12-2015

Bayesian Minimum Description Length Techniques for Multiple Changepoint Detection

Hewa Anuradha Priyadarshani
Clemson University, hpriyad@clemson.edu

Bayesian Minimum Description Length Techniques for Multiple Changepoint Detection

---

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Mathematics

---

by
Hewa Arachchige Anuradha Priyadarshani
December 2015

Accepted by:
Dr. Robert Lund, Committee Chair
Dr. Yingbo Li
Dr. Collin Gallagher
Dr. Christoper Mcmahan

Abstract

This dissertation develops a minimum description length (MDL) multiple changepoint detection procedure that allows for prior distributions. MDL methods, which are penalized likelihood techniques with penalties based on data description-length information principles, have been successfully applied to many recent multiple changepoint problems. This work shows how to modify the MDL penalty to account for various prior knowledge.

Our motivation lies in climatology. Here, a metadata record, which is a file listing times when a recording station physically moved, instrumentation was changed, etc., sometimes exists. While metadata records are notoriously incomplete, they permit the construct a prior distribution that helps detect changepoints. This allows both documented and undocumented changepoints to be analyzed in tandem. The method developed here takes into account 1) metadata, 2) reference series, 3) seasonal means, and 4) autocorrelations. Asymptotically, our estimated multiple changepoint configuration of monthly data is shown to be consistent. The methods are illustrated in the analysis of 114 years of monthly temperatures from Tuscaloosa, Alabama. The multivariate aspect of the methods allow maximum and minimum temperatures to be jointly studied.

A method for homogenizing daily temperature series is also developed. While daily temperatures have a complex structure, statistical techniques have been accu-

mulating that can now accommodate all of the salient characteristics of daily temperatures. The goal here is to combine these techniques in a reasonable manner for multiple changepoint identification in daily series; computational speed is key as a century of daily data has over 36,000 data points. Autocorrelation aspects are important since correlation can destroy changepoint techniques and sample correlations of day-to-day temperature anomalies are often as large as 0.7. While homogenized daily temperatures may not be as useful as homogenized monthly or yearly temperatures, homogenization done on a daily scale affords one greater statistical precision. It is relatively easy to visually discern two changepoints (breakpoints) two years apart with daily data, but virtually impossible to see them in annual series. The methods are applied to 46 years of daily data at South Haven, Michigan.

Dedication

This work is dedicated to my loving parents and my husband. Without their love and support, I would not be here today.

Acknowledgments

I'm indebted to my research advisor, Dr. Robert Lund for his excellent guidance and support. His advices, insightful thoughts, statistical knowledge and experience was a great support to achieve my goals. Also, I'm grateful to my co-advisor Dr. Yingbo Li, who always come up with fresh ideas to improve our research.

I would like to thank Dr. Colin Gallagher and Dr. Chris McMahan for dedicating their time to review my dissertation. I greatly appreciate their valuable comments. I'm thankful to Department of Mathematical Sciences and Dr. Robert Lund for providing the financial support during my graduate studies.

I'm really grateful to my loving husband, Buddhi. Without his great support and encouragement I'm most certainly would not be here.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 General Introduction

Climate time series often display artificial discontinuities induced by station relocations, gauge changes, observer changes, etc. Such times may induce statistical discontinuities in the record and are called changepoints (breakpoints). While changepoints can alter series variabilities, marginal distributions, or autocovariances, our focus here is on mean shifts. Mitchell (1953) estimates that US temperature series experience about six significant changes per century on average. Some, but not necessarily all, of these change times may induce mean shifts in the series. While some gauge change times, station relocation dates, and other events are written down (documented) in station history logs called metadata, these records are often incomplete — many changepoints are undocumented.

Undocumented changepoint identification is crucial in climate analysis (Potter (1981); Vincent (1998); Caussinus and Mestre (2004); Menne and Williams Jr. (2005, 2009)). The changepoint locations and mean shift sizes are essential aspects for making accurate inferences; in fact, Lund et al. (2001) show that changepoint

information is the most important issue to consider in realistically estimating a temperature trend at a fixed US station. Also, Lu and Lund (2007) and the references therein show that trends estimated from temperature series can be misleading when changepoint features are ignored. Once the changepoint times are identified, most statistical inference procedures are relatively straightforward.

Common methods to identify changepoints are segmentation and at most one changepoint (AMOC) methods. Workhorse AMOC procedures include the standard normal homogeneity (SNH) test, the nonparametric SNH test, and the two phase regression of Lund and Reeves (2002) and are reviewed in Reeves et al. (2007). Those procedures rely on the assumptions that the underlying regression form of the series is known and that the error terms are independent and identically distributed normal random variables. These assumptions are unrealistic for monthly or daily temperature series.

Any AMOC procedure can be turned into a multiple changepoint estimator via segmentation. In segmentation techniques, the time series is first classified as changepoint free or having a single changepoint. If one changepoint is declared, then the series is partitioned into two series about the changepoint time. Next, AMOC methods are applied to the two shorter series to examine for further changepoints. This procedure is repeated until all segments are changepoint free. The performance of segmentation techniques are often questionable. Li and Lund (2012) show that segmentation techniques fails to detect two changepoints that are located closely in time.

In the multiple changepoint literature, penalized likelihood techniques are ubiquitous. Caussinus and Mestre (2004) use a BIC-based penalized log-likelihood model to estimate the number of changepoints, their locations, and any outliers. Davis et al. (2006) propose an automatic procedure to segment non-stationary time series

into blocks of different autoregressive (AR) processes. The number of changepoints, their locations, and the orders of the AR models are found by optimizing a minimum description length (MDL) via a genetic algorithm. Lu et al. (2010) and Li and Lund (2012) develop MDL techniques to locate mean shifts in annual and monthly temperature series following Davis et al. (2006). Unlike AIC and BIC penalties, MDL penalties are not a multiple of the number of model parameters, but consider features of the changepoint configuration such as the number of changepoints and how far apart they are. Bayesian procedures can be viewed as penalized likelihoods: in the posterior distribution, the prior density acts as the penalty. Bayesian multiple changepoint authors include Carlin et al. (1992); Barry and Hartigan (1993); Chib (1998); Ray and Tsay (2002); Fearnhead (2006); Giordani and Kohn (2008); Hannart and Naveau (2009); Beaulieu et al. (2010); Lai and Xing (2011); Eckley et al. (2011); Fearnhead and Liu (2011); Hannart and Naveau (2012); Ko et al. (2015); Du et al. (2015).

This research seeks to identify all changepoint times in monthly and daily temperature records while accounting for the following four statistical features: correlation, a seasonal cycle, a reference series, and metadata. The only paper that consider all four aforementioned features in tandem are Li and Lund (2015). It is imperative to consider these features in the analysis of daily/monthly temperature series. Autocorrelation aspects are crucial in analyzing monthly and daily temperature series. For daily data, sample correlations of day-to-day temperatures are often as large as 0.7. Periodic mean cycles are clearly visible in both monthly and daily series. Since seasonal mean cycles add additional parameters to estimate, changepoint detection is harder in periodic series.

Li and Lund (2015) examine multiple changepoint detection using metadata, applying Bayesian statistical methods to detect changepoints in annual precipitation

data. Prior distributions for the number of changepoints and their locations are constructed from the metadata records. The prior distribution specifies that all times listed in the metadata are equally likely to be changepoints and that all times not listed in the metadata are equally likely to be changepoints. Also, metadata times are more likely to induce mean shifts than non-metadata times. This criteria allows to analyze documented and undocumented changepoints in tandem. We will expand on these techniques to handle metadata and correlation aspects.

Relative homogenization; that is, analyzing a target series minus reference series, is a common technique in climate homogenization (Menne and Williams Jr. (2005, 2009)). A reference series is a record from a station near to the target series, which is expected to be highly correlated with the target series. The idea is that both series should experience similar weather. Hence, subtracting the reference series from the target series helps reduce seasonal means and trends and illuminates artificial discontinuities. Changepoints are easier to see in target minus reference comparisons. We will be analyzing target minus reference series of monthly and daily temperatures.

In addition to univariate multiple changepoint detection, this work also considers joint changepoint detection in maximum (Tmax) and minimum (Tmin) temperatures. Daily temperatures are often defined as the average of Tmax and Tmin values. This enables use of spring-loaded thermometers. Such gages push high and low needles on the thermometer to daily extremes, hence reducing observer burdens to a daily task (monthly temperatures are the average of all daily temperatures within the month). Changepoint aspects in bivariate Tmax and Tmin series are considerably complex. Specifically, a station relocation might move the temperature gauge to a more sheltered location, where nighttime lows do not change but daytime highs decrease. A station relocation to a drier location can simultaneously increase daytime highs and reduce nighttime lows.

In this research, a bivariate autoregressive time series model for Tmax and Tmin is used to account for month-to-month autocorrelation. Changepoints are allowed to occur in either the Tmax or Tmin series by themselves, or in both series at the same time (these are called concurrent shifts). For concurrent shifts, the two means need not shift in the same direction. As concurrent changes are thought to occur more often than non-concurrent changes, our prior distributions are constructed to reflect this belief.

1.2 Daily Series

The aforementioned literature narrates changepoint methods for monthly and annual series. This dissertation presents a method to analyze daily temperature series. The changepoint literature for daily temperature data is less developed. Homogenized daily data is useful in trend, extremes and climate variability studies. Since a daily series contains many more points than a monthly or annual series, analysis using daily data should have a greater precision. On the other hand, analysis of daily data is more challenging due to the long series lengths and the number of parameters in the model. In fact, a simple model for daily temperature series contains more than 1095 (365×3) parameters.

Vincent and Zhang (2002) present a method to homogenize daily maximum and minimum temperatures over Canada. Their method homogenizes daily data based on the changepoints and subsequent adjustments found in monthly data. Daily temperature adjustments incorporate a linear interpolation, which preserve the long-term trend and variations present in monthly series. Della-Marta and Wanner (2006) propose a method to homogenize daily data, which is capable of adjusting the mean and higher order moments. Their method incorporates a non-linear model to estimate

the relationship between a target and reference series. Kuglitsch et al. (2009) present a quality control based homogenization method based on a penalized log-likelihood procedure and a non linear model. The break detection and correction methods there depend on the existence of a highly correlated reference series. The breakpoints are identified by applying Caussinus and Mestre (2004) methods to annually differenced series. The homogenization methods of Vincent and Zhang (2002), Della-Marta and Wanner (2006), and Kuglitsch et al. (2009) are based on the changepoints identified in corresponding annual or monthly series.

Climate series homogenization consists of two processes: 1) detect artificial inhomogeneities, and 2) adjust the data for inhomogeneities. Our work here focus on the first process: detection of artificial inhomogeneities. Our aim is to find the best changepoint model (number of changepoints and their locations) for a given series. This can be viewed as a model selection problem. Here, our model selection criteria is based on a modified MDL, which is refereed to as a Bayesian MDL.

1.3 Bayesian MDL

Our research develops a novel changepoint MDL method to detect multiple changepoints. MDL methods were introduced by Rissanen (1989) and are based on Kolmogorov's complexity theory and Shannon's work on coding (Hansen and Yu, 2001). Among a class of plausible models, the MDL principle seeks the model with the shortest so-called description length. The description length is the number of digits in a binary string used to encode the model (and data) for transmission. Better models should have shorter description lengths. For more background, the interested reader is referred to Hansen and Yu (2001) and Grünwald et al. (2005).

Our modified MDL is called a Bayesian MDL because (1) it accommodates

subjective knowledge from domain experts via prior specification and straightforward hyper-parameter elicitation, (2) it is essentially an empirical Bayes approach, which enjoys asymptotic model selection consistency and exhibits good performance in finite samples, and (3) it permits the use of stochastic model search algorithms from the Bayesian model selection literature to achieve efficient computation. Derivation of our Bayesian MDL is presented in Chapter 2.

The optimal changepoint model has the minimal Bayesian MDL. A naive approach to this minimization is to compute the Bayesian MDL for each possible model. Since this is not viable, an efficient optimization technique is required. Here, Bayesian model search algorithms (García-Donato and Martínez-Beneito, 2013) such as Markov chain Monte-Carlo approach or genetic algorithms (Goldberg and Holland, 1988) can be implemented to optimize Bayesian MDL. More details on these algorithms are in chapter 2 and 3.

This dissertation is organized as follows. Chapter 2 introduces Bayesian MDL techniques for univariate and bivariate monthly temperature series. Chapter 3 modifies the Bayesian MDL to accommodate daily temperature series. A simulation study and real temperature series analyses are included in both chapters.

Chapter 2

Bayesian MDL for Monthly Series

2.1 Data and Model

2.1.1 The Tuscaloosa data

Figure 2.1 plots a monthly Tmax and Tmin series from Tuscaloosa, Alabama (the target station) over the 114 year period January, 1901 — December, 2014. Lu et al. (2010) study average values of the series from 1901-2000. In Section 2.6, the Tmax and Tmin series will be analyzed from univariate and bivariate perspectives. The Tuscaloosa metadata list station relocations in November 1921, March 1939, June 1956, and May 1987; November 1956 and May 1987 are listed as instrumentation change times. While Lu et al. (2010) use the metadata to justify changepoint conclusions in hindsight, the metadata will be used in our detection procedure here, which substantially boosts detection power.

Our estimated changepoint configuration (justified in Section 2.6) is revealed in Figure 2.1. Estimated changepoint times are marked with vertical dashed lines. Mean shifts are difficult to see in series with large seasonal cycles, which are on the

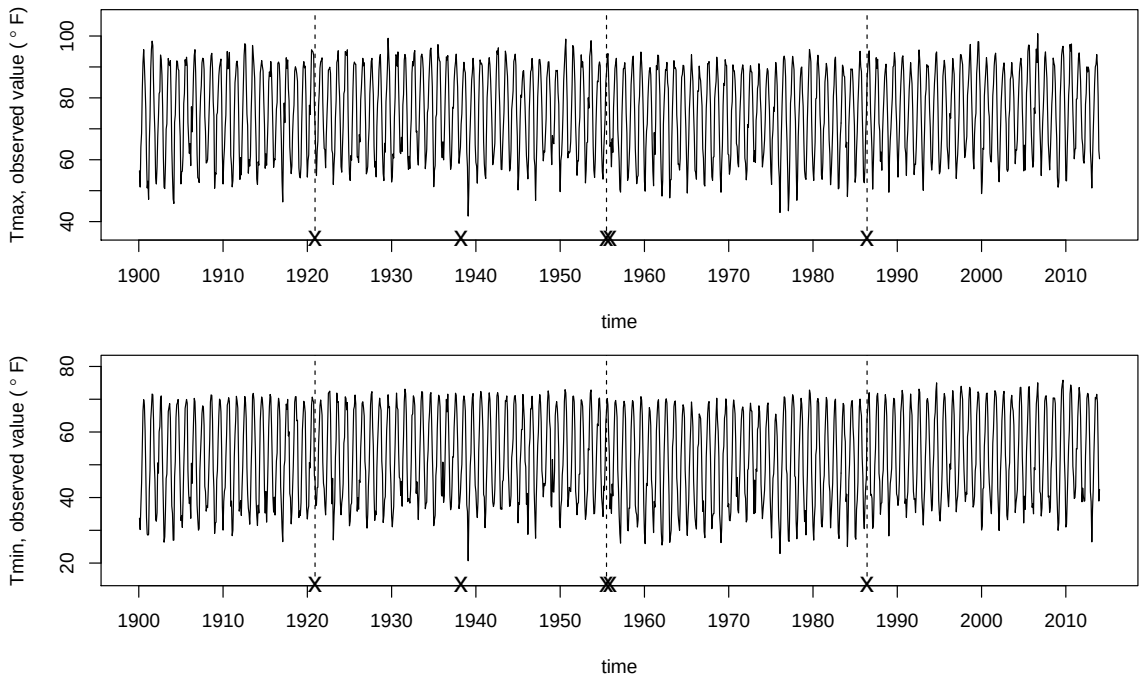

Figure 2.1: Tuscaloosa monthly Tmax (top panel) and Tmin (bottom panel) series. Metadata times are marked with crosses on the axis. Vertical dashed lines show estimated changepoint times from our methods.

order of 40 degrees Fahrenheit here. Each metadata time is marked by an X on the axis. Observe that all three of the identified changepoints occur at metadata times, and that all of them occur in both Tmax and Tmin series.

Following Lu et al. (2010), our reference is obtained by averaging three nearby stations: Aberdeen, MS; Greensboro, AL; and Selma, AL. By averaging multiple reference series (this is called a composite reference), impacts of mean shifts in any of the individual stations in the composite reference are minimized. Figure 2.2 plots the monthly target minus reference series and its estimated changepoint configuration. Now, 12 changepoint times, all of which are concurrent (occur at the same time in both Tmax and Tmin), emerge. In particular, the November 1956 changepoint shifts the Tmax series upwards and the Tmin series downwards. This configuration is examined further in Section 7.

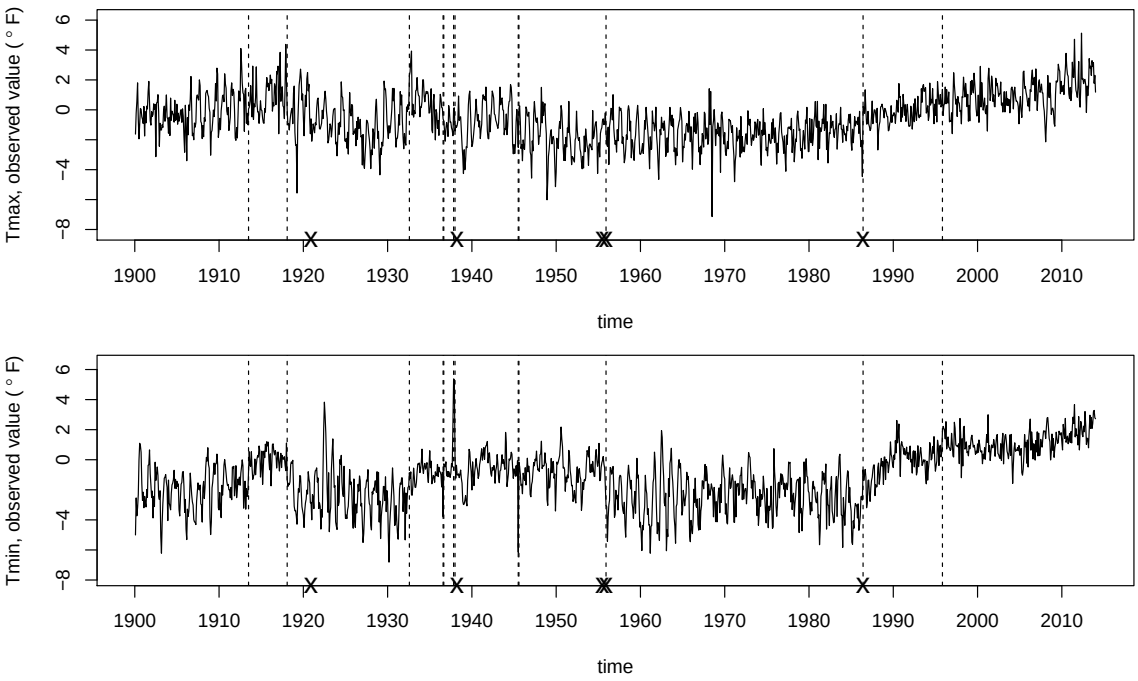

Figure 2.2: Target minus reference Tmax (top panel) and Tmin (bottom panel) series. Metadata times for Tuscaloosa are marked with crosses on the axis. Vertical dashed lines show estimated changepoint times from our methods.

2.1.2 A univariate model

Consider a univariate time series with data $\mathbf{X}_{1:N} = (X_1, X_2, \ldots, X_N)'$, where $t \in \{1, 2, \ldots, N\}$ denotes time. As our data is monthly, periodicities will be represented by writing time t as $t = (u-1)T+v$, where u denotes the year of the observation and $v \in \{1, \ldots, 12\}$ is the month of the observation. Here, the fundamental period is $T = 12$.

Suppose m changepoints partition the timeline $t \in \{1, 2, \ldots, N\}$ into $m + 1$ distinct regimes (segments). During the rth regime, $r \in \{1, 2, \ldots, m + 1\}$, the series has mean μ_r (neglecting the seasonal cycle). A model with autocorrelated errors

describing this situation is

$$X_t = s_{v(t)} + \mu_{r(t)} + \epsilon_t, \tag{2.1}$$

$$\epsilon_t = \sum_{j=1}^{p} \phi_j \epsilon_{t-j} + Z_t, \quad Z_t \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2). \tag{2.2}$$

Here, $v(t) = t - T\lfloor (t-1)/T \rfloor$ is the season corresponding to time t, where $\lfloor x \rfloor$ is the largest integer less than or equal to x. The monthly means are $s_1, s_2, \ldots, s_T$, the regime means are $\mu_1, \mu_2, \ldots, \mu_{m+1}$, and the regression errors $\{\epsilon_t\}_{t=1}^N$ are stationary but autocorrelated. In particular, $\{\epsilon_t\}_{t=1}^N$ is assumed to be a causal zero mean autoregression of order p driven by the independent and identically distributed noise $\{Z_t\}_{t=1}^N$. The parameters $\phi_1, \phi_2, \ldots, \phi_p$ are autoregressive coefficients and $\text{Var}(Z_t) = \sigma^2$. In climate applications, monthly averaged temperatures and the logarithm of annual precipitation are approximately normally distributed (Wilks, 2011). Hence, in further likelihood computations, Gaussianity is assumed.

Suppose that the m changepoints are located at the times $\tau_1 < \tau_2 < \cdots < \tau_m$. To avoid trite work with edge effects of the autoregression, we assume that no changepoints occur during the first p observations. For notation, let $\tau_0 = 1$ and $\tau_{m+1} = N + 1$. Then the regime indicator $r(t)$ in (2.1) has $r(t) = r$ when $\tau_{r-1} \leq t < \tau_r$. To ensure parameter identifiability, μ_1 is taken as zero; hence, $E[X_t] = s_{v(t)}$ when t lies in the first regime. The model in (2.1) and (2.2) contains the following unknown parameters: the number of changepoints m, the changepoint location times $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_m)'$, the seasonal means $\mathbf{s} = (s_1, s_2, \ldots, s_T)'$, the regime means $\boldsymbol{\mu} = (\mu_2, \mu_3, \ldots, \mu_{m+1})'$, the AR parameters $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_p)'$, and the white noise variance σ^2.

Following Li and Lund (2015), we denote the multiple changepoint configuration $(m; \boldsymbol{\tau})$ as an $(N-p)$-dimensional vector of zero/one indicators: $\boldsymbol{\eta} = (\eta_{p+1}, \eta_{p+2}, \ldots, \eta_N)'$,

where $\eta_t \in \{0, 1\}$ for $t \in \{p+1, \ldots, N\}$. In this notation, $\eta_t = 1$ means that time t is a changepoint; $\eta_t = 0$ means that time t is not a changepoint.

