Estimating the number of block boundaries from diagonal blockwise matrices without penalization

Vincent Brault, Maud Delattre, Tristan Mary-Huard, Céline Leduc

Scandinavian Journal of Statistics Theory and Applications

Scandinavian Journal of Statistics, Vol. 44: 563–580, 2017
doi: 10.1111/sjos.12266

Estimating the Number of Block Boundaries from Diagonal Blockwise Matrices Without Penalization

VINCENT BRAULT, MAUD DELATTRE, EMILIE LEBARBIER and CÉLINE LÉVY-LEDUC
UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay

TRISTAN MARY-HUARD
UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay

INRA, UMR 0320 / UMR 8120 Génétique Quantitative et Evolution-Le Moulon F-91190 Gif-sur-Yvette, France

ABSTRACT. In computational biology, numerous recent studies have been dedicated to the analysis of the chromatin structure within the cell by two-dimensional segmentation methods. Motivated by this application, we consider the problem of retrieving the diagonal blocks in a matrix of observations. The theoretical properties of the least squares estimators of both the boundaries and the number of blocks are investigated. More precisely, the contribution of the paper is to establish the consistency of these estimators. A surprising consequence of our results is that, contrary to the one-dimensional case, a penalty is not needed for retrieving the true number of diagonal blocks. Finally, the results are illustrated on synthetic data.

Key words: Model selection, Hi-C data, Segmentation

1. Introduction

Detecting change points in one-dimensional signals is a very important task that arises in many applications, ranging from electroencephalography to speech processing and network intrusion detection (Basseville & Nikiforov, 1993; Brodsky & Darkhovsky, 2000; Tartakovsky *et al.*, 2014). The aim of such approaches is to split a signal into several homogeneous segments according to some quantity. A large literature has been dedicated to the change-point detection issue for one-dimensional data. This problem may also have several applications when dealing with two-dimensional data.

One of the main situations in which this problem occurs is the detection of chromosomal regions having close spatial location in the nucleus of a cell. Detecting such regions provides valuable insight to understand the influence of chromosomal conformation on cell functioning.

More precisely, we will consider the problem of identifying the so-called *cis*-interactions between regions of a chromosome. In this context, n locations spatially ordered along a given chromosome are considered, the goal being to find clusters of adjacent locations that strongly interact. The elements $Y_{i,j}$ of a data matrix $\mathcal{Y}$ will then correspond to the interaction level between locations i and j of a chromosome, which can be measured using the recently developed HiC technologies (Dixon *et al.*, 2012). In this application, the signal – and consequently the data matrix – exhibits a strong structure: one should observe high signal levels within blocks of locations along the matrix diagonal and a signal that is close to some (low) baseline level everywhere else.

As shown in Lévy-Leduc *et al.* (2014), the identification of *cis*-interactions can be cast as a segmentation problem, where the goal is to identify diagonal blocks (or regions) with

homogeneous interaction levels. Thanks to the spatial repartition of these regions along the diagonal, the two-dimensional segmentation of the data matrix actually boils down to a particular one-dimensional segmentation. The dynamic programming algorithm originally proposed by Bellman (1961) is well known to provide the exact solution of the one-dimensional segmentation issue in the least squares sense. Therefore, we benefit from the data structure by avoiding both the computational burden and the approximation errors that come with heuristic methods used to solve the complex generic problem of two-dimensional segmentation.

While being able to handle large interaction data matrices from an algorithmic point of view, model selection (i.e. selecting the number of blocks K) remains an open question when dealing with such data. This is contrasted with the problem of one-dimensional signal segmentation, for which the properties of the estimators have been largely addressed, for instance, in Boysen *et al.* (2009), Lavielle and Moulines (2000), and Yao and Au (1989). In these approaches, the number of change points is usually performed thanks to a Schwarz-like penalty $\lambda_n K$ where λ_n is often calibrated on data, as in Lavielle (2005) and Lavielle and Moulines (2000), or a penalty $K(a + b\log(n/K))$ as in Lebarbier (2005) and Massart (2004), where a and b are data-driven as well.

The goal of the present paper is to prove the consistency of the estimators of both the boundaries and the number of blocks obtained by minimizing the (slightly modified) least squares criterion proposed by Lévy-Leduc *et al.* (2014). The proof relies on the strong structure of the data, which is of great help for the model selection issue and for the algorithmic aspects.

More precisely, we will prove that the *non-penalized* least squares estimators of the number of blocks are consistent.

The paper is organized as follows: Section 2 introduces the modelling of the data and the definition of the least squares estimators that will be considered throughout the article. The theoretical properties of the estimators are derived in Section 3 and illustrated on synthetic data in Section 4. A discussion is given in Section 5. The technical aspects of the proofs are detailed in Section 6 and in the supplementary material.

2. Statistical framework

2.1. Modelling

Let us consider $\mathcal{Y} = (Y_{i,j})_{1\leq i,j\leq n}$, a symmetric matrix of random variables. Because of the symmetry, we shall focus on its upper-triangular part denoted by $\mathbf{Y} = (Y_{i,j})_{1\leq i\leq j\leq n}$ where the $Y_{i,j}$ will be assumed to be independent and such that

$$Y_{i,j} = \mathbb{E}\left[Y_{i,j}\right] + \varepsilon_{i,j} = \mu_{i,j} + \varepsilon_{i,j},\ 1 \leq i \leq j \leq n. \tag{1}$$

The $\varepsilon_{i,j}$ satisfies the following assumption:

A 1. The $\varepsilon_{i,j}$ is assumed to be centred, i.i.d. and such that there exists a positive constant β such that for all $\nu \in \mathbb{R}$,

$$\mathbb{E}\left[e^{\nu\varepsilon_{11}}\right] \leq e^{\beta\nu^2}.$$

We shall moreover assume that the matrix of means $(\mu_{i,j})_{1\leq i\leq j\leq n}$ is block diagonal. More precisely, let $\boldsymbol{\tau}^\star = (\tau_0^\star, \tau_1^\star, \ldots, \tau_{K^\star}^\star)$ be a vector of break fractions such that $0 = \tau_0^\star < \tau_1^\star < \cdots < \tau_{K^\star}^\star = 1$. In what follows, the break fractions are fixed quantities: neither their number nor their positions change when n grows. The parameters $\mu_{i,j}$ are such that

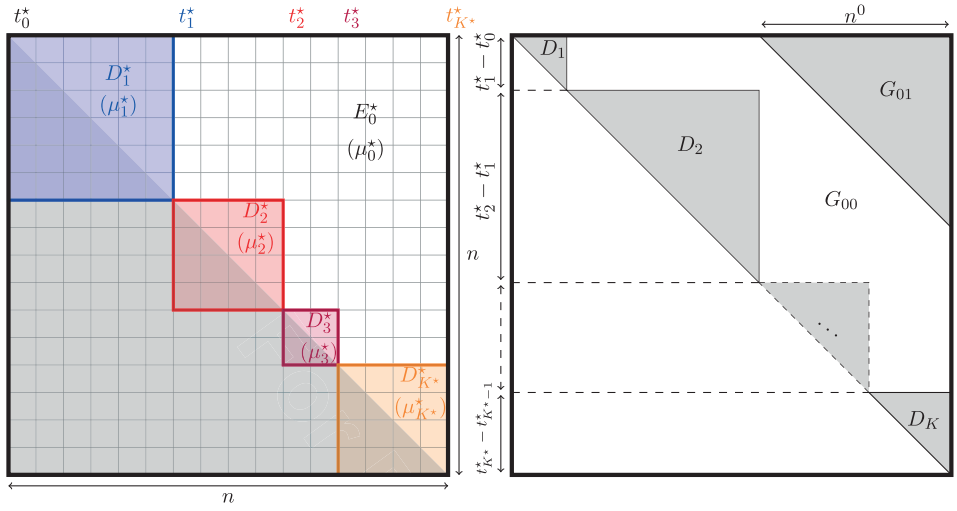

Fig. 1. Left: Example of a matrix $(\mu_{i,j})$ with $n = 16$ and $K^\star = 4$. Right: Illustration of the notations used in the estimation criterion. [Colour figure can be viewed at wileyonlinelibrary.com]

$$\begin{aligned} \mu_{i,j} &= \mu_k^\star \quad \text{if } (i,j) \in D_k^\star,\ k = 1, \ldots, K^\star, \\ &= \mu_0^\star \quad \text{if } (i,j) \in E_0^\star, \end{aligned} \tag{2}$$

where the (half) diagonal blocks $D_k^\star$ $(k = 1, \ldots, K^\star)$ are defined as follows:

$$D_k^\star = \{(i,j) : t_{k-1}^\star \le i \le j \le t_k^\star - 1\}, \tag{3}$$

where $t_k^\star = [n\tau_k^\star] + 1$ are thus such that $1 = t_0^\star < t_1^\star < \cdots < t_{K^\star}^\star = n + 1$, $[x]$ denoting the integer part of x. They stand for the true block boundaries, and $K^\star$ corresponds to the true number of blocks. In (2), $E_0^\star$ corresponds to the set of positions lying outside the diagonal blocks:

$$E_0^\star = \{(i,j) : 1 \le i \le j \le n\} \cap \left(\cup D_k^\star\right)^C, \tag{4}$$

where A^C denotes the complement of set A. An example of such a matrix is displayed in Fig. 1 (left).

The following will also be assumed for the true block sizes:

A2. For all ℓ, one has

$$0 < \Delta_{\boldsymbol{\tau}}^\star = \min_{k \in \{1, \ldots, K^\star\}} \left|\tau_k^\star - \tau_{k-1}^\star\right| \le |\tau_{\ell+1}^\star - \tau_\ell^\star| \le c,$$

where $c \in (0,1)$ is a known constant.

