}_H = \frac{1}{k}\sum_{i=1}^{k}(\log Y_{(i),[x_0]} - \log v_n).$$

As an alternative, Smith (1987) suggests fitting the differences between the largest k observations and the cutoff v_n to $F_{\text{GP}}(y;\xi,\sigma)$ and constructing the maximum likelihood estimator of ξ and σ. In particular, this estimator can be implemented as

$$(\hat{\xi}_{ML}, \hat{\sigma}_{ML}) = \arg\max_{\xi\in\Xi,\sigma\in\mathbb{R}^+} \sum_{i=1}^{k} \log(f_{\text{GP}}(Y_{(i),[x_0]} - v_n;\xi,\sigma)) \tag{15}$$

where $f_{\mathrm{GP}}(y;\xi,\sigma) = \partial F_{\mathrm{GP}}(y;\xi,\sigma)/\partial y$.

Both Hill's and Smith's estimators are root-k consistent and asymptotically normal, provided that v_n grows at a certain rate. Since Hill's estimator is only defined for positive tail indices, we restrict $\xi(x_0)$ to be positive. For notational ease, we write $\gamma(x) = 1/\xi(x)$ and $\tilde{\gamma}(x) = 1/\tilde{\xi}(x)$. The following additional conditions are imposed.

Condition 3.1 $1 - F_{Y|X=x}(y) = c(x)y^{-\gamma(x)}(1 + d(x)y^{-\tilde{\gamma}(x)} + r(x,y))$ uniformly as $y \to \infty$ where $c(\cdot) > 0$ and $d(\cdot)$ are continuously differentiable functions and uniformly bounded between 0 and ∞, $\gamma(\cdot) > 0$ and $\tilde{\gamma}(\cdot) > 0$ are continuously differentiable functions, and $r(x,y)$ is continuously differentiable with bounded derivatives w.r.t. both x and y and satisfies $\limsup_{y\to\infty} \sup_{x\in B_\delta(x_0)} |r(x,y)/y^{-\tilde{\gamma}(x)}| = 0$ for some δ.

Condition 3.2 $\sqrt{k}\tilde{\gamma}(x_0)d(x_0)v_n^{-\tilde{\gamma}(x_0)}/\left(\gamma\left(x_0\right) + \tilde{\gamma}\left(x_0\right)\right) \to \mu(x_0)$ for some constant $\mu(x_0) \in \mathbb{R}$.

Condition 3.1 is a second order condition on the GPD approximation. In particular, the conditional CDF is approximated by a GPD in the first order, and the parameter $\tilde{\xi}$ governs the approximation bias. This condition is commonly assumed to study unconditional tail problems (see, for example, Hall (1982), Smith (1987), and Chernozhukov (2005)). Condition 3.2 specifies the choice of the tail cutoff that leads to a non-degenerate asymptotic bias in the tail index estimator (cf. eq. (3.3) in Smith (1987)). This is seen in the following proposition.

Proposition 1 *Suppose Conditions 1.1, 1.4, 3.1 and 3.2 hold with $\xi(x_0) > 0$. Then*

$$\sqrt{k}(\hat{\xi}_H - \xi) \xrightarrow{d} \mathcal{N}\left(\mu_H, \xi^2\right),$$

and

$$\sqrt{k}(\hat{\xi}_{ML} - \xi) \xrightarrow{d} \mathcal{N}\left(\mu_{ML}, (1+\xi)^2\right)$$

where

$$\begin{aligned} \mu_H &= -\mu\xi \\ \mu_{ML} &= -\frac{\mu(1+\xi)\xi(1-\tilde{\gamma})}{1+\xi-\xi\tilde{\gamma}} \end{aligned}$$

and $(\xi, \mu, \tilde{\gamma})$ are evaluated at x_0.

Proposition 1 derives the asymptotic distributions of Hill's and Smith's estimators, which are identical to those established in Goldie and Smith (1987) and Smith (1987), respectively.

The bias terms are difficult to estimate since they involve the second order parameters $\mu(x_0)$ and $\tilde{\gamma}(x_0)$. A feasible CI for $\xi(x_0)$ is then constructed by choosing k of a smaller order than specified in Condition 3.2 so that the bias is asymptotically zero (cf. Theorem 2 in Hall (1982)). This is similar to the undersmoothing choice of the bandwidth in kernel estimation. The CIs are obtained accordingly by plugging in the tail index estimate for the variance.

4 Monte Carlo results

In this section we run Monte Carlo experiments to examine the small sample performance of the new approach. Section 4.1 considers the simple panel data $\{Y_{it}, X_{it}\}$ without any fixed effect. In Section 4.2, we compare the efficiency of the new approach with a kernel estimator, which essentially uses more than one NNs. In Sections 4.3, we impose the linear regression setup (9) with classic individual fixed effects or random coefficients. Finally in Section 4.4, we present the results on inference about the conditional tail index.

4.1 Conditional extreme quantile

We continue to consider the three examples in Section 2.2 as the data generating processes (DGPs). In all experiments, data are i.i.d. across i. The dependence structure across t is as follows.

1. **Joint Normal** $X_{it} = \rho X_{it-1} + u_{it}$ with $u_{it} \sim^{iid} \mathcal{N}(0, 1-\rho^2)$ and $X_{i1} \sim \mathcal{N}(0,1)$. $Y_{it} = r_{xy} X_{it} + \sqrt{1-r_{xy}^2} v_{it}$ where $v_{it} \sim^{iid} \mathcal{N}(0,1)$ and independent of u_{it}. Set $\rho = 0.5$ and $r_{xy} = 0.5$.

2. **Joint Student's t** (X_{it}, Y_{it}) is i.i.d. across t and distributed as $t_v(\mu, \Sigma)$ with $v = 3$, $\mu = [0,0]^\intercal$, and $\Sigma = [1, 0.5; 0.5, 1]$.

3. **Conditional Pareto** $X_{it} = \rho X_{it-1} + u_{it}$ with $u_{it} \sim^{iid} \mathcal{N}(0, (1-\rho^2))$ and $X_{i1} \sim \mathcal{N}(0,1)$. $Y_{it}|X_{it} = x \sim \text{Pa}(\xi(x))$, that is, $\mathbb{P}(Y_{it} \leq y | X_{it} = x) = 1 - y^{-1/\xi(x)}$ for $y \geq 1$ where $\xi(x) = x + 0.5$.

We construct CIs for $Q_{Y|X=x_0}(1-h/n)$ with $x_0 = 0$ and 1.65 (the 50% and 95% quantiles of X, respectively) and $h = 1$ and 5. The sample sizes n and T are either 200 or 500, with smaller combinations exercised in later experiments.

We compare three approaches: (i) the fixed-k approach (fixed-k) introduced in Section 2.2, (ii) quantile regression (QR), and (iii) bootstrapping the empirical quantile (Boot). More specifically, we produce the fixed-k CI using $k = 20$ in most cases if not specially noted. The space of ξ is restricted as $[-1/2, 1/2]$ in this and the following two sections where we target quantile. For the QR approach, we run a quantile regression of Y_{it} on X_{it} and a constant at the τ quantile for each i . The conditional quantile is estimated at $\hat{\beta}_{0i} + x_0\hat{\beta}_{1i}$ where $\hat{\beta}_{0i}$ and $\hat{\beta}_{1i}$ are the coefficient estimates using the i-th individual's observations. The CI is simply the 2.5% and 97.5% quantiles of these n estimates. The bootstrap CI is based on bootstrapping the empirical τ quantile in $\{Y_{i,[x_0]}\}_{i=1}^{n}$. The bootstrap size is 200.

Tables 1-3 depict the coverage probabilities (Cov) and the average lengths (Lgth) of the above three methods based on 500 simulation draws. The fixed-k approach performs very well in both coverage and length in all specifications. Regarding the QR method, since the conditional quantile is a linear function of X in the first DGP but not in the other two, the CIs based on QR perform well in the first DGP but deliver substantial undercoverage and longer length in the other two due to misspecification. The bootstrap approach is robust to misspecification but requires the asymptotic normal approximation, which performs well only in the mid-sample. This is why the bootstrap intervals exhibit more undercoverage for $h = 1$ than 5.

We end this subsection with a remark about the choice of k. A larger k leads to more tail observations and hence shorter confidence intervals, but is subject to a larger approximation bias due to including too many mid-sample data. This bias and variance trade off indicates that the choice of k is difficult, especially when n is only moderate. It is actually impossible to choose a uniformly best k allowing the underlying CDF to be flexible (see Theorem 1 of Müller and Wang (2017)). The CDFs in our Monte Carlo are all well behaved so that a k as large as 40% of the sample size performs well. This is seen in Table 3, which reports the numbers for $k = 20$ and 50.

4.2 Comparison to kernel smoothing

The new approach takes only one NN in each time series, which raises the question of efficiency loss. We answer this by comparing the fixed-k approach with the kernel smoothing method proposed by Gardes, Girard, and Lekina (2010). In particular, we first pool the panel data into a cross-sectional sample. Suppose the object of interest is still $Q_{Y|X=x_0}(\tau)$. We follow Gardes, Girard, and Lekina (2010) to pick the bin $B_{b_{nT}}(x_0)$ centered at x_0 with a bandwidth b_{nT}. Since there is no theoretical justification for the optimal choice of b_{nT}, we

Table 1: Finite sample performance of inference about conditional extreme quantile, no model specification

n	200 (97.5% quantile)				500 (99% quantile)			
T	200		500		200		500	
	Cov	Lgth	Cov	Lgth	Cov	Lgth	Cov	Lgth
				Joint Normal				
fixed-k	0.97	0.63	0.96	0.66	0.95	0.56	0.96	0.56
QR	1.00	0.63	1.00	0.41	1.00	0.89	1.00	0.56
Boot	0.97	0.64	0.91	0.61	0.88	0.58	0.95	0.55
				Joint Student's t				
fixed-k	0.96	1.35	0.96	1.47	0.95	1.62	0.94	1.63
QR	0.95	2.20	0.00	1.31	1.00	4.76	0.01	2.87
Boot	0.91	1.36	0.95	1.34	0.89	1.51	0.94	1.68
				Conditional Pareto				
fixed-k	0.96	7.65	0.97	7.14	0.98	15.8	0.97	11.6
QR	0.00	$>10^3$	0.00	$>10^3$	0.00	$>10^3$	0.00	$>10^3$
Boot	0.93	8.30	0.93	7.80	0.94	15.3	0.90	12.7

Note: Entries are coverages and lengths of the CIs for $Q_{Y|X=0}(1-5/n)$. See the main text for the description of the three approaches and the data generating processes. Confidence level is 5%. Based on 500 simulation draws.

Table 2: Finite sample performance of inference about conditional extreme quantile, no model specification

n	200 (97.5% quantile)				500 (99% quantile)			
T	200		500		200		500	
	Cov	Lgth	Cov	Lgth	Cov	Lgth	Cov	Lgth
				Joint Normal				
fixed-k	0.96	0.65	0.96	0.65	0.95	0.57	0.94	0.57
QR	1.00	1.28	1.00	0.80	1.00	1.81	1.00	1.13
BEQ	0.93	0.63	0.92	0.64	0.92	0.55	0.91	0.56
				Joint Student's t				
fixed-k	0.97	2.42	0.95	2.36	0.97	2.88	0.97	2.77
QR	1.00	3.53	1.00	2.26	1.00	6.23	1.00	3.94
BEQ	0.94	2.31	0.93	2.25	0.93	2.83	0.95	2.72
				Conditional Pareto				
fixed-k	0.95	9.30	0.97	7.45	0.84	16.3	0.94	12.5
QR	0.00	$>10^3$	0.00	$>10^3$	0.00	$>10^3$	0.00	$>10^3$
BEQ	0.95	12.5	0.96	8.91	0.80	26.8	0.93	15.2

Note: Entries are coverages and lengths of the CIs for $Q_{Y|X=1.65}(1-5/n)$. See the main text for the description of the three approaches and the data generating processes. Confidence level is 5%. Based on 500 simulation draws.

