Clustering Random Walk Time Series

GSI 2015 - Geometric Science of Information

Gautier Marti, Frank Nielsen, Philippe Very, Philippe Donnat

Hellebore Capital Management

ÉCOLE POLYTECHNIQUE UNIVERSITÉ PARIS-SACLAY

29 October 2015

Hellebore Capital Management

Context (data from www.datagrapple.com)

Hellebore
Capital Management

What is a clustering program?

Definition

Clustering is the task of grouping a set of objects in such a way that objects in the same group (cluster) are more similar to each other than those in different groups.

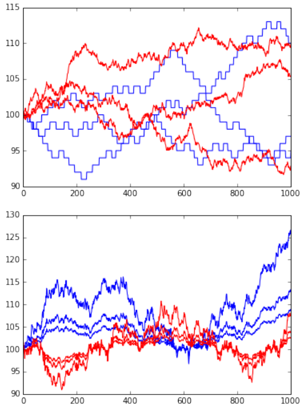
Example of a clustering program

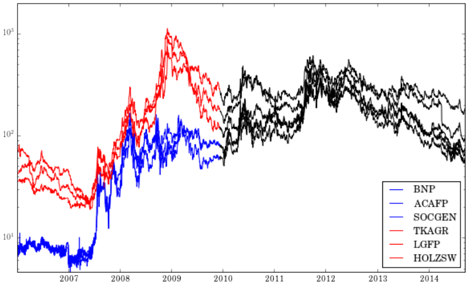
We aim at finding k groups by positioning k group centers $\{c_1, \ldots, c_k\}$ such that data points $\{x_1, \ldots, x_n\}$ minimize $\min_{c_1,\ldots,c_k} \sum_{i=1}^{n} \min_{j=1}^{k} d(x_i, c_j)^2$

But, what is the distance d between two random walk time series?

Hellebore
Capital Management

What are clusters of Random Walk Time Series?

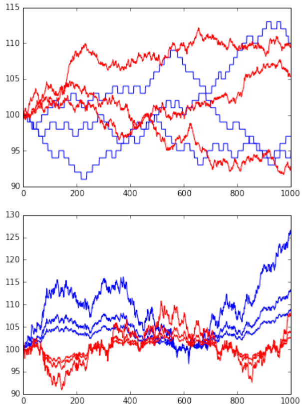


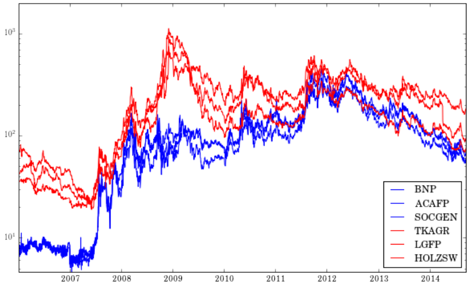


French banks and building materials CDS over 2006-2015

Hellebore Capital Management

What are clusters of Random Walk Time Series?





French banks and building materials CDS over 2006-2015

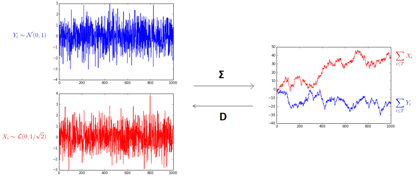
Hellebore Capital Management

Hellebore Capital Management

Geometry of RW TS $\equiv$ Geometry of Random Variables

i.i.d. observations:

$$
\begin{array}{llllll}
X_1 : & X_1^1, & X_1^2, & \ldots, & X_1^T \\
X_2 : & X_2^1, & X_2^2, & \ldots, & X_2^T \\
 & \ldots, & \ldots, & \ldots, & \ldots, & \ldots \\
X_N : & X_N^1, & X_N^2, & \ldots, & X_N^T
\end{array}
$$



Which distances $d(X_i, X_j)$ between dependent random variables?

