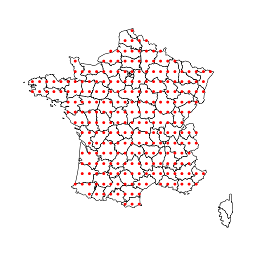
Clustering high dimensional data

Dominique Picard

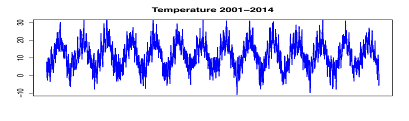
Université Paris-Diderot
LPMA

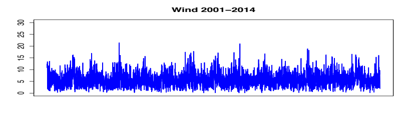
Joint works with V. Lefieux, M. Marchand, M. Mougeot, – A. Fischer

Arpege French Meteorological data



At $n = 259$ locations,
- Temperature and Wind
- for 14 years
- hourly sample rate
- $d = 122\ 712$ points for raw data
- Y data matrix (n x d)
- $n << d$





Objective and Questions :

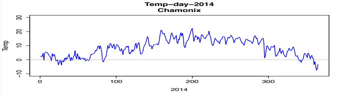
Goals

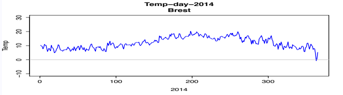
- Segmentation of the country into regions using meteorological data
- Temperature and/or Wind
- Study the Between Year variability

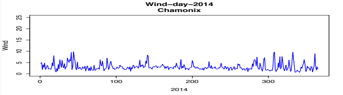
Methodological & Statistical Questions :

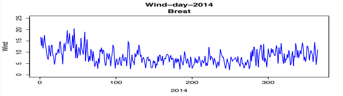
- High dimensional data $n = 259$, $d = 122\,712$, $d >> n$
- Features extraction, Smoothing
- Representation of the data
- Modeling
  - Mixtures, HMM
  - Spectral clustering
- Clustering algorithms :
  - Hierarchical clustering, Kmeans
  - Kernel clustering
  - Spectral clustering
  - Number of clusters, Smoothing

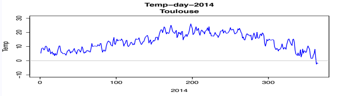
Wind and Temperature spots for 2014

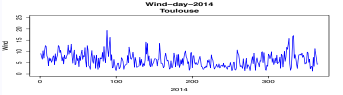




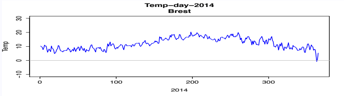








Natural-time aggregation smoothing



The data are observed hourly. It is commonly admitted to take

1. the average on a day : daily observed data $T = 365$ for one year.
2. the average on a week : daily observed data $T = 52$ for one year.
3. the average on a month : daily observed data $T = 12$ for one year.

Projection of the observations using a data driven orthonormal basis

X centered data matrix (n, d)
$n = 259,\ d >> n$ large

The Feature matrix (n, T) is computed by projection, $T << d$:

$$\boxed{Z = XU_T}$$

U_T is the matrix defined by the first eigenvectors of S, the Variance-Covariance matrix.

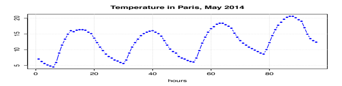
T chosen so that ? $\frac{\lambda_1 + \ldots + \lambda_T}{\Sigma_j \lambda_j} = \kappa_{pca}$ (0.95)

$\rightarrow$ Global linear method involving all the $n = 259$ spots to compute U_T
$\rightarrow$ Is U_T similar between years ?

Data are (in fact) functions of time regularly spaced.

$$X_t^i = f^i(t/d) + \epsilon_t^i,$$

f^i is unknow, $\epsilon^i \sim \mathcal{N}(0, \sigma^2)$, $t = 1, \ldots, d$.
Nonparametric estimation of f^i : $f^i = \sum_{\ell=1}^{T} \beta_\ell^i g_\ell$
with $\mathcal{D} = \{g_1, \ldots, g_p\}$ dictionary of functions.



