Journal of Multivariate Analysis 167 (2018) 319–330

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

A model selection approach for multiple sequence segmentation and dimensionality reduction

Bruno M. Castro [a,*], Renan B. Lemes [b], Jonatas Cesar [b], Tábita Hünemeier [b], Florencia Leonardi [c]

[a] Departamento de Estatística, Universidade Federal do Rio Grande do Norte, Brazil
[b] Instituto de Biociências, Universidade de São Paulo, Brazil
[c] Instituto de Matemática e Estatística, Universidade de São Paulo, Brazil

ARTICLE INFO

Article history:
Received 30 April 2018
Available online 22 May 2018

AMS 2010 subject classifications:
62G05
62G20

ABSTRACT

In this paper we consider the problem of segmenting n aligned random sequences of equal length m into a finite number of independent blocks. We propose a penalized maximum likelihood criterion to infer simultaneously the number of points of independence as well as the position of each point. We show how to compute exactly the estimator by means of a dynamic programming algorithm with time complexity $O(m^2 n)$. We also propose another method, called hierarchical algorithm, that provides an approximation to the estimator when the sample size increases and runs in time $O\{m \ln(m) n\}$. Our main theoretical results are the strong consistency of both estimators when the sample size n grows to infinity. We illustrate the convergence of these algorithms through some simulation examples and we apply the method to identify recombination hotspots in real SNPs data.

© 2018 Elsevier Inc. All rights reserved.

1. Introduction

The problem of multiple sequence segmentation and dimensionality reduction is of crucial importance for many applied areas, including the analysis of multiple alignments of DNA/RNA and Amino Acid (AA) sequences. In these cases, one of the main goals is to investigate some aspects of the genetic variation, for example, inferring which genomic regions can be considered putative hotspots of genetic recombination. Another application on practical ground is to look for small subregions in the sequences that are related to a phenotypic variable. One example of this is the genome-wide association studies (GWAS) of Single Nucleotide Polymorphisms (SNPs), where the interest is to find positions in the genome associated with a given phenotypic trait. Traditionally, this task is performed by making a simultaneous hypotheses test on each individual position or on small sub-windows of fixed length, as in the PLINK suite [6,20]. But considering all variables as mutually independent does not translate the intrinsic relations present in genomic data and can result in weak or spurious discoveries. This fact has led the community to develop methods that take into account the dependence between adjacent or even non-adjacent variables; see, e.g., [16].

Many other authors have also considered the problem of inferring local dependencies in data, using a wide range of probabilistic models. In a recent paper, Algama and Keith [1] present a detailed review about the most well-known sequence segmentation techniques and the models assumed in each case. Their list contains sliding window analysis [22], hidden Markov models [3,10], recursive segmentation algorithms [8,18] and multiple change-point analysis [9,21]. They also refer to other methods for sequence segmentation and pattern identification based on least squares estimation [13] or on wavelet

* Corresponding author.
E-mail address: bruno.monte19@gmail.com (B.M. Castro).

https://doi.org/10.1016/j.jmva.2018.05.006
0047-259X/© 2018 Elsevier Inc. All rights reserved.

analysis [23]. We refer the reader to the work [1] where a brief explanation of these methods is presented, and also other references for the problem of sequence segmentation are given.

Our main goal in this paper is to introduce a new approach for the problem of multiple sequence segmentation into independent blocks. We are interested in inferring the maximal set of points of independence, when the number of such points is unknown. To do this we propose a penalized maximum likelihood criterion to infer simultaneously the number of points of independence and their positions, for n aligned random sequences of equal length m. We show how to compute exactly this estimator by means of a dynamic programming algorithm and we prove its almost sure convergence to the true set of points of independence when the sample size n increases. In cases where the size m of the sequences is large, we propose a suboptimal but more efficient algorithm that also converges almost surely to the set of points of independence when the sample size n increases. The main advantage of our procedure is that we do not need to assume a fixed number of segments and the optimal number of points of independence can be learned from the data. Our method can be used to reduce drastically the dimensionality of the joint probability distributions from exponential to linear functions of the length of the sequences, given by m, somehow sharing the same objectives of correspondence analysis, a principal component method for nominal categorical data.

