King's College London

Department of Natural and Mathematical Science

Bayesian Models for Covariance Changepoint Detection

Author:
Student 1870087

Supervisor:
Dr. Davide Pigoli

A report submitted for the degree of

MSc in Non-Equilibrium Systems

September 4, 2019

Abstract

Herein, we investigate two parametric Bayesian changepoint detection algorithms, the Hidden Markov Model (HMM) and Product Partition Model (PPM) with applications to finding changes in the covariance structure in multivariate signals. Both models were ran on several synthetic and empirical data sets, including various distributions of correlated random variables (r.v.s), industry portfolio data, and rainfall observations. Bootstrapping methods were performed to determine thresholds for change-detection in the PPM. In-depth investigations revealed with high probability that the methods were indeed detecting changes in the covariance structure of the data. For comparison, we used a non-parametric binary segmentation (BS) algorithm to compare the results. We find in nearly all cases the non-parametric BS method outperforms the HMM and PPM in both computation time and accuracy.

Acknowledgements

This report is dedicated to Yanik Förster, Lewis Wright and Zac Walker.
Thank you for being my sources of strength throughout this year in CANES.

Contents

1 Introduction **4**

2 Background **5**

2.1 Review . . . 5

2.1.1 CUSUM . . . 5

2.1.2 Maximum Likelihood Estimation (MLE) . . . 6

2.1.3 Other Major Methods . . . 6

2.1.4 Applications in Non-Equilibrium Systems . . . 6

2.2 Data . . . 6

3 Methods **8**

3.1 Hidden Markov Model . . . 8

3.1.1 Baum-Welch Algorithm for Parameter Fitting . . . 8

3.1.2 Viterbi Algorithm for Determining State Changes . . . 11

3.2 Bayesian Product Partition Model . . . 11

3.2.1 Choice of Prior . . . 12

3.2.2 Bootstrapping Thresholds . . . 13

3.3 A Non-Parametric Model: Binary Segmentation . . . 13

3.4 Generating Correlated Random Variables . . . 14

4 Results **15**

4.1 Bootstrap Results for Bayesian PPM . . . 15

4.2 Comparison of Results: Synthetic Data . . . 15

4.3 Comparison of Results: Empirical Data . . . 21

5 Conclusion **26**

Bibliography **29**

List of Figures

2.1 Portfolio and Rainfall data signals 7

3.1 Negative binomial distribution for various k, p 13

4.1 Bayesian PPM bootstrap results 15
4.2 Bayesian PPM for Gaussian and Exponential correlated r.v.s 16
4.3 Hidden Markov model state probabilities for Gaussian, Exponential, and Cauchy correlated r.v.s 17
4.4 Binary segmentation for Gaussian, Exponential, and Cauchy correlated r.v.s . . . 18
4.5 Hidden Markov model state probabilities for Gaussian, Exponential, and Cauchy correlated r.v.s (2) 19
4.6 Bayesian PPM for Gaussian and Exponential correlated r.v.s (2) 19
4.7 Hidden Markov model state probabilities for Gaussian, Exponential, and Cauchy correlated r.v.s (2) 20
4.8 Hidden Markov model state probabilities for empirical data 21
4.9 Bayesian PPM for empirical data 22
4.10 BS detection of one changepoint in portfolio data 23
4.11 BS detection of five changepoints in portfolio data 24
4.12 BS detection of one changepoint in rainfall data 25

Chapter 1

Introduction

Given a sufficient amount of time, the dynamics of any system must change. A pendulum does not swing forever, boiling water is taken off the stove, and slowly the gas escapes the box. Our concept of equilibrium is by nature constrained by time, the underlying laws may stay the same but the system will at one point or another be shaken, kicked, or dismantled. For some processes, this spontaneous change is constantly occurring, and the system is characterized by transience.

Consider the life cycle of a white blood cell, which lasts about 170 minutes in rats[1]. During three hour period the cell acts in a myriad ways; moving throughout the organism, sending signaling molecules, and producing antibodies. Or the weather in Britain; oscillating unstably from sun to rain and back again. In each of these systems, we have segments of time that appear to have sustained behaviour (the weather or cell function *at this moment*), but taken over a lifetime we find an eclectic set of processes that cannot be described as a whole. What we seek to explore in this report is how, given dynamical information about our system, we can distinguish transitions between states as accurately as possible.

Knowing where these transitions are copiously expands our capability for analysis and inference. During the timescale where we know the process to be stable, we can utilize our abundance of equilibrium tools; if the system goes awry, one can determine at which point; and for systems that recurrently take on previous states, one can exploit the process through their prior knowledge. There may even be cases where we are unaware a change has even occurred, such that these methods could guard us from catastrophic loss.

Determining shifts in the mean and variance of signals has been well studied for nearly a century[2, 3]. However, some systems admit transitions between processes that cannot be inferred by looking at the individual variables, and instead modify the relationship *between* variables. Seizures are characterized by synchronization between neurons[4], changes in correlation structure of stocks can distinguish between bull and bear markets[5], and variation in gene expression may indicate various diseases[6].

In this report, we investigate changes in the correlation structure of two data sets using parametric Bayesian methods; the value-weighted returns of 10 industry portfolios, and Indian monsoon rainfall/El Niño-Southern Oscillation (ENSO) data for India and Brazil, along with several sets of correlated random variables (r.v.s) of different distributions. We seek to test how the Gaussian assumption holds in the presence of possibly non-Gaussian data, and explore the limitations of Bayesian based changepoint detection.

We begin with an overview of the changepoint detection problem, including its origins in process control, the major methods developed, online vs. offline and parametric vs. non-parametric approaches, and applications in non-equilibrium systems in Section 2, followed by a description of the empirical data used. Then, in Section 3, we introduce our two parametric methods, the Bayesian Product Partition Model (PPM) and Hidden Markov Model (HMM), along with a non-parametric binary segmentation algorithm for comparison. Our results are then presented and discussed in Section 4, with concluding statements in Section 5.

Chapter 2

Background

2.1 Review

Detection and estimation of change-points is a relatively recent advance in mathematical and engineering sciences, with its origins occuring only within the last century. The first explicit papers dealing with a change in processes is accredited to E.S. Page in his 1954 and 1955 articles outlining the general problem and cumulative sum (CUSUM) technique[2, 7]. He was focused on the problem of online change-point detection; finding process shifts in real-time for applications in quality control at industrial plants, extending off the original control-chart conception by Shewhart[8]. This is contrasted with the offline method of change-point detection, which attempts to determine process shifts within an already complete data-set.

The change-point problem has been stated in various ways. Most well known is the hypothesis testing formulation. Consider a set of observations $\boldsymbol{x} = \{x_1, ..., x_T\}$ with corresponding distributions $\boldsymbol{F} = \{F_1, ..., F_T\}$, we then seek to test null and alternate hypotheses:

$$H_0 : F_1 = F_2 = ... = F_T \tag{2.1}$$

$$H_A : F_1 = F_2 = ... = F_k \neq F_{k+1} = F_{k+2} = ... = F_T, \quad 1 < k < T \tag{2.2}$$

This is often extended to include the possibility of having multiple changepoints, possibly even returning to prior distributions. We could also phrase the problem as one of statistical homogeneity, and determine 1) whether a sample is homogeneous and 2) if not, whether it has homogeneous segments[9].

The general format of changepoint detection algorithms are consistent with the following structure: a **search method** attempts to minimize (or maximize) a **cost function** with respect to some **constraint**[3]. Here we outline some of the major developments in each of these areas.

2.1.1 CUSUM

We begin with the original CUSUM algorithm developed by Page in 1954, which works directly off the likelihood ratio of the hypotheses in Eqn. 2.2

$$L = \prod_{i=k}^{T} \frac{P_1(x_i)}{P_0(x_i)} \tag{2.3}$$

In the original conception[2], P_i is a Gaussian distribution with means μ_i; and the change is determined when the logarithm of the above (the cost function) passes some pre-defined threshold. The general properties of Eqn. 2.3 have been well studied [10, 11] with respect to optimal stopping criteria, minimizing false alarms, and using generalized distributions.

2.1.2 Maximum Likelihood Estimation (MLE)

Consider any signal comprised of independent random variables, $\boldsymbol{x} = \{x_1, ..., x_T\}$, sampled via a piece-wise distribution,

$$x_t \sim \begin{cases} F_1 & t_1 < t < t_2 \\ F_2 & t_2 < t < t_3 \\ ... \end{cases}$$

with $t_1, t_2, ...$ as locations of the change points. We can model the cost of a segment $t \in [i, j]$ of our time series generally as

$$L = -\max_{\theta} \sum_{t=i}^{j} \log F(x_t|\theta) \tag{2.4}$$

where we maximize over parameters θ of the distribution F. Our expectation is to find a location where the parameters change significantly. This can in practice be used with any distribution, the most natural choices modeling F as a Gaussian, where one can find shifts in the first two moments, or Poisson, to determine process-rate changes in any set of observations[12].

2.1.3 Other Major Methods

Commonly used with or instead of MLE and CUSUM methods are Bayesian approaches, first developed by A.F.M. Smith in 1975[13], and later evolved to include Markovian[14] and online methods[15]. Attempts at distribution-free changepoint detection (which still relied on knowing the initial distribution) appeared for CUSUM[16] and MLE[17] techniques. This soon led to the development of fully non-parametric approaches starting with using the Mann-Whitney two-sample test by Pettitt in 1979[18], and later spawning a number of rank based methods[19] [20] [21]. Since then, most changepoint detection algorithms have been given a non-parametric or Bayesian alternative[9].

Non-parametric methods have the advantage of not assuming an underlying distribution for the data, and instead rely only upon sampling statistics across segments. Thus, no additional error is introduced by making incorrect prior assumptions about the data.