How to choose T ? (more to come)

Here
We note $\hat{X}_{j_0}^i = \sum_{j=1}^{j_0} \hat{\beta}_{(j)}^i g_j$
with $|\hat{\beta}_{(1)}^i| \geq \ldots \geq |\hat{\beta}_{(n)}^i|$, and $\frac{||\hat{X}_{(j_0)}^i||^2}{||X^i||^2} \geq T_{NP}(= 0.95)$.

Choose k the number of clusters

Find the Arg min (in $C_1, \ldots, C_k$) of :

$$\sum_{r=1}^{k} \sum_{j \neq j', \in C_r} \| Y_j - Y_{j'} \|^2 = 2 \sum_{r=1}^{k} \sum_{j \in C_r} \| Y_j - \bar{Y}_r \|^2,$$

$$\bar{Y}_r = \frac{1}{|C_j|} \sum_{j \in C_r} Y_j.$$

Surrogate Model : Maximum Likelihood approach in a Mixture of Gaussian Variables

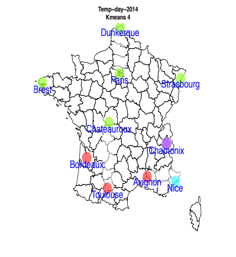
$$g(x|\theta) = \sum_{l=1}^{k} \alpha_l g_l(x|\theta_l)$$

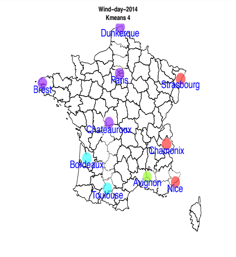
where α_l belongs to $[0, 1]$ and g_l is a gaussian density with expectation $\theta_l (\in \boldsymbol{R}^d)$ covariance matrix I_d.

Kmeans Clustering

Choose k the number of clusters

1. INPUT k centroids $\bar{Z}_1 \dots \bar{Z}_k$ (k points at random)
2. Compute $\sum_{k=1}^{K} \sum_{c(i)=k} ||Z_i - \bar{Z}_k||^2$
3. Reassign each item to its nearest cluster centroid
4. Update the cluster centroids after each assignment.
5. REPEAT 2,3,4 with **until** no further assignment of items takes place.





over 14 years, for different temporal aggregation levels

Data : 14 x one year of data,
Kmeans algorithm
Temperature :

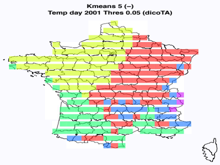
	day (365)	week (52)	month (12)
PCA 95%	5 (0)	4.9 (0.2)	4.7 (0.4)
NP Reg. Trigo	5 (0)	4.8 (0.4)	4.7 (0.4)
NP Reg. Haar	5 (0)	4.8 (0.4)	4.7 (0.4)

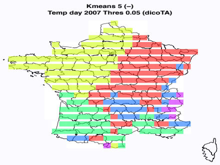
Wind :

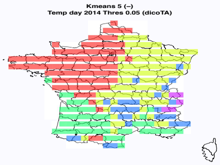
	day (365)	week (52)	month (12)
Pca 90%	4.15 (0.3)	4.23 (0.4)	4.31 (0.4)
NP Reg. Trigo	4.15 (0.3)	4 (0)	4.08 (0.2)
NP Reg. Haar	4.23 (0.4)	4.31 (0.4)	4.15 (0.3)

Segmentation for 2001, 2007, 2014 daily data, $n = 259$

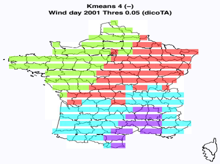
Temperature

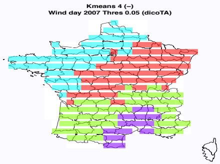


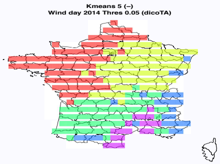




Wind







Questions

1. What is better : raw data or smoothing ?
2. What conditions ? (sparsity, separation of clusters...)
3. How to smooth ? Does usual adaptation methods work as well to detect clusters ?
4. On-line (signal by signal smoothing) or off-line smoothing (using a pre-process involving all the signals) ?
5. What are the rates ?