Moreover, the $\mu_k^\star$ satisfies the following assumption:

A3. $$\underline{\lambda}^{(0)} = \min_{1 \le k \le K^\star} \left|\mu_k^\star - \mu_0^\star\right| > 0.$$

2.2. Inference

In this framework, the inference consists in estimating both the number of blocks and the true break fraction vector $\boldsymbol{\tau}^\star$ (or equivalently the true boundary vector $\mathbf{t}^\star$). One strategy would be

to use the following least squares criterion:

$$\widehat{\mathbf{t}}_K^{\mathrm{LS}} \in \underset{\mathbf{t}\in\mathcal{A}_{n,K}^{\Delta_n}}{\mathrm{Argmin}} \left\{ \left[\sum_{k=1}^{K} \sum_{(i,j)\in D_k} \left(Y_{i,j} - \overline{Y}_{D_k}\right)^2 \right] + \sum_{(i,j)\in E_0} \left(Y_{i,j} - \overline{Y}_{E_0}\right)^2 \right\}, \tag{5}$$

where $\overline{Y}_{\mathcal{D}}$ is the empirical mean of the $Y_{i,j}$ when the indices (i,j) belong to $\mathcal{D}$, D_k and E_0 are defined as in (3) and (4) except that $\mathbf{t}^\star$ is replaced by $\mathbf{t}$ and K is the considered number of segments – $K^\star$ being unknown in practice. Moreover,

$$\begin{aligned} \mathcal{A}_{n,K}^{\Delta_n} = \{\mathbf{t} = (t_0, \ldots, t_K) : t_0 = 1 < t_1 < \cdots < t_K = n+1 \\ \text{and } \forall 1 \le k \le K,\ n\Delta_n \le t_k - t_{k-1} < cn\} \end{aligned} \tag{6}$$

is the set of admissible segmentations, where Δ_n denotes a positive sequence.

However, thanks to (A2), one can derive an unbiased estimator of $\mu_0^\star$ using the upper-right triangle part of the matrix $\mathcal{Y}$ denoted G_{01} and defined by

$$G_{01} = \{(i,j) : 1 \le i \le n^0, (n - n^0 + 1) \le j \le n\} \quad \text{with } n^0 = [(1-c)n]. \tag{7}$$

Indeed, the intersection between the blocks D_k and G_{01} will always be empty. Thus, we can split $E_0^\star$ into two disjoint sets $G_{00}^\star$ and G_{01} (see the right part of Fig. 1) as follows:

$$E_0^\star = G_{00}^\star \cup G_{01}. \tag{8}$$

Consequently, we will consider the following slightly modified least squares criterion:

$$\widehat{\mathbf{t}}_K \in \mathrm{Argmin}_{\mathbf{t}\in\mathcal{A}_{n,K}^{\Delta_n}} \mathcal{Q}_n^K(\mathbf{t}), \tag{9}$$

where

$$\mathcal{Q}_n^K(\mathbf{t}) = \left\{ \left[\sum_{k=1}^{K} \sum_{(i,j)\in D_k} \left(Y_{i,j} - \overline{Y}_{D_k}\right)^2 \right] + \sum_{(i,j)\in E_0} \left(Y_{i,j} - \overline{Y}_{G_{01}}\right)^2 \right\}. \tag{10}$$

Lastly, we will consider the following estimator of $K^\star$:

$$\widehat{K} = \mathrm{Argmin}_{1\le K\le K_{\max}} \mathcal{Q}_n^K\left(\widehat{\mathbf{t}}_K\right), \tag{11}$$

where $\widehat{\mathbf{t}}_K$ is defined in (9) and $K_{\max}$ is the maximal number of blocks considered.

Criterion (11) based on (10) has been proposed by Lévy-Leduc *et al.* (2014). The goal of our paper is to validate this latter approach theoretically. Note that the main difference between (5) and (10) is the estimation of $\mu_0^\star$ that is independent from the segmentation, because G_{01} is fixed. Hence, $\mu_0^\star$ can be estimated prior to the optimization of the criterion (10). As a consequence, this optimization can be performed by using the dynamic programming algorithm as explained in Lévy-Leduc *et al.* (2014).

3. Theoretical results

The goal of this section is to derive the consistency of $\widehat{K}$ and $\widehat{\boldsymbol{\tau}}$. To prove these results, we shall need the following assumption on Δ_n:

A4. $$\Delta_n \frac{\sqrt{n}}{(\log n)^{1/4}} \underset{n\to+\infty}{\longrightarrow} +\infty \quad \text{and} \quad \Delta_n \le \Delta_{\boldsymbol{\tau}}^\star, \text{ for large enough } n.$$

Theorem 1. Let $Y_{i,j}$ be defined by (1). Assume that (A1), (A2), (A3) and (A4) hold. Then $\widehat{K}$ defined in (11) is such that

$$\mathbb{P}\left(\widehat{K} \neq K^\star\right) \longrightarrow 0, \text{ as } n \to +\infty. \tag{12}$$

Remark 3.1. Observe that, contrary to classical statistical frameworks, $\widehat{K}$ is a consistent estimator of $K^\star$ even if it is obtained without any penalization.

Remark 3.2. In theorem 1, the estimator $\widehat{K}$ is defined as the minimizer of $Q_n^K(\widehat{\mathbf{t}}_K)$ where $\widehat{\mathbf{t}}_K$ is obtained by minimizing $Q_n^K(\mathbf{t})$ over the set $\mathcal{A}_{n,K}^{\Delta_n}$. If we are only interested in proving that $\mathbb{P}(\widehat{K} < K^\star) \to 0$, the minimization can be performed on the set $\mathcal{A}_{n,K}^{1/n}$ instead of $\mathcal{A}_{n,K}^{\Delta_n}$, that is, without any constraint on the minimal distance between two consecutive change points (see lemma 1 (i) later and lemmas 2, 3 and 4, which are given in Section 6).

Remark 3.3. Theorem 1 is valid under (A2) that implies that the number of observations within each segment increases linearly with n, because $t_k^\star = [n\tau_k^\star] + 1$. This assumption could be alleviated by assuming that $\Delta_{\boldsymbol{\tau}}^\star$ is no longer a constant. In that case, we shall need to assume that $\Delta_{\boldsymbol{\tau}}^\star n^{1/4}/(\log n)^{1/8}$ tends to infinity, as n tends to infinity.

Remark 3.4. The assumption $\Delta_n \gg (\log n)^{1/4}/\sqrt{n}$ of (A4) can be understood in the light of lemma 1 (ii) and (17) at the end of the proof of theorem 1. It is required to ensure the convergence to zero of the exponential inequalities of the random parts given in lemmas 2, 3 and 4.

This assumption is only required for proving that $\mathbb{P}\left(\widehat{K} > K^\star\right)$ tends to zero as n tends to infinity. As a consequence, when the number of blocks is known ($\widehat{K} = K^\star$), the break fractions consistency is obtained in our paper when $\Delta_n = 1/n$. Such a choice is impossible in the one-dimensional segmentation framework of Lavielle and Moulines (2000) because it is required that $n\Delta_n \to +\infty$ and $\Delta_n \to 0$, as n tends to infinity, in order to obtain the break fractions consistency when the number of breaks is known.

Remark 3.5. In practice, c has to be chosen in order to use the top right part of the matrix of observations to estimate the parameter $\mu_0^\star$. This choice can come either from a prior biological knowledge or from a simple visualization of the data. In the case of the analysis of HiC data, the size of the interaction diagonal blocks is expected to be small compared with the size of the chromosome, that is, the size of the data matrix. In this context, $c = 3/4$ can be safely chosen, as suggested in Lévy-Leduc *et al.* (2014). If the value of c is misspecified, the estimator of $\mu_0^\star$ is biased. The consistency result of theorem 3 still holds if (A3) is replaced by $\min_{1 \leq k \leq K^\star} \left|\mu_k^\star - \mathbb{E}(\bar{Y}_{G_{01}})\right| > 0$.

Sketch of proof of Theorem 1. In order to prove (12), we shall prove that $\mathbb{P}\left(\widehat{K} < K^\star\right)$ and $\mathbb{P}\left(\widehat{K} > K^\star\right)$ tend to zero as n tends to infinity. Note that

$$\mathbb{P}\left(\widehat{K} < K^\star\right) \leq \sum_{K=1}^{K^\star-1} \mathbb{P}\left(\widehat{K} = K\right) \text{ and } \mathbb{P}\left(\widehat{K} > K^\star\right) \leq \sum_{K=K^\star+1}^{K_{\max}} \mathbb{P}\left(\widehat{K} = K\right).$$

Hence, we shall prove that for $K < K^\star$ and $K > K^\star$,

$$\mathbb{P}\left(\widehat{K} = K\right) \longrightarrow 0, \text{ as } n \to +\infty.$$

Observe that by definition of $\widehat{K}$ given in (11),

$$\mathbb{P}\left(\widehat{K}=K\right) \leq \mathbb{P}\left(\min_{\mathbf{t}\in\mathcal{A}_{n,K}^{\Delta_n}} Q_n^K(\mathbf{t}) - \min_{\mathbf{t}\in\mathcal{A}_{n,K^\star}^{\Delta_n}} Q_n^{K^\star}(\mathbf{t}) \leq 0\right)$$
$$\leq \mathbb{P}\left(\min_{\mathbf{t}\in\mathcal{A}_{n,K}^{\Delta_n}} Q_n^K(\mathbf{t}) - Q_n^{K^\star}(\mathbf{t}^\star) \leq 0\right),$$

because, for large enough n, $\Delta_n \leq \Delta_{\boldsymbol{\tau}}^\star$ and hence, $\mathbf{t}^\star$ belongs to $\mathcal{A}_{n,K^\star}^{\Delta_n}$. Thus, we shall focus on

$$\mathbb{P}\left(\min_{\mathbf{t}\in\mathcal{A}_{n,K}^{\Delta_n}} J_n(\mathbf{t}) \leq 0\right),$$

where

$$J_n(\mathbf{t}) = \frac{2}{n(n+1)}\left(Q_n^K(\mathbf{t}) - Q_n^{K^\star}(\mathbf{t}^\star)\right). \tag{13}$$

We shall prove in the supplementary material that

$$J_n(\mathbf{t}) = \mathrm{B}_n(\mathbf{t}) + V_n(\mathbf{t}) + W_n(\mathbf{t}) + Z_n(\mathbf{t}), \tag{14}$$

where B_n, V_n, W_n and Z_n are defined by (20), (21), (22), (23) and (24) in Section 6. In (14), B_n corresponds to the deterministic part, and the other terms correspond to the random part of J_n.