2.1.4 Applications in Non-Equilibrium Systems

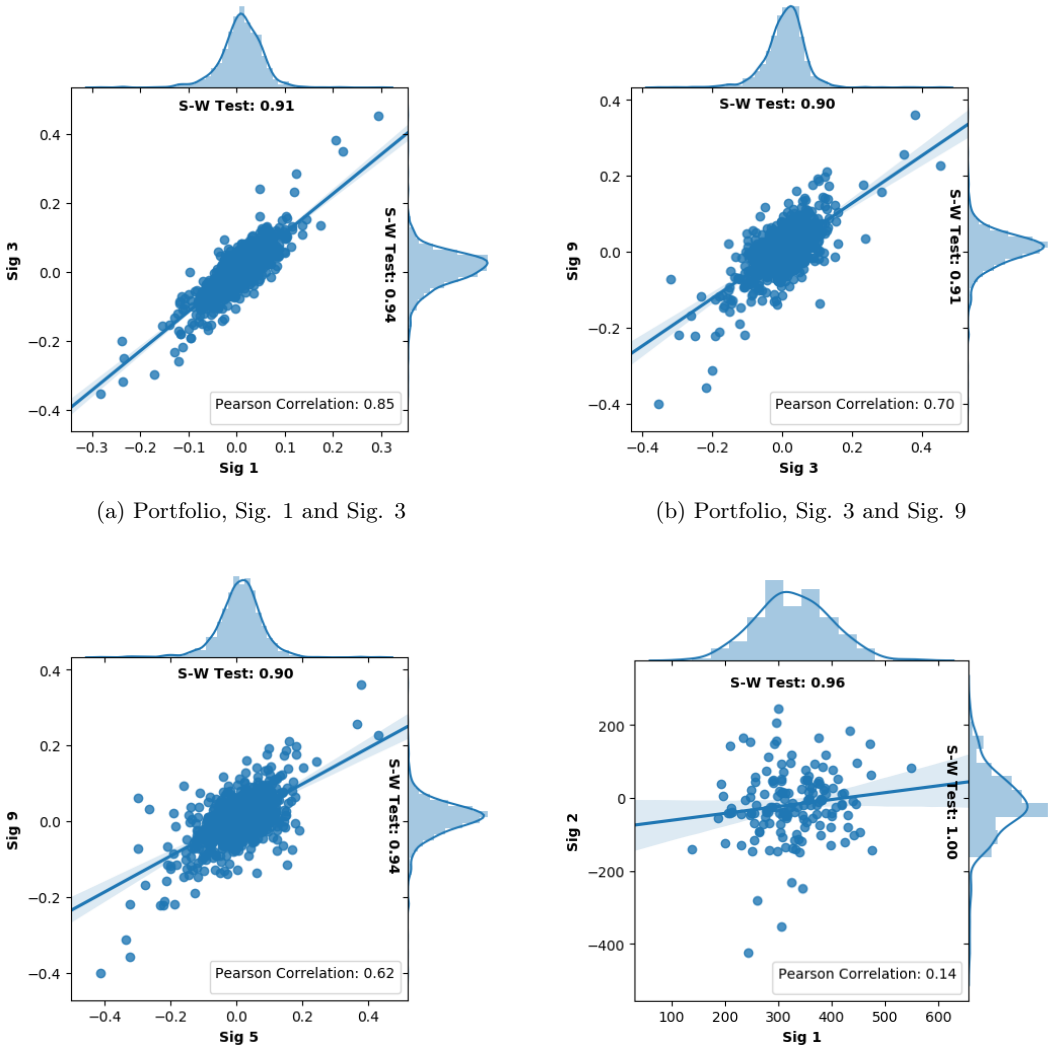
Non-equilibrium systems continuously evolve throughout time, exhibiting novel dynamics at different intervals. These may come about continuously or abruptly, but in either case we cannot represent the statistics as being homogeneous between different processes. Mean-shift changepoint detection is used extensively in the life sciences, measuring concentration changes in biological switches[22] or in ecological climate change[23]. Variance and covariance change detection is used extensively in financial applications[24] and EEG data[25]. These algorithms have even been applied detecting shifts in spending to counteract money laundering[26] and declines in violence with the introduction of new laws[27].

2.2 Data

Selected changepoint detection methods were applied three synthetic and two empirical sets of data, focusing on changes in covariance between signals. Synthetic data was two signals of Gaussian, Exponential, or Cauchy r.v.s, changing in correlation in equally spaced intervals, generated as described in Section 3.4.

The first empirical dataset is value weighted returns for 10 industry portfolios tracked over the period from July 1926 - June 2019, freely available from Kenneth French's homepage at `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html` [28]. The data was assessed with a Shapiro-Wilk test for normality, where we found the strongest deviation being signal 4 with falling into the 85% confidence range. To better fit the model to our Gaussian

assumptions, we took a log transform of the data, such that all signals fell into the 90% confidence interval for being normally distributed. Signals 1, 3, 5, and 9 are presented in Figure 2.1 (a-c) with their corresponding distribution and test statistic.



(a) Portfolio, Sig. 1 and Sig. 3

(b) Portfolio, Sig. 3 and Sig. 9

(c) Portfolio, Sig. 5 and Sig. 9

(d) Rainfall, Sig. 1 and Sig. 2

Figure 2.1: Distribution of values for (a-c) log-normalized value weighted returns for 4 of the 10 industry portfolios and (d) IMR/ENSO rainfall data, along with their corresponding Shapiro-Wilk test statistic and Pearson correlation with another signal.

The second set examined the correlation between the Indian monsoon rainfall (IMR) and El Niño Southern Oscillation (ENSO) data in Brazil over the period from 1849 - 2002. The former data was obtained from the Indian Institute of Tropical Meteorology[29], the latter from Todd Mitchell at the University of Washington[30]. Indian data was the monthly mean precipitation in the mountainous monsoon region. The Brazil data is a combination of two regions, Fortaleza and Quixeramobim, whose rainfall has been standardized, averaged together, and standardized again[30]. For each year, the monthly rainfall was averaged during the El Niño seasons for each region (April - June for Brazil, July - September for India), and a Shapiro-Wilk normality test was performed (Figure 2.1 (d)). It has been argued that there has been a shift in correlation between these two series around the 1980s, which is thought to be linked to climate change[31].

Chapter 3

Methods

Here we discuss our two main methods, HMMs and Bayesian PPMs, which we take to rely both on a prior and assumed distribution of the data. We then introduce a non-parametric binary segmentation algorithm that will serve to contrast the methods above. Finally, we discuss how the synthetic correlated data was produced.

3.1 Hidden Markov Model

Markov models (MMs) are defined as a set of states that randomly switch between one another. The core assumption in MMs is that transition probabilities only depend on the current state, not the past behaviour of dynamics nor initial conditions[32]. As one expects for a physical system, the different states correspond to different processes (and hence outputs). A hidden Markov model (HMM) is a doubly stochastic process; we assume our (probabilistic) outputs are generated by a Markovian process with unobservable states, and hence try to infer (probabilistic) transitions between the underlying processes from the observed data.

The intuition behind the method is clear, different signals are produced by different states, and transitions between states correspond to our changepoints. But, we run into one small issue; we know nothing about the system! How are we to determine which states produce which signal if we have no information about the number of states, nor their transition probabilities? In this case, we are going to have to make some assumptions in the form of a Bayesian prior; making our best estimate to what the initial, transition, and output emission probabilities are and how many states are assumed to exist. After initializing the initial parameters, we can perform maximum likelihood estimation yield a maximum *a posteriori* update. Then, using the updated parameters, can then determine the probability of observing a sequence of states.

3.1.1 Baum-Welch Algorithm for Parameter Fitting

The parameters for our HMM can be obtained via maximizing the likelihood to observe a certain sequence of observations $\{x_1, ..., x_N\}$ from a particular sequence of hidden states $\{y_1, ..., y_N\}$ given a set of model parameters λ:

$$\begin{aligned} P(x_1, ..., x_N|\lambda) &= \sum_{[y_1,...,y_N]} P(y_1, ..., y_N|\lambda) P(x_1, ..., x_N|y_1, ..., y_N, \lambda) \\ &= \sum_{[y_1,...,y_N]} P(y_1) \prod_{t=2}^{N} P(y_t|y_{t-1}) \prod_{t=1}^{N} P(x_t|y_t) \\ &= \sum_{[y_1,...,y_N]} P(y_1) P(x_1|y_t) \prod_{t=2}^{N} P(y_t|y_{t-1}) P(x_t|y_t) \end{aligned}$$

where we are implicitly conditional on model parameters λ. For a model where any hidden state $y \in \{y_1, ..., y_N\}$ can be in any of M states, this takes $N * M^N$ operations! Infeasible to

calculate for even small systems of no more than 100 observations. Luckily, there is a better technique to compute $P(x_1, ..., x_N|\lambda)$ that uses dynamic programming: the forward-backward algorithm. When used in conjunction with expectation-maximization, it is known as the Baum-Welch algorithm[33, 34]. Here we highlight the key steps of the algorithm, for complete reviews see [35] and [36].

Let y_t and x_t be the hidden and observed states at time t. The probability of being in state i at time t given an observed sequence of N data points is:

$$\begin{aligned} P(y_t = S_i|x_1, x_2, ..., x_N) &\propto P(y_t, x_1, x_2, ..., x_N) \\ &= P(y_t, x_1, x_2, ..., x_t)P(x_{t+1}, x_{t+2}, ..., x_N|y_t, x_1, x_2, ..., x_t) \\ &= P(y_t, x_1, x_2, ..., x_t)P(x_{t+1}, x_{t+2}, ..., x_N|y_t) \end{aligned}$$

Using the property that emissions (observations) are only conditional on the current state. The first term and second terms on the RHS we can determine through the forward and backward algorithms respectively.

Let $\alpha_t(i) = P(x_1, x_2, ..., x_t, y_t = S_i|\lambda)$ be the probability to observe the sequence $x_1, ..., x_t$ and be in state S_i at time t, given the parameters of the model λ. We can then find $\alpha_t(i)$ through recursive programming as follows:

Step 1: Initialize

$\alpha_1(i) = \pi_i e_i(x_1)$, $\forall i$, where $\pi_i = P(y_0 = S_i)$ are initial probabilities, and $e_i(t) = P(x_t|y_t = S_i)$ our emission probability for state i to output the value x_t.

Step 2: Recursion

We start by taking the marginal of the joint distribution:

$$\begin{aligned} P(y_t, x_1, ..., x_t) &= \sum_{j=1}^{M} P(y_t, y_{t-1} = S_j, x_1, ..., x_t) \\ &= \sum_{j=1}^{M} P(x_t|y_t, y_{t-1} = S_j, x_1, ..., x_{t-1})P(y_t|y_{t-1} = S_j, x_1, ..., x_{t-1})P(y_{t-1} = S_j, x_1, ..., x_{t-1}) \\ &= P(x_t|y_t) \sum_{j=1}^{M} P(y_t|y_{t-1} = S_j)P(y_{t-1} = S_j, x_1, ..., x_{t-1}) \end{aligned}$$

The first term in the second step we recognize as the output at time t, which is only dependent on the hidden state (i.e. the emission probability). Likewise, the second term is the state at time t, which is only conditional on the state at t-1. Finally, the last term is the result for the forward algorithm at time t-1, which provides the basis for the iterative algorithm.

The only issue is that this only yields the probability to observe the data set up to time t, in order to get the joint probability for all our observations we must use the backward algorithm. The form is very similar to the forward case, here we determine $\beta_t(i) = P(x_{t+1}, ..., x_N|y_t = S_i)$, i.e. the probability to observe values x_{t+1} to x_N given that at time t we are in state y_t.