Table 3: Finite sample performance of inference about conditional extreme quantile, no model specification

n	200 (99.5% quantile)				500 (99.8% quantile)			
T	200		500		200		500	
	Cov	Lgth	Cov	Lgth	Cov	Lgth	Cov	Lgth
	Joint Normal							
fixed-k(k=20)	0.95	1.82	0.96	1.83	0.97	1.69	0.96	1.70
QR	1.00	1.19	1.00	0.75	1.00	1.18	1.00	1.07
BEQ	0.63	0.62	0.64	0.59	0.64	0.57	0.65	0.59
	Joint Student's t							
fixed-k(k=20)	0.96	4.71	0.96	4.69	0.96	5.62	0.97	5.61
fixed-k(k=50)	0.94	3.91	0.92	3.90	0.95	4.85	0.92	4.73
QR	1.00	8.51	0.68	5.51	1.00	8.47	1.00	11.5
BEQ	0.62	2.01	0.60	2.02	0.63	2.57	0.61	2.56
	Conditional Pareto							
fixed-k(k=20)	0.98	27.6	0.98	26.1	0.94	48.1	0.97	40.5
QR	0.00	$>10^3$	0.00	$>10^3$	0.00	$>10^3$	0.00	$>10^3$
BEQ	0.71	25.9	0.63	30.4	0.78	76.2	0.77	43.9

Note: Entries are coverages and lengths of the CIs for $Q_{Y|X=0}(1-1/n)$. See the main text for the description of the three approaches and the data generating processes. Confidence level is 5%. Based on 500 simulation draws.

take the rule-of-thumb choice $c(nT)^{-1/5}$ with different values of the constant c. Now a certain choice of b_{nT} leads to a certain collection of $Y's$ whose paired $X's$ are in the bin $B_{b_{nT}}(x_0)$. Order these induced $Y's$ descendingly into $\{Y_{(1)} \geq Y_{(2)} \geq ... \geq Y_{(m)}\}$ where m denotes the local sample size determined by the bandwidth. Such local sample size is approximately nTb_n in the kernel smoothing (as opposed to n in our new approach).

Given the induced Y's, the conditional quantile is estimated as $\hat{Q}_{Y|X=x_0}(\tau) = Y_{(\lfloor(1-\tau)m\rfloor)}$, that is, the $\lfloor(1-\tau)m\rfloor$-th largest order statistics in the induced Y's where $\lfloor(1-\tau)m\rfloor$ denotes the integer part of $(1-\tau)\,m$. Gardes, Girard, and Lekina (2010) show that under $m(1-\tau) \to \infty$ and some other regularity conditions,

$$\sqrt{m(1-\tau)}\left(\frac{\hat{Q}_{Y|X=x_0}(\tau)}{Q_{Y|X=x_0}(\tau)} - 1\right) \xrightarrow{d} \mathcal{N}(0, 1/\xi_0^2(x_0)).$$

Then the CI of $Q_{Y|X=x_0}(\tau)$ is constructed by the delta method and plugging in some consistent estimator of ξ_0. One choice they propose is the Hill-type estimator

$$1/\hat{\xi} = \frac{1}{k-1}\sum_{i=1}^{k-1} i \log(Y_{(i)}/Y_{(i+1)}) \tag{16}$$

for some choice of $k < m$.

For comparison, we implement the fixed-k approach by using the panel data and the above kernel estimator by pooling the data. In particular, we implement the conditional

Table 4: Finite sample performance of inference about conditional extreme quantile, comparison with kernel method

T	50		100		200		500	
	Cov	Lgth	Cov	Lgth	Cov	Lgth	Cov	Lgth
fixed-k	0.97	21.1	0.97	19.8	0.97	16.8	0.98	15.3
NP(c=0.1)	0.91	50.1	0.89	30.0	0.94	24.4	0.93	14.6
NP(c=0.25)	0.94	33.1	0.96	19.0	0.93	13.3	0.96	9.15
NP(c=0.5)	0.93	17.1	0.94	12.7	0.93	9.28	0.95	6.36
NP(c=1)	0.93	13.9	0.90	9.84	0.89	7.14	0.88	4.77
NP(c=2)	0.37	15.6	0.24	10.1	0.14	6.86	0.11	4.18

Note: Entries are coverages and lengths of the CIs for $Q_{Y|X=0}(1-1/n)$ under the conditional Pareto DGP. See the main text for the description of the two approaches and details of the DGP. Confidence level is 5%. Based on 500 simulation draws.

Pareto DGP in the previous experiment with $n = 200$ and T ranging from 50 to 500. For the fixed-k CI, we set $k = 50$. For the kernel method, we implement $c \in \{0.1, 0.25, 0.5, 1, 2\}$ and set k (in the Hill-type index estimator (16)) as the largest integer less than or equal to $m/4$.

Table 4 presents the coverage and the length of the fixed-k and the kernel CIs. Several interesting observations are made. First, the kernel approach is sensitive to the choice of the bandwidth. In particular, a correct coverage relies on a narrow window of the bandwidth choice. A larger choice can lead to a substantial undercoverage since the smoothing bias dominates quickly in the tail. Second, when T is only moderately large (say 25 and 50), the fixed-k CIs are much shorter than the kernel one and both of them have good coverage properties. This is because the fully nonparametric method ignores the domain of attraction information, which is utilized by the fixed-k method. Third, when T is very large, say 500, choosing only one NN does incur an efficiency loss as we compare the length between the fixed-k and the kernel CIs. But such loss is approximately in a factor of 2 or 3 instead of $T^{1/2}$. This means a general covariate-dependent tail is very difficult to estimate in a fully nonparametric way.

In Table 5, we consider a two-dimensional standard normal X and generate Y_{it} by $Y_{it}|X_{it} = x \sim \pm\text{Pa}(\xi(x))$ with $\xi(x_1, x_2) = x_1 + x_2 + 0.5$. The kernel method is illustrated with $c \in \{0.5, 1, 2, 4\}$. All other choices of both methods remain unchanged as in Table 4. The results clearly suggest that the fixed-k method together with the NN choice dominates the kernel method in both coverage probabilities and length. In particular, the kernel method may suffer from the curse of dimensionality as the dimension of X increases.

As a final remark of this subsection, we also implement the standard kernel weighted

Table 5: Finite sample performance of inference about conditional extreme quantile, comparison with kernel method, two-dimensional X

T	50		100		200		500	
	Cov	Lgth	Cov	Lgth	Cov	Lgth	Cov	Lgth
fixed-k	0.96	21.0	0.97	17.2	0.96	16.5	0.96	16.3
NP(c=0.5)	0.56	65.8	0.73	60.6	0.73	54.3	0.81	53.8
NP(c=1)	0.83	75.0	0.80	60.8	0.93	60.9	0.91	32.0
NP(c=2)	0.96	66.1	1.00	36.0	0.97	25.2	0.97	16.5
NP(c=4)	0.74	54.5	0.47	35.4	0.23	22.8	0.16	13.2

Note: Entries are coverages and lengths of the CIs for $Q_{Y|X=0}(1-1/n)$ under the conditional Pareto DGP. See the main text for the description of the two approaches and details of the DGP. Confidence level is 5%. Based on 500 simulation draws.

quantile regression method designed for the mid-sample quantiles (cf. Chapter 10 of Li and Racine (2007)). Given a large T, the target $1-1/n$ conditional quantile is relatively in the mid-sample after pooling the panel data into a cross-sectional one, and hence the confidence interval based on asymptotic normality might work. However, unreported Monte Carlo simulations show that this method works only if T is substantially, say 5 times, larger than n. In our experiments, it is strictly dominated by the method proposed by Gardes, Girard, and Lekina (2010).

4.3 Conditional extreme quantile in linear model

In this section, we first consider the linear regression model $Y_{it} = \alpha_i + X_{it}\beta_0 + u_{it}$ and assume data are i.i.d. across i. For the time series dependence, we set $\alpha_i = T^{-1}\sum_{t=1}^{T} X_{it}$ and $X_{it} = \rho X_{it-1} + e_{it}$ with $e_{it} \sim^{iid} \mathcal{N}(0, (1-\rho^2))$ and $X_{i0} \sim \mathcal{N}(0,1)$. The distribution of u_{it} conditional on $X_{it} = x$ is as follows.

1. **Conditional Normal** $u_{it}|X_{it} = x \sim \mathcal{N}(0, 1+x^2)$.

2. **Conditional Student's t** $u_{it}|X_{it} = x \sim t(2+|x|)$.

3. **Conditional Pareto** $u_{it}|X_{it} = x \sim \pm\text{Pa}(\xi(x))$, that is, $\mathbb{P}(u_{it} \le y|X_{it} = x) = 1/2 + (1-(1+y)^{-1/\xi(x)})/2$ for $y \ge 0$, and $\mathbb{P}(u_{it} \le y|X_{it} = x) = (-y+1)^{-1/\xi(x)}/2$ for $y \le 0$ where $\xi(x) = x + 0.5$.

We use the same three approaches as in the Section 4.1 to construct CIs for the conditional extreme quantile $Q_{Y_{it}|X_{it}=x_0}(\tau) = Q_{\varepsilon_{it}|X_{it}=x_0}(\tau) + x_0\beta_0$, where ε_{it} denotes $\alpha_i + u_{it}$. In

Table 6: Finite sample performance of inference about conditional extreme quantile, non-dynamic model with random effects

n	200 (99.5% quantile)				100 (99% quantile)			
T	200		500		25		50	
	Cov	Lgth	Cov	Lgth	Cov	Lgth	Cov	Lgth
	Conditional Normal							
fixed-k w. LS	0.94	2.23	0.93	2.13	0.80	2.85	0.88	2.34
fixed-k w/o LS	0.93	2.22	0.92	2.14	0.53	3.17	0.76	2.62
QR	0.00	3.04	0.00	2.19	1.00	3.10	1.00	2.96
Boot	0.73	0.77	0.67	0.71	0.73	1.11	0.81	0.92
	Conditional Student's t							
fixed-k w. LS	0.95	15.3	0.94	15.5	0.93	10.0	0.96	10.4
fixed-k w/o LS	0.95	15.3	0.94	15.5	0.91	10.0	0.93	10.3
QR	1.00	26.5	1.00	7.98	0.99	11.0	1.00	15.0
Boot	0.56	11.5	0.60	11.0	0.47	6.32	0.51	6.01
	Conditional Pareto							
fixed-k w. LS	0.00	16.2	0.00	5.90	0.02	59.6	0.01	58.1
fixed-k w/o LS	0.97	18.8	0.97	16.9	0.95	16.0	0.96	15.4
QR	0.00	$>10^3$	0.00	$>10^3$	1.00	$>10^3$	1.00	$>10^3$
Boot	0.71	16.2	0.67	16.8	0.76	313	0.78	31.4

Note: Entries are coverages and lengths of the CIs for $Q_{Y|X=0}(1-1/n)$. See the main text for the description of different approaches and the data generating processes. Confidence level is 5%. Based on 500 simulation draws.

particular, the fixed-k approach is conducted in two ways: with or without using the standard within least squares estimator of β_0. For the former (fixed-k w. LS), we first estimate β_0 using the standard within estimator $\hat{\beta}$ and back out $\hat{\varepsilon}_{it} = Y_{it} - X_{it}\hat{\beta}$. Then we implement the steps in Section 2.2 to construct the CIs for the conditional quantile of ε_{it}. The CIs for $Q_{Y_{it}|X_{it}=x_0}(\tau)$ are obtained by adding back $x_0\hat{\beta}$. For the one ignoring the linear regression structure (fixed-k w/o LS), we directly use Y_{it} and X_{it} and apply Steps 1-3 in Section 2.2.

Table 6 presents the results for $n \in \{100, 200\}$ and $T \in \{25, 50, 200, 500\}$. Several interesting observations can be found. First, the error in the conditional t and conditional Pareto models does not have a finite variance when x_0 is 0, and hence the LS estimator of β_0 behaves poorly. This leads to a poor performance of the fixed-k approach if the linear regression model is utilized. This problem can be solved by using the least absolute deviation (LAD) estimator as shown in unreported results. In comparison, the fixed-k CIs without using the linear regression model always perform well given a large enough sample size. Second, the QR approach still suffers from undercoverage in all three specifications since the normal and the student's t DGPs have nonlinear heteroskedasticity and the conditional Pareto DGP violates the constant tail shape condition. Finally, the bootstrap method performs poorly if the extreme quantile under investigation is too far in the tail.