Simpler framework

1. 2 classes only
2. The change occurs on a time scale

We observe $Y_1, \dots, Y_n$ n independent signals.
Each signal is observed discretely, i.e. $Y_j = (Y_j^1, \dots, Y_j^d)$,
Gaussian clustering :
There exists a set $A \subset \{1, \dots, n\}$, and two regular vectors of $\mathbb{R}^d$ θ_- and θ_+ such that

$$\begin{aligned}
&Y_j = \theta_j + \eta_j,\ 1 \leq j \leq n, \quad \eta_j\ i.i.d. N(0, \sigma^2 I_d)\\
&\theta_j = \theta_-,\ \forall j \in A,\\
&\theta_j = \theta_+,\ \forall j \in A^c
\end{aligned}$$

Two classes K means algorithm

$$\hat{B} = ArgMin_{B \subset \{1,\ldots,n\}, n\epsilon \le \# B \le n(1-\epsilon)}$$

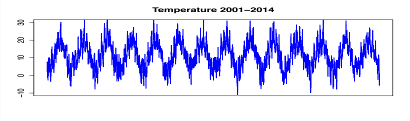
$$\left\{ \sum_{j \in B} \sum_{\ell \le d} (Y_j^\ell - \frac{1}{\# B} \sum_{j \in B} Y_j^\ell)^2 + \sum_{j \in B^c} \sum_{\ell \le d} (Y_j^\ell - \frac{1}{\# B^c} \sum_{j \in B^c} Y_j^\ell)^2 \right\}$$

Clustering with time scale :
There exits $0 < \tau < 1$ (change-point), and two regular vectors of $\boldsymbol{R}^d$: θ_- and θ_+ such that ,

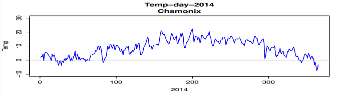
$$\theta_j = \theta_-,\ \forall j \leq n\tau$$
$$\theta_j = \theta_+,\ \forall j > n\tau$$

$$A = \{1, \ldots, n\tau\}$$

Example of time change classification



1. Only one spot (Chamonix)
2. The data are separated into different years $n = 14$
3. Each year has $d = 8760$ points of observation



1. We want to detect a change-point occurring at one precise year.

Key parameters

→ The change point τ

→ The energy of the change τ, $\Delta^2 := \|\theta_- - \theta_+\|^2$.

Two classes K means clustering algorithm in this context

$$\hat{\tau} = ArgMin_{t \in]\epsilon, 1-\epsilon[}$$

$$\left\{ \sum_{j \le nt} \sum_{\ell \le d} (Y_j^\ell - \frac{1}{nt} \sum_{j \le nt} Y_j^\ell)^2 + \sum_{j \ge nt+1} \sum_{\ell \le d} (Y_j^\ell - \frac{1}{n(1-t)} \sum_{j \ge nt+1} Y_j^\ell)^2 \right\}$$

Change point : Questions

- Our goal is to estimate τ and θ_-, θ_+.
- Does smoothing help ? How ?
- Sparsity conditions ?
- What are the different rates of convergence ?
- How to smooth optimally ?

Smoothing : simplified sparsity assumptions

For $s > 0$, we define

$$\Theta(s, L) := \{\theta \in \boldsymbol{R}^d, \ \sup_{K \in \mathbb{N}^*} K^{2s} \sum_{k \geq K} (\theta^k)^2 \leq L^2\}.$$

We will suppose that θ_- and θ_+ are in $\Theta(s, L)$.
$\rightarrow$ Again, this kind of sparsity reflects an ordering in the importance of the coefficients : the first ones are supposedly more important than the last ones.
$\rightarrow$ Possible extensions to other kind of sparsity like for $q < 1$,

$$\Theta(q, L) := \{\theta \in \boldsymbol{R}^d, \ \sum_k |\theta^k|^q \leq L\}.$$

