Smooth Clustering for high dimensional data

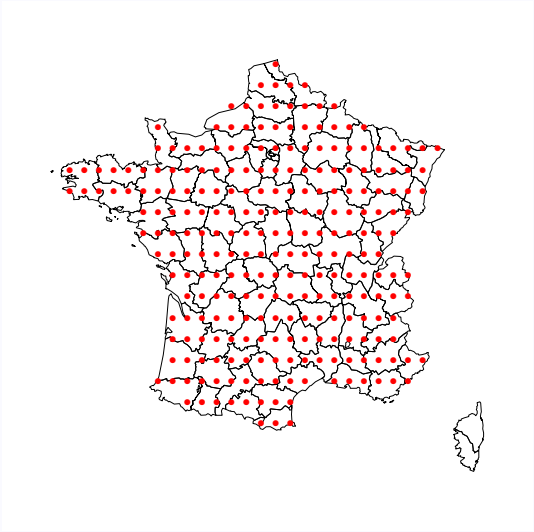
Dominique Picard

Université Paris-Diderot
LPMA

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

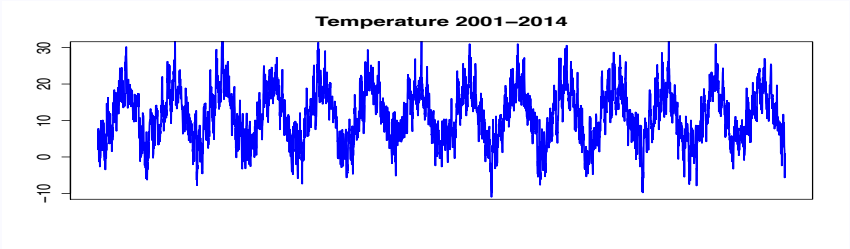
université
PARIS
DIDEROT

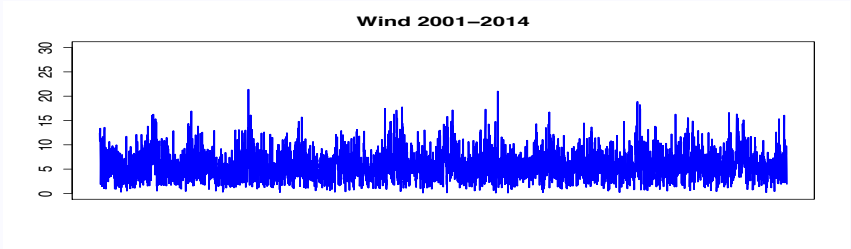
Arpege French Meteorological data



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)
- $n << d$



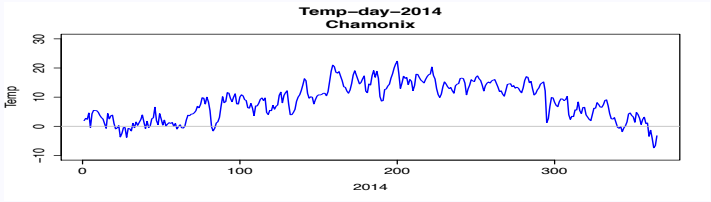


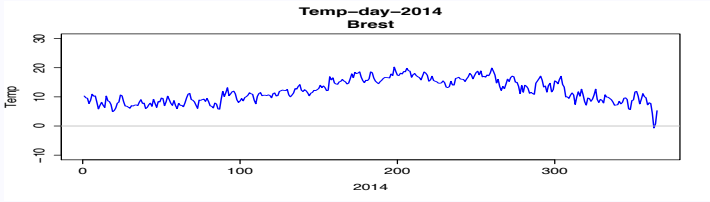
Objective and Questions :