Table 7: Finite sample performance of inference about large quantiles of the random coefficients

n	200				500 (99% quantile)			
T	10		20		10		20	
	Cov	Lgth	Cov	Lgth	Cov	Lgth	Cov	Lgth
	CIs for $Q_\alpha(1-5/n)$							
fixed-k	0.92	0.77	0.95	0.76	0.92	0.69	0.93	0.67
QR	0.84	1.13	0.90	1.07	0.88	2.96	0.94	2.94
Boot	0.89	0.81	0.93	0.76	0.81	0.70	0.91	0.69
	CIs for $Q_\beta(1-5/n)$							
fixed-k	0.91	0.81	0.96	0.76	0.86	0.69	0.96	0.67
QR	0.81	1.15	0.88	1.08	0.88	3.21	0.94	2.83
Boot	0.85	0.82	0.92	0.76	0.78	0.73	0.91	0.68
	CIs for $Q_\alpha(1-1/n)$							
fixed-k	0.91	2.32	0.93	2.13	0.87	2.10	0.94	1.96
QR	0.89	1.45	0.91	1.42	0.89	1.33	0.88	1.29
Boot	0.57	0.52	0.58	0.51	0.57	0.47	0.54	0.47
	CIs for $Q_\beta(1-1/n)$							
fixed-k	0.88	2.32	0.94	2.28	0.85	2.16	0.93	1.93
QR	0.90	1.52	0.91	1.52	0.86	1.41	0.88	1.29
Boot	0.57	0.55	0.58	0.54	0.55	0.51	0.58	0.45

Note: The entries are coverage and length of the confidence intervals based on (i) the fixed-k approach using the largest k=20 estimated coefficients, (ii) empirical quantile of the estimated coefficients with asymptotic normal approximation, and (iii) empirical quantile function of the estimated coefficients and bootstrap. Data are generated from $Y_{it} = \alpha_i + X_{it}\beta_i + u_{it}$ where $(\alpha_i, \beta_i, X_{it}, u_{it})^\intercal \sim^{iid} \mathcal{N}(0, I_4)$. The target is the 1-h/n quantile of α_i and β_i with $h = 1$ and 5, corresponding to 97.5%, 98%, 99%, and 99.8% quantiles given $n = 200$ and 500, respectively. Confidence level is 5%. Based on 500 simulation draws.

In Table 7, we study the CIs for high quantiles of α_i and β_i with data generated from $Y_{it} = \alpha_i + X_{it}\beta_i + u_{it}$ where $(\alpha_i, \beta_i, X_{it}, u_{it})^\intercal \sim^{iid} \mathcal{N}(0, I_4)$. The i.i.d. condition is across both i and t in this setting. In particular, we first estimate α_i and β_i by regressing Y_{it} on $(1, X_{it})^\intercal$ with T observations from individual i. Then we collect the estimators $\hat{\alpha}_i$ and $\hat{\beta}_i$ for all i and sort them descendingly to apply the fixed-k, QR, and bootstrap methods. The QR estimator is simply the empirical quantile among the estimators, whose asymptotic variance is estimated by the standard kernel density estimator with the rule-of-thumb bandwidth. The results suggest that the fixed-k approach with NN dominates the other two in both coverage and length, especially when the sample size is only moderate.

4.4 Conditional tail index

The last experiment examines the CIs of the conditional tail index. We consider the following three DGPs.

1. Joint Student's t (X_{it}, Y_{it}) is i.i.d. across i and t and is distributed as $t_v(\mu, \Sigma)$ with

$v = 2$, $\mu = [0, 0]^\intercal$, and $\Sigma = [1, 0.5; 0.5, 1]$.

2. **Conditional Pareto** $X_{it} = \rho X_{it-1} + u_{it}$ with $u_{it} \sim \mathcal{N}(0, (1 - \rho^2))$ and i.i.d. across i and t, and $X_{i1} \sim \mathcal{N}(0, 1)$. $Y_{it}|X_{it} = x \sim \mathrm{Pa}(1/\xi(x))$, that is, $\mathbb{P}(Y_{it} \leq y|X_{it} = x) = 1 - y^{-1/\xi(x)}$ for $y \geq 1$. Set $\xi(x) = x - x_0 + 0.5$.

3. **Independent F** $X_{it} = \rho X_{it-1} + u_{it}$ with $u_{it} \sim \mathcal{N}(0, (1 - \rho^2))$ and $X_{i1} \sim \mathcal{N}(0, 1)$. Y_{it} is F(4,4) and independent of X_{it}. Y_{it} and u_{it} are both i.i.d. across i and t.

The normal distribution is replaced with independent F(4,4), so that the true conditional tail index is 0.5 in all three designs when conditioned on $X = x_0$. We set $\Xi = [0, 1]$ since Hill's estimator is only defined for positive tail indices. Table 8 reports the coverage and length of the fixed-k CI (fixed-k) based on inverting (13) and that based on Hill's estimator (Hill) and the maximum likelihood estimator (MLE) as described in Section 3.2. We set $x_0 = 1.65$, $\rho = 0.5$ and $n = T = 1000$, and choose $k \in \{20, 50, 100, 200\}$. This is to make sure k is small relatively to n but still large enough for the increasing-k asymptotic approximation. As expected from the theoretical derivation, the fixed-k CIs deliver excellent coverage and length when k is small. As k grows, the MLE gradually performs better. A heuristic rule-of-thumb for choosing the MLE instead of the fixed-k CI is based on whether k is over 100 or not, provided n is substantially larger. Note that Hill's estimator is not location invariant and thus heavily relies on the Pareto tail approximation. This is why it has short length and precise coverage when the underlying distribution is exactly Pareto, but has large extents of undercoverage when the DGP is student's t or F.

5 Empirical application to US stock returns

Tail risk in stock returns has been an important topic in finance. See Backus, Chernov, and Martin (2011) and Bollerslev and Todorov (2011) among many others. Due to limited observations, tail features are usually difficult to study if using time series data only. Motivated by this issue, Kelly and Jiang (2014) use panel data on stock returns and assume the left tail of the i-th stock at period t has a time varying tail index $\xi_{it} = \lambda_t / a_i$. The λ_t term captures the dynamics that are shared by all assets and a_i measures the stock specific tail risk. We relax such ratio structure by considering a covariate-dependent tail index. In particular, we consider the stock size as the covariate.

We follow the convention to use monthly returns of NYSE/AMEX/NASDAQ stocks with share codes 10 and 11. To obtain a large T, we only use the stocks that are traded for more

Table 8: Small sample performance for inference about the conditional tail index

k	20		50		100		200	
	Cov	Lgth	Cov	Lgth	Cov	Lgth	Cov	Lgth
				Conditional Student's t				
fixed-k	0.97	0.76	0.96	0.69	0.92	0.53	0.81	0.37
Hill	0.88	0.41	0.83	0.26	0.86	0.18	0.96	0.14
MLE	0.76	0.80	0.86	0.69	0.86	0.55	0.75	0.38
				Conditional Pareto				
fixed-k	0.95	0.75	0.94	0.68	0.90	0.54	0.88	0.39
Hill	0.96	0.48	0.95	0.30	0.93	0.21	0.94	0.14
MLE	0.81	1.00	0.90	0.79	0.94	0.59	0.95	0.42
				Independent F				
fixed-k	0.96	0.75	0.94	0.69	0.92	0.54	0.94	0.39
Hill	0.97	0.49	0.94	0.31	0.63	0.24	0.02	0.19
MLE	0.77	0.79	0.90	0.70	0.92	0.56	0.94	0.41

Note: Entries are coverages and lengths of CIs on the tail index of the underlying conditional distribution, based on the largest k order statistics. See the main text for a description of the three approaches and the DGPs. Confidence level is 5%. Based on 500 Monte Carlo simulations.

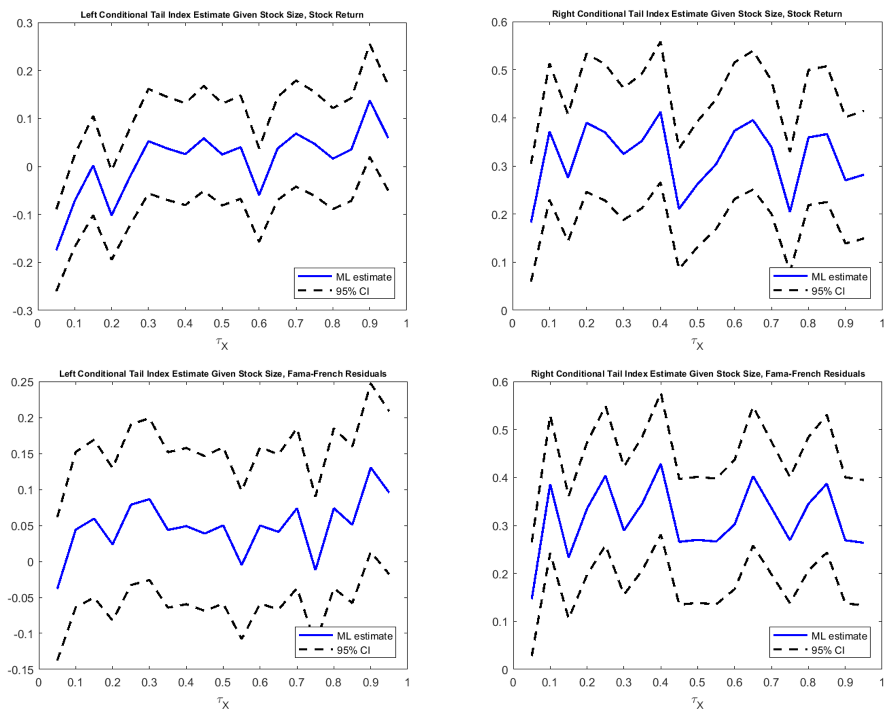
than 120 months. This leads to an unbalanced panel dataset with $n = 1744$ and T ranging from 121 to 1104. Given such a large n, we apply the MLE (15) with $k = 349$ (20% of n) and the corresponding CIs based on its asymptotic normality. Top panels in Figure 2 plot the estimated left and right conditional tail indices of stock returns given stock size equal to its τ_x unconditional quantile with $\tau_x \in [0.05, 0.95]$. Lower panels plot the same estimates and CIs based on the residuals of the Fama-French three-factor regression (see eq.(17) below). The results suggest that large stocks tend to exhibit heavier left tails, but such relation is weak for the right tail. This result is coherent with that of Chen, Hong, and Stein (2001), who use linear regression to find such pattern.

Next, we examine the conditional extreme quantiles. The first two rows in Figure 3 plot the QR estimates and the fixed-k CIs of the τ conditional quantiles of stock returns conditional on the τ_x quantile of the stock size. In particular, we present the results for $\tau_x \in \{0.05, 0.5, 0.95\}$ and $\tau = h/n$ and $1 - h/n$ for $\{1, 2, ..., 10\}$. The QR estimate is based on running quantile regression of the stock return on a constant and the stock size. The fixed-k CIs are based on $k = 100$. This is set for good coverage but possibly conservative length. The figure shows that the QR estimates are outside the fixed-k CIs for the left tail, but not for the right one, indicating that the left conditional extreme quantiles are highly nonlinear in the covariate but the right ones are close to be linear.

Finally, we examine whether a Fama-French three factor regression can fully explain the tail behavior (cf. Fama and French (1993)). In particular, we run the regression

$$R_{it} - R_{Ft} = \alpha_i + \beta_i(R_{Mt} - R_{Ft}) + s_i SMB_t + h_i HML_t + e_{it}, \tag{17}$$

Figure 2: Plots of the QR estimates and the fixed-k CIs of the conditional tail indices of stock returns and the Fama-French residuals given different stock sizes



Note: This figure plots the QR estimates (solid line) and 95% fixed-k CIs (dash line) of the left and right conditional tail indices of the stock returns (upper row) or the residuals from the Fama-French 3 factors regression (lower row). See the main context for details of these two approaches. Data are available from http://www.crsp.com.

where $(R_{it}, R_{Mt}, R_{Ft})^\intercal$ denote the monthly returns of the i-th stock, market, and risk free asset, respectively, and SMB_t and HML_t are the other two factors (cf. eq(4) in Fama and French (2015)). The coefficients $(\alpha_i, \beta_i, s_i, h_i)^\intercal$ are potentially varying across stocks. Lower two rows in Figure 3 plot the QR estimates and the fixed-k CIs of the τ conditional quantiles of the regression residuals in (17) conditional on different stock sizes. Similarly to the stock returns, the residuals still exhibit some tail dependence on the stock size, suggesting that the Fama-French factors are still insufficient to characterize the tail features as opposed to their success in fitting the mean. In Figure 4, we use the method introduced in Section 2.3 to construct the 95% fixed-k CIs of the extreme quantiles of the random coefficients $(\beta_i, s_i, h_i)^\intercal$, using the same dataset and tuning parameters as for the conditional quantiles. These figures clearly suggest that the factor coefficients are substantially varying across stocks, so that it might be misleading to consider them as fixed constants.

