QQ Plots, Random Sets and Data from a Heavy Tailed Distribution

Bikramjit Das
(joint work with Sidney Resnick)

Cornell University, Ithaca, NY

December 14, 2007

Introduction

Introduction

Preliminaries

Goodness of fit

- $X_1, \ldots, X_n$ iid sample. Does this come from some distribution F ?

Goodness of fit

- $X_1, \ldots, X_n$ iid sample. Does this come from some distribution F ?
- Answer: KS-statistics, Cramer-von Mises statistics,etc.

Goodness of fit

- $X_1, \ldots, X_n$ iid sample. Does this come from some distribution F ?
- Answer: KS-statistics, Cramer-von Mises statistics,etc.
- A visual method: *QQ plots*

Goodness of fit

- $X_1, \ldots, X_n$ iid sample. Does this come from some distribution F ?
- Answer: KS-statistics, Cramer-von Mises statistics,etc.
- A visual method: *QQ plots*
- QQ plots: Plotting the sample quantiles against the target theoretical quantiles. Let $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$ be the order statistics from this sample. Define

$$\mathbf{S}_n := \{(F^{\leftarrow}(\frac{i}{n+1}), X_{i:n}), \quad 1 \leq i \leq n\}.$$

Goodness of fit

- $X_1, \ldots, X_n$ iid sample. Does this come from some distribution F ?
- Answer: KS-statistics, Cramer-von Mises statistics,etc.
- A visual method: *QQ plots*
- QQ plots: Plotting the sample quantiles against the target theoretical quantiles. Let $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$ be the order statistics from this sample. Define

$$\mathbf{S}_n := \{(F^{\leftarrow}(\frac{i}{n+1}), X_{i:n}), \quad 1 \leq i \leq n\}.$$

- Conventional wisdom: The QQ plot becomes closer to a straight line $\mathbf{S} := \{(x, x); x \in \mathit{support}(F)\}$ as the sample size increases.

Our goal

- Confirm the conventional wisdom.

Our goal

- Confirm the conventional wisdom.
- Provide a result on the convergence of least squares line through a closed set of points, for example a QQ plot.

Our goal

- Confirm the conventional wisdom.
- Provide a result on the convergence of least squares line through a closed set of points, for example a QQ plot.
- Extend the above to the case of heavy-tailed random variables.

Our goal

- Confirm the conventional wisdom.
- Provide a result on the convergence of least squares line through a closed set of points, for example a QQ plot.
- Extend the above to the case of heavy-tailed random variables.
- Tail index estimation of a heavy-tailed distribution using the least squares line.

Closed sets and Fell topology

- $\mathcal{F}_d, \mathcal{G}_d$ and $\mathcal{K}_d$ are the classes of closed, open and compact subsets of $\mathbb{R}^d$ respectively.

Closed sets and Fell topology

- $\mathcal{F}_d, \mathcal{G}_d$ and $\mathcal{K}_d$ are the classes of closed, open and compact subsets of $\mathbb{R}^d$ respectively.

$$\mathcal{F}_{\mathbf{B}} = \{\mathbf{F} : \mathbf{F} \in \mathcal{F}, \quad \mathbf{F} \bigcap \mathbf{B} \neq \emptyset\}, \quad \mathcal{F}^{\mathbf{B}} = \{\mathbf{F} : \mathbf{F} \in \mathcal{F}, \quad \mathbf{F} \bigcap \mathbf{B} = \emptyset\}.$$

Closed sets and Fell topology

- $\mathcal{F}_d, \mathcal{G}_d$ and $\mathcal{K}_d$ are the classes of closed, open and compact subsets of $\mathbb{R}^d$ respectively.

$$\mathcal{F}_{\mathbf{B}} = \{\mathbf{F} : \mathbf{F} \in \mathcal{F}, \quad \mathbf{F} \bigcap \mathbf{B} \neq \emptyset\}, \quad \mathcal{F}^{\mathbf{B}} = \{\mathbf{F} : \mathbf{F} \in \mathcal{F}, \quad \mathbf{F} \bigcap \mathbf{B} = \emptyset\}.$$

- The space $\mathcal{F}$ can be topologized by the Fell topology which has as its subbase the families $\{\mathcal{F}^{\mathbf{K}}, \mathbf{K} \in \mathcal{K}\}$ and $\{\mathcal{F}_{\mathbf{G}}, \mathbf{G} \in \mathcal{G}\}$.

