Consistent change-point detection with kernels

Sylvain Arlot[1] (joint works with Alain Celisse[2], Damien Garreau[3] & Zaïd Harchaoui[4])

[1]Université Paris-Sud

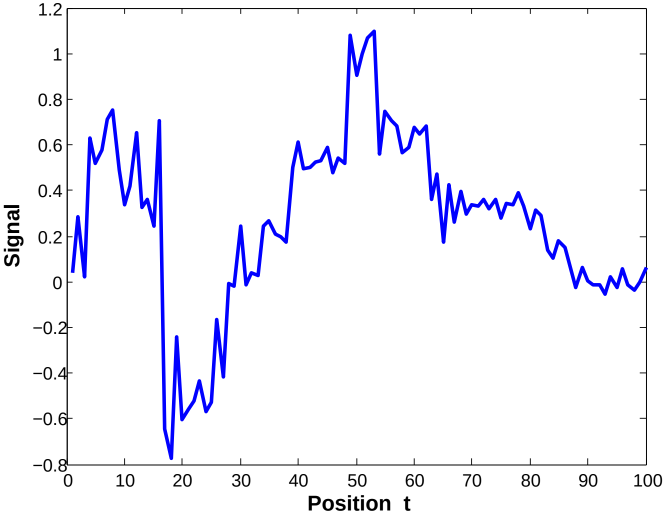
[2]Université Lille 1

[3]Max Planck Institute for Intelligent Systems, Tübingen

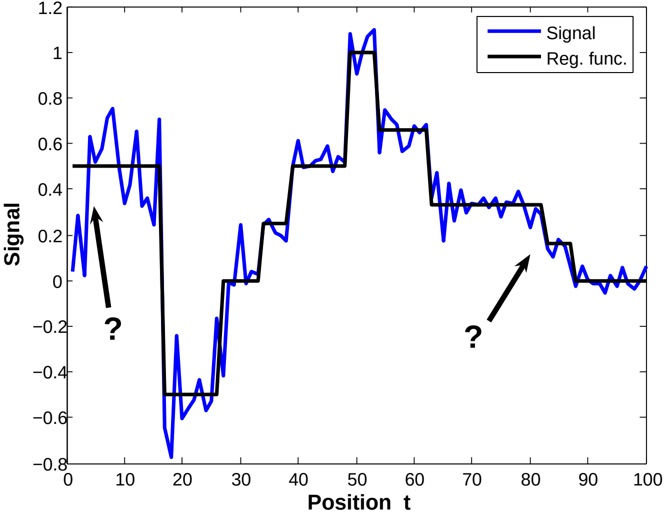
[4]University of Washington

ECARES Seminar, Université libre de Bruxelles, October 4, 2018

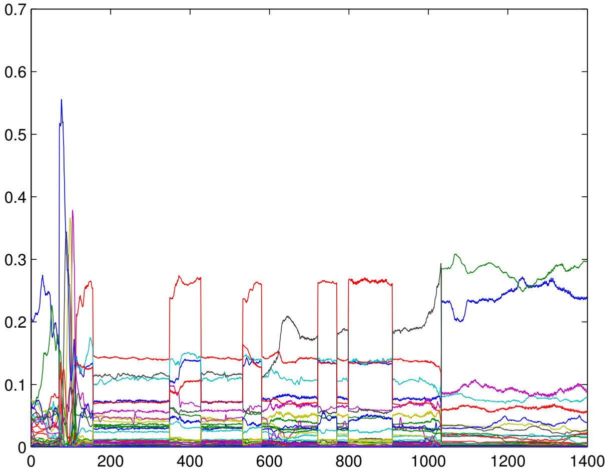
Example 1: 1-D signal



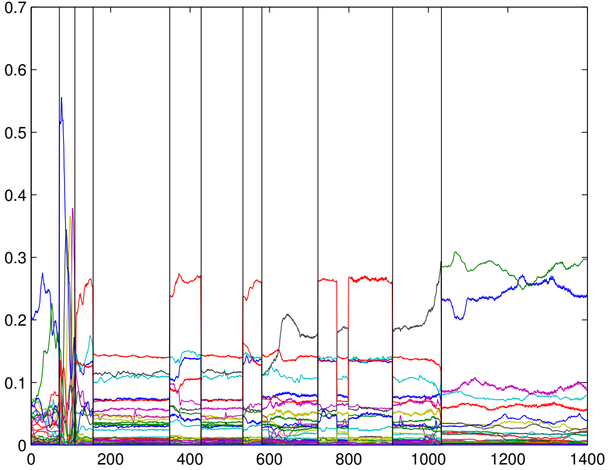
Example 1: 1-D signal: Find abrupt changes in the mean



Example 2: shot detection in a movie



Example 2: shot detection in a movie



The change-point problem

- Observation: $X_1, \ldots, X_n \in \mathcal{X}$ independent random variables ($\mathcal{X}$: arbitrary measurable set).
- P_{X_i}: distribution of X_i.