Hellebore Capital Management

Pitfalls of a basic distance

Let (X, Y) be a bivariate Gaussian vector, with $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$, $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ and whose correlation is $\rho(X, Y) \in [-1, 1]$.

$$\mathbb{E}[(X - Y)^2] = (\mu_X - \mu_Y)^2 + (\sigma_X - \sigma_Y)^2 + 2\sigma_X\sigma_Y(1 - \rho(X, Y))$$

Now, consider the following values for correlation:

- $\rho(X, Y) = 0$, so $\mathbb{E}[(X - Y)^2] = (\mu_X - \mu_Y)^2 + \sigma_X^2 + \sigma_Y^2$. Assume $\mu_X = \mu_Y$ and $\sigma_X = \sigma_Y$. For $\sigma_X = \sigma_Y \gg 1$, we obtain $\mathbb{E}[(X - Y)^2] \gg 1$ instead of the distance 0, expected from comparing two equal Gaussians.
- $\rho(X, Y) = 1$, so $\mathbb{E}[(X - Y)^2] = (\mu_X - \mu_Y)^2 + (\sigma_X - \sigma_Y)^2$.

Hellebore Capital Management

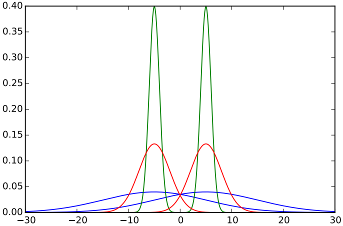
Pitfalls of a basic distance

Let (X, Y) be a bivariate Gaussian vector, with $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$, $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ and whose correlation is $\rho(X, Y) \in [-1, 1]$.

$$\mathbb{E}[(X - Y)^2] = (\mu_X - \mu_Y)^2 + (\sigma_X - \sigma_Y)^2 + 2\sigma_X \sigma_Y (1 - \rho(X, Y))$$

Now, consider the following values for correlation:

- $\rho(X, Y) = 0$, so $\mathbb{E}[(X - Y)^2] = (\mu_X - \mu_Y)^2 + \sigma_X^2 + \sigma_Y^2$. Assume $\mu_X = \mu_Y$ and $\sigma_X = \sigma_Y$. For $\sigma_X = \sigma_Y \gg 1$, we obtain $\mathbb{E}[(X - Y)^2] \gg 1$ instead of the distance 0, expected from comparing two equal Gaussians.
- $\rho(X, Y) = 1$, so $\mathbb{E}[(X - Y)^2] = (\mu_X - \mu_Y)^2 + (\sigma_X - \sigma_Y)^2$.



Probability density functions of Gaussians $\mathcal{N}(-5, 1)$ and $\mathcal{N}(5, 1)$, Gaussians $\mathcal{N}(-5, 3)$ and $\mathcal{N}(5, 3)$, and Gaussians $\mathcal{N}(-5, 10)$ and $\mathcal{N}(5, 10)$. Green, red and blue Gaussians are equidistant using L_2 geometry on the parameter space (μ, σ).

Hellebore Capital Management

Sklar's Theorem

Theorem (Sklar's Theorem (1959))

For any random vector $X = (X_1, \ldots, X_N)$ having continuous marginal cdfs P_i, $1 \leq i \leq N$, its joint cumulative distribution P is uniquely expressed as

$$P(X_1, \ldots, X_N) = C(P_1(X_1), \ldots, P_N(X_N)),$$

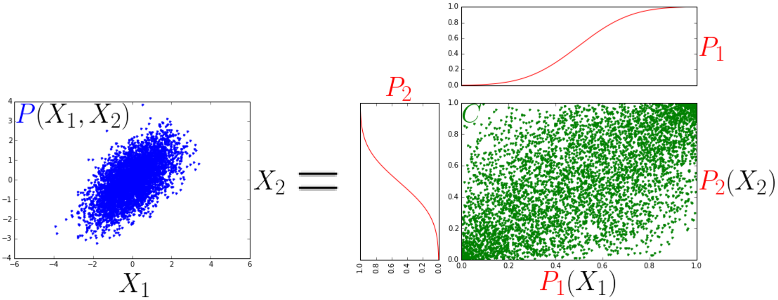
where C, the multivariate distribution of uniform marginals, is known as the copula of X.

