Consistent Change-point Detection with Kernels

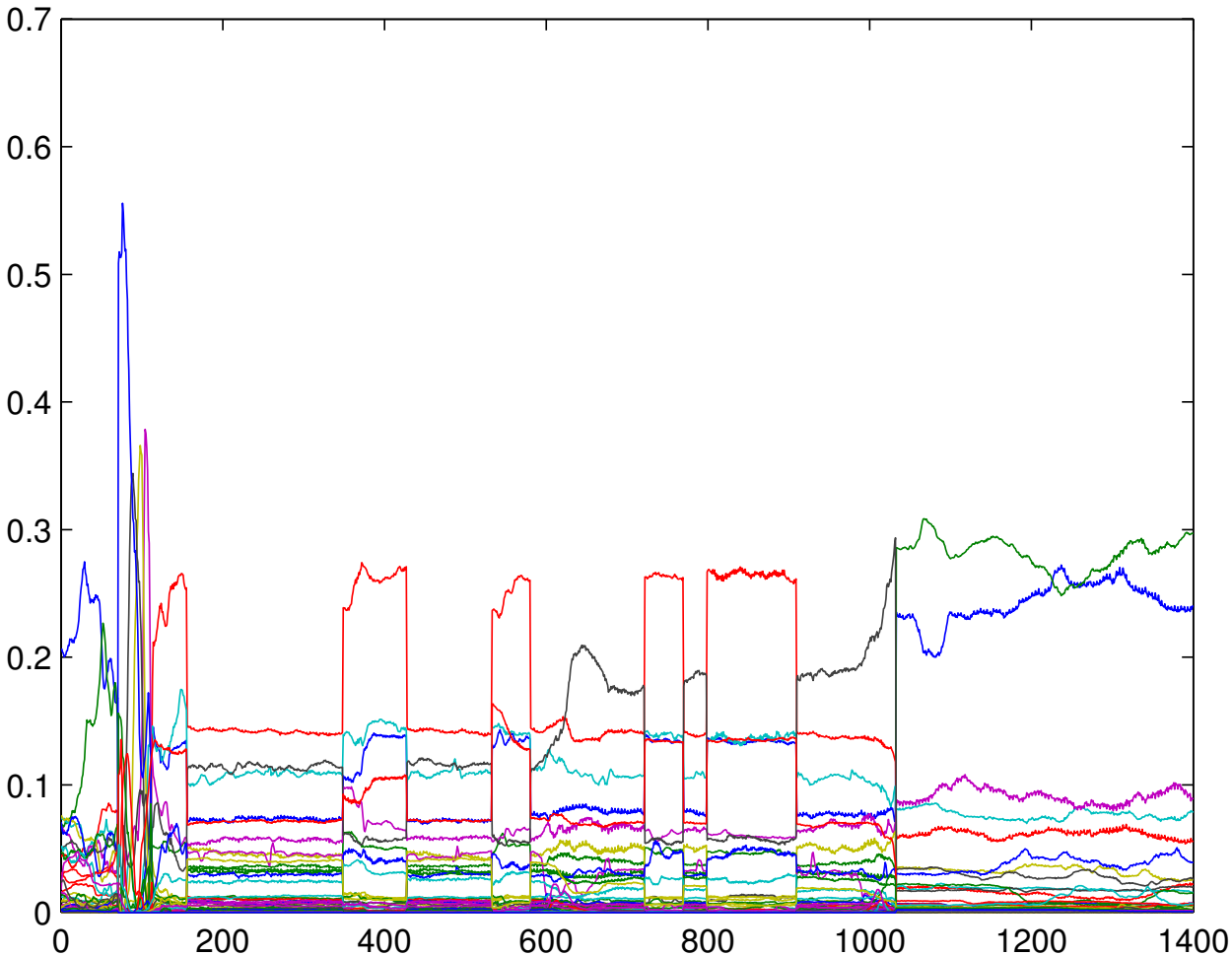
Damien Garreau [1] Sylvain Arlot [2]

