Supplementary material for Segmentation of the mean of heteroscedastic data via cross-validation

Sylvain Arlot and Alain Celisse

September 20, 2010

1 Proofs of theoretical results

For completeness, we report here proofs (or sketch of proofs) of the theoretical results stated in the main paper, that are coming from previous works.

1.1 Proof of Lemma 1 of the main paper

Let us first state the following lemma.

Lemma 1.

$$\mathbb{E}\left[S_{\lambda,1}^2\right] = n_\lambda \left(\sigma_\lambda^r\right)^2 + n_\lambda^2 \beta_\lambda^2 \,, \quad \text{and} \quad \mathbb{E}\left[S_{\lambda,2}\right] = n_\lambda \left(\left(\sigma_\lambda^r\right)^2 + \left(\sigma_\lambda^b\right)^2 + \beta_\lambda^2 \right) \,,$$

with

$$\begin{aligned}
S_{\lambda,1} &= \frac{\sum_{i=1}^n Y_i \mathbb{1}_\lambda(t_i)}{n_\lambda} \,, \quad \text{and} \quad S_{\lambda,2} = \frac{\sum_{i=1}^n Y_i^2 \mathbb{1}_\lambda(t_i)}{n_\lambda} \,, \\
\left(\sigma_\lambda^r\right)^2 &= \frac{\sum_{i=1}^n \sigma(t_i)^2 \mathbb{1}_\lambda(t_i)}{n_\lambda} \,, \quad \text{and} \quad \left(\sigma_\lambda^b\right)^2 = \frac{\sum_{i=1}^n \left(s(t_i) - \beta_\lambda\right)^2 \mathbb{1}_\lambda(t_i)}{n_\lambda} \,, \\
\beta_\lambda &= \frac{\sum_{i=1}^n s(t_i) \mathbb{1}_\lambda(t_i)}{n_\lambda} \,, \quad \text{and} \quad \widehat{\beta}_\lambda = \frac{\sum_{i=1}^n Y_i \mathbb{1}_\lambda(t_i)}{n_\lambda} \,,
\end{aligned}$$

where $n_\lambda = \operatorname{Card}\left(\{j \mid t_j \in \lambda\}\right)$.

With the notation of Lemma 1, simple calculations lead to

$$\begin{aligned}
\mathbb{E}\left[P_n \gamma\left(\widehat{s}_m\right)\right] &= \sum_{\lambda \in \Lambda_m} \left\{ \frac{1}{n} \sum_{i=1}^n \left(\sigma(t_i)^2 + \mathbb{E}\left[s(t_i) - \widehat{\beta}_\lambda \right]^2 \right) \mathbb{1}_\lambda(t_i) + \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[2\varepsilon_i \left(s(t_i) - \widehat{\beta}_\lambda \right) \right] \mathbb{1}_\lambda(t_i) \right\} \\
&= \sum_{\lambda \in \Lambda_m} \left\{ \frac{1}{n} \left(n_\lambda \left(\sigma_\lambda^r\right)^2 + \sum_{i=1}^n \mathbb{E}\left[s(t_i) - \widehat{\beta}_\lambda \right]^2 \mathbb{1}_\lambda(t_i) \right) - \frac{2}{n} \left(\sigma_\lambda^r\right)^2 \right\} \,.
\end{aligned}$$

Moreover, by use of Lemma 1, it comes

$$\mathbb{E}\left[s(t_i) - \widehat{\beta}_\lambda \right]^2 = \left[s(t_i) - s_m(t_i) \right]^2 + \frac{1}{n_\lambda} \left(\sigma_\lambda^r\right)^2 \,,$$

which enables to conclude

$$\mathbb{E}\left[P_n\gamma\left(\widehat{s}_m\right)\right] = \|s - s_m\|_n^2 - \frac{1}{n}\sum_{\lambda\in\Lambda_m}\left(\sigma_\lambda^r\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\sigma(t_i)^2 \ .$$

The expectation of the loss of $\widehat{s}_m$ results from similar calculations.

1.2 Proof of Theorem 1 of the main paper

Since the proof of Theorem 1 of the main paper is quite technical, it is not reproduced here. For the detailed calculations, we refer interested readers to the Ph.D. thesis of the second author [6, Section 3.5.3], which is freely available online.

1.3 Proof of Proposition 1 of the main paper

On the basis of Theorem 1 of the main paper and assuming that for every $\lambda \in \Lambda_m$, $n_\lambda \geq 3$ and $p \leq n_\lambda + 1$, it comes that the expectation of the Lpo risk is

$$\begin{aligned}
\mathbb{E}\left[\widehat{R}_{\mathrm{Lpo}_p}\right] &= \sum_{\lambda\in\Lambda_m}\frac{1}{p}\left(\left(\sigma_\lambda^r\right)^2 + \frac{n_\lambda}{n_\lambda - 1}\left(\sigma_\lambda^b\right)^2\right)\left[n_\lambda - 1 - \frac{n_\lambda(n-p)}{n} + \frac{n}{n-p}V_\lambda(1)V_\lambda(-1)\right] \cdot \\
&= \sum_{\lambda\in\Lambda_m}\left(\sigma_\lambda^r\right)^2\left[\frac{n_\lambda}{n} + \frac{1}{p}\left(\frac{n}{n-p}V_\lambda(1)V_\lambda(-1) - 1\right)\right] \\
&\quad + \sum_{\lambda\in\Lambda_m} n_\lambda\left(\sigma_\lambda^b\right)^2\frac{1}{p}\left[1 - \frac{n-p}{n}\frac{n_\lambda}{(n_\lambda-1)} + \frac{n}{n-p}\frac{V_\lambda(1)V_\lambda(-1)}{n_\lambda - 1}\right] \cdot
\end{aligned}$$

Thanks to Lemma 2, it comes that

$$\frac{1}{n-p} \leq \frac{1}{p}\left(\frac{n}{n-p}V_\lambda(1)V_\lambda(-1) - 1\right) \leq \frac{p}{n-p} + \frac{n}{n-p}\frac{1}{n_\lambda} + o\left(\frac{n}{n-p}\frac{1}{n_\lambda}\right) + o\left(\frac{p}{n-p}\right) \cdot$$

With the assumptions of Lemma 2 and assuming $p \to +\infty$ as n tends to infinity, one gets

$$\frac{1}{n-p} \leq \frac{1}{p}\left(\frac{n}{n-p}V_\lambda(1)V_\lambda(-1) - 1\right) \leq \frac{p}{n-p} + o\left(\frac{p}{n-p}\right) \cdot$$

Then,

$$\begin{aligned}
\frac{\sum_{\lambda\in\Lambda_m}\left(\sigma_\lambda^r\right)^2}{n-p} &\leq \sum_{\lambda\in\Lambda_m}\left(\sigma_\lambda^r\right)^2\left[\frac{n_\lambda}{n} + \frac{1}{p}\left(\frac{n}{n-p}V_\lambda(1)V_\lambda(-1) - 1\right)\right] - \frac{1}{n}\sum_{i=1}^{n}\sigma(t_i)^2 \\
&\leq \frac{p\left(\sum_{\lambda\in\Lambda_m}\left(\sigma_\lambda^r\right)^2 + o(1)\right)}{n-p} \ ,
\end{aligned}$$

under the additional assumption (**BN**) that $\sigma(t_i)^2$ is uniformly bounded.

