Offline change point detection for binary data via regularization methods

Lucas de Oliveira Prates

Dissertação/Tese apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Mestre em Estatística

Programa: Estatística

Orientador: Prof[a]. Florencia Graciela Leonardi

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES

São Paulo, fevereiro de 2021

Offline change point detection for binary data

Esta versão da dissertação/tese contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 12/04/2021. Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Profª. Drª. Florencia Graciela Leonardi (orientadora) - IME-USP
- Profª. Drª. Daniela Andrea Rodriguez - UBA
- Prof. Dr. Bruno Monte de Castro - UFRN

Agradecimentos

Apesar de levar meu nome, este trabalho não seria possível sem o apoio de diversas pessoas. Desta forma, gostaria de agradecer brevemente. Agradeço à meus pais Liana e Geraldo, meu irmão Thiago, e minha amiga Silvana, pelo amor, carinho e confiança ao longo de minha vida. Acima de tudo, agraço por privilegiar-me com a oportunidade estudar.

Agradeço à minha orientadora, Florencia, pela sua paciência, confiança, e direcionamento na pesquisa. Aos amigos de infância e de faculdade, com quem compartilhei memórias queridas.

Por fim, sou grato à todos que contribuíram com o conhecimento humano até então acumulado, especialmente em matemática, estatística e computação. O meu aprendizado é fruto do esforço coletivo de todas essas pessoas.

Resumo

Prates, L. O. **Detecção offline de pontos de mudança para dados binários via métodos de regularização**. 2021. 120 f. Dissertação - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2021.

Em análise de séries temporais, o problema de detecção de pontos de mudança consiste em estimar os tempos nos quais a distribuição de probabilidade sofre alguma alteração. Sob à hipótese de que os dados têm distribuição Bernoulli, o problema pode ser visto como estimar os tempos nos quais o parâmetro de probabilidade se altera. Neste trabalho, apresentaremos métodos estatísticos para estimar o número e a localização dos pontos de mudança quando os dados têm distribuição Bernoulli. Os métodos escolhidos foram verossimilhança penalizada, Fused LASSO e métodos baseados em validação cruzada. Provamos a consistência de alguns dos métodos propostos, e fornecemos um estudo de simulação para comparação de modelos. Por fim, aplicamos os modelos no problema de identificação de regiões de homozigose em arrays de SNPs.

Palavras-chave: detecção de pontos de mudança, regularização, SNPs.

Abstract

Prates, L. O. **Offline change point detection for binary data via regularization methods**. 2021. 120 f. Master's Thesis - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2021.

In time series analysis, change point detection consists in estimating the times in which the probability distribution changes. Under the assumption that the data is Bernoulli distributed, the problem can be seen as estimating the time in which the probability parameter changes. We will present statistical methods based on penalized likelihood to estimate the number and location of the change points. The chosen methods were penalized likelihood, Fused LASSO and methods based on cross validation. The consistency of some of the methods is proved, and a simulation study is provided to compare models. We then apply the models to the identification of regions of homozygosity in SNP arrays.

Keywords: change point detection, regularization, SNPs.

Contents

List of abbreviations

AIC	Akaike Information Criteria
BIC	Bayesian Information Criteria
CDF	Cumulative Distribution Function
CV	Cross Validation
CVDYNSEG	Cross Validation on regularization constant Algorithm
CVSEG	Cross Validation on the number of segments Algorithm
DPS	Dynamical Segmentation Algorithm
EPE	Expected Prediction Error
HS	Hierarhical Segmentation Algorithm
MLE	Maximum Likelihood Estimator
PL	Penalized Likelihood
PML	Penalized Maximum Likelihood

List of Figures

List of Tables

Chapter 1

Introduction

Humans observe with keen interest the processes of this dynamical universe aiming to describe and mimic them. While watching the process in its entirety, it might be only a few singularities, abrupt changes in the process states, that create a sparkle in the viewer's mind, providing fuel for his creativity. By carefully describing what these changes are and how they occur, he is able to unravel the whole. Furthermore, he could detect new changes, rapidly reacting after one occurs, or even predict when it will occur, taking actions to prevent or prepare himself.

Change point detection is a multidisciplinary field in statistics that studies changes in processes of very different natures. Framed in the context of time series analysis, it uses mathematics, statistics, and computer science to provide rigorous, reliable methodologies to detect and predict subtle and complex changes.

