A Bayesian Approach to Estimation of a Statistical Change-point in the Mean Parameter for High Dimensional Non-linear Time Series

Darrin Speegle[a] and Robert Steward[a]

[a]Saint Louis University, Department of Mathematics, Saint Louis, MO, USA

ABSTRACT

We propose a semiparametric approach to infer the existence of and estimate the location of a statistical change-point to a nonlinear high dimensional time series contaminated with an additive noise component. In particular, we consider a $p-$dimensional stochastic process of independent multivariate normal observations where the mean function varies smoothly except at a single change-point. Our approach first involves a dimension reduction of the original time series through a random matrix multiplication. Next, we conduct a Bayesian analysis on the empirical detail coefficients of this dimensionally reduced time series after a wavelet transform. We also present a means to associate confidence bounds to the conclusions of our results. Aside from being computationally efficient and straight forward to implement, the primary advantage of our methods is seen in how these methods apply to a much larger class of time series whose mean functions are subject to only general smoothness conditions.

Keywords: wavelet, random matrix, posterior distribution, Schwarz Information Criteria

1. INTRODUCTION

In this article we demonstrate how a technique involving a Bayesian analysis of wavelet detail coefficients from a multivariate time series may be applied in a high dimensional setting to detect and estimate the location of a statistical change-point in mean. The approach involves both a presentation of the so called Bayesian-wavelet change-point estimation equation, $p(\tau|\mathbf{D}^*)$, and pertinent random matrix dimensionality reduction techniques. Unlike other proposed methods in the multivariate case for the statistical change-point problem,[1–4] we relax the usual assumption that each element of the time series on common sides of the change-point location is identically distributed.

The multivariate change-point in high dimensions has applications in such areas as finance,[5] surveillance video,[6] and network traffic analysis.[7] Computational challenges increasingly complicate the change-point problem in high dimensional settings preventing direct applications of low dimensional techniques. While less work has been done on the change-point problem in the high dimensional setting Aston and Kirch,[8] Cho and Fryzlewicz,[9] and Xie et al.[10] have all proposed methods of detecting change-points in the high dimensional setting. As in the lower dimensional case, however, these proposed high dimensional change-point techniques generally do not apply without rather strict assumptions on the time series generating mean function. In particular our approach generalizes to the case where the true underling mean varies smoothly except possibly at a change-point location.

Consider a p-dimensional time series, $\mathbf{X}$, of independent observations $\{\mathbf{x}_i\}_i^N$ for $N \in \mathbb{N}$ where $\mathbf{x}_i$ is a p-dimensional vector, such that

$$\mathbf{x}_i \sim N_p(\boldsymbol{\mu}_i, \Sigma). \tag{1}$$

The mean parameter, $\boldsymbol{\mu}_i$, is assumed to be generated by a p-dimensional mean function, $g(\cdot)$, smoothly changing except possibly at a single point in time where the shift occurs. Let the symbol, τ, represent the change-point location where $\tau \in [1, N-1]$. The question becomes does a statistical change-point exist in the time series? If so, how closely can we estimate the location of this change-point? We assume Σ is an unknown but constant

Further author information:
D.S.: E-mail: speegled@su.edu
R.S.: E-mail: rstewa12@slu.edu

$p \times p$ covariance matrix throughout our time series. When considered in conjunction with its mean function, a particular observation of the time series may alternatively be expressed as

$$\mathbf{x}_i = g(i) + \boldsymbol{\varepsilon}_i \tag{2}$$

where

$$\boldsymbol{\varepsilon}_i \sim N_p(\mathbf{0}, \Sigma). \tag{3}$$

The important point is that the usual assumption of the data being independently and identically distributed is relaxed. Instead we assume only independence between time series elements and the rather mild regularity condition that g be piecewise smooth. Finally, we conclude this article by presenting examples from both simulated and real world data that demonstrate the effectiveness of this method to a wide range of practical problems.

2. BACKGROUND

In this section we review necessary background material that we later apply in our study of the statistical change-point problem in high dimensions. Firstly, we discuss important statistical properties of the discrete wavelet transform when applied to a multivariate time series contaminated with an additive multivariate normal noise component. Secondly, we also present known results from the literature on random matrix theory that will be the key tool in our dimensionality reduction of high dimensional data.

2.1 Discrete Wavelet Transform (DWT) Observations

The discrete wavelet transform[11–13] is a remarkable tool in signal analysis that has been the subject of extensive research in the past 25 years. Given a discrete noisy signal, our objective is to extract information about the true underling signal apart from the additive noise component. As Donoho[14] showed, the detail coefficients from smooth functions have sparse representation in the wavelet domain. 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.[15] Since detail coefficients represent signal change at different scales, on a localized scale we expect high level detail coefficients from smooth functions to have similar representations.[16]

Observe Figure 1 which displays two examples of a smooth one dimensional 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 property in Figure 1 illustrates the sparsity property of the DWT and is a phenomena that we find holds in general for any smoothly varying mean functions sharing a common change-point.

Next we consider the statistical properties of the detail coefficients for multivariate time series with an additive noise component as in equation (2). Suppose we observe a p-dimensional time series, $\mathbf{X}$, of N elements where each row represents a distinct element of the time series. We may transform $\mathbf{X}$ to the wavelet domain by taking a 1-dimensional DWT of $\mathbf{X}$ column by column and collecting the resulting vectors of detail coefficients, $\mathbf{d}^*_{jk}$, into a matrix $\mathbf{D}^*$. The dual subscript notation of each detail coefficient vector denotes the usual detail level (j) and translation (k) index[12] while the * denotes that this is an empirical detail coefficient vector that we actually observe. By the linearity of the DWT we may express each detail coefficient vector as the sum

$$\mathbf{d}^*_{jk} = \mathbf{d}_{jk} + \boldsymbol{\eta}_{jk} \tag{4}$$

where $\mathbf{d}_{jk}$ is the true (but unknown) detail coefficient of the mean function and $\boldsymbol{\eta}_{jk}$ is the DWT of the additive noise component, $\boldsymbol{\varepsilon}_i$. In the case of a Gaussian additive noise component, the statistical properties of random process is retained by the high level detail coefficients after a DWT. That is, if we assume $\boldsymbol{\varepsilon}_i$ is generated from a Gaussian process as in (3), then $\boldsymbol{\eta}_{jk}$ in equation (4) will also be Gaussian.[12,17] Alternatively, we may express each empirical detail coefficient vector as a random variable from the distribution

