Kernel change-point detection

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

[1]CNRS

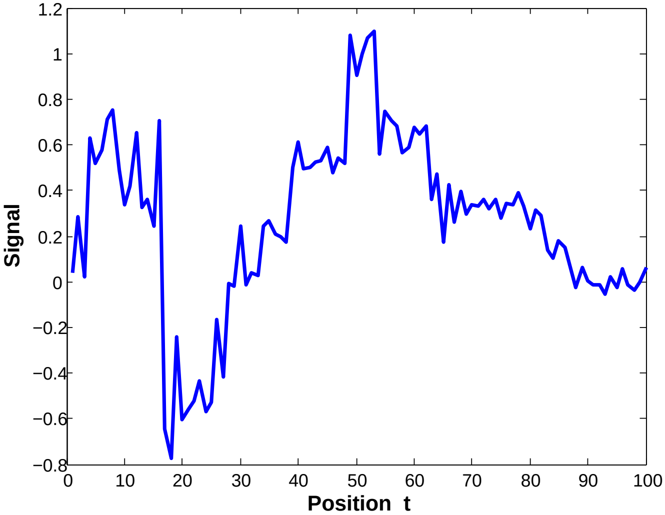
[2]École Normale Supérieure (Paris), DIENS, Équipe SIERRA

[3]Université Lille 1

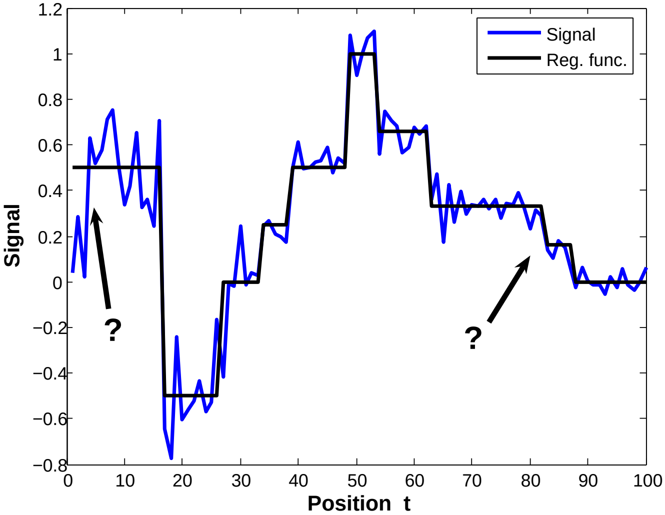
[4]INRIA Grenoble

Workshop Kernel methods for big data, Lille, 1st April 2014.

1-D signal (example)

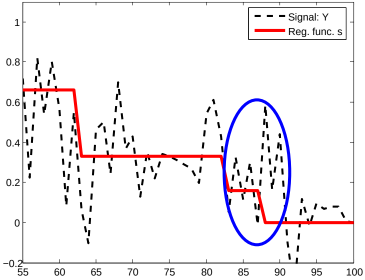


1-D signal (example): Find abrupt changes in the mean



Estimation rather than identification

With a finite sample, it is impossible to recover some change-points in noisy regions.



Purpose:

1. Estimate the regression function.
2. Use the quadratic loss $\ell(u, v) = \|u - v\|^2$.

Rk: Without too strong noise, recover all change-points.

Detect abrupt changes...

General purposes:

1. Detect changes in the whole distribution (not only in 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)

Detect abrupt changes...

General purposes:

1. Detect changes in the whole distribution (not only in 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)
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.

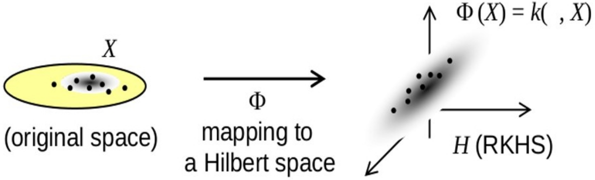
Detect abrupt changes...

General purposes:

1. Detect changes in the whole distribution (not only in 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)
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

Kernel and Reproducing Kernel Hilbert Space (RKHS)

- $\mathcal{X}$: initial input space.
- $X_1, \ldots, X_n$: initial observations.
- $k(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$: reproducing kernel ($\mathcal{H}$: RKHS).
- $\phi(\cdot) : \mathcal{X} \to \mathcal{H}$ s.t. $\phi(x) = k(x, \cdot)$: canonical feature map.



