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

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

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

Vincent Brault[1a,b], Sarah Ouadah[c], Laure Sansonnet[c], Céline Lévy-Leduc[c]

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

Abstract

We propose a novel nonparametric approach for estimating the location of block boundaries (change-points) of non-overlapping blocks in a random symmetric matrix which consists of random variables having their distribution changing from one block to the other. Our method is based on a nonparametric two-sample homogeneity test for matrices that we extend to the more general case of several groups. We first provide some theoretical results for the two associated test statistics and we explain how to derive change-point location estimators. Then, some numerical experiments are given in order to support our claims. Finally, our approach is applied to Hi-C data which are used in molecular biology for better understanding the influence of the chromosomal conformation on the cells functioning.

Keywords: Nonparametric tests, change-point estimation, Hi-C data

1. Introduction

Detecting and localizing changes in the distribution of random variables is a major statistical issue that arises in many fields such as the surveillance of industrial processes, see Basseville and Nikiforov (1993), the detection of anomalies in internet traffic data, see Tartakovsky et al. (2006) and Lévy-Leduc and Roueff (2009) or in molecular biology. In the latter field, several change-point detection methods have been designed for dealing with different kinds of data such as CNV (Copy Number Variation), see Picard et al. (2005) or Vert and Bleakley (2010), RNAseq data, see Cleynen et al. (2013) and more recently Hi-C data which motivated this work.

The Hi-C technology corresponds to one of the most recent chromosome conformation capture method that has been developed to better understand the influence of the chromosomal conformation on the cells functioning. This technology is based on a deep sequencing approach and provides read pairs corresponding to pairs of genomic loci that physically interacts in the nucleus, see Lieberman-Aiden et al. (2009). The raw measurements provided by Hi-C data are often summarized as a square matrix where each entry at row i and

[1] 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).

column j stands for the total number of read pairs matching in position i and position j, see Dixon et al. (2012) for further details. Blocks of different intensities arise among 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, which gives, as a by-product, a list of non-overlapping interacting chromosomic regions. It has to be noticed that this issue has already been addressed by Lévy-Leduc et al. (2014) 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 latter work, the authors use a parametric approach based on the maximization of the likelihood. In the following, we shall address the case where the non-overlapping blocks are not diagonal anymore by 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 having their distribution changing from one block to the other.

A large literature is dedicated to change-point detection and estimation in the very general mutivariate setting. To list but a few, Bai (2010) proposed a change-point estimation method in the case where both the number of observations and the number of vectors can go to infinity but with different rates. Horvath and Huskova (2012) 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 (2015) devised a parametric approach for identifying 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 (2015) proposed non parametric change-point tests in the very general high dimensional settings. Matteson and James (2014) devised a nonparametric change-point estimation procedure which allows them to retrieve change-points within n K-dimensional 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 introduced by Szekely and Rizzo (2005). Another approach based on ranks has also been proposed by Lung-Yut-Fong et al. (2015) in the same framework as Matteson and James (2014). More precisely, the approach proposed by Lung-Yut-Fong et al. (2015) consists in extending the classical Wilcoxon and Kruskal-Wallis statistics (Lehmann and D'Abrera (2006)) to the multivariate case.

In this paper, we propose a nonparametric change-point estimation approach based on nonparametric homogeneity tests generalizing the approach of Lung-Yut-Fong et al. (2015) to the case where we have to deal with large matrices instead of fixed multidimensional 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 go both to infinity, which is a very particular setting that has not been studied yet to the best of our knowledge.

The paper is organized as follows. We first propose in Sections 2.1 and 2.2 nonparametric homogeneity tests for two-samples and several samples, respectively. In Section 2.3,

we deduce from these tests a nonparametric procedure for estimating the block boundaries of a matrix of random variables having their distribution changing from one block to the other. These methodologies are then illustrated by some numerical experiments in Section 3. An application to real Hi-C data is also given in Section 4. Finally, the proofs of our theoretical results are given in Section 6.

2. Homogeneity tests and multiple change-point estimation

2.1. Two-sample homogeneity test

2.1.1. Statistical framework

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

Let n_1 be a given integer in $\{1, \ldots, n\}$. The goal of this section is to propose a statistic to test the null hypothesis (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 hypothesis (H_1): "$(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n_1)})$ has the distribution $\mathbb{P}_1$ and $(\mathbf{X}^{(n_1+1)}, \ldots, \mathbf{X}^{(n)})$ has the distribution $\mathbb{P}_2$, where $\mathbb{P}_1 \neq \mathbb{P}_2$". Note that the hypotheses (H_0) and (H_1) can be reformulated as follows. The null hypothesis (H_0) means that for all $i \in \{1, \ldots, n\}$, $X_{i,1}, \ldots, X_{i,n}$ are independent and identically distributed (i.i.d) random variables and the alternative hypothesis (H_1) means that there exists $i \in \{1, \ldots, n\}$ such that $X_{i,1}, \ldots, X_{i,n_1}$ have the distribution $\mathbb{P}_1^i$ and $X_{i,n_1+1}, \ldots, X_{i,n}$ have the distribution $\mathbb{P}_2^i$, with $\mathbb{P}_1^i \neq \mathbb{P}_2^i$.

For deciding whether (H_0) has to be rejected or not, we propose to use a test statistic inspired by the one designed by Lung-Yut-Fong et al. (2015) 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. More precisely, the test statistic that we propose for assessing the presence of the potential change n_1 is defined by

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

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

The great difference between our framework and the one considered by Lung-Yut-Fong et al. (2015) is that, in their framework, the vectors $\mathbf{X}^{(j)}$ are K-dimensional with K fixed whereas, in our framework, the vectors are n-dimensional where n may be large.

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

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

where

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

is the rank of $X_{i,j}$ among $(X_{i,1},\dots,X_{i,n})$. This alternative form of $U_{n,i}$ will be used in Section 2.2 in order to extend the two-sample homogeneity test to deal with the multiple sample case.

2.1.2. Theoretical results

If the cumulative distribution function of the $X_{i,j}$'s is assumed to be 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 tends to infinity under the null hypothesis (H_0). Note that the null hypothesis (H_0) assumes that for all $i \in \{1,\dots,n\}$, $X_{i,1},\dots,X_{i,n}$ are i.i.d random variables. Since we also assume that $\mathbf{X} = (X_{i,j})_{1\leq i,j\leq n}$ is a symmetric matrix such that the $X_{i,j}$'s are independent random variables when $i \geq j$, it implies that under the null hypothesis (H_0) all the rows i have the same distribution. Hence, all the $X_{i,j}$ such that $i \geq j$ are i.i.d.

Theorem 1. *Let $\boldsymbol{X} = (X_{i,j})_{1\leq i,j\leq n}$ be a symmetric matrix of random variables $X_{i,j}$ such that the $X_{i,j}$'s are i.i.d. when $i \geq j$. Assume that the cumulative distribution function 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,*

$$T_n(n_1) := n^{-1/2}\big(S_n(n_1) - \mathbb{E}(S_n(n_1)\big) = O_P(1) \text{ as } n \to \infty,$$

where

$$\mathbb{E}(S_n(n_1)) = \frac{n+1}{3}.$$

The proof of Theorem 1 is given in Section 6.1.

Observe that the assumptions under which Theorem 1 is established correspond to the null hypothesis (H_0) described in Section 2.1.1. Hence, we shall reject this null hypothesis when

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

where s is a threshold. A way of computing this threshold in practical situations will be given in Section 3.1.1.

2.2. Multiple-sample homogeneity test

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

2.2.1. Statistical framework

Let us assume that $\mathbf{X} = (X_{i,j})_{1 \leq i,j \leq n}$ is still a symmetric matrix such that the $X_{i,j}$'s are independent random variables when $i \geq j$. Let $0 = n_0 < n_1 < \ldots < n_L < n_{L+1} = n$ be L integers given in $\{1, \ldots, n-1\}$. We propose in this section a statistic to test the null hypothesis: "$(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n_1)})$, $(\mathbf{X}^{(n_1+1)}, \ldots, \mathbf{X}^{(n_2)})$, $\ldots$,$(\mathbf{X}^{(n_L+1)}, \ldots, \mathbf{X}^{(n)})$ have the same distribution" against the alternative hypothesis: "there exists $\ell \in \{1, \ldots, L\}$ such that $(\mathbf{X}^{(n_{\ell-1}+1)}, \ldots, \mathbf{X}^{(n_\ell)})$ has the distribution $\mathbb{P}_\ell$ and $(\mathbf{X}^{(n_\ell+1)}, \ldots, \mathbf{X}^{(n_{\ell+1})})$ has the distribution $\mathbb{P}_{\ell+1}$, where $\mathbb{P}_\ell \neq \mathbb{P}_{\ell+1}$".

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

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

with

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

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.

Let us observe that (5) can be seen as a natural extension of the classical Kruskal-Wallis statistic for univariate observations to the multivariate case, see (van der Vaart, 1998, 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

$$\begin{aligned}
&\frac{4}{n^2} \left[n_1 \sum_{i=1}^{n} \left(\frac{1}{n_1} \sum_{j=1}^{n_1} R_j^{(i)} - \frac{n+1}{2} \right)^2 + (n - n_1) \sum_{i=1}^{n} \left(\frac{1}{n - n_1} \sum_{j=n_1+1}^{n} R_j^{(i)} - \frac{n+1}{2} \right)^2 \right] \\
&= \frac{4}{n^2 n_1} \sum_{i=1}^{n} \left\{ \sum_{j=1}^{n_1} \left(R_j^{(i)} - \frac{n+1}{2} \right) \right\}^2 + \frac{4}{n^2 (n - n_1)} \sum_{i=1}^{n} \left\{ \sum_{j=n_1+1}^{n} \left(R_j^{(i)} - \frac{n+1}{2} \right) \right\}^2 \\
&= \frac{1}{n} \left[\sum_{i=1}^{n} n_1 \left\{ \frac{1}{\sqrt{n n_1 (n - n_1)}} \sum_{j=1}^{n_1} \left(R_j^{(i)} - \frac{n+1}{2} \right) \right\}^2 \right. \\
&\left. + \sum_{i=1}^{n} (n - n_1) \left\{ \frac{1}{\sqrt{n n_1 (n - n_1)}} \sum_{j=n_1+1}^{n} \left(R_j^{(i)} - \frac{n+1}{2} \right) \right\}^2 \right] = \sum_{i=1}^{n} U_{n,i}^2(n_1),
\end{aligned}$$

by using (2), which corresponds to (1).

2.2.2. Theoretical results

If the cumulative distribution function of the $X_{i,j}$'s is assumed to be 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 tends to infinity.