Closed sets and Fell topology

- $\mathcal{F}_d, \mathcal{G}_d$ and $\mathcal{K}_d$ are the classes of closed, open and compact subsets of $\mathbb{R}^d$ respectively.

$$\mathcal{F}_{\mathbf{B}} = \{\mathbf{F} : \mathbf{F} \in \mathcal{F}, \quad \mathbf{F} \bigcap \mathbf{B} \neq \emptyset\}, \quad \mathcal{F}^{\mathbf{B}} = \{\mathbf{F} : \mathbf{F} \in \mathcal{F}, \quad \mathbf{F} \bigcap \mathbf{B} = \emptyset\}.$$

- The space $\mathcal{F}$ can be topologized by the Fell topology which has as its subbase the families $\{\mathcal{F}^{\mathbf{K}}, \mathbf{K} \in \mathcal{K}\}$ and $\{\mathcal{F}_{\mathbf{G}}, \mathbf{G} \in \mathcal{G}\}$.
- The Fell topology on $\mathcal{F}$ is compact and metrizable (Beer(1993), Flaschmeyer(1963)).

A sequence $\{\mathbf{F}_n\}$ converges to $\mathbf{F}$ in $\mathcal{F}$ in the Fell topology if and only if

1. If an open set $\mathbf{G}$ hits $\mathbf{F}$, $\mathbf{G}$ hits all $\mathbf{F}_n$, provided n is sufficiently large.
2. If a compact set $\mathbf{K}$ is disjoint from $\mathbf{F}$, it is disjoint from $\mathbf{F}_n$ for all sufficiently large n.

A sequence $\{\mathbf{F}_n\}$ converges to $\mathbf{F}$ in $\mathcal{F}$ in the Fell topology if and only if

1. If an open set $\mathbf{G}$ hits $\mathbf{F}$, $\mathbf{G}$ hits all $\mathbf{F}_n$, provided n is sufficiently large.
2. If a compact set $\mathbf{K}$ is disjoint from $\mathbf{F}$, it is disjoint from $\mathbf{F}_n$ for all sufficiently large n.

Lemma (Matheron)

For $\mathbf{F}_n, \mathbf{F} \in \mathcal{F}, n \geq 1, \mathbf{F}_n \to \mathbf{F}$ as $n \to \infty$ if and only if:

- *For any $\mathbf{y} \in \mathbf{F}$, for all large n, $\exists\, \mathbf{y}_n \in \mathbf{F}_n$ s.t. $d(\mathbf{y}_n, \mathbf{y}) \to 0$ as $n \to \infty$.*
- *For any subsequence $\{n_k\}$, if $\mathbf{y}_{n_k} \in \mathbf{F}_{n_k}$ converges, then $\lim_{k\to\infty} \mathbf{y}_{n_k} \in \mathbf{F}$*

Furthermore, if $\sup_{j\geq 1} \sup\{\|\mathbf{x}\| : \mathbf{x} \in \mathbf{S}_j\} < \infty$ then sets $\mathbf{S}_n \to \mathbf{S}$ in $\mathcal{K}$.

This Lemma directly translates to almost sure convergence or convergence in probability of a sequence of random sets to a non-random limit.