Similarly, another use of Lemma 2 implies

$$\begin{aligned}
\frac{n}{(n-p)n_\lambda} \leq \frac{n}{(n-p)n_\lambda}V_\lambda(1)V_\lambda(-1) &\leq \frac{n}{(n-p)n_\lambda} + \frac{np}{(n-p)n_\lambda^2} + o\left(\frac{np}{(n-p)n_\lambda^2}\right) \\
&\quad + \frac{p(p-1)}{(n-p)n_\lambda} + o\left(\frac{p(p-1)}{(n-p)n_\lambda}\right) \cdot
\end{aligned}$$

After some calculations, it yields

$$n_\lambda \left(\sigma_\lambda^b\right)^2 \frac{1}{p}\left[1 - \frac{n-p}{n}\frac{n_\lambda}{(n_\lambda - 1)} + \frac{n}{n-p}\frac{V_\lambda(1)V_\lambda(-1)}{n_\lambda - 1}\right] = n_\lambda \left(\sigma_\lambda^b\right)^2 \left[\frac{1}{n} + o\left(\frac{1}{n}\right)\right] \quad .$$

Moreover, if one assumes (**BV**) that there exists $M > 0$ such that for evey i, $(s(t_i) - s_m(t_i))^2 \leq M$, then

$$\sum_{\lambda \in \Lambda_m} n_\lambda \left(\sigma_\lambda^b\right)^2 \frac{1}{p}\left[1 - \frac{n-p}{n}\frac{n_\lambda}{(n_\lambda - 1)} + \frac{n}{n-p}\frac{V_\lambda(1)V_\lambda(-1)}{n_\lambda - 1}\right] = \|s - s_m\|_n^2 + o\left(\frac{1}{n}\right) \quad ,$$

which concludes the proof.

1.4 Magnitude of $V_\lambda(1)V_\lambda(-1)$

Lemma 2. *Let $m \in \mathcal{M}_n$. Assume that some $K \in (0,1)$ exists such that for every $\lambda \in \Lambda_m$, $n_\lambda \geq Kn$, and $n_\lambda > p$ with $p^2/n \to 0$ as n tends to infinity. Then,*

$$1 \leq V_\lambda(1)V_\lambda(-1) \leq 1 + \frac{p}{n_\lambda} + o\left(\frac{p}{n_\lambda}\right) + \frac{p(p-1)}{n} + o\left(\frac{p(p-1)}{n}\right) \quad .$$

Lemma 2. The lower bound comes from Jensen's inequality, while the upper bound results from

$$\begin{aligned} V_\lambda(1)V_\lambda(-1) &\leq \frac{n_\lambda}{n_\lambda - p}\frac{n^p}{n(n-1)\dot{(}n-(p-1))} \\ &\leq 1 + \left(\frac{p}{n_\lambda} + \frac{p(p-1)}{n} + o\left(\frac{p}{n_\lambda} + \frac{p(p-1)}{n}\right)\right) \quad . \end{aligned}$$

□

2 Random frameworks

In order to assess the generality of the results of Section 5.3 of the main paper, the procedures $[\![\mathrm{ERM}, \mathrm{VF}_5]\!]$, $[\![\mathrm{Loo}, \mathrm{VF}_5]\!]$, $[\![\mathrm{Lpo}_{20}, \mathrm{VF}_5]\!]$, $[\![\mathrm{ERM}, \mathrm{BM}]\!]$, BGH, ZS, and PML have been compared in three random settings. The following process has been repeated $N = 10\,000$ times. First, piecewise constant functions s and σ are randomly chosen (see Section 2.2 for details). Then, given s and σ , a data sample $(t_i, Y_i)_{1 \leq i \leq n}$ is generated as described in Section 3.3.1 of the main paper, and the same collection of models is used. Finally, each procedure $\mathfrak{P}$ is applied to the sample $(t_i, Y_i)_{1 \leq i \leq n}$, and its loss $\|s - \widetilde{s}_{\mathfrak{P}}(P_n)\|_n^2$ is measured, as well as the loss of the oracle $\inf_{m \in \mathcal{M}_n}\left\{\|s - \widehat{s}_m\|_n^2\right\}$.

To summarize the results, the quality of each procedure is measured by the ratio

$$C_{\mathrm{or}}^{(R)}(\mathfrak{P}) = \frac{\mathbb{E}_{s,\sigma,\varepsilon_1,\ldots,\varepsilon_n}\left[\|s - \widetilde{s}_{\mathfrak{P}}(P_n)\|_n^2\right]}{\mathbb{E}_{s,\sigma,\varepsilon_1,\ldots,\varepsilon_n}\left[\inf_{m \in \mathcal{M}_n}\left\{\|s - \widehat{s}_m\|_n^2\right\}\right]} \quad .$$

The notation $C_{\mathrm{or}}^{(R)}(\mathfrak{P})$ differs from $C_{\mathrm{or}}(\mathfrak{P})$ to emphasize that each expectation includes the randomness of s and σ , in addition to the one of $(\varepsilon_i)_{1 \leq i \leq n}$.

Framework	A	B	C
$[\![\mathrm{ERM}, \mathrm{BM}]\!]$	4.68 ± 0.03	4.63 ± 0.03	7.34 ± 0.05
$[\![\mathrm{ERM}, \mathrm{VF}_5]\!]$	$\mathbf{4.67} \pm 0.03$	4.61 ± 0.03	6.49 ± 0.04
$[\![\mathrm{Loo}, \mathrm{VF}_5]\!]$	$\mathbf{4.63} \pm 0.03$	$\mathbf{4.52} \pm 0.03$	6.04 ± 0.04
$[\![\mathrm{Lpo}_{20}, \mathrm{VF}_5]\!]$	5.40 ± 0.03	5.25 ± 0.03	6.57 ± 0.05
$[\![\mathrm{Lpo}_{50}, \mathrm{VF}_5]\!]$	6.23 ± 0.03	5.94 ± 0.04	7.56 ± 0.05
BGH	4.85 ± 0.03	4.78 ± 0.03	7.70 ± 0.05
ZS	5.34 ± 0.04	5.25 ± 0.04	7.28 ± 0.05
PML	5.03 ± 0.03	4.99 ± 0.03	$\mathbf{4.84} \pm 0.04$

Table 1: Performance $C_{\mathrm{or}}^{(R)}(\mathfrak{P})$ of several model selection procedures $\mathfrak{P}$ in frameworks A, B, C with sample size $n = 100$. In each framework, $N = 10\,000$ independent samples have been considered. Next to each value is indicated the corresponding empirical standard deviation divided by $\sqrt{N}$.

2.1 Results of the simulation experiments

The results of this experiment—which are reported in Tables 1—mostly confirm the results of Section 5.3 of the main paper (except that all the frameworks are heteroscedastic here), that is, whatever p, $[\![\mathrm{Lpo}_p, \mathrm{VF}_5]\!]$ outperforms $[\![\mathrm{ERM}, \mathrm{VF}_5]\!]$, which strongly outperforms $[\![\mathrm{ERM}, \mathrm{BM}]\!]$.

Moreover, the difference between the performances of $[\![\mathrm{Lpo}_p, \mathrm{VF}_5]\!]$ and $[\![\mathrm{ERM}, \mathrm{VF}_5]\!]$ is the largest in setting C and the smallest in setting A. This fact confirms the interpretation given in Section 3 of the main paper for the failure of ERM for localizing a given number of change-points. Indeed, the main differences between frameworks A, B and C—which are precisely defined in Section 2.2—can be sketched as follows:

A. the partitions on which s is built is often close to regular, and σ is chosen independently from s.

B. the partitions on which s is built are often irregular, and σ is chosen independently from s.

C. the partitions on which s is built are often irregular, and σ depends on s, so that the noise-level is smaller where s jumps more often.