The data likelihood given a changepoint configuration $\boldsymbol{\eta}$ is now developed. Suppose that the changepoint configuration $\boldsymbol{\eta}$ contains $m = \sum_{t=p+1}^{N} \eta_t$ changepoints. Equation (2.1) has the regression representation

$$\mathbf{X}_{1:N} = \mathbf{A}_{1:N}\mathbf{s} + \mathbf{D}_{1:N}\boldsymbol{\mu} + \boldsymbol{\epsilon}_{1:N}, \tag{2.3}$$

where $\mathbf{A}_{1:N} \in \mathbb{R}^{N \times T}$ and $\mathbf{D}_{1:N} \in \mathbb{R}^{N \times m}$ are seasonal and regime indicators:

$$[\mathbf{A}_{1:N}]_{t,v} = \mathbf{1}(\text{time } t \text{ is in season } v), \quad v \in \{1, 2, \ldots, T\},$$
$$[\mathbf{D}_{1:N}]_{t,r-1} = \mathbf{1}(\text{time } t \text{ is in regime } r), \quad r \in \{2, 3, \ldots, m+1\},$$

and $\mathbf{1}(A)$ denotes the indicator of the event A. In (2.3), the subscript $1 : N$, or in general $t_1 : t_2$, signifies that only rows t_1 through t_2 are used in the quantities. The first $\tau_1 - 1$ rows of $\mathbf{D}_{1:N}$ are taken as zero for parameter identifiability. The white noise process $\{Z_t\}$ driving $\boldsymbol{\epsilon}_{1:N}$ is assumed independent and Gaussian with variance σ^2. This yields the distributional result

$$\boldsymbol{\epsilon}_{(p+1):N} - \sum_{j=1}^{p} \phi_j \boldsymbol{\epsilon}_{(p+1-j):(N-j)} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{N-p}), \tag{2.4}$$

where $\mathbf{I}_k$ denotes the $k \times k$ identity matrix. Now define

$$\widetilde{\mathbf{X}} = \mathbf{X}_{(p+1):N} - \sum_{j=1}^{p} \phi_j \mathbf{X}_{(p+1-j):(N-j)}, \tag{2.5}$$

$$\widetilde{\mathbf{A}} = \mathbf{A}_{(p+1):N} - \sum_{j=1}^{p} \phi_j \mathbf{A}_{(p+1-j):(N-j)}, \tag{2.6}$$

$$\widetilde{\mathbf{D}} = \mathbf{D}_{(p+1):N} - \sum_{j=1}^{p} \phi_j \mathbf{D}_{(p+1-j):(N-j)}, \tag{2.7}$$

and observe that (2.4) becomes

$$\widetilde{\mathbf{X}} - \widetilde{\mathbf{A}}\mathbf{s} - \widetilde{\mathbf{D}}\boldsymbol{\mu} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{N-p}). \tag{2.8}$$

Note that all terms superscripted with $\sim$ depend on the unknown AR parameter $\boldsymbol{\phi}$. To avoid trite work with autoregressive edge effects, a likelihood conditional on X_t for $t \in \{1, 2, \ldots, p\}$ is used. In the change of variable computations, the Jacobian $|\partial(\widetilde{\mathbf{X}} - \widetilde{\mathbf{A}}\mathbf{s} - \widetilde{\mathbf{D}}\boldsymbol{\mu})/\partial \mathbf{X}_{(p+1):N}| = 1$ and hence the likelihood has the multivariate normal form

$$f\left(\mathbf{X}_{(p+1):N} \mid \boldsymbol{\mu}, \mathbf{s}, \sigma^2, \boldsymbol{\phi}, \boldsymbol{\eta}, \mathbf{X}_{1:p}\right) = \left(2\pi\sigma^2\right)^{-\frac{N-p}{2}} e^{-\frac{1}{2\sigma^2}(\widetilde{\mathbf{X}} - \widetilde{\mathbf{A}}\mathbf{s} - \widetilde{\mathbf{D}}\boldsymbol{\mu})'(\widetilde{\mathbf{X}} - \widetilde{\mathbf{A}}\mathbf{s} - \widetilde{\mathbf{D}}\boldsymbol{\mu})}.$$

Innovations forms of the likelihood (Brockwell and Davis (1991)) could be used if one wants a moving-average or long-memory component in $\{\epsilon_t\}$.

2.1.3 A bivariate model for Tmax and Tmin series

To model the Tmax and Tmin series jointly, concatenate them via $\mathbf{X}_{1:N} = (\mathbf{X}'_{1:N,1}, \mathbf{X}'_{1:N,2})' \in \mathbb{R}^{2N}$, where $\mathbf{X}_{1:N,i} = (X_{1,i}, \ldots, X_{N,i})'$ is the observed record for Tmax ($i = 1$) or Tmin ($i = 2$). Each time in $\{p+1, p+2, \ldots, N\}$ is allowed to

be a changepoint in the Tmax or Tmin series, or both. A multiple changepoint configuration is denoted by $\boldsymbol{\eta} = (\boldsymbol{\eta}_1', \boldsymbol{\eta}_2')'$, where $\boldsymbol{\eta}_i = (\eta_{p+1,i}, \ldots, \eta_{N,i})' \in \{0,1\}^{N-p}$ is defined as in the univariate case.

Given a bivariate changepoint configuration $\boldsymbol{\eta}$, series i has $m_i = \sum_{t=p+1}^{N} \eta_{t,i}$ changepoints. As in the univariate case, the seasonal means are denoted by $\mathbf{s}_i = (s_{1,i}, s_{2,i}, \ldots, s_{T,i})' \in \mathbb{R}^T$; regime means are denoted by $\boldsymbol{\mu}_i = (\mu_{2,i}, \mu_{3,i}, \ldots, \mu_{m_i+1,i})' \in \mathbb{R}^{m_i}$. The seasonal indicator matrix $\mathbf{A}_{1:N,i} \in \mathbb{R}^{N \times T}$ and the regime indicator matrix $\mathbf{D}_{1:N,i} \in \mathbb{R}^{N \times m_i}$ are constructed analogously to their univariate counterparts.

The regression representation (2.3) holds for the bivariate case, with $\mathbf{s} = (\mathbf{s}_1', \mathbf{s}_2')'$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1', \boldsymbol{\mu}_2')'$, $\boldsymbol{\epsilon}_{1:N} = (\epsilon_{1:N,1}', \epsilon_{1:N,2}')'$ denoting the concatenated seasonal means, regime means, and regression errors, respectively. The seasonal and regime indicator matrices have the block diagonal forms

$$\mathbf{A}_{1:N} = \begin{bmatrix} \mathbf{A}_{1:N,1} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{1:N,2} \end{bmatrix}, \quad \mathbf{D}_{1:N} = \begin{bmatrix} \mathbf{D}_{1:N,1} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_{1:N,2} \end{bmatrix}. \tag{2.9}$$

Note that the seasonal indicators for Tmax and Tmin coincide, i.e., $\mathbf{A}_{1:N,1} = \mathbf{A}_{1:N,2}$, while $\mathbf{D}_{1:N,1}$ and $\mathbf{D}_{1:N,2}$ differ unless all changepoints are concurrent.

As the Tmax and Tmin data tend to fluctuate about the mean in tandem (positive correlation), the errors $\{\boldsymbol{\epsilon}_t\}$ need to be correlated. For this, a Gaussian vector autoregression (VAR) model of order p is employed:

$$\boldsymbol{\epsilon}_t = \sum_{j=1}^{p} \boldsymbol{\Phi}_j \boldsymbol{\epsilon}_{t-j} + \mathbf{Z}_t, \quad \mathbf{Z}_t \overset{\text{iid}}{\sim} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{2.10}$$

where $t \in \{p+1, p+2, \cdots, N\}$. Here, $\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_p$ are 2×2 VAR coefficient matrices. The VAR model allows for correlation both in time and between components.

To simplify calculations, a conditional likelihood given $\mathbf{X}_{1:p}$ is used to avoid

edge effects of the autoregression. As (2.8) holds after replacing $\sigma^2\mathbf{I}_{N-p}$ with $\mathbf{\Sigma}\otimes\mathbf{I}_{N-p}$, the likelihood of $\mathbf{X}_{(p+1):N}$, conditional on the initial observations $\mathbf{X}_{1:p}$ and model parameters, is (to a multiplicative constant)

$$\begin{aligned} & f(\mathbf{X}_{(p+1):N} \mid \mathbf{s}, \boldsymbol{\mu}, \mathbf{\Sigma}, \mathbf{\Phi}_1, \cdots, \mathbf{\Phi}_p, \boldsymbol{\eta}, \mathbf{X}_{1:p}) \\ \propto & |\mathbf{\Sigma}|^{-\frac{N-p}{2}} \exp\left[-\frac{1}{2}(\widetilde{\mathbf{X}} - \widetilde{\mathbf{A}}\mathbf{s} - \widetilde{\mathbf{D}}\boldsymbol{\mu})'(\mathbf{\Sigma}^{-1}\otimes\mathbf{I}_{N-p})(\widetilde{\mathbf{X}} - \widetilde{\mathbf{A}}\mathbf{s} - \widetilde{\mathbf{D}}\boldsymbol{\mu})\right]. \end{aligned} \tag{2.11}$$

Here, $\otimes$ denotes a Kronecker product and the terms $\widetilde{\mathbf{X}}, \widetilde{\mathbf{A}}, \widetilde{\mathbf{D}}$ are modified by replacing ϕ_j with $\mathbf{\Phi}_j \otimes \mathbf{I}_{N-p}$ in (2.5), (2.6), and (2.7) for $j \in \{1, 2, \ldots, p\}$. In the rest of the paper, the edge variables $\mathbf{X}_{1:p}$ are omitted in notation.

2.2 A Brief MDL Review

Multiple changepoint problems can be viewed from a model selection perspective, where each changepoint configuration $\boldsymbol{\eta}$ being a candidate model. Among all 2^{N-p} (univariate) or $2^{2(N-p)}$ (bivariate) changepoint configurations, our objective is to identify the configuration that optimizes a certain objective function. The objective function used here is a Bayesian MDL. According to MDL principle, competing probability models can be compared by their description lengths; the true data generating distribution (i.e., the true model) should have the shortest expected description length.

For a discrete random variable X taking values in $\mathcal{X}$ with probability mass function $f(x) = P(X = x)$, Shannon's Source Coding Theorem (Shannon, 1948) states that the encoding with code length

$$\mathcal{L}(X) = -\log_2[f(X)] \tag{2.12}$$

has the shortest expected description length. For example, if X is uniformly distributed over $\mathcal{X} = \{1, 2, \ldots, n\}$, then its MDL is $\mathcal{L}(X) = -\log_2(1/n) = \log_2(n)$. If $\mathbf{X}$ is a continuous variable in a k-dimensional space with density function $f(\cdot)$, then after discretizing each dimension into equal cells of size δ (often viewed as the machine precision), mimicking (2.12) gives the MDL

$$\mathcal{L}(\mathbf{X}) = -\log_2[f(\mathbf{X})\delta^k] = -\log_2[f(\mathbf{X})] - k\log_2(\delta). \tag{2.13}$$

Because k and δ do not vary with $\mathbf{X}$, the term $-k\log_2(\delta)$ does not affect comparison between different $\mathbf{X}$ and is often omitted. One can substitute the natural-based logarithm for the base two logarithm — this does not affect model comparisons since $\log_2(x)/\log(x)$ is constant.

Now suppose that a dataset $\mathbf{X} = (X_1, X_2, \ldots, X_N)'$, believed to be generated from a certain parametric model $\mathcal{M}$ with density $f(\mathbf{X} \mid \theta, \mathcal{M})$, is to be transmitted along with a possibly unknown parameter $\theta \in \Theta$. To transmit the data, two types of MDL approaches, the two-part MDL and the mixture MDL, are discussed in Hansen and Yu (2001).

2.2.1 Two-part MDLs

The two-part MDL (also called a two-stage MDL) considers the transmission of $\mathbf{X}$ and θ in two steps. If both the sender and receiver know θ, the MDL of $\mathbf{X}$ is

$$\mathcal{L}(\mathbf{X} \mid \theta, \mathcal{M}) = -\log[f(\mathbf{X} \mid \theta, \mathcal{M})].$$

Here, notations such as $\mathcal{L}(\cdot \mid \cdot)$ are adopted and are analogous to conditional distribution notations; this notation emphasizes dependence on $\mathcal{M}$ and θ. Should θ also

be unknown to the receiver, an additional cost of $\mathcal{L}(\theta \mid \mathcal{M})$ is incurred transmitting it. Hence, the two-part MDL becomes

$$\mathcal{L}(\mathbf{X}, \theta \mid \mathcal{M}) = \mathcal{L}(\mathbf{X} \mid \theta, \mathcal{M}) + \mathcal{L}(\theta \mid \mathcal{M}). \tag{2.14}$$

Suppose that the MDL in (2.14) is minimized at $\hat{\theta}$, an estimator of θ based on $\mathbf{X}$. If θ is a k-dimensional continuous parameter and $\hat{\theta}$ is a $\sqrt{n}$-consistent estimator of θ, then one can view the machine precision as $\delta = c/\sqrt{n}$, where c is a positive constant. Under a uniform encoder $\pi(\theta \mid \mathcal{M}) \propto 1$, the MDL in (2.13) needed to transmit θ (including $\hat{\theta}$) is hence

$$\mathcal{L}(\theta \mid \mathcal{M}) = -\log[\pi(\theta \mid \mathcal{M})] - k \log\left(\frac{c}{\sqrt{n}}\right) = \frac{k}{2}\log(n) - k\log(c), \tag{2.15}$$

which does not depend on θ. Hence, the maximum likelihood estimator (MLE) minimizes (2.14) and the two-part MDL coincides with the classic Bayesian Information Criteria (BIC) (Schwarz (1978)). Note that $\hat{\theta}$ need not be the MLE. In fact, any $\sqrt{n}$-consistent estimator $\hat{\theta}$ is justifiable.

Suppose there is a discrete set of candidate models. To account for model uncertainty, the two-part MDL can be modified to include an additional description length for the model, i.e.,

$$\mathcal{L}(\mathbf{X}, \hat{\theta}, \mathcal{M}) = \mathcal{L}(\mathbf{X}, \hat{\theta} \mid \mathcal{M}) + \mathcal{L}(\mathcal{M}), \tag{2.16}$$

where $\mathcal{L}(\mathcal{M}) = -\log[\pi(\mathcal{M})]$, and $\pi(\mathcal{M})$ denotes the prior distribution over the model space. The model with the smallest MDL in (2.16) is selected as the optimal model. Here, $\hat{\theta}$ is model dependent.

Two-part MDLs have been used in time series changepoint problems (Davis

et al., 2006, 2008; Lu et al., 2010; Li and Lund, 2012). However, for a finite sample size n, a drawback exists when the dimension of θ changes with the model, as is the multiple changepoint case. Consider a setting of two competing models $\mathcal{M}_1$ and $\mathcal{M}_2$, whose parameters θ_j are k_j-dimensional continuous parameters, respectively, for $j \in \{1, 2\}$. Model $\mathcal{M}_1$ is favored if $\mathcal{L}(\mathbf{X}, \hat{\theta}_1, \mathcal{M}_1) - \mathcal{L}(\mathbf{X}, \hat{\theta}_2, \mathcal{M}_2)$ is negative; else, model $\mathcal{M}_2$ is favored. From (2.14) and (2.16), $\mathcal{L}(\mathbf{X}, \hat{\theta}_1, \mathcal{M}_1) - \mathcal{L}(\mathbf{X}, \hat{\theta}_2, \mathcal{M}_2)$ contains the term

$$\begin{aligned} \mathcal{L}(\hat{\theta}_1 \mid \mathcal{M}_1) - \mathcal{L}(\hat{\theta}_2 \mid \mathcal{M}_2) = \; & \log[\pi(\hat{\theta}_2 \mid \mathcal{M}_2)] - \log[\pi(\hat{\theta}_1 \mid \mathcal{M}_1)] \\ & + \frac{k_1 - k_2}{2}[\log(n) - 2\log(c)]. \end{aligned} \tag{2.17}$$

When $k_1 \neq k_2$, this term depends on c, which is problematic as $\mathcal{L}(\mathbf{X}, \hat{\theta}_1, \mathcal{M}_1) - \mathcal{L}(\mathbf{X}, \hat{\theta}_2, \mathcal{M}_2)$ could be either positive or negative depending on the values of n and c. In this case, one cannot judge one model superior without knowledge of c. Thus, when the dimension of θ changes with $\mathcal{M}$, the two-part MDL in (2.16) has issues. This issue does not conflict with the asymptotic consistency of BIC or modified BIC (Zhang and Siegmund, 2007): as n increases, $\log(n)$ dominates the fixed constant $\log(c)$ in (2.17). We now consider mixture MDLs, which do not suffer from such problems.

2.2.2 Mixture MDLs

Suppose that θ for the model $\mathcal{M}$ is believed to have a prior distribution with density $\pi(\theta \mid \mathcal{M})$. The marginal likelihood of $\mathbf{X}$ averages the likelihood $f(\mathbf{X} \mid \theta, \mathcal{M})$ under the prior distribution of θ:

$$f(\mathbf{X} \mid \mathcal{M}) = \int_{\Theta} f(\mathbf{X} \mid \theta, \mathcal{M}) \pi(\theta \mid \mathcal{M}) d\theta. \tag{2.18}$$

The mixture MDL is the negative logarithm of the marginal likelihood:

$$\mathcal{L}(\mathbf{X} \mid \mathcal{M}) = -\log f(\mathbf{X} \mid \mathcal{M}) = -\log \int_{\Theta} f(\mathbf{X} \mid \theta, \mathcal{M}) \pi(\theta \mid \mathcal{M}) d\theta. \tag{2.19}$$

If the prior for θ depends on an unknown hyper-parameter τ, then a two-part MDL can be used to account for the additional description length needed to transmit τ. In this case, the overall mixture MDL, for any $\sqrt{n}$-consistent estimator of τ, is

$$\mathcal{L}(\mathbf{X}, \hat{\tau} \mid \mathcal{M}) = -\log \int_{\Theta} f(\mathbf{X} \mid \theta, \mathcal{M}) \pi(\theta \mid \hat{\tau}, \mathcal{M}) d\theta + \mathcal{L}(\hat{\tau} \mid \mathcal{M}). \tag{2.20}$$

The mixture MDL for the model $\mathcal{M}$ based on (2.16) and (2.20) is

$$\mathcal{L}(\mathbf{X}, \hat{\tau}, \mathcal{M}) = \mathcal{L}(\mathbf{X}, \hat{\tau} \mid \mathcal{M}) + \mathcal{L}(\mathcal{M}).$$

This is related to empirical Bayes (EB) approaches. If the prior probabilities of two models are the same, i.e., $\pi(\mathcal{M}_1) = \pi(\mathcal{M}_2)$, and the hyper-parameter τ is transmitted under a uniform encoder $\pi(\tau \mid \mathcal{M}_j) \propto 1$ for $j \in \{1, 2\}$, then the negative logarithm of their Bayes factor (Kass and Raftery, 1995) equals the difference of their mixture MDLs $\mathcal{L}(\mathbf{X}, \hat{\tau}_1, \mathcal{M}_1) - \mathcal{L}(\mathbf{X}, \hat{\tau}_2, \mathcal{M}_2)$. Similarly, in EB settings, although the estimator $\hat{\tau}$ is often chosen to maximize the marginal likelihood $f(\mathbf{X} \mid \tau, \mathcal{M})$, other estimators can be used (Carlin and Louis, 2000).

2.3 Bayesian MDLs (BMDLs)

Our main idea is to apply the mixture MDL to parameters such as $\boldsymbol{\mu}$ whose dimensions vary across models, and use the two-part MDL for other parameters. This section first introduces our prior choices on $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$, derives our BMDL criterion,

and then discusses computational strategies.

2.3.1 The prior distribution of changepoint configurations

Our prior distribution for the changepoint configuration $\pi(\boldsymbol{\eta})$ in the univariate case assumes that, in the absence of metadata, each time t has an equal probability ρ of being a changepoint, independently of other times, i.e.,

$$\eta_t \mid \rho \overset{\text{iid}}{\sim} \text{Bernoulli}(\rho), \quad t \in \{p+1, p+2, \ldots, N\}. \tag{2.21}$$

Chernoff and Zacks (1964), Yao (1984), and Barry and Hartigan (1993) use this prior; it is a reasonable choice in climate time series applications where knowledge beyond metadata is generally unavailable. For other applications, $\pi(\boldsymbol{\eta})$ can have different success probabilities during different regimes (Chib, 1998); correlation across different changepoint times can also be achieved (Li and Zhang, 2010).

As estimated changepoint configurations are sensitive to ρ, a hyper-prior is placed on it. Barry and Hartigan (1993) let ρ have a uniform prior on $(0, \rho_0)$, where $\rho_0 < 1$. For additional flexibility, the Beta distribution $\rho \sim \text{Beta}(a, b)$ will be used here. Due to Beta-Binomial conjugacy, the marginal prior density of $\boldsymbol{\eta}$ has the closed form

$$\pi(\boldsymbol{\eta}) = \int_0^1 \left[\prod_{t=p+1}^{N} \pi(\eta_t \mid \rho) \right] \pi(\rho) d\rho = \frac{\beta(a+m, b+N-p-m)}{\beta(a,b)}, \tag{2.22}$$

where $\pi(\eta_t \mid \rho) = \rho^{\eta_t}(1-\rho)^{1-\eta_t}$ for $\eta_t \in \{0,1\}$ and $\beta(\cdot,\cdot)$ denotes the Beta function. Beta-Binomial priors are common in the Bayesian model selection literature (Scott and Berger, 2010).

The Beta-Binomial prior can be tuned to accommodate subjective knowl-

edge from domain experts. For example, Mitchell (1953) estimates an average of six changes per century for United States temperature series; this long-term rate is 0.005 changepoints per month and can be produced with $a = 1$ and $b = 199$; with these parameters, $E(\rho) = a/(a+b) = 0.005$.

This prior can be modified to accommodate metadata. Suppose that during the times $\{p+1, p+2, \ldots, N\}$, there are $N^{(2)}$ documented times in the metadata and $N^{(1)} = N - p - N^{(2)}$ undocumented times. For notation, all quantities superscripted with (1) refer to undocumented times; quantities superscripted with (2) refer to documented times. Following Li and Lund (2015), we posit that the undocumented times have a Beta-Binomial$(a, b^{(1)})$ prior distribution, and independently, the documented times have a Beta-Binomial$(a, b^{(2)})$ prior. To make the metadata times more likely to induce true mean shift, we impose

$$E[\rho^{(1)}] = \frac{a}{a + b^{(1)}} < \frac{a}{a + b^{(2)}} = E[\rho^{(2)}].$$

Here, the parameter a is common to both documented and undocumented times. For monthly data, default values are $a = 1, b^{(1)} = 239$, and $b^{(2)} = 47$, making $E(\rho^{(1)}) = 0.0042$ and $E(\rho^{(2)}) = 0.0208$; *a priori*, a documented time is roughly five times as likely to be a changepoint as an undocumented time.

Arguing akin to (2.22), a changepoint configuration $\boldsymbol{\eta}$ with $m^{(2)}$ documented changepoints and $m^{(1)}$ undocumented changepoints ($m = m^{(1)} + m^{(2)}$) has marginal prior distribution (up to a normalizing constant)

$$\pi(\boldsymbol{\eta}) \propto \prod_{k=1}^{2} \Gamma\left(a + m^{(k)}\right) \Gamma\left(b^{(k)} + N^{(k)} - m^{(k)}\right) \tag{2.23}$$

after the Beta functions are written in their Gamma representations.

For the bivariate case, a hierarchical prior is put on $\boldsymbol{\eta}$ that encourages both documented changes and concurrent changes. For $t \in \{p+1, p+2, \ldots, N\}$, the indicator $\boldsymbol{\eta}_t = (\eta_{t,1}, \eta_{t,2})'$ takes values in one of the four categories: $(1,1)$, mean shifts in both Tmax and Tmin; $(1,0)$, a mean shift in Tmax but not in Tmin; $(0,1)$, a mean shift in Tmin but not in Tmax; and $(0,0)$, no mean shifts. A Dirichlet-Multinomial prior is put on $\boldsymbol{\eta}_t$:

$$\boldsymbol{\eta}_t \mid \boldsymbol{\rho} \overset{\text{iid}}{\sim} \text{Multinomial}(1; \boldsymbol{\rho}), \quad \boldsymbol{\rho} \sim \text{Dirichlet}(\boldsymbol{\alpha}), \tag{2.24}$$

where $\boldsymbol{\rho} = (\rho_1, \rho_2, \rho_3, \rho_4)'$ are the probabilities of the four categories, such that $0 < \rho_\ell < 1$ for $\ell \in \{1,2,3,4\}$, and $\sum_{\ell=1}^{4} \rho_\ell = 1$; $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)'$ are the Dirichlet parameters; $\alpha_\ell > 0$ for each ℓ.