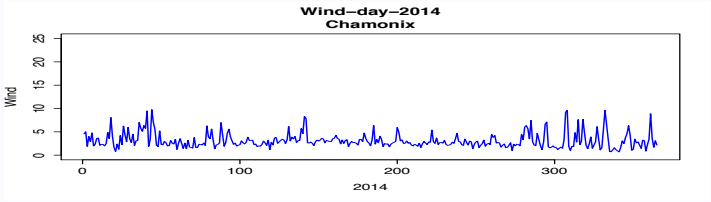
Goals

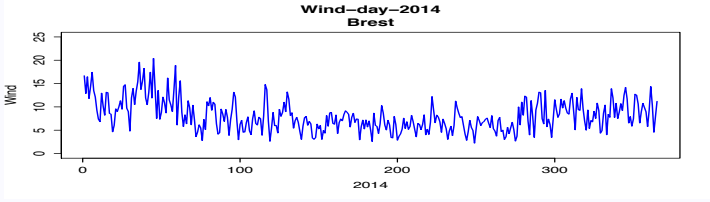
- Segmentation of the country into regions using meteorological data
- Temperature and/or Wind
- Study the Between Year variability

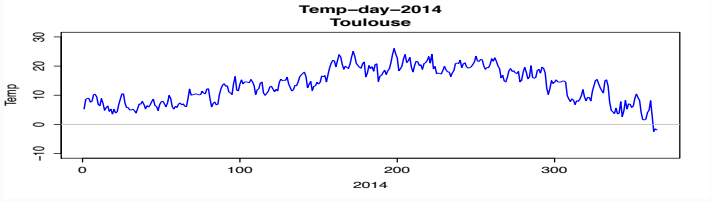
Wind and Temperature spots for 2014

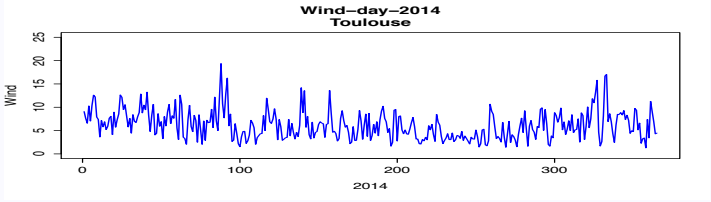






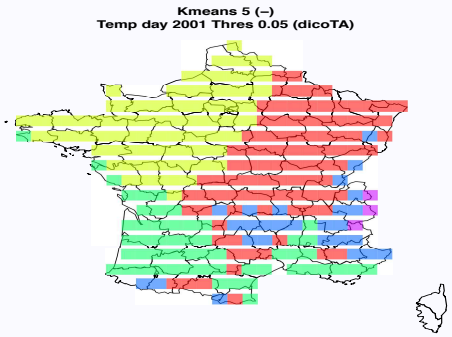


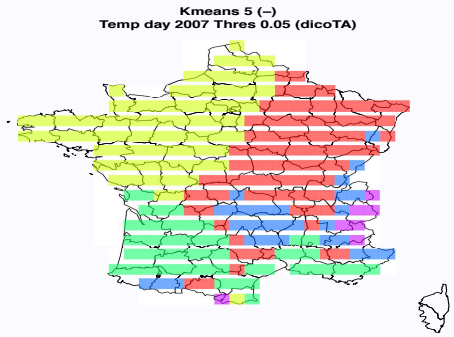


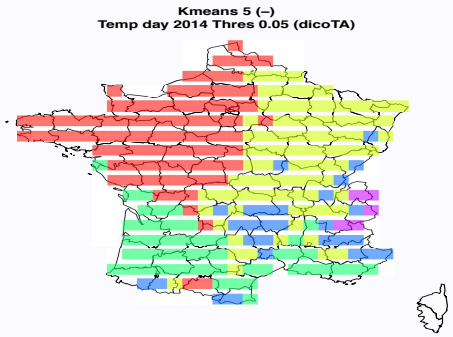


Segmentation for 2001, 2007, 2014 daily data

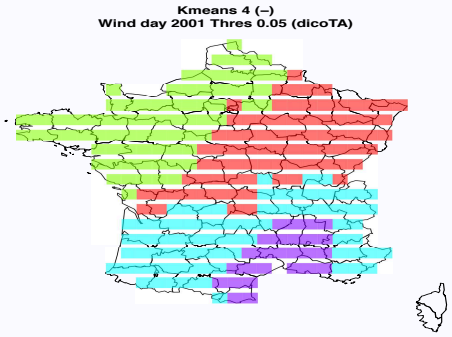