In other words, frameworks A, B and C have been built so that for any $D \in \mathcal{D}_n$, the largest variations over $\mathcal{M}_n(D)$ of $V(m)$ (defined by Eq. (7) of the main paper) occur in framework C, and the smallest variations occur in framework A. As a consequence, variations of the performance of $[\![\mathrm{ERM}, \mathrm{VF}_5]\!]$ compared to $[\![\mathrm{Lpo}_p, \mathrm{VF}_5]\!]$ according to the framework certainly come from the local overfitting phenomenon presented in Section 3 of the main paper.

2.2 Detailed definition of the random frameworks

Let us now detail how piecewise constant functions s and σ have been generated in the frameworks A, B and C. In each framework, s and σ are of the form

$$s(x) = \sum_{j=0}^{K_s-1} \alpha_j \mathbb{1}_{[a_j;a_{j+1})} + \alpha_{K_s} \mathbb{1}_{[a_{K_s};a_{K_s+1}]} \qquad \text{with } a_0 = 0 < a_1 < \cdots < a_{K_s+1} = 1$$

$$\sigma(x) = \sum_{j=0}^{K_\sigma-1} \beta_j \mathbb{1}_{[b_j;b_{j+1})} + \beta_{K_\sigma} \mathbb{1}_{[b_{K_\sigma};b_{K_\sigma+1}]} \qquad \text{with } b_0 = 0 < b_1 < \cdots < b_{K_\sigma+1} = 1$$

for some positive integers K_s, K_σ and real numbers $\alpha_0, \ldots, \alpha_{K_s} \in \mathbb{R}$ and $\beta_0, \ldots, \beta_{K_\sigma} > 0$.

Remark 1. The frameworks A, B and C depend on the sample size n, through the distribution of K_s, K_σ, and of the size of the intervals $[a_j; a_{j+1})$ and $[b_j; b_{j+1})$. This ensures that the signal-to-noise ratio remains rather small, so that the quadratic risk remains an adequate performance measure for change-point detection.

When the signal-to-noise ratio is larger (that is, when all jumps of s are much larger than the noise-level, and the number of jumps of s is small compared to the sample size), the change-point detection problem is of different nature. In particular, the number of change-points would be better estimated with procedures targeting identification (such as BIC, or even larger penalties) than efficiency (such as VFCV).

2.2.1 Framework A

In framework A, s and σ are generated as follows:

- K_s, the number of jumps of s, has uniform distribution over $\{3, \ldots, \lfloor\sqrt{n}\rfloor\}$.
- For $0 \le j \le K_s$,
$$a_{j+1} - a_j = \Delta^s_{\min} + \frac{(1 - (K_s+1)\Delta^s_{\min})U_j}{\sum_{k=0}^{K_s} U_k}$$
with $\Delta^s_{\min} = \min\{5/n, 1/(K_s+1)\}$ and $U_0, \ldots, U_{K_s}$ are i.i.d. with uniform distribution over $[0; 1]$.
- $\alpha_0 = V_0$ and for $1 \le j \le K_s$, $\alpha_j = \alpha_{j-1} + V_j$ where $V_0, \ldots, V_{K_s}$ are i.i.d. with uniform distribution over $[-1; -0.1] \cup [0.1; 1]$.
- K_σ, the number of jumps of σ, has uniform distribution in $\{5, \ldots, \lfloor\sqrt{n}\rfloor\}$.
- For $0 \le j \le K_\sigma$,
$$b_{j+1} - b_j = \Delta^\sigma_{\min} + \frac{(1 - (K_\sigma+1)\Delta^\sigma_{\min})U'_j}{\sum_{k=0}^{K_s} U'_k}$$
with $\Delta^\sigma_{\min} = \min\{5/n, 1/(K_\sigma+1)\}$ and $U'_0, \ldots, U'_{K_\sigma}$ are i.i.d. with uniform distribution over $[0; 1]$.
- $\beta_0, \ldots, \beta_{K_\sigma}$ are i.i.d. with uniform distribution over $[0.05; 0.5]$.

Two examples of a function s and a sample (t_i, Y_i) generated in framework A are plotted on Figure 1.

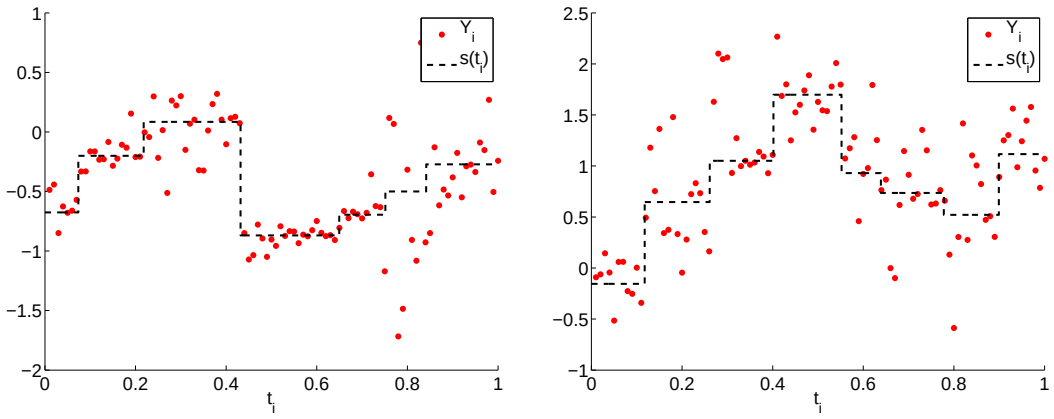

Figure 1: Random framework A: two examples of a sample $(t_i, Y_i)_{1 \leq i \leq 100}$ and the corresponding regression function s.

2.2.2 Framework B

The only difference with framework A is that $U_0, \ldots, U_{K_s}$ are i.i.d. with the same distribution as $Z = |10Z_1 + Z_2|$ where Z_1 has Bernoulli distribution with parameter $1/2$ and Z_2 has a standard Gaussian distribution. Two examples of a function s and a sample (t_i, Y_i) generated in framework B are plotted on Figure 2.

2.2.3 Framework C

The main difference between frameworks C and B is that $[0;1]$ is split into two regions: $a_{K_{s,1}+1} = 1/2$ and $K_s = K_{s,1} + K_{s,2} + 1$ for some positive integers $K_{s,1}, K_{s,2}$, and the bounds of the distribution of β_j are larger when $b_j \geq 1/2$ and smaller when $b_j < 1/2$. Two examples of a function s and a sample (t_i, Y_i) generated in framework C are plotted on Figure 3. More precisely, s and σ are generated as follows:

- $K_{s,1}$ has uniform distribution over $\{2, \ldots, K_{\max,1}\}$ with $K_{\max,1} = \lfloor \sqrt{n} \rfloor - 1 - \lfloor (\lfloor \sqrt{n} - 1 \rfloor)/3 \rfloor$.
- $K_{s,2}$ has uniform distribution over $\{0, \ldots, K_{\max,2}\}$ with $K_{\max,2} = \lfloor (\lfloor \sqrt{n} - 1 \rfloor)/3 \rfloor$.
- Let $U_0, \ldots, U_{K_s}$ be i.i.d. random variables with the same distribution as $Z = |10Z_1 + Z_2|$ where Z_1 has Bernoulli distribution with parameter $1/2$ and Z_2 has a standard Gaussian distribution.
- For $0 \leq j \leq K_{s,1}$,

$$a_{j+1} - a_j = \Delta_{\min}^{s,1} + \frac{(1 - (K_{s,1} + 1)\Delta_{\min}^{s,1})U_j}{\sum_{k=0}^{K_{s,1}} U_k}$$

  with $\Delta_{\min}^{s,1} = \min\{5/n, 1/(K_{s,1} + 1)\}$.