Grasping the concept of change is complex and context-dependent. However, this broad interpretation of what is to change allows the field to be useful to many different areas. Like any application in statistics, defining the concepts correctly and preparing the data might be more important than the method used.

In this Chapter, we begin by posing the basic idea behind change point detection. Then, we discuss some aspects of the problem considered in research and applications.

1.1 Problem Definition

Let $\mathcal{T}$ be an ordered set and consider we have a sequence of $\{X_{t_i}\}_{i=1}^{m}$, $m \in \mathbb{N}$ independent random variables where X_{t_i} has CDF F_i, $t_i \in \mathcal{T}$. Defining $1 : (m-1) = \{1, \ldots, m-1\}$, we are interested in obtaining the change point set C^* given by

$$C^* = \{c \in 1 : (m-1) | F_c \neq F_{c+1}\} \quad .$$

We can see this as blocks of random variables with the same distribution. Our task is to detect the blocks and the CDF of each block. The number of change points can either be known or unknown, resulting in very different approaches. We can also have multiple sequences of data, as will be the case in this work.

For the parametric approach, we usually define the CDFs as being elements of a family $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$, where the CDF is fully specified by the parameter θ. In this case, the problem translates to detecting when the changes in the parameter. For example, if the data comes from a Gaussian distribution with unknown mean and variance, then we could design methods to estimate the locations in which the mean changes.

1.2 Aspects of change point problems

Changes have multiple aspects to consider, and it might not be possible to devise a method that works irrespective of the problem. Here, we will discuss the possible scenarios briefly. The next

section introduces the most common mathematical formulation of the problem.

The first and most clear distinction is on single and multiple change point problems. A single change point problem is when we have at most one change, and our task is only to find its location if the change exists. This was the first studied model, presented on the pioneering work of Page [1954].

A multiple change point problem requires us to detect the location of possible multiple change points. Two different settings are possible here: known and unknown number of change points. For the second problem, the number of points may range from no change points at all up to a maximum number of change points established beforehand. This difficulty usually deteriorates the performance and run time, and it is much harder to prove their correctness.

Albeit most of the applications consider change points with respect to time, many methods are general enough to work with other variables and dimensions where an order relation is available. For instance, when analyzing a SNP array, we search change points in the base pairs' physical position, not on a time dimension.

Another important distinction is between online and offline change point detection. In online problems, we expect to continually receive new data to analyze in the dimension of interest. This type of problem is common in economics, quality control, and signal processing, and we can have many different objectives. We could try to react as fast as possible when noticing a change, predict the next change, estimate how long the current state will last, and so on.

In offline problems, also called retrospective analysis, we are interested in detecting changes in a phenomenon that already occurred or that has a fixed size. Again, there is a broad range of interesting questions that can help us better capture what happened. Can we associate the change with some other event known to have occurred near the change point estimated?

If we dive into mathematical and statistical assumptions, we can split the research further. Lee [2010] provides a comprehensive list on change point detection research up to 2010, summarising the number of papers considering the various problems and methodologies discussed.

Parametric and nonparametric settings have been advanced to solve problems such as detecting changes in the mean, variance and regression slope. Parametric change point detection was introduced by Page [1955], studying the Normal mean change point model. Chen and Gupta [2011] offers an introduction to parametric change point detection with a mathematically rigorous approach, presenting well-studied change point models such as changes in Normal mean, Normal variance, Poisson rate, binomial parameter, hazard rate, while providing applications in different areas.

A forking point between researches is the estimation approach. Bayesian, maximum likelihood, and regularizations estimation methods have been applied to online and offline problems, sometimes considering different aspects of the problem.

Regularized estimators are often used when the number of change points is unknown. Zou et al. [2014] used the BIC criterion for estimating change points combined with a nonparametric maximum likelihood approach. More recently, sparsity inducing regularizations have received more attention, with Levy-leduc and Harchaoui [2008] introducing the use of Lasso for change point detection.

For bayesian methods, the works of Raftery and Akman [1986] and Carlin et al. [1992] study offline single change point detection, modeling different aspects of the problem. In online problems, Adams and MacKay [2007] modeled the sequence run length, i.e., the time since the last change. An extension of the model by Agudelo-España et al. [2020] also models residual time, i.e., the time until the next change point.

