Nonparametric multiple change-point estimation for analyzing large Hi-C data matrices

Vincent Brault, Sarah Ouadah, Laure Sansonnet, Céline Lévy-Leduc

Nonparametric multiple change-point estimation for analyzing large Hi-C data matrices

Vincent Brault[a], Sarah Ouadah[b], Laure Sansonnet[b], Céline Lévy-Leduc[b]

[a] *Univ. Grenoble Alpes, CNRS, LJK, 38000 Grenoble, France*
[b] *UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay*

Abstract

We propose a novel nonparametric approach to estimate the location of block boundaries (change-points) of non-overlapping blocks in a random symmetric matrix which consists of random variables whose distribution changes from block to block. Our change-point location estimators are based on nonparametric homogeneity tests for matrices. We first provide some theoretical results for these tests. Then, we prove the consistency of our change-point location estimators. Some numerical experiments are also provided in order to support our claims. Finally, our approach is applied to Hi-C data which are used in molecular biology to study the influence of chromosomal conformation on cell function.

Keywords: Hi-C data, multiple change-point estimation, nonparametric estimation.

1. Introduction

Detecting and identifying the location of changes in the distribution of random variables is a major statistical issue that arises in many fields such as industrial process surveillance [2], anomalie detection in internet traffic data [15, 21], and molecular biology. In the latter field, several change-point detection methods have been designed to deal with different kinds of data such as Copy Number Variation or CNV [19, 23], RNAseq data [7], and more recently Hi-C data which motivated this work.

Hi-C technology is a recent chromosome conformation capture method that was developed to enhance our understanding of the influence of chromosomal conformation on cell function. This technology, which is based on a deep sequencing approach, provides read pairs corresponding to pairs of genomic loci that physically interact in the nucleus [16]. The raw measurements provided by Hi-C data are often summarized as a square matrix where entry (i, j) gives the total number of read pairs matching in positions i and j; see [8] for further details. Blocks of different intensities arise within this matrix, revealing interacting genomic regions among which some have already been confirmed to host co-regulated genes. The purpose of the statistical analysis is then to provide an efficient strategy to determine a decomposition of the matrix in non-overlapping blocks, yielding as a by-product a list of non-overlapping interacting chromosomic regions.

This issue has already been addressed by Lévy-Leduc et al. [14] in the particular framework where the mean of the observations changes from one diagonal block to the other and is constant everywhere else. In this work, the authors use a parametric maximum likelihood approach. In contrast, we will address here the case where the non-overlapping blocks are no longer diagonal using a nonparametric method. Our goal will thus be to design an efficient and nonparametric method to find the block boundaries, also called change-points, of non-overlapping blocks in large matrices which can be modeled as matrices of random variables whose distribution changes from one block to the next.

A large literature is devoted to change-point detection when both the number of observations and the number of vectors go to infinity at different rates. Horváth and Hušková [9] proposed a change-point detection approach also in the context where the number of observations and the number of vectors go to infinity but cannot be equal. Cho and Fryzlewicz [6] devised a parametric approach to identify multiple change-points in the second-order structure of a multivariate (possibly high-dimensional) time series based on localized periodograms and cross-periodograms computed on the original multivariate time series. Jirak [10] proposed nonparametric change-point tests in very general high-dimensional settings. Matteson and James [18] devised a nonparametric change-point estimation procedure which allows them to retrieve

change-points within n K-variate multivariate observations, where K is fixed and n may be large. It is based on the use of an empirical divergence measure derived from the divergence measure of Szekely and Rizzo [20]. Another approach based on ranks has been proposed by Lung-Yut-Fong et al. [17] in the same framework as [18]. Their approach consists in extending the classical Wilcoxon and Kruskal–Wallis statistics [13] to the multivariate case.

In this paper, we propose a nonparametric change-point estimation approach based on nonparametric homogeneity tests generalizing the approach of [17] to the case where we have to deal with large matrices instead of fixed vectors. Moreover, our methodology is adapted to our very specific problem where we have to process a large symmetric matrix $\mathbf{X} = (X_{i,j})_{1\leq i,j\leq n}$ such that the $X_{i,j}$s are independent random variables when $i \geq j$. Hence, in our case, the number of observations and the number of vectors are equal and both go to infinity. This specific setting has never been considered, so far as we know.

The paper is organized as follows. We first propose in Section 2 nonparametric homogeneity tests for two, and more than two, samples. In Section 3, we deduce from these tests a nonparametric procedure to estimate the block boundaries of a matrix of random variables whose distribution changes from block to block. The consistency of these change-point location estimators is established in Theorems 3–4. These methods are then illustrated by some numerical experiments in Section 4. An application to real Hi-C data is also given in Section 5. Finally, the proofs of our theoretical results are given in Section 7.

2. Homogeneity tests

In this section, we propose nonparametric homogeneity test statistics for two, and more than two, samples. These statistics will be used in Section 3 to estimate the location of block boundaries (change-points) of non-overlapping blocks in a random symmetric matrix.

2.1. Two-sample homogeneity test

Let $\mathbf{X} = (X_{i,j})_{1\leq i,j\leq n}$ be a symmetric matrix whose entries $X_{i,j}$ are independent random variables when $i \geq j$. Observe that $\mathbf{X}$ can be rewritten as $\mathbf{X} = (\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)})$, where $\mathbf{X}^{(j)} = (X_{1,j}, \ldots, X_{n,j})^\top$ denotes the jth column of $\mathbf{X}$.

Let n_1 be a given integer in $\{1, \ldots, n\}$. The purpose of this section is to propose a statistic to test the null hypothesis $\mathcal{H}_0$: "$(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n_1)})$ and $(\mathbf{X}^{(n_1+1)}, \ldots, \mathbf{X}^{(n)})$ are identically distributed random vectors" against the alternative $\mathcal{H}_1$: "$(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n_1)})$ has distribution $\mathbb{P}_1$ and $(\mathbf{X}^{(n_1+1)}, \ldots, \mathbf{X}^{(n)})$ has distribution $\mathbb{P}_2$, where $\mathbb{P}_1 \neq \mathbb{P}_2$". Hypothesis $\mathcal{H}_0$ means that for all $i \in \{1, \ldots, n\}$, $X_{i,1}, \ldots, X_{i,n}$ are independent and identically distributed (iid) random variables and while alternative $\mathcal{H}_1$ means that there exists $i \in \{1, \ldots, n\}$ such that $X_{i,1}, \ldots, X_{i,n_1}$ have distribution $\mathbb{P}^i_1$ and $X_{i,n_1+1}, \ldots, X_{i,n}$ have distribution $\mathbb{P}^i_2$, with $\mathbb{P}^i_1 \neq \mathbb{P}^i_2$.

To decide whether $\mathcal{H}_0$ should be rejected or not, we propose to use a test statistic inspired by the one designed by [17] which extends the well-known Wilcoxon–Mann–Whitney rank-based test to deal with multivariate data. Our statistical test can thus be seen as a way to decide whether n_1 can be considered as a potential change in the distribution of the $X_{i,j}$s or not.

The test statistic that we propose for assessing the presence of the potential change n_1 is defined by

$$S_n(n_1) = \sum_{i=1}^{n} U^2_{n,i}(n_1), \tag{1}$$

where

$$U_{n,i}(n_1) = \frac{1}{\sqrt{nn_1(n-n_1)}} \sum_{j_0=1}^{n_1} \sum_{j_1=n_1+1}^{n} h(X_{i,j_0}, X_{i,j_1}),$$

with $h(x, y) = \mathbf{1}_{\{x\leq y\}} - \mathbf{1}_{\{y\leq x\}}$.

Our framework is different from that of Lung-Yut-Fong et al. [17] because the vectors $\mathbf{X}^{(j)}$ they consider are K-variate with K fixed, while ours are n-dimensional where n may be large.

Note that the statistic $U_{n,i}$ can also be written using the rank $R^{(i)}_j$ of $X_{i,j}$ among $X_{i,1}, \ldots, X_{i,n}$. Indeed,

$$U_{n,i}(n_1) = \frac{2}{\sqrt{nn_1(n-n_1)}} \sum_{j_0=1}^{n_1} \left(\frac{n+1}{2} - R^{(i)}_{j_0} \right) = \frac{2}{\sqrt{nn_1(n-n_1)}} \sum_{j_1=n_1+1}^{n} \left(R^{(i)}_{j_1} - \frac{n+1}{2} \right), \tag{2}$$

where

$$R_j^{(i)} = \sum_{k=1}^{n} \mathbf{1}_{\{X_{i,k} \le X_{i,j}\}}. \tag{3}$$

Form (2) will be used below to extend the two-sample homogeneity test to the multiple sample case.

If the distribution of the $X_{i,j}$s is continuous, then the following theorem establishes that the test statistic $S_n(n_1)$ is properly normalized, namely $S_n(n_1)$ is bounded in probability as $n \to \infty$ under $\mathcal{H}_0$. Note that $\mathcal{H}_0$ assumes that for all $i \in \{1, \dots, n\}$, $X_{i,1}, \dots, X_{i,n}$ are iid random variables. Since we also assume that $\mathbf{X} = (X_{i,j})_{1 \le i,j \le n}$ is a symmetric matrix whose entries are independent random variables when $i \ge j$, it implies that under $\mathcal{H}_0$, all the rows i have the same distribution. Hence, all the $X_{i,j}$s such that $i \ge j$ are iid.

Theorem 1. *Let $\mathbf{X} = (X_{i,j})_{1 \le i,j \le n}$ be a symmetric matrix of random variables $X_{i,j}$ whose entries are iid when $i \ge j$. Assume that the distribution of the $X_{i,j}$s is continuous and that there exists $\tau_1 \in (0,1)$ such that $n_1/n \to \tau_1$ as $n \to \infty$. Then, as $n \to \infty$, $T_n(n_1) \equiv n^{-1/2}\left[S_n(n_1) - \mathrm{E}\left\{S_n(n_1)\right\}\right] = O_P(1)$, where $\mathrm{E}\left\{S_n(n_1)\right\} = (n+1)/3$.*

The proof of Theorem 1 is given in Section 7.1. Observe that the assumptions of Theorem 1 correspond to the null hypothesis $\mathcal{H}_0$ described in Section 2.1. Hence, we could reject $\mathcal{H}_0$ when

$$T_n(n_1) > s, \tag{4}$$

for some threshold s. A way of computing this threshold will be given in Section 4.1.

2.2. Multiple-sample homogeneity test

The purpose of this section is to extend the two-sample homogeneity test of the previous section to deal with the multiple sample case.

Let us assume that $\mathbf{X} = (X_{i,j})_{1 \le i,j \le n}$ is still a symmetric matrix whose entries are independent random variables when $i \ge j$. Let $0 = n_0 < \cdots < n_{L+1} = n$ be L integers given in $\{1, \dots, n-1\}$. We propose a statistic to test $\mathcal{H}_0$: "$(\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(n_1)})$, $(\mathbf{X}^{(n_1+1)}, \dots, \mathbf{X}^{(n_2)})$, …,$(\mathbf{X}^{(n_L+1)}, \dots, \mathbf{X}^{(n)})$ have the same distribution" against the alternative $\mathcal{H}_1$: "there exists $\ell \in \{1, \dots, L\}$ such that $(\mathbf{X}^{(n_{\ell-1}+1)}, \dots, \mathbf{X}^{(n_\ell)})$ has distribution $\mathbb{P}_\ell$ and $(\mathbf{X}^{(n_\ell+1)}, \dots, \mathbf{X}^{(n_{\ell+1})})$ has distribution $\mathbb{P}_{\ell+1}$, where $\mathbb{P}_\ell \ne \mathbb{P}_{\ell+1}$".

The homogeneity test presented in the previous section for two groups can be extended in order to deal with $L+1$ groups instead of 2 by using the following statistic:

$$S_n(n_1, \dots, n_L) = \frac{4}{n^2} \sum_{\ell=0}^{L} (n_{\ell+1} - n_\ell) \sum_{i=1}^{n} \left(\overline{R}_\ell^{(i)} - \frac{n+1}{2} \right)^2, \tag{5}$$

with

$$\overline{R}_\ell^{(i)} \quad = \quad \frac{1}{n_{\ell+1} - n_\ell} \sum_{j=n_\ell+1}^{n_{\ell+1}} R_j^{(i)}, \tag{6}$$

where the rank $R_j^{(i)}$ of $X_{i,j}$ is defined by (3) and $\overline{R}_\ell^{(i)}$ is its mean in the group ℓ.

Observe that (5) can be seen as a natural extension of the classical Kruskal–Wallis statistic for univariate observations to the multivariate case; see [22, p. 181].

Remark 1. *Note that when $L = 1$, $S_n(n_1)$ defined in* (5) *boils down to $S_n(n_1)$ defined in* (1) *since*

$$\frac{4}{n^2} \left\{ n_1 \sum_{i=1}^{n} \left(\frac{1}{n_1} \sum_{j=1}^{n_1} R_j^{(i)} - \frac{n+1}{2} \right)^2 + (n - n_1) \sum_{i=1}^{n} \left(\frac{1}{n - n_1} \sum_{j=n_1+1}^{n} R_j^{(i)} - \frac{n+1}{2} \right)^2 \right\}$$
$$= \frac{4}{n^2 n_1} \sum_{i=1}^{n} \left\{ \sum_{j=1}^{n_1} \left(R_j^{(i)} - \frac{n+1}{2} \right) \right\}^2 + \frac{4}{n^2 (n - n_1)} \sum_{i=1}^{n} \left\{ \sum_{j=n_1+1}^{n} \left(R_j^{(i)} - \frac{n+1}{2} \right) \right\}^2,$$

which, in view of (2)*, can be reexpressed as*

$$\sum_{i=1}^{n} U_{n,i}^2(n_1) = \frac{1}{n}\left[\sum_{i=1}^{n}(n-n_1)\left\{\frac{2}{\sqrt{nn_1(n-n_1)}}\sum_{j=1}^{n_1}\left(R_j^{(i)}-\frac{n+1}{2}\right)\right\}^2 + \sum_{i=1}^{n} n_1\left\{\frac{2}{\sqrt{nn_1(n-n_1)}}\sum_{j=n_1+1}^{n}\left(R_j^{(i)}-\frac{n+1}{2}\right)\right\}^2\right].$$

If the distribution of the $X_{i,j}$s is continuous, then the following theorem establishes that the test statistic $S_n(n_1,\ldots,n_L)$ is properly normalized, namely $S_n(n_1,\ldots,n_L)$ is bounded in probability as $n\to\infty$.

Theorem 2. *Let $\boldsymbol{X}=(X_{i,j})_{1\le i,j\le n}$ be a symmetric matrix of random variables $X_{i,j}$ whose entries are iid when $i\ge j$. Assume that distribution of the $X_{i,j}$s is continuous and that there exist $0<\tau_1<\cdots<\tau_L<1$ such that for all $\ell\in\{1,\ldots,L\}$, $n_\ell/n\to\tau_\ell$ as $n\to\infty$. Then, as $n\to\infty$, $n^{-1/2}\left[S_n(n_1,\ldots,n_L)-\mathrm{E}\left\{S_n(n_1,\ldots,n_L)\right\}\right]=O_P(1)$, with $\mathrm{E}\left\{S_n(n_1,\ldots,n_L)\right\}=L(n+1)/3$.*

The proof of Theorem 2 is given in Section 7.2. Note that the n_ℓs can be seen as the boundaries of groups of random variables having different distributions. We will explain in the next section how to derive from this theorem a methodology for estimating the n_ℓs when they are assumed to be unknown.