Hellebore Capital Management

Sklar's Theorem

Theorem (Sklar's Theorem (1959))

For any random vector $X = (X_1, \ldots, X_N)$ having continuous marginal cdfs P_i, $1 \leq i \leq N$, its joint cumulative distribution P is uniquely expressed as $P(X_1, \ldots, X_N) = C(P_1(X_1), \ldots, P_N(X_N))$, where C, the multivariate distribution of uniform marginals, is known as the copula of X.



The Copula Transform

Definition (The Copula Transform)

Let $X = (X_1, \ldots, X_N)$ *be a random vector with continuous marginal cumulative distribution functions (cdfs)* P_i, $1 \leq i \leq N$. *The random vector*

$$U = (U_1, \ldots, U_N) := P(X) = (P_1(X_1), \ldots, P_N(X_N))$$

is known as the copula transform.

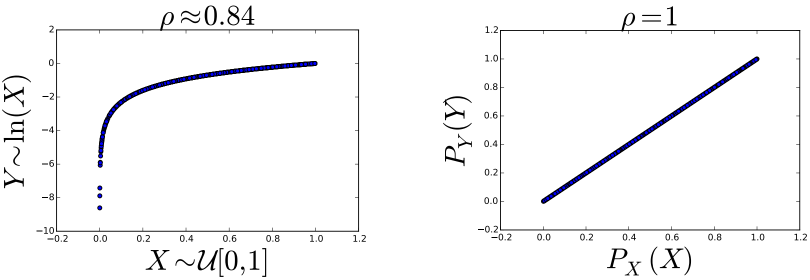
U_i, $1 \leq i \leq N$, are uniformly distributed on $[0, 1]$ (the probability integral transform): for P_i the cdf of X_i, we have $x = P_i(P_i{}^{-1}(x)) = \mathrm{Pr}(X_i \leq P_i{}^{-1}(x)) = \mathrm{Pr}(P_i(X_i) \leq x)$, thus $P_i(X_i) \sim \mathcal{U}[0, 1]$.

Hellebore
Capital Management

The Copula Transform

Definition (The Copula Transform)

Let $X = (X_1, \dots, X_N)$ be a random vector with continuous marginal cumulative distribution functions (cdfs) P_i, $1 \leq i \leq N$. The random vector $U = (U_1, \dots, U_N) := P(X) = (P_1(X_1), \dots, P_N(X_N))$ is known as the copula transform.



The Copula Transform invariance to strictly increasing transformation

Hellebore
Capital Management

Deheuvels' Empirical Copula Transform

Let $(X_1^t, \dots, X_N^t)$, $1 \leq t \leq T$, be T observations from a random vector $(X_1, \dots, X_N)$ with continuous margins. Since one cannot directly obtain the corresponding copula observations $(U_1^t, \dots, U_N^t) = (P_1(X_1^t), \dots, P_N(X_N^t))$, where $t = 1, \dots, T$, without knowing a priori $(P_1, \dots, P_N)$, one can instead

Definition (The Empirical Copula Transform)

- estimate the N empirical margins $P_i^T(x) = \frac{1}{T}\sum_{t=1}^T \mathbf{1}(X_i^t \leq x)$, $1 \leq i \leq N$, to obtain the T empirical observations

$$(\tilde{U}_1^t, \dots, \tilde{U}_N^t) = (P_1^T(X_1^t), \dots, P_N^T(X_N^t)).$$

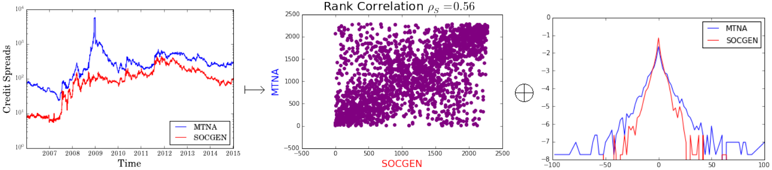
- Equivalently, since $\tilde{U}_i^t = R_i^t / T$, R_i^t being the rank of observation X_i^t, the empirical copula transform can be considered as the **normalized rank transform**.