A related approach is considered by Gwadera et al. [12], who present a method to determine the optimal number of segments in a sequence using a Variable Length Markov Chain (VLMC) model on each segment. They propose to use the Bayesian Information Criterion (BIC) and a variant of the Minimum Description Length (MDL) Principle to select the number of segments for the given sequence. Their method consists in estimating change points on a unique stationary sequence while ours looks for points of independence on non-stationary aligned sequences. In stark contrast to their approach, we do not need to assume a specific probabilistic model on each segment and we can estimate a general multivariate distribution on each segment. Moreover, Gwadera et al. [12] do not present a formal proof that their method succeeds to detect the number and position of the change-points.

This paper is organized as follows. In Section 2 we present background material, show how to compute the estimators, and state the main theoretical results. In Section 3 we report the results of simulations illustrating the performance of the segmentation method, and in Section 4 we show a practical application on real data. In Section 5 we discuss the results and in the Appendix we include the proofs of the theoretical results presented in Section 2.

2. Likelihood function and model selection

2.1. Notation and definitions

Let $\mathbf{X} = (X_1, \ldots, X_m)$ be a random vector taking values in $A_1 \times \cdots \times A_m$, where A_i is a finite alphabet for all $i \in \{1, \ldots, m\}$. The cardinal of the finite set A_i will be denoted by $|A_i|$. We say that $j \in \{1, \ldots, m-1\}$ is a point of independence for $\mathbf{X}$ if the random vectors $(X_1, \ldots, X_j)$ and $(X_{j+1}, \ldots, X_m)$ are independent.

Given two integers $r \leq s$, denote by $r:s$ the integer interval $r, \ldots, s$. We say $U_{r:s} \subset r : (s-1)$ is a maximal set of points of independence for the interval $r:s$ if no $v \in r : (s-1) \setminus U_{r:s}$ is a point of independence for $\mathbf{X}$. For each random vector $\mathbf{X}$ and each interval $r:s$ there is only one maximal set of points of independence; from now on this special set will be denoted by $U^*_{r:s}$. In the special case $r = 1, s = m$ we will simply write U^*.

Without loss of generality we will also suppose that the set $U_{r:s}$ is ordered; in this case $U_{r:s} = (u_1, \ldots, u_k)$ with $u_i < u_j$ if $i < j$. From $U_{r:s}$ it is possible to obtain the set of blocks of independent variables as the set $B(U_{r:s}) = \{I_1, \ldots, I_{k+1}\}$ of integer intervals given by $I_1 = r : u_1$, $I_i = (u_{i-1} + 1) : u_i$ for all $i \in \{2, \ldots, k\}$, and $I_{k+1} = (u_k + 1) : s$.

Given an integer interval $I = r:s$ denote by A^I the set of finite strings on $A_r \times \cdots \times A_s$ with positive probability, viz.

$$A^I = \{w \in A_r \times \cdots \times A_s : \Pr(w) > 0\}.$$

Assume we observe an iid sample $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ of $\mathbf{X}$, denoted by $\mathbf{x}$. Then, the likelihood function for the set U can be written as

$$L(U; \mathbf{x}) = \prod_{i=1}^{n} \prod_{I \in B(U)} \Pr(X_j = x_j^{(i)} : j \in I). \tag{1}$$

Denote by $\mathbf{x}_{r:s}$ the iid sample $\mathbf{x}_{r:s}^{(1)}, \ldots, \mathbf{x}_{r:s}^{(n)}$. Given a finite string $a_{r:s} \in A^{r:s}$, define

$$N(a_{r:s}) = \sum_{i=1}^{n} \mathbf{1}\{x_{r:s}^{(i)} = a_{r:s}\}.$$

Then $L(U; \mathbf{x})$ can be rewritten as

$$L(U; \mathbf{x}) = \prod_{I \in B(U)} \prod_{a_I \in A^I} \Pr(X_I = a_I)^{N(a_I)}. \tag{2}$$

Denote by $\widehat{\Pr}(a_I)$ the maximum likelihood estimators for the probabilities $\Pr(a_I)$, i.e., the values maximizing (2). It can be proved that for any interval I and any $a_I \in A^I$, the estimator $\widehat{\Pr}(a_I)$ is given, for all $a_I \in A^I$, by

$$\widehat{\Pr}(a_I) = N(a_I)/n.$$

Download English Version:

https://daneshyari.com/en/article/7546554

Download Persian Version:

https://daneshyari.com/article/7546554

Daneshyari.com