Consistent change-point detection with kernels

Sylvain Arlot¹ (joint works with Alain Celisse², Damien Garreau³ & Zaïd Harchaoui⁴)

¹Université Paris-Sud
²Université Lille 1
³Max Planck Institute for Intelligent Systems, Tübingen
⁴University 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:**

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

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:
     - 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:
     - 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:
     - 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

- $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ measurable is a positive semidefinite kernel if \( \forall x_1, \ldots, x_m \in \mathcal{X} \), the matrix \((k(x_i, x_j))_{1 \leq i, j \leq m}\) 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: \( k(x, y) = \exp(-\|x - y\|^2/(2h^2)) \),
  - $\chi^2$ kernel on $\Delta^d$: \( k(x, y) = \exp \left( -\frac{1}{h \cdot d} \sum_{i=1}^{d} \frac{(x_i - y_i)^2}{x_i + y_i} \right) \)
  - ...
The kernel least-squares criterion

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

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

- **Kernel least-squares criterion**: 

$$\hat{R}_n(\tau) := \frac{1}{n} \sum_{i=1}^{n} k(X_i, X_i)$$

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

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

(A., Celisse & Harchaoui, 2012)

\[
\hat{\tau} \in \arg\min_{\tau \in T_n} \left\{ \hat{R}_n(\tau) + \text{pen}(\tau) \right\}
\]

where pen is a function increasing with \( D_\tau \), such as:

\[
\text{pen}(\tau) = \frac{1}{n} \left( c_1 \log \left( \frac{n - 1}{D_\tau - 1} \right) + c_2 D_\tau \right)
\]

\[
\text{pen}(\tau) = \frac{D_\tau}{n} \left( c_1 \log \left( \frac{n}{D_\tau} \right) + c_2 \right)
\]

\[
\text{pen}(\tau) = \frac{c_1 D_\tau}{n}.
\]

For \( \mathcal{X} = \mathbb{R} \), linear kernel, Birgé & Massart (2001) and Lebarbier (2005) take

\[
\text{pen}(\tau) = \frac{\sigma^2 D_\tau}{n} \left[ c_1 \log \left( \frac{n}{D_\tau} \right) + c_2 \right].
\]
(Abstract) intuition on KCP

- KCP $\iff$ kernelized version of (penalized) least-squares change-point detection
(Abstract) intuition on KCP

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

- Canonical feature map $\Phi : x \in X \mapsto k(x, \cdot) \in \mathcal{H}$ reproducing kernel Hilbert space (RKHS)

- $Y_i = \Phi(X_i) \in \mathcal{H}$ are independent $\mathcal{H}$-valued r.v.

Consistent change-point detection with kernels
Sylvain Arlot
(Abstract) intuition on KCP

- KCP $\iff$ kernelized version of (penalized) least-squares change-point detection
- Canonical feature map $\Phi : \mathcal{X} \mapsto k(x, \cdot) \in \mathcal{H}$ 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 $\mu_i^* \in \mathcal{H}$ the “mean” of $Y_i$ 
  $\Rightarrow$ KCP detects jumps of the “mean” $\mu_i^*$ of $Y_i$
(Abstract) intuition on KCP

- KCP ⇔ 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.

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

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

- Remark: if $k$ is characteristic (eg, Gaussian kernel), $\mu_i^*$ characterizes $P_{X_i}$.
KCP for fixed $D$ (Harchaoui & Cappé, 2007)

$$\hat{\tau}(D) \in \arg\min_{\tau \in T_n^D} \{ \hat{R}_n(\tau) \}$$

- Dynamic programming algorithm
- No computation in $H$, only needs to compute the $k(X_i, X_j)$ (cost $C_k$)

Complexity of computing $(\hat{\tau}(D))_{1 \leq D \leq D_{\text{max}}}$:

- Time $O((C_k + D_{\text{max}})n^2)$ and space $O(D_{\text{max}}n)$

(Celisse et al., 2016).
Main assumptions

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

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

$\Rightarrow$ always satisfied for Gaussian and $\chi^2$ kernel.
**True segmentation $\tau^*$:**

$$
\mu^*_1 = \cdots = \mu^*_1 \neq \mu^*_{1+1} = \cdots = \mu^*_2 \neq \cdots \neq \mu^*_{D_{\tau^*}+1} = \cdots = \mu^*_n.
$$

**Smallest jump size:** $\Delta := \min_i |\mu^*_i \neq \mu^*_{i+1}| \|\mu^*_i - \mu^*_{i+1}\|_H$

(MMD, Gretton et al. 2006).

**Smallest segment length:** $\Lambda_{\tau} := \frac{1}{n} \min_{1 \leq \ell \leq D_{\tau}} |\tau_{\ell} - \tau_{\ell-1}|$.