In practice

```
x_transform = rankdata(x)/len(x)
```

Hellebore Capital Management

Generic Non-Parametric Distance



$$d_\theta^2(X_i, X_j) = \theta 3 \mathbb{E}\left[|P_i(X_i) - P_j(X_j)|^2\right] + (1-\theta)\frac{1}{2}\int_{\mathbf{R}}\left(\sqrt{\frac{dP_i}{d\lambda}} - \sqrt{\frac{dP_j}{d\lambda}}\right)^2 d\lambda$$

(i) $0 \leq d_\theta \leq 1$, (ii) $0 < \theta < 1$, d_θ metric,
(iii) d_θ is invariant under diffeomorphism

Hellebore
Capital Management

Generic Non-Parametric Distance

$$d_0^2 : \frac{1}{2} \int_{\mathbf{R}} \left(\sqrt{\frac{dP_i}{d\lambda}} - \sqrt{\frac{dP_j}{d\lambda}} \right)^2 d\lambda = \mathrm{Hellinger}^2$$

$$d_1^2 : 3\mathbb{E}\left[|P_i(X_i) - P_j(X_j)|^2 \right] = \frac{1-\rho_S}{2} = 2 - 6 \int_0^1 \int_0^1 C(u,v)\mathrm{d}u\mathrm{d}v$$

Remark:

If $f(x,\theta) = c_\Phi(u_1,\ldots,u_N;\Sigma)\prod_{i=1}^N f_i(x_i;\nu_i)$ then

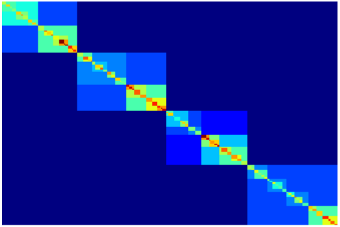
$$ds^2 = ds^2_{GaussCopula} + \sum_{i=1}^{N} ds^2_{margins}$$

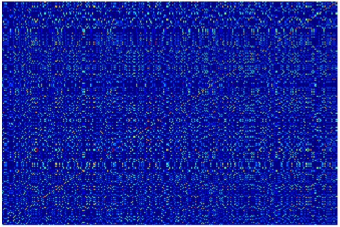
Hellebore
Capital Management

Hellebore Capital Management

The Hierarchical Block Model

A model of nested partitions





The nested partitions defined by the model can be seen on the distance matrix for a proper distance and the right permutation of the data points

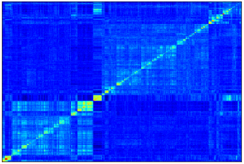
In practice, one observe and work with the above distance matrix which is identitical to the left one up to a permutation of the data

Results: Data from Hierarchical Block Model

Algo.	Distance	Adjusted Rand Index: Distrib	Correl	Correl+Distrib
HC-AL	$(1-\rho)/2$	0.00 ±0.01	0.99 ±0.01	0.56 ±0.01
	$\mathbb{E}[(X-Y)^2]$	0.00 ±0.00	0.09 ±0.12	0.55 ±0.05
	GPR $\theta = 0$	0.34 ±0.01	0.01 ±0.01	0.06 ±0.02
	GPR $\theta = 1$	0.00 ±0.01	0.99 ±0.01	0.56 ±0.01
	GPR $\theta = .5$	0.34 ±0.01	0.59 ±0.12	0.57 ±0.01
	GNPR $\theta = 0$	**1**	0.00 ±0.00	0.17 ±0.00
	GNPR $\theta = 1$	0.00 ±0.00	**1**	0.57 ±0.00
	GNPR $\theta = .5$	0.99 ±0.01	0.25 ±0.20	**0.95** ±0.08
AP	$(1-\rho)/2$	0.00 ±0.00	0.99 ±0.07	0.48 ±0.02
	$\mathbb{E}[(X-Y)^2]$	0.14 ±0.03	0.94 ±0.02	0.59 ±0.00
	GPR $\theta = 0$	0.25 ±0.08	0.01 ±0.01	0.05 ±0.02
	GPR $\theta = 1$	0.00 ±0.01	0.99 ±0.01	0.48 ±0.02
	GPR $\theta = .5$	0.06 ±0.00	0.80 ±0.10	0.52 ±0.02
	GNPR $\theta = 0$	**1**	0.00 ±0.00	0.18 ±0.01
	GNPR $\theta = 1$	0.00 ±0.01	**1**	0.59 ±0.00
	GNPR $\theta = .5$	0.39 ±0.02	0.39 ±0.11	**1**

