Statistica Sinica Preprint No: SS-2024-0196

Title	Distinguishing Forms of Asymptotic Dependence in Heavy Tailed Data
Manuscript ID	SS-2024-0196
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0196
Complete List of Authors	Tiandong Wang and Sidney I. Resnick
Corresponding Authors	Tiandong Wang
E-mails	td_wang@fudan.edu.cn

Notice: Accepted author version.

Statistica Sinica **24** (201?), 000-000
doi:http://dx.doi.org/10.5705/ss.20??.???

Distinguishing Forms of Asymptotic Dependence in Heavy Tailed Data

Tiandong Wang

Shanghai Center for Mathematical Sciences, Fudan University

Shanghai Academy of Artificial Intelligence for Science

Sidney I. Resnick

Cornell University

Abstract: In multivariate heavy tail estimation, the support of the limit measure provides information on the asymptotic dependence structure of the random vector with the heavy tail distribution. This asymptotic dependence structure may be difficult to discern, even in favorable cases of $\mathbb{R}_+^2$-valued data since exploratory methods can be ambiguous and heavily dependent on threshold choice. We restrict ourselves to techniques that help distinguish between the following asymptotic models for heavy tails on $\mathbb{R}_+^2$: (i) full dependence where the limit measure concentrates on a ray from the origin; (ii) strong dependence where the support of the limit measure is a proper connected subcone of the positive quadrant; (iii) weak dependence where the limit measure concentrates on the whole positive quadrant. We propose two test statistics, analyze their asymptotically normal behavior under full and not-full dependence, and discuss method implementation using bootstrap methods. The methodology is illustrated with both simulated and real data.

Keywords: Multivariate extremes, asymptotic dependence, hidden regular variation

1. Introduction

In multivariate heavy tail estimation, the support of the limit measure provides information on the dependence structure of the random vector with the heavy tail distribution (Lehtomaa and Resnick, 2020). However, even in simple circumstances in $\mathbb{R}_+^2$, the positive quadrant in two dimensions, exploratory methods such as scatter, diamond or density estimation plots may have trouble distinguishing between cases:

- *Full dependence* where the limit measure concentrates on a ray of the form $\{(x, y) \in \mathbb{R}_+^2 : y/x = m > 0\}$;
- *Strong dependence* (See Figure 2) where the support of the limit measure is a proper *connected* subcone of $\mathbb{R}_+^2$ of the form $\{(x, y) \in \mathbb{R}_+^2 : y/x \in [m_l, m_u] \subsetneq [0, 1]\}$;
- *Weak dependence* where the support of the limit measure is all of $\mathbb{R}_+^2$; and
- *Asymptotic independence* where the limit measure concentrates on the axes $\mathbb{R}_+ \times \{0\} \cup \{0\} \times \mathbb{R}_+$.

Estimation and visualization techniques reasonably detect lack of connectedness in the second bullet so we downplay that possibility. However, exploratory techniques can struggle to distinguish the bulleted cases, the most obvious reason being the requirement that data be thresholded according to the distance from the origin. Plots can look rather different depending on the choice of threshold. This is illustrated by diamond plots Das and Resnick (2017); Lehtomaa and Resnick (2020) in Figure 1 of Exxon (XOM) returns vs returns from Chevron (CVX) from January 04, 2016 to December 30, 2022. The data $\{(x_i, y_i); 1 \leq i \leq 1761\}$ is mapped to the L_1-unit sphere via $(x, y) \mapsto (x, y)/(|x| + |y|)$ and then subsetted by retaining only the points with k largest values of $(|x| + |y|)$ where $k = 100$ (left) or $k = 500$ (right). The two plots give different impressions of where the limit measure concentrates.

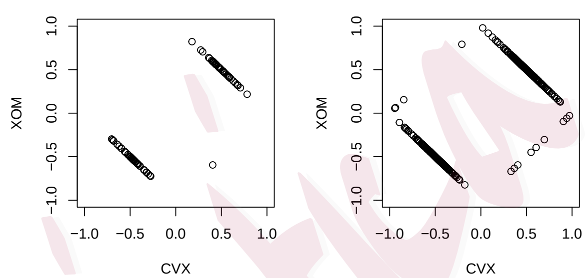

Figure 1: Left: Diamond plot using the 100 points furthest from the origin. Right: Diamond plot using 500 most remote points.

Threshold selection is addressed by a data-driven technique, popular in computer and network science (Clauset et al., 2009; Virkar and Clauset, 2014) and implemented in Csardi and Nepusz (2006) or Gillespie (2015). This technique is consistent (Bhattacharya et al., 2020; Drees et al., 2020) and may increase comfort level with threshold selection. However, the selection method offers no guarantee of optimal threshold choice and has the additional drawback of preventing tail estimators such as Hill from being asymptotically normal (Drees et al., 2020). Sensitivity of exploratory methods to threshold choice means test statistics capable of assisting choice of model from the bulleted list above would be welcome.

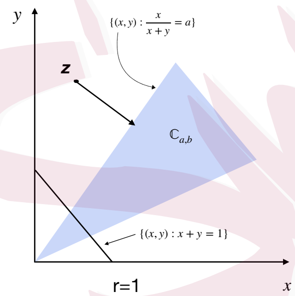

Figure 2: The concentration cone $\mathbb{C}_{a,b}$. Concentration on $\mathbb{C}_{a,b}$ is strong dependence. If $a = b$ we have full dependence.

One motivation for thinking about distinguishing between the bulleted cases above came from previous efforts to fit preferential attachment (PA) models of directed social networks to data consisting of in- and out-degree of each node. The classical PA model of directed edge growth (Krapivsky and Redner, 2001; Bollobás et al., 2003) when standardized to equal tail indices for each component gives a heavy tail model with limit measure concentrating on all of $\mathbb{R}_+^2$ (Samorodnitsky et al., 2016), the case of weak dependence. However, adding the reciprocity feature to the theoretical model leads to a limit measure that concentrates on a ray, the full dependence case (Wang and Resnick, 2022a,b; Cirkovic et al., 2023; Wang and Resnick, 2023). If there were a statistical test for full dependence, it would provide guidance on whether one needs to add the reciprocity feature to the model to obtain a satisfactory statistical fit for social network data.

We understand that network data or financial returns are not the same as iid observations but the theory in this paper starts with the basic case and assumes all observations come from a heavy tailed iid model by repeated sampling. Because preferential attachment network data does not satisfy the independent assumption, we reserve consideration of this case for a future project. Also, there is current interest in the topic of extremal clustering (Drees et al., 2021; Drees and Sabourin, 2021; Janßen and Wan, 2020; Fomichov and Ivanovs, 2023; Davis et al., 2023) which has some philosophical connections to the current work but is less focused on identifying correct asymptotic models.

We give two test statistics D_n and T_n which help distinguish full vs not-full asymptotic dependence and show the statistics are asymptotically normal but with different asymptotic variances, depending on the case. As is typical in heavy tail analysis, to get asymptotic normality with a constant centering for estimators requires not only the regular variation assumption for the underlying distribution but also second order regular variation (2RV) which controls deviations between a finite sample mean and an asymptotic mean; this is explained in de Haan and Ferreira (2006); Resnick (2007); de Haan and Resnick (1996); de Haan (1996); Peng (1998). Our multivariate version of the 2RV condition, (2.6) or (2.8), is similar to one used in Einmahl et al. (2021) but we gain some theoretical flexibility by posing the condition as $\mathbb{M}$-convergence of signed measures. Interestingly, 2RV coupled with multivariate regular variation with limit measure concentrating on a proper subset of the state space imply hidden regular variation (HRV) and this derived HRV is employed in our proofs. This is explained in Section 2. See also (de Haan and Resnick, 1993; de Haan and de Ronde, 1998; Resnick, 2002; Das and Kratz, 2020).

The proposed hypothesis testing framework is discussed in more detail in Section 5 and here is a sketch of our procedure where we also explain the necessity of using bootstrap techniques.

1.1 Guidance for distinguishing between cases.

Suppose we have heavy tailed data in $\mathbb{R}_+^2$ from the iid model $\{\boldsymbol{Z}_i = (X_i, Y_i) : 1 \le i \le n\}$ with L_1-polar coordinates $R_i = X_1 + Y_i$, $\Theta_i = X_i/(X_i+Y_i)$, $1 \le i \le n$. The limit measure of regular variation of $\mathbb{P}[\boldsymbol{Z}_1 \in \cdot]$ is $\eta(\cdot)$, the angular measure on $[0,1]$ is $S(\cdot)$, the Pareto measure is $\nu_\alpha(x,\infty) = x^{-\alpha}$ and for some regularly varying scaling function $b(t) \to \infty$, as $t \to \infty$, we have $tP[\big(R_1/b(t), \Theta_1\big) \in \cdot] \to \nu_\alpha \times S(\cdot)$. More complete explanations are in the next section.

Now consider the following tests where a, b, θ_0 are given (but later estimated). We begin in Step 1 by using a fairly crude distance based statistic D_n to perform a preliminary screen based on hypotheses formulated from exploratory plots or estimates of a, b.

Step 1: $H_0^{(1)} : S([a,b]) = 1$ vs $H_a^{(1)} : S([a,b]) < 1$.

