METHODS IN STATISTICAL CHANGE-POINT ANALYSIS

Robert M. Steward, B.S., M.A.

An Abstract Presented to the Faculty of the Graduate
School of Saint Louis University in Partial
Fulfillment of the Requirements for the
Degree of Doctor of Philosophy

2014

Abstract

We investigate the statistical change-point problem in a variety of settings. In particular, we develop a number of statistical models that may be applied to the problem of detecting the existence of a statistical change-point and then estimating its location in one and several dimensions. Of primary interest in the univariate case are statistical change-points in mean, variance, and multiple change-points. For the multivariate case we also investigate the statistical change-point inference and location problem, but also consider the question in which dimensions the change-point occurrred. Throughout the text we review various classical methods in change-point analysis that include maximum likelihood (MLE), Bayesian, and control charting methods. These existing methods are then compared against new methods we propose involving a Bayesian statistical analysis of the wavelet detail coefficients after a Discrete Wavelet Transform of our original time series. We find our new approach offers several advantages from the above classical methods. Furthermore, we present an application of our Bayesian-wavelet method of change-point estimation in conjunction with random matrix dimension reduction techniques for time series of very high dimension. Finally, we apply the method of Reversible Jump Markov Chain Monte Carlo to the multivariate change-point problem. Along with detailed derivations of theoretical results, we provide numerous numerical experiments and various real world applications for the methods we present.

METHODS IN STATISTICAL CHANGE-POINT ANALYSIS

Robert M. Steward, B.S., M.A.

A Dissertation Presented to the Faculty of the Graduate
School of Saint Louis University in Partial
Fulfillment of the Requirements for the
Degree of Doctor of Philosophy

2014

COMMITTEE IN CHARGE OF CANDIDACY:

Profesor Darrin Speegle,
Chairperson and Advisor

Professor Steven Rigdon

Associate Professor Michael Lamar

Dedication

DEDICATION TEXT

Acknowledgments

ACKNOWLEDGMENTS TEXT

Table of Contents

List of Figures

List of Tables

Chapter 1

Preliminaries

1.1 Introduction

In this dissertation we develop a number of statistical models and investigate various approaches to a variety of statistical change-point problems. A consistent theme throughout this work is the blending of statistical tools with concepts based in wavelet theory. Classical statistical approaches to the change-point problem such as Maximum Likelihood Estimation (MLE) and Bayesian methods offer established frameworks to model a variety of change-point problems. Meanwhile, the concepts of wavelet theory developed in the past 25 years provide an alternative perspective to view the data. This alternative perspective offers the possibility to pursue improved statistical change-point problem methods.

MLE and Bayesian methods are two of the most important parametric techniques that may be applied to the change-point problem [1]. As we will see in later examples and numerical experiments, these two methods may provide a means to detect and correctly estimate the location for even very subtle change-points in a time series. By their very construction, however, they are limited to cases in which certain assumptions concerning the statistical

distributional properties of the data remain valid. In many real world situations, however, the distribution assumptions of the data at hand may not be clear *a priori*. In such instances the assumptions upon which the MLE and Bayesian methods are built constrain the applicability of these approaches to more general contexts.

Wavelets and in particular the discrete wavelet transform (DWT) of a time series allow us to view the data from a wholly different context [2]. Many natural phenomena have a "sparse" representation in the wavelet domain. In particular, wavelets often represent smooth functions in an economical manner [3]. Furthermore, wavelets provide function change information simultaneously from both localized and collective perspectives [4]. Understanding these properties gives us a prior knowledge about our transformed data set and insights into the structure of the time series that we did not have before. Despite these (and many other) known wavelet properties that apply to a broad class of functions, the question remains how best to utilize wavelets in the change-point problem.

The prior knowledge we obtain through a DWT suggests a natural connection with the tools offered in Bayesian statistical analysis. Our goal is to thus capitalize on the powerful Bayesian methods in the change-point problem, but now in the wavelet domain. Doing so allows us to analyze time series for statistical change-points while making fewer assumptions governing the distribution form underlying the mean function. We defer a formal statement of the change-point problem to the next chapter and present here preliminary concepts that will be used throughout this dissertation. While extensive volumes exist on both wavelet theory and Bayesian statistics throughout the literature, our goal is to merely introduce the essential elements of these topics that we will later apply to the methods developed in the chapters to follow.

1.2 Wavelets

Throughout history people have endeavored to simplify complicated systems by breaking them apart into simpler constituent components [5]. Wavelets are one answer to this "representation problem" where we re-express a function in terms of simpler basis functions. The motivation for wavelets begins with a brief introduction to Fourier series where we borrow notation from Katznelson [6]. At the beginning of the 19th century Joseph Fourier recognized that a large class of functions could be expressed as a complex trigonometric series of the form

$$S \sim \sum_{-\infty}^{\infty} \hat{f}(n) e^{int}$$

where $n \in \mathbb{Z}$, $t \in [0, 2\pi)$, and $\hat{f}(n)$ is the Fourier coefficient defined below. For $f \in L^2[0, 2\pi)$, we have by definition

$$\int_0^{2\pi} |f|^2 dt < \infty.$$

For such a class of functions, we may express f in terms of its trigonemtric series as

$$S[f] \sim \sum_{-\infty}^{\infty} \hat{f}(n) e^{int}$$

where the symbol $\sim$ is used to denote equality *almost everywhere* with respect to the Lebesgue measure. The $\hat{f}(n)$'s denote the *Fourier coefficients* and are defined as

$$\hat{f}(n) = \frac{1}{2\pi} \int_0^{2\pi} f(t) e^{-int} dt. \tag{1.1}$$

The usefulness of this series representation of f is that we may build the original function from much simpler trigonometric functions. Furthermore, equation (1.1) gives us insight into the fundamental frequency components underpinning our function of interest. For example, if we are analyzing a signal composed of finitely many harmonic frequencies (frequencies of the form $1/n$, $2/n$, ...), $\hat{f}(n)$ will be nonzero at precisely those harmonic frequencies [7]. An important question then becomes how closely we can approximate f with a finite trigonometric series for nonharmonic signals. For many functions with good decay properties, often the number of terms is quite small. The investigation of Fourier coefficient properties has been the subject of extensive study in the field of harmonic analysis.

We may next extend the investigation of Fourier series of periodic functions to the study of those functions defined on the real line, $\mathbb{R}$. Many of the Fourier series results found on the circle have direct analogous results on the real line [6]. In this setting we define the Fourier transform of f as

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x) e^{-2\pi i \xi x} dx \quad \forall\ \xi \in \mathbb{R}.$$

For integrable functions, we may recover the original function from its inversion formula given by

$$f(x) = (\check{f})\hat{}\,(x) = \int_{-\infty}^{\infty} \hat{f}(\xi) e^{2\pi i \xi x} d\xi \quad \forall\ x \in \mathbb{R}. \tag{1.2}$$

Here we are really considering $f(x)$ in equation (1.2) as some representative from an equivalence class of integrable functions which are pointwise equal almost everywhere. Equation (1.2) tells us that any such integrable function may be represented as a continuous sum (integral) of terms of

the form $\hat{f}(\xi)e^{2\pi i\xi x}$ ranging over all possible frequencies $\xi \in \mathbb{R}$ [8]. In other words, the Fourier transform allows us to extract the underlying frequencies along with their respective weights of any integrable function.

From the standpoint of the statistical change-point problem, we are investigating the mean function properties of a time series (signal) contaminated with some stochastic random component (noise) in a very localized way. In particular, we are trying to detect an abrupt change to the true underlying mean function of the time series at a particular point in time. In this context, a limitation of the Fourier analysis approach is that the trigonometric basis functions have infinite support. Because of the support property of the sines and cosines, the Fourier approach does not allow us to easily distinguish how the frequency of a signal is changing in a localized way from its overall change [4]. So while the Fourier approach allows us to represent a complicated signal from a more simplified approach in terms basis functions, the particular characteristics of this representation system are not optimal for our problem at hand.

Why wavelets?

Wavelets are a popular analytical tool whose applicability continues to spread to a variety of scientific and engineering fields. The attractiveness of wavelets springs from both their simplicity in theory and flexibility in application [9]. From a statistical point of view, wavelets offer alternative methods in data smoothing, density estimation, and multiscale time series analysis [10]. The multiscale characteristic of wavelets is particularly important in our setting because of the flexibility it offers to various change-point problems we might encounter. Heuristically, this flexibility allows a researcher to focus in on fine features of the data, step back to look at generalized characteristics of the data, or combine both perspectives.

In our discussion on Fourier series, we saw how $L^2(\mathbb{R})$ functions may be represented as a sum of trigonometric functions. In wavelet theory, we now see how wavelets play analogous roles to the trigonometric basis functions previously used. Wavelets, however, offer a certain flexibility that the Fourier approach does not. For example, numerous wavelet families exist each with their own particular properties. Depending on the particular problem at hand, we may choose a wavelet with the best suited properties on a case by case basis.

While many variations of wavelet transforms exists, generally the transform involves dividing the original time series into "scaling coefficient" and "detail coefficient" components. The scaling coefficients represent varying degrees of time series smoothing or averaging while the details may be thought of as representing how much the function is changing at a certain resolution level. Figure 1.1 illustrates this concept by displaying a time series along with two example plots of corresponding scaling and detail coefficients. In some sense, wavelets offer a method to analyze the change of a time series through a lens that may be readily "zoomed in" or "zoomed out" as required [4].

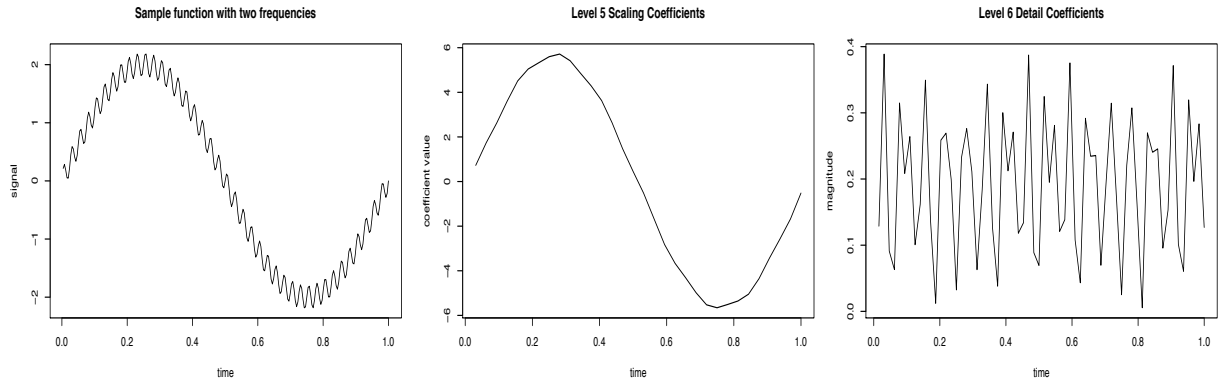

Figure 1.1: Original function (left), scaling coefficients at level 5 smoothing the time series (middle), and detail coefficients capturing how the time series is changing at detail level 6 (right).

As discussed above, Fourier representations of functions are a powerful diagnostic tool to break apart time series (signals) into constituent frequency

components. The drawback to the Fourier approach with our particular problem, however, is that the Fourier representation of a function may have poor or nonexistent converging properties at certain points of the function. Wavelets on the other hand, by dilations may be localized in scale (frequency) and by translations may also be localized in time [4]. In particular, at points of discontinuity we expect Fourier techniques to exhibit poor convergence properties while wavelets should easily be able to model the function at such points. It is precisely these points of discontinuity in the change-point problem that we are interested in locating. Hence, wavelets offer a promising alternative to Fourier analysis and serve as our motivation for applying them in our present context.

Wavelet fundamentals

While many excellent references on wavelets exist with varying degree of theoretical depth, we follow the presentation given by Ogden [11] and Nason [10] as we outline important elements of wavelet theory pertinent to our results. All wavelets we consider in this dissertation constitute an orthonormal basis of $L^2(\mathbb{R})$. In particular, we may approximate any $f \in L^2(\mathbb{R})$ arbitrarily close in the $L^2(\mathbb{R})$ sense as

$$f(t) = \sum_{k=-\infty}^{\infty} w_{0,k}\psi_{j,k} + \sum_{j=j_0}^{\infty}\sum_{k=-\infty}^{\infty} d_{j,k}\psi_{j,k} \tag{1.3}$$

where $w_{j,k}$ and $d_{j,k}$ are called the scaling and detail coefficients, respectively. Explicitly, $w_{j,k}$ and detail $d_{j,k}$ are given by

$$w_{j,k} =< f, \phi_{j,k} >= \int_{-\infty}^{\infty} f(t)\ \phi_{j,k}(t)dt$$

$$d_{j,k} =< f, \psi_{j,k} >= \int_{-\infty}^{\infty} f(t)\ \psi_{j,k}(t)dt.$$

Wavelets by definition are systems of dialations and translations and so each $w_{j,k}$ and $d_{j,k}$ are related to each by the so called "father" and "mother" wavelet

expressed as

$$\{\phi_{j,k}(t) = 2^{-j/2}\phi(2^{-j}t - k)\}_{j,k}$$

$$\{\psi_{j,k}(t) = 2^{-j/2}\psi(2^{-j}t - k)\}_{j,k}. \tag{1.4}$$

The simplest wavelet that we may explicitly express in closed form is the Haar wavelet given by

$$\phi(t) = \begin{cases} 1, & 0 \leq t < 1 \\ 0, & \text{otherwise} \end{cases} \qquad \psi(t) = \begin{cases} 1, & 0 \leq t < \frac{1}{2} \\ -1, & \frac{1}{2} \leq t < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Besides being orthonormal bases for $L^2(\mathbb{R})$, these wavelet systems also possess other special properties. By fixing a particular j in equation (1.4), we denote the spans of the scaling and wavelet functions as V_j and O_j, respectively, as k ranges through the integers. Each V_j is an approximation space for the next finer approximation space of spanning scaling functions, V_{j+1}, with the difference in information being precisely O_j. In particular V_j is orthogonal to O_j with the direct sum of these orthogonal subspaces equal to V_{j+1}, that is $V_{j+1} = V_j \oplus O_j$. Such a construction leads to the so called multiresolution analysis and allows us to approximate (smooth) a signal at various approximation levels while precisely keeping track of detail levels. The detail levels capture function change at these different resolutions and will play a key role in our analysis of the change-point problem.

In practice when working with actual data it is often more convenient to use the discrete wavelet transform (DWT). We start with a discrete time series $\mathbf{x} = \{x_i\}_{i=1}^{2^J}$ for some natural number J. Next we let j be an index across the DWT scales ranging from $J-1$ down to 0. When $j = J-1$ we produce 2^j scaling and detail coefficients. For the Haar DWT, the finest level of scaling (w_{jk})

and detail (d_{jk}) coefficients are computed by

$$w_{(J-1)k} = (x_{2k} + x_{2k-1})/\sqrt{2} \quad \text{and} \quad d_{(J-1)k} = (x_{2k} - x_{2k-1})/\sqrt{2}. \tag{1.5}$$

We then compute all subsequent levels of scaling and detail coefficients by the formulas

$$w_{(j-1)k} = (w_{j,2k} + w_{j,2k-1})/\sqrt{2} \quad \text{and} \quad d_{(j-1)k} = (w_{j,2k} - w_{j,2k-1})/\sqrt{2}. \tag{1.6}$$

This process terminates when $j = 0$ and we produce just a single ($2^0 = 1$) additional scaling and detail coefficient. After we take the DWT, we have in total $N - 1$ scaling and detail coefficients. Of course, if we chose a different wavelet, the above formulas would also be different. Irrespective of the wavelet we choose, however, certain properties will always hold. For example, for orthogonal wavelets such as the Haar wavelet, there will always exist an orthogonal matrix, say $\mathbf{W}$, such that if we perform a matrix multiplication $\mathbf{Wx}$, we obtain a vector representation of our detail coefficients $\mathbf{d}$ (plus the single lowest scaling coefficient). Since orthogonal matrices are invertible, we easily recover our original time series from this detail coefficient vector by the multiplication

$$\mathbf{x} = \mathbf{W}^{-1}\mathbf{d}.$$

While the matrix representation provides a mathematically convenient way to express the DWT, computationally we can obtain better numerical efficiency [10]. Mallat [12] developed his fast "pyramid algorithm" to compute the DWT which is generally what we find implemented in wavelet software.

1.3 Techniques in Bayesian Statistics

The second main theme of this dissertation involves the use of Bayesian statistics. While the formulation of Bayes Theorem actually predates frequentist methods based on the likelihood approach by more than a century and half, the popularity of the Bayesian approach has come much more recently [13]. Aside from past philosophical disagreements between frequentist and Bayesian statistical approaches concerning the subjectivity of Bayeisan methods was a more practical impediment. The Bayesian approach often involves computing high dimensional integrals that are analytically intractable [13]. While a method known as Monte Carlo Markov Chain (MCMC) was a known numerical strategy to abstract results from high dimensional integrals, the method required extensive computing power. It is therefore no coincidence that the popularity of Bayesian statistics coincided with the wide spread availability of computing power [14]. Ever since that time Bayesian methods have continued to gain considerable attention and now offer viable alternatives to the frequentist approach.

The Bayesian approach begins with Bayes Theorem. In its most familiar representation as introduced in a first course in probability, Bayes theorem takes the discrete form

$$p(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{\sum_{A_i} P(B|A_i)P(A_i))}$$

where A and B denote events from some probability space and the sum in the denominator of the right most equality represents some partition of this space. This form naturally extends to the case of continuous random variables. In the Bayesian context we apply Bayes theorem to probability distribution functions

and obtain a posterior distribution function of our parameter set given some observed data [14]. In this sense the Bayesian approach assess uncertainty using probability. Letting $\boldsymbol{\theta}$ represent our parameter vector and X some observed data set, we write the general form of our posterior distribution as

$$p(\boldsymbol{\theta}|X) = \frac{L(X|\boldsymbol{\theta})p_0(\boldsymbol{\theta})}{f(X)}.$$

Here $L(X|\boldsymbol{\theta})$ is the likelihood function from the traditional frequentist approach, p_0 is our prior distribution function, and $f(X)$ is the probability of observing the data set. As in the frequentist approach, the likelihood function plays a central role. Unlike the frequentist approach, the modeler has the ability to inject prior information by a particular choice of p_0. Finally, notice $f(X)$ has no dependence on the parameter vector. In many applications $f(X)$ is a high dimensional integral that is analytically intractable. Fortunately, this term is essentially just a normalizing constant that often can be safely ignored throughout preliminary calculations. Once a proportional form of the posterior distribution is found, the denominator term can then be reintroduced to normalize the distribution. We therefore usually consider the posterior distribution in its proportionality form [13]

$$p(\boldsymbol{\theta}|X) \propto L(X|\boldsymbol{\theta})p_0(\boldsymbol{\theta}).$$

The prior distribution component of the posterior distribution function gives the Bayesian approach a flexibility to data analysis that other methods do not possess. Throughout the subsequent chapters we employ both conjugate priors and noninformative priors. The application of a conjugate prior typically involves a compromise between constraints of the data set along with

mathematical convenience. For example, suppose we wanted to model the arrival rate for the number of customers entering a store over some time period. Firstly, assume we observe x customers over this time period. A reasonable likelihood approach in this case would be the Poisson likelihood function with associated rate parameter λ

$$L(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!},\ x = 0, 1, 2, \ldots$$

Our compromise between data constraints and mathematical convenience suggests a gamma distribution prior of the form

$$p_0(\lambda) = \frac{\lambda^{a-1}e^{-\lambda/b}}{\Gamma(a)b^a},\ \lambda > 0$$

where Γ is the gamma function and a and b are tuning parameters chosen by the modeler. On the one hand, the gamma distribution shares the same support as the likelihood function. Furthermore by the choice of tuning parameters a and b we obtain tremendous flexibility to reflect our prior belief of the value of λ. On the other hand, dropping multiplicative constants we may write our proportional posterior distribution function as a product of our likelihood and prior distributions

$$p(\lambda|x) \propto \lambda^{x+a-1}e^{-\lambda(b+1)/b}.$$

Our proportional distribution function is conveniently once again in the form of a gamma distribution now with parameters $x + a$ and $\frac{b}{1+b}$. Since $p(\lambda|x)$ is a probability distribution function and must integrate to 1, we immediately have our proportionality constant based on the form of a gamma distribution with these parameters. So our choice of a suitable conjugate prior in this case results in a known distributional form of our final posterior distribution function.

Of course a potential difficulty with the conjugate prior approach and for

that matter the use of any informative prior is the requirement to choose suitable tuning parameters. If prior parameter information is unavailable then there may be no reasonable way to select these tuning parameters. In such cases the best option may be to use so called noninformative priors such as a uniform prior or Jeffreys prior [15]. By the choice of such noninformative priors, the modeler injects the maximum amount of variance into the prior and does not prejudice the results *a priori*. Throughout, our subsequent analysis we employ both conjugate and noniformative priors as we construct our changepoint models.

1.4 Markov Chain Monte Carlo

Much of the power of the Bayesian approach relies on the ability to perform a high volume of intensive calculations [13]. In chapter 6 we present a new approach to the multidimensional change-point problem that requires high computational power using the Markov Chain Monte Carlo (MCMC) algorithm. The MCMC algorithm is widely considered to be one of the top ten most important algorithms of the 20th century [16]. In cases where analytical results are difficult or impossible to achieve, MCMC offers a strategy through random draws of a specified probability distribution to produce a Markov chain that converges to a distribution of interest [17].

The approach first requires us to choose initial parameter values. Next, we must construct a Markov chain with parameters in the state space of interest. Randomly drawing numbers from the state space via some predetermined proposal distribution, we update our parameter vector based only on the previous step in the chain (hence Markov) [14]. Under rather general conditions, ergodic theory says our chain will eventually converge to the true posterior distribution. The main conditions we require are the so called stationarity and reversibility

properties. The stationarity property simply says $p(X_{n+1}|X_n)$ does not depend on n. Related to this concept, the reversibility property ensures our chain maintains the same probabilistic properties running both forward and backwards in time. That is the probability distributions of $(X_k, X_{k+1}, \ldots X_{k+m})$ and $(X_{k+m}, \ldots X_{k+1}, X_k)$ are identical. One notes reversibility implies stationarity, but not vice versa [18].

There are many approaches to actually generating the chain itself. The method we utilize in later chapters is known as the Metroplis-Hastings (MH) algorithm. An important characteristic of this particular algorithm is that it only requires a function proportional to the posterior distribution function of interest. Let $L(\mathbf{D}|\boldsymbol{\theta})p_0(\boldsymbol{\theta})$ represent our proportional posterior distribution function where $\mathbf{D}$ is our observed data and $\boldsymbol{\theta}$ is our parameter vector of interest. Next, we must choose a proposal distribution $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$. The proposal distribution represents the "Monte Carlo" portion of the algorithm and randomly explores the probability space of the posterior distribution [13]. The proposal distribution should have support that matches the unknown posterior distribution and ideally samples the space in both an efficient and economical manner.

The Metroplis-Hastings Algorithm is then applied as follows.

1. Randomly generate a proposed parameter vector $\boldsymbol{\theta}'$ from the proposal distribution: $q(\cdot|\boldsymbol{\theta})$.

2. The probability of moving to the proposed parameter state $\boldsymbol{\theta}'$ is found by computing $\alpha = min\Big(\frac{L(\mathbf{D}|\boldsymbol{\theta}')p_0(\boldsymbol{\theta}')}{L(\mathbf{D}|\boldsymbol{\theta})p_0(\boldsymbol{\theta})}, 1\Big)$

3. Now generate another random variable $u \sim U(0, 1)$

4. Accept the move to the new parameter vector $\boldsymbol{\theta}$ if $\alpha > u$ otherwise stay at $\boldsymbol{\theta}$.

5. Repeat steps 1-4 until the chain converges. Once the chain converges repeat

steps 1-4 an additional number of times as required to draw inference from the posterior distribution function.

Ergodic theory states that so long as our mild conditions are met, the chain will eventually visit parameter states proportional to their relative values in the posterior distribution function [17]. In the above MH construction we see why only the proportional posterior distribution is relevant since the normalizing constant in the denominator of the posterior distribution function cancels in the ratio above. The question remains how long the chain must be run to achieve convergence. This question may be partially answered with suitable diagnostics, but is still susceptible to pseudoconvergence when the chain remains in only certain portions of the state space for particularly long periods of time [18]. Running the chain for as long as possible remains the most sure way to arrive at the best approximation for the true posterior distribution function.

1.5 Conclusion

The topics presented in this chapter will serve as the fundamental building blocks that we apply to the change-point problem developed in the subsequent chapters. While the material here was given at only a cursory level of detail, our intent was to merely familiarize the reader to those most important concepts essential to the theory later presented as we move forward. For a more in depth review of wavelets at an introductory level we recommend [11, 19] and for a more thorough exposition see for example [2, 9, 12, 20]. An excellent review of Bayesian statistics and MCMC theory can be found at [14, 17] or at a more advanced level [13, 21, 22] provides a good overview of theoretical details.

Chapter 2

The Univariate Case

2.1 Introduction

Chapter 2 considers the univariate change-point problem from the standpoint of a time series undergoing a shift in the mean function at some unknown time. In particular, we assume the time series is governed by a smooth underlying mean function except at a single point along with a Gaussian additive noise component. At some unknown time point, an abrupt change occurs to the mean function. The change-point problem is then how do we see through the additive noise component to correctly detect and identify the location of the change-point within the time series? Many applications of the change-point problem exist which we highlight in this and later chapters. Throughout this chapter in particular we emphasize applications to Statistical Process Control (SPC).

Common in the SPC literature are the so called "prospective" and "retrospective" change-point problems. The prospective change-point problem is concerned with detecting statistical changes in a near real time setting while the retrospective problem is a look back at what has already occurred [23]. While the change-point estimation methods in this chapter are natural for the retrospective

case, we present a strategy where our change-point estimation methods can also be used for the prospective change-point problem while working in conjunction with SPC control charts.

In this chapter we focus on a single permanent abrupt change to the mean function; however, the statistical change could also be gradual or temporary [7]. Furthermore, we need not limit ourselves to a single dimension or single change-point; although as we increase the complexity of the problem, the probability of detecting small changes also diminishes. These topics will be covered in later chapters. For our present purposes we consider only changes in the process mean which tends to be the most common problem in applications such as SPC and biosurveillance [24].

In this chapter we also present a Bayesian-wavelet method of change-point location estimation and compare it against two established change-point detection methods. We provide brief summaries of both the maximum likelihood estimation (MLE) method and standard Bayesian method in section 2 below. We then develop the Bayesian-wavelet method, originally proposed by Ogden [25]. In the chapters to follow we further develop both the Bayesian-wavelet method in a number of other contexts and extend this methods to the multivariate setting. Finally, we conduct a number of simulations under a variety of assumptions that attempt to replicate in part scenarios that are common in both SPC and biosurveillance applications.

2.2 Description of methods

2.2.1 Maximum Likelihood Estimation (MLE)

The MLE method is based on a hypothesis testing approach that seeks to maximize a likelihood function. Here, we borrow notation from Hawkins [26] as we develop this method. Consider a sequence of independently distributed random variables, $\{x_i\}_i^N$ for $N \in \mathbb{N}$ where:

$$x_i = \mu_1 + \varepsilon_i \qquad (1 \leq i \leq \tau)$$

$$x_i = \mu_2 + \varepsilon_i \qquad (\tau < i \leq N)$$

such that $\varepsilon_i \sim N(0, \sigma^2)$.

In this present context, the variance is assumed to be known, but we wish to determine the change-point, $\tau = 1, 2, ...N - 1$. To do so, we test the following null hypothesis

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_N$$

against the alternative hypothesis

$$H_a : \mu_1 = \mu_2 = \cdots = \mu_\tau \neq \mu_{\tau+1} = \cdots = \mu_N.$$

We use the following statistic

$$S_\tau = \sum_{i=1}^{\tau} (x_i - \bar{x}_\tau)^2 + \sum_{i=\tau+1}^{N} (x_i - \bar{x}'_\tau)^2,$$

where

$$\bar{x}_\tau = \sum_{i=1}^{\tau} x_i/\tau \quad \text{and} \quad \bar{x}'_\tau = \sum_{i=\tau+1}^{N} x_i/(N - \tau).$$

Letting $S = S_N$ and $\bar{x} = \bar{x}_N$, Hawkins shows that through an analysis of variance we may write

$$S = S_\tau + E_\tau,$$

where

$$E_\tau = \tau(\bar{x} - \bar{x}_\tau)^2 + (N - \tau)(\bar{x} - \bar{x}'_\tau)^2.$$

But this means minimizing S_τ is equivalent to maximizing E_τ. After some minor algebra, our likelihood ratio test statistic becomes:

$$W = \max_{1 \leq \tau < N} \sqrt{E_\tau}.$$

The τ that maximizes W is our point estimation of the time series change-point location.

2.2.2 Bayesian change-point analysis

A potential improvement over the classical MLE method to the change-point problem came by applying Bayesian techniques. The Bayesian approach enjoys the potential advantage of being capable of incorporating the modeler's prior information to obtain a probability density function. That is, while the MLE approach estimates the location within a time series of a single change-point, the Bayesian method offers a complete probability distribution of the change-point location.

For the following discussion we summarize univariate results presented by Pan and Rigdon [27] and originally developed by Smith [28]. The basic problem setup remains the same as in the above MLE case, except now we will further generalize by assuming unknown, but constant, variance. Here with the Bayesian approach, we first choose a prior distribution of our parameter of interest (either noninformative or otherwise) and derive a posterior distribution. For the time being we will take the most general approach and assume we have no prior

information about the change-point location. Similar to the MLE case we assume our series satisfies the following distributional properties:

$$x_i \sim N(\mu_1, \sigma^2) \text{ for } 1 \leq i \leq \tau$$
$$x_i \sim N(\mu_2, \sigma^2) \text{ for } \tau < i \leq N.$$

From normal frequentist methods and setting $\sigma^2 = 1/\rho$, our likelihood function L becomes a product of likelihood functions L_1 and L_2 from before and after the unknown change-point:

$$L_1(x_1, .., x_\tau | \mu_1, \rho, \tau) = \prod_{i=1}^{\tau} \sqrt{\frac{\rho}{2\pi}} \exp\left[-\frac{\rho}{2}(x_i - \mu_1)^2 \right]$$

and

$$L_2(x_{\tau+1}, .., x_N | \mu_2, \rho, \tau) = \prod_{i=\tau+1}^{N} \sqrt{\frac{\rho}{2\pi}} \exp\left[-\frac{\rho}{2}(x_i - \mu_2)^2 \right].$$

Firstly, we assume μ_1 is known or estimated from perhaps some set of burn-in observations. Next, we choose a normal distribution as a conjugate prior for μ_2 and a discrete uniform distribution as a noninformative prior for our change-point τ mathematically represented in the following way:

$$p_0(\mu_2) = \sqrt{\tfrac{\beta\rho}{2\pi}} \exp\left[-\tfrac{\beta\rho}{2}(\mu_2 - \alpha)^2 \right]$$

and

$$p_0(\tau) = \tfrac{1}{N-1}.$$

Treating the marginal density distribution term in the denominator as a constant, Bayes theorem takes the form

$$p(\mu_2, \rho, \tau) \propto L(x_1, .., x_\tau | \mu_1, \rho, \tau) p_0(\mu_2) p_0(\tau).$$

Then by standard techniques we may integrate out μ_2 and ρ to arrive at the marginal posterior distribution which we can maximize to find the mode of the change-point for a given data set [27]. We now express our posterior distribution in the following way:

$$p(k|x_1, x_2, ..., x_N) \propto \frac{2}{\sqrt{(N-\tau)}} \exp\left[-\frac{\rho}{2}\left(\sum_{i=1}^{\tau}(x_i - \mu_1)^2 + \sum_{i=\tau+1}^{N}(x_i - \bar{x}')^2 \right) \right]$$

where

$$\bar{x}' = \frac{1}{N-\tau} \sum_{i=\tau+1}^{N} x_i.$$

While in this case, we assume no prior knowledge of the change-point, in reality this may not always be the case. In the simulation portion below we will see how using a beta-binomial prior may improve the performance of the Bayesian change-point method when applied to wavelet details coefficients.

2.2.3 Bayesian change-point analysis in the Wavelet Domain

Wavelets have grown increasingly popular in the past 25 years in a variety of scientific fields. More recently their applications to a number of statistical applications have grown increasingly important in such areas as density estimation, data smoothing, and multi-scale time series analysis [10]. In particular, the wavelet transform has the attractive property of decorrelating data and representing data in one or more dimensions sparsely [29]. Here will review some preliminary wavelet properties and explain their application to change-point analysis [10].

Wavelets in the change-point problem

Given a discrete noisy signal, we can take its DWT and analyze the resulting detail coefficients to distinguish the signal from the statistical noise. Recall a function is said to be smooth if it possesses derivatives of all orders. Donoho [30] showed the DWT of a noisy signal with an underlying smoothly varying mean function results in a sparse representation of the detail coefficients

provided the signal to noise ratio is sufficiently high. In particular, the contribution of the signal to the high level detail coefficient magnitudes should be close to zero leaving the energy of the true signal concentrated in a relatively sparse number of low level detail coefficients representing overall signal change. The noise component of the original signal, however, does not have a sparse representation; rather, it is again transformed to noise after the DWT and spread throughout all resolution levels. We will exploit this difference between signal and noise detail representation in our change-point detection and estimation method described below. Explicitly we model our original time series as

$$x_i = g(i) + \varepsilon_i$$

where $g(\cdot)$ is our true underlying smooth (except possibly at a single change-point) mean function observed for a discrete number of equally spaced time intervals and ε_i is some additive noise component. Next, we take the DWT of our time series and obtain a "transformed" data model of the following form

$$d^*_{jk} = d_{jk} + \eta_{jk}$$

where d^*_{jk} is the empirical detail coefficient we actually observe. In the case of the Haar wavelet d^*_{jk} would be the computation results after recursively applying equations (1.5) and (1.6). Next, d_{jk} is the true (but unknown) detail coefficient of the underlying smooth mean function we wish to estimate. Finally, η_{jk} is the transformed additive noise component from the original time series that transforms again to noise [10].

If we assume ε_i is generated from a Gaussian process, then η_i will also be Gaussian [31]. One strategy to distinguish between detail coefficients due to noise from detail coefficients due to the true signal is to employ thresholding techniques. In practice there are many ways to threshold such data. Donoho

proposed the universal threshold defined as $\sigma\sqrt{2\log n}$, where σ is usually estimated by the sample standard deviation of the finest level of detail coefficients [30]. In the case of the change-point problem, this method suggests a straightforward approach to estimating the change-point location. The large detail coefficients should be directly attributed to the location of where the time series is changing most rapidly.

In the case when ε_i is not Gaussian, we still expect the largest detail coefficients to be attributed to the locations of the time series where the first moment is most rapidly changing. The question of how we threshold the coefficients, however, must be addressed differently than above. Antoniadis [32] proposed thresholding methods for heavy tailed exponential power noise distributions. Fryzlewicz [33] established a thresholding procedure requiring only the first and second moments to be linked by a variance function.

Here we see our first indication of why the wavelet approach to the change-point problem offers something that parametric methods such as MLE do not. In the case of a noisy step function, the only large detail coefficients should be due to the shift, however, we are not restricted to such simple functions anymore. For some other smooth function with a shift, we also would expect the largest detail coefficients to be attributed to the shift. Thus we may relax our assumptions on the mean function to include much more general situations.

Building on the work of Donoho and Johnstone [30], Wang [34] first connected these properties of the DWT to the change-point problem. He recognized that under suitable conditions the largest detail coefficients are the result of those places where the time series is changing most rapidly and probably not attributable to noise. Wang then hypothesized that the places where the time series is most rapidly changing may be due to a statistical change-point. In Figure 2.1 (top) we provide an example similar to Wang's [34] of some noisy time

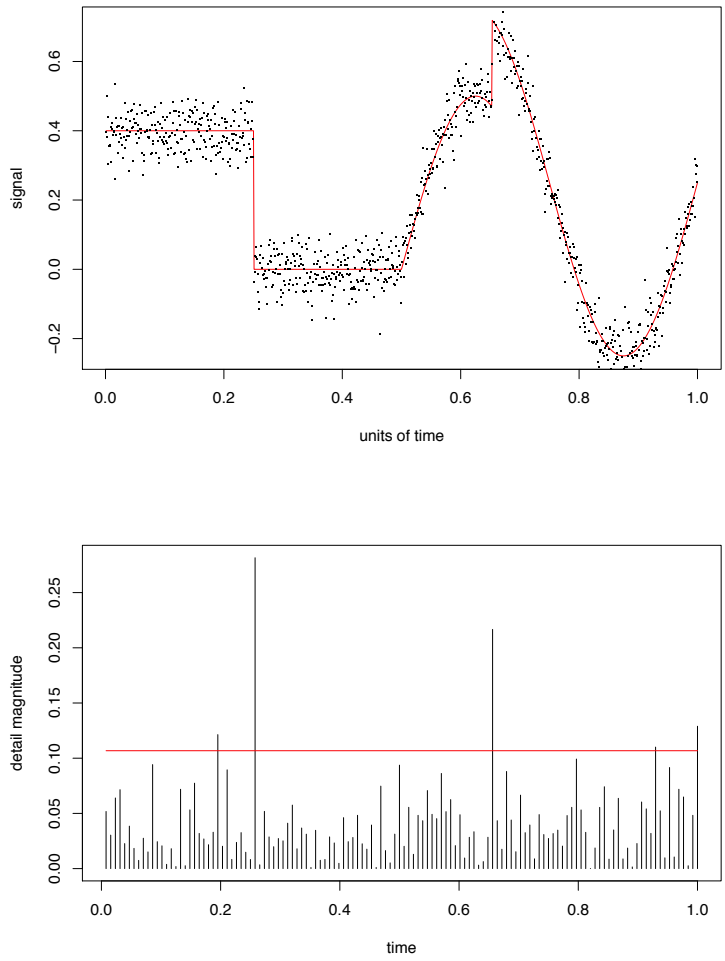

Figure 2.1: Stochastic process overlaid with true underlying function (top) and magnitude of corresponding detail coefficients at detail level 4 (bottom).

series which undergoes a shift at two separate time points. Analyzing, the detail coefficients we easily see in Figure 2.1 (bottom) how the largest detail coefficients closely correspond to these change-points. The red horizontal line in the Figure 2.1 (bottom) represents the universal threshold as defined above.

While Wang's method works well for relatively easy change-point problems such as the one depicted in Figure 2.1, it becomes much less reliable as the signal to noise ratio is decreased. Additionally, there is also the issue of determining which detail level (or levels) to use in the analysis. At this point, we would like to

combine the strengths of the MLE method with the strengths of the wavelet method. This is the subject of our next section.

Bayesian-wavelet approach to the change-point problem

Ogden [25] proposed the following method to estimate the location of a univariate time series where a change-point in mean occurs. In later chapters we extend this approach to study applications for the change-point inference problem, multiple change-point problem, and detecting a change-point in variance. Using our knowledge of detail coefficient properties, we apply Bayesian change-point methods to the DWT of our time series. Given a univariate time series $\{x_i\}_{i=1}^{N}$, our first step it apply some DWT to the time series thereby transferring our observations to the wavelet domain. We expect our detail coefficients to be centered around zero, but in our case we have additional prior knowledge, namely a shift in mean of the original time series has occurred. Since the additive noise component of the time series retains its statistical properties after the DWT, we have

$$d_{j,k}^{*}|(\tau\ ,\Delta\ ,\sigma^2) \sim N(\Delta q_{j,k}(\tau)\,,\,\sigma^2)$$

where we change notation slightly and perform a rescaling to let $\tau = t/N$ where t represents the true, but unknown, change-point. In other words, τ may be thought of as the change-point rescaled to the interval $(0, 1)$. Δ is introduced to denote the magnitude of the change-point shift. For a given change-point τ, the q function or the mean function in the above expression represents the wavelet transform of the true underlying mean function. With the Haar wavelet in mind,

the q function is defined as follows

$$q_{j,k}(\tau) = \frac{2^{j/2}}{n} \begin{cases} 2^{-j}k - \lfloor \tau n \rfloor, & 2^{-j} \leq \tau < 2^{-j}(k+1/2) \\ -(n2^{-j}(k+1) - \lfloor \tau n \rfloor), & 2^{-j}(k+1/2) \leq \tau < 2^{-j}(k+1). \end{cases} \tag{2.1}$$

Equipped with the wavelet transformed series and the assumed normal distribution of the empirical detail coefficients, we then proceed as above in the Bayesian change-point approach. We may now approximate the likelihood function as

$$p(d^*|\tau, \Delta, \sigma^2) = \prod_i \prod_j p(d_{ij}|\tau, \Delta, \sigma^2).$$

After applying Bayes theorem, we write the posterior distribution as:

$$p(\tau, \Delta, \sigma^2|d^*) \propto \prod_i \prod_j p(d_{ij}|\tau, \Delta, \sigma^2) p_0(\tau, \Delta, \sigma^2). \tag{2.2}$$

Since we are interested in the location of the change-point, Δ and σ are nuisance parameters that are not directly applicable to the problem at hand. Instead of equation (2.2), we would like to consider the marginal posterior distribution function by integrating out Δ and σ

$$p(\tau|d^*) \propto \int p(\tau, \Delta, \sigma^2|d^*) p_0(\sigma^2, \tau, \Delta).$$

In this case we assume σ is constant but unknown throughout the process. Mathematically, we find it convenient to assume a noninformative prior. That is we let

$$p_0(\sigma^2, \tau, \Delta) \propto \frac{1}{\sigma}.$$

Thus we write

$$p(\tau|d^*) \propto \int_0^\infty \int_{-\infty}^\infty \sigma^{-\frac{M}{2}} \exp(\frac{-\sum\sum(d_{jk} - \Delta q_{ij})^2}{2\sigma^2})\sigma^{-1} d\Delta d\sigma^2. \tag{2.3}$$

Expanding the quadratic term and then completing the square in terms of Δ, we integrate out Δ using the properties of the normal distribution

$$p(\tau|d^*) \propto \int_0^\infty \sigma^{2^{-(\frac{M}{4}+\frac{1}{2})}} (\sigma^2)^{\frac{1}{2}} (2\pi)^{-\frac{1}{2}} C^{-\frac{1}{2}} \exp(\frac{\frac{1}{2}(A - \frac{B^2}{C})}{\sigma^2}) d\sigma^2$$

where

$$A = \sum_{j \geq j_0} \sum_k d_{jk}^2, \;\; B = \sum_{j \geq j_0} \sum_k d_{jk} q_{jk}(\tau), \;\; C = \sum_{j \geq j_0} \sum_k q_{jk}^2.$$

Pulling out constants and combining exponents, we may arrange the integrand in the form of an inverse gamma distribution:

$$p(\tau|d^*) \propto C^{-\frac{1}{2}} \int_0^\infty (\sigma^2)^{-\frac{M}{4}} \exp(\frac{\frac{1}{2}(A - \frac{B^2}{C})}{\sigma^2}) d\sigma^2$$

$$= C^{-\frac{1}{2}} (A - \frac{B^2}{C})^{-(\frac{M}{4}-1)}. \tag{2.4}$$

We note this is almost the same expression Ogden states is the result of his calculations. Here, we computed a slightly different exponent term $-(\frac{M}{4} - 1)$ versus his $-(\frac{M}{4} - \frac{1}{2})$ although this has a negligible impact in practice.

