Dependencies in higher dimensions and in complex data structures

Irène Gijbels

KU Leuven
Department of Mathematics
Belgium

September 21–23, 2022

collaborations with Vojtech Kika, Marek Omelka, Charles University Prague

Steven De Keyser, KU Leuven

José Ameijeiras-Alonso, University of Santiago de Compostella, Spain

introduction

example: data on Environmental Quality Index

- available at website of United States Environmental Protection Agency `https://edg.epa.gov/data/Public/ORD/NHEERL/EQI`
- Lobdell *et al.* (2011)
- EQI: variables from five domains – air, water, land, built and sociodemographic
- 3141 observations (representing different locations across the US)

219 variables

consider here variables with more than 90% of unique values

restricting to 3 domains (air, land and water) ...

Table: List of selected variables. Units do not affect analysis and are thus omitted.

Domain	Dimension	Variables
Air	9	Acrolein, Acrylonitrile, Carbon disulfide, Chlorobenzene, Glycol ethers, Methanol, Methyl isobutyl ketone, Polycyclic organic matter/polycyclic aromatic hydrocarbons, Selenium compounds
Water	9	Ammonium, Calcium, Chloride, Magnesium, Nitrate, Potassium, Sodium, Sulfate (all in precipitation), Mercury (deposited)
Land	4	Lead, Zinc, Copper, Herbicides
Combined	22	All the above variables

bivariate association measures and copulas

standard measure of strength of pairwise dependence:

Pearson's correlation coefficient

$$\rho^{(\mathsf{P})}(X_1, X_2) = \frac{\mathsf{Cov}(X_1, X_2)}{\sqrt{\mathsf{Var}(X_1)\mathsf{Var}(X_2)}} = \frac{\mathsf{E}(X_1X_2) - \mathsf{E}(X_1)\,\mathsf{E}(X_2)}{\sqrt{\mathsf{Var}(X_1)\mathsf{Var}(X_2)}}$$

if (X_1, X_2) has a bivariate normal distribution function, then $\rho^{(\mathsf{P})}(X_1, X_2)$ completely characterizes the dependence structure

needs assumption of finite variances

$\Longrightarrow$ need for other association measures

Spearman's rho

Kendall's tau

...

many association measures for measuring **strength of dependency** between X_1 and X_2:

- **Kendall's tau**

with (Y_1, Y_2) an independent copy of (X_1, X_2)

$$\tau(X_1, X_2) = \Pr\left[(X_1 - Y_1)(X_2 - Y_2) > 0\right] - \Pr\left[(X_1 - Y_1)(X_2 - Y_2) < 0\right]$$

- **Spearman's rho**

with (Y_1, Y_2) and (Z_1, Z_2) two independent copies of (X_1, X_2)

$$\begin{aligned}\rho^{(\mathsf{S})}(X_1, X_2) &= 3\left\{\Pr\left[(X_1 - Y_1)(X_2 - Z_2) > 0\right] - \Pr\left[(X_1 - Y_1)(X_2 - Z_2) < 0\right]\right\} \\ &= \frac{\mathsf{Cov}\big(F_1(X_1), F_2(X_2)\big)}{\sqrt{\mathsf{Var}(F_1(X_1))\mathsf{Var}(F_2(X_2))}} = \rho^{(\mathsf{P})}\big(\underbrace{F_1(X_1)}_{=U_1}, \underbrace{F_2(X_2)}_{=U_2}\big)\end{aligned}$$

- **Gini's gamma**
- **Blomqvist beta**

.....

what about the dependence structure?

copula function

joint cumulative distribution function of (X_1, X_2) :

$$F(x_1, x_2) = \Pr(X_1 \leq x_1, X_2 \leq x_2) \qquad (x_1, x_2) \in \mathbb{R}^2$$

marginal cumulative distribution functions :
$F_1(x_1) = P(X_1 \leq x_1) \qquad F_2(x_2) = P(X_2 \leq X_2)$

Sklar's theorem (Sklar (1959)); $\exists$ function $C : [0,1]^2 \to [0,1]$ such that