Asset:

Enables to work with high-dimensional heterogeneous data.

Rk:
Estimators depend on the Gram matrix $K := \{ k(X_i, X_j) \}_{1 \leq i,j \leq n}$.

Model

Mapping of the initial data

$$\forall 1 \leq i \leq n, \quad Y_i = \phi(X_i) \in \mathcal{H} \ .$$

$$\longrightarrow (t_1, Y_1), \ldots, (t_n, Y_n) \in [0,1] \times \mathcal{H} : \quad \text{independent} \ .$$

Model

$$\forall 1 \leq i \leq n, \qquad Y_i = s_i^\star + \varepsilon_i \quad \in \mathcal{H} \ ,$$

where

- $s_i^\star \in \mathcal{H}$: mean element of P_{X_i} (distribution of X_i)

$$\langle s_i^\star, f \rangle_{\mathcal{H}} = \mathbb{E}_{X_i} \left[\langle \phi(X_i), f \rangle_{\mathcal{H}} \right], \quad \forall f \in \mathcal{H}.$$

- $\forall i$, $\varepsilon_i := Y_i - s_i^\star$ with $\mathbb{E}[\varepsilon_i] = 0$ and $v_i := \mathbb{E}\left[\lVert \varepsilon_i \rVert_{\mathcal{H}}^2 \right]$.

Model

$$\forall 1 \le i \le n, \qquad Y_i = s_i^\star + \varepsilon_i \quad \in \mathcal{H} \ ,$$

where

- $s_i^\star \in \mathcal{H}$: mean element of P_{X_i} (distribution of X_i)

$$\langle s_i^\star , f \rangle_{\mathcal{H}} = \mathbb{E}_{X_i} \left[\langle \phi(X_i), f \rangle_{\mathcal{H}} \right], \quad \forall f \in \mathcal{H}.$$

- $\forall i$, $\varepsilon_i := Y_i - s_i^\star$ with $\mathbb{E}\left[\varepsilon_i\right] = 0$ and $v_i := \mathbb{E}\left[\left\| \varepsilon_i \right\|_{\mathcal{H}}^2 \right]$.

Assumptions

1. $\max_i \left\| Y_i \right\|_{\mathcal{H}} \le M \quad a.s. \qquad$ **(Db)** .
2. $\max_i v_i \le v_{\max} \qquad$ **(Vmax)** .
3. $s^\star = \left(s_1^\star , \ldots , s_n^\star \right) \in \mathcal{H}^n$: piecewise constant.

$$\left\| s^\star - \mu \right\|^2 := \sum_{i=1}^n \left\| s_i^\star - \mu_i \right\|_{\mathcal{H}}^2 .$$

Goal: $\longrightarrow$ Estimate $s^\star$ to recover change-points.

Least-squares estimator

- Empirical risk minimizer over S_m (= model):

$$\widehat{s}_m \in \arg\min_{u \in S_m} \widehat{\mathcal{R}}_n(u) \quad \text{where} \quad \widehat{\mathcal{R}}_n(u) = \frac{1}{n} \left\| u - Y \right\|^2 = \frac{1}{n} \sum_{i=1}^{n} \left\| u_i - Y_i \right\|_{\mathcal{H}}^2 .$$

- Regressogram:

$$\widehat{s}_m = \sum_{\lambda \in m} \widehat{\beta}_\lambda \mathbb{1}_\lambda \qquad \widehat{\beta}_\lambda = \frac{1}{\operatorname{Card}\left\{ t_i \in \lambda \right\}} \sum_{t_i \in \lambda} Y_i .$$

Model selection

Models:

- $\mathcal{M}_n = \{ m, \text{ segmentation of } \{1, \dots, n\} \}, \quad D_m = \mathrm{Card}(m)$.
- $m \Leftrightarrow \{ I_1 = [0, t_{m_1}], I_2 = (t_{m_1}, t_{m_2}], \dots, I_{D_m} = (t_{m_{D_m - 1}}, 1] \}$.
- $S_m = \{ \mu : (t_1, \dots, t_n) \to \mathcal{H},$ piecewise const. on all $\lambda \in m \}$ $\Leftrightarrow$ subspace of $\mathcal{H}^n$.