The key advantage of this approach is that it should be less sensitive to the structure of the assumed underlying function which follows by [34], but work well with high noise as in the Bayesian scheme. Furthermore, because the DWT approximately decorrelates correlated data, we see the potential to investigate

change-points in ARIMA type series that would appear to violate the assumptions of the Bayesian and MLE approaches. One should also be aware of a technical problem that arises in this approach from requiring our original series to be a length of a power of 2. This may be remedied by various "padding" techniques. For example, we may append zeros or low white noise in front of the time series to obtain the required dyadic length [10].

2.3 Estimating the mean function change-point location for a time series with additive noise

Equipped now with the above theoretical results, we investigate their effectiveness in estimating the change-point location in a variety of different time series with applications in both SPC and biosurveillance. We will incrementally relax our assumptions about the true nature of our underlying function in the different scenarios and test the methods with varying signal to noise ratios ($SNR = \frac{|\Delta|}{\sigma^2}$). Throughout our analysis we will use the Haar wavelet and Daubechies 4-tap wavelet transformations and pad any series of non-dyadic length as described above.

When considering applications to prospective monitoring, there is in general a trade-off between false alarms and delaying the detection of a process change [35]. Once a control chart signals an alarm, we must now ask ourselves whether this is an actual alarm? If so, at which point did the process change? At this point a classic control chart offers little assistance to answer these questions and we must employ other methods to draw appropriate conclusions.

If we now employ a change-point estimation method to our process and find either with low probability that a change-point has occurred or the

change-point is in a very unlikely location due to our prior knowledge, we have evidence of a false alarm. On the other hand, we may find the change-point method returns a change-point location somewhere in agreement with where we might expect from our control chart analysis. Now, not only do we have further strong evidence of a bona fide statistical process change, but also a hypothesis pinpointing where exactly the change occurred.

2.3.1 Retrospective Case

We begin with the retrospective analysis problem setup as this is the key subproblem to the prospective case we consider shortly. We assume our time series $\{x_i\}$ is of the form

$$x_i = \mu_1 + \varepsilon_i \ (1 \leq i \leq \tau)$$
$$x_i = \mu_2 + \varepsilon_i \ (\tau < i \leq N)$$

where it is assumed $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. Without loss of generality, we fix $\mu_1 = 0$ for all simulations and vary μ_2 incrementally from 0.5 to 3. Furthermore, for uniformity purposes, we use a standard deviation value of 1 and adjust the SNR by changing the magnitude of the mean shift. We use a simulated time series of length 128 and randomly vary the change-point position in the simulations. Table 2.1 below reflects the performance of each method by indicating the number of successful change-point estimations for each method out 1000 simulations. We define a successful estimation as the method being able to estimate within 2 time units of the actual change-point. Figure 2.2 below shows two sample time series each with shifts at time point 85 to provide some intuition for the data being examined. Note that for an SNR of 3 units the change-point is obvious to the naked eye while being fairly obscured for an SNR of 1.

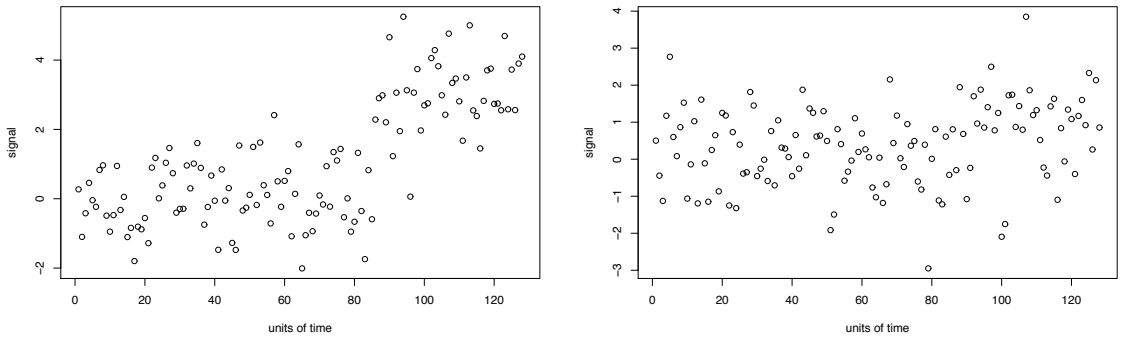

Figure 2.2: Two example time series with shifts of 3 (left) and 1 (right) at time point 85

Table 2.1: Results of 1000 simulations for the retrospective monitoring case. The table shows the number of successful change-point detections (i.e. estimated change-point is within 2 time units of actual change-point).

SNR	MLE	Bayesian	Bayesian-wavelet
0.5	243	245	249
1.0	609	605	593
1.5	824	817	799
2.0	949	927	921
2.5	978	956	959
3.0	991	985	981

As expected, Table 2.1 shows improved performance for all three methods as the SNR increases. Between Bayesian, MLE, and Bayesian-wavelet there is a slight, but consistent, difference favoring the MLE method in simulations involving larger shifts while no difference is observed with low SNR. In all cases we used the complete set of detail coefficients for the Bayesian-wavelet analysis. Based on these simulations, provided that the SNR was not too low, all three method are effective approaches to correctly estimate actual change-points.

While Table 2.1 indicates the Bayesian-wavelet approach is slightly less effective in estimating the true change-point location, we may easily modify the marginal posterior distribution function to reflect prior information. For example, suppose that we have prior information that the most likely location of the

change-point occurs in the latter 10% of the time series. In this case the noninformative uniform prior on the change-point location would not be optimal. Instead, consider the following beta-binomial distribution density function:

$$p_0(t|\alpha,\beta) = \binom{n}{t} \frac{B(t+\alpha, n-t+\beta)}{B(\alpha,\beta)} \tag{2.5}$$

where $B(\cdot)$ is the beta function.

This seems to be a potentially nice prior distribution since this is a discrete probability mass function for our time series with matching support of the likelihood function. Furthermore, this distribution has the flexibility with appropriately chosen parameters to favor different portions of the interval (0,1) as our prior knowledge dictates. We can rescale the above pmf over the interval (0,1) by letting $p_0(t|\alpha,\beta) = p_0(\tau|\alpha,\beta)$ for $\tau = t/n$. The derivation of this new posterior will be similar to the previous case since this distribution contains no parameters that will effect the integrations in the derivation. The new posterior may be written as

$$p(\tau|w) \propto p_0(\tau|\alpha,\beta) C^{-1/2} (A - \tfrac{B^2}{C})^{-(M-2)/4}$$

where p_0 is as in equation 2.5 while

$$A = \sum_{j \geq j_0} \sum_k d_{j,k}^2, \;\; B = \sum_{j \geq j_0} \sum_k d_{j,k} q_{j,k}(\tau) 2, \;\; C = \sum_{j \geq j_0} \sum_k q_{j,k}^2.$$

Table 2.2 represents simulation results comparing the three methods where the true change-point is randomly selected from the final 10% of the time series. By supposing we have prior information on this location of the change-point, we place a Beta-binomial prior distribution with parameters $\alpha = 10$ and $\beta = 2$. An example of this situation occurs using control charts as seen in the next section, where we expect alarms to occur in the latter portion of the time series. With our informative prior on the Bayesisan-wavelet marginal posterior equation, we now

Table 2.2: Results of 1000 simulations where the change-point location is chosen at random from the last 10% of the time series. The table shows the number of successful change-point detections (i.e. estimated change-point is within 2 time units of actual change-point). We now place a Beta-binomial prior distribution on the Bayesian-wavelet model with parameters $\alpha = 10$ and $\beta = 2$.

SNR	MLE	Bayesian	Bay-Wav-BetBin
0.5	205	238	320
1.0	575	590	625
1.5	802	795	820
2.0	924	915	931
2.5	973	961	963
3.0	989	978	987

see the Bayesian-wavelet method slightly outperforms the other methods with low SNR . When the SNR is high, the change-point problem becomes less difficult and prior information on the change-point location is less important for a correct change-point location estimation.

2.3.2 Prospective monitoring case

This problem represents a basic scenario in the SPC setting. Here we fundamentally change the problem from the retrospective case above, to that of monitoring an ongoing process with a standard Shewart chart using traditional Western Electric signaling rules in the hope of an early detection of a change-point. Once the chart signals an alarm, we implement each one of the above methods to estimate where the change-point actually occurred.

The simulated time series in Figure 2.3 has an actual change-point at point 91 as indicated by the red vertical line. The vertical green line at point 103 is where the control chart alarms to signal the process is out of control. From point 103 we then look back and employ our change-point algorithms and test the effectiveness of each method. For the purpose of these simulations the success

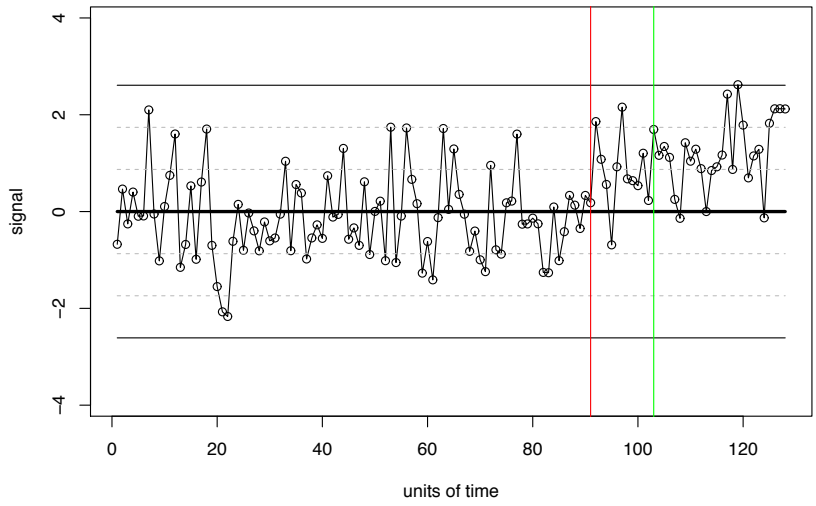

Figure 2.3: Example time series with a shift of 1 at point 91 that triggers an alarm at time point 103.

criteria is the same as above, where we consider only the single highest probability change-point from each method and count a success so long as the estimated and true points are within 2 time units of each other.

The prospective scheme requires two separate processes to both work correctly. Firstly, the control chart must signal an alarm at some time after the change-point. Next, after the alarm signals, the detection method also needs to then correctly pick out the change-point. Obviously, if by chance, the chart signals an alarm prior to the true change-point, the change-point detection methods cannot correctly estimate the true change-point. Additionally, if the step size is sufficiently small, the control chart may not signal an alarm after the shift at all. In either case, we will fail to detect the true change-point.

Table 2.3 summarizes the results from 1000 simulations where we denote separate results for the Bayesian-wavelet methods with both flat and Beta-binomial priors. Our results once again suggest each method works comparably. Our work to include the Beta-binomial prior for the Bayesian

Table 2.3: Results of 1000 simulations from the prospective monitoring case. Out of 1000 simulations, each number represents the number of successful change-point estimations. That is, the alarm signals sometime after the true change-point and then the detection method correctly estimates to within two time units the true change-point. Here, a "Good Alarm" denotes the control chart correctly signaling sometime after the true change-point which is a random point chosen after time point 75 but before the series ends at time point 128.

SNR	Good Alarms	MLE	Bayesian	Bay-Wav-flat	Bay-Wav-BetBin
0.5	624	166	190	122	133
1.0	904	592	587	555	559
1.5	910	748	742	740	733
2.0	907	853	828	836	836
2.5	914	898	883	878	877
3.0	922	921	914	922	906

wavelet-method offered no clear advantage for $\alpha = 2$ and $\beta = 1$. If, however, we had more confidence in our prior knowledge and could choose more informative α and β parameters, then perhaps this modified posterior distribution could be of value in certain cases. A key point, is that we are achieving comparable results with the Bayesian-wavelet method without the same strict assumptions on the time series that both the MLE and Bayesian methods make.

2.3.3 Prospective monitoring for a correlated data set with a single change-point

Now we shift our attention away from the independence assumption of the previous investigations. Here our simulations could fall more in line with chemical processing control problems or biosurveillance scenarios where the independence assumption is not valid. We run simulations on an AR1 process with various correlation and step size combinations as indicated in the tables below. Explicitly, we simulate from the following model:

$$Y_t = \phi Y_{t-1} + e_t, \qquad 1 \leq t < k$$
$$Y_t = \phi Y_{t-1} + e_t + \Delta, \qquad k \leq t < n$$

where, as in the previous scenarios, Δ is some step size occurring at the change-point and ϕ is the correlation coefficient of the time series.

Table 2.4: Out of 1000 simulations, each entry represents the number of successful change-point estimations. A successful change-point is considered the detection method correctly estimates to within two time units the true change-point.

ϕ	Shift Size	MLE	Bayesian	Bayesian-wavelet
.3	0.5	141	159	189
.3	1.0	454	458	473
.3	1.5	690	682	675
.3	2.0	856	826	864
.3	2.5	920	900	922
.3	3.0	978	979	973
.6	0.5	80	91	96
.6	1.0	203	234	237
.6	1.5	386	401	406
.6	2.0	601	616	615
.6	2.5	721	716	786
.6	3.0	818	810	835
.9	0.5	51	49	107
.9	1.0	78	80	160
.9	1.5	118	108	236
.9	2.0	159	172	363
.9	2.5	233	225	518
.9	3.0	334	348	625

We assume for this scenario that we are already in Phase II of the monitoring process, but that we would like to detect where the shift occurred. That is, our alarming mechanism has signaled an out of control state but we wish to find the highest probability of where the change occurred.

In principle the assumptions of both the MLE and Bayesian models are violated; however, we see in practice they perform fairly well until we use very highly correlated data. With highly correlated data Table 2.5 clearly breaks out the performance levels for the three change-point detection methods. The Bayesian-wavelet method applies the Bayesian method of the approximately uncorrelated data after a wavelet transform. We found the best results were achieved using only the finest level of detail coefficients in the wavelet analysis.

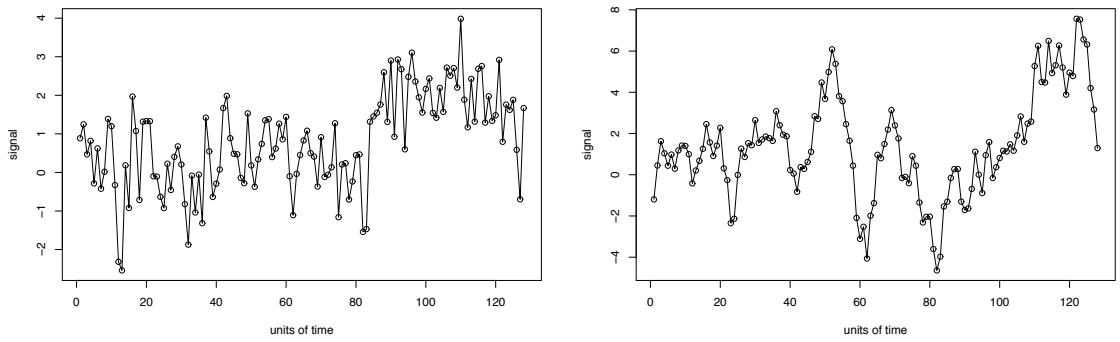

Figure 2.4: Example time series with shift sizes of 3 and $\sigma^2 = 1$ at time point 83 are given for ϕ=.3 (left), and .9 (right). Notice for ϕ=.3 the change-point is clear while for ϕ=.9 the change-point is obscured.

Simulations show that even with $\phi = 0.9$ the Bayesian-wavelet method performs fairly well provided the SNR is sufficiently high. On the other hand, the other two methods perform poorly in these cases.

2.3.4 Change-point of a piecewise smooth function

For this final scenario we consider the monitoring problem from a context that may have applications in biosurveillance or other situations where less is known about the true underlying function. Unlike the typical situation in SPC, in biosurveillance we are monitoring processes that we cannot directly control. As such, it is likely there exists a natural unknown pattern, perhaps annual or otherwise, that we must account for when detecting a change-point. For example, consider occurrences of influenza which may peak during influenza season and then taper off to a minimum value in the offseason before rising again. With this idea in mind, we assume we have a smooth underlying function that may represent the annual cycle of some similar type process with additive noise. Once again we assume we are in Phase II of the prospective monitoring phase and wish to detect the change-point after our control chart or other monitoring mechanism

has signaled an alarm.

Figure 2.5 shows three different example simulations setups where we adjust the standard deviation of the random noise component. We relax our success criteria slightly and now count successes as the true change-point being within two of the top two change-point estimations of the respective algorithms rather than just the single top estimation as above. Interestingly, the MLE and Bayesian method do not even accidently detect any change-points when the SNR is high. The underlying functional structure violates the assumptions of these methods in such a way that the algorithms consistently choose points near the peak of the time series sometime around the 50th time unit. Since we randomly choose change-points in the latter of half of the time series, these methods fail in each estimation. When the noise is increased, the time series more closely approximates a pure white noise process and these methods register a few successes by chance alone.

Table 2.5: Out of 1000 simulations, each entry represents the number of successful change-point estimations where a success is counted as the true change-point being within 2 time units of the top two estimations for each method.

σ^2	Shift Size	MLE	Bayesian	Bayesian-wavelet
.1	1	0	0	937
.3	1	0	0	872
.5	1	2	3	585
.7	1	3	5	454
1.0	1	12	15	299

The Bayesian-wavelet method on the other hand does well seeing through the underlying sine curve and picking out the correct change-point when the appropriate detail coefficients are applied. We found that using the four highest detail coefficients provided the best results across the spectrum of standard

deviations considered. As we noted in section 2.3.2, wavelets represent smooth functions sparsely. So by taking the DWT of this series, the detail coefficients tell us a lot of change is occurring at the change-point. The higher detail coefficients of 6, 5, and 4 contain very localized change-point information which is the characteristic of single abrupt change-point. When the noise we are trying to see through is not too great, Table 2.5 shows how this wavelet property nicely picks out the true change-point.

2.4 Conclusion

This chapter introduced the univariate statistical change-point problem under a variety of assumptions while providing applications throughout. Specifically, we examined the performance of Ogden's Bayesian-wavelet change-point detection method against MLE and Bayesian approaches. After introducing necessary theoretical background, simulations were performed to provide evidence for the strengths and deficiencies of each method. Simulation results verified that MLE and Bayesian methods both performed well when their assumptions remained valid for the time-series being tested; however, both methods were unreliable when detecting the change-point in highly correlated data or when the true underlying mean function varied smoothly.

The Bayesian-wavelet technique on the other hand performed well in each scenario we considered provided the SNR was not too low. Because the wavelet transform approximately decorrelates time-series data and represents smooth functions sparsely, applying Bayesian change-point techniques to wavelet details captured the true change-point in settings when both MLE and Bayesian methods failed. In situations where weaker assumptions about the time-series can

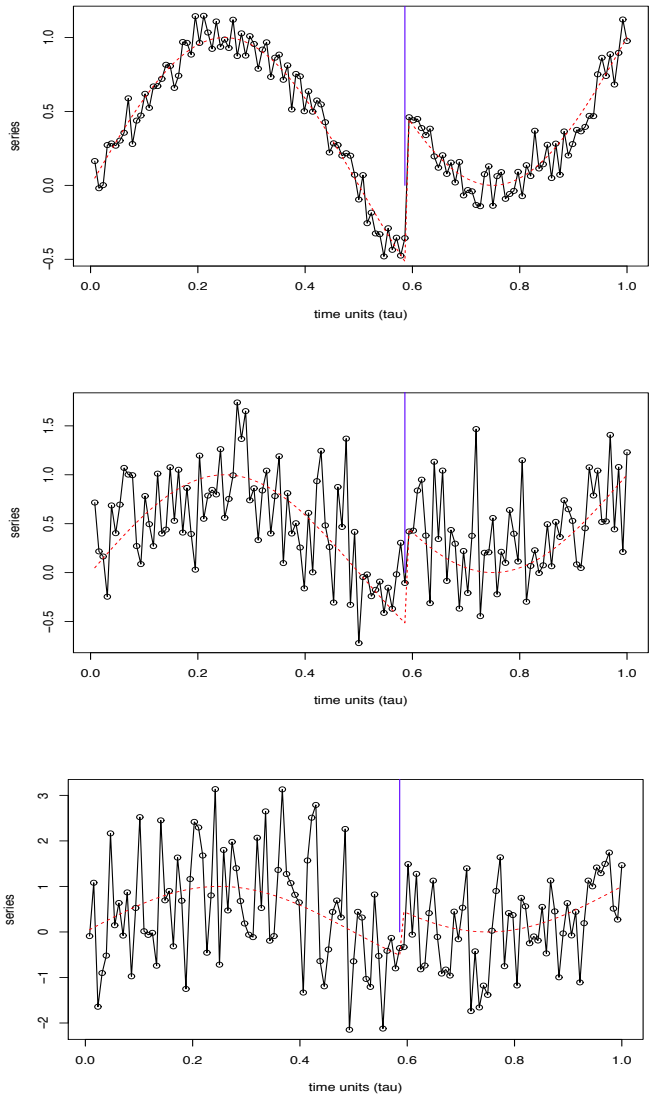

Figure 2.5: Three example time series with a smooth underlying function except at one shift point. The standard deviations from upper right and proceeding clockwise are .1, 0.5, and 1 respectively where the blue vertical line represents the actual change-point.

be made, applying this method could offer greater robustness by making fewer model assumptions. With this first most fundamental change-point case in hand, we now turn to the many related change-point problems in the chapters to follow.

Chapter 3

Further Univariate Case Applications

3.1 Introduction

In this chapter we once again consider a univariate time series, $X = \{x_i\}_{i=1}^{N}$ for $N \in \mathbb{N}$, with an additive noise component. In the previous chapter, we assumed a shift occurred in the mean function and the result of our analysis was an estimation of where the change-point occurred. In this chapter we build upon this previous work to answer questions about the existence of a change-point, estimating multiple change-points, and estimating the location of a variance change-point in a time series.

If a change-point in a time series is assumed, then we maximize the likelihood of our time series model for the most likely change-point location. This algorithmically straight forward process returns good results provided the SNR is not too low and the random component follows a normal distribution. We still would like to determine the existence of a change-point in the time series in the first place. Here MLE and Bayesian methods offer two fundamentally different approaches which we explore in section 3.2.3 below. Once we can answer the inference question, then we can extend our previous results to the multiple change-point problem as we demonstrate in section 3.4.

Finally in section 3.5, we develop a Bayesian-wavelet approach to analyzing the change-point of variance problem of a time series by transforming the series into the wavelet domain. While various maximum likelihood (MLE) and Bayesian methods for detection of change-point in variance exist, they typically require strict assumptions that limit their applicability when these assumptions are violated. As before, our more indirect approach involves a wavelet transformation of the time series and then a Bayesian analysis on the empirical detail coefficients. As we see below, we find some precision in change-point estimates may be lost but what is gained is an applicability to more general contexts.

3.2 Inference of a change-point in mean

In chapter 2 we demonstrated how statistical control charts may be used to signal the presence of a statistical change-point in a time series. In this section we demonstrate how the MLE, Bayesian, and Bayesian-wavelet results estimating the location of a change-point in a time series may also be applied to establish inference of a statistical change-point. The classical MLE approach is presented below to provide both background for the inference problem and to serve as a reference against which to compare the other methods [36, 26, 37]. We then develop Bayesian methods both in the classical sense and then using wavelet detail coefficients. Finally, numerical experiments are conducted to compare the effectiveness of each method using simulated data.

3.2.1 Maximum likelihood approach

With the maximum likelihood approach we wish to compare the null hypothesis

$$H_0 : \mu_1 = \mu_2 ... = \mu_N$$

against the alternative hypothesis

$$H_a : \mu_1 = \mu_2 \cdots = \mu_k \neq \mu_{k+1} \cdots = \mu_N$$

for some time series $\{x_i\}_i^N$, $N \in \mathbb{N}$. Each x_i is assumed normally distributed around μ_i sharing a known constant variance of σ^2. Under these hypotheses, we form two separate likelihood functions L_0 and L_a for the null and alternative hypothesis respectively:

$$L_0(x_1, .., x_k|\mu, \sigma) = \prod_{i=1}^{N} \sqrt{\frac{1}{2\pi\sigma^2}} \exp\Big[-\frac{1}{2\sigma^2}(x_i - \mu)^2 \Big]$$

and

$$L_a(x_1, .., x_k|\mu_1, \mu_2, \sigma, \tau)$$
$$= \prod_{i=1}^{\tau} \sqrt{\frac{1}{2\pi\sigma^2}} \exp\Big[-\frac{1}{2\sigma^2}(x_i - \mu_1)^2 \Big] \prod_{i=\tau+1}^{N} \sqrt{\frac{1}{2\pi\sigma^2}} \exp\Big[-\frac{1}{2\sigma^2}(x_i - \mu_2)^2 \Big].$$

We then let

$$\hat{\mu} = \tfrac{1}{N} \sum_{1}^{N} x_i \quad \hat{\mu}_1 = \tfrac{1}{\tau} \sum_{i=1}^{\tau} x_i \quad \hat{\mu}_2 = \tfrac{1}{N-\tau} \sum_{\tau+1}^{N} x_i$$

be our maximum likelihood estimators for μ, μ_1, and μ_2, respectively. By taking $2 \log \frac{L_a}{L_0}$ we obtain the ratio test statistic

$$W = \max_{1 \leq \tau < N} \sqrt{E_\tau}.$$

This is the same test statistic as previously presented in chapter 2. While obtaining the ratio test statistic is fairly straight forward, the more difficult task remains in establishing a p-value to either accept or reject H_0 at some desired confidence level. The derivation of the distribution of W requires a series of careful nonstandard arguments as presented by such authors as Worsley,

Hawkins, Chen and Gupta, and Csörgo and L. Horváth [36, 26, 37, 38]. Due to the required preparation and length of this derivation, we refer interested readers to one of these sources for complete details.

3.2.2 Bayesian approach

An alternative approach to classic hypothesis testing is the Bayesian model comparison approach. We note several criticisms of the frequentist method for hypothesis testing exist [13]. For example:

Models must be nested within one another.

By failing to accept H_a, we do not prove H_0 rather we simply fail to reject H_0.

The p-value itself is not a direct interpretation of "weight of evidence" rather a long term probability.

Computing a likelihood ratio probability distribution may be computationally intractable.

In the Bayesian paradigm, rather than conduct a hypothesis test, we compare two models, say M_1 and M_2. Given our observed data, we form an odds ratio using these two models to decide which model the observed data most strongly supports. Calculating the posterior odds of M_1 against the prior odds of M_1 we obtain the so called Bayes factor (BF). By assuming $P(M_2|x_1, ...x_N) = 1 - P(M_1|x_1, ...x_N)$, we may express the BF as:

$$BF = \frac{P(M_1|x_1, ...x_N)/P(M_2|x_1, ...x_N)}{P(M_1)/P(M_2)} = \frac{\frac{P(x_1,.....x_N|M_1)P(M_1)}{p(x_1,.....x_N)} / \frac{P(x_1,.....x_N|M_2)P(M_2)}{p(x_1,.....x_N)}}{P(M_1)/P(M_2)}$$
$$= \frac{P(x_1, ...x_N|M_1)}{P(x_1, ...x_N|M_2)}$$

We still need some way to interpret the Bayes factor. One often cited source is from a paper in 1995 where Kass [39] proposes the values given in Table 3.1.

Table 3.1: Bayes Factor Table

$\log(BF)$	Evidence against M_2
0 to 1/2	Not worth a bare mention
1/2 to 1	Substantial
1/2 to 2	Strong
> 2	Decisive

Applying the Bayes factor approach to the change-point problem, we let M_1 be the case with a change-point in mean at time point τ where $\tau \in \{1, 2, \ldots, N-1\}$. Furthermore, for M_1 we assume μ_1, μ_2, and σ are unknown. The M_2 case will be the no change model where we assume μ and σ are constant (but unknown) throughout the time series. We construct each model in turn.

Model 1 (M_1) (with change-point):

We build our model based on the assumption we do not have sufficient prior knowledge to use an informed prior. In this case, we may apply the improper noninformative prior and assume our prior distribution is equal to a constant. Mathematically for some constant K, we find it convenient to represent p_0 as

$$p_0(\mu_1, \mu_2, \sigma^2, \tau|M_1) = K\frac{1}{\sigma}.$$

Next we need to integrate out our unknown nuisance parameters μ_1, μ_2, and σ by evaluating the following integral

$$p(x_1, x_2, ...x_n, \tau|M_1) = \int p(x_1, x_2, ...x_n|M_1, \mu_1, \mu_2, \sigma^2, \tau) p_0(\mu_1, \mu_2, \sigma^2, \tau|M_1). \quad (3.1)$$

More explicitly,

$$\begin{aligned} p(x_1, x_2, ...x_n, \tau|M_1) = \int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty (2\pi\sigma^2)^{-N/2} \prod_{i=1}^{\tau} \exp\left[-\tfrac{1}{2\sigma^2}(x_i - \mu_1)^2 \right] \\ \prod_{i=\tau+1}^{N} \sqrt{\tfrac{1}{2\pi\sigma^2}} \exp\left[-\tfrac{1}{2\sigma^2}(x_i - \mu_2)^2 \right] \tfrac{1}{\sigma} d\mu_2 d\mu_1 d\sigma^2 \\ = (2\pi)^{-\frac{n}{2}+1} (\tau(n-\tau))^{-\frac{1}{2}} \Gamma(\frac{n}{2} + \frac{1}{2}) \\ \left[\frac{1}{2} (\sum_{i=1}^{i=\tau} (x_i - \bar{x}_\tau)^2 + \sum_{i=\tau+1}^{i=N} (x_i - \bar{x}_{n-\tau})^2) \right]^{-(\frac{n}{2}+\frac{1}{2})}. \end{aligned}$$

Model 2, M_2 (No change-point):

We again use the same noninformative prior as above. Then

$$\begin{aligned} p(x_1, x_2, ...x_n, \tau|M_1) &= \int p(x_1, x_2, ...x_n|M_1, \mu, \sigma^2, \tau) p_0(\mu, \sigma^2, \tau|M_1) \\ &= \int_0^\infty \int_{-\infty}^\infty \prod_{i=1}^{N} \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2 \right] \frac{1}{\sigma} d\mu d\sigma^2 \quad (3.2) \\ &= (2\pi)^{-\frac{n}{2}+\frac{1}{2}} (n)^{-\frac{1}{2}} \Gamma(\frac{n}{2} - 1) (\frac{1}{2} (\sum_{i=1}^{i=N} (x_i - \bar{x}_\tau)^2]^{-(\frac{n}{2}-1)} \end{aligned}$$

Taking the ratio of equations (3.1) and (3.2), we compute our Bayes factor from the observed data.

3.2.3 Bayesian-wavelet approach

We now compute the Bayes factor for the Bayesian-wavelet method. Once again let $\{x_i\}_{i=1}^N$ be our observed time series. We wish to compare Model 1 (M_1),

a change in mean has occurred at time τ, against Model 2 (M_2), no statistical change has occurred. Since we will be using the final expression in terms of a ratio, it is unclear *a priori* which constant terms will cancel out in the final analysis. Thus, we carefully rederive the original expression maintaining all terms along the way. Letting $\mathbf{d}$ represent the vector of detail coefficients after taking a DWT of our time series, we seek a closed form expression for

$$p(\mathbf{d}|M_1) = \int p(\mathbf{d}|M_1, \sigma^2, \tau, \Delta,)p_0(\sigma^2, \tau, \Delta|M_1)d\Delta d\sigma^2.$$

In this case we assume σ is constant but unknown throughout the process and find it convenient to assume a noninformative prior. That is we let

$$p_0(\sigma^2, \tau, \Delta|M_1) = K\frac{1}{\sigma}$$

where K is an unknown constant common to both models that later will cancel out in the BF ratio. Thus we compute:

$$p(\mathbf{d}|M_1) = \int_0^\infty \int_{-\infty}^\infty (2\pi)^{-\frac{M}{2}} \sigma^{-\frac{M}{2}} \exp(\frac{-\sum\sum(d_{j,k} - \Delta q_{i,j})^2}{2\sigma^2}) K\sigma^{-1} d\Delta d\sigma^2 \quad (3.3)$$

Expanding the quadratic term and then completing the square in terms of Δ, we integrate out Δ using the properties of the normal distribution, so

$$p(\mathbf{d}|M_1) = \int_0^\infty K(2\pi)^{-\frac{M}{2}} \sigma^{2^{-(\frac{M}{2}+\frac{1}{2})}} (\sigma^2)^{\frac{1}{2}} (2\pi)^{\frac{1}{2}} C^{-\frac{1}{2}} \exp(\frac{\frac{1}{2}(A - \frac{B^2}{C})}{\sigma^2}) d\sigma^2$$

where

$$A = \sum_{j\geq j_0}\sum_k d_{jk}^2, \;\; B = \sum_{j\geq j_0}\sum_k d_{jk}q_{j,k}, \;\; C = \sum_{j\geq j_0}\sum_k q_{jk}^2.$$

Pulling out constants and combining exponents, we may arrange the integrand in the form of an inverse gamma distribution:

$$
\begin{aligned}
p(\mathbf{d}|M_1) &= K(2\pi)^{\frac{-(M-1)}{2}} C^{-\frac{1}{2}} \int_0^\infty (2\pi)^{-\frac{M}{2}} (\sigma^2)^{-(\frac{M}{2}-1)-1} \exp(\frac{\frac{1}{2}(A-\frac{B^2}{C})}{\sigma^2}) d\sigma^2 \\
&= K(2\pi)^{\frac{-(M-1)}{2}} C^{-\frac{1}{2}} \Gamma(\frac{M}{2}-1)(\frac{1}{2}(A-\frac{B^2}{C}))^{-(\frac{M}{2}-1)}.
\end{aligned} \tag{3.4}
$$

We note this is almost the same expression as in chapter 2 only now with constant terms retained. Next we construct the case for M_2 where we assume both constant variance and constant mean through the time series using the same noninformative prior

$$
p(\mathbf{d}|M_2) = \int p(\mathbf{d}|M_2, \sigma^2, \Delta, \tau) p_0(\sigma^2, \tau, \Delta|M_2).
$$

We must account for each possible value of τ, so the explicit form we must compute becomes

$$
p(\mathbf{d}|M_2) = \sum_{\tau=1}^{N-1} \int_0^\infty \int_{-\infty}^\infty K(2\pi)^{-\frac{M}{2}} \sigma^{-\frac{M}{2}-1} \exp(\frac{-\sum\sum(d_{j,k}-\Delta q_{i,j})}{2\sigma^2}) d\Delta d\sigma^2. \tag{3.5}
$$

In M_2, however, we are assuming that $\Delta = 0$. Since the only term in the above expression with dependence on τ is q_{jk}, equation (3.5) simplifies and we obtain

$$
\begin{aligned}
p(\mathbf{d}|M_2) &= K(2\pi)^{-\frac{M}{2}} \int_0^\infty (\sigma^2)^{-(\frac{M}{2}-\frac{1}{2})-1} \exp(\frac{-\sum\sum d_{jk}^2}{2\sigma^2}) d\sigma^2 \\
&= K(2\pi)^{\frac{-(M)}{2}} \Gamma(\frac{M}{2}-\frac{1}{2})(\frac{1}{2}A)^{-(\frac{M}{2}-\frac{1}{2})}
\end{aligned} \tag{3.6}
$$

which follows from the form of the inverse gamma distribution. Finally, we combine the results from our M_1 and M_2 computations to obtain our Bayes

Factor (BF)

$$BF = \frac{\arg\max_{\tau} \sqrt{2\pi} C^{-\frac{1}{2}} \Gamma(\frac{M}{2} - 1)(\frac{1}{2}(A - \frac{B^2}{C}))^{-(\frac{M}{2}-1)}}{\Gamma(\frac{M}{2} - \frac{1}{2})(\frac{1}{2}A)^{-(\frac{M}{2}-\frac{1}{2})}}. \tag{3.7}$$

This final result will be what we implement for the simulation in the the next section.

3.3 Simulations: Inference of a change-point

We investigate the performance of detecting a statistical change-point in the mean function for each of the methods described in the previous section. In particular, we conduct a variety of simulations both with and without a statistical change-point present and then record both the number of correct detections and false positives for the Bayesian-wavelet, Bayesian, and MLE approaches. The two basic scenarios we examine are a constant mean function (except at the change-point) and a smoothly varying mean function (except at the change-point). In the first case we expect the MLE and Bayesian methods to outperform the Bayesian-wavelet method, since the data is generated from the exact assumption these models are based upon. On the other hand, when considering a more general mean function, we then expect the Bayesian-wavelet method to outperform the other methods. Observing Tables 3.2 and 3.3 in fact this is indeed what we find from our simulations.

With any Bayes factor approach a certain amount of subjectivity must be accepted when selecting the actual value to be used for choosing between the models. There will always exists a clear tradeoff between false positives (with $\Delta = 0$) against correct detections (with $\Delta \neq 0$). For these simulations, we decided a false positive was less desirable than a missed actual shift and so used a

higher BF threshold value of 2. There is a certain flexibility at the modeler's discretion in this case. By adjusting the Bayes factor cut-off value we could clearly be more sure of detecting actual shift changes, but at the cost of additional false positives.

We find the Bayesian-wavelet method is somewhat less effective in correctly inferring whether or not a change-point in mean has occurred than the MLE and pure Bayesian methods when the simulated data is generated from the exact assumptions from which the MLE and Bayesian are developed from. This difference in performance is really only seen, however, for the $SNR = 0.5$ case when the shift size is quite small. On the other hand, both the MLE and pure Bayesian methods are of no value when inferring the existence of a change-point from a noisy sine wave. Since both of these methods indicate a change-point almost always occurs (even in the absence of a shift), they become useless indicators of inferring a statistical change-point in this case. This result is not surprising since as we saw in chapter 2, the pure Bayesian and MLE methods were completely unable to determine the location of a change-point from a smoothly varying mean function in the first place. The Bayesian-wavelet method, however, performs quite well again in this scenario especially when the SNR is sufficiently high. We conclude that when assumptions pertaining to the true underlying mean function are uncertain, the Bayesian-wavelet method offers greater robustness than the classical MLE and pure Bayesian methods in correctly inferring the existence of a statistical change-point.

Table 3.2: Results of 1000 simulations for detecting the existence of a change-point in the mean function for a noisy time series where the true underlying mean function is either a constant or a step function (see Figure 3.1 for two examples). In the case of $SNR = 0$, the methods should "detect" as few shifts as possible. In the case with where a shift exists ($SNR > 0$), the larger the number of detections indicates the effectiveness of the method.

SNR	MLE	Bayesian	Bayesian-wavelet
0.0	51	63	48
0.5	439	421	410
1.0	963	971	956
1.5	1000	1000	1000
2.0	1000	1000	1000
2.5	1000	1000	1000

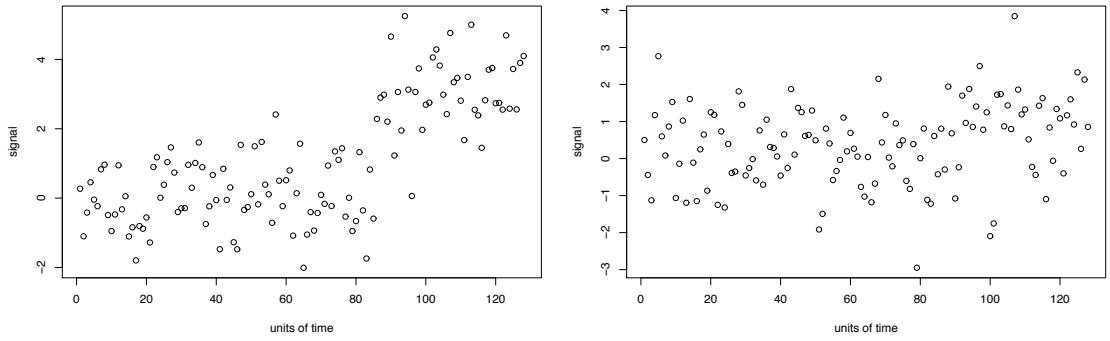

Figure 3.1: Two representative time series for simulation results presentend in Table 3.2. Here the time seires have an $SNR = 3$ (left) and $SNR = 1$ (right) each with the shift occurring at time point 85.