Step 1: Initialize (arbitrarily)

$\beta_N(i) = P(x_{N+1}|y_N = S_i) = 1$

Step 2: Recursion

Again, start with the marginal

$$\begin{aligned}
\beta_t(i) &= P(x_{t+1}, ..., x_N | y_t) \\
&= \sum_{j=1}^{M} P(x_{t+1}, ..., x_N, y_{t+1} = S_j | y_t) \\
&= \sum_{j=1}^{M} P(x_{t+2}, ..., x_N | x_{t+1}, y_t, y_{t+1} = S_j) P(x_{t+1}, y_{t+1} = S_j, y_t) \\
&= \sum_{j=1}^{M} P(x_{t+2}, ..., x_N | x_{t+1}, y_t, y_{t+1} = S_j) P(x_{t+1} | y_{t+1} = S_j, y_t) P(y_{t+1} = S_j | y_t) \\
&= \sum_{j=1}^{M} P(x_{t+2}, ..., x_N | y_{t+1} = S_j) P(x_{t+1} | y_{t+1} = S_j) P(y_{t+1} = S_j | y_t = S_i)
\end{aligned}$$

Where we have factored the distribution twice, then are left with three terms on the RHS. The first is the backward algorithm result for t+1, the second is the emission probability, and the last is the transition probability. With the results from the forward and backward algorithm, one can then find the posterior probability $P(y_t | x_1, ..., x_N)$, from which we can use the Baum-Welch algorithm to update our model parameters.

From the product rule we can determine the probability to occupy state i at time t:

$$\begin{aligned}
\gamma_i(t) &= P(y_t = S_i | x_1, ..., x_N, \lambda) \\
&= \frac{P(y_t = S_i, x_1, ..., x_N | \lambda)}{P(x_1, ..., x_N | \lambda)} \\
&= \frac{\alpha_i(t)\beta_i(t)}{\sum_{j=1}^{M} \alpha_j(t)\beta_j(t)}
\end{aligned}$$

given that we have a sequence of observations $x_1, ..., x_N$. Likewise, we can determine the joint probability to be in state i and j at times t and t+1 respectively as:

$$\begin{aligned}
\eta_{ij}(t) &= P(y_t = S_i, y_{t+1} = S_j | x_1, ..., x_N, \lambda) \\
&= \frac{P(y_t = S_i, y_{t+1} = S_j, x_1, ..., x_N | \lambda)}{P(x_1, ..., x_N | \lambda)} \\
&= \frac{\alpha_i(t) a_{ij} \beta_j(t+1) e_j(x_{t+1})}{\sum_{i,j=1}^{m} \alpha_i(t) a_{ij} \beta_j(t+1) e_j(x_{t+1})}
\end{aligned}$$

Where we have denoted the transition probability from i to j as a_{ij}. From these, we can now obtain our HMM parameters[36]:

$$\begin{aligned}
\pi_i* &= \gamma_i(1) \\
a_{ij}* &= \frac{\sum_{t=1}^{N-1} \eta_{ij}(t)}{\sum_{t=1}^{N-1} \gamma_i(t)}
\end{aligned}$$

Our emissions were modeled as Gaussian distributions, i.e. $e_i(x_t) = P(x_t | y_t = S_i) = Normal(x_t | \mu_i, \Sigma_i)$. The mean and variance of these can be estimated as follows[35]:

$$\begin{aligned}
\mu_{ik} &= \frac{\sum_{t=1}^{N} \gamma_i(t) x_t}{\sum_{t=1}^{N} \gamma_i(t)} \\
\Sigma_{ik} &= \frac{\sum_{t=1}^{N} \gamma_i(t)(x_t - \mu_i)(x_t - \mu_i)^T}{\sum_{t=1}^{N} \gamma_i(t)}
\end{aligned}$$

Thus, by starting with some prior estimate of our initial, transition, and emission probabilities, we can iteratively run the Baum-Welch algorithm until convergence to fit our model to the data.

3.1.2 Viterbi Algorithm for Determining State Changes

Now that we have the model parameters, we can ask the question: given a set of observations, which sequence of states best fits the data? This is typically taken as maximizing the complete sequence $P(y_1, ..., y_N|x_1, ..., x_N, \lambda)$, rather than optimizing over individual states. The fastest known method of computing this is the Viterbi[37] algorithm.

Let the highest probability path, based on the observations and model parameters, up to time t be denoted by:

$$\zeta_t(i) = \max_{y_1, y_2, ..., y_{t-1}} P(x_1, ..., x_t, y_1, ..., y_{t-1}, y_t = S_i|\lambda)$$

Where we have used $P(y_1, ..., y_N|x_1, ..., x_N, \lambda) \propto P(y_1, ..., y_N, x_1, ..., x_N, \lambda)$ and made our conditioning on λ implicit. We can then observe:

$$\begin{aligned}
\zeta_{t+1}(i) &= \max_{y_1, y_2, ..., y_t} P(x_1, ..., x_t, y_1, ..., y_t, x_{t+1}, y_{t+1} = S_i) \\
&= \max_{y_1, y_2, ..., y_t} P(x_{t+1}, y_{t+1} = S_i|x_1, ..., x_t, y_1, ..., y_t) P(x_1, ..., x_t, y_1, ..., y_t) \\
&= \max_{y_1, y_2, ..., y_t} P(x_{t+1}, y_{t+1} = S_i|y_t) P(x_1, ..., x_t, y_1, ..., y_t = S_k) \\
&= \max_k P(x_{t+1}, y_{t+1} = S_i|y_t = S_j) \max_{y_1, y_2, ..., y_{t-1}} P(x_1, ..., x_t, y_1, ..., y_t = S_k) \\
&= \max_k P(x_{t+1}, y_{t+1} = S_i) P(y_{t+1} = S_i|y_t = S_j) \zeta_t(k)
\end{aligned}$$

Thus, we again compute a simple recursive relationship $\zeta_{t+1}(i) = \max_j \zeta_t(j) a_{ji} b_i(x_{t+1})$ similar to the Baum-Welch procedure. These two algorithms allow us to start with an initial guess for our model parameters, which are then updated to then give us the underlying state sequence for the data. In this model, we take our priors to be composed of random uniform variables on (0,1), running the iterative process for 100,000 steps.

3.2 Bayesian Product Partition Model

Our secondary method is a Bayesian approach to the Product Partition Model (PPM)[38, 39]. A PPM describes the pdf of the distribution as being made up of several independent partitions, such that the complete distribution is given by:

$$P(x_{1:N}|\hat{t}) = \prod_{k=1}^{K} P_{\hat{t}_k}(x_{\hat{t}_k}) = \prod_{k=1}^{K} P(x_{\hat{t}_k}) \tag{3.1}$$

Where the pdf of our observed values, $x_{1:N}$, conditional on the partitions, $\hat{t}$, is the product of pdfs for the observations from individual segments. We make two assumptions regarding the partitions: 1) that they do not overlap and 2) each is parameterized by a multivariate Gaussian.

This method is similar to a HMM, in that we assume the signal takes on a variety of different states. However, unlike the HMM, we do not assume that we will be revisiting old states. Instead, each segment is assumed to be produced by a completely new set of parameters. In an infinite HMM[40], we also have an unbounded number of states, but also we can revisit old states as well. Though the infinite HMM cannot use the algorithms described in Section 3.1, nor the algorithms described here, and as such rely on much more computationally expensive methods.

Our goal is to compute the maximum *a posteriori* segmentation:

$$(K^*, \hat{t}^*) = \arg\max_{K, \hat{t}_{1:K}} P(x_{1:N}|\hat{t}_{1:K}) P(\hat{t}_{1:K})$$

using the recursive algorithm proposed by Fearnhead[41, 42] and extended to the multi-dimensional case by Xuan and Murphy[43]. The algorithm also relies on the forward-backward procedure, though instead of maximizing over parameters of the model, we instead maximize the probability to observe some segment $x_{i:j}$ given a model m (i.e. $P(x_{i:j}|m)$).

The approach is both Bayesian and parametric, thus to obtain the posterior,

$$P(\hat{t}_{1:K}|x_{1:N}) = \frac{P(x_{1:N}|\hat{t}_{1:K})P(\hat{t}_{1:K})}{P(x_{1:N})} \tag{3.2}$$

we require three inputs: 1) the observation likelihood $P(x_{i:j}|m)$, 2) a prior on the length and positions of the segments $P(l)$ and $P(\hat{t})$, and 3) a prior over the models $P(m)$[43]. The likelihood of observation was modeled by a multivariate Gaussian using the full-covariance model from Xuan and Murphy[43]. The analytic expression for the marginal, $P(x_{i:j}) = \int P(x_{i:j}|m')P(m')dm'$ is:

$$P(x_{i:j}) = \pi^{-\frac{(j-i+1)d}{2}} \frac{|V_0|^{N_0/2}}{|V_{(j-i+1)}|^{(N_0+(j-i+1))/2}} \frac{\Gamma_d(N_0/2)^{-1}}{\Gamma_d((N_0+(j-i+1))/2)^{-1}} \tag{3.3}$$

$$V_n = V_0 + \sum_{s=i}^{j} x_s x_s^T \tag{3.4}$$

$$\Gamma_d(t) = \pi^{d(d-1)/4} \prod_{i=0}^{t-1} \Gamma(t - \frac{i}{2}) \tag{3.5}$$

Where N_0, V_0 are the hyper-parameters for the inverse-Wishart prior of our covariance matrix, $\Sigma \sim IW(N_0, V_0)$, taken to be $N_0 = d$ and $V_0 = \hat{\sigma}^2 I$ for simplicity, where d is the dimension of the data and $\hat{\sigma}^2$ the empirical variance. In this report, we chose to exclude mean changes from our model.

Unfortunately, this method has several limitations. The first is computational, as the process involves computing 3 NxN matrices (where N is the number of data points) which uses an enormous amount of memory and time for large data sets. There are some online variations of the algorithm that circumvent this issue[15], though there was no straightforward way to adapt the method to the multivariate case. Second, for r.v.s with undefined or infinite variance, the algorithm could not construct positive semi-definite covariance matrices, which is fine given the Gaussian assumptions made. Thirdly[43], in higher dimensions the covariance matrix requires a significant amount of data for each segment. If the segments are small, we are bound to miss transitions to other partitions; though this can be solved by choosing a MLE prior for the covariance matrix.