3. Change-point estimation

We propose in this section to use the test statistic (5) to derive the location of the block boundaries $n_1^*<\cdots<n_L^*$. More precisely, we propose to estimate $(n_1^*,\ldots,n_L^*)$ as follows:

$$\widehat{\boldsymbol{n}}=(\widehat{n}_1,\ldots,\widehat{n}_L)\equiv \mathrm{Argmax}_{1\le n_1<\cdots<n_L<n}\, S_n(n_1,\ldots,n_L), \tag{7}$$

where $S_n(n_1,\ldots,n_L)$ is defined in (5). For all $\ell\in\{0,\ldots,L\}$, set

$$D_\ell^*=\{i\in\{1,\ldots,n\}\ :\ n_\ell^*+1\le i\le n_{\ell+1}^*\}. \tag{8}$$

3.1. Theoretical results

The following theorem establishes that the procedure provides a consistent estimator for the change-point location when $L=1$.

Theorem 3. *Let $\boldsymbol{X}=(X_{i,j})_{1\le i,j\le n}$ be a symmetric matrix whose entries are independent random variables when $i\ge j$ with a continuous distribution. Let $\mathbb{P}_0^0$ be the distribution of $X_{i,j}$ for $i,j\in D_0^*$, $\mathbb{P}_1^0$ be the distribution of $X_{i,j}$ for $i\in D_0^*$, $j\in D_1^*$, and $\mathbb{P}_1^1$ be the distribution of $X_{i,j}$ for $i,j\in D_1^*$ where $\mathbb{P}_0^0\neq\mathbb{P}_1^0$ or $\mathbb{P}_1^0\neq\mathbb{P}_1^1$. Assume that*

$$\Pr(X\le Y)\neq 1/2, \tag{9}$$

where $X\sim\mathbb{P}_0^0$ (or $\mathbb{P}_1^0$) and $Y\sim\mathbb{P}_0^0$ (or $\mathbb{P}_1^1$). Assume also that there exists $\tau_1^*\in(0,1)$ such that $n_1^*/n\to\tau_1^*$, as $n\to\infty$. Then, for all positive δ, as $n\to\infty$, $\Pr\left(|\widehat{n}_1-n_1^*|\ge n\delta\right)\to 0$, where $\widehat{n}_1$ is defined by (7) *when $L=1$.*

Remark 2. *Note that Assumption* (9) *in Theorem 3 is classic in the context of rank-based test statistics such as the Mann–Whitney test; see [22].*

The proof of Theorem 3 is given in Section 7.3. The following theorem extends the results of Theorem 3 to the case where $L>1$.

Theorem 4. *Let $\boldsymbol{X}=(X_{i,j})_{1\le i,j\le n}$ be a symmetric matrix whose entries are independent random variables when $i\ge j$ with a continuous distribution. Let $\mathbb{P}_{\ell_1}^{\ell_2}$ be the distribution of $X_{i,j}$ for $i\in D_{\ell_2}^*$ and $j\in D_{\ell_1}^*$ and F_{ℓ_2,ℓ_1} the associated cumulative distribution function, where the D_ℓ^*s are defined in* (8)*. Assume that for all ℓ*

in $\{1,\dots,L\}$*, there exists* $\tau_\ell^* \in (0,1)$ *such that* $n_\ell^*/n \to \tau_\ell^*$*, as* $n \to \infty$ *such that* $\Delta_\tau^* = \min_{0\le\ell\le L} |\tau_{\ell+1}^* - \tau_\ell^*| > 0$*. Assume also that, for all* $\ell_1 \in \{0,\dots,L-1\}$*, there exists* $\ell_4 \in \{0,\dots,L\}$ *such that*

$$\sum_{\ell_3=0}^{L} (\tau_{\ell_3+1}^* - \tau_{\ell_3}^*) \mathrm{E}\left\{F_{\ell_4,\ell_1}(X) - F_{\ell_4,\ell_1+1}(X)\right\} \neq 0, \tag{10}$$

where $X \sim \mathbb{P}_{\ell_3}^{\ell_4}$*. Then, for all positive* δ*,* $\Pr\left(\|\widehat{\boldsymbol{n}} - \boldsymbol{n}^*\|_\infty \ge n\delta\right) \to 0$*, as* $n \to \infty$*, where* $\widehat{\boldsymbol{n}}$ *is defined by* (7)*,* $\boldsymbol{n}^* = (n_1^*,\dots,n_L^*)$ *and* $\|\widehat{\boldsymbol{n}} - \boldsymbol{n}^*\|_\infty = \max_{0\le\ell\le L} |\widehat{n_\ell} - n_\ell^*|$*.*

Remark 3. *Observe that Assumption* (10) *holds for example when, for all* ℓ_1*, there exists* ℓ_4 *in* $\{0,\dots,L\}$*, such that, for all* x*,* $F_{\ell_4,\ell_1}(x) > F_{\ell_4,\ell_1+1}(x)$ *or for all* x*,* $F_{\ell_4,\ell_1}(x) < F_{\ell_4,\ell_1+1}(x)$*. This holds for instance in the case where the* $X_{i,j}$ *are Gaussian random variables having different means in two consecutive blocks.*

A sketch of the proof of Theorem 4 is given in Section 7.4.

3.2. Practical implementation

In practice, maximizing (7) directly is computationally prohibitive; the task's complexity grows exponentially with L. However, thanks to the additive structure of (5), it is possible to use a dynamic programming strategy as we will explain hereafter. We refer here to the classical dynamic programming approach described in [11] which can be traced back to the note of Bellman [4].

Let us introduce the following notations:

$$\Delta(n_\ell + 1 : n_{\ell+1}) = (n_{\ell+1} - n_\ell) \sum_{i=1}^{n} \left(\overline{R}_\ell^{(i)} - \frac{n+1}{2}\right)^2,$$

where $\overline{R}_\ell^{(i)}$ is defined by (6) and

$$I_L(p) = \max_{1\le n_1<\cdots<n_L<n_{L+1}=p} \sum_{\ell=0}^{L} \Delta(n_\ell + 1 : n_{\ell+1}), \tag{11}$$

for $L \in \{0,\dots,L_{\max}\}$ and $p \in \{1,\dots,n\}$, where $L_{\max}$ is assumed to be a known upper bound for the number of block boundaries. Observe that $I_L(p)$ satisfies the recursive formula

$$I_L(p) = \max_{n_L} \left\{I_{L-1}(n_L) + \Delta(n_L + 1 : p)\right\}, \tag{12}$$

which is proved in Section 7.5. Thus, for solving the optimization problem (7), we proceed as follows. We start by computing $\Delta(i : j)$ for all (i,j) such that $1 \le i \le j \le n$. All the $I_0(p)$ are thus available for $p \in \{1,\dots,n\}$. Then $I_1(p)$ is computed by using the recursion (12) and so on. Hence the complexity of our algorithm is $O(n^3)$.

Figure 1 shows the computational times in seconds associated with our multiple change-point estimation strategy based on the dynamic programming algorithm. The computational time of our procedure is seen to be polynomial. For instance, it takes 15 minutes to our algorithm to process a 500×500 matrix. Note that in the framework of univariate time series segmentation, the PELT procedure devised by [12] performs multiple change-point detection at a linear computational cost. It would be interesting to see if the computational burden of our procedure could be reduced by using an extension of their approach.

4. Numerical experiments

4.1. Statistical performance of the two-sample homogeneity test

We propose hereafter a procedure for calibrating the threshold s of the rejection region $T_n(n_1) > s$ defined in (4). To ensure that the two-sample homogeneity test is of level α, an estimation of the $1-\alpha$ quantile of $T_n(n_1)$ has to be provided. In the sequel, such an estimation is given in the case where $\alpha = 0.05$.

We generated 10,000 $n \times n$ symmetric matrices $\boldsymbol{X} = (X_{i,j})$ with $n \in \{50, 100, 500, 1000\}$. More precisely, the $(X_{i,j})_{i\ge j}$s are independent random variables having a zero mean standard Gaussian distribution, $\mathcal{N}(0,1)$,

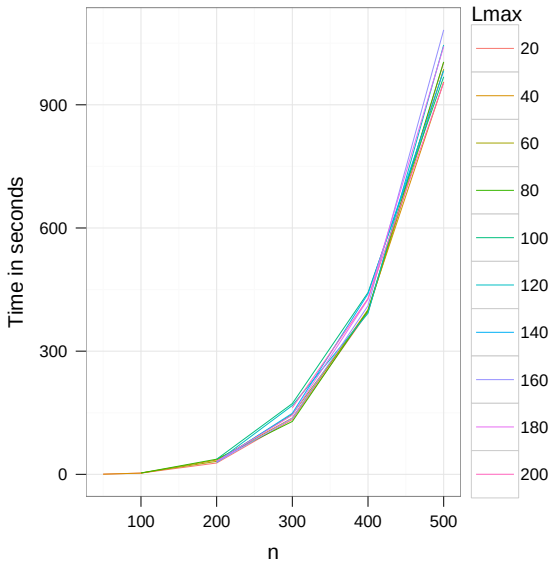

Figure 1: Computational times in seconds for the dynamic programming algorithm described in Section 3 as a function of n for different values of $L_{\max}$.

a Cauchy distribution with 0 and 1 location and scale parameters, $\mathcal{C}(0,1)$, respectively or an exponential distribution with parameter 2, $\mathcal{E}(2)$. We consider two values for n_1, namely $n_1 = \lfloor 0.1n \rfloor$ and $n_1 = \lfloor 0.5n \rfloor$, where $\lfloor x \rfloor$ denotes the integer part of x.

The empirical 0.95 quantiles of $T_n(n_1)$ are given in Table 1. We observe from this table that the empirical 0.95 quantiles do not seem to be sensitive neither to the values of n_1 and n nor to the distribution of the observations since they slightly vary around 0.8.

4.1.1. Power of the test statistic

In this section, we study the power of the two-sample homogeneity test defined in Section 3. We generated 10,000 $n \times n$ symmetric matrices $\boldsymbol{X} = (X_{i,j})$ split into four blocks defined as follows and $n \in \{50, 100, 500, 1000\}$. Let $\mathcal{I}_1 = \{(i,j) : 1 \leq j \leq i \leq n_1\}$, $\mathcal{I}_2 = \{(i,j) : 1 \leq j \leq n_1,\ n_1 + 1 \leq i \leq n\}$, and $\mathcal{I}_3 = \{(i,j) : n_1 + 1 \leq j \leq i \leq n\}$.

In the sequel, we assume that $(X_{i,j})_{(i,j)\in\mathcal{I}_1} \overset{\text{iid}}{\sim} \mathcal{L}_1$, $(X_{i,j})_{(i,j)\in\mathcal{I}_2} \overset{\text{iid}}{\sim} \mathcal{L}_2$ and $(X_{i,j})_{(i,j)\in\mathcal{I}_3} \overset{\text{iid}}{\sim} \mathcal{L}_3$ and we take the following values for n_1: $n_1 = \lfloor 0.1n \rfloor$ and $n_1 = \lfloor 0.5n \rfloor$.

Figure 2 displays the power curves of the two-sample homogeneity test defined in Section 2 when $\mathcal{L}_1 = \mathcal{L}_3 = \mathcal{N}(0,1)$ and $\mathcal{L}_2 = \mathcal{N}(\mu, 1)$, where $\mu \in \{0, 0.01, 0.02, \ldots, 0.99, 1\}$. We can see from this figure that for large values of n, our testing procedure appears to be powerful whatever the value of μ. For small values of n, we observe that the power of our testing procedure becomes higher as μ increases.

Table 1: Estimation of the empirical 0.95 quantiles of $T_n(n_1)$.

	$n_1 = \lfloor 0.1n \rfloor$			$n_1 = \lfloor 0.5n \rfloor$		
	$\mathcal{N}(0,1)$	$\mathcal{C}(0,1)$	$\mathcal{E}(2)$	$\mathcal{N}(0,1)$	$\mathcal{C}(0,1)$	$\mathcal{E}(2)$
$n = 50$	0.83	0.83	0.82	0.78	0.79	0.76
$n = 100$	0.81	0.8	0.82	0.78	0.8	0.78
$n = 500$	0.78	0.8	0.81	0.8	0.78	0.77
$n = 1000$	0.79	0.78	0.79	0.78	0.77	0.79

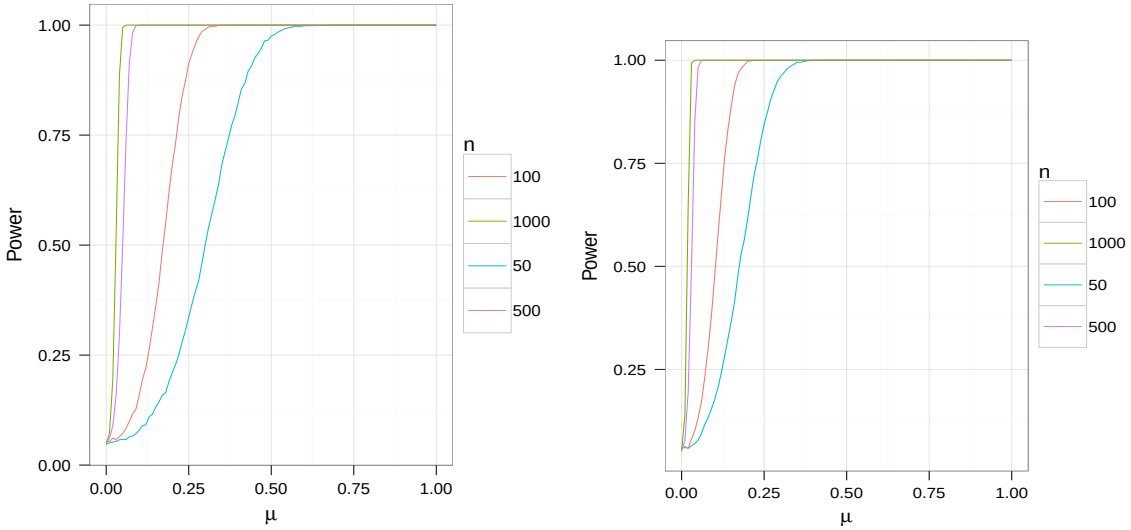

Figure 2: Power curves for the two-sample homogeneity test as a function of μ for different values of n, $n_1 = \lfloor 0.1n \rfloor$ (left) and $n_1 = \lfloor 0.5n \rfloor$ (right).

4.2. Statistical performance of the multiple change-point estimation procedure

In this section, we study the statistical performance of the multiple change-point estimation procedure described in Section 3. This method is implemented in the R package `MuChPoint` available on the Comprehensive R Archive Network (CRAN).

We generated 10,000 $n \times n$ symmetric matrices $\mathbf{X} = (X_{i,j})$ where $n \in \{50, 100, 200, 300, 400\}$ with different block configurations and $L = 10$ block boundaries (change-points). In the following, this value of L is assumed to be known.

We first consider the Block Diagonal configuration. In this case, the matrix consists of diagonal blocks of size $n/10$. Within each of these diagonal blocks, the $X_{i,j}$s such that $i \geq j$ are independent and have distribution $\mathcal{L}_1$. The $X_{i,j}$s lying in the extra-diagonal part of the lower triangular part of $\mathbf{X}$ are independent and have distribution $\mathcal{L}_2$, with $\mathcal{L}_2 \neq \mathcal{L}_1$. The upper triangular part of $\mathbf{X}$ is then derived by symmetry.

We also consider the Chessboard configuration. In this case, the matrix consists of non overlapping blocks of size $n/10$. The $X_{i,j}$s belonging to two blocks sharing a boundary have different distributions. This configuration implies that only two distributions $\mathcal{L}_1$ and $\mathcal{L}_2$ are at stake. The distribution of the upper left block is denoted by $\mathcal{L}_1$ in the sequel.