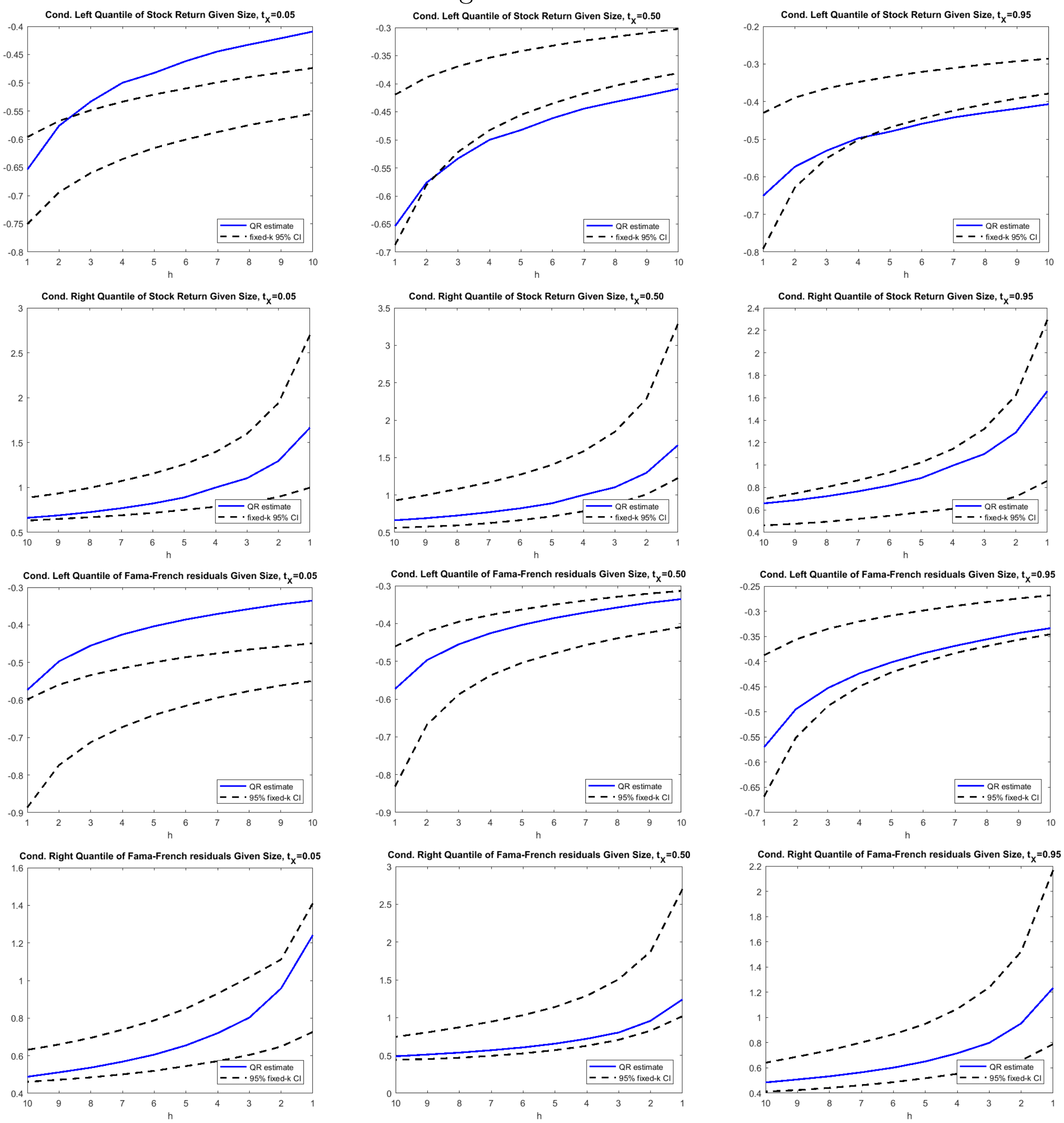
6 Concluding remarks

This paper develops a new framework on inference about conditional tail features using panel data. The key insight is that the induced order statistics in each time series can be treated as approximately stemming from the true conditional distribution, and the large order statistics among these induced values can then be used to study the conditional tail. By focusing on the induced order statistics, we essentially reduce the conditional tail problem into an unconditional one, so that existing approaches about unconditional tail features become applicable. Monte Carlo simulations show that the new method delivers excellent small sample performance in terms of coverage probability and length.

The new method is substantially more flexible than the extremal quantile regression because the latter assumes that the conditional extreme quantile is a parametric location-shift model, which is an empirical concern in some applications. If a linear regression model is imposed, the new method is easily combined with any existing consistent estimator of the structural parameter and applies to the tail features of the random coefficients.

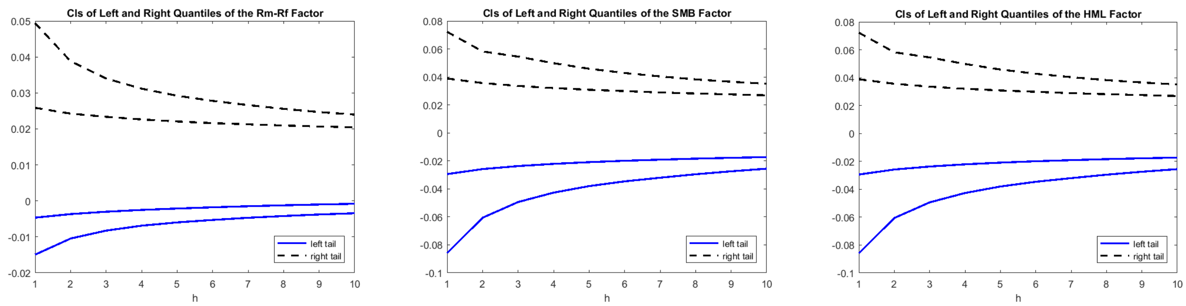
If a large cross-sectional sample is available, the econometrician can first randomly decompose the sample into a panel/repeated cross-sectional dataset and then apply the new method to the generated panel data.

Figure 3: Plots of the QR estimates and the fixed-k CIs of the conditional quantiles of stock returns and the Fama-French residuals given different stock sizes



Note: This figure plots the QR estimates (solid line) and the 95% fixed-k CIs (dash line) of the right $(1 - h/n)$ and left (h/n) conditional quantiles of the stock returns (upper two rows) or residuals from the Fama-French 3 factors regression (lower two rows), where $n = 1744$ and $h \in \{1, \ldots, 10\}$. See the main context for details of these two approaches. Data are available from http://www.crsp.com.

Figure 4: Plots of the fixed-k CIs of quantiles of the Fama-French factor coefficients



Note: This figure plots the 95% fixed-k CIs of the right $(1-h/n)$ and left (h/n) unconditional quantiles of the three coefficients in the Fama-French 3 factors regression, where $n = 1744$ and $h \in \{1, \ldots, 10\}$. Data are available from http://www.crsp.com.

A Appendix

A.1 Omitted details and primitive conditions in Section 2

This section provides more details about important implications of Condition 1.1 and primitive sufficient conditions for Condition 1.3. We first establish a lemma about the convergence of the NN using Condition 1.1. The proof is collected in the next subsection.

Lemma 1 *Under Condition 1.1, for each i and for some $\eta > 0$,*

$$\left|\left|X_{i,(x_0)} - x_0\right|\right| = o_{a.s.}(T^{-\eta}) \text{ and} \tag{18}$$

$$\mathbb{E}\left[\left|\left|X_{i,(x_0)} - x_0\right|\right|\right] = O\left(T^{-1/2}\right). \tag{19}$$

Next, we provide primitive conditions for Condition 1.3. The following conditions are sufficient. Recall that y_0 denotes the right end-point $\sup\{y, F_{Y|X=x_0}(y) < 1\}$. The notation is simpler if we use the following notations: $\gamma(\cdot) = 1/\xi(\cdot)$, g_i denotes the the partial derivative of a generic function $g(\cdot,\cdot)$ w.r.t. the i-th element, and g_{ij} the i,j-th cross derivative.

Condition B X_{it} has a compact support. $F_{Y|X=x}(y)$ satisfies either (i) $\xi(x) > 0$ and

$$1 - F_{Y|X=x}(y) = c(x)y^{-\gamma(x)}(1 + d(x)(y)^{-\tilde{\gamma}(x)} + r(x,y))$$

where $c(\cdot) > 0$ and $d(\cdot)$ are uniformly bounded between 0 and ∞ and continuously differentiable with uniformly bounded derivatives, $\gamma(\cdot) > 0$ and $\tilde{\gamma}(\cdot) > 0$ are continuously differentiable functions, and $r(x,y)$ is continuously differentiable with bounded derivatives w.r.t. both x and y, and satisfies for some $\delta > 0$

$$\limsup_{y \to y_0} \sup_{x \in B_\delta(x_0) \cap \{(x:\xi(x)>0\}} \left|r(x,y)/y^{-\tilde{\gamma}(x)}\right| \rightarrow 0,$$

$$\limsup_{y\to y_0} \sup_{x\in(B_\delta(x_0)\cap\{x:\xi(x)>0\}} \left| r_2(x,y)/(y^{-\tilde{\gamma}(x)-1}) \right| \quad \to \quad 0,$$

$$\limsup_{y\to y_0} \sup_{x\in B_\delta(x_0)\cap\{(x:\xi(x)>0\}} \left| r_1(x,y)/y^{-\tilde{\gamma}(x)} \right| \quad \to \quad 0,$$

$$\limsup_{y\to y_0} \sup_{(B_\delta(x_0)\cap\{x:\xi(x)>0\}} \left| r_{21}(x,y)/(y^{-\tilde{\gamma}(x)-1}) \right| \quad \to \quad 0;$$

or (ii) $\xi(x) = 0$ and

$$f_{Y|X=x}(y) = c(x)y^{\tilde{c}(x)} \exp(-d(x)\tilde{d}(y))(1 + r(x,y)),$$

where $c(\cdot) > 0$ and $d(\cdot) > 0$ are some continuously differential functions that are uniformly bounded between 0 and ∞, $\tilde{c}(\cdot)$ is continuously differentiable and uniformly bounded by -1 and ∞, and $\tilde{d}(y)$ is continuously differentiable and satisfies $C_1(\log y)^2 \le \tilde{d}(y) \le C_2 y^{C_3}$ for some constants $0 \le C_1, C_2, C_3 < \infty$. The remainder $r(x,y)$ is uniformly bounded and continuously differentiable w.r.t. both arguments with bounded derivatives, and satisfies that for some $\delta > 0$

$$\limsup_{y\to y_0} \sup_{x\in B_\delta(x_0)\cap\{x:\xi(x)=0\}} |\max\{r_1(x,y), r_2(x,y), r_{21}(x,y)\}| \to 0.$$

Condition B assumes that the error of approximating the true CDF with a generalized Pareto distribution consists of the leading terms $1+d(x)(y)^{-\tilde{\gamma}(x)}$ and $c(x)y^{\tilde{c}(x)} \exp(-d(x)\tilde{d}(y))$, respectively in the two cases with $\xi(x) > 0$ and $\xi(x) = 0$ and the remainder $r(x,y)$. Case (i) covers regularly varying tails, and are imposed by Smith (1982) to study unconditional problems. See also Hall (1982) and Smith (1987). Case (ii) covers slowly varying tails, including Gaussian ($\tilde{c}(x) = 0$ and $\tilde{d}(y) = y^2$), lognormal ($\tilde{c}(x) = -1$ and $\tilde{d}(y) = (\log y)^2$), and the exponential family ($\tilde{c}(x) = 0$ and $\tilde{d}(y) = y$). See, for example, Chapter B in de Haan and Ferreira (2007). Compared with those literature, we require a stronger version that the derivatives of $r(x,y)$ are uniformly bounded. This is to guarantee that the tail of $f_{Y|X=x_0}$ is also uniformly bounded. The compact support of X is imposed to simplify the proof (cf. Wang and Li (2013)). The following lemma establishes Condition 1.3 using Conditions 1.4 and B. Its proof is collected at the very end of this article.

Lemma 2 *If Condition 1.4 and Condition B hold, then Condition 1.3 holds, i.e., for $u_n = a_n y + b_n$ with any fixed $y > 0$, as $n \to \infty$ and $T \to \infty$*

(a) $\lim_{u_n\to y_0} \sup_{x\in B_{\eta_T}(x_0)} \left\| \frac{\partial F_{Y|X=x}(u_n)/\partial x}{1-F_{Y|X=x_0}(u_n)} \right\| = 0,$

(b) $\lim_{u_n\to y_0} \sup_{x\in B_{\eta_T}(x_0)} \left\| \frac{\partial f_{Y|X=x}(u_n)/\partial x}{f_{Y|X=x}(u_n)} \right\| = 0.$

To give a better sense of Condition B, we now show that it is satisfied by the three examples introduced in Section 2.2.

