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

Sylvain Arlot and Alain Celisse

April 8, 2009

1 Calibration of Birgé and Massart's penalization

Birgé and Massart's penalization makes use of the penalty

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

In a previous version of this work [6, Chapter 7], \hat{C} was defined as suggested in [7, 8], that is, $\hat{C} = 2\hat{K}_{\text{max,jump}}$ with the notation below. This yielded poor performances, which seemed related to the definition of \hat{C} . Therefore, alternative definitions for \hat{C} have been investigated, leading to the choice $\hat{C} = 2\hat{K}_{\text{thresh}}$ throughout the paper, where \hat{K}_{thresh} is defined by (2) below. The present appendix 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 noise-level, for instance,

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

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

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

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

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

The first approach follows from theoretical and experimental results [4, 8] which show that \hat{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

s.	σ .	$2\widehat{K}_{ m max.jump}$	$2\hat{K}_{\text{thresh.}}$	$\widehat{\sigma^2}$	$\sigma_{ m true}^2$
1	с	6.85 ± 0.12	3.91 ± 0.03	1.74 ± 0.02	2.05 ± 0.02
	pc,3	17.56 ± 0.15	13.08 ± 0.04	4.42 ± 0.04	10.43 ± 0.05
	\mathbf{s}	20.07 ± 0.31	9.41 ± 0.04	2.18 ± 0.03	1.66 ± 0.02
2	с	6.02 ± 0.03	5.27 ± 0.03	3.58 ± 0.02	3.54 ± 0.02
	pc,3	17.76 ± 0.10	20.12 ± 0.07	10.58 ± 0.07	16.64 ± 0.08
	\mathbf{s}	10.17 ± 0.05	9.69 ± 0.04	5.28 ± 0.03	10.95 ± 0.02
3	с	4.97 ± 0.02	4.39 ± 0.01	4.62 ± 0.01	4.21 ± 0.01
	pc,3	8.66 ± 0.03	8.47 ± 0.03	6.64 ± 0.02	8.00 ± 0.03
	\mathbf{s}	8.50 ± 0.04	7.59 ± 0.03	5.94 ± 0.02	15.50 ± 0.04
	А	7.52 ± 0.04	6.82 ± 0.03	4.86 ± 0.03	5.55 ± 0.03
	В	7.89 ± 0.04	7.21 ± 0.04	5.18 ± 0.03	5.77 ± 0.03
	\mathbf{C}	12.81 ± 0.08	13.49 ± 0.07	8.93 ± 0.06	12.44 ± 0.07

Table 1: Performance $C_{\rm or}(BM)$ with four different definitions of \widehat{C} (see text), in some of the simulation settings considered in the paper. 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} .

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 $\hat{C} = 2\hat{K}_{\max,\text{jump}}$ can even perform better than $\hat{C} = \sigma^2$ when $\hat{K}_{\max,\text{jump}}$ corresponds to the highest jump of $\hat{D}(K)$.

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

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

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

- $2\hat{K}_{\text{thresh.}}$ almost always beats $2\hat{K}_{\text{max.jump}}$, even in homoscedastic settings. This confirms some simulation results reported in [2].
- σ_{true}^2 often 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_{true}^2 (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.

The latter result actually comes from an artefact, which can be explained as follows. First,

$$\mathbb{E}\left[\widehat{\sigma^2}\right] = \frac{1}{n} \sum_{i=1}^n \sigma(t_i)^2 + \frac{1}{n} \sum_{i=1}^n \left(s(t_{2i}) - s(t_{2i-1})\right)^2 \ge \frac{1}{n} \sum_{i=1}^n \sigma(t_i)^2 = \sigma_{\text{true}}^2 \quad .$$

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 $s = s_1$, $n^{-1} \sum_i (s(t_{2i}) - s(t_{2i-1}))^2 = 0.04$ whereas σ_{true}^2 varies between 0.015 (when $\sigma = \sigma_{pc,1}$) to 0.093 (when $\sigma = \sigma_{pc,3}$). Nevertheless, $\widehat{\sigma}^2$ would not overestimate σ_{true}^2 at all in a very close setting: Shifting the jumps of s_1 by 1/100 is sufficient to make $n^{-1} \sum_i (s(t_{2i}) - s(t_{2i-1}))^2$ equal to zero, and the performances of BM with $\widehat{C} = \widehat{\sigma^2}$ would then be very close to the performances of BM with $\widehat{C} = \sigma_{\text{true}}$.