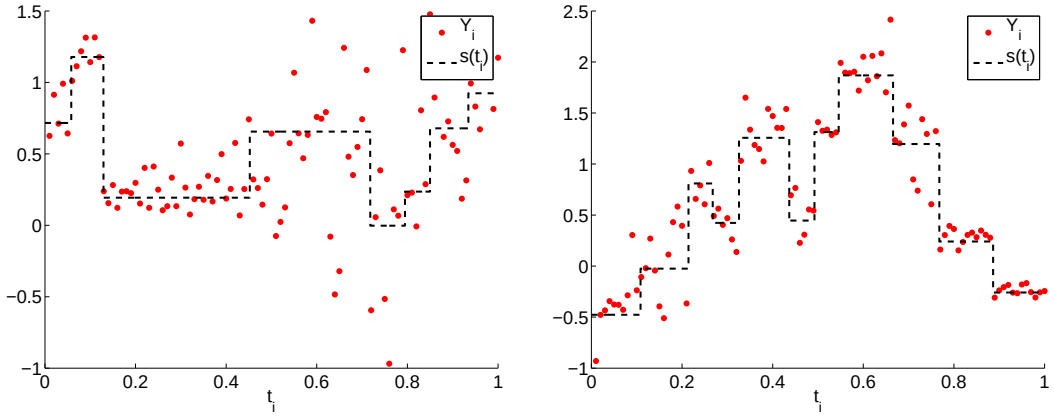

Figure 2: Random framework B: two examples of a sample $(t_i, Y_i)_{1\leq i\leq 100}$ and the corresponding regression function s.

- For $K_{s,1}+1 \leq j \leq K_s$,

$$a_{j+1} - a_j = \Delta_{\min}^{s,2} + \frac{(1-(K_{s,2}+1)\Delta_{\min}^{s,2})U_j}{\sum_{k=K_{s,1}+1}^{K_s} U_k}$$

  with $\Delta_{\min}^{s,2} = \min\{5/n, 1/(K_{s,2}+1)\}$.

- $\alpha_0 = V_0$ and for $1 \leq j \leq K_s$, $\alpha_j = \alpha_{j-1} + V_j$ where $V_0, \ldots, V_{K_s}$ are i.i.d. with uniform distribution over $[-1;-0.1] \cup [0.1;1]$.

- K_σ, $(b_{j+1} - b_j)_{0\leq j\leq K_\sigma}$ are distributed as in frameworks A and B.

- $\beta_0, \ldots, \beta_{K_\sigma}$ are independent.
  When $b_j < 1/2$, β_j has uniform distribution over $[0.025; 0.2]$.
  When $b_j \geq 1/2$, β_j has uniform distribution over $[0.1; 0.8]$.

3 Restrictions on the model collection with cross-validation

The use of CV induces some restrictions on the collection of models we consider.

First, when used for choosing the best segmentation of each dimension (Section 3 of the main paper), CV estimators require that each interval of the considered segmentation contains at least two observations: one belonging to the training sample, and the other to the validation sample. This implies a first constraint of the largest dimension: $D \leq n/2$. Note that there is only one possible segmentation with two points in each interval (with n even).

Second, when the estimation of the number $D-1$ of change-points is addressed in Section 4 of the main paper, $\mathrm{crit}_{\mathrm{VF}_V}(D)$ is computed by considering the models $\widetilde{S}_D((t_i)_{i\notin B_k})$ for $k = 1, \ldots, V$, which are non-empty only when $D \leq (n - \max_k\{\mathrm{Card}(B_k)\})/2 \approx$

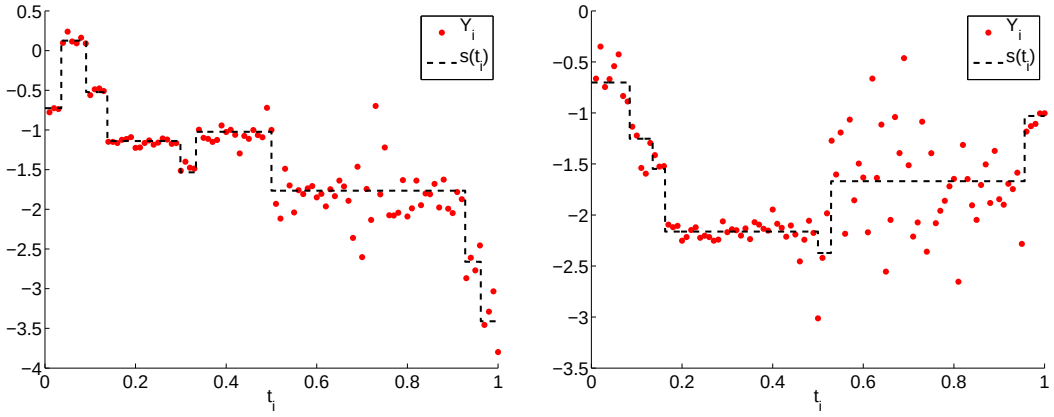

Figure 3: Random framework C: two examples of a sample $(t_i, Y_i)_{1 \leq i \leq 100}$ and the corresponding regression function s.

$n(V-1)/(2V)$. Indeed, at each round of the V-fold process, $\max_k \{\mathrm{Card}(B_k)\}$ points are removed from the whole sample. Since, moreover, Lpo is used at each round to choose the best segmentations, the largest possible dimension is $n(V-1)/(2V)$.

From a pratical point of view, the blocks $(B_k)_{1 \leq k \leq V}$ are always chosen such that $\forall i, k$, $\{t_i\,,\, t_{i+1}\} \cap B_k^c \neq \emptyset$, so that $\widetilde{S}_D((t_i)_{i \notin B_k})$ is not too different from $\widetilde{S}_D((t_i)_{1 \leq i \leq n})$. Despite this careful choice of $(B_k)_{1 \leq k \leq V}$, $\widetilde{S}_D((t_i)_{i \notin B_k})$ is still much smaller than $\widetilde{S}_D((t_i)_{1 \leq i \leq n})$ when D is close to $n(V-1)/(2V)$. Since the curves on Figure 4 of the main paper reflect the maximum *estimation error* over all the segmentations in $\widetilde{S}_D$ (up to the bias), the smaller $\widetilde{S}_D((t_i)_{i \notin B_k})$, the lower the value of the curve. This explains why $\mathrm{crit}_{\mathrm{VF}_V}(D)$ decreases for the largest values of D on Figure 4 of the main paper.

In practical applications, since this artefact could lead $\mathrm{crit}_{\mathrm{VF}_V}(D)$ to underestimate $\|s - \widehat{s}_{\widehat{m}(D)}\|_n^2$ for values of D close to $n(V-1)/(2V)$, we suggest to discard values of $D > 9n(V-1)/(20V)$. This restriction of the collection of models has been made, for all procedures, in the main paper. Since we took $V = 5$ and $n = 100$ in all our experiments, this leads to the choice

$$\mathcal{D}_n = \left\{1, \ldots, \frac{9n(V-1)}{20V}\right\} = \{1, \ldots, 36\}$$

instead of $\mathcal{D}_n = \{1, \ldots, 40\}$.

4 Calibration of Birgé and Massart's penalization

Birgé and Massart's penalization makes use of the penalty

$$\mathrm{pen}_{\mathrm{BM}}(D) := \frac{\widehat{C} D}{n}\left(5 + 2\log\left(\frac{n}{D}\right)\right) \quad .$$

In a previous version of this work [6, Chapter 7], $\widehat{C}$ was defined as suggested in [7, 8], that is, $\widehat{C} = 2\widehat{K}_{\mathrm{max.jump}}$ with the notation below. This yielded poor performances, which

seemed related to the definition of $\widehat{C}$. Therefore, alternative definitions for $\widehat{C}$ have been investigated, leading to the choice $\widehat{C} = \widehat{\sigma^2}$ defined by (1) below. The present section intends to motivate this choice.