First consider the joint normal distribution. Condition B.(ii) is satisfied by setting $c(x) = \sqrt{2\pi(1-\rho^2)}$, $d(x)=1$, $\tilde{d}(y) = y^2/(2(1-\rho^2))$, and $r(x,y) = \exp(2\rho x/y + \rho^2 x^2/y^2) - 1$. Second, for the conditional student's t distribution, Ding (2016) derives that the conditional PDF of Y given $X = x$ is

$$f_{Y|X=x}(y) = \frac{C}{\sigma(x)}\left(1 + \frac{(y-\rho x)^2}{(v+1)\,\sigma(x)^2}\right)^{-\frac{v+2}{2}}$$

for some constant C depending on v only and $\sigma(x) = \sqrt{(1-\rho^2)(v+x^2)/(v+1)}$. Then Condition B.(i) holds with $\gamma(x) = 1/(v+1)$, $c(x) \propto \sigma(x)^{v+1}, d(x) \propto \rho x, \tilde{\gamma}(x) = 1$, and $r(x,y) = O(y^{-2})$ for any $x \in \mathbb{R}$. Finally, for the conditional Pareto distribution, Taylor expansion yields

$$\begin{aligned} 1 - F_{Y|X=x}(y) &= y^{-1/x}(1+1/y)^{-1/x} \\ &= y^{-1/x}(1 - \frac{1}{xy} + O(\frac{1}{y^2})). \end{aligned}$$

Thus Condition B.(i) holds with $c(x) = 1, \gamma(x) = 1/x, d(x) = -1/x, \tilde{\gamma}(x) = 1$, and $r(x,y) = O(y^{-2})$ for x bounded below from 0.

A.2 Proofs

Proof of Theorem 1 By Corollary 1.2.4 and Remark 1.2.7 in de Haan and Ferreira (2007), the constants a_n and b_n can be chosen as follows. If $\xi(x_0) > 0$, we choose $a_n(\xi(x_0)) = Q_{Y|X=x_0}(1-1/n)$ and $b_n(\xi(x_0)) = 0$. If $\xi(x_0) = 0$, we choose $a_n(\xi(x_0)) = 1/(n f_{Y|X=x_0}(b_n(x_0)))$ and $b_n(\xi(x_0)) = Q_{Y|X=x_0}(1-1/n)$. By construction, these constants satisfy that $1 - F_{Y|X=x_0}(a_n(\xi(x_0))y + b_n(\xi(x_0))) = O(n^{-1})$ for any fixed $y > 0$ in both cases (cf. Chapter 1.1.2 in de Haan and Ferreira (2007)).

Now we prove Theorem 1 using the above introduced constants. We suppress $\xi(x_0)$ in $a_n(\cdot)$ and $b_n(\cdot)$ and consider $k = 1$ first. By strict stationarity across t (Condition 1.1), we have that for any generic argument v,

$$\begin{aligned} \mathbb{P}\left(Y_{i,[x_0]} \le v\right) &= \mathbb{E}_{X_{i,(x_0)}}\left[\mathbb{P}\left(Y_{i,[x_0]} \le v | X_{i,(x_0)}\right)\right] \\ &= \mathbb{E}_{X_{i,(x_0)}}\left[F_{Y|X=X_{i,(x_0)}}(v)\right]. \end{aligned} \tag{20}$$

Thus,

$$\begin{aligned} & \mathbb{P}\left(Y_{(1),[x_0]} \le a_n y + b_n\right) \\ = \; & F^n_{Y_{i,[x_0]}}(a_n y + b_n) \text{ (by i.i.d. across } i) \\ = \; & F^n_{Y|X=x_0}(a_n y + b_n)\left(\frac{\mathbb{P}\left(Y_{i,[x_0]} \le a_n y + b_n\right)}{F_{Y|X=x_0}(a_n y + b_n)}\right)^n \end{aligned}$$

$$
\begin{aligned}
&= F^n_{Y|X=x_0}(a_n y + b_n)\left(\frac{\mathbb{E}_{X_{i,(x_0)}}\left[F_{Y|X=X_{i,(x_0)}}(a_n y + b_n)\right]}{F_{Y|X=x_0}(a_n y + b_n)}\right)^n \text{ (by (20))} \\
&= F^n_{Y|X=x_0}(a_n y + b_n)\left(1+\frac{\mathbb{E}_{X_{i,(x_0)}}\left[F_{Y|X=X_{i,(x_0)}}(a_n y + b_n)\right]-F_{Y|X=x_0}(a_n y + b_n)}{F_{Y|X=x_0}(a_n y + b_n)}\right)^n \\
&\equiv A_n(y)\left(1+\frac{B_{n,T}(y)}{F_{Y|X=x_0}(a_n y + b_n)}\right)^n .
\end{aligned}
$$

By the standard EV theory and Condition 1.2, $A_n(y) \to G_\xi(y)$ as $n \to \infty$. Regarding $B_{n,T}(y)$, we derive that, for some $\dot{x}_i$ between $X_{i,(x_0)}$ and x_0 for each i, some open ball $B_{\eta_T}(x_0)$ centered at x_0 with radius $\eta_T = O(T^{-\eta})$, and some constant $0 < C < \infty$,

$$
\begin{aligned}
& |B_{n,T}(y)| \\
&=_{(1)} \mathbb{E}\left[\frac{\partial}{\partial x}F_{Y|X=x}(a_n y + b_n)|_{x=\dot{x}_i}\left(X_{i,(x_0)}-x_0\right)\right] \\
&\leq_{(2)} CT^{-\eta}\sup_{x\in B_{T^{-\eta}}(x_0)}\left\|\frac{\partial}{\partial x}F_{Y|X=x}(a_n y + b_n)\right\| \\
&\leq_{(3)} CT^{-\eta}n^{-1}\sup_{x\in B_{T^{-\eta}}(x_0)}\left\|\frac{\frac{\partial}{\partial x}F_{Y|X=x}(a_n y + b_n)}{1-F_{Y|X=x_0}(a_n y + b_n)}\right\| \\
&=_{(4)} o(n^{-1}),
\end{aligned}
$$

where equality(1) is by the mean value expansion; inequality(2) follows from that $X_{i,(x_0)} \in B_{\eta_T}(x_0)$ almost surely (Lemma 1); inequality(3) is by the fact that $1 - F_{Y|X=x_0}(a_n y + b_n) = O(1/n)$ and equality(4) is given by Conditions 1.3-1.4. Hence given $a_n y + b_n \to y_0$ and using Lemma 8.4.1 in Arnold, Balakrishnan, and Nagaraja (1992), we have

$$
\begin{aligned}
\left(1+\frac{B_{n,T}(y)}{F_{Y|X=x_0}(a_n y + b_n)}\right)^n &\leq \left(1+\frac{o(n^{-1})}{F_{Y|X=x_0}(a_n y + b_n)}\right)^n \\
&\to 1.
\end{aligned}
$$

The proof for $k = 1$ is then complete by the continuous mapping theorem.

Generalization to $k > 1$ is as follows. Consider $y_1 > y_2 > \cdots > y_k$. Chapter 8.4 in Arnold, Balakrishnan, and Nagaraja (1992) gives that

$$
\begin{aligned}
& \mathbb{P}\left(Y_{(1),[x_0]} \leq a_n y_1 + b_n, ..., Y_{(k),[x_0]} \leq a_n y_k + b_n\right) \\
&= F^{n-k}_{Y_{i,[x_0]}}(a_n y_k + b_n)\prod_{r=1}^{k}(n-r+1)\,a_n f_{Y_{i,[x_0]}}(a_n y_r + b_n) \text{ (by i.i.d. across } i) \\
&= \left[F^{n-k}_{Y|X=x_0}(a_n y_k + b_n)\prod_{r=1}^{k}(n-r+1)\,a_n f_{Y|X=x_0}(a_n y_r + b_n)\right]\times
\end{aligned}
$$

$$
\begin{aligned}
& \left[\left(\frac{\mathbb{P}\left(Y_{i,[x_0]} \le a_n y_k + b_n\right)}{F_{Y|X=x_0}\left(a_n y_k + b_n\right)}\right)^{n-k} \prod_{r=1}^{k} \frac{f_{Y_{i,[x_0]}}\left(a_n y_r + b_n\right)}{f_{Y|X=x_0}\left(a_n y_r + b_n\right)}\right] \\
\equiv \;& \tilde{A}_n \times \tilde{B}_{nT}.
\end{aligned}
$$

The convergence that $\tilde{A}_n \rightarrow G_\xi\left(y_k\right)\prod_{r=1}^{k}\{g_\xi\left(y_r\right)/G_\xi\left(y_k\right)\}$ is established by Theorem 8.4.2 in Arnold, Balakrishnan, and Nagaraja (1992). It now remains to show $\tilde{B}_{nT} \rightarrow 1$. First, $\left(\mathbb{P}\left(Y_{i,[x_0]} \le a_n y_k + b_n\right)/F_{Y|X=x_0}\left(a_n y_k + b_n\right)\right)^{n-k} \rightarrow 1$ is shown by the same argument as above in the $k=1$ case. Second, for any v

$$
\begin{aligned}
\frac{f_{Y_{i,[x_0]}}\left(v\right)}{f_{Y|X=x_0}\left(v\right)} &= \frac{\frac{\partial \mathbb{P}\left(Y_{i,[x_0]}\le v\right)}{\partial v}}{f_{Y|X=x_0}\left(v\right)} \\
&= \frac{\frac{\partial}{\partial v}\mathbb{E}_{X_{i,(x_0)}}\left[F_{Y|X=X_{i,(x_0)}}\left(v\right)\right]}{f_{Y|X=x_0}\left(v\right)} \text{ (by (20))} \\
&= \frac{\frac{\partial}{\partial v}\int F_{Y|X=x}\left(v\right) f_{X_{i,(x_0)}}\left(x\right) dx}{f_{Y|X=x_0}\left(v\right)} \\
&= \frac{\int \frac{\partial}{\partial v} F_{Y|X=x}\left(v\right) f_{X_{i,(x_0)}}\left(x\right) dx}{f_{Y|X=x_0}\left(v\right)} \quad \text{(by Leibniz's rule)} \\
&= \frac{\mathbb{E}_{X_{i,(x_0)}}\left[f_{Y|X=X_{i,(x_0)}}\left(v\right)\right]}{f_{Y|X=x_0}\left(v\right)},
\end{aligned}
$$

where applying Leibniz's rule is permitted by the assumption (Condition 1.3) that $f_{Y|X=x}\left(v\right)$ is uniformly continuous in x and v. Then similarly as bounding $B_{n,T}$ above, we use the mean value expansion under Condition 1.3, Lemma 1, and Conditions 1.3-1.4 to derive that for any $r \in \{1,...,k\}$ and some constant $0 < C < \infty$,

$$
\begin{aligned}
& \left|\frac{f_{Y_{i,[x_0]}}\left(a_n y_r + b_n\right)}{f_{Y|X=x_0}\left(a_n y_r + b_n\right)} - 1\right| \\
= \;& \left|\frac{\mathbb{E}_{X_{i,(x_0)}}\left[f_{Y|X=X_{i,(x_0)}}\left(v\right) - f_{Y|X=x_0}\left(a_n y_r + b_n\right)\right]}{f_{Y|X=x_0}\left(a_n y_r + b_n\right)}\right| \\
\le \;& \sup_{x \in B_{\eta_T}(x_0)} \left|\left|\frac{\partial f_{Y|X=x}\left(a_n y_r + b_n\right)/\partial x}{f_{Y|X=x_0}\left(a_n y_r + b_n\right)}\right|\right| \mathbb{E}\left[\left|\left|X_{i,(x_0)} - x_0\right|\right|\right] \\
\le \;& o(1) \times O\left(T^{-\eta}\right) \\
= \;& o(1).
\end{aligned}
$$

■

Proof of Corollary 1 By Corollary 1.2.4 and Remark 1.2.7 in de Haan and Ferreira (2007), the constants a_n and b_n can be chosen as follows. We introduce the case for α only, and the choice for β follows identically. If $\xi_\alpha > 0$, we choose $a_n(\xi_\alpha) = Q_\alpha(1-1/n)$ and $b_n(\xi_\alpha) = 0$, where recall $Q_\alpha(\cdot)$ denotes the quantile function of α_i. If $\xi_\alpha = 0$, we choose $a_n(\xi_\alpha) = 1/(nf_\alpha(b_n(\xi_\alpha)))$ and $b_n(\xi_\alpha) = Q_\alpha(1-1/n)$, where recall $f_\alpha(\cdot)$ denotes the PDF of α_i. By construction, these constants satisfy that $1 - F_\alpha(a_n(\xi_\alpha)y + b_n(\xi_\alpha)) = O(n^{-1})$ for any fixed $y > 0$ in both cases (cf. Chapter 1.1.2 in de Haan and Ferreira (2007)).