**Loss between segmentations $\tau^1, \tau^2 \in \mathcal{T}_n$:**

$$
d_{\infty,n}(\tau^1, \tau^2) := \frac{1}{n} \max_{1 \leq i \leq D_{\tau^1} - 1} \left\{ \min_{1 \leq j \leq D_{\tau^2} - 1} \left| \tau^1_i - \tau^2_j \right| \right\}
= \frac{1}{n} \max_{1 \leq i \leq D_{\tau^1} - 1} \left| \tau^1_i - \tau^2_i \right|
$$

if $D_{\tau^1} = D_{\tau^2}$ and $\tau^1, \tau^2$ “close”
Theorem (A. & Garreau, 2016)

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

$$
\Lambda_{\tau^*} > \nu_n(y) := \frac{148 D_{\tau^*} M^2}{\Delta^2} \cdot \frac{y + \log n + 1}{n}.
$$

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

$$
\forall \hat{\tau}(D_{\tau^*}) \in \arg\min_{\tau \in T_n D_{\tau^*}} \{ \hat{R}_n(\tau) \}, \quad d_{\infty,n}(\tau^*, \hat{\tau}(D_{\tau^*})) \leq \nu_n(y).
$$
Corollary (A. & Garreau, 2016, simplified result)

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

$$\frac{\Delta^2}{M^2} \gtrsim \frac{D_{\tau^*}}{\Lambda_{\tau^*}} \cdot \frac{\log n}{n}.$$

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

$$\forall \hat{\tau}(D_{\tau^*}) \in \text{argmin}\{\hat{R}_n(\tau)\}, \quad d_{\infty,n}(\tau^*, \hat{\tau}(D_{\tau^*})) \lesssim \frac{D_{\tau^*} M^2}{\Delta^2} \cdot \frac{\log n}{n}.$$

- $\frac{\Delta^2}{M^2} \approx$ 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: $Y = (Y_1, \ldots, Y_n) \in \mathcal{H}^n$, $\mu^* = (\mu_1^*, \ldots, \mu_n^*) \in \mathcal{H}^n$

- 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 $\hat{\mu}_\tau = \Pi_\tau Y$

and least-squares criterion:

$\hat{R}_n(\tau) = \frac{1}{n} \| Y - \hat{\mu}_\tau \|^2 = \frac{1}{n} \sum_{i=1}^n \| Y_i - (\hat{\mu}_\tau)_i \|_\mathcal{H}^2$
KCP: data-driven $D$ by model selection

- **Notation**: $Y = (Y_1, \ldots, Y_n) \in \mathcal{H}^n$, $\mu^* = (\mu_1^*, \ldots, \mu_n^*) \in \mathcal{H}^n$

- For any $\tau \in \mathcal{T}_n$, $\Pi_\tau : \mathcal{H}^n \rightarrow \mathcal{H}^n$ 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 $\hat{\mu}_\tau = \Pi_\tau Y$

and least-squares criterion:

$\hat{R}_n(\tau) = \frac{1}{n} \| Y - \hat{\mu}_\tau \|^2 = \frac{1}{n} \sum_{i=1}^{n} \| Y_i - (\hat{\mu}_\tau)_i \|^2_{\mathcal{H}}$

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

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

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

$$\forall \tau \in \mathcal{T}_n, \quad \text{pen}(\tau) \geq \text{pen}_{id}(\tau) := R(\hat{\mu}_\tau) - \hat{R}_n(\tau) + \text{cst} .$$
Oracle inequality for KCP

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

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

$$\forall \tau \in T_n, \quad \text{pen}(\tau) \geq \frac{CM^2}{n} \left[ \log \left( \frac{n - 1}{D_{\tau} - 1} \right) + D_{\tau} \right].$$

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

$$\forall \hat{\tau} \in \arg\min_{\tau \in T_n} \left\{ \hat{\mathcal{R}}_n(\tau) + \text{pen}(\tau) \right\},$$

$$\mathcal{R}(\hat{\mu}_{\hat{\tau}}) \leq 2 \inf_{\tau \in T_n} \left\{ \mathcal{R}(\hat{\mu}_{\tau}) + \text{pen}(\tau) \right\} + \frac{83yM^2}{n}.$$

- applies to $\text{pen}(\tau) = \frac{CM^2D_{\tau}}{n}$ if $C \geq 465 \log(n)$.
- $\mathcal{X} = \mathbb{R}$, linear kernel: Birgé & Massart (2001), Lebarbier (2005).
Change-point estimation performance of KCP

**Theorem (A. & Garreau, 2016)**

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

$$C_{\text{min}} := \frac{74}{3} (D_{\tau^*} + 1)(y + \log n + 1) < C < C_{\text{max}} := \frac{\Delta^2}{M^2} \frac{\Lambda_{\tau^*}}{6D_{\tau^*}} n.$$  

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

$$\forall \hat{\tau} \in \arg \min_{\tau \in \mathcal{T}_n} \left\{ \hat{\mathcal{R}}_n(\tau) + \frac{CM^2D_{\tau}}{n} \right\}, \quad D_{\hat{\tau}} = D_{\tau^*}$$

and

$$d_{\infty,n}(\tau^*, \hat{\tau}) \leq v_n(y) := \frac{148D_{\tau^*}M^2}{\Delta^2} \cdot \frac{y + \log n + 1}{n}.$$  