Random closed sets

- $(\Omega, \mathcal{A}, P^{'})$ - a complete probability space.
- $\sigma_{\mathcal{F}}$: Borel σ-algebra of subsets of $\mathcal{F}$ generated by the Fell topology.
- A *random closed set* $\mathbf{X} : \Omega \mapsto \mathcal{F}$ is a measurable mapping from $(\Omega, \mathcal{A}, P^{'})$ to $(\mathcal{F}, \sigma_{\mathcal{F}})$.
- $\{\mathbf{X}_n\}_{n\geq 1}$ weakly converges to a random closed set $\mathbf{X}$ with distribution P if

$$P_n(\mathcal{B}) = P^{'} \circ \mathbf{X}_n^{-1}(\mathcal{B}) \to P(\mathcal{B}) = P^{'} \circ \mathbf{X}^{-1}(\mathcal{B}), \qquad \text{as } n \to \infty,$$

for each $\mathcal{B} \in \sigma_{\mathcal{F}}$ such that $P(\partial\mathcal{B}) = 0$.

Alternative verification of Weak convergence

- Weak convergence in terms of sup-measures (Vervaat,1997) .
- $h : \mathbb{R}^d \mapsto \mathbb{R}_+ = [0, \infty)$. For $\mathbf{X} \subset \mathbb{R}^d$, define $h(\mathbf{X}) = \{h(\mathbf{x}) : \mathbf{x} \in \mathbf{X}\}$ and $h^\vee$ is the sup-measure generated by h defined by

$$h^\vee(\mathbf{X}) = \sup\{h(\mathbf{x}) : \mathbf{x} \in \mathbf{X}\}$$

Lemma (Choquet's theorem)

A sequence $\{\mathbf{X}_n\}_{n\geq 1}$ of random closed sets converges weakly to a random closed set $\mathbf{X}$ if and only if $\mathbb{E}h^\vee(\mathbf{X}_n)$ converges to $\mathbb{E}h^\vee(\mathbf{X})$ for every non-negative continuous function $h : \mathbb{R}^d \mapsto \mathbb{R}$ with a bounded support.

Convergence of QQ plots with a known target distribution

Proposition

Suppose $X_1, \ldots, X_n$ are iid with common distribution $F(\cdot)$ and $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$ are the order statistics from this sample. If F is strictly increasing and continuous on its support, then

$$\mathbf{S}_n := \{(F^{\leftarrow}(\frac{i}{n+1}), X_{i:n}); 1 \leq i \leq n\}$$

converges in probability to

$$\mathbf{S} := \{(x, x); x \in \mathit{support}(F)\}$$

in $\mathcal{F}_2$.

Idea of proof

- Prove for F being the $U(0,1)$ distribution. *Easy*. Can prove *a.s.* convergence.
- For general F, increasing and continuous, $F(X_1), F(X_2), \ldots, F(X_n)$ are iid and uniformly distributed on $[0,1]$. So,

$$\mathbf{U}_n := \{(\frac{i}{n+1}, F(X_{i:n})); 1 \leq i \leq n\} \stackrel{a.s.}{\to} \mathbf{U} = \{(x,x); 0 \leq x \leq 1\}.$$

- Use Choquet's theorem. Let F have compact support on $[a,b]$. Define the map $g : [0,1]^2 \mapsto [a,b]^2$ by

$$g(x,y) = (F^{\leftarrow}(x), F^{\leftarrow}(y)).$$

- Observe that $g(\mathbf{U}_n)$=$\mathbf{S}_n$ and $g(\mathbf{U}) = \mathbf{S}$.

- Extend to $\mathbb{R}^2$ continuously such that $g(\mathbf{U}_n) = g^*(\mathbf{U}_n)$ and $g(\mathbf{U}) = g^*(\mathbf{U})$.
- Let f be a non-negative continuous function on $\mathbb{R}^2$ with bounded support.

$$\begin{aligned}\mathbb{E}f^\vee(\mathbf{S}_n) &= \mathbb{E}f^\vee(g(\mathbf{U}_n)) = \mathbb{E}f^\vee(g^*(\mathbf{U}_n)) \\ &= \mathbb{E}(f \circ g^*)^\vee(\mathbf{U}_n) \to \mathbb{E}(f \circ g^*)^\vee(\mathbf{U}) \\ &= \mathbb{E}f^\vee(g^*(\mathbf{U})) = \mathbb{E}f^\vee(g(\mathbf{U})) = \mathbb{E}f^\vee(\mathbf{S}).\end{aligned}$$

Therefore $\mathbf{S}_n$ converges to $\mathbf{S}$ weakly and hence in probability.