$$\mathbf{d}_{jk}^{*} \sim N_p(\mathbf{d}_{jk}, \Sigma). \tag{5}$$

Now let $\mathbf{H}$ be a multivariate step function, also p-dimensional and of N elements in length, with a change-point at precisely the same time point as $\mathbf{X}$. Taking the DWT of $\mathbf{H}$ column by column and storing each resulting detail coefficient vector, $\mathbf{q}_{jk}$, we obtain the matrix $\mathbf{Q}$. By the properties of the DWT and as illustrated in Figure 1, we expect the true higher level detail coefficient vectors, $\mathbf{d}_{jk}$, to closely correspond to $\mathbf{q}_{jk}$ so long as $g(\cdot)$ is smooth. If $g(\cdot)$ is a multivariate step function these detail coefficient values will exactly correspond for all detail levels to within a multiplicative factor $\mathbf{\Delta} = [\delta_1, \delta_2, \ldots, \delta_p]$ corresponding to the size of the shift at the change-point. In particular, we model each empirical detail coefficient vector as

$$\mathbf{d}_{jk} \sim N_p(\mathbf{\Delta}\mathbf{q}_{jk}, \Sigma). \tag{6}$$

where $\mathbf{\Delta}\mathbf{q}_{jk} = [\delta_1 q_{jk,1}, \delta_2 q_{jk,2}, \ldots, \delta_p q_{jk,p}]$. This final representation as given by (6) will be the fundamental model assumption we make when developing the Bayesian-wavelet change-point estimation equation in section 3 below.

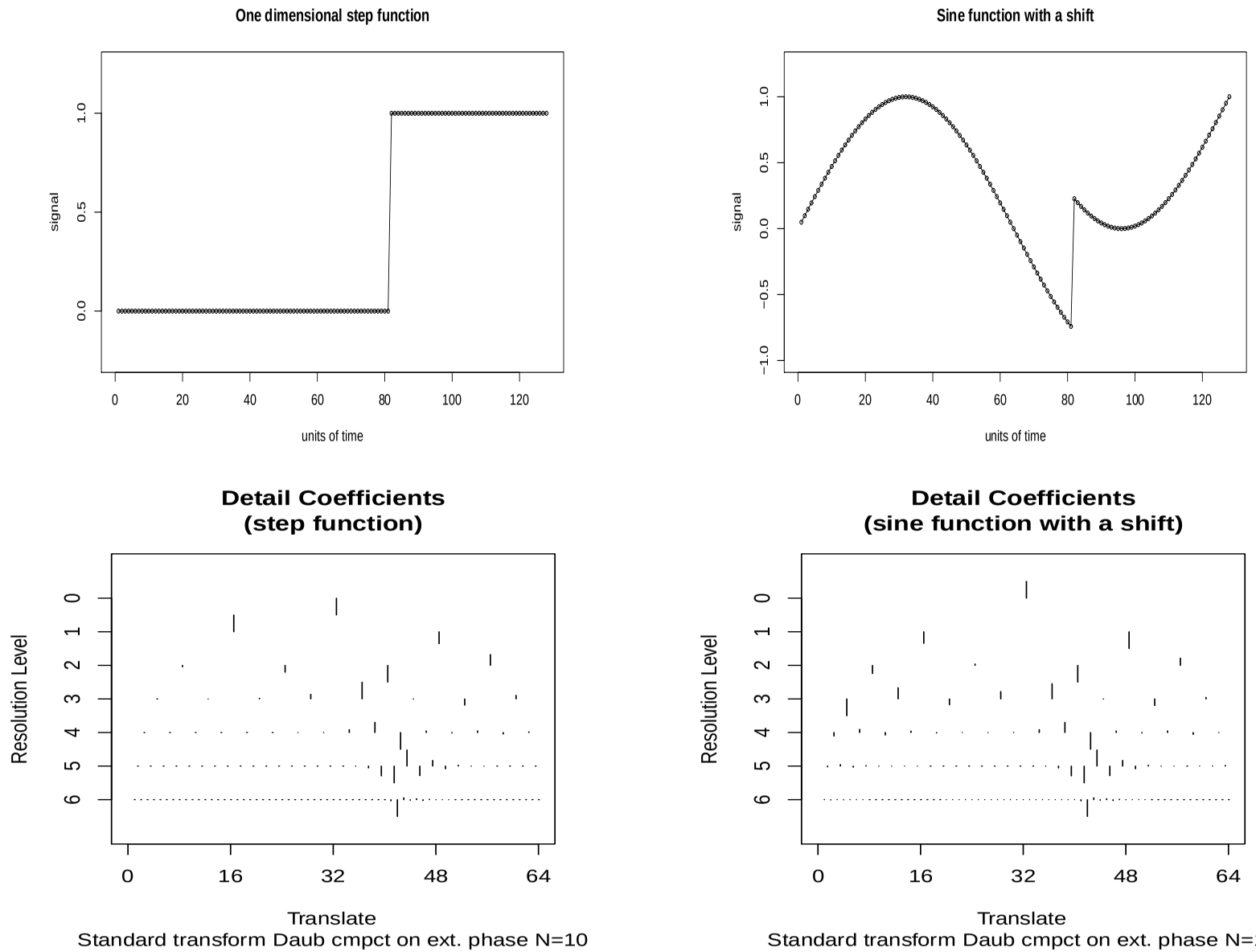


Figure 1: Two example mean functions with a change-point at time point 81 (top) along with their respective detail coefficients (bottom). Each detail level is normalized by its l^{∞}-norm. Notice at the finest three resolution levels the detail coefficients are essentially identical to each other.

2.2 Dimensionality reduction by way of random matrix multiplication

When encountering a time series of high dimensions, invariably a point is reached where direct computations on the data become impractical or impossible due to existing computing limitations. One method when dealing with the so called "curse of dimensionality" conceptually involves approximating the high dimensional time series on a computationally tractable lower dimensional manifold.[18] With any such approximation, information is almost always lost and the question becomes to what extent we may draw conclusions while working in this lower dimensional setting.[19] Random matrix dimensionality reduction offers an attractive practical approach in this case because of its ease and inexpensive means of implementation. Furthermore, known theoretical results exist in the literature for certain special cases that conveniently describe how confidence bounds may be associated with the results obtained from the lower dimensional manifold.