For these two configurations, we consider for $\mathcal{L}_1$ a $\mathcal{N}(1, \sigma^2)$, a $\mathcal{E}(2)$ or a $\mathcal{C}(1, a)$ distribution where $\sigma, a \in \{1, 2, 5\}$. The $\mathcal{L}_2$ distributions associated with each of them are $\mathcal{N}(0, \sigma^2)$, $\mathcal{E}(\lambda)$ and $\mathcal{C}(0, a)$ where $\lambda \in \{1, 0.5, 4\}$. We display in Figure 3 some examples of the Block Diagonal and Chessboard configurations for the Gaussian, exponential and Cauchy distributions. In these plots, large values are displayed in red and small values in blue.

In the Gaussian Chessboard configuration, Figure 4 displays the frequency of the number of times where each position in $\{1, \ldots, n-1\}$ has been estimated as a change-point. We can see from this figure that the true change-point positions are in general properly retrieved by our approach even in cases where the change-points are not easy to detect with the naked eye. However, we observe that when σ increases, some spurious change-points appear close to the true change-point positions.

We also compared our multiple change-point estimation strategy (`MuChPoint`) to the one devised by [18] (`ecp`), which is, to the best of our knowledge, the most recent approach proposed for solving this issue. The results are illustrated in Figures 5 and 6, which display the boxplots of the distance $\mathcal{D}$, defined in (13), between the change-points provided by these procedures in the Block Diagonal and Chessboard configurations for the Gaussian, exponential and Cauchy distributions. To use the `ecp` package, we have

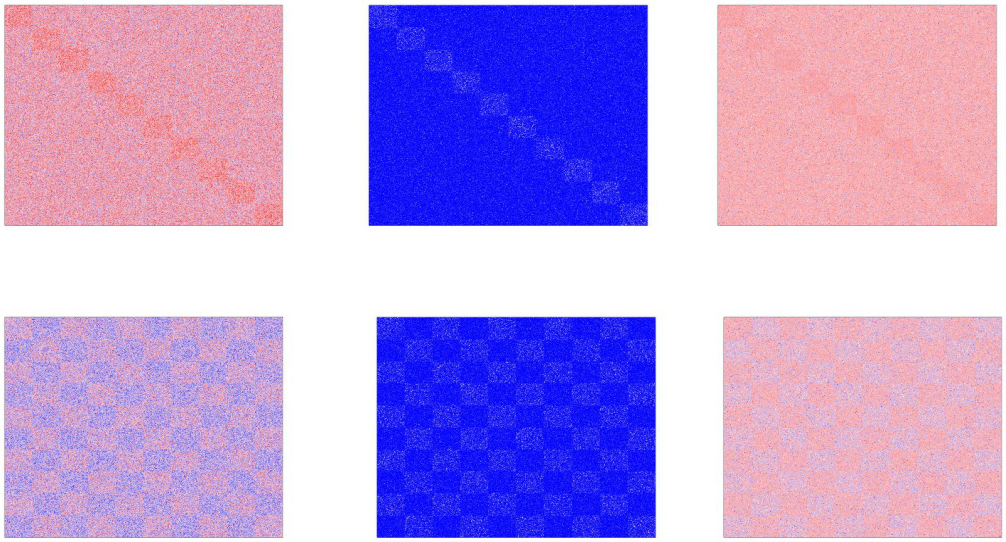

Figure 3: Examples of 400×400 matrices $\mathbf{X}$. Top: Block Diagonal configuration. Bottom: Chessboard configuration. Left: $\mathcal{L}_1 = \mathcal{N}(1,4)$, $\mathcal{L}_2 = \mathcal{N}(0,4)$, middle: $\mathcal{L}_1 = \mathcal{E}(2)$, $\mathcal{L}_2 = \mathcal{E}(1)$ and right: $\mathcal{L}_1 = C(1,1)$, $\mathcal{L}_2 = C(0,1)$.

to chose $\alpha \in (0,2]$ such that $\mathrm{E}(|\mathrm{X}|^{\alpha}) < \infty$. By default $\alpha = 1$ and we keep this value for the Gaussian and the exponential distributions but, for the Cauchy distribution, we need to have $\alpha < 1$; thus for this case we used $\alpha = 0.99$. These boxplots are obtained from 100 replications of $n \times n$ symmetric matrices where

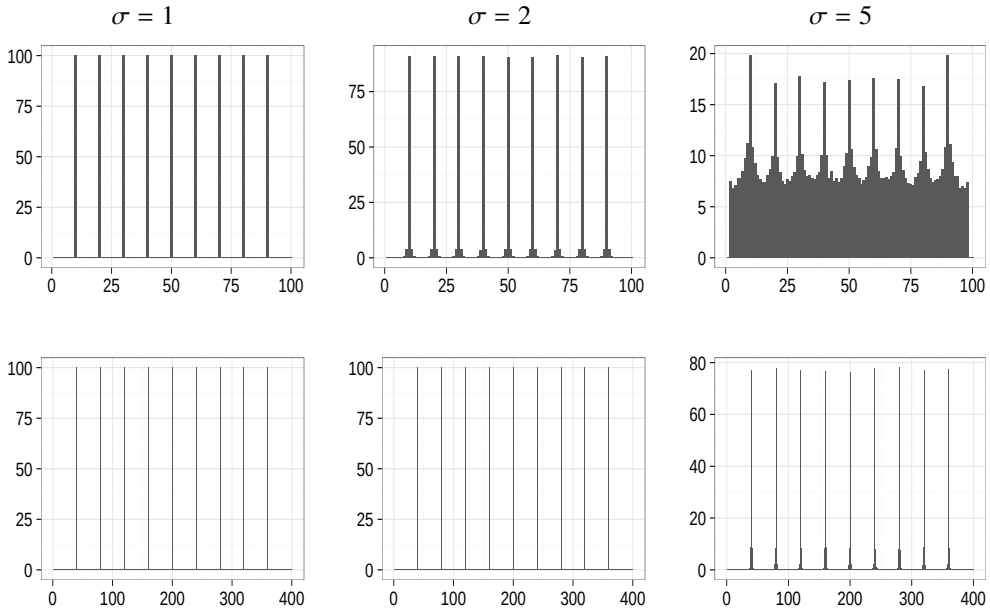

Figure 4: Barplots associated with the multiple change-point estimation procedure for $n = 100$ (top), $n = 400$ (bottom), $\mathcal{L}_1 = \mathcal{N}(1,\sigma^2)$ and $\mathcal{L}_2 = \mathcal{N}(0,\sigma^2)$ for different values of σ. The true positions of the change-points are located at the multiples of $n/10$.

$n \in \{50, 100, 200, 300, 400\}$. More precisely, the distance $\mathcal{D}$ is defined as

$$\mathcal{D}(\widehat{\mathbf{n}}, \mathbf{n}^*) = \frac{1}{n}\sqrt{\sum_{\ell=1}^{L}(\widehat{n}_\ell - n_\ell^*)^2}, \tag{13}$$

where $\mathbf{n}^* = (n_1^*, \ldots, n_L^*)$ denotes the vector of the true L change-point positions and $\widehat{\mathbf{n}} = (\widehat{n}_1, \ldots, \widehat{n}_L)$ its estimation either obtained by `MuChPoint` or `ecp`. Note that it actually corresponds to the usual ℓ_2-norm of the vector $\boldsymbol{\tau}^* - \widehat{\boldsymbol{\tau}}$ where $\boldsymbol{\tau}^* = (\tau_1^*, \ldots, \tau_L^*)$, $\widehat{\boldsymbol{\tau}} = (\widehat{\tau}_1, \ldots, \widehat{\tau}_L)$ with $n_\ell^* = \lfloor n\tau_\ell^* \rfloor$ and $\widehat{n}_\ell = \lfloor n\widehat{\tau}_\ell \rfloor$. In order to benchmark these methodologies, we provide to both of them the true value L of the number of change-points, which is here equal to 10.

We observe from Figures 5 and 6 that both approaches have similar statistical performance. However, `MuchPoint` performs better than `ecp` in the Cauchy case. In the Gaussian framework, the performance of `ecp` are a little bit better for small n and large σ.

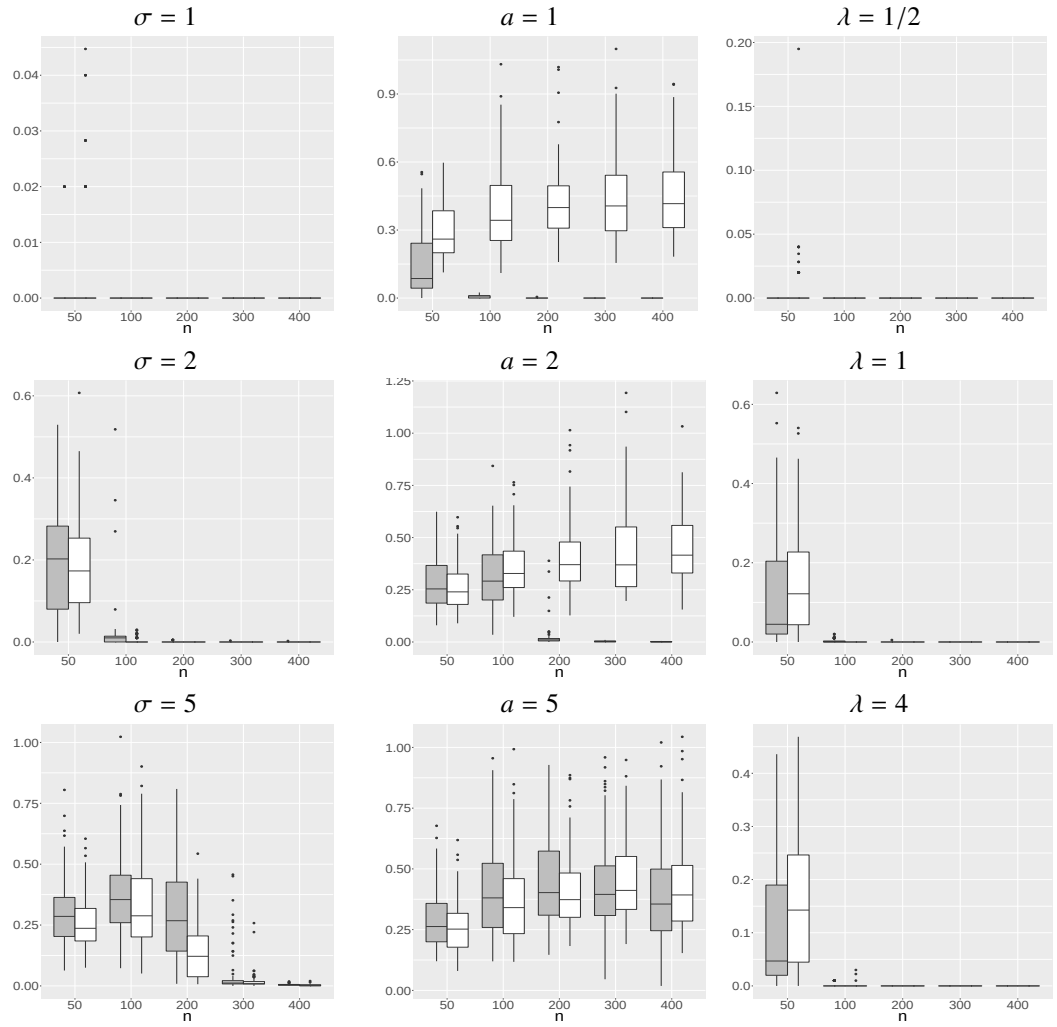

Figure 5: Boxplots of the distances $\mathcal{D}$ for `MuChPoint` and `ecp` in the Chessboard configuration. Left: $\mathcal{L}_1 = \mathcal{N}(1, \sigma^2)$, $\mathcal{L}_2 = \mathcal{N}(0, \sigma^2)$, middle: $\mathcal{L}_1 = \mathcal{C}(1, a)$, $\mathcal{L}_2 = \mathcal{C}(0, a)$ and right: $\mathcal{L}_1 = \mathcal{E}(2)$, $\mathcal{L}_2 = \mathcal{E}(\lambda)$ for different values of σ, λ and a. The boxplots associated with `MuChPoint` are displayed in gray and the ones of `ecp` in white.

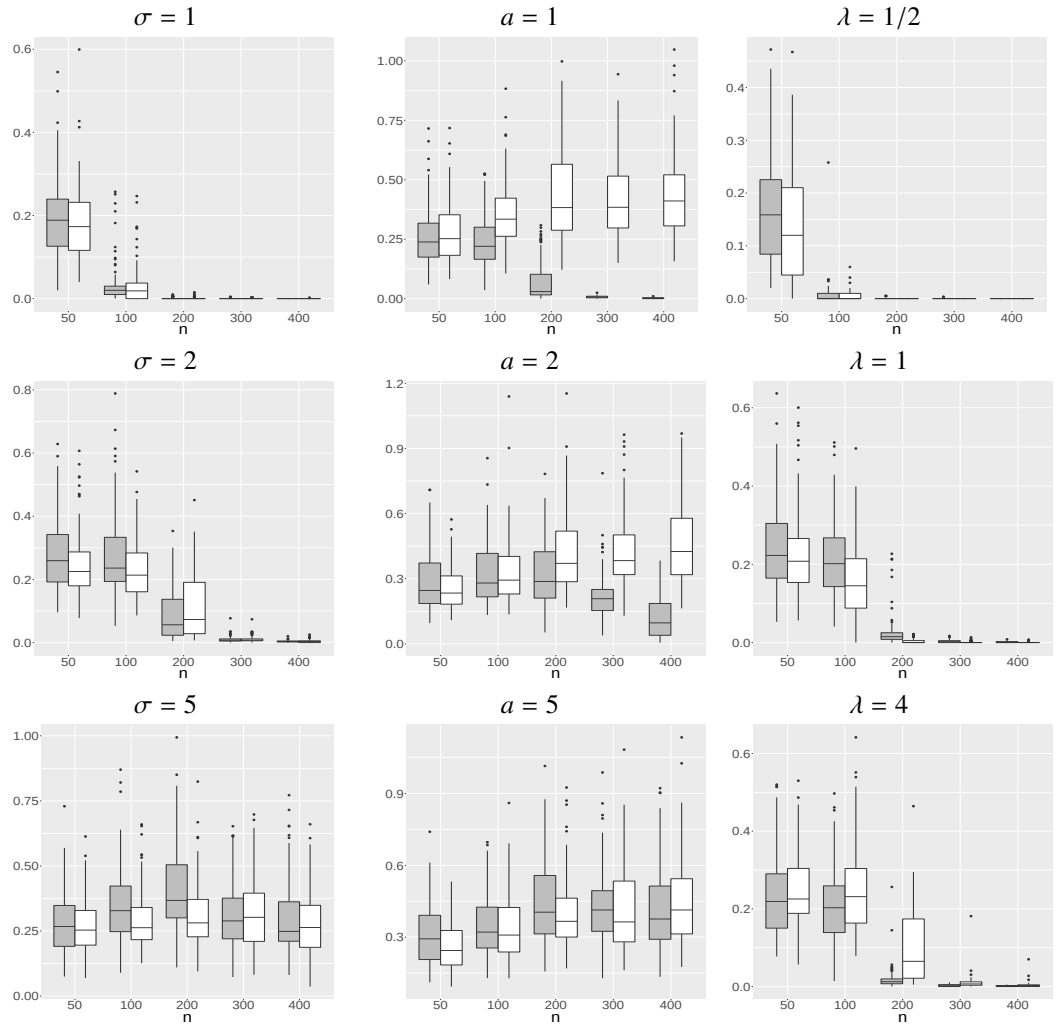

Figure 6: Boxplots of the distances $\mathcal{D}$ for `MuChPoint` and `ecp` in the Block Diagonal configuration. Left: $\mathcal{L}_1 = \mathcal{N}(1, \sigma^2)$, $\mathcal{L}_2 = \mathcal{N}(0, \sigma^2)$, middle: $\mathcal{L}_1 = C(1, a)$, $\mathcal{L}_2 = C(0, a)$ and right: $\mathcal{L}_1 = \mathcal{E}(2)$, $\mathcal{L}_2 = \mathcal{E}(\lambda)$ for different values of σ, λ and a. The boxplots associated with `MuChPoint` are displayed in gray and the ones of `ecp` in white.

