Optimal Copula Transport for Clustering Time Series

Gautier Marti[1,2], Frank Nielsen[2], Philippe Donnat[1]

[1]Hellebore Capital Limited & [2]Ecole Polytechnique

Clustering Time Series Which Dependence Measure? For Which Dependence?

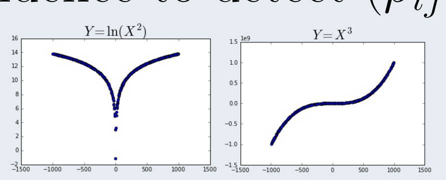
Many bivariate dependence measures are available. Usually, they aim at measuring:

- any deviation from independence,
- any deviation from co/counter-monotonicity.

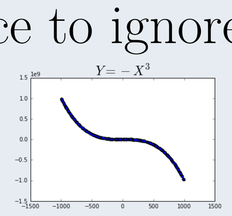
Motivation: What if

- we aim at specific dependence,
- and try to "ignore" some others?

Dependence to detect ($\rho_{ij} := 1$)



Dependence to ignore ($\rho_{ij} := 0$)



Problem: A dependence measure powerful enough to detect $y = f(x^2)$ will also detect $y = g(x)$, f increasing, g decreasing.

Copulas & Dependence

- Sklar's Theorem:

$$F(x_i, x_j) = C_{ij}(F_i(x_i), F_j(x_j))$$

- C_{ij}, the copula, encodes the dependence structure
- Fréchet-Hoeffding bounds:

$$\max\{u_i + u_j - 1, 0\} \leq C_{ij}(u_i, u_j) \leq \min\{u_i, u_j\}$$

- Bivariate dependence measures:
  - deviation from lower and upper bounds
    - Spearman's ρ_S, Gini's γ
  - deviation from independence $u_i u_j$
    - Spearman, Copula MMD, Schweizer-Wolff's σ, Hoeffding's Φ^2

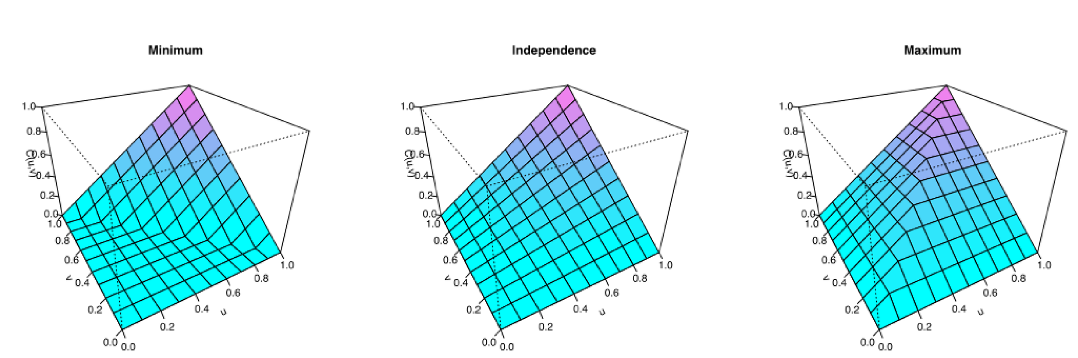


Figure 1: (left) lower-bound copula, (mid) independence copula, (right) upper-bound copula

Optimal Transport

Wasserstein metrics:

$$W_p^p(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{M \times M} d(x,y)^p \mathrm{d}\gamma(x,y)$$

In practice, the distance W_1 is estimated on discrete data by solving the following linear program with the Hungarian algorithm:

$$\begin{aligned} \mathrm{EMD}(s_1, s_2) := \min_f & \sum_{1 \leq k,l \leq n} \|p_k - q_l\| f_{kl} \\ \text{subject to} \quad & f_{kl} \geq 0,\ 1 \leq k,l \leq n, \\ & \sum_{l=1}^{n} f_{kl} \leq w_{p_k},\ 1 \leq k \leq n, \\ & \sum_{k=1}^{n} f_{kl} \leq w_{q_l},\ 1 \leq l \leq n, \\ & \sum_{k=1}^{n}\sum_{l=1}^{n} f_{kl} = 1. \end{aligned}$$