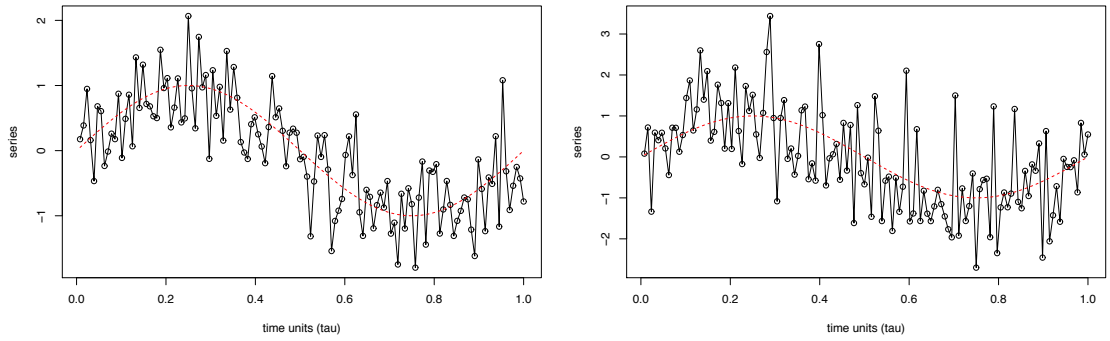

Figure 3.2: Two representative time series for simulation results presentend in Table 3.3. Here we have no shift present ($SNR = 0$) along with additive noise components of $\sigma = .5$ (left) and $\sigma = 1$ (right).

Table 3.3: Results of 1000 simulations from a noisy sine-wave with and without a shift, each entry represents the number of shift change detections (see Figures 3.2 and 3.3 for example time series). We adjust the variance of the addtive noise component and observe how this effects both the $SNR = 0$ and the $SNR > 0$ cases. In the case of $SNR = 0$, the method should "accidently" detect as few shifts as possible. In the case with a shift ($SNR > 0$), the larger the number of detections indicates the effectiveness of the method.

σ^2	Shift Size	MLE	Bayesian	Bayesian-wavelet
0.1	0	1000	1000	1
0.3	0	1000	1000	28
0.5	0	1000	1000	43
0.7	0	1000	1000	99
1.0	0	1000	1000	164
0.1	1	1000	1000	964
0.3	1	1000	1000	948
0.5	1	1000	1000	925
0.7	1	998	998	890
1.0	1	935	845	747

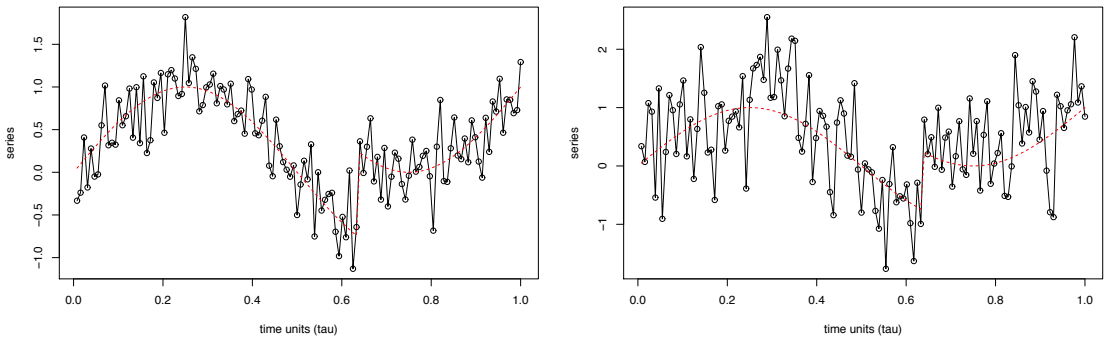

Figure 3.3: Two representative time series for simulation results presentend in Table 3.3. Here we have a shift of $\Delta = 1$ along with variances from the additive noise components of $\sigma = .3$ (left) and $\sigma = .7$ (right). In both figures the shift occurs at $\tau = .63$.

3.4 Multiple change-points in a time series

With our methods to both infer the existence of a change-point and to estimate its location, a natural next step is to analyze time series with multiple change-points. In this section we present two approaches to the multiple change-point problem by applying the Bayesian-wavelet change-point estimation equation. The first involves the so called "binary segmentation algorithm" while the second involves deriving a multiple change-point model from the onset. We

then discuss the strengths and limitations of each method.

3.4.1 Bayesian-wavelet method applied to the binary segmentation algorithm for multiple change-points

We once again consider a time series $\{x_i\}_{i=1}^{N}$, $N \in \mathbb{N}$ and assume

$$x_i \sim N(\mu_p, \sigma^2)\ k_p \leq i \leq k_{p+1}$$

where $\{k_p\}_{p=0}^{M}$, $M \in \mathbb{N}$ denotes the set of change-point locations of the time series. More explicitly, we assume

$$\sigma_1 = \sigma_2 ... = \sigma_N$$

but that

$$\mu_1 = \mu_2 ... = \mu_{k_1} \neq \mu_{k_1+1} = \mu_{k_1+2} ... = \mu_{k_2} \neq \ldots \neq \mu_{k_M} = \mu_{k_M+1} \ldots \mu_N.$$

Rather than take on this problem all at once we initially simplify the problem by setting up two competing models.

Model 1 (M_1): Constant variance and single change-point in mean, that is

$$\mu_1 = \mu_2 ... = \mu_k \neq \mu_{k+1} ... = \mu_N.$$

Model 2 (M_2): Constant variance and constant mean, that is

$$\mu_1 = \mu_2 ... = \mu_N.$$

In a previous section we discussed the Bayes factor (BF) and derived the following result for the above models

$$BF = \frac{\arg\max_{\tau} \sqrt{2\pi} C^{-\frac{1}{2}} \Gamma(\frac{M}{4} - 1) \frac{1}{2} (A - \frac{B^2}{C}))^{-(\frac{M}{4}-1)}}{\Gamma(\frac{M}{4} - \frac{1}{2}) \frac{1}{2} (A)^{-(\frac{M}{4}-\frac{1}{2})}}$$

where

$$A = \sum_{j \geq j_0} \sum_{k} d_{j,k}^2, \;\; B = \sum_{j \geq j_0} \sum_{k} d_{j,k} q_{j,k}(\tau), \;\; C = \sum_{j \geq j_0} \sum_{k} q_{j,k}^2$$

In the case of more than one change-point, this model setup clearly does not accurately reflect the true time series. Still, if more than one change-point in the mean exists, we do not expect our Bayes factor to favor the M_2 case. That is, should multiple change-points in the mean exist, we expect our Bayes-factor to still favor M_1.

Assuming for the moment that we actually have multiple change-points and our Bayes-factor indicates a change-point exists, we then estimate the location of the change-point with the results we derived in section 3.2.3. Explicitly, we say a change-point in mean occurs at time point k_i, where

$$k_i = \arg\max_{\tau} \left(C^{-\frac{1}{2}} (A - \tfrac{B^2}{C})^{-(\frac{M}{4}-1)} \right)$$

Now we segment our original time series into two separate time series $\{x_i\}_{i=1}^{k_i}$ and $\{x_i\}_{i=k_i+1}^{N}$ and repeat the above procedure on the segments $\{x_i\}_{i=1}^{k_i}$ and $\{x_i\}_{i=k_i+1}^{N}$ separately. If at any time the BF indicates no change-point, then we terminate the algorithm for that particular segment. When all segments terminate, we then have our estimated number of change-point along with their locations. This method is best illustrated through a variety of typical examples as we provide below.

Example 1: no change-point exists

Figure 3.4 is an example of a pure white noise process. Applying our algorithm, we compute a Bayes-factor for the time series. Our R output gives a Bayes factor of $BF = .916$, which per section 3.2.3 indicates M_1 is not favored

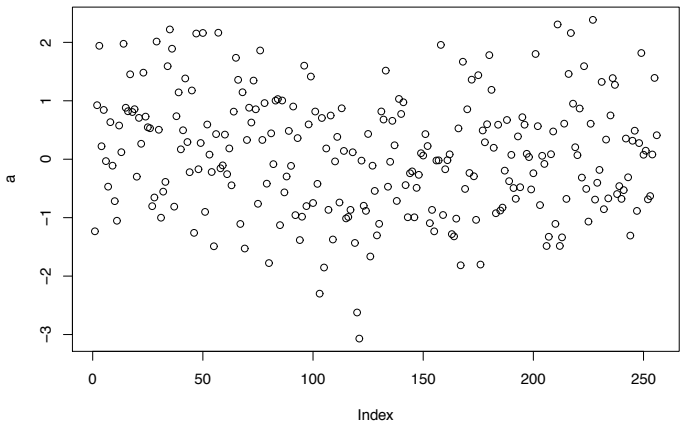

Figure 3.4: Example time series with no change-point in mean

over M_2. Thus we terminate our binary segmentation algorithm after just one step and correctly conclude no change-point in the series exists.

Example 2: one change-point exists

Figure 3.5 is an example of a time series with a single change-point of step size 1 at point 79. Applying our algorithm, we obtain the following results for the time series.

Iteration 1

- Bayes factor output from R for entire series $BF = 14.7$ which per section 3.2.3 indicates a change-point exists.
- Applying our change-point algorithm, we estimate a change-point occurs at point 79.

Iteration 2

- Bayes factor output from R for segments $\{x_i\}_{i=1}^{78}$ and $\{x_i\}_{i=80}^{190}$ are $BF = 1.7$ and $BF = 1.2$, respectively. These values are both below our

threshold value so we conclude no change-point exists in either segment.

Terminate algorithm.

Conclude (correctly) one change-point exists and is located at time point 79.

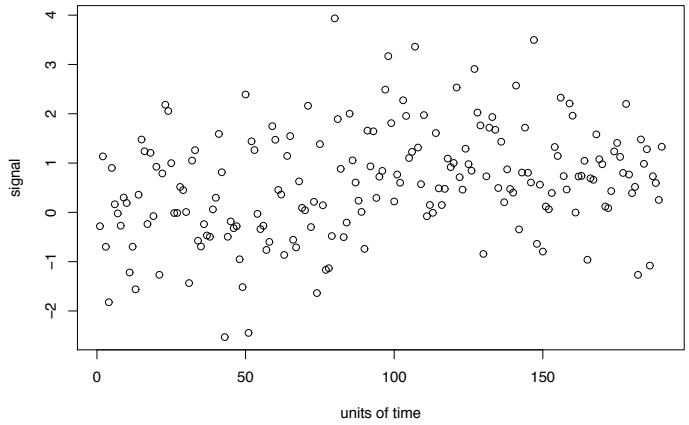

Figure 3.5: Example time series with one change-point in mean at point 79

Example 3: three change-point exists

Figure 3.6 is an example of a time series with three change-points at time points 37, 93, and 201. Applying the binary segmentation algorithm with Bayesian-wavelet method, we obtain the following results for the time series.

Iteration 1

- Bayes factor output from R for entire series $BF = 46$ which per section 3.2.3 indicates a change-point exists.
- Applying our change-point algorithm, we estimate a change-point occurs at time point 201.

Iteration 2

- Bayes factor output from R for segments $\{x_i\}_{i=1}^{196}$ and $\{x_i\}_{i=206}^{256}$ are $BF = 8.6$ and $BF = 1.5$, respectively. The first value indicates a change-point in $\{x_i\}_{i=1}^{195}$ while the second value indicates no change-point in $\{x_i\}_{i=205}^{256}$.
- Applying our change-point algorithm to $\{x_i\}_{i=1}^{196}$, we estimate a change-point occurs at point 39.

Iteration 3

- Bayes factor output from R for segments $\{x_i\}_{i=1}^{35}$ and $\{x_i\}_{i=44}^{195}$ are $BF = 1.7$ and $BF = 4.68$, respectively. The first value indicates no change-point in $\{x_i\}_{i=1}^{35}$ while the second value indicates a change-point in $\{x_i\}_{i=45}^{195}$.
- Applying our change-point algorithm to $\{x_i\}_{i=45}^{196}$, we estimate a change-point occurs at time point 94.

Iteration 4

- Bayes factor output from R for segments $\{x_i\}_{i=45}^{89}$ and $\{x_i\}_{i=99}^{196}$ are $BF = 1.95$ and $BF = 1.75$, respectively. Both values fall short of indicating change-points exist in their respective segments.

Terminate algorithm.

Conclude (almost correctly) three change-point exists and are located at time points 39, 94, and 201.

The above procedure works well provided two conditions are met. Firstly, as has been previously shown, when the signal-to-ratio (SNR) is not sufficiently

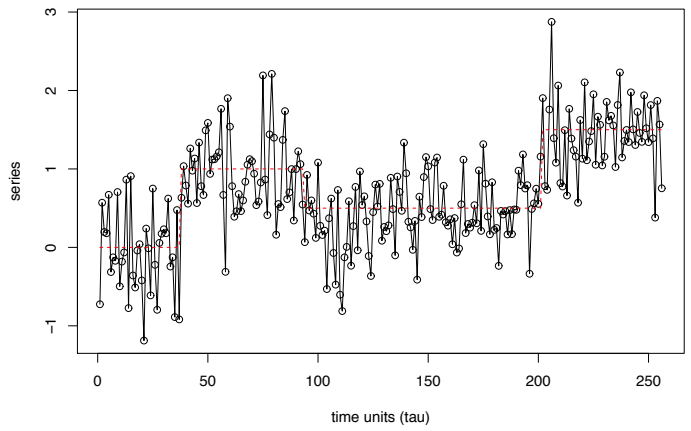

Figure 3.6: Time series with change-points at point 37 (shift 1), 93 (shift -.5), and 201 (shift 1).

high both the inference and estimation methods will fail. This generally holds for any procedural method (i.e. MLE or pure Bayesian) as any method will eventually breakdown at a some level of process variance. The previous chapter provides experimentally obtained SNR threshold values for when we should and should not expect good results from the Bayesian-wavelet estimation procedure.

A second condition that must be met is that the change-points cannot be "too close" together. A drawback to using wavelets as applied to the binary segmentation algorithm involves the issue of requiring a dyadic length time series before we can apply the discrete wavelet transform (DWT). Previously we overcame this problem by padding the beginning of the series with some low white noise. This method fails when applied above because we cannot assume that the initial points of each binary segment have mean zero. In fact, clearly, this will not generally be the case. One way to get around this problem is to estimate the segment mean function by some method from a sample of the observed data. With the mean function in hand, we then generate random noise around this estimated mean of the required length to pad against the segment. In such a way we achieve a dyadic length time series for each segment.

While all the change-points were randomly selected in the above examples, they were still contrived in such a way as to be rather far apart. For the amount of variance in the above examples, we found about 20 samples were required to obtain a sufficiently close mean approximation. The method consistently worked very well when there were 30 or more time units between change-points. When we have too few samples to estimate the mean, we found the estimate was often far enough off the actual mean so as to cause the algorithm to falsely identify another change-point where we inserted the padding.

3.4.2 A direct Bayesian-wavelet approach for time series with multiple change-points

The binary segmentation approach iteratively applied the Bayesian-wavelet change-point estimation method originally constructed from a single change-point model. The need to pad segments of the time series to meet the dyadic time series length requirement, however, limits the effectiveness of the binary segmentation algorithm when we cannot assume the time series change-points are sufficiently far apart. We address this short coming in this section by developing an alternative method designed with multiple change-points in mind from the onset. Deciding between various multiple change-point models can then be done with a model selection approach such as SIC. We explicitly construct the case for two change-points from which utilizing similar methodology can be extended to construct models with an arbitrary number of change-points.

Recall the construction of the change-point model began with the assumption that the wavelet details could be modeled as approximately distributed as

$$d_{j,k}|(\tau\ ,\Delta\ ,\sigma^2) \sim N(\Delta q_{j,k}(\tau)\,,\,\sigma^2)$$

where Δ is the unknown amount of shift at the change-point and q_{ij} are the wavelet details of the idealized mean function. Now suppose we have two change-points in the time series. We make the analogous assumption that our wavelet details can be modeled as

$$d_{j,k}|(\tau_1, \tau_2\ , \Delta_1, \Delta_2, \sigma^2) \sim N(\Delta_1 q_{1,j,k}(\tau) + \Delta_2 q_{2,j,k}(\tau)\,, \sigma^2).$$

A time series with a single change-point is generated from a stochastic process centered around a mean function with a single shift. The motivation behind this new method is that now the mean function containing two shifts can be equivalently expressed as the sum of two mean function each with one shift. The $q_{1,j,k}$'s and $q_{2,j,k}$'s represent the wavelet details from these two mean functions. Exploiting the linearity of the DWT, applying the same noninformative prior, and applying Bayes rule we now express the posterior distribution function as

$$p(\tau_1, \tau_2, \Delta_1, \Delta_2, \sigma^2|d^*) = K(2\pi)^{-\frac{M}{2}} \prod_i \prod_j p(d_{ij}|\tau_1, \tau_2, \Delta_1, \Delta_2, \sigma^2) p_0(\tau_1, \tau_2, \Delta_1, \Delta_2, \sigma^2)$$

where K is some constant of proportionality and m is the number of actual detail coefficients used in the analysis. As before we marginalize the posterior distribution by integrating out now in this case Δ_1, Δ_2, and σ^2. Now applying the definition of our posterior distrubtion we present the following detailed calculations

$$
\begin{aligned}
p(\tau_1,\tau_2,\Delta_1,\Delta_2,\sigma^2|d^*) &= K(2\pi)^{-\frac{m}{2}}\int_0^\infty\int_{-\infty}^\infty\int_{-\infty}^\infty \\
&\qquad \sigma^{-\frac{m}{2}}\mathrm{e}^{(\frac{-\sum\sum(d_{jk}-\Delta_1 q_{1,ij}-\Delta_2 q_{2,ij})^2}{2\sigma^2})}\sigma^{-1}d\Delta_1 d\Delta_2 d\sigma^2 \\
&= K(2\pi)^{-\frac{m-2}{2}}\int_0^\infty\int_{-\infty}^\infty\int_{-\infty}^\infty \\
&\qquad \sigma^{-\frac{m+1}{2}}\mathrm{e}^{\frac{-1}{\sigma^2}(C_1(\Delta_1-\frac{B_1-\Delta_2C_3}{C_1})^2+\frac{\Delta_2^2(C_2C_1-C_3^2)-2\Delta_2C_3B_1+B_1^2}{C_1}-2\Delta_2B_2+A)}d\Delta_1 d\Delta_2 d\sigma^2 \\
&= K(2\pi)^{-\frac{m}{2}}C_1^{-1/2}\int_0^\infty\int_{-\infty}^\infty \\
&\qquad \sigma^{-\frac{m}{2}}\mathrm{e}^{\frac{-1}{\sigma^2}(D(\Delta_2-\frac{B_2C_1C_3B_1}{DC_1})^2+A-\frac{B_1^2}{C_1}-\frac{(B_2C_1-B_3B_1)^2}{DC_1^2}}d\Delta_2 d\sigma^2 \\
&= K(2\pi)^{-\frac{m-1}{2}}(C_1C_2-C_3^2)^{-1/2}\int_0^\infty\int_{-\infty}^\infty \\
&\qquad \sigma^{-\frac{m-1}{2}}\mathrm{e}^{\frac{-1}{\sigma^2}(A-\frac{(B_2^2C_1+B_1^2C_2-2B_1B_2C_3)^2}{C_1C_2-C_3^2})}d\Delta_2 d\sigma^2 \\
&= K(2\pi)^{-\frac{m-1}{2}}(C_1C_2-C_3^2)^{-1/2}\Gamma(\frac{m-3}{2})(A-\tfrac{(B_2^2C_1+B_1^2C_2-2B_1B_2C_3)^2}{C_1C_2-C_3^2})^{-\frac{m-3}{2}}.
\end{aligned}
\tag{3.8}
$$

where

$$
A=\sum_{j\geq j_0}\sum_k d_{jk}^2,\ B_1=\sum_{j\geq j_0}\sum_k d_{jk}q_{1,jk}(\tau),\ B_2=\sum_{j\geq j_0}\sum_k d_{jk}q_{2,jk}(\tau),
$$
$$
C_1=\sum_{j\geq j_0}\sum_k q_{1,jk}^2,\ C_2=\sum_{j\geq j_0}\sum_k q_{2,jk}^2,\ C_3=\sum_{j\geq j_0}\sum_k q_{1,jk}q_{2,jk},\ D=\tfrac{C_2C_1-C_3^2}{C_1}.
$$

Given a time series with two change-points, we then estimate the location of these two change-points by computing $\arg\max_{\tau_1,\tau_2} p(\tau_1,\tau_2,\Delta_1,\Delta_2,\sigma^2|d^*)$. For example, consider Figure 3.7 where we generate a time series with two change-points. We indicate the mean function of this time series with the dashed red line. Applying, equation (3.8) we find change-points at time points 45 and 85 maximize the probability. Plotting the posterior distrubtion in Figure 3.8 we find a 95% credible region is given by $(42,46)\times(82,87)$.

The direct approach of estimating the location of multiple change-points in

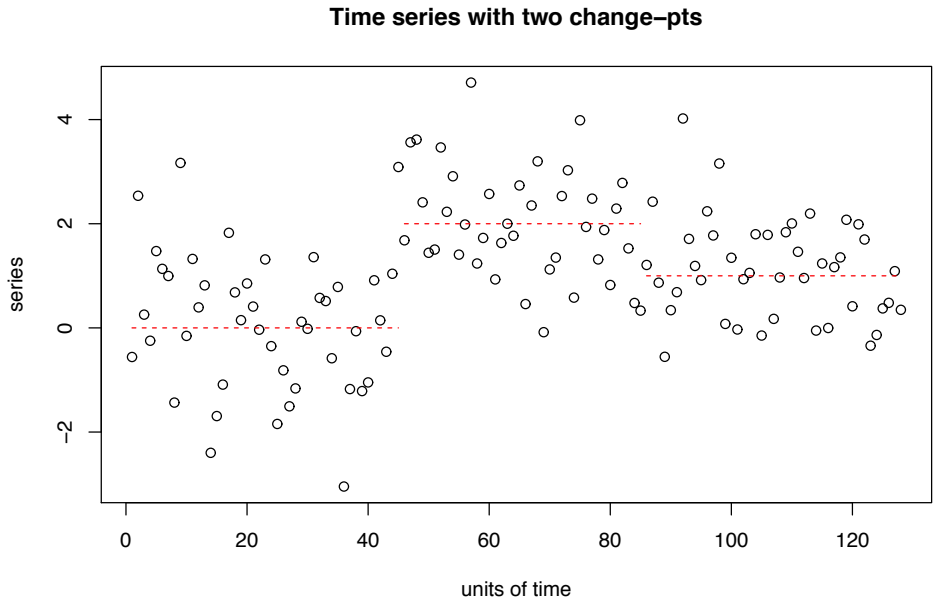

Figure 3.7: Time series with two change-points at time point 45 ($\Delta_1 = 2$) and time point 85 ($\Delta_2 = -1$).

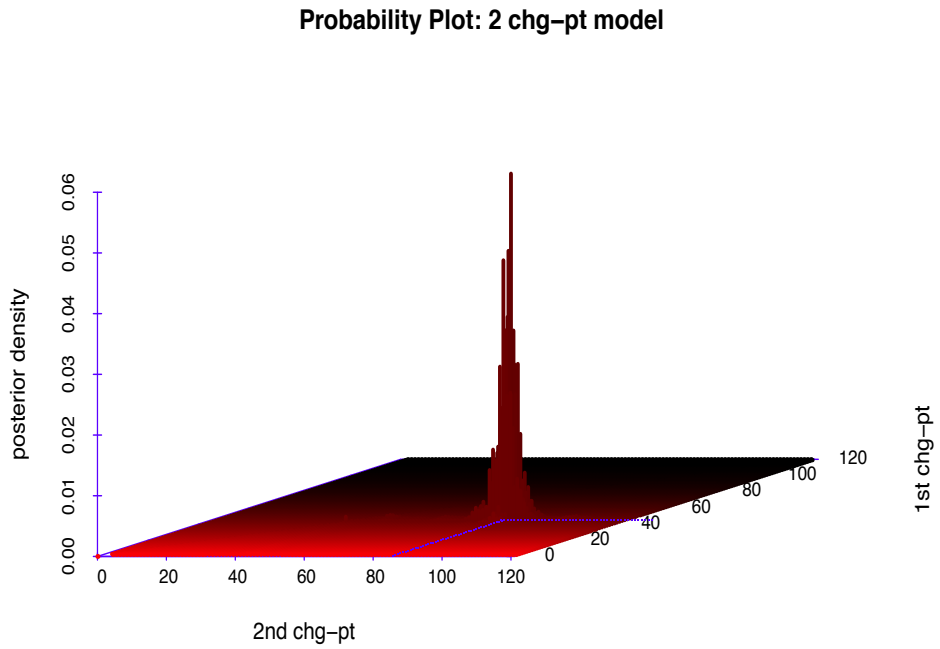

Figure 3.8: Posterior distribution of location of two change-point for the time series above.

this section has the clear advantage of not requiring padding for successive change-point estimations. The only limitation for estimating change-points in close proximity to each other is the SNR of the time series. Additionally, this method easily adapts to smoothly varying mean functions as well whereas the binary segmentation approach again runs into difficulties in this case because of the difficulty to pad for nonconstant mean functions. When deciding between models with 1 or 2 or more change-points an information criteria such as the Schwarz Information Criteria (SIC) can be used. One drawback to this direct method is the need to rederive a change-point model for the number of change-points in the time series. Analytically, these derivations become increasingly burdensome as the number of change-points increase. Secondly, computational difficulties also begin to arise in the presence of many change-points since the algorithm must compute all combinations of possible change-point locations before returning the change-point location combination with the highest probability.

3.5 Estimating the variance change-point location in a time series

In this section, we once again assume we observe some stochastic process $\{x_i\}_{i=1}^{N}$ for $N \in \mathbb{N}$. Here, we propose a method of analyzing the change-point of variance problem of a time series by transforming the series into the wavelet domain. While various maximum likelihood (MLE) and Bayesian methods exist for estimating the location of a change-point in variance, they typically require strict assumptions. These assumptions, however, limit their applicability to more general situations. A more indirect approach that potentially applies to more

general situations involves analyzing the time series after wavelet transformation. While some precision in change-point estimates may be lost, what is gained is an applicability in more general contexts.

In this section we derive in detail a Bayesian-wavelet approach to estimating the variance change-point location of a univariate time series. We then compare the effectiveness of this method against known MLE and Bayesian approaches when the underlying mean function is both a step function and then a smoothly varying mean function except a single point where the change-point occurs. While we omit the details of the derivation of the MLE and Bayesian methods, the formulation of these approaches are analogous to those details provided in chapter 2 for an estimation in the change of a mean function. Readers interested in details of these derivations are referred to [40, 41, 42, 43].

3.5.1 Bayesian-wavelet change-point in variance estimation

Given some observed time series, $\{x_i\}_{i=1}^{N}$ for $N \in \mathbb{N}$, we assume our data is generated from the following model:

$$x_i = g(i) + \varepsilon_i$$

where $g(\cdot)$ is some smoothly varying mean function. Furthermore, we assume the additive noise component to possess the following distributional properties:

$$\{\varepsilon_i\} \sim N(\mu, {\sigma_1}^2) \text{ for } 1 \leq i < \tau$$
$$\{\varepsilon_i\} \sim N(\mu, {\sigma_2}^2) \text{ for } \tau \leq i \leq N$$

Next we apply some discrete wavelet transform to our time series and recover $N-1$ detail coefficients of our original time series now in the wavelet domain. We express the elements of our transformed series in the following form:

$$d_i^* = d_i + \epsilon_i$$

where d_i^* is the empirical wavelet detail coefficient and d_i is the true detail coefficient. In the case of a constant mean function we expect the d_i's are approximately distributed around 0. In the case of a smoothly varying mean function, we expect the higher level detail coefficients also to be approximately distributed around zero. The ϵ_i component is the transformed additive noise component that transforms again to noise. With these assumptions in mind, each detail coefficients used in the model is described by one of the following two models:

$$d_{j,k}|(\sigma^2) \sim N(0\ ,{\sigma_1}^2)$$

or

$$d_{j,k}|(\sigma^2) \sim N(0\ ,{\sigma_2}^2).$$

We should point out that in the case of a nonconstant mean function, the lower level detail coefficients will not be in general be centered around 0. In this case we discard the lower level detail coefficients in the model and use only the high level coefficients.

Recall that each wavelet detail is produced by applying the DWT to certain elements of the original time series. For example, if we consider a simple 8 element time series $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8)$ and apply the Haar DWT we obtain 7 wavelet details $\mathbf{d} = (d_{1,1}, d_{1,2}, d_{1,3}, d_{1,4}, d_{2,1}, d_{2,2}, d_{3,1})$. Now only x_1 and x_2 are used to produce $d_{1,1}$, but x_5, x_6, x_7, x_8 are all used to produce $d_{2,2}$. From the context of our current change-point analysis this poses a problem because the change-point is somehow mixed up between all of our wavelet details. To address this problem, we define a function γ such that:

$$\gamma : \{d_{j,k}\} \to \mathbb{N}$$
$$d_{j,k} \mapsto J$$

where J indicates the average of all indices of the original time series that contributes to the value of wavelet detail rounded down to the next integer. For example, in the above time series $\gamma(d_{1,3}) = 5$. Applying, γ to the entire detail vector above, our sorted detail vector now becomes

$$\gamma(\mathbf{d}) = (d_{1,1}, d_{2,1}, d_{1,2}, d_{3,1}, d_{1,3}, d_{2,2}, d_{1,4}).$$

We may now more precisely state our model as:

$$d_{j,k}|(\tau, \sigma_1^2) \sim N(0, \sigma_1^2)\ 1 \leq \gamma(d_{j,k}) \leq \tau$$
$$d_{j,k}|(\tau, \sigma_2^2) \sim N(0, \sigma_2^2)\ 1 < \gamma(d_{j,k}) \leq N.$$

where τ represents the true but unknown change-point in variance of the time series. Letting $\mathbf{d}$ represent our vector of the wavelet details, we may write the likelihood function as:

$$p(\mathbf{d}|\tau, \sigma_1, \sigma_2) = (2\pi\sigma_1^2)^{-\tau/2} \exp\left[-\frac{1}{2\sigma_1^2}\sum_{i=1}^{\tau} d_{i,j,k}^2\right](2\pi\sigma_2^2)^{-(n-\tau)/2} \exp\left[-\frac{1}{2\sigma_2^2}\sum_{i=\tau+1}^{N} d_{i,j,k}^2\right]$$

where $i = \gamma(d_{j,k})$. Next, we need to place a prior distribution, $p_0(\tau, \sigma_1, \sigma_2)$, on our unknown parameter set $\{\tau, \sigma_1, \sigma_2\}$. Here, the modeler's subjectivity comes into play depending on experience, past known results, or other case specific factors. Often, our goal is to strike a balance between realistic assumptions and mathematical convenience. For the purposes of this analysis, we will assume no prior knowledge on the change-point location τ, but that each variance component follows a conjugate prior distribution of the inverse gamma distribution. Furthermore, we make the assumption that each of these unknowns

are independent from each other. Thus, we write our prior distribution as:

$$\begin{aligned} p_0(\tau, \sigma_1^2, \sigma_2^2 | \alpha_1, \beta_1, \alpha_2, \beta_2) &= p_0(\tau) p_0(\sigma_1 | \alpha_1, \beta_1) p_0(\sigma_2 | \alpha_2, \beta_2) \\ &= \frac{1}{N-1} \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)} (\sigma_1^2)^{-\alpha_1 - 1} e^{-\beta_1/\sigma_1^2} \frac{\beta_2^{\alpha_2}}{\Gamma(\alpha_2)} (\sigma_2^2)^{-\alpha_2 - 1} e^{-\beta_2/\sigma_2^2} \end{aligned}$$

where $\alpha_1, \beta_1, \alpha_2$ and β_2 are hyperparameters of the particular IG distribution chosen. From Bayes theorem we may express our posterior distribution as a proportionality in the form:

$$p(\tau, \sigma_1^2, \sigma_2^2 | \mathbf{d}, \alpha_1, \beta_1, \alpha_2, \beta_2) \propto p(\mathbf{d} | \tau, \sigma_1, \sigma_2) p_0(\tau, \sigma_1^2, \sigma_2^2 | \alpha_1, \beta_1, \alpha_2, \beta_2). \tag{3.9}$$

Dropping multiplicative constant terms, substituting, and minor rearranging gives our proportional posterior distribution:

$$\begin{aligned} p(\tau, \sigma_1^2, \sigma_2^2 | \mathbf{d}, \alpha_1, \beta_1, \alpha_2, \beta_2) \propto & (\sigma_1^2)^{-\frac{\tau}{2} - \alpha_1 - 1} (\sigma_2^2)^{\frac{(n-\tau)}{2} - \alpha_2 - 1} \\ & \exp\left[- \frac{\frac{1}{2} \sum_{i=1}^{\tau} d_{i,j,k}^2 - \beta_1}{\sigma_1^2} \right] \exp\left[- \frac{\frac{1}{2} \sum_{i=\tau+1}^{N} d_{i,j,k}^2 - \beta_2}{\sigma_1^2} \right]. \end{aligned}$$

Since we are interested in determining τ, we would like to obtain a form of the marginal proportional posterior distribution by integrating out σ_1 and σ_2. That

is, we perform the following calculation:

$$p(\tau|\mathbf{d},\alpha_1,\beta_1,\alpha_2,\beta_2) \propto \int_0^\infty \int_0^\infty (\sigma_1^2)^{-\frac{\tau}{2}-\alpha_1-1} \exp\left[-\frac{\frac{1}{2}\sum_{i=1}^{\tau} d_{i,j,k}^2 - \beta_1}{\sigma_1^2}\right] (\sigma_2^2)^{-\frac{(n-\tau)}{2}-\alpha_2-1} \exp\left[-\frac{\frac{1}{2}\sum_{i=\tau+1}^{N} d_{i,j,k}^2 - \beta_2}{\sigma_2^2}\right] d\sigma_2^2 d\sigma_1^2.$$

We notice this is in the form of the inverse gamma distribution which has a known result. Therefore, we write our final expression as:

$$p(\tau|\mathbf{w},\alpha_1,\beta_1,\alpha_2,\beta_2) \propto \Gamma(\frac{\tau}{2}+\alpha_1)\Gamma(\frac{(n-\tau)}{2}+\alpha_2) \left[-\frac{1}{2}\sum_{i=1}^{\tau} d_{i,j,k}^2 - \beta_1\right]^{-(\frac{\tau}{2}+\alpha_1)} \left[-\frac{1}{2}\sum_{i=\tau+1}^{N} d_{i,j,k}^2 - \beta_2\right]^{-\frac{(n-\tau)}{2}-\alpha_2}$$

As a final point we provide a note about the hyperparameters $\alpha_1, \beta_1, \alpha_2, \beta_2$. Firstly, should we not want bias our result in anyway, we could choose 0 as the value for all our hyperparamters. This would correspond to a "flat" gamma distribution or an uninformative prior distribution. Typically, however, we either have data in hand that we are trying to analyze and already have some notion of the variance of the process or at least a plausible distribution from which the variance may be drawn from. In our simulation below we set $\alpha_1 = \alpha_2 = 2$ and $\beta_1 = \beta_2 = 1$. This roughly corresponds to the prior belief that the initial standard deviation of the process is drawn from a gamma distribution with mean of 1, but no additional assumptions are made as to how the standard deviation changes.

3.6 Simulation results

With our main result derived above, we would like to compare this approach to other known methods for detecting a shift in variance. We compare the above Bayesian-wavelet approach against both a classical MLE method and a purely Bayesian approach with similar hyperparameter values. Figure 3.9 gives the reader some intuition for the problem at hand by illustrating a typical time series in which we wish to estimate the variance change-point location.

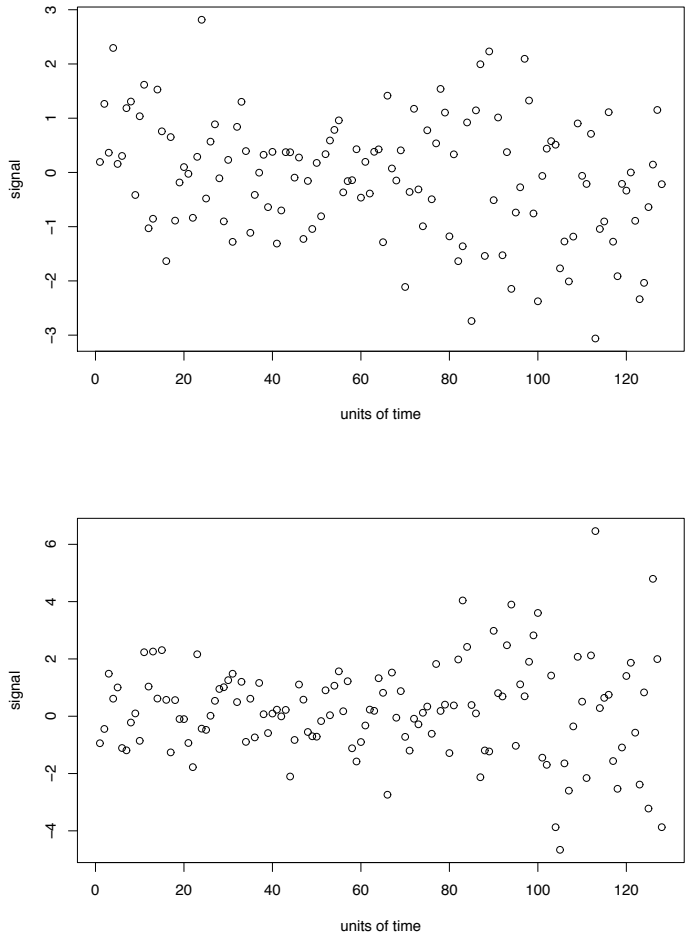

Figure 3.9: Two example time series with standard deviation shifts from 1 to 1.5 (left) and 1 to 2.5 (right) at time point 80.

Table 3.4 below provides results from 4 separate simulations. For both the MLE and Bayesian methods, we define a "success" as the method correctly estimating within two time units the actual change-point location. For the Bayesian-wavelet method we use two different criteria to measure success. One criteria is the same as above where a success is correctly estimating within two time units the actual change-point. For the second criteria we count a success as correctly estimating within three time units the actual change-point. We note, that the wavelet transform partially smooths the original time series resulting in some loss of fidelity in the data's structure. By comparing the Bayesian-wavelet against the two known methods with a more relaxed success criteria, we find the same information is retained under the transform only now in a somewhat more diffuse representation. We conclude that while the MLE and pure Bayesian methods perform almost identically as seen by table 3.4, the Bayesian-wavelet method performs somewhat worse or at least less precise.

Table 3.4: Results of 1000 simulations for a noisy constant function

Variance Shift	MLE	Bayesian	Bayesian-wavelet
.5	331	295	306
1	602	595	584
1.5	785	780	753
2	848	849	824

Number of successful change-point detections (i.e. estimated change-point is within three time units of actual change-point).

While the Bayesian-wavelet method might lack some precision, it may be directly applied in more general cases. Consider, a noisy sine wave such as in figure 3.10 that also undergoes a variance shift at some unknown unit of time. The DWT of such a function retains approximately constant standard deviation for the highest wavelet detail coefficients. By applying the Bayesian change-point technique directly to these level of detail coefficients, table 3.5 shows we are fairly

successful at detecting shifts in variance provided the shift is sufficiently large.

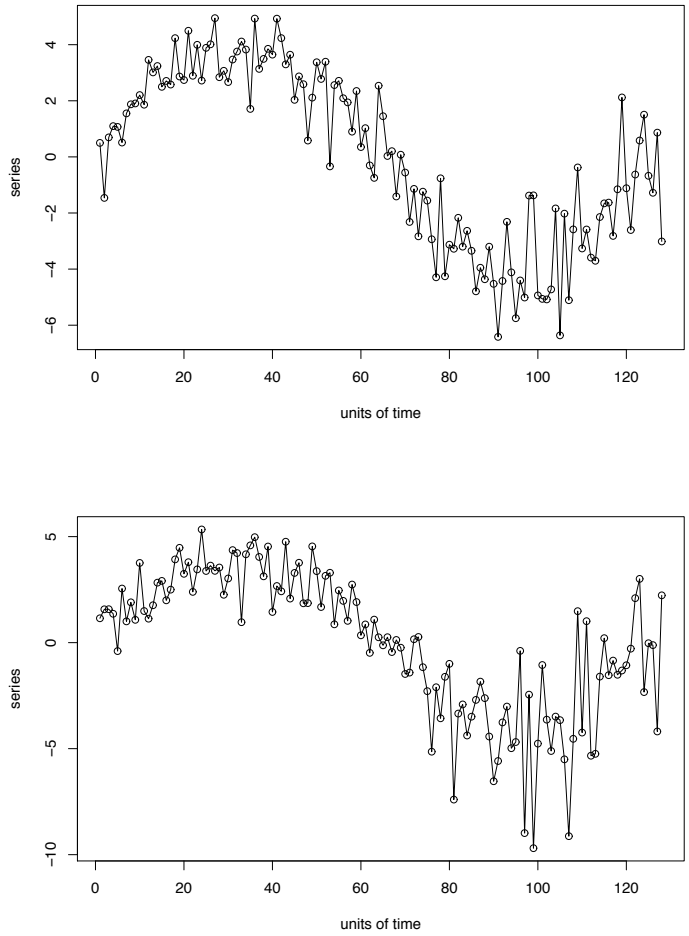

Figure 3.10: Two example noisy sine waves with standard deviation shifts from 1 to 1.5 (left) and 1 to 2.5 (right) at time point 80.

3.7 Conclusion

In this chapter we extended the formulation of the Bayesian-wavelet change-point estimation equation from chapter 1 in several ways. Beginning with a derivation of models both with and without a change-point, we obtained a

Table 3.5: Results of 1,000 simulations for a change-point in variance where the underlying mean function is a sine wave

Variance Shift	MLE	Bayesian	Bayesian-wavelet
.5	22	17	285
1	12	23	557
1.5	42	15	727
2	90	63	795

Number of successful change-point detections (i.e. estimated change-point is within 3 time units of actual change-point).

Bayes factor formula providing us with a model selection mechanism. This formulation allowed us to apply the binary segmentation algorithm to the multiple change-point problem. The binary segmentation algorithm approach to the multiple change-point problem, however, was not with some shortcomings. To address these issues, we developed an alternative method to the multiple change-point problem that incorporated a multiple change-point assumption in the time series from the onset. Finally, we developed a method to detect a statistical change-point to the variance of a time series. Through these various approaches, we compared our method against known methods through simulation experiments. We found that in most cases the Bayesian-wavelet method performed comparably to the MLE and pure Bayesian methods when the mean function was as step function, but that the Bayesian-wavelet method outperformed these other methods in more general cases when the mean function smoothly changed.