Previous works (Lavielle & Moulines, 2000, among many others): real case ($\mathcal{H} = \mathbb{R}$) only (with dependent data).
**Corollary (A. & Garreau, 2016, simplified result)**

**Assume: $\mathcal{H}$ separable,** $\{D_b\}$ and

$$D_{\tau^*} \log n \lesssim C \lesssim \frac{\Delta^2}{M^2} \frac{\Lambda_{\tau^*}}{D_{\tau^*}} n.$$  

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

$$\forall \hat{\tau} \in \arg\min_{\tau \in \mathcal{T}_n} \left\{ \hat{R}_n(\tau) + \frac{CM^2 D_{\tau}}{n} \right\}, \quad D_{\hat{\tau}} = D_{\tau^*}$$

and

$$d_{\infty,n}(\tau^*, \hat{\tau}) \lesssim \frac{D_{\tau^*} M^2}{\Delta^2} \cdot \frac{\log n}{n}.$$  

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

- **Notation:** \( \varepsilon = Y - \mu^* \in \mathcal{H}^n \)
- **Ideal penalty:**

\[
\text{pen}_{id}(\tau) := R(\hat{\mu}_\tau) - \hat{R}_n(\tau) + \frac{1}{n} \| \varepsilon \|^2 \\
= \frac{2}{n} \langle \Pi_\tau \mu^* - \mu^*, \varepsilon \rangle + \frac{2}{n} \| \Pi_\tau \varepsilon \|^2 \\
= -L_\tau \text{(linear term)} \quad \quad \quad = Q_\tau \text{(quadratic term)}
\]

- **Concentration** for \( L_\tau \) and \( Q_\tau \) around their expectations

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

- **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 $(D\mathbf{b})$. Then, for every $\tau \in T_n$, $x > 0$:

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

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

Proof ideas:

- Pinelis-Sakhanenko’s inequality ($\|\sum_{i \in \lambda} \varepsilon_i\|_{\mathcal{H}}$).
- Bernstein’s inequality (upper bounding moments).
Introduction

Concentration of the linear term

Proposition

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

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

for every $\theta > 0$.

Proof: Bernstein’s inequality.
Identification of change-points: proof ideas

\[ \hat{\tau} \in \arg\min_{\tau \in \mathcal{T}_n} \{ \hat{\mathcal{R}}_n(\tau) + \text{pen}(\tau) \} \]

- **Empirical risk:**

\[ \hat{\mathcal{R}}_n(\tau) = \frac{1}{n} \left\| \mu^* - \Pi_\tau \mu^* \right\|^2 + \frac{2}{n} \langle \mu^* - \Pi_\tau \mu^*, \varepsilon \rangle - \frac{1}{n} \left\| \Pi_\tau \varepsilon \right\|^2 + \frac{1}{n} \left\| \varepsilon \right\|^2 \]

\[ = A_\tau \text{ (approximation)} \quad = L_\tau \text{ (linear term)} \quad = Q_\tau \text{ (quadratic term)} \quad \text{(constant)} \]

- **Previous concentration inequalities for** \( L_\tau, Q_\tau \).

- **Deterministic bounds on** \( A_\tau \):

\[ D_\tau < D_{\tau^*} \implies \frac{1}{n} A_\tau \geq \frac{1}{2} \Lambda_{\tau^*} \Delta^2 \quad \text{(for showing } D_\hat{\tau} \geq D_{\tau^*}) \]

\[ \frac{1}{n} A_\tau \geq \frac{1}{2} \min\left\{ \Lambda_{\tau^*}, d_\infty, n(\tau^*, \tau) \right\} \Delta^2 \quad \text{(for } \hat{\tau}(D_{\tau^*})) \]
Constant mean and variance: the distribution of $X_i$ is chosen among $\mathcal{B}(0.5)$, $\mathcal{N}(0.5, 0.25)$ and $\mathcal{E}(0.5)$. 

![Graph showing data distribution and change points](image)
Constant mean and variance: results \((D_{\tau^*})\)

KCP with \(D_{\tau^*}\) known.
Constant mean and variance: results ($\hat{D}$)

- Linear
- Hermite
- Gaussian

KCP with $\hat{D}$ data-driven.
Histogram-valued data

\(X_i \in d\)-dimensional simplex, Dirichlet distribution \((p^\ell_1, \ldots, p^\ell_d)\) on the \(\ell\)-th segment, with \(p^\ell_i\) independent \(\sim U([0, 0.2])\).
Histogram-valued data: results ($D_{\tau^*}$)

\begin{align*}
\chi^2 \text{ kernel} & \quad \text{Gaussian kernel} \\
\text{KCP with } D_{\tau^*} \text{ known.}
\end{align*}
Histogram-valued data: results ($\hat{D}$)

\[ \chi^2 \text{ kernel} \quad \text{Gaussian kernel} \]

KCP with $\hat{D}$ 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.