Dependencies in higher dimensions and in complex data structures by Irène Gijbels

discussed by Fabrizio Durante

Università del Salento, Lecce, Italy
PRIN 2017 "Stochastic Models for Complex Systems"
fabrizio.durante@unisalento.it
sites.google.com/site/fbdurante

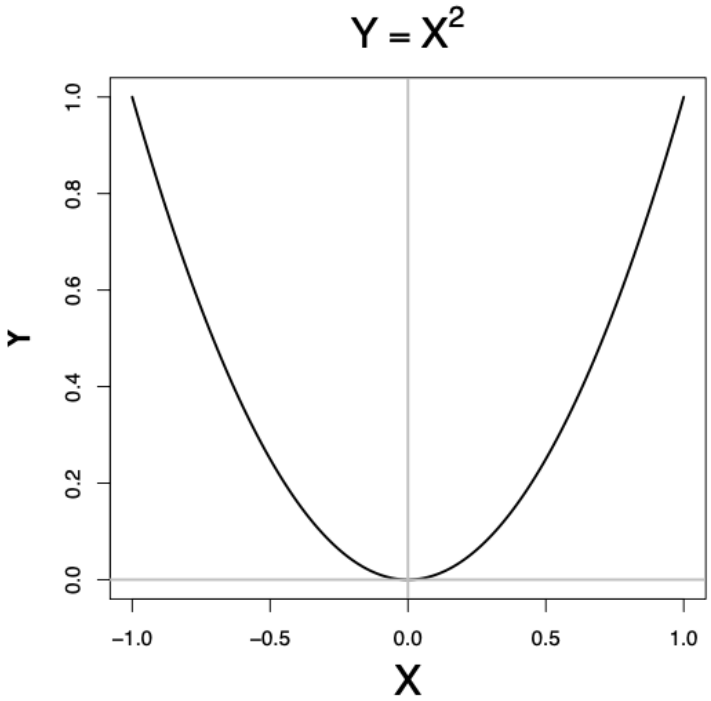
Statistical methods and models for complex data, Padova, September 2022

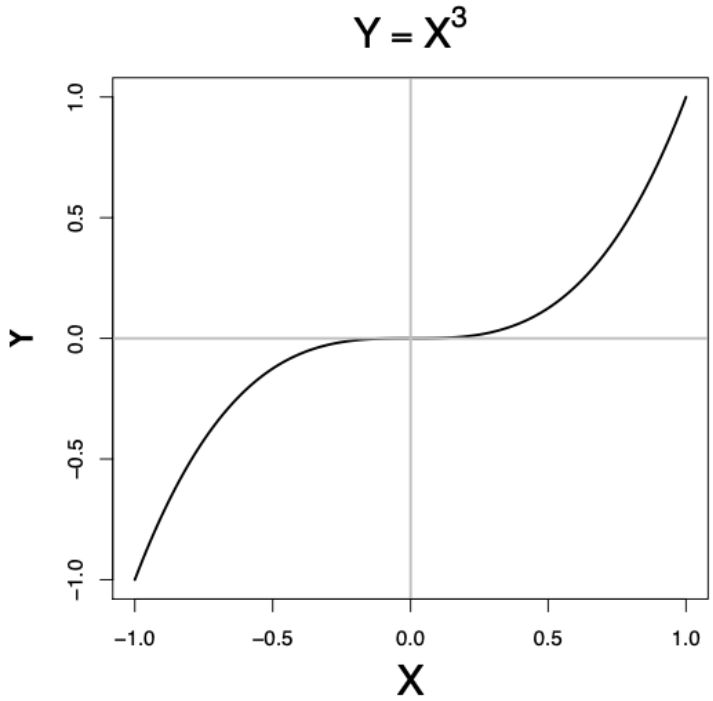
Association and dependence

Copulas are a fairly complex way to describe dependence. Thus, it may be convenient to simplify to a numerical summary of dependence.

- Association: The extent up to which large (small) values of X go together with large (small) values of Y (e.g., Concordance).
- Dependence: The extent up to which the outcome of Y is predictable from the outcome of X.

Association and dependence





Left: perfectly dependent, but not associated. Right: perfectly dependent and associated.

From mathematical soundness ...

A) $\delta(\xi, \eta)$ *is defined for any pair of random variables* ξ *and* η, *neither of them being constant with probability* 1.