Strategy:

$$(S_m)_{m \in \mathcal{M}_n} \longrightarrow (\widehat{s}_m)_{m \in \mathcal{M}_n} \longrightarrow \widehat{s}_{\widehat{m}} \quad ???$$

Oracle model: $m^\star \in \mathrm{argmin}_{m \in \mathcal{M}_n} \|s^\star - \widehat{s}_m\|^2$.

Model selection

Models:

- $\mathcal{M}_n = \{\, m, \text{ segmentation of } \{1, \ldots, n\} \,\}, \quad D_m = \mathrm{Card}(m)$.
- $m \Leftrightarrow \{\, I_1 = [0, t_{m_1}],\ I_2 = (t_{m_1}, t_{m_2}], \ldots,\ I_{D_m} = (t_{m_{D_m - 1}}, 1] \,\}$.
- $S_m = \{\, \mu : (t_1, \ldots, t_n) \to \mathcal{H},$ piecewise const. on all $\lambda \in m \,\}$
  $\Leftrightarrow$ subspace of $\mathcal{H}^n$.

Strategy:

$$(S_m)_{m \in \mathcal{M}_n} \quad \longrightarrow \quad (\widehat{s}_m)_{m \in \mathcal{M}_n} \quad \longrightarrow \quad \widehat{s}_{\widehat{m}} \quad ???$$

Oracle model: $m^\star \in \mathrm{argmin}_{m \in \mathcal{M}_n} \| s^\star - \widehat{s}_m \|^2$.
Goal: Oracle inequality (in expectation, or with large probability):

$$\| s^\star - \widehat{s}_{\widehat{m}} \|^2 \leq C \inf_{m \in \mathcal{M}_n} \left\{ \| s^\star - \widehat{s}_m \|^2 + R(m, n) \right\}$$

Choose $(D-1)$ change-points...

Assumption: (Harchaoui & Cappé (2007))

The number $(D-1)$ of change-points is known.

Question:

Find the locations of the $(D-1)$ change-points? (D is given).

Strategy:

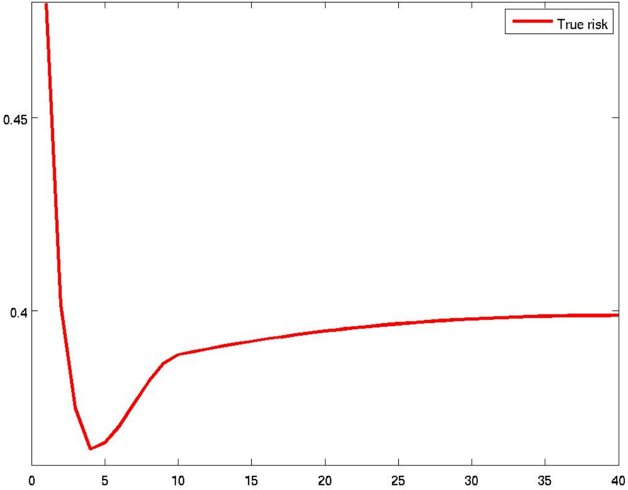
The "best" segmentation in D pieces is obtained by applying the ERM algorithm over $\bigcup_{D_m=D} S_m$:

ERM algorithm:

$$\widehat{m}_{\mathrm{ERM}}(D) = \operatorname*{argmin}_{m | D_m = D} \widehat{\mathcal{R}}_n\left(\widehat{s}_m\right) .$$

Rk: Based on dynamic programming.

Quality of the segmentations



Elementary calculations

Ideal criterion: (Π_m: orthog. proj. operator onto S_m)

$$\left\| s^\star - \widehat{s}_m \right\|^2 = \left\| s^\star - \Pi_m s^\star \right\|^2 + \left\| \Pi_m \varepsilon \right\|^2 \quad .$$

Empirical risk:

$$\left\| Y - \widehat{s}_m \right\|^2 = \left\| s^\star - \Pi_m s^\star \right\|^2 - \left\| \Pi_m \varepsilon \right\|^2 + 2 \left\langle (I - \Pi_m) s^\star , \, \varepsilon \right\rangle + \left\| \varepsilon \right\|^2 .$$