$$F(x_1, x_2) = C(F_1(x_1), F_2(x_2)) \qquad (x_1, x_2) \in \mathbb{R}^2$$

if $F_1(x)$ and $F_2(x)$ are continuous in x, then C is **unique**

C fully describes the dependence structure between X_1 and X_2

C is joint distribution function of $U_1 = F_1(X_1)$ and $U_2 = F_2(X_2)$

copula and association measures

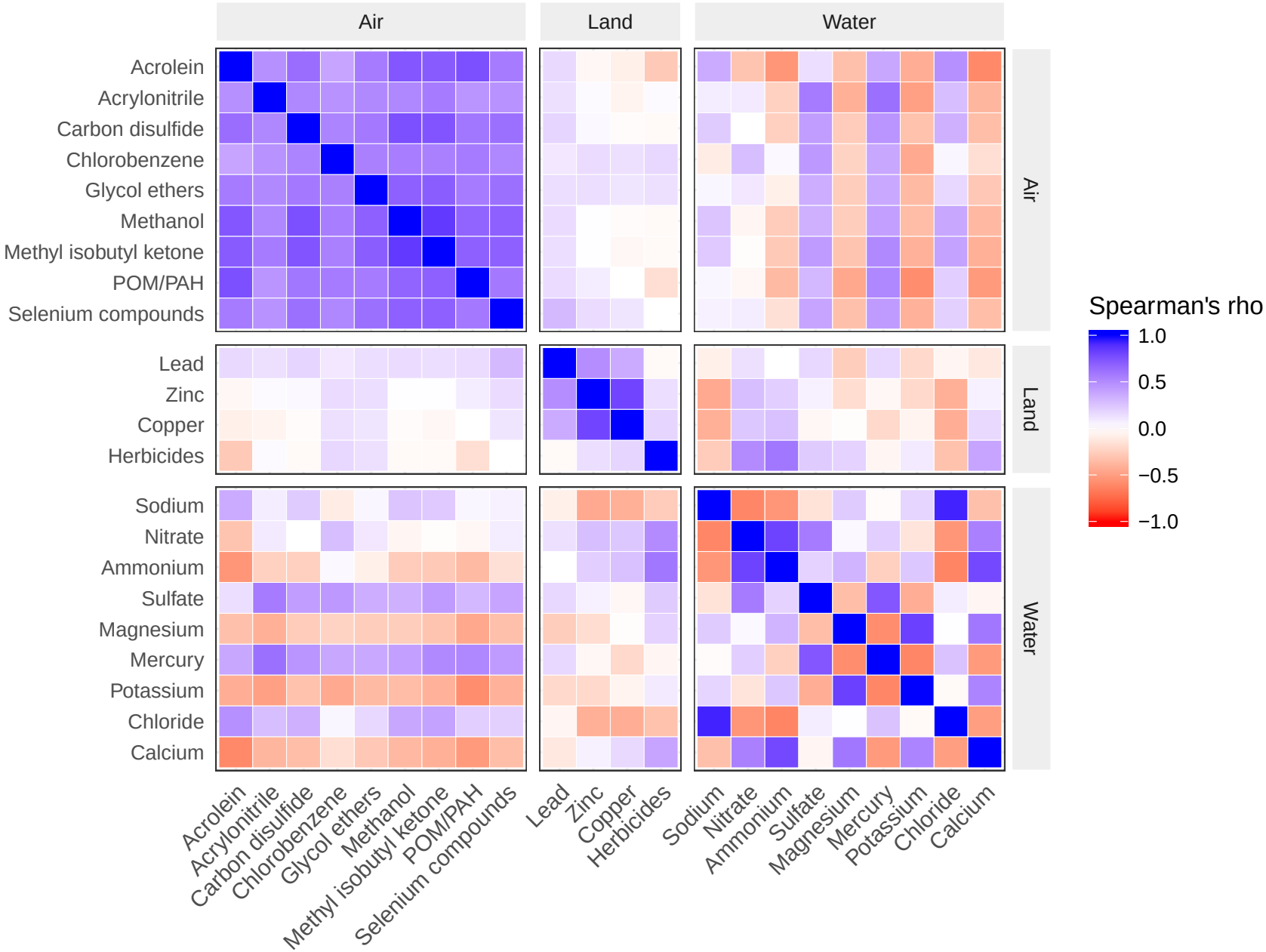
copula	association measure	
C	Kendall's tau	$\tau = 4 \iint C(u_1, u_2)\, dC(u_1, u_2) - 1$
	Spearman's rho	$\rho = 12 \iint C(u_1, u_2)\, du_1\, du_2 - 3$
	Blomqvist's beta	$\beta = 4\, C(0.5, 0.5) - 1$
	Gini's index	$\gamma = 4 \left[\int_0^1 C(u, 1-u) du - \int_0^1 [u - C(u, u)] du \right]$
	...	...

any of these association measures:

functional of the bivariate copula C $(= C_2)$

$\kappa_2(C_2)$

example: heatmap of the empirical pairwise Spearman's rho values



multivariate association measures

d-variate random vector $\mathbf{X} = (X_1, \ldots, X_d)^\top$

aim: how to measure association between the d components ?
which properties ?

what happens if d increases, and eventually tends to infinity ?

F: joint cumulative distribution function of $\mathbf{X}$

$F_1, \ldots, F_d$: continuous marginal distribution functions of $X_1, \ldots, X_d$

applying Sklar's theorem: $\exists$ function $C_d : [0,1]^d \to [0,1]$ such that

$$F(x_1, \ldots, x_d) = C_d(F_1(x_1), \ldots, F_d(x_d)), \quad (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$$

C_d joint cumul. distrib. funct. of $\boldsymbol{U} = (U_1, \ldots U_d)^\top = (F_1(X_1), \ldots, F_d(X_d))^\top$:

$$C_d(\boldsymbol{u}) = \Pr(\boldsymbol{U} \leq \boldsymbol{u}) \qquad \boldsymbol{u} = (u_1, \ldots, u_d)^\top \in [0,1]^d$$

multivariate association measure:

$$\begin{array}{rccl} \kappa_d: & \mathsf{Cop}(d) & \longrightarrow & \mathbb{R} \\ & C_d & \longrightarrow & \kappa_d(C_d) \end{array} \qquad \mathsf{Cop}(d) = \text{set of all } d\text{-variate copulas}$$

what do we expect as properties of such a function ?

some concepts and notations: for a copula C_d of $\boldsymbol{U}$, define

- **survival function** $\overline{C}_d(\boldsymbol{u}) = \Pr(\boldsymbol{U} > \boldsymbol{u})$
- the **survival copula** C_d^S is defined as the copula of $\mathbf{1} - \boldsymbol{U}$, that is

$$C_d^S(\boldsymbol{u}) = \Pr(\mathbf{1} - \boldsymbol{U} \leq \boldsymbol{u}) = \Pr(\boldsymbol{U} > \mathbf{1} - \boldsymbol{u}) = \overline{C}_d(\mathbf{1} - \boldsymbol{u})$$

- **concordance ordering** for copulas $A_d, B_d \in \mathsf{Cop}(d)$:

$$A_d \preceq_C B_d \iff \forall \boldsymbol{u} \in [0,1]^d : A_d(\boldsymbol{u}) \leq B_d(\boldsymbol{u}) \text{ and } \overline{A}_d(\boldsymbol{u}) \leq \overline{B}_d(\boldsymbol{u})$$

- **reflections** of the d-dimensional unit cube $[0,1]^d$:

  a mapping $\xi : [0,1]^d \to [0,1]^d$ is a reflection if $\xi(\boldsymbol{u}) = \boldsymbol{v}$ where for $i \in \{1, \ldots, d\}$ we have $v_i = u_i$ or $v_i = 1 - u_i$

  set of all reflections: $\mathcal{R}_d$

  for any reflection $\xi \in \mathcal{R}_d$, we can define a new copula C_d^{ξ}

  - reflection with respect to all components:

    will be denoted as σ, that is $\sigma(\boldsymbol{u}) = \mathbf{1} - \boldsymbol{u}$

    if C_d is copula of $(X_1, \ldots, X_d)$, then C_d^{σ} is copula of $(-X_1, \ldots, -X_d)$

axioms for multivariate association measures

Rényi (1952), Scarsini (1984), Taylor (2007), Schmid it et al. (2010), ...

(A_1) (Normalization) $\kappa_d(M_d) = 1, \kappa_d(\Pi_d) = 0$.