Suppose that the changepoint configuration $\boldsymbol{\eta}$ has m_ℓ times in category ℓ. Due to the Dirichlet-multinomial conjugacy, the marginal prior of $\boldsymbol{\eta}$ has a closed form after integrating $\boldsymbol{\rho}^{(1)}$ and $\boldsymbol{\rho}^{(2)}$ out:

$$\pi(\boldsymbol{\eta}) \propto \prod_{k=1}^{2} \prod_{\ell=1}^{4} \Gamma\left(\alpha_\ell^{(k)} + m_\ell^{(k)}\right). \tag{2.25}$$

Our choice of the hyper-parameter $\boldsymbol{\alpha}$ reflects our belief that concurrent changepoints are more likely to occur than an independent scenario. By (2.24), the ratios between the prior expectations satisfy $E(\rho_1) : E(\rho_2) : E(\rho_3) : E(\rho_4) = \alpha_1 : \alpha_2 : \alpha_3 : \alpha_4$. If changepoints in the Tmax and Tmin series at time t are independent events, then $\rho_1 = P(\eta_{t,1} = 1, \eta_{t,2} = 1) = P(\eta_{t,1} = 1)P(\eta_{t,2} = 1) = (\rho_1 + \rho_2)(\rho_1 + \rho_3)$. Hence, to encourage concurrent shifts, it is assumed that concurrent changepoints occur more

often than in independent settings. This is done by choosing $\boldsymbol{\alpha}$ such that

$$E[\rho_1] = \frac{\alpha_1}{\sum_{\ell=1}^{4} \alpha_\ell} > \frac{\alpha_1 + \alpha_2}{\sum_{\ell=1}^{4} \alpha_\ell} \frac{\alpha_1 + \alpha_3}{\sum_{\ell=1}^{4} \alpha_\ell} = E[\rho_1 + \rho_2]E[\rho_1 + \rho_3].$$

In addition, between the univariate and bivariate models, we match the prior means of the probabilities of no changepoints, i.e.,

$$\frac{b}{a+b} = \frac{\alpha_4}{\sum_{\ell=1}^{4} \alpha_\ell}.$$

After consulting climatologists, default hyper-parameters are set to $\boldsymbol{\alpha}^{(1)} = (3/7, 2/7, 2/7, 239)'$ and $\boldsymbol{\alpha}^{(2)} = (3/7, 2/7, 2/7, 47)'$ for monthly data.

2.3.2 The prior distribution of regime means

For a changepoint configuration with $m > 0$ changepoints, the regime means $\boldsymbol{\mu}$ are posited to have independent normal prior distributions. For the univariate model, this is

$$\boldsymbol{\mu} \mid \sigma^2, \boldsymbol{\eta} \sim \mathrm{N}(\mathbf{0}, \nu\sigma^2 \mathbf{I}_m); \tag{2.26}$$

for the bivariate model, distributions obey

$$\boldsymbol{\mu} \mid \boldsymbol{\Sigma}, \boldsymbol{\eta} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Omega}), \quad \boldsymbol{\Omega} = \nu \ \mathrm{diag}\left(\underbrace{\sigma_1^2, \ldots, \sigma_1^2}_{m_1}, \underbrace{\sigma_2^2, \ldots, \sigma_2^2}_{m_2} \right), \tag{2.27}$$

where σ_1^2 and σ_2^2 are the diagonal entries of the white noise covariance $\boldsymbol{\Sigma}$. Here, ν is a pre-specified non-negative parameter that is relatively large so that the variances of the regime means are large multiples of the white noise variances. As in the sensitivity analysis in Du et al. (2015), our experience suggests that model selection results are stable under a wide range of ν. Our default takes $\nu = 5$.

In fact, $\pi(\boldsymbol{\mu})$ can be any continuous distribution. For example, if mean shifts can be large, heavy-tailed distributions such as Student-t may be preferable. When $\boldsymbol{\mu}$ cannot be tractably integrated out, inferences can be based on posterior samples, drawn by trans-dimensional MCMC algorithms such as the reversible-jump (Green, 1995). In the rest of the paper, for computational efficiency, the conjugate priors in (2.26) and (2.27) are used, under which the (conditional) marginal likelihoods have closed forms.

2.3.3 The BMDL expressions

We now obtain a BMDL for each changepoint model $\boldsymbol{\eta}$. As derivations for the univariate and bivariate cases are similar, work is shown only for the univariate model. Recall that the mixture MDL is applied to the dimensionally varying parameter $\boldsymbol{\mu}$, and the two-part MDL is applied to the other model parameters.

For a changepoint configuration $\boldsymbol{\eta}$ with $m > 0$, the conditional marginal likelihood has the closed form

$$\begin{aligned} f(\mathbf{X}_{(p+1):N} \mid \mathbf{s}, \sigma^2, \boldsymbol{\phi}, \boldsymbol{\eta}) &= \int_{\mathbb{R}^m} f\left(\mathbf{X}_{(p+1):N} \mid \boldsymbol{\mu}, \mathbf{s}, \sigma^2, \boldsymbol{\phi}, \boldsymbol{\eta}\right) \pi(\boldsymbol{\mu} \mid \sigma^2, \boldsymbol{\eta}) d\boldsymbol{\mu} \\ &= (2\pi\sigma^2)^{-\frac{N-p}{2}} \nu^{-\frac{m}{2}} \left| \widetilde{\mathbf{D}}'\widetilde{\mathbf{D}} + \frac{\mathbf{I}_m}{\nu} \right|^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(\widetilde{\mathbf{X}}-\widetilde{\mathbf{A}}\mathbf{s})'\widetilde{\mathbf{B}}(\widetilde{\mathbf{X}}-\widetilde{\mathbf{A}}\mathbf{s})}, \end{aligned}$$

where the notation has

$$\widetilde{\mathbf{B}} = \mathbf{I}_{N-p} - \widetilde{\mathbf{D}} \left(\widetilde{\mathbf{D}}'\widetilde{\mathbf{D}} + \frac{\mathbf{I}_m}{\nu} \right)^{-1} \widetilde{\mathbf{D}}'. \tag{2.28}$$

If $\mathbf{s}, \sigma^2, \boldsymbol{\phi}$, and $\boldsymbol{\eta}$ are known, the mixture MDL in (2.19) is simply

$$\mathcal{L}(\mathbf{X}_{(p+1):N} \mid \mathbf{s}, \sigma^2, \boldsymbol{\phi}, \boldsymbol{\eta}) = -\log[f(\mathbf{X}_{(p+1):N} \mid \mathbf{s}, \sigma^2, \boldsymbol{\phi}, \boldsymbol{\eta})]. \tag{2.29}$$

A two-part MDL will be used to quantify the cost of transmitting $\mathbf{s}, \sigma^2, \boldsymbol{\phi}$, and the model $\boldsymbol{\eta}$. The optimal $\mathbf{s}$ and σ^2 have closed forms:

$$\hat{\mathbf{s}} = \arg\min_{\mathbf{s}} \mathcal{L}(\mathbf{X}_{(p+1):N} \mid \mathbf{s}, \sigma^2, \boldsymbol{\phi}, \boldsymbol{\eta}) = (\widetilde{\mathbf{A}}'\widetilde{\mathbf{B}}\widetilde{\mathbf{A}})^{-1}(\widetilde{\mathbf{A}}'\widetilde{\mathbf{B}}\widetilde{\mathbf{X}}), \tag{2.30}$$

$$\begin{aligned}\hat{\sigma}^2 &= \arg\min_{\sigma^2} \mathcal{L}(\mathbf{X}_{(p+1):N} \mid \hat{\mathbf{s}}, \sigma^2, \boldsymbol{\phi}, \boldsymbol{\eta}) \\ &= \frac{1}{N-p}\widetilde{\mathbf{X}}' \left[\widetilde{\mathbf{B}} - \widetilde{\mathbf{B}}\widetilde{\mathbf{A}} \left(\widetilde{\mathbf{A}}'\widetilde{\mathbf{B}}\widetilde{\mathbf{A}}\right)^{-1} \widetilde{\mathbf{A}}'\widetilde{\mathbf{B}}\right] \widetilde{\mathbf{X}}.\end{aligned} \tag{2.31}$$

These estimators depend on $\boldsymbol{\phi}$. After plugging (2.30) and (2.31) into (2.29), the $\boldsymbol{\phi}$ that minimizes $\mathcal{L}(\mathbf{X}_{(p+1):N} \mid \hat{\mathbf{s}}, \hat{\sigma}^2, \boldsymbol{\phi}, \boldsymbol{\eta})$ is intractable. In general, likelihood estimators for autoregressive models do not have closed forms. Hence, simple Yule-Walker moment estimators, which are asymptotically most efficient and $\sqrt{n}$-consistent under the true changepoint model, are used. There is little difference between moment and likelihood estimators for autoregressions; Brockwell and Davis (1991) discuss this issue in detail.

In the linear model (2.3), the ordinary least squares residuals are

$$\boldsymbol{\epsilon}^{\text{ols}}_{1:N} = (\mathbf{I}_N - \mathcal{P}_{[\mathbf{A}_{1:N}\ \mathbf{D}_{1:N}]})\mathbf{X}_{1:N}, \tag{2.32}$$

where $\mathcal{P}_{[\mathbf{A}_{1:N}\ \mathbf{D}_{1:N}]}$ is the orthogonal projection matrix onto the linear space spanned by the columns of $\mathbf{A}_{1:N}$ and $\mathbf{D}_{1:N}$. The sample autocovariance of the residuals at lag $h \in \{0, 1, \ldots, p\}$ are

$$\hat{\gamma}(h) = \frac{1}{N}\sum_{t=h+1}^{N} \epsilon^{\text{ols}}_t \epsilon^{\text{ols}}_{t-h}. \tag{2.33}$$

The Yule-Walker estimator of $\boldsymbol{\phi}$ is

$$\hat{\boldsymbol{\phi}} = \hat{\boldsymbol{\Gamma}}_p^{-1}\hat{\boldsymbol{\gamma}}_p, \tag{2.34}$$

where $\hat{\boldsymbol{\gamma}}_p = (\hat{\gamma}(1), \hat{\gamma}(2), \ldots, \hat{\gamma}(p))'$ and $\hat{\boldsymbol{\Gamma}}_p$ is a $p \times p$ matrix whose (i, j)th entry is

$\hat{\gamma}(|i-j|)$. This matrix is invertible whenever the data are non-constant (Brockwell and Davis, 1991). Next, the Yule-Walker estimator $\hat{\boldsymbol{\phi}}$ is substituted for $\boldsymbol{\phi}$ in (2.5), (2.6), (2.7), and (2.28). The resulting quantities are denoted by $\widehat{\mathbf{X}}$, $\widehat{\mathbf{A}}$, $\widehat{\mathbf{D}}$, and $\widehat{\mathbf{B}}$, respectively. In particular, $\widehat{\mathbf{X}}$ contains estimated one-step-ahead prediction residuals (innovations). Hence, up to a constant, (2.29) is

$$\begin{aligned}\mathcal{L}(\mathbf{X}_{(p+1):N} \mid \hat{\mathbf{s}}, \hat{\sigma}^2, \hat{\boldsymbol{\phi}}, \boldsymbol{\eta}) =& \frac{N-p}{2} \log\left(\widehat{\mathbf{X}}' \left[\widehat{\mathbf{B}} - \widehat{\mathbf{B}}\widehat{\mathbf{A}}\left(\widehat{\mathbf{A}}'\widehat{\mathbf{B}}\widehat{\mathbf{A}}\right)^{-1}\widehat{\mathbf{A}}'\widehat{\mathbf{B}}\right]\widehat{\mathbf{X}}\right) \\ &+ \frac{m}{2}\log(\nu) + \frac{1}{2}\log\left(\left|\widehat{\mathbf{D}}'\widehat{\mathbf{D}} + \frac{\mathbf{I}_m}{\nu}\right|\right).\end{aligned}$$

By (2.15), under the uniform encoder, the additional costs to transmit $\hat{\mathbf{s}}, \hat{\sigma}^2$, and $\hat{\boldsymbol{\phi}}$ are (up to constants)

$$\mathcal{L}(\hat{\mathbf{s}} \mid \boldsymbol{\eta}) = \frac{T}{2}\log(N-p),\ \mathcal{L}(\hat{\sigma}^2 \mid \boldsymbol{\eta}) = \frac{1}{2}\log(N-p),\ \mathcal{L}(\hat{\boldsymbol{\phi}} \mid \boldsymbol{\eta}) = \frac{p}{2}\log(N-p). \quad (2.35)$$

These costs are constant across models and hence can be omitted from the MDL. Furthermore, based on (2.23), the MDL of $\boldsymbol{\eta}$ is

$$\mathcal{L}(\boldsymbol{\eta}) = -\log[\pi(\boldsymbol{\eta})] = -\sum_{k=1}^{2}\log\left[\Gamma\left(a + m^{(k)}\right)\Gamma\left(b^{(k)} + N^{(k)} - m^{(k)}\right)\right].$$

Using (2.14) and (2.16) and omitting the terms in (2.35), the BMDL to transmit the data $\mathbf{X}_{(p+1):N}$, the model $\boldsymbol{\eta}$, and its parameters is

$$\mathrm{BMDL}(\boldsymbol{\eta}) = \mathcal{L}(\mathbf{X}_{(p+1):N} \mid \hat{\mathbf{s}}, \hat{\sigma}^2, \hat{\boldsymbol{\phi}}, \boldsymbol{\eta}) + \mathcal{L}(\boldsymbol{\eta}). \quad (2.36)$$

For a model with $m > 0$ changepoints, its BMDL is hence

$$\begin{aligned}\text{BMDL}(\boldsymbol{\eta}) = &\frac{N-p}{2}\log\left\{\widehat{\mathbf{X}}'\left[\widehat{\mathbf{B}} - \widehat{\mathbf{B}}\widehat{\mathbf{A}}\left(\widehat{\mathbf{A}}'\widehat{\mathbf{B}}\widehat{\mathbf{A}}\right)^{-1}\widehat{\mathbf{A}}'\widehat{\mathbf{B}}\right]\widehat{\mathbf{X}}\right\} + \frac{m}{2}\log(\nu) \\ &+\frac{1}{2}\log\left(\left|\widehat{\mathbf{D}}'\widehat{\mathbf{D}} + \frac{\mathbf{I}_m}{\nu}\right|\right) - \sum_{k=1}^{2}\log\left[\Gamma\left(a + m^{(k)}\right)\Gamma\left(b^{(k)} + N^{(k)} - m^{(k)}\right)\right].\end{aligned} \tag{2.37}$$

For the no changepoint model ($m = 0$), denoted by $\boldsymbol{\eta}_\emptyset$, the above needs modification since it does not involve $\boldsymbol{\mu}$. Skipping the mixture MDL step and arguing as above produce

$$\mathcal{L}(\mathbf{X}_{(p+1):N} \mid \hat{\mathbf{s}}, \hat{\sigma}^2, \hat{\boldsymbol{\phi}}, \boldsymbol{\eta}_\emptyset) = \frac{N-p}{2}\log\left\{\widehat{\mathbf{X}}'\left[\mathbf{I}_{N-p} - \widehat{\mathbf{A}}\left(\widehat{\mathbf{A}}'\widehat{\mathbf{A}}\right)^{-1}\widehat{\mathbf{A}}'\right]\widehat{\mathbf{X}}\right\}.$$

Hence, the BMDL for the model $\boldsymbol{\eta}_\emptyset$ is

$$\begin{aligned}\text{BMDL}(\boldsymbol{\eta}_\emptyset) = &\frac{N-p}{2}\log\left\{\widehat{\mathbf{X}}'\left[\mathbf{I}_n - \widehat{\mathbf{A}}\left(\widehat{\mathbf{A}}'\widehat{\mathbf{A}}\right)^{-1}\widehat{\mathbf{A}}'\right]\widehat{\mathbf{X}}\right\} \\ &- \sum_{k=1}^{2}\log\left[\Gamma(a)\,\Gamma\left(b^{(k)} + N^{(k)}\right)\right].\end{aligned} \tag{2.38}$$

Past MDL authors (Davis et al., 2006; Lu et al., 2010) use formulas containing the term $\log(m)$, which is problematic for the null model $\boldsymbol{\eta}_\emptyset$ where $m = 0$. The BMDL in (2.38) resolves this issue.

For the bivariate case where $m_1 + m_2 > 0$, the conditional marginal likelihood, after integrating $\boldsymbol{\mu}$ out, retains a closed form:

$$\begin{aligned}&f(\mathbf{X}_{(p+1):N} \mid \mathbf{s}, \boldsymbol{\Sigma}, \boldsymbol{\Phi}_{1:p}, \boldsymbol{\eta}) \\ &\propto |\boldsymbol{\Sigma}|^{-\frac{N-p}{2}}\,|\boldsymbol{\Omega}|^{-\frac{1}{2}}\left|\widetilde{\mathbf{D}}'(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{I}_{N-p})\widetilde{\mathbf{D}} + \boldsymbol{\Omega}^{-1}\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(\widetilde{\mathbf{X}}-\widetilde{\mathbf{A}}\mathbf{s})'\widetilde{\mathbf{B}}(\widetilde{\mathbf{X}}-\widetilde{\mathbf{A}}\mathbf{s})},\end{aligned} \tag{2.39}$$

where $\widetilde{\mathbf{B}}$ is modified to

$$\begin{aligned}\widetilde{\mathbf{B}} =&(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_{N-p})\times \\ &\left\{\mathbf{I}_{2(N-p)} - \widetilde{\mathbf{D}}\left[\widetilde{\mathbf{D}}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_{N-p})\widetilde{\mathbf{D}} + \boldsymbol{\Omega}^{-1}\right]^{-1}\widetilde{\mathbf{D}}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_{N-p})\right\}.\end{aligned}$$

The least squares estimator of $\tilde{\mathbf{s}}$ that optimizes (2.39) is unaltered from (2.30). However, after plugging $\hat{\mathbf{s}}$ back in (2.39), the maximum likelihood estimators of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_p$ again do not have closed forms. Therefore, Yule-Walker estimators are used.

To find Yule-Walker estimators for the the time series regression in (2.3) and (2.9), generalized least squares residuals of the mean fit, denoted by $\boldsymbol{\epsilon}^{\text{gls}}_{1:N} = ((\boldsymbol{\epsilon}^{\text{gls}}_{1:N,1})', (\boldsymbol{\epsilon}^{\text{gls}}_{1:N,2})')' \in \mathbb{R}^{2N}$, are computed via

$$\boldsymbol{\epsilon}^{\text{gls}}_{1:N} = \left\{\mathbf{I}_{2N} - \mathbf{G}\left[\mathbf{G}'\left(\hat{\boldsymbol{\Gamma}}^{\text{ols}}(0)^{-1} \otimes \mathbf{I}_N\right)\mathbf{G}\right]^{-1}\mathbf{G}'\left(\hat{\boldsymbol{\Gamma}}^{\text{ols}}(0)^{-1} \otimes \mathbf{I}_N\right)\right\}\mathbf{X}_{1:N}, \quad (2.40)$$

where the design matrix is

$$\mathbf{G} = \begin{bmatrix} \mathbf{A}_{1:N,1} & \mathbf{D}_{1:N,1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_{1:N,2} & \mathbf{D}_{1:N,2} \end{bmatrix}.$$

Here $\hat{\boldsymbol{\Gamma}}^{\text{ols}}(0) = N^{-1}\sum_{t=1}^{N}\boldsymbol{\epsilon}^{\text{ols}}_t(\boldsymbol{\epsilon}^{\text{ols}}_t)'$ is a 2×2 covariance matrix of the ordinary (unweighted) least squares residuals $\boldsymbol{\epsilon}^{\text{ols}}_t = (\epsilon^{\text{ols}}_{t,1}, \epsilon^{\text{ols}}_{t,2})'$, where $\epsilon^{\text{ols}}_{t,1}$ and $\epsilon^{\text{ols}}_{t,2}$ are computed analogously to (2.32) with the design matrices $[\mathbf{A}_{1:N,1}\ \mathbf{D}_{1:N,1}]$ and $[\mathbf{A}_{1:N,2}\ \mathbf{D}_{1:N,2}]$, respectively.

The sample autocovariances at lag $h \in \{0, 1, \ldots, p\}$ of the generalized least

squares residuals $\boldsymbol{\epsilon}_t^{\text{gls}} = (\epsilon_{t,1}^{\text{gls}}, \epsilon_{t,2}^{\text{gls}})', t \in \{1, 2, \ldots, N\}$ are computed as:

$$\hat{\boldsymbol{\Gamma}}(h) = \frac{1}{N} \sum_{t=h+1}^{N} \boldsymbol{\epsilon}_t^{\text{gls}} (\boldsymbol{\epsilon}_{t-h}^{\text{gls}})'.$$

The Yule-Walker estimators in the bivariate setting are

$$\left(\hat{\boldsymbol{\Phi}}_1, \ldots, \hat{\boldsymbol{\Phi}}_p\right) = \left(\widehat{\boldsymbol{\Gamma}}(1), \ldots, \widehat{\boldsymbol{\Gamma}}(p)\right) \begin{bmatrix} \widehat{\boldsymbol{\Gamma}}(0) & \widehat{\boldsymbol{\Gamma}}(1) & \cdots & \widehat{\boldsymbol{\Gamma}}(p-1) \\ \widehat{\boldsymbol{\Gamma}}(1)' & \widehat{\boldsymbol{\Gamma}}(0) & \cdots & \widehat{\boldsymbol{\Gamma}}(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\boldsymbol{\Gamma}}(p-1)' & \widehat{\boldsymbol{\Gamma}}(p-2)' & \cdots & \widehat{\boldsymbol{\Gamma}}(0) \end{bmatrix}^{-1},$$

$$\widehat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Gamma}}(0) - \sum_{j=1}^{p} \hat{\boldsymbol{\Phi}}_j \hat{\boldsymbol{\Gamma}}(j)'.$$

After plugging $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\Phi}}_1, \ldots, \widehat{\boldsymbol{\Phi}}_p$ back into the likelihood, the terms $\widetilde{\mathbf{X}}, \widetilde{\mathbf{A}}, \widetilde{\mathbf{D}}, \widetilde{\mathbf{B}}, \boldsymbol{\Omega}$, which depend on $\boldsymbol{\Sigma}$ and $\boldsymbol{\Phi}_1, \cdots, \boldsymbol{\Phi}_p$ are denoted by $\widehat{\mathbf{X}}, \widehat{\mathbf{A}}, \widehat{\mathbf{D}}, \widehat{\mathbf{B}}, \widehat{\boldsymbol{\Omega}}$. Hence, the Bayesian MDL for $\boldsymbol{\eta}$ is (up to a constant)

$$\begin{aligned} & \text{BMDL}(\boldsymbol{\eta}) \qquad (2.41) \\ = \; & -\sum_{k=1}^{2} \sum_{\ell=1}^{4} \log \left[\Gamma \left(\alpha_\ell^{(k)} + m_\ell^{(k)} \right) \right] + \frac{N-p}{2} \log \left(\left| \widehat{\boldsymbol{\Sigma}} \right| \right) + \frac{1}{2} \sum_{i=1}^{2} m_i \log(\nu \hat{\sigma}_i^2) \\ & + \frac{1}{2} \log \left(\left| \widehat{\mathbf{D}}' (\widehat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_{N-p}) \widehat{\mathbf{D}} + \widehat{\boldsymbol{\Omega}}^{-1} \right| \right) + \frac{1}{2} \widehat{\mathbf{X}}' \left[\widehat{\mathbf{B}} - \widehat{\mathbf{B}} \widehat{\mathbf{A}} \left(\widehat{\mathbf{A}}' \widehat{\mathbf{B}} \widehat{\mathbf{A}} \right)^{-1} \widehat{\mathbf{A}}' \widehat{\mathbf{B}} \right] \widehat{\mathbf{X}}. \end{aligned}$$

The null model $\boldsymbol{\eta}_{\emptyset}$ has BMDL

$$\begin{aligned} & \mathrm{BMDL}(\boldsymbol{\eta}_{\emptyset}) \qquad (2.42) \\ = \; & -\sum_{k=1}^{2}\sum_{\ell=1}^{4}\log\left[\Gamma\left(\alpha_{\ell}^{(k)}\right)\right] + \frac{N-p}{2}\log\left(\left|\widehat{\boldsymbol{\Sigma}}\right|\right) + \frac{1}{2}\widehat{\mathbf{X}}'(\widehat{\boldsymbol{\Sigma}}^{-1}\otimes\mathbf{I}_{N-p})\widehat{\mathbf{X}} \\ & -\frac{1}{2}\widehat{\mathbf{X}}'\left\{(\widehat{\boldsymbol{\Sigma}}^{-1}\otimes\mathbf{I}_{N-p})\widehat{\mathbf{A}}\left[\widehat{\mathbf{A}}'(\widehat{\boldsymbol{\Sigma}}^{-1}\otimes\mathbf{I}_{N-p})\widehat{\mathbf{A}}\right]^{-1}\widehat{\mathbf{A}}'(\widehat{\boldsymbol{\Sigma}}^{-1}\otimes\mathbf{I}_{N-p})\right\}\widehat{\mathbf{X}}. \end{aligned}$$

2.3.4 BMDL optimization

The optimal changepoint model $\hat{\boldsymbol{\eta}}$ has the smallest BMDL score. If a BMDL for each model can be computed, one selects the model with the smallest BMDL. However, exhaustively searching the changepoint configuration space is formidable. Even in the univariate case, the total number of models, 2^{N-p}, is extremely large. Genetic algorithms are used by Davis et al. (2006) and Lu et al. (2010) to overcome this hurdle. Here, a Markov chain Monte Carlo approach is developed.