Theorem 2. *Let $\boldsymbol{X} = (X_{i,j})_{1 \leq i,j \leq n}$ be a symmetric matrix of random variables $X_{i,j}$ such that the $X_{i,j}$'s are i.i.d when $i \geq j$. Assume that the cumulative distribution function of the $X_{i,j}$'s is continuous and that there exist $0 < \tau_1 < \tau_2 < \ldots < \tau_L < 1$ such that for all $\ell \in \{1, \ldots, L\}$, $n_\ell / n \to \tau_\ell$ as $n \to \infty$. Then,*

$$n^{-1/2}\big(S_n(n_1, \ldots, n_L) - \mathbb{E}\left[S_n(n_1, \ldots, n_L)\right]\big) = O_P(1) \text{ as } n \to \infty,$$

with

$$\mathbb{E}\left[S_n(n_1, \ldots, n_L)\right] = \frac{L(n+1)}{3}.$$

The proof of Theorem 2 is given in Section 6.2

Note that the n_ℓ's can be seen as the boundaries of groups of random variables having different distributions. We shall explain in the next section how to derive from this theorem a methodology for estimating the n_ℓ's when they are assumed to be unknown.

2.3. Change-point estimation

We propose in this section to use the test statistic (5) defined in Section 2.2 to derive the location of the block boundaries $n_1^* < n_2^* < \cdots < n_L^*$. More precisely, we propose to estimate $(n_1^*, n_2^*, \cdots, n_L^*)$ as follows:

$$(\widehat{n}_1, \cdots, \widehat{n}_L) := \text{Argmax}_{1 \leq n_1 < \ldots < n_L < n} \, S_n(n_1, \ldots, n_L), \tag{7}$$

where $S_n(n_1, \ldots, n_L)$ is defined in (5).

2.3.1. Theoretical results

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

Theorem 3. *Let $\boldsymbol{X} = (X_{i,j})_{1 \leq i,j \leq n}$ be a symmetric matrix such that the $X_{i,j}$'s are independent random variables when $i \geq j$ with a continuous cumulative distribution function. Let $\mathbb{P}_0^0$ be the distribution of $X_{i,j}$ for $i, j \in \{1, \ldots, n_1^*\}$, $\mathbb{P}_1^0$ the distribution of $X_{i,j}$ for $i \in \{1, \ldots, n_1^*\}$, $j \in \{n_1^* + 1, \ldots, n\}$ and $\mathbb{P}_1^1$ the distribution of $X_{i,j}$ for $i, j \in \{n_1^* + 1, \ldots, n\}$ where $\mathbb{P}_0^0 \neq \mathbb{P}_1^0$ or $\mathbb{P}_1^0 \neq \mathbb{P}_1^1$. Assume that*

$$\mathbb{P}(X \leq Y) \neq \frac{1}{2}, \tag{8}$$

where $X \sim \mathbb{P}_0^0$ (or $\mathbb{P}_1^0$) and $Y \sim \mathbb{P}_1^0$ (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 tends to infinity. Then, for all positive δ,

$$\mathbb{P}\big(|\widehat{n}_1 - n_1^*| \geq n\delta\big) \to 0, \text{ as } n \to \infty,$$

where $\widehat{n}_1$ is defined by (7) when $L = 1$.

Remark 2. Note that Assumption (8) in Theorem 3 is classical in the context of rank based test statistics such as the Mann-Whitney test (see van der Vaart (1998)).

2.3.2. Practical implementation

In practice, directly maximizing (7) is computationally prohibitive as it corresponds to a task which complexity exponentially grows with L. However, thanks to the additive structure of (5), it is possible to use a dynamic programming strategy as we shall explain hereafter. We refer here to the classical dynamic programming approach described in Kay (1993) which can be traced back to the note of Bellman (1961).

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

$$I_L(p) = \max_{1 \leq n_1 < \cdots < n_L < n_{L+1} = p} \sum_{\ell=0}^{L} \Delta(n_\ell + 1 : n_{\ell+1}), \tag{9}$$

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

$$I_L(p) = \max_{n_L} \left\{ I_{L-1}(n_L) + \Delta(n_L + 1 : p) \right\} , \tag{10}$$

which is proved in Section 6.4. Thus, for solving the optimization problem (7), we proceed as follows. We start by computing the $\Delta(i : j)$ for all (i, j) such that $1 \leq i \leq j \leq n$. All the $I_0(p)$ are thus available for $p = 1, \ldots, n$. Then $I_1(p)$ is computed by using the recursion (10) and so on. Hence the complexity of our algorithm is $O(n^3)$.

Figure 1 displays the computational times in seconds associated with our multiple change-point estimation strategy based on the dynamic programming algorithm. We observe from this figure the polynomial computational time of our procedure. For instance, it takes 15 minutes to our algorithm for processing a 500×500 matrix. Note that in the framework of univariate time series segmentation, the PELT procedure devised by Killick et al. (2012) performs multiple change-point detection at a linear computational cost. It could be interesting to see if the computational burden of our procedure could be reduced by using an extension of their approach.

3. Numerical experiments

3.1. Statistical performance of the two-sample homogeneity test

3.1.1. Practical calibration of the rejection region

We propose hereafter a procedure for calibrating the threshold s of the rejection region $T_n(n_1) > s$ defined in (4). For ensuring that the two-sample homogeneity test is of level α,

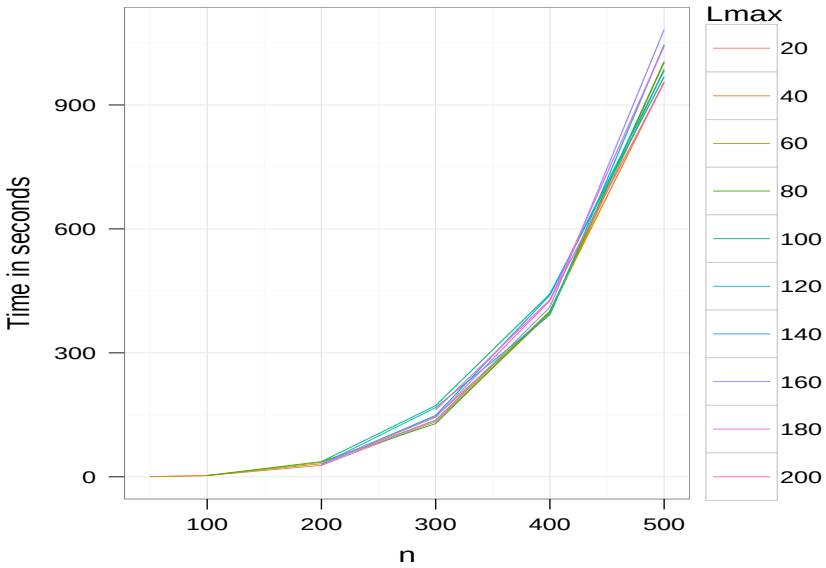

Figure 1: Computational times in seconds for the dynamic programming algorithm described in Section 2.3 as a function of n for different values of $L_{\max}$.

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 10^4 $n \times n$ symmetric matrices $\boldsymbol{X} = (X_{i,j})$ with $n \in \{50, 100, 500, 1000\}$. More precisely, the $(X_{i,j})_{i \geq j}$'s are independent random variables distributed as a zero mean standard Gaussian distribution ($\mathcal{N}(0,1)$), a Cauchy distribution with 0 and 1 location and scale parameters ($\mathcal{C}au(0,1)$), respectively or an Exponential distribution of parameter 2 ($\mathcal{E}xp(2)$). We shall consider two values for n_1: $n_1 = \lfloor 0.1n \rfloor$ 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.

	$n_1 = \lfloor 0.1n \rfloor$			$n_1 = \lfloor 0.5n \rfloor$		
	$\mathcal{N}(0,1)$	$\mathcal{C}au(0,1)$	$\mathcal{E}xp(2)$	$\mathcal{N}(0,1)$	$\mathcal{C}au(0,1)$	$\mathcal{E}xp(2)$
$n = 50$	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

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

3.1.2. Power of the test statistic

In this section, we study the power of the two-sample homogeneity test defined in Section 2.1.1. We generated 10^4 $n \times n$ 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 $(X_{i,j})_{(i,j)\in\mathcal{I}_1} \overset{iid}{\sim} \mathcal{L}_1$, $(X_{i,j})_{(i,j)\in\mathcal{I}_2} \overset{iid}{\sim} \mathcal{L}_2$ and $(X_{i,j})_{(i,j)\in\mathcal{I}_3} \overset{iid}{\sim} \mathcal{L}_3$ 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 2.1.1 in the case where $\mathcal{L}_1 = \mathcal{L}_3 = \mathcal{N}(0,1)$ and $\mathcal{L}_2 = \mathcal{N}(\mu, 1)$ where μ belongs to the set $\{0, 0.01, 0.02, \ldots, 0.99, 1\}$.

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 our testing procedure is all the more powerful that μ is large.

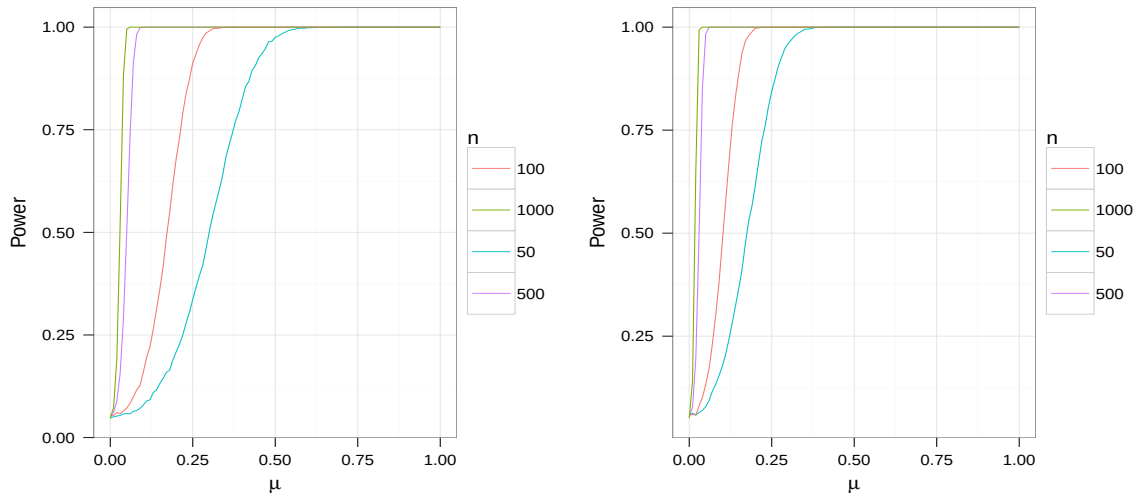

Figure 2: Power curves for the two-sample homogeneity test as a function of μ for different values of n, $n_1 = \lfloor 0.1n \rfloor$ (left) and $n_1 = \lfloor 0.5n \rfloor$ (right).

3.2. 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 2.3. This method is implemented in the R package MuChPoint, which will be available on the Comprehensive R Archive Network (CRAN).

We generated 10^4 $n \times n$ 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).

We shall 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 the 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 the distribution $\mathcal{L}_2$, which is assumed to be different from $\mathcal{L}_1$. The upper triangular part of $\mathbf{X}$ is then derived by symmetry.

We shall 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 shall consider for $\mathcal{L}_1$ a $\mathcal{N}(1, \sigma^2)$, a $\mathcal{E}xp(2)$ or a $\mathcal{C}au(1, a)$ distribution where σ and a are in $\{1, 2, 5\}$. The $\mathcal{L}_2$ distributions associated with each of them are $\mathcal{N}(0, \sigma^2)$, $\mathcal{E}xp(\lambda)$ and $\mathcal{C}au(0, a)$ 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.