Failure to reject $H_0^{(1)}$ could be due to $S(\cdot)$ concentrating at $\theta_0 \in [a,b]$ so we proceed to test for full dependence using an asymptotically normal statistic T_n:

Step 2: $H_0^{(2)}: S(\{\theta_0\}) = 1$ vs $H_a^{(2)}: S([0,1] \setminus \{\theta_0\}) > 0$.

However, since the distance-based test in Step 1 is crude, if either

(a) the test statistic T_n rejects $H_0^{(2)}$ in favor of $H_a^{(2)}$ or

(b) we erroneously accepted $H_0^{(1)}$ even though $S(\cdot)$ concentrates on an interval containing $[a,b]$ as a sub-interval (see for example the data example in Section 5.3.2,

then we want to test for strong vs weak dependence, so we move to Step 3:

Step 3: $H_0^{(3)}: \text{supp}S(\cdot) = [a,b]$ vs $H_a^{(3)}: \text{supp}S(\cdot) = [0,1]$.

Here supp$S(\cdot)$ means the support of the measure $S(\cdot)$ which we assume is connected. This third step relies on a variance comparison.

The asymptotically normal statistics D_n and T_n are specified in Sections 3 and 4. More detail on methodology is in Section 5.1.

2. Background: Multivariate, Hidden and Second Order Regular Variation

Here is a review of notation and concepts necessary for formulating and proving results. We particularize the concept of regular variation of measures on a complete, separable metric space $\mathbb{X}$ for the case of $\mathbb{X} = \mathbb{R}_+^2$

where visualization is most informative (Lindskog et al. (2014); Hult and Lindskog (2006); Das et al. (2013); Kulik and Soulier (2020); Basrak and Planinić (2019); Resnick (2024)).

Recall $\{\boldsymbol{Z}_i = (X_i, Y_i); 1 \le i \le n\}$ are iid random vectors of $\mathbb{R}_+^2$ sampled from a regularly varying distribution. Based on observing these vectors, we analyze asymptotic dependence.

2.1 Multivariate regular variation.

We begin with the concept of $\mathbb{M}$-convergence of measures which is extended to $\mathbb{M}$-convergence of signed measures in Section 2.3 on 2RV.

Definition 2.1. For a closed subcone $\mathbb{C}$ of $\mathbb{X}$, let $\mathbb{M}(\mathbb{X} \setminus \mathbb{C})$ be the set of Borel measures on $\mathbb{X} \setminus \mathbb{C}$ which are finite on sets bounded away from $\mathbb{C}$, and $\mathcal{C}(\mathbb{X} \setminus \mathbb{C})$ be the set of continuous, bounded, non-negative functions on $\mathbb{X} \setminus \mathbb{C}$ whose supports are bounded away from $\mathbb{C}$. Then for $\mu_n, \mu \in \mathbb{M}(\mathbb{X} \setminus \mathbb{C})$, we say $\mu_n \to \mu$ in $\mathbb{M}(\mathbb{X} \setminus \mathbb{C})$, if $\int f \mathrm{d}\mu_n \to \int f \mathrm{d}\mu$ for all $f \in \mathcal{C}(\mathbb{X} \setminus \mathbb{C})$.

Without loss of generality (Lindskog et al., 2014), we can take functions in $\mathcal{C}(\mathbb{X} \setminus \mathbb{C})$ to be uniformly continuous. The modulus of continuity of a uniformly continuous function $f : \mathbb{R}_+^p \mapsto \mathbb{R}_+$ is

$$\Delta_f(\delta) = \sup\{|f(\boldsymbol{x}) - f(\boldsymbol{y})| : d(\boldsymbol{x}, \boldsymbol{y}) < \delta\} \tag{2.1}$$

where $d(\cdot, \cdot)$ is an appropriate metric on the domain of f. Uniform continuity means $\lim_{\delta \to 0} \Delta_f(\delta) = 0$.

Denote by RV_c, the class of regularly varying functions $f : \mathbb{R}_+ \mapsto \mathbb{R}_+$ with index $c \in \mathbb{R}$ and write $f \in RV_c$. The formal definition of multivariate regular variation (MRV) of distributions for the classical case $\mathbb{X} = \mathbb{R}_+^2$ and $\mathbb{C} = \{\mathbf{0}\}$ is next.

Definition 2.2. The distribution $\mathbb{P}[\boldsymbol{Z}_1 \in \cdot\,]$ of a random vector $\boldsymbol{Z}_1 = (X_1, Y_1)$ on $\mathbb{R}_+^2$, is (standard) regularly varying on $\mathbb{R}_+^2 \setminus \{\mathbf{0}\}$ with index $\alpha > 0$ if there exists some regularly varying scaling function $b(t) \in \mathrm{RV}_{1/\alpha}$ and a not identically zero limit measure $\eta(\cdot) \in \mathbb{M}(\mathbb{R}_+^2 \setminus \{\mathbf{0}\})$ such that as $t \to \infty$,

$$t\mathbb{P}[\boldsymbol{Z}_1/b(t) \in \cdot\,] \to \eta(\cdot), \qquad \text{in } \mathbb{M}(\mathbb{R}_+^2 \setminus \{\mathbf{0}\}). \tag{2.2}$$

It is sometimes convenient to write $\mathbb{P}[\boldsymbol{Z}_1 \in \cdot\,] \in \mathrm{MRV}(\alpha, b(t), \eta, \mathbb{R}_+^2 \setminus \{\mathbf{0}\})$.

2.1.1 Cartesian to polar and back.

When analyzing the asymptotic dependence between components of a bivariate random vector $\boldsymbol{Z}$ satisfying (2.2), it is often informative to make a polar coordinate transform and consider the transformed points located on the L_1-unit sphere

$$(x, y) \mapsto \left(\frac{x}{x+y}, \frac{y}{x+y}\right), \tag{2.3}$$

after thresholding the data according to the L_1 norm. The plot of such data is the (positive-quadrant) diamond plot. (Figure 1 is the 4-quadrant version using data whose components could be positive or negative under the map $(x, y) \mapsto (x, y)/(|x| + |y|)$.) In $\mathbb{R}_+^2$, the convenient version of the L_1-polar coordinate transformation is $T : \mathbb{R}_+^2 \setminus \{\mathbf{0}\} \mapsto (\mathbb{R}_+ \setminus \{0\}) \times [0, 1]$, defined by

$$T(x, y) = \big(x + y, x/(x + y)\big) = (r, \theta).$$

The inverse transformation from polar to Cartesian coordinates is $(r, \theta) \mapsto (r\theta, r(1-\theta))$. The map T disintegrates $\eta(\cdot)$ in (2.2) into the product measure

$$\eta \circ T^{-1}(\cdot) = \nu_\alpha \times S(\cdot),$$

where $S(\cdot)$ can be taken to be a probability measure on $[0, 1]$ called the angular measure, and $\nu_\alpha(\cdot)$ is the Pareto measure with $\nu_\alpha(x, \infty) = x^{-\alpha}$, $x > 0$.

2.2 Hidden regular variation.

Denote by $\mathbb{C}_{a,b}$ the subcone of $\mathbb{R}_+^2$ such that (see Figure 2 for a visual aid)

$$\mathbb{C}_{a,b} = \{(x, y) \in \mathbb{R}_+^2 : \theta := x/(x + y) \in [a, b]\}, \quad 0 \le a \le b \le 1.$$

When the limit measure of regular variation $\eta(\cdot)$ concentrates on a proper subcone $\mathbb{C}_{a,b} \subset \mathbb{X} = \mathbb{R}_+^2$ of the full state space, we may improve estimates of probabilities in the complement of the subcone, if there is a second *hidden* regular variation regime after removing the subcone.

Definition 2.3. The random vector $\boldsymbol{Z}_1$ in $\mathbb{R}_+^2$ has a distribution that is regularly varying on $\mathbb{R}_+^2 \setminus \{\mathbf{0}\}$ and has hidden regular variation on $\mathbb{R}_+^2 \setminus \mathbb{C}_{a,b}$

if there exist $0 < \alpha \leq \alpha_0$, scaling functions $b(t) \in RV_{1/\alpha}$ and $b_0(t) \in RV_{1/\alpha_0}$ with $b(t)/b_0(t) \to \infty$ and limit measures η concentrating on $\mathbb{C}_{a,b}$ and another limit measure η_0, such that

$$\mathbb{P}(\boldsymbol{Z}_1 \in \cdot) \in \mathrm{MRV}(\alpha, b(t), \eta, \mathbb{R}_+^2 \setminus \{\mathbf{0}\}) \cap \mathrm{MRV}(\alpha_0, b_0(t), \eta_0, \mathbb{R}_+^2 \setminus \mathbb{C}_{a,b}). \tag{2.4}$$

It is sometimes useful to characterize HRV through the *generalized polar coordinate transformation* for $\mathbb{R}_+^2 \setminus \mathbb{C}_{a,b}$, assuming use of an associated metric $d(\cdot,\cdot)$ satisfying $d(cx, cy) = cd(x, y)$ for scalars $c > 0$. The metric $d(\cdot,\cdot)$ that we use in practice is a scaled L_1-metric. When using generalized polar coordinates with respect to the forbidden zone $\mathbb{C}_{a,b}$, we define $\aleph_{\mathbb{C}_{a,b}} := \{\boldsymbol{z} \in \mathbb{R}_+^2 \setminus \mathbb{C}_{a,b} : d(\boldsymbol{z}, \mathbb{C}_{a,b}) = 1\}$, the locus of points at distance 1 from $\mathbb{C}_{a,b}$. Then the generalized polar coordinates are specified through the transformation, $\mathrm{GPOLAR} : \mathbb{R}_+^2 \setminus \mathbb{C}_{a,b} \mapsto (0, \infty) \times \aleph_{\mathbb{C}_{a,b}}$ with

$$\mathrm{GPOLAR}(\boldsymbol{z}) = \left(d(\boldsymbol{z}, \mathbb{C}_{a,b}), \frac{\boldsymbol{z}}{d(\boldsymbol{z}, \mathbb{C}_{a,b})} \right). \tag{2.5}$$