The remainder of the proof is based on lemma 1, which is proved in Section 6.2 and which provides a lower bound for the deterministic part of J_n, and on lemmas 2, 3 and 4, given in Section 6, which provide deviation inequalities for the random terms of J_n.

Lemma 1. *Let $B_n(\mathbf{t})$ be defined by (20) and (21), then*

(i) if $K < K^\star$,

$$\min_{\mathbf{t}\in\mathcal{A}_{n,K}^{1/n}} B_n(\mathbf{t}) \geq \frac{\underline{\lambda}^{(0)2}}{64}\left(\Delta_{\boldsymbol{\tau}}^\star\right)^4,$$

(ii) if $K > K^\star$,

$$\min_{\mathbf{t}\in\mathcal{A}_{n,K}^{\Delta_n}} B_n(\mathbf{t}) \geq \frac{\underline{\lambda}^{(0)2}}{4}\Delta_n^2,$$

(iii) if $K = K^\star$, for all positive δ,

$$\min_{\left\{\mathbf{t}\in\mathcal{A}_{n,K}^{1/n}, \|\mathbf{t}-\mathbf{t}^\star\|_\infty > n\delta\right\}} B_n(\mathbf{t}) \geq \frac{\underline{\lambda}^{(0)2}}{32}\min\left(\Delta_{\boldsymbol{\tau}}^\star/2, \delta\right)\left(\Delta_{\boldsymbol{\tau}}^\star\right)^3, \tag{15}$$

where $\Delta_{\boldsymbol{\tau}}^\star$ is defined in (A2), $\underline{\lambda}^{(0)}$ is defined in (A3) and $\mathcal{A}_{n,K}^{\Delta_n}$ is defined in (6). $\mathcal{A}_{n,K}^{1/n}$ is a particular case with $\Delta_n = 1/n$ and

$$\left\|\mathbf{t}-\mathbf{t}^\star\right\|_\infty = \max_{0\leq k\leq K^\star}\left|t_k - t_k^\star\right|. \tag{16}$$

Thus,

$$\mathbb{P}\left(\min_{\mathbf{t}\in\mathcal{A}_{n,K}^{\Delta_n}} J_n(\mathbf{t}) \le 0\right) \le \mathbb{P}\left(\min_{\mathbf{t}\in\mathcal{A}_{n,K}^{\Delta_n}} [\mathrm{B}_n(\mathbf{t}) + V_n(\mathbf{t}) + W_n(\mathbf{t}) + Z_n(\mathbf{t})] \le 0\right).$$

The right-hand side of the previous inequality is bounded by

$$\mathbb{P}\left(-\min_{\mathbf{t}\in\mathcal{A}_{n,K}^{\Delta_n}} V_n(\mathbf{t}) - \min_{\mathbf{t}\in\mathcal{A}_{n,K}^{\Delta_n}} W_n(\mathbf{t}) - \min_{\mathbf{t}\in\mathcal{A}_{n,K}^{\Delta_n}} Z_n(\mathbf{t}) \ge \min_{\mathbf{t}\in\mathcal{A}_{n,K}^{\Delta_n}} \mathrm{B}_n(\mathbf{t})\right).$$

For bounding this term, we shall use lemma 1 (ii).

For $K > K^\star$, we obtain

$$\begin{aligned}\mathbb{P}\left(\min_{\mathbf{t}\in\mathcal{A}_{n,K}^{\Delta_n}} J_n(\mathbf{t}) \le 0\right) &\le \mathbb{P}\left(-\min_{\mathbf{t}\in\mathcal{A}_{n,K}^{\Delta_n}} V_n(\mathbf{t}) \ge \frac{\underline{\lambda}^{(0)2}}{12}\Delta_n^2\right)\\ &+ \mathbb{P}\left(-\min_{\mathbf{t}\in\mathcal{A}_{n,K}^{\Delta_n}} W_n(\mathbf{t}) \ge \frac{\underline{\lambda}^{(0)2}}{12}\Delta_n^2\right) + \mathbb{P}\left(-\min_{\mathbf{t}\in\mathcal{A}_{n,K}^{\Delta_n}} Z_n(\mathbf{t}) \ge \frac{\underline{\lambda}^{(0)2}}{12}\Delta_n^2\right).\end{aligned} \tag{17}$$

By lemmas 2, 3 and 4, we conclude that

$$\mathbb{P}\left(\widehat{K} = K\right) \underset{n\to+\infty}{\longrightarrow} 0,$$

for $K > K^\star$. The case $K < K^\star$ can be proved by following the same lines. □

Remark 3.6. We can observe from theorem 1 that adding a penalty term is not necessary for obtaining a consistent estimator of the number of diagonal blocks. This may be surprising because, in the one-dimensional case, it is proved in theorem 9 of Lavielle and Moulines (2000) that a penalty term is required. More precisely, the main difference between our two-dimensional framework and the one-dimensional case is the behaviour of the deterministic part of our criterion B_n: it is lower bounded whatever the value of K ($K \ge K^\star$ or $K < K^\star$), as proved in lemma 1. The key point of the proof is that when $K > K^\star$, some entries of the diagonal blocks will have their mean estimated by $\overline{Y}_{G_{01}}$ rather than their proper associated mean that is the empirical mean of the observations lying in the corresponding diagonal block [Fig. 6 (right panel)].

On the contrary, in the one-dimensional case, a penalty term of the type $\beta_n K$ is necessary to obtain such a lower bound when $K \ge K^\star$. In the case where $K < K^\star$, a lower bound for B_n is obtained without penalization. For further details, see the proof of theorem 9 in Lavielle and Moulines (2000).

Theorem 2. Assume that the assumptions of theorem 1 hold, and then, for all $\delta > 0$,

$$\mathbb{P}\left(\left\|\mathbf{t}^\star - \widehat{\mathbf{t}}_{\widehat{K}}\right\|_{\mathcal{H}} > n\delta\right) \underset{n\to+\infty}{\longrightarrow} 0, \tag{18}$$

where $\widehat{\mathbf{t}}_{\widehat{K}}$ is defined in (9) and (11) and $\|\cdot\|_{\mathcal{H}}$ denotes the Hausdorff distance defined by

$$\left\|\mathbf{t}^\star - \widehat{\mathbf{t}}_K\right\|_{\mathcal{H}} = \max\left[\max_{0\le k\le K^\star}\min_{0\le \ell\le K}\left|t_k^\star - \widehat{t}_\ell\right|, \max_{0\le \ell\le K}\min_{0\le k\le K^\star}\left|t_k^\star - \widehat{t}_\ell\right|\right].$$

Observe that (18) can be rewritten as $\mathbb{P}\left(\left\|\boldsymbol{\tau}^\star - \widehat{\boldsymbol{\tau}}_{\widehat{K}}\right\|_{\mathcal{H}} > \delta\right) \underset{n\to+\infty}{\longrightarrow} 0$, where $\widehat{\boldsymbol{\tau}}_{\widehat{K}} = \widehat{\mathbf{t}}_{\widehat{K}}/n$.

Sketch of proof of Theorem 2. Observe that

$$\mathbb{P}\left(\left\|\mathbf{t}^\star - \widehat{\mathbf{t}}_{\widehat{K}}\right\|_{\mathcal{H}} > n\delta\right) = \mathbb{P}\left(\left\{\left\|\mathbf{t}^\star - \widehat{\mathbf{t}}_{\widehat{K}}\right\|_{\mathcal{H}} > n\delta\right\} \cap \left\{\widehat{K} \neq K^\star\right\}\right)$$
$$+ \mathbb{P}\left(\left\{\left\|\mathbf{t}^\star - \widehat{\mathbf{t}}_{\widehat{K}}\right\|_{\mathcal{H}} > n\delta\right\} \cap \left\{\widehat{K} = K^\star\right\}\right) \leq \mathbb{P}\left(\widehat{K} \neq K^\star\right) + \mathbb{P}\left(\|\widehat{\mathbf{t}}_{K^\star} - \mathbf{t}^\star\|_\infty > n\delta\right),$$

where $\|\widehat{\mathbf{t}}_{K^\star} - \mathbf{t}^\star\|_\infty$ is defined in (16) because $\|\widehat{\mathbf{t}}_{\widehat{K}} - \mathbf{t}^\star\|_\infty = \|\widehat{\mathbf{t}}_{\widehat{K}} - \mathbf{t}^\star\|_{\mathcal{H}}$ when $\widehat{K} = K^\star$. By theorem 1, proving (18) amounts to proving that

$$\mathbb{P}\left(\max_{0 \leq k \leq K^\star} \left|t_k^\star - \widehat{t}_k\right| > n\delta\right) \to 0, \text{ as } n \to +\infty.$$

Observe that

$$\mathbb{P}\left(\max_{1 \leq k \leq K^\star} \left|t_k^\star - \widehat{t}_k\right| > n\delta\right) \leq \mathbb{P}\left(\min_{\left\{\mathbf{t} \in \mathcal{A}_{n,K^\star}^{1/n}, \|\mathbf{t}-\mathbf{t}^\star\|_\infty > n\delta\right\}} J_n(\mathbf{t}) \leq 0\right).$$

Using the same arguments as those used in the proof of theorem 1, the proof follows from the decomposition of J_n given by (14), the lower bound (15) of lemma 1 and the deviation inequalities for the random terms given by lemmas 2, 3 and 4. □

4. Numerical experiments

The goal of this section is to illustrate the theoretical results obtained in Section 3. For an application of our method to real data, we refer the reader to Lévy-Leduc *et al.* (2014).

