MURI 3.6

Phyllis Wan

Columbia University

March 6, 2015

Set-up

- Let $\mathbf{X} \sim RV(\alpha)$, $R = ||\mathbf{X}||$, $\Theta = \mathbf{X}/||\mathbf{X}||$
- For any $\{r_n\} \uparrow \infty$,

$$P\left[\left(\frac{R}{r_n}, \Theta\right) \in \cdot | R > r_n\right] \xrightarrow{w} \nu_\alpha \times S, \text{ as } n \to \infty. \tag{1}$$

Set-up

- Let $\mathbf{X} \sim RV(\alpha)$, $R = ||\mathbf{X}||$, $\Theta = \mathbf{X}/||\mathbf{X}||$
- For any $\{r_n\} \uparrow \infty$,

$$P\left[\left(\frac{R}{r_n}, \Theta\right) \in \cdot | R > r_n\right] \stackrel{w}{\to} \nu_\alpha \times S, \text{ as } n \to \infty. \tag{1}$$

- Goal: Given iid data $\mathbf{X_1}, \mathbf{X_2} \ldots$, to find r such that when $R > r$,

$$R \perp\!\!\!\perp \Theta$$

approximately.

Distance Covariance: a Measure of Dependence

- Random variables $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$,

$$X \perp\!\!\!\perp Y \iff f_{X,Y} = f_X f_Y.$$

Distance Covariance: a Measure of Dependence

- Random variables $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$,

$$X \perp\!\!\!\perp Y \iff f_{X,Y} = f_X f_Y.$$

- Distance covariance (Székely et al., 2007)

$$\mathcal{V}^2(X,Y) = \int \left|f_{X,Y}(s,t) - f_X(s) f_Y(t)\right|^2 \cdot w(s,t) ds dt,$$

where

$$w(s,t) = \frac{1}{c_p c_q |t|_p^{1+p} |s|_q^{1+q}}.$$

Distance Covariance: a Measure of Dependence

- Random variables $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$,

$$X \perp\!\!\!\perp Y \iff f_{X,Y} = f_X f_Y.$$

- Distance covariance (Székely et al., 2007)

$$\mathcal{V}^2(X, Y) = \int |f_{X,Y}(s,t) - f_X(s)f_Y(t)|^2 \cdot w(s,t)dsdt,$$

  where

$$w(s,t) = \frac{1}{c_p c_q |t|_p^{1+p} |s|_q^{1+q}}.$$

- If $E|X| + E|Y| < \infty$, then

$$\mathcal{V}^2(X, Y) < \infty.$$

Distance Covariance: a Measure of Dependence

- Random variables $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$,

$$X \perp\!\!\!\perp Y \iff f_{X,Y} = f_X f_Y.$$

- Distance covariance (Székely et al., 2007)

$$\mathcal{V}^2(X, Y) = \int |f_{X,Y}(s,t) - f_X(s) f_Y(t)|^2 \cdot w(s,t) ds dt,$$

where

$$w(s,t) = \frac{1}{c_p c_q |t|_p^{1+p} |s|_q^{1+q}}.$$

- If $E|X| + E|Y| < \infty$, then

$$\mathcal{V}^2(X, Y) < \infty.$$

- Distance correlation:

$$\mathcal{R}^2(X, Y) = \frac{\mathcal{V}^2(X, Y)}{\sqrt{\mathcal{V}^2(X, X)\mathcal{V}^2(Y, Y)}}.$$

Distance Covariance: a Measure of Dependence

- Given $\{(X_k, Y_K), k = 1, \dots, n\}$, define

$$a_{kl} = |X_k - X_l|_p, \;\; \bar{a}_{k\cdot} = \frac{1}{n}\sum_{l=1}^{n} a_{kl}, \;\; \bar{a}_{\cdot l} = \frac{1}{n}\sum_{k=1}^{n} a_{kl}, \;\; \bar{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl}$$

$$A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot}$$

  Define B_{kl} similarly with $b_{kl} = |Y_k - Y_l|_q$'s.