The first result we consider is due to Johnson and Lindenstrauss and describes how the magnitude of a high dimensional vector may change after a random matrix dimension reduction.[20] Let $\mathbf{X} = \{\mathbf{x}_i\}_i^N$ denote a $p-$dimensional time series of length N with a multivariate additive Gaussian noise component and a shift of $\mathbf{\Delta}^X$ at an unknown change-point location, τ. Next we define our dimensionality reduction linear transformation (or interchangeably a matrix) $\mathbf{P}$. Explicitly, we write $\mathbf{P} : \mathbb{R}^p \to \mathbb{R}^d$ where $d \leq p$ such that

$$\mathbf{P} = \frac{1}{\sqrt{d}}\mathbf{A}. \tag{7}$$

Here $\mathbf{A}$ is a random matrix with each entry, a_{ij}, randomly distributed as $a_{ij} \sim N(0,1)$. If we perform a matrix multiplication we obtain a dimensionally reduced time series $\mathbf{Y}^T = \mathbf{P}\mathbf{X}^T$ with a shift vector $\mathbf{\Delta}^Y$ also at time point τ. With the above notation, the Johnson and Lindenstrauss lemma applied in this setting takes the following form.

Lemma 2.1. *Letting $||\cdot||^2$ denote the squared magnitude of a vector and requiring $0 < \varepsilon < 1$, then so long as $d > \frac{2\log\alpha}{\log(1-\varepsilon)-\varepsilon}$, we are at least $100(1-\alpha)\%$ confident that $||\mathbf{\Delta}^Y||^2 \geq (1-\varepsilon)||\mathbf{\Delta}^X||^2$.*

The next result we consider concerns what happens to the largest expected singular value, s_{max}, of a random matrix $\mathbf{P}$ (where $\mathbf{P}$ is in the form given above). In section 3 we will see how this in turn relates to largest singular value of the covariance matrix associated with the additive noise component after a random matrix dimension reduction. Based on Gordon's theorem for Gaussian matrices and Slepian's inequality, Vershynin[21] proves the following proposition.

Proposition 2.2. *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 $\mathbf{P}$.

The difficulty of the statistical change-point problem in any dimension ultimately depends on both the size of the shift at the change-point location and the structure of the covariance of the additive noise component. Clearly larger shifts are easier to detect than smaller shifts while more variance increases the difficulty of correctly estimating the change-point location. In the context of the high dimensional statistical change-point problem, the above two results in this section will be important to better understand the scope and effectiveness of the methods developed below.

3. GENERAL RESULTS

Here we outline our general results that we later apply to the high dimensional change-point problem. We may consider the shift in the mean function at a particular time as the signal within our time series we seek to detect. This shift, however, is obscured by the additive noise component of the time series itself. As a means of quantifying the difficulty of estimating the change-point location in a given time series, we define a form of the signal-to-noise ratio (SNR) as follows:

$$SNR = ||\boldsymbol{\Delta}||/(p^{1/4}\sqrt{\lambda_{max}}) \tag{8}$$

where as usual $||\boldsymbol{\Delta}||$ denotes the magnitude of the shift, p denotes the dimension of the time series, and λ_{max} denotes the maximum eigenvalue from the covariance matrix of the additive noise component. This definition is not crucial to the results to follow; however, it is a useful metric for determining the difficulty of correctly estimating the location of a change-point in a time series. For example, using an $SNR \geq 3$ rule of thumb, we have found through extensive simulation studies for time series of 100 dimensions and below that the Bayesian-wavelet estimation equation correctly estimates the change-point location well over 99% of the time.[22]

3.1 A Model for the Multidimensional Change-point Problem

We begin by deriving a statistical model characterizing the detail coefficients from a multivariate time series of arbitrary dimension. Ogden[23] developed a similar approach for the statistical change-point problem although his method applied only to 1-dimensional time series. Firstly, from Bayes Theorem we obtain its functional analog, the so called posterior distribution function, expressed as

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

where

$$m(X) = \int L(X|\boldsymbol{\theta})p_0(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

Here $\boldsymbol{\theta}$ is the unknown parameter vector, $L(X|\boldsymbol{\theta})$ is the likelihood of observing X given $\boldsymbol{\theta}$, p_0 is the prior distribution of $\boldsymbol{\theta}$, and $m(X)$ is the probability of observing the data set. Notice $m(X)$ does not depend on the parameter vector and essentially becomes a normalizing constant that we can safely discard in our initial analysis and reintroduce as necessary later.

After transforming our multivariate time series to the wavelet domain as previously explained, we observe a matrix of detail coefficient vectors which we denote as $\mathbf{D}^*$. Our parameter vector is comprised of the change-point location τ, the shift to the mean function at the change-point $\boldsymbol{\Delta}$, and the covariance of our additive noise component Σ. There are many possible prior distributions that we may place on our parameter vector depending on what knowledge is known beforehand. For both mathematical convenience and to simultaneously keep our approach as general as possible, we apply Jeffrey's noninformative prior expressed as $p_0(\Sigma, \Delta, \tau) \propto |\Sigma|^{-1/2}$.[24] Modeling our empirical detail coefficient vectors as given by equation (6) and letting f represent a p-dimensional multivariate normal distribution our posterior distribution may now be expressed as

$$\begin{aligned} p(\tau, \Delta, \Sigma|\mathbf{D}^*) &\propto \prod_j \prod_k f(\mathbf{d}^*_{jk}|\Sigma, \tau, \Delta)p_0(\tau, \Delta, \Sigma) \\ &\propto |\Sigma|^{-m/2} \exp\left[-\frac{1}{2} \sum_j \sum_k (\mathbf{d}^*_{jk} - \Delta\mathbf{q}_{jk})^T \Sigma^{-1} (\mathbf{d}^*_{jk} - \Delta\mathbf{q}_{jk}) \right] |\Sigma|^{-1/2}. \end{aligned}$$

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

$$\begin{aligned} p(\tau|\mathbf{D}^*) \propto &\int_{PD(p)} \int_{\mathbb{R}^p} \\ &|\Sigma|^{-(m+1)/2} \exp\left[-\frac{1}{2} \sum_j \sum_k (\mathbf{d}^*_{jk} - \Delta\mathbf{q}_{jk})^T \Sigma^{-1} (\mathbf{d}^*_{jk} - \Delta\mathbf{q}_{jk}) \right] d\Delta d\Sigma. \end{aligned} \tag{9}$$

where $PD(p)$ represents the space of p-dimensional positive definite matrices. After performing the calculation in equation (17) which we provide in the appendix, our marginalized posterior probability function becomes

$$p(\tau|\mathbf{D}^*) \propto C^{-\frac{1}{2}} \left| \sum_j \sum_k \mathbf{d}_{jk}^* \mathbf{d}_{jk}^{*T} - \frac{1}{C} BB^T \right|^{-(m-p-1)/2} \tag{10}$$

where

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

Equation (10) is the Bayesian-wavelet change-point estimation equation. After reintroducing the normalizing constant it becomes a complete marginalized probability distribution where 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. Furthermore, we may calculate any desired credible interval with this estimate to express our degree of confidence in this change-point location estimate.