Examples

(a) $F \sim \exp(\alpha)$ with > 0, i.e., $F(x) = 1 - e^{-\alpha x}$, $x > 0$.Then

$$\{(-\frac{1}{\alpha}\log(1 - \frac{i}{n+1}), X_{i:n}); 1 \leq i \leq n\} \xrightarrow{P} \{(x, x) : 0 \leq x < \infty\}.$$

(b) $F \sim Pareto(\alpha)$ with $\alpha > 0$, i.e., $F(x) = 1 - x^{-\alpha}, x > 1$. Then

$$\{(-\log(1 - \frac{i}{n+1}), \log X_{i:n}); 1 \leq i \leq n\} \xrightarrow{P} \{(x, \frac{x}{\alpha}) : 0 \leq x < \infty\}.$$

Regular variation and slow variation

Definition (Regular variation)

A measurable function $U(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ is regularly varying at ∞ with index $\rho \in \mathbb{R}$ if for $x > 0$

$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^\rho. \tag{4.1}$$

We write $U \in RV_\rho$. When $\rho = 0$ we call $U(\cdot)$ *slowly varying* and denote it by $L(\cdot)$. For $\rho \in \mathbb{R}$, we can always write $U \in RV_\rho$ as $U(x) = x^\rho L(x)$, where $L(\cdot)$ is slowly varying.

A univariate distribution is said to be heavy-tailed or regularly varying with some index α if $\bar{F} \in RV_{-\alpha}$. The generic example is:

- $X \sim$ Pareto (α) where $F(x) = 1 - x^{-\alpha}, x > 1, \alpha > 0$.

Convergence of QQ plots for Regularly varying distributions

Proposition

$X_1, X_2, \ldots, X_n$ i.i.d. from F where $\bar{F} \in RV_{-\alpha}$. We have $X_{(1)} \geq X_{(2)} \geq \ldots \geq X_{(n)}$. Define

$$\mathbf{S}_n' = \{(-\log \frac{j}{k}, \log \frac{X_{(j)}}{X_{(k)}}); 1 \leq j \leq k\},$$

$$\mathbf{T} = \{(x, \frac{x}{\alpha}); 0 \leq x < \infty\}.$$

Assume $k = k(n) \to \infty$ and $k/n \to 0$ as $n \to \infty$. Then as $n \to \infty$

$$\mathbf{S}_n' \xrightarrow{P} \mathbf{T}.$$

Idea of Proof

$$\mathbf{S}_n^{'} = \{(-\log t, \log \frac{X_{([kt])}}{X_{(k)}}); t \in \{\frac{1}{k}, \ldots, \frac{k-1}{k}, 1\}\},$$

$$\mathbf{T} = \{(x, \frac{x}{\alpha}); x \geq 0\} = \{(-\log t, -\frac{1}{\alpha}\log t); 0 < t \leq 1\}.$$

Idea of Proof

$$\mathbf{S}_n^{'} = \{(-\log t, \log \frac{X_{([kt])}}{X_{(k)}}); t \in \{\frac{1}{k}, \ldots, \frac{k-1}{k}, 1\}\},$$

$$\mathbf{T} = \{(x, \frac{x}{\alpha}); x \geq 0\} = \{(-\log t, -\frac{1}{\alpha}\log t); 0 < t \leq 1\}.$$

Lemma

$0 \leq x(\cdot) \in C(0,1]$ *is continuous on* $(0,1]$ *and strictly decreasing with* $\lim_{\epsilon \downarrow 0} x(\epsilon) = \infty$. $y_n(\cdot) \in D_l(0,1]$ *and* $y(\cdot) \in C(0,1]$ *and* $y_n \to y$ *locally uniformly on* $(0,1]$. *Then for* $k = k(n) \to \infty$,

$$\mathbf{F}_n := \{\big(x(\frac{j}{k}), y_n(\frac{j}{k})\big); 1 \leq j \leq k\} \to \{\big(x(t), y(t)\big); 0 < t \leq 1\} =: \mathbf{F}$$

in $\mathcal{F}_2$.