4.1. Simulation framework

We generated Gaussian diagonal block matrices according to model (1) with $\mu_k^\star = 1$ for the $K^\star = 5$ diagonal blocks and $\mu_0^\star = 0$ for different values of n ($n \in \{500, 1500\}$). The change-point locations are $(\tau_0^\star, \ldots, \tau_5^\star) = (0, 0.07, 0.2, 0.4, 0.67, 1)$; hence, $\Delta_{\boldsymbol{\tau}}^\star = 0.07$. We shall use different values for the standard deviation σ of the $\varepsilon_{i,j}$: $\sigma \in \{1, \ldots, 10\}$. For each case, 500 matrices were simulated, and the procedure was tested. Examples of such matrices are displayed in Fig. 2 for different values of σ.

The results that are presented next have been obtained by using the R package HiCseg that is available on the CRAN. In this package, the values of Δ_n and c are fixed and equal to $2/n$ and $3/4$, respectively.

4.2. Statistical performance

Performance of the statistical procedure. We first consider the problem of estimating the true number of blocks $K^\star$, and provide some insight about the consistency of our procedure without penalty, outlined in remark 3.1. Boxplots of the estimated number of change points are displayed in Fig. 3 for n in $\{500, 1500\}$ and for different values of σ.

On the one hand, we observe that for high signal-to-noise ratios, the true value of $K^\star$ is retrieved by our procedure. On the other hand, when the signal-to-noise ratio becomes very low, $K^\star$ is not properly estimated. In this situation, $K^\star$ is overestimated, which is in accordance with what occurs in the one-dimensional case where a non-penalized procedure would result in a systematic overestimation of $K^\star$. However, when n increases, the value of σ from which this overestimation occurs is unsurprisingly larger. One way to improve the estimation of $K^\star$ would perhaps be to use a penalized procedure. This will be the subject of a future work (see Section 5 for further details).

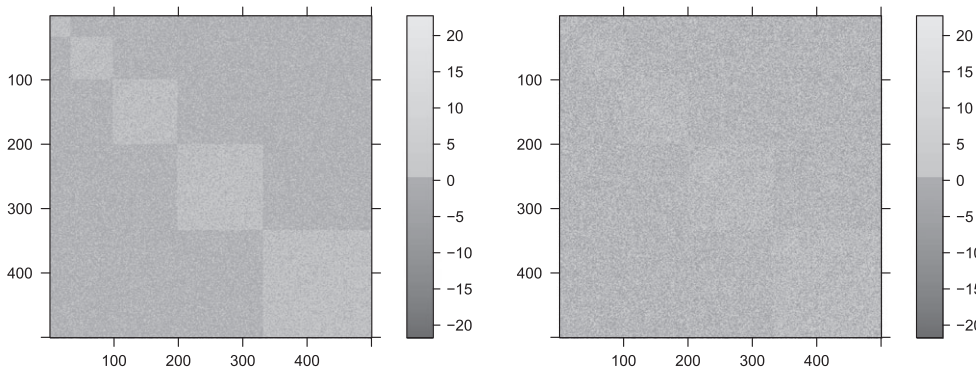

Fig. 2. Examples of simulated matrices following model (1) with $(\tau_0^\star, \ldots, \tau_5^\star) = (0, 0.07, 0.2, 0.4, 0.67, 1)$ and $n = 500$ for two values of σ: $\sigma = 1$ (left) and $\sigma = 4$ (right).

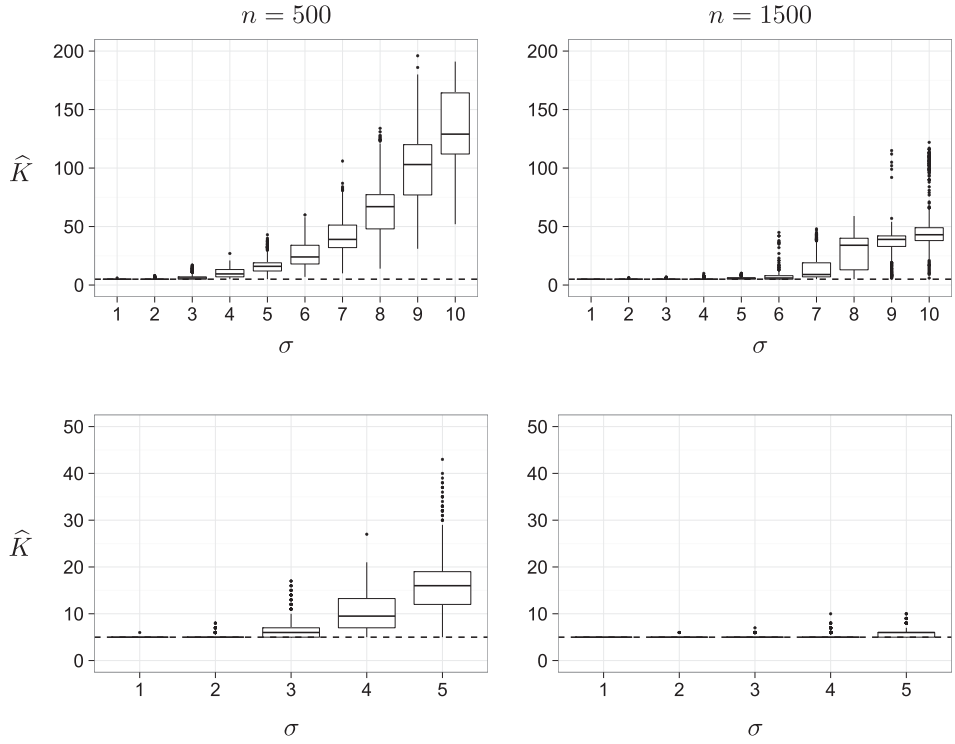

Fig. 3. Top: Boxplots of the estimations of $K^\star = 5$ as a function of the standard deviation σ for $n = 500$ (left) and $n = 1500$ (right). The dashed line corresponds to the true value of $K^\star$. Bottom: Same plots with the x-axis values restricted to $\{1, \ldots, 5\}$.

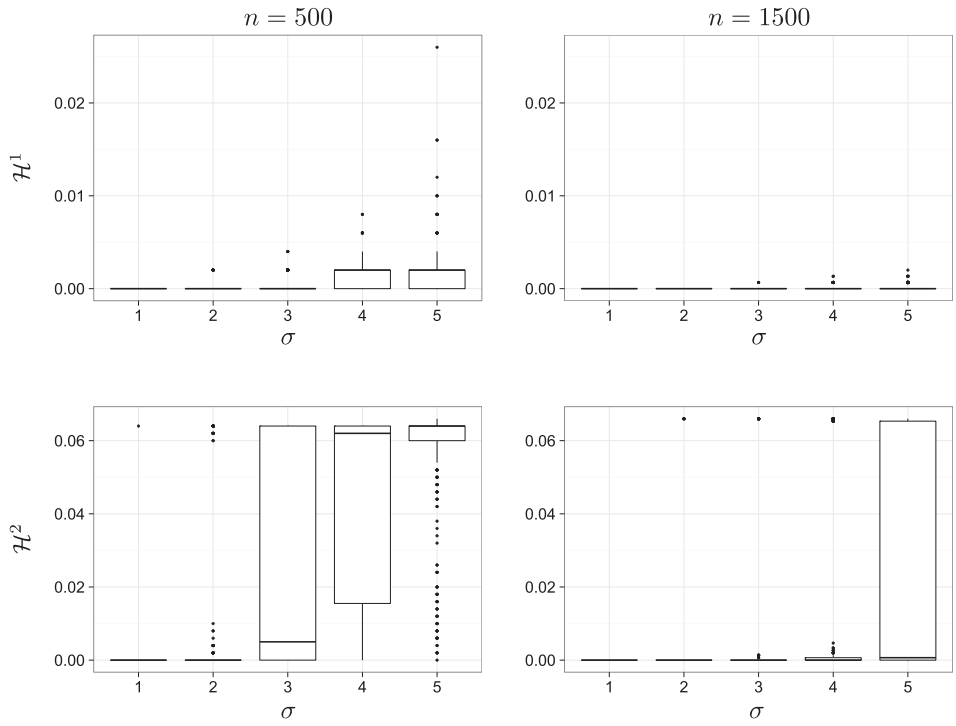

Fig. 4. Boxplots of the two parts of the Hausdorff distance: $\mathcal{H}^1$ (top) and $\mathcal{H}^2$ (bottom) for $\boldsymbol{n = 500}$ (left) and $\boldsymbol{n = 1500}$ (right). For each case, the boxplots are displayed as a function of σ.