Chapter 4

The Multivariate Case

4.1 Introduction

Now we turn to the multivariate case of the change-point problem. We propose a semiparametric approach to estimate the existence and location of a statistical change-point to a nonlinear multivariate time series contaminated with an additive noise component. In particular, we consider a $p-$dimensional stochastic process of independent observations $\{\mathbf{x}_i\}_i^N$ for $N \in \mathbb{N}$ such that $\mathbf{x}_i \sim N_p(\boldsymbol{\mu}_i, \Sigma)$ where $\boldsymbol{\mu}_i$ smoothly varies except at a single change-point. Our approach involves conducting a Bayesian analysis on the empirical detail coefficients of the original time series after a wavelet transform. If the mean function of our time series may be expressed as a multivariate step function, we find our Bayesian-wavelet method performs comparably with classical parametric methods such as maximum likelihood (MLE). The advantage of our multivariate change-point method is seen in how it applies to a much larger class of functions that require only general smoothness assumptions.

4.2 Background

In some sense the multivariate change-point problem could be considered a less difficult change-point problem than in the univariate case. In the multivariate setting we have more information in the form of extra dimensions to include in our analysis than when working with just a single dimension. Without the presence of covariance in our time series, we could simply analyze each component of the time series separately with univariate techniques. This method has the drawback that if estimation discrepancies of the change-point location exist between the various dimensions, then some methodology must still be used after the initial analysis for a final estimation of the change-point location. Of course another difficulty with this method is that often we encounter covariance in the multivariate setting [44]. For these reasons and others we need a method that simultaneously evaluates across dimensions for the change-point location and takes into account any covariance present.

There appears to be a gap in the change-point literature that addresses the change-point problem for nonlinear multivariate time series. Classical parametric approaches such as maximum likelihood (MLE) and Bayesian methods exist to detect and estimate the location of one or more statistical change-points in multivariate time series [45]. Many variations of such parametric approaches exist for detecting multivariate statistical change-points [37, 46, 44, 47], but invariably these methods require strict assumptions on the time series mean function. Müller [48] developed an approach to detect discontinuties in derivative of smooth function using left and right one sided kernel smoothers for one dimensional functions. More recently Ciuperca [49] and Battaglia [50] have results for detecting change-points in nonlinear time series. As with other methods in the nonlinear change-point problem, however, both Ciuperca's and Battaglia's

approaches apply in only in the one dimensional case. Matteson [51] developed a fully nonparametric approach for estimating the location of multiple change-point in a multivariate data. While their work is perhaps the method most closely applying to the change-point problem in this article, their method still only applies to data sets where the mean function is piecewise constant.

The multivariate change-point problem is an important problem that has direct applications in a surprising number of otherwise seemingly unrelated fields. In statistical process control (SPC), the multivariate change-point problem is important to quickly detect and estimate changes in many industrial processes [52]. The Department of Transportation has applied the multivariate change-point problem to estimate statistical change-points around a speed limit increase from 55 mph to 65 mph [37]. Additional applications occur in such unrelated fields as biosurveillance, financial market analysis, and hydrology to name a few [53, 54]. When encountering the change-point problem for real world multivariate data often imposing strict assumptions on the time series may be impractical. Unfortunately, in the multivariate time series setting there have not been many other good options. In this article we propose a method that attempts to bridge this gap by developing a multivariate change-point detection and estimation method that applies when strict assumptions about the true underlying mean function cannot be made.

4.3 Description of methods

4.3.1 Maximum likelihood approach

Classical likelihood methods remain one of the primary workhorses in modern day statistics and also apply to the multivariate change-point problem.

In this section, we present the important elements of the standard multivariate change-point MLE approach as outlined by Horváth [46] and Chen [37] to serve as a reference for when we compare the likelihood method and the Bayesian-wavelet method in section 4.6.

Consider a multivariate time series of independent elements $\{\mathbf{x}_i\}_i^N$ for $N \in \mathbb{N}$ such that $\mathbf{x}_i \sim N_p(\boldsymbol{\mu}_i, \Sigma)$. We wish to compare the null hypothesis

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_N$$

against the alternative hypothesis

$$H_a : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_\tau \neq \boldsymbol{\mu}_{\tau+1} = \cdots = \boldsymbol{\mu}_N$$

where $\boldsymbol{\mu}_i \in \mathbb{R}^p$, $p \in \mathbb{N}$. Under the two hypotheses above, we form two separate likelihood functions L_0 and L_a for the null and alternative hypotheses respectively:

$$L_0(\mathbf{x}_1, .., \mathbf{x}_N | \boldsymbol{\mu}, \Sigma) = (2\pi)^{-Np/2} |\Sigma|^{-N/2} \exp\left[-\frac{1}{2} \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right]$$

and

$$L_a(\mathbf{x}_1, .., \mathbf{x}_N | \boldsymbol{\mu}_\tau, \boldsymbol{\mu}'_\tau, \Sigma) = (2\pi)^{-Np/2} |\Sigma|^{-N/2}$$
$$\exp\left[-\frac{1}{2} \Big(\sum_{i=1}^{\tau} (\mathbf{x}_i - \boldsymbol{\mu}_\tau)^T \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_\tau) + \sum_{i=1+\tau}^{N} (\mathbf{x}_i - \boldsymbol{\mu}'_\tau)^T \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}'_\tau) \Big) \right].$$

We then apply the usual MLE's for $\boldsymbol{\mu}$, $\boldsymbol{\mu}_\tau$, $\boldsymbol{\mu}'_\tau$, and Σ namely

$$\hat{\boldsymbol{\mu}} = \tfrac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i, \quad \hat{\boldsymbol{\mu}}_\tau = \tfrac{1}{\tau} \sum_{i=1}^{\tau} \mathbf{x}_i, \quad \hat{\boldsymbol{\mu}}'_\tau = \tfrac{1}{N-\tau} \sum_{i=\tau+1}^{N} \mathbf{x}_i,$$
$$\widehat{\Sigma}_0 = \tfrac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T,$$

and

$$\widehat{\Sigma}_a = \tfrac{1}{N}\Big(\sum_{i=1}^{\tau}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_\tau)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_\tau)^T + \sum_{i=\tau+1}^{N}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}'_\tau)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}'_\tau)^T\Big)$$

where $\widehat{\Sigma}_0$ and $\widehat{\Sigma}_a$ denote our Σ estimates under H_0 and H_a, respectively. In the H_0 case, if we regard the scalar term in the exponent as a 1×1 matrix, then we can use the identity $tr(MN) = tr(NM)$, where M and N are matrices, to equate

$$\begin{aligned}
\sum_{i=1}^{N}\Big(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T\widehat{\Sigma}_0^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})\Big) &= \sum_{i=1}^{N} tr\Big((\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T\widehat{\Sigma}_0^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})\Big) \\
&= \sum_{i=1}^{N} tr\Big(\widehat{\Sigma}_0^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T\Big) \\
&= tr\Big(NI_p\Big) \\
&= Np.
\end{aligned}$$

Thus,

$$L_0(\mathbf{x}_1, \ldots, \mathbf{x}_k|\hat{\boldsymbol{\mu}}, \widehat{\Sigma}_0) = (2\pi)^{-Np/2}|\widehat{\Sigma}_0|^{-N/2}\exp(-\tfrac{1}{2}Np).$$

Similarly for the case of the alternative hypothesis:

$$L_a(\mathbf{x}_1, .., \mathbf{x}_k|\hat{\boldsymbol{\mu}}_\tau, \hat{\boldsymbol{\mu}}'_\tau, \widehat{\Sigma}_a) = (2\pi)^{-Np/2}|\widehat{\Sigma}_a|^{-N/2}\exp(-\tfrac{1}{2}Np).$$

The value of τ that maximizes the likelihood function also maximizes Hotelling's T^2 test statistic defined as

$$T^2_\tau = \frac{\tau(N-\tau)}{N}(\hat{\boldsymbol{\mu}}_\tau - \hat{\boldsymbol{\mu}}'_\tau)^T W^{-1}(\hat{\boldsymbol{\mu}}_\tau - \hat{\boldsymbol{\mu}}'_\tau)$$

where W represents the pooled sample covariance matrix [37] and is defined as

$$W = \tfrac{1}{N-2}\Big(\sum_{i=1}^{\tau}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_\tau)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_\tau)^T + \sum_{i=\tau+1}^{N}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}'_\tau)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}'_\tau)^T\Big).$$

We would then reject H_0 in favor of H_a when $T_\tau^2 > p$, where the p-value is some constant corresponding to a desired confidence level. Determining the value of p and the associated probability level requires a derivation of the distribution of $\arg\max_\tau T_\tau^2$ under H_0 as presented by Worsley [36]. For now, we are assuming a change-point in the mean vector has occurred. Then based on the above analysis, we estimate this change-point occurs at time

$$\arg\max_\tau \frac{\tau(N-\tau)}{N}(\hat{\boldsymbol{\mu}}_\tau - \hat{\boldsymbol{\mu}}'_\tau)^T W^{-1}(\hat{\boldsymbol{\mu}}_\tau - \hat{\boldsymbol{\mu}}'_\tau).$$

This formulation will be what we implement in the simulations to follow when we compare the MLE method against the Bayesian-wavelet method.

4.3.2 Bayesian-wavelet approach

The DWT allows us to analyze a time series at varying resolution levels and stores the resulting details of smooth functions in a similar way. Observe Figure 4.1 which displays two examples of a smooth time series mean function except at a change-point at time point 81 (top) along with the respective detail coefficient values (bottom). Notice that the detail coefficient values are essentially identical for the finest three resolution levels despite the fact that the mean functions are quite different. Even resolution levels 4 closely compare. While the coefficient values at the lowest three resolution levels do begin to diverge, 118 of the total 127 detail coefficients in this 128 element time series closely agree. The phenomena in Figure 4.1 illustrates the sparsity property of the DWT and holds in general for any smoothly varying mean functions which share a common change-point.

From Chapter 2 we know any additive noise component of a time series is again transformed to an additive noise component after a DWT. Ogden [25]

exploited these properties of the DWT by proposing a method for estimating the change-point location of a one dimensional time series by applying Bayesian techniques in the wavelet domain. We generalize a similar methodology to now an arbitrary dimensional time series and extend the approach to answer the inference question. We consider a multi-dimensional time series of independent observations $\{\mathbf{x}_i\}_i^N$ for $N \in \mathbb{N}$ where $\mathbf{x}_i$ is a p-dimensional vector, such that

$$\mathbf{x}_i \sim N_p(\boldsymbol{\mu}_i, \Sigma). \tag{4.1}$$

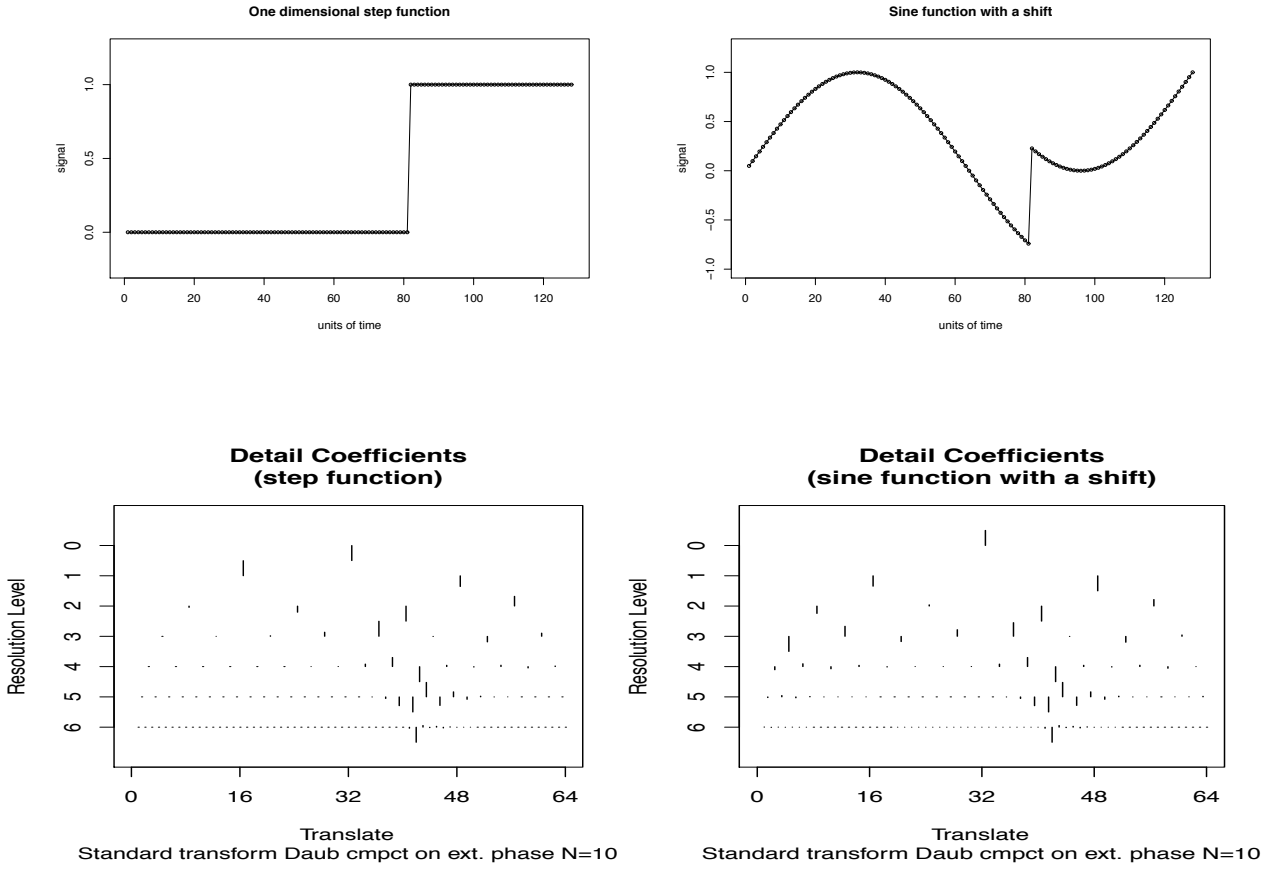

Figure 4.1: Two example mean functions with a change-point at time point 81 (top) along with their respective detail coefficients (bottom). Each detail level is normalized by its l^∞-norm. Notice at the finest three resolution levels the detail coefficients are essentially identical to each other.

In the typical case where Bayesian or MLE techniques are applied to the

multivariate change-point problem, $\boldsymbol{\mu}_i$ is assumed to be a p-dimensional step function. For our more general analysis, however, $\boldsymbol{\mu}_i$ is assumed to be generated by a p-dimensional function, $g(\cdot)$, smoothly changing except at a single point in time where the shift occurs. Throughout this article, we denote the unknown time series change-point location with the symbol τ. We also assume Σ is an unknown but constant $p \times p$ covariance matrix throughout our time series. A particular observation of the time series takes the form

$$\mathbf{x}_i = g(i) + \boldsymbol{\varepsilon}_i$$

where

$$\boldsymbol{\varepsilon}_i \sim N_p(\mathbf{0}, \Sigma).$$

Next we let the $N \times p$ matrix $\mathbf{X}$ represent our time series where each row represents an observation at a particular time. Additionally, we introduce the idealized $N \times p$ matrix, $\mathbf{H}$, which we compare against $\mathbf{X}$:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\tau-1,1} & x_{\tau-1,2} & \dots & x_{\tau-1,p} \\ x_{\tau 1} & x_{\tau 2} & \dots & x_{\tau p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{Np} \end{pmatrix}_{N \times p} \quad \text{and} \quad \mathbf{H} = \begin{pmatrix} 0 & 0 & \dots & \dots & 0 \\ 0 & 0 & \dots & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \dots & 0 \\ 1 & 1 & \dots & \dots & 1 \\ \vdots & \ddots & \ddots & \vdots & \\ 1 & 1 & \dots & \dots & 1 \end{pmatrix}_{N \times p}$$

We assume our time series is of dyadic length, that is, of length $N = 2^J$ for some $J \in \mathbb{N}$. While this appears to be a restrictive requirement, in practice there are several padding techniques that remedy this apparent difficulty [10]. For

example we might simply concatenate low level statistical noise to the front end of the time series to achieve the required dyadic length if we have data available from the in control time series state. Another method is to reflect the time series elements of sufficient length to obtain the require dyadic length. For example, a data set with six elements $(x_1, x_2, x_3, x_4, x_5, x_6)$ could be modified as $(x_3, x_2, x_1, x_2, x_3, x_4, x_5, x_6)$ to achieve the required dyadic length. The latter approach is what we will apply for our practical example in section 4.5.2.

We now take a one-dimensional discrete wavelet transform (DWT) of both $\mathbf{X}$ and $\mathbf{H}$ column by column which produces two $(N-1) \times p$ matrices in the wavelet domain $\mathbf{D}$ and $\mathbf{Q}$. We can normalize each detail level by its l^∞ norm which has the effect weighting coefficients from different resolution levels equally. With l^∞ normalized detail levels our subsequent analysis becomes more sensitive to changes detected by the finest resolution levels. In section 4.5 we apply our algorithm both with and without normalized detail coefficients. When the rows of zeroes and ones of $\mathbf{H}$ exactly correspond to the rows of $\mathbf{X}$ before and after change-point, the rows of $\mathbf{D}$ and $\mathbf{Q}$ will closely relate to each other in a meaningful manner as we describe below. Since the statistical properties of the additive noise component of the time series are retained after a one dimensional DWT, it can easily be shown using the linearity of the DWT that the expected covariance matrix after the transform remains Σ.

Notationally, we index our detail matrices to emphasize the detail levels of each row. More explicitly, supposing our time series is of length 2^J, we denote a $p-$dimensional detail coefficient as $\mathbf{d}_{jk} = (d_{jk,1}, d_{jk,2}, \ldots, d_{jk,p})$ where j represents a particular detail level and k the translation index at the given detail level. We

then express the DWT of $\mathbf{X}$ and $\mathbf{Q}$ as the matrices $\mathbf{D}$ and $\mathbf{Q}$ where,

$$\mathbf{D} = \begin{pmatrix} d_{01,1} & d_{01,2} & \dots & d_{01,p} \\ d_{11,1} & d_{11,2} & \dots & d_{11,p} \\ d_{12,1} & d_{12,2} & \dots & d_{12,p} \\ d_{21,1} & d_{21,2} & \dots & d_{21,p} \\ \vdots & \vdots & \ddots & \vdots \\ d_{jk,1} & d_{jk,2} & \dots & d_{jk,p} \\ \vdots & \vdots & \ddots & \vdots \\ d_{(J-1)\frac{J}{2},1} & d_{(J-1)\frac{J}{2},2} & \dots & d_{(J-1)\frac{J}{2},p} \end{pmatrix}_{(N-1)\times p}$$

and

$$\mathbf{Q} = \begin{pmatrix} q_{01,1} & q_{01,2} & \dots & q_{01,p} \\ q_{11,1} & q_{11,2} & \dots & q_{11,p} \\ q_{12,1} & q_{12,2} & \dots & q_{12,p} \\ q_{21,1} & q_{21,2} & \dots & q_{21,p} \\ \vdots & \vdots & \ddots & \vdots \\ q_{jk,1} & q_{jk,2} & \dots & q_{jk,p} \\ \vdots & \vdots & \ddots & \vdots \\ q_{(J-1)\frac{J}{2},1} & q_{(J-1)\frac{J}{2},2} & \dots & q_{(J-1)\frac{J}{2},p} \end{pmatrix}_{(N-1)\times p.}$$

Next, we define $\Delta = [\delta_1, \delta_2, \dots, \delta_p]$ as the amount our mean function shifts at the unknown change-point. It is important to note that here Δ is not a vector, rather a set of coefficients. We use the [ ] notation to distinguish this from say $\mathbf{q}_{11} = (q_{11,1}, q_{11,2}, \dots q_{11,p})$ which is a p-dimensional vector. So in particular, we define, $\Delta\mathbf{q}_{11} = (\delta_1 q_{11,1}, \delta_2 q_{11,2}, \dots \delta_p q_{11,p})$ using element-by-element scalar multiplication.

We know the additive noise component of the original time series is again transformed to an additive noise component after the dimension-by-dimension DWT is taken of the original time series. Furthermore, as illustrated in Figure 4.1, we know that at least for the finest level detail vectors that the true detail vector values should very closely match the detail vectors of $\Delta\mathbf{Q}$. In the case when the mean function of our time series is a multivariate step function, all true detail vectors will match the detail vectors of $\Delta\mathbf{Q}$. In the more general case where the true underlying mean function is unknown, we will ultimately retain only the finest level detail vectors in our final analysis. With these properties in mind, those retained empirical detail coefficient vectors, $\mathbf{d}^*_{jk}$, may therefore be modeled as

$$\mathbf{d}^*_{jk} \sim N_p(\mathbf{d}_{jk}, \Sigma) = N_p(\Delta\mathbf{q}_{jk}, \Sigma)$$

where $\mathbf{d}_{jk}$ is the true detail vector while $\mathbf{d}^*_{jk} = (d_{jk1}, d_{jk2}, \ldots d_{jkp})$ and $\mathbf{q}_{jk} = (q_{jk,1}, q_{jk,2}, \ldots q_{jk,p})$ are the jk rows of the matrices $\mathbf{D}$ and $\mathbf{Q}$, respectively. Using Bayes theorem, our posterior distribution of τ, Δ, and Σ takes the form of the product of our likelihood and prior distribution, $L(\mathbf{d}_{j,k}|\tau, \Delta, \Sigma)$ and $p_0(\tau, \Delta, \Sigma)$, respectively, as

$$p(\tau, \Delta, \Sigma|\mathbf{D}) \propto \prod_j \prod_k L(\mathbf{d}_{jk}|\Sigma, \tau, \Delta) p_0(\tau, \Delta, \Sigma). \tag{4.2}$$

In equation (5.1) we use the double index notation to emphasize that we are taking the product over distinct detail coefficients by their resolution and translation indices. In our model Σ is a constant but unknown covariance matrix. Following the discussion above, any prior covariance matrix information we have in our original time series directly applies after our transform. For example, we could put a Wishart distribution as an informative prior on Σ if we have sufficient

prior knowledge of Σ and its associated distribution parameters. For the most general case, however, we will apply Jeffrey's noninformative prior given as $p_0(\Sigma, \Delta, \tau) \propto |\Sigma|^{-1/2}$. We also note, that implicit in this prior is that we assign equal probability to the change-point location throughout the time series. Our posterior distribution takes the form

$$p(\tau, \Delta, \Sigma|\mathbf{D}) \propto |\Sigma|^{-m/2}\exp\left[-\tfrac{1}{2}\sum_j \sum_k (\mathbf{d}_{jk} - \Delta\mathbf{q}_{jk})^T\Sigma^{-1}(\mathbf{d}_{jk} - \Delta\mathbf{q}_{jk})\right]|\Sigma|^{-1/2}$$

where m represents the actual number of detail coefficients used in the analysis. In general, $\mathbf{\Delta}$ and Σ are unknown nuisance parameters. We therefore would like to integrate out these terms and use instead the marginal posterior distribution function

$$p(\tau|\mathbf{D}) \propto \int_{\mathbb{R}^{p+}\times\mathbb{R}^{p+}}\int_{\mathbb{R}^p} |\Sigma|^{-(m+1)/2}\exp\left[-\frac{1}{2}\sum_j \sum_k (\mathbf{d}_{jk} - \Delta\mathbf{q}_{jk})^T\Sigma^{-1}(\mathbf{d}_{jk} - \Delta\mathbf{q}_{jk})\right] d\Delta d\Sigma. \tag{4.3}$$

At first glance equation (4.3) appears to pose a difficult analytical problem because for each jk combination, we must integrate over a set of coefficients of the $\mathbf{q}_{jk}$ vector. We can remedy this issue by simply reversing the roles of $\mathbf{q}_{jk}$ and Δ. Notice by how $\mathbf{Q}$ is defined, that all the elements of any $\mathbf{q}_{jk}$ are identical. With this observation in mind, we let q_{jk} be a scalar representative for a given row of $\mathbf{Q}$ corresponding to the value of each element in that particular row. Next we let $\mathbf{\Delta}$ represent a vector of the mean function shift at the change-point in the natural way. With this change of notation in hand, we may equivalently write

equation (4.3) as

$$p(\tau|\mathbf{D}) \propto \int_{\mathbb{R}^{p+}\times\mathbb{R}^{p+}} \int_{\mathbb{R}^p} |\Sigma|^{-(m+1)/2} \exp\left[-\frac{1}{2} \sum_j \sum_k (\mathbf{d}_{jk} - q_{jk}\boldsymbol{\Delta})^T \Sigma^{-1} (\mathbf{d}_{jk} - q_{jk}\boldsymbol{\Delta}) \right] d\boldsymbol{\Delta} d\Sigma. \tag{4.4}$$

Expanding the exponent of equation (4.4) we obtain

$$\begin{aligned} p(\tau|\mathbf{D}) &\propto \int_{\mathbb{R}^{p+}\times\mathbb{R}^{p+}} \int_{\mathbb{R}^p} |\Sigma|^{-(m+1)/2} \\ &\times \exp\left[-\frac{1}{2} \sum_j \sum_k \left(\mathbf{d}_{jk}^T \Sigma^{-1} \mathbf{d}_{jk} + q_{jk} \boldsymbol{\Delta}^T \Sigma^{-1} q_{jk} \boldsymbol{\Delta} - q_{jk} \boldsymbol{\Delta}^T \Sigma^{-1} \mathbf{d}_{jk} - \mathbf{d}_{jk}^T \Sigma^{-1} q_{jk} \boldsymbol{\Delta} \right) \right] d\boldsymbol{\Delta} d\Sigma \\ &= \int_{\mathbb{R}^{p+}\times\mathbb{R}^{p+}} \int_{\mathbb{R}^p} |\Sigma|^{-(m+1)/2} \exp\left[-\frac{1}{2} \left(A + C \boldsymbol{\Delta}^T \Sigma^{-1} \boldsymbol{\Delta} - \boldsymbol{\Delta}^T \Sigma^{-1} B - B^T \Sigma^{-1} \boldsymbol{\Delta} \right) \right] d\boldsymbol{\Delta} d\Sigma \end{aligned} \tag{4.5}$$

where

$$A = \sum_j \sum_k \mathbf{d}_{jk}^T \Sigma^{-1} \mathbf{d}_{jk}, \;\; B = \sum_j \sum_k q_{ij} \mathbf{d}_{jk}, \;\; B^T = \sum_j \sum_k q_{ij} \mathbf{d}_{jk}^T, \; C = \sum_j \sum_k q_{ij}^2.$$

Continuing from equation (4.5) we provide the following detailed calculations:

$$
\begin{aligned}
p(\tau|\mathbf{D}) \propto & \int_{\mathbb{R}^{p+}\times\mathbb{R}^{p+}} \int_{\mathbb{R}^p} |\Sigma|^{-(m+1)/2} \\
& \exp\left[-\frac{C}{2}\left(\frac{A}{C} + \boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta} - \boldsymbol{\Delta}^T\Sigma^{-1}\frac{B}{C} - \frac{B^T}{C}\Sigma^{-1}\boldsymbol{\Delta}\right)\right] d\boldsymbol{\Delta}\, d\Sigma \\
= & \int_{\mathbb{R}^{p+}\times\mathbb{R}^{p+}} |\Sigma|^{-(m+1)/2} \exp\left[-\frac{1}{2}\left(A - \frac{1}{C}B^T\Sigma^{-1}B\right)\right] \\
& \int_{\mathbb{R}^p} \exp\left[-\frac{C}{2}\left((\boldsymbol{\Delta} - \frac{B}{C})^T\Sigma^{-1}(\boldsymbol{\Delta} - \frac{B}{C})\right)\right] d\boldsymbol{\Delta}\, d\Sigma \\
= & \int_{\mathbb{R}^{p+}\times\mathbb{R}^{p+}} |\Sigma|^{-(m+1)/2} |\Sigma|^{1/2} C^{-1/2} \exp\left[-\frac{1}{2}\left(A - \frac{1}{C}B^T\Sigma^{-1}B\right)\right] d\Sigma \\
= & \int_{\mathbb{R}^{p+}\times\mathbb{R}^{p+}} \Sigma^{-m/2} C^{-1/2} \exp\left[-\frac{1}{2}\left(\sum_j\sum_k \mathbf{d}_{jk}^T\Sigma^{-1}\mathbf{d}_{jk} - \frac{1}{C}B^T\Sigma^{-1}B\right)\right] d\Sigma \\
= & \int_{\mathbb{R}^{p+}\times\mathbb{R}^{p+}} \Sigma^{-m/2} C^{-1/2} \exp\left[-\frac{1}{2}\left(tr(\Sigma^{-1}\sum_j\sum_k \mathbf{d}_{jk}\mathbf{d}_{jk}^T - \frac{1}{C}BB^T)\right)\right] d\Sigma \\
\propto & \, C^{-\frac{1}{2}} \left| \sum_j\sum_k \mathbf{d}_{jk}\mathbf{d}_{jk}^T - \frac{1}{C}BB^T \right|^{-(m-p-1)/2}
\end{aligned}
\tag{4.6}
$$

where in the last step equation (4.6) follows by dropping multiplicative constants and applying the known form of the Wishart distribution. One can show that the univariate case as derived in chapter 2 is simply the $p = 1$ case in equation (4.6).

Formally we estimate the change-point of the time series as $\arg\max_\tau p(\tau|\mathbf{D})$. In particular there are $N-1$ possible values of τ and a maximum value always exists. Notice equation (4.6) is neither wavelet nor detail level specific. Depending on what we know (or do not know) about the time series, different wavelet and detail level combinations may be more appropriate. Depending on the true underlying mean function of the time series, we found through simulation studies the choice of wavelet had a minor, but noticeable, effect on correctly estimating the change-point location. In the simplest case, when the mean function is represented by a multivariate step function, studies

show it is also the simplest wavelet (i.e. the Haar wavelet) that performs marginally better. In the case of a smoothly varying mean function, the Daubechies wavelet with ten vanishing moments became the best choice in correctly estimating the change-point location.

We also need to decide which detail levels to apply. This decision is fairly straightforward depending on what is known about the true mean function. In general the more applicable detail vectors that we can use in equation (4.6), the more confidence we will be able to attribute to our conclusions. So long as the mean function is smooth except at the change-point location then our model assumptions apply and at least the finest two to three detail levels should be applied. If more information about the mean is available, it may be optimal to use more detail levels. For example in the case of a multivariate step function, all detail levels should be applied.

4.4 Bayesian-wavelet approach to detecting the existence of a change-point

We determine the existence of a change-point by taking a model selection approach and applying a form of the Schwarz Information Criteria (SIC). Let M_1 denote that a single change-point occurs in the mean function of our time series and let M_2 denote the model where no change occurs. We first compute the likelihood of observing either of these two models given the observed data. Since it is unclear which constants will cancel in the ratio of these two models, we must retain them throughout the calculations. Then similar to our previous derivation of equation (4.6), we obtain the following likelihood for M_1:

$$P(\mathbf{D}|M_1) = K(2\pi)^{(p-mp)/2}(2)^{mp/2}\Gamma_p(\tfrac{m}{2})C^{-1/2}\left|\sum_j\sum_k \mathbf{d}_{jk}\mathbf{d}_{jk}^T - \frac{1}{C}BB^T\right|^{-(m-p-1)/2} \tag{4.7}$$

where $\Gamma_p(\cdot)$ is the multivariate gamma function defined as

$$\Gamma_p(x) = \pi^{p(p-1)/4}\sum_{i=1}^{p}\Gamma[x + (1-i)/2],$$

K is a constant common to both models, and all other terms are as previously defined.

In M_2, calculations are simplified since $\mathbf{\Delta}$ is assumed to be the $p-$dimensional zero vector. Once again adopting a similar approach as before we obtain the likelihood of observing our data under M_2:

$$P(\mathbf{D}|M_2) = K(2\pi)^{-mp/2}(2)^{(m+1)p/2}\Gamma_p(\tfrac{(m+1)}{2})|\sum_j\sum_k \mathbf{d}_{jk}\mathbf{d}_{jk}^T|^{-(m-p)/2}. \tag{4.8}$$

We note the difference in the number of free parameters in M_1 and M_2 is $k_2 - k_1 = p$, namely the dimension of $\mathbf{\Delta}$. This suggests a form of the SIC

$$\Delta(SIC) = -2(\log P(\mathbf{D}|M_2) - \log P(\mathbf{D}|M_1)) + (k_2 - k_1)\log N.$$

For our multi-dimensional change-point problem we maximize equation (4.7) for τ to obtain our final result

$$\Delta(SIC) = -2(\log(P(\mathbf{D}|M_2)) - \log P(\mathbf{D}|M_1)) + p\log(N) \tag{4.9}$$

where equation (4.9) implicitly assumes equal probabilty of realizing either M_1 or M_2. In certain instances the modeler may have reason to favor one model over the other and so the prior odds ratio of the two models would not be 1. Recall, the posterior odds ratio may be expressed as

$$\begin{aligned} \frac{P(M_1|\mathbf{D})}{P(M_2|\mathbf{D})} &= \frac{P(\mathbf{D}|M_1)}{P(\mathbf{D}|M_2)} \frac{P(M_1)}{P(M_2)} \\ &= \text{Bayes Factor} \times \frac{P(M_1)}{P(M_2)}. \end{aligned} \tag{4.10}$$

We may modify equation (4.9) to incorporate a prior belief to *a priori* favor one model over the other. In our setting this may be accomplished by substituting the data dependent terms in equation (4.9) with -2 times the log of equation (4.10). For the later examples and simulations we provide, we note that each model is given equal weight and equation (4.9) is implemented in its present form.

Our selection process is now a straightforward calculation of $\Delta(SIC)$. We select the no change model when $\Delta(SIC) < 0$ and infer a change-point exists in the time series when $\Delta(SIC) > 0$. We note slightly positive values (i.e. $\Delta(SIC) \leq 3$) should be treated with caution. Although the change-point model is favored in such cases, the evidence is not particularly strong. Values computed farther from zero (i.e. $\Delta(SIC) > 3$) denote strong evidence of the existence of a change-point with more assurance obtained with larger computed values.

4.5 Examples

4.5.1 Illustrative Example

We provide an illustrative example to demonstrate how our Bayesian-wavelet approach to the multivariate change-point problem easily adapts to various mean functions. For this example, we simulate data from a

three dimensional normal distribution centered around 0 for the first 85 elements of the time series and then introduce a shift of 1 unit in the first and third dimensions for the remaining 43 elements. The covariance matrix remains constant throughout the time series and has .25 on all diagonal elements and 0 on all off diagonal elements. Figure 4.2 depicts the time series of our time series where the shift in the first and third dimensions is visually evident.

Applying a classical MLE approach to this time series correctly returns time point 85 as the estimated change-point location. A pure Bayesian such as the one described by Perreault [44] approach also returns time point time point 85 as the estimated change-point location along with a 95% credible interval of 84-86. Applying our Bayesian-wavelet approach we first calculate the SIC using equation (4.9) to determine the existence of a change-point. Equation (4.9) returns a value of 53.25 providing us with near certainty that a change-point exists in the data. Estimating the change-point location with equation (4.6) also correctly returns the change-point location at time point 85 with a 95% credible interval of 84-87 (see Figure 4.3).

To illustrate the power of the Bayesian-wavelet approach, suppose we now impose one period of a sine wave on the same data set in each dimension. This new data set now represents the scenario where our time series in nonlinear. Figure 4.4 depicts this new time series where we see the change-point at time point 85 is much more obscured. Applying both the MLE and pure Bayesian approaches to this time series with a nonlinear mean function both return meaningless results as the assumptions upon which they are based are now violated. Directly imputing the new time series in the MLE algorithm, for example, incorrectly estimates the change-point location at time 63.

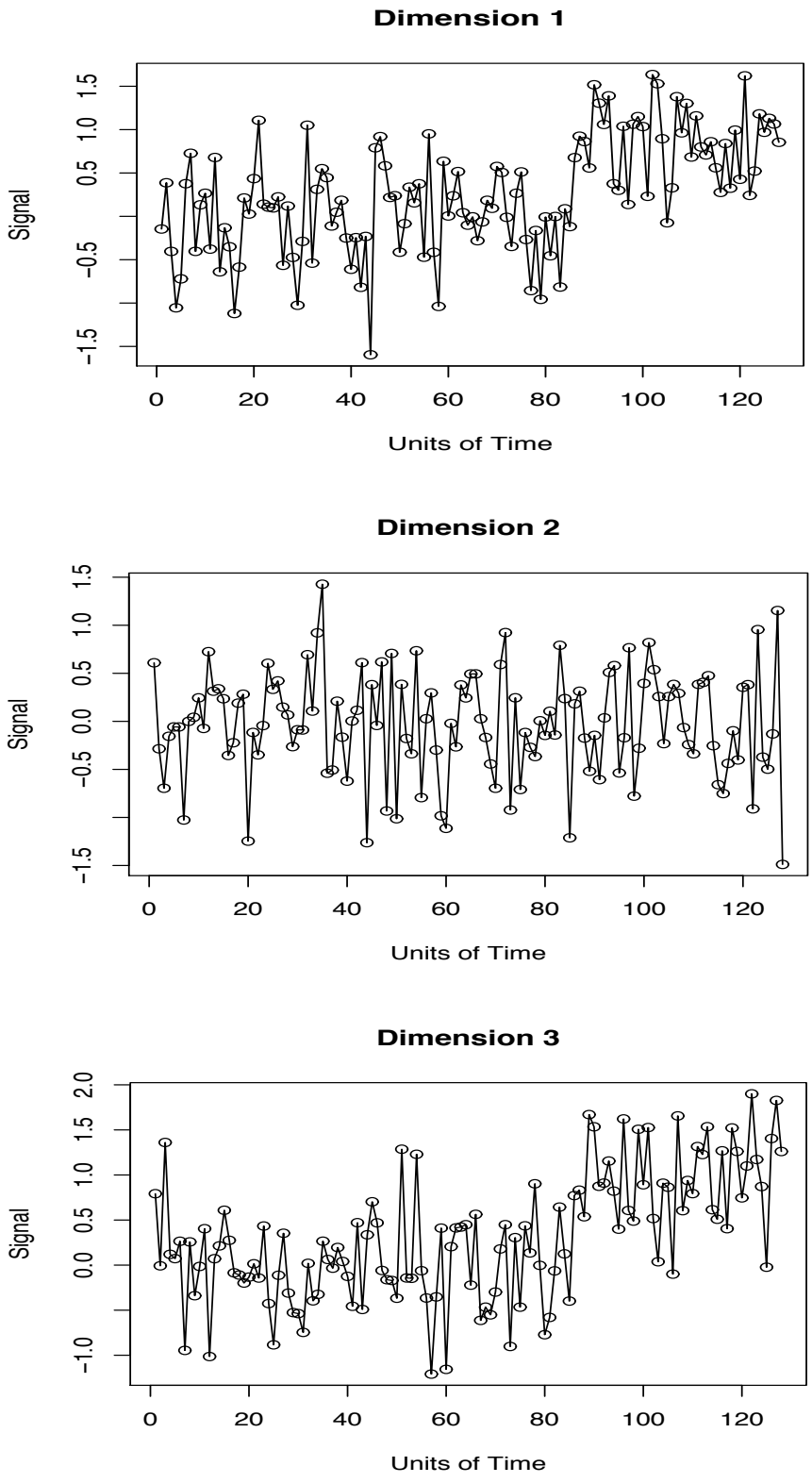

Figure 4.2: A three dimensional time series where the mean function is a three dimensional step function. In particular a shift occurs at time point 85 in the first and third dimensions.

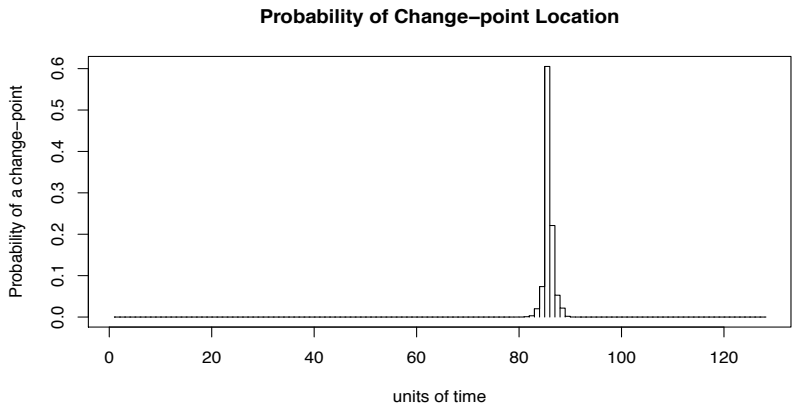

Figure 4.3: Marginal posterior distribution from the time series in Figure 4.2 with concentrated probability at the correct change-point at time point 85.

Our Bayesian-wavelet approach, however, easily adapts to this more complicated situation. Using the four highest detail coefficient levels we calculate an SIC of 12.5 indicating the presence of a change-point in the time series. Maximizing equation (4.6) for τ correctly estimates the change-point location once again at time point 85. Figure 4.5 displays the relative probabilities for the change-point location with a slightly less concentrated 95% credible interval of 82-88.

4.5.2 Practical Example

We present a practical example implementing the methods developed in this article involving six hydrological sequences in the Northern Québec Labrador region as represented in Figure 4.6. In particular, we analyze the streamflow in units of $1/(\text{km}^2 \times \text{s})$ measured in the springs from 1957 to 1995. It has been noted that a perceptible general decrease in streamflow seemed to occur in the 1980's in this region. The regional proximity of the rivers suggests a likely relationship between the rivers, but the specific covariance structure is unclear *a priori*.