- Empirical distance covariance

$$\mathcal{V}_n^2(X, Y) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl} B_{kl}$$

Distance Covariance: a Measure of Dependence

- Given $\{(X_k, Y_K), k = 1, \ldots, n\}$, define

$$a_{kl} = |X_k - X_l|_p, \ \ \bar{a}_{k\cdot} = \frac{1}{n}\sum_{l=1}^{n} a_{kl}, \ \ \bar{a}_{\cdot l} = \frac{1}{n}\sum_{k=1}^{n} a_{kl}, \ \ \bar{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl}$$

$$A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot}$$

  Define B_{kl} similarly with $b_{kl} = |Y_k - Y_l|_q$'s.

- Empirical distance covariance

$$\mathcal{V}_n^2(X, Y) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl} B_{kl}$$

- If $E|X| + E|Y| < \infty$, then

$$\mathcal{V}_n^2(X, Y) = \int |\hat{f}_{X,Y}(s,t) - \hat{f}_X(s)\hat{f}_Y(t)|^2 \cdot w(s,t) ds dt,$$

Distance Covariance: a Measure of Dependence

- If $E|X| + E|Y| < \infty$, then

$$\mathcal{V}_n^2(X, Y) \overset{a.s.}{\to} \mathcal{V}^2(X, Y)$$

Distance Covariance: a Measure of Dependence

- If $E|X| + E|Y| < \infty$, then

$$\mathcal{V}_n^2(X, Y) \overset{a.s.}{\to} \mathcal{V}^2(X, Y)$$

- Additionally, if X and Y are independent, then

$$n\mathcal{V}_n^2(X, Y) \overset{w}{\to} \int |\zeta(s, t)|^2 \cdot w(s, t) dsdt,$$

where ζ is a zero-mean Gaussian process.

Distance Covariance: a Measure of Dependence

- If $E|X| + E|Y| < \infty$, then

$$\mathcal{V}_n^2(X, Y) \overset{a.s.}{\to} \mathcal{V}^2(X, Y)$$

- Additionally, if X and Y are independent, then

$$n\mathcal{V}_n^2(X, Y) \overset{w}{\to} \int |\zeta(s, t)|^2 \cdot w(s, t) dsdt,$$

  where ζ is a zero-mean Gaussian process.

- Can be used as a test for independence.

Threshold Selection

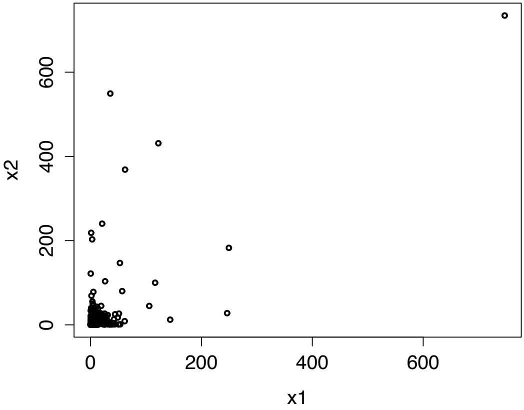


Figure : $(X_1, X_2) \sim$ Bivariate Logistic with parameter $\beta = 0.7$,$n = 1000$

Threshold Selection

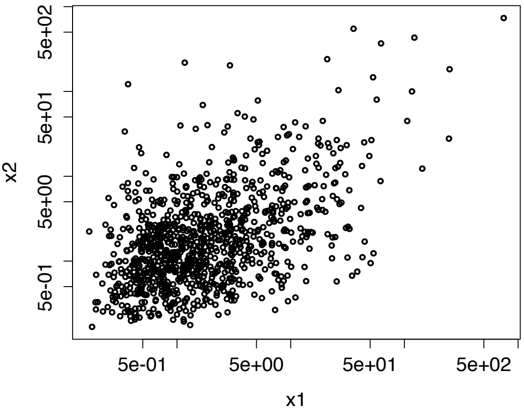


Figure : $(X_1, X_2) \sim$ Bivariate Logistic with parameter $\beta = 0.7, n = 1000$

Threshold Selection

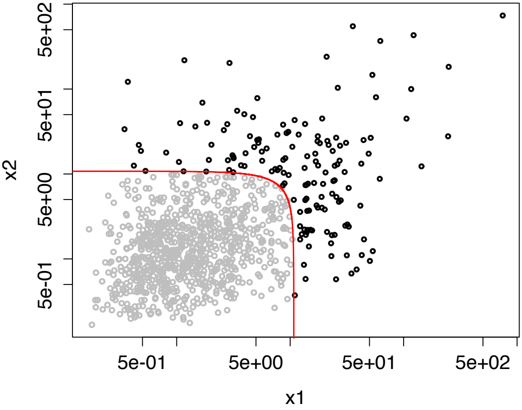


Figure : $(X_1, X_2) \sim$ Bivariate Logistic with parameter $\beta = 0.7, n = 1000$