$\Rightarrow$ find where are the abrupt changes in the sequence $P_{X_1}, \ldots, P_{X_n}$?

Notation:

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

segmentation (of $\{1, \ldots, n\}$) into $D_\tau = D \in \{1, \ldots, n\}$ segments.

Challenges for (multiple) change-point detection

1. Detect changes in the whole distribution (not only the mean)
   - Mean:
     - homoscedastic: Birgé & Massart (2001), Comte & Rozenholc (2002, 2004), Baraud, Giraud & Huet (2010)...
     - heteroscedastic: A. & Celisse (2011)
   - Mean and variance: Picard et al. (2007), Fryzlewicz and Subba Rao (2014)
   - Full distribution: Zou et al. (2014) in $\mathbb{R}$, Matteson & James (2014) in $\mathbb{R}^d$

Challenges for (multiple) change-point detection

1. Detect changes in the whole distribution (not only the mean)
   - Mean:
     - homoscedastic: Birgé & Massart (2001), Comte & Rozenholc (2002, 2004), Baraud, Giraud & Huet (2010)...
     - heteroscedastic: A. & Celisse (2011)
   - Mean and variance: Picard et al. (2007), Fryzlewicz and Subba Rao (2014)
   - Full distribution: Zou et al. (2014) in $\mathbb{R}$, Matteson & James (2014) in $\mathbb{R}^d$
2. High-dimensional data of different nature:
   - Vectorial: measures in $\mathbb{R}^d$, curves (sound recordings,...)
   - Non vectorial: phenotypic data, graphs, DNA sequence,...
   - Both vectorial and non vectorial data.

Challenges for (multiple) change-point detection

1. Detect changes in the whole distribution (not only the mean)
   - Mean:
     - homoscedastic: Birgé & Massart (2001), Comte & Rozenholc (2002, 2004), Baraud, Giraud & Huet (2010)...
     - heteroscedastic: A. & Celisse (2011)
   - Mean and variance: Picard et al. (2007), Fryzlewicz and Subba Rao (2014)
   - Full distribution: Zou et al. (2014) in $\mathbb{R}$, Matteson & James (2014) in $\mathbb{R}^d$
2. High-dimensional data of different nature:
   - Vectorial: measures in $\mathbb{R}^d$, curves (sound recordings,...)
   - Non vectorial: phenotypic data, graphs, DNA sequence,...
   - Both vectorial and non vectorial data.
3. Efficient algorithm allowing to deal with large data sets

Kernels: a quick reminder

- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ measurable is a positive semidefinite kernel if $\forall x_1, \ldots, x_m \in \mathcal{X}$, the matrix $(k(x_i, x_j))_{1 \leqslant i,j \leqslant m}$ is positive semidefinite.

- Examples:
  - linear kernel: $k(x,y) = \langle x, y \rangle$,
  - polynomial kernel: $k(x,y) = (1 + \langle x, y \rangle)^p$,
  - Gaussian kernel: $k(x,y) = \exp(- \lVert x - y \rVert^2 / (2h^2))$,
  - χ^2 kernel on Δ^d: $k(x,y) = \exp\left(- \frac{1}{h \cdot d} \sum_{i=1}^d \frac{(x_i - y_i)^2}{x_i + y_i} \right)$
  - ...

The kernel least-squares criterion

- Least-squares criterion (when $\mathcal{X} = \mathbb{R}$): $\forall \tau \in \mathcal{T}_n := \bigcup_{D \geqslant 1} \mathcal{T}_n^D$,

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

- Kernel least-squares criterion:

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

- The two definitions coincide when $\mathcal{X} = \mathbb{R}$ and $k(x, y) = xy$.