In Figure 2, the length of the arrow from $\boldsymbol{z}$ to the boundary of $\mathbb{C}_{a,b}$ is a scaled version of $d(\boldsymbol{z}, \mathbb{C}_{a,b})$. For a probability measure $S_0(\cdot)$ on $\aleph_{\mathbb{C}_{a,b}}$, the generalized polar coordinates allow re-writing the second MRV in (2.4) as

$$t\mathbb{P}\left[\left(\frac{d(\boldsymbol{Z}_1, \mathbb{C}_{a,b})}{b_0(t)}, \frac{\boldsymbol{Z}_1}{d(\boldsymbol{Z}_1, \mathbb{C}_{a,b})} \right) \in \cdot \right] \to (\nu_{\alpha_0} \times S_0)(\cdot)$$

in $\mathbb{M}\big((\mathbb{R}_+ \setminus \{0\}) \times \aleph_{\mathbb{C}_{a,b}}\big)$. In particular $\mathbb{P}[d(\boldsymbol{Z}_1, \mathbb{C}_{a,b}) > x] \in RV_{-\alpha_0}$ is a lighter tail than $\mathbb{P}[\|\boldsymbol{Z}_1\| > x] \in RV_{-\alpha}$. See Das et al. (2013) and Lindskog et al. (2014) for details.

2.3 Second order regular variation.

In one dimension, second order regular variation (2RV) controls deviation of finite sample means from asymptotic means and allows asymptotic normality for estimators such as the Hill estimator. Our test statistics are derived from two dimensional tail empirical measures and so it is natural to expect that asymptotic normality requires two dimensional second order regular variation conditions and the condition we use is similar to what appears in Einmahl et al. (2021).

2.3.1 The second order condition.

There are several ways to state this condition which strengthens multivariate regular variation. The first uses $\mathbb{M}$-convergence. We need a function $A \in RV_{-\rho}$, $\rho > 0$, and a signed measure $\chi(\cdot)$ which is not identically 0 and is the difference of two measures in $\mathbb{M}\big((\mathbb{R}_+ \setminus \{0\}) \times [0,1]\big)$, such that in $\mathbb{M}\big((\mathbb{R}_+ \setminus \{0\}) \times [0,1]\big)$,

$$\frac{1}{A(t)}\left(t\mathbb{P}\Big((R/b(t),\Theta)\in\cdot\Big) - \nu_\alpha \times S(\cdot)\right) \to \chi(\cdot), \tag{2.6}$$

meaning that evaluation of the signed measure on the left at a function $f \in \mathcal{C}\big(\big((\mathbb{R}_+ \setminus \{0\}) \times [0,1]\big)$ converges to the evaluation $\chi(f)$; or in symbols

$$\frac{1}{A(t)}\left(t\mathbb{E}f\big(R/b(t),\Theta\big) - \iint_{\mathbb{R}_+\setminus\{0\}\times[0,1]} f(r,\theta)\nu_\alpha(dr)S(d\theta)\right) \to \chi(f). \tag{2.7}$$

The second way to phrase condition (2.6) which looks more like convergence of distribution functions is

$$\frac{1}{A(t)}\Big(t\mathbb{P}\Big(\frac{R}{b(t)} > r, \Theta \le \theta\Big) - r^{-\alpha}S[0,\theta]\Big) \to \chi\big((r,\infty)\times[0,\theta]\big) \tag{2.8}$$

locally uniform in $r \in (0,\infty)$ for each $\theta \in [0,1]$ where the limit is specified before (2.6).

If $f_1(r) \in \mathbb{M}(\mathbb{R}_+ \setminus \{0\})$, set $f(r,\theta) := f_1(r)\theta \in \mathbb{M}\big((\mathbb{R}_+ \setminus \{0\}) \times [0,1]\big)$ and inserting this into (2.7) gives

$$\frac{1}{A(t)}\Big(t\mathbb{E}\Theta f_1\big(R/b(t)\big) - \int_{[0,1]} \theta S(d\theta)\nu_\alpha(f_1)\Big) \to \int_{(0,\infty)\times[0,1]} \theta f_1(r)\chi(dr,d\theta). \tag{2.9}$$

or in convergence of signed measures formulation,

$$\frac{1}{A(t)}\Big(t\mathbb{E}\Theta\epsilon_{R/b(t)}(\cdot) - \int_{[0,1]} \theta S(d\theta)\nu_\alpha(\cdot)\Big) \to \iint_{(\cdot)\times[0,1]} \theta\chi(dr,d\theta). \tag{2.10}$$

Note that (2.6) and (2.8) are formulated so they can be marginalized and therefore the regularly varying distribution of R is 2RV in one dimension. Also, straightforward extension of (2.9) allows control of the expectation of Θ on the set where R is large. If we set

$$v(t) = E\Theta_1 \mathbf{1}_{[R_1 > t]}, \quad \mu_S = \int_{[0,1]} \theta S(d\theta),$$

then (2.10) gives as $t \to \infty$,

$$\frac{tv(b(t)x) - \mu_S x^{-\alpha}}{A(t)} \to h(x) := \iint_{((x,\infty))\times[0,1]} \theta\chi(dr,d\theta). \tag{2.11}$$

which leads to the more traditional form of the 2RV condition for $v(t)$, namely: for $b^{\leftarrow}(\cdot)$ denoting the generalized inverse of $b(\cdot)$, we have

$$\lim_{s\to\infty} \frac{\frac{v(sx)}{v(s)} - x^{-\alpha}}{A \circ b^{\leftarrow}(s)} = h(x)/\mu_S, \tag{2.12}$$

where $A \circ b^{\leftarrow} \in RV_{-\rho\alpha}$ and the limit function $h(x)$ must be of the form (de Haan and Stadtmueller (1996); Peng (1998); de Haan and Ferreira (2006)),

$$h(x) = cx^{-\alpha}\Big(\frac{1 - x^{-\rho\alpha}}{\rho\alpha}\Big), \quad x > 0,\, c \neq 0.$$

2.3.2 2RV and HRV

We discuss why the second order condition (2.6) together with the assumption $S([a,b]) = 1$ for $[a,b] \subsetneq [0,1]$ implies HRV. The essentials of the argument in the context of asymptotic independence are in (de Haan and de Ronde, 1998; Resnick, 2002).

Theorem 2.1 (2RV can imply HRV)**.** Assume the 2RV condition (2.6) or (2.8) hold and $S([a,b]) = 1$ for $[a,b] \subsetneq [0,1]$. Set $U(t) = t/A(t) \in RV_{1+\rho}$, so that $U^{\leftarrow}(t) \in RV_{1/(1+\rho)}$ and therefore

$$b_0(t) := b \circ U^{\leftarrow}(t) \in RV_{1/(\alpha(1+\rho))},\, \rho > 0. \tag{2.13}$$

Then provided $\chi(\cdot)$ is not identically 0 on $(0,\infty) \times ([0,1] \setminus [a,b])$,

$$\mathbb{P}\left[(R,\Theta) \in \cdot\right] \in \mathrm{MRV}(\alpha(1+\rho), b_0(t), \chi(\cdot), (\mathbb{R}_+ \setminus \{0\}) \times ([0,1] \setminus [a,b])). \tag{2.14}$$

The proof of Theorem 2.1 is given in Section S2 of the supplement.

3. Testing the existence of strong dependence

For strong convergence, we assume that $0 \le a \le b \le 1$ fixed, with $[a,b] \subsetneq [0,1]$ and $S([a,b]) = 1$. The condition $\theta = x/(x+y) \in [a,b]$ translates to

$$(x,y) \in \{(u,v) \in \mathbb{R}_+^2 : v/u \in [b^{-1}-1, a^{-1}-1]\} \equiv \mathbb{C}_{a,b}.$$

So the closed cone $\mathbb{C}_{a,b}$ is the set of first quadrant points between the two rays $y = m_u x$ and $y = m_l x$, $x > 0$, where the slopes are $m_u = a^{-1} - 1$, $m_l = b^{-1} - 1$ and since $a \le b$, we have $m_u \ge m_l$. Define the scaled distance from $\boldsymbol{z} = (x,y) \in \mathbb{R}_+^2$ to $\mathbb{C}_{a,b}$ (see Figure 2) as

$$d((x,y), \mathbb{C}_{a,b}) := \max\left\{(b^{-1}-1)x - y, y - (a^{-1}-1)x, 0\right\}. \tag{3.15}$$

Note that when (x,y) is above cone $\mathbb{C}_{a,b}$ so that $y/x > m_u$ and thus $y > (a^{-1}-1)x$, $d((x,y),\mathbb{C}_{a,b}) = y-(a^{-1}-1)x$. When (x,y) is below the cone $\mathbb{C}_{a,b}$ so that $y/x < m_l$ and $y < (b^{-1}-1)x$, $d((x,y),\mathbb{C}_{a,b}) = (b^{-1}-1)x - y$. When $(x,y) \in \mathbb{C}_{a,b}$, $d((x,y),\mathbb{C}_{a,b}) = 0$. When $\mathbb{C}_{a,b} = \{(x,y) : y = (\theta_0^{-1} - 1)x, x > 0\}$ because $S\{\theta_0\} = 1$, then $d((x,y),\mathbb{C}_{a,b}) = |(\theta_0^{-1}-1)x - y|$.