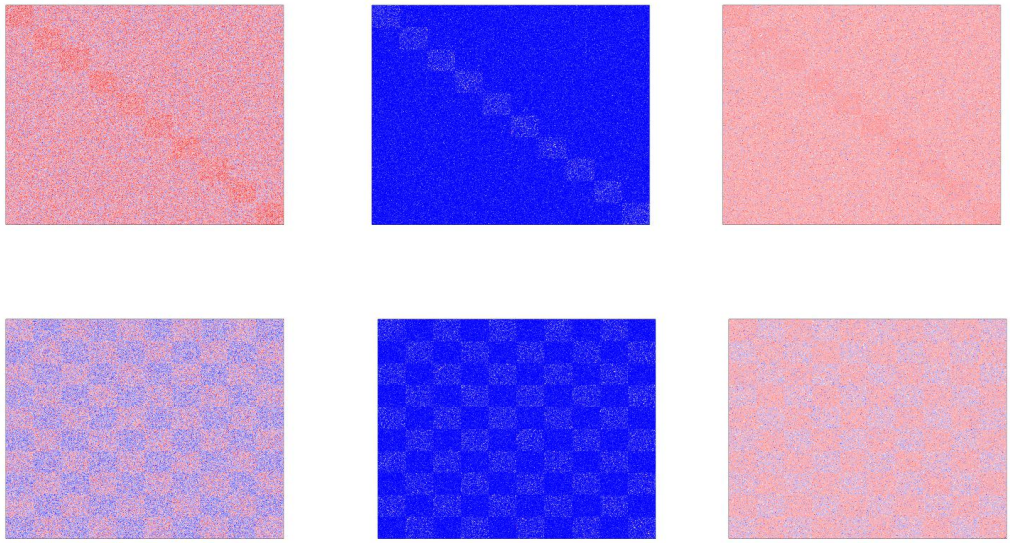

Figure 3: Examples of 400×400 matrices $\mathbf{X}$. Top: Block Diagonal configuration. Bottom: Chessboard configuration. Left: $\mathcal{L}_1 = \mathcal{N}(1, 4)$, $\mathcal{L}_2 = \mathcal{N}(0, 4)$, middle: $\mathcal{L}_1 = \mathcal{E}xp(2)$, $\mathcal{L}_2 = \mathcal{E}xp(1)$ and right: $\mathcal{L}_1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

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 in the cases where σ increases, some spurious change-points appear close to the true change-point positions.

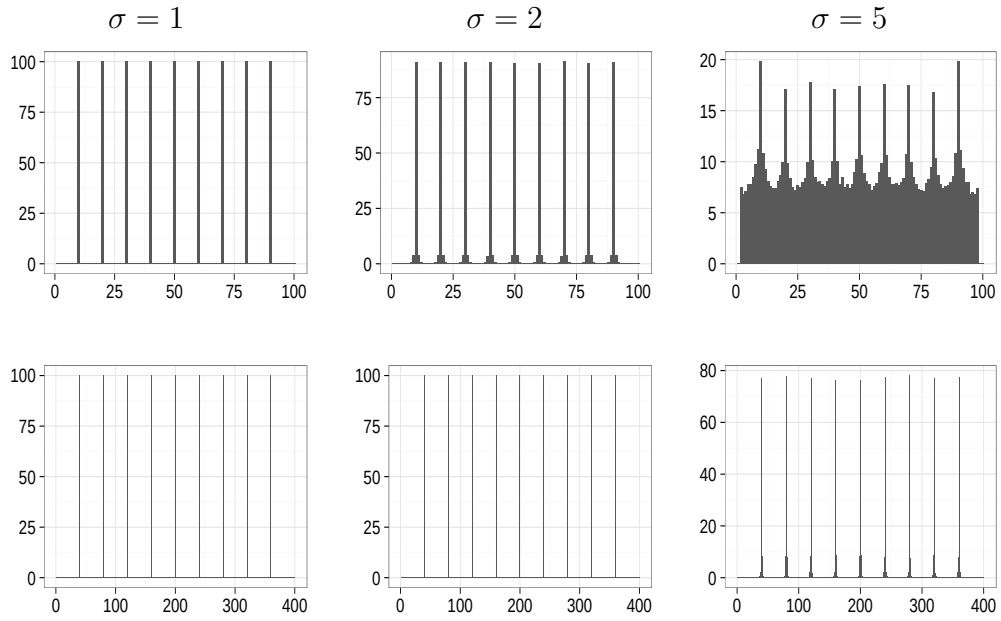

Figure 4: Barplots associated with the multiple change-point estimation procedure for $n = 100$ (top), $n = 400$ (bottom), $\mathcal{L}_1 = \mathcal{N}(1, \sigma^2)$ and $\mathcal{L}_2 = \mathcal{N}(0, \sigma^2)$ for different values of σ. The true positions of the change-points are located at the multiples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

We also compared our multiple change-point estimation strategy (MuChPoint) to the one devised by Matteson and James (2014) (ecp), which is, to the best of our knowledge, the most recent approach proposed for solving this issue. The results are gathered in Figures 5 and 6 which display the boxplots of the distance D, defined in (11), 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 to chose $\alpha \in (0; 2]$ such that $\mathbb{E}\left[|X|^{\alpha}\right] < +\infty$. 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 $n \in \{50, 100, 200, 300, 400\}$. More precisely, the distance D is defined as follows

$$D(\widehat{\mathbf{n}}, \mathbf{n}^*) = \frac{1}{n}\sqrt{\sum_{k=1}^{K^*}(\widehat{n}_k - n_k^*)^2}, \tag{11}$$

where $\mathbf{n}^* = (n_1^*, \ldots, n_{K^*}^*)$ denotes the vector of the true K^* change-point positions and $\widehat{\mathbf{n}} = (\widehat{n}_1, \ldots, \widehat{n}_{K^*})$ 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 $\boldsymbol{\tau}^* = (\tau_1^*, \ldots, \tau_{K^*}^*)$, $\widehat{\boldsymbol{\tau}} = (\widehat{\tau}_1, \ldots, \widehat{\tau}_{K^*})$ with $n_k^* = \lfloor n\tau_k^* \rfloor$ and $\widehat{n}_k = \lfloor n\widehat{\tau}_k \rfloor$. In order to benchmark these methodologies, we provide to both of them the true value K^* of the number of change-points, which is here equal to 10.

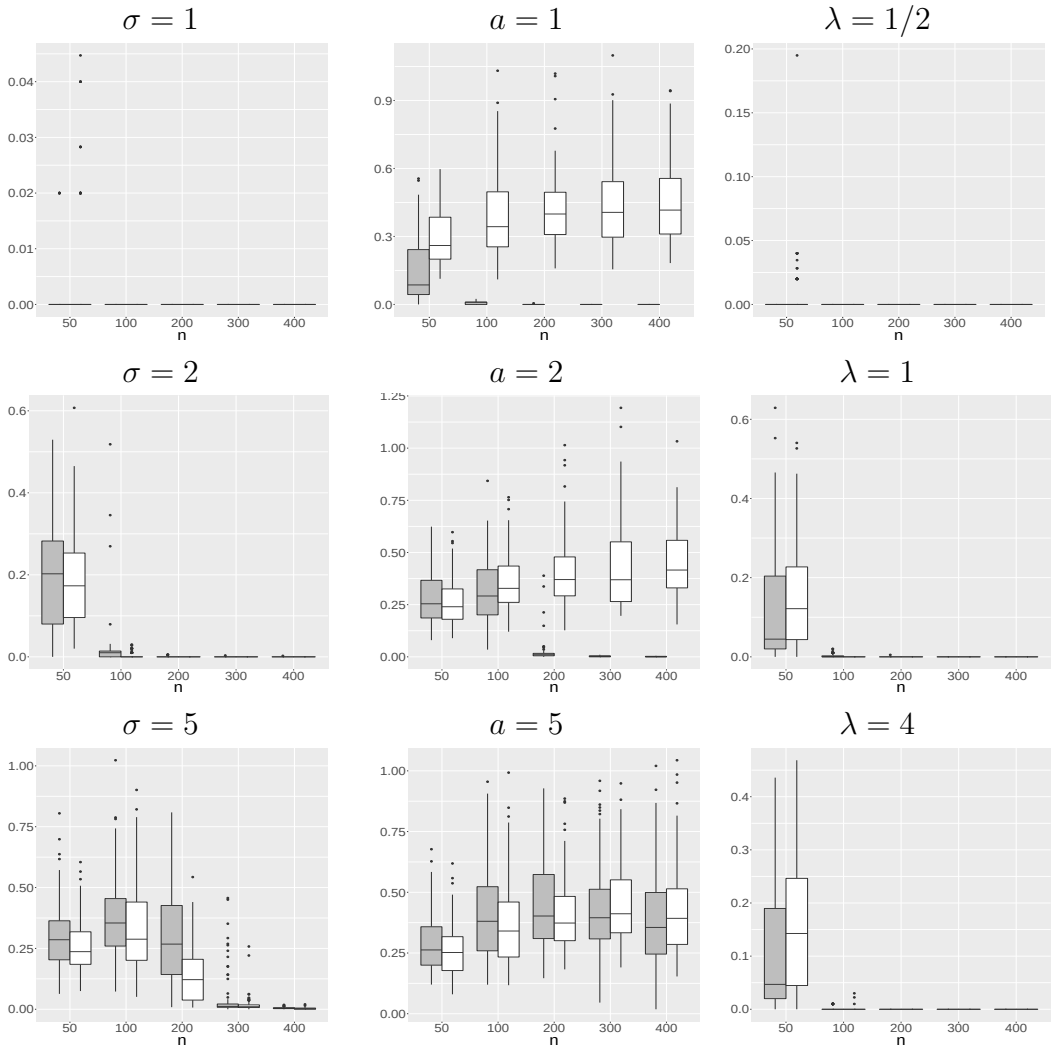

Figure 5: Boxplots of the distances D for MuChPoint and ecp in the Chessboard configuration. Left: $\mathcal{L}_1 = \mathcal{N}(1, \sigma^2)$, $\mathcal{L}_2 = \mathcal{N}(0, \sigma^2)$, middle: $\mathcal{L}_1 = \mathcal{C}au(1, a)$, $\mathcal{L}_2 = \mathcal{C}au(0, a)$ and right: $\mathcal{L}_1 = \mathcal{E}xp(2)$,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

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.