Connections between BMDL and empirical Bayes (EB) allow us to efficiently explore the model space by visiting a relatively small number of promising models. The univariate BMDL in (2.36) is equivalent to the negative logarithm of an EB estimator of the posterior probability of $\boldsymbol{\eta}$ under the prior distributions in (2.23) and (2.26):

$$p_{\mathrm{EB}}(\boldsymbol{\eta} \mid \mathbf{X}_{(p+1):N}) \propto \; \pi(\boldsymbol{\eta}) \int_{\mathbb{R}^m} f\left(\mathbf{X}_{(p+1):N} \mid \boldsymbol{\mu}, \hat{\mathbf{s}}, \hat{\sigma}^2, \hat{\boldsymbol{\phi}}, \boldsymbol{\eta}\right) \pi(\boldsymbol{\mu} \mid \hat{\sigma}^2, \boldsymbol{\eta}) d\boldsymbol{\mu}.$$

A similar result holds in bivariate cases. As a BMDL approach is tractable, Bayesian stochastic model search algorithms can be used; see García-Donato and Martínez-Beneito (2013) and the references therein.

Here, the Metropolis-Hastings algorithm in George and McCulloch (1997) is

modified by intertwining two types of proposals: a component-wise flipping at a random location and a simple random swapping between a changepoint and a non-changepoint. This algorithm is described in detail in Li and Lund (2015) and is implemented by the R package `BayesMDL` in the supplementary material.

2.4 Asymptotic Consistency of the BMDL

This section shows that in univariate cases, the true changepoint model has the smallest BMDL under infill asymptotics as $N \to \infty$. Infill asymptotics assume that the number of observations between all changepoints tends to infinity and have been previously studied in the multiple changepoint detection literature. For example, Davis et al. (2006) prove that when the true value of m is known, their MDL for piecewise autoregressive processes is a consistent model selector. Du et al. (2015) prove a similar result for their marginal likelihood maximizer under independent observations, while relaxing the condition that the true m is known. Here, an analogous result for our BMDL is proven, which allows autocorrelation, seasonality, and an unknown true value of m.

A relative changepoint configuration of m changepoints is denoted by $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_m)'$, where $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_m < 1$. Here, time is scaled to $[0, 1]$ by mapping time t to t/N. For example, $\lambda_1 = 0.1$ means the first changepoint occurs 10% into the data record. For the edges, set $\lambda_0 = 0$ and $\lambda_{m+1} = 1$. For a fixed N, the rth changepoint location τ_r can be recovered from $\boldsymbol{\lambda}$ via $\tau_r = \lfloor \lambda_r N \rfloor$. For $r \in \{1, 2, \ldots, m+1\}$, the length of the rth regime $N_r = \lfloor \lambda_r N \rfloor - \lfloor \lambda_{r-1} N \rfloor$ satisfies

$$\lim_{N \to \infty} \frac{N_r}{N} = \lambda_r - \lambda_{r-1}. \tag{2.43}$$

For any $\boldsymbol{\lambda}$, no changepoints occur in $t \in \{1, 2, \ldots, p\}$ when N is large.

Suppose, the relative changepoint configuration is $\boldsymbol{\lambda}^0 = (\lambda_1^0, \lambda_2^0, \ldots, \lambda_{m^0}^0)'$ in truth. True parameter values are superscripted with zero. Our goal is to identify $\boldsymbol{\lambda}^0$ over many models. In fact, for a (fixed) large integer M, all relative changepoint configurations in

$$\boldsymbol{\Lambda} = \{\boldsymbol{\lambda} : 0 \leq m \leq M, \min_{r=1,2,\ldots,m+1} \lambda_r - \lambda_{r-1} \geq d\}$$

are considered, where d is a small positive constant, smaller than $\lambda_r^0 - \lambda_{r-1}^0$ for all $r \in \{1, 2, \ldots, m^0 + 1\}$. We assume that $m^0 \leq M$; hence, $\boldsymbol{\lambda}^0 \in \boldsymbol{\Lambda}$. Between the true model $\boldsymbol{\lambda}^0$ and any other model $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$, the pairwise difference of their BMDLs in (2.37) or (2.38) is used to decide which model is favorable.

Theorem 2.4.1. *In the univariate case, for any relative changepoint configuration* $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$*, if* $\boldsymbol{\lambda} \neq \boldsymbol{\lambda}^0$*, then as* $N \to \infty$*,*

$$BMDL\,(\boldsymbol{\lambda}) - BMDL\,\left(\boldsymbol{\lambda}^0\right) \xrightarrow{P} \infty. \tag{2.44}$$

More specifically, if all relative changepoints in $\boldsymbol{\lambda}^0$ *are contained in* $\boldsymbol{\lambda}$*, then the BMDL difference in* (2.44) *is* $O_P(\log N)$*; otherwise, it is* $O_P(N)$*.*

A proof of Theorem 2.4.1 is provided in Appendix A. This theorem shows that asymptotically, the true relative changepoint model $\boldsymbol{\lambda}^0$ achieves the smallest BMDL in probability among all competing models in $\boldsymbol{\Lambda}$. An implication of the result is that it is possible to consistently identify the true changepoint configuration in the limit. While this result is proven for the univariate case, a similar bivariate result is expected.

2.5 A Simulation Study

This section studies the BMDL's changepoint detection performance under finite samples via simulation. Our simulation parameters are selected to roughly mimic the Tuscaloosa data. Specifically, the bivariate error series $\{\boldsymbol{\epsilon}_t\}$ follows a zero mean Gaussian VAR model with $p = 3$. The VAR parameters are taken as

$$\boldsymbol{\Phi}_1 = \begin{pmatrix} 0.2 & 0.02 \\ 0.02 & 0.2 \end{pmatrix}, \boldsymbol{\Phi}_2 = \begin{pmatrix} 0.1 & 0.01 \\ 0.01 & 0.1 \end{pmatrix}, \boldsymbol{\Phi}_3 = \begin{pmatrix} 0.05 & 0.005 \\ 0.005 & 0.05 \end{pmatrix},$$

and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 9 & 2 \\ 2 & 9 \end{pmatrix}.$$

In each of 1000 independent runs, 50 year monthly Tmax and Tmin series ($N = 600$) are simulated with $m = 3$ changepoints in each series. For the Tmax series, mean shifts are placed at times $150, 300$, and 450. The regime means have form $\boldsymbol{\mu}_1 = (0, \Delta, 2\Delta, 3\Delta)'$ where $\Delta > 0$ will be varied. For the Tmin series, mean shifts are placed at times $150, 300$, and 375. The regime means are $\boldsymbol{\mu}_2 = (0, -\Delta, \Delta, 0)'$. Here, Tmax has monotonic "up, up, up" shifts of equal shift magnitudes; Tmin shifts in a "down, up, down" fashion and the second shift is twice as large as the other two shifts. The shifts at times 150 and 300 are concurrent in both series; the shift at time 150 moves Tmax upwards and Tmin downwards.

Seasonal means are set to $\mathbf{s} = (0, 3, 10, 18, 26, 33, 36, 36, 31, 20, 8, 2)'$ in both series. Seasonal mean parameters are not critical, but the Δ parameter controlling the mean shift size is. Our detection powers will be reported under different signal to noise ratios, which is defined as $\kappa = \Delta/\sigma$. We will examine $\kappa \in \{1, 1.5, 2\}$, where $\sigma = 3$. For metadata, a record containing four documented changes at the times

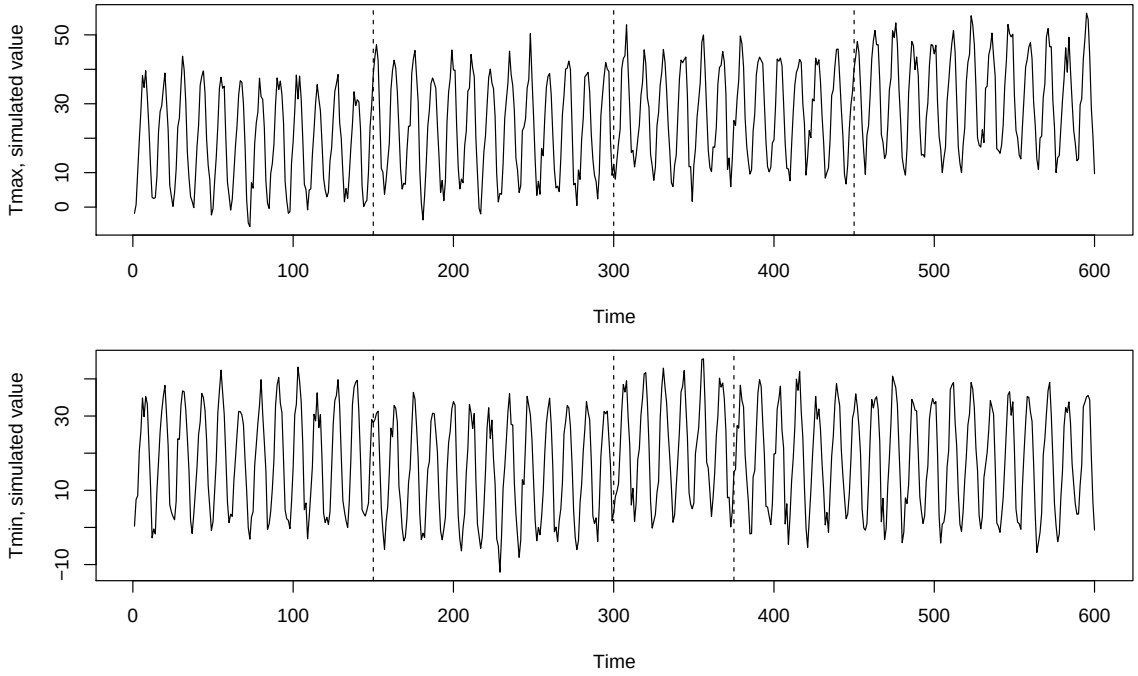

Figure 2.3: A simulated dataset with three changepoints in Tmax (top panel) and three changepoints in Tmin (bottom panel). Vertical dashed lines mark the true changepoint times.

$75, 150, 250$, and 550 is posited. Among the documented times, only time 150 is a true changepoint. A simulated series with $\kappa = 1.5$ is shown in Figures 2.3. Figure 2.4 shows the same series after subtraction of sample monthly means.

2.5.1 Univariate simulations

Each Tmax and Tmin series is fitted via univariate BMDL methods, once without metadata and once with it. In each fit, a Metropolis-Hastings chain of 100,000 iterations is generated. The optimal multiple changepoint model is taken as the one with the smallest BMDL. All hyper-parameters are set to default values.

For Tmax series, empirical detection percentages (at their exact times) are reported in the top half of Table 2.1. Since the three shifts are of equal size Δ, the detection rates should be similar, as the top panel in Figure 2.5 confirms. The

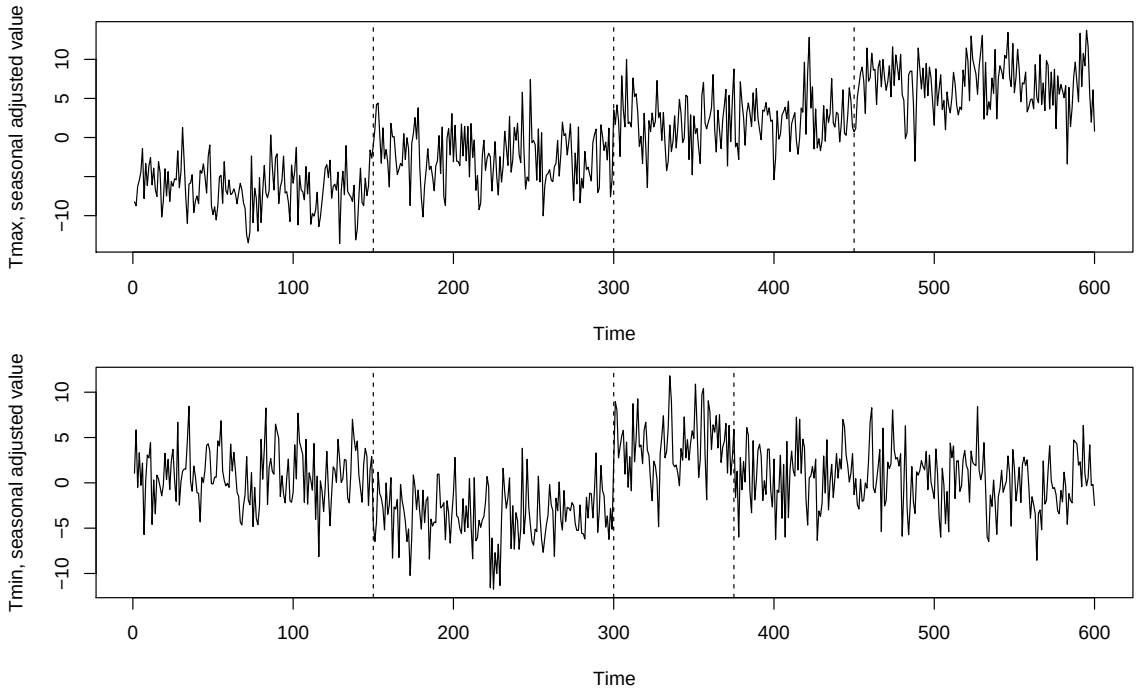

Figure 2.4: The Figure 3 series after subtracting sample monthly means. Vertical dashed lines mark the true changepoint times.

average false detection rate of non-changepoint times is very low; when $\kappa = 1$, this false positive rate is only 0.4%.

Use of metadata substantially increases detection power. In Figure 2.5, the true documented change at time 150 is detected 76.2% of the time when metadata is used, more than twice as high (36.1%) when metadata is eschewed. Moreover, times near the changepoint at time 150 are less likely to be flagged as changepoints. Our prior belief that metadata times are more likely to be changepoints is important, especially when the mean shift is small: when $\kappa = 1$, using metadata increases the detection rate of the time 150 changepoint from 15.2% to 57.4%. On the other hand, Figure 2.5 shows that using metadata does not substantially increase false positives (the prior distribution likely does not overwhelm the data). Table 2.1 shows that the average detection rates of the three metadata times that do not induce mean shifts

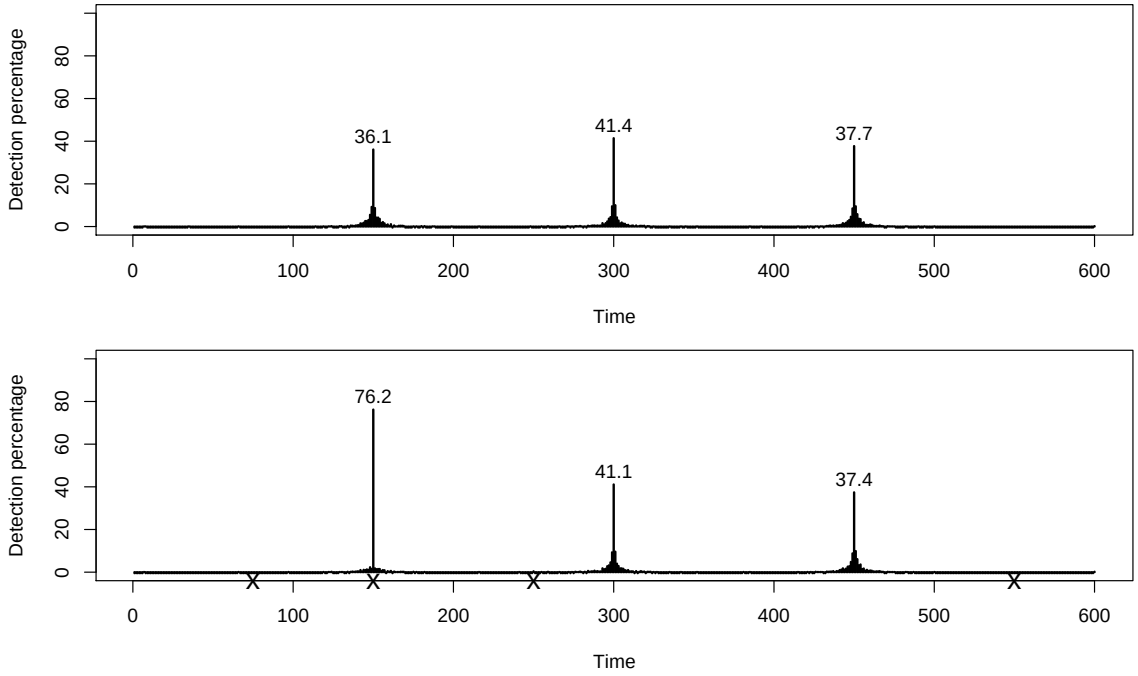

Figure 2.5: Detection times and percentage of changepoints in Tmax series using univariate BMDL methods. The top panel ignores the four metadata times; the bottom panel uses the metadata. Metadata times are marked as crosses on the axis. The results are aggregated from 1000 independent simulated Tmax series simulated with $\kappa = 1.5$.

— times 75, 250, and 550 — are similar to the overall false positive rates; the latter even drop after using metadata.

The number of estimated changepoints is also studied. Table 2.2 reports, with or without metadata, the correct number of changepoints ($m = 3$) is estimated in more than 60% of the runs when $\kappa = 1$, and in more than 98% of the runs when $\kappa = 1.5$ and $\kappa = 2$. Metadata use slightly increases accuracy.

For Tmin series, the non-monotonic shift aspect (down, up, down) that trouble AMOC binary segmentation approaches (Li and Lund, 2012) is well handled by our BMDL method. The top half of Table 2.3 shows that when metadata is ignored, the larger shift at time 300 is more easily detected than the two smaller shifts at times 150 and 375. When metadata is used, the detection rate of the shift at time 150 is

Table 2.1: Changepoint detection percentage for Tmax, aggregated from 1000 simulated series.

		True positive			False positive		
κ	Metadata	$t = 150$	$t = 300$	$t = 450$	average	avg(meta)	$t = 375$
Univariate							
1.0	no	15.2	15.5	16.4	0.4	0.0	0.1
	yes	57.4	17.4	15.3	0.3	0.4	0.0
1.5	no	36.1	41.4	37.7	0.3	0.0	0.0
	yes	76.2	41.1	37.4	0.2	0.1	0.0
2.0	no	54.3	59.5	57.4	0.2	0.0	0.0
	yes	84.1	59.1	57.6	0.2	0.0	0.0
Bivariate							
1.0	no	36.5	55.2	11.4	0.4	0.0	8.3
	yes	60.7	54.5	11.5	0.3	0.0	6.8
1.5	no	66.7	82.9	33.9	0.2	0.0	10.8
	yes	81.1	82.2	34.2	0.2	0.0	7.3
2.0	no	84.7	94.8	55.6	0.1	0.0	6.2
	yes	92.1	93.5	55.9	0.1	0.0	3.7

comparable to the detection rate of the mean shift at time 300, which is twice as large, but is not a metadata time. False positive rates are uniformly low. Table 2.4 shows that when $\kappa = 1$, the correct number of changepoints ($m = 3$) is estimated over 76% of the time; when $\kappa = 1.5$ and $\kappa = 2$, this rate increases to over 98%.

2.5.2 Bivariate fits

Each bivariate series is fitted by a MCMC chain of 50,000 iterations — once without metadata, and once with metadata. Metadata impacts are similar to the univariate case, increasing the detection rates of the true metadata times and also slightly decreasing overall false positive rates (see the bottoms of Tables 2.1 and 2.3).

As concurrent shifts are believed more likely to occur, bivariate methods should enhance detection power of concurrent changepoints. Figure 2.6 shows the bivariate detection rates when $\kappa = 1.5$. At time 150, where Tmax (Tmin) shifts Δ ($-\Delta$),

Table 2.2: Empirical percentage of estimated number of changepoints m for Tmax, aggregated from 1000 independent simulated series.

κ	Metadata	0	1	2	3	4	5	≥ 6
Univariate								
1.0	no	0.0	4.4	33.0	60.0	2.5	0.1	0.0
	yes	0.0	2.7	32.6	62.8	1.9	0.0	0.0
1.5	no	0.0	0.0	0.0	98.0	2.0	0.0	0.0
	yes	0.0	0.0	0.0	98.4	1.6	0.0	0.0
2.0	no	0.0	0.0	0.0	98.2	1.8	0.0	0.0
	yes	0.0	0.0	0.0	98.3	1.6	0.1	0.0
Bivariate								
1.0	no	0.0	0.2	1.9	78.0	18.7	1.2	0.0
	yes	0.0	0.0	4.1	80.2	14.9	0.8	0.0
1.5	no	0.0	0.0	0.0	71.3	27.9	0.8	0.0
	yes	0.0	0.0	0.0	83.0	16.0	0.7	0.3
2.0	no	0.0	0.0	0.0	87.9	11.6	0.5	0.0
	yes	0.0	0.0	0.0	93.4	6.1	0.5	0.0

the bivariate BMDL increases the univariate detection rate from about 77% to above 81%. At time 300, where Tmax (Tmin) shifts by Δ (2Δ), the detection rate for Tmax increases from 41.1% to 82.2%. Tables 2.1 and 2.3 show that detection power gains under the bivariate approach are greater for small κ: when $\kappa = 1$, without metadata, the bivariate BMDL increases detection rates at time 150 from 15.2% to 36.5% for Tmax, and from 18.8% to 36.2% for Tmin. Furthermore, the detection rate at time 300 for Tmax increases from 15.5% to 52.2%.

An interesting phenomenon is observed: bivariate methods improve univariate methods more when the concurrent shifts move the series in opposite directions. For example, at time 150, where the mean of Tmax rises and the mean of Tmin drops, the bivariate approach increases the detection rates for both Tmax and Tmin. In contrast, at time 300, where Tmax and Tmin both shift upwards, bivariate methods substantially improve Tmax detection, whose absolute shift size is Δ; however, it hardly improves Tmin detection, where the mean shift is larger (2Δ). This phe-

Table 2.3: Changepoint detection percentage for Tmin, aggregated from 1000 simulated series.