We first establish the convergence of $\mathbf{A}$. By the standard EV theory, Condition 2.1 (α_i is i.i.d.) and Condition 2.2 ($F_\alpha \in \mathcal{D}\left(G_{\xi_\alpha}\right)$) imply

$$\left(\frac{\alpha_{(1)} - b_n(\xi_\alpha)}{a_n(\xi_\alpha)}, ..., \frac{\alpha_{(k)} - b_n(\xi_\alpha)}{a_n(\xi_\alpha)}\right)^\intercal \xrightarrow{d} \mathbf{V}\left(\xi_\alpha\right), \tag{21}$$

where $\mathbf{V}(\xi_\alpha)$ is jointly EV distributed with tail index ξ_α.

Let $I = (I_1, \ldots, I_k) \in \{1, \ldots, T\}^k$ be the k random indices such that $\alpha_{(j)} = \alpha_{I_j}$, $j = 1, \ldots, k$, and let $\hat{I}$ be the corresponding indices such that $\hat{\alpha}_{(j)} = \hat{\alpha}_{\hat{I}_j}$. Then the convergence of $\mathbf{A}$ follows from (21) once we establish $|\hat{\alpha}_{\hat{I}_j} - \alpha_{I_j}| = o_p(a_n(\xi_\alpha))$ for $j = 1, \ldots, k$. We consider $k = 1$ for simplicity and the argument for a general k is very similar. Denote $\varepsilon_i \equiv \hat{\alpha}_i - \alpha_i$.

Consider the case with $\xi_\alpha > 0$. The part in Condition 2.3 for $\xi_\alpha > 0$ yields that

$$\begin{aligned} \sup_i |\varepsilon_i| &= \sup_i \left| \bar{X}_i^\intercal \left(\beta_i - \hat{\beta}_i\right) + \bar{u}_i \right| \\ &\leq \sup_i \left|\left|\bar{X}_i\right|\right| \sup_i \left|\left|\beta_i - \hat{\beta}_i\right|\right| + \sup_i |\bar{u}_i| \\ &= o_p(1). \end{aligned}$$

Given this, we have that, on one hand, $\hat{\alpha}_{\hat{I}} = \max_i\{\alpha_i + \varepsilon_i\} \leq \alpha_I + \sup_i |\varepsilon_i| = \alpha_I + o_p(1)$; and on the other hand, $\hat{\alpha}_{\hat{I}} = \max_i\{\alpha_i + \varepsilon_i\} \geq \max_i\{\alpha_i + \min_i\{\varepsilon_i\}\} \geq \alpha_I + \min_i\{\varepsilon_i\} \geq \alpha_I - \sup_i |\varepsilon_i| = \alpha_I - o_p(1)$. Therefore, $\left|\hat{\alpha}_{\hat{I}} - \alpha_I\right| \leq o_p(1) = o_p(a_n\left(\xi_\alpha\right))$ since $a_n(\xi_\alpha) \to \infty$.

Consider the case with $\xi_\alpha = 0$. Corollary 1.2.4 in de Haan and Ferreira (2007) implies that $a_n\left(\xi_\alpha\right) = f_\alpha\left(Q_\alpha(1-1/n)\right)$. Thus, the part in Condition 2.3 for $\xi_\alpha = 0$ implies that

$$\begin{aligned} \frac{1}{a_n\left(\xi_\alpha\right)} \sup_i |\varepsilon_i| &\leq \frac{\sup_i \left|\left|\bar{X}_i\right|\right| \sup_i \left|\left|\beta_i - \hat{\beta}_i\right|\right| + \sup_i |\bar{u}_i|}{f_\alpha\left(Q_\alpha(1-1/n)\right)} \\ &= o_p(1). \end{aligned}$$

Then the same argument as above yields that $\left|\hat{\alpha}_{\hat{I}} - \alpha_I\right| \leq O_p\left(\sup_i |\varepsilon_i|\right) = o_p(a_n\left(\xi_\alpha\right))$.

Now we establish the convergence of $\mathbf{B}$. Recall that we focus on, without loss of generality, the first component of β_i, so that $(\beta_{(1)}, ..., \beta_{(k)})^\intercal$ denote the largest k elements in the first components of $\{\beta_i\}_{i=1}^n$. Conditions 2.1 and 2.2 imply that

$$\left(\frac{\beta_{(1)} - b_n\left(\xi_\beta\right)}{a_n\left(\xi_\beta\right)}, ..., \frac{\beta_{(k)} - b_n\left(\xi_\beta\right)}{a_n\left(\xi_\beta\right)}\right)^\intercal \xrightarrow{d} \mathbf{V}\left(\xi_\beta\right).$$

Condition 2.3 and the similar argument for **A** complete the proof. ■

Proof of Proposition 1 To derive the asymptotic distributions of Hill's and Smith's estimators, we use Theorem 4.3.1 in Goldie and Smith (1987) and Proposition 3.1 in Smith (1987). Both results require $\mathbb{P}\left(Y_{i,[x_0]} \leq y\right)$ to satisfy their SR2 condition (cf. pp.1179 in Smith (1987)), which we establish now. Since our Condition 3.1 implies that $F_{Y|X=x_0}$ satisfies this SR2 condition with $\phi(v_n) = v_n^{-\tilde{\gamma}(x_0)}$ (cf. pp.1181 in Smith (1987)), it then suffices to show that for any y,

$$\begin{aligned} & \left|\frac{1-\mathbb{P}\left(Y_{i,[x_0]} \leq yv_n\right)}{1-\mathbb{P}\left(Y_{i,[x_0]} \leq v_n\right)} - \frac{1-F_{Y|X=x_0}(yv_n)}{1-F_{Y|X=x_0}(v_n)}\right| \\ \leq & \left|\frac{\mathbb{P}\left(Y_{i,[x_0]} \leq yv_n\right) - F_{Y|X=x_0}(yv_n)}{1-\mathbb{P}\left(Y_{i,[x_0]} \leq v_n\right)}\right| \\ & + \left|\frac{\mathbb{P}\left(Y_{i,[x_0]} \leq v_n\right) - F_{Y|X=x_0}(v_n)}{1-\mathbb{P}\left(Y_{i,[x_0]} \leq v_n\right)}\right| \times \left|\frac{1-F_{Y|X=x_0}(yv_n)}{1-F_{Y|X=x_0}(v_n)}\right| \\ \leq & \; o(\phi\left(v_n\right)). \end{aligned} \tag{22}$$

We show the first item in (22) is uniform $o(\phi(v_n))$. The second one follows similarly since Condition 3.1 implies $(1-F_{Y|X=x_0}(yv_n))/(1-F_{Y|X=x_0}(v_n)) = 1 + O(\phi(v_n))$.

First, the argument in pp.1181 in Smith (1987) and Condition 3.2 yield that $v_n = O\left(n^{1/(\gamma(x_0)+2\tilde{\gamma}(x_0))}\right)$. Then by Conditions 3.1 and 1.4, we derive that for some constant $C > 0$

$$\begin{aligned} & \sup_{x \in B_{\eta_T}(x_0)} \left\|\frac{\partial F_{Y|X=x}(v_n)/\partial x}{1-F_{Y|X=x}(v_n)}\right\| \\ \leq & \sup_{x \in B_{\eta_T}(x_0)} \left\|\frac{c_1(x)}{c(x)} - \log\left(v_n\right)\gamma_1(x) + Cd_1(x)v_n^{-\tilde{\gamma}(x)} \right. \\ & \left. -Cd(x)v_n^{-\tilde{\gamma}(x)}\tilde{\gamma}_1(x)\log(v_n) + Cr_1(x,v_n)\right\| \\ = & \; O(\log v_n) = O\left(\log n\right) \end{aligned} \tag{23}$$

and

$$\begin{aligned} \sup_{x \in B_{\eta_T}(x_0)} \left|\frac{1-F_{Y|X=x}(v_n)}{1-F_{Y|X=x_0}(v_n)}\right| &= \sup_{x \in B_{\eta_T}(x_0)} v_n^{-\gamma(x)+\gamma(x_0)} T^{-\eta} \log v_n \\ &= O\left(\exp\left(T^{-\eta}\log v_n\right) T^{-\eta}\log v_n\right) \\ &= O\left(n^{-\eta}\log n\right). \end{aligned} \tag{24}$$

Finally, apply the mean value expansion, Lemma 1, Condition 1.4, and (23)-(24) to obtain that for any $y > 0$

$$\left|\frac{\mathbb{P}\left(Y_{i,[x_0]} \leq yv_n\right) - F_{Y|X=x_0}(yv_n)}{1-\mathbb{P}\left(Y_{i,[x_0]} \leq v_n\right)}\right|$$

$$
\begin{aligned}
&= \left| \frac{\mathbb{E}_{X_{i,(x_0)}}\left[F_{Y|X=X_{i,(x_0)}}(yv_n) - F_{Y|X=x_0}(yv_n)\right]}{1-\mathbb{E}_{X_{i,(x_0)}}\left[F_{Y|X=X_{i,(x_0)}}(v_n)\right]} \right| \\
&\leq C\mathbb{E}\left[||X_{i,(x_0)} - x_0||\right] \frac{\sup_{x\in B_{\eta_T}(x_0)} ||\frac{\partial}{\partial x} F_{Y|X=x}(yv_n)||}{1-F_{Y|X=x_0}(v_n) + O(T^{-1/2} v_n^{-\gamma(x_0)} \log v_n)} \\
&= O(T^{-1/2}T^{-\eta}(\log n)^2) \\
&= o(\phi(v_n)),
\end{aligned}
$$

which establishes the SR2 condition. Given this SR2 condition and the fact that our Condition 3.2 is sufficient for eq.(4.3.3) in Goldie and Smith (1987) and eq.(3.2) in Smith (1987), the arguments in Theorem 4.3.1 in Goldie and Smith (1987) and Proposition 3.1 in Smith (1987) hold and complete the proof. ■

Proof of Lemma 1 We first prove (18). The subscript i is suppressed for notional ease. Define $D_t = ||X_t - x_0||$ for $t \in \{1, \ldots, T\}$, which is still strictly stationary and β-mixing. By Berbee's lemma (enlarging the probability space as necessary), the process $\{D_t\}$ can be coupled with a process $\{D_t^*\}$ that satisfies the following three properties: (i) $Z_i \equiv \{D_{(i-1)\times q_T+1}, \ldots, D_{i\times q_T}\}$ and $Z_i^* \equiv \{D^*_{(i-1)\times q_T+1}, \ldots, D^*_{i\times q_T}\}$ are identically distributed for all $i \in \{1, \ldots, k_T\}$, where Z_i^* is the same decomposition of $\{D_t^*\}$ as Z_i and $k_T \times q_T = T$; (ii) $\mathbb{P}(Z_i^* \neq Z_i) \leq \beta(q_T)$ for all $i \in \{1, \ldots, k_T\}$; and (iii) $\{Z_1^*, Z_3^*, \ldots\}$ are independent and $\{Z_2^*, Z_4^*, \ldots\}$ are independent (cf. Lemma 2.1 in Berbee (1987) and Proposition 2 in Doukhan, Massart, and Rio (1995)). Suppose k_T is an even integer for simplicity and define U_i^* as i.i.d. standard uniform random variable. Then these properties yield that

$$
\begin{aligned}
&\mathbb{P}\left(\min_{t\in\{1,\ldots,T\}}\{D_t\} > \varepsilon T^{-\eta}\right) \\
&= \mathbb{P}\left(\min_{t\in\{1,\ldots,T\}}\{D_t\} > \varepsilon T^{-\eta}, \{D_t\}_{t=1}^T = \{D_t^*\}_{t=1}^T\right) \\
&\quad + \mathbb{P}\left(\min_{t\in\{1,\ldots,T\}}\{D_t\} > \varepsilon T^{-\eta}, \{D_t\}_{t=1}^T \neq \{D_t^*\}_{t=1}^T\right) \\
&\leq_{(1)} \mathbb{P}\left(\min_{t\in\{2q_T,4q_T\ldots,k_Tq_T\}}\{D_t^*\} > \varepsilon T^{-\eta}\right) + \mathbb{P}\left(\{D_t\}_{t=1}^T \neq \{D_t^*\}_{t=1}^T\right) \\
&\leq_{(2)} \mathbb{P}\left(\min_{i\in\{1,2,\ldots,k_T/2\}}\{U_i^*\} > F_D\left(\varepsilon T^{-\eta}\right)\right) + \mathbb{P}\left(\{D_t\}_{t=1}^T \neq \{D_t^*\}_{t=1}^T\right) \\
&\leq_{(3)} (1-CT^{-\eta})^{k_T/2} + k_T\beta\left(q_T\right),
\end{aligned}
$$

where inequality(1) follows by considering the first elements in all even blocks, which are independent by property(iii) above, inequality(2) follows from the CDF transformation, and inequality(3) follows from the CDF of the standard uniform distribution and properties (ii) and (iii) above.