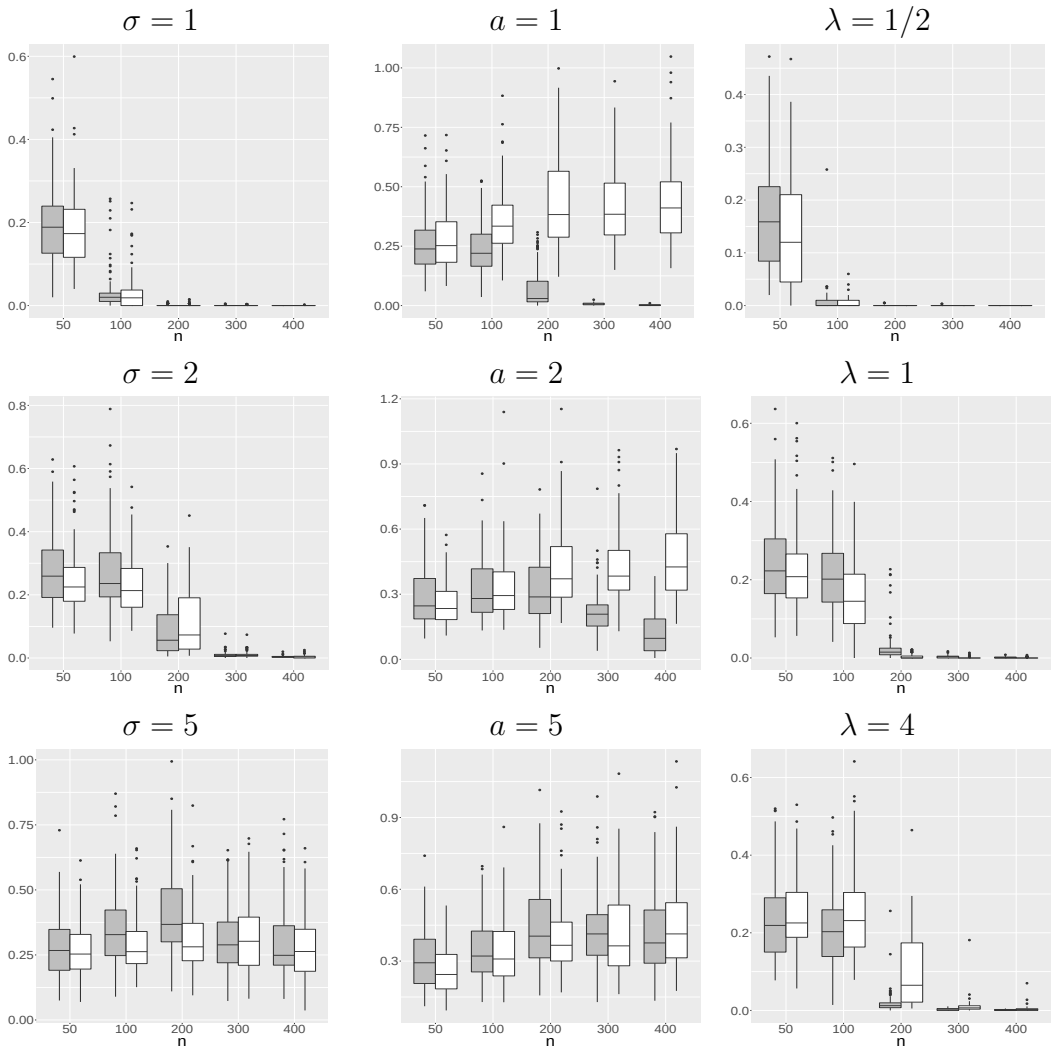

Figure 6: Boxplots of the distances D for MuChPoint and ecp in the Block Diagonal configuration. Left: $\mathcal{L}_1 = \mathcal{N}(1, \sigma^2)$, $\mathcal{L}_2 = \mathcal{N}(0, \sigma^2)$, middle: $\mathcal{L}_1 = \mathcal{C}au(1, a)$, $\mathcal{L}_2 = \mathcal{C}au(0, a)$ and right: $\mathcal{L}_1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

4. 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. (2012). This technology provides read pairs corresponding to pairs of genomic loci that physically interacts in the nucleus, see Lieberman-Aiden et al. (2009) 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. (2009) considered a Mb resolution, whereas Dixon et al. (2012) 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. (2012) 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 $n = 1534$.

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.

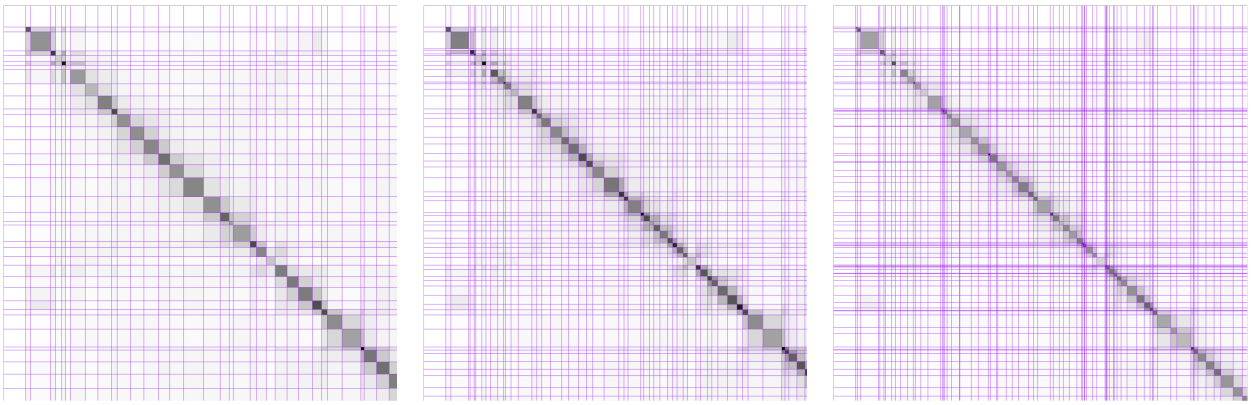

Figure 7: Estimated matrices $\widehat{\mathbf{X}}$ for different number of estimated change-points: 35 (left), 55 (middle) and 75 (right).

In order to further compare our approach with the one proposed by Dixon et al. (2012), we computed the two parts of the Hausdorff distance which is defined by

$$d\left(\widehat{\boldsymbol{n}}_B, \widehat{\boldsymbol{n}}\right) = \max\left(d_1\left(\widehat{\boldsymbol{n}}_B, \widehat{\boldsymbol{n}}\right), d_2\left(\widehat{\boldsymbol{n}}_B, \widehat{\boldsymbol{n}}\right)\right) \ , \tag{12}$$

where $\widehat{\boldsymbol{n}}$ and $\widehat{\boldsymbol{n}}_B$ are the change-points found by our approach and Dixon et al. (2012), respectively. In (12),

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

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.

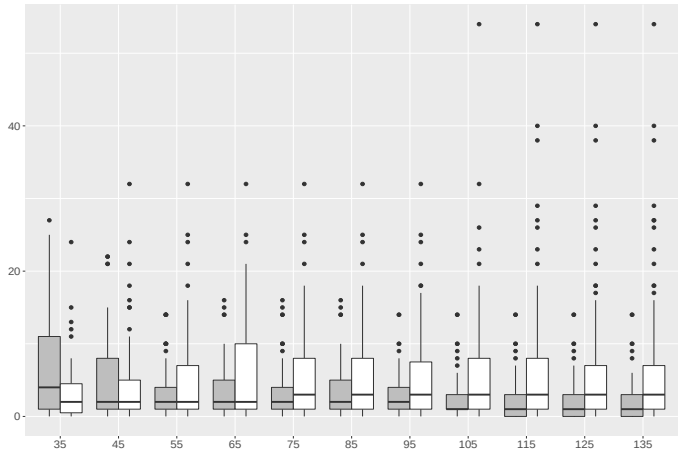

Figure 8: Boxplots for the infimum parts of the Hausdorff distances d_1 (white) and d_2 (gray) between the change-points found by Dixon et al. (2012) and our approach for different values of the estimated number of change-points.

For comparison purpose, we used the R package Capushe which implements a model selection approach based on the slope heuristics theory and described in Baudry et al. (2012). 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 by 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 Dixon et al. (2012), that is equal to 85, the positions of the block boundaries are very close as displayed in Figure 11.

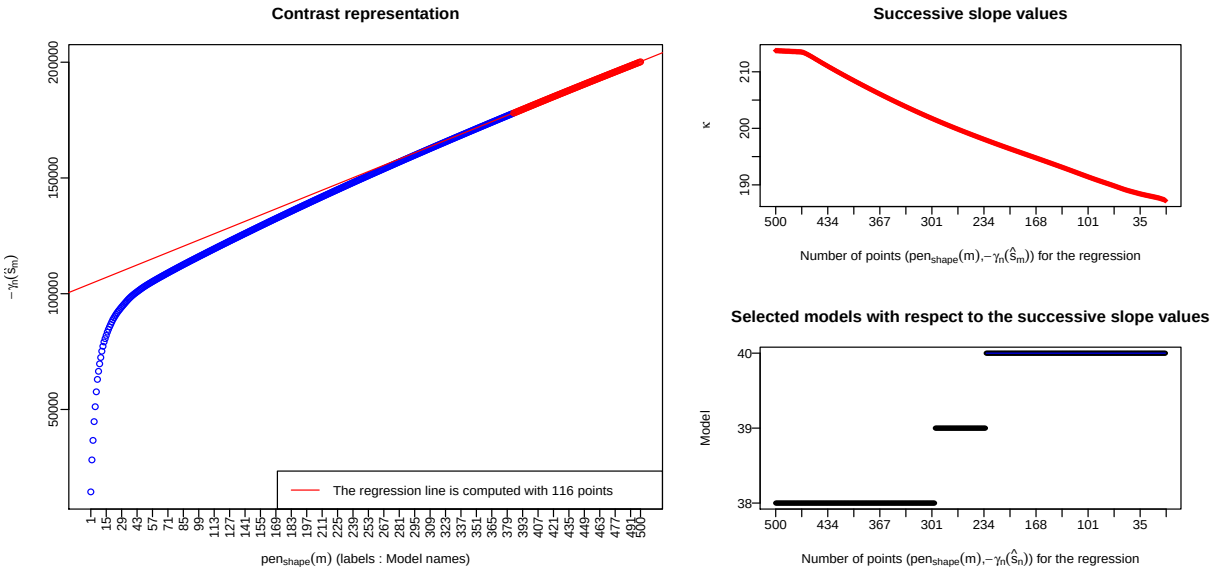

Figure 9: Outputs of the R package Capushe.

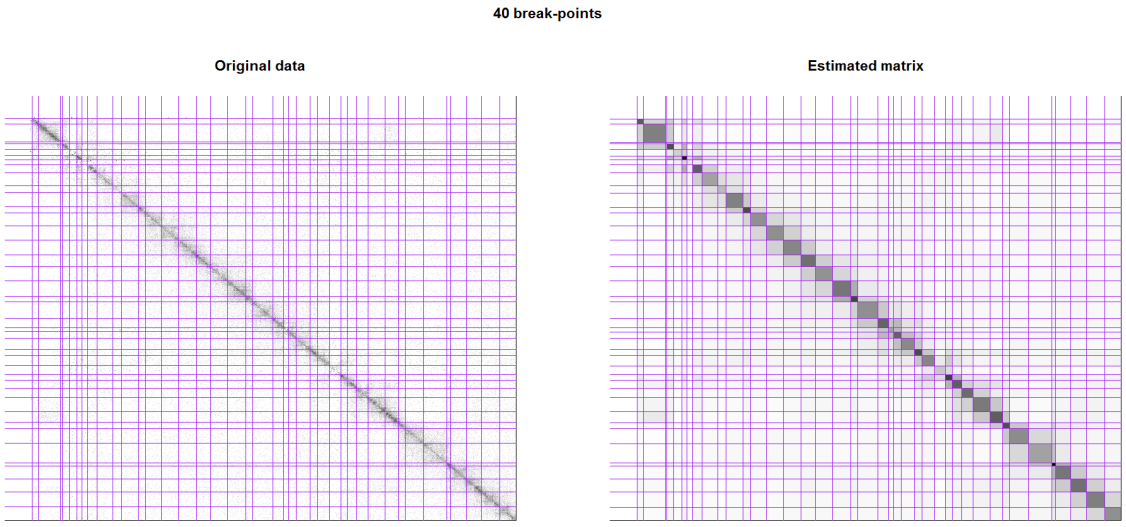

Figure 10: Estimated matrix $\widehat{\mathbf{X}}$ when L is estimated by using the R package Capushe.