[1]Inria, DI ENS

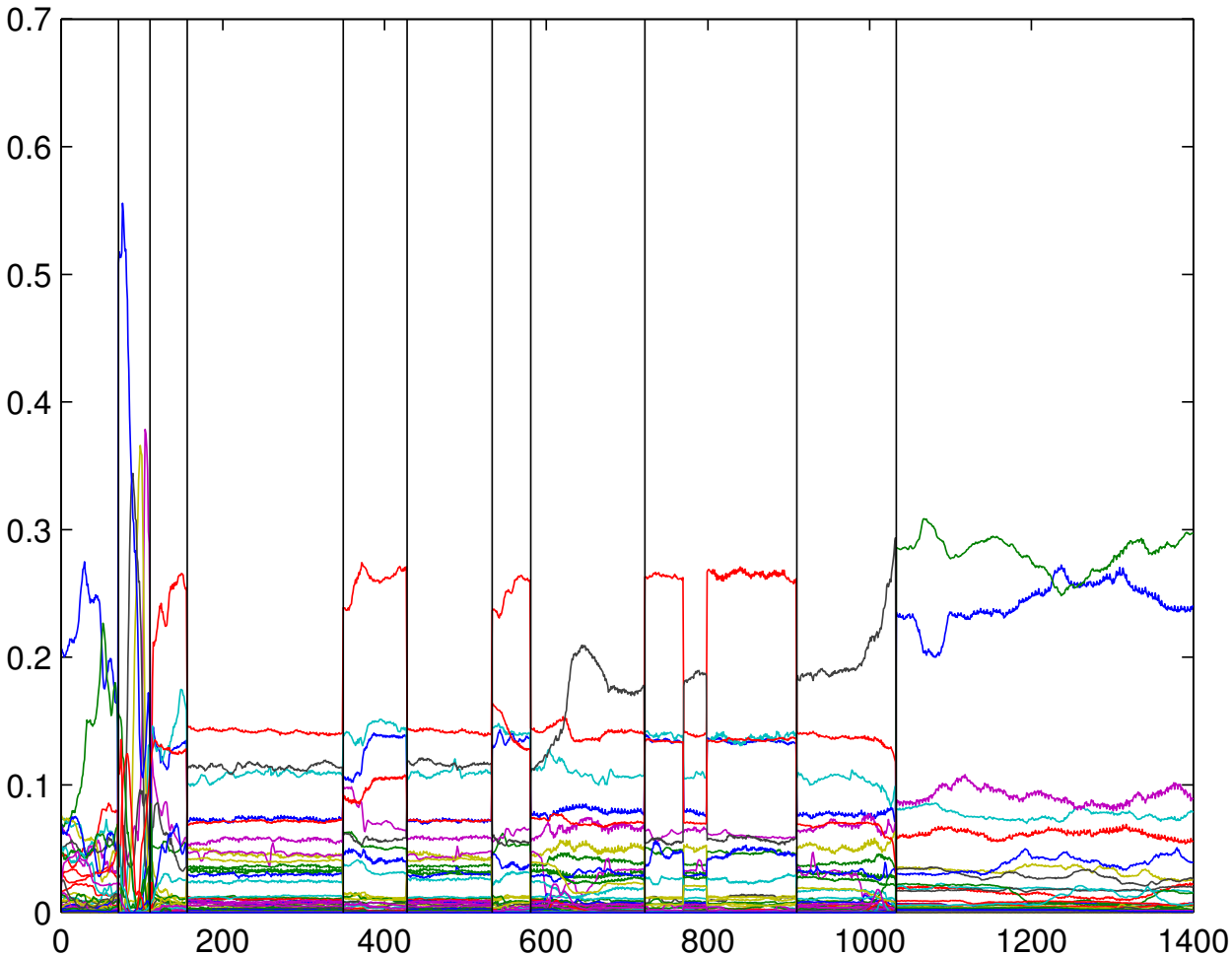
[2]Université Paris-Sud, Laboratoire de Mathématiques d'Orsay

April 6, 2016

An example: shot detection in a movie



An example: shot detection, cont.



Goals

We want to:

- detect abrupt changes in the *distribution* of the data
- deal with *interesting* (structured) data: each point is a curve, a graph, a histogram, a persistence diagram...

The change-point problem

- $\mathcal{X}$ arbitrary (measurable) set, $n < +\infty$, and $X_1, \ldots, X_n \in \mathcal{X}$ sequence of independent random variables.
- $\forall i \in \{1, \ldots, n\}$, P_{X_i} the distribution of X_i.

The change-point problem can be formalized as follows:

- Given $(X_i)_{1 \leq i \leq n}$, we want to find the locations of the abrupt changes in the sequence $P_{X_1}, \ldots, P_{X_n}$.