Kernel change-point detection (KCP)

(A., Celisse & Harchaoui, 2012)

$$\widehat{\tau} \in \operatorname*{argmin}_{\tau \in \mathcal{T}_n} \left\{ \overbrace{\widehat{\mathcal{R}}_n(\tau)}^{\substack{\text{kernel}\\ \text{least-squares}\\ \text{criterion}}} + \underbrace{\mathrm{pen}(\tau)}_{\substack{\text{penalty}\\ \text{function}}} \right\}$$

where pen is a function increasing with D_τ, such as:

$$\mathrm{pen}(\tau) = \frac{1}{n} \left(c_1 \log \binom{n-1}{D_\tau - 1} + c_2 D_\tau \right)$$

$$\mathrm{pen}(\tau) = \frac{D_\tau}{n} \left(c_1 \log \left(\frac{n}{D_\tau} \right) + c_2 \right)$$

$$\mathrm{pen}(\tau) = \frac{c_1 D_\tau}{n} .$$

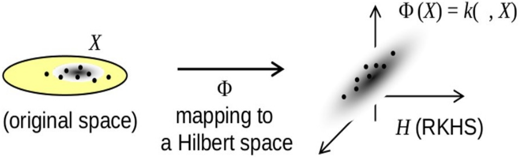
For $\mathcal{X} = \mathbb{R}$, linear kernel, Birgé & Massart (2001) and Lebarbier (2005) take $\mathrm{pen}(\tau) = \frac{\sigma^2 D_\tau}{n} \left[c_1 \log \left(\frac{n}{D_\tau} \right) + c_2 \right]$.

(Abstract) intuition on KCP

- KCP $\Leftrightarrow$ kernelized version of (penalized) least-squares change-point detection

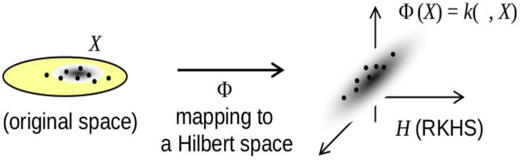
(Abstract) intuition on KCP

- KCP $\Leftrightarrow$ kernelized version of (penalized) least-squares change-point detection
- Canonical feature map $\Phi : x \in \mathcal{X} \mapsto k(x, \cdot) \in \mathcal{H}$ reproducing kernel Hilbert space (RKHS)
- $Y_i = \Phi(X_i) \in \mathcal{H}$ are independent $\mathcal{H}$-valued r.v.



(Abstract) intuition on KCP

- KCP $\Leftrightarrow$ kernelized version of (penalized) least-squares change-point detection
- Canonical feature map $\Phi : x \in \mathcal{X} \mapsto k(x, \cdot) \in \mathcal{H}$ reproducing kernel Hilbert space (RKHS)
- $Y_i = \Phi(X_i) \in \mathcal{H}$ are independent $\mathcal{H}$-valued r.v.

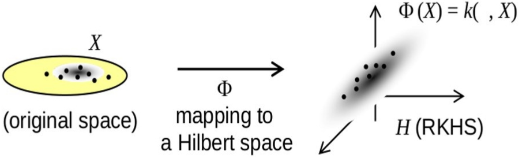


- $\mathbb{E}[\sqrt{k(X_i, X_i)}] < \infty \Rightarrow$ can define $\mu_i^\star \in \mathcal{H}$ the "mean" of Y_i

$\Rightarrow$ KCP detects jumps of the "mean" $\mu_i^\star$ of Y_i

(Abstract) intuition on KCP

- KCP $\Leftrightarrow$ kernelized version of (penalized) least-squares change-point detection
- Canonical feature map $\Phi : x \in \mathcal{X} \mapsto k(x, \cdot) \in \mathcal{H}$ reproducing kernel Hilbert space (RKHS)
- $Y_i = \Phi(X_i) \in \mathcal{H}$ are independent $\mathcal{H}$-valued r.v.



- $\mathbb{E}[\sqrt{k(X_i, X_i)}] < \infty \Rightarrow$ can define $\mu_i^\star \in \mathcal{H}$ the "mean" of Y_i

$\Rightarrow$ KCP detects jumps of the "mean" $\mu_i^\star$ of Y_i

- Remark: if k is characteristic (eg, Gaussian kernel), $\mu_i^\star$ characterizes P_{X_i}.