Two main approaches have been considered in the literature for defining $\widehat{C}$ in the penalty $\mathrm{pen}_{\mathrm{BM}}$:

- Use $\widehat{C} = \widehat{\sigma^2}$ any estimate of the (average) noise-level, for instance,

$$\widehat{\sigma^2} := \frac{1}{n} \sum_{i=1}^{n/2} (Y_{2i} - Y_{2i-1})^2 \quad , \tag{1}$$

  assuming n is even and $t_1 < \cdots < t_n$.

- Use Birgé and Massart's *slope heuristics*, that is, compute the sequence

$$\widehat{D}(K) := \arg\min_{D \in \mathcal{D}_n} \left\{ P_n \gamma(\widehat{s}_{\widehat{m}_{\mathrm{ERM}}(D)}) + \frac{KD}{n} \left(5 + 2 \log\left(\frac{n}{D} \right) \right) \right\} \quad ,$$

  find the (unique) $K = \widehat{K}_{\mathrm{jump}}$ at which $\widehat{D}(K)$ jumps from large to small values, and define $\widehat{C} = 2\widehat{K}_{\mathrm{jump}}$.

The first approach follows from theoretical and experimental results [4, 8] which show that $\widehat{C}$ should be close to σ^2 when the noise-level is constant; (1) is a classical estimator of the variance used for instance by Baraud [3] for model selection in a different setting.

The optimality (in terms of oracle inequalities) of the second approach has been proved for regression with *homoscedastic Gaussian noise* and possibly exponential collections of models [5], as well as in a heteroscedastic framework with polynomial collections of models [2]. In the context of change-point detection with *homoscedastic data*, Lavielle [7] and Lebarbier [8] showed that $\widehat{C} = 2\widehat{K}_{\mathrm{max.jump}}$ can even perform better than $\widehat{C} = \sigma^2$ when $\widehat{K}_{\mathrm{max.jump}}$ corresponds to the highest jump of $\widehat{D}(K)$.

Alternatively, it was proposed in [2] to define $\widehat{C} = 2\widehat{K}_{\mathrm{thresh.}}$ where

$$\widehat{K}_{\mathrm{thresh.}} := \min \left\{ K \text{ s.t. } \widehat{D}(K) \leq D_{\mathrm{thresh.}} := \left\lfloor \frac{n}{\ln(n)} \right\rfloor \right\} \quad . \tag{2}$$

These three definitions of $\widehat{C}$ have been compared with $\widehat{C} = \sigma^2_{\mathrm{true}} := n^{-1} \sum_{i=1}^{n} \sigma(t_i)^2$ in the settings of the paper. The results are reported in Table 2. The main conclusions are the following.

- $2\widehat{K}_{\mathrm{thresh.}}$ almost always beats $2\widehat{K}_{\mathrm{max.jump}}$, even in homoscedastic settings. This confirms some simulation results reported in [2].

- σ^2_{true} almost always beats slope heuristics-based definitions of $\widehat{C}$, but not always, as previously noticed by Lebarbier [8]. Differences of performance can be huge (in particular when $\sigma = \sigma_s$), but not always in favour of σ^2_{true} (for instance, when $s = s_3$).

- $\widehat{\sigma^2}$ yields significantly better performance than σ^2_{true} in most settings (but not all), with huge margins in some heteroscedastic settings.

$s.$	$\sigma.$	$2\widehat{K}_{\text{max.jump}}$	$2\widehat{K}_{\text{thresh.}}$	$\widehat{\sigma^2}$	σ^2_{true}
1	c	17.12 ± 0.27	5.87 ± 0.03	**1.70** ± 0.02	1.97 ± 0.02
	pc,1	134.35 ± 1.50	11.00 ± 0.03	**1.24** ± 0.02	7.24 ± 0.04
	pc,2	35.60 ± 0.33	11.20 ± 0.03	**3.10** ± 0.03	7.63 ± 0.04
	pc,3	23.60 ± 0.20	11.28 ± 0.03	**3.81** ± 0.04	7.67 ± 0.04
	s	46.36 ± 0.48	10.14 ± 0.03	2.08 ± 0.02	**1.63** ± 0.02
2	c	6.34 ± 0.04	6.97 ± 0.03	3.58 ± 0.02	**3.54** ± 0.02
	pc,1	15.01 ± 0.08	17.63 ± 0.05	**6.87** ± 0.06	11.74 ± 0.06
	pc,2	15.25 ± 0.07	17.94 ± 0.05	**9.25** ± 0.06	13.00 ± 0.06
	pc,3	15.13 ± 0.07	17.59 ± 0.05	**8.79** ± 0.06	12.66 ± 0.07
	s	8.80 ± 0.04	10.05 ± 0.03	**4.76** ± 0.03	11.00 ± 0.02
3	c	5.17 ± 0.03	4.68 ± 0.01	4.67 ± 0.01	**4.19** ± 0.01
	pc,1	7.37 ± 0.11	5.98 ± 0.02	**4.55** ± 0.02	5.09 ± 0.02
	pc,2	7.15 ± 0.03	6.83 ± 0.02	**5.90** ± 0.02	5.95 ± 0.02
	pc,3	7.16 ± 0.02	7.19 ± 0.02	**6.24** ± 0.02	6.31 ± 0.02
	s	8.81 ± 0.08	7.38 ± 0.02	**5.64** ± 0.02	15.13 ± 0.04
4	c	17.73 ± 0.28	5.92 ± 0.03	2.05 ± 0.02	**1.97** ± 0.02
	pc,2	35.50 ± 0.34	11.40 ± 0.03	**5.54** ± 0.04	7.77 ± 0.04
A		7.13 ± 0.04	7.34 ± 0.03	**4.68** ± 0.03	4.84 ± 0.03
B		6.84 ± 0.04	7.16 ± 0.03	**4.63** ± 0.03	4.74 ± 0.03
C		10.43 ± 0.06	12.25 ± 0.06	**7.34** ± 0.05	8.92 ± 0.05

Table 2: Performance $C_{\text{or}}(\text{BM})$ with four different definitions of $\widehat{C}$ (see text), in some of the simulation settings considered in the paper. See the text for the definition of s_4. In each setting, $N = 10\,000$ independent samples have been generated. Next to each value is indicated the corresponding empirical standard deviation divided by $\sqrt{N}$, measuring the uncertainty of the estimated performance.

The latter result actually comes from an artefact, which can be explained by the bias of $\widehat{\sigma^2}$ as an estimator of σ^2_{true}. Indeed,

$$\mathbb{E}\left[\widehat{\sigma^2}\right] = \frac{1}{n}\sum_{i=1}^{n}\sigma(t_i)^2 + \frac{1}{n}\sum_{i=1}^{n/2}\left(s(t_{2i}) - s(t_{2i-1})\right)^2 \geq \frac{1}{n}\sum_{i=1}^{n}\sigma(t_i)^2 = \sigma^2_{\text{true}} \ . \tag{3}$$

The difference between these expectations is not negligible in all the settings of the paper. For instance, when $n = 100$, $t_i = i/n$ and $s = s_1$, $n^{-1}\sum_{i=1}^{n}(s(t_{2i}) - s(t_{2i-1}))^2 = 0.04$ whereas σ^2_{true} varies between 0.015 (when $\sigma = \sigma_{pc,1}$) and 0.093 (when $\sigma = \sigma_{pc,3}$). To stress the influence of the bias on the performance, we introduced s_4, which is based on s_1 with its four jumps shifted by 1/200 to the right (that is, with jumps at 0.205, 0.405, 0.605 and 0.805). When $s = s_4$, $\widehat{\sigma^2}$ is an unbiased estimator of σ^2_{true}, which significantly deteriorates the performances of BM (see Table 2), while the performances of other procedures do not change.