5. Application to real data

In this section, we apply our methodology to publicly available Hi-C data (`http://chromosome.sdsc.edu/mouse/hi-c/download.html`) already studied by Dixon et al. [8]. This technology provides read pairs corresponding to pairs of genomic loci that physically interact in the nucleus;1 see [16] for further details. The raw measurements provided by Hi-C data is therefore a list of pairs of locations along the chromosome, at the nucleotide resolution. These measurements are often summarized by a symmetric matrix $\boldsymbol{X}$, where each entry $X_{i,j}$ corresponds the total number of read pairs matching in position i and position j, respectively. Positions refer here to a sequence of non-overlapping windows of equal sizes covering the genome. The number of windows may vary from one study to another; Lieberman-Aiden et al. [16] considered a Mb resolution, whereas Dixon et al. [8] went deeper and used windows of 40kb (called hereafter the resolution).

In the sequel, we analyze the interaction matrices of Chromosome 19 of the mouse cortex at a resolution 40 kb and we compare the location of the estimated change-points found by our approach with those obtained by Dixon et al. [8] on the same data since no ground truth is available. In this case, the matrix that has to be processed is a $n \times n$ symmetric matrix where $n = 1534$.

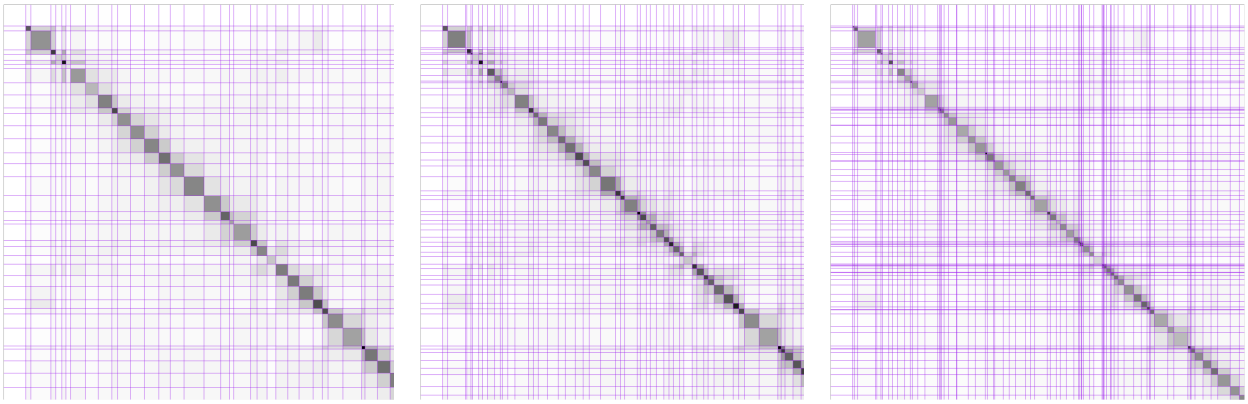

Figure 7: Estimated matrices $\widehat{\mathbf{X}}$ for different number of estimated change-points: 35 (left), 55 (middle) and 75 (right).

We display in Figure 7 the estimated matrix $\widehat{\mathbf{X}}$ obtained by using our strategy for various numbers of estimated change-points. This estimated matrix is a block-wise constant matrix for which the block boundaries are estimated by using MuChPoint and the values within each block correspond to the empirical mean of the observations lying in it. We can see from this figure that both the diagonal and the extra diagonal blocks are properly retrieved even when the number of estimated change-points is not that large.

In order to further compare our approach with the one proposed by Dixon et al. [8], we computed the two parts of the Hausdorff distance which is defined by

$$d\left(\widehat{\boldsymbol{n}}_B, \widehat{\boldsymbol{n}}\right) = \max\left\{d_1\left(\widehat{\boldsymbol{n}}_B, \widehat{\boldsymbol{n}}\right), d_2\left(\widehat{\boldsymbol{n}}_B, \widehat{\boldsymbol{n}}\right)\right\}, \tag{14}$$

where $\widehat{\boldsymbol{n}}$ and $\widehat{\boldsymbol{n}}_B$ are the change-points found by our approach and [8], respectively. In (14),

$$d_1\left(\mathbf{a}, \mathbf{b}\right) = \sup_{b \in \mathbf{b}} \inf_{a \in \mathbf{a}} |a - b|, \quad d_2\left(\mathbf{a}, \mathbf{b}\right) = d_1\left(\mathbf{b}, \mathbf{a}\right).$$

More precisely, Figure 8 displays the boxplots of the d_1 and d_2 parts of the Hausdorff distance without taking the supremum in white and gray for different values of the estimated number of change-points, respectively.

For comparison purpose, we used the R package `Capushe` which implements a model selection approach based on the slope heuristics theory and described in [3]. It can be used here to estimate the number of change-points L. According to the outputs of this package which are given in Figure 9, L is estimated to be 40. The corresponding estimated matrix $\widehat{\mathbf{X}}$ is displayed in Figure 10.

When the number of estimated change-points considered in our methodology is on a par with the one of [8], i.e., equal to 85, the positions of the block boundaries are very close as displayed in Figure 11.

6. Conclusion

In this paper, we designed a novel nonparametric method for retrieving the block boundaries of non-overlapping blocks in large matrices modeled as symmetric matrices of random variables having their distribution changing from one block to the other. Our approach is implemented in the R package `MuChPoint` which will be available from the Comprehensive R Archive Network (CRAN). In the course of this study, we have shown that our method, inspired by a generalization of nonparametric multiple sample tests to multivariate data, has two main features which make it very attractive. First, it is a nonparametric approach which performs very well from a practical point of view. Second, its computational burden makes its use possible on large Hi-C data matrices.

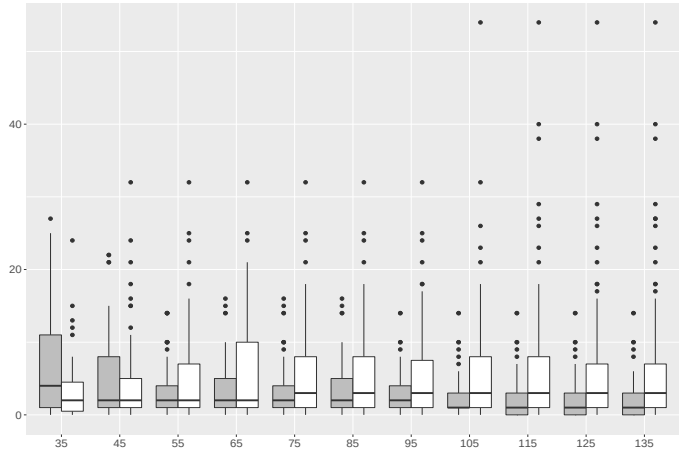

Figure 8: Boxplots for the infimum parts of the Hausdorff distances d_1 (white) and d_2 (gray) between the change-points found by [8] and our approach for different values of the estimated number of change-points.

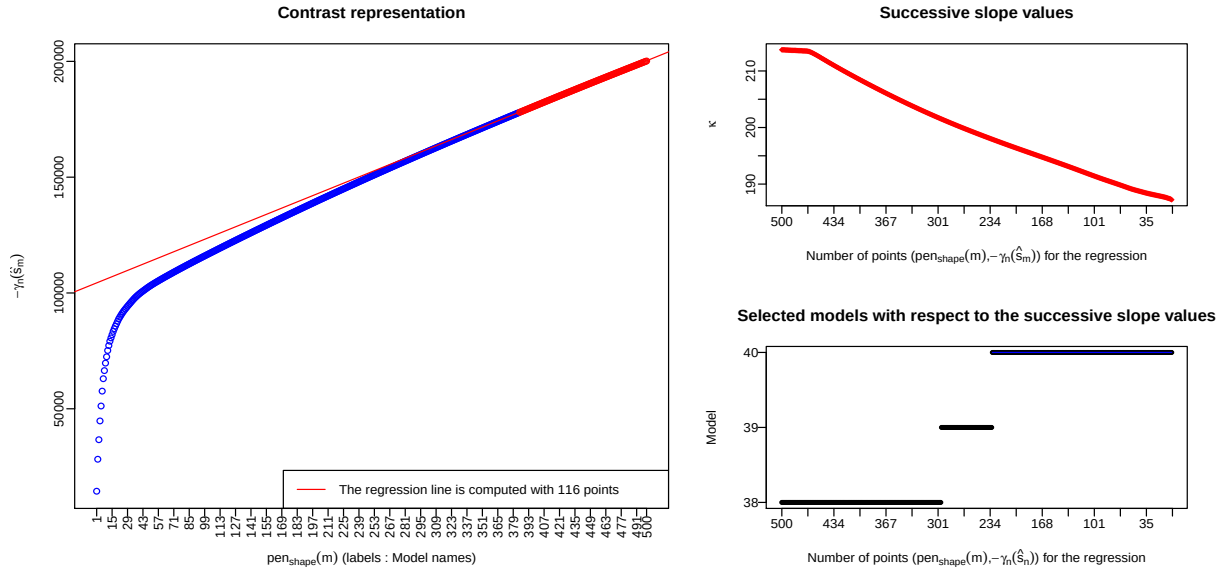

Figure 9: Outputs of the R package `Capushe`.

7. Proofs

In this section, we prove Theorems 1–3 and Eq. (12). We also give a sketch of the proof of Theorem 4. The proofs of the theorems given below use technical lemmas established in Section 7.5.

7.1. Proof of Theorem 1

To prove Theorem 1, we first compute the expectation of $S_n(n_1)$, viz.

$$\mathrm{E}\{S_n(n_1)\} = \sum_{i=1}^{n} \mathrm{E}\{U_{n,i}^2(n_1)\} = \frac{1}{nn_1(n-n_1)} \sum_{i=1}^{n} \mathrm{E}\left[\left\{\sum_{j_0=1}^{n_1} \sum_{j_1=n_1+1}^{n} h(X_{i,j_0}, X_{i,j_1})\right\}^2\right]$$

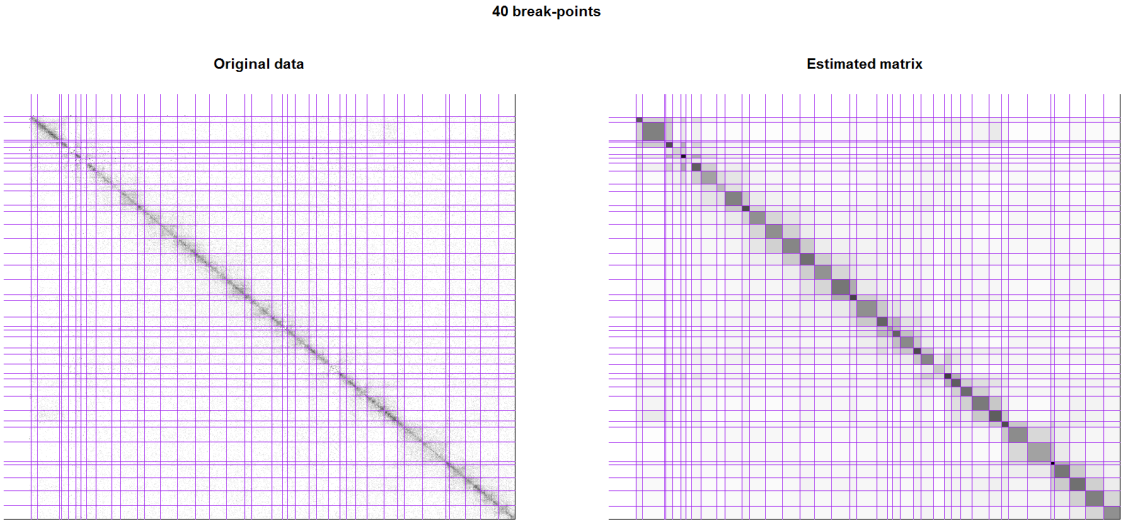

Figure 10: Estimated matrix $\widehat{\mathbf{X}}$ when L is estimated by using the R package `Capushe`.

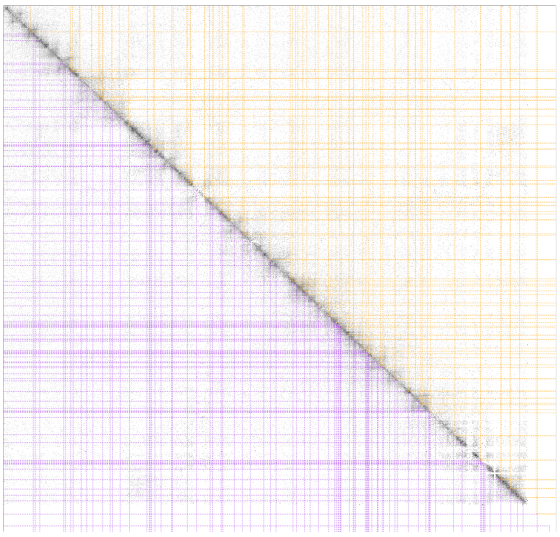

Figure 11: Topological domains detected by Dixon et al. [8] (upper triangular part of the matrix) and by our method (lower triangular part of the matrix).

$$= \frac{1}{nn_1(n-n_1)} \sum_{i=1}^{n} \sum_{1 \le j_0, k_0 \le n_1} \sum_{n_1+1 \le j_1, k_1 \le n} \mathrm{E}\{h(X_{i,j_0}, X_{i,j_1}) h(X_{i,k_0}, X_{i,k_1})\},$$

which can be expanded as follows:

$$\frac{1}{nn_1(n-n_1)} \sum_{i=1}^{n} \left[\sum_{j_0=1}^{n_1} \sum_{j_1=n_1+1}^{n} \mathrm{E}\{h^2(X_{i,j_0}, X_{i,j_1})\} + \sum_{j_0=1}^{n_1} \sum_{n_1+1 \le j_1 \ne k_1 \le n} \mathrm{E}\{h(X_{i,j_0}, X_{i,j_1}) h(X_{i,j_0}, X_{i,k_1})\} \right.$$

$$+ \sum_{1\le j_0\neq k_0\le n_1} \sum_{j_1=n_1+1}^{n} \mathrm{E}\{h(X_{i,j_0},X_{i,j_1})h(X_{i,k_0},X_{i,j_1})\} + \sum_{1\le j_0\neq k_0\le n_1} \sum_{n_1+1\le j_1\neq k_1\le n} \mathrm{E}\{h(X_{i,j_0},X_{i,j_1})h(X_{i,k_0},X_{i,k_1})\}\Bigg].$$

Using Lemma 1, we get

$$\mathrm{E}\{S_n(n_1)\} = \frac{1}{nn_1(n-n_1)} \sum_{i=1}^{n} \left\{ n_1(n-n_1) + \frac{1}{3}\, n_1(n-n_1)(n-n_1-1) + \frac{1}{3}\, n_1(n_1-1)(n-n_1) \right\} = \frac{n+1}{3}.$$