B) $\delta(\xi, \eta) = \delta(\eta, \xi)$.

C) $0 \leqq \delta(\xi, \eta) \leqq 1$.

D) $\delta(\xi, \eta) = 0$ *if and only if* ξ *and* η *are independent.*

E) $\delta(\xi, \eta) = 1$ *if*[1] *there is a strict dependence between* ξ *and* η, *i. e. either* $\xi = g(\eta)$ *or* $\eta = f(\xi)$ *where* $g(x)$ *and* $f(x)$ *are Borel-measurable functions.*

F) *If the Borel-measurable functions* $f(x)$ *and* $g(x)$ *map the real axis in a one-to-one way onto itself,* $\delta(f(\xi), g(\eta)) = \delta(\xi, \eta)$.

G) *If the joint distribution of* ξ *and* η *is normal, then* $\delta(\xi, \eta) = |\mathbf{R}(\xi, \eta)|$ *where* $\mathbf{R}(\xi, \eta)$ *is the correlation coefficient of* ξ *and* η.

(Rényi, 1959)

Long history that continues until now:
Distance correlation (Székely et al., 2007)
MIC (Reshef et al., 2011)
Chatterjee's measure (Chatterjee, 2020; Shi, Drton and Han, 2022)
...to name only a few!

... to statistical convenience

(a) **Distribution-freeness**: the (limiting) distribution of the correlation coefficient under the hypothesis of independence should not depend on the marginal distributions of X and Y;

(b) **Consistency**: the correlation coefficient should consistently estimate a measure of dependence that is 0 if and only if X is independent of Y within a fairly large distribution family of (X, Y);

(c) **Statistical efficiency**: the test of independence based on the correlation coefficient should have nontrivial power over root-n neighborhoods of "smooth" parametric models;

(d) **Computational efficiency**: there should exist a nearly linear-time algorithm to compute the correlation coefficient.

(Han, 2021)

Any favourite measure given the above list of properties?

About axioms for variable clustering

(A8) (Independent Component Addition) For X_{d+1} independent of $(X_1, \ldots, X_d)^\top$
$\kappa_d(C_d) > \kappa_{d+1}(C_{d+1}) > 0 \quad \text{or} \quad \kappa_d(C_d) < \kappa_{d+1}(C_{d+1}) < 0$
or $\kappa_d(C_d) = \kappa_{d+1}(C_{d+1}) = 0.$

For variable clustering, consider also

(P1) (Comonotone Component Addition) For X_{d+1} comonotone with at least one element of $(X_1, \ldots, X_d)^\top$

$$\kappa_d(C_d) \leq \kappa_{d+1}(C_{d+1}).$$

Gijbels et al. (2021): *Duplication of one component (or more generally adding a conical combination of all components) cannot decrease the association measure.*

😟 Spearman's ρ and Gini's *gamma* fail to satisfy (P1) (see Gijbels et al., 2021; Fuchs et al., 2021).

About axioms for variable clustering

(P2) If one variable X_{d+1} is added to $(X_1, \ldots, X_d)^\top$, then

$$\psi_1(\kappa_d(C_d), \cdot) \leq \kappa_{d+1}(C_{d+1}) \leq \psi_2(\kappa_d(C_d), \cdot)$$

for suitable mappings ψ_1, ψ_2.

Gijbels et al. (2021): *What is the effect of adding $d_2 - d$ (with $d_2 > d$) arbitrary components?*

For variable (hierarchical) clustering, consider also

(P2') (Reducibility) If $\kappa_2(X_1, X_2) \geq \max\{\kappa_2(X_1, Y), \kappa_2(X_2, Y)\}$, then $\kappa_3(X_1, X_2, Y) \leq \kappa_2(X_1, X_2)$.

Are there interesting examples of monotone convergence of $d \mapsto (\kappa_d(C_d))_{d \in \mathbb{N}}$?

About estimation

- *Quality of estimator κ_n depends on that of the empirical copula $\widehat{C}_n$.*

   Are there improvements when smooth empirical copula versions (Beta, Bernstein, etc.) are used?

- *Problem when estimating TDC's: need to deal with the limits.*

   Non-, Semi- (EV), Fully parametric? Any suggestion? Does only the sample size matter?

Thanks again Irène for your talk