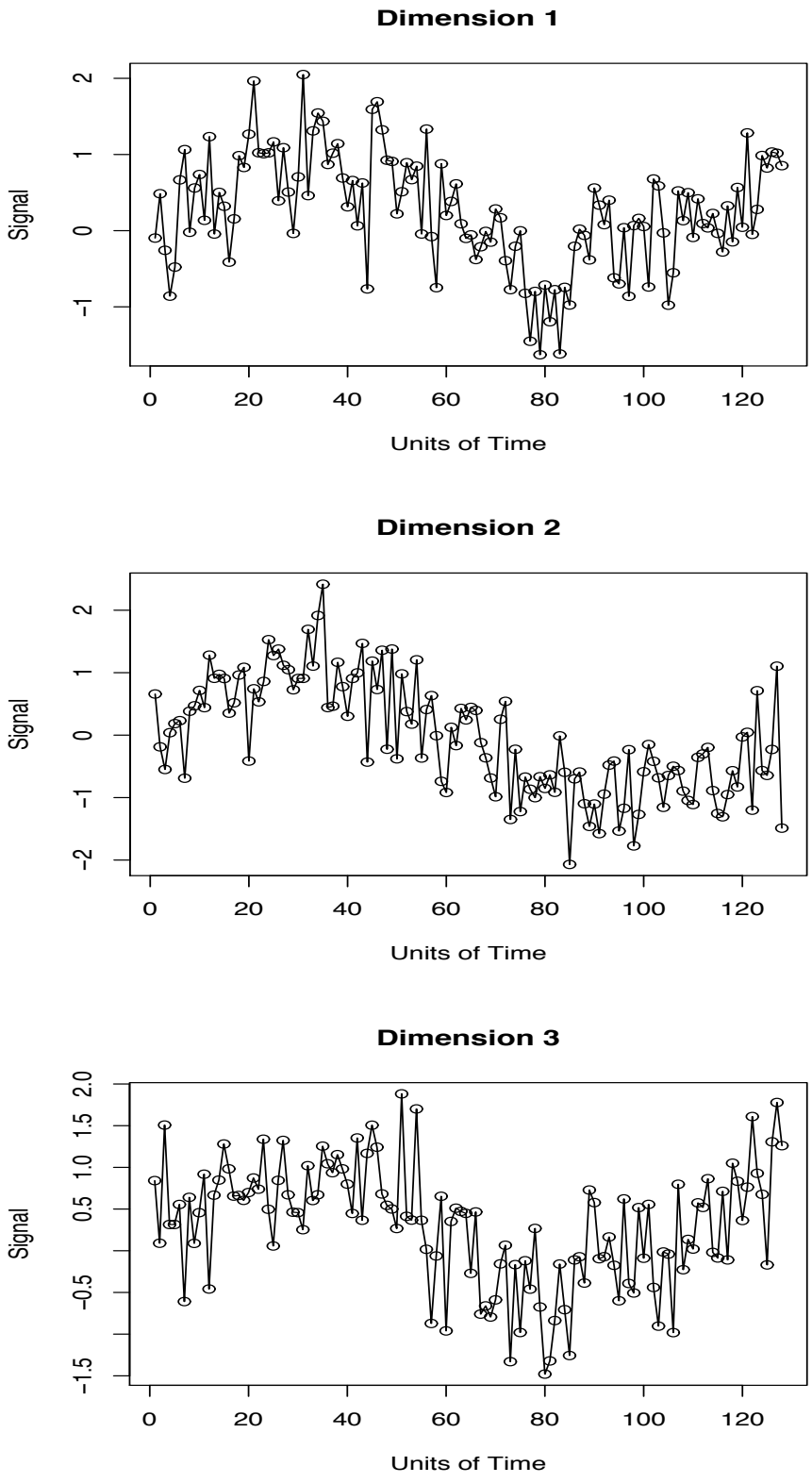

Figure 4.4: This is the same data set at in Figure 4.2 only now with one period of the trigonometric function $\sin(2\pi t/128)$ added to the elements in each dimension.

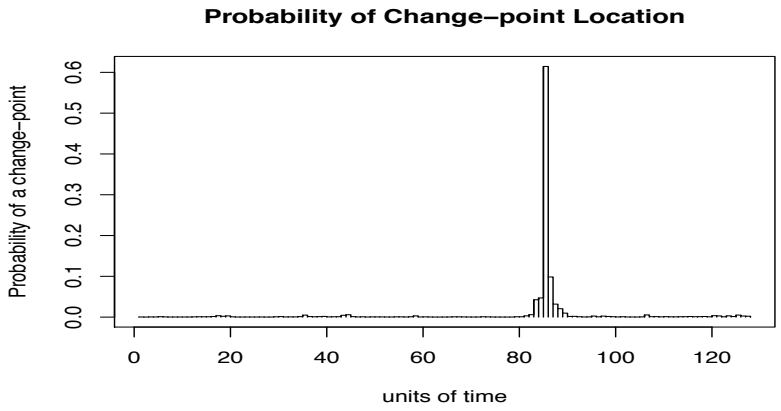

Figure 4.5: Marginal posterior distribution from the time series in Figure 4.4 with concentrated probability at the correct change-point at time point 85.

Hence, a multivariate analysis certainly appears more appropriate than six individual river univariate studies. The assertion is that due to causes attributed to perhaps climate change or other regional factors, a change-point in streamflow has occurred. Applying our methods, we would like to determine whether or not our methods support this assertion and if so estimate the change-point year.

Perreault [44] originally applied a retrospective Bayesian change-point analysis to this data set. The principle advantage of our Bayesian-wavelet method over Perreault's pure Bayesian approach to this data set is that our method applies even if the true underlying mean function is not a step function. Perreault spends considerable time justifying rather strict assumptions on the data and the choice of hyperparameters used in the model. While Perreault's analysis appears largely valid in this case, the strict assumptions required by such a pure Bayesian approach limit its applicability in more general contexts and often make conclusions less compelling. With the Bayesian-wavelet approach, however, we have no need to illicit informative priors for the mean vectors both before and after the unknown change-point nor for the covariance matrix to construct our model. As discussed above, we require only that the true underlying mean

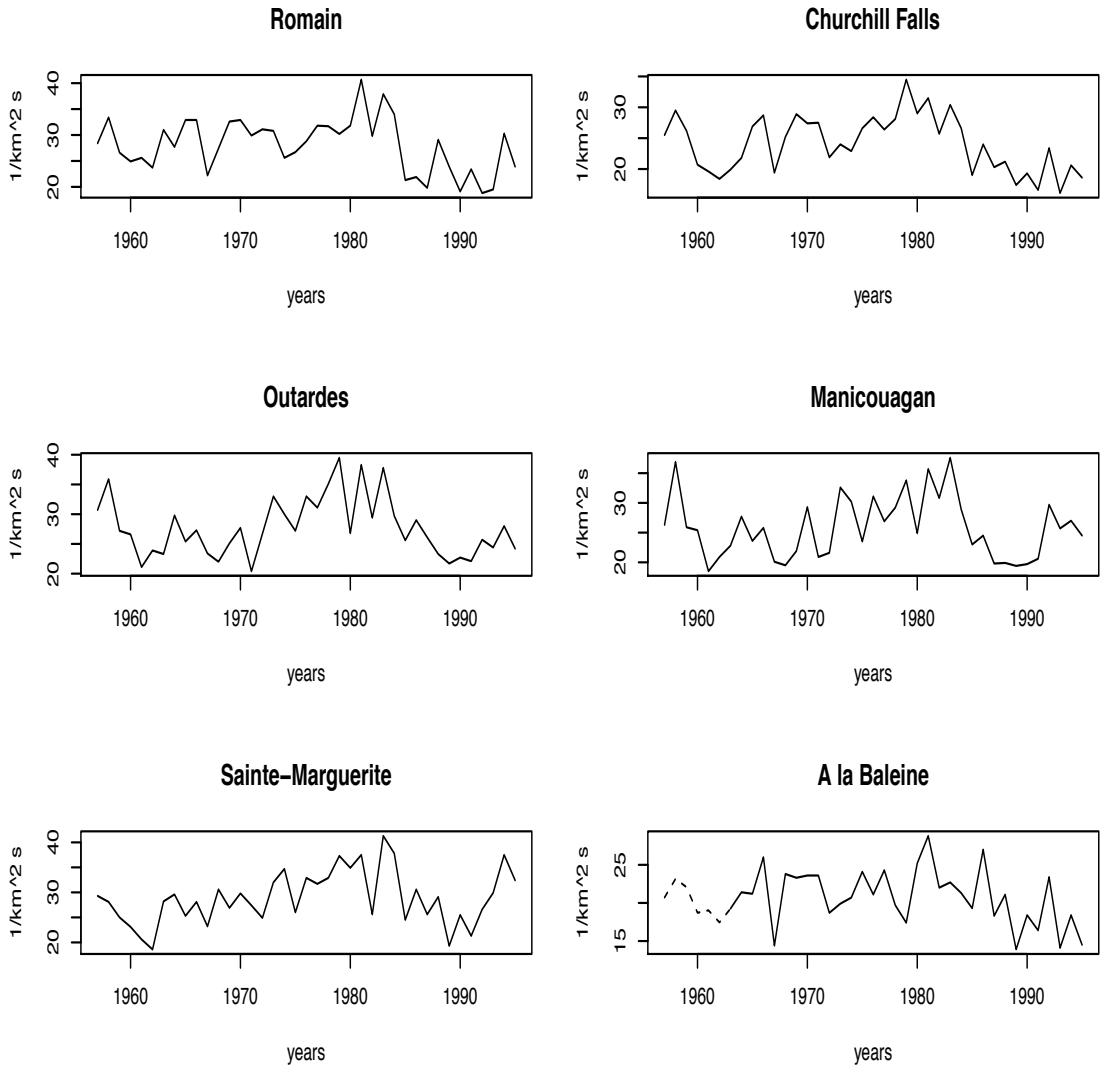

Figure 4.6: Plots of riverflows of six rivers in the Northern Québec Labrador region. The dashed lines for À la Baliene are years river flows are estimated from a linear regression since the actual data is unavailable.

function be smooth except at the single change-point and that the random component be normally distributed.

To begin our analysis we note measurements for À la Baliene are unavailable from the years 1957-1962 inclusive. To handle this discrepancy we took two different approaches. In the first case we simply analyzed the data for the common years from 1963-1995 inclusive. In the second approach, we treat river flows for À la Baliene as a dependent variable and perform a linear regression for the years with complete data against the other five rivers. With the linear model in hand, we estimate river flows for À la Baliene for the years 1957-1962 from the linear model using the data from the other rivers with complete data sets. The dashed line in Figure 4.6 for À la Baliene represents these estimated values. After a comparison of our analyses we find very similar results are obtained in both cases. As such we present results from only the latter case.

We implement the Daubechies 10-tap wavelet since it has known properties particularly well suited to detect abrupt time series change [55]. Based on Perreault's analysis, the mean function is some unknown multivarate step function. If this property actually holds, we should be able to apply all detail levels with Bayesian-wavelet and arrive at the same answer. Standardizing detail coefficients as described in section 4.3.2, we thus apply apply all detail coefficients in our analysis. Finally, we note this time series is not a power of two as required to apply any DWT. We remedy this situation by simply reflecting the beginning of the time series to achieve the required dyadic length.

With our wavelet parameters in hand, we next must determine whether or not a statistical change-point in the mean vector even exists in our data set. A computation of the SIC returns a value 14.53 which represents strong evidence for the existence of a statistical change-point. Next, we estimate the location of the change-point by maximizing the Bayesian-wavelet change-point equation for τ.

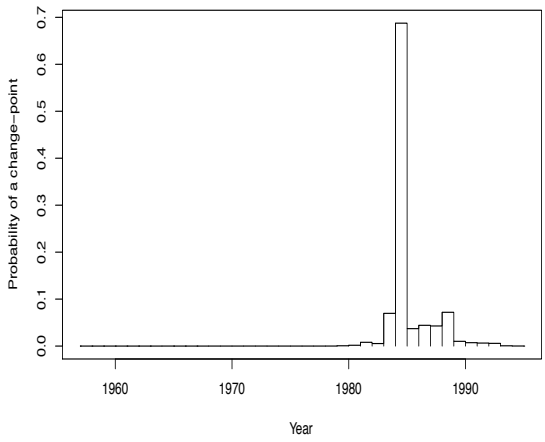

Figure 4.7: Posterior distribution of a change-point for six hydrological sequence in the Northern Québec Labrador region.

This returns the year 1984 as the posterior mode with posterior probability of nearly .70. Furthermore, we note a .90 credible interval ranges around this estimation of the change-point location from 1983-1986 (see Figure 4.7). We note these results are similar to Perreault, who also estimated the change-point year as 1984, but with a slightly different .90 credible interval of 1982-1985 [44].

4.6 Simulations

4.6.1 Multivariate step mean function

For the simulations in this section we generate multivariate time series with an underlying mean function represented as a multivariate step function. The time series length in each case is 128 where the change-point is randomly selected somewhere in the middle 90% of the time series elements. Furthermore, we generate simulated data for two separate covariance matrices $\Sigma_1 = I$, the identity covariance matrix, and then Σ_2 a covariance matrix with 1's along the diagonal and .5's on all off diagonal elements. We record the percentage each

method correctly estimates the change-point location with two time units of the true change-point location for each 1,000 simulation run.

Table 4.1: Perecentage of change-point estimations within two time units of actual change-point after 1,000 simulations. In all cases the initial mean vector is $\boldsymbol{\mu}_\tau = (0, 0, \ldots, 0)$ and then shifts to $\boldsymbol{\mu}'_\tau = (\delta, \delta, \ldots, \delta)$. BW indicates the Bayesian-wavelet approach and MLE indicates the maximum likelihood estimation approach. Simulations are conducted with two covariance matrices, the identity covariance matrix ($\mathbf{I}$) and a covariance matrix with 1's along the diagonal and .5's on all off diagonal elements (Σ_1).

			Time Series Dimension						
Chgt Pt method	δ	Σ	2	5	10	25	50	75	100
BW	0.5	$\mathbf{I}$	.36	.60	.79	.98	.98	.98	.93
MLE	0.5	$\mathbf{I}$	.37	.59	.79	.94	.98	.98	.92
BW	1.0	$\mathbf{I}$	.77	.96	.99	1.00	1.00	1.00	1.00
MLE	1.0	$\mathbf{I}$	.77	.96	.99	1.00	1.00	1.00	1.00
BW	1.5	$\mathbf{I}$	.95	.99	1.00	1.00	1.00	1.00	1.00
MLE	1.5	$\mathbf{I}$	.96	.99	1.00	1.00	1.00	1.00	1.00
BW	2.0	$\mathbf{I}$	.99	1.00	1.00	1.00	1.00	1.00	1.00
MLE	2.0	$\mathbf{I}$	.99	1.00	1.00	1.00	1.00	1.00	1.00
BW	0.5	Σ_1	.21	.14	.20	.13	.07	.06	.05
MLE	0.5	Σ_1	.22	.16	.21	.13	.08	.07	.05
BW	1.0	Σ_1	.60	.56	.71	.62	.41	.24	.12
MLE	1.0	Σ_1	.60	.56	.71	.63	.41	.26	.14
BW	1.5	Σ_1	.85	.81	.89	.87	.76	.57	.32
MLE	1.5	Σ_1	.84	.81	.90	.88	.76	.57	.31
BW	2.0	Σ_1	.96	.97	.98	.97	.91	.88	.54
MLE	2.0	Σ_1	.96	.97	.98	.97	.91	.88	.53

Before we can begin our simulations, we must decide which detail levels and which wavelet to apply. In the case of a stationary time series with a single-change point there is no underlying trend to the time series contributing to time series change except the single change-point. In this case, the change information contained in the detail coefficients only pertains to the change-point itself. Hence, we apply all wavelet details in our simulations. For the wavelet function itself we present results from the Haar wavelet although applying the Daubechies 10-tap wavelet yielded similar results.

We compared the effectiveness of the Bayesian-wavelet and MLE methods for estimating the change-point location by varying the dimension, jump size, and covariance matrix of our time series. It is interesting to note that the simulation results suggest there is very little difference between the MLE and Bayesian-wavelet approach in correctly estimating the change-point. Furthermore, what differences may exist become less important the higher we go in dimension. Thus, we obtain comparable results to the MLE method with our Bayesian-wavelet method without the same stringent MLE time series assumptions.

4.6.2 Multivariate piecewise smooth function with a single mean function shift

We see the Bayesian-wavelet method performed similarly to the MLE method in estimating change-points for multivariate step functions such as those considered in section 4.6.1. These comparable results were obtained even though the simulations of section 4.6.1 were constructed from the precise assumptions upon which the MLE method was built. We next investigate how these methods perform when the underlying mean function does not conform to a multivariate step function. In particular, since the Bayesian-wavelet method requires only the underlying mean function to be smooth except at the change-point, we consider a multivariate time series with a nonconstant smoothly varying mean function.

We generate time series with a smoothly varying mean function except at a single change-point. Specifically, we set the initial mean vector to $\boldsymbol{\mu}_\tau = (\sin(\frac{2\pi t}{128}), \sin(\frac{2\pi t}{128}), \ldots, \sin(\frac{2\pi t}{128}))$ and then after the change point the mean vector becomes $\boldsymbol{\mu}'_\tau = (\sin(\frac{2\pi t}{128}) + 1, \sin(\frac{2\pi t}{128}) + 1, \ldots, \sin(\frac{2\pi t}{128}) + 1)$. That is the shift vector is $\boldsymbol{\Delta} = (1, 1, \ldots, 1)$ for all simulations. We then incrementally adjust the

variance of the additive noise by changing the diagonal terms of the covariance matrix. We set our covariance matrix equal to the identity multiplied by the constant σ^2 as given in Table 4.2. The change-point is randomly selected from the middle 90% of the time series and the Daubechies 10-tap wavelet is applied using the four highest details coefficients.

Table 4.2: Percentage each method estimates the change-point location within 2 time units of true change-point location where each run represents 1,000 simulations. In all cases the initial mean vector is $\boldsymbol{\mu}_\tau = (\sin(\frac{2\pi t}{128}), \sin(\frac{2\pi t}{128}), \ldots, \sin(\frac{2\pi t}{128}))$ and then shifts to $\boldsymbol{\mu}'_\tau = (\sin(\frac{2\pi t}{128}) + 1, \sin(\frac{2\pi t}{128}) + 1, \ldots, \sin(\frac{2\pi t}{128}) + 1)$. BW indicates the Bayesian-wavelet approach and MLE indicates the maximum likelihood estimation approach. Throughout the simulations the covariance matrix used is the identity multiplied by σ^2.

		Time Series Dimension						
Chgt Pt method	σ^2	2	4	6	8	10	25	50
BW	0.2	.98	.99	1.00	1.00	1.00	1.00	1.00
MLE	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
BW	0.4	.88	.99	.99	1.00	1.00	1.00	1.00
MLE	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
BW	0.6	.71	.92	.99	.99	.99	1.00	1.00
MLE	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0
BW	0.8	.60	.87	.94	.97	.99	1.00	1.00
MLE	0.8	.01	.01	0.0	0.0	0.0	0.0	0.0
BW	1.0	.53	.78	.89	.96	.97	1.00	1.00
MLE	1.0	.02	.01	0.0	0.0	0.0	0.0	0.0
BW	1.2	.44	.70	.89	.90	.95	.99	1.00
MLE	1.2	.01	0.0	0.0	0.0	0.0	0.0	0.0
BW	1.4	.39	.62	.76	.83	.89	.99	.99
MLE	1.4	.01	.01	0.0	0.0	0.0	0.0	0.0
BW	1.6	.31	.54	.69	.79	.87	.98	.99
MLE	1.6	.01	.00	0.0	.00	0.0	0.0	0.0

Simulation results provide evidence that the Bayesian-wavelet method does well seeing through additive noise component of the time series and estimating the true change-point location. Applying equation (4.6) exactly as we did in section 4.6.1 only now with just the higher detail coefficients and the Daubechies 10-tap wavelet, we have a method that easily adapts to estimate

change-points in a very different time series. Method such as MLE or a pure Bayesian that make strict assumptions on the true underlying mean function do not share this same flexibility. We see the underlying form of the oscillating mean function violates the MLE assumptions in such a way that this method has no ability to correctly estimate the change-point location. Only in the lower dimensional cases with high variance when the time series more closely resembles pure noise, does the MLE register a few correct estimates by chance alone. In the other cases the geometry of the time series forces the MLE method away from the true change-point location.

4.7 Conclusion

Another interesting aspect of the Bayesian-wavelet approach is how it may be used as an indirect tool to validate certain data set assumptions. When parametric methods such as MLE or pure Bayesian models are applied to infer and estimate the location of a single change-point in a multivariate time series, the true underlying mean function is typically a multivariate step function. In principle using all detail levels of our Bayesian-wavelet method should always return very nearly identical change-point location estimates in such cases. If a discrepancy exists between the above parametric methods with our Bayesian-wavelet method, then either the time series signal to noise ratio is not sufficiently high or the model assumptions are simply not valid.

Our multivariate Bayesian-wavelet approach for detecting statistical change-points performs comparably with the classical MLE method when the true mean function of the time series is a multivariate step function. The advantage to our approach is seen in how our method also easily extends to more general situations. The simulations demonstrate how the MLE method fails when its

model assumptions become invalid, but also show how the Bayesian-method still performs well. We chose a multivariate trigonometric function as an example in our simulations, but the detail coefficient properties we exploited here could be applied equally well to many other such piecewise smooth multivariate functions. We thus find that the Bayesian-wavelet method affords the modeler valuable flexibility in more general situations and could serve as a valuable diagnostic tool in the setting of the multivariate change-point problem.

Chapter 5

Applications in Dimensionality Reduction

5.1 Introduction

In the previous chapter we extended the Bayesian-wavelet change-point estimation equation, $p(\tau|\mathbf{X})$, to the multivariate setting. While $p(\tau|\mathbf{X})$ applies to arbitrary dimensional time series in theory, in practice computational difficulties almost always arise as we encounter "the curse of dimensionality" when the dimension of the time series becomes very large [56]. One approach to overcome the problems associated with working with high dimensional data is to apply dimension reduction techniques [57]. The hope is that the time series projected into some lower dimensional space still retains the salient characteristics of the original time series only now in a computationally tractable dimension [58].

Two of the most popular dimensional reduction strategies are principal component analysis (PCA) and random projections. In the mean square sense, PCA is superior to random projections as we observe less vector distortion after the projection [59]. One significant PCA drawback, however, is that computationally PCA itself may become infeasible in high dimensions and thus

defeat the purpose of employing the approach in the first place [60]. Random projections, on the other hand, require much less computationally expensive operations, namely, the generation of a random matrix and then a matrix multiplication. While we must accept some additional uncertainty in drawing conclusions from the randomly projected data, the benefit is that we can apply this method to much higher dimensional data sets.

In this chapter, we investigate the applicability of our Bayesian-wavelet change-point estimation equation to high dimensional data. After a dimension reduction through a random projection, we apply $p(\tau|\mathbf{X})$ to the now dramatically reduced dimensional space. Since dimension reduction always comes at the price of information loss, the question becomes how confident are we of still correctly estimating the change-point location. In what follows, we develop a conservative lower confidence bound to correctly estimate the change-point location given both the shift vector and covariance matrix of the original time series along with the dimension of our projection space.

5.2 Notation and definitions

We begin by providing a list of definitions along with associated notation that will be used throughout this chapter. We also provide assumptions that our subsequent analysis will be based upon.

- $\mathbf{T}$ is a dimension preserving linear transformation (or interchangeably a matrix) such that $\mathbf{T} : \mathbb{R}^p \to \mathbb{R}^p$
- $\mathbf{P}$ is a dimension reducing linear transformation (or also interchangeably a matrix) such that $\mathbf{P} : \mathbb{R}^p \to \mathbb{R}^d$ where $d \leq p$.
- $\mathbf{X} = \{\mathbf{x_i}\}_i^n$ and $\mathbf{Y} = \{\mathbf{y_i}\}_i^n$ are $p-$dimensional and $d-$dimensional time

series respectively, both of length n with a multivariate additive gaussian noise component. Depending on the context, we write $\mathbf{T}\mathbf{X}^T = \mathbf{Y}^T$ or $\mathbf{P}\mathbf{X}^T = \mathbf{Y}^T$. Furthermore, we assume the true underling mean function of $\mathbf{X}$ is a multivariate step function.

· $\mathbf{\Delta}$ is the shift vector that the mean function of our time series undergoes at an unknown time, τ.

· We use superscript notation to denote either the original time series or the transformed time series. For example, Σ^X denotes the covariance matrix of the original time series. $\mathbf{\Delta}^Y$ denotes the shift vector of the transformed time series, etc.

· From empirical observations we define a rule of thumb for the signal-to-noise ratio: $SNR = ||\mathbf{\Delta}||/(p^{1/4}\sqrt{\lambda_{max}})$ for a p dimensional time series.

· $p(\tau|\mathbf{X})$ is the Bayesian-wavelet marginal posterior distribution function here evaluated at change-point τ for time series $\mathbf{X}$. Recall, the explicit formula

$$p(\tau|\mathbf{X}) = C^{-\frac{1}{2}} \left| \sum_j \sum_k \mathbf{d}_{jk}\mathbf{d}_{jk}^T - \frac{1}{C} BB^T \right|^s$$

where

$$B = \sum_j \sum_k q_{ij}\mathbf{d}_{jk},\ (B)^T = \sum_j \sum_k q_{ij}\mathbf{d}_{jk}^T,\ C = \sum_j \sum_k q_{ij}^2, \text{ and } s = -\tfrac{m-p-1}{2}.$$

5.3 Confidence bound after a random projection

Given the covariance structure, Σ^X, and the shift vector, $\mathbf{\Delta}^X$, of a high dimensional time series, we seek to establish a lower confidence bound for correctly estimating the change-point location of such a time series. Our

approach is to first standardize an arbitrary time series in a manner that will be clear below. The purpose of this standardization step is twofold. Firstly, the standardization step allows us to more clearly compare the difficulty of the change-point problem for different time series. Secondly, the standardization step transforms our time series into a form where we can take advantage of known random projection results.

Claim. If $\mathbf{d}_{ij}^X$ and $\mathbf{d}_{ij}^Y$ represent the detail coefficient vectors of a time series at some corresponding level and scale, ij, both before and after a linear transformation, $\mathbf{T}$, then $\mathbf{T}\mathbf{d}_{ij}^X = \mathbf{d}_{ij}^Y$.

Proof. We may represent our time series by an $n \times p$ matrix $\mathbf{X}$:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

Here each row represents a distinct element from the time series. We may represent our linear transformation, $\mathbf{T}$, by a $p \times p$ matrix of full rank

$$\mathbf{T} = \begin{pmatrix} t_{11} & t_{12} & \dots & t_{1p} \\ t_{21} & t_{22} & \dots & t_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ t_{p1} & t_{p2} & \dots & t_{pp} \end{pmatrix}$$

So we represent the linear transformation by the matrix multiplication $\mathbf{T}\mathbf{X}^T$ to obtain $\mathbf{Y}^T$ where

$$\mathbf{Y}^T = \begin{pmatrix} t_{11}x_{11} + \cdots + t_{1p}x_{1p} & t_{11}x_{21} + \cdots + t_{1p}x_{2p} & \dots & t_{11}x_{n1} + \cdots + t_{1p}x_{np} \\ t_{21}x_{11} + \cdots + t_{2p}x_{1p} & t_{21}x_{21} + \cdots + t_{2p}x_{2p} & \dots & t_{21}x_{n1} + \cdots + t_{2p}x_{np} \\ \vdots & \vdots & \ddots & \vdots \\ t_{p1}x_{11} + \cdots + t_{pp}x_{1p} & t_{p1}x_{21} + \cdots + t_{pp}x_{2p} & \dots & t_{p1}x_{n1} + \cdots + t_{pp}x_{np} \end{pmatrix}$$

With our transformed time series in hand, we take its transpose and then apply the DWT column by column of $\mathbf{Y}$. Noting the linearity of the DWT we obtain a matrix of detail vectors

$$\mathbf{D^Y} = \begin{pmatrix} t_{11}d_{11} + \cdots + t_{1p}d_{1p} & t_{21}d_{11} + \cdots + t_{2p}d_{1p} & \dots & t_{p1}d_{11} + \cdots + t_{pp}d_{1p} \\ t_{11}d_{21} + \cdots + t_{1p}d_{2p} & t_{21}d_{21} + \cdots + t_{2p}d_{2p} & \dots & t_{p1}d_{21} + \cdots + t_{pp}d_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ t_{11}d_{r1} + \cdots + t_{1p}d_{rp} & t_{21}d_{r1} + \cdots + t_{2p}d_{rp} & \dots & t_{p1}d_{r1} + \cdots + t_{pp}d_{rp} \end{pmatrix}.$$

We note the d_{ij}'s are exactly the same as those had we applied the DWT directly to the original time series. It follows $\mathbf{T}\mathbf{d}_{ij}^X = \mathbf{d}_{ij}^Y$ as claimed.

□

With the above claim in hand, we prove the following proposition which will be useful later when "standardizing" the time series in this chapter.

Theorem 5.3.1. *Given a linear transform, $\mathbf{T} \in GL_p(\mathbb{R})$ as in 5.2, then $p(\tau|\mathbf{X}) = p(\tau|\mathbf{Y})\ \forall\ \tau$. That is, the Bayesian-wavelet change-point estimation equation is invariant under dimension preserving linear transformations.*

Proof. As presented in chapter 3, we write our full model for the transformed time series as

$$p(\Sigma^Y, \tau, \mathbf{\Delta}^Y|\mathbf{Y}) = \frac{\prod_j \prod_k p(\mathbf{d}_{j,k}^Y|\Sigma^Y, \tau, \mathbf{\Delta})^Y p_0(\Sigma^Y, \mathbf{\Delta}^Y, \tau)}{m(\mathbf{Y})}. \tag{5.1}$$

In chapter 3, we only considered the proportionality form of this model which is all we required in our previous analysis. If we take care to include the denominator, however, the linear transformation introduces constant terms that ultimately cancel out. We handle the numerator and the denominator separately.

We first marginalize the numerator of equation 5.1. From calculations similar to those explicitly shown in the chapter 3, we obtain:

$$\int_{\mathbb{R}^{p+}\times\mathbb{R}^{p+}}\int_{\mathbb{R}^p}\prod_j\prod_k p(\mathbf{d}_{j,k}^Y|\Sigma^Y,\tau,\boldsymbol{\Delta})^Y p_0(\Sigma^Y,\boldsymbol{\Delta}^Y,\tau)d\boldsymbol{\Delta}d\Sigma^Y$$

$$=(2\pi)^{\frac{p-mp}{2}}(2)^{\frac{mp}{2}}\Gamma_p(\tfrac{m}{2})C^{-\frac{1}{2}}\left|\sum_j\sum_k \mathbf{d}_{jk}^Y\mathbf{d}_{jk}^{Y\,T}-\frac{1}{C}B^Y B^{Y^T}\right|^s. \tag{5.2}$$

By the above claim, we have

$$\mathbf{d}_{jk}^Y=\mathbf{T}\mathbf{d}_{jk}^X.$$

This means for a given jk we have

$$\mathbf{d}_{jk}^Y\mathbf{d}_{jk}^{Y\,T}-\frac{1}{C}B^Y B^Y T=\mathbf{T}\mathbf{d}_{jk}^X\mathbf{d}_{jk}^{X\,T}\mathbf{T}^T-\frac{1}{C}\mathbf{T}B^X B^{X^T}\mathbf{T}^T=\mathbf{T}(\mathbf{d}_{jk}\mathbf{d}_{jk}^{X\,T}-\frac{1}{C}B^X B^{X^T})\mathbf{T}^T$$

Applying this to equation (5.2)

$$(2\pi)^{\frac{p-mp}{2}}(2)^{\frac{mp}{2}}\Gamma_p(\tfrac{m}{2})C^{-\frac{1}{2}}\left|\sum_j\sum_k \mathbf{d}_{jk}^Y\mathbf{d}_{jk}^{Y\,T}-\frac{1}{C}B^Y B^{Y^T}\right|^s$$

$$=(2\pi)^{\frac{p-mp}{2}}(2)^{\frac{mp}{2}}\Gamma_p(\tfrac{m}{2})C^{-\frac{1}{2}}\left|\mathbf{T}\Big(\sum_j\sum_k \mathbf{d}_{jk}^X\mathbf{d}_{jk}^{X\,T}-\frac{1}{C}B^X B^{X^T}\Big)\mathbf{T}^T\right|^s$$

$$=(2\pi)^{\frac{p-mp}{2}}(2)^{\frac{mp}{2}}\Gamma_p(\tfrac{m}{2})C^{-\frac{1}{2}}|\mathbf{T}|^s\left|\Big(\sum_j\sum_k \mathbf{d}_{jk}^X\mathbf{d}_{jk}^{X\,T}-\frac{1}{C}B^X B^{X^T}\Big)\right|^s|\mathbf{T}^T|^s$$

$$=(2\pi)^{\frac{p-mp}{2}}(2)^{\frac{mp}{2}}\Gamma_p(\tfrac{m}{2})C^{-\frac{1}{2}}|\mathbf{T}|^{2s}\left|\Big(\sum_j\sum_k \mathbf{d}_{jk}^X\mathbf{d}_{jk}^{X\,T}-\frac{1}{C}B^X B^{X^T}\Big)\right|^s.$$

Next consider the denominator. Our calculation of $p(\mathbf{X}|\Sigma,\tau,\Delta)$ is similar to above except we must integrate over all our parameters, namely

$$m(\mathbf{Y})=\int L(\mathbf{D}^Y|\Sigma^Y,\boldsymbol{\Delta}^Y,\tau)p_0(\Sigma,\boldsymbol{\Delta}^Y,\tau)d\Sigma^Y d\boldsymbol{\Delta}^Y d\tau$$

$$= \sum_{\tau=2}^{N} (2\pi)^{\frac{p-mp}{2}} (2)^{\frac{mp}{2}} \Gamma_p(\tfrac{m}{2}) C^{-\frac{1}{2}} |\mathbf{T}|^{2s} \left| \Big(\sum_j \sum_k \mathbf{d}_{jk}^X \mathbf{d}_{jk}^{X^T} - \frac{1}{C} B^X B^{X^T} \Big) \right|^s$$

$$= |\mathbf{T}|^{2s} \sum_{\tau=2}^{N} (2\pi)^{\frac{p-mp}{2}} (2)^{\frac{mp}{2}} \Gamma_p(\tfrac{m}{2}) C^{-\frac{1}{2}} \left| \Big(\sum_j \sum_k \mathbf{d}_{jk}^X \mathbf{d}_{jk}^{X^T} - \frac{1}{C} B^X B^{X^T} \Big) \right|^s$$

Clearly the constant term, $|\mathbf{T}|^{2s}$, introduced by the linear transformation in the numerator and denominator cancel out. Therefore we have $p(\tau|\mathbf{X}) = p(\tau|\mathbf{Y}) \ \forall \ \tau$. □

The next proposition recalls covariance matrix diagonalization properties. Since we obtain our diagonal covariance matrix by linear transformations, theorem 5.3.1 ensures us that our ability to estimate the change-point location of our time series remains unchanged.

Proposition 5.3.2. *Given a multivariate time series* $\mathbf{X}$ *with covariance matrix* Σ^X*, then there exists a linear transformation,* $\mathbf{T}$*, such that* $\mathbf{T}\mathbf{X}^T = \mathbf{Y}^T$ *with an associated diagonal covariance matrix,* Σ^Y*. Furthermore,* Σ^Y *has a maximum eigenvalue of* $\lambda_{max} = 1$*.*

Proof. Recall that any covariance matrix is diagonalizable [61]. Therefore, there exists an orthogonal matrix, say $\mathbf{R}$, that diagonalizes Σ^X. In particular, for some diagonal matrix $\mathbf{D}$ and some rotation operator $\mathbf{R}$ we have

$$\Sigma^X = \mathbf{R}^T \mathbf{D} \mathbf{R}.$$

Now say $\mathbf{Z}^T = \mathbf{R}\mathbf{X}^T$, then by definition of the covariance matrix we may write the covariance matrix of our transformed time series as

$$\Sigma^Z = E[(\mathbf{Z}^T - E[\mathbf{Z}^T])(\mathbf{Z}^T - E[\mathbf{Z}^T])^T]$$

$$= E[(\mathbf{R}\mathbf{X}^T - E[\mathbf{R}\mathbf{X}^T])(\mathbf{R}\mathbf{X}^T - E[\mathbf{R}\mathbf{X}^T])^T]$$

$$= \mathbf{R}E[(\mathbf{X}^T - E[\mathbf{X}^T])(\mathbf{X}^T - E[\mathbf{X}^T])^T]\mathbf{R}^T$$

$$= \mathbf{R}\Sigma^X\mathbf{R}^T$$

$$= \mathbf{D}$$

Let $\lambda_{max}(\mathbf{D})$ denote the maximum eigenvalue of $\mathbf{D}$ and $\mathbf{T} = \frac{1}{\sqrt{\lambda_{max}(\mathbf{D})}}\mathbf{R}$. Set $\mathbf{T}\mathbf{X}^T = \mathbf{Y}^T$. We see the associated covariance matrix, $\mathbf{\Sigma}^Y$, is also diagonal, now with $\lambda_{max}(\Sigma^Y) = 1$ as desired.

□

Since we are trying to understand the efficacy of our Bayesian-wavelet change-point estimation equation, we need a way of comparing its application to different time series. The upshot of theorem 5.3.1 and proposition 5.3.2 is that from the standpoint of $p(\tau|\mathbf{X})$, for an arbitrary time series, there exists an equivalent time series with the covariance structure of $\mathbf{\Sigma}^Y$ as in proposition 5.3.2. That is, if we can understand the likelihood of $p(\tau|\mathbf{X})$ correctly estimating the change-point location for multi-dimensional time series with covariance structure of $\mathbf{\Sigma}^Y$, then in principal can we understand the likelihood of $p(\tau|\mathbf{X})$ correctly estimating the change-point location for all time series.

If the maximum eigenvalue of our covariance matrix is 1, in general all that can be said about the rest of the spectrum is that any other eigenvalue, λ, has a value such that $0 < \lambda \leq 1$. Since our purpose here is to develop a lower confidence bound for $p(\tau|\mathbf{X})$ correctly estimating the change-point location, we take the conservative approach and assume the covariance matrix is simply the identity. This assumption injects the maximum allowable variance into the time series making it only more difficult for $p(\tau|\mathbf{X})$ to correctly estimate the change-point location. Although this assumption results in some loss of precision when establishing subsequent confidence bounds, we may employ known results

from random matrix theory thereby simplifying calculations. In the remainder of this chapter we assume our original time series $\mathbf{X}$ has already been standardized with covariance matrix $\Sigma^X = I$, where I represents the identity matrix.

With the covariance matrix standardized to the identity of our original time series, we next focus on a dimensionality reduction of our time series. Letting $\mathbf{\Delta}^X$ represent the shift vector of our original time series, the next proposition tells us what happens to the shift vector after being acted upon by a dimensionality reducing linear transformation, $\mathbf{P}$.