Second, overpenalization turns out to improve the results of BM in most of the heteroscedastic settings considered in the paper. The reason for this phenomenon is illustrated by the right panel of Figure 4. Indeed, pen_{BM} is a poor penalty when data are heteroscedastic, underpenalizing dimensions close to the oracle but overpenalizing the largest dimensions (remember that $\hat{C} = 2\hat{K}_{\text{thresh.}}$ on Figure 4). 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 different consequences in other settings, such as $(s_1, \sigma_s \text{ or } (s_3, \sigma_c)$. Nevertheless, even choosing \hat{C} using both P_n and s, $(\text{crit}_{BM}(D))_{D>0}$ remains a poor estimate of $\left(\left\| s - \hat{s}_{\widehat{m}_{\text{ERM}}(D)} \right\|_n^2 \right)_{D>0}$ in most heteroscedastic settings (even up to an additive constant).

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 1 could be misleading. This leads to the remaining choice $\widehat{C} = 2\widehat{K}_{\text{thresh.}}$ which has been used throughout the paper, although this calibration method may certainly be improved.

Results of Table 1 for $\hat{C} = \sigma_{\text{true}}^2$ indicate how far the performances of pen_{BM} could be improved without overpenalization. According to Tables 4 and 5, BM with $\hat{C} = \sigma_{\text{true}}^2$ only has significantly better performances than [[ERM, VF₅]] or [[Loo, VF₅]] in the three homoscedastic settings and in setting (s_1, σ_s) .

Finally, overpenalization could be used to improve BM, but 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.

2 Random frameworks generation

The purpose of this appendix is to detail how piecewise constant functions s and σ have been generated in the frameworks A, B and C of Section 5.3. In each framework, s and σ are of the form

$$s(x) = \sum_{j=0}^{K_s - 1} \alpha_j \mathbb{1}_{[a_j; a_{j+1})} + \alpha_{K_s} \mathbb{1}_{[a_{K_s}; a_{K_s + 1}]} \qquad \text{with } a_0 = 0 < a_1 < \dots < a_{K_s} = 1$$
$$\sigma(x) = \sum_{j=0}^{K_\sigma - 1} \beta_j \mathbb{1}_{[b_j; b_{j+1})} + \beta_{K_\sigma} \mathbb{1}_{[b_{K_\sigma}; b_{K_\sigma + 1}]} \qquad \text{with } b_0 = 0 < b_1 < \dots < b_{K_\sigma} = 1$$

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).

2.1 Framework A

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

- K_s , the number of jumps of s, has uniform distribution over $\{3, \ldots, |\sqrt{n}|\}$.
- For $0 \le j \le K_s$,

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

with $\Delta_{\min}^s = \min \{5/n, 1/(K_s+1)\}$ and U_0, \ldots, U_{K_s} are i.i.d. with uniform distribution over [0, 1].

- $\alpha_0 = V_0$ and for $1 \leq j \leq K_s$, $\alpha_j = \alpha_{j-1} + V_j$ where V_0, \ldots, V_{K_s} are i.i.d. with uniform distribution over $[-1; -0.1] \cup [0.1; 1]$.
- K_{σ} , the number of jumps of σ , has uniform distribution in $\{5, \ldots, |\sqrt{n}|\}$.
- For $0 \leq j \leq K_{\sigma}$,

$$b_{j+1} - b_j = \Delta_{\min}^{\sigma} + \frac{(1 - (K_{\sigma} + 1)\Delta_{\min}^{\sigma})U'_j}{\sum_{k=0}^{K_s} U'_k}$$

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

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.



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



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

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

• For $K_{s,1} + 1 \leq j \leq K_s$,

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

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} , $(b_{j+1} b_j)_{0 \le j \le K_{\sigma}}$ 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 $b_j \ge 1/2$, β_j has uniform distribution over [0.1; 0.8].

3 Additional results from the simulation study

In the next pages are presented extended versions of the Tables of the main paper, as well as an extended version of Table 1 (Table 7).