Using generalized polar coordinates, the HRV assumption for the distribution of $\boldsymbol{Z}_1 = (X_1, Y_1)$ on $\mathbb{R}_+^2 \setminus \mathbb{C}_{a,b}$ reads

$$t\mathbb{P}\left(\left(\frac{d\big(\boldsymbol{Z}_1, \mathbb{C}_{a,b}\big)}{b_0(t)}, \frac{\boldsymbol{Z}_1}{d(\boldsymbol{Z}_1, \mathbb{C}_{a,b})}\right) \in \cdot\right) \longrightarrow \nu_{\alpha_0} \times S_0(\cdot)$$

in $\mathbb{M}((\mathbb{R}_+ \setminus \{0\}) \times \aleph_{\mathbb{C}_{a,b}})$ and in particular $\mathbb{P}[d\big(\boldsymbol{Z}_1, \mathbb{C}_{a,b}\big) > x] \in RV_{-\alpha_0}$ and assuming 2RV from the previous section, $\alpha_0 = \alpha(1+\rho)$.

Let $\{\boldsymbol{Z}_j = (X_j, Y_j) : j \geq 1\}$ be iid, set $R_j := X_j + Y_j$, and define $\boldsymbol{Z}_i^* = (X_i^*, Y_i^*)$ to be the vector such that $X_i^* + Y_i^*$ is the i-th largest order statistic of $\{R_j : 1 \leq j \leq n\}$, which we denote $R_{(i)}$. Consider the following hypotheses: for fixed and known (for now) $0 < a \leq b < 1$, $[a, b] \subsetneq [0, 1]$,

$$H_0^{(1)} : S([a,b]) = 1, \qquad H_a^{(1)} : S([a,b]) < 1. \tag{3.16}$$

We now propose a test statistic for (3.16). Define

$$\begin{aligned} D_n :=& \frac{1}{k(n)} \sum_{i=1}^{k(n)} \left(1 + \frac{d\big(\boldsymbol{Z}_i^*, \mathbb{C}_{a,b}\big)}{R_{(k(n))}} \right) \log \frac{R_{(i)}}{R_{(k(n))}} \\ =& H_{k(n),n} + \frac{1}{k(n)} \sum_{i=1}^{k_n} \left(\frac{d\big(\boldsymbol{Z}_i^*, \mathbb{C}_{a,b}\big)}{R_{(k(n))}} \right) \log \frac{R_{(i)}}{R_{(k(n))}} \end{aligned} \tag{3.17}$$

where $H_{k(n),n}$ is the Hill estimator of $1/\alpha$ applied to $\{R_j, 1 \leq j \leq n\}$ based on $k(n)$ upper order statistics. Of course, D_n depends on a, b but this dependence is suppressed in the notation. The proposal of D_n is motivated by the reasoning that if $S([a,b]) = 1$, then the second term of (3.17) should be small so that D_n must have similar asymptotic behavior as $H_{k(n),n}$.

A consistent estimator of a, b is suggested in Section 5.1.1 but for now we assume a, b are fixed and known. Theorem 3.1 gives the asymptotic normality of D_n.

Theorem 3.1. Assume the 2RV condition (2.8) holds, $\alpha_0 \equiv \alpha(1+\rho) > 1$, $b_0(t)$ is defined in (2.13) and $\{k(n)\}$ is an intermediate sequence (i.e. $k(n) \to \infty$, $n/k(n) \to \infty$, $n \to \infty$) satisfying

$$\sqrt{k(n)} \frac{b_0(n/k(n))}{b(n/k(n))} \to 0, \qquad n \to \infty. \tag{3.18}$$

Under $H_0^{(1)}$ as given in (3.16), we have

$$\sqrt{k(n)}(D_n - 1/\alpha) \Rightarrow \frac{1}{\alpha}N(0,1). \tag{3.19}$$

Note that the condition $\alpha_0 = \alpha(1+\rho) > 1$ is mild as it is rare in practice for tails to be so heavy that $\alpha < 1$. The proof of Theorem 3.1 is based on asymptotic normality of the tail empirical measure, and is in Section S3 of the supplement. For treatments explaining the need for the second order condition, see (Resnick, 2007, Section 9.1) or de Haan and Ferreira (2006). Here we give some remarks.

First, under $H_0^{(1)}$, for $\boldsymbol{Z}_i$ corresponding to large R_i, the distance from $\boldsymbol{Z}_i$ to $\mathbb{C}_{a,b}$ should be small with high probability and therefore D_n should be close to the Hill estimator which is asymptotically normal. The proof of Theorem 3.1 shows that when $S[a,b] = 1$,

$$\sqrt{k(n)}(D_n - H_{k(n),n}) = \frac{1}{\sqrt{k(n)}}\sum_{i=1}^{k_n}\left(\frac{d\big(\boldsymbol{Z}_i^*, \mathbb{C}_{a,b}\big)}{R_{(k(n))}}\right)\log\frac{R_{(i)}}{R_{(k(n))}} \Rightarrow 0, \tag{3.20}$$

as $n \to \infty$. Equation (3.20) thus confirms that under strong dependence, distance to $\mathbb{C}_{a,b}$ must be negligible. However, (3.20) also means that in (3.19) we cannot replace $1/\alpha$ by the plug-in estimate, necessitating the use of bootstrap methods in Section 5.1.2.

In addition, Theorem 3.1 suggests that for fixed a, b, we reject $H_0^{(1)}$ in (3.16) if $|D_n - 1/\alpha| > 1.96/(\alpha/\sqrt{k(n)})$. If we choose too wide an interval $[a,b] \subsetneq [0,1]$, then the test statistic D_n becomes closer to $H_{k,n}$ as more data points are included in $\mathbb{C}_{a,b}$. Failure to reject for the fixed interval means also that one fails to reject for any bigger interval. So using only D_n, we cannot distinguish whether the support of $S(\cdot)$ is in $[a, b]$ or a subset of $[a, b]$ and, in particular, if we fail to reject $H_0^{(1)}$, it could be the support is $\{\theta_0\}$ for some $\theta_0 \in [a, b]$. Therefore, in the next section, we give another test statistic that helps decide whether $\mathbb{P}[\boldsymbol{Z}_1 \in \cdot\,]$ is asymptotically fully or strongly dependent.

4. Full vs strong dependence

Now consider the hypothesis test as formed in Step 2:

$$H_0^{(2)} : S(\{\theta_0\}) = 1 \qquad H_a^{(2)} : S([0, 1] \setminus \{\theta_0\}) > 0. \tag{4.21}$$

where $\theta_0 \in [a, b]$, and to capitalize on hidden regular variation resulting from 2RV, we need the assumption that $[a, b] \subsetneq [0, 1]$ is a proper subset of $[0, 1]$. Since $\theta_0 \in [a, b]$ and D_n given in Theorem 3.1 is unable to distinguish between the two hypotheses in (4.21), we propose another test statistic. Let $\Theta_i^* := X_i^*/(X_i^* + Y_i^*)$ be the concomitant of $R_{(i)}$, and define

$$T_n := \frac{\sum_{i=1}^{k(n)} \Theta_i^* \log \frac{R_{(i)}}{R_{(k(n))}}}{\sum_{i=1}^{k(n)} \Theta_i^*}. \tag{4.22}$$

The next two results recommend we distinguish between strong and full dependence by assessing the asymptotic variance of T_n. Under $H_0^{(2)}$ the asymptotic variance of T_n is $1/\alpha^2$ but under $H_a^{(2)}$ the asymptotic variance is strictly greater than $1/\alpha^2$.

4.1 Full dependence

The next Theorem 4.1 is posed under the assumption $H_0^{(2)}$ in (4.21) of full dependence where the limit angular measure concentrates at a point $\theta_0 \in (0,1)$.

Theorem 4.1. Assume $H_0^{(2)}$ holds and the angular measure $S(\cdot) = \epsilon_{\theta_0}(\cdot)$, for $\theta_0 \in (0,1)$. Suppose the 2RV condition in (2.6) holds with $A(t) \in RV_{-\rho}$, $\rho > 0$. Define $b_0(t)$ as in (2.13) so $b_0(t) \in RV_{1/(\alpha(1+\rho))}$ and $\alpha_0 = \alpha(1+\rho)$. Let $\{k(n)\}$ be an intermediate sequence satisfying (3.18), then we have

$$\sqrt{k(n)}\left(T_n - \frac{1}{\alpha}\right) \Rightarrow N(0, 1/\alpha^2). \tag{4.23}$$

The proof is reserved for Section S4 of the supplementary material but is somewhat truncated since the proof of Theorem 4.2 is similar. Additionally, the proof of Theorem 4.1 indicates that when $H_0^{(2)}$ holds, we have in $\mathbb{R}$,

$$\sqrt{k(n)}(T_n - H_{k(n),n}) \Rightarrow 0, \tag{4.24}$$

precluding the use of the plug-in estimator for $1/\alpha$. Tests based on T_n will require bootstrap methods.