		True positive			False positive		
κ	Metadata	$t = 150$	$t = 300$	$t = 375$	average	avg(meta)	$t = 450$
Univariate							
1.0	no	18.8	52.3	14.3	0.3	0.0	0.0
	yes	61.3	52.8	14.1	0.2	0.1	0.0
1.5	no	36.7	84.3	39.2	0.2	0.0	0.0
	yes	77.3	84.6	38.2	0.2	0.0	0.0
2.0	no	58.3	95.4	56.4	0.2	0.0	0.0
	yes	85.3	95.4	56.1	0.1	0.0	0.0
Bivariate							
1.0	no	36.2	55.3	10.2	0.4	0.0	9.6
	yes	60.1	54.9	9.5	0.3	0.0	8.7
1.5	no	66.4	83.4	34.2	0.3	0.0	21.3
	yes	81.2	83.0	33.0	0.2	0.0	15.2
2.0	no	84.8	95.1	54.9	0.2	0.0	32.1
	yes	92.0	94.8	57.8	0.1	0.0	16.2

nomenon is explainable: Tmax and Tmin are positively correlated series. Hence, concurrent shifts in the same direction may be misattributed to positively correlated errors; this cannot happen when the the two series shift in opposite directions.

Overall, while bivariate detection does not induce more false positives, it tends to flag more false positives at locations where the mean in the other series shifts. Figure 2.6 shows that at time 375, a changepoint time in Tmin but not in Tmax, a false detection rate of 7.3% for Tmax is obtained. At time 450, a changepoint in Tmax but not Tmin, a false detection rate of 15.2% is obtained for Tmin. These false positive rates likely degrade inferences at nearby changepoints; for example, at time 450 for Tmax and time 375 for Tmin, detection rates are 34.2% and 33.0%, respectively, slightly lower than the 37.4% and 38.2% reported in the univariate case. Finally, the bottom halves of Tables 2.2 and 2.4 show that bivariate approaches tend to overestimate m, which differs from univariate methods.

Table 2.4: Empirical percentage of estimated number of changepoints m for Tmin, aggregated from 1000 independent simulated series.

κ	Metadata	0	1	2	3	4	5	≥ 6
Univariate								
1.0	no	6.5	1.8	14.0	76.3	1.2	0.2	0.0
	yes	5.7	0.9	14.7	78.1	0.5	0.1	0.0
1.5	no	0.0	0.0	0.0	98.1	1.6	0.3	0.0
	yes	0.0	0.0	0.2	98.4	1.2	0.2	0.0
2.0	no	0.0	0.0	0.0	98.5	1.4	0.1	0.0
	yes	0.0	0.0	0.0	98.8	1.1	0.1	0.0
Bivariate								
1.0	no	0.8	0.0	2.2	76.3	19.6	1.1	0.0
	yes	1.0	0.2	3.9	78.2	15.6	1.1	0.0
1.5	no	0.0	0.0	0.0	41.2	56.6	2.1	0.1
	yes	0.0	0.0	0.0	63.9	34.5	1.1	0.5
2.0	no	0.0	0.0	0.0	42.5	56.2	1.1	0.2
	yes	0.0	0.0	0.0	72.8	26.5	0.7	0.0

2.6 The Tuscaloosa Data

The monthly Tuscaloosa data in Section 2.1.1 will now be analyzed. Results for univariate and bivariate BMDLs, with and without metadata, will be reported. All hyper-parameters are set to default values and $p = 2$ is judged appropriate. Justifying the AR order further, Figure 2.7 plots sample autocorrelation of residuals fitted by univariate BMDL methods with $p = 2$ with pointwise 95% confidence bands. Almost all residual autocorrelations lie inside the confidence bands.

To ensure MCMC convergence in the search algorithm, for each fit, 50 Markov chains are generated from different starting points, each containing one million (univariate) or 100,000 (bivariate) iterations. Among all changepoint models visited by the 50 Markov chains, the one with the smallest BMDL is reported as the optimal model.

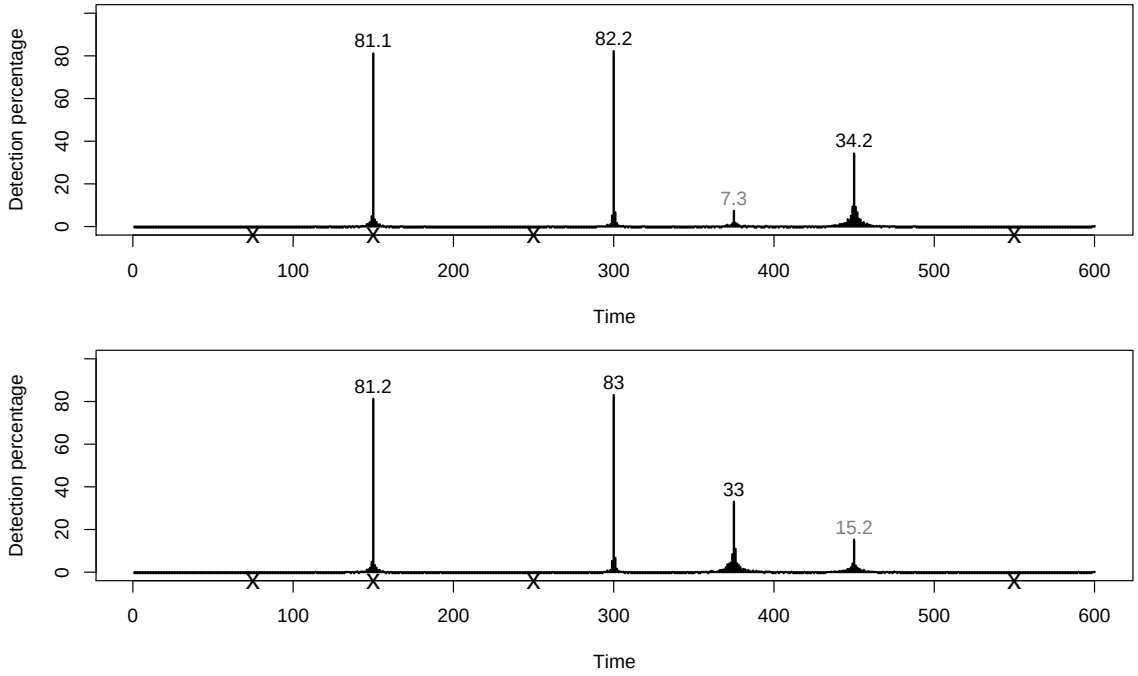

Figure 2.6: Detection percentages of Tmax (top panel) and Tmin (bottom panel) using bivariate BMDL methods with metadata (metadata times are marked as crosses on the axis). Numerical percentages on the graphic are for detection at "their exact time". The results are aggregated from 1000 independent Tmax series simulations with $\kappa = 1.5$.

2.6.1 Univariate fits

The top half of Table 2.5 displays detected changepoints for the univariate series without using our reference series. When metadata is ignored, Tmax has two estimated changepoints and Tmin has three; of these, only January 1990 is a concurrent change. Another changepoint is approximately concurrent — March 1957 for Tmax and July 1957 for Tmin. The 1918 changepoint flagged for Tmin is close to the station relocation in November 1921; the station relocation in June 1956 and the equipment change in November 1956 are near the two estimated changepoints in 1957. The metadata time in May 1987 is about three years from the concurrent changepoints flagged in January 1990. Of course, when metadata is ignored,

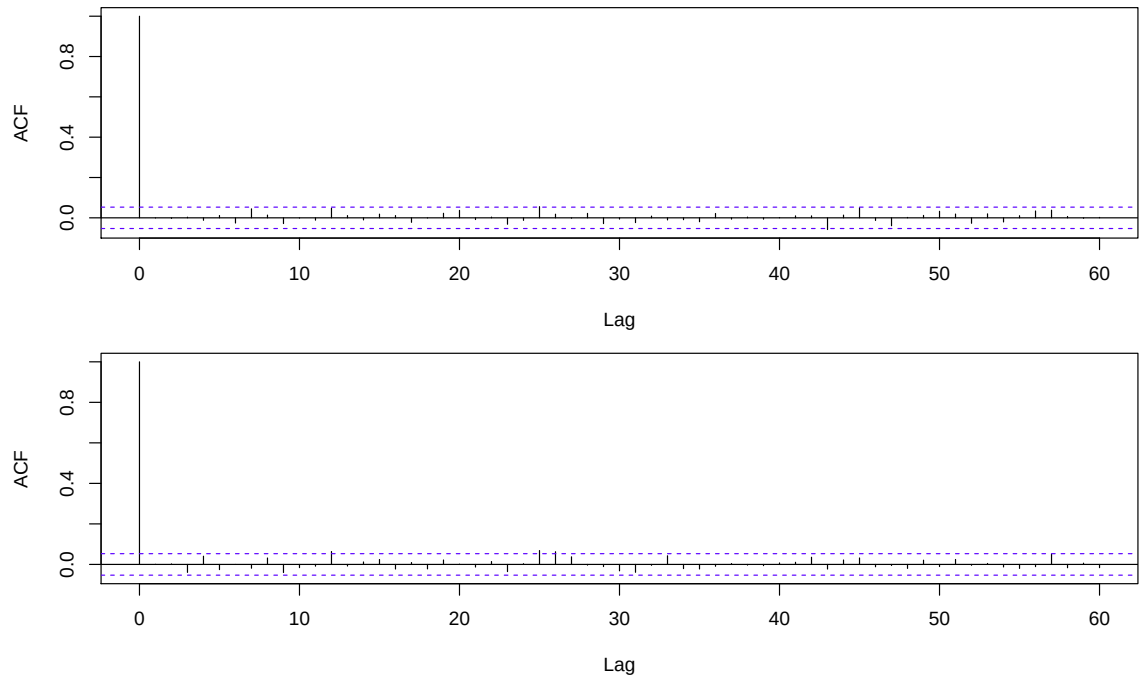

Figure 2.7: Sample model residual autocorrelations for Tmax (top panel) and Tmin (bottom panel), fitted using the univariate BMDL with metadata and $p = 2$.

estimated changepoint times may not coincide (exactly) with metadata times.

Repeating the above analysis with metadata (this still ignores our reference series), two changepoints are found in Tmax and three in Tmin. The estimated changepoint times now coincide with metadata times. Only the May 1987 changepoint is concurrent. Between Tmax and Tmin, the two estimated changepoints in 1956 (i.e., the two metadata times in 1956) are just a few months apart. As parameter estimates are similar with or without metadata, only estimates for the optimal changepoint model using metadata are reported. For Tmax, estimated regime means are (standard errors in parentheses) $\hat{\mu}_2 = -1.50$ (0.24) and $\hat{\mu}_3 = 0.66$ (0.25) (recall that $\mu_1 = 0$); estimated AR(2) coefficients are $\hat{\phi}_1 = 0.21, \hat{\phi}_2 = 0.05$, and $\hat{\sigma}^2 = 11.59$. For Tmin, the estimated parameters are $\hat{\mu}_2 = 1.76$ (0.21), $\hat{\mu}_3 = -1.06$ (0.22), $\hat{\mu}_4 = 2.35$ (0.24), $\hat{\phi}_1 = 0.18, \hat{\phi}_2 = 0.05$, and $\hat{\sigma}^2 = 10.81$. The concurrent May 1987 changepoint shifts both

Table 2.5: Estimated changepoints for the Tuscaloosa data.

Metadata	Series	Estimated changepoints
Univariate		
no	Tmax	1957 Mar, 1990 Jan
	Tmin	1918 Feb, 1957 Jul, 1990 Jan
yes	Tmax	1956 Nov, 1987 May
	Tmin	1921 Nov, 1956 Jun, 1987 May
Bivariate		
no	Tmax	1918 Feb, 1957 Jul, 1988 Jul
	Tmin	1918 Feb, 1957 Jul, 1988 Jul
yes	Tmax	1921 Nov, 1956 Jun, 1987 May
	Tmin	1921 Nov, 1956 Jun, 1987 May

series to warmer regimes.

2.6.2 Bivariate fits

The analyses are repeated using both series in tandem. Three changepoints are detected in both series, with or without metadata, and all are concurrent (see the bottom half of Table 2.5). Figure 2.1 shows the optimal bivariate BMDL changepoint configuration. When metadata is used, all estimated changepoint times migrate to metadata times. Comparing to the univariate results, the bivariate approach yields the same changepoint configuration for Tmin; for Tmax, a new changepoint in November 1921 is flagged and the November 1956 changepoint moves to June 1956, thus becoming a concurrent change. For this changepoint configuration, the estimated VAR parameters are

$$\widehat{\mathbf{\Phi}}_1 = \begin{pmatrix} 0.21 & -0.01 \\ -0.02 & 0.20 \end{pmatrix}, \quad \widehat{\mathbf{\Phi}}_2 = \begin{pmatrix} 0.06 & -0.02 \\ -0.04 & 0.08 \end{pmatrix},$$

and

$$\widehat{\mathbf{\Sigma}} = \begin{pmatrix} 11.56 & 8.13 \\ 8.13 & 10.81 \end{pmatrix}.$$

Finally, the target minus reference series are analyzed using the bivariate BMDL and Tuscaloosa's metadata record. Climatologists trust target minus reference analyses more than target analyses alone because the target minus reference comparison reduces variabilities and trends. As shown in Figure 2.2, the optimal changepoint configuration for the difference series contains 12 concurrent changes: June 1914, January 1919, July 1933, July 1937, August 1937, October 1938, December 1938, June 1946, July 1946, November 1956, May 1987, and October 1996. Among them, the 1956 and 1987 changepoints are in the metadata; the two changepoints in 1938 are close to the 1939 station relocation. The changepoints in 1919, 1933, and 1990 are also flagged by Lu et al. (2010). One of the shifts, November 1956, moves the Tmax series warmer and the Tmin series colder.

Some of the changepoints may be due to typos in the raw record. Specifically, the October and December 1938 changepoints are likely a recording error, whereby the October and November 1938 Tmin values in the target minus reference series appear to be abnormally high. While these series have been quality checked, some errors still occur. This conjecture is made because the three references stations lie in various directions from Tuscaloosa; climatologically, series to the north and west of Tuscaloosa should be cooler and those to the south and east should be warmer. In this case, Tuscaloosa was significantly warmer than all references. Similar statements may apply to the two "outlier" changepoints in 1937, and the two changepoints in 1946, where the Tmin records for Tuscaloosa are lower than those for all three reference stations. It is interesting that MDL methods pick up these outliers.

It is natural to flag more changepoints in the target minus reference series

than the target series alone. An ideal reference series should have the same trend and seasonal cycles as the target series and be free of artificial mean shifts. This said, we do not assume that the target minus reference comparison completely removes the monthly mean cycle; indeed, Liu et al. (2015) shows that this is seldom the case. Reference series selection is a problem currently studied by climatologists. As our reference series averages three neighbor stations, mean shifts in any of the reference records may induce shifts in the target minus reference series. For example, the estimated changepoint in 1914 is close to the 1915 metadata time of the Aberdeen reference. This said, averaging three neighbors should help mitigate the effects of changepoints in any individual reference series.

2.7 Discussion

This chapter developed a multiple changepoint detection approach amenable to metadata. MDL penalization methods were modified to accommodate various prior distributional specifications. The theory was used to detect mean shifts in univariate autoregressive processes with seasonal means, and then extended to bivariate VAR settings. The methods have several advantages, including simple parameter elicitation, asymptotic consistency, and efficient computation.

The approach can be extended to accommodate more flexible error structures including moving averages, periodic autoregressions, and more than two series. The methods could also be tailored to categorical data. With count data, the likelihood could be Poisson-based. With a conjugate Gamma prior, the resulting marginal likelihoods will again have closed forms. There is no technical difficulty in allowing a background linear trend, or even piecewise linear trends. This said, linear trends can be mistaken for multiple mean shifts should trends be present and ignored (Li and

Lund, 2015).

Non-MCMC stochastic search methods are possible. The genetic algorithms popular in multiple changepoint MDL optimizations can also be used to minimize the BMDL. When no global parameters exist in the likelihood (i.e., independent observations, no seasonal cycle, error variance known), dynamic programming techniques can further accelerate computational speed.

Chapter 3

Homogenization of Daily Temperature Series

3.1 Multiple Changepoint Models for Daily Data

Our object of interest is a daily temperature series. Such series display autocorrelation, seasonal means, trends, and possible multiple mean shifts at changepoint times. A model that captures the above features will now be devised. We consider data $\mathbf{X} = (X_1, \ldots, X_N)'$ recorded daily. Here, $N = dT$, where $T = 365$ is the period of the series and d is the number of years of data. We assume data for d complete years to avoid trite work. The season (day of year) is indexed by $\nu \in \{1, 2, \ldots, T\}$. The notation $X_{nT+\nu}$ refers to the observation during the νth day of the nth year, for years $n = 0, 1, \ldots, d-1$. With daily data, leap year observations are omitted to enforce the period $T = 365$.

Our fundamental model is a linear regression with seasonality, trends, multiple

possible mean shifts, and periodic random errors:

$$X_{nT+\nu} = \mu_\nu + \alpha(nT+\nu) + \delta_{nT+\nu} + \varepsilon_{nT+\nu}. \tag{3.1}$$

Here, μ_ν is the mean temperature on day ν (neglecting the trend and mean shifts). We assume that the linear trend parameter, α, is time-homogeneous; other trend structures can be accommodated, but this is seldom necessary when examining target minus reference series as the subtraction greatly reduces any trends. The time-ordered changepoints are denoted by $\tau_1 < \tau_2 < \cdots < \tau_m$, where m is the unknown number of changepoints. The changepoint structure can be described by a binary indicator vector $\boldsymbol{\eta} = (\eta_2, \ldots, \eta_N)'$, with

$$\eta_t = \begin{cases} 1, & \text{if time } t \text{ is a changepoint} \\ 0, & \text{otherwise} \end{cases}.$$

This model enables 2^{n-1} distinct changepoint models.

The m changepoints in $\boldsymbol{\eta}$ partition the series into $m+1$ distinct regimes. The kth regime consists of the observations from the times t with $\tau_{k-1} \leq t < \tau_k$ for $k = 1, 2, \ldots, m+1$. We take $\tau_0 = 0$ and $\tau_{m+1} = N+1$ for edge notations. In regime $j \geq 2$, Δ_j is how much the mean has shifted relative to the first regime (neglecting the seasonal cycle and the trend). For parameter identifiability, $\Delta_1 = 0$ is imposed.

Define a vector $\boldsymbol{\Delta} = (\Delta_2, \ldots, \Delta_{m+1})'$ whose entries are the mean shift parameters that need to be estimated. The component $\{\delta_{nT+\nu}\}$ in (3.1) has the shift structure

$$\delta_t = \begin{cases} \Delta_1 = 0, & \tau_0 \le t < \tau_1 \\ \Delta_2, & \tau_1 \le t < \tau_2 \\ \vdots & \\ \Delta_{m+1}, & \tau_m \le t < \tau_{m+1} \end{cases} .$$

The shift from regime j to regime $j+1$ changes mean temperatures by $\Delta_{j+1} - \Delta_j$ degrees.

The model errors $\{\epsilon_{nT+\nu}\}$ have zero mean and are correlated. Daily temperatures are in fact heavily correlated, with consecutive days often having correlation on the order of 0.7. As winter temperatures are some 2 to 5 times more variable than summer temperatures in the United States (as measured by standard deviation), a first order periodic autoregressive time series (PAR(1)) (Lund et al. (1995)) will be used for the regression errors. A PAR(1) time series indeed has autocorrelation and periodic variances. A time series $\{\epsilon_t\}$ is said to be a PAR(1) series with zero mean if it satisfies the seasonal difference equation

$$\epsilon_{nT+\nu} = \phi(\nu)\epsilon_{nT+\nu-1} + Z_{nT+\nu}. \tag{3.2}$$

Here, $\phi(\nu)$ is the autoregressive parameter during day ν. In (3.2), $\{Z_{nT+\nu}\}$ is zero mean periodic white noise with variance $\sigma^2(\nu) = \mathrm{Var}(Z_{nT+\nu})$.

Our model has the changepoint parameters $m; \tau_1, \tau_2, \ldots, \tau_m$, the seasonal means $\mu_1, \ldots, \mu_T$, the linear trend α, the mean shifts $\Delta_2, \ldots, \Delta_{m+1}$ and the time series PAR(1) parameters $\phi(1), \ldots, \phi(T)$ and $\sigma^2(1), \ldots, \sigma^2(T)$. In next section, we introduce a Bayesian MDL objective function, which is subsequently minimized to estimate an optimal configuration.

3.2 Bayesian Minimum Description Lengths (BMDLs)

This section develops an objective function that can be minimized to obtain an estimate of the unknown changepoint configuration $\boldsymbol{\eta}$. The task is similar to the derivation in Li et al. (2015) for monthly data; however, because such derivations are lengthy, we will only list the end objective function. This said, the derivation needed is slightly different from Li et al. (2015) since $\{\epsilon_t\}$ has periodic components here.

The MDL principle will be used as our model selection criteria. An MDL objective function is a penalized likelihood with a smart penalty tailored to the changepoint problem. The MDL penalty has an analogous role to Akaike Information Criterion (AIC) and the Bayesian information criterion (BIC) penalties, but differs from them in that it is more than a simple multiple of the number of changepoint parameters. In fact, the MDL penalty depends on how far the changepoints lie from one and other. The MDL penalty was developed in Rissanen (1989) from information theory. Among a class of plausible models, the MDL principle seeks the model with the shortest so-called description length. Better models should have shorter description lengths. For more background, see Hansen and Yu (2001) and Grünwald et al. (2005). The MDL principle has been utilized in climate changepoint detection (Davis et al. (2006); Li and Lund (2012); Lu et al. (2010)). In fact, Li et al. (2015) developed a new MDL technique (a Bayesian MDL) that uses metadata. Here, these methods are modified for daily data.

For a given changepoint configuration $\boldsymbol{\eta}$ with at least one changepoint, the Bayesian MDL objective function is

$$\text{BMDL}(\boldsymbol{\eta}) = \frac{m}{2}\log(\kappa g^2) + \frac{1}{2}\sum_{t=1}^{N}\log(\sigma^2(t)) + \frac{1}{2}\log(|\mathbf{B}|) + \frac{1}{2}\sum_{t=1}^{N}\frac{Y_t^2}{\sigma^2(t)} - \frac{1}{2}\mathbf{b}'\mathbf{B}^{-1}\mathbf{b}$$
$$- \log\left[\Gamma(1+m_1)\Gamma(\beta_1+n_1-m_1)\Gamma(1+m_2)\Gamma(\beta_2+n_2-m_2)\right]. \qquad (3.3)$$

In the above, $\Gamma(x)$ is the Gamma function at the argument x, the logarithm is natural-based, and $|\mathbf{B}|$ is the determinant of the matrix $\mathbf{B}$. The optimal changepoint configuration is obtained as the $\boldsymbol{\eta}$ that minimizes $\text{BMDL}(\boldsymbol{\eta})$. For each candidate changepoint configuration $\boldsymbol{\eta}$, the mean shift, trend, and time series parameters are not too difficult to optimally estimate. That this procedure works will be shown in our forthcoming simulation section.