3.2.1 Choice of Prior

Following Fearnhead[41], we used a negative binomial distribution for the prior on segment lengths between two changepoints:

$$P(l) = \binom{l-k}{k-1} p^k (1-p)^{l-k} \tag{3.6}$$

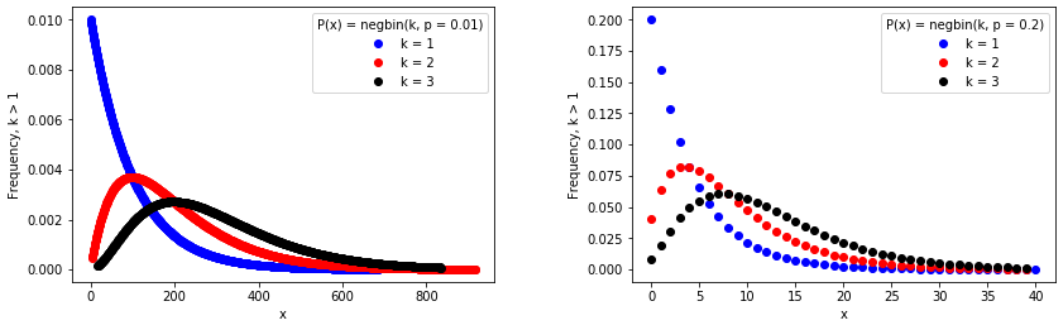
$$\tag{3.7}$$

The positions for the K changepoints can then be determined by estimating from the most likely lengths:

$$P(\hat{t}_{1:K}) = P(\text{First length} = \tau_1) \prod_{j=2}^{K} P(\text{Length} = \tau_j - \tau_{j-1}) P(\text{No more change-points}) \tag{3.8}$$

$$= P_0(l = \tau_1) \prod_{j=2}^{K} P(l = \tau_j - \tau_{j-1})(1 - \sum_{s=1}^{N-\tau_K} P(l = s)) \tag{3.9}$$

Though, if you have an idea of how the change-points should be distributed, i.e. many rapid or few slow transitions, the prior can be more accurately chosen. Below in Figure 3.1 is the negative binomial distribution for various k and p. Lower values of k and higher p increase the probability of observing short segments. For the size of our data sets we obtained best results for k = 2 and p = 0.01. Higher values of p contributed to more false positives at the beginning of the series, while higher k made it more likely to yield false negatives.



(a) p = 0.01 (b) p = 0.20

Figure 3.1: The negative binomial distribution for k = [1, 2, 3], and p = [0.01, 0.20]. High value of k and low values of p correspond favors less changepoints and longer segments.

3.2.2 Bootstrapping Thresholds

The Bayesian PPM computes posterior probabilities that a change occured at time t, rather than a discrete state value like the HMM. As such, we needed a method to discern which points were legitimate shifts in the parameters and which were artifacts. Hence, we utilized a bootstrapping procedure detailed in Carlstein et al. 1994 [44]. Bootstrapping is a non-parametric method of statistical inference for testing observations with random samples of the observed data[45]. We took 1000+ random permutations of the data, and on each ran the Bayesian PPM. For each run we obtained the maximum posterior probability that a changepoint was detected. We then took our confidence interval to be bounded by the lower 95% of the values obtained over random permutations of the data.

3.3 A Non-Parametric Model: Binary Segmentation

We have decided to contrast our parametric models with a non-parametric binary segmentation (BS) approach, arguably the most popular method of changepoint detection[46]. BS works on the hypothesis in Eqn. 2.2 assuming there is one change in the data, which is determined through the minimization of a cost function:

$$\tau_1 = \arg\min_{1<t<N} C(x_{1:t}) + C(x_{t+1:N}) \tag{3.10}$$

Multiple changepoints are detected by splitting the signal into segments at each prior result, and running the algorithm on each individual segment. The process stops once the predefined maximum number of points have been found, or the cost function exceeds some threshold. For our purposes, finding change in the covariance structure, we have chosen two cost functions:

$$C_{normal}(x_{i:j}) = (j-i)\log|\hat{\Sigma}_{i:j}|$$

$$C_{kernel}(x_{i:j}) = (j-i) - \frac{1}{j-i}\sum_{r,s=i+1}^{j} \exp -\gamma||x_r - x_s||^2$$

The first detects changes in Gaussian processes, where $\hat{\Sigma}_{i:j}$ is the empirical covariance matrix over the interval (i,j). The second computes changes in the kernalized mean, where γ is the inverse median value on (i,j). The binary segmentation algorithm and associated cost functions were implemented using the python Ruptures package.

3.4 Generating Correlated Random Variables

In our simulated data, we faced the problem of generating random variables (r.v.s) from a non-normal distribution that are correlated according to a specified coefficient. Our method of data generation for two signals is as follows:

Algorithm 1 Generating Correlated Non-Normal R.V.s

procedure GEN_DATA($dist, corr[array], n\_samples, n\_segments$)
  Initialize Data Matrix : D_mat
  for $i \leftarrow 1, n\_segments$ **do**
    $samp_1 \leftarrow dist(n\_samples)$ ▷ Generate samples from distribution
    $samp_2 \leftarrow dist(n\_samples)$
    $samp_c \leftarrow zeros(n\_samples)$
    $p \leftarrow Uniform([0, 1], n\_samples)$
    for $j \leftarrow 1, n\_samples$ **do**
      if p[j] $<$ corr **then**
        $samp_c[j] = samp_1[j]$ ▷ Draw from '1' w/ prob $= |corr|$
      else
        $samp_c[j] = samp_2[j]$ ▷ Draw from '2' w/ prob $= 1 - |corr|$
      end if
    end for
    if $corr[i] < 0$ **then**
      $samp_c = -samp_c$
    end if
    $D\_mat[, sig1] \leftarrow samp_1 + noise$ ▷ Add Gaussian Noise
    $D\_mat[, sig2] \leftarrow samp_c + noise$
  end for
end procedure

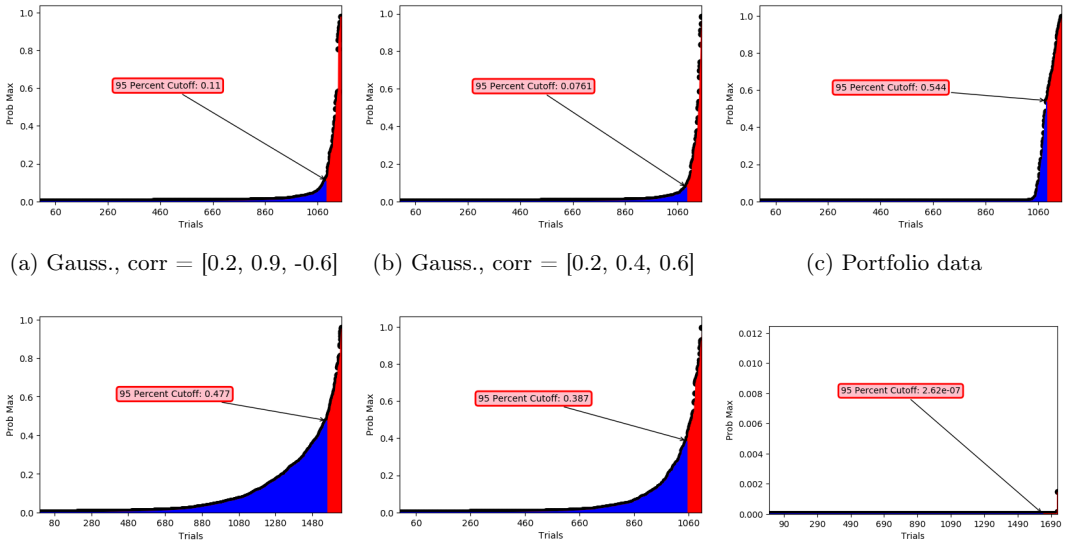
The result is a matrix of r.v.s, with each segment of length $n\_samples$ having close to the specified correlation between signals. We have added Gaussian noise to increase the variability of the data, and reduce any over-fitting by our models. Exponential r.v.s were given a shape parameter of 2.

Chapter 4

Results

4.1 Bootstrap Results for Bayesian PPM

Our bootstrapping thresholds were determined over 1000+ runs across each dataset, in each run using a random permutation of the original data. All results are shown in Figure 4.1, with the blue shaded region representing values that fall below our 95% confidence interval. For many runs, we obtain high changepoint probabilities even on permuted data, hence illustrating that detections may be simply anomalies. However, for most runs this is not the case, and one should be confident values surpassing the threshold are worth investigating.



(a) Gauss., corr = [0.2, 0.9, -0.6] (b) Gauss., corr = [0.2, 0.4, 0.6] (c) Portfolio data

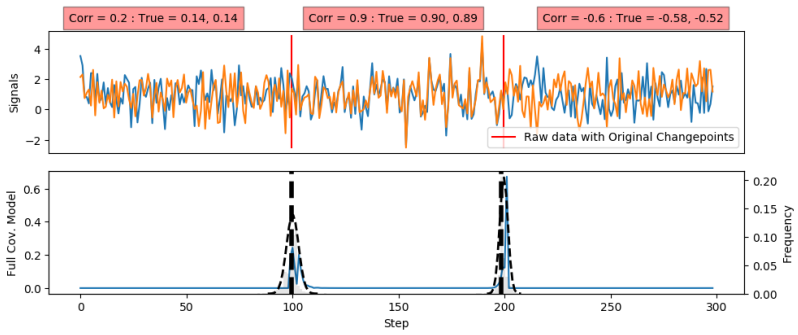
(d) Exp., corr = [0.2, 0.9, -0.6] (e) Exp., corr = [0.0, 0.5, 1.0] (f) Rainfall data

Figure 4.1: Bootstrapping results for the Bayesian PPM using the method described in Section 3.2.2. 'corr' indicates the correlation between signals in each segment.

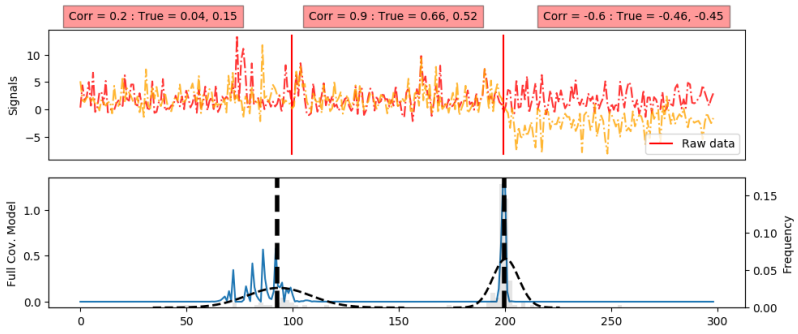
4.2 Comparison of Results: Synthetic Data

Our three synthetic data sets were two dimensional signals of random variables drawn from an Exponential, Gaussian, and Cauchy distribution that changed correlation at predefined points, as described in Sections 2.2 and 3.4. Results are presented in Figures 4.3, 4.2, and 4.4 for the HMM, PPM, and BS algorithms respectively. For each method, we generated 200+ samples of random data with the desired correlation, and generated histograms of the changepoint locations which were fit to Gaussians as shown by dashed black lines. For HMMs, changepoints were determined

when there was a transition between three states shown in purple, blue, and green. In the Bayesian PPM, we counted each step that surpassed the bootstrap cutoff.