Notations

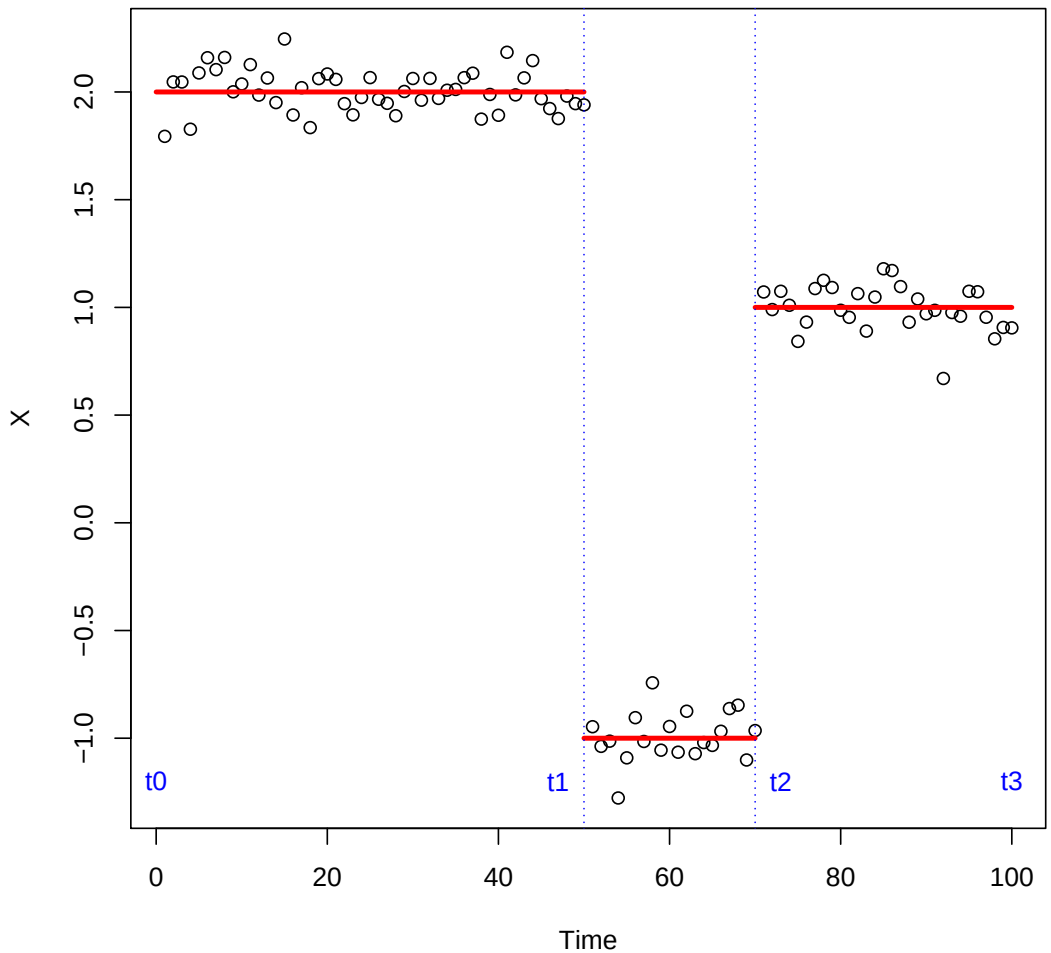
- Take any $D \in \{1, \dots, n+1\}$, the set of sequences of $D-1$ change-points is defined by

$$\mathcal{T}_n^D := \{(\tau_0, \dots, \tau_D) \in \mathbb{N}^{D+1},\ 0 = \tau_0 < \tau_1 < \cdots < \tau_D = n\}.$$

- $\tau_1, \dots, \tau_{D-1}$ are the *change-points*, τ is a *segmentation* of $\{1, \dots, n\}$ into D_τ segments.
- $\tau^\star$ the true segmentation, $D^\star = D_{\tau^\star}$ the true number of change-points.