The reason for this change of performance is that overpenalization improves the results of BM in most of the considered heteroscedastic settings, as shown by the right panel of Figure 4 of the main paper. Indeed, pen_{BM} is a poor penalty when data are heteroscedastic, underpenalizing dimensions close to the oracle but overpenalizing the largest dimensions. Then, in a setting like $(s_2, \sigma_{pc,3})$ multiplying pen_{BM} by a factor $C_{\text{over}} > 1$ helps decreasing the selected dimension; the same cause has opposite consequences in other settings, such as (s_1, σ_s) or (s_3, σ_c). Nevertheless, even choosing $\widehat{C}$ using both P_n and s, $(\text{crit}_{\text{BM}}(D))_{D>0}$ remains a poor estimate of $\left(\left\|s - \widehat{s}_{\widehat{m}_{\text{ERM}}(D)}\right\|_n^2\right)_{D>0}$ in most heteroscedastic settings (even up to an additive constant). Choosing the overpenalization factor from data is a difficult problem, especially without knowing *a priori* whether the signal is homoscedastic or heteroscedastic. This question deserves a specific extensive simulation experiment. To be completely fair with CV methods, such an experiment should also compare BM with overpenalization to V-fold penalization [1] with overpenalization, for choosing the number of change-points.

To conclude, pen_{BM} with $\widehat{C} = \widehat{\sigma^2}$ is not a reliable change-point detection procedure, and the apparently good performances observed in Table 2 could be misleading.

Results of Table 2 for $\widehat{C} = \sigma^2_{\text{true}}$ indicate how far the performances of BM could be improved without overpenalization. According to Tables 5 and 1, BM with $\widehat{C} = \sigma^2_{\text{true}}$ only has significantly better performances than $[\![\text{ERM}, \text{VF}_5]\!]$ or $[\![\text{Loo}, \text{VF}_5]\!]$ in the three homoscedastic settings and in setting (s_1, σ_s).

5 Additional results from the simulation study

In the next pages are presented extended versions of the Tables of the main paper, and two additional Figures.

References

[1] Sylvain Arlot. V-fold cross-validation improved: V-fold penalization, February 2008. arXiv:0802.0566v2.

[2] Sylvain Arlot and Pascal Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10:245–279 (electronic), 2009.

[3] Yannick Baraud. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6:127–146 (electronic), 2002.

[4] Lucien Birgé and Pascal Massart. Gaussian model selection. *J. European Math. Soc.*, 3(3):203–268, 2001.

[5] Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.

[6] Alain Celisse. *Model selection via cross-validation in density estimation, regression and change-points detection*. PhD thesis, University Paris-Sud 11, December 2008. http://tel.archives-ouvertes.fr/tel-00346320/.

[7] Marc Lavielle. Using penalized contrasts for the change-point problem. *Signal Proces.*, 85:1501–1510, 2005.

[8] Émilie Lebarbier. Detecting multiple change-points in the mean of a Gaussian process by model selection. *Signal Proc.*, 85:717–736, 2005.

$s.$	$\sigma.$	ERM	Loo	Lpo_{20}	Lpo_{50}	PML
1	c	**1.56** ± 0.01	**1.56** ± 0.01	**1.56** ± 0.01	**1.55** ± 0.01	**1.57** ± 0.01
	pc,1	**1.12** ± 0.01	**1.14** ± 0.02	**1.14** ± 0.02	**1.13** ± 0.01	**1.14** ± 0.01
	pc,2	**1.82** ± 0.02	**1.81** ± 0.02	**1.82** ± 0.02	**1.80** ± 0.02	1.85 ± 0.02
	pc,3	**2.07** ± 0.02	**2.07** ± 0.02	**2.07** ± 0.02	**2.07** ± 0.02	2.19 ± 0.02
	s	**1.51** ± 0.02	**1.51** ± 0.02	**1.51** ± 0.02	**1.50** ± 0.02	1.56 ± 0.02
2	c	**2.87** ± 0.01	**2.89** ± 0.01	2.90 ± 0.01	2.96 ± 0.01	3.02 ± 0.01
	pc,1	1.33 ± 0.02	1.15 ± 0.02	1.14 ± 0.01	1.11 ± 0.01	**1.04** ± 0.01
	pc,2	2.91 ± 0.02	2.21 ± 0.02	2.15 ± 0.02	2.02 ± 0.02	**1.28** ± 0.01
	pc,3	3.14 ± 0.03	2.52 ± 0.02	2.47 ± 0.02	2.36 ± 0.02	**1.44** ± 0.01
	s	**2.98** ± 0.01	**2.98** ± 0.01	**3.00** ± 0.01	3.08 ± 0.01	3.17 ± 0.01
3	c	**3.18** ± 0.01	3.25 ± 0.01	3.29 ± 0.01	3.44 ± 0.01	3.71 ± 0.01
	pc,1	3.04 ± 0.02	2.70 ± 0.02	2.71 ± 0.02	2.79 ± 0.02	**2.29** ± 0.02
	pc,2	4.14 ± 0.02	3.69 ± 0.02	3.72 ± 0.02	3.86 ± 0.02	**3.12** ± 0.01
	pc,3	4.44 ± 0.02	3.98 ± 0.02	4.00 ± 0.02	4.14 ± 0.02	**3.24** ± 0.01
	s	3.92 ± 0.01	**3.72** ± 0.01	3.75 ± 0.01	3.88 ± 0.01	3.81 ± 0.01
A		**3.28** ± 0.02	**3.25** ± 0.02	**3.28** ± 0.02	3.39 ± 0.02	3.42 ± 0.02
B		3.21 ± 0.02	**3.15** ± 0.02	**3.18** ± 0.02	3.28 ± 0.02	3.38 ± 0.02
C		4.27 ± 0.03	3.79 ± 0.03	3.79 ± 0.03	3.78 ± 0.03	**3.14** ± 0.02

Table 3: Average performance $C_{\text{or}}\left([\![\mathfrak{P}, \text{Id}]\!]\right)$ for change-point detection procedures $\mathfrak{P}$ among ERM , Loo and Lpo_p with $p = 20$ and $p = 50$. Several regression functions s and noise-level functions σ have been considered, each time with $N = 10\,000$ independent samples. Next to each value is indicated the corresponding empirical standard deviation divided by $\sqrt{N}$, measuring the uncertainty of the estimated performance.

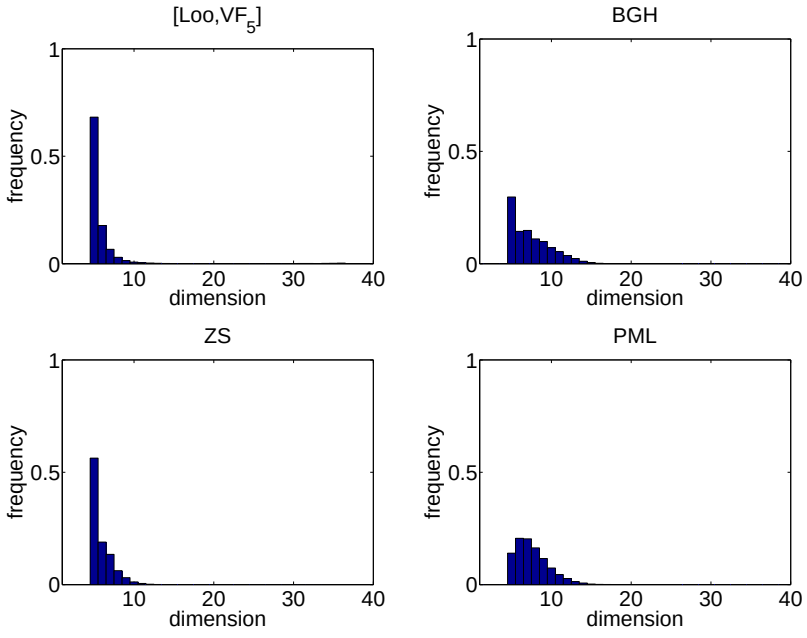

Figure 4: Empirical distribution of the selected dimension $\widehat{D}$ over $N = 10\,000$ independent samples, for 4 procedures, in setting $(s_1, \sigma_{pc,3})$. The true model has dimension $D = 5$.