(a) Gaussian r.v.s



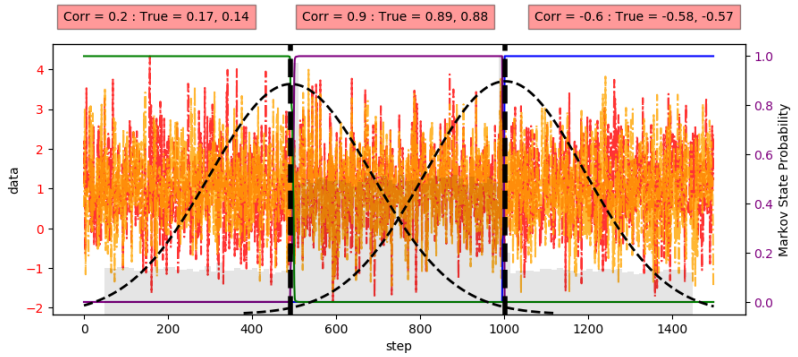
(b) Exponential r.v.s

Figure 4.2: Bayesian PPM for correlated (a) Gaussian and (b) Exponential random variables. The bottom figure shows the posterior probability of a changepoint for the current run (blue) and the histogram of results for previous runs, with a dashed black line for the most likely transition over 200 runs. The ideal, Pearson, and Spearman correlations for each segment are boxed in red.

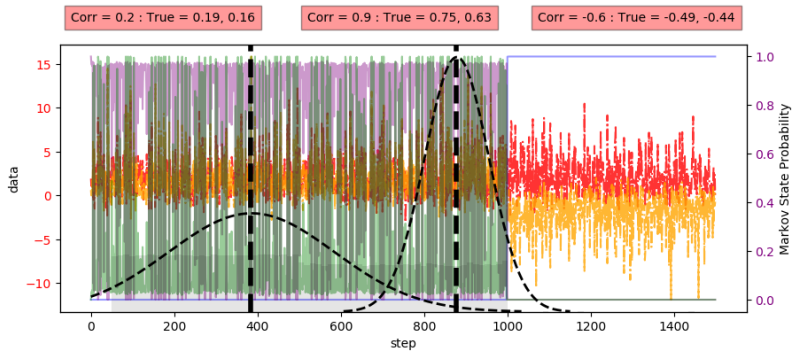
Gaussian distributed r.v.s perform the best in all models, although in the HMM, there is a large variance in the distribution of state transitions. This is due to some trials having fast oscillations between states, as in the case of the Exponential distribution in Figure 4.3 (b). We take this to be resultant from three causes: 1) the size and 2) locations of the segments, and 3) our use of random priors. HMMs have a large number of unknown parameters, including the initial, transition, and emission probabilities for all states. Without sufficient data to observe for each state, the emission probabilities cannot be accurately assessed. Likewise, without multiple transitions between states, the transition probabilities will often be inaccurate. It has been suggested that these problems can be alleviated by using multiple sequences of data (i.e. training sets)[47], rather than fitting to the entire set of observations. Additionally, the large phase space for the parameters also means that the system may become stuck in a local minimum, which necessitates the use of relevant priors.

For exponential variables, our results begin to falter for all methods. The HMM struggles to predict the changepoint between the first two regions, though is accurate at finding where the correlation turns negative. We take this to be due to the significant change in mean and variance at this point, rather than the algorithm detecting the covariance change. Likewise, the Bayesian PPM and BS algorithm are able to detect the shift to negative correlation, however detection around the first changepoint is weaker than in the Gaussian case.

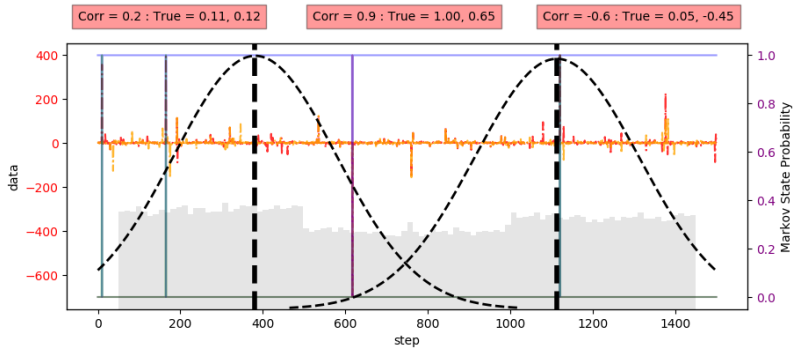
Finally, for Cauchy r.v.s, the parametric approaches fail completely. The HMM has minor discontinuities in the transition histogram around the changepoints, but otherwise is a nearly uniform distribution. For the Bayesian PPM, the algorithm completely was unable to generate a suitable covariance matrix for most trials with the Cauchy distribution, and was discarded entirely. On the other hand, the non-parametric binary segmentation approach performed very well. The normal cost method, although heavily skewed, produced semi-accurate estimation of the changepoints; wheras the rbf cost function obtained exceptional results.



(a) Gaussian r.v.s

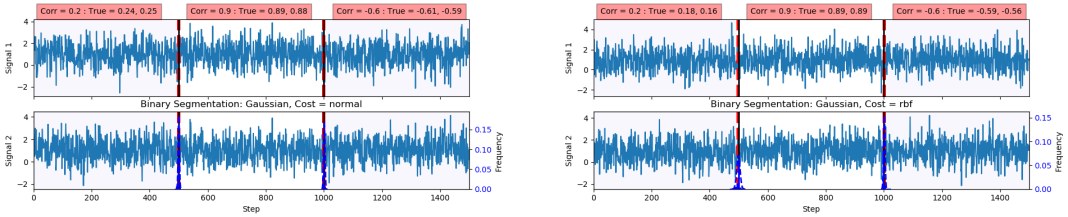


(b) Exponential r.v.s



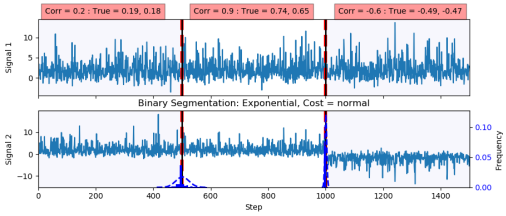
(c) Cauchy r.v.s

Figure 4.3: State occupation probabilities for (a) Gaussian, (b) Exponential, and (c) Cauchy correlated r.v.s for a 3 state HMM. The ideal, Pearson, and Spearman correlations for each segment are boxed in red. Black lines represent the distribution of transitions that occurred during 1000 runs of the model, corresponding to the underlying histograms. Dashed lines indicate the most likely transitions. True changepoints are at 500 and 1000 for all distributions.

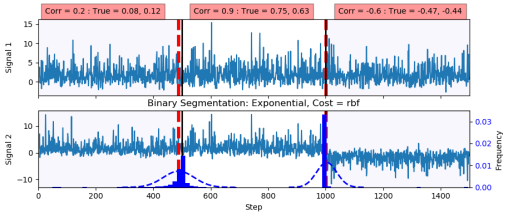


(a) Gaussian r.v.s, cost = normal

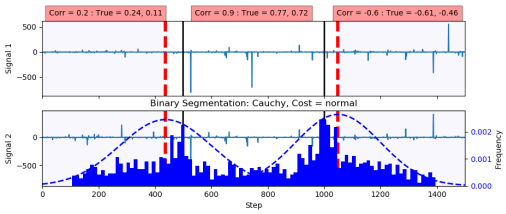
(b) Gaussian r.v.s, cost = rbf



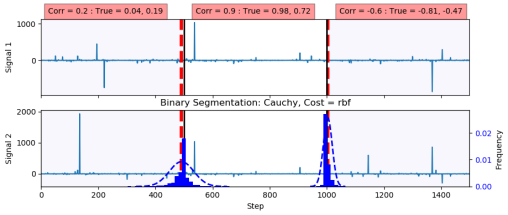
(c) Exponential r.v.s, cost = normal



(d) Exponential r.v.s, cost = rbf



(e) Cauchy r.v.s, cost = normal

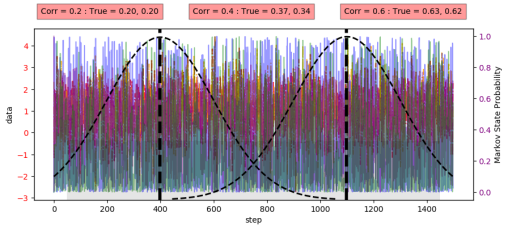


(f) Cauchy r.v.s, cost = rbf

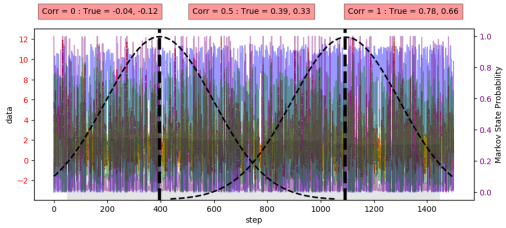
Figure 4.4: Binary segmentation for correlated (a,b) Gaussian, (c,d) Exponential, and (e,f) Cauchy r.v.s. The ideal, Pearson, and Spearman correlations for each segment are boxed in red. A histogram of results for 1000 runs, fitted to a Gaussian, indicates the most likely changepoint locations estimated by the model. True changepoints are shown by black lines.

The correlations used for the comparisons, [0.2, 0.9, -0.6], had been chosen with the aim of being large enough to detect. However, there is no guarantee that empirical data should have such correlations. Hence, we performed the computations again, using smaller changes between segments. For the Cauchy variables, we attempted to increase the correlation, hoping our HMM would be able better classify the segments. Results are presented in Figures 4.5, 4.6, and 4.7

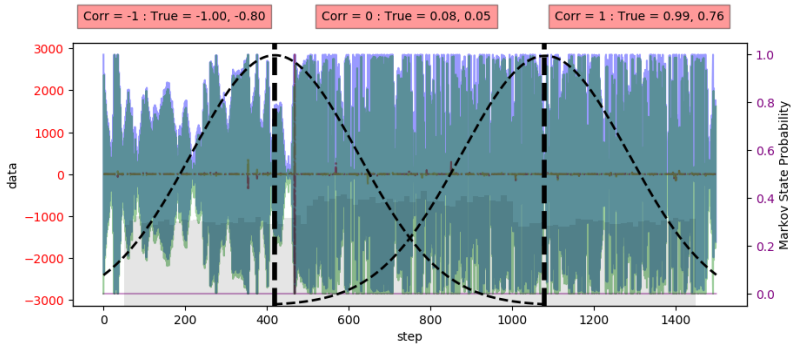
The HMM terribly represents the data, with state changes being uniform over the entire signal. It appears the algorithm was not able to determine the best sequence without an informative prior. Additionally, because the Markov assumption disregards the role of time, there is no penalty to having multiple switches within a region as there is for the Bayesian PPM. In the Cauchy case, we see slight step-wise behaviour in the histogram at the changepoints, but without definitive results. Comparatively, the Bayesian PPM and BS method performed well for closely correlated variables, albiet a higher variance in the results. In this case, the low number of samples used in the PPM compared to the BS algorithm may have contributed to its inaccuracy, as finer detail, and thus more information, is need. To test this, we ran the PPM again with 1500 observations for the exponential case over 10 trials (Figure 4.6 (c)), and found no significant increase in accuracy. Had more trials been taken, perhaps we could yield comparable results, but the BS algorithm was orders of magnitude faster in terms of computation time, and thus of better practical use.