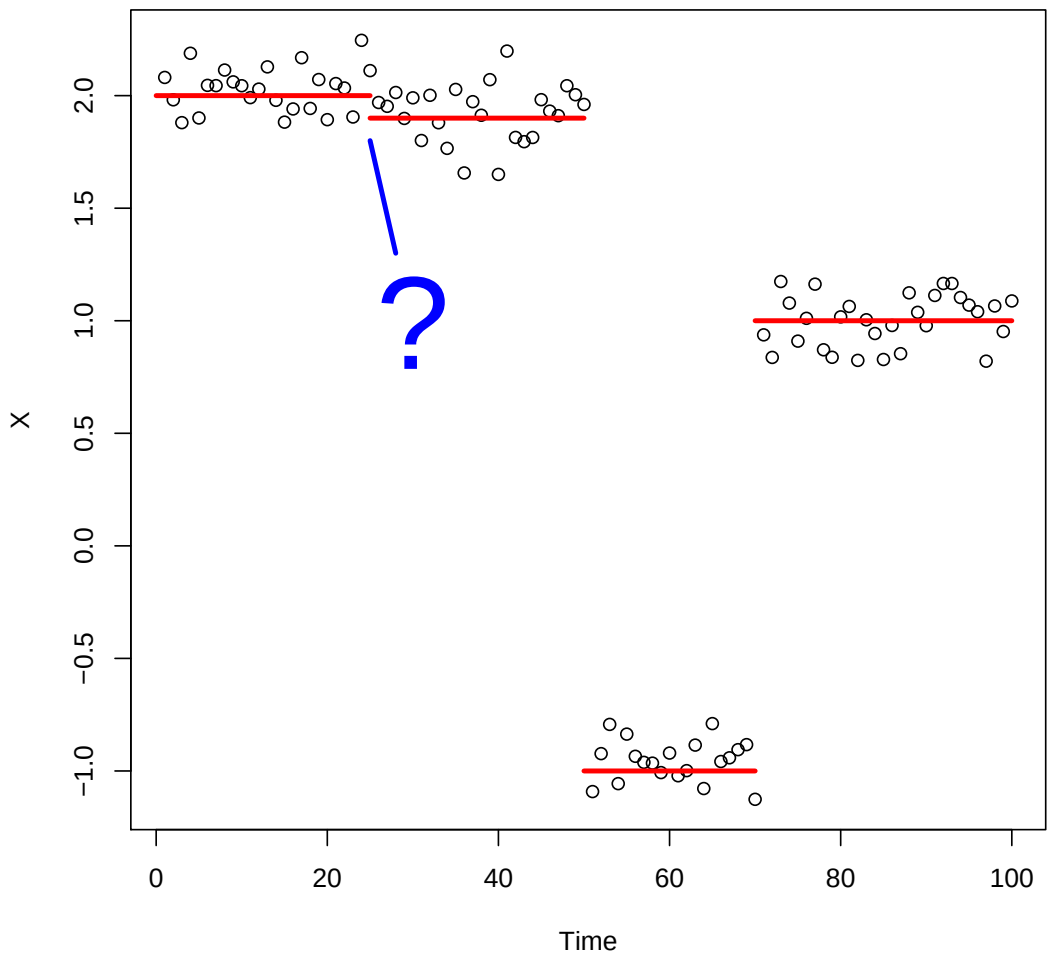
In pictures

Here, $\mathcal{X} = \mathbb{R}$, $D^\star = 3$, and $\tau^\star = (0, 50, 70, 100)$.



In pictures, cont.

It is not an easy question:



Summary

- With finite sample size, it is *not easy* to recover the true change-points in presence of noise.
- When $\mathcal{X} = \mathbb{R}^d$ and the changes occur in the first moments of the distribution, the problem has already received considerable attention, cf. [Basseville and Nikiforov, 1993].
- Kernel change-point detection can tackle more subtle changes / less conventional data.

Plan

Kernels: a quick reminder

- Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a semidefinite kernel.
- k is a measurable function s.t. $\forall x_1, \ldots, x_m \in \mathcal{X}$, the matrix $(k(x_i, x_j))_{1 \leq i,j \leq m}$ is *positive semi-definite*. Think *inner product*.
- Examples include
  - the *linear* kernel $k(x, y) = \langle x, y \rangle$,
  - the *Gaussian* kernel $k(x, y) = \exp(-\|x - y\|^2 / (2h^2))$,
  - the *histogram* kernel $k(x, y) = \sum_{k=1}^{p} \min(x_k, y_k)$,
  - ...

The kernel least-squares criterion

- Intuition: least-squares criterion

$$\widehat{\mathcal{R}}_n(\tau) := \frac{1}{n} \sum_{\ell=1}^{D} \sum_{i=\tau_{\ell-1}+1}^{\tau_\ell} (X_i - \overline{X}_{\tau_{\ell-1}+1,\tau_\ell})^2.$$

- Define

$$\widehat{\mathcal{R}}_n(\tau) := \frac{1}{n} \sum_{i=1}^{n} k(X_i, X_i) - \\ - \frac{1}{n} \sum_{\ell=1}^{D} \left[\frac{1}{\tau_\ell - \tau_{\ell-1}} \sum_{i=\tau_{\ell-1}+1}^{\tau_\ell} \sum_{j=\tau_{\ell-1}+1}^{\tau_\ell} k(X_i, X_j) \right].$$

- This is just a kernelized version, the two definitions coincide when $\mathcal{X} = \mathbb{R}$ and $k(x, y) = xy$.