$s.$	$\sigma.$	Id	VF_5	BM	BGH
1	c	1.56 ± 0.01	2.38 ± 0.02	**1.70** ± 0.02	1.85 ± 0.02
	pc,1	1.12 ± 0.01	2.67 ± 0.03	**1.24** ± 0.02	5.83 ± 0.05
	pc,2	1.82 ± 0.02	3.20 ± 0.03	**3.10** ± 0.03	6.38 ± 0.05
	pc,3	2.07 ± 0.02	**3.42** ± 0.03	3.81 ± 0.04	6.51 ± 0.04
	s	1.51 ± 0.02	2.73 ± 0.03	**2.08** ± 0.02	3.83 ± 0.03
2	c	2.87 ± 0.01	3.99 ± 0.02	3.58 ± 0.02	**3.52** ± 0.02
	pc,1	1.33 ± 0.02	**3.64** ± 0.05	6.87 ± 0.06	9.50 ± 0.07
	pc,2	2.91 ± 0.02	**5.62** ± 0.05	9.25 ± 0.06	10.13 ± 0.07
	pc,3	3.14 ± 0.03	**5.94** ± 0.06	8.79 ± 0.06	9.77 ± 0.07
	s	2.98 ± 0.01	**4.34** ± 0.03	4.76 ± 0.03	4.88 ± 0.03
3	c	3.18 ± 0.01	**4.31** ± 0.02	4.67 ± 0.01	4.47 ± 0.01
	pc,1	3.04 ± 0.02	4.65 ± 0.02	**4.55** ± 0.02	4.92 ± 0.02
	pc,2	4.14 ± 0.02	**5.82** ± 0.02	5.90 ± 0.02	5.93 ± 0.02
	pc,3	4.44 ± 0.02	**6.13** ± 0.02	6.24 ± 0.02	6.31 ± 0.02
	s	3.92 ± 0.01	**5.61** ± 0.02	**5.64** ± 0.02	**5.63** ± 0.02
A		3.28 ± 0.02	**4.67** ± 0.03	**4.68** ± 0.03	4.85 ± 0.03
B		3.21 ± 0.02	**4.61** ± 0.03	**4.63** ± 0.03	4.78 ± 0.03
C		4.27 ± 0.03	**6.49** ± 0.04	7.34 ± 0.05	7.70 ± 0.05

Table 4: Performance $C_{\mathrm{or}}\left(\llbracket \mathrm{ERM}, \mathfrak{P} \rrbracket\right)$ for $\mathfrak{P} = \mathrm{Id}$ (that is, choosing the dimension $D^\star := \arg\min_{D \in \mathcal{D}_n} \left\{ \left\| s - \widehat{s}_{\widehat{m}_{\mathrm{ERM}}(D)} \right\|_n^2 \right\}$), $\mathfrak{P} = \mathrm{VF}_V$ with $V = 5$, $\mathfrak{P} = \mathrm{BM}$ or $\mathfrak{P} = \mathrm{BGH}$. Several regression functions s and noise-level functions σ have been considered, each time with $N = 10\,000$ independent samples. Next to each value is indicated the corresponding empirical standard deviation divided by $\sqrt{N}$, measuring the uncertainty of the estimated performance.

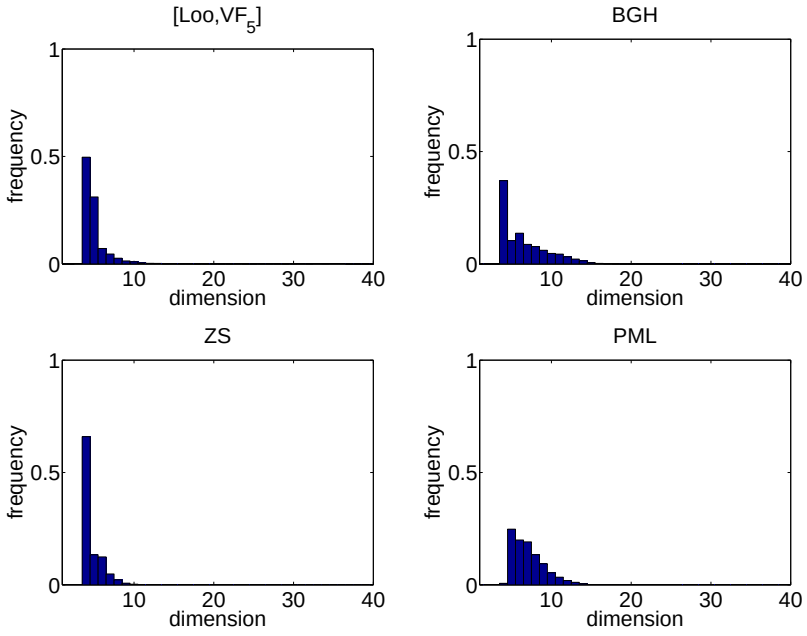

Figure 5: Same as Figure 4 in setting $(s_2, \sigma_{pc,2})$. The true model has dimension $D = 5$.

(s, σ)	(s_1, σ_c)	$(s_1, \sigma_{pc,1})$	$(s_1, \sigma_{pc,2})$	$(s_1, \sigma_{pc,3})$	(s_1, σ_s)
$[\![\mathrm{Loo}, \mathrm{VF}_5]\!]$	2.40 ± 0.02	2.64 ± 0.03	3.17 ± 0.03	**3.40** ± 0.03	2.59 ± 0.03
$[\![\mathrm{Lpo}_{20}, \mathrm{VF}_5]\!]$	2.58 ± 0.02	3.78 ± 0.04	3.29 ± 0.03	3.55 ± 0.04	2.86 ± 0.03
$[\![\mathrm{Lpo}_{50}, \mathrm{VF}_5]\!]$	2.69 ± 0.03	3.97 ± 0.04	3.60 ± 0.04	3.77 ± 0.04	3.15 ± 0.04
$[\![\mathrm{ERM}, \mathrm{VF}_5]\!]$	2.38 ± 0.02	2.67 ± 0.03	3.20 ± 0.03	**3.42** ± 0.03	2.73 ± 0.03
$[\![\mathrm{ERM}, \mathrm{BM}]\!]$	**1.70** ± 0.02	**1.24** ± 0.02	**3.10** ± 0.03	3.81 ± 0.04	**2.08** ± 0.02
$[\![\mathrm{ERM}, \mathrm{BM}_{\mathrm{thr}}]\!]$	5.87 ± 0.03	11.00 ± 0.03	11.20 ± 0.03	11.28 ± 0.03	10.14 ± 0.03
BGH	1.85 ± 0.02	5.83 ± 0.05	6.38 ± 0.05	6.51 ± 0.04	3.83 ± 0.03
ZS	**1.71** ± 0.02	2.20 ± 0.03	3.92 ± 0.04	4.41 ± 0.04	2.46 ± 0.03
PML	2.79 ± 0.02	2.48 ± 0.03	3.42 ± 0.03	3.97 ± 0.04	2.72 ± 0.03