(a) Gaussian r.v.s

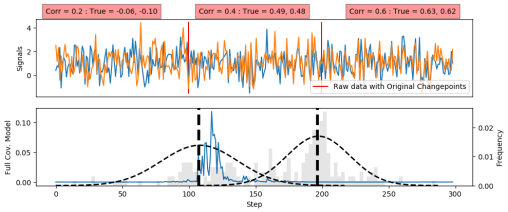


(b) Exponential r.v.s

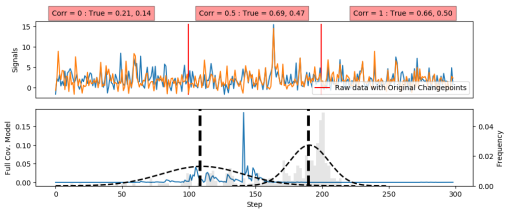


(c) Cauchy r.v.s

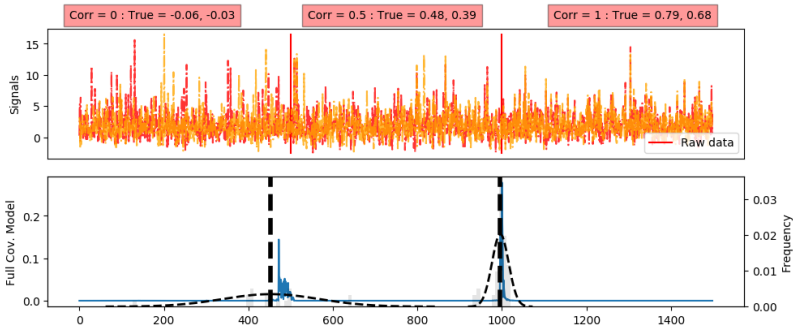
Figure 4.5: Similar to Figure 4.3, but with more extreme correlations to test the efficacy of the HMM. For (a) Gaussian and (b) Exponential the method fails completely. In (c), the more extreme correlation shifts did pronounce the step-wise behaviour more clearly, though the HMM still produces poor results.



(a) Gaussian r.v.s

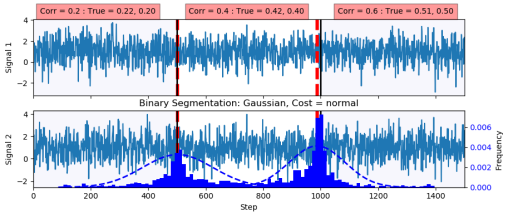


(b) Exponential r.v.s

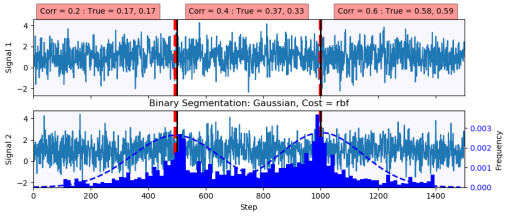


(c) Exponential r.v.s with 1500 samples

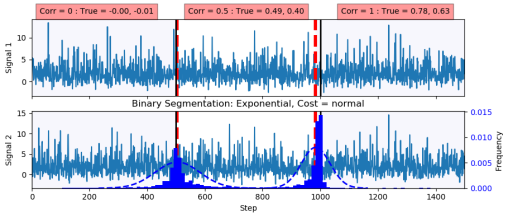
Figure 4.6: Figure 4.2 but with closer correlation changes for (a) Gaussian and (b,c) Exponential r.v.s. In (c) we considered 1500 data points, though only sampled over 10 runs.



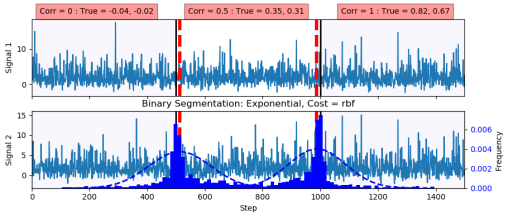
(a) Gaussian r.v.s, cost = normal



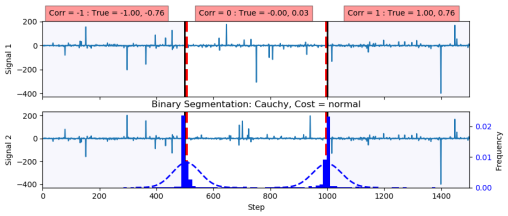
(b) Gaussian r.v.s, cost = rbf



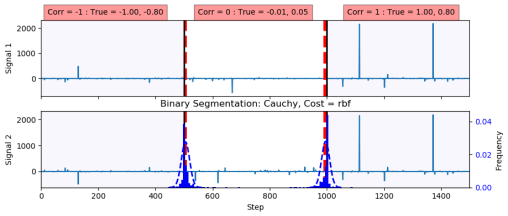
(c) Exponential r.v.s, cost = normal



(d) Exponential r.v.s, cost = rbf



(e) Cauchy r.v.s, cost = normal

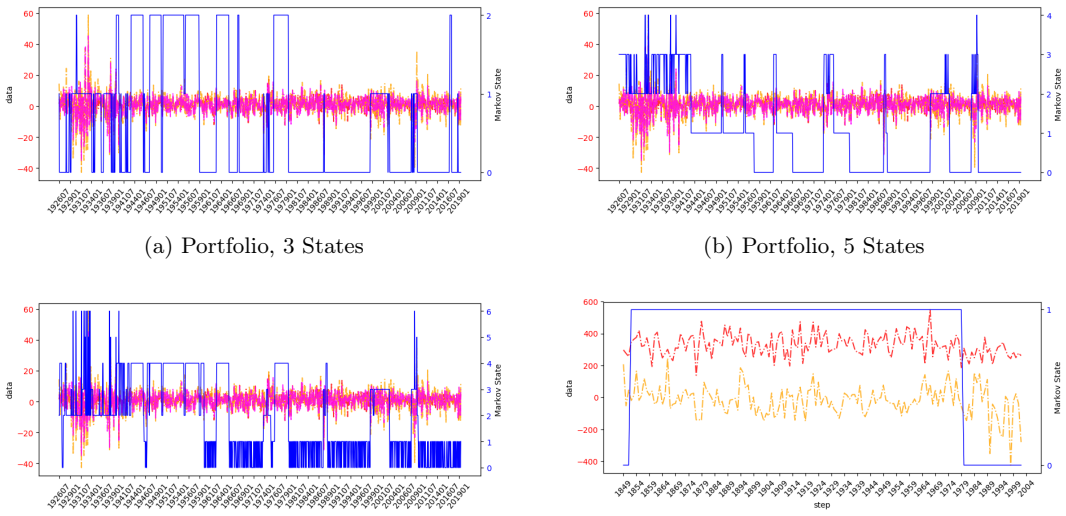


(f) Cauchy r.v.s, cost = rbf

Figure 4.7: Essentially equivalent to Figure 4.4 but with reduced changes in correlations between segments for (a,b) Gaussian and (c,d) Exponential r.v.s, and increased for (e,f) Cauchy r.v.s.

4.3 Comparison of Results: Empirical Data

With an unknown number and location of changepoints for the portfolio data, we simply compared our results across each method and with those obtained from a generalized graphical model (GGM) of the same data in Xuan and Murphy (2007) [43], choosing to use 3, 5, and 7 hidden states in the HMM, and 1-7 changepoints in BS. In the HMM, a segment was considered to be portions of the dataset which a) oscillated between two states or b) was generally in one state. The inferred changepoints are presented in Table 4.1, with good agreement between the separate HMMs and Xuan's GGM method.



(a) Portfolio, 3 States

(b) Portfolio, 5 States

(c) Portfolio, 7 States

(d) Rainfall, 2 States

Figure 4.8: HMM on the (a-c) portfolio and (d) rainfall data described in Section 2.2 using a) 3, b) 5, c) 7, and d) 2 states.

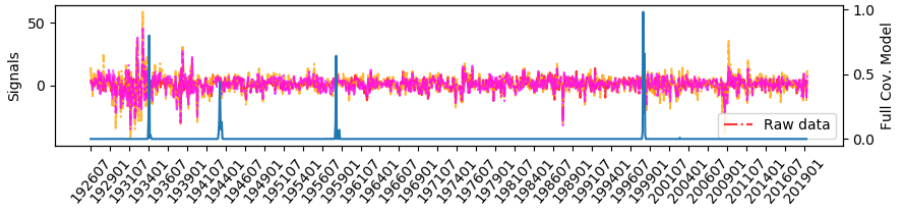
For the Bayes PPM, we have detected changes in July 1931, January 1941, October 1955, and January 1966; however, the second does not exceed the threshold determined by the bootstrap procedure. BS yields the dates February 1929, January 1934, November 1939, January 1966, and March 1977 in the case of five changepoints using the rbf cost function. The algorithm experienced high clustering of results around the beginning of the dataset when using the normal cost function; where it seemed to be picking up large, frequent changes in variance. This tended to occur for the rbf method as well beyond five changepoints, and hence provided a reasonable cutoff.