4.2 Strong dependence

Under strong dependence, the asymptotic variance of T_n is strictly larger than $1/\alpha^2$.

Theorem 4.2. Assume strong dependence exists such that supp$S(\cdot) = [a, b]$. Suppose also the 2RV condition in (2.6) holds with a limiting signed measure $\chi(\cdot)$ and $A(t) \in RV_{-\rho}$, $\rho > 0$. Define $b_0(t)$ as in (2.13), so $b_0(t) \in RV_{1/(\alpha(1+\rho))}$ and $\alpha_0 = \alpha(1+\rho)$. As before, $\{k(n)\}$ is an intermediate sequence satisfying (3.18). Define

$$\mu := \int_a^b x S(\mathrm{d}x), \qquad \sigma^2 := \int_a^b (x-\mu)^2 S(\mathrm{d}x),$$

and under strong dependence assumption $H_a^{(2)}$, we have

$$\sqrt{k(n)}\left(T_n - \frac{1}{\alpha}\right) \Rightarrow N\left(0, \frac{1}{\alpha^2}(1+\sigma^2/\mu^2)\right). \tag{4.25}$$

Proofs use a functional central limit theorem for row sums of a triangular array of $D[0,1]$-functions (Pollard, 1990, Theorem 10.6) that generalizes the sequential result of Hahn (1978) and are discussed in Section S5 of the supplement. We also give details for the proof of Theorem 4.2 in the supplementary material since it showcases the key steps to show Theorem 4.1.

5. Implementation of testing

Applying the test statistics to data requires estimating a minimal length interval $[a,b]$ containing the support of the angular measure. On the one hand, choosing an unnecessarily wide interval $[a,b]$ leads D_n to conclude $S([a,b]) = 1$ but only shows the support is a subset of $[a,b]$. Also making $[a,b]$ too wide may mean there are few points in $[0,1] \setminus [a,b]$, so that even if the true support of S is $[0,1]$, we could falsely accept the existence of strong dependence. On the other hand, fixing an excessively narrow interval $[a,b]$ may lead to D_n inaccurately rejecting existence of strong dependence.

We begin with a method for estimating a, b and then proceed to bootstrap methods for implementing the tests. This is followed in Sections 5.2 and 5.3 by illustrations using simulated and real data.

5.1 Methodology

5.1.1 Estimating $[a, b]$

We estimate a, b as the minimizer of an objective function $g_n(a, b)$ subject to the constraint $0 \le a \le b \le 1$ where

$$g_n(a,b) := (b-a) + \sqrt{k(n)}\left|D_n - H_{k(n),n}\right| \tag{5.26}$$

The first part of the objective function, $b - a$, favors a narrow interval $[a, b]$ the second part requires a wide enough interval $[a, b]$ so that $\left|D_n - H_{k(n),n}\right| \approx 0$. Hence, by minimizing g_n, we obtain an estimated interval $[\hat{a}, \hat{b}]$ of reasonable length and satisfying $\left|D_n - H_{k(n),n}\right| \approx 0$. In practice, the `constrOptim` function in R suffices for the minimization.

Theorem 5.1 gives the consistency of $\hat{a}$ and $\hat{b}$ for $\alpha > 1$.

Theorem 5.1. Suppose the support of S is $[a, b]$, $\alpha > 1$ and the intermediate sequence $\{k(n)\}$ satisfies (3.18). Let $\hat{a}$ and $\hat{b}$ be the minimizer of (5.26). Then as $n \to \infty$,

$$\hat{a} \xrightarrow{p} a, \qquad \hat{b} \xrightarrow{p} b.$$

The proof of Theorem 5.1 is provided in Section S6 of the supplement. In fact, the consistency result in Theorem 5.1 also holds if we redefine for some $\lambda > 0$,

$$g_n(s,t) = (t-s) + \lambda\sqrt{k(n)}\left|D_n - H_{k(n),n}\right|. \quad (5.27)$$

Note that when λ is large, the penalty term in (5.27) suggests the optimization problem favor a wide estimated interval.

5.1.2 Bootstrap methods

Formulating tests based on either Theorem 3.1 or 4.1 requires knowing the values of α, a, b, which, however, is unlikely to be true for real datasets. Substitution methods suggest replacing α with the corresponding Hill estimator, $1/H_{k(n),n}$ and investigating the effect on the limit distribution but this will not work here due to (3.20) and (4.24). Thus, we propose bootstrap methods to implement the proposed tests and try the approach on simulated and real datasets. We will report elsewhere on justifications for the bootstrap methods and here we show numerical experiments that suggest its applicability.

Assume an intermediate sequence $\{k(n)\}$ is chosen. According to Feigin and Resnick (1997) and Chapter 6.4 of Resnick (2007), bootstrapping of heavy-tailed phenomena requires taking a bootstrap sample size $m = m(n) \approx n/k(n)$ so that $m(n)/n \to 0$ and $m(n) \to \infty$. Let $\{I_1(n), \ldots, I_m(n)\}$ be iid discrete uniform random variables on $\{1, \ldots, n\}$, independent from $\{(X_i, Y_i) : i \geq 1\}$. We construct a bootstrap resample of size m by

$$\boldsymbol{Z}_{I_j(n)} = (X_{I_j(n)}, Y_{I_j(n)}), \qquad j = 1, \ldots, m.$$

Define $R^{\text{boot}}_{(i)}$ as the i-th largest order statistic among $\{R_{I_j(n)} \equiv X_{I_j(n)} + Y_{I_j(n)} : 1 \le j \le m\}$, and let $\boldsymbol{Z}^*_{I_i(n)} = (X^*_{I_i(n)}, Y^*_{I_i(n)})$ be the pair of random variables such that $X^*_{I_i(n)} + Y^*_{I_i(n)} \equiv R^{\text{boot}}_{(i)}$.

(1) Test $H_0^{(1)}$. For the test in (3.16), we first solve (5.27) with a proper choice of λ using the whole sample of size n to estimate the support of the angular measure, $[\hat{a}, \hat{b}]$, from the sample. Then we obtain $\mathbb{C}_{\hat{a},\hat{b}} := \{(x, y) \in \mathbb{R}^2_+ : \hat{a} \le x/(x + y) \le \hat{b}\}$ and

$$\widehat{D}_m = \frac{1}{k(m)} \sum_{i=1}^{k(m)} \left(1 + \frac{d\big(\boldsymbol{Z}^*_{I_i(n)}, \mathbb{C}_{\hat{a},\hat{b}}\big)}{R^{\text{boot}}_{(k(m))}}\right) \log \frac{R^{\text{boot}}_{(i)}}{R^{\text{boot}}_{(k(m))}}.$$

Conditioning on the original sample, we presume from Theorem 3.1 that for large n, $\sqrt{k(m)}\left(\widehat{D}_m - H_{k(n),n}\right) \approx N(0, H^2_{k(n),n})$. Therefore, we use the z-test by computing the Hill estimator $H_{k(n),n}$ from the full sample and reject $H_0^{(1)}$ in (3.16) if

$$\left|\widehat{D}_m - H_{k(n),n}\right| > 1.96 \frac{H_{k(n),n}}{\sqrt{k(m)}}. \tag{5.28}$$

In practice we would generate B bootstrap samples and reject if more than 5% satisfy (5.28).

(2) Full vs strong dependence. For the test in (4.21), generate B bootstrap resamples indexed by $t = 1, \dots, B$. For each t, let $R^{\text{boot}}_{(i),t}$ be the i-largest order statistic in the t-th resample; $\Theta^*_{i,t}$ is the corresponding concomitant. Compute the corresponding test statistics for each resample,

$$T_m^{(t)} = \frac{\sum_{i=1}^{k(m)} \Theta^*_{i,t} \log \frac{R^{\text{boot}}_{(i),t}}{R^{\text{boot}}_{(k(m)),t}}}{\sum_{i=1}^{k(m)} \Theta^*_{i,t}}, \qquad t = 1, \dots, B.$$

Based on Theorem 4.1, we presume under $H_0^{(2)}$ that conditional on the original sample, $\sqrt{k(m)}\left(T_m^{(t)} - H_{k(n),n}\right)$ is approximately normal with mean 0 and variance $H^2_{k(n),n}$ for large n. Using all B resamples, we obtain the bootstrap estimate of the standard error of T_n:

$$\mathrm{SE}_{\mathrm{boot}}(m) := \left(\frac{1}{B-1}\sum_{t=1}^{B}\left(T_m^{(t)} - \bar{T}_m\right)^2\right)^{1/2},$$

where $\bar{T}_m = \frac{1}{B}\sum_{t=1}^{B} T_m^{(t)}$. Then due to the presumed asymptotic normality of $T_m^{(t)}$, we use the chi-square test for normal variance and reject $H_0^{(2)}$ in (4.21) if

$$k(m)\frac{\mathrm{SE}^2_{\mathrm{boot}}(m)}{H^2_{k(n),n}} > \chi^2_{0.95,B-1}/(B-1),$$

where $\chi^2_{0.95,B-1}$ denotes the 95% quantile of a chi-square distribution with $B-1$ degrees of freedom.