For $1 \leq T \leq d$, let us consider
$\rightarrow$ T smooth data : $Y_j(T) = (Y_j^1, \ldots, Y_j^T)$ instead of
$Y_j = Y_j(d) = (Y_j^1, \ldots, Y_j^d)$,

$$\hat{\tau}(T) = ArgMin_{t \in]\epsilon, 1-\epsilon[}$$
$$\left\{ \sum_{j \leq nt} \sum_{\ell \leq T} (Y_j^\ell - \frac{1}{nt} \sum_{j \leq nt} Y_j^\ell)^2 + \sum_{j \geq nt+1} \sum_{\ell \leq T} (Y_j^\ell - \frac{1}{n(1-t)} \sum_{j \geq nt+1} Y_j^\ell)^2 \right\}$$

Misclassification rate

1. How big is $|\hat{\tau}(T) - \tau|$?
   In a general context : $Max\{\#\{\hat{A}^c \cap A\}, \#\{\hat{A} \cap A^c\}\}$?

2. How does this depend on T, s, $\Delta^2 = \|\theta_- - \theta_+\|^2$?

Proposition

Under the conditions above ($\Theta(s, L)$). If we stop the observation at $T \leq d : Y_j(T) = (Y_j^1, \ldots, Y_j^T)$, and we assume that there exists a constant R

$$\Delta^2 \geq R[T^{-2s} \vee \frac{\sigma^2 T}{n}],$$

then there exists constants c_2, c_3, such that for any κ,

$$P(|\hat{\tau}(T) - \tau| \geq \kappa \frac{\sigma^2 T}{n\Delta^2}) \leq 2n[\exp\{-c_2 RT\} + \exp\{-c_3 \kappa T\}].$$

Corollary

Under the conditions above, for

$$T_s := [\frac{n}{\sigma^2}]^{\frac{1}{1+2s}},$$

if there exists R such that,

$$\Delta^2 \geq R[\frac{\sigma^2}{n}]^{\frac{2s}{1+2s}},$$

then there exists constants c_2, c_3, such that for any κ,

$$P(|\hat{\tau}(T) - \tau| \geq \kappa[\frac{n}{\sigma^2}]^{\frac{-2s}{1+2s}}\Delta^{-2}) \leq 2n[\exp\{-c_2 R T_s\} + \exp\{-c_3 \kappa T_s\}].$$

Comments

$$\Delta^2 = \|\theta_- - \theta_+\|^2 \geq RT^{-2s} \vee \frac{\sigma^2 T}{n}, \; Rate \; \frac{\sigma^2 T}{n\Delta^2}$$

1. It is natural that the rate of convergence for τ is decreasing in Δ.
2. Advantage to smoothing : the rate is better
3. The conditions on Δ are less readable.
4. Condition $\Delta^2 \gtrsim [T^{-2s}]$, is necessary for identifiability : otherwise $\sum_{l \leq T}(\theta_-^l - \theta_+^1)^2$ may be arbitrarily close to zero, leading to a model on the $Y_j(T)$'s observations in which τ has no proper meaning and cannot be estimated.
5. Condition : $\Delta^2 \gtrsim [\frac{\sigma^2 T}{n}]$ is necessary for the MLE to converge.

Comparison

$$\Delta^2 = \|\theta_- - \theta_+\|^2 \geq RT^{-2s} \vee \frac{\sigma^2 T}{n}, \; Rate \; \frac{\sigma^2 T}{n\Delta^2}$$

1. In fact, the conditions on Δ^2 are less restrictive, with better rates, as soon as T decreases subject to the condition $T^{-2s} \lesssim \Delta^2$.
2. This leads to minimize $\frac{\sigma^2 T}{n}$ subject to $T^{-2s} \lesssim \Delta^2$ $\rightarrow T_{opt} = T_s := [\frac{n}{\sigma^2}]^{\frac{1}{1+2s}}$