To illustrate the performance of our procedure in terms of the estimation of change-point location, Fig. 4 displays the boxplots of the two parts of the Hausdorff distance defined by

$$\begin{aligned} \left\| \mathbf{t}^\star - \widehat{\mathbf{t}}_{\widehat{K}} \right\|_{\mathcal{H}^1} &= \max_{0 \le k \le K^\star} \min_{0 \le \ell \le \widehat{K}} \left| t_k^\star - \widehat{t}_\ell \right|, \\ \left\| \mathbf{t}^\star - \widehat{\mathbf{t}}_{\widehat{K}} \right\|_{\mathcal{H}^2} &= \max_{0 \le \ell \le \widehat{K}} \min_{0 \le k \le K^\star} \left| t_k^\star - \widehat{t}_\ell \right|. \end{aligned} \tag{19}$$

We observe from this figure that when $K^\star$ is overestimated, the true change points are nevertheless recovered well ($\|\cdot\|_{\mathcal{H}^1}$ is close to 0), the other estimated change points being spurious ones ($\|\cdot\|_{\mathcal{H}^2}$ is large). As proved in theorem 2, this phenomenon is less visible when n becomes large.

Effect of a poor estimation of $\mu_0^\star$. We study the behaviour of our segmentation procedure when $\mu_0^\star$ is poorly estimated that may occur, for instance, when the constant c appearing in (7) is too small. To this end, we generated data in which the mean of the $n_0 \times n_0$ top right part of the observation matrix is modified, where n_0 is defined in (2). More precisely, the mean of this part is equal to $\mu_0^\star + \omega$, where $\omega \in \{0.2, 0.4, 0.6, 0.8\}$. The results are displayed in Fig. 5. We can see from this figure that when the value of $\mu_0^\star + \omega$ is close to the values of the means of the diagonal blocks, our procedure tends to overestimate $K^\star$. This phenomenon is less visible when n is large (Fig. 6).

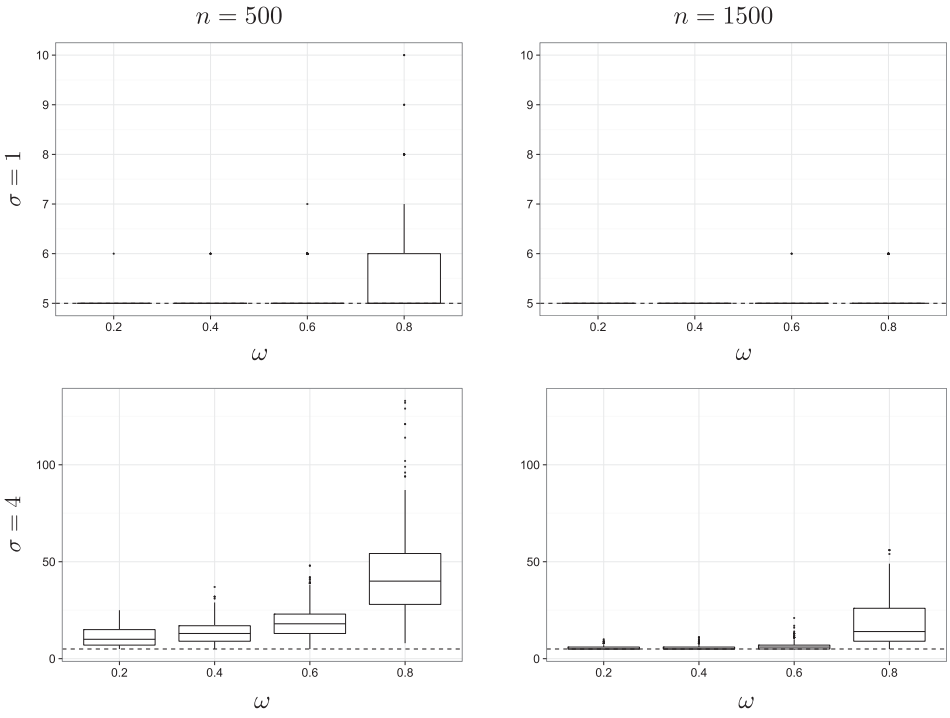

Fig. 5. Boxplots of $\widehat{K}$ for $\sigma = 1$ (top) and $\sigma = 4$ (bottom) for $n = 500$ (left) and $n = 1500$ (right). For each case, the boxplots are displayed as a function of ω. The dashed line corresponds to the true value of $K^\star$.

5. Discussion

In this paper, we established that the (slightly modified) least squares estimators for the number of blocks and their boundaries in a block diagonal matrix are consistent. Note that the obtained results are non-standard in the sense that we proved that penalizing the least squares criterion is not required to obtain a consistent estimator of the number of diagonal blocks. This has to be contrasted with the one-dimensional case, where it is well known that a penalization is required to ensure consistency, see, for instance, Lavielle and Moulines (2000). More precisely, a close look at the proof of theorem 9 in Lavielle and Moulines (2000) shows that a penalty is required to discard models such that $K > K^\star$. This comes from the fact that in the one-dimensional setting when $K > K^\star$, the deterministic part B_n of J_n vanishes for all segmentations $\mathbf{t}$ satisfying $\|\mathbf{t}^\star - \mathbf{t}\|_{\mathcal{H}^1} = 0$ (i.e. for all segmentations $\mathbf{t}$ nested in the true segmentation $\mathbf{t}^\star$). This bias term being null, a penalty term has to be added to the criterion to compensate the stochastic deviations of the random terms in J_n. In the two-dimensional setting, the deterministic part B_n does not vanish when $K > K^\star$ – as proved in lemma 1 – ensuring consistency.

The framework that we have chosen for proving our results consists in assuming that the observations are independent and that the size of the observation matrix is large (asymptotic framework), which is adapted to the analysis of HiC experiments. From a practical point of view, the independence assumption is not always satisfied, for instance, when the observation matrix is a correlation or a similarity matrix, see, for example, Dehman *et al.* (2015) and Ioanna Delatola *et al.* (2015). Hence, relaxing the independence assumption to retrieve diagonal block boundaries in such cases would be a natural extension of this paper.

Moreover, it could be interesting to see if adding a penalty term to our criterion would improve the rates of convergence of our estimators or would allow us to alleviate our assumptions. This will be the subject of a future work.

6. Proofs

6.1. Definition of B_n, V_n, W_n and Z_n

We define hereafter B_n, V_n, W_n and Z_n that appear in (14) by

$$\mathrm{B}_n(\mathbf{t}) = \mathrm{B}_n^D(\mathbf{t}) + \mathrm{B}_n^0(\mathbf{t}),\ V_n(\mathbf{t}) = V_n^D(\mathbf{t}) + V_n^0(\mathbf{t}),\ W_n = W_n^D(\mathbf{t}) + W_n^0(\mathbf{t}), \tag{20}$$

and

$$\begin{aligned} \mathrm{B}_n^D(\mathbf{t}) &= \frac{2}{n(n+1)} \left(\sum_{k=1}^{K} \sum_{(i,j)\in D_k} \left(\mathbb{E}\left[Y_{i,j}\right] - \mathbb{E}\left[\overline{Y}_{D_k}\right] \right)^2 \right), \\ \mathrm{B}_n^0(\mathbf{t}) &= \frac{2}{n(n+1)} \sum_{(i,j)\in G_{00}} \left(\mathbb{E}\left[Y_{i,j}\right] - \mathbb{E}\left[\overline{Y}_{G_{01}}\right] \right)^2, \end{aligned} \tag{21}$$

$$\begin{aligned} V_n^D(\mathbf{t}) &= \frac{2}{n(n+1)} \left[\sum_{k=1}^{K^\star} \frac{\left(\sum_{(i,j)\in D_k^\star} \varepsilon_{i,j}\right)^2}{|D_k^\star|} - \sum_{k=1}^{K} \frac{\left(\sum_{(i',j')\in D_k} \varepsilon_{i',j'}\right)^2}{|D_k|} \right], \\ V_n^0(\mathbf{t}) &= \frac{2}{n(n+1)} \frac{1}{|G_{01}|^2} \left(\sum_{(i,j)\in G_{01}} \varepsilon_{i,j} \right)^2 \left(|G_{00}| - |G_{00}^\star| \right), \end{aligned} \tag{22}$$

$$\begin{aligned} W_n^D(\mathbf{t}) &= \frac{4}{n(n+1)} \left[\sum_{k=1}^{K^\star} \left(\sum_{(i,j)\in D_k^\star} \varepsilon_{i,j} \right) \mu_k^\star - \sum_{k=1}^{K} \left[\left(\sum_{(i',j')\in D_k} \varepsilon_{i',j'} \right) \mathbb{E}\left[\overline{Y}_{D_k}\right] \right] \right], \\ W_n^0(\mathbf{t}) &= \frac{4}{n(n+1)} \mu_0^\star \left(\sum_{(i,j)\in G_{00}^\star} \varepsilon_{i,j} - \sum_{(i,j)\in G_{00}} \varepsilon_{i,j} \right), \end{aligned} \tag{23}$$