Proposition 5.3.3. *Given a dimensionality reduction by linear transformation operator,* $\mathbf{P}$*, along with a shift vector,* $\mathbf{\Delta}^X$*, then the shift vector in the projected space becomes* $\mathbf{P}\mathbf{\Delta}^X = \mathbf{\Delta}^Y$*.*

Proof. Supposing the change-point of the time series occurs at time τ, let $\boldsymbol{\mu}_\tau^X$ and $\boldsymbol{\mu}_{\tau+1}^X$ represent the mean vector of the time series both before and after the change-point of the original time series, respectively. Similarly so for $\boldsymbol{\mu}_\tau^Y$ and $\boldsymbol{\mu}_{\tau+1}^Y$ in the projected space. We have by definition of the shift vector both before and after the transformation

$$\mathbf{\Delta}^X = \boldsymbol{\mu}_{\tau+1}^X - \boldsymbol{\mu}_\tau^X$$

and

$$\mathbf{\Delta}^Y = \boldsymbol{\mu}_{\tau+1}^Y - \boldsymbol{\mu}_\tau^Y.$$

We may represent $\mathbf{P}$ as a $d \times p$ matrix, so we obtain our dimensionally reduced

time series, $\mathbf{Y}$, by a matrix multiplication

$$\mathbf{PX}^T = \begin{pmatrix} p_{11} & p_{21} & \dots & p_{1p} \\ p_{12} & p_{22} & \dots & p_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1d} & p_{p2} & \dots & p_{dp} \end{pmatrix} \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \dots & x_{pn} \end{pmatrix}.$$

$$= \begin{pmatrix} \sum_{i=1}^{p} p_{i1}x_{i1} & \sum_{i=1}^{p} p_{i1}x_{i2} & \dots & \sum_{i=1}^{p} p_{i1}x_{in} \\ \sum_{i=1}^{p} p_{i2}x_{i1} & \sum_{i=1}^{p} p_{i2}x_{i2} & \dots & \sum_{i=1}^{p} p_{i2}x_{in} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{p} p_{id}x_{i1} & \sum_{i=1}^{p} p_{id}x_{i2} & \dots & \sum_{i=1}^{p} p_{id}x_{in} \end{pmatrix}$$

$$= \mathbf{Y}^T$$

Then using the definition of the expected value, we have

$$\boldsymbol{\Delta}^Y = \boldsymbol{\mu}^Y_{\tau+1} - \boldsymbol{\mu}^Y_{\tau} = \begin{pmatrix} \mu^Y_{\tau+1,1} \\ \mu^Y_{\tau+1,2} \\ \vdots \\ \mu^Y_{\tau+1,d} \end{pmatrix} - \begin{pmatrix} \mu^Y_{\tau,1} \\ \mu^Y_{\tau,2} \\ \vdots \\ \mu^Y_{\tau,d} \end{pmatrix}$$

$$= E\begin{pmatrix} \frac{1}{n-(\tau+1)} \sum_{j=\tau+1}^{n} y_{j1} \\ \frac{1}{n-(\tau+1)} \sum_{j=\tau+1}^{n} y_{j2} \\ \vdots \\ \frac{1}{n-(\tau+1)} \sum_{j=\tau+1}^{n} y_{jd} \end{pmatrix} - E\begin{pmatrix} \frac{1}{\tau} \sum_{j=1}^{\tau} y_{j1} \\ \frac{1}{\tau} \sum_{j=1}^{\tau} y_{j2} \\ \vdots \\ \frac{1}{\tau} \sum_{j=1}^{\tau} y_{jd} \end{pmatrix}$$

$$= E\begin{pmatrix} \dfrac{1}{n-(\tau+1)} \displaystyle\sum_{j=\tau+1}^{n}\sum_{i=1}^{p} p_{i1}x_{ij} \\ \dfrac{1}{n-(\tau+1)} \displaystyle\sum_{j=\tau+1}^{n}\sum_{i=1}^{p} p_{i2}x_{ij} \\ \vdots \\ \dfrac{1}{n-(\tau+1)} \displaystyle\sum_{j=\tau+1}^{n}\sum_{i=1}^{p} p_{id}x_{ij} \end{pmatrix} - E\begin{pmatrix} \dfrac{1}{\tau} \displaystyle\sum_{j=1}^{\tau}\sum_{i=1}^{p} p_{i1}x_{ij} \\ \dfrac{1}{\tau} \displaystyle\sum_{j=1}^{\tau}\sum_{i=1}^{p} p_{i2}x_{ij} \\ \vdots \\ \dfrac{1}{\tau} \displaystyle\sum_{j=1}^{\tau}\sum_{i=1}^{p} p_{id}x_{ij} \end{pmatrix}$$

$$= E\begin{pmatrix} \displaystyle\sum_{i=1}^{p} p_{i1} \sum_{j=\tau+1}^{n} \dfrac{x_{ij}}{n-(\tau+1)} \\ \displaystyle\sum_{i=1}^{p} p_{i2} \sum_{j=\tau+1}^{n} \dfrac{x_{ij}}{n-(\tau+1)} \\ \vdots \\ \displaystyle\sum_{i=1}^{p} p_{id} \sum_{j=\tau+1}^{n} \dfrac{x_{ij}}{n-(\tau+1)} \end{pmatrix} - E\begin{pmatrix} \displaystyle\sum_{i=1}^{p} p_{i1} \sum_{j=1}^{\tau} \dfrac{x_{ij}}{\tau} \\ \displaystyle\sum_{i=1}^{p} p_{i2} \sum_{j=1}^{\tau} \dfrac{x_{ij}}{\tau} \\ \vdots \\ \displaystyle\sum_{i=1}^{p} p_{id} \sum_{j=1}^{\tau} \dfrac{x_{ij}}{\tau} \end{pmatrix}$$

$$= \begin{pmatrix} p_{11} & p_{21} & \dots & p_{1p} \\ p_{12} & p_{22} & \dots & p_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1d} & p_{p2} & \dots & p_{dp} \end{pmatrix} E\begin{pmatrix} \displaystyle\sum_{j=\tau+1}^{n} \dfrac{x_{1j}}{n-(\tau+1)} \\ \displaystyle\sum_{j=\tau+1}^{n} \dfrac{x_{2j}}{n-(\tau+1)} \\ \vdots \\ \displaystyle\sum_{j=\tau+1}^{n} \dfrac{x_{pj}}{n-(\tau+1)} \end{pmatrix} - \begin{pmatrix} p_{11} & p_{21} & \dots & p_{1p} \\ p_{12} & p_{22} & \dots & p_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1d} & p_{p2} & \dots & p_{dp} \end{pmatrix} E\begin{pmatrix} \displaystyle\sum_{j=1}^{\tau} \dfrac{x_{1j}}{\tau} \\ \displaystyle\sum_{j=1}^{\tau} \dfrac{x_{2j}}{\tau} \\ \vdots \\ \displaystyle\sum_{j=1}^{\tau} \dfrac{x_{pj}}{\tau} \end{pmatrix}.$$

$$= \mathbf{P}(\boldsymbol{\mu}^X_{\tau+1} - \boldsymbol{\mu}^X_{\tau})$$

$$= \mathbf{P}\boldsymbol{\Delta}^X$$

□

So we determine our shift vector after a linear dimensionality reduction simply by a matrix multiplication. For the remainder of this chapter we assume a particular form of our $d \times n$ projection matrix, $\mathbf{P}$, namely

$$\mathbf{P} = \frac{1}{\sqrt{d}}\mathbf{A} \tag{5.3}$$

where $\mathbf{A}$ is a random matrix with each entry, a_{ij}, randomly distributed as $a_{ij} \sim N(0,1)$.

Next observe that if the covariance matrix is the identity, I, (or cI for $c \in \mathbb{R}$), then it is only the magnitude of $\mathbf{\Delta}^X$ that effects $p(\tau|\mathbf{X})$. This is another consequence of theorem 5.3.1. We note a rotational transform leaves the covariance matrix unchanged while the components of $\mathbf{\Delta}^X$ can be rotated into any proportion across dimensions so long as the magnitude remains unchanged.

If $\mathbf{\Delta}^X$ is known and the projection matrix is random, we can use similar calculations as in proposition 5.3.3 to determine the expected value of $||\mathbf{\Delta}^Y||^2$. Letting $\mathbf{\Delta}^X = (\delta_1^X, \delta_2^X, \ldots, \delta_p^x)^T$ and $\mathbf{\Delta}^Y = (\delta_1^Y, \delta_2^Y, \ldots, \delta_d^Y)^T$, we have our next proposition.

Proposition 5.3.4. *Given a random matrix,* $\mathbf{P}$*, as in equation (5.3) and* $\mathbf{\Delta}^Y$ *as in proposition 5.3.3. then* $E||\mathbf{\Delta}^Y||^2 = ||\mathbf{\Delta}^X||^2$

Proof. $E||\mathbf{\Delta}^Y||^2 = E[(\mathbf{\Delta}^Y)^T(\mathbf{\Delta}^Y] = E[(\mathbf{\Delta}^X)^T\mathbf{P}^T\mathbf{P}(\mathbf{\Delta}^X]$

$$= E\left(\delta_1^X, \delta_2^X, \ldots \delta_p^X\right) \frac{1}{\sqrt{d}} \begin{pmatrix} a_{11} & a_{21} & \ldots & a_{d1} \\ a_{12} & a_{22} & \ldots & a_{d2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1p} & a_{2p} & \ldots & a_{dp} \end{pmatrix} \frac{1}{\sqrt{d}} \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1p} \\ a_{21} & a_{22} & \ldots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{d1} & a_{d2} & \ldots & a_{dp} \end{pmatrix} \begin{pmatrix} \delta_1^X \\ \delta_2^X \\ \vdots \\ \delta_p^X \end{pmatrix}$$

$$= \left(\delta_1^X, \delta_2^X, \ldots \delta_p^X\right) \frac{1}{d} \begin{pmatrix} E\Big[\sum_{j=1}^d a_{1j}^2\Big] & E\Big[\sum_{j=1}^d a_{1j}a_{2j}\Big] & \ldots & E\Big[\sum_{j=1}^d a_{1j}a_{dj}\Big] \\ E\Big[\sum_{j=1}^d a_{2j}a_{1j}\Big] & E\Big[\sum_{j=1}^d a_{2j}^2\Big] & \ldots & E\Big[\sum_{j=1}^d a_{2j}a_{dj}\Big] \\ \vdots & \vdots & \ddots & \vdots \\ E\Big[\sum_{j=1}^d a_{dj}a_{1j}\Big] & E\Big[\sum_{j=1}^d a_{dj}a_{2j}\Big] & \ldots & E\Big[\sum_{j=1}^d a_{dj}^2\Big] \end{pmatrix} \begin{pmatrix} \delta_1^X \\ \delta_2^X \\ \vdots \\ \delta_p^X \end{pmatrix}$$

$$= \left(\delta_1^X, \delta_2^X, \dots \delta_p^X\right) I \begin{pmatrix} \delta_1^X \\ \delta_2^X \\ \vdots \\ \delta_p^X \end{pmatrix}$$

$$= ||\Delta^X||^2$$

□

The difficulty of the change-point problem depends not only on the shift vector, but also on the covariance structure. We, therefore, need to understand what happens to our covariance matrix under a random projection as well. Since we are assuming the covariance matrix of our original time series has been transformed in such a way that $\Sigma^X = I$, this step is straightforward.

Proposition 5.3.5. *Under our random projection, the expected covariance matrix of our transformed time series is*

$$E[\Sigma^Y] = \begin{pmatrix} \frac{p}{d} & 0 & \dots & 0 \\ 0 & \frac{p}{d} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{p}{d} \end{pmatrix}$$

Proof. We have by our above definitions

$$E[\Sigma^Y] = E[\mathbf{P}\Sigma^X\mathbf{P}^T] = E[\tfrac{1}{\sqrt{d}}\mathbf{A} I \tfrac{1}{\sqrt{d}}\mathbf{A}^T]$$

$$= E\left[\frac{1}{d}\begin{pmatrix} a_{11} & a_{21} & \dots & a_{p1} \\ a_{12} & a_{22} & \dots & a_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1d} & a_{p2} & \dots & a_{pd} \end{pmatrix}\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1d} \\ a_{21} & a_{22} & \dots & a_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \dots & a_{pd} \end{pmatrix}\right]$$

$$= \frac{1}{d} \begin{pmatrix} E\Big[\sum_{i=1}^{p} a_{i1}^2\Big] & E\Big[\sum_{i=1}^{p} a_{i1}a_{i2}\Big] & \dots & E\Big[\sum_{i=1}^{p} a_{i1}a_{id}\Big] \\ E\Big[\sum_{i=1}^{p} a_{i2}a_{i1}\Big] & E\Big[\sum_{i=1}^{p} a_{i2}^2\Big] & \dots & E\Big[\sum_{i=1}^{p} a_{i2}a_{id}\Big] \\ \vdots & \vdots & \ddots & \vdots \\ E\Big[\sum_{i=1}^{p} a_{id}a_{i1}\Big] & E\Big[\sum_{i=1}^{p} a_{id}a_{i2}\Big] & \dots & E\Big[\sum_{i=1}^{p} a_{id}a_{id}\Big] \end{pmatrix}$$

$$= \begin{pmatrix} \frac{p}{d} & 0 & \dots & 0 \\ 0 & \frac{p}{d} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{p}{d} \end{pmatrix}$$

Where the last step follows by the properties of the χ^2 distribution and expectation of independent normal random variables.

□

From propositions 5.3.4 and 5.3.5 we attain expected values for the shift vector and covariance matrix structure under a dimensionality reduction by linear transformation. Next we need to establish confidence bounds for these expected values. The Johnson-Lindenstrauss lemma establishes the existence of a dimension reduction mapping with only minor distortion in the Euclidian metric sense. Several proofs of the Johnson-Lindenstrauss lemma exist in the literature [62, 63, 64]. Vempala [65] establishes theorem 5.3.6 and then takes a union bound to prove the Johnson-Lindenstrauss lemma. We modify the proof slightly to obtain somewhat better results in terms of how far we can reduce the dimension of our original time series and still maintain the same confidence bound. We begin with the following claim.

Claim.

$$\frac{||\sqrt{d}\mathbf{P}\Delta^X||^2}{||\mathbf{\Delta}^X||^2} \sim \chi^2_{(d)}$$

where $\chi^2_{(d)}$ is the χ^2 distribution with d degrees of freedom.

Proof. Firstly, observe $\langle\sqrt{d}\mathbf{P}_j, \mathbf{\Delta}^X\rangle/||\mathbf{\Delta}^X||\sim N(0,1)$ where $\mathbf{P}_j$ represents the jth row of the projection matrix $\mathbf{P}$. This follows since

> $\langle\sqrt{d}\mathbf{P}_j, \mathbf{\Delta}^X\rangle/||\mathbf{\Delta}^X||$ is the sum of normally distributed random variables and hence must also be normally distrubted.

$$E[\langle\sqrt{d}\mathbf{P}_j, \mathbf{\Delta}^X\rangle/||\mathbf{\Delta}^X||] = \tfrac{1}{||\mathbf{\Delta}^X||}(\delta_1^X E[a_{j1}] + \delta_1^X E[a_{j2}] + \ldots \delta_p^X E[a_{jp}]) = 0$$

$$Var(\langle\sqrt{d}\mathbf{P}_j, \mathbf{\Delta}^X\rangle/||\mathbf{\Delta}^X||) = E[(\sqrt{d}\langle\mathbf{P}_j, \mathbf{\Delta}^X\rangle/||\mathbf{\Delta}^X||)^2] - E[\langle\sqrt{d}\mathbf{P}_j, \mathbf{\Delta}^X/||\mathbf{\Delta}^X||\rangle]^2$$
$$= \tfrac{1}{||\mathbf{\Delta}^X||^2}(\textstyle\sum_{1\leq l,m\leq p}\delta_l^X\delta_m^X E[a_{jl}a_{jm}] - 0 = \tfrac{||\mathbf{\Delta}^X||^2}{||\mathbf{\Delta}^X||^2} = 1$$

Next observe

$$||\sqrt{d}\mathbf{P}\mathbf{\Delta}^X||^2 = \sum_{j=1}^{d}\langle\sqrt{d}\mathbf{P}_j, \mathbf{\Delta}^X\rangle^2$$

Therefore as the sum of squares of standard normal random variables, we have

$$||\sqrt{d}\mathbf{P}_j\mathbf{\Delta}^X||^2/||\mathbf{\Delta}^X||^2\sim \chi^2_{(d)}$$

as claimed.

□

The above claim helps establish the following proposition which is a key component of the Johnson-Lindenstrauss lemma.

Theorem 5.3.6. *Let $||\mathbf{\Delta}^X||^2$ be the squared magnitude of our shift vector for a high dimensional time series and let $\mathbf{P}$ be a $d \times n$ dimension reduction matrix. Then so long as $d > \frac{2\log\alpha}{\log(1-\varepsilon)-\varepsilon}$, we are at least $100(1-\alpha)\%$ confident that $||\mathbf{\Delta}^Y||^2 \geq (1-\varepsilon)||\mathbf{\Delta}^X||^2$.*

Proof.

$$
\begin{aligned}
P(||\mathbf{\Delta}^Y||^2 < (1-\varepsilon)||\mathbf{\Delta}^X||^2) &= P(||\mathbf{P}\mathbf{\Delta}^X||^2 < (1-\varepsilon)||\mathbf{\Delta}^X||^2) \\
&= P(||\frac{1}{\sqrt{d}}\mathbf{A}\mathbf{\Delta}^X||^2 < (1-\varepsilon)||\mathbf{\Delta}^X||^2) \\
&= P(\frac{||\mathbf{A}\mathbf{\Delta}^X||^2}{||\mathbf{\Delta}^X||^2} < d(1-\varepsilon)) \\
&= P(\chi^2_{(d)} < d(1-\varepsilon)) \\
&= P(\sum_{k=1}^{d} a_k^2) < d(1-\varepsilon) \text{ where } a_k \sim N(0,1) \\
&= P(e^{t\sum_{k=1}^{d} a_k^2)} < e^{td(1-\varepsilon)} \text{ where } t > 0 \\
&\leq E[e^{t\sum_{k=1}^{d} a_k^2)}]/e^{td(1-\varepsilon)} \text{ by Markov's inequality} \\
&= E[(e^{ta^2)})]^k/e^{td(1-\varepsilon)} \text{ by } a_k's \text{ are } iid \\
&= (1-2t)^{-d/2}/e^{td(1-\varepsilon)} \text{ MGF of } \chi^2 \text{ distribution}
\end{aligned}
\tag{5.4}
$$

We note the last equation holds for $0 < t < \frac{1}{2}$ by the properties of the moment generating function for a χ^2 distribution. Since we are seeking a lower bound, we can minimize the above result for t with basic calculus techniques. Explicitly, we solve for t

$$\frac{d}{dt}(1-2t)^{-d/2}/e^{td(1-\varepsilon)} = 0$$

$$-d(1-\varepsilon)e^{-td(1-\varepsilon)}(1-2t)^{-d/2} + de^{-td(1-\varepsilon)}(1-2t)^{-d/2-1} = 0$$

$$\frac{1}{1-\varepsilon} = 1-2t$$

$$t = \frac{\varepsilon}{2(1-\varepsilon)}.$$

Now substituting this value into the above we get

$$P(||\mathbf{\Delta}^Y||^2 < (1-\varepsilon)||\mathbf{\Delta}^X||^2) \leq (1-\varepsilon)^{d/2}e^{-d\varepsilon/2}$$

A more convenient form for our subsequent calculations uses the common exponent terms.

$$(1-\varepsilon)^{d/2}e^{-d\varepsilon/2} = (e^{\log(1-\varepsilon)}e^{-\varepsilon})^{d/2}$$

Notice this bound does not depend upon the dimension of our original time series. So if we want to be $1-\alpha$ confident that $||\mathbf{\Delta}_y||^2$ is of magnitude $(1-\epsilon)||\mathbf{\Delta}_x||^2$ or greater we can consider

$$1-\alpha \geq 1-(e^{\log(1-\varepsilon)}e^{-\varepsilon})^{d/2}$$

If we rearrange this inequality to isolate d we obtain

$$\frac{2\log\alpha}{\log(1-\varepsilon)-\varepsilon} \geq d$$

In other words if

$$\frac{2\log\alpha}{\log(1-\varepsilon)-\varepsilon} < d$$

then we are at least $1-\alpha$ confident that $||\mathbf{\Delta}_y||^2$ is of magnitude $(1-\epsilon)||\mathbf{\Delta}_x||^2$ or greater. □

Establishing some control over the magnitude of the shift vector after a dimension reduction is only half the problem. We also need to understand how the covariance structure is affected. While in proposition 5.3.5 we established the form of our expected covariance matrix, $\mathbf{\Sigma}^Y$, in the lower dimensional space, we provided no associated confidence level. In particular, to apply our signal to noise ratio rule of thumb, we need the value of the largest expected eigenvalue, λ_{max}, of $\mathbf{\Sigma}^Y$ along with a confidence level for this value. Recalling our proof of proposition 5.3.5, we note this is equivalent to finding the expected value and confidence level

for the largest singular value, s_{max}, of $\mathbf{P}$. Here we can appeal to known results from random matrix theory.

Proposition 5.3.7. *Given a random $d \times n$ dimension reduction matrix, $\mathbf{P}$, then with probability of at least $1 - 2\exp(-t^2/2)$ and for every $t \geq 0$*

$$s_{max} \leq \sqrt{\frac{p}{d}} + 1 + \frac{t}{\sqrt{d}}$$

where s_{max} denotes the largest singular value of $\mathbf{P}$.

Proof. Vershynin [66] provides a proof of this proposition which is based on Gordon's theorem for Gaussian matrices and Slepian's inequality. □

From the empirical studies provided in chapter 3, we found for an $SNR > 3$ that simulations resulted in 1000 out of 1000 correct change-point estimations. These experiments were conducted for a variety parameter combinations for dimensions 100 and below always yielding the same results. This SNR value is actually quite conservative as very good results are also observed for SNR values much lower than 3 for many cases. Maintaining a conservative approach, however, we choose the value $SNR = 3$ as our benchmark to judge with near certainty that $p(\tau|\mathbf{X})$ correctly estimates the true change-point location.

Applying this to our problem at hand, if our dimension reduction results in a 100 dimensional or lower time series with a calculated SNR of 3.0 or greater, then we say with very high confidence that $p(\tau|\mathbf{X})$ correctly estimates the change-point location. The question still remains, however, how certain are we of our computed SNR? By assuming independence between the shift vector and covariance structure, a lower confidence bound for our SNR simply becomes the product of our computed lower confidence bounds for the size of the shift vector and largest eigenvalue of the covariance matrix after the dimensionality reduction by linear transformation, respectively.

5.4 Putting it all together

Recall, the goal of this chapter has been to estimate the change-point location of a high dimensional time series by applying our Bayesian-wavelet change-point estimation equation, $p(\tau|\mathbf{X})$. When we directly apply $p(\tau|\mathbf{X})$ to very high dimensional time series, however, we encounter numerical difficulties requiring alternative methods. Our approach to handle high dimensional time series situations has been to randomly project our time series into a lower more computationally tractable dimension. While the dimension reduction step is straight forward, we needed to develop a confidence level for correctly estimating the change-point location in the now dimensionally reduced space. The following summary succinctly ties our propositions and theory together to show how we acquire this confidence level.

- By theorem 5.3.1 and proposition 5.3.2 we "standardize" the covariance matrix of our time series. This standardization establishes a baseline to compare the difficulty of the change-point against for any time series and allows us to use known dimension reduction results that we use for later establishing confidence levels.

- By proposition 5.3.3, we obtain a formula for the shift vector in the dimensionally reduced space.

- Proposition 5.3.4 and theorem 5.3.6 are used to obtain the expected value of the shift value after a dimension reduction along wth a confidence bound for its squared magnitude.

- By propositions 5.3.5 and 5.3.7 we obtain our expected covariance matrix and an upper confidence bound for our largest singular value after the

dimension reduction.

· We compute our SNR based on the shift vector, dimension of the reduced time series, and largest singular value of the covariance matrix of our time series. We observe our SNR rule of thumb obtained by empirical evidence as presented in Chapter 4. If the calculated SNR is 3 or greater we may assume we correctly estimate the change-point location of a time series.

· We obtain a confidence level of our SNR by multiplying the confidence bound results obtained from theorem 5.3.6 and proposition 5.3.7.

5.5 Worked Numerical Example

Consider a time series of length 128 which consists of 256×256 pixel images of John Lennon. We generate an additive noise component with the identity matrix covariance structure. We may consider this sequence of images as a $256^2 = 65,536$ dimensional time series which in practice is far too high a dimension to directly apply $p(\tau|\mathbf{X})$ to estimate the change-point. Next, we inject a change-point at time point 95 indicated by the small dark box in the lower left side of the right image in Figure 5.1. The question now becomes, how far can we dimensionally reduce the time series and still correctly estimate the correct change-point location? We provide detailed calculations for how far we can apply our theory to estimate the change-point in the this high dimensional time series.

· The small dark box in the right figure corresponds to a shift vector magnitude of $\mathbf{\Delta}^X = 350$.

· Let $\alpha_1 = .025$ and $\varepsilon = .1$. We compute the lowest dimension to which we can reduce our time series and still be $100(1 - \alpha_1)\%$ confident that our new

shift vector will be least $(1-\varepsilon)350^2 \approx 332^2$ in magnitude or greater. Applying theorem 5.3.6 we make the following calculation:

$$\frac{2\log\alpha}{\log(1-\varepsilon)-\varepsilon} \approx 36$$

- Next we say we want to be 97.5% confident of our largest eigenvalue from the dimensionally reduced covariance matrix. Applying proposition 5.3.7 we make the following calculation for t

$$t = \sqrt{-2\log(1-.975)/2} \approx 3$$

- Applying proposition 5.3.7 once more we compute s_{max} as

$$s_{max} \leq \sqrt{\frac{256^2}{36}} \approx 43 \tag{5.5}$$

- We compute our SNR as $\frac{332}{36^{1/4}43} \approx 3.15$.

- Based on our empirical rule of thumb, because the $SNR > 3$, we have very high confidence of correctly estimating the change-point location in the time series after a dimension reduction.

- The posterior distribution plot correctly indicates the change-point at time point 95. Notice the sharpness of the peak consistent with what we expect with such a high SNR.

5.6 Conclusion

In this chapter we applied the Bayesian-wavelet change-point estimation equation to high dimensional data. While numerical complications prevent a

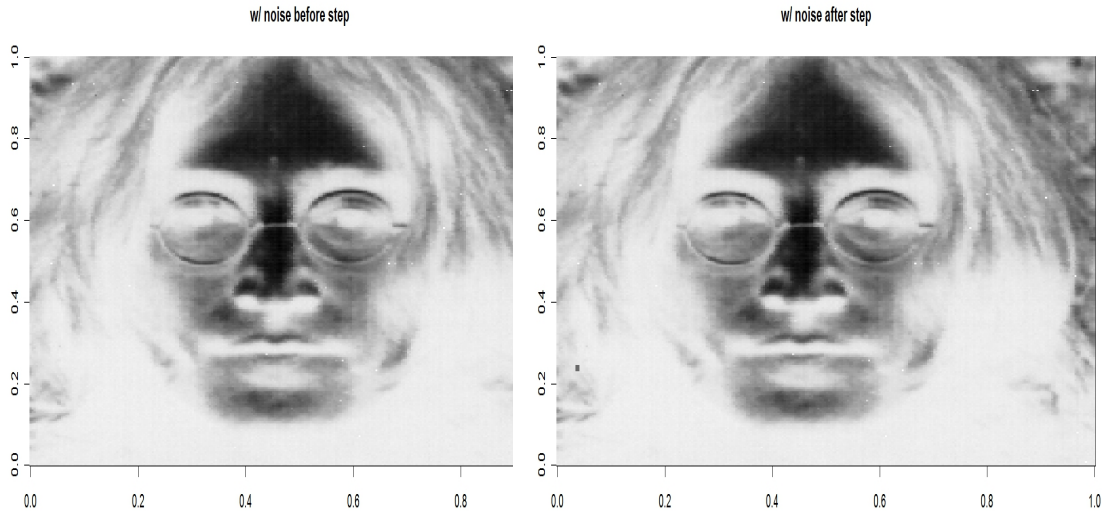

Figure 5.1: Notice the small dark box on the right picture which is the shift.

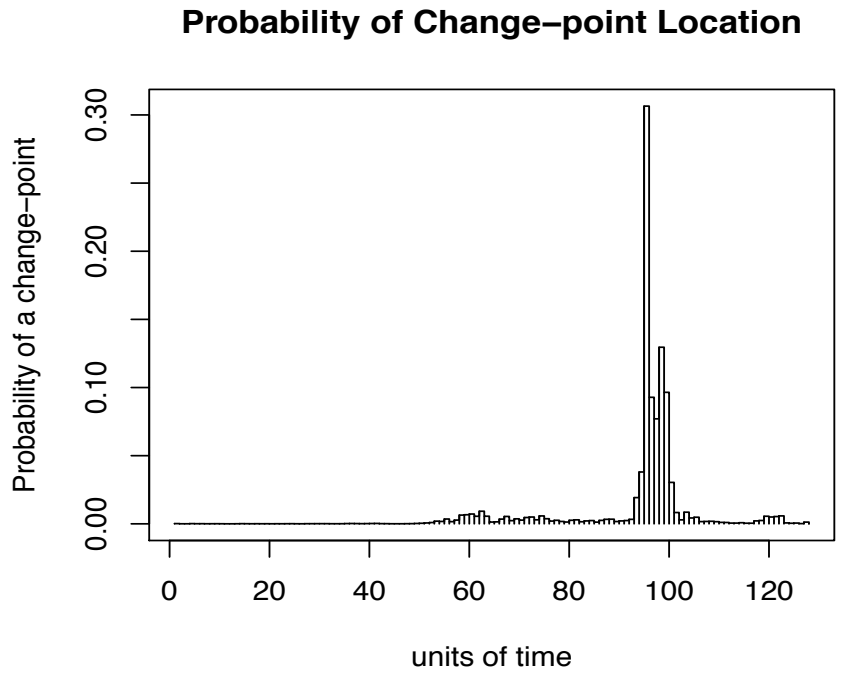

Figure 5.2: Posterior plot of change-point location correctly estimating the change-point at point 95 with an 80% credible interval of 90-100.

direct application to very high dimensional time series, we are able to project down to a lower dimension space with a random projection and apply $p(\tau|\mathbf{X})$ in

this approximation space. A detailed proof was given establishing a lower confidence bound for our certainty of correctly estimating the change-point location with $p(\tau|\mathbf{X})$.

In fact the confidence bound we provide is very conservative. Experimental evidence suggests in reality we can usually project down to even lower dimensional spaces with at least as high of confidence bound for our computed SNR. A theoretical limitation is that there are not a lot of nonasymptotic random matrix results known. To take advantage of what is known while being assured of our results, we make a few conservative assumptions along the way. In particular, after our standardization step we assume the maximum amount of variance into our covariance structure by assuming $\Sigma^X = I$. While this simplifies subsequent random projection calculations, we pay the price of some precision loss to our confidence bounds.

Despite our assumptions, we still find we can very accurately estimate the change-point location for rather small shifts in high dimensional time series. In very high dimensional space what may visually appear to be a very small shift could still be somewhat large in the Euclidian sense. In our high dimensional time series example, we represented a time series with a sequence of 256×256 pixel image of John Lennon. While the shift appeared very small to the naked eye, in the Euclidian sense the shift was still large enough to be largely retained in a projection all the way down to 36 dimensions. In fact, this example was rather contrived so that a shift would be visible for sake of illustration. Other examples are available where the shift is essentially invisible upon visual inspection and yet still easily and accurately detected by $p(\tau|\mathbf{X})$.

As a final note, we point out that our results in this chapter are largely theoretical. In practice we likely will know neither the shift vector magnitude nor the covariance structure of the high dimensional time series in advance. Often,

however, either historical data or prior knowledge allows us to make certain estimations or reasonable assumptions about the covariance structure of the data. With an estimation of the covariance structure in hand, then we standardize the time series. Our results then tell us how large the magnitude of the shift vector must be to accurately estimate the change-point location in the time series. In this sense the properties of the Bayesian-wavelet change-point estimation equation provide a useful diagnostic tool to estimate the change-point location of a high dimensional time series.

Chapter 6

Reversible Jump Markov Chain Monte Carlo in the multidimensional change-point problem

6.1 Introduction

For our final chapter we once again investigate the multivariate change-point problem, only now we apply a fully Bayesian approach. Throughout this chapter we present applications of this method to statistical process control (SPC). When a multivariate control chart raises an out-of-control signal, several diagnostic questions arise. When did the change occur? Which components or quality characteristics changed? For those components whose mean shifted, what are the new values for the mean? While methods exist for addressing these questions individually, we present a Bayesian approach that addresses all three questions in a single model. We employ Markov chain Monte Carlo (MCMC) methods in a Bayesian analysis that can be used in a unified approach to the diagnostics questions for multivariate charts. We demonstrate how a reversible jump Markov chain Monte Carlo (RJMCMC) approach can be used to infer (1)

the change point, (2) the change model (i.e., which components changed), and (3) post-change estimates of the mean.

6.2 Background

In this section we outline the important theoretical underpinnings of our model based on results originally presented by [67]. We begin by considering $M_k,\ k = 1, 2, \ldots, l$ possible models and denote the parameter vector for model M_k as $\boldsymbol{\theta}_k$. The parameter $\boldsymbol{\theta}_k$ will contain the change point τ along with the values of whatever other parameters are considered unknown (this could include the mean vector after the change, the mean vector before and after after the change, or the covariance matrix). The dimension of $\boldsymbol{\theta}_k$ can vary, as it depends on how many components have shifted. For example, if the before-shift parameters are assumed known, then for a four-dimensional process where only the first component changed we would have $\boldsymbol{\theta}_1 = \left(\tau, \theta^{(1)}_{1,\text{after}}\right)$ whereas for a four-dimensional process where all four components shifted we would have $\boldsymbol{\theta}_2 = \left(\tau, \theta^{(2)}_{1,\text{after}}, \ldots, \theta^{(2)}_{4,\text{after}}\right)$. Note that $\theta^{(1)}_{1,\text{after}}$ and $\theta^{(2)}_{1,\text{after}}$ represent the after-shift values for the mean of the first component under M_1 (shift in component 1 only) and model 2 (shift in all four components); these are different parameters.

Then given model M_k and observed data $\mathbf{D}$ a Bayesian analysis produces a posterior distribution, $\pi_k(\boldsymbol{\theta}_k|\mathbf{D}, M_k)$ which is expressed as

$$\pi_k(\boldsymbol{\theta}_k|\mathbf{D}, M_k) = \frac{L(\mathbf{D}|\boldsymbol{\theta}_k, M_k)p_0(\boldsymbol{\theta}_k|M_k)}{p_k(\mathbf{D}|M_k)}$$

where L is the likelihood function and p_0 is a prior distribution for the parameter vector $\boldsymbol{\theta}_k$ given M_k. Furthermore, p_k is the probability of observing $\mathbf{D}$ given M_k (which does not depend upon $\boldsymbol{\theta}_k$). We note that $p_k(\mathbf{D}|M_k)$ is often a high dimensional integral obtained by marginalizing over the parameter space that essentially becomes a normalizing constant. We can extend this approach to a

framework that simply allows us to treat our model, M_k, as an additional parameter. Thus, given our data and model space we obtain

$$\pi(M_k, \boldsymbol{\theta}_k|\mathbf{D}) \propto L(\mathbf{D}|M_k, \boldsymbol{\theta}_k)p_0(M_k, \boldsymbol{\theta}_k) = L(\mathbf{D}|M_k, \boldsymbol{\theta}_k)p_0(\boldsymbol{\theta}_k|M_k)p_0(M_k). \quad (6.1)$$

If the parameter state space for a given model is $\mathbb{S}_k$, then we express our transdimensional model state space as $\mathbb{S} = \bigcup_{k=1}^{l}\{M_k\} \times \mathbb{S}_k$. The goal is to sample from $\mathbb{S}$ in an MCMC fashion to produce a chain that converges to the posterior distribution. From this chain, we then draw parameter inference using familiar Bayesian analysis methods [68]. If the current state of our chain is $x = (M_k, \boldsymbol{\theta}_k)$, then we apply the following reversible jump Markov chain Monte Carlo (RJMCMC) algorithm (whose details we make clear throughout the article) to move to the next state $x' = (M_{k'}, \boldsymbol{\theta}_{k'})$:

1. Choose initial conditions, $x_0 = (M_{k_0}, \boldsymbol{\theta}_{k_0})$.

2. Perform a standard within model Metropolis-Hastings update of only the parameter vector $\boldsymbol{\theta}_{k_0}$. Still within model k_0, label this new state x.

3. Propose to "jump" to model $M_{k'}$ with conditional probability density $j(M_{k'}|M_k)$.

4. Generate a random vector u from a predetermined probability distribution g (note the dimension of u will depend upon both the dimension of the current and proposed models). This step accounts for the difference in dimensions as the chain jumps between models and makes our chain reversible throughout $\mathbb{S}$. Note, u' and g' analogously denote a random vector and probability distribution within the proposed model which we apply below in equation (6.2).

5. Propose a new state x' from a deterministic differentiable bijective function $h(x, u)$.

6. Accept the proposed move with probability $\alpha(x, x')$.

7. Repeat steps 2-6 until the chain converges. Once the chain converges repeat an additional number of times as necessary to obtain information about the posterior distribution.

Ergodic theory requires our chain to be aperiodic, irreducible and for the detailed balance equation to be satisfied [67]. These requirements will ensure probability equilibrium for moving between states and (eventual) convergence of our chain. Given we are in state space $x = (M_k, \boldsymbol{\theta}_k)$ and propose to move to state space $x' = (M_{k'}, \boldsymbol{\theta}_{k'})$, Green [67] showed the detailed balance requirement is satisfied if the following holds

$$\int_{\mathcal{S}_k \times \mathcal{S}_{k'}} \pi(x) j(M_k|M_{k'}) g(u) \alpha(x, x') dx du = \int_{\mathcal{S}_k \times \mathcal{S}_{k'}} \pi(x') j(M_{k'}|M_k) g'(u') \alpha(x', x) dx' du'. \tag{6.2}$$

That is, for our chain to possess the necessary convergence and mixing properties for parameter estimation, equation (6.2) must hold. By applying a simple change of variables calculation, we see equation (6.2) holds if

$$\pi(x) j(M_k|M_{k'}) g(u) \alpha(x, x') = \pi(x') j(M_{k'}|M_k) g'(u') \alpha(x', x) \left| \frac{\partial(x', u')}{\partial(x, u)} \right|$$

where the right-most term denotes the absolute value of the determinant of the Jacobian. This suggests a ratio form for the usual Metroplis-Hastings calculation. Namely, we set our probability for accepting the proposed move as

$$\alpha(x, x') = \min \left\{ 1, \frac{\pi(x') j(M_{k'}|M_k) g'(u')}{\pi(x) j(M_k|M_{k'}) g(u)} \left| \frac{\partial(x', u')}{\partial(x, u)} \right| \right\}. \tag{6.3}$$

We recall a necessary condition for our transformation to be bijective (i.e., one-to-one and onto) is that the determinant of the Jacobian must be nonzero [69]. This condition in turn requires a dimension matching property to hold between our current and proposed states. Explicitly, if we let n_k and r_k represent the dimensions of our current state and random vector, respectively, and $n_{k'}$ and $r_{k'}$ the dimension of our proposed state and dimension of the random vector for the reverse move, we require

$$n_k + r_k = n_{k'} + r_{k'}. \tag{6.4}$$

For the dimension matching criteria condition to be met, we must choose a suitable probability density function, g, to sample our random vector from. Additionally, we also need a corresponding bijective function, h, to map from our current to proposed state [68]. Often obtaining suitable g and h represents the main challenge for a successful implementation of RJMCMC methods [70]. In our application of RJMCMC to the multivariate change-point problem, we demonstrate how a particularly simple choice of g and h efficiently samples $\mathcal{S}$ and provides accurate parameter inference.

We now have all the theoretical machinery we will need to apply RJMCMC to the multivariate change-point problem. Although a complete understanding of the above formalism requires a careful review of the article by Green [67], the implementation of RJMCMC can be straightforward.

6.3 RJMCMC in the multivariate change-point problem

Consider a multivariate time series of independent elements

$$\mathbf{x}_i \sim N_p(\boldsymbol{\mu}_i, \Sigma), \quad 1, 2, \ldots, N$$

where $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_\tau \neq \boldsymbol{\mu}_{\tau+1} = \ldots \cdots = \boldsymbol{\mu}_N$ for some unknown τ.

We choose mean vector representatives $\boldsymbol{\mu}_{\tau,k}$ and $\boldsymbol{\mu}_{\tau+1,k}$ to denote our initial mean vector and post-change mean vector for model k, respectively. Perhaps more precisely we could write $\boldsymbol{\mu}_{\tau_k,k}$ and $\boldsymbol{\mu}_{\tau_k+1,k}$ instead of $\boldsymbol{\mu}_{\tau,k}$ and $\boldsymbol{\mu}_{\tau+1,k}$ since the change-point parameter is model specific, but we choose to suppress the additional subscript k throughout the remainder of this article for ease of notation. Additionally, we assume we have a constant covariance structure throughout the time series. Through our analysis, we seek estimates for the the model, the change-point location, the mean vector both before and after the change-point, and the covariance matrix. Explicitly, the parameters of interest of our time series are:

k: An index parameter denoting a particular model

M_K: The true model of the time series

τ_k: Change-point location of the time series for model k

$\boldsymbol{\mu}_{\tau,k}$: Mean vector before the change-point for model k

$\boldsymbol{\mu}_{\tau+1,k}$: Mean vector after the change-point for model k

Σ: Covariance matrix of our time series (assumed constant throughout)

In this section our aim is to develop an explicit expression for equation (6.1) that we will later implement with real data in section 6.4. We note that equation (6.1) may be equivalently expressed as

$$\pi(M_k, \boldsymbol{\theta}_k|\mathbf{D}) \propto \pi_k(\boldsymbol{\theta}_k|\mathbf{D}) p_0(M_k) \tag{6.5}$$

where π_k denotes the posterior distribution of $\boldsymbol{\theta}_k$ within M_k. While developing an expression for π_k represents the majority of our effort in this section, we mention two possible prior distributions we could put on the model space. For the first and simpler case we assume no prior knowledge of the model space and prescribe equal probability to each model. If we choose such a flat prior, equation (6.5) becomes

$$\pi(M_k, \boldsymbol{\theta}_k|\mathbf{D}) \propto \pi_k(\boldsymbol{\theta}_k|\mathbf{D}). \tag{6.6}$$

We note that the above "flat prior" is in some sense not so flat. There are substantially more models that undergo a shift in $p/2$ dimensions than models that undergo shifts in fewer dimensions (or higher dimensions). This difference becomes ever more pronounced as we increase the dimension of our data sets. For example in the 8-dimensional case there are 70 models that undergo a shift in four dimensions, but only 28 models that undergo a shift in two dimensions. Chipman [71] suggests an alternative model space prior that takes into account dimensionality by weighing parsimonious models possibly more heavily. In this case Chipman [71] offers one possible model prior that has nonconstant probabilities

$$p_0(M_k) = w^{q_{M_k}}(1-w)^{p-q_{M_k}}, \quad M_k = 1, 2, \ldots, l, \tag{6.7}$$

where p represents the dimension of our data set, q_{M_k} represents the number of dimensions the mean vector has shifted for M_k, and w is a hyperparamter used to weight the models. Notice that if we set $w = \frac{1}{2}$, then equation (6.7) corresponds exactly to a flat prior and the posterior is then given by (6.6) as expressed above. In other situations we might expect any shift that occurs to the mean vector occurs in relatively few dimensions. In this case selecting values for w less than $\frac{1}{2}$ would be appropriate as more weight is given to parsimonious models. For example, the modeler may choose to more heavily weight the no change-point

model. Our subsequent discussion assumes prior knowledge of the model space is unavailable and hence we place a flat prior on the model space. If sufficient prior knowledge exists, however, our posterior distribution may be easily modified to incorporate prior model space information.

6.3.1 Known: $\boldsymbol{\mu}_\tau$ and Σ. Unknown: $\boldsymbol{\mu}_{\tau+1}$, τ, and M_K.

We use this case to establish the theory presented above to our problem at hand. Here we provide detailed calculations that to a large extent will also apply to later examples. By establishing first the simplest case, the method is then easily applied by relaxing assumptions as required for the more complicated change-point problems to follow.

Posterior distribution within M_k

We begin by developing an explicit form for $\pi_k(\boldsymbol{\mu}_{\tau+1,k}, \tau_k | \boldsymbol{\mu}_{\tau,k}, \Sigma, \mathbf{D})$ where $\mathbf{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. We consider only the proportionality form since, as we shall see later, the denominator, $P(\mathbf{D})$, eventually cancels out in our analysis. Thus, we seek an explicit expression for

$$\begin{aligned} \pi_k(\boldsymbol{\theta}_k \mid \boldsymbol{\mu}_{\tau,k}, \Sigma, \mathbf{D}) &\propto L(\mathbf{D}|\boldsymbol{\mu}_{\tau,k}, \boldsymbol{\mu}_{\tau+1,k}, \tau_k, \Sigma) p_0(\boldsymbol{\mu}_{\tau+1,k}, \tau_k) \\ &= L(\mathbf{D}|\boldsymbol{\mu}_{\tau,k}, \boldsymbol{\mu}_{\tau+1,k}, \tau_k, \Sigma) p_0(\boldsymbol{\mu}_{\tau+1,k}) p_0(\tau_k) \end{aligned} \tag{6.8}$$

where the right most equality follows by assuming independence of the after-shift mean vector and the change-point location.