Temperature

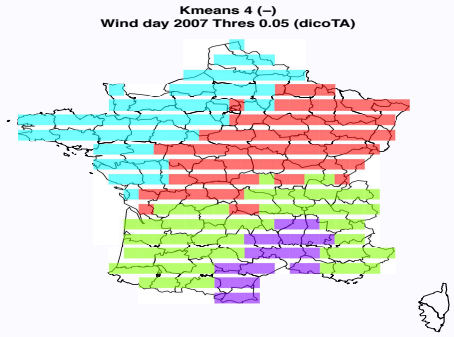


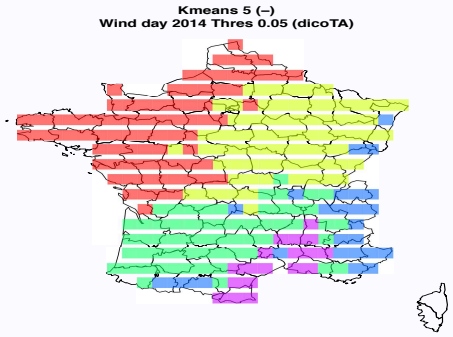




Wind







3 steps

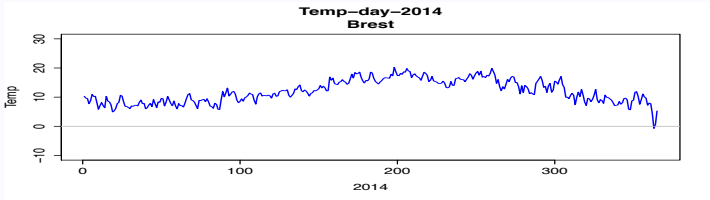
There are generally 3 important steps (linked in fact) :

1. Representation of the data
2. Smoothing
3. Selection procedure

Representation of the data

- Ad-hoc representations
- PCA
- Functional representation
- Kernel clustering
- Spectral clustering
- ...

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.

PCA-reduction

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, $T << d$:

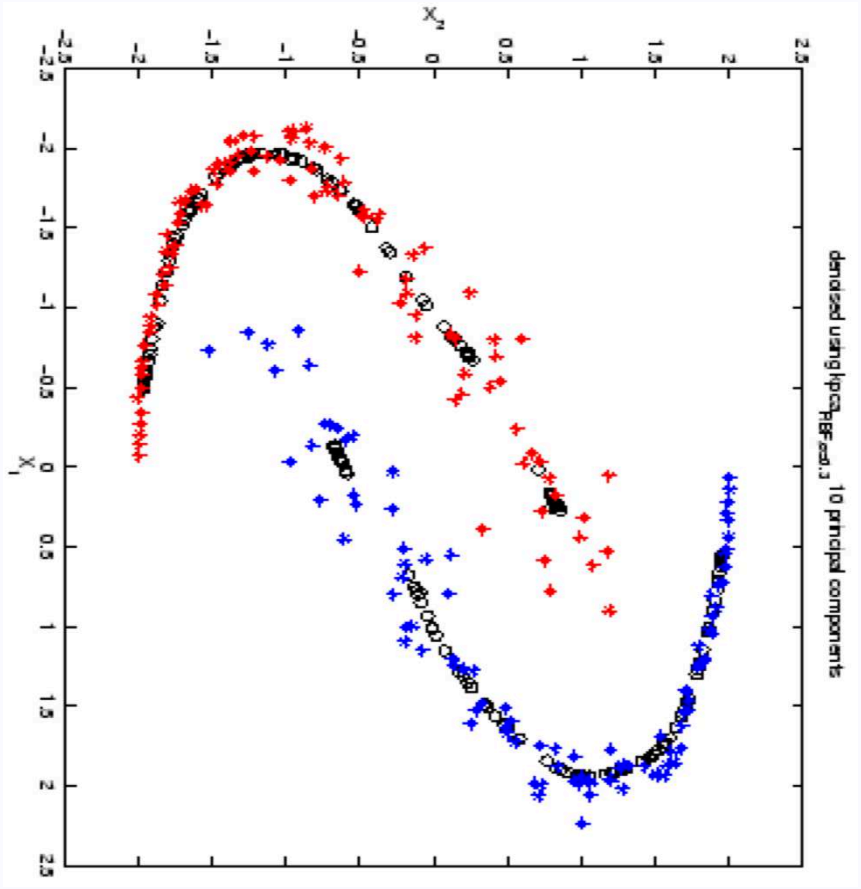
$$\boxed{Z = XU_T}$$

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

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

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

PCA might no be adapted

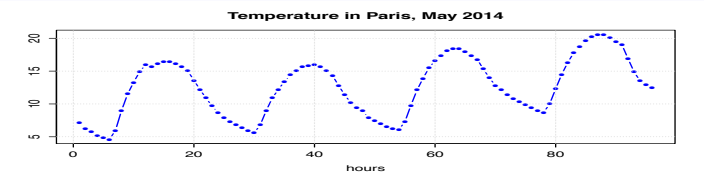


Functional smoothing

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 $\frac{||\hat{X}_{(j_0)}^i||^2}{||X^i||^2} \geq T_{NP} (= 0.95)$.

Selection procedure : Kmeans Clustering

Choose k the number of clusters

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

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

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

Best prediction

Questions

1. What is better : raw data (d) or smoothing ($T \leq d$) ?
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

Two classes model

We observe $Y_1, \ldots, Y_n$ n independent signals.
Each signal is observed discretely, i.e. $Y_j = (Y_j^1, \ldots, Y_j^d)$,
Gaussian clustering :
There exists a set $A \subset \{1, \ldots, n\}$, 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

$$\hat{B} = ArgMin_{B \subset \{1,\ldots,n\},}$$

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

Simplified two classes model : time change classification

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

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

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

Two classes K means clustering algorithm in this context

$$\hat{\tau} = ArgMin_{t\in(0,1)}$$

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

Change point : Questions

- Find the rate of convergence for τ using k-means.
- 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

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

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. (PCA, functional representations)
$\rightarrow$ Possible extensions to other kind of sparsity like for $q < 1$,

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

Clustering algorithm : MLE - Kmeans

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
$Y_j = Y_j(d) = (Y_j^1, \ldots, Y_j^d)$,

$$\hat{\tau}(T) = \text{arg min}_{k \in \{2,\ldots,n-2\}} \sum_{i=1}^{k} \sum_{j=1}^{T} \left(Y_i^j - \frac{1}{k} \sum_{i=1}^{k} Y_i^j \right)^2 +$$

$$\sum_{i=k+1}^{n} \sum_{j=1}^{T} \left(Y_i^j - \frac{1}{n-k} \sum_{i=k+1}^{n} Y_i^j \right)^2.$$

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, $\Delta^2 = \|\theta_- - \theta_+\|^2$?

Change point classification rate

$$\Delta^2 := \sum_{j=1}^{d} (\theta_-^j - \theta_+^j)^2 = \|\theta_+ - \theta_-\|^2.$$

We also define, for $T \leq d$,

$$\Delta_T^2 := \sum_{j=1}^{T} (\theta_-^j - \theta_+^j)^2, \quad \Psi_n(T, \Delta_T) = [\frac{\sigma^2}{n\Delta_T^2} \vee \frac{\sigma^4 T}{(n\Delta_T^2)^2}].$$