- Set

$$x(t) = -\log t,$$

$$Y_n(t) = \log \frac{X_{(\lceil kt \rceil)}}{X_{(k)}}, \quad y(t) = -\frac{1}{\alpha} \log t, \qquad 0 < t \leq 1.$$

- Set

$$x(t) = -\log t,$$
$$Y_n(t) = \log \frac{X_{(\lceil kt \rceil)}}{X_{(k)}}, \quad y(t) = -\frac{1}{\alpha} \log t, \qquad 0 < t \leq 1.$$

- We have $Y_n \xrightarrow{P} y$, in $D_l(0,1]$, the left continuous functions on $(0,1]$ with finite right limits, metrized by the Skorohod metric(Resnick,2006).

- Set

$$x(t) = -\log t,$$
$$Y_n(t) = \log \frac{X_{(\lceil kt \rceil)}}{X_{(k)}}, \quad y(t) = -\frac{1}{\alpha} \log t, \qquad 0 < t \leq 1.$$

- We have $Y_n \xrightarrow{P} y$, in $D_l(0,1]$, the left continuous functions on $(0,1]$ with finite right limits, metrized by the Skorohod metric(Resnick,2006).
- Suppose $\{n''\}$ is a subsequence. There exists a further subsequence $\{n'\} \subset \{n''\}$ such that $Y_{n'} \xrightarrow{a.s.} y$, in $D_l(0,1]$.

- Set

$$x(t) = -\log t,$$
$$Y_n(t) = \log \frac{X_{(\lceil kt \rceil)}}{X_{(k)}}, \quad y(t) = -\frac{1}{\alpha} \log t, \qquad 0 < t \leq 1.$$

- We have $Y_n \xrightarrow{P} y$, in $D_l(0,1]$, the left continuous functions on $(0,1]$ with finite right limits, metrized by the Skorohod metric(Resnick,2006).
- Suppose $\{n''\}$ is a subsequence. There exists a further subsequence $\{n'\} \subset \{n''\}$ such that $Y_{n'} \xrightarrow{a.s.} y$, in $D_l(0,1]$.
- This convergence is locally uniform because of continuity of y in $(0,1]$. Hence, as $n \to \infty$, $\mathbf{S}'_{n'} \xrightarrow{a.s.} \mathbf{T}$ in $\mathcal{F}$ and therefore, $\mathbf{S}'_n \xrightarrow{P} \mathbf{T}$ in $\mathcal{F}_2$.

The Least squares Line

- Suppose $\mathbf{S}_n$ is a closed of points converging to the straight line $\mathbf{S}$. Does $LS(\mathbf{S}_n) \longrightarrow$ slope of $\mathbf{S}$?

The Least squares Line

- Suppose $\mathbf{S}_n$ is a closed of points converging to the straight line $\mathbf{S}$. Does $LS(\mathbf{S}_n) \rightarrow$ slope of $\mathbf{S}$?
- No. Need a small assumption.

The Least squares Line

- Suppose $\mathbf{S}_n$ is a closed of points converging to the straight line $\mathbf{S}$. Does $LS(\mathbf{S}_n) \rightarrow$ slope of $\mathbf{S}$?
- No. Need a small assumption. Consider the following case:

$$\mathbf{S}_n = \{(\frac{i}{n}, 0), -n \leq i \leq n; (\frac{1}{n}(1 + \frac{j}{2^n}), \frac{1}{n}(1 + \frac{j}{2^n})), 0 \leq j \leq 2^n\}$$

and

$$\mathbf{S} = [-1, 1] \times \{0\}.$$

The Least squares Line

- Suppose $\mathbf{S}_n$ is a closed of points converging to the straight line $\mathbf{S}$. Does $LS(\mathbf{S}_n) \to$ slope of $\mathbf{S}$?
- No. Need a small assumption. Consider the following case:

$$\mathbf{S}_n = \{(\frac{i}{n}, 0), -n \leq i \leq n; (\frac{1}{n}(1 + \frac{j}{2^n}), \frac{1}{n}(1 + \frac{j}{2^n})), 0 \leq j \leq 2^n\}$$

and

$$\mathbf{S} = [-1, 1] \times \{0\}.$$

- Slope of $\mathbf{S} = 0$ but $LS(\mathbf{S}_n) \to 1$.

