Computational Statistics and Data Analysis 97 (2016) 71–86

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

An exact approach to Bayesian sequential change point detection

Eric Ruggieri *, Marcus Antonellis

Department of Mathematics and Computer Science, College of the Holy Cross, 1 College Street, Worcester, MA 01610, USA

HIGHLIGHTS

- Algorithm quickly updates its inference in linear time with each new observation.
- Derives uncertainty bounds on the number and location of change points in a data set.
- Explores potential detection criteria associated with posterior distribution.
- Simulation studies show high detection rate, low false positive rate.
- Analysis of two real data sets illustrate wide range of potential applications.

ARTICLE INFO

Article history:
Received 20 October 2014
Received in revised form 3 November 2015
Accepted 18 November 2015
Available online 2 December 2015

Keywords:
Dynamic programming
Exact Bayesian inference
Global temperature anomalies
Multiple change point
Piecewise regression

ABSTRACT

Change point models seek to fit a piecewise regression model with unknown breakpoints to a data set whose parameters are suspected to change through time. However, the exponential number of possible solutions to a multiple change point problem requires an efficient algorithm if long time series are to be analyzed. A sequential Bayesian change point algorithm is introduced that provides uncertainty bounds on both the number and location of change points. The algorithm is able to quickly update itself in linear time as each new data point is recorded and uses the exact posterior distribution to infer whether or not a change point has been observed. Simulation studies illustrate how the algorithm performs under various parameter settings, including detection speeds and error rates, and allow for comparison with several existing multiple change point algorithms. The algorithm is then used to analyze two real data sets, including global surface temperature anomalies over the last 130 years.

1. Introduction

Long time series are often heterogeneous in nature. As Chopin (2007, p. 349) notes, 'the assumption that an observed time series follows the same fixed stationary model over a very long period is rarely realistic.' Since the inability to recognize a regime change in a data set can have a detrimental effect on an algorithm's predicative performance, the ultimate goal of change point analysis is to fit a piecewise regression model to a data set where the exact timing of these regime changes is unknown. A 'change point' is defined as an abrupt shift in the parameters of a model. Common examples include detecting a change in the mean, variance or trend of the response variable. Ideally, one would want to identify exactly when this 'change point' occurs so that each regime can be separately fit by an appropriate model.

* Corresponding author. Tel.: +1 508 793 2400.
E-mail addresses: eruggier@holycross.edu (E. Ruggieri), mjanto15@g.holycross.edu (M. Antonellis).

http://dx.doi.org/10.1016/j.csda.2015.11.010
0167-9473/© 2015 Elsevier B.V. All rights reserved.

Change point models have been applied in a wide range of settings including finance (Chopin, 2007), climate (Ruggieri et al., 2009; Gallagher et al., 2012; Ruggieri, 2013), biology (Fearnhead and Liu, 2007), earthquake data (Grigg and Spiegelhalter, 2008), DNA segmentation (Liu and Lawrence, 1999), historical time series such as average annual wage growth (Western and Kleykamp, 2004), and in other areas where long sequences of data are available. A Bayesian approach to the change point problem is particularly appealing for two reasons. First, a Bayesian approach does not rely on the asymptotic assumptions about test statistics that are present in frequentist algorithms, which can be problematic in situations where the parametric models considered are restricted to a finite, possibly small interval of time (Chopin, 2007). Second, a Bayesian approach will allow one to quantify uncertainty both in the number and the positioning of the change points.

Here, we seek to describe an efficient and exact Bayesian change point model that can quickly update itself as each new observations is recorded (known as sequential change point detection). Each data point will require an update that has linear time complexity, giving the algorithm an overall complexity that is quadratic in the length of the data set. This stands in stark contrast to the exponential time required by a brute force approach. Once a new change point has been detected, the algorithm samples directly from the exact posterior distribution on the number of change points, their locations, and the parameters of the regression model, yielding uncertainty estimates for each of these quantities.

The rest of the article is organized as follows. Section 2 describes some of the existing change point techniques, with a special focus on sequential Bayesian methods. Following the literature review, Section 3 presents an overview of the Bayesian Change Point algorithm of Ruggieri (2013) that will be used as the foundation for the sequential method presented in this paper. In Section 4, we describe how the Bayesian Change Point algorithm of Ruggieri (2013) can be modified to handle sequential observations. In Section 5, the new sequential change point detector tested on simulated data sets as well as two real data sets and compare the results to several existing multiple change point algorithms. Section 6 provides discussion and conclusions.

2. Background

Due to the vast array of change point algorithms that have been developed, a full review of existing change point models is beyond the scope of this paper. Instead, the focus will be on giving a brief overview of some of the prominent categories of change point algorithms, emphasizing those which are most similar to what is proposed in this paper. Specifically, attention will focus on highlighting the similarities and differences among the most popular Bayesian sequential change point algorithms.

2.1. Batch vs. sequential change point algorithms

There are two main branches of change point detection algorithms, batch and sequential. In the batch setting, the data set is fixed and an algorithm will retrospectively look for change points in a time series. Given a time series with N observations, there are approximately $\binom{N}{k}$ ways to place k change points in the time series, partitioning the data set into $k+1$ segments. Because the number of solutions grows exponentially in the number of change points, an exact solution will require an efficient algorithm in order to be practical. On the frequentist side, solutions to the change point problem in regression often focus on minimizing squared error. For example, dynamic programming algorithms (e.g. Auger and Lawrence, 1989; Bai and Perron, 2003; Ruggieri et al., 2009) reduce the complexity to quadratic in the number of observations and are guaranteed to find the optimal solution. Another popular approach is binary segmentation (Scott and Knott, 1974), which is a greedy approach that recursively splits a data set at each identified change point until only homogeneous segments remain. A recent adaptation is that of Fryzlewicz (2014), whose wild binary segmentation algorithm uses a localized (rather than global) cumulative sum statistic to recursively split the data set. One shortcoming of these frequentist algorithms is that they are unable to quantify the uncertainty associated with their solutions, both in the number and locations of the change points. On the Bayesian side, Gibbs sampling (e.g. Carlin et al., 1992; Stephens, 1994; Western and Kleykamp, 2004) and MCMC (e.g. Barry and Hartigan, 1993; Green, 1995; Chib, 1998; Lavielle and Lebarbier, 2001) approximations have dominated the probabilistic solutions to the change point problem, although convergence issues exist (Fearnhead, 2006; Whiteley et al., 2011). Specifically, MCMC procedures (e.g. Stephens, 1994; Lavielle and Lebarbier, 2001) which update change point locations one at a time or condition on latent variables associated with each segment (Chib, 1998) can be slow mixing due to the strong correlations in the target distribution (Whiteley et al., 2011). Alternatively, Fearnhead (2006) and Ruggieri (2013) both avoid these approximations by directly sampling from the posterior distribution through the use of dynamic programming-like recursions.

Conversely, sequential change point algorithms need to be able to handle a constant stream of new observations and make a decision about whether or not a change has occurred based only on the data observed to that point. Essentially, a sequential change point algorithm strives to make a ‘quick’ update of the inference as each new observation becomes available, rather than having to re-analyze the entire data set. The ultimate goal is to be able to quickly detect a change in the system, subject to some tolerable limit on the risk of a false positive (i.e. false detection of a change point). An appropriate balance has to be struck since the desire to detect a change point quickly can lean to a high risk of false positives, while avoiding false alarms too strenuously can cause a longer delay between the actual occurrence of the change and its detection by the algorithm. If a change has been detected, then the algorithm can then stop to make an inference about its location