Computational Statistics and Data Analysis 68 (2013) 129–140

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Fast estimation of posterior probabilities in change-point analysis through a constrained hidden Markov model

The Minh Luong *, Yves Rozenholc, Gregory Nuel

MAP5, Université Paris Descartes, 45 rue des Saints-Pères, 75006 Paris, France

HIGHLIGHTS

- We describe a method to assess uncertainty in a set of prespecified change-points.
- It obtains exact estimates of posterior probability of locations without resampling.
- A constrained hidden Markov model estimates probabilities in linear time.
- Methods are implemented in the R package `postCP`, available on CRAN.
- Simulations showed comparable loss to Bayesian implementation in estimating means.

ARTICLE INFO

Article history:
Received 30 July 2012
Received in revised form 21 June 2013
Accepted 21 June 2013
Available online 29 June 2013

Keywords:
Change-point estimation
Segmentation
Posterior distribution of change-points
Constrained hidden Markov model
Forward–backward algorithm
Fast computation

ABSTRACT

The detection of change-points in heterogeneous sequences is a statistical challenge with applications across a wide variety of fields. In bioinformatics, a vast amount of methodology exists to identify an ideal set of change-points for detecting Copy Number Variation (CNV). While considerable efficient algorithms are currently available for finding the best segmentation of the data in CNV, relatively few approaches consider the important problem of assessing the uncertainty of the change-point location. Asymptotic and stochastic approaches exist but often require additional model assumptions to speed up the computations, while exact methods generally have quadratic complexity which may be intractable for large data sets of tens of thousands points or more. A hidden Markov model, with constraints specifically chosen to correspond to a segment-based change-point model, provides an exact method for obtaining the posterior distribution of change-points with linear complexity. The methods are implemented in the R package `postCP`, which uses the results of a given change-point detection algorithm to estimate the probability that each observation is a change-point. The results include an implementation of `postCP` on a publicly available CNV data set ($n = 120$). Due to its frequentist framework, `postCP` obtains less conservative confidence intervals than previously published Bayesian methods, but with linear complexity instead of quadratic. Simulations showed that `postCP` provided comparable loss to a Bayesian MCMC method when estimating posterior means, specifically when assessing larger scale changes, while being more computationally efficient. On another high-resolution CNV data set ($n = 14{,}241$), the implementation processed information in less than one second on a mid-range laptop computer.

1. Introduction

The detection of *change-points* in heterogeneous sequences is a statistical challenge with many applications in fields such as finance, reliability, signal analysis, neurosciences and biology (Pinkel et al., 1998; Snijders et al., 2001). In bioinformatics in

* Corresponding author. Tel.: +33 6 84 75 02 14.
E-mail addresses: luongtm@yahoo.com, the-minh.luong@parisdescartes.fr (T.M. Luong), yves.rozenholc@parisdescartes.fr (Y. Rozenholc), gregory.nuel@parisdescartes.fr (G. Nuel).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.06.020

particular, a vast amount of methodology (Olshen et al., 2004; Fridlyand et al., 2004; Hupé et al., 2004) exists for identifying an ideal set of change-points in data from array Comparative Genomic Hybridization (aCGH) techniques, in order to identify Copy Number Variation (CNV).

A typical expression of the change-point problem is as follows, given data $X = (X_1, X_2, \ldots, X_n)$ of real-valued observations, $(S_1, \ldots, S_n)$ corresponding segment indices of the observations, and $\mathscr{M}_K$ as the set of all possible combinations of S for fixed $K \geqslant 2$ number of segments. The goal is to find the best partitioning $S \in \mathscr{M}_K$ into K non-overlapping intervals, assuming that the distribution is homogeneous within each of these intervals.

For K segments of contiguous observations, the *segment-based model* expresses the distribution of X given a segmentation $S \in \mathscr{M}_K$ as

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

where $g_{\theta_k}(\cdot)$ is the parametric, or emission, distribution (e.g.: normal or Poisson) of the observed data with parameter θ_k, $\theta = (\theta_1, \ldots, \theta_K)$ is the global parameter, and S_i is the segment index at position i. For example, if $n = 5, K = 2$, and the distribution changes from position 2 to 3, then $S = (1, 1, 2, 2, 2)$.