KCP for fixed D (Harchaoui & Cappé, 2007)

$$\widehat{\tau}(D) \in \operatorname*{argmin}_{\tau \in \mathcal{T}_n^D} \left\{ \widehat{\mathcal{R}}_n(\tau) \right\}$$

- Dynamic programming algorithm
- No computation in $\mathcal{H}$, only needs to compute the $k(X_i, X_j)$ (cost $\mathcal{C}_k$)

- Complexity of computing $\left(\widehat{\tau}(D)\right)_{1 \leqslant D \leqslant D_{\max}}$:

  time $\mathcal{O}\left((\mathcal{C}_k + D_{\max}) n^2 \right)$ and space $\mathcal{O}(D_{\max} n)$

  (Celisse et al., 2016).

Main assumptions

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

$$\exists M < +\infty, \forall i \in \{1, \ldots, n\}, \qquad k(X_i, X_i) \leqslant M^2 \text{ a.s.} \quad \textbf{(Db)}$$

$\Rightarrow$ always satisfied for Gaussian and χ^2 kernel.

$D = D_{\tau^\star}$ known: notations

- True segmentation $\tau^\star$:

$$\mu_1^\star = \cdots = \mu_{\tau_1^\star}^\star \neq \mu_{\tau_1^\star+1}^\star = \cdots = \mu_{\tau_2^\star}^\star \neq \quad \cdots \quad \neq \mu_{\tau_{D_{\tau^\star}-1}^\star+1}^\star = \cdots = \mu_n^\star \,.$$

- Smallest jump size: $\underline{\Delta} := \min_{i \,/\, \mu_i^\star \neq \mu_{i+1}^\star} \left\lVert \mu_i^\star - \mu_{i+1}^\star \right\rVert_{\mathcal{H}}$ (MMD, Gretton et al. 2006).
- Smallest segment length: $\underline{\Lambda}_\tau := \frac{1}{n} \min_{1 \leqslant \ell \leqslant D_\tau} \left| \tau_\ell - \tau_{\ell-1} \right|$.
- Loss between segmentations $\tau^1, \tau^2 \in \mathcal{T}_n$:

$$\mathrm{d}_{\infty,n}(\tau^1, \tau^2) := \frac{1}{n} \max_{1 \leqslant i \leqslant D_{\tau^1}-1} \left\{ \min_{1 \leqslant j \leqslant D_{\tau^2}-1} \left| \tau_i^1 - \tau_j^2 \right| \right\}$$
$$= \frac{1}{n} \max_{1 \leqslant i \leqslant D_{\tau^1}-1} \left| \tau_i^1 - \tau_i^2 \right| \qquad \text{if } D_{\tau^1} = D_{\tau^2} \text{ and } \tau^1, \tau^2 \text{ “close”}$$

$D = D_{\tau^\star}$ known: estimation of change-points locations

Theorem (A. & Garreau, 2016)

Assume: $\mathcal{H}$ separable, $(\mathbf{Db})$, $y > 0$ and

$$\underline{\Lambda}_{\tau^\star} > v_n(y) := \frac{148 D_{\tau^\star} M^2}{\underline{\Delta}^2} \cdot \frac{y + \log n + 1}{n} .$$

Then, with probability $1 - \mathrm{e}^{-y}$,

$$\forall \widehat{\tau}(D_{\tau^\star}) \in \operatorname*{argmin}_{\tau \in \mathcal{T}_n^{D_{\tau^\star}}} \left\{ \widehat{\mathcal{R}}_n(\tau) \right\}, \qquad \mathrm{d}_{\infty,n}\left(\tau^\star, \widehat{\tau}(D_{\tau^\star})\right) \leqslant v_n(y) .$$

$D = D_{\tau^\star}$ known: estimation of change-points locations (2)

Corollary (A. & Garreau, 2016, simplified result)

Assume: $\mathcal{H}$ *separable,* (**Db**) *and* $\frac{\underline{\Delta}^2}{M^2} \gtrsim \frac{D_{\tau^\star}}{\underline{\Lambda}_{\tau^\star}} \cdot \frac{\log n}{n}$.

Then, with probability $1 - n^{-2}$,

$$\forall \widehat{\tau}(D_{\tau^\star}) \in \operatorname*{argmin}_{\tau \in \mathcal{T}_n^{D_{\tau^\star}}} \{\widehat{\mathcal{R}}_n(\tau)\}, \qquad \mathrm{d}_{\infty,n}(\tau^\star, \widehat{\tau}(D_{\tau^\star})) \lesssim \frac{D_{\tau^\star} M^2}{\underline{\Delta}^2} \cdot \frac{\log n}{n} .$$

- $\frac{\underline{\Delta}^2}{M^2} \approx$ signal-to-noise ratio.
- Matches minimax lower bound $\log(n)/n$ (Brunel, 2014).
- Remark: $\log(n)$ factor not necessary in the standard "asymptotic" setting (Korostelev & Tsybakov, 2012).