Choosing k_T as the largest even integer no larger than $2T^{1/3}$ and using Condition 1.1 again yield that

$$\begin{aligned}
& \sum_{T=1}^{\infty} \mathbb{P}\left(\min_{t\in\{1,\ldots,T\}}\{D_t\} > \varepsilon T^{-\eta}\right) \\
\leq\ & \sum_{T=1}^{\infty}(1-cT^{-\eta})^{T^{1/3}} + \sum_{T=1}^{\infty} T^{1/3} O\left(T^{-4/3-2\varepsilon}\right) \\
<\ & \infty \text{ for any } \eta \in (0,1/3).
\end{aligned}$$

Then $T^{\eta}\left|\left|X_{(x_0)} - x_0\right|\right| = o_{a.s.}(1)$ is implied by Borel Cantelli Lemma. The convergence of $\sum_{T=1}^{\infty}(1 - cT^{-\eta})^{T^{1/3}}$ is checked by the ratio test that $\lim_{T\to\infty}(1-c(T+1)^{-\eta})^{(T+1)^{1/3}}/(1-cT^{-\eta})^{T^{1/3}} < 1$. Thus, (18) holds with any $\eta \in (0,1/3)$.

Now we prove (19). Perform the same coupling argument as above and consider the minimum value within each block Z_i (and Z_i^*), denoted $\min\{Z_i\}$ (and $\min\{Z_i^*\}$). Let E_T denote the event that $\{D_t\}_{t=1}^T = \{D_t^*\}_{t=1}^T$. The above three properties and (18) yield that for some constant $C > 0$,

$$\begin{aligned}
& \mathbb{E}[\left|\left|X_{(x_0)} - x_0\right|\right|] \\
&= \mathbb{E}\left[\min_{t\in\{1,\ldots,T\}}\{D_t\}\mathbf{1}[E_T]\right] + \mathbb{E}\left[\min_{t\in\{1,\ldots,T\}}\{D_t\}\mathbf{1}[E_T^c]\right] \\
&\leq_{(1)} \mathbb{E}\left[\min_{i\in\{2,4,\ldots,k_T\}}\{\min\{Z_i\}\}\mathbf{1}[E_T]\right] + CT^{-\eta}\mathbb{E}\left[\mathbf{1}[E_T^c]\right] \\
&\leq_{(2)} \mathbb{E}\left[\min_{i\in\{2,4,\ldots,k_T\}}\{\min\{Z_i^*\}\}\right] + CT^{-\eta}k_T\beta\left(q_T\right) \\
&\leq_{(3)} \mathbb{E}\left[\min_{i\in\{2,4,\ldots,k_T\}}\{D_{i\times q_T}^*\}\right] + CT^{-\eta}k_T\beta\left(q_T\right),
\end{aligned}$$

where inequality(1) follows from considering even blocks only and (18), inequality(2) follows from property(ii) above, and inequality(3) follows from the fact that $\min\{Z_i^*\} \leq D_{i\times q_T}^*$ (the minimum value within the block Z_i^* is less than or equal to the last element in that block).

The second term in the last step above is $o(T^{-1/2})$ by setting $q_T = k_T$ equal to the largest even integer no larger than $T^{1/2}$. Regarding the first item above, notice that $\min_{i\in\{1,\ldots,k_T\}}\{D_{i\times q_T}^*\}$ is the sample minimum of $k_T/2$ random samples from some CDF $F_D(\cdot)$, which has the bounded lower end-point 0. Condition 1.1 implies that $F_D(\cdot)$ is continuously differentiable and monotonically increasing in a neighborhood of zero. Then we have

$$\begin{aligned}
\mathbb{E}\left[\min_{i\in\{2,4,\ldots,k_T\}}\{D_{i\times q_T}^*\}\right] &= \mathbb{E}\left[F_D^{-1}\left(\min_{i\in\{2,4,\ldots,k_T\}}\{U_i^*\}\right)\right] \\
&=_{(1)} \mathbb{E}\left[\left(1/f_D\left(F_D^{-1}(\dot{u})\right)\right)\min_{i\in\{2,4,\ldots,k_T\}}\{U_i^*\}\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq_{(2)} C\mathbb{E}\left[\min_{i\in\{2,4,\ldots,k_T\}}\{U_i^*\}\right] \\
&=_{(3)} O(k_T^{-1}),
\end{aligned}
$$

where equality(1) follows from mean value expansion with some $\dot{u}$ between 0 and $\min_{i\in\{2,4,\ldots,k_T\}}\{U_i^*\}$, inequality(2) follows from the fact that $f_D(\cdot)$ is uniformly bounded away from 0 in a neighborhood of zero, which is implied by Condition 1.1 again, and equality(3) follows from Theorem 5.3.1 in de Haan and Ferreira (2007) since U_i^* is i.i.d. standard uniform distribution with the tail index -1. So (19) is established by setting k_T equal to the largest even integer no larger than $T^{1/2}$ again. ■

Proof of Lemma 2 The proof is different for $\xi(x_0) >$ or $= 0$. We first consider the positive $\xi(x_0)$ case. Recall that $B_{\eta_T}(x_0)$ denotes an open ball centered at x_0 with radius $\eta_T = T^{-\eta}$, where η is determined in Lemma 1. For (a), $\inf_{x\in B_{\eta_T}(x_0)}\xi(x) > 0$ if T is large enough. This is feasible given the continuity of $\xi(\cdot)$. Then by the chain rule and the condition (Condition B.(i)) that

$$
1 - F_{Y|X=x}(y) = c(x)y^{-\gamma(x)}(1 + d(x)(y)^{-\tilde{\gamma}(x)} + r(x,y)), \tag{25}
$$

we have

$$
\begin{aligned}
&\frac{\partial F_{Y|X=x}(y)/\partial x}{1 - F_{Y|X=x}(y)} \\
= \ & \frac{c_1(x)}{c(x)} - \gamma_1(x)\log y + \frac{d_1(x)y^{-\tilde{\gamma}(x)}}{1 + d(x)(y)^{-\tilde{\gamma}(x)} + r(x,y)} \\
&- \frac{d(x)y^{-\tilde{\gamma}(x)}\tilde{\gamma}_1(x)\log y}{1 + d(x)(y)^{-\tilde{\gamma}(x)} + r(x,y)} + \frac{r_1(x,y)}{1 + d(x)(y)^{-\tilde{\gamma}(x)} + r(x,y)}.
\end{aligned}
$$

Recall that

$$
\begin{aligned}
u_n &= a_n y + b_n \\
&= O(Q_{Y|X=x_0}(1-1/n)) \\
&= O(n^{\xi(x_0)})
\end{aligned} \tag{26}
$$

(cf. Corollary 1.2.4 and Remark 1.2.11 in de Haan and Ferreira (2007)). Then after applying the triangle inequality and the smoothness and boundedness of $c(\cdot)$, $d(\cdot)$, and $\gamma(\cdot)$ (Condition B.(i)), we have that for some constant $C > 0$,

$$
\begin{aligned}
&\sup_{x\in B_{\eta_T}(x_0)}\left\|\frac{\partial F_{Y|X=x}(u_n)/\partial x}{1 - F_{Y|X=x}(u_n)}\right\| \qquad (27)\\
\leq \ & \sup_{x\in B_{\eta_T}(x_0)}\left\|\frac{c_1(x)}{c(x)} - \log(u_n)\gamma_1(x) + Cd_1(x)(u_n)^{-\tilde{\gamma}(x)}\right.
\end{aligned}
$$

$$
\begin{aligned}
&\quad -Cd(x)u_n^{-\tilde{\gamma}(x)}\tilde{\gamma}(x)\log(u_n) + Cr_1(x,u_n)\Big\| \\
&= O(\log(u_n)) \text{ (by Condition B.(i))} \\
&= O(\log n). \text{ (by (26))}
\end{aligned}
$$

By (25) again, we have

$$
\begin{aligned}
&\sup_{x\in B_{\eta_T}(x_0)}\left|\frac{1-F_{Y|X=x}(u_n)}{1-F_{Y|X=x_0}(u_n)}\right| \qquad (28)\\
\le\ & \sup_{x\in B_{\eta_T}(x_0)}\left|u_n^{-\gamma(x)+\gamma(x_0)}\right| \sup_{x\in B_{\eta_T}(x_0)}\left|\frac{c(x)}{c(x_0)}\right| \sup_{x\in B_{\eta_T}(x_0)}\left|\frac{1+d(x)(y)^{-\tilde{\gamma}(x)}+r(x,u_n)}{1+d(x_0)(y)^{-\tilde{\gamma}(x_0)}+r(x_0,u_n)}\right| \\
\le\ & C\exp\left(\sup_{x\in B_{\eta_T}(x_0)}\log\left(u_n^{-\gamma(x)+\gamma(x_0)}\right)\right) \quad \text{(by Condition B.(i))} \\
=\ & C\exp\left(O(T^{-\eta}\log(u_n))\right) \\
=\ & C\exp(O(T^{-\eta}\log n) \qquad \text{(by (26))} \\
=\ & O(1). \qquad \text{(by Condition 1.4)}
\end{aligned}
$$

Then part (a) follows by combining (27) and (28) and using $O(T^{-\eta})\times O(\log n) = o(1)$ by Condition 1.4 again.

For (b), Condition B.(i) implies that

$$
\begin{aligned}
f_{Y|X=x}(y) &= -c(x)\gamma(x)(y)^{-\gamma(x)-1}(1+d(x)(y)^{-\tilde{\gamma}(x)}+r(x,y)) \qquad (29)\\
&\quad +c(x)(y)^{-\gamma(x)}(-d(x)y^{-\tilde{\gamma}(x)-1}\tilde{\gamma}(x)+r_2(x,y)).
\end{aligned}
$$

A similar argument as above with Conditions B.(i) and 1.4 yields

$$
\sup_{x\in B_{\eta_T}(x_0)}\left\|\frac{\partial f_{Y|X=x}(u_n)/\partial x}{f_{Y|X=x}(u_n)}\right\| \le O(\log(u_n)) = O(\log n)
$$

and

$$
\sup_{x\in B_{\eta_T}(x_0)}\left\|\frac{f_{Y|X=x}(u_n)}{f_{Y|X=x_0}(u_n)}\right\| \le C\exp\left(\sup_{x\in B_{\eta_T}(x_0)}\log\left(u_n^{-\gamma(x)+\gamma(x_0)}\right)\right) = O(1),
$$

which yield part (b) by using Condition 1.4 again.

Now it remains to prove (a) and (b) for $\xi(x_0)=0$. Note that $u_n = O\left(Q_{Y|X=x_0}(1-1/n)\right)$, which is at most of the order $\exp(\Phi^{-1}(1-1/n)) = \exp(\sqrt{2\log n})$ by the condition $C_1(\log y)^2 \le \tilde{d}(y) \le C_2 y^{C_3}$.