Our next objective is to explain what the parameters in (3.3) represent. Toward this, the prediction residuals $\{Y_t\}_{t=1}^N$ are computed from

$$Y_t = [X_t - \mu_t - \alpha t] - \phi(t)[X_{t-1} - \mu_{t-1} - \alpha(t-1)],$$

with the convention that $Y_1 = X_1 - \mu_1 - \alpha$. During the jth regime,

$$a_j = \sum_{t=\tau_{j-1}}^{\tau_j-1}\frac{1}{\sigma^2(t)} + \sum_{t=\tau_{j-1}}^{\tau_j-1}\frac{\phi^2(t+1)}{\sigma^2(t+1)} - 2\sum_{t=\tau_{j-1}+1}^{\tau_j-1}\frac{\phi(t)}{\sigma^2(t)},$$

$$b_j = \sum_{t=\tau_{j-1}}^{\tau_j-1}\frac{Y(t)}{\sigma^2(t)} - \sum_{t=\tau_{j-1}}^{\tau_j-1}\frac{Y_{t+1}\phi(t+1)}{\sigma^2(t+1)},$$

and $c_j = \phi(\tau_{j-1})/\sigma^2(\tau_{j-1})$. Also, $\mathbf{b} = (b_2, \ldots, b_{m+1})$, and $\mathbf{B}$ is an $m \times m$ symmetric matrix with form

$$\mathbf{B} = \begin{pmatrix} a_2 + \frac{1}{\kappa g^2} & -c_3 & 0 & \cdots & 0 \\ -c_3 & a_3 + \frac{1}{\kappa g^2} & -c_4 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -c_{m+1} & a_{m+1} + \frac{1}{\kappa g^2} \end{pmatrix}.$$

Finally, m_1 and m_2 are the number of undocumented and documented changepoints, respectively. Thus, the total number of changepoints is $m = m_1 + m_2$. Moreover, n_2 is the number of metadata points and $n_1 = N - n_2$. The parameter $\kappa > 0$ is any large constant; it is thought of in units of standard deviation. Our default takes $\kappa = 5$.

Li et al. (2015) show that the MDL in (3.3) can be written as $\text{BMDL}(\boldsymbol{\eta}) = \mathcal{L}(\mathbf{X} \mid \boldsymbol{\eta}) + \mathcal{L}(\boldsymbol{\eta})$. The first five terms in (3.3) are $\mathcal{L}(\mathbf{X} \mid \boldsymbol{\eta})$, and the last term is $\mathcal{L}(\boldsymbol{\eta})$:

$$\begin{aligned} \mathcal{L}(\mathbf{X} \mid \boldsymbol{\eta}) &= \frac{m}{2}\log(\kappa g^2) + \frac{1}{2}\sum_{t=1}^{N}\log(\sigma^2(t)) + \frac{1}{2}\log|\mathbf{B}| + \frac{1}{2}\sum_{t=1}^{N}\frac{Y_t^2}{\sigma^2(t)} - \frac{1}{2}\mathbf{b}'\mathbf{B}^{-1}\mathbf{b}, \\ \mathcal{L}(\boldsymbol{\eta}) &= -\log\left[\Gamma(1+m_1)\Gamma(\beta_1 + n_1 - m_1)\Gamma(1+m_2)\Gamma(\beta_2 + n_2 - m_2)\right]. \end{aligned}$$

In these equations, means shifts are assumed normal and independent: $\boldsymbol{\Delta} \sim \text{N}(\mathbf{0}, \kappa g^2 \mathbf{I}_m)$, where $g^2 = [\prod_{\nu=1}^{T}\sigma^2(\nu)]^{\frac{1}{T}}$ is the geometric mean of $\sigma^2(1), \ldots, \sigma^2(T)$. The description length of the changepoint configuration $\boldsymbol{\eta}$ is $\mathcal{L}(\boldsymbol{\eta})$. This is where metadata is used. Elaborating, a beta-binomial prior is put on $\boldsymbol{\eta}$. This prior assumes that 1) each undocumented time is a changepoint with probability ρ_1, that 2) each documented time is a changepoint with probability ρ_2, and 3) documented times are more likely than undocumented times to be changepoints: $\rho_2 > \rho_1$. In the absence of information beyond the metadata record, changepoints declarations at all distinct time points are assumed to be statistically independent. In a Bayesian hierarchical fashion, the parameter ρ_1 is modeled as Beta(1,β_1) random variate; ρ_2 is modeled as

a Beta(1,β_2) variate. Our default values take $\beta_1 = 365/.06$ and $\beta_2 = 4$. This makes $E[\rho_1] = 1/(1+365/.06) \approx .06/365$ (approximately six changepoints per century) and $E[\rho_2] = 1/(1+4) = 0.2$ (one out of every five metadata times induces a true mean shift). As one will see in our next simulation section, it is not important to specify these parameters exactly.

The BMDL for the changepoint configuration with no changepoints, denoted by $\boldsymbol{\eta}_0$, is

$$\text{BMDL}(\boldsymbol{\eta}_0) = \frac{1}{2}\sum_{t=1}^{N}\log(\sigma^2(t)) + \frac{1}{2}\sum_{t=1}^{N}\frac{Y_t^2}{\sigma^2(t)} - \log\left[\Gamma(\beta_1+n_1)\Gamma(\beta_2+n_2)\right]. \quad (3.4)$$

This allows one to compare models with changepoints to models with no changepoints and fixes an issue with the methods in Davis et al. (2006); Li and Lund (2012), where the term $\ln(m)$ arises (this is undefined for the no changepoint configuration when $m = 0$). The first two terms in (3.4) are the description length of the data and the last term is the description length of the changepoint configuration. In next section, we discuss minimizing the BMDL over all possible changepoint configurations, which yields our estimated changepoint model.

When computing a BMDL for changepoint configuration, $\phi(\nu)$ and $\sigma^2(\nu)$ are replaced by their Yule-Walker estimators (Lund et al., 1995). The seasonal means $\mu_1, \ldots, \mu_T$ and the linear trend α are also replaced by their estimates, which are computed via ordinary least squares.

3.3 BMDL Minimization

The best changepoint configuration is the one(s) that minimizes the BMDL score. A naive approach to finding this configuration is to perform an exhaustive search. Such an approach requires $\binom{N-1}{m}$ BMDL model fits when $\boldsymbol{\eta}$ has m changepoints at unknown locations. Summing this count over $m = 0, 1, \ldots, N-1$ and applying the binomial theorem shows that there are 2^{N-1} distinct BMDL evaluations to perform in an exhaustive search. Thus, for a century of daily data, 2^{36500} BMDL evaluations need to be conducted, an impossible task on even the world's fastest computers. Hence, an efficient optimization algorithm is needed to find the best model. For this, a genetic algorithm (GA), which is an intelligent random walk search that is unlikely to visit suboptimal changepoint configurations, is devised to perform the minimization.

GAs are popular optimization tools (Goldberg and Holland (1988)), inspired by natural selection and genetics. Like Darwin's theory of evolution, GAs have aspects of genetic evolution that allow the fittest models to survive in a stochastic random walk search. GAs usually converge to global optimums. Beasley et al. (1993) compares GAs to traditional optimization methods such as gradient step and search methods and simulated annealing.

GAs encode each model as a chromosome. Here, a chromosome is represented by a binary indicator vector $\boldsymbol{\eta} = (\eta_2, \ldots, \eta_N)$ as in Section 2. The number of changepoints is $m = \sum_{t=2}^{N} \eta_t$ and the changepoint locations $\tau_1, \ldots, \tau_m$ are the nonzero positions in $\boldsymbol{\eta}$. The GA begins with a randomly generated initial population of chromosomes (as described below) and evaluates the BMDL score at each generated chromosome. The GA then simulates successive generations of chromosomes via a series of operations: parent selection, crossover, and mutation. Chromosomes

with smaller BMDLs are viewed as fitter and are more likely to bear children. From each generation, two chromosomes (parent chromosomes) are selected. These parent chromosomes are combined (crossover) in a manner described below to form a new chromosome called a child. The child's chromosome is allowed to mutate (in a manner described below) before joining the next generation. This process is repeated until a preset number of children are produced for the generation. The resulting population of children is referred to as the next generation. A pre-specified number of generations are often simulated. If done right, the overall fitness of the population, which is the BMDL score of the fittest individual in the generation, converges to the best possible model. Details to implement this algorithm are now given.

Initial Generation

An initial population often simply simulates a set of chromosomes at random. Here, each position in a chromosome is allowed to be a changepoint with some preset probability. For daily data, this probability is set to be 6/36500 (following Mitchell (1953), this is 6 changepoints per century). While small generation sizes could induce premature convergence, larger generation sizes slow the algorithm down. After some experimentation, a generation size of 150 was used here.

Parent Selection

Once the initial generation is simulated, parent (mother and father chromosomes) are selected to breed. To generate fitter offspring, a parent selection technique is needed. Such a technique should be more likely to choose the fitter individuals to breed children. Several selection mechanisms are listed in Beasley et al. (1993). Here, a linear ranking is used to select the parents from the 150 chromosomes. First, the 150

chromosomes' BMDL score is ranked in descending order; the chromosome with the highest BMDL has rank 1 and the chromosome with the smallest BMDL has rank 150. Parents are chosen with probabilities proportional to their ranks: if the rank of the ith chromosome is R_i, it is selected as a father with probability $R_i / \sum_{j=1}^{150} R_j = R_i/11,325$. The most fit chromosome has a 0.1324 chance of being selected as father; the least fit chromosome has a 0.00008809 chance. Mothers are then selected in the same way from all non-father chromosomes.

Crossover

Crossover mechanisms combine mother and the father chromosomes in a random manner to generate a child chromosome. The child chromosome ideally contains changepoint characteristics of both parents. Our crossover mechanism allows changepoints in either parent to be changepoints of the child. The general idea is best illustrated with an example: suppose the mother chromosome is $\boldsymbol{\eta}_1 = (0, 1, 0, 1, 0, 0)$ and the father chromosome is $\boldsymbol{\eta}_2 = (0, 0, 0, 1, 0, 1)$. Here, $N = 6$, the mother has changepoints at times 2 and 4, and the father has changepoints at times 4 and 6. The child chromosome is first set to have changepoints of either mother or father: $(0, 1, 0, 1, 0, 1)$. At this point the child can have more chromosomes than the mother or father. Hence, some of the child's chromosomes are randomly discarded. With the aforementioned child chromosome, a fair coin is flipped three times (one at each of the three changepoint times) and all changepoints with tails are discarded. If the resulting sequence is heads, tails, heads, then the second changepoint at time 4 is discarded and the resulting chromosome becomes $(0, 1, 0, 0, 0, 1)$.

Since the number of distinct changepoint configurations is enormous, changepoint locations are perturbed to speed algorithm convergence. Here, the location of

any changepoint is shifted via an integer-valued random variable with zero mean. To execute this, two independent Poisson random numbers D_1 and D_2 are generated at each changepoint time; the changepoint's location is shifted $D_1 - D_2$ time units. For example, a chromosome containing three changepoints might see Poisson differences of $-1, 0$, and 3, respectively. Then the first changepoint is shifted downward one day, the second changepoint time is not shifted, and the third changepoint time is shifted upward three days. Should any of the shifted times be less than day 1 or more than day N, the changepoint is eliminated. Choosing the best Poisson parameter (λ) is tricky. In early generations, a larger λ is needed to explore new changepoint locations; in later generations, a smaller value of λ is preferred to slightly tune the likely good changepoint configurations being explored in the current models. The λ parameter is described further below.

Mutation

Each child is allowed to mutate after crossover. Mutation changes randomly selected bits of each chromosome. If mutation is not allowed, the GA can hone in to a local BMDL minimum; with mutation, radically different chromosomes are continually explored. Mutation essentially ensures the exploration of whole changepoint configuration space, maintaining a diversity of the chromosome population and preventing pre-mature algorithmic convergence. Our mutation mechanism selects a random number of locations and flips the changepoint of the child at each of these selected locations. For example, if position 100 is chosen for mutation and is not a changepoint in the child, it is flipped to a changepoint; should time 100 already be a changepoint, it is flipped to a non-changepoint. In our algorithm, each time is allowed to mutate independently with very small probability (described below). In

many chromosomes, no mutation occurs.

Islands and Migration

There are 2^{N-1} (day one cannot be a changepoint) distinct changepoint configurations in a daily series of length N. Hence, the performance of a conventional GA is relatively slow. To speed the convergence, we implement an island GA (Davis (1991)) that allows some of the fitter chromosomes to migrate between islands. In an island GA, populations are divided into several subpopulations, called islands. GAs are run simultaneously on every island. The islands are largely isolated, but migrations occur between islands every periodic now and again. This allows very fit chromosomes to change islands. Migration increases chromosome diversity and prevents the algorithm from converging to a local BMDL minimums. The migration policy depends the number of islands, the migration rate (number of individuals to migrate), and the migration interval (the frequency of migrations). Similar to Lu et al. (2010), our migration policy replaces the least-fit individual on each island by the best-fit individual of a randomly selected other island, once every five generations.

The GA is terminated when a prescribed stopping criteria is reached. The most frequently used stopping criteria are that a pre-specified maximum number of generations are reached, or the lack of improvement in the most fit member of successive generations. The most fit chromosome of the last generation (among all islands) is taken as the estimated changepoint configuration.

GA convergence depends on parameters such as the number of islands, the generation size of each island, the mutation probability, and the Poisson parameter λ. One does not have to tune these parameters optimally to get good results; however, an efficient algorithm is usually appreciated. In our work in the next sections, the

following parameter settings were used: 1) with 46 years of daily data, two islands of size 75 were used, the mutation probability was set to 0.0001 and $\lambda = 50$. For 10 years of daily data, two islands of size 25 were used, the mutation probability was set to 0.001, and $\lambda = 10$.

3.4 A Simulation Study

This section presents a simulation study to assess the performance of our methods. We simulated one thousand series, each containing 10 years of daily data ($N = 3650$). For application realism, the daily means and linear trend were set to be those estimated values of the South Haven, Michigan daily temperature series analyzed in the next section. The parameters of the PAR(1) model were set to those estimated in the target minus reference series of the next section. Smooth sine curves were then fitted to these seasonal parameters. Figure 3.1 graphically displays these parameters. In each simulated series, the metadata record is posited changes at the five times $456, 913, 1521, 3194$ and 3346.

As a control run, one thousand series were simulated with no changepoints and our methods were applied. An island GA with two islands was used to optimize the BMDL. The results estimate 962 series with no changepoints, 33 series with one changepoint, and five series with two changepoints. The false-alarm rate (3.8%) is reasonably low.

Next, one thousand series were simulated with three true changepoints at times $\tau_1 = 913$ (03-July-year 3), $\tau_2 = 1825$ (01-January-year 5), and $\tau_3 = 2700$ (26-May-year 8). Here, the first changepoint is also a metadata point. The mean shifts by $1.5, 2$ and 3 degrees F at the three changepoint times, respectively. Figure 3.2 displays a simulated series and its corresponding seasonally adjusted series, where

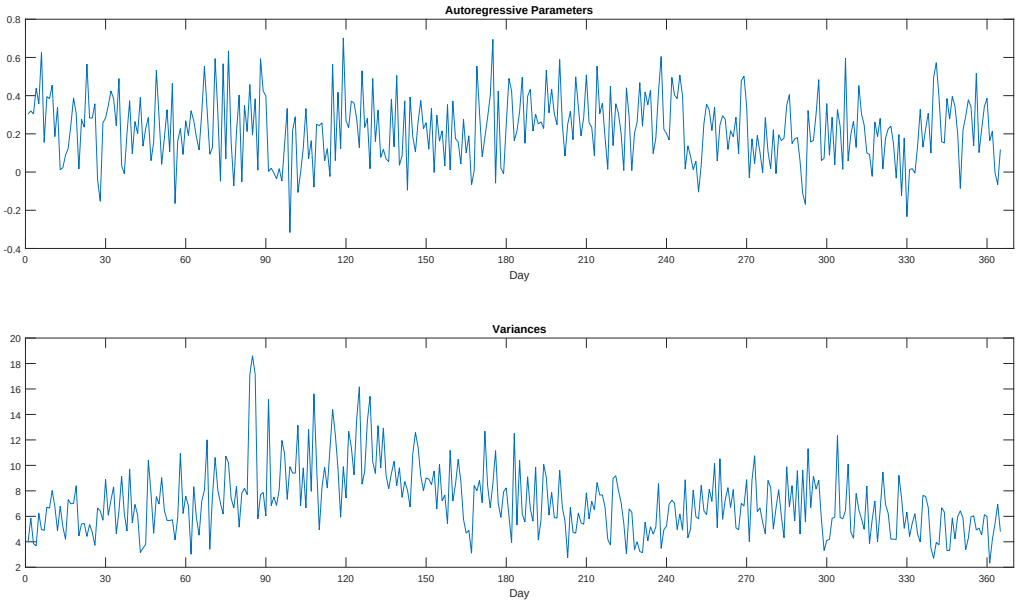

Figure 3.1: Autoregressive coefficients and periodic variances of the target-reference series.

daily means are subtracted. The metadata points are marked as crosses on the axis. Detection percentages are displayed in Figure 3.3. As expected, larger mean shifts make changepoint detection easier: the third changepoint has the largest detection count. Although the shift at τ_2 is larger than the shift at τ_1, the detection counts are not considerably different. This is attributed to two reasons. First, since τ_1 is a metadata point, τ_1 will be easier to flag as a changepoint, all other things being equal. Second, τ_1 occurs during summer, which is a season with less variability than τ_2 (a winter temperature). The higher winter variability makes changepoints harder to detect (this scenario is explored further below). Among the 1000 simulated series, the GA estimates the true number of changepoints correctly in 800 of the series. In the remaining 200 cases, 152 of the series were estimated to have 2 changepoints, 36 series to have one changepoint, and 12 series to have 4 changepoints.

Finally, one thousand series with two changepoints at $\tau_1 = 749$ (January 20th-

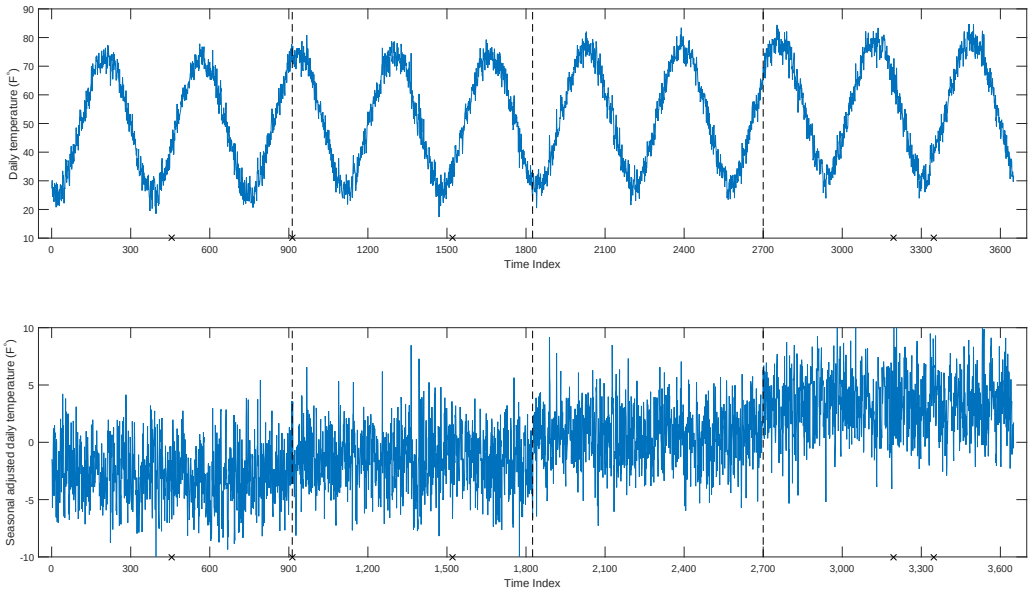

Figure 3.2: A simulated daily temperature series with three changepoints. The bottom plot shows the same series with the daily sample mean subtracted. Vertical dashed demarcate the three mean shift at times 913, 1825, and 2700.

year 3) and $\tau_2 = 2755$ (July 20th-year 3) were simulated. Both changepoints are posited to be undocumented and shift the mean $1.5°F$ upwards. Figure 3.4 displays a simulated series and corresponding seasonally adjusted series. The histogram of detection percentages is displayed in Figure 3.5. The detection rate of the January changepoint is indeed 13% lower than the detection rate of the July changepoint. Thus, winter changepoint detection is harder than summer changepoint detection. The true number of changepoints (two) were correctly estimated in 797 runs; 104 series were estimated with one changepoint, 12 series to have 3 changepoints, and 87 series to have no changepoints. Again, the true number of changepoints were estimated in about 80% of the runs.

In all simulation, an island GA with two islands of size 25 was applied to each

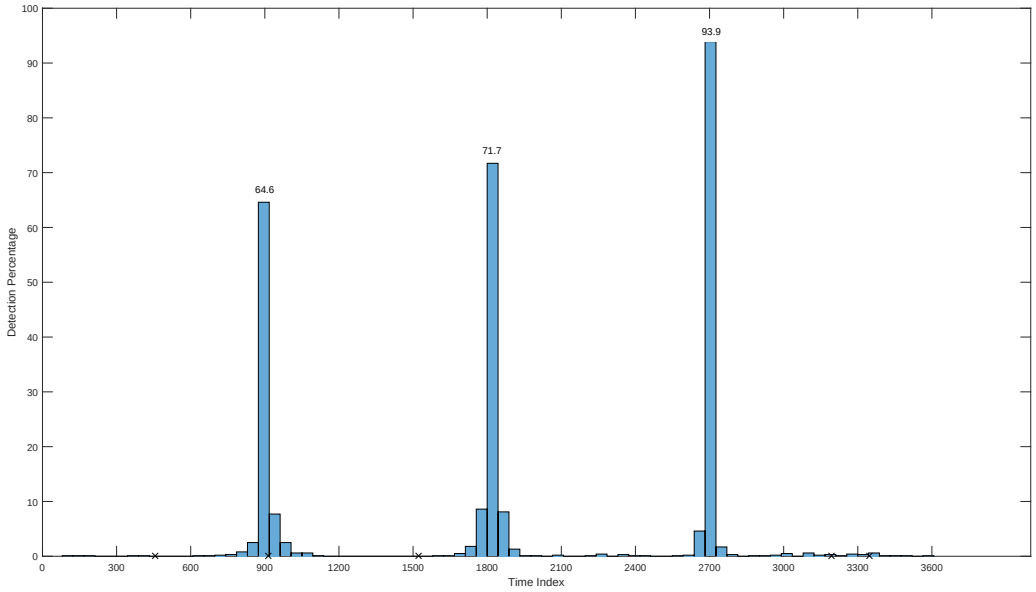

Figure 3.3: Detection rates. The true mean shifts are at times 913, 1825, and 2700. The detection rates spike around the true mean shift times, implying effective detection.

simulated series. Island GA conveged in less than 200 generations. For 200 iterations, the computational time was about 9 minutes.

3.5 South Haven, Michigan Analysis

Figure 3.6 depicts average daily temperatures at South Haven, Michigan from 1953-01-01 to 1998-12-31 (46 years). The bottom plots shows seasonally adjusted temperatures where a daily mean has been subtracted. Leap year data was omitted; hence, there are 365×46 $(16,790)$ data points. The periodic mean cycle of the daily temperatures in Figure 3.6 is evident; however, it is difficult to visually see changepoints in these plots. To illuminate mean shifts and to lessen trends and seasonal cycles, a reference series is often used (Menne and Williams Jr., 2005, 2009).