We let $f_{\tau,k}(\mathbf{x}_i \mid \boldsymbol{\mu}_{\tau,k}, \Sigma)$ and $f_{\tau+1,k}(\mathbf{x}_i | \boldsymbol{\mu}_{\tau+1,k}, \Sigma)$ denote the multivariate normal probability density function both before and after the change-point for model k, respectively. By the independence assumption of our time series, the

likelihood function becomes

$$L_k(\mathbf{D}|\boldsymbol{\mu}_{\tau,k}, \boldsymbol{\mu}_{\tau+1,k}, \tau_k, \Sigma) = \prod_{i=1}^{\tau} f_{\tau,k}(\mathbf{x}_i|\boldsymbol{\mu}_{\tau,k}, \Sigma) \prod_{i=\tau+1}^{N} f_{\tau+1,k}(\mathbf{x}_i|\boldsymbol{\mu}_{\tau+1,k}, \Sigma). \quad (6.9)$$

We place a multivariate normal prior on $\boldsymbol{\mu}_{\tau+1,k}$ and a discrete uniform prior on τ where $\tau \in \{1, 2, \ldots, N-1\}$. We denote M_1 as the no change-point model. For M_k, $k > 1$, the mean vector undergoes a shift in exactly d_k dimensions where $d_k \in \{1, 2, \ldots, p\}$. We introduce the notation $\tilde{\boldsymbol{\mu}}_{\tau+1,k}$ to denote just the components of $\boldsymbol{\mu}_{\tau+1,k}$ that undergo the unknown shift. If we assume a $N(\boldsymbol{\Lambda}_k|, \mathbf{P}_k)$ prior on $\tilde{\boldsymbol{\mu}}_{\tau+1,k}$, then our complete prior distribution is simply the product

$$p_{0,k}(\tilde{\boldsymbol{\mu}}_{\tau+1,k}, \tau) = 2\pi^{d_k/2}\mathbf{P}_k^{-1/2} \exp\left[-\frac{1}{2}(\tilde{\boldsymbol{\mu}}_{\tau+1,k} - \boldsymbol{\Lambda}_k)^T \mathbf{P}_k^{-1}(\tilde{\boldsymbol{\mu}}_{\tau+1,k} - \boldsymbol{\Lambda}_\mathbf{k}) \right] \frac{1}{N-1} \quad (6.10)$$

where $\mathbf{P}_k$ and $\boldsymbol{\Lambda}_k$ are parameters chosen by the modeler. Thus by substituting equations (6.9) and (6.10) into equation (6.8) we obtain our model specific posterior distribution. We also point out that for the M_1 case that the posterior distribution simplifies becoming directly proportional to equation (6.9) with $\boldsymbol{\mu}_{\tau,k} = \boldsymbol{\mu}_{\tau,k+1}$.

Within model Metropolis Hastings parameter update

After choosing initial values for the Markov chain, our first task is to update the state within the same model. Given we are in model M_k, we write our current parameter vector as $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_{\tau+1,k}, \tau_k)$. As above, let d_k represent the number of dimensions our mean vector shifts and let $\tilde{\boldsymbol{\mu}}_{\tau+1,k}$ denote the mean vector containing just the components of $\boldsymbol{\mu}_{\tau+1,k}$ that undergo the unknown shift.

We then propose parameter $\boldsymbol{\theta}'_k$ using the following proposal distributions.

$$\tilde{\boldsymbol{\mu}}'_{\tau+1,k} \sim N_p(\tilde{\boldsymbol{\mu}}_{\tau+1,k}, \Sigma_{d_k})$$

$$\tau'_k \sim DU(E)$$

where DU denotes the discrete uniform probability distribution over the set E while Σ_{d_k} and E are tuning parameters chosen by the modeler. For example, we might choose Σ_{d_k} to be a submatrix of Σ corresponding to the change components of M_k and $E = (\tau_k - 2, \tau_k - 1, \tau_k + 1, \tau_k + 2)$. From $\tilde{\boldsymbol{\mu}}'_{\tau+1,k}$ we obtain $\boldsymbol{\mu}'_{\tau+1,k}$ in the natural way since these additional components of $\boldsymbol{\mu}'_{\tau+1,k}$ are already known in M_k by how we defined $\boldsymbol{\mu}'_{\tau+1,k}$ in the first place. Thus, we obtain $\boldsymbol{\theta}'_k = (\boldsymbol{\mu}'_{\tau+1,k}, \tau'_k)$. We accept the move to $\boldsymbol{\theta}'_k$ with probability

$$a(\boldsymbol{\theta}_k, \boldsymbol{\theta}'_k) = \min\left(1, \frac{\pi_k(\boldsymbol{\theta}'_k|\mathbf{D})}{\pi_k(\boldsymbol{\theta}_k|\mathbf{D})}\right). \tag{6.11}$$

Here we see how the unknown denominator in the respective posterior distribution functions cancel out and why we only considered the proportional form of the posterior distribution. We note the technical point that should our proposal distribution propose a parameter value outside the range of possible values (e.g. $\tau_k = N$), we accept this move with zero probability. For more details on the standard Metropolis-Hastings algorithm at an introductory level see [14] or for a more thorough exposition see, for example, [13].

Model jumping and the detailed balance equation

Once we complete the within model parameter update, we next go to the transmodel parameter update step. Here we require a proposal distribution that eventually allows us to sample the entire model space. Ideally, such a distribution

samples this space both efficiently and as simply as possible. With these criteria in mind, we present the following methodology.

Given that we are in model M_k, we propose to move to model $M_{k'}$ with probability distribution $j(\cdot)$. From M_k, let $\mathcal{M}_k$ denote the set of all possible models we allow jumps to and let $j \sim DU(\mathcal{M}_k)$. Next, we introduce the notation $\mathcal{F}_k$ to denote the dimensions in model k in which a shift in the mean vector occurs (these would be the indices of the components of $\tilde{\boldsymbol{\mu}}_{\tau+1,k}$ used in 3.1.2). Let $\#$ denote the cardinality of a set, so that $\#\mathcal{F}_k = d_k$ where d_k is as defined above. We say $M_{k'} \in \mathcal{M}_k$ if and only if both the following hold

$$\#\mathcal{F}_{k'} = d_k \pm 1$$

$$\mathcal{F}_{k'} \subset \mathcal{F}_k \text{ or } \mathcal{F}_k \subset \mathcal{F}_{k'}.$$

There is nothing particularly special about how we define $\mathcal{M}_k$ except that it does meet all our initial criteria and allows us to sample from the parameter space in a particularly simple manner. We have just two possible model proposal cases for which need to define g and h. For example, suppose $p = 4$ and M_k represents the model for which only components 1 and 4 shift. In this case $\mathcal{F}_k = \{1, 4\}$ and so $\#\mathcal{F}_k = 2$. Then according to our criteria $\#\mathcal{F}_{k'} = 1$ or $\#\mathcal{F}_{k'} = 3$. In the first case, we allow proposal model $M_{k'}$ such that $\mathcal{F}_{k'} = \{1\}$ or $\mathcal{F}_{k'} = \{4\}$. In the second case we allow proposal model $M_{k'}$ such that $\mathcal{F}_{k'} = \{1, 2, 4\}$ or $\mathcal{F}_{k'} = \{1, 3, 4\}$. Thus, in this example, $\mathcal{M}_k$ contains four proposal models.

$\#\mathcal{F}_{k'} < \#\mathcal{F}_k$ case

In moving from M_k to M_k', we have a clear equivalent set of parameters. These models differ in only a single component, say, i. As discussed above by how we allow model jumps, there is a single component i in model M_k that undergoes

a shift that does not shift in M'_k. Formally we also need to generate a random variable from a probability distribution g' to satisfy the dimension matching criteria, but this only enters into the calculations for the $\#\mathcal{F}_{k'} > \#\mathcal{F}_k$ case discussed below. Letting

$$\boldsymbol{\theta}_k = (\boldsymbol{\mu}_{\tau+1,k}, \tau_k)$$

we note that except for the single full model, not all the components of $\boldsymbol{\mu}_{\tau+1,k}$ are unknown. Thus we define

$$\tilde{\boldsymbol{\theta}}_k = (\tilde{\boldsymbol{\mu}}_{\tau+1,k}, \tau_k) = ((\mu^1_{\tau+1,k}, \mu^2_{\tau+1,k}, \ldots, \mu^i_{\tau+1,k}, \ldots, \mu^{d_k}_{\tau+1,k}), \tau_k).$$

To be more precise we could denote the components of $\tilde{\boldsymbol{\mu}}_{\tau+1,k}$ as $\tilde{\mu}^i_{\tau+1,k}$ to distinguish them from the components of $\boldsymbol{\mu}_{\tau+1,k}$ since in general they will not correspond. For ease of notation, however, we suppress the $\sim$ symbol over the components of $\tilde{\boldsymbol{\mu}}_{\tau+1,k}$ since the meaning of the components will always be clear by context. We define our function h as

$$h : \mathbb{R}^{d_k} \times \mathbb{R} \times \mathbb{R}^{r_k} \to \mathbb{R}^{d_{k'}} \times \mathbb{R} \times \mathbb{R}^{r_{k'}} \tag{6.12}$$

$$\tilde{\boldsymbol{\theta}}_k = (\tilde{\boldsymbol{\mu}}_{\tau+1,k}, \tau_k) \mapsto \tilde{\boldsymbol{\theta}}_{k'} = (\tilde{\boldsymbol{\mu}}_{\tau+1,k'}, \tau_{k'}, u)$$

where u denotes a 1-dimension random variable from a normal probability distribution. In this case we have $d_{k'} = d_k - 1$, $r_k = 0$, and $r_{k'} = 1$, so we see the dimension matching condition in (6.4) is met. Explicitly, we have the mapping

$$\begin{aligned}
\tilde{\boldsymbol{\mu}}_{\tau+1,k} &= ((\mu^1_{\tau+1,k}, \mu^2_{\tau+1,k}, \ldots, \mu^i_{\tau+1,k}, \ldots, \mu^p_{\tau+1,k})) \\
&\mapsto ((\mu^1_{\tau+1,k'}, \mu^2_{\tau+1,k'}, \ldots, \mu^{-i}_{\tau+1,k'}, \ldots, \mu^p_{\tau+1,k'})) \\
&= ((\mu^1_{\tau+1,k}, \mu^2_{\tau+1,k}, \ldots, \mu^{-i}_{\tau+1,k}, \ldots, \mu^p_{\tau+1,k})) \\
&= \tilde{\boldsymbol{\mu}}_{\tau+1,k'}
\end{aligned} \tag{6.13}$$

where the superscript $-i$ is used to denote an omitted component. Furthermore,

$$\begin{aligned} \tau_k &\mapsto \tau_{k'} \\ &= \tau_k, \end{aligned} \tag{6.14}$$

that is, the change-point parameter remains constant when jumping between models. The extra component in the original model maps as

$$\mu^i_{\tau+1,k} \mapsto u. \tag{6.15}$$

Thus we have componentwise equality in this mapping except in the single dimension where no shift occurs in the proposed model. Clearly h defines a bijective function and the absolute value of the determinant of the Jacobian trivially evaluates to 1. We then recover our new parameter $\boldsymbol{\theta}_{k'}$ from $\tilde{\boldsymbol{\theta}}_{k'}$ as previously explained.

$\#\mathcal{F}_{k'} > \#\mathcal{F}_k$ case

Here we map via the inclusion map into the higher dimensional model. We then generate a random variable that maps to the extra dimension to satisfy the dimension matching condition and hence the detailed balance equation. We use similar notation as above except now we move in the opposite the direction. Thus h maps

$$h : \mathbb{R}^{d_k} \times \mathbb{R} \times \mathbb{R}^{r_k} \to \mathbb{R}^{d_{k'}} \times \mathbb{R} \times \mathbb{R}^{r_{k'}} \tag{6.16}$$

$$\tilde{\boldsymbol{\theta}}_k = (\tilde{\boldsymbol{\mu}}_{\tau+1,k}, \tau_k, u) \mapsto \tilde{\boldsymbol{\theta}}_{k'} = (\tilde{\boldsymbol{\mu}}_{\tau+1,k'}, \tau_{k'}).$$

In this case clearly $r_k = 1$, $r_{k'} = 0$, and $d_{k'} = d_k + 1$. The componentwise mapping and the recovery of $\boldsymbol{\theta}_{\tau+1,k'}$ from $\tilde{\boldsymbol{\theta}}_{\tau+1,k'}$ are similar to the details given above. Essentially, the only difference is how h maps to the extra dimension. Let i represent the coordinate in $M_{k'}$ that we cannot map to via the inclusion map. From either the most recent visit to this model (or from the initial conditions if

this is the first visit to this model), say the value in coordinate i is ϕ. We generate the random variable u via a normal distribution:

$$u \sim N(\phi, \sigma^2)$$

where σ is a tuning parameter chosen by the modeler. Then we have the mapping

$$\begin{aligned} &((\mu^1_{\tau+1,k}, \mu^2_{\tau+1,k}, \ldots, \mu^{-i}_{\tau,k}, \ldots, \mu^p_{\tau+1,k}), \tau_k, u) \\ &\quad \mapsto ((\mu^1_{\tau+1,k'}, \mu^2_{\tau+1,k'}, \ldots, u, \ldots, \mu^p_{\tau+1,k'}), \tau_{k'}) \\ &\quad = ((\mu^1_{\tau+1,k}, \mu^2_{\tau+1,k}, \ldots, u, \ldots, \mu^p_{\tau+1,k}), \tau_k) \end{aligned} \tag{6.17}$$

Once again we see the dimension matching condition is satisfied and the determinant of the Jacobian evaluates to 1.

Probability of jumping to a new model

With all details in hand we are finally ready to compute the probability of accepting a proposed model. We have from equation (6.3) in section 1 that our probability of accepting the proposed jump is

$$\alpha(x, x') = \min\left\{1, \frac{\pi(x')j(M_{k'}|M_k)g'(u')}{\pi(x)j(M_k|M_{k'})g(u)}\left|\frac{\partial(x', u')}{\partial(x, u)}\right|\right\}. \tag{6.18}$$

There are just two generic cases to consider. In the case when $\#\mathcal{F}_{k'} < \#\mathcal{F}_k$ we have no random variables to generate, so that equation (6.18) becomes

$$\alpha(x, x') = \min\left\{1, \frac{\pi(x')j(M_{k'}|M_k)}{\pi(x)j(M_k|M_{k'})}\right\} \tag{6.19}$$

and in the case when $\#\mathcal{F}_{k'} > \#\mathcal{F}_k$ we have a single random variable to generate,

so that equation (6.18) becomes

$$\alpha(x, x') = \min\left\{1, \frac{\pi(x')j(M_{k'}|M_k)}{\pi(x)j(M_k|M_{k'})g(u)}\right\} \tag{6.20}$$

where we note the determinant of the Jacobian in both cases is equal to 1 as previously discussed. We summarize the methodology detailed above with a specific algorithm based on the general algorithm given in section 6.2.

1. Choose initial conditions, $x_0 = (M_{k_0}, \boldsymbol{\theta}_{k_0})$.

2. Perform a within model Metropolis-Hastings update of $\boldsymbol{\theta}_{k_0}$ with the methodology detailed in section 6.3.1 with acceptance probability given by equation (6.11).

3. Propose to "jump" to model $M_{k'}$ with conditional probability density $j(M_{k'}|M_k) = DU(\mathcal{M}_k)$ as explained in section 6.3.1. In particular $M_{k'}$ differs from M_k by having either one more or one fewer mean shift components.

4. Generate a one dimensional random "vector" u from $N(\phi, \sigma^2)$.

5. Depending on the outcome of step 3 we will now either be in the $\#\mathcal{F}_{k'} < \#\mathcal{F}_k$ case or the $\#\mathcal{F}_{k'} > \#\mathcal{F}_k$ case. We will then propose a new state x' from either equation (6.12) or equation (6.16) appropriately.

6. Accept the proposed move with probability given by either equation (6.19) or (6.20) depending on whether we are proposing a jump down or up a dimension, respectively.

6.3.2 Known: $\boldsymbol{\mu}_\tau$. Unknown: Σ, $\boldsymbol{\mu}_{\tau+1}$, τ, and M_K.

For this more general case we will see how the majority of our work in the previous section directly applies in this case. We once again assume $\boldsymbol{\mu}_\tau$ is known,

but that Σ is unknown. In this case, the result of our analysis is an estimate of $\boldsymbol{\mu}_{\tau+1}$, τ and Σ.

Posterior distribution within M_k

We once again develop an explicit form for our posterior distribution $\pi_k(\boldsymbol{\theta}_k \mid \boldsymbol{\mu}_{\tau,k}, \mathbf{D})$. Assuming prior parameter independence, we can write

$$\begin{aligned} \pi_k(\boldsymbol{\theta}_k \mid \boldsymbol{\mu}_{\tau,k}, \mathbf{D}) &\propto L(\mathbf{D}|\boldsymbol{\mu}_{\tau+1,k}, \tau_k, \Sigma) p_0(\boldsymbol{\mu}_{\tau+1,k}, \tau_k, \Sigma) \\ &= L(\mathbf{D}|\boldsymbol{\mu}_{\tau+1,k}, \tau_k, \Sigma) p_0(\boldsymbol{\mu}_{\tau+1,k}) p_0(\tau_k) p_0(\Sigma) \end{aligned} \tag{6.21}$$

We first examine the likelihood portion of the posterior distribution function. Let $f_{\tau,k}(\mathbf{x}_i|\boldsymbol{\mu}_{\tau,k}, \Sigma)$ and $f_{\tau+1,k}(\mathbf{x}_i|\boldsymbol{\mu}_{\tau+1,k}, \Sigma)$ denote the multivariate normal probability distribution function both before and after the change-point for M_k, respectively. By independence, the likelihood function becomes the product

$$L_k(\mathbf{x}_1, .., \mathbf{x}_N|\boldsymbol{\mu}_{\tau,k}, \boldsymbol{\mu}_{\tau+1,k}, \tau_k, \Sigma) = \prod_{i=1}^{\tau} f_{\tau,k}(\mathbf{x}_i|\boldsymbol{\mu}_{\tau,k}, \Sigma) \prod_{i=\tau+1}^{N} f_{\tau+1,k}(\mathbf{x}_i|\boldsymbol{\mu}_{\tau+1,k}, \Sigma).$$

Dropping multiplicative constant terms and expanding the exponent we have

$$\begin{aligned} &L_k \\ &\propto |\Sigma|^{-N/2} \exp\Big(-\frac{1}{2}\Big[\sum_{i=1}^{\tau}(\mathbf{x}_i - \boldsymbol{\mu}_{\tau,k})^T\Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_{\tau,k}) + \sum_{i=\tau+1}^{N}(\mathbf{x}_i - \boldsymbol{\mu}_{\tau,k})^T\Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_{\tau,k})\Big]\Big) \\ &= |\Sigma|^{-N/2} \exp\Big(-\frac{1}{2} tr\Big(\Sigma^{-1}A\Big)\Big) \end{aligned} \tag{6.22}$$

where

$$A = \sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^T - 2\tau\boldsymbol{\mu}_{\tau,k}\bar{\mathbf{x}}_{\tau}^T - 2(N-\tau)\boldsymbol{\mu}_{\tau,k+1}\bar{\mathbf{x}}_{\tau+1}^T + \tau\boldsymbol{\mu}_{\tau,k}\boldsymbol{\mu}_{\tau,k}^T + (N-\tau)\boldsymbol{\mu}_{\tau+1,k}\boldsymbol{\mu}_{\tau,k}^T. \tag{6.23}$$

Next, we choose a prior for $\boldsymbol{\mu}_{\tau+1,k}$ in the form of a multivariate normal distribution and maintain a uniform prior for τ. We may then express p_0 as

$$\begin{aligned} p_0(\boldsymbol{\mu}_{\tau+1,k}, \tau_k, \Sigma) &\propto p_{0,k}(\tilde{\boldsymbol{\mu}}_{\tau+1,k}, \tau, \Sigma) \\ &\propto \mathbf{P}_k^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{\mu}_{\tau+1,k} - \boldsymbol{\Lambda}_k)^T \mathbf{P}_k^{-1}(\boldsymbol{\mu}_{\tau+1,k} - \boldsymbol{\Lambda}_\mathbf{k})\right] \end{aligned} \tag{6.24}$$

where $\mathbf{P}_k$ and $\boldsymbol{\Lambda}_k$ are chosen by the modeler. Since Σ is an unknown parameter, we would like to integrate it out. The rearrangement of our likelihood suggests that a Wishart prior distribution would allow us to integrate out the covariance term. Thus

$$\begin{aligned} \pi_k(\boldsymbol{\mu}_{\tau,k}, \boldsymbol{\mu}_{\tau,k+1}, \tau_k \mid \mathbf{D}) &\propto \int_{\mathbb{R}^p \times \mathbb{R}^p} \pi_k(\boldsymbol{\mu}_{\tau,k}, \boldsymbol{\mu}_{\tau,k+1}, \tau_k, \Sigma | \mathbf{D}) d\Sigma \\ &\propto \int_{\mathbb{R}^p \times \mathbb{R}^p} L(\mathbf{D} | \boldsymbol{\mu}_{\tau,k}, \boldsymbol{\mu}_{\tau+1,k}, \tau_k, \Sigma) p_0(\boldsymbol{\mu}_{\tau+1,k}, \tau_k, \Sigma) d\Sigma \\ &\propto |A|^{-(n-p)/2} \mathbf{P}_k^{-1/2} \exp\,[\, -\frac{1}{2}(\boldsymbol{\mu}_{\tau+1,k} - \boldsymbol{\Lambda}_k)^T \mathbf{P}_k^{-1}(\boldsymbol{\mu}_{\tau+1,k} - \boldsymbol{\Lambda}_\mathbf{k})]. \end{aligned} \tag{6.25}$$

Equation (6.25) represents our posterior distribution marginalized over the covariance matrix and replaces equation (6.8) from the case in section 6.3.1. We note in the M_1 case that the prior distribution reduces to a constant. Once again, because of the ratio in the Metropolis-Hasting step, common multiplicative constants ultimately cancel out. Hence, this proportional form is all that we require for subsequent parameter inference.

Completing the case from 3.2 and estimating Σ

The remainder of this case follows nearly verbatim as presented in 3.1 and so we omit these details. We close with a final note that although we integrated out Σ in our analysis, we could still estimate the covariance matrix upon

completion of this example. With our estimates of $\hat{\boldsymbol{\mu}}_\tau$, $\hat{\boldsymbol{\mu}}_{\tau+1}$, $\hat{\tau}$, and M_K in hand, we can estimate Σ by its MLE, namely

$$\widehat{\Sigma} = \frac{1}{N}\Big(\sum_{i=1}^{\hat{\tau}}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\tau,K})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\tau,K})^T + \sum_{i=\hat{\tau}+1}^{N}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\tau+1,K})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\tau+1,K})^T\Big). \quad (6.26)$$

If our model assumptions hold and our estimates of $\hat{\boldsymbol{\mu}}_\tau$, $\hat{\boldsymbol{\mu}}_{\tau+1}$, $\hat{\tau}$, and M_K are accurate, then we expect $\widehat{\Sigma}$ to closely approximate the true Σ as well.

6.3.3 Unknown: $\boldsymbol{\mu}_\tau$, Σ, $\boldsymbol{\mu}_{\tau+1}$, τ, and M_K

In this final most general case, we combine the results from both our previous two examples. Here we not only need to revisit the form of a model specific posterior distribution, but also discuss our modified model jumping approach. As we shall see, we have more random variables to generate and thus to account for to maintain the detailed balance equation.

Posterior distribution within M_k

Once again we require an explicit proportional form for $\pi_k(\boldsymbol{\mu}_{\tau,k}, \boldsymbol{\mu}_{\tau+1,k}, \tau_k|\mathbf{D})$. We account for our additional unknown parameter, with an additional term in the prior distribution function

$$\pi_k(\boldsymbol{\theta}_k|\mathbf{D}) \propto L(\mathbf{D}|\boldsymbol{\mu}_{\tau,k}, \boldsymbol{\mu}_{\tau+1,k}, \tau_k, \Sigma)p_0(\boldsymbol{\mu}_{\tau,k}, \boldsymbol{\mu}_{\tau+1,k}, \tau_k, \Sigma)$$

$$= L(\mathbf{D}|\boldsymbol{\mu}_{\tau,k}, \boldsymbol{\mu}_{\tau+1,k}, \tau_k, \Sigma)p_0(\boldsymbol{\mu}_{\tau,k})p_0(\boldsymbol{\mu}_{\tau+1,k})p_0(\tau_k)p_0(\Sigma).$$

We write our likelihood function as above

$$L(\mathbf{D}|\boldsymbol{\mu}_{\tau,k}, \boldsymbol{\mu}_{\tau+1,k}, \tau_k, \Sigma) \propto |\Sigma|^{-N/2}\exp\Big(-\frac{1}{2}tr\Big(\Sigma^{-1}A\Big)\Big)$$

where A is the matrix given in (6.23). We place multivariate normal priors on $\boldsymbol{\mu}_{\tau,k}$ and $\boldsymbol{\mu}_{\tau+1,k}$. Additionally, we place a Wishart prior on Σ and a discrete uniform prior on τ as in section 3.2. We assume for M_k, that the mean vector undergoes a shift in exactly d_k dimensions. As before $\tilde{\boldsymbol{\mu}}_{\tau,k}$ and $\tilde{\boldsymbol{\mu}}_{\tau+1,k}$ denote the components of $\boldsymbol{\mu}_{\tau,k}$ and $\boldsymbol{\mu}_{\tau+1k}$ that undergo the shift. We write our prior distribution as

$$
\begin{aligned}
&p_{0,k}(\tilde{\boldsymbol{\mu}}_{\tau,k}, \tilde{\boldsymbol{\mu}}_{\tau+1,k}, \tau, \Sigma) \\
&\quad \propto |\mathbf{P}_{\tau,k}|^{-1/2} \exp\left[-\tfrac{1}{2}(\tilde{\boldsymbol{\mu}}_{\tau,k} - \boldsymbol{\Lambda}_{\tau,k})^T \mathbf{P}_{\tau,k}^{-1}(\tilde{\boldsymbol{\mu}}_{\tau,k} - \boldsymbol{\Lambda}_{\tau,k})\right] \\
&\qquad \times |\mathbf{P}_{\tau+1,k}|^{-1/2} \exp\left[-\tfrac{1}{2}(\tilde{\boldsymbol{\mu}}_{\tau+1,k} - \boldsymbol{\Lambda}_{\tau+1,k})^T \mathbf{P}_{\tau+1,k}^{-1}(\tilde{\boldsymbol{\mu}}_{\tau+1,k} - \boldsymbol{\Lambda}_{\tau+1,k})\right] \frac{1}{N-1}
\end{aligned}
$$

where $\mathbf{P}_{\tau,k}$, $\mathbf{P}_{\tau+1,k}$, $\boldsymbol{\Lambda}_{\tau,k}$, and $\boldsymbol{\Lambda}_{\tau+1,k}$ are hyperparameters chosen by the modeler. Integrating out Σ we have our marginalized posterior distribution function:

$$
\pi_k(\boldsymbol{\theta}_k|\mathbf{D}) \propto |A|^{-(n-p)/2} p_{0,k}(\tilde{\boldsymbol{\mu}}_{\tau,k}, \tilde{\boldsymbol{\mu}}_{\tau+1,k}, \tau, \Sigma). \tag{6.27}
$$

Equation (6.27) will be the engine that drives both the within model parameter updates as well as the model jumping step. As previously noted, the posterior distribution for the no change-point model special case, M_1, reduces to a proportionality form and becomes

$$
\pi_k(\boldsymbol{\theta}_k|\mathbf{D}) \propto |A|^{-(n-p)/2}. \tag{6.28}
$$

Model jumping and the detailed balance equation

For this step of the algorithm, M_k, $j(\cdot)$, $\mathcal{M}_k$, $\tilde{\boldsymbol{\mu}}_{\tau,k}$, $\tilde{\boldsymbol{\mu}}_{\tau+1,k}$, $\mathcal{F}_k$, and d_k are all as previously defined in the cases above.

$\#\mathcal{F}_{k'} < \#\mathcal{F}_k$ case

Given parameter state $\boldsymbol{\theta}_k$ in model M_k we propose parameter $\boldsymbol{\theta}_{k'}$ in model $M_{k'}$. Here $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_{\tau,k}, \boldsymbol{\mu}_{\tau+1,k}, \tau_k)$. As before we use $\tilde{\boldsymbol{\theta}}_k$ to emphasize the unknown components of the our current model state. Then

$$\tilde{\boldsymbol{\theta}}_k = (\tilde{\boldsymbol{\mu}}_{\tau,k}, \tilde{\boldsymbol{\mu}}_{\tau+1,k}, \tau_k) = ((\mu^1_{\tau,k}, \mu^2_{\tau,k}, \ldots, \mu^i_{\tau,k}, \ldots, \mu^{d_k}_{\tau,k}), (\mu^1_{\tau+1,k}, \mu^2_{\tau+1,k}, \ldots, \mu^i_{\tau+1,k}, \ldots, \mu^{d_k}_{\tau+1k,}), \tau_k)$$

Our function h maps as

$$h : \mathbb{R}^{d_k} \times \mathbb{R}^{d_k} \times \mathbb{R} \times \mathbb{R}^{r_k} \to \mathbb{R}^{d_{k'}} \times \mathbb{R}^{d_{k'}} \times \mathbb{R} \times \mathbb{R}^{r'_k}$$

$$\tilde{\boldsymbol{\theta}}_k = \left(\tilde{\boldsymbol{\mu}}_{\tau,k}, \tilde{\boldsymbol{\mu}}_{\tau+1,k}, \tau_k\right) \mapsto \tilde{\boldsymbol{\theta}}_{k'} = (\tilde{\boldsymbol{\mu}}_{\tau,k'}, \tilde{\boldsymbol{\mu}}_{\tau+1,k'}, \tau_{k'}, u)$$

where u denotes a 2-dimension random vector and $d_{k'} = d_k - 1$. In this case $r_k = 0$ and $r'_k = 2$, so we see the dimension matching condition in 6.4 is met. The componentwise mapping is similar to that which was explicitly shown before. One can check that the Jacobian evaluates to 1. We then recover our new parameter $\boldsymbol{\theta}_{k'}$ from $\tilde{\boldsymbol{\theta}}_{k'}$.

$\#\mathcal{F}_{k'} > \#\mathcal{F}_k$

Similar to before, the only other case we need to consider is for $\#\mathcal{F}_{k'} > \#\mathcal{F}_k$. We map the common components via the inclusion map into the higher dimensional model and then map random variables to the extra dimensions satisfying the dimension matching condition. Our function h maps between the spaces

$$h : \mathbb{R}^{d_k} \times \mathbb{R}^{d_k} \times \mathbb{R} \times \mathbb{R}^{r_k} \to \mathbb{R}^{d_{k'}} \times \mathbb{R}^{d_{k'}} \times \mathbb{R} \times \mathbb{R}^{r'_k}$$

$$\tilde{\boldsymbol{\theta}}_k = (\tilde{\boldsymbol{\mu}}_{\tau,k}, \tilde{\boldsymbol{\mu}}_{\tau+1,k}, \tau_k) \mapsto \tilde{\boldsymbol{\theta}'}_k = (\tilde{\boldsymbol{\mu}}_{\tau+1,k'}, \tilde{\boldsymbol{\mu}}_{\tau+1,k'}, \tau_{k'}, u)$$

Here u is a two-dimensional random variable, so $d_{k'} = d_k + 1$, $r_k = 2$ and $r'_k = 0$. The componentwise mapping and the recovery of $\boldsymbol{\theta}_{\tau+1,k'}$ from $\tilde{\boldsymbol{\theta}}_{\tau+1,k'}$ are similar to the details given above. Essentially, the only difference is how h maps to the extra dimensions. Thus we need to generate a two dimensional random vector, u. We write u as

$$u = (u_\tau, u_{\tau+1})$$

and generate u_τ and $u_{\tau+1}$ from the following normal distributions:

$$u_\tau \sim N(\phi_\tau, \sigma_\tau^2)$$

$$u_{\tau+1} \sim N(\phi_{\tau+1}, \sigma_{\tau+1}^2).$$

Here ϕ_τ and $\phi_{\tau+1}$ correspond to either the initial value or the most recent value of this component in model $M_{k'}$. The variance components, σ_τ and $\sigma_{\tau+1}$, are tuning parameters chosen by the modeler.

6.4 Numerical Examples

For a given dimension, p, there are 2^p different models to consider. Except for the no change model, each model contains from 2 to $p + 1$ parameters that must be estimated, namely the change-point location and the shift amount in each dimension. For all the examples to follow we label the models according to the following scheme: M_1 is the no change-point model. M_2 through M_{p+1} contain a change in only a single component from 1 through p, respectively. M_{p+2} contains a shift in components 1 and 2 and so on. We continue in this manner until M_{2^p} which contains a change in every component of the mean vector. It follows that as the dimensionality of the problem increases the number of

parameters also begins to dramatically increase. Although this does present some computational difficulties for high dimensional data sets, we note that so long as the signal to noise ratio is sufficiently high, the jumping step rather quickly finds the correct model. Because the overwhelming majority of the models are never visited in this case, we end up safely ignoring the vast majority of the models and hence their respective parameters.

6.4.1 Case Study 1: A Phase I Application with a Temperature Data Set from an Industrial Boiler

6.5 displays a table where the first 24 rows are temperature readings from eight different burners on an industrial grade boiler. The burner temperatures on the boiler are to be monitored to detect deviations from an in-control state with the goal of detecting a "cold spot" on the boiler, should one develop [52]. Observing each temperature sequence suggests the data appears to be approximately stationary. Mason provides this same data set only originally with an additional outlier reading at time point 9. Once this outlier is removed, Mason demonstrates through the use of a standard Hotelling T^2 multivariate control chart and a Q-Q plot that this data set appears to be an example of an in-control homogeneous data set. Mason concludes these temperature readings are in-control and thus approximately follow a multivariate normal distribution [52] with the following statistics computed from the data set

$$\bar{X} = (524.7, 513.5, 539.4, 521.4, 503.6, 512.5, 478.6, 477.3)^T \tag{6.29}$$

and

$$\widehat{\Sigma} = \left\{ \begin{array}{rrrrrrrr} 53.45 & .84 & 25.44 & 30.00 & 19.70 & -2.23 & 20.07 & 00.30 \\ .84 & 5.04 & 0.31 & 0.66 & .26 & 3.22 & .67 & 3.60 \\ 25.44 & 3.31 & 18.59 & 14.11 & 07.28 & 1.90 & 8.38 & 2.77 \\ 30.00 & 2.66 & 14.11 & 20.85 & 12.76 & -.2.22 & 12.16 & .25 \\ 19.70 & .26 & 7.28 & 12.76 & 11.11 & -1.65 & 10.49 & -.25 \\ -2.23 & 3.22 & 1.91 & -2.22 & -1.65 & 04.69 & -.71 & 3.88 \\ 20.07 & .67 & 8.38 & 12.16 & 10.49 & -.71 & 11.65 & .59 \\ .30 & 3.60 & 2.77 & .25 & -.25 & 3.88 & .59 & 4.02 \end{array} \right\}. \tag{6.30}$$

Alternatively, we may instead conduct a Phase I analysis on this data set by applying the RJMCMC multivariate change-point approach to determine whether or not the data is in-control. Since in this case we assume no parameter information is available beforehand, the approach detailed in section 6.3.3 applies. Following the derivation given in section 6.3.3, we place a multivariate normal prior on the mean vectors both before and after the change-point ($\boldsymbol{\mu}_{\tau,k}$ and $\boldsymbol{\mu}_{\tau+1,k}$) and a noninformative discrete uniform on the change-point location, τ_k, for each model, M_k.

Since we assume we are working in the absence of other more informative information, we set identical hyperparameters for the distributions of $\boldsymbol{\mu}_{\tau,k}$ and $\boldsymbol{\mu}_{\tau+1,k}$ for a given model, M_k. In particular for a given model denote by $\boldsymbol{\Lambda}_k$ and $\mathbf{P}_k$ the components of $\bar{X}$ and $\widehat{\Sigma}$, given by (6.29) and (6.30), that are assumed to undergo a shift in M_k. For example, consider M_{12} which is the model corresponding to shifts in the first and fourth dimensions only. In this case we have

$$\boldsymbol{\Lambda}_{12} = (524.7, 521.4)^T \quad \text{and} \quad \mathbf{P}_{12} = \begin{pmatrix} 53.45 & 30.00 \\ 30.00 & 20.85 \end{pmatrix}.$$

Finally, for the initial conditions we arbitrarily set $\boldsymbol{\mu}_{\tau+1,k} = \boldsymbol{\Lambda}_k$, $\tau_k = 12$, and $k = 100$. Figure 6.1 graphically depicts relative model visits after running algorithm where we see the no change-point model is heavily favored. The other most frequently visited models, as expected, mostly contain shifts in one or two components. We conclude from our analysis that the no-change model is clearly the most probable which corresponds to the in-control state. Once the data is determined to be in-control, subsequent parameter estimation becomes trivial to calculate.

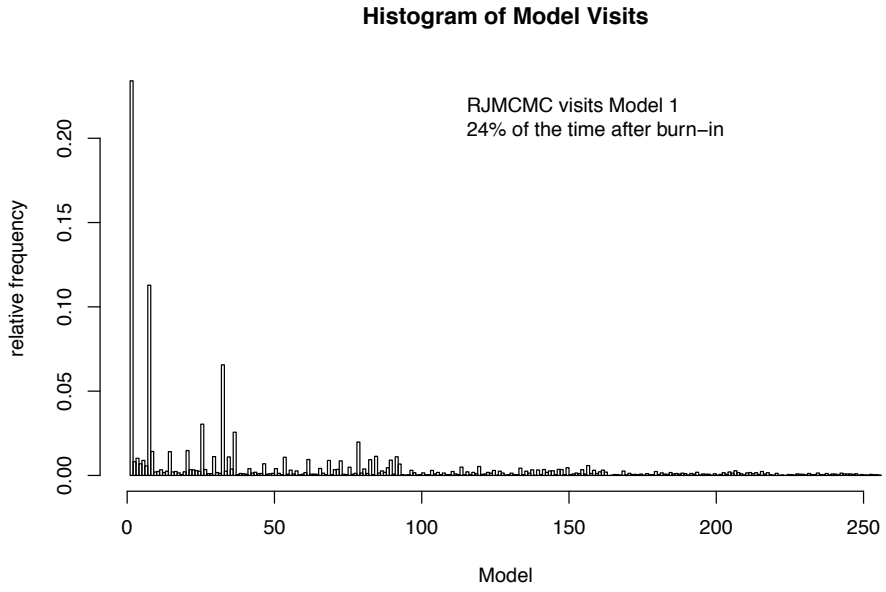

Figure 6.1: The histogram above depicts the relative number of visits to the possible change-point models. We see M_1 (the no change-point model) is visited 24% of the time which corresponds to more than twice the frequency of the second most visited model.

For illustrative purposes suppose instead we are given all 32 elements of the data set in 6.5. Elements 25-32 were generated by drawing from a multivariate normal distribution with the covariance matrix given by (6.30) and a mean vector corresponding to a shift in one standard deviation from equation (6.29) in the third, fifth, and eighth dimensions. Explicitly, the mean vector after

the change-point is given by

$$\bar{X}_{\tau+1} = (524.7, 513.5, 543.6869, 521.4, 506.9588, 512.5, 478.6, 479.2554)^T. \quad (6.31)$$

Making identical assumptions as in the preceding case, we compute our hyperparameter values for our prior distributions by taking the sample mean and covariance matrix of the entire out of control data set. As before since we assume no better information is available, we assign the same priors to the mean vectors both before and after the change-point. Figure 6.2 is a histogram depicting the relative model visits. M_{80} which corresponds to a shift in the third, fifth, and eighth dimensions is by the far the most visited model representing 28% of the total model visits. The RJMCMC change-point method also simultaneously provides estimates for all other parameters. Simulation results correctly estimate the change-point location to be at time point 24 with a 95% credible interval of 22-25 while the remaining parameter estimates are given in Table 6.1. One notes the mean vector component values in the dimensions not undergoing a shift are simply the sample mean of the respective component. The sample covariance matrix may then be easily computed using equation (6.26).

Table 6.1: Parameter estimates for model 80 corresponding to a shift in the 3rd, 5th, and 8th dimensions before the model estimated change-point at time point 24 (top) and after the change-point (bottom).

Dim	Estimated value	95% cred. interval	True Value
3	539.7	539.3-540.1	539.4
5	503.6	503.3-504.6	503.6
8	477.2	477.0-477.9	477.3
3	542.6	540.1-543.8	543.7
5	506.6	505.3-507.0	507.0
8	478.6	477.1-479.5	479.3

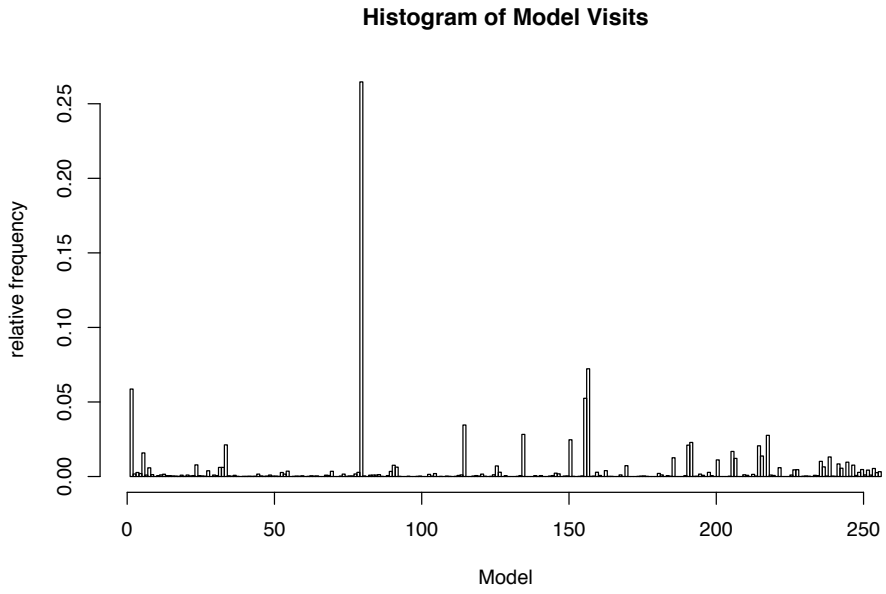

Figure 6.2: The histogram above depicts the relative number of visits to the possible change-point models. We see M_{80} (model with shifts in the third, fifth, and eighth components) is visited 28% of the time which corresponds to about four times the frequency of the second most visited model.