Table 4.1: Detected changepoints for 3, 5, and 7 state HMMs on portfolio data

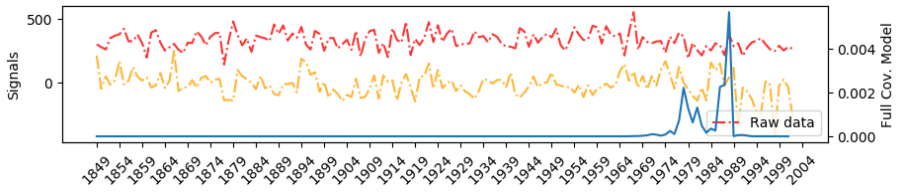
3 States	5 States	7 States	Xuan GGM
Dec 1928	—	—	1930
—	—	—	—
Aug 1940	Aug 1943	Dec 1940	1942
—	—	—	—
Jan 1959	—	Nov 1957	1959
—	—	—	—
—	July 1966	—	—
Jan 1976	June 1971	Nov 1974	1974
Mar 1979	Apr 1980	May 1979	—
May 1996	—	Jan 1997	1996
May 2002	—	May 2003	—

The aggregation of results for the three models are given in Table 4.2, with values in green indicating agreement between all models. Highlighted values may correlate to the Wall Street

crash of 1929 and the end of the Great Depression in 1939, but we are wary about making spurious correlations, which is one major difficulty in changepoint detection. When a change is found, immediately one can try to reason about it and interpret the results, even if the event is an anomaly. All we can say is that at several dates the models agree, and there seems to be no real outliers other than the 2002-2003 values for the HMMs that were absent from other methods.



(a) Portfolio



(b) Rainfall

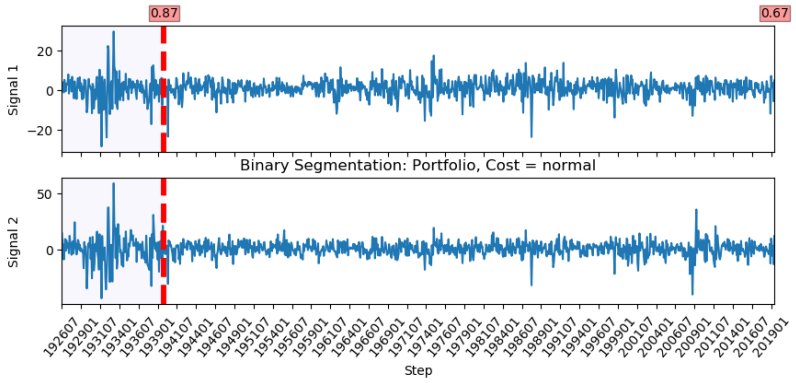
Figure 4.9: Bayesian PPM results for (a) portfolio and (b) rainfall datasets. Blue lines represent the posterior probability for a changepoint at that location.

Table 4.2: Identified changepoints across all methods considered, green text indicates common detections.

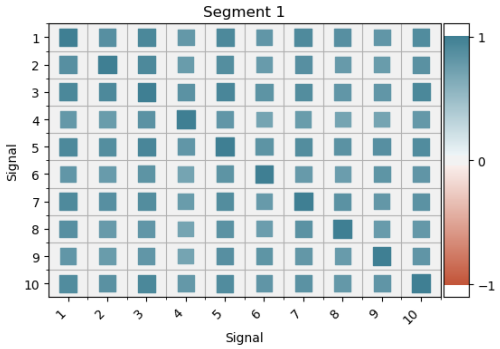
HMM	1928	—	1940-1943	1957-1959	1966	1971-1974	1976	1979-1980	1996-1997	2002-2003
Bayes PPM	1931	—	1941	1955	—	—	—	—	1966	—
BS	1929	1934	1939	—	1966	—	—	1977	—	—

For the rainfall data, we detected changepoints at 1980, (1977 and 1987), and 1978 for the 2-State HMM, PPM, and BS algorithm respectively. Our aim was to study whether these two series, the IMR and ENSO data, had a change in covariance between them. To see whether this was what the algorithms were determining, we computed the mean, variance, and correlation matrix for each signal within each segment. From Figure 4.12 (b-e), large changes in both the mean and variance for the two signals is found across the segments. There is a slight shift from positive to negative correlation, but remarkably minor. Thus, we conclude that our tests were inconclusive at determining a covariance change between the two series, and were most likely dominated by the mean and variance shifts.

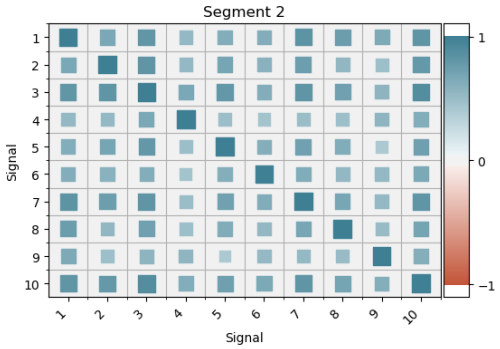
Likewise, we considered the same procedure for the portfolio data in Figure 4.11 (b-e). Similarly, it seems the first two changepoints are due to large shifts in mean and variance, though the next three segments have little change in these parameters. Looking at the covariance matrix between segments 5 and 6, there is a clear visual difference between each segment. With the small changes in mean and variance, we can say that it is probable that the covariance change has a large effect on labeling this specific point as a transition.



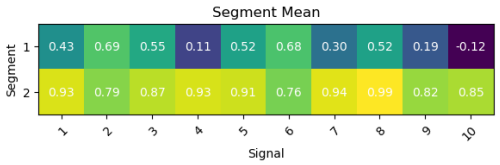
(a) Portfolio BS changepoint detection



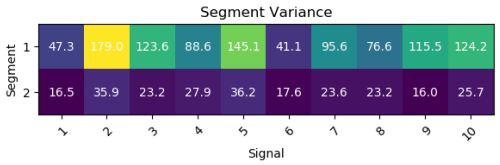
(b) Correlation matrix on segment 1



(c) Correlation matrix on segment 1

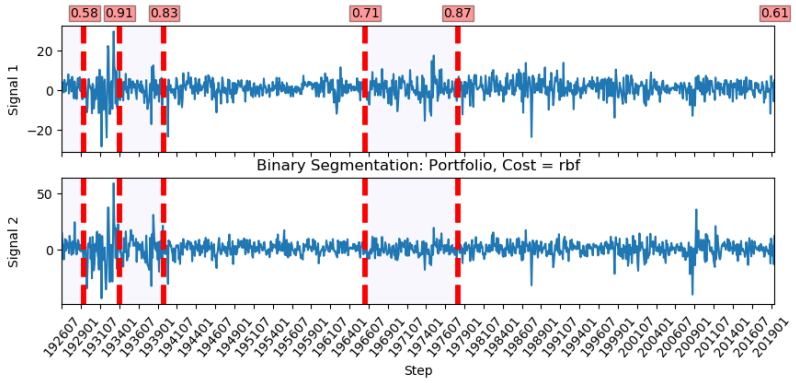


(d) Means across segments

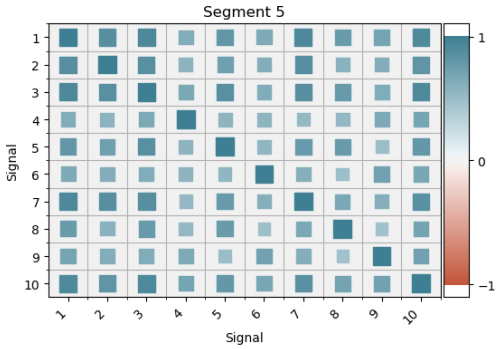


(e) Variances across segments

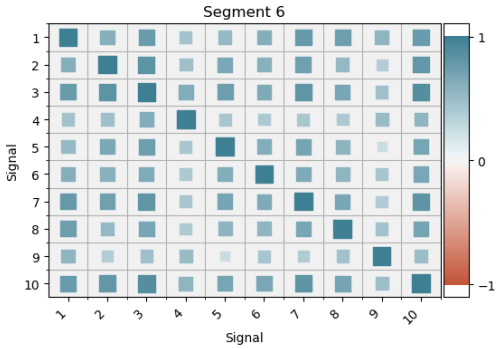
Figure 4.10: BS detection of a single changepoint in portfolio data, showing two of the ten signals. Red labels above (a) indicate the correlation of the first two signals in the prior segment. (b,c) are correlation matricies between all ten signals in the two segments. (d,e) show mean and variance for each signal across both segments.



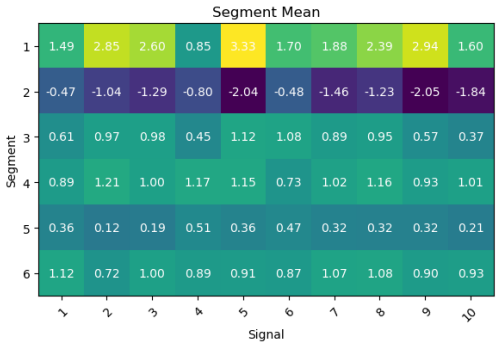
(a) Portfolio BS changepoint detection



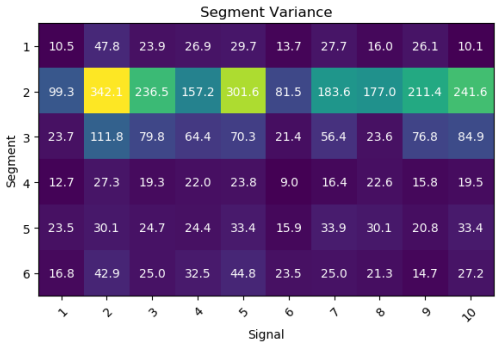
(b) Correlation matrix on segment 5



(c) Correlation matrix on segment 6

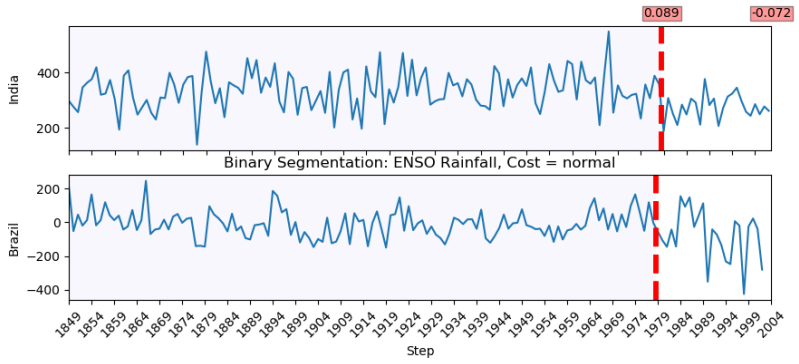


(d) Means across segments

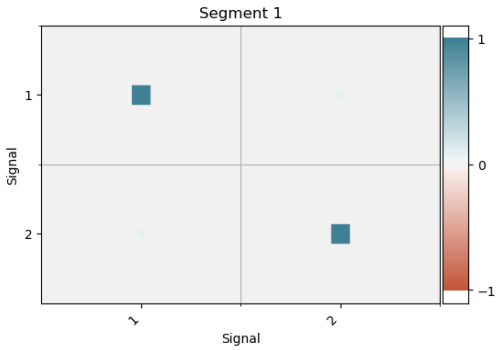


(e) Variances across segments

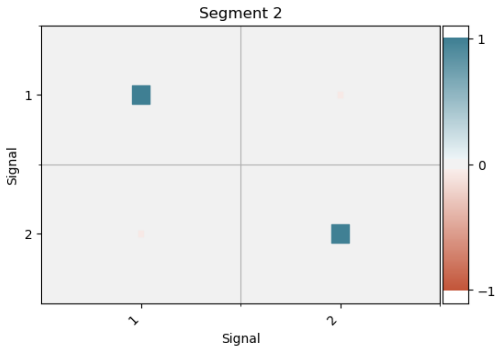
Figure 4.11: BS changepoint detection splits the data into 6 segments, with labels above (a) the correlation of the first two signals in the prior segment. (b,c) are correlation matrices between all 10 signals in segments 5 and 6 respectively. (d,e) are the mean and variance for each individual signal in every segment.