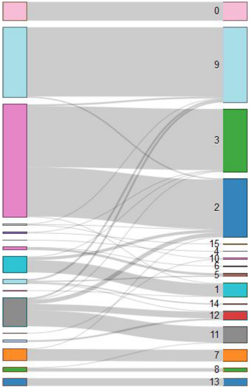
Hellebore Capital Management

Results: Application to Credit Default Swap Time Series

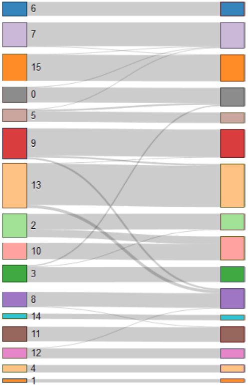


Distance matrices computed on CDS time series exhibit a **hierarchical block structure**

Marti, Very, Donnat, Nielsen IEEE ICMLA 2015



(un)Stability of clusters with L_2 distance



Stability of clusters with the proposed distance

Hellebore Capital Management

Consistency

Definition (Consistency of a clustering algorithm)

A clustering algorithm $\mathcal{A}$ is consistent with respect to the Hierarchical Block Model defining a set of nested partitions $\mathcal{P}$ if the probability that the algorithm $\mathcal{A}$ recovers all the partitions in $\mathcal{P}$ converges to 1 when $T \to \infty$.

Definition (Space-conserving algorithm)

A space-conserving algorithm does not distort the space, i.e. the distance D_{ij} between two clusters C_i and C_j is such that

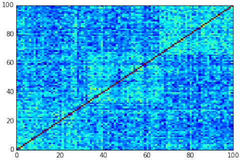
$$D_{ij} \in \left[\min_{x \in C_i, y \in C_j} d(x,y), \max_{x \in C_i, y \in C_j} d(x,y) \right].$$

Hellebore
Capital Management

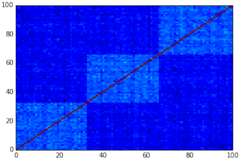
Consistency

Theorem (Consistency of space-conserving algorithms (Andler, Marti, Nielsen, Donnat, 2015))

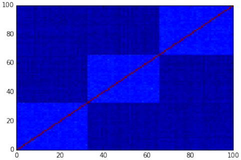
Space-conserving algorithms (e.g., Single, Average, Complete Linkage) are consistent with respect to the Hierarchical Block Model.



$T = 100$



$T = 1000$



$T = 10000$

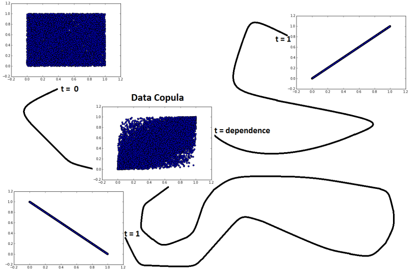
Hellebore
Capital Management

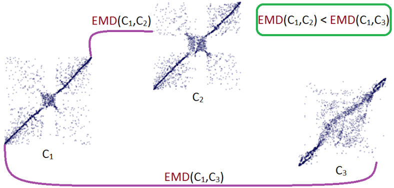
Hellebore Capital Management

Discussion and questions?

Avenue for research:

- distances on (copula,margins)
- clustering using multivariate dependence information
- clustering using multi-wise dependence information





Optimal Copula Transport for Clustering Multivariate Time Series, Marti, Nielsen, Donnat, 2015

Hellebore Capital Management