Our approach to the inference of a statistical change-point involves a model selection approach and an application of the Schwarz information criteria (SIC). Let M_1 denote the model 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 these two models given the observed data. Applying $\arg\max_\tau p(\tau|\mathbf{D}^*)$ to equation (10) and retaining all constants throughout the derivation, we obtain the probability of observing $\mathbf{D}^*$ given M_1:

$$P(\mathbf{D}^*|M_1) = K(2\pi)^{(p-mp)/2}(2)^{mp/2}\Gamma_p(\tfrac{m}{2})C^{-1/2} \left| \sum_j \sum_k \mathbf{d}_{jk}^* \mathbf{d}_{jk}^{*T} - \frac{1}{C} BB^T \right|^{-(m-p-1)/2} \tag{11}$$

where K is a constant common to both models and $\Gamma_p(\cdot)$ is the multivariate gamma function defined as

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

Next we compute the probability of observing $\mathbf{D}^*$ given no statistical change-point exists in the data. The derivation of this probability involves a similar approach to our previous work in obtaining equation (10) only here the shift, $\mathbf{\Delta}$, at the change-point is zero. The probability of observing of observing $\mathbf{D}^*$ given M_2 is expressed as:

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

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

$$\begin{aligned} \Delta(SIC) &= -2(\log P(\mathbf{D}^*|M_2) - \log P(\mathbf{D}^*|M_1)) + (k_2 - k_1)\log N \\ &= -2(\log(P(\mathbf{D}^*|M_2)) - \log P(\mathbf{D}^*|M_1)) + p \log N. \end{aligned} \tag{12}$$

where equation (12) implicitly assumes equal probability of realizing either M_1 or M_2. In certain instances the modeler may have reason to favor one model over the other in which case the prior odds ratio may easily be adjusted according to this additional information.

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.

3.2 Applications in Dimensionality Reduction

While the Bayesian-wavelet change-point estimation equation, $p(\tau|\mathbf{D}^*)$, applies to arbitrary dimensional time series in theory, in practice computational difficulties prevent a direct application of this method for high dimensional time series. We overcome this limitation by firstly dimensionally reducing the high dimensional time series through a random projection to a computationally tractable dimension. Next, we apply the Bayesian-wavelet change-point estimation equation to this reduced dimensional space to estimate the change-point location. Since dimension reduction normally comes at the price of information loss, the question becomes how confident are we of still correctly estimating the change-point location in this lower dimension. Here we may appeal to known results from random matrix theory provided in section 2.2 that may be applied to this question as well.

For ease of notation, we suppress * to indicate "empirical" for the remainder of this article and let $\mathbf{d}^X$, $\mathbf{d}^Y$, $\mathbf{D}^X$, and $\mathbf{D}^Y$ denote empirical detail coefficient vectors and matrices of time series $\mathbf{X}$ and $\mathbf{Y}$, respectively. Now suppose we have a multivariate time series $\mathbf{X}$ along with a dimension preserving linear transformation $\mathbf{T}$, then we may define a new multivariate time series as $\mathbf{Y}^T = \mathbf{T}\mathbf{X}^T$. The question becomes is the Bayesian-wavelet change-point estimation equation invariant to linear transformations? The following theorem answers this question in the affirmative.

Theorem 3.1. *Given a linear transform, $\mathbf{T} \in GL_p(\mathbb{R})$, then $p(\tau|\mathbf{D}^X) = p(\tau|\mathbf{D}^Y)\ \forall\ \tau$ where $\mathbf{Y}^T = \mathbf{T}\mathbf{X}^T$. That is, the Bayesian-wavelet change-point estimation equation is invariant under dimension preserving linear transformations.*

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

$$p(\Sigma^Y, \tau, \mathbf{\Delta}^Y|\mathbf{D}^Y) = \frac{\prod_j \prod_k p(\mathbf{d}^Y_{j,k}|\Sigma^Y, \tau, \mathbf{\Delta}^Y) p_0(\Sigma^Y, \mathbf{\Delta}^Y, \tau)}{m(\mathbf{D}^Y)}. \tag{13}$$

Letting $s = \frac{-(m-p)}{2}$, we consider the numerator of equation (13) where by equation (11) is equal to:

$$= K(2\pi)^{\frac{p-mp}{2}}(2)^{\frac{mp}{2}}\Gamma_p(\tfrac{m}{2})C^{-\frac{1}{2}}\left|\sum_j\sum_k \mathbf{d}^Y_{jk}{\mathbf{d}^Y_{jk}}^T - \frac{1}{C}B^Y{B^Y}^T\right|^s \tag{14}$$

after the nuisance parameters are integrated out. By the linearity of the DWT, we have

$$\mathbf{d}^Y_{jk} = \mathbf{T}\mathbf{d}^X_{jk}.$$

This means for a given jk we have

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

Applying this to equation (14)

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

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

$$\begin{aligned} m(\mathbf{D}^Y) &= \int L(\mathbf{D}^Y|\Sigma^Y, \mathbf{\Delta}^Y, \tau) p_0(\Sigma, \mathbf{\Delta}^Y, \tau) d\Sigma^Y d\mathbf{\Delta}^Y d\tau \\ &= \sum_{\tau=1}^{N-1} (2\pi)^{\frac{p-mp}{2}} (2)^{\frac{mp}{2}} \Gamma_p(\tfrac{m}{2}) C^{-\frac{1}{2}} |\mathbf{T}|^{2s} \left| \Big(\sum_j \sum_k \mathbf{d}_{jk}^X \mathbf{d}_{jk}^{X\,T} - \frac{1}{C} B^X B^{X\,T} \Big) \right|^s \\ &= |\mathbf{T}|^{2s} \sum_{\tau=1}^{N-1} (2\pi)^{\frac{p-mp}{2}} (2)^{\frac{mp}{2}} \Gamma_p(\tfrac{m}{2}) C^{-\frac{1}{2}} \left| \Big(\sum_j \sum_k \mathbf{d}_{jk}^X \mathbf{d}_{jk}^{X\,T} - \frac{1}{C} B^X B^{X\,T} \Big) \right|^s \end{aligned}$$

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

3.2.1 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 using equation (10). 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.

With theorem 3.1 in hand, we have the following proposition which will be useful later when "standardizing" the time series in this chapter. Proposition 3.2 recalls covariance matrix diagonalization properties. Since we obtain our diagonal covariance matrix by linear transformations, theorem 3.1 ensures us that our ability to estimate the change-point location of our time series remains unchanged.

