Detecting independence via Lévy metrics

⊕	B. Böttcher	2018	MSTA
	M. Keller-Ressel	2019	AoS
⊕⊕	B. Böttcher	2020	Open Stats.

Key Words distance covariance

distance multivariance

Szekely, Rizzo, Bakirov

≈ 2007/2009

2 r.v.

n r.v.

RS et al.

§ Problem

- measure (in-)dependence of data sets X, Y
- high dim ⇝ computation!
- $\dim X \neq \dim Y$

classical correlation

Pearson corr. → COV=0 ⇏ ⫫

2 rv only

same dim

Szeleky:

1° $f_Z(u) = \mathbb{E}\, e^{iuZ}$ char fn.

$$f_{(X,Y)}(s,t) - f_X(s) f_Y(t) \overset{!}{=} 0$$

2° $$V^2(X,Y) = \iint \left| f_{X,Y}(s,t) - f_X(s) f_Y(t) \right|^2 \underbrace{\omega(s,t)\,ds\,dt}_{\text{weight}}$$

$$\omega(s,t) = \underbrace{c_{\alpha,m} |s|^{-\alpha-m}}_{\text{Lévy measure, } \alpha\text{-stable}} c_{\alpha,n} |t|^{-\alpha-n}$$

3° $$V^2(X,Y) = \mathbb{E}\, |X-X'|^{\alpha} |Y-Y'|^{\alpha} + \mathbb{E}|X-X'|^{\alpha}\, \mathbb{E}|Y-Y'|^{\alpha} - 2\,\mathbb{E}\, |X'-\tilde{X}|^{\alpha} |Y'-\tilde{Y}|^{\alpha}$$

$$\boxed{\begin{array}{l} (X',Y'),\ (X,Y) \text{ iid copies} \\ \tilde{X} \sim X,\ \tilde{Y} \sim Y \ \oplus\ \tilde{X} \perp\!\!\!\perp \tilde{Y} \end{array}}$$

4° $V^2(X,Y) < \infty$ if $\mathbb{E}|X|^{\alpha}, \mathbb{E}|Y^{\alpha}| < \infty$

$V^2(X,Y) = 0 \iff X \perp\!\!\!\perp Y$

$\exists$ estimator for "$\perp\!\!\!\perp$"

$$V_N^2(\vec{X}, \vec{Y}) = T_1 + T_2 - T_3$$

$$T_1 = \frac{1}{N^2} \sum_{k,l} |x_k - x_l|^{\alpha} |y_k - y_l|^{\alpha}$$

$$T_2 = \frac{1}{N^2} \frac{1}{N^2} \sum_{k,l} |x_k - x_l|^{\alpha} \sum_{k,l} |y_k - y_l|^{\alpha}$$

$$T_3 = \frac{1}{N^3} \sum_{k,l,m} |x_k - x_l|^\alpha \, |y_k - y_m|^\alpha$$

$\exists$ good properties: consistent

+ approx. properties

§ First generalization $|\cdot|^\alpha \rightsquigarrow \psi(\cdot)$

$c_{\alpha,m} |s|^{-\alpha-m}$ Lévy measure

$$|x|^\alpha = \cancel{c_{\alpha,m}} \int (1-\cos xs) \underbrace{\frac{ds \, c_{\alpha,m}}{|s|^{\alpha+m}}}, \quad x \in \mathbb{R}^m$$

$\downarrow$ Lévy - Khintchine

$$\psi(x) = \int (1-\cos xs)\, \mu(ds) + \frac{1}{2} x \, Q x \geq 0$$

$\int s^2 \wedge 1 \, \mu(ds) < \infty$

$m \times m$ matrix pos semidef

- char. exponent of a Lévy process (symm.)
- continuous neg. definite function

  Schoenberg 1938

  - $d_\psi(x,x') := \sqrt{\psi(x-x')}$

    metric in $\mathbb{R}^m$

  - $\|\psi\|_\infty < \infty \iff \mu(\mathbb{R}^m) < \infty$

    $Q \equiv 0$

BB

- $\mathbb{E}\,\psi(X-z) = \mathbb{E}\,\psi(\tilde{X}-z) \quad \forall z$
  $\Longrightarrow X \sim \tilde{X}$ then ψ is "characterizing"
  $\Updownarrow$
  $\operatorname{supp}\mu = \mathbb{R}^m \setminus \{0\}$