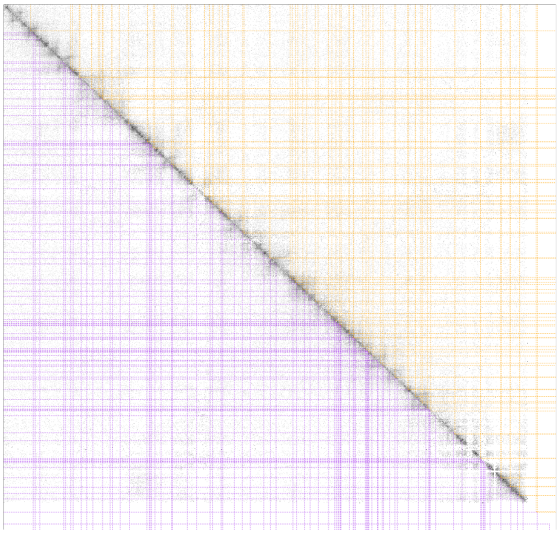

Figure 11: Topological domains detected by Dixon et al. (2012) (upper triangular part of the matrix) and by our method (lower triangular part of the matrix).

5. 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. Firstly, it is a nonparametric approach which showed very good statistical performances from a practical point of view. Secondly, its computational burden makes its use possible on large Hi-C data matrices.

6. Proofs

In this section, we prove Theorems 1, 2, 3 and Equation (10). The proofs of the theorems given below use technical lemmas established in Section 7.

6.1. Proof of Theorem 1

For proving Theorem 1, we first compute the expectation of $S_n(n_1)$.

$$
\begin{aligned}
\mathbb{E}\left[S_n(n_1)\right] &= \sum_{i=1}^{n} \mathbb{E}\left[U_{n,i}^2(n_1)\right] \\
&= \frac{1}{nn_1(n-n_1)} \sum_{i=1}^{n} \mathbb{E}\left[\left(\sum_{j_0=1}^{n_1} \sum_{j_1=n_1+1}^{n} h(X_{i,j_0}, X_{i,j_1})\right)^2\right] \\
&= \frac{1}{nn_1(n-n_1)} \sum_{i=1}^{n} \sum_{1\le j_0,k_0\le n_1} \sum_{n_1+1\le j_1,k_1\le n} \mathbb{E}\left[h(X_{i,j_0}, X_{i,j_1}) h(X_{i,k_0}, X_{i,k_1})\right] \\
&= \frac{1}{nn_1(n-n_1)} \sum_{i=1}^{n} \Bigg\{ \sum_{j_0=1}^{n_1} \sum_{j_1=n_1+1}^{n} \mathbb{E}\left[h^2(X_{i,j_0}, X_{i,j_1})\right] \\
&\quad + \sum_{j_0=1}^{n_1} \sum_{n_1+1\le j_1\ne k_1\le n} \mathbb{E}\left[h(X_{i,j_0}, X_{i,j_1}) h(X_{i,j_0}, X_{i,k_1})\right] \\
&\quad + \sum_{1\le j_0\ne k_0\le n_1} \sum_{j_1=n_1+1}^{n} \mathbb{E}\left[h(X_{i,j_0}, X_{i,j_1}) h(X_{i,k_0}, X_{i,j_1})\right] \\
&\quad + \sum_{1\le j_0\ne k_0\le n_1} \sum_{n_1+1\le j_1\ne k_1\le n} \mathbb{E}\left[h(X_{i,j_0}, X_{i,j_1}) h(X_{i,k_0}, X_{i,k_1})\right] \Bigg\}.
\end{aligned}
$$

By using Lemma 1, we get that

$$
\begin{aligned}
\mathbb{E}\left[S_n(n_1)\right] &= \frac{1}{nn_1(n-n_1)} \sum_{i=1}^{n} \Bigg\{ n_1(n-n_1) + \frac{1}{3} n_1(n-n_1)(n-n_1-1) \\
&\qquad\qquad + \frac{1}{3} n_1(n_1-1)(n-n_1) \Bigg\} \\
&= 1 + \frac{n-n_1-1}{3} + \frac{n_1-1}{3} = \frac{n+1}{3}.
\end{aligned}
$$

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

$S_n(n_1)$ as follows:

$$\begin{aligned}
S_n(n_1) - \mathbb{E}\left[S_n(n_1)\right] &= \frac{1}{nn_1(n-n_1)} \sum_{i=1}^{n} \left(\sum_{j_0=1}^{n_1} \sum_{j_1=n_1+1}^{n} h(X_{i,j_0}, X_{i,j_1}) \right)^2 - \frac{n+1}{3} \\
&= \frac{1}{nn_1(n-n_1)} \left\{ \sum_{i=1}^{n} \sum_{j_0=1}^{n_1} \sum_{j_1=n_1+1}^{n} \left[h^2(X_{i,j_0}, X_{i,j_1}) - 1 \right] \right. \\
&\quad + \sum_{i=1}^{n} \sum_{j_0=1}^{n_1} \sum_{n_1+1 \le j_1 \ne k_1 \le n} [h(X_{i,j_0}, X_{i,j_1}) h(X_{i,j_0}, X_{i,k_1}) - 1/3] \\
&\quad + \sum_{i=1}^{n} \sum_{1 \le j_0 \ne k_0 \le n_1} \sum_{j_1=n_1+1}^{n} [h(X_{i,j_0}, X_{i,j_1}) h(X_{i,k_0}, X_{i,j_1}) - 1/3] \\
&\quad \left. + \sum_{i=1}^{n} \sum_{1 \le j_0 \ne k_0 \le n_1} \sum_{n_1+1 \le j_1 \ne k_1 \le n} h(X_{i,j_0}, X_{i,j_1}) h(X_{i,k_0}, X_{i,k_1}) \right\} \\
&=: \frac{1}{nn_1(n-n_1)} \left\{ A + B + C + D \right\},
\end{aligned}$$

where each term of this equality is centered. First, we observe that $A = 0$ a.s. (almost surely) by Assertion (ii) of Lemma 1.

By using the Markov inequality we get that for all $\varepsilon > 0$,

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

By using the Cauchy-Schwarz inequality, we thus get that

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

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-hand 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 smaller or equal to 5 have to be considered. Moreover, note that

$$\left|(h(x,y)h(z,t) - 1/3) \times (h(x',y')h(z',t') - 1/3)\right| \leq 16/9 \leq 2,$$

for all $x, y, z, t, x', y', z', t'$. Hence we get that, for all $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{B}{\sqrt{n}}\right| > \frac{6n^3}{\varepsilon}\right) \leq \frac{\varepsilon}{6n^{7/2}} \sum_{i=1}^{n} 2n^{5/2} = \varepsilon/3. \tag{13}$$

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

$$\mathbb{P}\left(\left|\frac{C}{\sqrt{n}}\right| > \frac{6n^3}{\varepsilon}\right) \leq \varepsilon/3. \tag{14}$$

By using the Markov and the Cauchy-Schwarz inequalities as previously, we get that, for all $\varepsilon > 0$,

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

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 greater than 8 and smaller than 10 by Assertion (v) of Lemma 1. Only the cases where the cardinality of the set is smaller than 7 have to be considered. Observe moreover that

$$\left|h(x,y)h(z,t)h(x',y')h(z',t')\right| \leq 1, \text{ for all } x, y, z, t, x', y', z', t' \in \mathbb{R}.$$

Therefore, for all $\varepsilon > 0$, we get,

$$\mathbb{P}\left(\left|\frac{D}{\sqrt{n}}\right| > \frac{3n^3}{\varepsilon}\right) \leq \frac{\varepsilon}{3n^{7/2}} \times n^{7/2} = \varepsilon/3. \tag{15}$$

Finally, by combining (13), (14) and (15), we obtain that, for all $\varepsilon > 0$,

$$\mathbb{P}\left(nn_1(n-n_1) \times \frac{|S_n(n_1) - \mathbb{E}\left[S_n(n_1)\right]|}{\sqrt{n}} > \frac{15n^3}{\varepsilon}\right) \quad \leq \quad \varepsilon,$$

which can be rewritten as

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

Since we assumed that $n_1/n \to \tau_1$ as $n \to \infty$, we get that

$$\frac{S_n(n_1) - \mathbb{E}\left[S_n(n_1)\right]}{\sqrt{n}} = O_{\mathbb{P}}(1),$$

which concludes the proof of Theorem 1.

6.2. Proof of Theorem 2

Let us start with the computation of the expectation of $S_n\left(n_1, \ldots, n_L\right)$. First observe that, for any $i \in \{1, \ldots, n\}$ and $\ell \in \{0, \ldots, L\}$,

$$\begin{aligned}
\left(\overline{R}_\ell^{(i)} - \frac{n+1}{2}\right)^2 &= \left(\frac{1}{n_{\ell+1} - n_\ell} \sum_{j=n_\ell+1}^{n_{\ell+1}} R_j^{(i)} - \frac{n+1}{2}\right)^2 \\
&= \frac{1}{(n_{\ell+1} - n_\ell)^2} \sum_{j=n_\ell+1}^{n_{\ell+1}} \left(R_j^{(i)} - \frac{n+1}{2}\right)^2 \\
&\quad + \frac{1}{(n_{\ell+1} - n_\ell)^2} \sum_{n_\ell+1 \leq j \neq j' \leq n_{\ell+1}} \left(R_j^{(i)} - \frac{n+1}{2}\right)\left(R_{j'}^{(i)} - \frac{n+1}{2}\right) \\
&= \frac{1}{(n_{\ell+1} - n_\ell)^2} \left(\sum_{j=n_\ell+1}^{n_{\ell+1}} A_j^{(i)} + \sum_{n_\ell+1 \leq j \neq j' \leq n_{\ell+1}} B_{jj'}^{(i)}\right),
\end{aligned} \tag{16}$$

where

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

By using the definition (6) of $R_j^{(i)}$, we get that,

$$\begin{aligned} A_j^{(i)} &= \left(\sum_{k=1}^{n} \mathbb{1}_{\{X_{i,k}\leq X_{i,j}\}} - \frac{n+1}{2}\right)^2 = \left(1 + \sum_{\substack{k=1\\k\neq j}}^{n} \mathbb{1}_{\{X_{i,k}\leq X_{i,j}\}} - \frac{n+1}{2}\right)^2 \\ &= \left(\sum_{\substack{k=1\\k\neq j}}^{n} \left(\mathbb{1}_{\{X_{i,k}\leq X_{i,j}\}} - \frac{1}{2}\right)\right)^2 \\ &= \sum_{\substack{k=1\\k\neq j}}^{n} g(X_{i,k}, X_{i,j})^2 + \sum_{\substack{k=1\\k\neq j}}^{n} \sum_{\substack{k'=1\\k'\neq k\\k\neq j}}^{n} g(X_{i,k}, X_{i,j}) g(X_{i,k'}, X_{i,j}), \end{aligned} \tag{17}$$