(A_2) (Continuity) If $\lim_{m\to\infty} C_{d,m}(\boldsymbol{u}) = C_d(\boldsymbol{u}), \forall \boldsymbol{u} \in [0,1]^d$, then $\lim_{m\to\infty} \kappa_d(C_{d,m}) = \kappa_d(C_d)$.

(A_3) (Permutation invariance) $\kappa_d(C_d^\pi) = \kappa_d(C_d)$ for every permutation π.

(A_4) (Ordering) If $C_{d,1} \preceq_C C_{d,2}$, then $\kappa_d(C_{d,1}) \leq \kappa_d(C_{d,2})$.

(A_5) (Duality) $\kappa_d(C_d^\sigma) = \kappa_d(C_d)$.

(A_6) (Reflection principle) $\sum_{\xi\in\mathcal{R}_d} \kappa_d(C_d^\xi) = 0$.

(A_7) (Transition property) There exists a constant r_{d-1} such that

$$\kappa_d(C_d) + \kappa_d(C_d^{\sigma_1}) = r_{d-1}\kappa_{d-1}(C_{d-1}^{(-1)})$$

(A_8) (Independent component addition) For X_{d+1} independent of $(X_1, \ldots, X_d)^\top$

$$\kappa_d(C_d) > \kappa_{d+1}(C_{d+1}) > 0, \quad \text{or} \quad \kappa_d(C_d) < \kappa_{d+1}(C_{d+1}) < 0,$$
$$\text{or} \quad \kappa_d(C_d) = \kappa_{d+1}(C_{d+1}) = 0.$$

how to get to multivariate association meaures ?

does an association measure satisfy these expected, minimal, properties ?

which other properties do they exhibit?

two main approaches towards getting to multivariate association measures:

- pairwise approach: rely on looking at values of bivariate association measures
- copula approach: generalize the bivariate structure to multivariate setting

pairwise approach

- κ_2: bivariate association measure
- create a d-variate association measure as an average of all pairwise measures, i.e.

$$\kappa_d^{\mathsf{PW}}(C_d) = \frac{1}{\binom{d}{2}} \sum_{1 \le i < j \le d} \kappa_2(C_2^{i,j}) \qquad \text{with } C_2^{i,j} \text{ copula of } (X_i, X_j)^\top$$

- does such an approach fullfills the desired properties?

Proposition

Let $\kappa_2 : \mathrm{Cop}(2) \to \mathbb{R}$ be a bivariate measure of association
Define for $d \in \{2, 3, \dots\}$, the measure $\kappa_d^{\mathsf{PW}} : \mathrm{Cop}(d) \to \mathbb{R}$, and set $r_{d-1} = 2(d-2)/d$ in axiom (A_7).
Then $\{(\kappa_d^{\mathsf{PW}}, r_d)\}_{d=3}^{\infty}$ fulfils axioms (A_1) to (A_8).

pairwise approach leads to a 'reasonable' multivariate association measure if the initial bivariate measure is 'reasonable' ...

but : pairwise-constructed measures always assign value 0 to vectors having pairwise independent components no matter if there is an association of higher order ...

copula approach

• multivariate Spearmans's rho

- recall bivariate Spearman's rho:

$$\rho(X_1, X_2) = \frac{\mathsf{cov}(U_1, U_2)}{\sqrt{\mathsf{Var}(U_1)}\sqrt{\mathsf{Var}(U_2)}}$$

- this can be expressed as

$$\begin{aligned}
\rho(C_2) &= \frac{\int_{[0,1]^2} u_1 u_2 \mathrm{d}C_2(u_1, u_2) - (1/2)^2}{\sqrt{(1/12)}\sqrt{(1/12)}} \\
&= \frac{\int_{[0,1]^2} \Pi_2(\boldsymbol{u})\mathrm{d}C_2(\boldsymbol{u}) - \int_{[0,1]^2} \Pi_2(\boldsymbol{u})\mathrm{d}\Pi_2(\boldsymbol{u})}{\underbrace{\int_{[0,1]^2} \Pi_2(\boldsymbol{u})\mathrm{d}M_2(\boldsymbol{u})}_{=1/3} - \underbrace{\int_{[0,1]^2} \Pi_2(\boldsymbol{u})\mathrm{d}\Pi_2(\boldsymbol{u})}_{=1/4}} \qquad \boldsymbol{u} = (u_1, u_2)^\top
\end{aligned}$$

where

$\Pi_2(\boldsymbol{u}) = u_1 u_2$ independence copula

$M_2(\boldsymbol{u}) = \min(u_1, u_2)$ comonotonicity copula, Fréchet upper bound

- a first multivariate generalization (Wolff (1980)) is

$$\begin{aligned}\rho_1(C_d) &= \frac{\int_{[0,1]^d} C_d(\boldsymbol{u})\mathrm{d}\boldsymbol{u} - \int_{[0,1]^d} \Pi_d(\boldsymbol{u})\mathrm{d}\boldsymbol{u}}{\underbrace{\int_{[0,1]^d} M_d(\boldsymbol{u})\mathrm{d}\boldsymbol{u}}_{=1/(d+1)} - \underbrace{\int_{[0,1]^d} \Pi_d(\boldsymbol{u})\mathrm{d}\boldsymbol{u}}_{=1/2^d}} \\ &= h_\rho(d)\left\{2^d \int_{[0,1]^d} C_d(\boldsymbol{u})\mathrm{d}\boldsymbol{u} - 1\right\}\end{aligned}$$

with $h_\rho(d) = (d+1)/\{2^d - (d+1)\}$

- or start from alternative expression in bivariate case, obtained after integration by parts (Schmid *et al.* (2007))

$$\rho(C_2) = \frac{\int_{[0,1]^2} C_2(\boldsymbol{u})\mathrm{d}\Pi_2(\boldsymbol{u}) - \int_{[0,1]^2} \Pi_2(\boldsymbol{u})\mathrm{d}\Pi_2(\boldsymbol{u})}{\int_{[0,1]^2} M_2(\boldsymbol{u})\mathrm{d}\Pi_2(\boldsymbol{u}) - \int_{[0,1]^2} \Pi_2(\boldsymbol{u})\mathrm{d}\Pi_2(\boldsymbol{u})} = \frac{\int [C_2(\boldsymbol{u}) - \Pi_2(\boldsymbol{u})]\,\mathrm{d}\Pi_2(\boldsymbol{u})}{\int [M_2(\boldsymbol{u}) - \Pi_2(\boldsymbol{u})]\,\mathrm{d}\Pi_2(\boldsymbol{u})}$$

standardized average distance between C_2 and Π_2

- leads to an alternative multivariate generalization (Joe (1990), ...)
- two generalizations of bivariate Spearman's rho:

$$\rho_1(C_d) = h_\rho(d)\left\{2^d \int_{[0,1]^d} C_d(\boldsymbol{u})\mathrm{d}\boldsymbol{u} - 1\right\}$$

$$\rho_2(C_d) = h_\rho(d)\left\{2^d \int_{[0,1]^d} \Pi_d(\boldsymbol{u})\mathrm{d}C_d(\boldsymbol{u}) - 1\right\}$$

which of these generalizations is preferable ?