(s, σ)	(s_2, σ_c)	$(s_2, \sigma_{pc,1})$	$(s_2, \sigma_{pc,2})$	$(s_2, \sigma_{pc,3})$	(s_2, σ_s)
$[\![\mathrm{Loo}, \mathrm{VF}_5]\!]$	4.02 ± 0.02	3.14 ± 0.04	4.95 ± 0.05	5.24 ± 0.05	4.32 ± 0.03
$[\![\mathrm{Lpo}_{20}, \mathrm{VF}_5]\!]$	4.10 ± 0.02	3.34 ± 0.05	5.00 ± 0.05	5.26 ± 0.05	4.49 ± 0.03
$[\![\mathrm{Lpo}_{50}, \mathrm{VF}_5]\!]$	4.44 ± 0.03	4.41 ± 0.06	5.20 ± 0.05	5.54 ± 0.06	4.86 ± 0.03
$[\![\mathrm{ERM}, \mathrm{VF}_5]\!]$	3.99 ± 0.02	3.64 ± 0.05	5.62 ± 0.05	5.94 ± 0.06	4.34 ± 0.03
$[\![\mathrm{ERM}, \mathrm{BM}]\!]$	3.58 ± 0.02	6.87 ± 0.06	9.25 ± 0.06	8.79 ± 0.06	4.76 ± 0.03
$[\![\mathrm{ERM}, \mathrm{BM}_{\mathrm{thr}}]\!]$	6.97 ± 0.03	17.63 ± 0.05	17.94 ± 0.05	17.59 ± 0.05	10.05 ± 0.03
BGH	**3.52** ± 0.02	9.50 ± 0.07	10.13 ± 0.07	9.77 ± 0.07	4.88 ± 0.03
ZS	3.62 ± 0.02	3.36 ± 0.05	6.50 ± 0.05	6.92 ± 0.06	**4.16** ± 0.02
PML	4.34 ± 0.02	**2.35** ± 0.03	**2.73** ± 0.03	**2.84** ± 0.03	4.32 ± 0.02

(s, σ)	(s_3, σ_c)	$(s_3, \sigma_{pc,1})$	$(s_3, \sigma_{pc,2})$	$(s_3, \sigma_{pc,3})$	(s_3, σ_s)
$[\![\mathrm{Loo}, \mathrm{VF}_5]\!]$	4.42 ± 0.02	4.22 ± 0.02	5.24 ± 0.02	5.59 ± 0.02	**5.35** ± 0.02
$[\![\mathrm{Lpo}_{20}, \mathrm{VF}_5]\!]$	4.61 ± 0.02	4.62 ± 0.02	5.47 ± 0.02	5.78 ± 0.02	5.55 ± 0.02
$[\![\mathrm{Lpo}_{50}, \mathrm{VF}_5]\!]$	5.06 ± 0.02	4.84 ± 0.02	5.86 ± 0.02	6.25 ± 0.02	6.03 ± 0.02
$[\![\mathrm{ERM}, \mathrm{VF}_5]\!]$	**4.31** ± 0.02	4.65 ± 0.02	5.82 ± 0.02	6.13 ± 0.02	5.61 ± 0.02
$[\![\mathrm{ERM}, \mathrm{BM}]\!]$	4.67 ± 0.01	4.55 ± 0.02	5.90 ± 0.02	6.24 ± 0.02	5.64 ± 0.02
$[\![\mathrm{ERM}, \mathrm{BM}_{\mathrm{thr}}]\!]$	4.68 ± 0.01	5.98 ± 0.02	6.83 ± 0.02	7.19 ± 0.02	7.38 ± 0.02
BGH	4.47 ± 0.01	4.92 ± 0.02	5.93 ± 0.02	6.31 ± 0.02	5.63 ± 0.02
ZS	5.46 ± 0.02	4.44 ± 0.02	6.63 ± 0.02	6.61 ± 0.02	6.31 ± 0.02
PML	5.05 ± 0.02	**3.70** ± 0.03	**4.67** ± 0.03	**4.99** ± 0.03	5.52 ± 0.02

Table 5: Performance $C_{\mathrm{or}}(\mathfrak{P})$ for several change-point detection procedures $\mathfrak{P}$. '$\mathrm{BM}_{\mathrm{thr}}$' refers to BM with $\widehat{C} = 2\widehat{K}_{\mathrm{thresh.}}$ (see Section 4), whereas 'BM' correspond to $\widehat{C} = 2\widehat{\sigma^2}$ as everywhere in the main paper. Several regression functions s and noise-level functions σ have been considered, each time with $N = 10\,000$ independent samples. Next to each value is indicated the corresponding empirical standard deviation.

(s, σ)	(s_4, σ_c)	$(s_4, \sigma_{pc,2})$
$[\![\text{Loo}, \text{VF}_5]\!]$	2.36 ± 0.02	**3.17** ± 0.03
$[\![\text{Lpo}_{20}, \text{VF}_5]\!]$	2.53 ± 0.02	3.32 ± 0.03
$[\![\text{Lpo}_{50}, \text{VF}_5]\!]$	2.64 ± 0.03	3.62 ± 0.04
$[\![\text{ERM}, \text{VF}_5]\!]$	2.38 ± 0.02	3.24 ± 0.03
$[\![\text{ERM}, \text{BM}]\!]$	2.05 ± 0.02	5.54 ± 0.04
$[\![\text{ERM}, \text{BM}_{\text{thr}}]\!]$	5.92 ± 0.03	11.40 ± 0.03
BGH	1.85 ± 0.02	6.54 ± 0.05
ZS	**1.70** ± 0.02	4.00 ± 0.04
PML	2.77 ± 0.02	3.55 ± 0.04

Table 6: Same as Table 5 with $s = s_4$.

(s, σ)	$(s_2, \sigma_{pc,2})$	$(s_2, \sigma_{pc,3})$	$(s_3, \sigma_{pc,2})$	$(s_3, \sigma_{pc,3})$
$[\![\text{Loo}, \text{VF}_5]\!]$	**4.47** ± 0.05	**4.69** ± 0.06	**4.80** ± 0.03	**5.11** ± 0.03
$[\![\text{Lpo}_{20}, \text{VF}_5]\!]$	4.62 ± 0.06	4.88 ± 0.06	5.03 ± 0.03	5.29 ± 0.03
$[\![\text{Lpo}_{50}, \text{VF}_5]\!]$	5.22 ± 0.07	5.54 ± 0.07	5.45 ± 0.03	5.83 ± 0.03
$[\![\text{ERM}, \text{VF}_5]\!]$	5.98 ± 0.07	6.31 ± 0.07	5.82 ± 0.04	6.22 ± 0.04
$[\![\text{ERM}, \text{BM}]\!]$	10.81 ± 0.09	10.31 ± 0.09	6.09 ± 0.04	6.45 ± 0.04
$[\![\text{ERM}, \text{BM}_{\text{thr}}]\!]$	17.83 ± 0.09	17.26 ± 0.08	6.60 ± 0.03	6.99 ± 0.03
BGH	11.67 ± 0.09	11.15 ± 0.09	5.94 ± 0.03	6.42 ± 0.04
ZS	9.34 ± 0.09	9.29 ± 0.09	6.60 ± 0.04	6.83 ± 0.04
PML	5.04 ± 0.06	5.00 ± 0.06	5.17 ± 0.03	5.40 ± 0.03

Table 7: Same as Table 5 with ε_i having an exponential distribution.