Most important slide of the talk

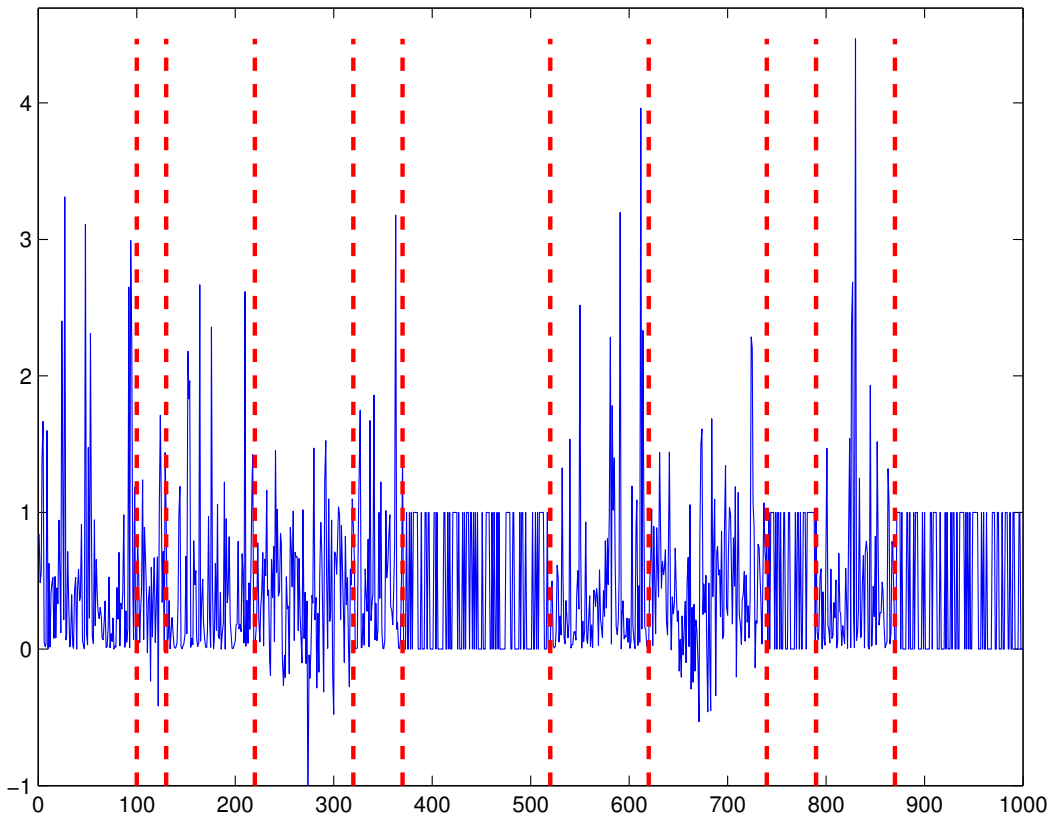
We investigate the properties of

$$\widehat{\tau} \in \underset{\tau \in \mathcal{T}_n}{\arg\min} \left\{ \overbrace{\widehat{\mathcal{R}}_n(\tau)}^{least-squares\ criterion} + \underbrace{\text{pen}(\tau)}_{penalty\ function} \right\},$$

where pen is a function increasing with D_τ.