For (a), we decompose $B_{\eta_T}(x_0)$ into $B_{\eta_T}(x_0)\cap\{x:\xi(x)>0\}$ and $B_{\eta_T}(x_0)\cap\{x:\xi(x)=0\}$, and then

$$
\sup_{x\in B_{\eta_T}(x_0)}\left\|\frac{\partial F_{Y|X=x}(u_n)/\partial x}{1-F_{Y|X=x_0}(u_n)}\right\|
$$

$$\leq \max\left\{ \sup_{x\in B_{\eta_T}(x_0)\cap\{x:\xi(x)>0\}} \left|\left| \frac{\partial F_{Y|X=x}(u_n)/\partial x}{1-F_{Y|X=x_0}(u_n)} \right|\right|, \sup_{x\in B_{\eta_T}(x_0)\cap\{x:\xi(x)=0\}} \left|\left| \frac{\partial F_{Y|X=x}(u_n)/\partial x}{1-F_{Y|X=x_0}(u_n)} \right|\right| \right\}. \quad (30)$$

For the first item in (30), Conditions 1.1 and B.(i) imply that $\partial F_{Y|X=x}(u_n)/\partial x = O(u_n^{-\gamma(x)}\log u_n)$ and $\gamma(x) = 1/\xi(x) = O\left(1/\xi'(\dot{x})T^{\eta}\right) \geq O(T^{\eta})$ where $\dot{x}$ is within $B_{\eta_T}(x_0)$. Thus, Condition 1.4 and the fact that $1-F_{Y|X=x_0}(u_n) = O(n^{-1})$ yield that for any $x \in B_{\eta_T}(x_0)\cap\{x:\xi(x)>0\}$,

$$\begin{aligned}
& \left|\left| \frac{\partial F_{Y|X=x}(u_n)/\partial x}{1-F_{Y|X=x_0}(u_n)} \right|\right| \\
= \ & O\left(n \times u_n^{-\gamma(x)}\log u_n\right) \\
= \ & O\left(\exp\left(\log n - \gamma(x)\log u_n + \log\left(\log u_n\right)\right)\right) \\
\leq \ & O(\exp\left(\log n - T^{\eta}\log u_n + \log\left(\log u_n\right)\right)) \\
= \ & o(1).
\end{aligned}$$

For the second term in (30), apply Leibniz's rule and Condition B.(ii) to obtain

$$\begin{aligned}
& \sup_{x\in B_{\eta_T}(x_0)\cap\{x:\xi(x)=0\}} \left|\left| \frac{\partial F_{Y|X=x}(u_n)/\partial x}{1-F_{Y|X=x_0}(u_n)} \right|\right| \\
\leq \ & \sup_{x\in B_{\eta_T}(x_0)\cap\{x:\xi(x)=0\}} Cn\int_{u_n}^{y_0} y^{C_3} f_{Y|X=x}(y)dy \\
\leq \ & Cn\int_{u_n}^{y_0} y^{C_3+\bar{C}_T}\exp(-\underline{D}_T\left(\log y\right)^2)dy \qquad (31) \\
= \ & Cn\int_{\log u_n}^{y_0} \exp(-\underline{D}_T s^2 + (C_3+\bar{C}_T+1)s)ds \text{ (by change of variables)} \\
= \ & O(1),
\end{aligned}$$

where we denote $\bar{C}_T = \sup_{x\in B_{\eta_T}(x_0)}\tilde{c}(x) < \infty$ and $\underline{D}_T = \inf_{x\in B_{\eta_T}(x_0)} d(x) > 0$, and the last equation follows from that u_n is at most of the order $\exp(\sqrt{2\log n})$ and the fact that the $1-1/n$ quantile of a normal distribution is $O(\sqrt{\log(n)})$.

For (b), we similarly derive

$$\begin{aligned}
\sup_{x\in B_{\eta_T}(x_0)} \left|\left| \frac{\partial f_{Y|X=x}(u_n)/\partial x}{f_{Y|X=x_0}(u_n)} \right|\right| \leq \max\Bigg\{ & \sup_{x\in B_{\delta_T}(x_0)\cap\{x:\xi(x)>0\}} \left|\left| \frac{\partial f_{Y|X=x}(u_n)/\partial x}{f_{Y|X=x_0}(u_n)} \right|\right|, \\
& \sup_{x\in B_{\delta_T}(x_0)\cap\{x:\xi(x)=0\}} \left|\left| \frac{\partial f_{Y|X=x}(u_n)/\partial x}{f_{Y|X=x_0}(u_n)} \right|\right| \Bigg\}.
\end{aligned}$$

Using (29) and Condition B.(i), we have $\left|\left|\partial f_{Y|X=x}(u_n)/\partial x\right|\right| = O\left(u_n^{-\gamma(x)-1}\gamma(x)\right) + O\left(u_n^{-\gamma(x)}\right)$ when $\xi(x)>0$. By Condition B.(ii) and under $\xi(x_0)=0$, we have $1/f_{Y|X=x_0}(u_n) \leq Cu_n\exp\left(\bar{D}_T C_2 u_n^{C_3}\right)$ where we denote $\bar{D}_T = \sup_{x\in B_{\eta_T}(x_0)} d(x) > 0$. Thus for any $x \in B_{\delta_T}(x_0) \cap \{x : \xi(x) > 0\}$,

$$\begin{aligned}
& \left\| \frac{\partial f_{Y|X=x}(u_n)/\partial x}{f_{Y|X=x_0}(u_n)} \right\| \\
\leq\; & C u_n \exp\left(\bar{D}_T C_2 u_n^{C_3}\right)\left(u_n^{-\gamma(x)-1}\gamma(x) + u_n^{-\gamma(x)}\right) \\
=\; & C u_n \exp\left(\bar{D}_T C_2 u_n^{C_3} - (\gamma(x)+1)\log(u_n) + \log\gamma(x)\right) \\
& + u_n C \exp\left(\bar{D}_T C_2 u_n^{C_3} - \gamma(x)\log(u_n)\right) \\
\leq\; & C u_n \exp\left(\bar{D}_T C_2 u_n^{C_3} - T^{-\eta}\log(u_n) + \log\gamma(x)\right) \\
=\; & o(1),
\end{aligned}$$

where the last line follows from Condition 1.4 and the fact that u_n is at most of the order $\exp(\sqrt{2\log n})$.

The second term is bounded by

$$\begin{aligned}
& \sup_{x \in B_{\eta_T}(x_0)} \left\| \frac{c_1(x)}{c(x)} + \frac{1}{u_n}\tilde{c}_1(x) + d_1(x)\tilde{d}(u_n) + \frac{r_1(x,u_n)}{1+r(x,u_n)} \right\| \\
\leq\; & O(u_n^{C_3}) \leq O((\log(n))^{C_3/2}).
\end{aligned}$$

Thus (b) for $\xi(x_0) = 0$ is established. ■

References

Abrevaya, J. (2001): "The effects of demographics and maternal Behavior on the distribution of birth outcomes," *Empirical Economics*, 26, 247–257.

Adrian, T., and M. K. Brunnermeier (2016): "CoVaR," *American Economic Review*, 106(7), 1705–1741.

Arnold, B. C., N. Balakrishnan, and H. H. N. Nagaraja (1992): *A First Course in Order Statistics*. Siam.

Backus, D., M. Chernov, and I. Martin (2011): "Disasters implied by equity index options," *The Journal of Finance*, 66(6), 1969–2012.

Beare, B., and A. A. Toda (2017): "Geometrically Stopped Markovian Random Growth Processes and Pareto Tails," *arXiv:1712.01431*.

BEIRLANT, J., E. JOOSSENS, AND J. SEGERS (2004): "Discussion of "Generalized Pareto Fit to the Society of Actuaries' Large Claims Database" by A Cebrian, M Denuit and P Lambert," *North American Actuarial Journal*, 8, 108–111.

BERBEE, H. (1987): "Convergence Rates in the Strong Law for Bounded Mixing Sequences," *Probability Theory and Related Fields*, 74, 255–270.

BOLLERSLEV, T., AND V. TODOROV (2011): "Tails, fears, and risk premia," *The Journal of Finance*, 66(6), 2165–2211.

CHEN, J., H. HONG, AND J. C. STEIN (2001): "Forecasting crashes: trading volume, past returns, and conditional skewness in stock prices," *Journal of Financial Economics*, 61, 345–381.

CHERNOZHUKOV, V. (2005): "Extremal Quantile Autoregression," *The Annals of Statistics*, 33(2), 806–839.

CHERNOZHUKOV, V., AND I. FERNÁNDEZ-VAL (2011): "Inference for extremal conditional quantile models, with an application to market and birthweight Risks," *The Review of Economic Studies*, 78, 559–589.

CHERNOZHUKOV, V., I. FERNÁNDEZ-VAL, AND T. KAJI (2017): "Extremal Quantile Regression: An Overview," *arXiv: 1612.06850.*

CHERNOZHUKOV, V., AND L. UMANTSEV (2001): "Conditional value-at-risk: Aspects of modeling and estimation," *Empirical Economics*, 26(1), 271–293.

COLES, S. (2001): *An Introduction to Statistical Modeling of Extreme Values*. Springer, London.

DAOUIA, A., L. GARDES, AND S. GIRARD (2013): "On Kernel Smoothing for Extremal Quantile Regression," *Bernoulli*, 19, 2557–2589.

DE HAAN, L., AND A. FERREIRA (2007): *Extreme Value Theory: An Introduction.* Springer Science and Business Media, New York.

DING, P. (2016): "On the Conditional Distribution of the Multivariate t Distribution," *The American Statistician*, 70, 293–295.

DOUKHAN, P., P. MASSART, AND E. RIO (1995): "Invariance principles for absolutely regular empirical processes," *Annales de l'I.H.P. Probabilités et statistiques*, 31(2), 393–427.

ENGLE, R. F., AND S. MANGANELLI (2004): "CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles," *Journal of Business & Economic Statistics*, 22(4), 367–381.

FAMA, E. F., AND K. R. FRENCH (1993): "Common Risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, 33, 3–56.

——— (2015): "A five-factor asset pricing model," *Journal of Financial Economics*, 116, 1–22.

FISHER, R. A., AND L. H. C. TIPPETT (1928): *"Limiting forms of the frequency distribution of the largest or smallest member of a sample." Mathematical Proceedings of the Cambridge Philosophical Society.* Cambridge University Press.

GABAIX, X., J. LASRY, P. LIONS, AND B. MOLL (2016): "The Dynamics of Inequality," *Econometrica*, 85(6), 2071–2111.

GARDES, L., S. GIRARD, AND A. LEKINA (2010): "Functional Nonparametric Estimation of Conditional Extreme Quantiles," *Journal of Multivariate Analysis*, 101, 419–433.

GARDES, L., A. GUILLOU, AND A. SCHORGEN (2012): "Estimating the Conditional Tail Index by Integrating a Kernel Conditional Quantile Estimator," *Journal of Statistical Planning and Inference*, 142, 1586–1598.

GNEDENKO, B. V. (1943): "Sur la distribution limite du terme maximum d'une serie aléatoire," *Annals of Mathematics*, 44, 423–453.

GOLDIE, C. M., AND R. L. SMITH (1987): "Slow Variation with Remainder: Theory and Applications," *the Quarterly Journal of Mathematics*, 38, 45–71.

HALL, P. (1982): "On Some Simple Estimates of an Exponent of Regular Variation," *Journal of Royal Statistic Society, Series B*, 44(1), 37–42.

HILL, B. M. (1975): "A Simple General Approach to Inference about the Tail of a Distribution," *Annals of Statistics*, 3(5), 1163–1174.

HSIAO, C., AND M. PESARAN (2004): "Random Coefficient and Panel Data Models," .

JONES, C. I., AND J. KIM (2018): "A Schumpeterian Model of Top Income Inequality," *Journal of Political Economy*, 126(5), 1785–1826.

Kelly, B., and H. Jiang (2014): "Tail Risk and Asset Prices," *The Review of Financial Studies*, 27(10), 2841–2871.

Koenker, R., and G. S. Bassett (1978): "Regression Quantiles," *Econometrica*, 46, 33–50.

Koenker, R., and K. Hallock (2001): "Quantile Regression: An introduction," *Journal of Economic Perspectives*, 15, 143–156.

Li, Q., and J. S. Racine (2007): *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

Martins-Filho, C., F. Yao, and M. Torero (2018): "Nonparametric Estimation of Conditional Valuat-at-Risk and Expected Shortfall Based on Extreme Value Theory," *Econometric Theory*, 34, 23–67.

Müller, U. K., and Y. Wang (2017): "Fixed-k Asymptotic Inference about Tail Properties," *the Journal of the American Statistical Association*, 112, 1134–1143.

Pickands, III, J. (1975): "Statistical inference using extreme order statistics," *Annals of Statistics*, 3(1), 119–131.

Piketty, T., and E. Saez (2003): "Income inquality in the United States, 1913-1998," *The Quarterly Journal of Economics*, 118(1), 1–41.

Smith, R. L. (1982): "Uniform rates of convergence in Extreme-Value theory," *Advances in Applied Probability*, 13(3), 600–622.

——— (1987): "Estimating Tails of Probability Distributions," *Annals of Statistics*, 15, 1174–1207.

Toda, A. A. (2019): "Wealth Distribution with Random Discount Factors," *Journal of Monetary Economics*, 104, 101–113.

Wang, H., and D. Li (2013): "Estimation of Extreme Conditional Quantiles Through Power Transformation," *Journal of the American Statistical Association*, 108(503), 1062–1074.

Wang, H., and C. L. Tsai (2009): "Tail Index Regression," *Journal of the American Statistical Association*, 104, 1233–1240.

Wang, Y., and Z. Xiao (2019): "Nearly Unbiased Estimation of Tail Features with Complete, Censored, or Truncated Data," *Working Paper*.

Wooldridge, J. M. (2002): *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, Cambridge, Massachusetts.

——— (2005): "Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models," *Review of Economics and Statistics*, 87, 385–390.

Yang, S. S. (1977): "General distribution theory of the concomitants of order Statistics," *The Annals of Statistics*, 5(5), 996–1002.