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

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] L. Birgé and P. 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. oai:tel.archives-ouvertes.fr:tel-00346320_v1.
- [7] M. Lavielle. Using penalized contrasts for the change-point problem. Signal Proces., 85:1501–1510, 2005.
- [8] E. Lebarbier. Detecting multiple change-points in the mean of a Gaussian process by model selection. Signal Proc., 85:717–736, 2005.

s.	σ .	ERM	Loo	Lpo_{20}	Lpo_{50}
1	с	$\textbf{1.59}\pm0.01$	1.60 ± 0.02	1.58 ± 0.01	$\textbf{1.58}\pm0.01$
	$_{\rm pc,1}$	1.04 ± 0.01	1.06 ± 0.01	1.06 ± 0.01	1.06 ± 0.01
	pc,2	1.89 ± 0.02	1.87 ± 0.02	1.87 ± 0.02	1.87 ± 0.02
	pc,3	2.05 ± 0.02	2.05 ± 0.02	$\textbf{2.05} \pm 0.02$	2.07 ± 0.02
	\mathbf{s}	1.54 ± 0.02	1.52 ± 0.02	1.52 ± 0.02	1.51 ± 0.02
2	с	2.88 ± 0.01	2.93 ± 0.01	2.93 ± 0.01	2.94 ± 0.01
	pc,1	1.31 ± 0.02	1.16 ± 0.02	1.14 ± 0.02	$\textbf{1.11}\pm0.01$
	pc,2	2.88 ± 0.02	2.24 ± 0.02	2.19 ± 0.02	$\textbf{2.13} \pm 0.02$
	pc,3	3.09 ± 0.03	2.52 ± 0.03	2.48 ± 0.03	2.32 ± 0.03
	\mathbf{s}	$\textbf{3.01} \pm 0.01$	$\textbf{3.03}\pm0.01$	3.05 ± 0.01	3.13 ± 0.01
3	с	$\textbf{3.18} \pm 0.01$	3.25 ± 0.01	3.29 ± 0.01	3.44 ± 0.01
	pc,1	3.00 ± 0.01	2.67 ± 0.02	2.68 ± 0.02	2.77 ± 0.02
	pc,2	4.06 ± 0.02	$\textbf{3.63} \pm 0.02$	$\textbf{3.64} \pm 0.02$	3.78 ± 0.02
	pc,3	4.41 ± 0.02	$\textbf{3.97} \pm 0.02$	4.00 ± 0.02	4.11 ± 0.02
	\mathbf{s}	4.02 ± 0.01	$\textbf{3.82}\pm0.01$	3.85 ± 0.01	3.98 ± 0.01

Table 2: Average performance $C_{\text{or}}(\llbracket \mathfrak{P}, \operatorname{Id} \rrbracket)$ 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.

s.	σ.	Oracle	VF_5	BM
1	с	1.59 ± 0.01	5.40 ± 0.05	$\textbf{3.91}\pm0.03$
	$_{\rm pc,1}$	1.04 ± 0.01	$\textbf{11.96} \pm 0.03$	12.85 ± 0.04
	$_{\rm pc,2}$	1.89 ± 0.02	$\textbf{6.43}\pm0.05$	13.03 ± 0.04
	pc,3	2.05 ± 0.02	$\textbf{4.96} \pm 0.05$	13.08 ± 0.04
	\mathbf{s}	1.54 ± 0.02	$\textbf{7.33} \pm 0.06$	9.41 ± 0.04
2	с	2.88 ± 0.01	$\textbf{4.51}\pm0.03$	5.27 ± 0.03
	$_{\rm pc,1}$	1.31 ± 0.02	11.67 ± 0.09	19.36 ± 0.07
	$_{\rm pc,2}$	2.88 ± 0.02	$\textbf{6.58} \pm 0.06$	19.82 ± 0.07
	pc,3	3.09 ± 0.03	$\textbf{6.66} \pm 0.06$	20.12 ± 0.07
	\mathbf{S}	3.01 ± 0.01	$\textbf{5.21}\pm0.04$	9.69 ± 0.40
3	с	3.18 ± 0.01	4.41 ± 0.02	$\textbf{4.39}\pm0.01$
	$_{\rm pc,1}$	3.00 ± 0.01	$\textbf{4.91} \pm 0.02$	6.50 ± 0.02
	$_{\rm pc,2}$	4.06 ± 0.02	$\textbf{5.99} \pm 0.02$	7.86 ± 0.03
	pc,3	4.41 ± 0.02	$\textbf{6.32}\pm0.02$	8.47 ± 0.03
	\mathbf{s}	4.02 ± 0.01	5.97 ± 0.03	7.59 ± 0.03