Proposition 3.2. *(Whitening Transform) Given a multivariate time series* $\mathbf{X}$ *with covariance matrix* Σ^X*, there exists a linear transformation,* $\mathbf{T}$*, such that* $\mathbf{T}\mathbf{X}^T = \mathbf{Y}^T$ *with an associated covariance matrix,* Σ^Y*, such that* Σ^Y *is the identity matrix.*

Proof. This is a known result. See, for example, Bell.[25]

The upshot of theorem 3.1 and proposition 3.2 is that from the standpoint of $p(\tau|\mathbf{D}^X)$, for an arbitrary time series, there exists an equivalent time series with the covariance structure of $\mathbf{\Sigma}^Y$ as in proposition 3.2. So in particular, for the purpose of our analysis all we need to consider are those time series with an identity covariance matrix structure.

Our purpose here is to develop a lower confidence bound for $p(\tau|\mathbf{D}^X)$ after a dimension reduction for correctly estimating the change-point location. By assuming the covariance matrix is simply the identity, we may employ known results from random matrix theory thereby simplifying calculations. For this reason, we assume our original time series $\mathbf{X}$ has already been standardized with covariance matrix $\Sigma^X = I$, where I represents the identity matrix.

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 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

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

and

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

We may represent $\mathbf{P}$ as a $d \times p$ matrix with entries p_{ij}, so we obtain our dimensionally reduced time series, $\mathbf{Y}$, by a matrix multiplication

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

Then using the definition of the expected value we have

$$\begin{aligned}
\boldsymbol{\Delta}^Y &= \boldsymbol{\mu}_{\tau+1}^Y - \boldsymbol{\mu}_\tau^Y \\
&= E\begin{pmatrix} y_{(\tau+1),1} \\ y_{(\tau+1),2} \\ \vdots \\ y_{(\tau+1),d} \end{pmatrix} - E\begin{pmatrix} y_{\tau 1} \\ y_{\tau 2} \\ \vdots \\ y_{\tau d} \end{pmatrix} = E\begin{pmatrix} \sum_{i=1}^p p_{1i} x_{(\tau+1),i} \\ \sum_{i=1}^p p_{2i} x_{(\tau+1),i} \\ \vdots \\ \sum_{i=1}^p p_{di} x_{(\tau+1),i} \end{pmatrix} - E\begin{pmatrix} \sum_{i=1}^p p_{1i} x_{\tau i} \\ \sum_{i=1}^p p_{2i} x_{\tau i} \\ \vdots \\ \sum_{i=1}^p p_{di} x_{\tau i} \end{pmatrix} \\
&= \mathbf{P}(\boldsymbol{\mu}_{\tau+1}^X - \boldsymbol{\mu}_\tau^X) = \mathbf{P}\boldsymbol{\Delta}^X
\end{aligned}$$

So we determine our shift vector after a linear dimension reduction simply by a matrix multiplication. For the remainder of this article we assume $\mathbf{P}$ is a $d \times n$ projection matrix of the form given by equation (7).

As a consequence of theorem 3.1, observe that if the covariance is diagonal, then it is only the magnitude of $\boldsymbol{\Delta}^X$ that effects $p(\tau|\mathbf{X})$. If $\boldsymbol{\Delta}^X$ is known and the projection matrix is random, we can use similar calculations as in proposition 3.3 to determine the expected value of $||\boldsymbol{\Delta}^Y||^2$. Letting $\boldsymbol{\Delta}^X = (\delta_1^X, \delta_2^X, \ldots, \delta_p^x)^T$ and $\boldsymbol{\Delta}^Y = (\delta_1^Y, \delta_2^Y, \ldots, \delta_d^Y)^T$, we have our next proposition.

Proposition 3.4. *Given a random matrix, $\mathbf{P}$, as in equation (7) and $\boldsymbol{\Delta}^Y$ as in proposition 3.3. then $E||\boldsymbol{\Delta}^Y||^2 = ||\boldsymbol{\Delta}^X||^2$*

Proof. This proof simply involves taking the expected value of a vector after performing a matrix multiplication and is therefore omitted.

The difficulty of estimating the location of a statistical change-point in a high dimensional time series not only depends the characteristics of 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 3.5. *Under our random projection, the expected covariance matrix of our transformed time series is*

$$E[\Sigma^Y] = \frac{p}{d} I$$

where I is the identity matrix.

Proof. A direct calculation by our above definitions gives

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

From propositions 3.4 and 3.5 we obtain expected values for the shift vector magnitude and covariance matrix structure of our transformed time series after a dimensionality reduction by linear transformation. From lemma 2.1, we further obtain confidence bounds for the magnitude of the shift vector after a dimension reduction. In particular, recall from section 2 that if $\mathbf{P}$ is a $d \times n$ dimension reduction matrix, then so long as

$$d > \frac{2\log\alpha_1}{\log(1-\varepsilon)-\varepsilon} \tag{15}$$

we are at least $100(1-\alpha_1)\%$ confident that $||\mathbf{\Delta}^Y||^2 \geq (1-\varepsilon)||\mathbf{\Delta}^X||^2$.

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 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. We note this is equivalent to finding the expected value and confidence level for the largest singular value, s_{max}, of $\mathbf{P}$. Recalling proposition 2.2 we have with probability of at least $\alpha_2 = 1 - 2\exp(-t^2/2)$ and for every $t \geq 0$

$$s_{max} \leq \sqrt{\frac{p}{d}} + 1 + \frac{t}{\sqrt{d}} \tag{16}$$

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

3.2.2 Discussion

We now bring all the results together in the context of inferring the existence of and location of a change-point in a high dimensional time series. Given a high dimensional time series with the conditions listed in section 1, suppose we are interested in determining the location of a statistical change-point. The first step is to perform a dimension reduction by way of random matrix multiplication to a computationally tractable lower dimension of say 100 or less. We may then apply equation (12) to infer the existence of change-point and apply equation (10) to estimate the location of the change-point along with an associated credible interval.

The next question becomes how far can we dimensionally reduce the time series and how confident can we be of our results? Here we appeal to results from simulation studies and theoretical results from random matrix theory. Firstly, if we calculate a value $SNR > 3$ with equation (8), then for most practical purposes we may assume we are guaranteed of both correctly inferring the existence of a change-point and estimating its location within a time series. In fact this SNR value is actually quite conservative as very good results are also observed for SNR values much lower than 3 for the majority of cases as well.

The question still remains, however, how certain are we of our computed SNR after a random matrix dimension reduction? 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. Explicitly, these respective confidence bounds are determined from lemma 2.1 and proposition 2.2 based on the dimension of the original time series, the dimension of the dimensionally reduced time series, and the desired confidence level.