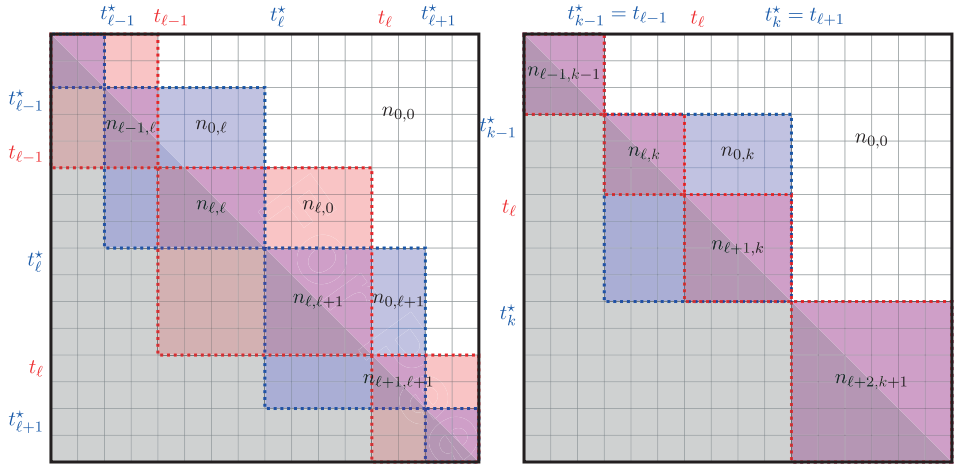

Fig. 6. Left: $K < K^\star$. Right: $K > K^\star$. [Colour figure can be viewed at wileyonlinelibrary.com]

$$Z_n(\mathbf{t}) = \frac{4}{n(n+1)} \frac{1}{|G_{01}|} \left(\sum_{(i,j)\in G_{01}} \varepsilon_{i,j} \right) \left[\left(\sum_{(i,j)\in G_{00}^{\star}} \varepsilon_{i,j} - \sum_{(i,j)\in G_{00}} \varepsilon_{i,j} \right) - \sum_{(i,j)\in G_{00}} \left(\mathbb{E}\left[Y_{i,j}\right] - \mu_0^{\star}\right) \right]. \tag{24}$$

In the equations, $G_{00}^{\star}$ and G_{01} are defined in (8) and (7), and G_{00} has the same definition as $G_{00}^{\star}$ except that $\mathbf{t}^{\star}$ is replaced by $\mathbf{t}$.

6.2. *Proof of Lemma 1*

We shall first rewrite B_n^D and B_n^0 defined by (21). Let us first denote by

$$n_{k,\ell} = \left| D_k \cap D_\ell^{\star} \right|, \tag{25}$$

the number of observations that belong to the intersection of the two blocks D_k and $D_\ell^{\star}$ (with the convention that $D_0 = G_{00}$ and $D_0^{\star} = G_{00}^{\star}$) and

$$n_k = \sum_{\ell=0}^{K^{\star}} n_{k,\ell} \quad \text{and} \quad n_\ell^{\star} = \sum_{k=0}^{K} n_{k,\ell}.$$

Because $\mathbb{E}\left[\overline{Y}_{G_{01}}\right] = \mu_0^{\star}$, $G_{00} \subset \left(\bigcup_{\ell=0}^{K^{\star}} D_\ell{}^{\star}\right)$ and $\mathbb{E}\left[Y_{i,j}\right] = \mu_k^{\star}$, for all $(i,j) \in D_k^{\star}$, we obtain

$$\begin{aligned} \mathrm{B}_n^0(\mathbf{t}) &= \frac{2}{n(n+1)} \sum_{(i,j)\in G_{00}} \left(\mathbb{E}\left[Y_{i,j}\right] - \mu_0^{\star}\right)^2 = \frac{2}{n(n+1)} \sum_{\ell=0}^{K^{\star}} \sum_{(i,j)\in G_{00}\cap D_\ell^{\star}} \left(\mathbb{E}\left[Y_{i,j}\right] - \mu_0^{\star}\right)^2 \\ &= \frac{2}{n(n+1)} \sum_{\ell=0}^{K^{\star}} n_{0,\ell} \left(\mu_\ell^{\star} - \mu_0^{\star}\right)^2. \end{aligned} \tag{26}$$

Because $|D_k| = \sum_{\ell=0}^{K^{\star}} \left|D_k \cap D_\ell^{\star}\right| = \sum_{\ell=0}^{K^{\star}} n_{k,\ell} = n_k$,

$$\mathbb{E}\left[\overline{Y}_{D_k}\right] = \frac{1}{n_k} \sum_{(i,j)\in D_k} \mathbb{E}\left[Y_{i,j}\right] = \frac{1}{n_k} \sum_{\ell=0}^{K^{\star}} \sum_{(i,j)\in D_k\cap D_\ell^{\star}} \mathbb{E}\left[Y_{i,j}\right] = \frac{1}{n_k} \sum_{\ell=0}^{K^{\star}} \mu_\ell^{\star} n_{k,\ell}, \tag{27}$$

where we use for all $k \in \{1,\ldots,K\}$, $D_k \subset \left(\bigcup_{\ell=0}^{K^{\star}} D_\ell{}^{\star}\right)$. Thus,

$$\sum_{(i,j)\in D_k}\left(\mathbb{E}\left[Y_{i,j}\right]-\mathbb{E}\left[\overline{Y}_{D_k}\right]\right)^2=\frac{1}{n_k^2}\sum_{(i,j)\in D_k}\left(n_k\mathbb{E}\left[Y_{i,j}\right]-\sum_{\ell'=0}^{K^\star}\mu^\star_{\ell'}n_{k,\ell'}\right)^2$$

$$=\frac{1}{n_k^2}\sum_{\ell=0}^{K^\star}\sum_{(i,j)\in D_k\cap D^\star_\ell}\left(n_k\mathbb{E}\left[Y_{i,j}\right]-\sum_{\ell'=0}^{K^\star}\mu^\star_{\ell'}n_{k,\ell'}\right)^2=\frac{1}{n_k^2}\sum_{\ell=0}^{K^\star}n_{k,\ell}\left[\sum_{\ell'=0}^{K^\star}n_{k,\ell'}\left(\mu^\star_\ell-\mu^\star_{\ell'}\right)\right]^2$$

$$=\frac{1}{n_k^2}\sum_{\ell=0}^{K^\star}\sum_{\ell_1=0}^{K^\star}\sum_{\ell_2=0}^{K^\star}n_{k,\ell}n_{k,\ell_1}n_{k,\ell_2}\left(\mu^\star_\ell-\mu^\star_{\ell_1}\right)\left(\mu^\star_\ell-\mu^\star_{\ell_2}\right)$$

$$=\frac{1}{n_k^2}\sum_{\ell=0}^{K^\star}\sum_{\ell_1=0}^{K^\star}n_{k,\ell}n_{k,\ell_1}\left(\mu^\star_\ell-\mu^\star_{\ell_1}\right)\sum_{\ell_2=0}^{K^\star}n_{k,\ell_2}\left(\mu^\star_\ell-\mu^\star_{\ell_2}\right)$$

$$=\frac{1}{n_k}\sum_{\ell=0}^{K^\star}\sum_{\ell_1=0}^{K^\star}n_{k,\ell}n_{k,\ell_1}\left({\mu^\star_\ell}^2-\mu^\star_{\ell_1}\mu^\star_\ell\right)-\frac{1}{n_k^2}\sum_{\ell_2=0}^{K^\star}n_{k,\ell_2}\mu^\star_{\ell_2}\underbrace{\sum_{\ell=0}^{K^\star}\sum_{\ell_1=0}^{K^\star}n_{k,\ell}n_{k,\ell_1}\left(\mu^\star_\ell-\mu^\star_{\ell_1}\right)}_{=0}$$

$$=\frac{1}{n_k}\sum_{\ell=0}^{K^\star}\sum_{\ell_1=0}^{K^\star}n_{k,\ell}n_{k,\ell_1}\left({\mu^\star_\ell}^2-\mu^\star_{\ell_1}\mu^\star_\ell\right)=\frac{1}{2n_k}\sum_{\ell=0}^{K^\star}\sum_{\ell'=0}^{K^\star}n_{k,\ell}n_{k,\ell'}\left(\mu^\star_\ell-\mu^\star_{\ell'}\right)^2.$$

Hence,

$$\mathrm{B}_n^D(\mathbf{t})=\frac{1}{n(n+1)}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{\ell=0}^{K^\star}\sum_{\ell'=0}^{K^\star}n_{k,\ell}n_{k,\ell'}\left(\mu^\star_\ell-\mu^\star_{\ell'}\right)^2. \tag{28}$$

Case $K<K^\star$ and $t\in\mathcal{A}^{1/n}_{n,K}$. Observe that $\mathrm{B}_n(\mathbf{t})\geq\mathrm{B}_n^D(\mathbf{t})$.