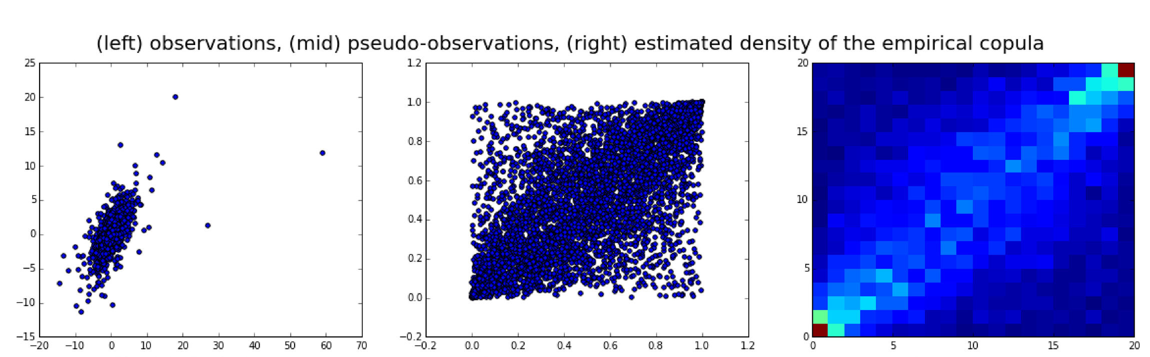
It is called the Earth Mover Distance (EMD) in the CS literature.

EMD between Copulas

- Probability integral transform of a variable x_i:

$$F_T(x_i^k) = \frac{1}{T}\sum_{t=1}^{T} I(x_i^t \leq x_i^k),$$

i.e. computing the ranks of the realizations, and normalizing them into [0,1]



Why the Earth Mover Distance?

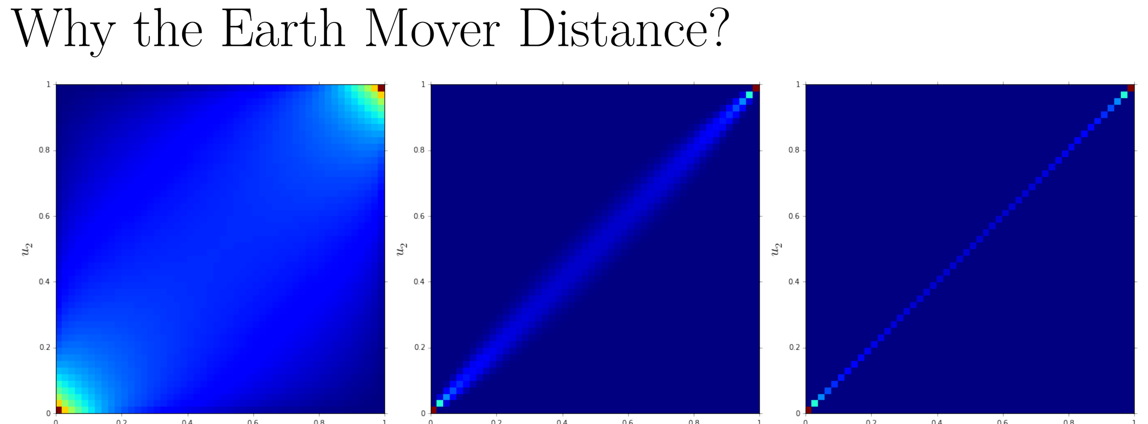


Figure 3: Copulas C_1, C_2, C_3 encoding a correlation of $0.5, 0.99, 0.9999$ respectively; *Which pair of copulas is the nearest?* For Fisher-Rao, Kullback-Leibler, Hellinger and related divergences: $\mathrm{D}(C_1, C_2) \leq \mathrm{D}(C_2, C_3)$; $\mathrm{EMD}(C_2, C_3) \leq \mathrm{EMD}(C_1, C_2)$

A target-oriented dependence coefficient

- Build the independence copula C_{ind}
- Build the target-dependence copulas $\{C_k\}_k$
- Compute the empirical copula C_{ij} from x_i, x_j

$$\mathtt{TDC}(C_{ij}) = \frac{\mathrm{EMD}(C_{\mathrm{ind}}, C_{ij})}{\mathrm{EMD}(C_{\mathrm{ind}}, C_{ij}) + \min_k \mathrm{EMD}(C_{ij}, C_k)}$$

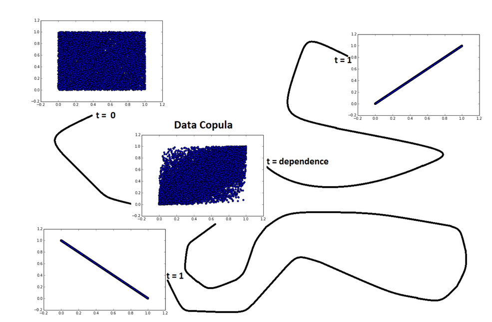


Figure 2: Dependence is measured as the relative distance from independence to the nearest target-dependence

Benchmark: Power of Estimators

Our coefficient can robustly target complex dependence patterns such as the ones displayed in Fig. 4.