- both ρ_1 and ρ_2 satisfy $(A_1) - (A_7)$
  except for the duality axiom (A_5) (see Schmid *et al.* (2010))

  (A_8) is satisfied for ρ_1 and ρ_2 (G. Kika & Omelka (2021))

multivariate Kendall's tau

- bivariate Kendall's tau:

$$\tau(X_1, X_2) = \Pr\{(X_1 - Y_1)(X_2 - Y_2) > 0\} - \Pr\{(X_1 - Y_1)(X_2 - Y_2) < 0\}$$

  where (X_1, X_2) has copula C_2

- can be expressed as

$$\tau(C_2) = 4 \int_{[0,1]^2} C_2(\boldsymbol{u}) \mathrm{d}C_2(\boldsymbol{u}) - 1$$

- can be generalized as (Nelsen (1996), ...)

$$\tau(C_d) = \frac{1}{2^{d-1} - 1} \left\{ 2^d \int_{[0,1]^d} C_d(\boldsymbol{u}) \mathrm{d}C_d(\boldsymbol{u}) - 1 \right\}$$

$\tau(C_d)$ satisfies $(A_1) - (A_8)$

multivariate Gini's gamma

- bivariate Gini's gamma

$$\gamma(C_2) = 2 \int_{[0,1]^2} (M_2(\boldsymbol{u}) + W_2(\boldsymbol{u}) + \overline{M}_2(\boldsymbol{u}) + \overline{W}_2(\boldsymbol{u}))\mathrm{d}C_2(\boldsymbol{u}) - 2$$

$M_2(u_1, u_2) = \min(u_1, u_2)$	comonotonicity copula	Fréchet upper bound
$W_2(u_1, u_2) = \max(u_1 + u_2 - 1, 0)$	countermonotonicity copula	Fréchet lower bound

$$W_2 \leq C \leq M_2$$

- first multivariate generalization (Behboodian *et al.* (2007))

$$\gamma_1(C_d) = \frac{1}{b(d) - a(d)} \left(\int_{[0,1]^d} (M_d(\boldsymbol{u}) + W_d(\boldsymbol{u}) + \overline{M}_d(\boldsymbol{u}) + \overline{W}_d(\boldsymbol{u}))\mathrm{d}C_d(\boldsymbol{u}) - a(d) \right)$$

where

$$a(d) = \frac{2}{d+1} + \frac{1}{(d+1)!} + \sum_{j=0}^{d} (-1)^j \binom{d}{j} \frac{1}{(j+1)!} \qquad b(d) = 2 - \sum_{j=1}^{d-1} \frac{1}{2j}$$

- alternative expression for bivariate Gini's gamma:

$$\gamma(C_2) = 8 \int_{[0,1]^2} C_2(\boldsymbol{u}) \mathrm{d} \left(\frac{M_2(\boldsymbol{u}) + W_2(\boldsymbol{u})}{2} \right) - 2$$

- leads to multivariate generalization (Taylor(2007))

$$\gamma_2(C_d) = \frac{2^d}{2^{d-1}-1} \left(\int_{[0,1]^d} (C_d(\boldsymbol{u}) + C_d^S(\boldsymbol{u})) \mathrm{d} \left(\frac{1}{2^d} \sum_{\xi \in \mathcal{R}_d} M_d^{\xi}(\boldsymbol{u}) \right) - \frac{1}{2^{d-1}} \right)$$

- can conveniently be expressed in terms of reflections

$$\gamma_2(C_d) = \frac{1}{2^{d-1}-1} \left(\sum_{\xi \in \mathcal{R}_d} \int_0^1 (C_d(\xi(u,\ldots,u)) + \overline{C}_d(\xi(u,\ldots,u))) \mathrm{d}u - 2 \right)$$

which involves only calculation of one-dimensional integrals!

which of these two generalizations is preferable ?

- two generalizations of bivariate Gini's gamma:

$$\gamma_1(C_d) = \frac{1}{b(d) - a(d)} \left(\int_{[0,1]^d} (M_d(\boldsymbol{u}) + W_d(\boldsymbol{u}) + \overline{M}_d(\boldsymbol{u}) + \overline{W}_d(\boldsymbol{u})) \mathrm{d}C_d(\boldsymbol{u}) - a(d) \right)$$

$$\gamma_2(C_d) = \frac{1}{2^{d-1} - 1} \left(\sum_{\xi \in \mathcal{R}_d} \int_0^1 (C_d(\xi(u, \ldots, u)) + \overline{C}_d(\xi(u, \ldots, u))) \mathrm{d}u - 2 \right)$$

- has been shown that (G., Kika & Omelka (2021))
  - γ_1: satisfies $(A_1) - (A_3)$

    does NOT satisfy duality axiom (A_5) and (A_8)
  - γ_2: satisfies $(A_1) - (A_8)$

    simpler expression (only one-dimensional integrals)

---

bivariate $\Longrightarrow$ multivariate association measures

not so straightforward; which version matters ...

multivariate association measures in increasing dimension

what happens with association measures in case of growing dimension ?