Because $K<K^\star$, $t_K-t^\star_K=t^\star_{K^\star}-t^\star_K\geq n\Delta^\star_{\boldsymbol{\tau}}$. Hence, $\{k,t_k-t^\star_k\geq n\Delta^\star_{\boldsymbol{\tau}}/2\}\neq\emptyset$. Let $\ell=\min\{k,t_k-t^\star_k\geq\Delta^\star_{\boldsymbol{\tau}}/2\}$, then $\ell\geq 1$ and

$$t_{\ell-1}\leq t^\star_\ell-n\Delta^\star_{\boldsymbol{\tau}}/2\leq t^\star_\ell+n\Delta^\star_{\boldsymbol{\tau}}/2\leq t_\ell.$$

By definition of $\Delta^\star_{\boldsymbol{\tau}}$,

$$\begin{aligned}n_{\ell,\ell}=|D_\ell\cap D^\star_\ell|&\geq\min\{(t^\star_\ell-t_{\ell-1})(t^\star_\ell-t_{\ell-1}+1)/2,(t^\star_\ell-t^\star_{\ell-1})(t^\star_\ell-t^\star_{\ell-1}+1)/2\}\\&\geq\left(n\Delta^\star_{\boldsymbol{\tau}}\right)^2/8,\end{aligned} \tag{29}$$

and

$$n_{\ell,0}\geq\min\{(t_\ell-t^\star_\ell)(t^\star_\ell-t_{\ell-1}),(t^\star_{\ell+1}-t^\star_\ell)(t^\star_\ell-t^\star_{\ell-1})\}\geq\left(n\Delta^\star_{\boldsymbol{\tau}}\right)^2/4. \tag{30}$$

Thus, using (29) and (30), we obtain

$$\mathrm{B}_n(\mathbf{t})\geq\frac{1}{n(n+1)n_\ell}\left[n_{\ell,\ell}n_{\ell,0}\left(\mu^\star_0-\mu^\star_\ell\right)^2\right]\geq\frac{\underline{\lambda}^{(0)2}}{n(n+1)n_\ell}\frac{\left(n\Delta^\star_{\boldsymbol{\tau}}\right)^4}{32}\geq\frac{\left(\Delta^\star_{\boldsymbol{\tau}}\right)^4\underline{\lambda}^{(0)2}}{64}.$$

Because $n_\ell\leq n(n+1)/2$.

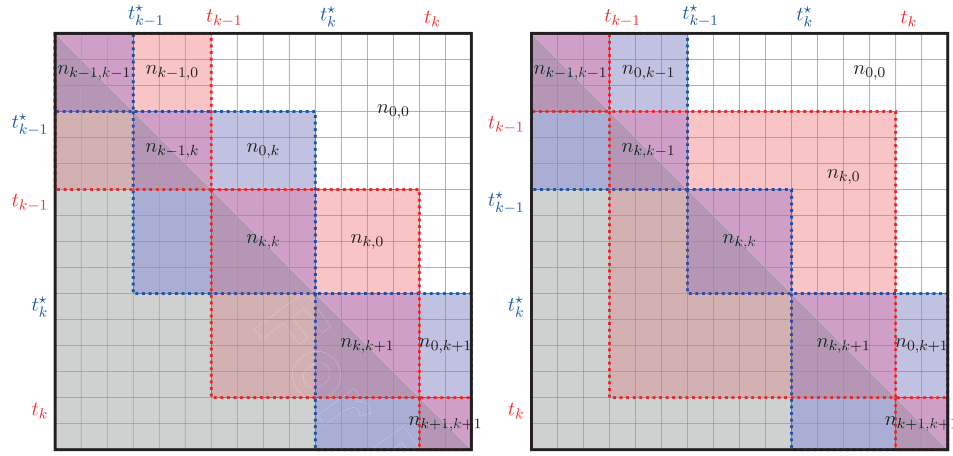

Fig. 7. $K = K^\star$ and $\|\mathbf{t} - \mathbf{t}^\star\|_\infty < \frac{n\Delta^\star_\tau}{2}$. Left: $t^\star_{k-1} < t_{k-1} < t^\star_k < t_k$. Right: $t_{k-1} < t^\star_{k-1} < t^\star_k < t_k$. [Colour figure can be viewed at wileyonlinelibrary.com]

Case $K > K^\star$ *and* $\mathbf{t} \in \mathcal{A}^{\Delta_n}_{n,K}$. We have

$$\mathrm{B}_n(\mathbf{t}) \geq \mathrm{B}^0_n(\mathbf{t}) \geq \frac{2}{n(n+1)} n_{0,k} \left(\mu^\star_k - \mu^\star_0\right)^2 \geq \frac{2}{n(n+1)}\, n_{0,k}\, \underline{\lambda}^{(0)2}$$

for any $k \in \{0, \ldots, K^\star\}$. Because $\mathbf{t} \in \mathcal{A}^{\Delta_n}_{n,K}$, there exists $\ell \in \{1, \ldots, K-1\}$ such that for all $k \in \{0, \ldots, K^\star\}$

$$\left|t^\star_k - t_\ell\right| > \frac{n\Delta_n}{2},$$

(otherwise, it will imply that $K \leq K^\star$). Moreover, let us choose k such that $t^\star_{k-1} + n\Delta_n/2 < t_\ell < t^\star_k - n\Delta_n/2$ then

$$n_{0,k} \geq \left(t_\ell - t^\star_{k-1}\right)\left(t^\star_k - t_\ell\right) \geq \left(\frac{n\Delta_n}{2}\right)^2.$$

This leads to

$$\mathrm{B}_n(\mathbf{t}) \geq \frac{1}{4}\underline{\lambda}^{(0)2}\Delta_n^2.$$

Case $K = K^\star$ *and* $t \in \mathcal{A}^{1/n}_{n,K}$, $\|\mathbf{t} - \mathbf{t}^\star\|_\infty > n\delta$. We have

$$B_n(t) \geq \frac{1}{n(n+1)} \frac{1}{n_\ell} n_{\ell,\ell'} n_{\ell,0} (\mu^\star_0 - \mu^\star_{\ell'})^2 \tag{31}$$

for every $\ell \in \{1, \ldots, K\}$ and every $\ell' \in \{1, \ldots, K^\star\}$. Then, we shall consider two cases: (i) $\|\mathbf{t} - \mathbf{t}^\star\|_\infty < \frac{n\Delta^\star_\tau}{2}$ and (ii) $\|\mathbf{t} - \mathbf{t}^\star\|_\infty \geq \frac{n\Delta^\star_\tau}{2}$.

(i) $\|\mathbf{t} - \mathbf{t}^\star\|_\infty < \frac{n\Delta^\star_\tau}{2}$.

We shall assume that $t_k - t^\star_k = \|\mathbf{t} - \mathbf{t}^\star\|_\infty > 0$.

There are two possible configurations (Fig. 7). If $t^\star_{k-1} < t_{k-1} < t^\star_k < t_k$, then, by definition of $\Delta^\star_\tau$, we obtain

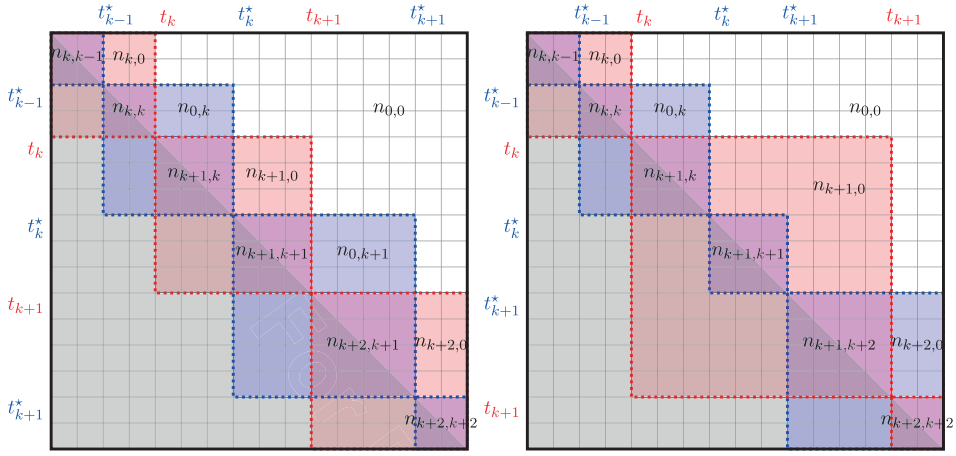

Fig. 8. $K = K^\star$ and $\|\mathbf{t}-\mathbf{t}^\star\|_\infty \geq \frac{n\Delta^\star_{\boldsymbol{\tau}}}{2}$. Left: $t_k < t^\star_k < t^\star_{k+1} < t_{k+1}$. Right: $t_k < t^\star_k < t_{k+1} < t^\star_{k+1}$. [Colour figure can be viewed at wileyonlinelibrary.com]

$$n_{k,k} = \frac{(t^\star_k - t_{k-1})(t^\star_k - t_{k-1}+1)}{2} \geq \frac{\left(\overbrace{(t^\star_k - t^\star_{k-1})}^{\geq n\Delta^\star_{\boldsymbol{\tau}}} - \overbrace{(t_{k-1} - t^\star_{k-1})}^{\leq \|\mathbf{t}-\mathbf{t}^\star\|_\infty}\right)(t^\star_k - t_{k-1}+1)}{2} \geq \frac{(n\Delta^\star_{\boldsymbol{\tau}})^2}{8}. \tag{32}$$

Otherwise, if $t_{k-1} < t^\star_{k-1} < t^\star_k < t_k$, we obtain

$$n_{k,k} = \frac{(t^\star_k - t^\star_{k-1})(t^\star_k - t^\star_{k-1}+1)}{2} \geq \frac{(n\Delta^\star_{\boldsymbol{\tau}})^2}{2} \geq \frac{(n\Delta^\star_{\boldsymbol{\tau}})^2}{8}. \tag{33}$$

Then, by using the aforementioned decomposition of $(t^\star_k - t_{k-1})$, we obtain

$$\begin{aligned} n_{k,0} &\geq (t^\star_k - t_{k-1})(t_k - t^\star_k), \\ &\geq \left(\underbrace{(t^\star_k - t^\star_{k-1})}_{\geq n\Delta^\star_{\boldsymbol{\tau}}} - \underbrace{(t_{k-1} - t^\star_{k-1})}_{\leq \|\mathbf{t}-\mathbf{t}^\star\|_\infty}\right) \underbrace{(t_k - t^\star_k)}_{=\|\mathbf{t}-\mathbf{t}^\star\|_\infty} \geq \frac{\Delta^\star_{\boldsymbol{\tau}}}{2} n^2\delta. \end{aligned} \tag{34}$$

By choosing $(\ell = k, \ell' = k)$ in (31), and by using (32), (33) and (34), we obtain

$$B_n(t) \geq \frac{1}{n(n+1)} \frac{1}{n_k} \frac{(n\Delta^\star_{\boldsymbol{\tau}})^2}{8} \frac{\Delta^\star_{\boldsymbol{\tau}}}{2} n^2 \delta \underline{\lambda}^{(0)2} \geq \frac{(\Delta^\star_{\boldsymbol{\tau}})^3}{32} \delta \underline{\lambda}^{(0)2}.$$

(ii) $\|\mathbf{t}-\mathbf{t}^\star\|_\infty \geq \frac{n\Delta^\star_{\boldsymbol{\tau}}}{2}$.