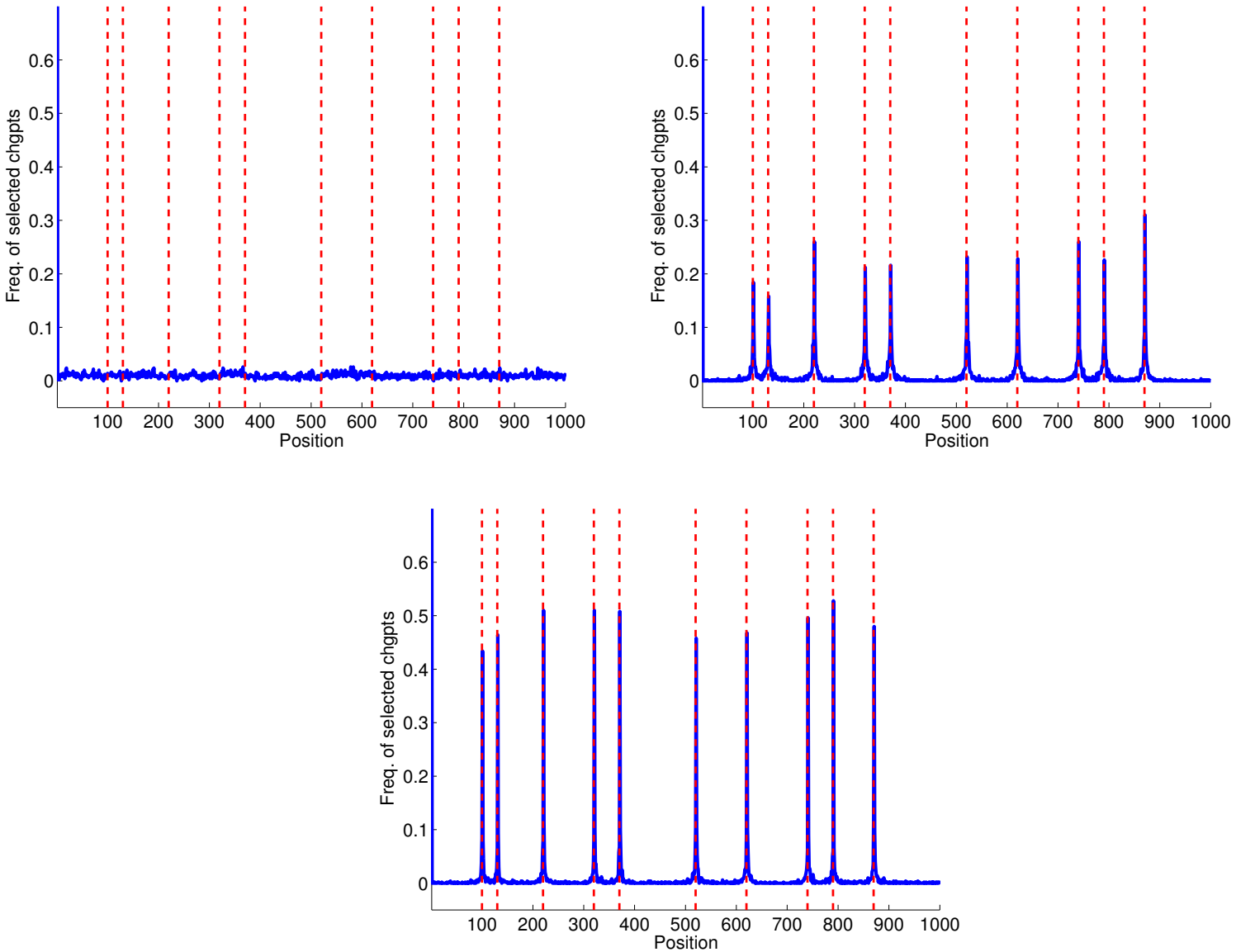
Constant mean and variance

Constant mean and variance: the distribution of X_i is chosen among $\mathcal{B}(0.5)$, $\mathcal{N}(0.5, 0.25)$ and $\Gamma(1, 0.5)$.



(courtesy of [Arlot et al., 2012])

Constant mean and variance, cont.



Linear, Hermite, and Gaussian kernels (courtesy of [Arlot et al., 2012]).

Plan

More notations

- Along with the kernel k comes a reproducing kernel Hilbert space $\mathcal{H}$ endowed with $\langle \cdot, \cdot \rangle_{\mathcal{H}}$.
- There exists a mapping $\Phi : \mathcal{X} \to \mathcal{H}$ such that, for any $x, y \in \mathcal{X}$, $k(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}}$.
- The algorithm is looking for breaks in the "mean" of $Y_i := \Phi(X_i) \in \mathcal{H}$.
- Whenever possible, define $\mu_i^\star$ the mean of Y_i; it satisfies

$$\forall g \in \mathcal{H}, \quad \langle \mu_i^\star, g \rangle_{\mathcal{H}} = \mathbb{E}\left[g(X_i) \right] = \mathbb{E}\left[\langle Y_i, g \rangle_{\mathcal{H}} \right].$$

- We write $Y_i = \mu_i^\star + \varepsilon_i$.

Hypothesis

- $\mathcal{H}$ is separable.
- Bounded data/kernel:

$$\exists M \in (0, +\infty), \quad \forall 1 \leq i \leq n, \quad k(X_i, X_i) \leq M^2. \qquad \text{(Db)}$$

- Finite variance:

$$\forall 1 \leq i \leq n, \quad v_i := \mathbb{E}\left[\|\varepsilon_i\|_{\mathcal{H}}^2\right] \leq V < +\infty. \qquad \text{(V)}$$

Under (Db), an oracle inequality has been proven.
$\longrightarrow$ See [Arlot et al., 2012] for the result.

Dimension selection, light version

- Assume that (Db) holds true;
- Suppose that $\mathrm{pen}(\cdot)$ is "large enough";
- Suppose that $\underline{\Delta}^2 \times \underline{\Gamma}$ is "large enough", where $\underline{\Delta} := \inf_{\mu_i^\star \neq \mu_{i+1}^\star} \|\mu_i^\star - \mu_{i+1}^\star\|_{\mathcal{H}}$ is the size of the smallest jump in $\mathcal{H}$, and $\underline{\Gamma}$ depends only on the geometry of $\tau^\star$;
- Then, with high probability, $D_{\widehat{\tau}} = D^\star$.

If k is characteristic, we recover *all* the changes in P_{X_i}.

Dimension selection

Theorem

Let y be a positive number. Assume that (Db) holds true and that

$$\forall \tau \in \mathcal{T}_n, \quad \text{pen}(\tau) = \frac{C D_\tau M^2}{n} \left(1 + \sqrt{2 \left(4 + y + \log \frac{n}{D_\tau} \right)} \right)^2 ,$$

with $C \geq (2D^\star + 1)(5 + y + \log D^\star)$. Suppose that

$$\underline{\Delta}^2 \times \underline{\Gamma} \gtrsim \frac{C D^\star M^2}{n} \left(y + \log \frac{n}{D^\star} \right) .$$

Then $\mathbb{P}\left(D_{\widehat{\tau}} = D^\star \right) \geq 1 - \mathrm{e}^{-y}$.