- difficult to study in full generality
- study in 2 settings: Archimedean copulas
  meta-elliptical copulas

Archimedian copulas:

* C_d is a d-dimensional Archimedean copula if for any $\boldsymbol{u} \in [0,1]^d$ it permits the representation

$$C_d(\boldsymbol{u}) = \psi\left[\psi^{-1}(u_1) + \cdots + \psi^{-1}(u_d)\right]$$

for some Archimedean generator ψ

$\psi : [0,\infty) \to [0,1]$ a nonincreasing and continuous function
satisfying $\psi(0) = 1$, $\lim_{x\to\infty} \psi(x) = 0$
is strictly decreasing on $[0, \inf\{x : \psi(x) = 0\})$

• Archimedian copulas and multivariate association measures in increasing dimension

- **Spearman's rho** in increasing dimension Wysocki (2015) showed

$$\lim_{d\to\infty} \rho_1(C_d) = c_1 \in [0,1] \quad \Longleftrightarrow \quad \lim_{d\to\infty} (d+1) \int_{[0,1]^d} C_d(\boldsymbol{u}) \mathrm{d}\boldsymbol{u} = c_1$$

$$\lim_{d\to\infty} \rho_2(C_d) = c_2 \in [0,1] \quad \Longleftrightarrow \quad \lim_{d\to\infty} (d+1) \int_{[0,1]^d} \Pi_d(\boldsymbol{u}) \mathrm{d}C_d(\boldsymbol{u}) = c_2$$

  unknown: whether strict positive c_1 and c_2 are possible (achievable)

- **Kendall's tau** in increasing dimension Wysocki (2015) showed

$$\lim_{d\to\infty} \tau(C_d) = 0$$

  or Archimedean copulas are not able to carry any association (measured with Kendall's tau) in very high dimension

- **Gini's gamma** in increasing dimension
  - ◇ established explicit expression for $\gamma_2(\cdot)$ in terms of generator
  - ◇ shown that (Kika, G. & Omelka (2021))

$$\lim_{d\to\infty} \gamma_2(C_d) = 0$$

• meta-elliptical copulas and multivariate association measures in increasing dimension

meta-elliptical copulas are copulas with elliptical contours

- Genest *et al.* (2011): all meta-elliptical copulas with the same correlation matrix also share the same Kendall's tau value

  Kendall's tau depends only on the correlation matrix

- more precisely: C_d copula of $\boldsymbol{U} = (U_1, \ldots U_d)^\top$

  meta-elliptical copula with correlation matrix $\boldsymbol{R} = (\varrho_{i,j})$, with $\varrho_{i,j} \in [-1, 1]$; then

$$\tau(C_d) = \frac{1}{2^{d-1} - 1}\{-1 + 2^d \Pr(\boldsymbol{Z} \geq \boldsymbol{0})\}$$

  $\boldsymbol{Z}$ is d-variate mean zero normal random vector with correlation matrix $\boldsymbol{R}$

- what happens if d increases ?
- so far: answer is known only for some specific correlation structures

- correlation matrix $\boldsymbol{R}$ of the form

$$\varrho_{i,j} = \lambda_i \lambda_j, \quad i \neq j \qquad \lambda_j \in [-1, 1]$$

covers an equicorrelated correlation matrix

for which $\varrho_{i,j} = \varrho$ for all $i \neq j$, with $\varrho \in (-1/(d-1), 1)$

- correlation matrix $\boldsymbol{R}$ of a banded type and $(m+2)$-diagonal (with $m \in \mathbb{Z}_{>0}$), and takes the form

$$\varrho_{i,j} = \begin{cases} 1, & \text{if } i = j \\ c_{i,j} & \text{if } |i-j| \in \{1, \cdots, m\} \\ 0, & \text{if } |i-j| > m \end{cases}$$

where $c_{i,j} = c_{j,i} \in [-1, 1]$ are constants, not all zero, and such that $\boldsymbol{R}$ is a correlation matrix

$$\begin{pmatrix} 1 & * & 0 & 0 & \dots & 0 \\ * & 1 & * & 0 & \dots & 0 \\ 0 & * & 1 & * & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & * & 1 & * & 0 \\ 0 & \dots & 0 & * & 1 & * \\ 0 & \dots & 0 & 0 & * & 1 \end{pmatrix}$$

case $(m = 1)$: tridiagonal matrix; only non-zero values on main diagonal and 2 adjacent diagonals

for both correlation structures Kendall's tau tends to zero as d tends to infinity

Theorem

Let $\{C_d\}$ be a sequence of d-dimensional meta-elliptical copulas with a correlation matrix $\boldsymbol{R} = (\varrho_{i,j})$

(i) first correlation structure: the λ_j satisfy the assumption that there exists $\lambda_0 < 1$ such that $\lambda_j \leq \lambda_0$ for all $j \in \mathbb{Z}_{>0}$

or

(ii) second correlation structure

in both cases, it holds that $\lim_{d\to\infty} \tau(C_d) = 0$

what can happen in case of other correlation structures?

nonparametric estimation of multivariate association measures

$\boldsymbol{X}$ with copula C_d **unknown** (C_d joint distribution of $(\underbrace{F_1(X_1)}_{=U_1}, \ldots, \underbrace{F_d(X_d)}_{=U_d})$)

marginal distributions $F_1, \cdots, F_d$ **unknown**

$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ random sample from $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$

where $\boldsymbol{X}_i = (X_{1,i}, \ldots X_{d,i})^\top$ for $i \in \{1, \ldots n\}$

nonparametric estimation of multivariate association measures

- empirical marginal distribution function F_j for the j-th component

$$\widehat{F}_{j,n}(x) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{1}(X_{j,i} \leq x)$$

  lead to pseudo-observations $\widehat{U}_{j,i} = \widehat{F}_{j,n}(X_{j,i})$ from $U_j = F_j(X_j)$
- empirical copula (recall: $C(\boldsymbol{u}) = \Pr(\boldsymbol{U} \leq \boldsymbol{u})$)

$$\widehat{C}_n(\boldsymbol{u}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(\widehat{U}_{1,i} \leq u_1, \ldots, \widehat{U}_{d,i} \leq u_d)$$

- and empirical survival function (recall: $\overline{C}_d(\boldsymbol{u}) = \Pr(\boldsymbol{U} > \boldsymbol{u})$)

$$\widehat{\overline{C}}_n(\boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\widehat{U}_{1,i} > u_1, \ldots, \widehat{U}_{d,i} > u_d)$$