Proposition

Let us assume conditions [edge-out] and [Gaussian errors]. For any $\gamma > 0$ there exists constants $\kappa(\gamma, \epsilon)$, and $c(\gamma, \epsilon)$ such that,

$$\frac{n\Delta_T^2}{\sigma^2} \geq c(\gamma, \epsilon) \log n, \quad \lambda \geq \kappa(\gamma, \epsilon) \log n, \text{ then}$$

$$P\Big(|\hat{\tau}(T) - \tau| \geq \lambda \Psi_n(T, \Delta_T)\Big) \leq n^{-\gamma}.$$

Change point framework rate for τ

$$\Delta_T^2 := \sum_{j=1}^{T} (\theta_-^j - \theta_+^j)^2, \quad \Psi_n(T, \Delta_T) = [\frac{\sigma^2}{n\Delta_T^2} \vee \frac{\sigma^4 T}{(n\Delta_T^2)^2}].$$

- Note that for this result, no sparsity conditions on θ_+ and θ_- are needed.
- Using Korostelev and Lepski (MMS 2008), $\Psi_n(T, \Delta_T)$ is the minimax rate in this framework. Compared to their result, we are apparently loosing a logarithmic factor (contained in λ).
- But it is important to stress that in the paper above, the bound ϵ was supposed to be known, whereas our estimator $\hat{\tau}(T)$ is adaptive in ϵ.

Comments : Taking $T = d$

$$\Delta^2 := \sum_{j=1}^{d} (\theta_-^j - \theta_+^j)^2 = \|\theta_+ - \theta_-\|^2. \quad \Psi_n(d, \Delta_d) = [\frac{\sigma^2}{n\Delta^2} \vee \frac{\sigma^4 d}{(n\Delta^2)^2}].$$

- This rate is 'typically' composed of two different regimes : a 'dimension-free one' $\frac{\sigma^2}{n\Delta^2}$, and a 'dimension-depending' $\frac{\sigma^4 d}{(n\Delta^2)^2}$ (deteriorating with the dimension).
- If $c(\gamma, \epsilon)\frac{\sigma^2 \ln(n)}{n} \leq \Delta^2 < \frac{\sigma^2 d}{n}$, the rate of convergence is $\frac{\sigma^4 d}{(n\Delta^2)^2}$,
- if $\Delta^2 \geq \frac{\sigma^2 d}{n} \vee c(\gamma, \epsilon)\frac{\sigma^2 \ln(n)}{n}$, it is $\frac{\sigma^2}{n\Delta^2}$.
  - Taking $T = d$ (so raw data), allows to obtain the best rate $\frac{\sigma^2}{n\Delta^2}$. Taking a smaller T could lead to a reduction of Δ_T damaging the rate.
  - However this condition is quite restrictive when d is large
- Try to replace $\Psi_n(T, \Delta_T)$ by $\Psi_n(T, \Delta) = [\frac{\sigma^2}{n\Delta^2} \vee \frac{\sigma^4 T}{(n\Delta^2)^2}]$.
- Possible using regularity assumptions.

Comments

$$\Delta^2 := \sum_{j=1}^{d} (\theta_-^j - \theta_+^j)^2 = \|\theta_+ - \theta_-\|^2. \quad \Psi_n(T, \Delta_T) = [\frac{\sigma^2}{n\Delta_T^2} \vee \frac{\sigma^4 T}{(n\Delta_T^2)^2}].$$

- Without assumptions on the behavior of the parameters θ_+ and θ_-, there is not much to hope about the way Δ_T is increasing in T. $[\Theta(s, L)] \implies$ for T such that $\Delta^2 \geq 8L^2 T^{-2s}$, then Δ_T and Δ are comparable, in the sense that $\Delta_T^2 \geq \Delta^2/2$.
- If Δ_T and Δ are comparable, then $\Psi_n(T, \Delta_T) \sim \Psi_n(T, \Delta)$ is composed of two regimes
  - a good one $\frac{\sigma^2}{n\Delta^2}$ for $T \leq \frac{n\Delta^2}{\sigma^2}$,
  - and a slow one $\frac{\sigma^4 T}{(n\Delta^2)^2}$ for larger T's.