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

$$\mathbb{E}\left[A_j^{(i)}\right] = \frac{1}{4}(n-1) + \frac{1}{12}(n-1)(n-2) = \frac{(n-1)(n+1)}{12}. \tag{18}$$

Then, we decompose $B_{jj'}^{(i)}$ in the four following terms.

$$\begin{aligned}
B_{jj'}^{(i)} &= \left(R_j^{(i)} - \frac{n+1}{2}\right)\left(R_{j'}^{(i)} - \frac{n+1}{2}\right) \\
&= \left(\sum_{k=1}^{n} \mathbb{1}_{\{X_{i,k}\leq X_{i,j}\}} - \frac{n+1}{2}\right)\left(\sum_{k'=1}^{n} \mathbb{1}_{\{X_{i,k'}\leq X_{i,j'}\}} - \frac{n+1}{2}\right) \\
&= \left(1 + \sum_{\substack{k=1\\k\neq j}}^{n} \mathbb{1}_{\{X_{i,k}\leq X_{i,j}\}} - \frac{n+1}{2}\right)\left(1 + \sum_{\substack{k'=1\\k'\neq j'}}^{n} \mathbb{1}_{\{X_{i,k'}\leq X_{i,j'}\}} - \frac{n+1}{2}\right) \\
&= \left(\sum_{\substack{k=1\\k\neq j}}^{n} \left(\mathbb{1}_{\{X_{i,k}\leq X_{i,j}\}} - \frac{1}{2}\right)\right)\left(\sum_{\substack{k'=1\\k'\neq j'}}^{n} \left(\mathbb{1}_{\{X_{i,k'}\leq X_{i,j'}\}} - \frac{1}{2}\right)\right) \\
&= \sum_{\substack{k=1\\k\neq j}}^{n}\sum_{\substack{k'=1\\k'\neq j'}}^{n} g(X_{i,k},X_{i,j})g(X_{i,k'},X_{i,j'}) \\
&= g(X_{i,j'},X_{i,j})g(X_{i,j},X_{i,j'}) \\
&\quad + \sum_{\substack{k=1\\k\neq j\\k\neq j'}}^{n} g(X_{i,k},X_{i,j})g(X_{i,j},X_{i,j'}) \\
&\quad + \sum_{\substack{k'=1\\k'\neq j'\\k'\neq j}}^{n} g(X_{i,j'},X_{i,j})g(X_{i,k'},X_{i,j'}) \\
&\quad \sum_{\substack{k=1\\k\neq j\\k\neq j'}}^{n}\sum_{\substack{k'=1\\k'\neq j'\\k'\neq j}}^{n} g(X_{i,k},X_{i,j})g(X_{i,k'},X_{i,j'}) \\
&=: B_1 + B_2 + B_3 + B_4. \qquad (19)
\end{aligned}$$

By Lemma 2, we obtain that

$$\mathbb{E}\left[B_1\right] = -\frac{1}{4},\ \mathbb{E}\left[B_2\right] = \mathbb{E}\left[B_3\right] = -\frac{n-2}{12} \text{ and } \mathbb{E}\left[B_4\right] = \frac{n-2}{12},$$

since the only term in the sum defining B_4 having a non null expectation is the one for which $k = k'$. Hence,

$$\mathbb{E}\left[B_{jj'}^{(i)}\right] = -\frac{1}{4} - 2\times\frac{n-2}{12} + \frac{n-2}{12} = -\frac{1}{4} - \frac{n-2}{12} = -\frac{n+1}{12}. \qquad (20)$$

By (16), (18) and (20),

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

By (5), we get that

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

Now we focus on the asymptotic behavior of $S_n\left(n_1,\ldots,n_L\right)$. For this, we decompose the centered version of $S_n\left(n_1,\ldots,n_L\right)$ as follows.

$$\begin{aligned}
&S_n\left(n_1,\ldots,n_L\right) - \mathbb{E}\left[S_n\left(n_1,\ldots,n_L\right)\right] \\
&= \frac{4}{n^2}\sum_{\ell=0}^{L}\left(n_{\ell+1}-n_\ell\right)\sum_{i=1}^{n}\left(\overline{R}_\ell^{(i)} - \frac{n+1}{2}\right)^2 - \frac{L(n+1)}{3} \\
&= \frac{4}{n^2}\sum_{\ell=0}^{L}\left(n_{\ell+1}-n_\ell\right)\sum_{i=1}^{n}\left[\frac{1}{(n_{\ell+1}-n_\ell)^2}\left(\sum_{j=n_\ell+1}^{n_{\ell+1}}\left(A_j^{(i)} - \mathbb{E}\left[A_j^{(i)}\right]\right)\right.\right. \\
&\qquad\qquad \left.\left. + \sum_{n_\ell+1\le j\neq j'\le n_{\ell+1}}\left(B_{jj'}^{(i)} - \mathbb{E}\left[B_{jj'}^{(i)}\right]\right)\right)\right] \\
&= \frac{4}{n^2}\sum_{\ell=0}^{L}\frac{1}{n_{\ell+1}-n_\ell}\sum_{i=1}^{n}\sum_{t=1}^{7} Z_i^{(t)},
\end{aligned}$$

where $A_j^{(i)}$ and $B_{jj'}^{(i)}$ are defined in (17) and (19), and the $Z_i^{(t)}$ are defined as follows:

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

Then, we get that, for all $M > 0$,

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

Using the Markov inequality we get that

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

By using the Cauchy-Schwarz inequality we obtain that

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

We shall now give upper bounds for $\mathbb{E}\left[\left(\sum_{i=1}^{n}Z_i^{(t)}\right)^2\right]$ for all $t\in\{1,\ldots,7\}$. First, by using Assertion (ii) of Lemma 2, we get

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

Then,

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

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 (iii) of Lemma 2. Hence, only the cases where the cardinality of this set is smaller or equal to 7 have to be considered. Since

$$\left|\left(g(x,y)g(z,t)-\frac{1}{12}\right)\left(g(x',y')g(z',t')-\frac{1}{12}\right)\right|\leq 1/9\leq 1,$$

for all $x,y,z,t,x',y',z',t'\in\mathbb{R}$, we get that,

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

By using similar arguments and Assertion (iv) of Lemma 2, we get that

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

By using similar arguments as those used for bounding $\mathbb{E}\left[\left(\sum_{i=1}^{n} Z_i^{(2)}\right)^2\right]$ and by Assertion (iii) of Lemma 2, we get that $\mathbb{E}\left[g(X,Y)g(Y,Z)\right] = -\mathbb{E}\left[g(X,Y)g(Z,Y)\right] = -1/12$. Hence,

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

By using Assertion (ii) of Lemma 2, we obtain that

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=1}^{n} Z_i^{(3)}\right)^2\right] &= \sum_{i=1}^{n}\sum_{i'=1}^{n}\mathbb{E}\left[Z_i^{(3)}Z_{i'}^{(3)}\right] \\ &= \sum_{i=1}^{n}\sum_{i'=1}^{n}\sum_{n_\ell+1\leq j\neq j'\leq n_{\ell+1}}\sum_{n_\ell+1\leq r\neq r'\leq n_{\ell+1}}\mathbb{E}\Bigg[\left\{g(X_{i,j'},X_{i,j})g(X_{i,j},X_{i,j'})+\frac{1}{4}\right\} \\ &\qquad \left\{g(X_{i'r'},X_{i'r})g(X_{i'r},X_{i'r'})+\frac{1}{4}\right\}\Bigg] = 0\end{aligned}$$

Finally,

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=1}^{n} Z_i^{(7)}\right)^2\right] &= \sum_{i=1}^{n}\sum_{i'=1}^{n}\mathbb{E}\left[Z_i^{(7)}Z_{i'}^{(7)}\right] \\ &= \sum_{i=1}^{n}\sum_{i'=1}^{n}\sum_{n_\ell+1\leq j\neq j'\leq n_{\ell+1}}\sum_{\substack{k=1\\k\neq j\\k\neq j'}}^{n}\sum_{\substack{k'=1\\k'\neq j'\\k'\neq j\\k'\neq k}}^{n}\sum_{n_\ell+1\leq r\neq r'\leq n_{\ell+1}}\sum_{\substack{s=1\\s\neq r\\s\neq r'}}^{n}\sum_{\substack{s'=1\\s'\neq r'\\s'\neq r\\s'\neq s}}^{n}\mathbb{E}\Big[g(X_{i,k},X_{i,j})g(X_{i,k'},X_{i,j'}) \\ &\qquad g(X_{i's},X_{i'r})g(X_{i's'},X_{i'r'})\Big].\end{aligned}$$

The above expectation is null when the the cardinality of the set of indices $\{i,i',j,j',k,k',r,r',s,s'\}$ is equal or greater than 8, by using Assertion (i) of Lemma 2. Observe moreover that

$$|g(x,y)g(z,t)g(x',y')g(z',t')|\,| \leq 1/16 \leq 1,$$

for all $x,y,z,t,x',y',z',t' \in \mathbb{R}$. Therefore, we get,

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

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

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

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

6.3. Proof of Theorem 3

For all $\delta > 0$, let us define

$$\mathcal{C}_{n_1^*,\delta} = \left\{n_1 \in \{1,\ldots,n-1\} : |n_1 - n_1^*| \geq n\delta\right\}.$$

Note that

$$\begin{aligned}
&\mathbb{P}\big(|\hat{n}_1 - n_1^*| \geq n\delta\big)\\
&\leq \mathbb{P}\left(\max_{n_1\in\mathcal{C}_{n_1^*,\delta}}\left\{S_n(n_1) - S_n(n_1^*)\right\} \geq 0\right)\\
&\leq \mathbb{P}\left(\max_{n_1\in\mathcal{C}_{n_1^*,\delta}}\left\{S_n(n_1) - S_n(n_1^*) - \mathbb{E}\left[S_n(n_1) - S_n(n_1^*)\right] + \mathbb{E}\left[S_n(n_1) - S_n(n_1^*)\right]\right\} \geq 0\right)\\
&\leq \mathbb{P}\left(\max_{n_1\in\mathcal{C}_{n_1^*,\delta}}\left\{S_n(n_1) - S_n(n_1^*) - \mathbb{E}\left[S_n(n_1) - S_n(n_1^*)\right]\right\} \geq -\max_{n_1\in\mathcal{C}_{n_1^*,\delta}}\left\{\mathbb{E}\left[S_n(n_1) - S_n(n_1^*)\right]\right\}\right).
\end{aligned}$$