Distance between segmentations

- We consider only segmentation with the same number of segments $D^\star$.
- Several possibilities, equivalent under assumptions regarding

$$\underline{\Lambda}_\tau := \frac{1}{n} \min_{\lambda \in \tau} |\lambda| \,.$$

- We focus on

$$\mathrm{d}_\infty(\tau^1, \tau^2) := \max_{1 \leq i \leq D^\star - 1} \left| \tau_i^1 - \tau_i^2 \right| .$$

Localization of the change-points, light version

- Assume that $D^\star$ is known and that (V) holds true.
- Take $\delta_n > 0$, and choose $\widehat{\tau}$ in

$$\mathcal{T}_n^{D^\star}(\delta_n) := \{\tau \in \mathcal{T}_n,\ \underline{\Lambda}_\tau \geq \delta_n,\ D_\tau = D^\star\}.$$

- Then, for any $0 < x < \underline{\Lambda}_{\tau^\star}$,

$$\mathbb{P}\left(\frac{1}{n}\mathrm{d}_\infty(\widehat{\tau}, \tau^\star) \geq x\right) \lesssim \frac{V}{nx}\left(\frac{1}{\delta_n} + \frac{1}{x}\right).$$

- This goes to 0 whenever $\delta_n \to 0$ and $n\delta_n \to +\infty$.

Localization of the change-points

Theorem

Assum that (V) *holds true. Take* $\delta_n > 0$ *and choose*

$$\widehat{\tau} \in \underset{\tau \in \mathcal{T}_n^{D^\star}(\delta_n)}{\arg\min} \left\{ \widehat{\mathcal{R}}_n(\tau) \right\}.$$

Suppose that $\delta_n \leq \underline{\Lambda}_{\tau^\star}$. *Then, for any* $0 < x \leq \underline{\Lambda}_{\tau^\star}$,

$$\mathbb{P}\left(\mathrm{d}_\infty(\widehat{\tau}, \tau^\star) \geq x \right) \lesssim \frac{V D^\star}{n x \underline{\Delta}^2} \left(\frac{1}{\delta_n} + \frac{(D^\star)^3 \overline{\Delta}^2}{x \underline{\Delta}^2} \right).$$

For instance: take $\delta_n = n^{-1/2}$: $\mathrm{d}_\infty(\hat{\tau}, \tau^\star) = o_P(n^{-1/2})$.

Plan

Take away message

- Kernelized version of the change-point detection procedure of [Lebarbier, 2005].
- Detection of changes in the distribution, not only the first moments.
- Possible to deal with structured data more efficiently.
- Under reasonable assumptions and for a class of penalty functions,
  - we dispose of an oracle inequality
  - the procedure is consistent
  - it recovers the true localization of the change-points

Future work

- Exchange the hypothesis and *still* prove our results (in progress);
- Tackle dependency structures within the X_is as in [Lavielle and Moulines, 2000];
- Learn how to choose the kernel;
- Find interesting data!

Thank you for your attention!

References I


Arlot, S., Celisse, A., and Harchaoui, Z. (2012).
Kernel change-point detection.
arXiv preprint arXiv:1202.3878.


Basseville, M. and Nikiforov, I. (1993).
Detection of abrupt changes: theory and application.
Prentice Hall Englewood Cliffs.


Lavielle, M. and Moulines, E. (2000).
Least-squares estimation of an unknown number of shifts in a time series.
Journal of time series analysis, 21(1):33–59.


Lebarbier, É. (2005).
Detecting multiple change-points in the mean of a Gaussian process by model selection.
Signal Proces., 85:717–736.

Bonus: elements of proof (dimension selection)

Main idea: $\forall \tau \in \mathcal{T}_n$ s.t. $D_\tau \neq D^\star$, w.h.p.,

$$\mathcal{R}(\tau) + \text{pen}(\tau) > \mathcal{R}(\tau^\star) + \text{pen}(\tau^\star).$$

Since $\widehat{\tau}$ minimizes $\mathcal{R}(\cdot) + \text{pen}(\cdot)$, $\widehat{D} = D^\star$ w.h.p..

Bonus: elements of proof (dimension selection)

Main idea: $\forall \tau \in \mathcal{T}_n$ s.t. $D_\tau \neq D^\star$, w.h.p.,

$$\mathcal{R}(\tau) + \mathrm{pen}(\tau) > \mathcal{R}(\tau^\star) + \mathrm{pen}(\tau^\star).$$

Since $\widehat{\tau}$ minimizes $\mathcal{R}(\cdot) + \mathrm{pen}(\cdot)$, $\widehat{D} = D^\star$ w.h.p..

$$\mathcal{R}(\tau) = \frac{1}{n}\left\|\mu^\star - \mu^\star_\tau\right\|^2 + \frac{2}{n}\left\langle \mu^\star - \mu^\star_\tau, \varepsilon \right\rangle - \frac{1}{n}\left\|\Pi_\tau \varepsilon\right\|^2 + \frac{1}{n}\left\|\varepsilon\right\|^2,$$

where $\mu^\star_\tau$ is the projection of $\mu^\star$ on the subset of $\mathcal{H}^n$ "constant on the segments of τ", i.e.,

$$F_\tau := \left\{ f \in \mathcal{H}^n,\ f_{\tau_{\ell-1}+1} = \cdots = f_{\tau_\ell}\ \forall 1 \leq \ell \leq D_\tau \right\}.$$

Bonus: elements of proof (dimension selection)