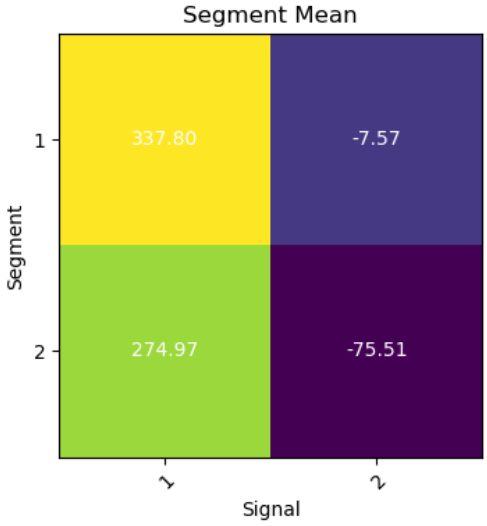
(a) Rainfall BS changepoint detection



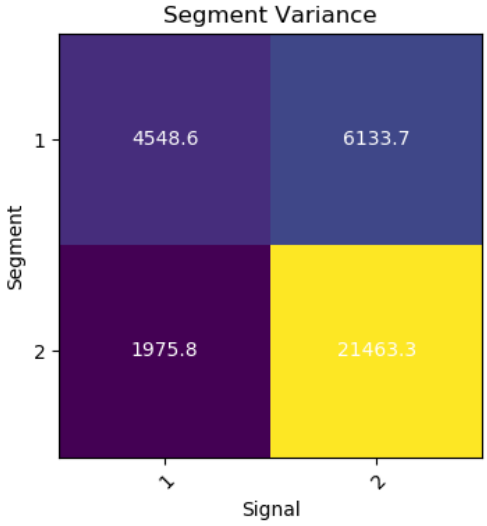
(b) Correlation matrix on segment 1



(c) Correlation matrix on segment 2



(d) Means across segments



(e) Variances across segments

Figure 4.12: BS detection of a single changepoint in rainfall data, with red labels above (a) indicating the correlation in the previous segment. (b,c) visually represent the correlation matrices for both segments. In (d,e) we have the mean and variance for each signal in both segments.

Chapter 5

Conclusion

Two parametric Bayesian approaches to changepoint detection have been thoroughly investigated. We find that for several synthetic datasets, these methods fare far worse than a simple non-parametric binary segmentation algorithm in both computational complexity and accuracy of detection. Out of all methods, the HMM produced the least convincing results. However, we suspect this is for three reasons: 1) the size of the segments are too small to accurately infer the model parameters without an informative prior, 2) there were not enough transitions between states, and hence the transition matrix could not be well defined, and 3) random priors caused the system to become stuck in local minima. The use of a maximum *a posteriori* prior and additional training sets is thought to alleviate these issues. The Bayesian PPM produced decent results for all but the heavy tailed distributed data, even with a minimal information prior, though the computational complexity of the algorithm is a very limiting aspect for its continued application, especially when BS already outperforms it in accuracy.

We have shown that these methods are all capable to some extent of detecting changes purely in the correlation structure between signals in synthetic data, going beyond univariate methods that only detect shifts in mean and variance. In empirical data, we have considered the mean, variance, and covariance between segments, finding that indeed our methods are able to detect changes in the covariance structure. However, these correlation changes easily overshadowed by shifts in the first two moments of the signals; future research may investigate the possibility of methods that can determine changes in the correlation structure whilst excluding those of individual mean and variance changes.

Bibliography

[1] Donald C Van Dyke and Rex L Huff. Life span of white blood cells as measured in irradiated parabiotic rats. *American Journal of Physiology-Legacy Content*, 165(2):341–347, 1951.

[2] Ewan S Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

[3] Charles Truong, Laurent Oudre, and Nicolas Vayatis. A review of change point detection methods. *arXiv preprint arXiv:1801.00718*, 2018.

[4] Wilder Penfield and Herbert Jasper. Epilepsy and the functional anatomy of the human brain. 1954.

[5] Andrew Ang and Geert Bekaert. International asset allocation with regime shifts. *The Review of Financial Studies*, 15(4):1137–1187, 2002.

[6] Ashish Bhan, David J Galas, and T Gregory Dewey. A duplication growth model of gene expression networks. *Bioinformatics*, 18(11):1486–1493, 2002.

[7] ES Page. A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3/4):523–527, 1955.

[8] Mark Best and Duncan Neuhauser. Walter a shewhart, 1924, and the hawthorne factory. *BMJ Quality & Safety*, 15(2):142–143, 2006.

[9] E Brodsky and Boris S Darkhovsky. *Nonparametric methods in change point problems*, volume 243. Springer Science & Business Media, 2013.

[10] Gary Lorden et al. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42(6):1897–1908, 1971.

[11] Albert N Shiryaev. On optimum methods in quickest detection problems. *Theory of Probability & Its Applications*, 8(1):22–46, 1963.

[12] Venkata Jandhyala, Stergios Fotopoulos, Ian MacNeill, and Pengyu Liu. Inference for single and multiple change-points in time series. *Journal of Time Series Analysis*, 34(4):423–446, 2013.

[13] AFM Smith. A bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika*, 62(2):407–416, 1975.

[14] Til Aach and André Kaup. Bayesian algorithms for adaptive change detection in image sequences using markov random fields. *Signal Processing: Image Communication*, 7(2):147–160, 1995.

[15] Ryan Prescott Adams and David JC MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

[16] CA McGilchrist and KD Woodyer. Note on a distribution-free cusum technique. *Technometrics*, 17(3):321–325, 1975.

[17] Ashish Sen and Muni S Srivastava. On tests for detecting change in mean. *The Annals of statistics*, pages 98–108, 1975.

[18] AN Pettitt. A non-parametric approach to the change-point problem. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(2):126–135, 1979.

[19] L Allison Jones-Farmer, Victoria Jordan, and Charles W Champ. Distribution-free phase i control charts for subgroup location. *Journal of Quality Technology*, 41(3):304–316, 2009.

[20] Changsoon Park, Changsoon Park, Marion R Reynolds Jr, and Marion R Reynolds Jr. Nonparametric procedures for monitoring a location parameter based on linear placement statistics. *Sequential Analysis*, 6(4):303–323, 1987.

[21] David McDonald. A cusum procedure based on sequential ranks. *Naval Research Logistics (NRL)*, 37(5):627–646, 1990.

[22] Elisabeth Vieth. Fitting piecewise linear regression functions to biological responses. *Journal of applied physiology*, 67(1):390–396, 1989.

[23] Brian Beckage, Lawrence Joseph, Patrick Belisle, David B Wolfson, and William J Platt. Bayesian change-point analyses in ecology. *New Phytologist*, 174(2):456–467, 2007.

[24] Makram Talih and Nicolas Hengartner. Structural learning with time-varying components: tracking the cross-section of financial time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):321–341, 2005.

[25] Zaid Harchaoui and Olivier Cappé. Retrospective mutiple change-point estimation with kernels. In *2007 IEEE/SP 14th Workshop on Statistical Signal Processing*, pages 768–772. IEEE, 2007.

[26] Xuan Liu, Pengzhu Zhang, and Dajun Zeng. Sequence matching for suspicious activity detection in anti-money laundering. In *International Conference on Intelligence and Security Informatics*, pages 50–61. Springer, 2008.

[27] Yanjun Xu, Qingzhao Yu, Richard Scribner, Katherine Theall, Scott Scribner, and Neal Simonsen. Multilevel spatiotemporal change-point models for evaluating the effect of an alcohol outlet control policy on changes in neighborhood assaultive violence rates. *Spatial and spatio-temporal epidemiology*, 3(2):121–128, 2012.

[28] Kenneth R. French. Data library, 2019. [Online; accessed 10-August-2019].

[29] Indian Institute of Tropical Meteorology. Core-monsoon india rainfall, 1871-2006. [Online; accessed 12-August-2019].

[30] Todd Mitchell. Northeast brazil rainfall anomaly index, 2002. [Online; accessed 12-August-2019].

[31] Alessandra Giannini, John CH Chiang, Mark A Cane, Yochanan Kushnir, and Richard Seager. The enso teleconnection to the tropical atlantic ocean: Contributions of the remote and local ssts to rainfall variability in the tropical americas. *Journal of Climate*, 14(24):4530–4544, 2001.

[32] Paul A Gagniuc. *Markov chains: from theory to implementation and experimentation*. John Wiley & Sons, 2017.

[33] Leonard E Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, 41(1):164–171, 1970.

[34] Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563, 1966.

[35] Jeff A Bilmes et al. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, 4(510):126, 1998.

[36] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[37] AJ Viterbi. Error bounds for convolutionalcodes and an asymmetrically optimum decoding algorithm.-ieee transactions on information theory. 1967.

[38] Daniel Barry and John A Hartigan. Product partition models for change point problems. *The Annals of Statistics*, pages 260–279, 1992.

[39] Daniel Barry and John A Hartigan. A bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88(421):309–319, 1993.

[40] Matthew J Beal, Zoubin Ghahramani, and Carl E Rasmussen. The infinite hidden markov model. In *Advances in neural information processing systems*, pages 577–584, 2002.

[41] Paul Fearnhead. Exact bayesian curve fitting and signal segmentation. *IEEE Transactions on Signal Processing*, 53(6):2160–2166, 2005.

[42] Paul Fearnhead. Exact and efficient bayesian inference for multiple changepoint problems. *Statistics and computing*, 16(2):203–213, 2006.

[43] Xiang Xuan and Kevin Murphy. Modeling changing dependency structure in multivariate time series. In *Proceedings of the 24th international conference on Machine learning*, pages 1055–1062. ACM, 2007.

[44] Edward G Carlstein, Hans-Georg Müller, and David Siegmund. Change-point problems. IMS, 1994.

[45] Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer, 1992.

[46] Rebecca Killick, Paul Fearnhead, and Idris A Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

[47] Chandralika Chakraborty and PH Talukdar. Issues and limitations of hmm in speech processing: a survey. *International Journal of Computer Applications*, 141(7):975–8887, 2016.