Threshold Selection

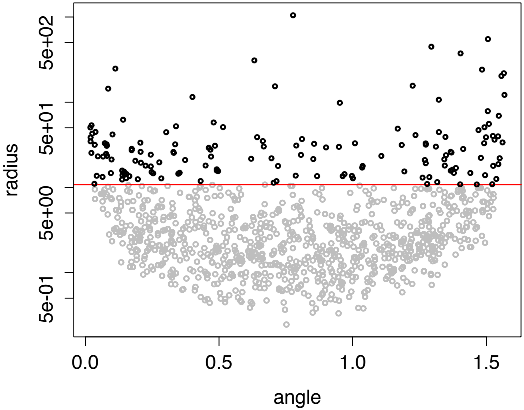


Figure : $(X_1, X_2) \sim$ Bivariate Logistic with parameter $\beta = 0.7, n = 1000$

Threshold Selection

- Let $(i_1, \ldots, i_{k_r})$ be the indices of the data points with $R > r$
- Test independence between $(R_{i_1}, \ldots, R_{i_{k_r}})$ and $(\Theta_{i_1}, \ldots, \Theta_{i_{k_r}})$ using distance correlation
- Choose r when the independence become significant

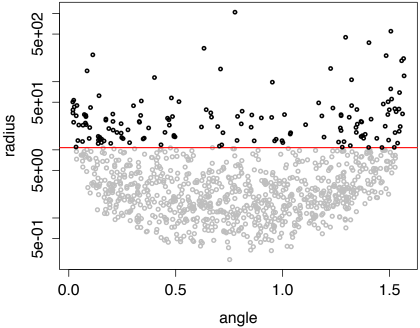


Figure : $(X_1, X_2) \sim$ Bivariate Logistic with parameter $\beta = 0.7, n = 1000$

Threshold Selection

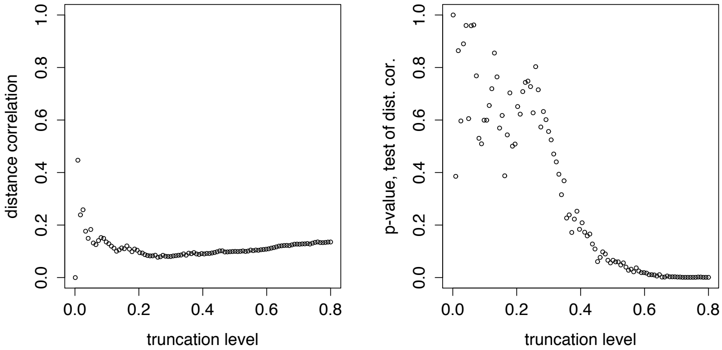


Figure : Distance correlation and p-value of test of independence of R and Θ vs. truncation level for $(X_1, X_2) \sim$ Bivariate Logistic with parameter $\beta = 0.7$, $n = 1000$

P-value Path

- A set of truncation levels $q_1, \ldots, q_k$
- P-value of test of independence of (R, Θ) at each truncation level $p_1, \ldots, p_k$
- How can we identify/approximate the independence threshold?

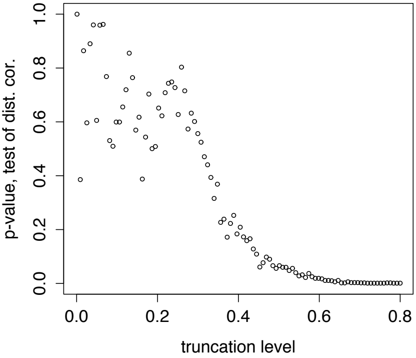


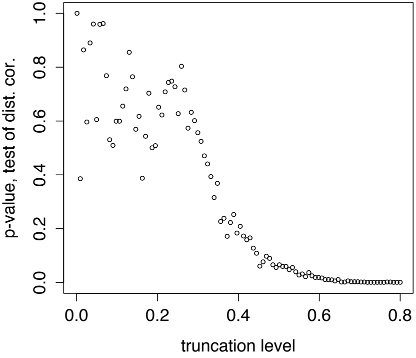
Figure : $(X_1, X_2) \sim$ Bivariate Logistic with parameter $\beta = 0.7, n = 1000$