Change point framework rate for τ

Theorem

We assume conditions [edge-out], and $[\Theta(s, L)]$.
For any $\gamma > 0$, *there exist constants* $\kappa(\gamma, \epsilon)$ *and* $c(\gamma, \epsilon)$ *such that, if*

$$\Delta^2 \geq \left[2c(\gamma, \epsilon) \frac{\sigma^2 \ln(n)}{n} \vee 8L^2 T^{-2s} \right], \quad \lambda \geq \kappa(\gamma, \epsilon) \ln(n),$$

then

$$P\Big(|\hat{\tau}(T) - \tau| \geq \lambda \Psi_n(T, \Delta) \Big) \leq n^{-\gamma}.$$

If, now,

$$\Delta^2 \geq \left[2c(\gamma, \epsilon) \frac{\sigma^2 \ln(n)}{n} \vee 8L^2 T^{-2s} \vee \frac{\sigma^2 T}{n} \right], \quad \lambda \geq \kappa(\gamma, \epsilon) \ln(n), \tag{1}$$

$$P\Big(|\hat{\tau}(T) - \tau| \geq \lambda \frac{\sigma^2}{n\Delta^2} \Big) \leq n^{-\gamma}.$$

Corollary

Optimizing in T leads to $T_{opt} \sim T_s := \left(\frac{8L^2 n}{\sigma^2}\right)^{\frac{1}{1+2s}}$.

Corollary

Under the conditions above, for any $\gamma > 0$, there exist constants $\kappa(\gamma, \epsilon)$ and $c(\gamma, \epsilon)$ such that, if

$$\Delta^2 \geq \left[2c(\gamma,\epsilon)\frac{\sigma^2 \ln(n)}{n} \vee \left(\frac{\sigma^2}{n}\right)^{\frac{2s}{1+2s}} \left(8L^2\right)^{\frac{1}{1+2s}}\right], \quad \lambda \geq \kappa(\gamma,\epsilon)\ln(n), \tag{2}$$

$$P\left(|\hat{\tau}(T_s) - \tau| \geq \lambda \frac{\sigma^2}{n\Delta^2}\right) \leq n^{-\gamma}.$$

Discussion

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

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 can damage the rate of misclassification.

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

Adaptative choice for T

1. First, using the complete data set (so off-line), we will create surrogate data, estimating a parameter β of regularity s. These data will be used to finding an optimal $\hat{T}$.
2. Estimating the regularity of a signal is impossible without important extraneous assumptions
3. But adaptive procedures are producing - Lepski's procedure- a smoothing parameter $\hat{T}$ such that $\hat{T} \leq T_s$, with overwhelming probability .
4. This is not enough in our case. However, fortunately, Lepski's procedure, also controls the bias of the procedure, assuring that $\Delta^2 \leq 2\Delta^2_{\hat{T}}$ with large probability.

Adaptative choice for T

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

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

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's smoother (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 θ_+ and θ_- belong to $\Theta(s, L)$. We suppose that there exists a constant $\alpha > 0$ such that

$$\frac{n}{\sigma^2} \geq \alpha \ln d.$$

$$\hat{T} := \min \left\{ k \geq 1 : \forall d \geq j \geq m \geq k, \sum_{\ell=m}^{j} (Z^\ell)^2 \leq C_{\mathcal{L}} j \frac{\sigma^2}{n} \ln(d \vee n) \right\}.$$

Then, for any $\gamma > 0$, there exist 2 constants $R(\gamma, \epsilon)$ and $\kappa(\gamma, \epsilon)$ such that if $\Delta^2 \geq R \left(\frac{\sigma^2 \ln(d \vee n)}{n} \right)^{\frac{2s}{1+2s}}$, and $\lambda \geq \kappa \log n$, then,

$$P\left(|\hat{\tau}(\hat{T}) - \tau| \geq \lambda \frac{\sigma^2}{n\Delta^2} \right) \leq n^{-\gamma}.$$

To do...