4. NUMERICAL EXAMPLES

4.1 Illustrative Example

Consider a time series of length 128 which consists of 256×256 pixel images of John Lennon where we add a random multivariate normal component with identity covariance structure to each image of the time series. 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 $p(\tau|\mathbf{D}^X)$ 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 2. We consider the question how far can we dimensionally reduce the time series and still be at least 95% confident of correctly finding the change-point location?

Firstly, we note the small dark box in Figure 2 (right) corresponds to a shift vector magnitude of $||\mathbf{\Delta}^X|| = 350$. For $\alpha_1 = .025$ and $\epsilon = .1$, we apply equation (15) to conclude that after a random matrix dimension reduction down to $d = 36$, we are at least 97.5% confident that $||\mathbf{\Delta}^Y|| \geq 332$. Now letting $\alpha_2 = .025$, we apply equation (16) to conclude that after the same random matrix dimension reduction $s_{max} \leq 44$. Computing the SNR using equation (8), we combine these results to conclude with at least 95% confidence that $SNR > 3$.

Based on our empirical rule of thumb of $SNR > 3$, we have very high confidence of correctly estimating the change-point location in the time series after a dimension reduction. We verify this result by generating a random matrix and dimensionally reducing the sequence of images down to 36 dimensions and applying equation (10) to estimate the change-point location. Observing the posterior probability plot in Figure 3, we find highest probability change-point is in fact tightly concentrated around the true change-point location at time point 95.

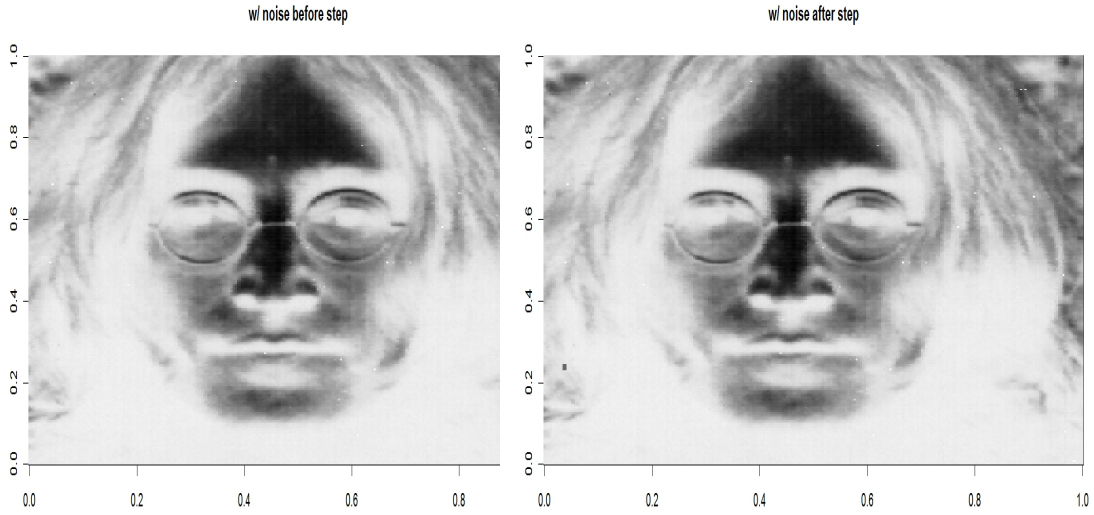


Figure 2: Notice the small dark box on the right picture which is the shift.

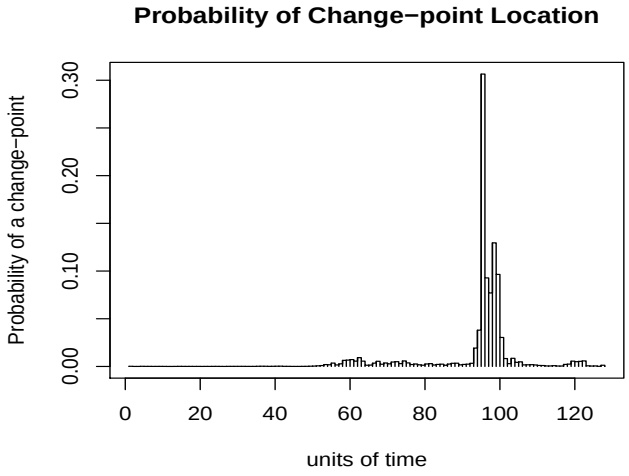


Figure 3: Posterior plot of change-point location correctly estimating the change-point at point 95 with an 80% credible interval of 90-100.

4.2 Practical Example

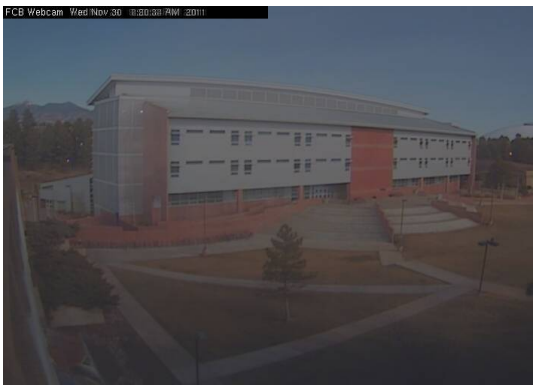
The archive of many outdoor scenes (AMOS) is large data set comprising long-term time lapse photos from over 32,000 publically available stationary cameras from around the world. Many AMOS data set related research projects are currently underway related to such problems as geolocation, camera calibration, and the automatic annotation of objects in a scene (cite website). The AMOS data set projects typically work under the assumption that the camera do in fact remain station. Occasionally, however, a camera shift does inadvertently occur for a variety of reasons thus potentially contaminating the image data. In such cases it would be helpful to identify if a shift has occurred in the data set and if so at what time point.

We apply the change-point inference and location estimation methods developed in this paper to four of these camera data sets; one with a known shift and three others chosen at random. To account for such factors as the variability of ambient light within a 24 hour period and small image shifts we process the images with the following algorithm.

1. Average the images within each calendar day
2. Extract edges of images by a convolution involving the Laplacian operator
3. Convert each image to a vector and sort the values from highest to lowest
4. Apply a random matrix dimension reduction down to 10 dimensions.

With the dimensionally reduced time series, we apply equation (12) to determine the existence of a shift followed by equation (10) to estimate the change-point location if a shift is detected. Working with the images from AMOS, we found a $\Delta SIC < 3$ threshold, which was very effective in detecting the presence of a small change-point with simulated data sets,[22] was too low a cut-off with this actual data set for large shifts. A clear interpretation is that small shifts are typically present in any AMOS image sequence even though they may not be shifts with which we are particularly concerned. Instead a value of $\Delta SIC < 30$ appears to be a better threshold for the large shifts we are interested in detecting. By setting our threshold value at 30, minor shifts in the image sequences are ignored while large shifts will still be detected. Choosing 128 consecutive available days in each data set process the images according to the above algorithm. Example images averaged of a single calendar from both the beginning and end of each data set are displayed in Figure 4 below. Notice the shift in the final pair of adjacent images.