Corollary

Under the conditions above, for

$$T_s := [\frac{n}{\sigma^2}]^{\frac{1}{1+2s}},$$

if there exists R such that,

$$\Delta^2 \geq R[\frac{\sigma^2}{n}]^{\frac{2s}{1+2s}},$$

then there exists constants c_2, c_3, such that for any κ,

$$P(|\hat{\tau}(T) - \tau| \geq \kappa[\frac{n}{\sigma^2}]^{\frac{-2s}{1+2s}}\Delta^{-2}) \leq 2n[\exp\{-c_2 R T_s\} + \exp\{-c_3 \kappa T_s\}].$$

- $$\Delta^2 \gtrsim [\frac{n}{\sigma^2}]^{\frac{-2s}{1+2s}}, \qquad \text{Rate } [\frac{n}{\sigma^2}]^{\frac{-2s}{1+2s}} \Delta^{-2}$$
- Rate and conditions could seem quite poor, but observe that very often σ^2 is of the form $\frac{\sigma_0^2}{d}$.
- In this case

$$\Delta^2 \gtrsim [\frac{nd}{\sigma_0^2}]^{\frac{-2s}{1+2s}}, \qquad \text{Rate } [\frac{nd}{\sigma_0^2}]^{\frac{-2s}{1+2s}} \Delta^{-2}$$

$$T_{opt} = T_s := [\frac{nd}{\sigma_0^2}]^{\frac{1}{1+2s}}$$

Choice of T : on-line ? off-line ?

- In particular case where σ^2 is of the form $\frac{\sigma_0^2}{d}$ the optimal smoothing is

$$T_{opt} = T_s := [\frac{nd}{\sigma_0^2}]^{\frac{1}{1+2s}}$$

  This proves that any (on-line) adaptive smoothing on each individual signal Y_j (thresholding or whatever) would give a rate -at best- of the form :

$$T_{opt} = T_s := [\frac{d}{\sigma_0^2}]^{\frac{1}{1+2s}}$$

  $\rightarrow$ loosing the factor n in the rate of misclassification.

- Meaning that the adaptive smoothing needs to be performed globally (off-line)

Adaptative choice for T

Form the following (off-line) pseudo-data in $\mathbb{R}^d$: $Z(1)$

$$Z^\ell(1) = \frac{1}{n}\sum_{j=1}^{n} Y_j^\ell - \frac{2}{n}\sum_{j=1}^{n/2} Y_j^\ell, \ell = 1, \dots, d$$

It has as mean

$$(1-\tau)[\theta_+ - \theta_-]\mathbb{I}\{\tau \geq 1/2\} + \tau[\theta_+ - \theta_-]\mathbb{I}\{\tau < 1/2\},$$

Adaptative choice for T

Consider the Lepski smoothers (c is a tuning constant)

$$\hat{T} := \min\{k,\ \sum_{\ell=k'}^{l} [Z^{\ell}(1)]^2 \leq cl\frac{\sigma^2}{n}\log[d \vee n],\ \forall l \geq k' \geq k\},$$

Adaptative choice for T

Theorem

We assume that $\theta_{\pm}$, is in $\Theta(s, L)$. We suppose that there exists a contant $a > 0$ such that

$$\frac{n}{\sigma^2} \geq a \log d.$$

Then, if there exists a constant $R = R(L, \epsilon)$ such that

$$\|\theta_- - \theta_+\|^2 = \Delta^2,\ \Delta^2 \geq R[\frac{\sigma^2 \log[d \vee n]}{n}]^{\frac{2s}{1+2s}}, \tag{1}$$

then for any γ,

$$P(|\hat{\tau}(\hat{T}) - \tau| \geq \kappa[\frac{\sigma^2 \log[d \vee n]}{n}]^{\frac{2s}{1+2s}} \Delta^{-2}) \leq [d \vee n]^{-\gamma}. \tag{2}$$

Adaptation rate for θ_- and θ_+, case $\sigma^2 = \frac{\sigma_0^2}{d}$

$\rightarrow$ We first detect the change using the procedure above, using $\hat{T}$