In order to derive the asymptotic behavior of $S_n(n_1)$ we write the centered version of $S_n(n_1)$ as

$$S_n(n_1) - \mathrm{E}\{S_n(n_1)\} = \frac{1}{nn_1(n-n_1)} \sum_{i=1}^{n} \left\{ \sum_{j_0=1}^{n_1} \sum_{j_1=n_1+1}^{n} h(X_{i,j_0},X_{i,j_1}) \right\}^2 - \frac{n+1}{3} \equiv \frac{1}{nn_1(n-n_1)} (A+B+C+D),$$

where

$$A = \sum_{i=1}^{n} \sum_{j_0=1}^{n_1} \sum_{j_1=n_1+1}^{n} \{h^2(X_{i,j_0},X_{i,j_1}) - 1\}, \quad B = \sum_{i=1}^{n} \sum_{j_0=1}^{n_1} \sum_{n_1+1\le j_1\neq k_1\le n} \{h(X_{i,j_0},X_{i,j_1})h(X_{i,j_0},X_{i,k_1}) - 1/3\},$$

$$C = \sum_{i=1}^{n} \sum_{1\le j_0\neq k_0\le n_1} \sum_{j_1=n_1+1}^{n} \{h(X_{i,j_0},X_{i,j_1})h(X_{i,k_0},X_{i,j_1}) - 1/3\},$$

$$D = \sum_{i=1}^{n} \sum_{1\le j_0\neq k_0\le n_1} \sum_{n_1+1\le j_1\neq k_1\le n} h(X_{i,j_0},X_{i,j_1})h(X_{i,k_0},X_{i,k_1})$$

in which each term is centered. First, we observe that $A = 0$ a.s. (almost surely) by Assertion (ii) of Lemma 1. Using Markov's inequality, we see that, for all $\varepsilon > 0$,

$$\begin{aligned}\Pr(|B/\sqrt{n}| > 6n^3/\varepsilon) &\le \varepsilon n^{-7/2}\mathrm{E}(|B|)/6 \\ &\le \frac{\varepsilon}{6n^{7/2}} \sum_{i=1}^{n} \mathrm{E}\left[\left| \sum_{j_0=1}^{n_1} \sum_{n_1+1\le j_1\neq k_1\le n} \{h(X_{i,j_0},X_{i,j_1})h(X_{i,j_0},X_{i,k_1}) - 1/3\} \right| \right].\end{aligned}$$

Using the Cauchy–Schwarz inequality, we thus deduce that

$$\begin{aligned}&\Pr(|B/\sqrt{n}| > 6n^3/\varepsilon) \\ &\quad\le \frac{\varepsilon}{6n^{7/2}} \sum_{i=1}^{n} \left\{ \mathrm{E}\left(\left[\sum_{j_0=1}^{n_1} \sum_{n_1+1\le j_1\neq k_1\le n} \{h(X_{i,j_0},X_{i,j_1})h(X_{i,j_0},X_{i,k_1}) - 1/3\} \right]^2 \right) \right\}^{1/2} \\ &\quad= \frac{\varepsilon}{6n^{7/2}} \sum_{i=1}^{n} \Bigg(\sum_{1\le j_0,j_0'\le n_1} \sum_{n_1+1\le j_1\neq k_1\le n} \sum_{n_1+1\le j_1'\neq k_1'\le n} \mathrm{E}\Big[\{h(X_{i,j_0},X_{i,j_1})h(X_{i,j_0},X_{i,k_1}) - 1/3\} \\ &\qquad\qquad \times \{h(X_{i,j_0'},X_{i,j_1'})h(X_{i,j_0'},X_{i,k_1'}) - 1/3\} \Big] \Bigg)^{1/2}.\end{aligned}$$

By Assertion (iii) of Lemma 1, the above expectation is equal to zero when the cardinality of the set of indices $\{j_0, j_0', j_1, j_1', k_1, k_1'\}$ equals 6. Indeed, the right- and left-hand side of the product in the expectation are independent in that case. Thus, only the cases where the cardinality of the set is at most 5 need be considered. Moreover, note that, for all $x, y, z, t, x', y', z', t'$, $|\{h(x,y)h(z,t) - 1/3\} \times \{h(x',y')h(z',t') - 1/3\}| \le 16/9 \le 2$. Hence we get that, for all $\varepsilon > 0$,

$$\Pr(|B/\sqrt{n}| > 6n^3/\varepsilon) \le \frac{\varepsilon}{6n^{7/2}} \sum_{i=1}^{n} 2n^{5/2} = \varepsilon/3. \tag{15}$$

Using similar arguments, we find that, for all $\varepsilon > 0$,

$$\Pr(|C/\sqrt{n}| > 6n^3/\varepsilon) \leq \varepsilon/3. \tag{16}$$

Using Markov's inequality and the Cauchy–Schwarz inequality as previously, we get that, for all $\varepsilon > 0$,

$$\begin{aligned}
\Pr(|D/\sqrt{n}| > 3n^3/\varepsilon) &\leq \varepsilon n^{-7/2}\mathrm{E}\left(|D|\right)/3 \\
&\leq \frac{\varepsilon}{3n^{7/2}}\mathrm{E}\left\{\left|\sum_{i=1}^{n}\sum_{1\leq j_0\neq k_0\leq n_1}\sum_{n_1+1\leq j_1\neq k_1\leq n} h(X_{i,j_0},X_{i,j_1})h(X_{i,k_0},X_{i,k_1})\right|\right\} \\
&\leq \frac{\varepsilon}{3n^{7/2}}\left(\mathrm{E}\left[\left\{\sum_{i=1}^{n}\sum_{1\leq j_0\neq k_0\leq n_1}\sum_{n_1+1\leq j_1\neq k_1\leq n} h(X_{i,j_0},X_{i,j_1})h(X_{i,k_0},X_{i,k_1})\right\}^2\right]\right)^{1/2} \\
&= \frac{\varepsilon}{3n^{7/2}}\Big[\mathrm{E}\Big\{\sum_{i=1}^{n}\sum_{1\leq j_0\neq k_0\leq n_1}\sum_{n_1+1\leq j_1\neq k_1\leq n} h(X_{i,j_0},X_{i,j_1})h(X_{i,k_0},X_{i,k_1}) \\
&\qquad\qquad \times\sum_{i'=1}^{n}\sum_{1\leq j_0'\neq k_0'\leq n_1}\sum_{n_1+1\leq j_1'\neq k_1'\leq n} h(X_{i',j_0'},X_{i',j_1'})h(X_{i',k_0'},X_{i',k_1'})\Big\}\Big]^{1/2}.
\end{aligned}$$

The above expectation is equal to zero when the cardinality of $\{i, i', j_0, j_0', k_0, k_0', j_1, j_1', k_1, k_1'\}$ is between 8 and 10 inclusively by Assertion (v) of Lemma 1. Only the cases where the cardinality of the set is at most 7 need be considered. Observe moreover that, for all $x, y, z, t, x', y', z', t' \in \mathbb{R}$, $|h(x,y)h(z,t)h(x',y')h(z',t')| \leq 1$. Therefore, for all $\varepsilon > 0$, we get

$$\Pr(|D/\sqrt{n}| > 3n^3/\varepsilon) \leq \frac{\varepsilon}{3n^{7/2}} \times n^{7/2} = \varepsilon/3. \tag{17}$$

Finally, by combining (15), (16) and (17), we obtain that, for all $\varepsilon > 0$,

$$\Pr\left[\frac{|S_n(n_1) - \mathrm{E}\{S_n(n_1)\}|}{\sqrt{n}} > \frac{15n^2}{\varepsilon n_1(n-n_1)}\right] \leq \varepsilon.$$

Since we assumed that $n_1/n \to \tau_1$ as $n \to \infty$, we get that $[S_n(n_1) - \mathrm{E}\{S_n(n_1)\}]/\sqrt{n} = O_{\mathbb{P}}(1)$, which concludes the proof of Theorem 1. □

7.2. Proof of Theorem 2

Start with the computation of $\mathrm{E}\{S_n(n_1,\ldots,n_L)\}$. First observe that, for any $i \in \{1,\ldots,n\}$ and $\ell \in \{0,\ldots,L\}$,

$$\left(\overline{R}_\ell^{(i)} - \frac{n+1}{2}\right)^2 = \left(\frac{1}{n_{\ell+1}-n_\ell}\sum_{j=n_\ell+1}^{n_{\ell+1}} R_j^{(i)} - \frac{n+1}{2}\right)^2 = \frac{1}{(n_{\ell+1}-n_\ell)^2}\left(\sum_{j=n_\ell+1}^{n_{\ell+1}} A_j^{(i)} + \sum_{n_\ell+1\leq j\neq j'\leq n_{\ell+1}} B_{jj'}^{(i)}\right), \tag{18}$$

where

$$A_j^{(i)} = \left(R_j^{(i)} - \frac{n+1}{2}\right)^2, \quad B_{jj'}^{(i)} = \left(R_j^{(i)} - \frac{n+1}{2}\right)\left(R_{j'}^{(i)} - \frac{n+1}{2}\right).$$

Using Definition (6) of $R_j^{(i)}$, we find

$$A_j^{(i)} = = \left\{\sum_{\substack{k=1\\k\neq j}}^{n}\left(\mathbf{1}_{\{X_{i,k}\leq X_{i,j}\}} - \frac{1}{2}\right)\right\}^2 = \sum_{\substack{k=1\\k\neq j}}^{n} g(X_{i,k},X_{i,j})^2 + \sum_{\substack{k=1\\k\neq j}}^{n}\sum_{\substack{k'=1\\k'\neq k\\k\neq j}}^{n} g(X_{i,k},X_{i,j})g(X_{i,k'},X_{i,j}), \tag{19}$$

where $g(x,y) = \mathbf{1}_{x\leq y} - 1/2$ and, by Assertions (i) and (ii) of Lemma 2, we get

$$\mathrm{E}(A_j^{(i)}) = \frac{1}{4}(n-1) + \frac{1}{12}(n-1)(n-2) = \frac{(n-1)(n+1)}{12}. \tag{20}$$

Then, we decompose $B^{(i)}_{jj'}$ into four terms as follows:

$$\begin{aligned} B^{(i)}_{jj'} &= \left(R^{(i)}_j - \frac{n+1}{2}\right)\left(R^{(i)}_{j'} - \frac{n+1}{2}\right) = \left\{\sum_{\substack{k=1\\k\neq j}}^{n}\left(\mathbf{1}_{\{X_{i,k}\leq X_{i,j}\}} - \frac{1}{2}\right)\right\}\left\{\sum_{\substack{k'=1\\k'\neq j'}}^{n}\left(\mathbf{1}_{\{X_{i,k'}\leq X_{i,j'}\}} - \frac{1}{2}\right)\right\} \\ &= \sum_{\substack{k=1\\k\neq j}}^{n}\sum_{\substack{k'=1\\k'\neq j'}}^{n} g(X_{i,k}, X_{i,j})g(X_{i,k'}, X_{i,j'}) \equiv B_1 + B_2 + B_3 + B_4, \end{aligned} \tag{21}$$

where

$$B_1 = g(X_{i,j'}, X_{i,j})g(X_{i,j}, X_{i,j'}), \quad B_2 = \sum_{\substack{k=1\\k\neq j\\k\neq j'}}^{n} g(X_{i,k}, X_{i,j})g(X_{i,j}, X_{i,j'}),$$

$$B_3 = \sum_{\substack{k'=1\\k'\neq j'\\k'\neq j}}^{n} g(X_{i,j'}, X_{i,j})g(X_{i,k'}, X_{i,j'}), \quad B_4 = \sum_{\substack{k=1\\k\neq j\\k\neq j'}}^{n}\sum_{\substack{k'=1\\k'\neq j'\\k'\neq j}}^{n} g(X_{i,k}, X_{i,j})g(X_{i,k'}, X_{i,j'})$$

We deduce from Lemma 2 that $\mathrm{E}(B_1) = -1/4$, $\mathrm{E}(B_2) = \mathrm{E}(B_3) = -(n-2)/12$ and $\mathrm{E}(B_4) = (n-2)/12$, since all terms in the sum defining B_4 have zero expectation except when $k = k'$. Hence,

$$\mathrm{E}(B^{(i)}_{jj'}) = -\frac{1}{4} - 2\times\frac{n-2}{12} + \frac{n-2}{12} = -\frac{1}{4} - \frac{n-2}{12} = -\frac{n+1}{12}. \tag{22}$$

By (18), (20) and (22),

$$\begin{aligned} \mathrm{E}\left\{\left(\overline{R}^{(i)}_\ell - \frac{n+1}{2}\right)^2\right\} &= \frac{1}{(n_{\ell+1}-n_\ell)^2}\left\{\sum_{j=n_\ell+1}^{n_{\ell+1}}\frac{(n-1)(n+1)}{12} - \sum_{n_\ell+1\leq j\neq j'\leq n_{\ell+1}}\frac{(n+1)}{12}\right\} \\ &= \frac{1}{(n_{\ell+1}-n_\ell)}\frac{(n-1)(n+1)}{12} - \frac{(n_{\ell+1}-n_\ell)(n_{\ell+1}-n_\ell-1)}{(n_{\ell+1}-n_\ell)^2}\times\frac{(n+1)}{12} \\ &= \frac{1}{(n_{\ell+1}-n_\ell)}\left\{\frac{(n-1)(n+1)}{12} - \frac{(n+1)(n_{\ell+1}-n_\ell-1)}{12}\right\}. \end{aligned}$$

By (5), we get that

$$\begin{aligned} \mathrm{E}\{S_n(n_1,\ldots,n_L)\} &= \frac{4}{n^2}\sum_{\ell=0}^{L}(n_{\ell+1}-n_\ell)\sum_{i=1}^{n}\mathrm{E}\left\{\left(\overline{R}^{(i)}_\ell - \frac{n+1}{2}\right)^2\right\} \\ &= \frac{4}{n}\sum_{\ell=0}^{L}\left\{\frac{(n-1)(n+1)}{12} - \frac{(n+1)(n_{\ell+1}-n_\ell-1)}{12}\right\} \\ &= \frac{4(n+1)}{12n}\{(L+1)(n-1) - (n-L-1)\} = \frac{L(n+1)}{3}. \end{aligned}$$