Appling equation (12) to the first three sequence of images returns values of 2.63, 9.99, and 11.39 respectively. With the $\Delta SIC < 30$ threshold, we conclude no large shift occurs in these image sequences. In the final image sequence, equation (12) returns a value of 89.92 implying the existence of a large shift. Next applying equation (10) to this final image sequence returns a change-point location at day 66 with 95% credible interval of [65,69]. The true change-point location appears to be at day 68.





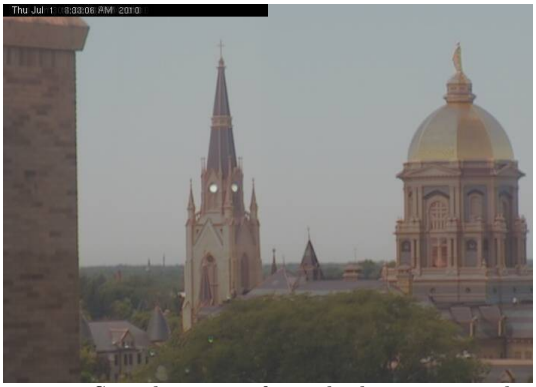
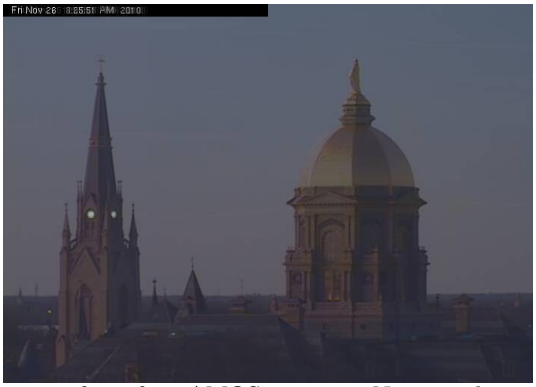



Figure 4: Sample images from the beginning and end of image sequences from four AMOS cameras. Notice only in the bottom image is a large shift apparent.

5. CONCLUSION

In this paper we developed a method to both detect and estimate a statistical change-point in a high dimensional nonlinear time series. The primary advantage of the Bayesian-wavelet change-point estimation equation, $p(\tau|\mathbf{D}^X)$, over other known methods in change-point analysis is seen in how it easily adapts to cases where the true underlying mean function varies smoothly except possibly at the change-point location. Since numerical complications prevent us from directly applying $p(\tau|\mathbf{D}^X)$ to high dimensional time series, we first project the original time series down to a computationally tractable approximation space and then apply $p(\tau|\mathbf{D}^X)$. Finally section 3.2 provides a theoretical answer to the question of how far can we dimensionally reduce the original time series and still be able to be reasonably certain of detecting and estimating the change-point location.

Our change-point detection method involved a model selection approach. As we saw in both the illustrative and the practical example involving the AMOS data set, ΔSIC may be set to detect either very small shifts or to ignore small shifts and only capture large shifts. In cases where a data set is fairly clean in the sense that the additive noise component closely conforms to a constant multivariate normal distribution, simulation results from[22] support using a value of $\Delta SIC < 3$ as a cut-off. Actual data sets such as the AMOS images often have various anomalies such as outliers or small shifts that in fact represent statistical change-points albeit not the ones the modeler is interested in discovering. In such cases the flexibility of our approach allows the modeler to simply adjust the ΔSIC cut-off point to a level where only change-points of interest are detected.

As a final point we note these methods may be extended to the case of multiple change-points. A straight forward approach involves an application of the so called binary segmentation algorithm. That is we may apply the following algorithm for some threshold value C

1. Perform a random matrix multiplication to dimensionally reduce the time series to a computationally tractable dimension.
2. Apply eqauation (12) to the dimensionally reduced time series.

   If $\Delta(SIC) < C$ terminate the algorithm and conclude time series has no change-points.
3. Apply equation (10) and record change point location τ.
4. Segment the original time series into two time series from elements 1 through τ and $\tau + 1$ through N.
5. Return to step 2 for each segment.

The algorithm runs until all segments terminate.

Appendix

We derive equation (10) beginning with the posterior distribution

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

Here, m represents the actual number of detail coefficients used in the analysis. We integrate out $\mathbf{\Delta}$ and Σ to obtain the the marginal posterior distribution function

$$p(\tau|\mathbf{D}) \propto \int_{PD(p)} \int_{\mathbb{R}^p} |\Sigma|^{-(m+1)/2} \exp\left[-\frac{1}{2} \sum_j \sum_k (\mathbf{d}_{jk} - \Delta\mathbf{q}_{jk})^T \Sigma^{-1} (\mathbf{d}_{jk} - \Delta\mathbf{q}_{jk}) \right] d\Delta d\Sigma. \tag{17}$$

where $PD(p)$ represents the space of p-dimensional positive definite matrices.

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 (17) as

$$p(\tau|\mathbf{D}) \propto \int_{PD(p)} \int_{\mathbb{R}^p} |\Sigma|^{-(m+1)/2} \exp\Big[-\frac{1}{2} \sum_j \sum_k (\mathbf{d}_{jk} - q_{jk}\mathbf{\Delta})^T \Sigma^{-1} (\mathbf{d}_{jk} - q_{jk}\mathbf{\Delta}) \Big] d\mathbf{\Delta} d\Sigma. \tag{18}$$