Main idea: $\forall \tau \in \mathcal{T}_n$ s.t. $D_\tau \neq D^\star$, w.h.p.,

$$\mathcal{R}(\tau) + \text{pen}(\tau) > \mathcal{R}(\tau^\star) + \text{pen}(\tau^\star).$$

Since $\widehat{\tau}$ minimizes $\mathcal{R}(\cdot) + \text{pen}(\cdot)$, $\widehat{D} = D^\star$ w.h.p..

$$\mathcal{R}(\tau) = \frac{1}{n}\left\|\mu^\star - \mu^\star_\tau\right\|^2 + \frac{2}{n}\left\langle \mu^\star - \mu^\star_\tau, \varepsilon \right\rangle - \frac{1}{n}\left\|\Pi_\tau \varepsilon\right\|^2 + \frac{1}{n}\left\|\varepsilon\right\|^2,$$

where $\mu^\star_\tau$ is the projection of $\mu^\star$ on the subset of $\mathcal{H}^n$ "constant on the segments of τ", i.e.,

$$F_\tau := \left\{ f \in \mathcal{H}^n,\ f_{\tau_{\ell-1}+1} = \cdots = f_{\tau_\ell}\ \forall 1 \leq \ell \leq D_\tau \right\}.$$

We are reduced to show that if $D_\tau \neq D^\star$, w.h.p.,

$$\frac{1}{n}\left\|\mu^\star - \mu^\star_\tau\right\|^2 + \frac{2}{n}\left\langle \mu^\star - \mu^\star_\tau, \varepsilon \right\rangle + \frac{1}{n}\left\|\Pi_{\tau^\star} \varepsilon\right\|^2 - \frac{1}{n}\left\|\Pi_\tau \varepsilon\right\|^2 > \\ > \text{pen}(\tau^\star) - \text{pen}(\tau).$$

Elements of proof, cont.

We control each term $\forall \tau$, with probability $1 - e^{-x}$:

- the linear term: $|\langle \mu^\star - \mu_\tau^\star, \varepsilon \rangle| \lesssim \theta \left\| \mu^\star - \mu_\tau^\star \right\|^2 + \frac{1}{\theta} M^2 x$,
- the quadratic term:
  $\left\| \Pi_\tau \varepsilon \right\|^2 - \mathbb{E}\left[\left\| \Pi_\tau \varepsilon \right\|^2 \right] \lesssim (x + \sqrt{x D_\tau}) M^2$,
- $\mathrm{pen}(\tau) - \mathrm{pen}(\tau^\star)$ via technical lemmas.

Elements of proof, cont.

Union bound,

$$\mathbb{P}\left(\bigcap_{\tau\in\mathcal{T}_n}\Omega_\tau\right)\geq 1-4\sum_{\tau\in\mathcal{T}_n}\mathrm{e}^{-x_\tau}\,.$$

Recall $\#\,\mathcal{T}_n^D=\binom{n-1}{D-1}\leq(n\,\mathrm{e}\,/d)^d$.

$$\begin{aligned}\sum_{\tau\in\mathcal{T}_n}\mathrm{e}^{-x_\tau}&\leq\sum_{d=1}^{n}\exp\left(d+d\log n-d\log d-4d-dy-d\log n+d\log d\right)\\&=\sum_{d=1}^{n}\exp(d(-3-y))=\sum_{d=1}^{n}\left(\frac{\mathrm{e}^{-y}}{\mathrm{e}^3}\right)^d\\&=\mathrm{e}^{-3-y}\,.\frac{1-(\mathrm{e}^{-3-y})^n}{1-\mathrm{e}^{-3-y}}\leq\mathrm{e}^{-y}\,/4.\end{aligned}$$

This part *fails* if we do not have bounded assumption.

Bonus: a word about the concentration result

Lemma

For every $x > 0$, with probability $\geq 1 - \mathrm{e}^{-x}$,

$$\|\Pi_\tau \varepsilon\|^2 - \mathbb{E}\left[\|\Pi_\tau \varepsilon\|^2\right] \leq \frac{14M^2}{3}\left(x + 2\sqrt{2xD_\tau}\right).$$

Proof.

Write $\|\Pi_\tau \varepsilon\|^2$ as $\sum_{1 \leq \ell \leq D_\tau} T_\ell$, a sum of independant random variables, where $T_\ell := \frac{1}{\tau_\ell - \tau_{\ell-1}} \left\| \sum_{j=\tau_{\ell-1}+1}^{\tau_\ell} \varepsilon_j \right\|^2$. Apply Bernstein's inequality. The tricky part is to check that the moment conditions for Bernstein are satisfied. Idea: Write $\mathbb{E}\left[T_\ell^q\right]$ as an integral depending upon $\mathbb{P}\left(\left\| \sum_{j=\tau_{\ell-1}+1}^{\tau_\ell} \varepsilon_j \right\| \geq y\right)$ for which Pinelis-Sakhanenko's inequality gives an upper-bound. □