Convergence of the Least Squares line

Proposition

Suppose we have a sequence of sets
$\mathbf{F}_n := \{(x_i(n), y_i(n)) : 1 \leq i \leq k_n\} \in \mathcal{K}_2$, each consisting of k_n points, which converge to a bounded line segment $\mathbf{F} \in \mathcal{K}_2$ with slope m where $|m| < \infty$. Then,

$$LS(\mathbf{F}_n) \to LS(\mathbf{F}) = m$$

provided $k_n \to \infty$ and the following condition holds: $\exists\ \delta > 0$, such that

$$p_\delta^n := \frac{\#\Big(\{(\bar{x}_n - \delta, \bar{x}_n + \delta) \times (\bar{y}_n - \delta, \bar{y}_n + \delta)\} \bigcap \mathbf{F}_n\Big)}{\#\mathbf{F}_n} \to p_\delta \in [0, 1).$$

Corollary

Corollary

If $\bar{x}_n \to \mu_x < \infty$ *and* $\bar{y}_n \to \mu_y < \infty$, *as* $n \to \infty$, *then the previous Proposition 5.1 holds if we replace* $(\bar{x}_n, \bar{y}_n)$ *in the condition by* (μ_x, μ_y). *So we assume that* $\exists\ \delta > 0$ *such that*

$$p_\delta^n = \frac{\#\{(\mu_x - \delta, \mu_x + \delta) \times (\mu_y - \delta, \mu_y + \delta)\} \bigcap \mathbf{F}_n}{\#\mathbf{F}_n} \to p_\delta \in [0, 1).$$

This is true for almost sure and in probability convergence. Hence the result holds for QQ plots of a distribution with finite support.

Tail index estimator in heavy-tailed distributions

- How to estimate the tail index parameter, α ?

Tail index estimator in heavy-tailed distributions

- How to estimate the tail index parameter, α ?
- Hill estimator for $\frac{1}{\alpha}$:

$$H_{k,n} = \frac{1}{k} \sum_{i=1}^{k} \frac{X_{(i)}}{X_{(k+1)}}$$

Tail index estimator in heavy-tailed distributions

- How to estimate the tail index parameter, α ?
- Hill estimator for $\frac{1}{\alpha}$:

$$H_{k,n} = \frac{1}{k} \sum_{i=1}^{k} \frac{X_{(i)}}{X_{(k+1)}}$$

- The slope of the least squares line through the QQ plot made by the upper k_n largest order statistics is a consistent estimator of $1/\alpha$. See Kratz, Resnick(1996), Beirlant, Vynckier, Teugels(1996).

Connecting ideas

Proposition

$X_1, X_2, \ldots, X_n$ non-negative are iid F where $\bar{F} \in RV_{-\alpha}$.Also $X_{(1)} \geq X_{(2)} \geq \ldots \geq X_{(n)}$.Again set

$$\mathbf{S}_n^{'} = \{(-\log \frac{j}{k_n}, \log \frac{X_{(j)}}{X_{(k_n)}}); j = 1, \ldots, k_n\} \quad \text{and}$$

$$\mathbf{T} = \{(x, \frac{x}{\alpha}); x \geq 0\}.$$

Then,

$$LS(\mathbf{S}_n^{'}) \xrightarrow{P} \frac{1}{\alpha} = LS(\mathbf{T}),$$

as $k := k_n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$.

- For samples from a continuous and strictly increasing distribution F, the QQ plot converges to a straight line in probability.
- For heavy tailed random variables we have a similar result.
- The LS line converges unless we have a clustering of points to a single point in the limit set.
- Reaffirming that the LS estimate is weakly consistent for $1/\alpha$ in the heavy-tailed case.