Elementary calculations

Ideal criterion: (Π_m: orthog. proj. operator onto S_m)

$$\|s^\star - \widehat{s}_m\|^2 = \|s^\star - \Pi_m s^\star\|^2 + \|\Pi_m \varepsilon\|^2 \quad .$$

Empirical risk:

$$\|Y - \widehat{s}_m\|^2 = \|s^\star - \Pi_m s^\star\|^2 - \|\Pi_m \varepsilon\|^2 + 2\left\langle (I - \Pi_m) s^\star , \, \varepsilon \right\rangle + \|\varepsilon\|^2 \, .$$

Expectations ($v_\lambda = \frac{1}{\mathrm{Card}(\lambda)} \sum_{i \in \lambda} v_i$)

$$\mathbb{E}\left[\|s^\star - \widehat{s}_m\|^2\right] = \|s^\star - \Pi_m s^\star\|^2 + \sum_{\lambda \in m} v_\lambda \quad ,$$

$$\mathbb{E}\left[\|Y - \widehat{s}_m\|^2\right] = \|s^\star - \Pi_m s^\star\|^2 - \sum_{\lambda \in m} v_\lambda + Cst \quad ,$$

Conclusion:

$\longrightarrow$ ERM prefers models with large $\sum_{\lambda \in m} v_\lambda$ (overfitting).

Choose the number of change-points

From $\left\{ \widehat{s}_{\widehat{m}_D} \right\}_D$, choose D amounts to choose the "best model".

Ideal penalty:

$$m^\star \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\| s^\star - \widehat{s}_m \right\|^2 = \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \left\| Y - \widehat{s}_m \right\|^2 + \mathrm{pen}_{\mathrm{id}}(m) \right\} ,$$

with $\mathrm{pen}_{\mathrm{id}}(m) =: 2 \left\| \Pi_m \varepsilon \right\|^2 - 2 \left\langle (I - \Pi_m) s^\star , \varepsilon \right\rangle$.

Strategy

1. Concentration inequalities for linear and quadratic terms.
2. Derive a tight upper bound $\mathrm{pen} \geq \mathrm{pen}_{\mathrm{id}}$ with high probability.

Previous work:

Birgé & Massart (2001): Gaussian assumption + real valued functions.

$\longrightarrow$ cannot be extended to Hilbert framework.

Concentration of the linear term

Theorem (Linear term)

Assume **(Db)**–**(Vmax)** *hold true.*
Then, for every segmentation $m \in \mathcal{M}_n$*, for every* $x > 0$ *with probability at least* $1 - 2e^{-x}$*,*

$$\left| \langle \Pi_m s^\star - s^\star, \varepsilon \rangle \right| \leq \theta \left\| \Pi_m s^\star - s^\star \right\|^2 + \left(\frac{v_{\max}}{\theta} + \frac{4M^2}{3} \right) x \enspace ,$$

for every $\theta > 0$*.*

Concentration of the quadratic term

Theorem (Quadratic term)

Assume (**Db**)–(**Vmax**), *and*

$$\exists \kappa \geq 1, \quad 0 < \frac{M^2}{\kappa} \leq \min_i v_i \qquad \textbf{(Vmin)} \quad .$$

Then, for every $m \in \mathcal{M}_n$, $x > 0$, *and* $\theta \in (0,1]$,

$$\left| \left\| \Pi_m \varepsilon \right\|^2 - \mathbb{E}\left[\left\| \Pi_m \varepsilon \right\|^2 \right] \right| \leq \theta \mathbb{E}\left[\left\| \Pi_m s^\star - \widehat{s}_m \right\|^2 \right] + \theta^{-1} L(\kappa) v_{\max} x \quad ,$$

with probability at least $1 - 2e^{-x}$, *where* $L(\kappa)$ *is a constant.*

Idea of the proof:

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

Oracle inequality

Theorem

Assume (**Db**)-(**Vmin**)-(**Vmax**) *and define*

$$\widehat{m} \in \operatorname*{argmin}_{m} \left\{ \frac{1}{n} \| Y - \widehat{s}_m \|^2 + \mathrm{pen}(m) \right\} \quad ,$$

where $\mathrm{pen}(m) = \frac{v_{\max} D_m}{n} \left[C_1 \ln\left(\frac{n}{D_m}\right) + C_2 \right]$ *for constants* $C_1, C_2 > 0$. *Then, for every* $x \geq 1$, *with probability at least* $1 - 2e^{-x}$,

$$\frac{1}{n} \| s^{\star} - \widehat{s}_{\widehat{m}} \|^2 \leq \Delta_1 \inf_{m} \left\{ \frac{1}{n} \| s^{\star} - \widehat{s}_m \|^2 + \mathrm{pen}(m) \right\} + \frac{\Delta_2 v_{\max} x}{n} \quad ,$$

where $\Delta_1 \geq 1$ *and* $\Delta_2 > 0$ *are absolute constants.*

In Birgé & Massart (2001), $\mathrm{pen}(m) = \frac{\sigma^2 D_m}{n} \left[c_1 \ln\left(\frac{n}{D_m}\right) + c_2 \right]$.

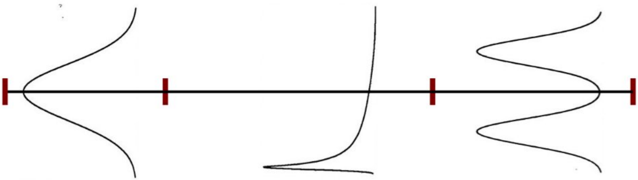
Model selection procedure

$$\mathrm{pen}(m) = \frac{v_{\max} D_m}{n} \left[C_1 \ln\left(\frac{n}{D_m} \right) + C_2 \right] = \mathrm{pen}(D_m) \quad .$$

Algorithm

1. For every $1 \leq D \leq D_{\max}$,
$$\widehat{m}_D \in \operatorname*{argmin}_{m,\ D_m = D} \left\{ \left\| Y - \widehat{s}_m \right\|^2 \right\} \quad ,$$
2. Define
$$\widehat{D} = \operatorname*{argmin}_{D} \left\{ \frac{1}{n} \left\| Y - \widehat{s}_{\widehat{m}_D} \right\|^2 + \frac{v_{\max} D}{n} \left[C_1 \ln\left(\frac{n}{D} \right) + C_2 \right] \right\} \quad .$$
where C_1, C_2: computed by simulation experiments.
3. Final estimator:
$$\widehat{s}_{\widehat{m}} =: \widehat{s}_{\widehat{m}_{\widehat{D}}} \, .$$

Changes in the distribution (synthetic data)

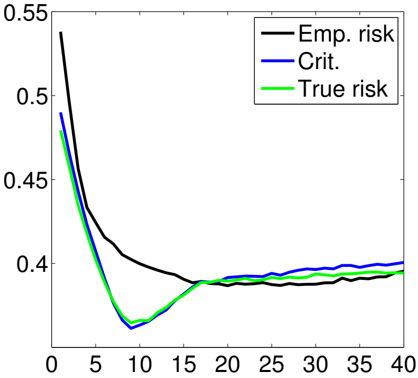


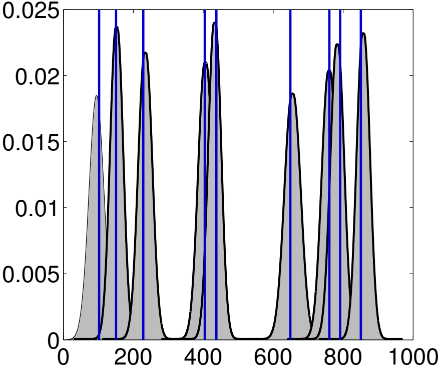
Description:

1. $n = 1\,000$, $D^* - 1 = 9$, $N_{rep} = 100$.
2. In each segment, observations generated according to one distribution within a pool of 10 distributions with same mean and variance.
3. Kernel-based approach enables to distinguish them (higher order moments)
4. Gaussian kernel: $k_h(x, y) = \exp\left[-\left\| x - y \right\|^2 / (2h^2) \right]$.

Changes in the distribution (synthetic data), cont.