- nonparametric estimator of an association measure $\kappa(C)$

$$\widehat{\kappa}_n = \kappa(\widehat{C}_n)$$

quality of estimator $\widehat{\kappa}_n$ depends on that of $\widehat{C}_n$ (for the copula C_d)

uniform convergence of empirical copula process
(Tsukahara (2005), Schmid *et al.* (2007), Segers (2012), ...)

joint weak convergence of empirical copula process and empirical survival copula process

- nonparametric estimators for Spearman's rho (Schmid *et al.* (2007))
- nonparametric estimator for Kendall's tau (Genest *et al.* (2011)):

$$\widehat{\tau}_n = \frac{1}{2^{d-1}-1}\left\{\frac{2^d}{n(n-1)}\left[\sum_{i\neq j}\mathbb{1}\left(\boldsymbol{X}_i \leq \boldsymbol{X}_j\right)\right]-1\right\}$$

asymptotic normality result: as $n \to \infty$

$$\sqrt{n}(\widehat{\tau}_n - \tau(C)) \xrightarrow{D} \mathcal{N}(0, \sigma_\tau^2)$$

where $$\sigma_\tau^2 = \left(\frac{2^d}{2^{d-1}-1}\right)^2 \mathrm{Var}(C(\boldsymbol{U}) + \overline{C}(\boldsymbol{U}))$$

- nonparametric estimator for Gini's gamma (Kika, G. & Omelka (2021)):

$$\sqrt{n}\left(\widehat{\gamma}_{2n} - \gamma_2(C)\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma_{\gamma_2}^2\right) \qquad \text{with } \sigma_{\gamma_2}^2 = ...$$

standard errors of the estimators

- get approximate standard errors by bootstrap procedures
- rely on asymptotic normality results, but focus on the main terms in the i.i.d. representations ...
- for example for CI for the multivariate Kendall's tau:
  - variance $\mathrm{Var}(C(\boldsymbol{U}) + \overline{C}(\boldsymbol{U}))$ can be estimated, using ideas from U-statistics, by the sample variance of $W_1, \ldots, W_n$, i.e.

$$\widehat{\sigma}_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (W_i - \overline{W}_n)^2, \quad \text{where} \quad \overline{W}_n = \frac{1}{n} \sum_{i=1}^{n} W_i$$

with

$$W_i = \frac{1}{n-1} \sum_{j=1, j \neq i}^{n} \mathbb{1}(\boldsymbol{X}_j \leq \boldsymbol{X}_i) + \frac{1}{n-1} \sum_{j=1, j \neq i}^{n} \mathbb{1}(\boldsymbol{X}_j \geq \boldsymbol{X}_i)$$

  - the standard error of $\widehat{\tau}_n$ is estimated by $\frac{2^d}{2^{d-1}-1} \frac{\widehat{\sigma}_n}{\sqrt{n}}$

real data application: data on Environmental Quality Index (EQI)

Table: Estimated multivariate association measures for EQI dataset (bootstrap SE).

Domain	d	$\widehat{\rho}_{3n}$	$\widehat{\tau}_n$	$\widehat{\gamma}_{2n}$
Air	9	0.44 (0.008)	0.29 (0.007)	0.29 (0.006)
Water	9	0.04 (0.004)	0.03 (0.002)	0.02 (0.002)
Land	4	0.30 (0.011)	0.21 (0.008)	0.23 (0.009)
Combined	22	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)
AirPlus	12	**0.23** (0.008)	**0.15** (0.005)	**0.14** (0.006)

Domain	d	$\widehat{\rho}_n^{\mathrm{PW}}$	$\widehat{\tau}_n^{\mathrm{PW}}$	$\widehat{\gamma}_n^{\mathrm{PW}}$
Air	9	0.60 (0.007)	0.44 (0.006)	0.49 (0.006)
Water	9	0.05 (0.005)	0.04 (0.004)	0.03 (0.004)
Land	4	0.33 (0.011)	0.24 (0.008)	0.27 (0.009)
Combined	22	0.11 (0.003)	0.09 (0.002)	0.09 (0.002)
AirPlus	12	**0.50** (0.007)	**0.36** (0.005)	**0.40** (0.006)

AirPlus: all 9 variables from the air domain, extended with 3 variables from the water domain (Sulfate, Mercury and Chloride)

multivariate tail coefficients

- bivariate lower and upper tail coefficients:

$$\lambda_L(C_2) = \lim_{u \searrow 0} \mathsf{P}(U_2 \leq u | U_1 \leq u) = \lim_{u \searrow 0} \mathsf{P}(U_1 \leq u | U_2 \leq u) = \lim_{u \searrow 0} \frac{C_2(u,u)}{u}$$

$$\lambda_U(C_2) = \lim_{u \nearrow 1} \mathsf{P}(U_2 > u | U_1 > u) = \lim_{u \nearrow 1} \mathsf{P}(U_1 > u | U_2 > u) = \lim_{u \nearrow 1} \frac{1 - 2u + C_2(u,u)}{1-u}$$

- multivariate tail coefficients in the literature ?
  - Frahm's extremal dependence coefficient (Frahm (2006)):
    with $U_{\max} = \max(U_1, \ldots, U_d)$ and $U_{\min} = \min(U_1, \ldots, U_d)$

$$\epsilon_L(C_d) = \lim_{u \searrow 0} \mathsf{P}(U_{\max} \leq u | U_{\min} \leq u) = \lim_{u \searrow 0} \frac{C_d(u, \ldots, u)}{1 - \overline{C}_d(u, \ldots, u)}$$

$$\epsilon_U(C_d) = \lim_{u \nearrow 1} \mathsf{P}(U_{\min} > u | U_{\max} > u) = \lim_{u \nearrow 1} \frac{\overline{C}_d(u, \ldots, u)}{1 - C_d(u, \ldots, u)}$$

    coefficients are not equal to λ_L, λ_U respectively in the bivariate case:

$$\epsilon_L(C_2) = \frac{\lambda_L(C_2)}{2 - \lambda_L(C_2)} \qquad \epsilon_U(C_2) = \frac{\lambda_U(C_2)}{2 - \lambda_U(C_2)}$$

different type of tail dependence coefficient ...