While other methods exist for determining whether or not a data set is in control, they generally either require some assumptions about the parameters of the data set or provide an incomplete picture. For example a Hotelling chart may not be helpful without a historical data set (HDS) to accurately compute the T^2 statistic. This problem also applies, albeit to a lesser extent, to both the popular multivariate CUSUM and multivariate EWMA control charts as the test statistic is difficult justify without an HDS.

Sullivan and Woodall proposed a Phase I method to determine whether or not a data set is in control without an HDS by defining their normalized control chart statistic $\hat{y}[m_1]$ where m_1 denotes the possible change-point location [72]. Figure 6.3 depicts their method applied to the same 32 element boiler temperature data set with a change-point introduced at time point 24. We see while their method does correctly indicate the data set is not in-control, this

approach incorrectly gives the highest probability of a change-point far away from the true change-point at time point 24. Another shortcoming with all these methods is that they provide at best an incomplete picture of parameter estimates. The RJMCMC change-point approach on the other hand may be equally well applied to both in-control and out-of-control data while simultaneously providing complete parameter inference.

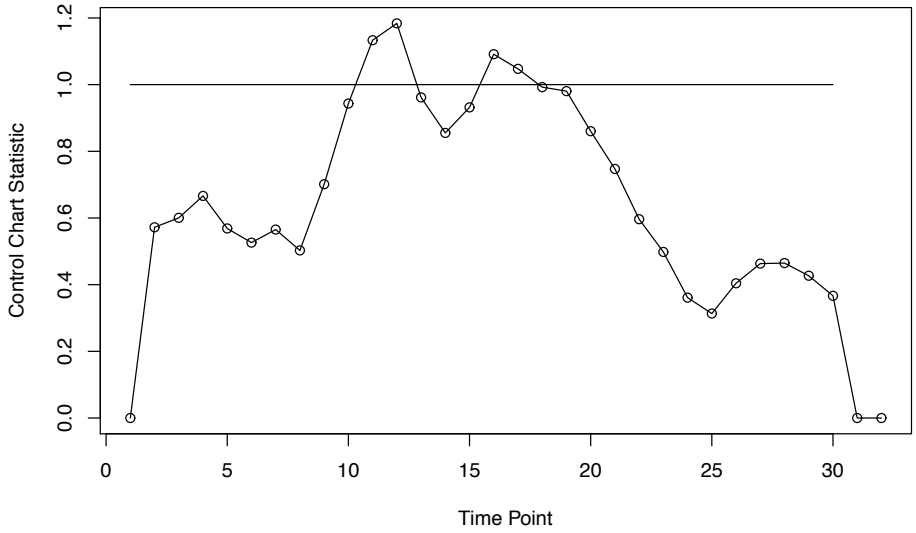

Figure 6.3: Sullivan and Woodal type control chart applied to the boiler temperature data set with a change-point at time point 24. The normalized UCL suggests the data is not in control, but at a time point away from time point 24.

6.4.2 Case Study 2: A Phase II Diagnostic Tool

In this example we analyze a simulated data set for the purposes of illustrating how the RJMCMC change-point method may also be applied as a Phase II diagnostic tool. Consider a six dimensional 100 element simulated data set given in 6.5. The data is initially generated around a mean vector of $\boldsymbol{\mu}_\tau = (0, 0, 0, 0, 0, 0)$ using a covariance matrix with 1's along the diagonal and .3's on all off diagonal elements. We inject a change-point at time point 80 generating

the remaining elements around the new mean vector $\boldsymbol{\mu}_{\tau+1} = (0, 0, 0, 0, .75, 2)$ while maintaining a constant covariance matrix.

Now suppose we are in Phase II monitoring incoming data with a Hotelling T^2 control chart beginning at time point 1 of the data set. We make the usual Phase II assumption that a historical data set is available providing us with the in-control data set parameters given above. Using these parameters we may generate the familiar T^2 test statistic and obtain our upper control limit (UCL) with the chi-squared distribution. Observing Figure 6.4 we see the control chart signals an alarm at the 99% confidence level (UCL=16.8) at time point 91. We now stop the monitoring process and perform diagnostics on data set elements 1 through 91 with the RJMCMC change-point method as described in section 6.3.2.

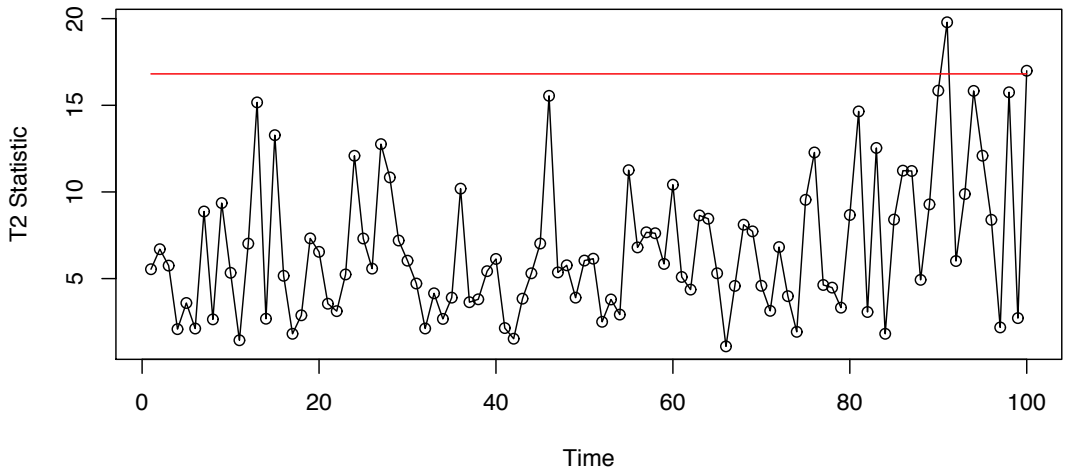

Figure 6.4: A T^2 control chart applied to the simulated data given in A.2. A change-point to the mean vector occurs at time point 80 and the control chart signals an alarm at the 99% confidence level at time point 91.

The first step in setting up the method detailed in section 6.3.2 is deciding the prior distributions to use. Although we could reasonably weight the latter portion of the data set for the change-point location, we instead take the more

conservative approach and place a flat prior on the change-point location. Next we need to choose the hyperparameters for our prior distribution on the new mean vector $\boldsymbol{\mu}_{\tau+1,k}$. Other than assuming normality for the prior distributions, we assume little additional prior information is available. Firstly, we compute the sample mean vector for the entire 91 data set observations. Next, for each model, M_k, we assign the components of the sample mean vector assumed to have undergone a change for M_k to the model mean vector hyperparameter, $\boldsymbol{\Lambda}_k$. Additionally, for the covariance hyperparameter we first compute the sample covariance from the entire data set and then obtain $\mathbf{P}_k$ as submatrices of the sample covariance matrix in a similar fashion as described in section 6.4.1. For this six dimensional case there are $2^6 = 64$ different possible models. We label these models as explained in the opening of section 6.4. The model labeling is unimportant except to the extent that we can keep track of models along with their particular assumptions. In this case M_{22} corresponds to a shift in the fifth and sixth dimensions and represents the true model that reflecting the data set.

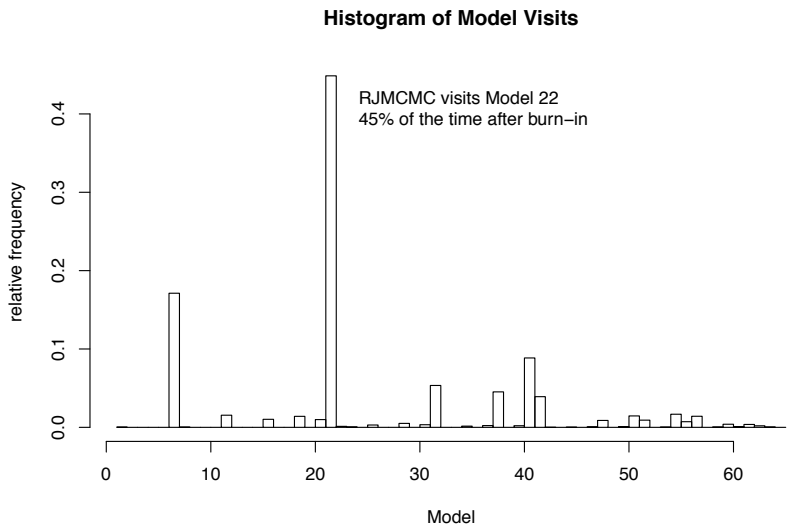

Figure 6.5: The correct model (M_{22}) is visited by the chain 31% of the time. Other highly visited models (in order of relative frequency) are models 7, 41, 32, and 38. Each of these other models differ from the correct model in just a single dimension and together with M_{22} account for over 80% of total model visits.

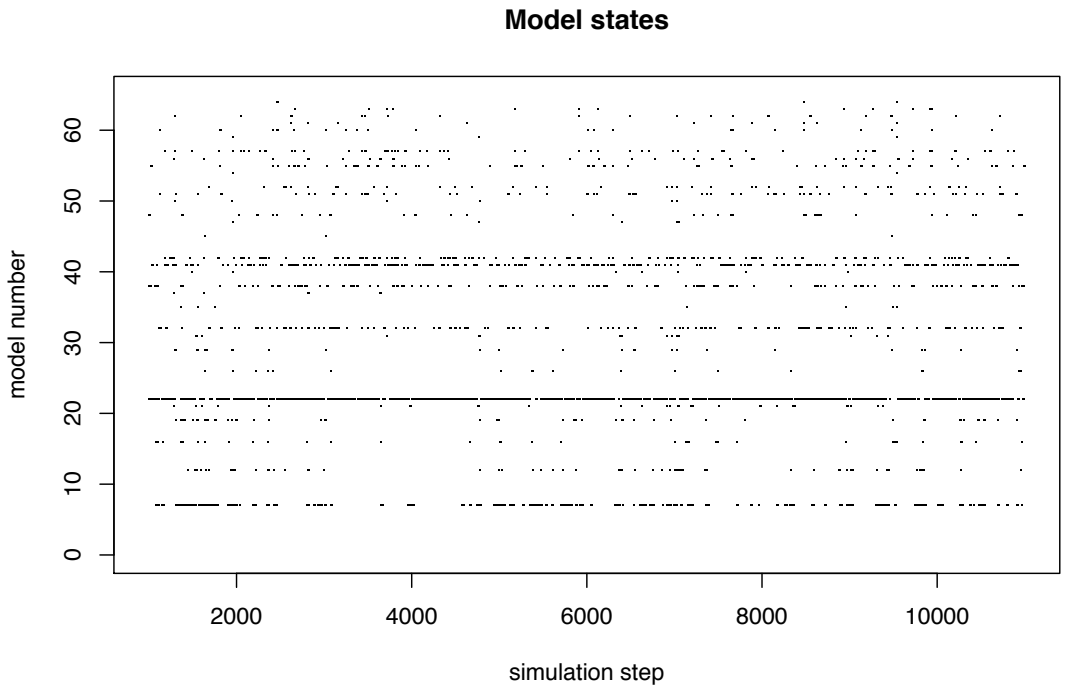

Figure 6.6: A progression of chain model visits as a function of time (simulation step). Notice M_{22} with 31% of the visits appears to be nearly a solid line. M_7 (17%), M_{41} (9%), M_{32} (5%), and M_{38} (4%) are also prominently featured.

The only other step we must accomplish is to choose an initial value to begin the RJMCMC process. For this example, we start by randomly selecting M_{40} as our initial model, namely only a shift in the third, fourth, and sixth dimensions occurs. We set the initial value of $\boldsymbol{\mu}_{\tau+1} = (0, 0, 0, 0, 0, 0)$ and the initial change-point location at time point 50. A better initial value guess may help the chain converge more quickly, but ultimately ergodic theory tells us the chain must eventually converge regardless of the initial value choice. With these preliminaries accomplished, we are finally set to implement the algorithm.

We run the chain for 10,000 iterations after a 1,000 iteration burn-in period. First observe Figure 6.5 which counts the number of visits to each of the possible 64 models for the final 10,000 iterations of the chain. We see the RJMCMC algorithm ignores many of the possible models and spends 45% of its time in the correct model, namely M_{22}. We gain further insight into the chain's

progression with Figure 6.6. While M_{22} is clearly the most visited model, other models that differ from M_{22} by only a single dimension also are also prominently featured in Figure 6.6 as well. For example, Model 41 represents the case for a change in dimensions 3, 5, and 6 captures 15% of the total chain visits. The

Table 6.2: Parameter estimates for M_{22} corresponding to a shift in the 5th and 6th dimensions at time point 80.

Dim	Estimated value	90% cred interval	True Value
5	.71	.65-.91	.75
6	1.96	1.73-2.1	2.0

RJMCMC change-point approach simultaneously returns estimates for each parameter for each respective model. From M_{22}, the change-point location is correctly estimated at time point 80 with a 95% credible interval from 76-81 while the remaining parameter estimates are given in Table 6.2.

To further illustrate how the chain walks through the probability space consider Figure 6.7 which depicts how the change-point parameter of each model varies against the chain model. A small amount of additive noise is added to the parameter values to provide a perspective on which model state combinations are the most frequently visited. We see that after the burn-in period many of the models are not visited at all. Those models that are visited mostly correspond with the true model in all but a single dimension of the mean vector after the change-point. Note also that Figure 6.7 illustrates the conditional nature of the model inference. For instance, if M_{41} (change in the mean components of 3, 5, and 6 only) is inferred, then a change-point of at time point 78 is more likely than a change-point at time point 77. On the other hand, if M_7 is inferred (a change in the mean components of 6 only) then time point 77 is a somewhat more likely change-point location than at time point 78.

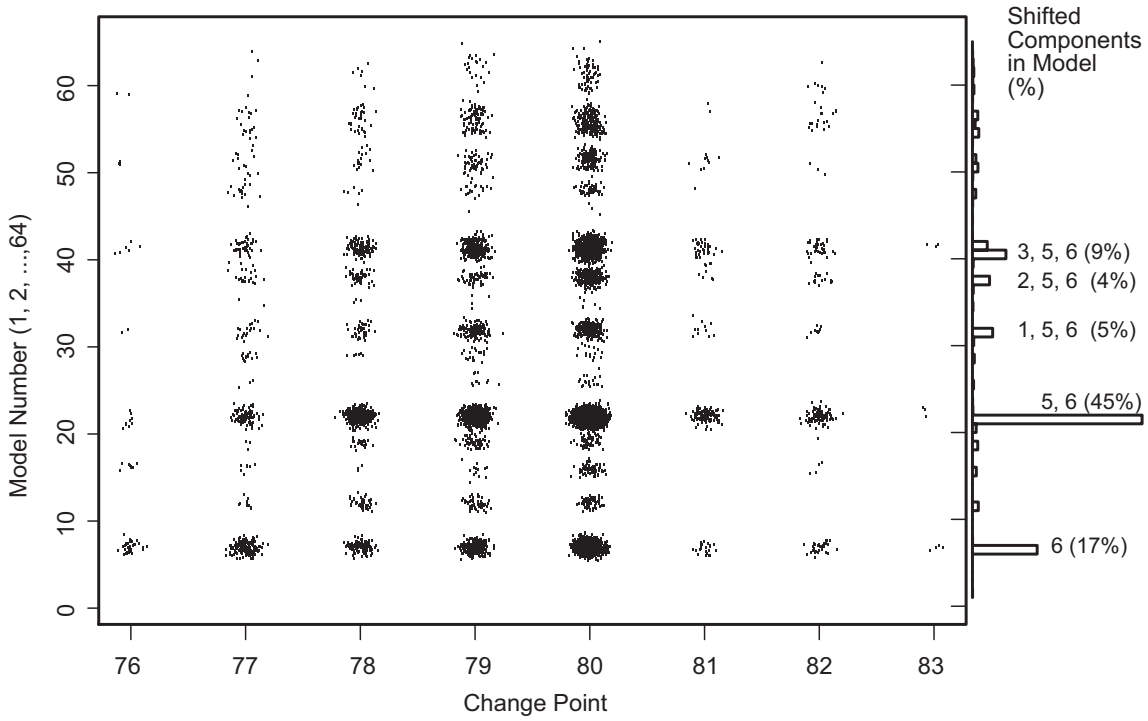

Figure 6.7: A plot of the change-point parameter against the model parameter. As a small amount of additive noise is added to the values to provide perspective on the most often visited parameter combinations.

While to our knowledge there does not exist another approach to the multivariate change-point problem that simultaneously incorporates complete parameter estimates into a single probability model, many methods exist in conjunction with control charts to partially answer these same questions. For example, to investigate which components of the mean vector have undergone a shift, we might employ the method proposed by Mason [73] to decompose the T^2 statistic into constituent contributing components of the T^2 statistic value. Calculating the T^2 statistic at time point 91 gives a value of 19.87 and corresponds to the first element of the data set signaling an alarm in Figure 6.4.

Mason *et. al.* [73] show in their paper the two terms of the T^2 decomposition of primary interest for determining the contribution a particular dimension to the T^2 statistic are T_j^2 and $T_{j\cdot 1,\ldots j+1,\ldots p}^2$. The first term corresponds to the unadjusted contribution of a single dimension j to the overall T^2 statistic

value. The second term, $T^2_{j\cdot 1,\ldots j+1,\ldots p}$, involves first adjusting the other $p-1$ variables and then calculating the adjusted contribution of dimension j to the T^2 value. We refer the reader to their paper for complete details on how these values are calculated, but list our calculations for these terms in Table 6.3 applied to the data set in this example. Notice at the 95% confidence level the T^2 decomposition method captures the contribution of dimension 6, but misses the more subtle contribution of dimension 5 that shifts by .75. As demonstrated above, the RJMCMC method correctly captures the change in both of these dimensions along with providing a complete probability distribution for parameter inference.

Variable (X_j)	T^2_j	$T^2_{j\cdot 1,\ldots j+1,\ldots p}$
X_1	0.04	0.63
X_2	0.09	0.04
X_3	1.88	0.54
X_4	0.08	0.87
X_5	2.80	0.69
X_6	8.10^*	5.58^*

Table 6.3: Decomposition of T^2 statistic. * indicates the respective decomposition component is significant at the 95% confidence level based on a one-sided upper critical value of 3.99

6.5 Conclusion

In this article we presented a detailed methodology for applying an RJMCMC approach to the multivariate change-point problem. By establishing the example in section 6.3.1, we then showed how we could extend the approach to more general situations. Our approach capitalizes on two strengths of the RJMCMC method, namely flexibility tailored to the structure of a given problem and minimal analysis to establish the method validity. We then applied this approach to both an actual data set with an artificial change-point introduced

and a simulated data set. In each example the algorithm favored the true model as evidenced by the frequency of model visits. Furthermore, the true parameter values in each case were also very closely estimated. Another strength of this method is that it is not particularly sensitive to the choice of the prior distributions placed on the model parameters. We obtained the prior hyperparameters simply by matching the components of change in a particular model with the corresponding components of the sample mean vector and sample covariance matrix computed from the entire data set.

As the number of dimensions increases, the number of potential parameters also begins to dramatically increase. In the case where the signal-to-noise ratio is sufficiently high, we find the majority of possible models are never visited. For example in Case Study 2 many of the possible 64 models were never visited by the chain. When the change-point is less pronounced, however, many more, if not all, models may be visited by the chain. This was the situation in Case Study 1 where nearly every model was visited and thus required us to keep track of over 1,200 parameters. In even higher dimensional cases the number of parameters will inevitably produce ever greater computational challenges. In our experiments, we successfully implemented the algorithm for cases up to 13 dimensions before encountering significant computational difficulties. This is by no means a dimensional ceiling as both improved programming efficiency and greater computational power could push this dimension higher as necessary.

This method seems particularly well suited to complement existing statistical quality control processes involving multivariate data. Even when the data may not be normally distributed, often there are techniques (such as through a well-chosen transformation) to approximate the data as such. While control charts can signal an alarm indicating an out-of-control condition in the data, they often do not provide a complete picture as what has occurred. Once

we have a data set with a statistical change-point, the RJMCMC approach to the multivariate change-point problem appears to be a valuable diagnostic tool to assist in the subsequent analysis. By sampling throughout the probability space, our chain efficiently marches along toward the correct model by (usually) accepting and rejecting moves toward the true parameter state appropriately. Once the chain converges the resulting empirical posterior probability distribution may then be used for parameter estimation to provide a more complete picture of the data set of interest.

Appendix: Chapter 6 Case Study Data

The first 24 observations of Table 4 represent the original boiler temperature data in the first analysis of Case Study 1 as presented in Chapter 6. The final 8 observations of the table were randomly generated around a multivariate normal distribution with parameters given by (6.31) and (6.30) and are rounded to the nearest whole number.

Table 5 is the simulated data set used for Case Study 2 in Chapter 6. At time point 91 the control chart signals. At this point, we replicate the scenario where the industrial process is halted and proceed to conduct a retrospective analysis on data points 1 through 91 to determine the location of the change-point, the components that have changed, and updated parameters for the out of control data set.

Table 4: Boiler Temperature Data Set

Obs	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6	Dim 7	Dim 8
1	507	516	527	516	499	512	472	477
2	512	513	533	518	502	510	476	475
3	520	512	537	518	503	512	480	477
4	520	514	538	516	504	517	480	479
5	530	515	542	525	504	512	481	477
6	528	516	541	524	505	514	482	480
7	522	513	537	518	503	512	479	477
8	527	509	537	521	504	508	478	472
9	530	512	538	524	507	512	482	477
10	530	512	541	525	507	511	482	476
11	527	513	541	523	506	512	481	476
12	529	514	542	525	506	512	481	477
13	522	509	539	518	501	510	476	475
14	532	515	545	528	507	511	481	478
15	531	514	543	525	507	511	482	477
16	535	514	542	530	509	511	483	477
17	516	515	537	515	501	516	476	481
18	514	510	532	512	497	512	471	476
19	536	512	540	526	509	512	482	477
20	522	514	540	518	497	514	475	478
21	520	514	540	518	501	514	475	478
22	526	517	546	522	502	516	477	480
23	527	514	543	523	502	512	475	476
24	529	518	544	525	504	516	479	481
25	529	514	544	524	510	514	480	482
26	536	514	552	526	510	514	484	480
27	520	517	545	521	503	515	475	482
28	514	513	540	515	505	514	477	479
29	529	513	543	527	511	508	482	477
30	529	513	546	524	509	511	478	478
31	521	515	539	520	506	514	479	480
32	519	510	539	515	502	510	474	477

Table 5: Once the alarm signals, we conduct an analysis on elements 1 through 91.

Obs.	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6	Obs.	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6
1	1.10	-0.01	-0.83	-1.18	0.54	-0.65	51	-1.71	0.30	-0.41	0.43	-0.71	-1.37
2	-0.54	1.28	1.12	0.67	1.36	-0.67	52	-0.22	0.25	-0.98	-0.48	-0.92	-0.96
3	-0.40	-0.32	-0.62	-2.22	-0.90	-1.24	53	-1.51	0.38	0.05	0.07	-0.35	-0.82
4	0.21	1.00	-0.27	-0.50	0.43	0.22	54	-0.99	-0.05	0.17	0.07	0.16	1.01
5	-0.65	0.13	-0.15	0.77	-0.58	1.08	55	0.06	-1.85	-1.65	1.46	-0.60	-0.76
6	0.46	-0.28	0.04	0.96	0.56	-0.41	56	-0.23	-0.99	-0.59	-2.05	-1.94	-0.79
7	0.72	0.77	0.75	-1.94	-0.71	0.52	57	0.33	-1.46	-0.35	-0.97	-0.15	-2.19
8	-0.69	0.06	0.50	-0.22	0.12	1.07	58	0.06	1.18	-1.85	-0.42	0.52	-0.41
9	1.80	0.04	2.07	1.17	-0.13	1.73	59	-0.46	0.47	-0.76	1.04	-1.09	-1.12
10	-0.66	0.23	-0.81	-0.59	0.69	1.33	60	-1.56	-0.32	0.23	-0.52	-0.74	-2.79
11	-0.09	-0.23	-0.19	0.42	0.84	0.45	61	-0.14	0.19	1.10	0.37	-1.46	-0.18
12	-2.13	-0.42	-1.57	-1.32	-1.33	-0.19	62	1.60	0.53	1.27	0.58	1.56	0.89
13	0.14	-2.43	-2.40	0.51	-1.81	-1.39	63	-0.50	-0.31	0.85	2.28	1.21	0.63
14	-0.81	-0.45	-1.00	-0.19	-0.19	0.65	64	-0.76	-1.39	0.70	-2.09	-0.60	0.21
15	1.18	2.41	1.80	2.44	1.50	-0.42	65	-0.31	-1.35	-0.24	-1.24	-1.29	0.57
16	-0.29	0.03	1.48	0.93	-0.28	1.25	66	0.83	-0.23	-0.15	-0.01	0.13	-0.09
17	-0.26	-0.58	-0.91	-0.03	-0.60	0.39	67	-0.14	-0.61	0.42	0.62	-0.07	1.63
18	-0.43	0.72	-0.59	0.38	-0.90	0.23	68	0.61	-1.28	-0.83	-0.79	1.11	1.06
19	-0.15	0.90	1.01	-1.67	0.18	0.76	69	0.23	-0.53	-0.02	-1.02	0.93	-1.96
20	-0.27	-1.61	1.21	-0.13	-0.91	-0.31	70	0.05	-0.37	1.08	-1.34	-0.33	-0.48
21	-0.85	0.88	-0.23	0.61	-0.09	-0.76	71	-0.67	-0.74	-0.79	0.90	-0.19	-0.01
22	-0.98	0.96	0.27	-0.07	-0.34	-0.38	72	1.26	0.48	0.22	0.65	2.10	1.76
23	0.13	-0.77	-1.40	-0.21	-1.80	-0.79	73	-1.02	0.40	0.29	-0.03	1.25	0.36
24	-0.27	1.41	-2.42	0.00	0.76	-0.13	74	-1.00	-0.03	0.30	-0.39	0.29	0.25
25	1.77	0.67	1.38	0.82	2.22	1.65	75	-2.08	-0.88	-0.51	-0.19	0.39	1.27
26	0.56	-1.07	0.71	1.02	1.02	-0.48	76	0.98	0.05	1.78	0.68	2.68	-0.67
27	1.38	-1.71	0.49	-0.70	-1.60	-1.57	77	-0.31	0.19	1.36	0.27	-1.10	0.08
28	1.96	1.04	-0.86	-0.44	-0.82	1.27	78	0.63	0.88	0.50	-0.39	-1.03	0.87
29	1.76	0.49	2.27	0.26	0.60	0.46	79	-0.92	0.33	0.01	-1.33	-0.08	0.12
30	0.05	0.09	0.19	1.35	-1.52	-0.21	80	-1.67	0.36	1.42	1.17	0.14	0.35
31	-1.08	-0.27	-0.57	-1.52	0.53	-0.98	81	-1.64	-1.04	-0.39	-0.46	0.46	2.44
32	-0.41	-0.04	-1.35	-0.58	-0.51	-0.59	82	-0.02	0.88	0.97	0.27	0.96	1.32
33	1.08	1.52	0.83	1.35	-0.01	0.80	83	0.22	0.91	0.38	0.63	0.28	3.36
34	-0.77	-0.21	-0.73	-0.61	0.68	0.29	84	0.72	-0.23	0.19	0.89	0.63	0.69
35	-0.14	-0.84	-0.46	-0.72	-1.54	-1.44	85	-0.64	-0.43	0.79	0.06	1.27	2.05
36	0.80	2.88	1.02	0.26	-0.28	0.62	86	-0.04	1.18	-0.43	-0.78	1.00	2.46
37	-0.39	0.26	-1.01	-0.71	-1.40	-1.01	87	0.34	-1.34	-0.36	0.09	1.29	2.17
38	1.02	0.65	-0.52	1.34	0.04	0.42	88	0.16	-0.33	0.84	0.07	1.51	1.38
39	1.11	1.86	0.38	1.52	0.96	0.15	89	-0.30	1.64	1.54	1.72	-0.09	1.76
40	1.07	1.13	-0.82	-0.78	0.77	-0.35	90	0.86	-0.99	0.94	-1.17	0.01	2.80
41	-1.38	-0.57	-0.65	-0.75	-0.27	-0.53	91	-0.19	0.31	1.41	-0.27	1.82	3.73
42	0.95	0.65	-0.02	-0.07	0.48	0.51	92	-0.13	-1.11	-0.04	1.07	0.49	1.38
43	0.72	0.74	1.36	1.16	-0.37	0.77	93	0.49	0.19	1.67	-0.15	0.54	2.65
44	-0.46	-1.69	-1.30	0.44	-0.14	-0.10	94	-2.18	-0.70	-1.62	-0.04	1.07	1.58
45	-0.68	-2.11	-1.08	-0.99	0.12	-1.66	95	0.41	-0.17	-0.46	-0.92	0.22	2.75
46	3.02	1.26	0.36	2.03	0.35	2.65	96	-0.36	-0.24	-0.53	-2.01	-0.47	1.33
47	-0.85	-1.73	0.68	-0.34	0.12	-0.04	97	0.31	0.59	1.09	0.84	0.78	1.13
48	1.26	0.62	0.59	0.62	2.13	-0.05	98	0.74	0.89	1.54	2.08	0.41	3.63
49	0.93	-0.06	0.04	0.09	1.08	-0.92	99	0.10	0.67	0.37	1.03	1.00	1.27
50	0.04	-0.88	-0.43	0.07	1.12	1.50	100	0.01	-1.66	-1.84	-0.12	1.82	1.56

Bibliography

[1] A. Mostaf and A. Ghorbal, "Bayesian and Non-Bayesian Analysis for Random Change Point Problem using Standard Computer Packages," *International Journal of Mathematical Archive*, vol. 10, pp. 1963–1979, 2011.

[2] S. Mallat, *A Wavelet Tour of Signal Processing The Sparse Way, 3rd Edition*. Burlington,MA 01803: Academic Press, 2009.

[3] E. Càndex and M. Wakin, "An Introduction to Compressive Sampling," *IEEE Signal Processing Mgazine*, vol. 21, pp. 21–30, 2008.

[4] B. Vidakovic, *Statistical Modeling by Wavelets*. 222 Rosewood Drive, Canvers, MA 01923: John Wiley & Sons, 1999.

[5] V. Temlyakov, *Greedy Approximation*. Cambridge, United Kingdom: Cambridge University Press, 2011.

[6] Y. Katznelson, *An Introduction to Harmonic Analysis 3rd Edition*. The Edinburgh Building, Cambridge CB2 8RU, UK: Cambridge Mathematical Library, 2004.

[7] J. Cryer and K. Chan, *Time Series Analysis with Applications in R*. New York, NY 10013: Springer Science+Business Media, LLC, 2008.

[8] C. Heil, *A Basis Theory Primer*. Atlanta, GA 30332: Birkhauser, 1998.

[9] I. Daubechies, *Ten Lectures on Wavelets*. 3600 University City Science Center, Philadelphia, Pennsylvania 19104: Society for Industrial and Applied Mathematics, 1992.

[10] G. Nason, *Wavelet Methods in Statistics with R*. 233 Spring Street, New York, NY 10013: Springer Science+Business Media, LLC, 2008.

[11] R. Ogden, *Essential Wavelets for Statistical Applications and Data Analysis*. 675 Massachusetts Avenue, Cambridge, MA 02139: Birkhauser, 1997.

[12] S. Mallat, "Theory for multiresolution signal decomposition: The wavelet representaion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[13] B. Carlin and T. Louis, *Bayesian Methods for Data Analysis 3rd edition*. 222 Rosewood Dr., Danvers, MA 01923: CRC Press, 2009.

[14] J. Kruschke, *Doing Bayesian Data Analysis*. 30 Corporate Dr, Burlington, MA 01803: Academic Press/Elsevier, 2011.

[15] M. Ghosh, "Objective Priors: An Introduction for Frequentists," *Statistical Science*, vol. 26, pp. 187–202, 201.

[16] A. D. C. Andrieu, N. Freita, "An Introduction to MCMC for Machine Learning," *Machine Learning*, vol. 50, pp. 5–43, 2003.

[17] C. Geyer, "Practical Markov Chain Monte Carlo," *Statistical Science*, vol. 7, pp. 473–511, 1992.

[18] C. Geyer, *Introduction to Markov chain Monte Carlo. In Handbook of Markov Chain Monte Carlo*. Boca Raton, FL: CRC Press, 2011.

[19] A. Graps, "An Introduction to Wavelets," *IEEE Computer Society*, vol. 2, 1995.

[20] J. Albert, *Bayesian Computation with R*. New York, NY: Springer, 2007.

[21] K. Koch, *Introduction to Bayesian Statistics*. New York, NY: Springer, 2007.

[22] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin, *Bayesian Data Analysis*. Boca Raton, FL: CRC Press, 2014.

[23] P. Qiu, *Introduction to Statistical Process Control*. Boca Raton, FL: CRC Press, 2014.

[24] D. Montgomery, *Introduction to Statistical Quality Control, 6th Edition*. Hoboken, NJ 07030: John Wiley & Sons, Inc, 2009.

[25] R. Ogden and J. Lynch, "Bayesian Analysis of Change-Point Models," *Lecture Notes in Statistics*, vol. 141, pp. 67–82, 1999.

[26] D. Hawkins, "Testing a Sequence of Observations for a Shift in Location," *Journal of the American Statistical Association*, vol. 72, pp. 180–186, 1977.

[27] R. Pan and S. Rigdon, "A Bayesian Approach to Change Point Estimation in Multivariate SPC," *Journal of Quality Technology*, vol. 44, pp. 231–248, 2012.

[28] A. Smith, "A Bayesian Approach to Inference about a Change-point in a Sequence of Random Variables," *Biometrika*, vol. 62, pp. 407–416, 1975.

[29] I. Johnstone and B. Silverman, "Wavelet Threshold Estimators for Data Correlated Noise," *Journal of Royal Statistics Society B*, vol. 59, pp. 319–351, 1997.

[30] D. Donoho and I. Johnstone, "Ideal Spatial Adaption by Wavelet Shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.

[31] K. Mardia, J. Kent, and J. Bibby, *Multivariate Analysis*. New York: Academic Press, 1979.

[32] A. Antoniadis, D. Leporini, and J. Pesquet, "Wavelet thresholding for some classes of non-Gaussian noise," *Statistica Neerlandica*, vol. 56, no. 4, pp. 434–453, 2002.

[33] P. Fryzlewicz, "Data driven wavelet-fisz methodology for nonparametric function estimation," *Electronic Journal of Statistics*, vol. 2, pp. 863–896, 2008.

[34] Y. Wang, "Jump and Sharp Cusp Detection by Wavelets," *Biometrika*, vol. 82, pp. 385–97, 1995.

[35] S. F. et al., "Performance Metrics for Surveillance Schemes," *Quality Engineering*, vol. 20(4), pp. 451–464, 2008.

[36] K. Worsley, "On the Likelihood Ratio Test for a Shift in Location of Normal Populations," *Journal of American Statistical Association*, vol. 74, pp. 365–367, 1979.

[37] J. Chen and A. Gupta, *Parametric Statistical Change Point Analysis*. 233 Spring Street, New York, NY 10013: Birkhauser, 2012.

[38] M. Csorgo and L. Horváth, *Limit Theorems in Change-Point Analysis*. Baffins Lane, Chichester West Sussez P019 IUD, England: John Wiley & Sons, 1997.

[39] R. Kass and A. Raftery, "Bayes Factors," *Journal of the American Statistical Association*, vol. 90, pp. 772–779, 1995.

[40] J. Chen and A. Gupta, "Testing and Locating Variance Change-points with Application to Stock Prices," *Journal of the American Statistical Association*, vol. 92, pp. 739–747, 1997.

[41] C. Inclán, "Detection of Multiple Changes of Variance Using Posterior Odds," *Journal of Business and Economics Statistics*, vol. 11, pp. 189–300, 1993.

[42] B. Abraham and W. Wei, "Inferences about the Parameters of a Time Series Model with Changing Variance," *Technometrics*, vol. 21, pp. 313–320, 1979.

[43] W. Davis, "Robust Methods for Detection of Shifts of the Innovation Variance of a Time Series," *Metrika*, vol. 11, pp. 183–194, 1984.

[44] L. Perreault, E. Parent, J. Bernier, B. Bobée, and E. Parent, "Retrospective multivariate Bayesian change-point analysis: A simultaneous single change in the mean of several hydrological sequences," *Journal of Multivariate Analysis*, vol. 235, pp. 221–241, 2000.

[45] K. Zamba and D. Hawkins, "A Multivariate Change-point Model for Change in Mean Vector and/or Covariance Structure," *Journal of Quality Technology*, vol. 41, no. 3, 2009.

[46] L. Horváth and P. Kokoszka, "Testing for changes in multivariate dependent observations with an application to temperature changes," *Journal of Multivariate Analysis*, vol. 68, pp. 96–119, 1999.

[47] Y. Son and S. Kim, "Bayesian single change point detection in a sequence of multivariate normal observations," *Statistics*, vol. 39, no. 5, pp. 373–387, 2005.

[48] H. G. Müller, "Change-points in nonparametric regression analysis," *The Annals of Statistics*, vol. 20, pp. 737–761, 1992.

[49] G. Ciuperca, "Estimating Nonlinear Regression with and without Change-points by the LAD Method," *Ann Inst Stat Math*, vol. 63, pp. 717–743, 2011.

[50] F. Battaglia and M. K. Protopapas, "Multiregime Models for Nonlinear Nonstationary Time Series," *Computational Statistics*, vol. 27, pp. 319–341, 2012.

[51] D. S. Matteson and N. A. James, "A nonparametric approach for multiple change point analysis of multivariate data," *Journal of the American Statistical Association*, vol. 109, pp. 334–345, 2014.

[52] R. Mason and J. Young, *Multivariate Statistical Process Control with Industrial Applications*. 3600 University City Science Center, Philadelphia, PA 19104-2688: Society for Industrial and Applied Mathematics, 2002.

[53] L. Perreault, J. Bernier, B. Bobée, and E. Parent, "Change-point analysis in hydrometeorological time series. part 1. the normal model revisted," *Journal of Multivariate Analysis*, vol. 235, pp. 221–241, 2000.

[54] M. Wagner, *Handbook of Bioserveillance*. 30 Corporate Drive, Suite 400, Burlington, MA 01803: Elsevier Academic Press, 2006.

[55] A. Jensen and A. la Cour-Harbo, *Ripples in Mathematics: the Discrete Wavelet Transform*. Springer, 2001.

[56] R. Bellman, *Adaptive Control Processes: a Guided Tour (Vol. 4)*. 41 William Street, Princeton, New Jersey 08540: Princeton University Press, 1961.

[57] K. Fukumizu, F. Bach, and M. Jordan, "Dimensionality Reduction for Supervised Learning with Reproducing Kernel Hilbert Spaces," *Journal Of Machine Learning Research*, vol. 5, pp. 73–99, 2004.

[58] I. Fodor, "A Survey of Dimension Reduction Techniques," 2002.

[59] I. Jolliffe, *Principal Component Analysis, Second Edition*. 175 fifth Ave, New York, New York 10010: Springer-Verlag, 2002.

[60] E. Bingham and H. Mannila, "Random Projection in Dimensionality Reduction: Applications to Image and Text Data," *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 245–250, 2001.

[61] R. Horn and C. Johnson, *Matrix Analysis*. 32 Avenue of the Americas, New York, NY: Cambridge University Press, 2012.

[62] P. Frankl and H. Maehara, "The Johnson-Lindenstrauss Lemma and the Sphericity of some Graphs," *Journal of Combinatorial Theory*, vol. 44, pp. 355–362, 1988.

[63] S. Dasgupta and A. Gupta, "An Elementary Proof of a Theorem of Johnson and Lindenstrauss," *Random Structures and Algorithms*, vol. 22, pp. 60–65, 2003.

[64] N. Linial, F. London, and Y. Rabinovich, "The Geometry of Graphs and some of its Algorithmic Applications," *Combinatorica*, vol. 15, pp. 215–245, 1995.

[65] S. Vempala, *The Random Projection Method*. 201 Charles St., Providence, RI 02904: The American Mathematical Society, 2004.

[66] R. Vershynin, "Introduction to the Non-Asymptotic Analysis of Random Matrices," 2010.

[67] P. Green, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, vol. 82, pp. 711–732, 1995.

[68] D. Hastie and P. Green, "Model choice using reversible jump Markov chain Monte Carlo," *Statistica Neerlandica*, vol. 66, pp. 309–338, 2012.

[69] L. Jeffrey, *Manifolds and Differential Geometry*. 201 Charles St, Providence, RI 02904: American Mathematical Society, 2009.

[70] G. Givens and J. Hoeting, *Compuational Statstics*. 111 River St. Hoboken, NJ 07030: John Wiley & Sons, 2013.

[71] H. Chipman, E. George, R. McCulloch, M.C., D. Foster, and R. Stine, "The Practical Implementation of Bayesian Model Selection," *Lecture Notes-Monograph Series*, vol. 38, pp. 65–134, 2001.

[72] J. H. Sullivan and W. H. Woodall, "Change-point Detection of Mean Vector or Covariance Matrix Shifts using Multivariate Individual Observations," *IIE Transactions*, vol. 32, pp. 537–549, 2000.

[73] R. L. Mason, N. D. Tracy, and J. C. Young, "Decomposition of t^2 for Multivariate Control Chart Interpretation," *Journal of Quality Technology*, vol. 27, pp. 99–108, 1995.

Vita Auctoris

Vita Auctoris goes here