By Proposition 1 given below,

$$\begin{aligned}
&\mathbb{P}\big(|\hat{n}_1 - n_1^*| \geq n\delta\big)\\
&\leq \mathbb{P}\left(\max_{n_1\in\mathcal{C}_{n_1^*,\delta}}\left\{S_n(n_1) - S_n(n_1^*) - \mathbb{E}\left[S_n(n_1) - S_n(n_1^*)\right]\right\} \geq \kappa' n^2\delta\right)
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{P}\big(|\hat{n}_1 - n_1^*| \geq n\delta\big) \\
&\leq \mathbb{P}\left(\max_{n_1 \in \mathcal{C}_{n_1^*,\delta}} \Big| S_n(n_1) - S_n(n_1^*) - \mathbb{E}\left[S_n(n_1) - S_n(n_1^*)\right] \Big| \geq \kappa' n^2 \delta\right) \\
&\leq \mathbb{P}\left(\max_{n_1 \in \mathcal{C}_{n_1^*,\delta}} \Big| S_n(n_1) - \mathbb{E}\left[S_n(n_1)\right] \Big| \geq \frac{\kappa'}{2} n^2 \delta\right) \\
&\quad + \mathbb{P}\left(\Big| S_n(n_1^*) - \mathbb{E}\left[S_n(n_1^*)\right] \Big| \geq \frac{\kappa'}{2} n^2 \delta\right) \\
&\leq \sum_{n_1 \in \mathcal{C}_{n_1^*,\delta}} \mathbb{P}\left(\Big| S_n(n_1) - \mathbb{E}\left[S_n(n_1)\right] \Big| \geq \frac{\kappa'}{2} n^2 \delta\right) \\
&\quad + \mathbb{P}\left(\Big| S_n(n_1^*) - \mathbb{E}\left[S_n(n_1^*)\right] \Big| \geq \frac{\kappa'}{2} n^2 \delta\right).
\end{aligned}$$

We mimic the proof of Theorem 1 to obtain that

$$\begin{aligned}
&\sum_{n_1 \in \mathcal{C}_{n_1^*,\delta}} \mathbb{P}\left(\left| S_n(n_1) - \mathbb{E}\left[S_n(n_1)\right] \right| \geq \frac{\kappa'}{2} n^2 \delta\right) \\
&\leq \sum_{n_1 \in \mathcal{C}_{n_1^*,\delta}} \frac{2}{\kappa' n^2 \delta} \mathbb{E}\left[\left| S_n(n_1) - \mathbb{E}\left[S_n(n_1)\right] \right|^2\right]^{1/2} \\
&\leq \sum_{n_1 \in \mathcal{C}_{n_1^*,\delta}} \frac{2n^{7/2}}{\kappa' n^2 \delta n n_1 (n - n_1)} = \frac{2\sqrt{n}}{\kappa' \delta} \sum_{n_1 \in \mathcal{C}_{n_1^*,\delta}} \frac{1}{n_1 (n - n_1)}.
\end{aligned}$$

Observing that

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

as n tends to infinity, concludes the proof.

In the following, we establish the proposition that we use for proving Theorem 3.

Proposition 1. *Under the assumptions of Theorem 3, there exists a positive constant κ, such that*

$$\mathbb{E}\left[S_n(n_1) - S_n(n_1^*)\right] = -\kappa n |n_1^* - n_1| \left(1 + \varepsilon_n(n_1)\right),$$

where $\max_{n_1 \in \mathcal{C}_{n_1^*,\delta}} |\varepsilon_n(n_1)| \to 0$, as n tends to infinity.

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

$$\begin{aligned}
\mathbb{E}\left[S_n(n_1^*)\right] &= \sum_{i=1}^{n} \mathbb{E}\left[U_{n,i}^2(n_1^*)\right] \\
&= \frac{1}{nn_1^*(n-n_1^*)} \sum_{i=1}^{n} \mathbb{E}\left[\left(\sum_{j_0=1}^{n_1^*} \sum_{j_1=n_1^*+1}^{n} h(X_{i,j_0}, X_{i,j_1})\right)^2\right] \\
&= \frac{1}{nn_1^*(n-n_1^*)} \sum_{i=1}^{n} \sum_{1\le j_0,k_0\le n_1^*} \sum_{n_1^*+1\le j_1,k_1\le n} \mathbb{E}\left[h(X_{i,j_0}, X_{i,j_1})h(X_{i,k_0}, X_{i,k_1})\right] \\
&= \frac{1}{nn_1^*(n-n_1^*)} \sum_{i=1}^{n} \left\{ \sum_{j_0=1}^{n_1^*} \sum_{j_1=n_1^*+1}^{n} \mathbb{E}\left[h^2(X_{i,j_0}, X_{i,j_1})\right] \right. \\
&\quad + \sum_{j_0=1}^{n_1^*} \sum_{n_1^*+1\le j_1\ne k_1\le n} \mathbb{E}\left[h(X_{i,j_0}, X_{i,j_1})h(X_{i,j_0}, X_{i,k_1})\right] \\
&\quad + \sum_{1\le j_0\ne k_0\le n_1^*} \sum_{j_1=n_1^*+1}^{n} \mathbb{E}\left[h(X_{i,j_0}, X_{i,j_1})h(X_{i,k_0}, X_{i,j_1})\right] \\
&\quad \left. + \sum_{1\le j_0\ne k_0\le n_1^*} \sum_{n_1^*+1\le j_1\ne k_1\le n} \mathbb{E}\left[h(X_{i,j_0}, X_{i,j_1})h(X_{i,k_0}, X_{i,k_1})\right] \right\} \\
&= \frac{A+B}{nn_1^*(n-n_1^*)},
\end{aligned}$$

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. Let us 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

$$
\begin{aligned}
A = \sum_{i=1}^{n_1^*} \Bigg\{ & \sum_{j_0=1}^{n_1^*} \sum_{j_1=n_1^*+1}^{n} \mathbb{E}\left[h^2(X_{i,j_0}, X_{i,j_1})\right] + \sum_{j_0=1}^{n_1^*} \sum_{n_1^*+1 \le j_1 \ne k_1 \le n} \mathbb{E}\left[h(X_{i,j_0}, X_{i,j_1}) h(X_{i,j_0}, X_{i,k_1})\right] \\
& + \sum_{1 \le j_0 \ne k_0 \le n_1^*} \sum_{j_1=n_1^*+1}^{n} \mathbb{E}\left[h(X_{i,j_0}, X_{i,j_1}) h(X_{i,k_0}, X_{i,j_1})\right] \\
& + \sum_{1 \le j_0 \ne k_0 \le n_1^*} \sum_{n_1^*+1 \le j_1 \ne k_1 \le n} \mathbb{E}\left[h(X_{i,j_0}, X_{i,j_1}) h(X_{i,k_0}, X_{i,k_1})\right] \Bigg\} \\
= n_1^* \Big\{ & n_1^*(n-n_1^*)\mathbb{E}\left[h^2(W,Y)\right] + n_1^*(n-n_1^*)(n-n_1^*-1)\mathbb{E}\left[h(W,Y)h(W,Y^{(1)})\right] \\
& + n_1^*(n_1^*-1)(n-n_1^*)\mathbb{E}\left[h(W,Y)h(W^{(1)},Y)\right] \\
& + n_1^*(n_1^*-1)(n-n_1^*)(n-n_1^*-1)\mathbb{E}\left[h(W,Y)h(W^{(1)},Y^{(1)})\right] \Big\}.
\end{aligned}
$$

In the same manner, we can see that,

$$
\begin{aligned}
B = (n-n_1^*) \Big\{ & n_1^*(n-n_1^*)\mathbb{E}\left[h^2(Y,Z)\right] + n_1^*(n-n_1^*)(n-n_1^*-1)\mathbb{E}\left[h(Y,Z)h(Y,Z^{(1)})\right] \\
& + n_1^*(n_1^*-1)(n-n_1^*)\mathbb{E}\left[h(Y,Z)h(Y^{(1)},Z)\right] \\
& + n_1^*(n_1^*-1)(n-n_1^*)(n-n_1^*-1)\mathbb{E}\left[h(Y,Z)h(Y^{(1)},Z^{(1)})\right] \Big\}.
\end{aligned}
$$

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 $n_1 < n_1^*$.

$$\begin{aligned}
\mathbb{E}\left[S_n(n_1)\right] &= \frac{1}{nn_1(n-n_1)} \sum_{i=1}^{n} \mathbb{E}\left[\left(\sum_{j_0=1}^{n_1} \sum_{j_1=n_1+1}^{n} h(X_{i,j_0}, X_{i,j_1})\right)^2\right] \\
&= \frac{1}{nn_1(n-n_1)} \sum_{i=1}^{n} \sum_{1\le j_0,k_0\le n_1} \sum_{n_1+1\le j_1,k_1\le n} \mathbb{E}\left[h(X_{i,j_0}, X_{i,j_1}) h(X_{i,k_0}, X_{i,k_1})\right] \\
&= \frac{1}{nn_1(n-n_1)} \sum_{i=1}^{n} \Bigg\{ \sum_{j_0=1}^{n_1} \sum_{j_1=n_1+1}^{n} \mathbb{E}\left[h^2(X_{i,j_0}, X_{i,j_1})\right] \\
&\quad + \sum_{j_0=1}^{n_1} \sum_{n_1+1\le j_1\ne k_1\le n} \mathbb{E}\left[h(X_{i,j_0}, X_{i,j_1}) h(X_{i,j_0}, X_{i,k_1})\right] \\
&\quad + \sum_{1\le j_0\ne k_0\le n_1} \sum_{j_1=n_1+1}^{n} \mathbb{E}\left[h(X_{i,j_0}, X_{i,j_1}) h(X_{i,k_0}, X_{i,j_1})\right] \\
&\quad + \sum_{1\le j_0\ne k_0\le n_1} \sum_{n_1+1\le j_1\ne k_1\le n} \mathbb{E}\left[h(X_{i,j_0}, X_{i,j_1}) h(X_{i,k_0}, X_{i,k_1})\right] \Bigg\} \\
&= \frac{C+D}{nn_1(n-n_1)},
\end{aligned}$$