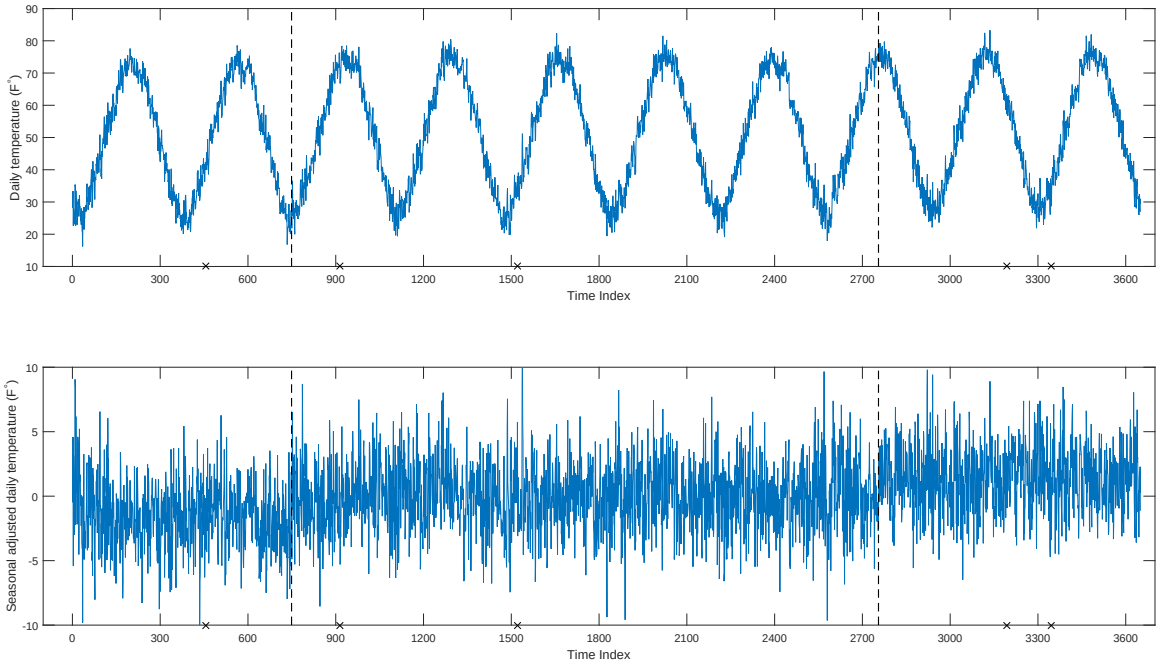

Figure 3.4: A simulated temperature series with two changepoints. The bottom plot shows the same series with daily mean subtracted. Vertical dashed lines demarcate the two mean shifts at times 749 and 2755.

For the South Haven target series, reference series are available from the nearby stations at Shelby, Benton Harbor, and Pellston Regional Airport. We use Benton Harbor series as our reference since it is located on the eastern coast of Lake Michigan, like South Haven. Figure 3.7 shows average daily temperature series and seasonally adjusted temperatures at Benton Harbor.

The records at South Haven and Benton Harbor are mostly complete, but do have a few sporadic missing data points (less than 1.3 % of the total record). For simplicity, missing data was infilled in our four series (maximums and minimums at the target and reference stations). To do this, a first-order vector autoregressive was fitted to the four series in tandem. Missing data were infilled as the best one-step ahead linear predictor. For example, if the maximum temperature of the reference

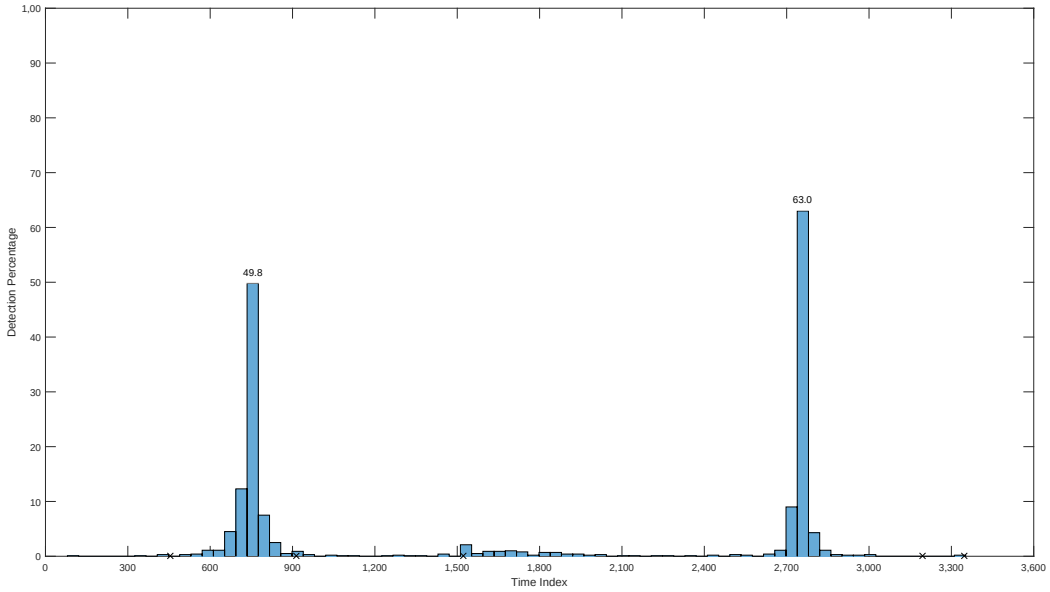

Figure 3.5: Detection rates. The true mean shifts are at times 749 and 2755. The detection rates spike around the true mean shift times, implying effective detection.

series at time t was missing, this point was estimated by its best one-step-ahead linear predictor from all non-missing observations of the other three series at times t, $t-1$, and $t+1$. Runs of missing values were infilled one at a time.

Figure 3.8 plots the difference of daily average temperatures (daily average temperatures are the average of maximum and minimum temperatures during the day) at South Haven and Benton Harbor. The graph appears to have some mean shifts, possibly attributable to either station. The metadata record of South Haven lists a change in equipment on 1990-08-22; the metadata record of Benton Harbor documents two station relocations in 1993-12-08 and 1996-06-19. These three times were used as metadata times in the analysis. The island GA algorithm with two islands, 75 changepoint models on each island, and 2000 generation iterations converged to a changepoint configuration with 13 changepoints. The run time of this was about 19

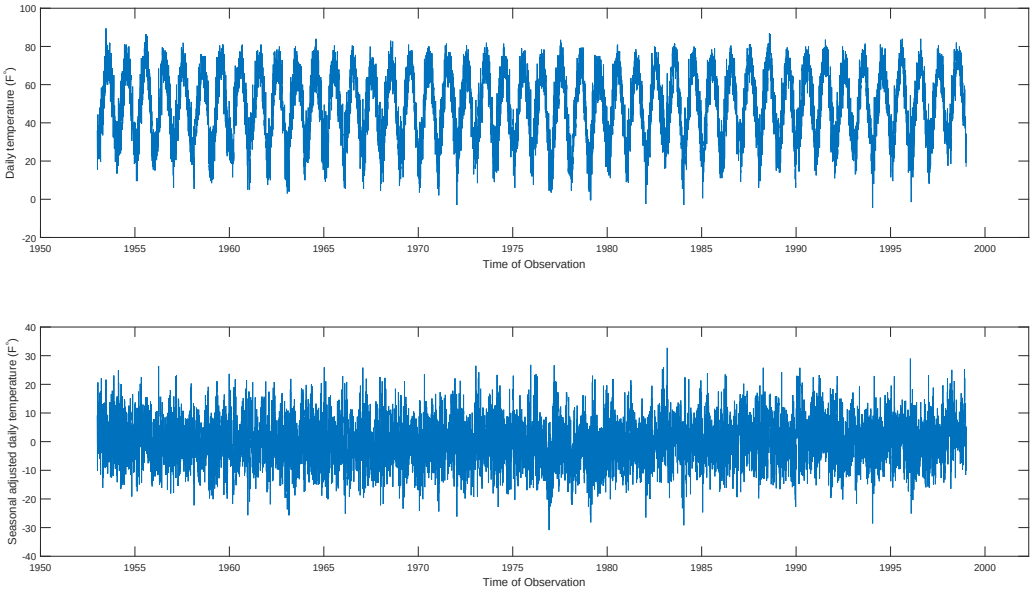

Figure 3.6: The South Haven daily average temperature series.

hours. The changepoint times and corresponding mean shifts are displayed in Table 1; figure 3.8 displays the fitted mean shift structure of the target minus reference series. Among the 13 flagged changepoints, only the 1993-12-26 changepoint is close to a metadata time (1993-12-08). Other metadata points have seemingly not induced significant mean shifts. The estimated PAR(1) autoregressive coefficients and their periodic variances are displayed in Figure 3.1. The estimated linear trend parameter is $-0.2208°F$ per century. Seven of the shift move the series to colder regimes and six to warmer regimes.

To complement the daily analysis, the annual target minus reference temperatures in Figure 3.9 were analyzed. Here, a multiple changepoint model with time-homogeneous autoregressive errors of order 1 was fitted to the data. A GA was used to minimize the annual BMDL score and revealed six changepoints at $1955, 1956, 1992, 1994, 1995,$ and 1997. Figure 3.9 shows the changepoints of the an-

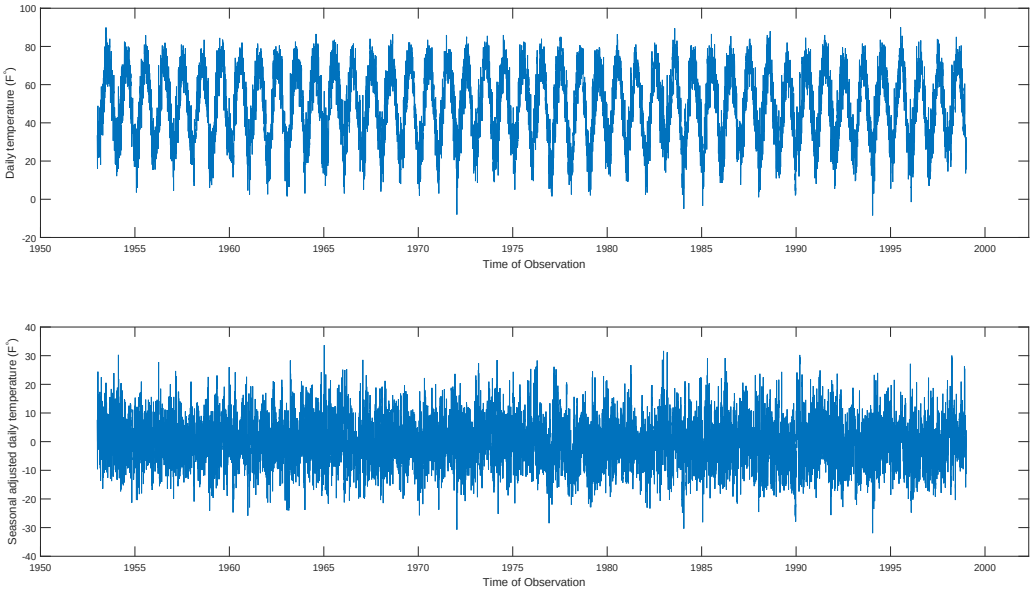

Figure 3.7: The Benton Harbor daily average temperature series.

nual target-reference series. While 13 changepoints were found in the daily series, only 6 changepoints were flagged in the annual analysis. One sees the extra precision induced by examining daily series.

3.6 Comments

This paper modified the Bayesian MDL techniques of Li et al. (2015) to accommodate daily temperature series. A Bayesian MDL score was minimized to estimate the best changepoint configuration. An island version of a GA was used as an numerical optimization tool. The MDL score here accounts for trends, seasonal means, autocorrelation, and seasonal variabilities. Identifying changepoints in daily data is challenging due to the long series length and the large number of model parameters.

The mean shift magnitudes in our model are non-seasonal — the mean shift

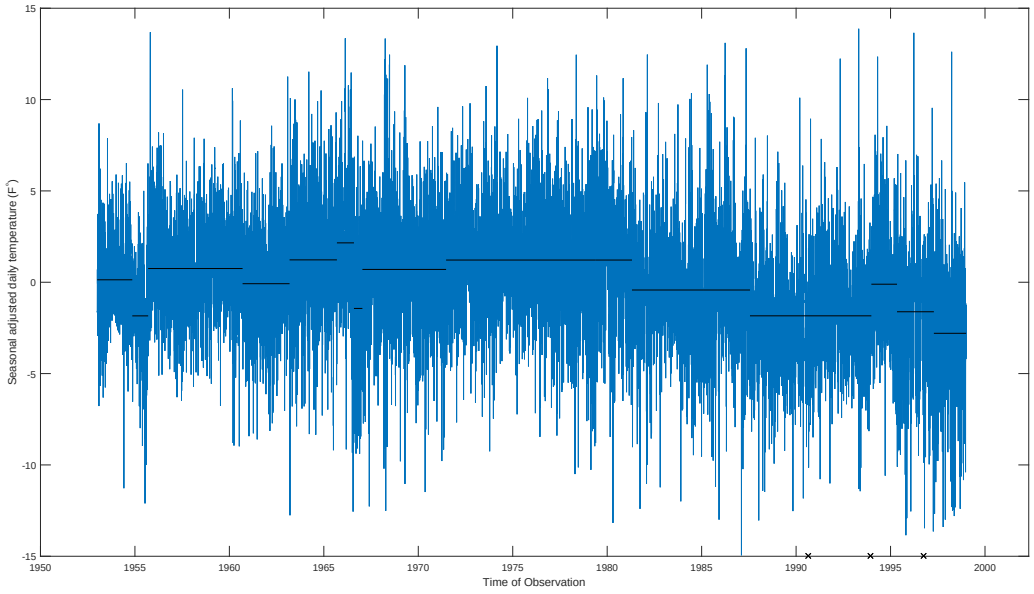

Figure 3.8: The South Haven minus the Benton Harbor series. The estimated change-point structure is superimposed on the graph and reveals 13 mean shifts of interest.

changes temperature on all days the same amount. Should one expect a seasonal mean shift structure (say with winter shifts being larger than summer shift), this could be allowed in the modeling procedure, although it would take significant work to accommodate such a structure.

While our study examined temperature series, our methods can be applied to other climatic series with a non-Gaussian likelihood. For example, Poisson-based likelihoods could be used for count series such as the monthly number of snow or thunderstorm days. While this research only considers univariate series, the methods could be modified to analyze multiple daily series as in Li et al. (2015).

Further improvements in computational speed of the algorithm are possible by further tinkering with the GA parameters. Such methods would be desirable to apply the methods to all $\binom{n}{2}$ pairwise differences in a network of n temperature series.

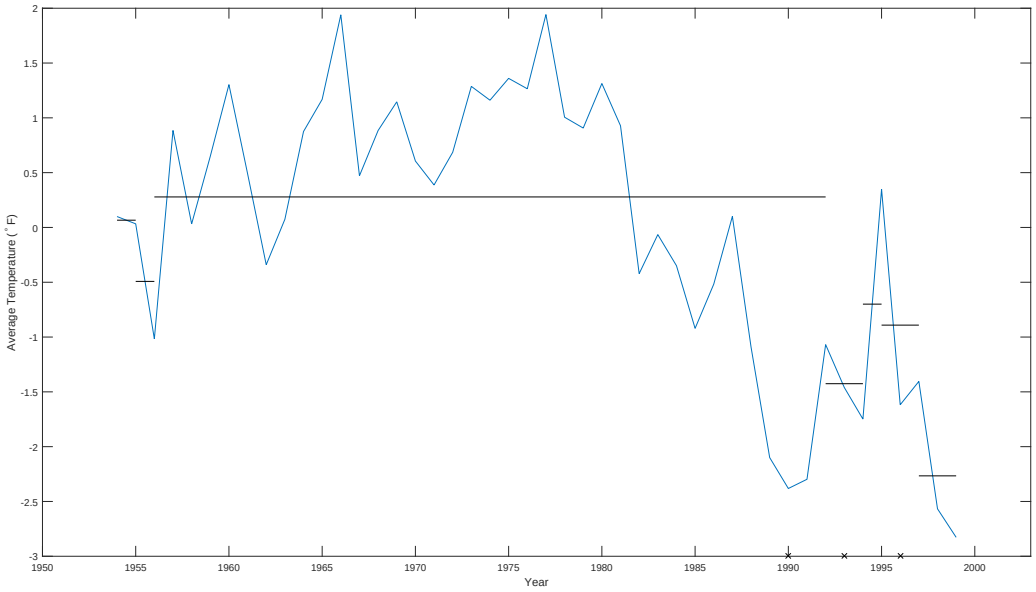

Figure 3.9: Annual South Haven minus Benton Harbor series with optimal change-point configuration superimposed.

Markov Chain Monte Carlo methods could also be used to help identify the optimal model; these techniques are developed in Li et al. (2015).

Appendices

Appendix A Proof of Theorem 2.4.1

A.1 Asymptotic behavior of the Yule-Walker estimator $\hat{\phi}$

To prove Theorem 2.4.1, the asymptotic limit of the Yule-Walker estimator in (2.34) is needed. For a sample size N, the observations obey the true changepoint model $\boldsymbol{\lambda}^0$ in (2.3):

$$\mathbf{X} = \mathbf{A}\mathbf{s} + \mathbf{D}^0\boldsymbol{\mu}^0 + \boldsymbol{\epsilon}. \tag{5}$$

For notation, the symbols $\mathbf{s}, \sigma^2, \boldsymbol{\phi}$ refer to the true parameters in $\boldsymbol{\lambda}^0$. Moreover, the subscript $1:N$ is omitted wherever there is no ambiguity. In (5), $\boldsymbol{\epsilon}$ is a zero-mean causal AR(p) series as formulated in (2.4).

For any relative changepoint configuration (model) $\boldsymbol{\lambda}$, suppose that $\boldsymbol{\eta}$ is the corresponding changepoint configuration under the sample size N. From (2.32), the ordinary least squares residual vector $\boldsymbol{\epsilon}^{\text{ols}}$ of the linear model in (2.3) is

$$\begin{aligned} \boldsymbol{\epsilon}^{\text{ols}} &= (\mathbf{I}_N - \mathcal{P}_{[\mathbf{A}\ \mathbf{D}]})\mathbf{X} \\ &= (\mathbf{I}_N - \mathcal{P}_{[\mathbf{A}\ \mathbf{D}]})(\mathbf{A}\mathbf{s} + \mathbf{D}^0\boldsymbol{\mu}^0 + \boldsymbol{\epsilon}) \\ &= (\mathbf{I}_N - \mathcal{P}_{[\mathbf{A}\ \mathbf{D}]})(\mathbf{D}^0\boldsymbol{\mu}^0 + \boldsymbol{\epsilon}). \end{aligned} \tag{6}$$

Here, the regime matrix $\mathbf{D}$ depends on $\boldsymbol{\eta}$ and may not necessarily equal $\mathbf{D}^0$.

Lemma A.1. *For each relative changepoint configuration $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ and $t \in \{1, 2, \ldots, N\}$, when N is large, each entry of $\boldsymbol{\epsilon}^{ols}$ can be expressed as $\epsilon_t^{ols} = \delta_t + W_t$, where*

$$\delta_t = \mu^0_{r^0(t)} - \bar{\mu}^0_{r(t)}, \quad W_t = \epsilon_t - \bar{\epsilon}_{r(t)} - \bar{\epsilon}_{v(t)} + \bar{\epsilon}. \tag{7}$$

Here, in regime ℓ of the changepoint configuration $\boldsymbol{\lambda}$, $\bar{\mu}_\ell^0 = (N_\ell)^{-1}\sum_{t\in\mathcal{R}_\ell}\mu_t^0$ is the average of the true mean parameters, N_ℓ is the number of time points in this regime, and $\mathcal{R}_\ell$ is the set of all time points in this regime. Likewise, $\bar{\epsilon}_\ell$ is the average of errors in regime ℓ, $\bar{\epsilon}_v$ is the average of errors in season v, and $\bar{\epsilon}$ is the average of all errors.

Proof. Because of (6), our main effort is to study the projection residual $\mathbf{I}_N - \mathcal{P}_{[\mathbf{A}\ \mathbf{D}]}$ under large N. Since the two column spaces spanned by $(\mathbf{I}_N - \mathcal{P}_{\mathbf{D}})\mathbf{A}$ and $\mathbf{D}$ are perpendicular, Theorem B.45 in Christensen (2002, pp. 411) gives $\mathcal{P}_{[(\mathbf{I}_N-\mathcal{P}_{\mathbf{D}})\mathbf{A}\ \mathbf{D}]} = \mathcal{P}_{(\mathbf{I}_N-\mathcal{P}_{\mathbf{D}})\mathbf{A}} + \mathcal{P}_{\mathbf{D}}$. Therefore,

$$\mathbf{I}_N - \mathcal{P}_{[\mathbf{A}\ \mathbf{D}]} = \mathbf{I}_N - \mathcal{P}_{[(\mathbf{I}_N-\mathcal{P}_{\mathbf{D}})\mathbf{A}\ \mathbf{D}]} = \mathbf{I}_N - \mathcal{P}_{(\mathbf{I}_N-\mathcal{P}_{\mathbf{D}})\mathbf{A}} - \mathcal{P}_{\mathbf{D}}. \tag{8}$$

Here, the term $\mathcal{P}_{(\mathbf{I}_N-\mathcal{P}_{\mathbf{D}})\mathbf{A}}$ is expanded as

$$\mathcal{P}_{(\mathbf{I}_N-\mathcal{P}_{\mathbf{D}})\mathbf{A}} = (\mathbf{I}_N - \mathcal{P}_{\mathbf{D}})\mathbf{A}\left[\mathbf{A}'(\mathbf{I}_N - \mathcal{P}_{\mathbf{D}})\mathbf{A}\right]^{-1}\mathbf{A}'(\mathbf{I}_N - \mathcal{P}_{\mathbf{D}}). \tag{9}$$

For any $n \in \mathbb{N}$, let $\mathbf{0}_n$ be the n-dimensional vector containing all zero entries, $\mathbf{1}_n$ be the n-dimensional vector containing whose entries are all unity, and $\mathbf{J}_n$ as the $n \times n$ matrix whose entries are all unity, i.e., $\mathbf{J}_n = \mathbf{1}_n\mathbf{1}_n'$.