$\rightarrow \hat{\tau} = \hat{\tau}(\hat{T})$.

$$\hat{\tau}(\hat{T}) = ArgMin_{t \in]\epsilon, 1-\epsilon[}$$
$$\left\{ \sum_{j \leq nt} \sum_{\ell \leq \hat{T}} (Y_j^\ell - \frac{1}{nt} \sum_{j \leq nt} Y_j^\ell)^2 + \sum_{j \geq nt+1} \sum_{\ell \leq \hat{T}} (Y_j^\ell - \frac{1}{n(1-t)} \sum_{j \geq nt+1} Y_j^\ell)^2 \right\}$$

$\rightarrow$ Then we estimate $\theta_\pm$, with the following procedure :

$$\hat{\theta}_\pm^k := \hat{\theta}_{\pm,k} I\{k \leq \hat{T}^*\}$$

$$\hat{\theta}_{-,k} := \frac{1}{n\hat{\tau}} \sum_{j=1}^{n\hat{\tau}} Y_j^k \quad \hat{\theta}_{+,k} := \frac{1}{n(1-\hat{\tau})} \sum_{j=n\hat{\tau}+1}^{n} Y_j^k$$

$$\hat{T}^* := \min\{k, \sum_{k+1}^{l} [\hat{\theta}_{-,l}]^2 \leq c(lep) l \frac{\sigma^2}{n} \log[d \vee n], \forall l \geq k+2\}.$$

Adaptation rate for θ_- and θ_+

Theorem

- *With the estimates defined above, then, for $0 < s$, with the property $\Theta(s, L)$, we have*

$$\mathbb{E}\|\hat{\theta}_\pm - \theta_\pm\|_2^2 \lesssim \{\frac{nd}{\log[n \vee d]}\}^{\frac{-2s}{1+2s}} \tag{3}$$

Minimax rate - No condition on Δ^2 needed.

How to choose the number of clusters ?

Many methods already in the literature :
Calinsky et al. 1974, Gap Statistic Friedman et al. 2000, ... Most of them based on :

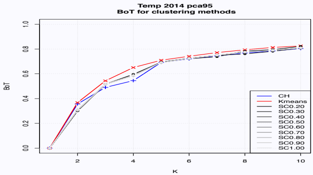
Variance Decomposition :
$T = W_k + B_k$

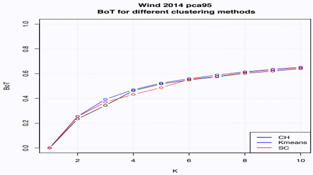
Total	$T = \frac{1}{n}\sum_i \|\|X_i - \bar{X}\|\|^2$
Between	$B_k = \frac{1}{n}\sum_k n_k \|\|\bar{X}_k - \bar{X}\|\|^2$
Within	$W_k = \frac{1}{n}\sum_k \sum_{i_k}^{n_k} \|\|X_k(i_k) - \bar{X}_k\|\|^2$

Quantification/ modeling indicator ratio :

$$\boxed{\rho_k = \frac{B_k}{T} \in [0,1]}$$

k_0 number of clusters :
with $\Delta_k = \rho_{k+1} - \rho_k$
$k_0 = \arg\min\{k,\ \Delta_k < 5\%\}$





over 14 years, for different temporal aggregation levels

Data : 14 x one year of data,
Kmeans algorithm with $\rho_k < 5\%$ criteria
Temperature :

	day (365)	week (52)	month (12)
PCA 95%	5 (0)	4.9 (0.2)	4.7 (0.4)
NP Reg. Trigo	5 (0)	4.8 (0.4)	4.7 (0.4)
NP Reg. Haar	5 (0)	4.8 (0.4)	4.7 (0.4)

Wind :

	day (365)	week (52)	month (12)
Pca 90%	4.15 (0.3)	4.23 (0.4)	4.31 (0.4)
NP Reg. Trigo	4.15 (0.3)	4 (0)	4.08 (0.2)
NP Reg. Haar	4.23 (0.4)	4.31 (0.4)	4.15 (0.3)