- Conditions ellipsoid $\rightarrow$ usual sparsity
- Change point $\rightarrow$ general clustering
- 2 classes $\rightarrow$ k classes
- Lower bounds
- Combine with dictionary search
- ...

-

Thank you for your attention

Recipe for proof

$$\hat{\tau}(T) = \mathsf{arg\ min}_{t \in \{\frac{2}{n}, \ldots, \frac{n-2}{n}\}} K^T(t).$$

$$K^T(t) = \min_{x_-, x_+} L(t, x_-, x_+) - L(\tau, 0, 0).$$

$$L(t, x_-, x_+) = \sum_{i=1}^{nt} \sum_{j=1}^{T} (Y_i^j - \theta_-^j - x_-^j)^2 + \sum_{i=nt+1}^{n} \sum_{j=1}^{T} (Y_i^j - \theta_+^j - x_+^j)^2.$$

$$K^T(t) = -\sum_{j=1}^{T} \sigma^2 V_j^2(t) - \sum_{j=1}^{T} \sigma^2 W_j^2(t) + \Delta_T^2 \frac{(nt - n\tau)n\tau}{nt} + 2N_1(t) - 2N_2(t),$$

How to choose the number of clusters ?

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

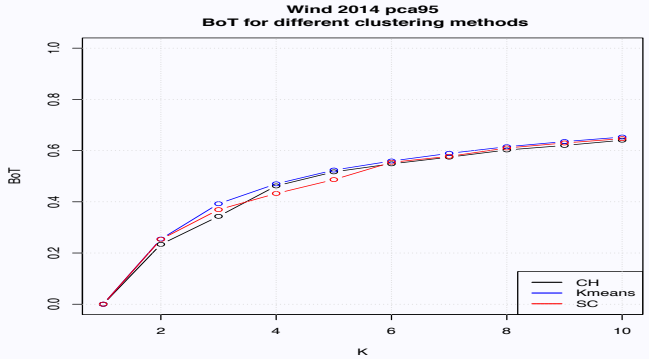
Many methods already in the literature : Calinsky et al. 1974, Gap Statistic Friedman et al. 2000, ... Most of them based on :

Variance Decomposition :
$T = W_k + B_k$

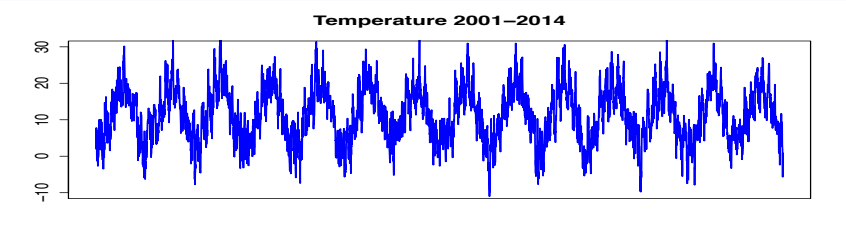
Total	$T = \frac{1}{n}\sum_i \|\|X_i - \bar{X}\|\|^2$
Between	$B_k = \frac{1}{n}\sum_k n_k \|\|\bar{X}_k - \bar{X}\|\|^2$
Within	$W_k = \frac{1}{n}\sum_k \sum_{i_k}^{n_k} \|\|X_k(i_k) - \bar{X}_k\|\|^2$

Quantification/ modeling indicator ratio :

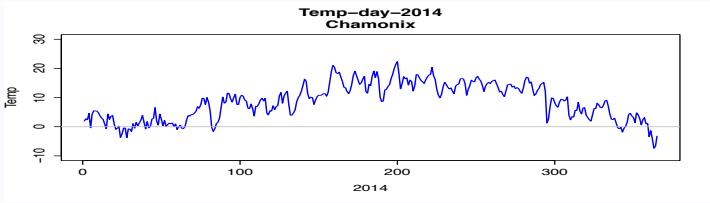
$$\boxed{\rho_k = \frac{B_k}{T} \in [0,1]}$$



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.

Stability of the number of clusters
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)