Table 3: Performance $C_{\text{or}}(\llbracket \text{ERM}, \mathfrak{P} \rrbracket)$ for $\mathfrak{P} = \text{Id}$ (that is, choosing the dimension $D^* := \arg\min_{D \in \mathcal{D}_n} \left\{ \left\| s - \widehat{s}_{\widehat{m}_{\text{ERM}}(D)} \right\|_n^2 \right\}$), $\mathfrak{P} = \text{VF}_V$ with V = 5 or $\mathfrak{P} = \text{BM}$. 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.

s.	σ .	$\llbracket \text{ERM}, \text{VF}_5 \rrbracket$	$\llbracket \text{Loo}, \text{VF}_5 \rrbracket$	$\llbracket \mathrm{Lpo}_{20}, \mathrm{VF}_5 \rrbracket$	$\llbracket Lpo_{50}, VF_5 \rrbracket$	$\llbracket \text{ERM}, \text{BM} \rrbracket$
1	с	5.40 ± 0.05	5.03 ± 0.05	5.10 ± 0.05	5.24 ± 0.05	$\textbf{3.91}\pm0.03$
	$_{\rm pc,1}$	11.96 ± 0.03	10.25 ± 0.03	10.28 ± 0.03	10.66 ± 0.04	12.85 ± 0.04
	$_{\rm pc,2}$	6.43 ± 0.05	5.83 ± 0.05	5.99 ± 0.05	6.20 ± 0.05	13.03 ± 0.04
	pc,3	4.96 ± 0.05	$\textbf{4.82}\pm0.04$	$\textbf{4.79} \pm 0.05$	5.02 ± 0.05	13.08 ± 0.04
	\mathbf{s}	7.33 ± 0.06	6.82 ± 0.05	6.99 ± 0.06	$\textbf{6.91}\pm0.06$	9.41 ± 0.04
2	с	$\textbf{4.51} \pm 0.03$	$\textbf{4.55}\pm0.03$	$\textbf{4.50}\pm0.03$	4.73 ± 0.03	5.27 ± 0.03
	$_{\rm pc,1}$	11.67 ± 0.09	10.26 ± 0.08	10.29 ± 0.08	10.45 ± 0.09	19.36 ± 0.07
	$_{\rm pc,2}$	6.58 ± 0.06	5.85 ± 0.06	5.85 ± 0.06	$\textbf{5.49} \pm 0.06$	19.82 ± 0.07
	pc,3	6.66 ± 0.06	5.81 ± 0.06	5.74 ± 0.06	$\textbf{5.66} \pm 0.06$	20.12 ± 0.06
	\mathbf{s}	$\textbf{5.21}\pm0.04$	5.19 ± 0.03	5.17 ± 0.03	5.51 ± 0.04	9.69 ± 0.04
3	с	4.41 ± 0.02	4.54 ± 0.02	4.62 ± 0.02	4.94 ± 0.02	$\textbf{4.39}\pm0.01$
	$_{\rm pc,1}$	4.91 ± 0.02	$\textbf{4.40} \pm 0.02$	$\textbf{4.44} \pm 0.02$	4.69 ± 0.02	6.50 ± 0.02
	$_{\rm pc,2}$	5.99 ± 0.02	5.34 ± 0.02	5.42 ± 0.02	5.75 ± 0.02	7.86 ± 0.03
	pc,3	6.32 ± 0.02	5.74 ± 0.02	5.81 ± 0.02	6.24 ± 0.02	8.47 ± 0.03
	\mathbf{S}	5.97 ± 0.02	5.72 ± 0.02	5.86 ± 0.02	6.07 ± 0.02	7.59 ± 0.03

Table 4: Performance $C_{\rm or}(\mathfrak{P})$ for several change-point detection procedures \mathfrak{P} . 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.

Framework	А	В	\mathbf{C}
$\llbracket \text{ERM}, \text{BM} \rrbracket$	6.82 ± 0.03	7.21 ± 0.04	13.49 ± 0.07
$\llbracket \text{ERM}, \text{VF}_5 \rrbracket$	4.78 ± 0.03	5.09 ± 0.03	7.17 ± 0.05
$\llbracket \text{Loo}, \text{VF}_5 \rrbracket$	$\textbf{4.65}\pm0.03$	$\textbf{4.88} \pm 0.03$	6.61 ± 0.05
$\llbracket Lpo_{20}, VF_5 \rrbracket$	4.78 ± 0.03	$\textbf{4.91} \pm 0.03$	$\textbf{6.49} \pm 0.05$
$\llbracket Lpo_{50}, VF_5 \rrbracket$	4.97 ± 0.03	5.18 ± 0.04	6.69 ± 0.05

Table 5: Performance $C_{\text{or}}^{(R)}(\mathfrak{P})$ 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} .