KCP: data-driven D by model selection

- Notation: $Y = (Y_1, \ldots, Y_n) \in \mathcal{H}^n$, $\mu^\star = (\mu^\star_1, \ldots, \mu^\star_n) \in \mathcal{H}^n$
- For any $\tau \in \mathcal{T}_n$, $\Pi_\tau : \mathcal{H}^n \to \mathcal{H}^n$ orthogonal projection onto $F_\tau = \{(f_1, \ldots, f_n) \in \mathcal{H}^n \,/\, f_{\tau_{\ell-1}+1} = \cdots = f_{\tau_\ell} \; \forall \ell = 1, \ldots, D_\tau\}$

$\Rightarrow$ Least-squares estimator $\widehat{\mu}_\tau = \Pi_\tau Y$
and least-squares criterion:
$\widehat{\mathcal{R}}_n(\tau) = \frac{1}{n} \| Y - \widehat{\mu}_\tau \|^2 = \frac{1}{n} \sum_{i=1}^n \| Y_i - (\widehat{\mu}_\tau)_i \|_{\mathcal{H}}^2$

KCP: data-driven D by model selection

- Notation: $Y = (Y_1, \ldots, Y_n) \in \mathcal{H}^n$, $\mu^\star = (\mu^\star_1, \ldots, \mu^\star_n) \in \mathcal{H}^n$
- For any $\tau \in \mathcal{T}_n$, $\Pi_\tau : \mathcal{H}^n \to \mathcal{H}^n$ orthogonal projection onto $F_\tau = \{(f_1, \ldots, f_n) \in \mathcal{H}^n \,/\, f_{\tau_{\ell-1}+1} = \cdots = f_{\tau_\ell} \; \forall \ell = 1, \ldots, D_\tau\}$

$\Rightarrow$ Least-squares estimator $\widehat{\mu}_\tau = \Pi_\tau Y$
and least-squares criterion:
$\widehat{\mathcal{R}}_n(\tau) = \frac{1}{n} \| Y - \widehat{\mu}_\tau \|^2 = \frac{1}{n} \sum_{i=1}^n \| Y_i - (\widehat{\mu}_\tau)_i \|^2_{\mathcal{H}}$

- Quadratic risk of $\mu \in \mathcal{H}^n$:

$$\mathcal{R}(\mu) = \frac{1}{n} \| \mu - \mu^\star \|^2 = \frac{1}{n} \sum_{i=1}^n \| \mu_i - \mu^\star_i \|^2_{\mathcal{H}} \; .$$

- Usual approach for model selection: take a penalty such that

$$\forall \tau \in \mathcal{T}_n, \qquad \mathrm{pen}(\tau) \geqslant \mathrm{pen}_{\mathrm{id}}(\tau) := \mathcal{R}(\widehat{\mu}_\tau) - \widehat{\mathcal{R}}_n(\tau) + \mathrm{cst} \, .$$

Oracle inequality for KCP

Theorem (A., Celisse & Harchaoui, 2012–2016)

Assume: $\mathcal{H}$ separable, (**Db**)*,* $y > 0$*,* $C \geqslant 119$ *and*

$$\forall \tau \in \mathcal{T}_n, \qquad \mathrm{pen}(\tau) \geqslant \frac{CM^2}{n}\left[\log\binom{n-1}{D_\tau - 1} + D_\tau\right] .$$

Then, with probability $1 - \mathrm{e}^{-y}$*,*

$$\forall \widehat{\tau} \in \operatorname*{argmin}_{\tau \in \mathcal{T}_n} \left\{ \widehat{\mathcal{R}}_n(\tau) + \mathrm{pen}(\tau) \right\} ,$$

$$\mathcal{R}(\widehat{\mu}_{\widehat{\tau}}) \leqslant 2 \inf_{\tau \in \mathcal{T}_n} \left\{ \mathcal{R}(\widehat{\mu}_\tau) + \mathrm{pen}(\tau) \right\} + \frac{83yM^2}{n} .$$

- applies to $\mathrm{pen}(\tau) = \dfrac{CM^2 D_\tau}{n}$ if $C \geqslant 465 \log(n)$.
- $\mathcal{X} = \mathbb{R}$, linear kernel: Birgé & Massart (2001), Lebarbier (2005).