P-value Path

Change point detection

- Look at the before and after mean ratio at each level i:

$$r_i = \frac{\frac{1}{i}\sum_{j=1}^{i} p_j}{\frac{1}{k-i}\sum_{j=i+1}^{k} p_j}$$

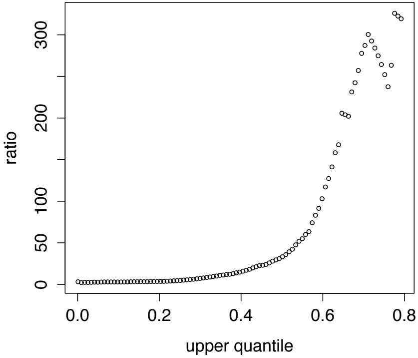


P-value Path

Change point detection

- Look at the before and after mean ratio at each level i:

$$r_i = \frac{\frac{1}{i}\sum_{j=1}^{i} p_j}{\frac{1}{k-i}\sum_{j=i+1}^{k} p_j}$$



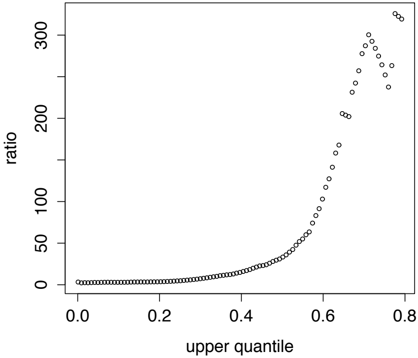
P-value Path

Change point detection

- Look at the before and after mean ratio at each level i:

$$r_i = \frac{\frac{1}{i}\sum_{j=1}^{i} p_j}{\frac{1}{k-i}\sum_{j=i+1}^{k} p_j}$$

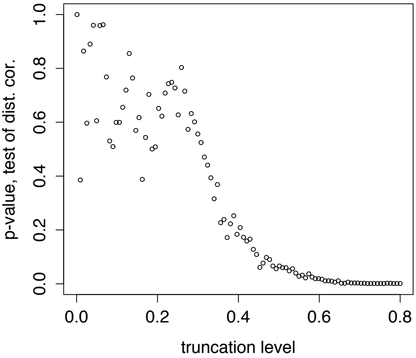
- Select the level when the ratio starts increasing significantly



P-value Path

Piecewise linear spline fitting

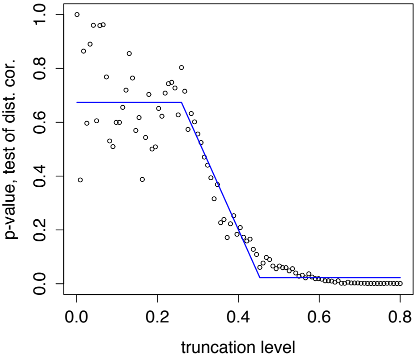
- Fit a sloped step function to the p-value path



P-value Path

Piecewise linear spline fitting

- Fit a sloped step function to the p-value path



Stock Price Return Data

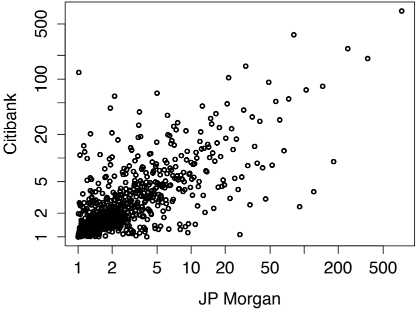


Figure : Rank-transformed weekly stock price return from JP Morgan vs. Citibank

Stock Price Return Data

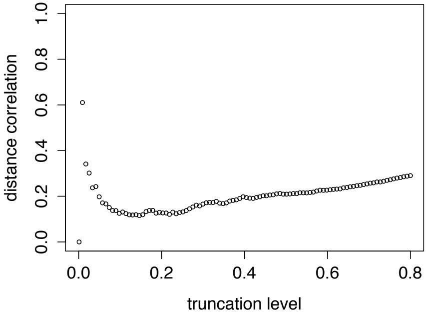


Figure : Distance correlation vs. truncation level of polar coordinates of the stock return data