Results

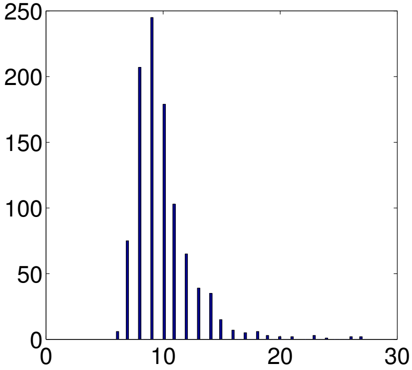




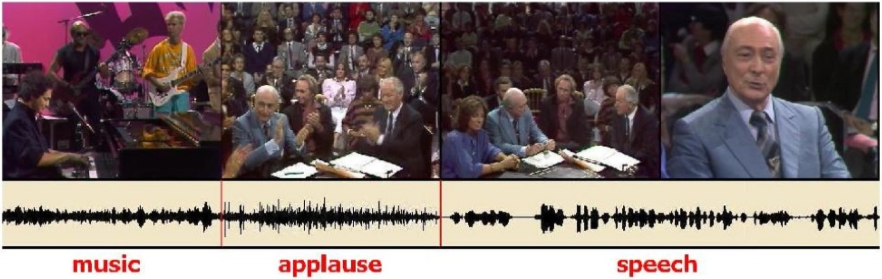
Hausdorff distance: 0.053 ± 0.006

Changes in the distribution (synthetic data), cont.

Results: estimated number of change-points



"Le grand échiquier", 70s-80s French talk show



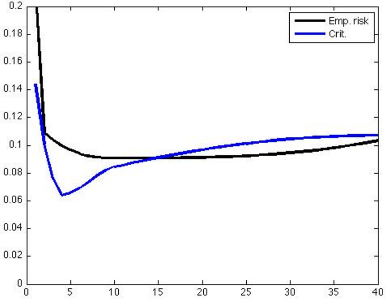
- Audio and video recordings.
- Audio: different situations can be distinguished from sound recordings (music, applause, speech,. . .).
- Video: different video scenes can be distinguished by their backgrounds or specific actions of people (clapping hands, discussing,. . .).

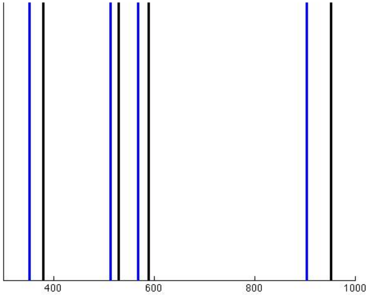
Audio signal

Description:

- $n = 500$, $D^* - 1 = 4$.
- At each t_i, one observes a multivariate vector of dimension 12.
- Gaussian kernel: $k_h(x,y) = \exp\left[- \|x-y\|^2 / (2h^2) \right]$.

Results: Hausdorff distance 0.079 ± 0.006



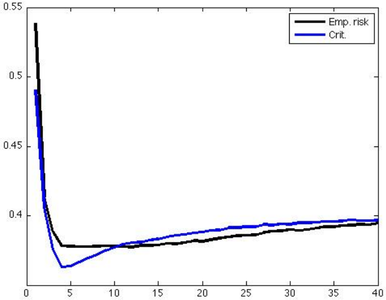


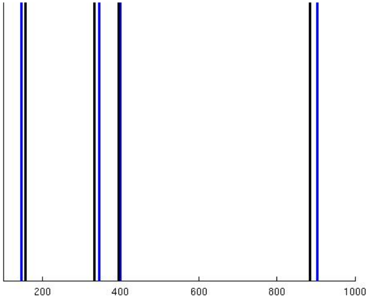
Video sequence

Description:

- $n = 10\,000$, $D^* - 1 = 4$.
- Each image summarized by a histogram with 1 024 bins.
- χ^2 kernel: $k_d(x,y) = \sum_{i=1}^{d} \frac{(x_i - y_i)^2}{x_i + y_i}$.

Results: Hausdorff distance 0.093 ± 0.007





Conclusion

Take-home message:

- Change-point detection algorithm for possibly high-dimensional or complex data
- Data-driven choice of the number of change-points
- Non-asymptotic oracle inequality (guarantee on the risk)
- Experiments: changes in less usual properties of the distribution, audio or video data

Conclusion

Take-home message:

- Change-point detection algorithm for possibly high-dimensional or complex data
- Data-driven choice of the number of change-points
- Non-asymptotic oracle inequality (guarantee on the risk)
- Experiments: changes in less usual properties of the distribution, audio or video data

Open questions:

1. Influence of the choice of kernel
2. Data-driven choice of the kernel
3. Relax the assumption on the variance
4. Extend our model selection theorem to other regression settings