Change-point estimation performance of KCP

Theorem (A. & Garreau, 2016)

Assume: $\mathcal{H}$ *separable,* $(\mathbf{Db})$, $y > 0$ *and*

$$C_{\min} := \frac{74}{3}(D_{\tau^\star} + 1)(y + \log n + 1) \;<\; C \;<\; C_{\max} := \frac{\underline{\Delta}^2}{M^2} \frac{\underline{\Lambda}_{\tau^\star}}{6 D_{\tau^\star}} n \, .$$

Then, with probability $1 - \mathrm{e}^{-y}$*:*

$$\forall \widehat{\tau} \in \operatorname*{argmin}_{\tau \in \mathcal{T}_n} \left\{ \widehat{\mathcal{R}}_n(\tau) + \frac{C M^2 D_\tau}{n} \right\} , \qquad D_{\widehat{\tau}} = D_{\tau^\star}$$

and $$\mathrm{d}_{\infty,n}(\tau^\star, \widehat{\tau}) \leqslant v_n(y) := \frac{148 D_{\tau^\star} M^2}{\underline{\Delta}^2} \cdot \frac{y + \log n + 1}{n} \, .$$

Previous works (Lavielle & Moulines, 2000, among many others): real case ($\mathcal{H} = \mathbb{R}$) only (with dependent data).

Change-point estimation performance of KCP (2)

Corollary (A. & Garreau, 2016, simplified result)

*Assume: $\mathcal{H}$ separable, (**Db**) and*

$$D_{\tau^\star} \log n \lesssim C \lesssim \frac{\underline{\Delta}^2}{M^2} \frac{\underline{\Lambda}_{\tau^\star}}{D_{\tau^\star}} n \, .$$

Then, with probability $1 - n^{-2}$:

$$\forall \widehat{\tau} \in \operatorname*{argmin}_{\tau \in \mathcal{T}_n} \left\{ \widehat{\mathcal{R}}_n(\tau) + \frac{C M^2 D_\tau}{n} \right\}, \qquad D_{\widehat{\tau}} = D_{\tau^\star}$$

$$\text{and} \qquad \mathrm{d}_{\infty,n}(\tau^\star, \widehat{\tau}) \lesssim \frac{D_{\tau^\star} M^2}{\underline{\Delta}^2} \cdot \frac{\log n}{n} \, .$$

- $\frac{\underline{\Delta}^2}{M^2} \approx$ signal-to-noise ratio.
- Lower bound on C: $\log(n)$ necessary (Birgé & Massart, 2007).

Oracle inequality: proof ideas

- Notation: $\varepsilon = Y - \mu^\star \in \mathcal{H}^n$
- Ideal penalty:

$$\mathrm{pen}_{\mathrm{id}}(\tau) := \mathcal{R}(\widehat{\mu}_\tau) - \widehat{\mathcal{R}}_n(\tau) + \frac{1}{n} \|\varepsilon\|^2$$
$$= \frac{2}{n} \underbrace{\langle \Pi_\tau \mu^\star - \mu^\star , \varepsilon \rangle}_{= -L_\tau \text{(linear term)}} + \frac{2}{n} \underbrace{\|\Pi_\tau \varepsilon\|^2}_{= Q_\tau \text{ (quadratic term)}}$$

- Concentration for L_τ and Q_τ around their expectations

$\Rightarrow$ show that $\mathrm{pen}(\tau) \geqslant \mathrm{pen}_{\mathrm{id}}(\tau)$ simultaneously for all $\tau \in \mathcal{T}_n$, with probability $\geqslant 1 - \mathrm{e}^{-y}$.

- Previous work (Birgé & Massart, 2001): Gaussian assumption + real-valued functions $\Rightarrow$ does not apply to RKHS case.

Concentration of the quadratic term

Proposition (A., Celisse & Harchaoui, 2012–2016)

Assume: $\mathcal{H}$ separable and **(Db)**. *Then, for every $\tau \in \mathcal{T}_n$, $x > 0$:*

$$\left\lVert \Pi_\tau \varepsilon \right\rVert^2 - \mathbb{E}\left[\left\lVert \Pi_\tau \varepsilon \right\rVert^2 \right] \leqslant \frac{14M^2}{3}\left(x + 2\sqrt{2x D_\tau}\right) ,$$

with probability at least $1 - \mathrm{e}^{-x}$.