- Li's tail dependence parameter (Li (2009), De Luca & Rivieccio (2012),...)

  consider a subset of indices: $\emptyset \neq I_h \subset \{1, \ldots, d\}$ such that $|I_h| = h$ and $J_{d-h} = \{1, \ldots, d\} \backslash I_h$

  lower and upper tail dependence parameters:

$$\lambda_L^{I_h|J_{d-h}}(C_d) = \lim_{u \searrow 0} \mathrm{P}(U_i \leq u, \forall i \in I_h | U_j \leq u, \forall j \in J_{d-h})$$

$$\lambda_U^{I_h|J_{d-h}}(C_d) = \lim_{u \nearrow 1} \mathrm{P}(U_i > u, \forall i \in I_h | U_j > u, \forall j \in J_{d-h})$$

  heavily depends on the choice of the set I_h

  includes the bivariate tail coefficients as special cases

  ($h = 1$, $I_1 = \{1\}$ and $J_1 = \{2\}$)

- Schmid's and Schmidt's tail dependence measure (Schmid & Schmidt (2007))

- ....

additional study:

- in case of extreme value copulas: tail dependence measures in terms of Pickands dependence function
- tail dependence using subvectors (compare with pairwise approach)
- set of desirable for multivariate tail coefficients
- investigate the properties of the tail coefficients, starting from these
- how do they behave when d increases ?

..., Kika, G. & Omelka (2020), ...

- how to estimate them nonparametrically ?

let's have a look at one multivariate tail coefficient ....

- problem when estimating tail coefficients: need to deal with the limits
- Frahm's lower tail and upper tail dependence

$$\epsilon_L(C) = \lim_{u \searrow 0} \frac{C_d(u, \ldots, u)}{1 - \overline{C}_d(u, \ldots, u)}$$

- consider $u_n > 0$
  nonparametric estimator of ϵ_L is

$$\widehat{\epsilon}_L(u_n) = \frac{\widehat{C}_n(u_n, \ldots, u_n)}{1 - \widehat{\overline{C}}_n(u_n, \ldots, u_n)}$$

  assumptions on sequence: such that $u_n \to 0$ and $nu_n \to \infty$, as $n \to \infty$
- in univariate setting: nu_n can be interpreted as number of extreme observations
- **how to choose u_n ?** (G. Kika & Omelka (2022))

application in variable clustering

suppose d objects, denoted $X_1, \ldots, X_d$

n observations on these objects

aim: find clusters in the d objects

- using agglomerative clustering methods
- see e.g. Di Lascio, Durante & Pappadà (2017), ...
  - start: each variable is one cluster (of size 1)
  - in each step: two clusters are merged together
  - example of approach: using linkage method
    - for each pair (X_j, X_k) calculate a bivariate similarity measure s_{jk}
    - translate this into dissimilarity measures: $g : [-1.1] \to [0, +\infty)$ decreasing function with $g(1) = 0$

$$\text{diss}_{j,k} = g(s_{j,k})$$

examples: $g(x) = \sqrt{1-x}$, or $g(x) = \sqrt{1-x^2}$

    - linkage method: describes how dissimilarity between 2 clusters can be based on dissimilarities among elements of the cluster

approach using multivariate dissimilarities: see Fuchs, Lascio & Durante (2021) ....

using multivariate dissimilarities ....

- two vectors $\boldsymbol{Y}$ and $\underline{Z}$, of dimensions d_1 and d_2
  consider $(\boldsymbol{Y}^\top, \boldsymbol{Z}^\top)^\top$, of dimension say d, with copula C_d
- example of dissimilarity function

$$\text{diss}_\tau(C_d) = \left(\frac{2^{d-1}-1}{2^d}\right)(1-\tau(C_d))$$

- (iterative) clustering procedure
  - ⋆ suppose r clusters in the current iteration
  - ⋆ for each pair of clusters $\boldsymbol{M}_k$ and $\boldsymbol{M}_\ell$, $k, \ell = 1, \ldots, r$, $k \neq \ell$, calculate dissimilarity
  - ⋆ merge 2 clusters with smallest dissimilarity measure
  - ⋆ continue to do this until all variables are in one cluster
- stopping rules can be considered, e.g. impose an upper bound for dissimilarity (no longer merge when this is exceeded)