<u>Def</u> $X \in \mathbb{R}^m$, $Y \in \mathbb{R}^n$, μ, ν Lévy meas. cndf ψ, φ. Set

$$V^2(X,Y) = \| f_{(X,Y)} - f_X \otimes f_Y \|^2_{L^2(\mu\otimes\nu)} \in [0,\infty]$$

<u>Thm</u> a) $V^2(X,Y) < \infty$ if $\mathbb{E}\,\psi(X), \mathbb{E}\,\varphi(Y) < \infty$

$$\leq 16\, \mathbb{E}\,\psi(X)\, \mathbb{E}\,\varphi(Y)$$

b) $\operatorname{supp}\mu = \mathbb{R}^m \setminus \{0\}$, $\operatorname{supp}\nu = \mathbb{R}^n \setminus \{0\}$ $\Longrightarrow \left[V^2(X,Y) = 0 \iff X \perp\!\!\!\perp Y \right]$

c) $$\boxed{\begin{aligned} V^2(X,Y) = {} & \mathbb{E}\,\psi(X_1 - X_4)\varphi(Y_1 - Y_4) \\ & + \mathbb{E}\,\psi(X_1 - X_4)\, \mathbb{E}\,\varphi(Y_1 - Y_4) \\ & - 2\mathbb{E}\,\psi(X_1 - X_2)\varphi(Y_1 - Y_3) \\ & (X_i, Y_i) \text{ iid copies of } (X,Y) \end{aligned}}$$

here we need finiteness, <u>no</u> "$\infty - \infty$"

§ Estimator

$\vec{X}, \vec{Y}$ empirical r.v's

samples $(x_i, y_i)_{i=1,\dots,N}$

$$V_N^2(\vec{X},\vec{Y}) = \frac{1}{N^2}\sum_{i,k=1}^{N} \psi(x_i - x_k)\,\varphi(y_i - y_k)$$

$$+ \underbrace{\frac{1}{N^2}\sum_{i,j}\psi(x_i - x_j)}_{a_N}\ \underbrace{\frac{1}{N^2}\sum_{k,l}\varphi(y_k - y_l)}_{b_N}$$

$$- \frac{2}{N^3}\sum_{i,j,m}\psi(x_i - x_j)\,\varphi(y_i - y_m)$$

⊕ algebra

$$= c_N \sum_{\underline{ij}} + c_N' \sum_{\underline{ijl}} + c_N'' \sum_{\underline{ijl,k}}$$

distinct

U-statistics!

Lemma $V_N^2(\vec{X},\vec{Y}) = \frac{1}{N^2}\operatorname{trace}(B^T A)$

$$A = C^T a C, \quad B = C^T b C$$

$$a = (-\psi(x_k - x_l))_{k,l}$$
$$b = (-\varphi(y_k - y_l))_{k,l}$$

distance matrices

$$\boxed{C = \begin{pmatrix} 1-1/N & & -1/N \\ & \ddots & \\ -1/N & & 1-1/N \end{pmatrix}}$$

<u>Thm 2</u> $\mathbb{E}\,\psi(X) < \infty, \quad \mathbb{E}\,\varphi(Y) < \infty$

$$V_N^2(\vec{X}, \vec{Y}) \xrightarrow[\text{a.s.}]{N \uparrow \infty} V^2(X, Y)$$

<u>consistency</u>.

