Signal Processing 98 (2014) 233–242

Contents lists available at ScienceDirect

Signal Processing

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

Fast estimation of the Integrated Completed Likelihood criterion for change-point detection problems with applications to Next-Generation Sequencing data

Alice Cleynen [a,b,*], The Minh Luong [c], Guillem Rigaill [d], Gregory Nuel [c]

[a] AgroParisTech, UMR 518 MIA, 16, rue Claude Bernard, 75005 Paris, France
[b] INRA, UMR 518 MIA, 16, rue Claude Bernard, 75005 Paris, France
[c] MAP5 - UMR CNRS 8145, Université Paris Descartes, Paris, France
[d] URGV INRA-CNRS-Université d'Évry Val d'Essonne, 2 Rue Gaston Crémieux, 91057 Evry Cedex, France

ARTICLE INFO

Article history:
Received 23 June 2013
Received in revised form
20 November 2013
Accepted 26 November 2013
Available online 10 December 2013

Keywords:
Hidden Markov model
Integrated Completed Likelihood
Model selection
Negative binomial
Segmentation

ABSTRACT

In this paper, we consider the Integrated Completed Likelihood (ICL) as a useful criterion for estimating the number of changes in the underlying distribution of data, specifically in problems where detecting the precise location of these changes is the main goal. The exact computation of the ICL requires $\mathcal{O}(Kn^2)$ operations (with K the number of segments and n the number of data-points) which is prohibitive in many practical situations with large sequences of data. We describe a framework to estimate the ICL with $\mathcal{O}(K^2n)$ complexity. Our approach is general in the sense that it can accommodate any given model distribution. We checked the run-time and validity of our approach on simulated data and demonstrate its good performance when analyzing real Next-Generation Sequencing (NGS) data using a negative binomial model. Our method is implemented in the R package `postCP` and available on the CRAN repository.

1. Introduction

The estimation of the number of segments is a central aspect in change-point methodology. For instance, in the context of Comparative Genomic Hybridization-array or Next-Generation Sequencing experiments, identifying the number and corresponding location of segments is crucial as the segments may relate to a biological event of interest. This theoretically complex problem can be handled in the more general context of model selection, leading to the use of *ad hoc* procedures in practical situations.

Among the procedures are the use of classical criteria based on penalized likelihoods such as the Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC or SIC, [1]). However, when choosing the number of segments, the BIC criterion uses a Laplace approximation requiring differentiability conditions of the likelihood function which are not satisfied by the model due to the discrete nature of the change-points, and thus which may not be appropriate when the number of observations in each segment is unequal and unknown. These criteria also tend to overestimate the number of segments as they do not account for the contiguous nature of observations within a segment. With this issue in mind, [2] proposed a modified BIC criterion using a Brownian motion model with changing drift for the specific case of normal data.

For this reason, there has been an extensive literature influenced by Birgé and Massart [3] which proposes new penalty shapes and constants in order to select a lower number of segments in the profile. The idea is to approximate the true distribution of the data by a piece-wise constant distribution and to choose the model that, within

[*] Corresponding author.
E-mail address: alice.cleynen@agroparistech.fr (A. Cleynen).

0165-1684/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.sigpro.2013.11.029

a set of models (*i.e.* among a set of piece-wise constant distributions), performs closest to the true value by deriving a tight upper bound on the estimation error. In this case, the bound is the variance term of the risk of the estimator. This leads to penalties that generally depend only on the number of segments K, and whose constants can be chosen adaptively to the data [4,5]. However, a large proportion of those methods focused on normal data, and are not applicable to count datasets modeled by the Poisson or the negative binomial distributions.

Other approaches for model selection appearing in the literature include sequential likelihood ratio tests [6] and Bayesian approaches through estimating model posterior probabilities by various MCMC methods [7–10]. However, the Bayesian approaches are often computationally intensive as they require re-sampling.

In the context of incomplete data models (e.g. mixture model for clustering) [11] proposed a model selection criterion accounting for both observed and unobserved variables based on the Integrated Completed Likelihood (ICL): $\sum_S \mathbb{P}(S|X) \log \mathbb{P}(S|X)$ where X are the observations and S are the corresponding (unknown) clustering membership. The goal of the ICL is to select a relevant number of clusters and it has been shown to be more robust than BIC to violation of some of the mixture model assumptions. Importantly a relevant number of clusters (or in our case segments) can differ from the true number of clusters. In other words, certain clusters may not be relevant if their corresponding location or length cannot be precisely identified due to their size or the difference between adjacent clusters being very small. In this situation, imposing the true number of segments on the change-point estimates could result in locations not corresponding to a true shift in the distribution; it is thus advantageous to identify a smaller number of segments with more certain change-point locations.