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. Observe that

$$
\begin{aligned}
C = \sum_{i=1}^{n_1^*} \Bigg\{ & \sum_{j_0=1}^{n_1} \sum_{j_1=n_1+1}^{n_1^*} \mathbb{E}\left[h^2(X_{i,j_0}, X_{i,j_1})\right] + \sum_{j_0=1}^{n_1} \sum_{j_1=n_1^*+1}^{n} \mathbb{E}\left[h^2(X_{i,j_0}, X_{i,j_1})\right] \\
& + \sum_{j_0=1}^{n_1} \sum_{n_1+1 \le j_1 \ne k_1 \le n_1^*} \mathbb{E}\left[h(X_{i,j_0}, X_{i,j_1}) h(X_{i,j_0}, X_{i,k_1})\right] \\
& + \sum_{j_0=1}^{n_1} \sum_{n_1^*+1 \le j_1 \ne k_1 \le n} \mathbb{E}\left[h(X_{i,j_0}, X_{i,j_1}) h(X_{i,j_0}, X_{i,k_1})\right] \\
& + 2 \sum_{j_0=1}^{n_1} \sum_{j_1=n_1+1}^{n_1^*} \sum_{k_1=n_1^*+1}^{n} \mathbb{E}\left[h(X_{i,j_0}, X_{i,j_1}) h(X_{i,j_0}, X_{i,k_1})\right] \\
& + \sum_{1 \le j_0 \ne k_0 \le n_1} \sum_{j_1=n_1+1}^{n_1^*} \mathbb{E}\left[h(X_{i,j_0}, X_{i,j_1}) h(X_{i,k_0}, X_{i,j_1})\right] \\
& + \sum_{1 \le j_0 \ne k_0 \le n_1} \sum_{j_1=n_1^*+1}^{n} \mathbb{E}\left[h(X_{i,j_0}, X_{i,j_1}) h(X_{i,k_0}, X_{i,j_1})\right] \\
& + \sum_{1 \le j_0 \ne k_0 \le n_1} \sum_{n_1+1 \le j_1 \ne k_1 \le n_1^*} \mathbb{E}\left[h(X_{i,j_0}, X_{i,j_1}) h(X_{i,k_0}, X_{i,k_1})\right] \\
& + \sum_{1 \le j_0 \ne k_0 \le n_1} \sum_{n_1^*+1 \le j_1 \ne k_1 \le n} \mathbb{E}\left[h(X_{i,j_0}, X_{i,j_1}) h(X_{i,k_0}, X_{i,k_1})\right] \\
& + 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} \mathbb{E}\left[h(X_{i,j_0}, X_{i,j_1}) h(X_{i,k_0}, X_{i,k_1})\right] \Bigg\} \\
= n_1^* \Big\{ & n_1(n_1^* - n_1) \mathbb{E}\left[h^2(W, W^{(1)})\right] + n_1(n - n_1^*) \mathbb{E}\left[h^2(W, Y)\right] \\
& n_1(n_1^* - n_1)(n_1^* - n_1 - 1) \mathbb{E}\left[h(W, W^{(1)}) h(W, W^{(2)})\right] \\
& + n_1(n - n_1^*)(n - n_1^* - 1) \mathbb{E}\left[h(W, Y) h(W, Y^{(1)})\right] \\
& + 2n_1(n_1^* - n_1)(n - n_1^*) \mathbb{E}\left[h(W, W^{(1)}) h(W, Y)\right] \\
& + n_1(n_1 - 1)(n_1^* - n_1) \mathbb{E}\left[h(W, W^{(1)}) h(W^{(2)}, W^{(1)})\right] \\
& + n_1(n_1 - 1)(n - n_1^*) \mathbb{E}\left[h(W, Y) h(W^{(1)}, Y)\right] \\
& + n_1(n_1 - 1)(n_1^* - n_1)(n_1^* - n_1 - 1) \mathbb{E}\left[h(W, W^{(1)}) h(W^{(2)}, W^{(3)})\right] \\
& + n_1(n_1 - 1)(n - n_1^*)(n - n_1^* - 1) \mathbb{E}\left[h(W, Y) h(W^{(1)}, Y^{(1)})\right] \\
& + 2n_1(n_1 - 1)(n_1^* - n_1)(n - n_1^*) \mathbb{E}\left[h(W, W^{(1)}) h(W^{(2)}, Y)\right] \Big\}.
\end{aligned}
$$

In the same manner, we can see that

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

where the absolute values of the above expectations are bounded by 1.

By Lemma 1, we get that

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

Hence,

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

Note that $\mathbb{E}\left[h(W,Y)\right] \neq 0$ or $\mathbb{E}\left[h(Y,Z)\right] \neq 0$ by (8) since

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

Hence, there exists a positive constant κ, such that

$$\mathbb{E}\left[S_n(n_1) - S_n(n_1^*)\right] = -\kappa n(n_1^* - n_1)\left(1 + \varepsilon_n(n_1)\right),$$

where $\max_{n_1 \in \mathcal{C}_{n_1^*,\delta}} |\varepsilon_n(n_1)| \to 0$, as n tends to infinity, since

$$1 - \frac{n_1^*}{n} \leq \frac{n - n_1^*}{n - n_1} \leq 1$$

and $n_1^*/n \to \tau_1^*$.

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

6.4. Proof of Equation (10)

By (9),

$$I_0(p) = \max_{1\leq n_1 = p} \Delta(1:n_1) = \Delta(1:p)$$

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 (10) when $L = 1$. By (9),

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

By using the previous expression of $I_1(p)$, we get that

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

which is (10) when $L = 2$ and so on, which gives (10).

7. Technical lemmas

Lemma 1. *Let h be defined by $h(x,y) = \mathbb{1}_{\{x\leq y\}} - \mathbb{1}_{\{y\leq x\}}$. Then,*

(i) $\mathbb{E}\left[h(X,Y)\right] = 0$,

(ii) $h^2(X,Y) = 1$ *a.s.*,

(iii) $\mathbb{E}\left[h(X,Y)h(X,Z)\right] = 1/3$,

(iv) $\mathbb{E}\left[h(X,Y)h(Z,Y)\right] = 1/3$,

(v) $\mathbb{E}\left[h(X,Y)h(Z,T)\right] = 0$,

where X, Y, Z and T are i.i.d. random variables having a continuous distribution function.

Proof. (i) Let X and Y be i.i.d. random variables with cumulative distribution function F. We have:

$$\mathbb{E}\left[h(X,Y)\right] = \mathbb{E}\left[\mathbb{1}_{\{X\leq Y\}}\right] - \mathbb{E}\left[\mathbb{1}_{\{Y\leq X\}}\right] = \mathbb{E}\left[1 - 2F(X)\right] = 0,$$

where we used that $F(X)$ is a uniform random variable on $[0,1]$.

(ii) For all $x \neq y$ in $\mathbb{R}$, $h^2(x,y) = \left(\mathbb{1}_{\{x\leq y\}} - \mathbb{1}_{\{y\leq x\}}\right)^2 = \mathbb{1}_{\{x\leq y\}} + \mathbb{1}_{\{y\leq x\}} - 2\mathbb{1}_{\{x\leq y\}}\mathbb{1}_{\{y\leq x\}} = 1$. Consequently, $h^2(X,Y) = 1$ a.s..

(iii) Let X, Y and Z be i.i.d. random variables with cumulative distribution function F. We have:

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

where we used that $F(X)$ is a uniform random variable on $[0,1]$.

(iv) Since $\mathbb{E}\left[h(X,Y)h(Z,Y)\right] = \mathbb{E}\left[h(Y,X)h(Y,Z)\right] = 1/3$, the result comes from (iii).

(v) By independance of (X,Y) with (Z,T),

$$\mathbb{E}\left[h(X,Y)h(Z,T)\right] = \mathbb{E}\left[h(X,Y)\right]\mathbb{E}\left[h(Z,T)\right] = 0.$$

□

Lemma 2. *Let us define the function g as $g(x,y) = \mathbb{1}_{\{x\leq y\}} - \frac{1}{2}$. Let X, Y and Z be i.i.d. random variables having a continuous distribution function. Then*

(i) $\mathbb{E}\left[g(X,Y)\right] = 0$,

(ii) $g(X,Y)^2 = \frac{1}{4}$ *a.s.*,

(iii) $\mathbb{E}\left[g(X,Y)g(Z,Y)\right] = \frac{1}{12}$,

(iv) $\mathbb{E}\left[g(X,Y)g(X,Z)\right] = \frac{1}{12}$.

Proof. (i) $\mathbb{E}\left[g(X,Y)\right] = \mathbb{E}\left[F(Y)\right] - 1/2 = 0$, since $F(Y)$ is a uniform random variable on $[0,1]$.

(ii) For all x, y in $\mathbb{R}$, $g(x,y)^2 = \left(\mathbb{1}_{\{x\leq y\}} - \frac{1}{2}\right)^2 = \mathbb{1}_{\{x\leq y\}} + \frac{1}{4} - \mathbb{1}_{\{x\leq y\}} = \frac{1}{4}$. Consequently, $g^2(X,Y) = \frac{1}{4}$ a.s..

(iii) Let X, Y and Z be i.i.d. random variables with cumulative distribution function F. We have:

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

where we used that $F(X)$ is a uniform random variable on $[0,1]$.

(iv) Note that

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

by (iii).

□

Bai, J., 2010. Common breaks in means and variances for panel data. Journal of Econometrics 157 (1), 78 – 92, nonlinear and Nonparametric Methods in Econometrics.

Basseville, M., Nikiforov, I. V., 1993. Detection of Abrupt Changes: Theory and Applications. Prentice-Hall.

Baudry, J.-P., Maugis, C., Michel, B., 2012. Slope heuristics: overview and implementation. Statistics and Computing 22 (2), 455–470.

Bellman, R., 1961. On the approximation of curves by line segments using dynamic programming. Communications of the ACM 4 (6), 284.

Cho, H., Fryzlewicz, P., 2015. Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 77 (2), 475–507.

Cleynen, A., Dudoit, S., Robin, S., 2013. Comparing segmentation methods for genome annotation based on rna-seq data. Journal of Agricultural, Biological, and Environmental Statistics 19 (1), 101–118.

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 (7398), 376–380.

Horvath, L., Huskova, M., 2012. Change-point detection in panel data. Journal of Time Series Analysis 33 (4), 631–648.

Jirak, M., 12 2015. Uniform change point tests in high dimension. The Annals of Statistics 43 (6), 2451–2483.

Kay, S., 1993. Fundamentals of statistical signal processing: detection theory. Prentice-Hall, Inc.

Killick, R., Fearnhead, P., Eckley, I. A., 2012. Optimal detection of changepoints with a linear computational cost. Journal of the American Statistical Association 107 (500), 1590–1598.

Lehmann, E. L., D'Abrera, H. J., 2006. Nonparametrics: statistical methods based on ranks. Springer New York.

Lévy-Leduc, C., Delattre, M., Mary-Huard, T., Robin, S., 2014. Two-dimensional segmentation for analyzing HiC data. Bioinformatics 30 (17), 386–392.

Lévy-Leduc, C., Roueff, F., 2009. Detection and localization of change-points in high-dimensional network traffic data. Ann. Applied Statist. 3 (2), 637–662.

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., et al., 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. science 326 (5950), 289–293.

Lung-Yut-Fong, A., Lévy-Leduc, C., Cappé, O., 2015. Homogeneity and change-point detection tests for multivariate data using rank statistics. Journal de la Société Française de Statistique 156 (4), 133–162.

Matteson, D. S., James, N. A., 2014. A nonparametric approach for multiple change point analysis of multivariate data. Journal of the American Statistical Association 109 (505), 334–345.

Picard, F., Robin, S., Lavielle, M., Vaisse, C., Daudin, J.-J., 2005. A statistical approach for array CGH data analysis. BMC Bioinformatics 6 (1), 27.

Szekely, J. G., Rizzo, L. M., 2005. Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method. Journal of Classification 22 (2), 151–183.

Tartakovsky, A., Rozovskii, B., Blazek, R., Kim, H., 2006. 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 (9), 3372 – 3382.

van der Vaart, A. W., 1998. Asymptotic statistics. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Vert, J., Bleakley, K., 2010. Fast detection of multiple change-points shared by many signals using group LARS. In: Advances in Neural Information Processing Systems 23. pp. 2343–2351.