Journal of Multivariate Analysis 165 (2018) 143–165

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

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, France*

ARTICLE INFO

Article history:
Received 20 May 2016
Available online 23 December 2017

AMS subject classifications:
62H10
62G20
62H12
62H15
62G10

Keywords:
Hi-C data
Multiple change-point estimation
Nonparametric estimation

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.

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 [1], anomaly detection in internet traffic data [14,20], 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 [18,22], RNAseq data [6], 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 [15]. 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 [7] 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. [13] 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

[*] Corresponding author.
E-mail address: vincent.brault@univ-grenoble-alpes.fr (V. Brault).

https://doi.org/10.1016/j.jmva.2017.12.005

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 [8] 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 [5] 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 [9] proposed nonparametric change-point tests in very general high-dimensional settings. Matteson and James [17] 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 [19]. Another approach based on ranks has been proposed by Lung-Yut-Fong et al. [16] in the same framework as [17]. Their approach consists in extending the classical Wilcoxon and Kruskal–Wallis statistics [12] to the multivariate case.

In this paper, we propose a nonparametric change-point estimation approach based on nonparametric homogeneity tests generalizing the approach of [16] 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 (i.i.d.) 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 $\mathbb{P}_1^i$ and $X_{i,n_1+1}, \ldots, X_{i,n}$ have distribution $\mathbb{P}_2^i$, with $\mathbb{P}_1^i \neq \mathbb{P}_2^i$.

To decide whether $\mathcal{H}_0$ should be rejected or not, we propose to use a test statistic inspired by the one designed by [16] 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

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

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

Our framework is different from that of Lung-Yut-Fong et al. [16] 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 $R_j^{(i)}$ of $X_{i,j}$ among $X_{i,1}, \ldots, X_{i,n}$. Indeed,

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