(3) Strong vs weak dependence. When testing for strong vs weak dependence, we rely on Theorem 4.2 and define $\widetilde{\Theta}_j := \Theta_j \mathbf{1}_{\{\Theta_j\in[a,b]\}}$, $\widetilde{R}_j := R_j \mathbf{1}_{\{\Theta_j\in[a,b]\}}$. Let $\widetilde{R}_{(i)}$ be the i-th largest order statistic of $\{\widetilde{R}_j : 1 \le j \le n\}$, and $\widetilde{\Theta}_i^*$ be the concomitant of $\widetilde{R}_{(i)}$. By assuming $0/0 \equiv 1$ we define also

$$\widetilde{T}_n := \frac{\sum_{i=1}^{k(n)} \widetilde{\Theta}_i^* \log\left(\frac{\widetilde{R}_{(i)}}{\widetilde{R}_{(k(n))}} \vee 1\right)}{\sum_{i=1}^{k(n)} \widetilde{\Theta}_i^*}. \tag{5.29}$$

For $[a,b] \subsetneq [0,1]$, we want to test strong vs weak dependence, i.e.

$$H_0^{(3)} : \mathrm{supp} S(\cdot) = [a,b] \qquad \text{v.s.} \qquad H_a^{(3)} : \mathrm{supp} S(\cdot) = [0,1]. \tag{5.30}$$

Under $H_0^{(3)}$, $\widetilde{T}_n$ must have the same asymptotic distribution as T_n. Here we apply the bootstrap method again to test whether T_n and $\widetilde{T}_n$ have the same asymptotic variance. Again estimate $[\hat{a}, \hat{b}]$ from (5.27). To obtain the t-th resample, we generate $\{I_{1,t}(n), \ldots, I_{m,t}(n)\}$ iid discrete uniform random variables on $\{1, \ldots, n\}$, and compute

$$\widetilde{\Theta}_{i,t} := \Theta_{I_{i,t}(n)} \mathbf{1}_{\{\Theta_{I_{i,t}(n)} \in [\hat{a}, \hat{b}]\}}, \qquad \widetilde{T}_m^{(t)} = \frac{\sum_{i=1}^{k(m)} \widetilde{\Theta}^*_{i,t} \log\left(\frac{\widetilde{R}_{(i),t}}{\widetilde{R}_{(k(m)),t}} \vee 1\right)}{\sum_{i=1}^{k(m)} \widetilde{\Theta}^*_{i,t}}.$$

We repeat the bootstrap resampling scheme twice to obtain $T_m^{(1)}, \ldots, T_m^{(B)}$, $\widetilde{T}_m^{(1)}, \ldots, \widetilde{T}_m^{(B)}$. Due to the presumed asymptotic normality of $T_m^{(t)}$, we use the F-test to compare variances, and reject $H_0^{(3)}$ if

$$\frac{\frac{1}{B-1} \sum_{t=1}^{B} \left(T_m^{(t)} - \bar{T}_m\right)^2}{\frac{1}{B-1} \sum_{s=1}^{B} \left(\widetilde{T}_m^{(s)} - \bar{\widetilde{T}}_m\right)^2} > F_{0.975, B-1, B-1} \quad \text{or} \quad < F_{0.025, B-1, B-1}, \tag{5.31}$$

where $\bar{\widetilde{T}}_m = \frac{1}{B} \sum_{t=1}^{B} \widetilde{T}_m^{(t)}$ and $F_{p,B-1,B-1}$ denotes the $100p\%$-percentile of an F-distribution with numerator and denominator degrees of freedom both equal to $B - 1$.

5.2 Simulation study

5.2.1 Example

Consider a simulated data example as below. Set $a = 0.25$, $b = 0.75$. Suppose $R_1 \sim$ Pareto(2), $R_2 \sim$ Pareto(6), $Z \sim$ Beta(0.1, 0.1), $\Theta_2 \sim$ Unif($[0, 1] \setminus [a, b]$), and $B \sim$ Bernoulli(0.5). Assume the random variables are all independent, and let $\Theta_1 := a + (b - a)Z^2$. Define

$$X := BR_1\Theta_1 + (1 - B)R_2\Theta_2$$

$$Y := BR_1(1 - \Theta_1) + (1 - B)R_2(1 - \Theta_2).$$

By construction, (X, Y) is MRV on $\mathbb{R}_+^2 \setminus \{\mathbf{0}\}$ with tail parameter $\alpha = 2$. The second order condition (2.6) also holds since

$$\frac{1}{t^{-1}} \left(t\mathbb{P}\left[\left(\frac{R}{b(t)}, \Theta \right) \in \cdot \right] - p\nu_2 \times \mathbb{P}[\Theta_1 \in \cdot] \right) \to (1-p)\nu_6 \times \mathbb{P}[\Theta_2 \in \cdot].$$

Furthermore, for $\mathbb{C}_{a,b} = \{(x, y) \in \mathbb{R}_+^2 : x/(x + y) \in [0.25, 0.75]\}$, the vector (X, Y) has HRV on $\mathbb{R}_+^2 \setminus \mathbb{C}_{a,b}$ with tail parameter $\alpha_0 = 6$.

We then generate $n = 2{,}500$ iid observations from the distribution of (X, Y). Applying the minimum distance method in Clauset et al. (2009) to $\{|x_j| + |y_j| : 1 \le j \le 2{,}500\}$ chooses $k(n) = 226 \approx \lceil 10.6n^{0.39} \rceil$ (which satisfies (3.18)), and the Hill plot (cf. left panel of Figure 3) shows a stable pattern around such a chosen $k(n)$ (Resnick, 2007, Chapter 4.4). Then thresholding with $k(n) = 226$ yields the histogram of angles in the middle panel of Figure 3. The histogram describes the dependence structure of (X, Y), and is consistent with the construction that $[a, b] = [0.25, 0.75]$.

We then estimate $[a, b]$ by solving (5.27) using the `constrOptim` function in `R` with $\lambda = 1$. To assess the consistency of the estimated $\hat{a}$ and $\hat{b}$, we generate 100 simulated samples, each of which consists of $n = 2{,}500$ iid observations, leading to small MSE values of 2.87×10^{-12} and 1.12×10^{-8}, respectively.

Next, set $m = \lceil 5n/k(n) \rceil$ and $k(m) = \lceil 2m^{0.39} \rceil$. For each of the 100 simulated samples of size 2,500, we obtain the corresponding $[\hat{a}, \hat{b}]$, and generate $B = 200$ bootstrap resamples. For each resample, compute $\widehat{D}_m$

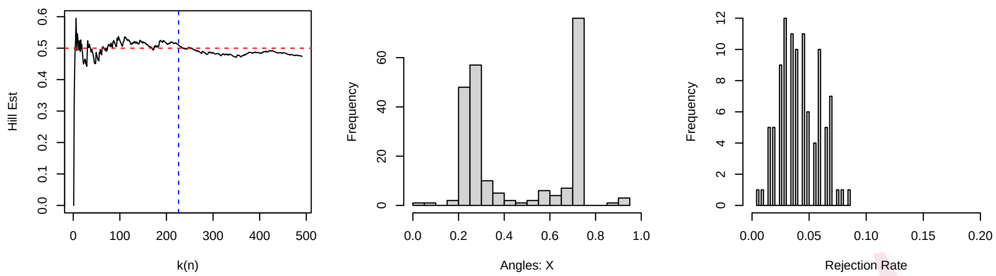

Figure 3: Left: Hill plot of $|x|+|y|$; the red line denotes $1/\alpha = 0.5$ and the blue line denotes $k(n) = 226$. Middle: Histogram of Θ_1 with $n = 2{,}500$ and $k(n) = 226$. Right: Histogram of rejection rates of $H_0^{(1)} : S([\hat{a}, \hat{b}]) = 1$ by $\widehat{D}_m$ among 100 simulated trials.

and use (5.28) to test

$$H_0^{(1)} : S([\hat{a}, \hat{b}]) = 1, \qquad H_a^{(1)} : S([\hat{a}, \hat{b}]) < 1.$$

For every simulated trials, we look at the rejection rate among the 200 bootstrap resamples, and plot the histogram in the right panel of Figure 3. Although 22 of the 100 simulated trials are rejecting $H_0^{(1)} : S([\hat{a}, \hat{b}]) = 1$, our numerical results also give an average rejection rate as 0.0427 with a standard deviation of 0.0195. When we further test $H_0^{(3)}$ using (5.31), only 8% of the trials reject the null, thus confirming the existence of strong dependence on $\mathbb{R}_+^2 \setminus \mathbb{C}_{\hat{a},\hat{b}}$. To check the existence of full dependence, we compute $\mathrm{SE}_{\mathrm{boot}}(m)$ based on the bootstrap resamples, and among the 100 trials, 35% of them reject $H_0^{(2)}$, showing a reasonable power of the proposed test using $T_m^{(t)}$.

5.2.2 Power analysis

Now we consider another simulated example of asymptotic weak dependence to examine the power of all three tests proposed in the paper. Define two independent random variables $B_0 \sim$ Bernoulli(0.1) and $Z_0 \sim$ Beta(1, 2), both of which are independent from all other random variables specified in the previous example. Then redefine $\Theta_1 := B_0(a + (b - a)Z) + (1 - B_0)Z_0$, and keep all other setup identical to the previous section.