Proof ideas:

- Pinelis-Sakhanenko's inequality ($\lVert \sum_{i \in \lambda} \varepsilon_i \rVert_{\mathcal{H}}$).
- Bernstein's inequality (upper bounding moments).

Concentration of the linear term

Proposition

Assume: $\mathcal{H}$ separable and **(Db)**. *Then, for every $\tau \in \mathcal{T}_n$, $x > 0$, with probability at least $1 - 2\mathrm{e}^{-x}$:*

$$\left| \langle \Pi_\tau \mu^\star - \mu^\star , \, \varepsilon \rangle \right| \leqslant \theta \left\| \Pi_\tau \mu^\star - \mu^\star \right\|^2 + \left(\frac{1}{2\theta} + \frac{4}{3} \right) M^2 x \quad ,$$

for every $\theta > 0$.

Proof: Bernstein's inequality.

Identification of change-points: proof ideas

$$\widehat{\tau} \in \operatorname*{argmin}_{\tau \in \mathcal{T}_n} \left\{ \widehat{\mathcal{R}}_n(\tau) + \mathrm{pen}(\tau) \right\}$$

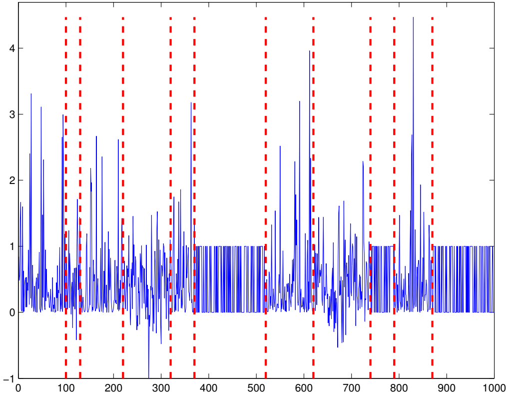
- Empirical risk:

$$\widehat{\mathcal{R}}_n(\tau) = \frac{1}{n} \underbrace{\left\| \mu^\star - \Pi_\tau \mu^\star \right\|^2}_{= A_\tau (\text{approximation})} + \frac{2}{n} \underbrace{\langle \mu^\star - \Pi_\tau \mu^\star , \, \varepsilon \rangle}_{= L_\tau (\text{linear term})} - \frac{1}{n} \underbrace{\left\| \Pi_\tau \varepsilon \right\|^2}_{= Q_\tau \ (\text{quadratic term})} + \frac{1}{n} \underbrace{\left\| \varepsilon \right\|^2}_{(\text{constant})}$$

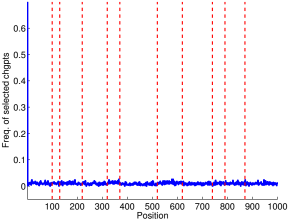
- Previous concentration inequalities for L_τ, Q_τ.
- Deterministic bounds on A_τ:
  $D_\tau < D_{\tau^\star} \ \Rightarrow \ \frac{1}{n} A_\tau \geqslant \frac{1}{2} \underline{\Lambda}_{\tau^\star} \underline{\Delta}^2$ (for showing $D_{\widehat{\tau}} \geqslant D_{\tau^\star}$)
  $\frac{1}{n} A_\tau \geqslant \frac{1}{2} \min \left\{ \underline{\Lambda}_{\tau^\star} , \, \mathrm{d}_{\infty,n}(\tau^\star, \tau) \right\} \underline{\Delta}^2$ (for $\widehat{\tau}(D_{\tau^\star})$)

Constant mean and variance: data

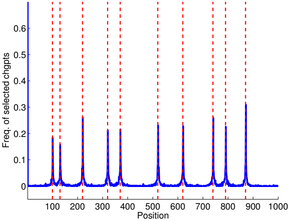
Constant mean and variance: the distribution of X_i is chosen among $\mathcal{B}(0.5)$, $\mathcal{N}(0.5, 0.25)$ and $\mathcal{E}(0.5)$.