Now we focus on the asymptotic behavior of $S_n(n_1,\ldots,n_L)$. For this, we first write

$$S_n(n_1,\ldots,n_L) - \mathrm{E}\{S_n(n_1,\ldots,n_L)\} = \frac{4}{n^2}\sum_{\ell=0}^{L}(n_{\ell+1}-n_\ell)\sum_{i=1}^{n}\left(\overline{R}^{(i)}_\ell - \frac{n+1}{2}\right)^2 - \frac{L(n+1)}{3},$$

which we decompose as

$$\frac{4}{n^2}\sum_{\ell=0}^{L}\frac{1}{n_{\ell+1}-n_\ell}\sum_{i=1}^{n}\sum_{t=1}^{7}Z^{(t)}_i =$$

$$\frac{4}{n^2}\sum_{\ell=0}^{L}(n_{\ell+1}-n_\ell)\sum_{i=1}^{n}\left(\frac{1}{(n_{\ell+1}-n_\ell)^2}\left[\sum_{j=n_\ell+1}^{n_{\ell+1}}\{A_j^{(i)}-\mathrm{E}(A_j^{(i)})\}+\sum_{n_\ell+1\le j\ne j'\le n_{\ell+1}}\{B_{jj'}^{(i)}-\mathrm{E}(B_{jj'}^{(i)})\}\right]\right),$$

where $A_j^{(i)}$ and $B_{jj'}^{(i)}$ are defined in (19) and (21), and the $Z_i^{(t)}$ are defined as follows:

$$\begin{aligned}
Z_i^{(1)} &= \sum_{j=n_\ell+1}^{n_{\ell+1}}\sum_{\substack{k=1\\k\ne j}}^{n}\left\{g(X_{i,k},X_{i,j})^2-\frac{1}{4}\right\},\\
Z_i^{(2)} &= \sum_{j=n_\ell+1}^{n_{\ell+1}}\sum_{\substack{k=1\\k\ne j}}^{n}\sum_{\substack{k'=1\\k'\ne k\\k'\ne j}}^{n}\left\{g(X_{i,k},X_{i,j})g(X_{i,k'},X_{i,j})-\frac{1}{12}\right\},\\
Z_i^{(3)} &= \sum_{n_\ell+1\le j\ne j'\le n_{\ell+1}}\left\{g(X_{i,j'},X_{i,j})g(X_{i,j},X_{i,j'})+\frac{1}{4}\right\},\\
Z_i^{(4)} &= \sum_{n_\ell+1\le j\ne j'\le n_{\ell+1}}\sum_{\substack{k=1\\k\ne j\\k\ne j'}}^{n}\left\{g(X_{i,k},X_{i,j})g(X_{i,j},X_{i,j'})+\frac{1}{12}\right\},\\
Z_i^{(5)} &= \sum_{n_\ell+1\le j\ne j'\le n_{\ell+1}}\sum_{\substack{k'=1\\k'\ne j'\\k'\ne j}}^{n}\left\{g(X_{i,j'},X_{i,j})g(X_{i,k'},X_{i,j'})+\frac{1}{12}\right\},\\
Z_i^{(6)} &= \sum_{n_\ell+1\le j\ne j'\le n_{\ell+1}}\sum_{\substack{k=1\\k\ne j\\k\ne j'}}^{n}\left\{g(X_{i,k},X_{i,j})g(X_{i,k},X_{i,j'})-\frac{1}{12}\right\},\\
Z_i^{(7)} &= \sum_{n_\ell+1\le j\ne j'\le n_{\ell+1}}\sum_{\substack{k=1\\k\ne j\\k\ne j'}}^{n}\sum_{\substack{k'=1\\k'\ne j'\\k'\ne j\\k'\ne k}}^{n}g(X_{i,k},X_{i,j})g(X_{i,k'},X_{i,j'}).
\end{aligned}$$

It follows that, for all $M>0$,

$$\begin{aligned}
\Pr\left[\left|\frac{S_n(n_1,\ldots,n_L)-\mathrm{E}\{S_n(n_1,\ldots,n_L)\}}{\sqrt{n}}\right|>M\right] &\le \sum_{\ell=0}^{L}\sum_{t=1}^{7}\Pr\left\{\frac{4}{n^2}\frac{1}{n_{\ell+1}-n_\ell}\left|\sum_{i=1}^{n}Z_i^{(t)}\right|>\frac{M\sqrt{n}}{7(L+1)}\right\}\\
&\le \sum_{\ell=0}^{L}\sum_{t=1}^{7}\Pr\left\{\left|\sum_{i=1}^{n}Z_i^{(t)}\right|>\frac{M(n_{\ell+1}-n_\ell)n^{5/2}}{28(L+1)}\right\}.
\end{aligned}$$

Using Markov's inequality, we deduce that

$$\Pr\left[\left|\frac{S_n(n_1,\ldots,n_L)-\mathrm{E}\{S_n(n_1,\ldots,n_L)\}}{\sqrt{n}}\right|>M\right]\le\sum_{\ell=0}^{L}\sum_{t=1}^{7}\frac{28(L+1)}{M(n_{\ell+1}-n_\ell)n^{5/2}}\mathrm{E}\left(\left|\sum_{i=1}^{n}Z_i^{(t)}\right|\right).$$

Calling on the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}
&\Pr\left[\left|\frac{S_n(n_1,\ldots,n_L)-\mathrm{E}\{S_n(n_1,\ldots,n_L)\}}{\sqrt{n}}\right|>M\right]\\
&\qquad\le\sum_{\ell=0}^{L}\sum_{t=1}^{7}\frac{28(L+1)}{M(n_{\ell+1}-n_\ell)n^{5/2}}\left[\mathrm{E}\left\{\left(\sum_{i=1}^{n}Z_i^{(t)}\right)^2\right\}\right]^{1/2}.
\end{aligned}$$

We now bound $\mathrm{E}\{(\sum_{i=1}^n Z_i^{(t)})^2\}$ for each $t \in \{1, \ldots, 7\}$. First, Assertion (i) of Lemma 2 implies

$$\mathrm{E}\left\{\left(\sum_{i=1}^n Z_i^{(1)}\right)^2\right\} = \sum_{i=1}^n \sum_{i'=1}^n \mathrm{E}(Z_i^{(1)} Z_{i'}^{(1)})$$
$$= \sum_{i=1}^n \sum_{i'=1}^n \sum_{j=n_\ell+1}^{n_{\ell+1}} \sum_{\substack{k=1 \\ k\neq j}}^n \sum_{r=n_\ell+1}^{n_{\ell+1}} \sum_{\substack{s=1 \\ s\neq r}}^n \mathrm{E}\left[\left\{g(X_{i,k}, X_{i,j})^2 - \frac{1}{4}\right\}\left\{g(X_{i's}, X_{i'r})^2 - \frac{1}{4}\right\}\right] = 0.$$

Then,

$$\mathrm{E}\left\{\left(\sum_{i=1}^n Z_i^{(2)}\right)^2\right\} = \sum_{i=1}^n \sum_{i'=1}^n \mathrm{E}(Z_i^{(2)} Z_{i'}^{(2)})$$
$$= \sum_{i=1}^n \sum_{i'=1}^n \sum_{j=n_\ell+1}^{n_{\ell+1}} \sum_{\substack{k=1 \\ k\neq j}}^n \sum_{\substack{k'=1 \\ k'\neq k \\ k\neq j}}^n \sum_{r=n_\ell+1}^{n_{\ell+1}} \sum_{\substack{s=1 \\ s\neq r}}^n \sum_{\substack{s'=1 \\ s'\neq s \\ s\neq r}}^n \mathrm{E}\left[\left\{g(X_{i,k}, X_{i,j})g(X_{i,k'}, X_{i,j}) - \frac{1}{12}\right\}\right.$$
$$\left.\times\left\{g(X_{i's}, X_{i'r})g(X_{i's'}, X_{i'r}) - \frac{1}{12}\right\}\right].$$

The above expectation is equal to zero when the cardinality of the set of indices $\{i, i', j, k, k', r, s, s'\}$ equals 8 by Assertion (ii) of Lemma 2. Hence, only the cases where the cardinality of this set is at most 7 need be considered. Since $|\{g(x,y)g(z,t) - 1/12\}\{g(x',y')g(z',t') - 1/12\}| \leq 1/9 \leq 1$ for all $x, y, z, t, x', y', z', t' \in \mathbb{R}$, we get

$$\mathrm{E}\left\{\left(\sum_{i=1}^n Z_i^{(2)}\right)^2\right\} \leq n^7.$$

Using similar arguments and Assertion (iii) of Lemma 2, we also find

$$\mathrm{E}\left\{\left(\sum_{i=1}^n Z_i^{(6)}\right)^2\right\} \leq n^7.$$

Calling on similar arguments as those used for bounding $\mathrm{E}\{(\sum_{i=1}^n Z_i^{(2)})^2\}$ and by Assertion (ii) of Lemma 2, we see that $\mathrm{E}\{g(X,Y)g(Y,Z)\} = -\mathrm{E}\{g(X,Y)g(Z,Y)\} = -1/12$. Hence,

$$\mathrm{E}\left\{\left(\sum_{i=1}^n Z_i^{(4)}\right)^2\right\} \leq n^7 \quad \text{and} \quad \mathrm{E}\left\{\left(\sum_{i=1}^n Z_i^{(5)}\right)^2\right\} \leq n^7.$$

Using Assertion (i) of Lemma 2, we obtain

$$\mathrm{E}\left\{\left(\sum_{i=1}^n Z_i^{(3)}\right)^2\right\} = \sum_{i=1}^n \sum_{i'=1}^n \mathrm{E}(Z_i^{(3)} Z_{i'}^{(3)})$$
$$= \sum_{i=1}^n \sum_{i'=1}^n \sum_{n_\ell+1\leq j\neq j'\leq n_{\ell+1}} \sum_{n_\ell+1\leq r\neq r'\leq n_{\ell+1}} \mathrm{E}\left[\left\{g(X_{i,j'}, X_{i,j})g(X_{i,j}, X_{i,j'}) + \frac{1}{4}\right\}\right.$$
$$\left.\times\left\{g(X_{i'r'}, X_{i'r})g(X_{i'r}, X_{i'r'}) + \frac{1}{4}\right\}\right] = 0.$$

Finally,

$$\mathrm{E}\left\{\left(\sum_{i=1}^n Z_i^{(7)}\right)^2\right\} = \sum_{i=1}^n \sum_{i'=1}^n \mathrm{E}(Z_i^{(7)} Z_{i'}^{(7)})$$

$$= \sum_{i=1}^{n}\sum_{i'=1}^{n} \sum_{n_\ell+1\le j\ne j'\le n_{\ell+1}} \sum_{\substack{k=1\\k\ne j\\k\ne j'}}^{n}\sum_{\substack{k'=1\\k'\ne j'\\k'\ne j\\k'\ne k}}^{n} \sum_{n_\ell+1\le r\ne r'\le n_{\ell+1}} \sum_{\substack{s=1\\s\ne r\\s\ne r'}}^{n}\sum_{\substack{s'=1\\s'\ne r'\\s'\ne r\\s'\ne s}}^{n} \mathrm{E}\{g(X_{i,k},X_{i,j})g(X_{i,k'},X_{i,j'})$$
$$g(X_{i's},X_{i'r})g(X_{i's'},X_{i'r'})\}.$$

The above expectation is zero when the the cardinality of the set of indices $\{i,i',j,j',k,k',r,r',s,s'\}$ is 8 or more by Assertion (i) of Lemma 2. Observe moreover that $|g(x,y)g(z,t)g(x',y')g(z',t')\| \le 1/16 \le 1$, for all $x,y,z,t,x',y',z',t' \in \mathbb{R}$. Therefore, we get,

$$\mathrm{E}\left\{\left(\sum_{i=1}^{n} Z_i^{(7)}\right)^2\right\} \le n^7.$$

Thus, we obtain that, for all $M > 0$,

$$\Pr\left[\left|\frac{S_n(n_1,\ldots,n_L) - \mathrm{E}\{S_n(n_1,\ldots,n_L)\}}{\sqrt{n}}\right| > M\right] \le \frac{1}{M}\sum_{\ell=0}^{L}\frac{5\times 28(L+1)n^{7/2}}{(n_{\ell+1}-n_\ell)n^{5/2}}.$$

Since for any ℓ, $n/(n_{\ell+1}-n_\ell)$ converges to $1/(\tau_{\ell+1}-\tau_\ell)$, the right-hand side of the above inequality tends to 0 when $M\to\infty$, which concludes the proof. □

7.3. Proof of Theorem 3

For all $\delta > 0$, let us define $C_{n_1^*,\delta} = \{n_1 \in \{1,\ldots,n-1\} : |n_1 - n_1^*| \ge n\delta\}$. Note that

$$\begin{aligned}
&\Pr\left(|\hat{n}_1 - n_1^*| \ge n\delta\right)\\
&\quad\le \Pr\left[\max_{n_1\in C_{n_1^*,\delta}}\{S_n(n_1) - S_n(n_1^*)\} \ge 0\right]\\
&\quad\le \Pr\left[\max_{n_1\in C_{n_1^*,\delta}}[S_n(n_1) - S_n(n_1^*) - \mathrm{E}\{S_n(n_1) - S_n(n_1^*)\} + \mathrm{E}\{S_n(n_1) - S_n(n_1^*)\}] \ge 0\right]\\
&\quad\le \Pr\left[\max_{n_1\in C_{n_1^*,\delta}}[S_n(n_1) - S_n(n_1^*) - \mathrm{E}\{S_n(n_1) - S_n(n_1^*)\}] \ge -\max_{n_1\in C_{n_1^*,\delta}}\{\mathrm{E}\{S_n(n_1) - S_n(n_1^*)\}\}\right].
\end{aligned}$$

By Proposition 1 given below,

$$\Pr\left(|\hat{n}_1 - n_1^*| \ge n\delta\right) \le \Pr\left[\max_{n_1\in C_{n_1^*,\delta}}[S_n(n_1) - S_n(n_1^*) - \mathrm{E}\{S_n(n_1) - S_n(n_1^*)\}] \ge \kappa' n^2\delta\right]$$

for large enough n for some positive constant κ'. Hence,

$$\begin{aligned}
\Pr\left(|\hat{n}_1 - n_1^*| \ge n\delta\right) &\le \Pr\left[\max_{n_1\in C_{n_1^*,\delta}}|S_n(n_1) - S_n(n_1^*) - \mathrm{E}\{S_n(n_1) - S_n(n_1^*)\}| \ge \kappa' n^2\delta\right]\\
&\le \Pr\left[\max_{n_1\in C_{n_1^*,\delta}}|S_n(n_1) - \mathrm{E}\{S_n(n_1)\}| \ge \frac{\kappa'}{2}n^2\delta\right]\\
&\quad + \Pr\left[|S_n(n_1^*) - \mathrm{E}\{S_n(n_1^*)\}| \ge \frac{\kappa'}{2}n^2\delta\right]\\
&\le \sum_{n_1\in C_{n_1^*,\delta}}\Pr\left(\left|S_n(n_1) - \mathrm{E}\{S_n(n_1)\}\right| \ge \frac{\kappa'}{2}n^2\delta\right)\\
&\quad + \Pr\left[\left|S_n(n_1^*) - \mathrm{E}\{S_n(n_1^*)\}\right| \ge \frac{\kappa'}{2}n^2\delta\right].
\end{aligned}$$

We mimic the proof of Theorem 1 to obtain

$$\sum_{n_1\in C_{n_1^*,\delta}} \Pr\left[|S_n(n_1) - \mathrm{E}\{S_n(n_1)\}| \geq \frac{\kappa'}{2} n^2\delta\right] \leq \sum_{n_1\in C_{n_1^*,\delta}} \frac{2}{\kappa' n^2\delta}\mathrm{E}\left[|S_n(n_1) - \mathrm{E}\{S_n(n_1)\}|^2\right]^{1/2}$$
$$\leq \sum_{n_1\in C_{n_1^*,\delta}} \frac{2n^{7/2}}{\kappa' n^2\delta n n_1(n-n_1)} = \frac{2\sqrt{n}}{\kappa'\delta} \sum_{n_1\in C_{n_1^*,\delta}} \frac{1}{n_1(n-n_1)}.$$

Observing that

$$\sum_{n_1\in C_{n_1^*,\delta}} \frac{1}{n_1(n-n_1)} \leq \frac{1}{n}\left(\sum_{n_1=1}^{n-1}\frac{1}{n_1} + \sum_{n_1=1}^{n-1}\frac{1}{n-n_1}\right) = \frac{2}{n}\sum_{n_1=1}^{n-1}\frac{1}{n_1} \sim 2\,\frac{\ln(n)}{n},$$

as $n \to \infty$, we can conclude. □

It remains to show Proposition 1, used above to prove Theorem 3.

Proposition 1. *Under the assumptions of Theorem 3, there exists a positive constant κ, such that* $\mathrm{E}\{S_n(n_1) - S_n(n_1^*)\} = -\kappa n|n_1^* - n_1|\{1 + \varepsilon_n(n_1)\}$, *where* $\max_{n_1\in C_{n_1^*,\delta}} |\varepsilon_n(n_1)| \to 0$, *as* $n \to \infty$.