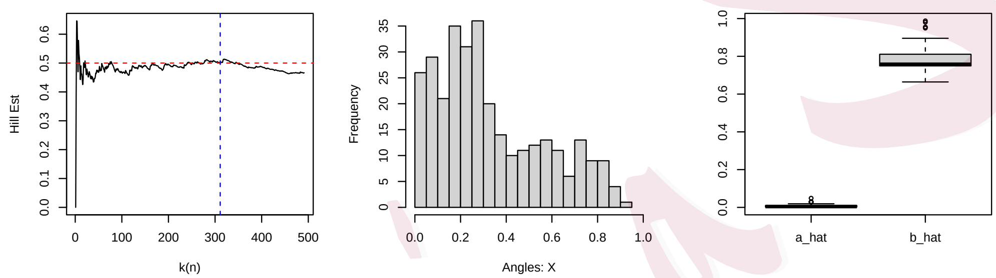

Figure 4: Left: Hill plot of $|x| + |y|$, and the red line denotes $1/\alpha = 0.5$. Middle: Histogram of Θ_1 with $n = 2{,}500$ and $k(n) = 311$. Right: Boxplot of $\hat{a}$ and $\hat{b}$ among 100 simulated trials.

We again generate 100 simulated samples with $n = 2{,}500$, and the left panel of Figure 4 gives the Hill plot from one specific sample. The red dashed line represents $1/\alpha = 0.5$, and the blue line marks $k(n) = 311$, chosen by the minimum distance method (Clauset et al., 2009). Given the stable shape of the Hill plot, we proceed by keeping $k(n) = 311$. Then the histogram of angles based on the thresholded data is presented in the middle panel of Figure 4, and the boxplot in the right panel of Figure 4 gives distributions of the estimated $\hat{a}$ and $\hat{b}$. Different from the previous case, we see that when asymptotic weak dependence exists, values of $\hat{b}$ vary a lot and are different from the true value of 1.

Next, for each sample, we obtain $B = 200$ resamples of size $m = \lceil 5n/k(n) \rceil$, and set $k(m) = \lceil 2m^{0.39} \rceil$. The rejection rates for $H_0^{(i)}$, $i = 1, 2, 3$, are 33%, 74%, and 23%, respectively. Similar to the previous case, the proposed test of asymptotic full dependence ($H_0^{(2)}$) shows a good power when we assume asymptotic weak dependence. When testing asymptotic strong dependence, the boxplot in Figure 4 suggests that the estimated interval $[\hat{a}, \hat{b}]$ can be so wide that one may fail to detect significant differences in the variances, thus leading to a lower power while testing $H_0^{(1)}$ and $H_0^{(3)}$.

5.3 Real data examples

We now consider the application of the bootstrap method to real data. We download the daily adjusted stock prices of Chevron (CVX), Exxon (XOM) and Apple (AAPL) during the time period from January 04, 2016 to December 30, 2022. To lessen the possible serial dependence of stock returns, we compute the log returns of these three stocks using their every-other-day prices. The acf plots in Figure 5 show little serial dependence for all three stocks. This leads to a reduced dataset of $n = 880$ observations for each stock. We realize a stylized fact about such data does not accord with the independence assumption but the acf-plots encourage us to continue with the analysis.

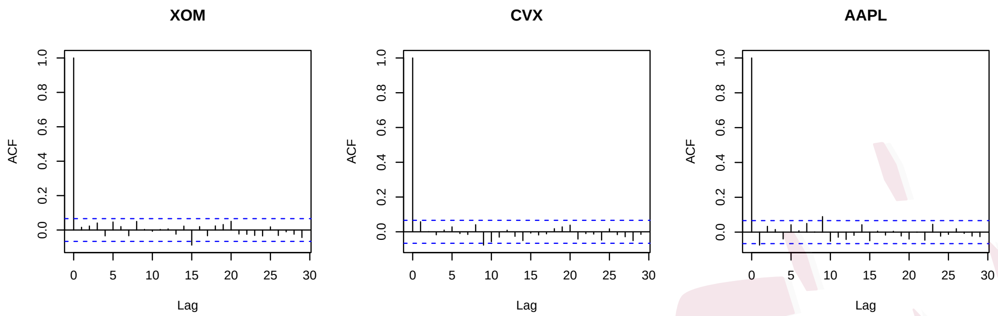

Figure 5: Acf plots for the log returns of every-other-day stock prices.

5.3.1 CVX vs XOM

In the left panel of Figure 6, we present the scatter plot of the log returns of CVX and XOM. To understand the dependence structure between absolute log returns of CVX and XOM, we also graph the histogram of $|x|/(|x|+|y|)$ in the right panel of Figure 6, where the threshold is chosen as $k(n) = 97$. The threshold $k(n) = 97$ is again suggested by the minimum distance method (Clauset et al., 2009), and the stable shape of Hill plot in the right panel of Figure 6 confirms the choice.

Based on the histogram of angles, we want to test whether there exists asymptotic strong or full dependence. Hence, we pick a large $\lambda = 4$ when solving (5.27), which gives estimates $\hat{a} = 0.257$ and $\hat{b} = 0.853$ (estimates

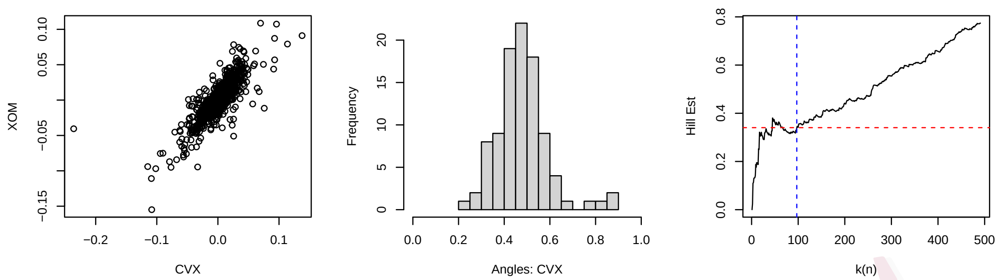

Figure 6: CVX vs XOM. Left: Scatter plot of CVX and XOM returns. Middle: Histogram of angles (absolute returns of CVX) with $k(n) = 97$. Right: Hill plot of $|x| + |y|$.

remain the same for $\lambda \geq 4$). Additionally, we have

$$\frac{1}{k(n)} \sum_{i=1}^{k(n)} \Theta_i^* = 0.482 \equiv \widehat{\theta}_0,$$

where $\widehat{\theta}_0$ will serve as a candidate when testing for full dependence. Then generate 200 bootstrap resamples with $m = \lceil 6n/k(n) \rceil$ and $k(m) = \lceil 2m^{0.4} \rceil$ to test $H_0^{(1)} : S([0.257, 0.853]) = 1$ v.s. $H_a^{(1)} : S([0.257, 0.853]) < 1$. For each bootstrap resample, we compute the corresponding test statistic $\widehat{D}_m$, and see that only 2.5% of the 200 bootstrap trials reject $H_0^{(1)}$. In addition, consider strong vs weak dependence (i.e. $H_0^{(3)}$ vs $H_a^{(3)}$), and calculate

$$\frac{\frac{1}{B-1}\sum_{t=1}^{B}\left(T_m^{(t)} - \bar{T}_m\right)^2}{\frac{1}{B-1}\sum_{t=1}^{B}\left(\tilde{T}_m^{(t)} - \bar{\tilde{T}}_m\right)^2} = 0.975 \in [0.757, 1.321].$$

Therefore, we accept the existence of strong dependence and conclude $S([0.257, 0.853]) = 1$.

To distinguish between full and strong dependence, we obtain

$$k(m)\frac{\mathrm{SE}^2_{\mathrm{boot}}(m)}{H^2_{k(n),n}} = 1.038 < \chi^2_{0.95,199}/199 = 1.204.$$

So we fail to reject the hypothesis of full dependence, i.e. $H_0^{(2)} : S(\{0.482\}) = 1$. Hence, we conclude that the absolute returns of CVX and XOM show full asymptotic dependence.

5.3.2 CVX vs AAPL

We now inspect the dependence structure between absolute returns of CVX and AAPL. The minimum distance method and the Hill plot (right panel of Figure 7) together suggest choosing $k(n) = 101$. We give the scatter plot and the histogram of $|x|/(|x| + |y|)$ in the left and middle panels Figure 7, respectively.

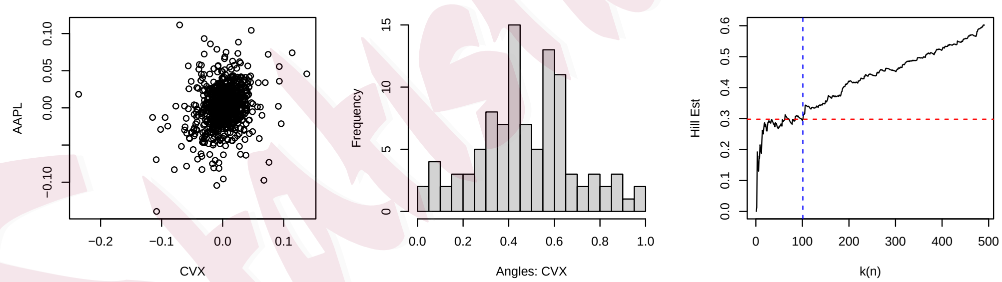

Figure 7: CVX vs AAPL. Left: Scatter plot of CVX and AAPL returns. Middle: Histogram of angles (absolute returns of CVX) with $k(n) = 101$. Right: Hill plot of $|x| + |y|$.