<u>Thm 3</u> $\mathbb{E}\,\psi(X) < \infty, \quad \mathbb{E}\,\varphi(Y) < \infty$

$$\mathbb{E}\,\log(1+|X|^2)^{1+\varepsilon} < \infty$$

$$\mathbb{E}\,\log(1+|Y|^2)^{1+\varepsilon} < \infty$$

$$X \perp\!\!\!\perp Y \Rightarrow N\,V_N^2(\vec{X}, \vec{Y}) \xrightarrow{d} \iint |G(s,t)|^2 \mu(ds)\,\lambda(dt)$$

of centered Gaussian field.

<u>Corollary</u> $\mathbb{E}\,\psi(X) < \infty, \quad \mathbb{E}\,\varphi(Y) < \infty$

a) $X \perp\!\!\!\perp Y$ + log moment

$$\Rightarrow \frac{N V_N^2}{a_N b_N} \xrightarrow{d} Q$$

later by BB

b) $X \not\perp\!\!\!\perp Y$: $\dfrac{N V_N^2}{a_N b_N} \xrightarrow{P} \infty$

§ More than 2 r.v. — multivariance

$\rho = \rho_1 \otimes \dots \otimes \rho_n$ $\quad$ n Lévy measures.

in $\mathbb{R}^{d_1}, \dots, \mathbb{R}^{d_n}$, $\quad \psi_1, \dots, \psi_n$

$$M_\rho^2(X_1, \dots, X_n) = \left\| \mathbb{E} \prod_{k=1}^{n} \left(e^{i X_k \cdot} - f_{X_k}(\cdot) \right) \right\|^2_{L^2(\rho)}$$

Issue 1 $\quad M_\rho^2 = 0 \not\Longleftarrow X_1, \dots, X_n$ indep.

Π $X_1 \equiv c$, $X_2, \dots, X_n$ any choice Π

Issue 2 $\quad X_1, \dots, X_n$ indep. $\iff$ $\forall m$ $\forall \{i_1, \dots, i_m\}$, $m < n$

$\underbrace{X_{i_1}, \dots, X_{i_m} \text{ indep.}}_{m\text{-independence}}$

$$\sum_m \binom{n}{m} \approx 2^n \text{ choices!}$$

total distance covariance $\quad ?$

$$M_\rho^2(X_1, \dots, X_n) = \sum_{\substack{1 \le i_1 < \dots < i_m \le n \\ 2 \le m \le n}} M^2_{\bigotimes_{j=1}^{m} \rho_{i_j}}(X_{i_1}, \dots, X_{i_m})$$

$$\prod_{k=1}^{n} (1 + x_k) = 1 + x_1 + \dots + x_n + \sum_{\%} x_{i_1} \cdots x_{i_m}$$

Thm 4 TFAE

- $X_1, \dots, X_n$ indep.
- $\overline{M}_\rho(X_1, \dots, X_n) = 0$

BB 2020

$u = 1$

Thm 5 $\rho_i \rightsquigarrow \psi_i$, $\mathbb{E}\boxed{\psi_i(X_i)} < \infty$

$$\overline{M}^2_\rho(X_1, \dots, X_n) = \mathbb{E} \prod_{i=1}^{n} \left(1 + \Psi_i(X_i, X_i')\right)$$

$\mathbb{E}\left[|X_i|^{\alpha_i \cdot X}\right]$

not $\boxed{S\alpha_i S}$

$$\Psi_i(X_i, X_i') =$$

$$= -\psi(X_i - X_i') + \mathbb{E}\left(\psi(X_i - X_i') \mid X_i\right) + \mathbb{E}\left(\psi(X_i - X_i') \mid X_i'\right) - \mathbb{E}\,\psi(X_i - X_i')$$

X_i not id, any rv

$(X_1, \dots, X_n)$, $(X_1', \dots, X_n')$ iid

$\rightsquigarrow$ $\exists$ consistent estimator

$\exists$ computationally fast

$\exists$ same failure as in $n=2$

$\exists$ detect dependence clusters

i.e.

$(X_1, \dots, X_n)$ dependent

$\rightsquigarrow$ BB 2020.

$$|x_i - x_j|^{\alpha}$$

$$\psi(x_i - x_j)$$