The ICL has been studied through simulation and real data studies: [11–13] and it was implemented for its ability to select a relevant model for example in [14–18] From a more theoretical perspective, [19] showed that the ICL is an approximation of the conditional classification likelihood which is shown to be consistent in the same preprint. Rigaill et al. [18] proposed the use of the ICL criterion in the multiple change-point detection context. Hence, the segmentation S can be considered as a set of unobserved variables in the sense that the segment-labels of each datapoint are not known. In this context, we can select the number of segments as

$$\hat{K} = \arg\min_K \text{ICL}(K) \quad \text{where } \text{ICL}(K) = -\log \mathbb{P}(X, K) + \mathcal{H}(K), \tag{1}$$

with $\mathcal{H}(K) = -\sum_{S \in \mathcal{M}_K} \mathbb{P}(S|X,K) \log \mathbb{P}(S|X,K)$, and $\mathcal{M}_K$ representing the set of all segmentations of the signal in K segments.

The entropy term $\mathcal{H}(K)$ can be viewed as an intrinsic penalty to quantify the reliability of a given model with K segments by characterizing the separation of the observations in different segments. In other words, for fixed K segments, the entropy $\mathcal{H}(K)$ thus will be lower when the best segmentation provides a much better fit compared to other segmentations with the same number of segments, hence favoring models which provide the most evidence of similarity within the detected segments. While other penalized likelihood approaches are designed to select the most likely number of segments by relying on approximation of posterior probabilities or oracle inequalities, the ICL criterion aims at selecting the number of segments with the lowest uncertainty.

In the segmentation context, where each segment has its own specific level, an exact algorithm with $\mathcal{O}(Kn^2)$ complexity computes the ICL in a Bayesian framework. In a simulation study, [18] showed that the ICL performed better than standard model selection criteria such as BIC or Deviance Information Criterion (DIC). However, the quadratic complexity and numerical precision restrict the use of this Bayesian ICL to relatively small profiles.

In the context of Hidden Markov Models (HMMs), it is well known that the posterior distribution $\mathbb{P}(S|X, K; \Theta_K)$, where Θ_K is the set of parameters from the K emission distributions, can be efficiently computed using standard forward–backward recursions with $\mathcal{O}(K^2 n)$ complexity [20]. However, the HMM requires that emission parameters take their values in a limited set of levels which are recurrently visited by the underlying hidden process.

Recently, Luong et al. [21] proposed a constrained HMM which corresponds exactly to a segmentation model and which reduces the complexity of the forward–backward algorithm to $\mathcal{O}(Kn)$.

In this paper we suggest a computation of the ICL conditionally to the segment parameters based on this constrained HMM and we propose a fast two-step procedure to compute this conditional ICL criterion with $\mathcal{O}(K^2 n)$ complexity in order to select the number of segments within a set of change-point data. First, we specify a range of possible K number of change-points, from one to a user-defined $K_{\max}$. We estimate the parameters of the segmentation in K segments through the use of any initial segmentation algorithm, and given these estimates, we compute the ICL for each value of K in the range. Second, we select the K which minimizes the ICL criterion. In essence, our conditional ICL explores only one aspect of the segmentation uncertainty, the position of the change-points, and ignores the uncertainty due to the segment parameters.

Section 2 describes the ICL estimation procedure through the use of the constrained hidden Markov model and Section 3 validates the approach by presenting the results of different simulations for detecting the correct number of change-points. Finally, Section 4 is a discussion of our method supported by a comparison with a few segmentation algorithms on data-sets simulated by resampling real RNA-Seq data, and an illustration on the original dataset from an experiment on a chromosome from the yeast species from the same study.

2. Integrated Completed Likelihood criterion estimation using a constrained HMM

In this paper we use the following *segment-based model* for the distribution of X given a segmentation $S \in \mathcal{M}_K$:

$$\mathbb{P}(X|S, K; \Theta_K) = \prod_{i=1}^{n} g_{\theta_{S_i}}(X_i) = \prod_{k=1}^{K} \prod_{i:S_i = k} g_{\theta_k}(X_i) \tag{2}$$