Different from the previous case, setting $\lambda = 4$ gives $[\hat{a}, \hat{b}] = [0.011, 0.928]$, which is already a wide interval. As before, choose also that $m = \lceil 6n/k(n) \rceil$ and $k(m) = \lceil 2m^{0.4} \rceil$. When testing

$$H_0^{(1)} : S([0.011, 0.928]) = 1 \quad \text{vs} \quad H_a^{(1)} : S([0.011, 0.928]) < 1,$$

we compute $\widehat{D}_m$ for each of the 200 resamples and 8.5% of them rejects $H_0^{(1)}$. In addition, generate two sets of 200 bootstrap resamples to test

$$H_0^{(3)} : \text{supp} S(\cdot) = [0.011, 0.928] \quad \text{vs} \quad H_a^{(3)} : \text{supp} S(\cdot) = [0, 1].$$

This gives a test statistic

$$\frac{\frac{1}{B-1}\sum_{t=1}^{B}\left(T_m^{(t)} - \bar{T}_m\right)^2}{\frac{1}{B-1}\sum_{t=1}^{B}\left(\widetilde{T}_m^{(t)} - \bar{\widetilde{T}}_m\right)^2} = 0.709 \notin [0.757, 1.321],$$

indicating the existence of weak dependence. We therefore conclude that considering the absolute returns of CVX and AAPL, the support of the angular measure is likely to be $[0, 1]$. Thus, underlying economic conditions bind both securities but not in such a way that large return changes are heavily dependent.

6. Final comments

The implementation Section 5 relies on bootstrap methods which need theoretical justification. We are completing such justifications and will report elsewhere. The data analyses in Sections 5.2, 5.3 utilizing bootstrap methods seem quite reasonable and promising and offer information about what range of models are consistent with data.

We approach finite sample inference about what is essentially an asymptotic model with caution and modest goals. We will broaden investigations of this paper to extend methodology from the classical iid setting to network data analysis and financial returns.

Supplementary Materials

The online supplementary materials contain additional simulation results and technical details for all theoretical results in the main paper.

Acknowledgements

T. Wang gratefully acknowledges National Natural Science Foundation of China Grant 12301660 and Science and Technology Commission of Shanghai Municipality Grant 23JC1400700.

The first author also thanks Shanghai Institute for Mathematics and Interdisciplinary Sciences (SIMIS) for their financial support. This research was partly funded by SIMIS under grant number SIMIS-ID-2024-WE. T. Wang is grateful for the resources and facilities provided by SIMIS, which were essential for the completion of this work.

References

Basrak, B. and H. Planinić (2019). A note on vague convergence of measures. *Statist. Probab. Lett. 153*, 180–186.

Bhattacharya, A., B. Chen, and R. van der Hofstad, B. Zwart (2020). Consistency of the PLFit estimator for power-law data. ArXiv eprint: 2002.06870.

Bollobás, B., C. Borgs, J. Chayes, , and O. Riordan (2003). Directed scale-free graphs. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms (Baltimore, MD, 2003)*, pp. 132–139. ACM, New York.

Cirkovic, D., T. Wang, and S. Resnick (2023). Preferential attachment with reciprocity: Properties and estimation. *Journal of Complex Networks 11*(5), cnad031.

Clauset, A., C. Shalizi, and M. Newman (2009). Power-law distributions in empirical data. *SIAM Rev. 51*(4), 661–703.

Csardi, G. and T. Nepusz (2006). The igraph software package for complex network research. *InterJournal, Complex Systems 1695*(5), 1–9.

Das, B. and M. Kratz (2020). Risk concentration under second order regular variation. *Extremes 23*(3), 381–410.

Das, B., A. Mitra, and S. Resnick (2013). Living on the multi-dimensional edge: Seeking hidden risks using regular variation. *Advances in Applied Probability 45*(1), 139–163.

Das, B. and S. Resnick (2017). Hidden regular variation under full and strong asymptotic dependence. *Extremes 20*(4), 873–904.

Davis, R., L. Fernandes, and K. Fokianos (2023). Clustering multivariate time series using energy distance. *Journal of Time Series Analysis 44*(5-6), 487–504.

de Haan, L. (1996). von Mises-type conditions in second order regular variation. *J. Math. Anal. Appl. 197*(2), 400–410.

de Haan, L. and J. de Ronde (1998). Sea and wind: multivariate extremes at work. *Extremes 1*(1), 7–46.

de Haan, L. and A. Ferreira (2006). *Extreme Value Theory: An Introduction*. New York: Springer-Verlag.

de Haan, L. and S. Resnick (1993). Estimating the limit distribution of multivariate extremes. *Stochastic Models 9*(2), 275–309.

de Haan, L. and S. Resnick (1996). Second-order regular variation and rates of convergence in extreme-value theory. *Ann. Probab. 24*(1), 97–124.

de Haan, L. and U. Stadtmueller (1996). Generalized regular variation of second order. *J. Aust. Math. Soc., Ser. A 61*(3), 381–395.

Drees, H., A. Janßen, and S. Neblung (2021). Cluster based inference for extremes of time series. *Stochastic Process. Appl. 142*, 1–33.

Drees, H., A. Janßen, and Resnick, S.I., Wang, T. (2020). On a minimum distance procedure for threshold selection in tail analysis. *Siam J. Math. Data Sci.*, 75–102.

Drees, H. and A. Sabourin (2021). Principal component analysis for multivariate extremes. *Electron. J. Stat. 15*(1), 908–943.

Einmahl, J., F. Yang, and C. Zhou (2021). Testing the multivariate regular variation model. *J. Bus. Econom. Statist. 39*(4), 907–919.

Feigin, P. and S. Resnick (1997). Linear programming estimators and bootstrapping for heavy tailed phenomena. *Adv. in Appl. Probab. 29*, 759–805.

Fomichov, V. and J. Ivanovs (2023). Spherical clustering in detection of groups of concomitant extremes. *Biometrika 110*(1), 135–153.

Gillespie, C. (2015). Fitting heavy tailed distributions: The poweRlaw package. *Journal of Statistical Software 64*(2), 1–16. http://www.jstatsoft.org/v64/i02/.

Hahn, M. (1978). Central limit theorems in $D[0,1]$. *Z. Wahrsch. Verw. Gebiete 44*(2), 89–101.

Hult, H. and F. Lindskog (2006). Regular variation for measures on metric spaces. *Publ. Inst. Math. (Beograd) (N.S.) 80(94)*, 121–140.

Janßen, A. and P. Wan (2020). k-means clustering of extremes. *Electron. J. Stat. 14*(1), 1211–1233.

Krapivsky, P. and S. Redner (2001). Organization of growing random networks. *Physical Review E 63*(6), 066123:1–14.

Kulik, R. and P. Soulier (2020). *Heavy-Tailed Time Series.* Springer Series in Operations Research and Financial Engineering. New York, NY: Springer.

Lehtomaa, J. and S. Resnick (2020). Asymptotic independence and support detection techniques for heavy-tailed multivariate data. *Insurance: Mathematics and Economics 93*, 262 – 277.

Lindskog, F., S. Resnick, and J. Roy (2014). Regularly varying measures on metric spaces: Hidden regular variation and hidden jumps. *Probab. Surv. 11*, 270–314.

Peng, L. (1998). *Second Order Condition and Extreme Value Theory.* Ph. D. thesis, Tinbergen Institute, Erasmus University, Rotterdam.

Pollard, D. (1990). *Empirical Processes: Theory and Applications.* NSF-CBMS Regional Conference Series in Probability and Statistics.

Resnick, S. (2002). Hidden regular variation, second order regular variation and asymptotic independence. *Extremes 5*(4), 303–336 (2003).

Resnick, S. (2007). *Heavy Tail Phenomena: Probabilistic and Statistical Modeling.* Springer Series in Operations Research and Financial Engineering. New York: Springer-Verlag. ISBN: 0-387-24272-4.

Resnick, S. (2024). *The Art of Finding Hidden Risks; Hidden Regular Variation in the 21st Century.* Switzerland: Springer. ISBN: 978-3-031-57598-3.

Samorodnitsky, G., S. Resnick, and Towsley, D, Davis, R, Willis, A, Wan, P (2016, March). Nonstandard regular variation of in-degree and out-degree in the preferential attachment model. *Journal of Applied Probability 53(1)*, 146–161.

Virkar, Y. and A. Clauset (2014). Power-law distributions in binned empirical data. *Ann. Appl. Stat. 8*(1), 89–119.

Wang, T. and S. Resnick (2022a). Asymptotic dependence of in- and out-degrees in a preferential attachment model with reciprocity. *Extremes 1*, 2.

Wang, T. and S. Resnick (2022b). Measuring reciprocity in a directed preferential attachment network. *Adv. in Appl. Probab. 54*(3), 718–742.

Wang, T. and S. Resnick (2023). Random networks with heterogeneous reciprocity. *Extremes*.

Fudan University, and Shanghai Academy of Artificial Intelligence for Science

E-mail: td_wang@fudan.edu.cn

Cornell University

E-mail: sir1@cornell.edu