Other techniques, such as kernel-based methods and sliding windows, have also been considered. Harchaoui and Cappe [2007] presents the use of kernels for offline change point detection, and Bouchikhi et al. [2020] investigates the usage of a kernel-based algorithm for online problems.

In practice, most estimators proposed are not so easily computed. Frequently, the estimators are formulated as solutions to optimization problems. Different paradigms and algorithms are considered for calculations. Usually, a there is a trade-off: use exact but more time-consuming search methods or fast greed search methods for an approximation. The choice usually depends on the aspects of the problem at hand. Chapter 3 discusses some paradigms and algorithms that can be applied to

change point detection.

1.3 Applications

Since change point detection is so multidisciplinary and broad, selective reviews aimed at the statistical community have been published. Niu et al. [2016] shows classical and modern applications of change point detection, and poses the problem mathematically for different settings. Truong et al. [2020] presents a more thorough discussion, giving a complete description of a broad approach to change point problems. Here, we discuss briefly some real problems in which change point detection methods were applied.

Based on the pioneering work of Page [1954], the well known CUSUM is one of the widest studied method applied to several fields. Williams et al. [1992] presents a first application of the CUSUM to medicine, and the work of Li et al. [2018] expands the method for data stream anomaly detection and apply it to industrial data.

An early application of change point detection in meteorology is given in Cobb [1978], which works on the single change point problem using maximum likelihood and apply his results to the Nile River data set. The data consists of annual volume measures of the Nile River at Aswan, from 1871 to 1970, and the goal was to understand if there was an abrupt change in rainfall regime near the beginning of the last century. Reeves et al. [2007] provides a review on change point methods applied to climate change.

With the exponential growth of computer power, applications using genomics, videos, and audios data sources are becoming available and feasible. In genetics, Castro et al. [2018] uses a regularized approach to identify recombination hotspots using SNP arrays, regions of the chromosome with higher recombination rates. Celisse et al. [2018] uses kernel-based methods to identify DNA copy number alterations, which have been associated with diseases in humans.

Tahmasbi and Rezaei [2008] presents a change point model for GARCH models in speech recognition tasks, trying to identify intervals of speech and non-speech. Application on satellite image time series was provided by Verbesselt et al. [2010], investigating the land cover variation over time.

As examples of online problems, Agudelo-España et al. [2020] uses their online bayesian model to monitor reliably sleep stages using EEG/EMG data. Tartakovsky et al. [2006] shows that change point detection can be used in cybersecurity to identify attacks on networks using network traffic data.

The examples provided are only a few in a range of possibilities of the applications of change point detection. However, there is a critique that the performance of most methods is evaluated on simulated data or on small time series data sets with unreliable ground truth. To assess that problem, van den Burg and Williams [2020] created a data designed to evaluate change point detection algorithms. They also run several models and present a benchmark, comparing the performance of the methods.

Chapter 2

Conclusions

In this work, we have presented the change point detection problem, defining it mathematically and viewing briefly some applications. We focused on the offline binary change point problem, with multiple and unknown change points, studying regularized likelihood methods and their consistency.

For the computation of the estimators, we showed that dynamical programming, hierarchical segmentation and disciplined convex programming can be applied. We provided pseudocodes for the estimators, and performed simulations studies to answer some raised questions. We provided an R package to fit all the models proposed in this work.

Finally, we tackled the problem of detecting ROH Islands in genetics, explaining the basic idea behind the problem and one possible way to frame it as a change point detection. We compared it to PLINK and obtained results that are related, but not identical.

Bibliography

Adams, R. P. and MacKay, D. J. C. [2007], 'Bayesian online changepoint detection'. 2

Agudelo-España, D., Gomez-Gonzalez, S., Bauer, S., Schölkopf, B. and Peters, J. [2020], 'Bayesian online prediction of change points'. 2, 3

Bouchikhi, I., Ferrari, A., Richard, C. and Bourrier, A. [2020], 'Online change-point detection with kernels'. 2

Carlin, B. P., Gelfand, A. E. and Smith, A. F. M. [1992], 'Hierarchical bayesian analysis of change-point problems', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **41**(2), 389–405.
URL: *https://rss.onlinelibrary.wiley.com/doi/abs/10.2307/2347570* 2

Castro, B. M., Lemes, R. B., Cesar, J., Hünemeier, T. and Leonardi, F. [2018], 'A model selection approach for multiple sequence segmentation and dimensionality reduction', *Journal of Multivariate Analysis* **167**, 319–330. 3