Because $K = K^\star$, there exists k such that $t^\star_k - t_k \geq n\frac{\Delta^\star_{\boldsymbol{\tau}}}{2}$ and $t_{k+1} - t^\star_k \geq n\frac{\Delta^\star_{\boldsymbol{\tau}}}{2}$ (otherwise, this would imply that $K > K^\star$). As shown previously, there are two possible cases, either $t_k < t^\star_k < t^\star_{k+1} < t_{k+1}$ or $t_k < t^\star_k < t_{k+1} < t^\star_{k+1}$ (Fig. 8).

If $t_k < t^\star_k < t^\star_{k+1} < t_{k+1}$, we obtain, by definition of $\Delta^\star_{\boldsymbol{\tau}}$,

$$n_{k+1,k+1} = \frac{(t^\star_{k+1} - t^\star_k)(t^\star_{k+1} - t^\star_k + 1)}{2} \geq \frac{(n\Delta^\star_{\boldsymbol{\tau}})^2}{2} \tag{35}$$

and

$$n_{k+1,0} \geq (t^\star_{k+1} - t^\star_k)(t^\star_k - t_k) \geq (n\Delta^\star_{\boldsymbol{\tau}})\frac{(n\Delta^\star_{\boldsymbol{\tau}})}{2}. \tag{36}$$

If $t_k < t^\star_k < t_{k+1} < t^\star_{k+1}$, we obtain

$$n_{k+1,k+1} = \frac{(t_{k+1} - t^\star_k)(t_{k+1} - t^\star_k + 1)}{2} \geq \frac{1}{2}\left(\frac{n\Delta^\star_{\boldsymbol{\tau}}}{2}\right)^2 \tag{37}$$

and

$$n_{k+1,0} \geq (t_{k+1} - t^\star_k)(t^\star_k - t_k) \geq \left(\frac{(n\Delta^\star_{\boldsymbol{\tau}})}{2}\right)^2. \tag{38}$$

By choosing ($\ell = \ell' = k+1$) in (31), and by using (35), (36), (37) and (38), we obtain

$$B_n(t) \geq \frac{(\Delta^\star_{\boldsymbol{\tau}})^4}{32}\underline{\lambda}^{(0)2}.$$

6.3. Deviation inequalities

Lemma 2. *For all $\alpha > 0$,*

$$\mathbb{P}\left(-\min_{\mathbf{t}\in\mathcal{A}^{1/n}_{n,K}} V_n(\mathbf{t}) \geq \alpha\right) \leq n(n+1)e^{-\frac{n(n+1)\alpha}{16K\beta}} + 2e^{-\frac{|G_{01}|\alpha}{8\beta}},$$

where V_n is defined by (20) and (22) and $\mathcal{A}^{1/n}_{n,K}$ is defined in (6) with $\Delta_n = 1/n$. Moreover, if $\alpha = \alpha_n$ is such that $\alpha_n n^2/\log(n) \to \infty$ and $\alpha_n|G_{0,1}| \to \infty$, as n tends to infinity, then

$$\mathbb{P}\left(-\min_{\mathbf{t}\in\mathcal{A}^{1/n}_{n,K}} V_n(\mathbf{t}) \geq \alpha_n\right) \to 0, \text{ as } n \to +\infty.$$

The proof is given in the supplementary material.

Lemma 3. *Let W_n be defined by (20) and (23), then there exists $C_1 > 0$ such that for all $\alpha > 0$*

$$\mathbb{P}\left(-\min_{\mathbf{t}\in\mathcal{A}^{1/n}_{n,K}} W_n(\mathbf{t}) \geq \alpha\right) \leq C_1 n^{4K_{\max}} \exp\left[-\frac{\alpha^2 n(n+1)}{128\,\beta\,(K+1)^2\,(K^\star+1)^2\,\overline{\lambda}^2}\right],$$

where $\overline{\lambda} = \sup_{k\neq\ell}\left|\mu^\star_k - \mu^\star_\ell\right|$ and $\mathcal{A}^{1/n}_{n,K}$ is defined in (6) with $\Delta_n = 1/n$. Moreover, if $\alpha = \alpha_n$ is such that $\alpha_n^2 n^2/\log(n) \to \infty$, as n tends to infinity, then

$$\mathbb{P}\left(-\min_{\mathbf{t}\in\mathcal{A}^{1/n}_{n,K}} W_n(\mathbf{t}) \geq \alpha_n\right) \to 0, \text{ as } n \to \infty.$$

The proof is given in the supplementary material.

Lemma 4. *For all $\alpha > 0$ and $\gamma > 0$,*

$$\mathbb{P}\left(-\min_{\mathbf{t}\in\mathcal{A}^{1/n}_{n,K}} Z_n(\mathbf{t}) \geq \alpha\right) \leq 2e^{-\frac{|G_{01}|\gamma^2}{4\beta}} + 2C_1 n^{4K_{\max}} e^{-\frac{\alpha^2 n(n+1)}{512\gamma^2\beta}} + 2e^{-\frac{|G_{01}|\alpha^2 n^2}{32\overline{\lambda}^2\beta}},$$

where Z_n is defined by (24), $\mathcal{A}^{1/n}_{n,K}$ is defined in (6) with $\Delta_n = 1/n$ and $\overline{\lambda} = \sup_{k\neq\ell}\left|\mu^\star_k - \mu^\star_\ell\right|$.

Moreover, if $\alpha = \alpha_n$ *is such that* $\alpha_n^2 n^2 / \log(n) \to \infty$ *and* $\alpha_n^2 n^2 |G_{0,1}| \to \infty$*, as n tends to infinity, then*

$$\mathbb{P}\left(- \min_{\mathbf{t} \in \mathcal{A}_{n,K}^{1/n}} Z_n(\mathbf{t}) \geq \alpha_n \right) \to 0, \text{ as } n \to \infty.$$

The proof is given in the supplementary material.

Acknowledgement

Vincent Brault would like to thank the French National Research Agency ANR which supported this research through the ABS4NGS project (ANR-11-BINF-0001-06).

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.

References

Basseville, M. & Nikiforov, I. (1993). *Detection of Abrupt Changes: Theory and Application*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Bellman, R. (1961). On the approximation of curves by line segments using dynamic programming for Reference bellman, (1961). *Commun. ACM* **4**, 1–18.

Boysen, L., Kempe, A., Liebscher, V., Munk, A. & Wittich, O. (2009). Consistencies and rates of convergence of jump-penalized least squares estimators. *Ann. Statist.* **37**, 157–183.

Brodsky, B. & Darkhovsky, B. (2000). *Non-parametric Statistical Diagnosis: Problems and Methods*, Kluwer Academic Publishers, Netherlands.

Dehman, A., Ambroise, C. & Neuvial, P. (2015). Performance of a blockwise approach in variable selection using linkage disequilibrium information. *BMC Bioinformatics* **16**, 148. doi:10.1186/s12859-015-0556-6.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380.

Ioanna Delatola, E, Lebarbier, E., Mary-Huard, T., Radvanyi, F., Robin, S. & Wong, J. (2015). SegCorr: A statistical procedure for the detection of genomic regions of correlated expression. *ArXiv e-prints*.

Lavielle, M. (2005). Using penalized contrasts for the change-point problem. *Signal Processing* **85**, 1501–1510.

Lavielle, M. & Moulines, E. (2000). Least-squares estimation of an unknown number of shifts in a time series. *J. Time Ser. Anal.* **21**, 33–59.

Lebarbier, E. (2005). Detecting multiple change-points in the mean of a Gaussian process by model selection. *Signal Processing* **85**, 717–736.

Lévy-Leduc, C., Delattre, M., Mary-Huard, T. & Robin, S. (2014). Two-dimensional segmentation for analyzing HiC data. *Bioinformatics* **30**, 386–392.

Massart, P. (2004). A non asymptotic theory for model selection. In *European Congress of Mathematics: Stockholm, June 27-July 2, 2004*, 309–324, Stockholm.

Tartakovsky, A., Nikiforov, I. & Basseville, M. (2014). *Sequential Analysis: Hypothesis Testing and Changepoint Detection*, Chapman & Hall/CRC, Taylor and Francis Group, Boca Raton, FL, USA.

Yao, Y. & Au, S. T. (1989). Least-squares estimation of a step function. *Sankhya Ser. A* **51**, 370–381.

Received June 2015, in final form May 2016

Vincent Brault, UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay.

E-mail: vincent.brault@agroparistech.fr