Expanding the exponent of equation (18) we obtain

$$\begin{aligned} p(\tau|\mathbf{D}) &\propto \int_{PD(p)} \int_{\mathbb{R}^p} |\Sigma|^{-(m+1)/2} \\ &\quad \times \exp\Big[-\frac{1}{2} \sum_j \sum_k \Big(\mathbf{d}_{jk}^T \Sigma^{-1} \mathbf{d}_{jk} + q_{jk} \mathbf{\Delta}^T \Sigma^{-1} q_{jk} \mathbf{\Delta} - q_{jk} \mathbf{\Delta}^T \Sigma^{-1} \mathbf{d}_{jk} - \mathbf{d}_{jk}^T \Sigma^{-1} q_{jk} \mathbf{\Delta} \Big) \Big] d\mathbf{\Delta} d\Sigma \\ &= \int_{PD(p)} \int_{\mathbb{R}^p} |\Sigma|^{-(m+1)/2} \exp\Big[-\frac{1}{2} \Big(A + C \mathbf{\Delta}^T \Sigma^{-1} \mathbf{\Delta} - \mathbf{\Delta}^T \Sigma^{-1} B - B^T \Sigma^{-1} \mathbf{\Delta} \Big) \Big] d\mathbf{\Delta} d\Sigma \end{aligned} \tag{19}$$

where

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

Continuing from equation (19) we provide the following detailed calculations:

$$\begin{aligned} p(\tau|\mathbf{D}) &= \int_{PD(p)} \int_{\mathbb{R}^p} |\Sigma|^{-(m+1)/2} \exp\Big[-\frac{C}{2} \Big(\frac{A}{C} + \mathbf{\Delta}^T \Sigma^{-1} \mathbf{\Delta} - \mathbf{\Delta}^T \Sigma^{-1} \frac{B}{C} - \frac{B^T}{C} \Sigma^{-1} \mathbf{\Delta} \Big) \Big] d\mathbf{\Delta} d\Sigma \\ &= \int_{PD(p)} |\Sigma|^{-(m+1)/2} \exp\Big[-\frac{1}{2} \Big(A - \frac{1}{C} B^T \Sigma^{-1} B \Big) \Big] \int_{\mathbb{R}^p} \exp\Big[-\frac{C}{2} \Big((\mathbf{\Delta} - \frac{B}{C})^T \Sigma^{-1} (\mathbf{\Delta} - \frac{B}{C}) \Big) \Big] d\mathbf{\Delta} d\Sigma \\ &= \int_{PD(p)} |\Sigma|^{-(m+1)/2} |\Sigma|^{1/2} C^{-1/2} \exp\Big[-\frac{1}{2} \Big(A - \frac{1}{C} B^T \Sigma^{-1} B \Big) \Big] d\Sigma \\ &= \int_{PD(p)} \Sigma^{-m/2} C^{-1/2} \exp\Big[-\frac{1}{2} \Big(\sum_j \sum_k \mathbf{d}_{jk}^T \Sigma^{-1} \mathbf{d}_{jk} - \frac{1}{C} B^T \Sigma^{-1} B \Big) \Big] d\Sigma \\ &= \int_{PD(p)} \Sigma^{-m/2} C^{-1/2} \exp\Big[-\frac{1}{2} \Big(tr\big(\Sigma^{-1} \sum_j \sum_k \mathbf{d}_{jk} \mathbf{d}_{jk}^T - \frac{1}{C} B B^T \big) \Big) \Big] d\Sigma \\ &\propto C^{-\frac{1}{2}} \Big| \sum_j \sum_k \mathbf{d}_{jk} \mathbf{d}_{jk}^T - \frac{1}{C} B B^T \Big|^{-(m-p-1)/2} \end{aligned} \tag{20}$$

where in the last step equation (10) follows by dropping multiplicative constants and applying the known form of the Wishart distribution.

REFERENCES

[1] Chen, J. and Gupta, A., [*Parametric Statistical Change Point Analysis*], Birkhauser, New York, NY (2012).

[2] Horváth, L. and Kokoszka, P., "Testing for changes in multivariate dependent observations with an application to temperature changes," *Journal of Multivariate Analysis* **68**, 96–119 (1999).

[3] Perreault, L., Parent, E., Bernier, J., Bobée, B., and Parent, E., "Retrospective multivariate Bayesian change-point analysis: A simultaneous single change in the mean of several hydrological sequences," *Journal of Multivariate Analysis* **235**, 221–241 (2000).

[4] Son, Y. and Kim, S., "Bayesian single change point detection in a sequence of multivariate normal observations," *Statistics* **39**(5), 373–387 (2005).

[5] Fan, J., Lv, J., and Qi, L., "Sparse high dimensional models in economics," *Annual review of economics* **3**, 291 (2011).

[6] Lee, K. and Kriegman, D., "Online learning of probabilistic appearance manifolds for video-based recognition and tracking," in [*Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*], **1**, 852–859, IEEE (2005).

[7] Lévy-Leduc, C. and Roueff, F., "Detection and localization of change-points in high-dimensional network traffic data," *The Annals of Applied Statistics* , 637–662 (2009).

[8] Aston, J. and Kirch, C., "Change points in high dimensional settings," *arXiv preprint arXiv:1409.1771* (2014).

[9] Cho, H. and Fryzlewicz, P., "Multiple-change-point detection for high dimensional time series via sparsified binary segmentation," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (2014).

[10] Xie, Y., Huang, J., and Willett, R., "Change-point detection for high-dimensional time series with missing data," *Selected Topics in Signal Processing, IEEE Journal of* **7**(1), 12–27 (2013).

[11] Daubechies, I., [*Ten Lectures on Wavelets*], Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania (1992).

[12] Nason, G., [*Wavelet Methods in Statistics with R*], Springer Science+Business Media, LLC, New York, NY (2008).

[13] Ogden, R., [*Essential Wavelets for Statistical Applications and Data Analysis*], Birkhauser, Cambridge, MA (1997).

[14] Donoho, D. and Johnstone, I., "Ideal Spatial Adaption by Wavelet Shrinkage," *Biometrika* **81**, 425–455 (1994).

[15] Mallat, S., "Theory for Multiresolution Signal Decomposition: The Wavelet Representaion," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**(7), 674–693 (1989).

[16] Vidakovic, B., [*Statistical Modeling by Wavelets*], John Wiley & Sons, Canvers, MA (1999).

[17] Mardia, K., Kent, J., and Bibby, J., [*Multivariate Analysis*], Academic Press, New York (1979).

[18] Fukumizu, K., Bach, F., and Jordan, M., "Dimensionality Reduction for Supervised Learning with Reproducing Kernel Hilbert Spaces," *Journal Of Machine Learning Research* **5**, 73–99 (2004).

[19] Fodor, I., "A Survey of Dimension Reduction Techniques," (2002).

[20] Vempala, S., [*The Random Projection Method*], The American Mathematical Society, Providence, RI (2004).

[21] Vershynin, R., "Introduction to the Non-Asymptotic Analysis of Random Matrices," (2010).

[22] Steward, R., "Statistical methods in change-point analysis," *Dissertation, St. Louis University* (2015).

[23] Ogden, R. and Lynch, J., "Bayesian Analysis of Change-Point Models," *Lecture Notes in Statistics* **141**, 67–82 (1999).

[24] Ghosh, M., "Objective Priors: An Introduction for Frequentists," *Statistical Science* **26**, 187–202 (201).

[25] Bell, A. J. and Sejnowski, T. J., "The independent components of natural scenes are edge filters," *Vision research* **37**(23), 3327–3338 (1997).