Proof. We first compute the expectation of $S_n(n_1^*)$, viz.

$$\mathrm{E}\{S_n(n_1^*)\} = \sum_{i=1}^{n} \mathrm{E}\{U_{n,i}^2(n_1^*)\} = \frac{1}{nn_1^*(n-n_1^*)} \sum_{i=1}^{n} \mathrm{E}\left[\left\{\sum_{j_0=1}^{n_1^*}\sum_{j_1=n_1^*+1}^{n} h(X_{i,j_0}, X_{i,j_1})\right\}^2\right]$$
$$= \frac{1}{nn_1^*(n-n_1^*)} \sum_{i=1}^{n} \sum_{1\leq j_0,k_0\leq n_1^*} \sum_{n_1^*+1\leq j_1,k_1\leq n} \mathrm{E}\{h(X_{i,j_0}, X_{i,j_1})h(X_{i,k_0}, X_{i,k_1})\}.$$

Rewrite the latter expression as

$$\frac{1}{nn_1^*(n-n_1^*)} \sum_{i=1}^{n}\left[\sum_{j_0=1}^{n_1^*}\sum_{j_1=n_1^*+1}^{n} \mathrm{E}\{h^2(X_{i,j_0}, X_{i,j_1})\} \quad + \sum_{j_0=1}^{n_1^*}\sum_{n_1^*+1\leq j_1\neq k_1\leq n} \mathrm{E}\{h(X_{i,j_0}, X_{i,j_1})h(X_{i,j_0}, X_{i,k_1})\}\right.$$
$$\left. + \sum_{1\leq j_0\neq k_0\leq n_1^*}\sum_{j_1=n_1^*+1}^{n} \mathrm{E}\{h(X_{i,j_0}, X_{i,j_1})h(X_{i,k_0}, X_{i,j_1})\} + \sum_{1\leq j_0\neq k_0\leq n_1^*}\sum_{n_1^*+1\leq j_1\neq k_1\leq n} \mathrm{E}\{h(X_{i,j_0}, X_{i,j_1})h(X_{i,k_0}, X_{i,k_1})\}\right]$$

and decompose this expression in the form

$$\mathrm{E}\{S_n(n_1^*)\} = \frac{A+B}{nn_1^*(n-n_1^*)},$$

where A corresponds to the sum over i which goes from 1 to n_1^* and B to the sum from n_1^*+1 to n. Introduce the independent random variables W, Y and Z, such that $W \sim \mathbb{P}_0^0$, $Y \sim \mathbb{P}_1^0 = \mathbb{P}_0^1$ and $Z \sim \mathbb{P}_1^1$ and denote $W^{(1)}, W^{(2)}, W^{(3)}, Y^{(1)}, Y^{(2)}, Y^{(3)}, Z^{(1)}, Z^{(2)}, Z^{(3)}$ their respective independent copies. Observe that

$$A = \sum_{i=1}^{n_1^*}\left[\sum_{j_0=1}^{n_1^*}\sum_{j_1=n_1^*+1}^{n} \mathrm{E}\{h^2(X_{i,j_0}, X_{i,j_1})\} + \sum_{j_0=1}^{n_1^*}\sum_{n_1^*+1\leq j_1\neq k_1\leq n} \mathrm{E}\{h(X_{i,j_0}, X_{i,j_1})h(X_{i,j_0}, X_{i,k_1})\}\right.$$
$$\left. + \sum_{1\leq j_0\neq k_0\leq n_1^*}\sum_{j_1=n_1^*+1}^{n} \mathrm{E}\{h(X_{i,j_0}, X_{i,j_1})h(X_{i,k_0}, X_{i,j_1})\} + \sum_{1\leq j_0\neq k_0\leq n_1^*}\sum_{n_1^*+1\leq j_1\neq k_1\leq n} \mathrm{E}\{h(X_{i,j_0}, X_{i,j_1})h(X_{i,k_0}, X_{i,k_1})\}\right]$$
$$= n_1^*\Big[n_1^*(n-n_1^*)\mathrm{E}\{h^2(W,Y)\} + n_1^*(n-n_1^*)(n-n_1^*-1)\mathrm{E}\{h(W,Y)h(W,Y^{(1)})\}$$
$$+ n_1^*(n_1^*-1)(n-n_1^*)\mathrm{E}\{h(W,Y)h(W^{(1)},Y)\} + n_1^*(n_1^*-1)(n-n_1^*)(n-n_1^*-1)\mathrm{E}\{h(W,Y)h(W^{(1)},Y^{(1)})\}\Big].$$

In the same manner, we can see that

$$B = (n-n_1^*)\Big[n_1^*(n-n_1^*)\mathrm{E}\{h^2(Y,Z)\} + n_1^*(n-n_1^*)(n-n_1^*-1)\mathrm{E}\{h(Y,Z)h(Y,Z^{(1)})\}$$
$$+ n_1^*(n_1^*-1)(n-n_1^*)\mathrm{E}\{h(Y,Z)h(Y^{(1)},Z)\} + n_1^*(n_1^*-1)(n-n_1^*)(n-n_1^*-1)\mathrm{E}\{h(Y,Z)h(Y^{(1)},Z^{(1)})\}\Big].$$

Note that all the absolute values of the above expectations in A and B are bounded by 1 by the definition of the function h. Then we compute the expectation of $S_n(n_1)$ in the case where $n_1 < n_1^*$, viz.

$$\mathrm{E}\{S_n(n_1)\} = \frac{1}{nn_1(n-n_1)}\sum_{i=1}^{n}\mathrm{E}\left[\left\{\sum_{j_0=1}^{n_1}\sum_{j_1=n_1+1}^{n}h(X_{i,j_0},X_{i,j_1})\right\}^2\right]$$
$$= \frac{1}{nn_1(n-n_1)}\sum_{i=1}^{n}\sum_{1\le j_0,k_0\le n_1}\sum_{n_1+1\le j_1,k_1\le n}\mathrm{E}\{h(X_{i,j_0},X_{i,j_1})h(X_{i,k_0},X_{i,k_1})\}.$$

Rewrite the latter expression as

$$\frac{1}{nn_1(n-n_1)}\sum_{i=1}^{n}\Bigg[\sum_{j_0=1}^{n_1}\sum_{j_1=n_1+1}^{n}\mathrm{E}\{h^2(X_{i,j_0},X_{i,j_1})\} + \sum_{j_0=1}^{n_1}\sum_{n_1+1\le j_1\ne k_1\le n}\mathrm{E}\{h(X_{i,j_0},X_{i,j_1})h(X_{i,j_0},X_{i,k_1}\}$$
$$+\sum_{1\le j_0\ne k_0\le n_1}\sum_{j_1=n_1+1}^{n}\mathrm{E}\{h(X_{i,j_0},X_{i,j_1})h(X_{i,k_0},X_{i,j_1})\}$$
$$+\sum_{1\le j_0\ne k_0\le n_1}\sum_{n_1+1\le j_1\ne k_1\le n}\mathrm{E}\{h(X_{i,j_0},X_{i,j_1})h(X_{i,k_0},X_{i,k_1})\}\Bigg],$$

and decompose this expression in the form

$$\mathrm{E}\{S_n(n_1)\} = \frac{C+D}{nn_1(n-n_1)},$$

where C corresponds to the sum over i which goes from 1 to n_1^* and D to the sum from n_1^*+1 to n. Next,

$$C = \sum_{i=1}^{n_1^*}\Bigg[\sum_{j_0=1}^{n_1}\sum_{j_1=n_1+1}^{n_1^*}\mathrm{E}\{h^2(X_{i,j_0},X_{i,j_1})\} + \sum_{j_0=1}^{n_1}\sum_{j_1=n_1^*+1}^{n}\mathrm{E}\{h^2(X_{i,j_0},X_{i,j_1})\}$$
$$+\sum_{j_0=1}^{n_1}\sum_{n_1+1\le j_1\ne k_1\le n_1^*}\mathrm{E}\{h(X_{i,j_0},X_{i,j_1})h(X_{i,j_0},X_{i,k_1})\} + \sum_{j_0=1}^{n_1}\sum_{n_1^*+1\le j_1\ne k_1\le n}\mathrm{E}\{h(X_{i,j_0},X_{i,j_1})h(X_{i,j_0},X_{i,k_1})\}$$
$$+2\sum_{j_0=1}^{n_1}\sum_{j_1=n_1+1}^{n_1^*}\sum_{k_1=n_1^*+1}^{n}\mathrm{E}\{h(X_{i,j_0},X_{i,j_1})h(X_{i,j_0},X_{i,k_1})\} + \sum_{1\le j_0\ne k_0\le n_1}\sum_{j_1=n_1+1}^{n_1^*}\mathrm{E}\{h(X_{i,j_0},X_{i,j_1})h(X_{i,k_0},X_{i,j_1})\}$$
$$+\sum_{1\le j_0\ne k_0\le n_1}\sum_{j_1=n_1^*+1}^{n}\mathrm{E}\{h(X_{i,j_0},X_{i,j_1})h(X_{i,k_0},X_{i,j_1})\} + \sum_{1\le j_0\ne k_0\le n_1}\sum_{n_1+1\le j_1\ne k_1\le n_1^*}\mathrm{E}\{h(X_{i,j_0},X_{i,j_1})h(X_{i,k_0},X_{i,k_1})\}$$
$$+\sum_{1\le j_0\ne k_0\le n_1}\sum_{n_1^*+1\le j_1\ne k_1\le n}\mathrm{E}\{h(X_{i,j_0},X_{i,j_1})h(X_{i,k_0},X_{i,k_1})\} + 2\sum_{1\le j_0\ne k_0\le n_1}\sum_{j_1=n_1+1}^{n_1^*}\sum_{k_1=n_1^*+1}^{n}\mathrm{E}\{h(X_{i,j_0},X_{i,j_1})h(X_{i,k_0},X_{i,k_1})\}\Bigg]$$

so that

$$C = n_1^*\Big[n_1(n_1^*-n_1)\mathrm{E}\{h^2(W,W^{(1)})\} + n_1(n-n_1^*)\mathrm{E}\{h^2(W,Y)\} + n_1(n_1^*-n_1)(n_1^*-n_1-1)\mathrm{E}\{h(W,W^{(1)})h(W,W^{(2)})\}$$
$$+ n_1(n-n_1^*)(n-n_1^*-1)\mathrm{E}\{h(W,Y)h(W,Y^{(1)})\} + 2n_1(n_1^*-n_1)(n-n_1^*)\mathrm{E}\{h(W,W^{(1)})h(W,Y)\}$$

$$
\begin{aligned}
&+ n_1(n_1-1)(n_1^*-n_1)\mathrm{E}\{h(W,W^{(1)})h(W^{(2)},W^{(1)})\} + n_1(n_1-1)(n-n_1^*)\mathrm{E}\{h(W,Y)h(W^{(1)},Y)\}\\
&+ n_1(n_1-1)(n_1^*-n_1)(n_1^*-n_1-1)\mathrm{E}\{h(W,W^{(1)})h(W^{(2)},W^{(3)})\}+\\
&+ n_1(n_1-1)(n-n_1^*)(n-n_1^*-1)\mathrm{E}\{h(W,Y)h(W^{(1)},Y^{(1)})\}\\
&+ 2n_1(n_1-1)(n_1^*-n_1)(n-n_1^*)\mathrm{E}\{h(W,W^{(1)})h(W^{(2)},Y)\}\Big].
\end{aligned}
$$

In the same manner, we can see that

$$
\begin{aligned}
D = (n-n_1^*)\Big[&n_1(n_1^*-n_1)\mathrm{E}\{h^2(Y,Y^{(1)}\} + n_1(n-n_1^*)\mathrm{E}\{h^2(Y,Z)\} + n_1(n_1^*-n_1)(n_1^*-n_1-1)\mathrm{E}\{h(Y,Y^{(1)})h(Y,Y^{(2)})\}\\
&+ n_1(n-n_1^*)(n-n_1^*-1)\mathrm{E}\{h(Y,Z)h(Y,Z^{(1)})\} + 2n_1(n_1^*-n_1)(n-n_1^*)\mathrm{E}\{h(Y,Y^{(1)})h(Y,Z)\}\\
&+ n_1(n_1-1)(n_1^*-n_1)\mathrm{E}\{h(Y,Y^{(1)})h(Y^{(2)},Y^{(1)}\} + n_1(n_1-1)(n-n_1^*)\mathrm{E}\{h(Y,Z)h(Y^{(1)},Z)\}\\
&+ n_1(n_1-1)(n_1^*-n_1)(n_1^*-n_1-1)\mathrm{E}\{h(Y,Y^{(1)})h(Y^{(2)},Y^{(3)})\}\\
&+ n_1(n_1-1)(n-n_1^*)(n-n_1^*-1)\mathrm{E}\{h(Y,Z)h(Y^{(1)},Z^{(1)})\}\\
&+ 2n_1(n_1-1)(n_1^*-n_1)(n-n_1^*)\mathrm{E}\{h(Y,Y^{(1)})h(Y^{(2)},Z)\}\Big],
\end{aligned}
$$

where the absolute values of the above expectations are bounded by 1. From Lemma 1, we get that

$$
\begin{aligned}
\mathrm{E}\{S_n(n_1)-S_n(n_1^*)\} &= \frac{C+D}{nn_1(n-n_1)} - \frac{A+B}{nn_1^*(n-n_1^*)}\\
&= \frac{(n_1^*-n_1)(n_1^*-2)}{3(n-n_1)} - \frac{n_1^*(n-n_1^*-1)(n_1^*-n_1)}{n(n-n_1)}\mathrm{E}\{h(W,Y)h(W,Y^{(1)})\}\\
&\quad + \frac{2n_1^*(n_1^*-n_1)(n-n_1^*)}{n(n-n_1)}\mathrm{E}\{h(W,W^{(1)})h(W,Y)\}\\
&\quad - \frac{n_1^*(n_1^*-n_1)(n-1)}{n(n-n_1)}\mathrm{E}\{h(W,Y)h(W^{(1)},Y)\}\\
&\quad - \frac{n_1^*(n-n_1^*-1)(n_1^*-n_1)(n-1)}{n(n-n_1)}\mathrm{E}\{h(W,Y)h(W^{(1)},Y^{(1)})\}\\
&\quad - \frac{(n-n_1^*)(n-n_1^*-1)(n_1^*-n_1)}{n(n-n_1)}\mathrm{E}\{h(Y,Z)h(Y,Z^{(1)})\}\\
&\quad + \frac{2(n-n_1^*)^2(n_1^*-n_1)}{n(n-n_1)}\mathrm{E}\{h(Y,Y^{(1)})h(Y,Z)\}\\
&\quad - \frac{(n-n_1^*)(n_1^*-n_1)(n-1)}{n(n-n_1)}\mathrm{E}\{h(Y,Z)h(Y^{(1)},Z)\}\\
&\quad - \frac{(n-n_1^*)(n-n_1^*-1)(n_1^*-n_1)(n-1)}{n(n-n_1)}\mathrm{E}\{h(Y,Z)h(Y^{(1)},Z^{(1)})\}.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\mathrm{E}\{S_n(n_1)-S_n(n_1^*)\} = -n(n_1^*-n_1)\Big[&\frac{n_1^*(n-n_1^*-1)(n-1)}{n^2(n-n_1)}\mathrm{E}\{h(W,Y)\}^2\\
&+ \frac{(n-n_1^*)(n-n_1^*-1)(n-1)}{n^2(n-n_1)}\mathrm{E}\{h(Y,Z)\}^2\\
&+ \frac{(2-n_1^*)}{3n(n-n_1)} + \frac{n_1^*(n-n_1^*-1)}{n^2(n-n_1)}\mathrm{E}\{h(W,Y)h(W,Y^{(1)})\}\\
&- 2\frac{n_1^*(n-n_1^*)}{n^2(n-n_1)}\mathrm{E}\{h(W,W^{(1)})h(W,Y)\} + \frac{n_1^*(n-1)}{n^2(n-n_1)}\mathrm{E}\{h(W,Y)h(W^{(1)},Y)\}\\
&+ \frac{(n-n_1^*)(n-n_1^*-1)}{n^2(n-n_1)}\mathrm{E}\{h(Y,Z)h(Y,Z^{(1)})\} - 2\frac{(n-n_1^*)^2}{n^2(n-n_1)}\mathrm{E}\{h(Y,Y^{(1)})h(Y,Z)\}\\
&+ \frac{(n-n_1^*)(n-1)}{n^2(n-n_1)}\mathrm{E}\{h(Y,Z)h(Y^{(1)},Z)\}\Big].
\end{aligned}
$$

Note that $\mathrm{E}\{h(W,Y)\} \neq 0$ or $\mathrm{E}\{h(Y,Z)\} \neq 0$ by (9) since

$$\mathrm{E}\{h(W,Y)\} = \mathrm{E}\left[\mathbf{1}_{\{W\leq Y\}} - \mathbf{1}_{\{Y\leq W\}}\right] = \Pr(W \leq Y) - \Pr(Y \leq W) = 2\Pr(W \leq Y) - 1 \neq 0.$$

Hence, there exists a positive constant κ, such that $\mathrm{E}\{S_n(n_1) - S_n(n_1^*)\} = -\kappa n(n_1^* - n_1)\{1 + \varepsilon_n(n_1)\}$, where $\max_{n_1 \in C_{n_1^*,\delta}} |\varepsilon_n(n_1)| \to 0$, as $n \to \infty$, since $1 - n_1^*/n \leq (n - n_1^*)/(n - n_1) \leq 1$ and $n_1^*/n \to \tau_1^*$.