This paper describes an exact method for obtaining the posterior distribution of the segmentation $\mathbb{P}(S|X;\theta)$ in linear time. Introducing a prior distribution $\mathbb{P}(S)$ on any $S \in \mathscr{M}_K$ obtains the following expression for $\mathbb{P}(S|X;\theta)$:

$$\mathbb{P}(S|X;\theta) = \frac{\mathbb{P}(X|S;\theta)\mathbb{P}(S)}{\sum_{R} \mathbb{P}(X|R;\theta)\mathbb{P}(R)}. \tag{2}$$

To simplify the above expression, a uniform prior sets $\frac{1}{\mathbb{P}(S)} = \binom{n-1}{K-1} = |\mathscr{M}_K|$.

A common alternative to the above segmentation procedure is to consider an unsupervised hidden Markov model (HMM). Assuming that S is a Markov chain of hidden states, this approach (Rabiner, 1989) can be thought of as being *level-based*, where the parameter of the kth segment takes its value in the set of $L \geqslant 1$ levels: $\{\theta_1, \theta_2, \ldots, \theta_L\}$. This simply is equivalent to the model defined by Eq. (1), with the noticeable difference that $S \in \{1, 2, \ldots, L\}^n$. With this level-based approach $K \geqslant L$ in general, and the HMM is unconstrained in the sense that transitions are possible between any pair of states. Similar to the segment-based model, the choice of L is critical and is usually addressed through penalized criteria. The conventional HMM is an appropriate model when the conditional distribution within a given segment of contiguous observations may be shared among other segments. While the unconstrained HMM is preferable in many practical situations, the segment-based model as described in this paper requires fewer assumptions and is thus a more general model.

A convenient feature of these HMM approaches is in computing efficiently the posterior distribution $\mathbb{P}(S|X;\theta)$ in $O(L^2 n)$ using classical forward–backward recursions (Durbin et al., 1998), making them suitable for handling large data sets. This paper focuses on using a computationally efficient exact procedure to characterize the uncertainty $\mathbb{P}(S|X;\theta)$ of the estimated change-point locations using a hidden Markov model adapted to the conventional segmentation model as previously described. We exploit the effectiveness of the level-based HMM approach through a constrained HMM corresponding *exactly* to the above segment-based model, providing a fast algorithm for computing $\mathbb{P}(S|X;\theta)$.

We develop this posterior distribution procedure as the uncertainty in change-point assessment in practical applications becomes more challenging from a computational point of view. For example, emerging high-throughput technologies are producing increasingly large amounts of data for CNV detection. For finding the exact posterior distribution of change-points $\mathbb{P}(S|X;\theta)$, Guédon (2007) suggested an algorithm in $O(Kn^2)$, while Rigaill et al. (2011) considered the same quantity in a Bayesian framework with the same complexity. However, the complexity of these approaches provides for very slow processing for large data sets with sequences of tens of thousands or more and with ten or more change-points in the data.

Other estimates generally focus on asymptotic behavior whose conditions are delicate due to the discreteness of the problem (Bai and Perron, 2003; Muggeo, 2003), on bootstrap techniques (Hušková and Kirch, 2008), estimating waiting time distributions in HMMs (Aston et al., 2012) and on stochastic methods such as particle filtering (Fearnhead and Clifford, 2003), recursive sampling (Lai et al., 2008), and Markov chain Monte Carlo (Erdman and Emerson, 2008). Furthermore, many of the faster stochastic algorithms assume a normal error structure to speed up the estimation procedures and are thus more difficult to adapt to non-normal data (Lai et al., 2008; Erdman and Emerson, 2008).

This paper details our procedure to characterize the uncertainty of a pre-specified set of change-points by an exact method in linear time. As such, the procedure is intended to complement any segmentation obtained by existing detection methods, by estimating the posterior probability of the location of each change-point. It can also be used for further applications such as model selection. While the underlying model from our segmentation approach is different from that of a conventional HMM, we apply constraints specifically to enable the use of efficient HMM algorithms for estimating posterior probabilities of interest. Furthermore, our frequentist approach obtains these probabilities in linear time, without resampling.

Section 2 presents a summary of current change-point detection methods, the constrained HMM algorithm and a description of the accompanying R statistical package, Section 3 implements the methods on a published array CGH data set and compares with published results, Section 4 presents examples of simple simulated data sets with comparisons between