Stock Price Return Data

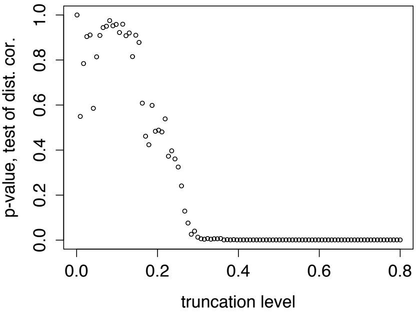


Figure : P-value of independence test vs. truncation level of the stock return data

Stock Price Return Data

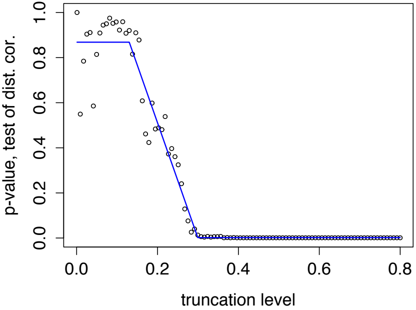


Figure : P-value of independence test vs. truncation level of the stock return data

Stock Price Return Data

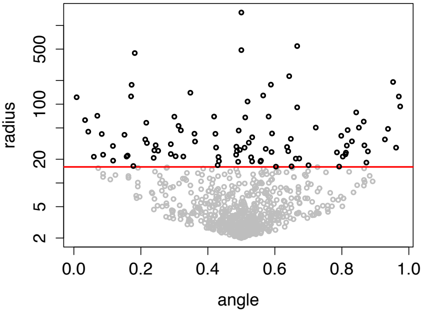


Figure : Radial vs. Angula measure after L_2-norm polar coordinate transformation with conservative threshold estimate

Stock Price Return Data

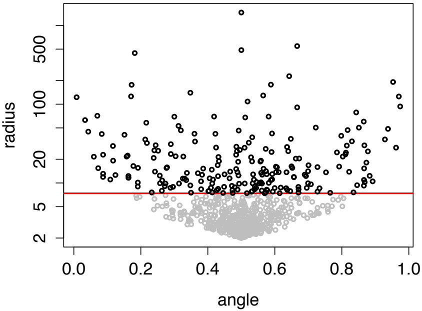


Figure : Radial vs. Angula measure after L_2-norm polar coordinate transformation with a more generous estimate

Theoretical Results

Proposition 1

Let $\{(\mathbf{X}_{in}, \mathbf{Y}_{in})\}_{1,\ldots,T_n} \in \mathbb{R}^p \times \mathbb{R}^q$, $n \to \infty$, such that

$$(\mathbf{X}_{in}, \mathbf{Y}_{in}) \overset{iid}{\sim} P_{X_n,Y_n},$$

where $T_n \sim B(n, p_n)$, and

$$P_{X_n,Y_n} \overset{d}{\to} P_{X,Y}.$$

Assume that

$$E|\mathbf{X}_n| < M_X < \infty, \;\; E|\mathbf{Y}_n| < M_Y < \infty,$$

and

$$np_n \to \infty.$$

Denote $\mathbf{X}_n = (\mathbf{X}_{1n}, \ldots, \mathbf{X}_{T_n n})$, $\mathbf{Y}_n = (\mathbf{Y}_{1n}, \ldots, \mathbf{Y}_{T_n n})$. Then

$$\mathcal{V}_n(\mathbf{X}_n, \mathbf{Y}_n) \overset{p}{\to} \mathcal{V}(X, Y).$$

Theoretical Results

Proposition 2

In addition, if X and Y are independent. And <some other conditions here>. Then

$$n \cdot \mathcal{V}_n(\mathbf{X}_n, \mathbf{Y}_n) \xrightarrow{d} ||\zeta(s,t)||^2,$$

where $\zeta(s,t)$ is a zero-mean Gaussian process.

Theoretical Results

Proposition 2

In addition, if X and Y are independent. And <some other conditions here>. Then

$$n \cdot \mathcal{V}_n(\mathbf{X}_n, \mathbf{Y}_n) \xrightarrow{d} ||\zeta(s,t)||^2,$$

where $\zeta(s,t)$ is a zero-mean Gaussian process.

- A possible condition is

$$\begin{aligned} \sqrt{n} \cdot (X_n - X) &\xrightarrow{L_1} 0 \\ \sqrt{n} \cdot (Y_n - Y) &\xrightarrow{L_1} 0 \end{aligned}$$

- Second-order regular variation?