example: data on Environmental Quality Index

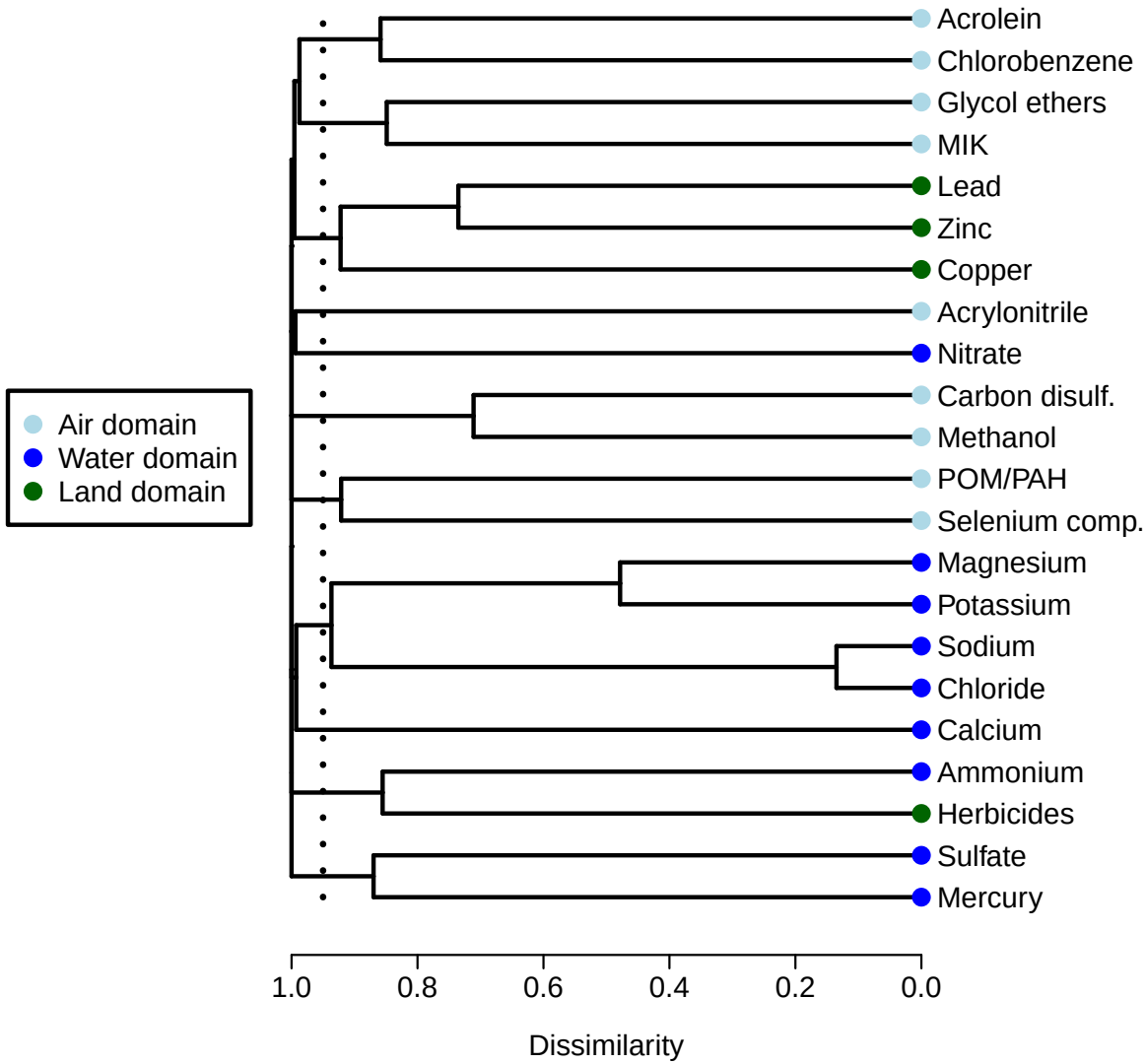


Figure: Clustering of chemical substances in EQI by extreme high concentraties. Vertical dotted line: dissimilarity of 0.95.

11 clusters; mostly of size 2; 1 cluster of size 3; 1 cluster of size 4; ...

measuring associations between random vectors

.... based on comparing clusters pairwise ....

what is we want to look at say k clusters simultaneaously ?

- need association measure between k random vectors, not only 2
- how to measure association between k random vectors ?

some recent approaches:

- ϕ-divergences (cf information theory) (De Keyser & G. (2022), ...)
- based on optimal transport ideas (Mordant & Segers (2022), ...)
- ....

measuring associations in more complex (different) data structures

- what if random variables are not real values, but curves, images, ... ?
- what if data live on a circle, a torus, ...?
- ...

how to measure (strength of) dependence in these cases?

for circular data: bivariate case

- random angles Θ_1, Θ_2, with joint (circular) cum. distr, function F and marginals F_1 and F_2
- a bivariate circula is defined as
  (Kato *et al.* (2022), Ameijeiras-Alonso & G. (2022), ...)

$$F(\theta_1, \theta_2) = C\left(2\pi F_1(\theta_1) - \pi, \ldots, 2\pi F_d(\theta_d) - \pi\right) \quad (\theta_1, \theta_2) \in [-\pi, \pi)^2$$

with $C : [-\pi, \pi)^2 \mapsto [0, 1]$, a bivariate circular distribution whose marginals are circular uniform distributions
(circular periodicity, i.e. $F_j(\theta + 2\pi) - F_j(\theta) = 1$)

- consider

$$\mathbf{X}_1 = (\cos(\Theta_1), \sin(\Theta_1))^\top \qquad \mathbf{X}_2 = (\cos(\Theta_2), \sin(\Theta_2))^\top$$

  $\mathbf{X}_1, \mathbf{X}_2$ values in $[-1,1]^2 \subset \mathbb{R}^2$

- since circular uniform marginals: for $s \in \{1,2\}$

$$E(\mathbf{X}_s) = \mathbf{0}^\top \qquad \text{and} \qquad E\left(\mathbf{X}_s \mathbf{X}_s^\top\right) = \mathbf{I}_2/2$$

  where $\mathbf{I}_2$ is the identity matrix of size 2

- furthermore

$$\begin{aligned}
&\boldsymbol{\Sigma}_{12} \\
&= E\left(\mathbf{X}_1 \mathbf{X}_2^\top\right) - E(\mathbf{X}_1)\, E(\mathbf{X}_2)^\top \\
&= \begin{pmatrix} \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \sin\theta_1 \sin\theta_2\, \mathbf{C}(\boldsymbol{\theta_1, \theta_2}) d\theta_1 d\theta_2 - 1 & -\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \cos\theta_1 \sin\theta_2\, \mathbf{C}(\boldsymbol{\theta_1, \theta_2}) d\theta_1 d\theta_2 \\ -\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \cos\theta_1 \sin\theta_2\, \mathbf{C}(\boldsymbol{\theta_1, \theta_2}) d\theta_1 d\theta_2 & \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \cos\theta_1 \cos\theta_2\, \mathbf{C}(\boldsymbol{\theta_1, \theta_2}) d\theta_1 d\theta_2 \end{pmatrix}
\end{aligned}$$

available dependence measures in terms of this matrix ....

notation: $\lambda_1(\boldsymbol{\Sigma})$ and $\lambda_2(\boldsymbol{\Sigma})$: eigenvalues of matrix $\boldsymbol{\Sigma}$

Table: *Dependence measures when applied to circulas.*

Expression	Reference
$\tau_{\mathsf{FL}} = 4\det(\boldsymbol{\Sigma}_{12})$	Lee & Fisher (1983)
$\tau_{\mathsf{JW}} = \sqrt{\max\{\lambda_1(4\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{12}^\top), \lambda_2(4\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{12}^\top)\}}$	Johnson (1977)
$\tau_{\mathsf{JM}} = \mathsf{trace}(4\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{12}^\top)$	Jupp (1980)
$\tau_{\mathsf{R}} = 2\mathsf{sign}(\det(\boldsymbol{\Sigma}_{12}))\min\{\lvert\lambda_1(\boldsymbol{\Sigma}_{12})\rvert, \lvert\lambda_2(\boldsymbol{\Sigma}_{12})\rvert\}$	Rivest (1982)

further issues

- what happens with dependence structures/strengths when variables are added to a vector:
  - an independent vector
  - or a canonical combination of components already included .. ?
  - or a set of arbitrary components ?
  -
- what are reasonable association measures, in particular in high dimensions ?
- how to measure dependence for complex data structures ?
-
-

Thank you