Framework	А	В	\mathbf{C}
$\llbracket \text{ERM}, \text{BM} \rrbracket$	9.04 ± 0.12	11.62 ± 0.14	21.21 ± 0.31
$\llbracket \text{ERM}, \text{BM}_{\widehat{\sigma}} \rrbracket$	5.34 ± 0.10	6.24 ± 0.11	11.48 ± 0.22
$\llbracket \text{ERM}, \text{VF}_5 \rrbracket$	5.10 ± 0.11	5.92 ± 0.11	7.31 ± 0.14
$\llbracket \text{Loo}, \text{VF}_5 \rrbracket$	$\textbf{4.90} \pm 0.11$	$\textbf{5.63} \pm 0.11$	$\textbf{6.89} \pm 0.16$
$\llbracket Lpo_{20}, VF_5 \rrbracket$	$\textbf{4.88} \pm 0.10$	$\textbf{5.55} \pm 0.10$	$\textbf{6.82} \pm 0.15$
$[\![\mathrm{Lpo}_{50},\mathrm{VF}_5]\!]$	5.11 ± 0.11	$\textbf{5.49} \pm 0.10$	7.14 ± 0.15

Table 6: Performance $C_{\text{or}}^{(R)}(\mathfrak{P})$ of several model selection procedures \mathfrak{P} in frameworks A, B, C with sample size n = 200. In each framework, N = 1, 000 independent samples have been considered. Next to each value is indicated the corresponding empirical standard deviation divided by \sqrt{N} .

s.	σ .	$2\widehat{K}_{\max,\mathrm{jump}}$	$2\hat{K}_{\text{thresh.}}$	$\widehat{\sigma^2}$	$\sigma_{ m true}^2$
1	с	6.85 ± 0.12	3.91 ± 0.03	1.74 ± 0.02	2.05 ± 0.02
	$_{\rm pc,1}$	70.97 ± 1.18	12.85 ± 0.04	1.13 ± 0.02	10.20 ± 0.05
	pc,2	23.74 ± 0.26	13.03 ± 0.04	3.55 ± 0.04	10.43 ± 0.05
	pc,3	17.56 ± 0.15	13.08 ± 0.04	4.42 ± 0.04	10.43 ± 0.05
	s	20.07 ± 0.31	9.41 ± 0.04	2.18 ± 0.03	1.66 ± 0.02
2	с	6.02 ± 0.03	5.27 ± 0.03	3.58 ± 0.02	3.54 ± 0.02
	$_{\rm pc,1}$	17.83 ± 0.10	19.36 ± 0.07	8.52 ± 0.06	15.62 ± 0.08
	$_{\rm pc,2}$	17.63 ± 0.10	19.82 ± 0.07	10.77 ± 0.07	16.56 ± 0.08
	pc,3	17.76 ± 0.10	20.12 ± 0.07	10.58 ± 0.07	16.64 ± 0.08
	\mathbf{s}	10.17 ± 0.05	9.69 ± 0.04	5.28 ± 0.03	10.95 ± 0.02
3	с	4.97 ± 0.02	4.39 ± 0.01	4.62 ± 0.01	4.21 ± 0.01
	$_{\rm pc,1}$	7.18 ± 0.03	6.50 ± 0.02	4.52 ± 0.02	6.70 ± 0.03
	$_{\rm pc,2}$	8.14 ± 0.03	7.86 ± 0.03	6.22 ± 0.02	7.55 ± 0.03
	pc,3	8.66 ± 0.03	8.47 ± 0.03	6.64 ± 0.02	8.00 ± 0.03
	\mathbf{S}	8.50 ± 0.04	7.59 ± 0.03	5.94 ± 0.02	15.50 ± 0.04
	А	7.52 ± 0.04	6.82 ± 0.03	4.86 ± 0.03	5.55 ± 0.03
	В	7.89 ± 0.04	7.21 ± 0.04	5.18 ± 0.03	5.77 ± 0.03
	\mathbf{C}	12.81 ± 0.08	13.49 ± 0.07	8.93 ± 0.06	12.44 ± 0.07

Table 7: Performance $C_{\rm or}(BM)$ with four different definitions of \hat{C} (see text), in some of the simulation settings considered in the paper. 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} .