We conclude the proof using similar arguments in the case where $n_1 > n_1^*$. □

7.4. Sketch of proof of Theorem 4

The proof of Theorem 4 is similar to that of Theorem 3. The main difference is the computation of the following expectation:

$$\begin{aligned}&\mathrm{E}\{S_n(n_1^*,\ldots,n_L^*) - S_n(n_1,\ldots,n_L)\}\\ &\quad= \frac{2}{n^2}\sum_{\ell=0}^{L}\sum_{\ell_1=0}^{L}\sum_{\ell_2=0}^{L}\sum_{\ell_4=0}^{L}\frac{n_{\ell,\ell_1}n_{\ell,\ell_2}|D_{\ell_4}^*|}{n_{\ell+1} - n_\ell}\left(\sum_{\ell_3=0}^{L}|D_{\ell_3}^*|\mathrm{E}\{F_{\ell_4,\ell_1}(X) - F_{\ell_4,\ell_2}(X)\}\right)^2\{1 + \varepsilon(n_1,\ldots,n_L)\},\end{aligned} \tag{23}$$

where $|A|$ denotes the cardinality of the set A, $X \sim \Pr_{\ell_3}^{\ell_4}$, $\sup_{(n_1,\ldots,n_L)}\varepsilon(n_1,\ldots,n_L) \to 0$ and

$$n_{\ell,\ell'} = |\{i \in \{1,\ldots,n\} : n_{\ell'}^* + 1 \leq i \leq n_{\ell'+1}^* \text{ and } n_\ell + 1 \leq i \leq n_{\ell+1}\}|.$$

Let $\delta > 0$. Define $C_{\boldsymbol{n}^*,\delta} = \{\boldsymbol{n} = (n_1,\ldots,n_L) : \|\boldsymbol{n} - \boldsymbol{n}^*\|_\infty \geq n\delta\}$. Using similar arguments as those given in (iii) of Lemma 1 in [5], under Assumption (10), there exists $\kappa > 0$ such that

$$\min_{\boldsymbol{n}\in C_{\boldsymbol{n}^*,\delta}}\mathrm{E}\{S_n(n_1^*,\ldots,n_L^*) - S_n(n_1,\ldots,n_L)\} \geq \kappa n^2 \tag{24}$$

for large enough n. The detailed proof of (23) and (24) is given in the Online Supplement. □

7.5. Proof of Eq. (12)

By (11), $I_0(p) = \max_{1\leq n_1=p}\Delta(1:n_1) = \Delta(1:p)$ and

$$I_1(p) = \max_{1\leq n_1<n_2=p}\{\Delta(1:n_1) + \Delta(n_1+1:p)\} = \max_{1\leq n_1<n_2=p}\{I_0(n_1) + \Delta(n_1+1:p)\},$$

which is (12) when $L = 1$. By (11),

$$I_2(p) = \max_{1\leq n_1<n_2<n_3=p}\{\Delta(1:n_1) + \Delta(n_1+1:n_2) + \Delta(n_2+1:p)\}.$$

Using the previous expression of $I_1(p)$, we get

$$I_2(p) = \max_{1<n_2<p}\{I_1(n_2) + \Delta(n_2+1:p)\},$$

which is (12) when $L = 2$. Following the same lines of reasoning, we can derive (12) for $L = 3, 4, \ldots$ □

Appendix: Technical lemmas

Lemma 1. *Let h be defined by $h(x,y) = \mathbf{1}_{\{x\leq y\}} - \mathbf{1}_{\{y\leq x\}}$. Then (i) $\mathrm{E}\{h(X,Y)\} = 0$; (ii) $h^2(X,Y) = 1$ a.s.; (iii) $\mathrm{E}\{h(X,Y)h(X,Z)\} = 1/3$; (iv) $\mathrm{E}\{h(X,Y)h(Z,Y)\} = 1/3$; (v) $\mathrm{E}\{h(X,Y)h(Z,T)\} = 0$; where X, Y, Z and T are iid random variables whose common distribution is continuous.*

Proof. (i) Let X and Y be iid random variables with cumulative distribution function F. We have $\mathrm{E}\{h(X,Y)\} = \mathrm{E}(\mathbf{1}_{\{X\leq Y\}}) - \mathrm{E}(\mathbf{1}_{\{Y\leq X\}}) = \mathrm{E}\{1 - 2F(X)\} = 0$, where we used that $F(X)$ is $\mathcal{U}[0,1]$. To prove (i), let, for all $x \neq y$ in $\mathbb{R}$, $h^2(x,y) = (\mathbf{1}_{\{x\leq y\}} - \mathbf{1}_{\{y\leq x\}})^2 = \mathbf{1}_{\{x\leq y\}} + \mathbf{1}_{\{y\leq x\}} - 2\mathbf{1}_{\{x\leq y\}}\mathbf{1}_{\{y\leq x\}} = 1$. Consequently, $h^2(X,Y) = 1$ a.s. Turning to (iii), let X, Y and Z be iid random variables with cumulative distribution function F. We have

$$\mathrm{E}\{h(X,Y)h(X,Z)\} = \mathrm{E}\{(\mathbf{1}_{\{X\leq Y\}} - \mathbf{1}_{\{Y\leq X\}})(\mathbf{1}_{\{X\leq Z\}} - \mathbf{1}_{\{Z\leq X\}})\}$$

$$
\begin{aligned}
&= \mathrm{E}\left(\mathbf{1}_{\{X\leq Y\}}\mathbf{1}_{\{X\leq Z\}}\right) - \mathrm{E}\left(\mathbf{1}_{\{X\leq Y\}}\mathbf{1}_{\{Z\leq X\}}\right) - \mathrm{E}\left(\mathbf{1}_{\{Y\leq X\}}\mathbf{1}_{\{X\leq Z\}}\right) + \mathrm{E}\left(\mathbf{1}_{\{Y\leq X\}}\mathbf{1}_{\{Z\leq X\}}\right) \\
&= \mathrm{E}[\{1-F(X)\}^2] - 2[\mathrm{E}\{F(X)\} - \mathrm{E}\{(X)^2\}] + \mathrm{E}\{F(X)^2\} \\
&= 1/3 - 2\,(1/2 - 1/3) + 1/3 = 1/3,
\end{aligned}
$$

where we used the fact that $F(X)$ is $\mathcal{U}[0,1]$. Statement (iv) stems from (iii), since $\mathrm{E}\{h(X,Y)h(Z,Y)\} = \mathrm{E}\{h(Y,X)h(Y,Z)\} = 1/3$. As for (v), by independence of (X,Y) with (Z,T), we have $\mathrm{E}\{h(X,Y)h(Z,T)\} = \mathrm{E}\{h(X,Y)\}\,\mathrm{E}\{h(Z,T)\} = 0$. □

Lemma 2. *Let us define the function g as $g(x,y) = \mathbf{1}_{\{x\leq y\}} - 1/2$. Let X, Y and Z be iid random variables whose common distribution is continuous. Then (i)* $\mathrm{E}\{g(X,Y)\} = 0$*; (ii)* $g(X,Y)^2 = 1/4$ *a.s.; (iii)* $\mathrm{E}\{g(X,Y)g(Z,Y)\} = 1/12$*; (iv)* $\mathrm{E}\{g(X,Y)g(X,Z)\} = 1/12$.

Proof. To show (i), note that $\mathrm{E}\{g(X,Y)\} = \mathrm{E}\{F(Y)\} - 1/2 = 0$ because $F(Y)$ is $\mathcal{U}[0,1]$. To show (ii), observe that for all x, y in $\mathbb{R}$, $g(x,y)^2 = (\mathbf{1}_{\{x\leq y\}} - 1/2)^2 = \mathbf{1}_{\{x\leq y\}} + 1/4 - \mathbf{1}_{\{x\leq y\}} = 1/4$. Thus $g^2(X,Y) = 1/4$ a.s. Turning to (iii), let X, Y and Z be iid random variables with distribution F. We have

$$
\begin{aligned}
\mathrm{E}\{g(X,Y)g(Z,Y)\} &= \mathrm{E}\left\{\left(\mathbf{1}_{\{X\leq Y\}} - \frac{1}{2}\right)\left(\mathbf{1}_{\{Z\leq Y\}} - \frac{1}{2}\right)\right\} = \mathrm{E}\left(\mathbf{1}_{\{X\leq Y\}}\mathbf{1}_{\{Z\leq Y\}}\right) - \frac{1}{2}\,\mathrm{E}\left(\mathbf{1}_{\{Z\leq Y\}}\right) - \frac{1}{2}\,\mathrm{E}\left(\mathbf{1}_{\{X\leq Y\}}\right) + \frac{1}{4} \\
&= \mathrm{E}\{F(Y)^2\} - \mathrm{E}\{F(Y)\} + \frac{1}{4} = \frac{1}{3} - \frac{1}{2} + \frac{1}{4} = \frac{1}{12},
\end{aligned}
$$

where we used that $F(X)$ is $\mathcal{U}[0,1]$. Finally, for (iv) note that, by (iii),

$$
\begin{aligned}
\mathrm{E}\{g(X,Y)g(X,Z)\} &= \mathrm{E}\left\{\left(\mathbf{1}_{\{X\leq Y\}} - \frac{1}{2}\right)\left(\mathbf{1}_{\{X\leq Z\}} - \frac{1}{2}\right)\right\} = \mathrm{E}\left\{\left(1 - \mathbf{1}_{\{Y\leq X\}} - \frac{1}{2}\right)\left(1 - \mathbf{1}_{\{Z\leq X\}} - \frac{1}{2}\right)\right\} \\
&= \mathrm{E}\left\{\left(\frac{1}{2} - \mathbf{1}_{\{Y\leq X\}}\right)\left(\frac{1}{2} - \mathbf{1}_{\{Z\leq X\}}\right)\right\} = \mathrm{E}\{g(Y,X)g(Z,X)\} = \frac{1}{12}.
\end{aligned}
$$

This concludes the proof. □

Acknowledgments

Vincent Brault would like to thank the Agence nationale de la recherche (ANR) for financial support through the ABS4NGS project (ANR-11-BINF-0001-06).

References

[1] J. Bai, Common breaks in means and variances for panel data, J. Econom. 157 (2010) 78–92.
[2] M. Basseville, I.V. Nikiforov, Detection of Abrupt Changes: Theory and Applications, Prentice-Hall, 1993.
[3] J.-P. Baudry, C. Maugis, B. Michel, Slope heuristics: Overview and implementation, Statist. Comput. 22 (2012) 455–470.
[4] R. Bellman, On the approximation of curves by line segments using dynamic programming, Commun. ACM 4 (1961) 284.
[5] V. Brault, M. Delattre, E. Lebarbier, T. Mary-Huard, C. Lévy-Leduc, Estimating the number of block boundaries from diagonal blockwise matrices without penalization, Scand. J. Statistics, 2017.
[6] H. Cho, P. Fryzlewicz, Multiple-change-point detection for high dimensional time series via sparsified binary segmentation, J. Roy. Statist. Soc. Ser. B 77 (2015) 475–507.
[7] A. Cleynen, S. Dudoit, S. Robin, Comparing segmentation methods for genome annotation based on RNA-seq data, J. Agric. Biol. Environ. Statist. 19 (2013) 101–118.
[8] J.R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J.S. Liu, B. Ren, Topological domains in mammalian genomes identified by analysis of chromatin interactions, Nature 485 (2012) 376–380.
[9] L. Horváth, M. Hušková, Change-point detection in panel data, J. Time Series Anal. 33 (2012) 631–648.
[10] M. Jirak, Uniform change point tests in high dimension, Ann. Statist. 43 (2015) 2451–2483.
[11] S. Kay, Fundamentals of Statistical Signal Processing: Detection Theory, Prentice-Hall, 1993.
[12] R. Killick, P. Fearnhead, I.A. Eckley, Optimal detection of changepoints with a linear computational cost, J. Amer. Statist. Assoc. 107 (2012) 1590–1598.
[13] E.L. Lehmann, H.J. D'Abrera, Nonparametrics: Statistical Methods Based on Ranks, Springer, New York, 2006.
[14] C. Lévy-Leduc, M. Delattre, T. Mary-Huard, S. Robin, Two-dimensional segmentation for analyzing HiC data, Bioinformatics 30 (2014) 386–392.
[15] C. Lévy-Leduc, F. Roueff, Detection and localization of change-points in high-dimensional network traffic data, Ann. Appl. Statist. 3 (2009) 637–662.

[16] E. Lieberman-Aiden, N.L. Van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B.R. Lajoie, P.J. Sabo, M.O. Dorschner, R. Sandstrom, B. Bernstein, M.A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L.A. Mirny, E.S. Lander, J. Dekker, Comprehensive mapping of long-range interactions reveals folding principles of the human genome, Science 326 (2009) 289–293.

[17] A. Lung-Yut-Fong, C. Lévy-Leduc, O. Cappé, Homogeneity and change-point detection tests for multivariate data using rank statistics, J. Soc. Fr. Statist. 156 (2015) 133–162.

[18] D.S. Matteson, N.A. James, A nonparametric approach for multiple change point analysis of multivariate data, J. Amer. Statist. Assoc. 109 (2014) 334–345.

[19] F. Picard, S. Robin, M. Lavielle, C. Vaisse, J.-J. Daudin, A statistical approach for array CGH data analysis, BMC Bioinformatics 6 (2005) 27.

[20] J.G. Szekely, L.M. Rizzo, Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method, J. Classif. 22 (2005) 151–183.

[21] A. Tartakovsky, B. Rozovskii, R. Blazek, H. Kim, A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods, IEEE Trans. Signal Process. 54 (2006) 3372–3382.

[22] A.W. van der Vaart, Asymptotic Statistics, Cambridge University Press, 1998.

[23] J. Vert, K. Bleakley, Fast detection of multiple change-points shared by many signals using group LARS, In: Adv. Neural Inform. Proc. Systems 23 (2010) 2343–2351.