Celisse, A., Marot, G., Pierre-Jean, M. and Rigaill, G. [2018], 'New efficient algorithms for multiple change-point detection with reproducing kernels', *Computational Statistics & Data Analysis* **128**, 200 – 220.
URL: *http://www.sciencedirect.com/science/article/pii/S0167947318301683* 3

Chen, J. and Gupta, A. K. [2011], *Parametric statistical change point analysis: with applications to genetics, medicine, and finance*, Springer Science & Business Media. 2

Cobb, G. W. [1978], 'The problem of the nile: Conditional solution to a changepoint problem', *Biometrika* **65**(2), 243–251. 3

Harchaoui, Z. and Cappe, O. [2007], Retrospective mutiple change-point estimation with kernels, *in* '2007 IEEE/SP 14th Workshop on Statistical Signal Processing', pp. 768–772. 2

Lee, T.-S. [2010], 'Change-point problems: Bibliography and review', *Journal of Statistical Theory and Practice* **4**(4), 643–662.
URL: *https://doi.org/10.1080/15598608.2010.10412010* 2

Levy-leduc, C. and Harchaoui, Z. [2008], Catching change-points with lasso, *in* J. Platt, D. Koller, Y. Singer and S. Roweis, eds, 'Advances in Neural Information Processing Systems', Vol. 20, Curran Associates, Inc., pp. 617–624.
URL: *https://proceedings.neurips.cc/paper/2007/file/e5841df2166dd424a57127423d276bbe-Paper.pdf* 2

Li, G., Wang, J., Liang, J. and Yue, C. [2018], 'The application of a double cusum algorithm in industrial data stream anomaly detection', *Symmetry* **10**(7), 264. 3

Niu, Y. S., Hao, N. and Zhang, H. [2016], 'Multiple change-point detection: A selective overview', *Statist. Sci.* **31**(4), 611–623.
URL: *https://doi.org/10.1214/16-STS587* 3

Page, E. S. [1954], 'Continuous inspection schemes', *Biometrika* **41**(1/2), 100–115.
URL: *http://www.jstor.org/stable/2333009* 2, 3

Page, E. S. [1955], 'A test for a change in a parameter occurring at an unknown point', *Biometrika* **42**(3-4), 523–527.
URL: *https://doi.org/10.1093/biomet/42.3-4.523* 2

Raftery, A. E. and Akman, V. E. [1986], 'Bayesian analysis of a poisson process with a change-point', *Biometrika* **73**(1), 85–89.
URL: *http://www.jstor.org/stable/2336274* 2

Reeves, J., Chen, J., Wang, X. L., Lund, R. and Lu, Q. Q. [2007], 'A Review and Comparison of Changepoint Detection Techniques for Climate Data', *Journal of Applied Meteorology and Climatology* **46**(6), 900–915.
URL: *https://doi.org/10.1175/JAM2493.1* 3

Tahmasbi, R. and Rezaei, S. [2008], 'Change point detection in garch models for voice activity detection', *IEEE Transactions on Audio, Speech, and Language Processing* **16**(5), 1038–1046. 3

Tartakovsky, A. G., Rozovskii, B. L., Blazek, R. B. and Hongjoong Kim [2006], 'A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods', *IEEE Transactions on Signal Processing* **54**(9), 3372–3382. 3

Truong, C., Oudre, L. and Vayatis, N. [2020], 'Selective review of offline change point detection methods', *Signal Processing* **167**, 107299.
URL: *http://dx.doi.org/10.1016/j.sigpro.2019.107299* 3

van den Burg, G. J. J. and Williams, C. K. I. [2020], 'An evaluation of change point detection algorithms'. 3

Verbesselt, J., Hyndman, R., Newnham, G. and Culvenor, D. [2010], 'Detecting trend and seasonal changes in satellite image time series', *Remote Sensing of Environment* **114**(1), 106 – 115.
URL: *http://www.sciencedirect.com/science/article/pii/S003442570900265X* 3

Williams, S., Parry, B. and Schlup, M. [1992], 'Quality control: An application of the cusum', *British Medical Journal* **304**, 1359–1361. 3

Zou, C., Yin, G., Feng, L. and Wang, Z. [2014], 'Nonparametric maximum likelihood approach to multiple change-point problems', *Ann. Statist.* **42**(3), 970–1002.
URL: *https://doi.org/10.1214/14-AOS1210* 2