- x-axis measures the noise added to the sample
- y-axis measures the frequency the coefficient is able to discern between the dependent sample and the independent one
- Basic check: no coefficient can discern between the "dependent" sample (with no dependence) and the independent sample.

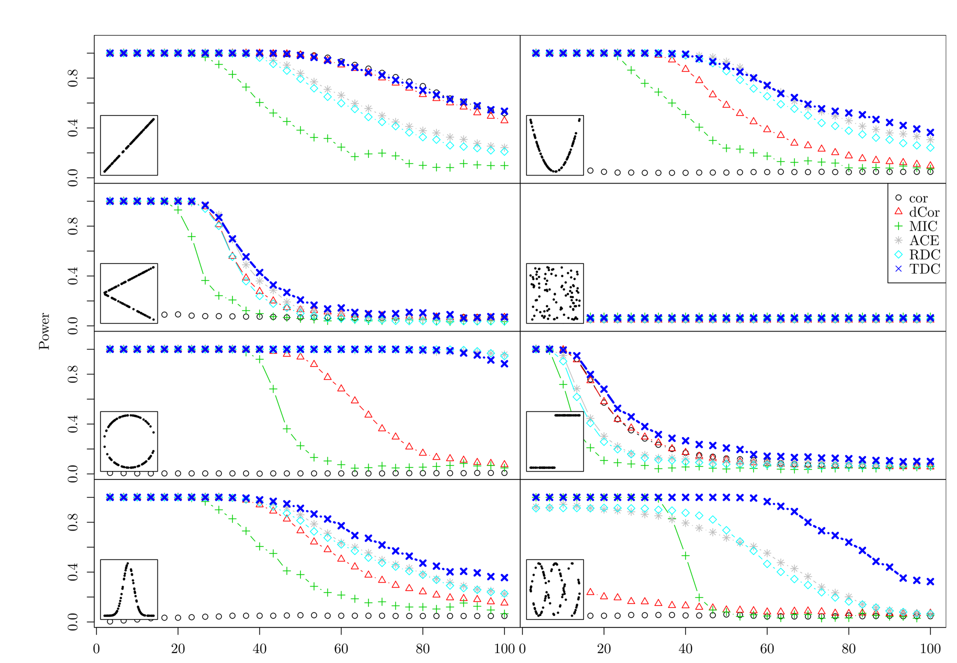


Figure 4: Dependence estimators power as a function of the noise for several deterministic patterns + noise. Their power is the percentage of times that they are able to distinguish between dependent and independent samples.

Clustering of Credit Default Swaps

- We use the two targets from Fig. 2
- Clustering distance: $D_{ij} = \sqrt{(1 - \mathtt{TDC}(C_{ij}))/2}$

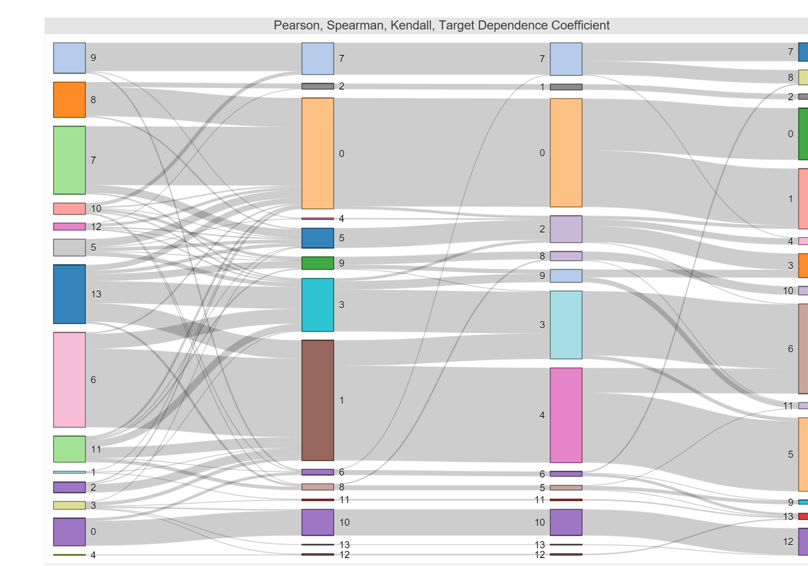


Figure 5: Impact of different measures on clusters

Conclusion

The methodology presented is

- non-parametric, robust, deterministic.

It has some scalability issues:

- in dimension, non-parametric density estimation;
- in time, EMD is costly to compute.

Approximation schemes or parametric modelling can alleviate these issues.

Information

- Web: www.datagrapple.com
- Email: gautier.marti@helleborecapital.com

Hellebore Capital Ltd

ÉCOLE POLYTECHNIQUE
UNIVERSITÉ PARIS-SACLAY