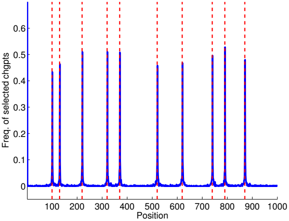
Constant mean and variance: results ($D_{\tau^\star}$)



Linear



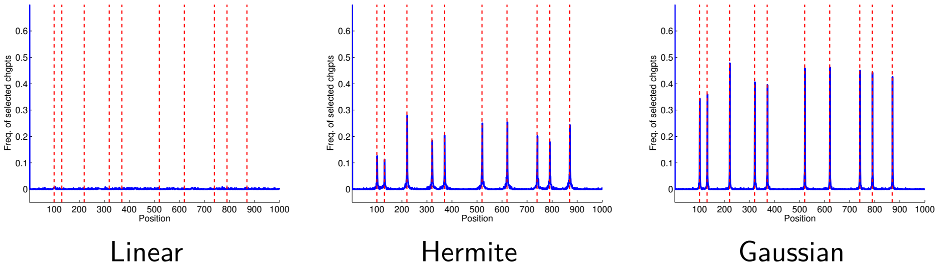
Hermite



Gaussian

KCP with $D_{\tau^\star}$ known.

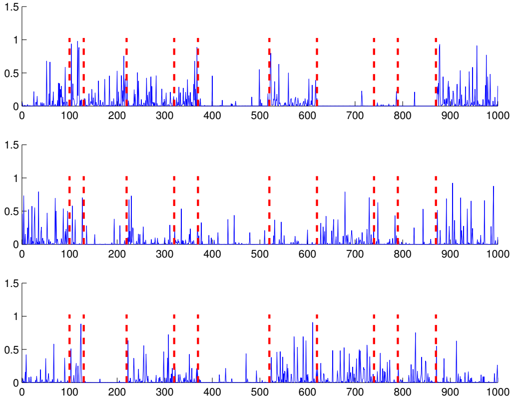
Constant mean and variance: results ($\widehat{D}$)



Gaussian kernel; KCP with $\widehat{D}$ data-driven.

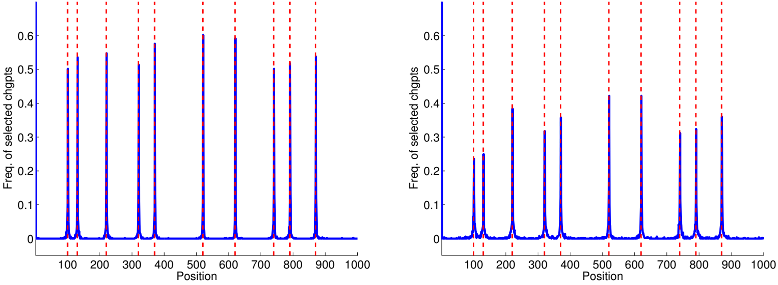
Histogram-valued data

$X_i \in d$-dimensional simplex, Dirichlet distribution $(p_1^\ell, \ldots, p_d^\ell)$ on the ℓ-th segment, with p_i^ℓ independent $\sim \mathcal{U}([0, 0.2])$.



(first three coordinates)

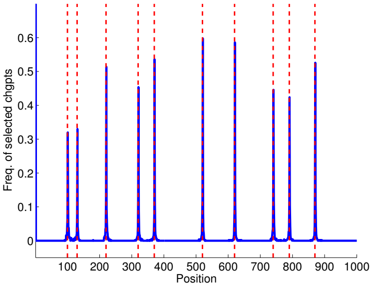
Histogram-valued data: results ($D_{\tau^\star}$)



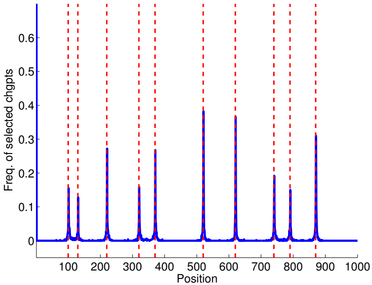
χ^2 kernel Gaussian kernel

KCP with $D_{\tau^\star}$ known.

Histogram-valued data: results ($\widehat{D}$)



χ^2 kernel



Gaussian kernel

KCP with $\widehat{D}$ data-driven.

Conclusion

Take-home message:

- Kernelized version of penalized least-squares change-point detection (eg, Lebarbier, 2005).
- Detection of changes in the distribution, not only the first moments.
- Can deal with structured data.
- Under reasonable assumptions and for a class of penalty functions:
  - oracle inequality;
  - identifies the correct number of change-points;
  - estimates at the correct rate the change-points locations.

Future work:

- Unbounded data/kernel.
- Dependent data?
- Learn how to choose the kernel.