For $v \in \{1, 2, \ldots, T\}$, suppose there are $k(v, \ell)$ time points in regime ℓ that are also in season v. Equation (2.43) shows that N_ℓ increases linearly with N; hence, so does $k(v, \ell)$. Moreover, when N is large, inside each regime, the seasonal counts $k(v, \ell)$ are equal except for edge effects, i.e., $k(v, \ell)/N_\ell \approx 1/T$ for all seasons v. We will ignore these edge effects in the ensuing calculations. Proceeding under this simplification,

the vth column in $\mathbf{A}$, denoted by $\mathbf{A}_v$, under the projection $\mathcal{P}_{\mathcal{D}}$, becomes

$$\mathcal{P}_{\mathcal{D}}\mathbf{A}_v = \left(\mathbf{0}'_{N_1}, \frac{k(v,2)}{N_2}\mathbf{1}'_{N_2}, \ldots, \frac{k(v,m+1)}{N_{m+1}}\mathbf{1}'_{N_{m+1}}\right)' = \left(\mathbf{0}'_{N_1}, \frac{1}{T}\mathbf{1}'_{N-N_1}\right)'. \tag{10}$$

We can now obtain an expression for $\mathbf{A}'(\mathbf{I}_N - \mathcal{P}_{\mathbf{D}})\mathbf{A}$. To do this, for $u, w \in \{1, 2, \ldots, T\}$,

$$\begin{aligned}
[\mathbf{A}'(\mathbf{I}_N - \mathcal{P}_{\mathbf{D}})\mathbf{A}]_{u,w} &= \mathbf{A}'_u\mathbf{A}_w - (\mathcal{P}_{\mathcal{D}}\mathbf{A}_u)'(\mathcal{P}_{\mathcal{D}}\mathbf{A}_w) \\
&= \begin{cases} \frac{N}{T^2}(T - (1-\lambda_1)), & \text{if } u = w, \\ -\frac{N}{T^2}(1-\lambda_1), & \text{if } u \neq w, \end{cases}
\end{aligned}$$

and it follows that $\mathbf{A}'(\mathbf{I}_N - \mathcal{P}_{\mathbf{D}})\mathbf{A} = NT^{-2}(T\mathbf{I}_T - (1-\lambda_1)\mathbf{J}_T)$. The inverse of this matrix can be verified as

$$[\mathbf{A}'(\mathbf{I}_N - \mathcal{P}_{\mathbf{D}})\mathbf{A}]^{-1} = \frac{1}{N}\left(T\mathbf{I}_T + \frac{1-\lambda_1}{\lambda_1}\mathbf{J}_T\right).$$

Plugging the inverse into (9) and denoting $\mathcal{Q}_{\mathbf{D}} = \mathbf{I}_N - \mathcal{P}_{\mathbf{D}}$ give

$$\begin{aligned}
\mathcal{P}_{(\mathbf{I}_N - \mathcal{P}_{\mathbf{D}})\mathbf{A}} &= \frac{1}{N}(\mathcal{Q}_{\mathbf{D}}\mathbf{A})\left(T\mathbf{I}_T + \frac{1-\lambda_1}{\lambda_1}\mathbf{J}_T\right)(\mathcal{Q}_{\mathbf{D}}\mathbf{A})' \\
&= \frac{T}{N}(\mathcal{Q}_{\mathbf{D}}\mathbf{A})(\mathcal{Q}_{\mathbf{D}}\mathbf{A})' + \frac{1-\lambda_1}{N\lambda_1}(\mathcal{Q}_{\mathbf{D}}\mathbf{A}\mathbf{1}_T)(\mathcal{Q}_{\mathbf{D}}\mathbf{A}\mathbf{1}_T)'.
\end{aligned} \tag{11}$$

For simplicity, we assume that regime ℓ starts with season 1, ends with season T, and contains n_ℓ full cycles. Using $n = \sum_{r=1}^{m+1} n_r$ and (10) gives

$$\mathcal{Q}_{\mathbf{D}}\mathbf{A} = \left(\begin{array}{c} \mathbf{1}_{n_1} \otimes \mathbf{I}_T \\ \hdashline \mathbf{1}_{n-n_1} \otimes \left(\mathbf{I}_T - \frac{1}{T}\mathbf{J}_T\right) \end{array}\right), \quad \mathcal{Q}_{\mathbf{D}}\mathbf{A}\mathbf{1}_T = \left(\begin{array}{c} \mathbf{1}_{N_1} \\ \hdashline \mathbf{0}_{N-N_1} \end{array}\right).$$

Hence, quadratic forms of these matrices are

$$(\mathcal{Q}_{\mathbf{D}}\mathbf{A})(\mathcal{Q}_{\mathbf{D}}\mathbf{A})' = \left(\begin{array}{c:c} \mathbf{J}_{n_1} \otimes \mathbf{I}_T & \mathbf{J}_{n_1\times(n-n_1)} \otimes \left(\mathbf{I}_T - \frac{1}{T}\mathbf{J}_T\right) \\ \hdashline \mathbf{J}_{(n-n_1)\times n_1} \otimes \left(\mathbf{I}_T - \frac{1}{T}\mathbf{J}_T\right) & \mathbf{J}_{n-n_1} \otimes \left(\mathbf{I}_T - \frac{1}{T}\mathbf{J}_T\right) \end{array}\right),$$

and

$$(\mathcal{Q}_{\mathbf{D}}\mathbf{A}\mathbf{1}_T)(\mathcal{Q}_{\mathbf{D}}\mathbf{A}\mathbf{1}_T)' = \left(\begin{array}{c:c} \mathbf{J}_{N_1} & \mathbf{0} \\ \hdashline \mathbf{0} & \mathbf{0} \end{array}\right).$$

Plugging these into (11) produces

$$\mathcal{P}_{(\mathbf{I}_N - \mathcal{P}_{\mathbf{D}})\mathbf{A}} = \frac{1}{N_1}\left(\begin{array}{c:c} \mathbf{J}_{N_1} & \mathbf{0} \\ \hdashline \mathbf{0} & \mathbf{0} \end{array}\right) + \frac{T}{N}\mathbf{J}_n \otimes \mathbf{I}_T - \frac{1}{N}\mathbf{J}_N.$$

Since $\mathcal{P}_{\mathbf{D}}$ is block-diagonal of form

$$\mathcal{P}_{\mathbf{D}} = \operatorname{diag}\left(\mathbf{0}_{N_1\times N_1}, \frac{\mathbf{J}_{N_2}}{N_2}, \ldots, \frac{\mathbf{J}_{N_{m+1}}}{N_{m+1}}\right),$$

we have

$$\mathbf{I}_N - \mathcal{P}_{[\mathbf{A}\ \mathbf{D}]} = \mathbf{I}_N - \operatorname{diag}\left(\frac{\mathbf{J}_{N_1}}{N_1}, \frac{\mathbf{J}_{N_2}}{N_2}, \ldots, \frac{\mathbf{J}_{N_{m+1}}}{N_{m+1}}\right) - \frac{T}{N}\mathbf{J}_n \otimes \mathbf{I}_T + \frac{1}{N}\mathbf{J}_N.$$

Therefore, for $t \in \{1, 2, \ldots, N\}$, the tth entries of the vectors in (6) are

$$\begin{aligned} W_t &= [(\mathbf{I}_N - \mathcal{P}_{[\mathbf{A}\ \mathbf{D}]})\boldsymbol{\epsilon}]_t = \epsilon_t - \bar{\epsilon}_{r(t)} - \bar{\epsilon}_{v(t)} + \bar{\epsilon}, \\ \delta_t &= [(\mathbf{I}_N - \mathcal{P}_{[\mathbf{A}\ \mathbf{D}]})\mathbf{D}^0\boldsymbol{\mu}^0]_t = \mu^0_{r^0(t)} - \bar{\mu}^0_{r(t)}. \end{aligned}$$

□

It is not hard to see that $\delta_t = 0$ for all $t = 1, 2, \ldots, N$ if and only if all relative changepoints in $\boldsymbol{\lambda}^0$ are contained in $\boldsymbol{\lambda}$ (denoted by $\boldsymbol{\lambda} \supset \boldsymbol{\lambda}^0$). For any changepoint configuration $\boldsymbol{\lambda}$, as N tends to infinity, the average $N^{-1}\sum_{t=h+1}^{N} \delta_t\delta_{t-h}$ converges to a constant that does not depend on the lag $h \in \{0, 1, \ldots, p\}$. This is because for any lag h, $\delta_t = \delta_{t-h}$ for all $t \in \{1, 2, \ldots, N\}$, except for at most $(m + m^0)h \leq (m + m^0)p$ times near the changepoints in $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}^0$. Hence, as $N \to \infty$, $N^{-1}\sum_{t=h+1}^{N} \delta_t\delta_{t-h}$ converges to its limit at rate $O(1/N)$. We denote this limit as

$$\delta^2 \stackrel{\text{def}}{=} \lim_{N\to\infty} \frac{1}{N} \sum_{t=h+1}^{N} \delta_t\delta_{t-h}. \tag{12}$$

Since (12) holds for $h = 0$, $\delta^2 \geq 0$.

To quantify the asymptotic limit of the Yule-Walker estimator $\hat{\boldsymbol{\phi}}$, let $\boldsymbol{\gamma}_p = (\gamma(1), \gamma(2), \ldots, \gamma(p))'$ and $\boldsymbol{\Gamma}_p$ be a $p \times p$ matrix with (i, j)th entry $\gamma(|i - j|)$.

Proposition A.1. *Under the relative changepoint configuration $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ (which may or may not be the true changepoint configuration), for $h \in \{0, 1, \ldots, p\}$, as $N \to \infty$, the lag-h sample autocovariance obeys*

$$\hat{\gamma}(h) = \gamma(h) + \delta^2 + O_P\left(\frac{1}{N}\right), \tag{13}$$

and the Yule-Walker estimator $\hat{\boldsymbol{\phi}} = \hat{\boldsymbol{\Gamma}}_p^{-1}\hat{\boldsymbol{\gamma}}_p$ *obeys*

$$\hat{\boldsymbol{\phi}} = \left(\boldsymbol{\Gamma}_p + \delta^2 \mathbf{J}_p\right)^{-1}\left(\boldsymbol{\gamma}_p + \delta^2 \mathbf{1}_p\right) + O_P\left(\frac{1}{N}\right). \tag{14}$$

Moreover, if and only if $\boldsymbol{\lambda} \supset \boldsymbol{\lambda}^0$, $\delta^2 = 0$ *and* $\hat{\boldsymbol{\phi}} \to \boldsymbol{\phi}$ *as* $N \to \infty$.

Proof. Since the AR(p) errors are assumed causal, we may write

$$\epsilon_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} \tag{15}$$

for some weights $\{\psi_j\}_{j=0}^{\infty}$, where $\sum_{j=0}^{\infty} |\psi_j| < \infty$. Since $W_t = \epsilon_t - \bar{\epsilon}_{r(t)} - \bar{\epsilon}_{v(t)} + \bar{\epsilon}$, one can write W_t as a linear combination of all Z_ts up to and before time N:

$$W_t = \sum_{j=-\infty}^{\infty} \psi_j^{(t)} Z_{t-j},$$

where

$$\psi_j^{(t)} = \psi_j - \frac{\sum_{k:r(k)=r(t)} \psi_{k-t+j}}{N_{r(t)}} - \frac{\sum_{l:v(l)=v(t)} \psi_{l-t+j}}{N/T} + \frac{\sum_{u=1}^{N} \psi_{u-t+j}}{N}. \tag{16}$$

Here, $\psi_j = 0$ when $j < 0$, implying that $\psi_j^{(t)} = 0$ if $j < t - N$.

The asymptotic limit of the sample autocovariances can now be derived:

$$\begin{aligned}\hat{\gamma}(h) &= \frac{1}{N}\sum_{t=h+1}^{N} \epsilon_t^{\text{ols}}\epsilon_{t-h}^{\text{ols}} \\ &= \frac{1}{N}\sum_{t=h+1}^{N} (W_t + \delta_t)(W_{t-h} + \delta_{t-h}) \\ &= \frac{1}{N}\sum_{t=h+1}^{N} (W_t W_{t-h} + \delta_{t-h} W_t + \delta_t W_{t-h} + \delta_t \delta_{t-h}).\end{aligned} \tag{17}$$

Arguing as in Proposition 7.3.5 of Brockwell and Davis (1991, pp. 232) gives

$$\frac{1}{N}\sum_{t=h+1}^{N} W_t W_{t-h} = \frac{1}{N}\sum_{t=h+1}^{N}\sum_{j=-\infty}^{\infty} \psi_j^{(t)}\psi_{j-h}^{(t-h)} Z_{t-j}^2 + O_P\left(\frac{1}{N}\right). \tag{18}$$

From $\sum_{j=0}^{\infty}|\psi_j| < \infty$ and (16), it is not hard to show that $\left\{\psi_j^{(t)}\psi_{j-h}^{(t-h)}\right\}_{j=-\infty}^{\infty}$ is absolutely convergent for each t and h. Since $\{Z_t\}$ is IID, with $E[Z_t^2] = \sigma^2$, the weak law of large numbers (WLLN) for linear processes (Brockwell and Davis, 1991, pp. 208, Proposition 6.3.10) gives

$$\frac{1}{N}\sum_{t=h+1}^{N} W_t W_{t-h} = \frac{1}{N}\sum_{t=h+1}^{N}\sum_{j=-\infty}^{\infty} \psi_j^{(t)}\psi_{j-h}^{(t-h)}\sigma^2 + O_P\left(\frac{1}{N}\right).$$

From (16), one can show that

$$\sup_t \sum_{\ell=-\infty}^{\infty} |\psi_\ell^{(t-h)} - \psi_\ell^{(t)}| < \infty.$$

Hence,

$$\frac{1}{N}\sum_{t=h+1}^{N} W_t W_{t-h} = \frac{1}{N}\sum_{t=h+1}^{N}\sum_{j=-\infty}^{\infty} \psi_j\psi_{j-h}\sigma^2 + O_P\left(\frac{1}{N}\right).$$

Now using that $\gamma(h) = \sigma^2\sum_{\ell=-\infty}^{\infty}\psi_\ell\psi_{\ell-h}$ gives

$$\frac{1}{N}\sum_{t=h+1}^{N} W_t W_{t-h} = \frac{N-h}{N}\gamma(h) + O_P\left(\frac{1}{N}\right) = \gamma(h) + O_P\left(\frac{1}{N}\right).$$

This identifies the limit of the first term in the bottom line of (17). For the second and third terms, apply the WLLN again to see that these terms converge to zero in probability at rate $O_P(1/N)$. Hence, as $N \to \infty$,

$$\hat{\gamma}(h) = \gamma(h) + \frac{1}{N}\sum_{t=h+1}^{N} \delta_t \delta_{t-h} + O_P\left(\frac{1}{N}\right) = \gamma(h) + \delta^2 + O_P\left(\frac{1}{N}\right),$$

which proves (14). □

A.2 Proof of asymptotic consistency of the univariate BMDL

To simplify the BMDL formulae in (2.37) and (2.38), we first establish some asymptotic results for its terms.

Lemma A.2. *Under any changepoint configuration* $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ *with* $m > 0$*, as* $N \to \infty$*,*

$$\frac{1}{N}\widehat{\mathbf{X}}'\left[\widehat{\mathbf{B}} - \widehat{\mathbf{B}}\widehat{\mathbf{A}}\left(\widehat{\mathbf{A}}'\widehat{\mathbf{B}}\widehat{\mathbf{A}}\right)^{-1}\widehat{\mathbf{A}}'\widehat{\mathbf{B}}\right]\widehat{\mathbf{X}} = \frac{1}{N}\widehat{\mathbf{X}}'\left(\mathbf{I}_N - \mathcal{P}_{[\widehat{\mathbf{A}}\ \widehat{\mathbf{D}}]}\right)\widehat{\mathbf{X}} + O_P\left(\frac{1}{N}\right); \quad (19)$$

for the model with $m = 0$*,*

$$\frac{1}{N}\widehat{\mathbf{X}}'\left[\mathbf{I}_N - \widehat{\mathbf{A}}\left(\widehat{\mathbf{A}}'\widehat{\mathbf{A}}\right)^{-1}\widehat{\mathbf{A}}'\right]\widehat{\mathbf{X}} = \frac{1}{N}\widehat{\mathbf{X}}'\left(\mathbf{I}_N - \mathcal{P}_{[\widehat{\mathbf{A}}\ \widehat{\mathbf{D}}]}\right)\widehat{\mathbf{X}}. \quad (20)$$

Furthermore, under any changepoint model $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$*,*

$$\frac{1}{N}\widehat{\mathbf{X}}'\left(\mathbf{I}_N - \mathcal{P}_{[\widehat{\mathbf{A}}\ \widehat{\mathbf{D}}]}\right)\widehat{\mathbf{X}} = \hat{\gamma}(0) - \hat{\boldsymbol{\gamma}}_p'\hat{\boldsymbol{\Gamma}}_p^{-1}\hat{\boldsymbol{\gamma}}_p + O_P\left(\frac{1}{N}\right). \tag{21}$$

Finally, as a function of δ^2*,*

$$f(\delta^2) \stackrel{def}{=} \hat{\gamma}(0) - \hat{\boldsymbol{\gamma}}_p'\hat{\boldsymbol{\Gamma}}_p^{-1}\hat{\boldsymbol{\gamma}}_p \tag{22}$$

is strictly increasing.

Proof. Under any changepoint configuration $\boldsymbol{\lambda}$ with $m > 0$, we will first show that (19) holds. Since $\hat{\boldsymbol{\phi}}$ has the limit in (14), it is not hard to show that as N tends to infinity, $\widehat{\mathbf{D}}'\widehat{\mathbf{D}}/N$ and $\widehat{\mathbf{D}}'\widehat{\mathbf{X}}/N$ converges in probability to a $m \times m$ positive definite matrix and an m-dimensional vector, respectively, both at rates $O_P(1/N)$. In the prior of $\boldsymbol{\mu}$, the parameter ν is a constant; hence,

$$\begin{aligned}
\frac{1}{N}\widehat{\mathbf{X}}'\widehat{\mathbf{B}}\widehat{\mathbf{X}} =& \frac{1}{N}\widehat{\mathbf{X}}'\left[\mathbf{I}_{N-p} - \widehat{\mathbf{D}}\left(\widehat{\mathbf{D}}'\widehat{\mathbf{D}} + \frac{\mathbf{I}_m}{\nu}\right)^{-1}\widehat{\mathbf{D}}'\right]\widehat{\mathbf{X}} \\
=& \frac{\widehat{\mathbf{X}}'\widehat{\mathbf{X}}}{N} - \frac{\widehat{\mathbf{X}}'\widehat{\mathbf{D}}}{N}\left(\frac{\widehat{\mathbf{D}}'\widehat{\mathbf{D}}}{N} + \frac{\mathbf{I}_m}{N\nu}\right)^{-1}\frac{\widehat{\mathbf{D}}'\widehat{\mathbf{X}}}{N} \\
=& \frac{\widehat{\mathbf{X}}'\widehat{\mathbf{X}}}{N} - \frac{\widehat{\mathbf{X}}'\widehat{\mathbf{D}}}{N}\left(\frac{\widehat{\mathbf{D}}'\widehat{\mathbf{D}}}{N}\right)^{-1}\frac{\widehat{\mathbf{D}}'\widehat{\mathbf{X}}}{N} + O_P\left(\frac{1}{N}\right) \\
=& \frac{1}{N}\widehat{\mathbf{X}}'\left[\mathbf{I}_{N-p} - \widehat{\mathbf{D}}\left(\widehat{\mathbf{D}}'\widehat{\mathbf{D}}\right)^{-1}\widehat{\mathbf{D}}'\right]\widehat{\mathbf{X}} + O_P\left(\frac{1}{N}\right) \\
=& \frac{1}{N}\widehat{\mathbf{X}}'\left(\mathbf{I}_{N-p} - \mathcal{P}_{\widehat{\mathbf{D}}}\right)\widehat{\mathbf{X}} + O_P\left(\frac{1}{N}\right).
\end{aligned}$$

Similar arguments give

$$\frac{1}{N}\widehat{\mathbf{X}}'\widehat{\mathbf{B}}\widehat{\mathbf{A}} = \frac{1}{N}\widehat{\mathbf{X}}'\left(\mathbf{I}_{N-p} - \mathcal{P}_{\widehat{\mathbf{D}}}\right)\widehat{\mathbf{A}} + O_P\left(\frac{1}{N}\right),$$
$$\frac{1}{N}\widehat{\mathbf{A}}'\widehat{\mathbf{B}}\widehat{\mathbf{A}} = \frac{1}{N}\widehat{\mathbf{A}}'\left(\mathbf{I}_{N-p} - \mathcal{P}_{\widehat{\mathbf{D}}}\right)\widehat{\mathbf{A}} + O_P\left(\frac{1}{N}\right).$$

Hence, the left hand side of (19) has the limit

$$\begin{aligned}
&\frac{1}{N}\widehat{\mathbf{X}}'\left[\widehat{\mathbf{B}} - \widehat{\mathbf{B}}\widehat{\mathbf{A}}\left(\widehat{\mathbf{A}}'\widehat{\mathbf{B}}\widehat{\mathbf{A}}\right)^{-1}\widehat{\mathbf{A}}'\widehat{\mathbf{B}}\right]\widehat{\mathbf{X}} \\
&= \frac{1}{N}\widehat{\mathbf{X}}'\left[\mathbf{I} - \mathcal{P}_{\widehat{\mathbf{D}}} - \mathcal{P}_{(\mathbf{I}_{N-p}-\mathcal{P}_{\widehat{\mathbf{D}}})\widehat{\mathbf{A}}}\right]\widehat{\mathbf{X}} + O_P\left(\frac{1}{N}\right) \\
&= \frac{1}{N}\widehat{\mathbf{X}}'\left(\mathbf{I}_{N-p} - \mathcal{P}_{[\widehat{\mathbf{A}}\ \widehat{\mathbf{D}}]}\right)\widehat{\mathbf{X}} + O_P\left(\frac{1}{N}\right),
\end{aligned}$$

where the last equality follows from (8).

Next, we will show that (21) holds for any $\boldsymbol{\lambda}$ with $m > 0$. For notational simplicity, for any $j \in \{0, 1, \ldots, p\}$, matrices formed by the rows of $\mathbf{A}$ and $\mathbf{D}$ are denoted by

$$\mathbf{A}_j \stackrel{\text{def}}{=} \mathbf{A}_{(p+1-j):(N-j)}, \quad \mathbf{D}_j \stackrel{\text{def}}{=} \mathbf{D}_{(p+1-j):(N-j)}.$$

Since both $\widehat{\mathbf{A}}$ and $\mathbf{A}_j$ are $(N-p) \times T$ matrices and each column in $\widehat{\mathbf{A}}$ can be written as a linear combination of the columns in $\mathbf{A}_j$, the corresponding column spaces agree: $C(\widehat{\mathbf{A}}) = C(\mathbf{A}_j)$. Therefore, $\mathcal{P}_{\widehat{\mathbf{A}}} = \mathcal{P}_{\mathbf{A}_j}$ for $j \in \{1, \ldots, p\}$. Now define

$$\boldsymbol{\Delta}_j = \mathbf{D}_j - \frac{\widehat{\mathbf{D}}}{1 - \hat{\phi}_1 - \hat{\phi}_2 - \cdots - \hat{\phi}_p}. \tag{23}$$

The denominator in (23) cannot be zero since a$1 - \sum_{k=1}^{p}\hat{\phi}_k \neq 0$ for any Yule-Walker

estimates (Brockwell and Davis, 1991).

Since there are at most $2m(p+h)$ non-zero entries in $\boldsymbol{\Delta}_j$, and none of these entries depend on N, $\boldsymbol{\Delta}_j'\boldsymbol{\Delta}_j = O_P(1)$. In addition, for any N-dimensional vectors $\boldsymbol{\alpha}$ whose entries do not depend on N, $\boldsymbol{\alpha}'\boldsymbol{\Delta}_j = O_P(1)$. Using (23),

$$\begin{aligned}\frac{\widehat{\mathbf{D}}'\left(\mathbf{I}_N - \mathcal{P}_{\widehat{\mathbf{A}}}\right)\widehat{\mathbf{D}}}{N\left(1-\sum_{k=1}^{p}\hat{\phi}_k\right)^2} &= \frac{1}{N}(\mathbf{D}_j - \boldsymbol{\Delta}_j)'\left(\mathbf{I}_N - \mathcal{P}_{\widehat{\mathbf{A}}}\right)(\mathbf{D}_j - \boldsymbol{\Delta}_j) \\ &= \frac{\mathbf{D}_j'\left(\mathbf{I}_N - \mathcal{P}_{\widehat{\mathbf{A}}}\right)\mathbf{D}_j}{N} + O_P\left(\frac{1}{N}\right),\end{aligned}$$

and

$$\begin{aligned}\frac{\boldsymbol{\alpha}'\left(\mathbf{I}_N - \mathcal{P}_{\widehat{\mathbf{A}}}\right)\widehat{\mathbf{D}}}{N\left(1-\sum_{k=1}^{p}\hat{\phi}_k\right)} &= \frac{1}{N}\boldsymbol{\alpha}'\left(\mathbf{I}_N - \mathcal{P}_{\widehat{\mathbf{A}}}\right)(\mathbf{D}_j - \boldsymbol{\Delta}_j) \\ &= \frac{\boldsymbol{\alpha}'\left(\mathbf{I}_N - \mathcal{P}_{\widehat{\mathbf{A}}}\right)\mathbf{D}_j}{N} + O_P\left(\frac{1}{N}\right).\end{aligned}$$

Therefore, for any $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^N$ whose entries do not depend on N,

$$\begin{aligned}&\frac{1}{N}\boldsymbol{\alpha}'\mathcal{P}_{\left(\mathbf{I}_N - \mathcal{P}_{\widehat{\mathbf{A}}}\right)\widehat{\mathbf{D}}}\boldsymbol{\beta} \\ &= \frac{\boldsymbol{\alpha}'\left(\mathbf{I}_N - \mathcal{P}_{\widehat{\mathbf{A}}}\right)\widehat{\mathbf{D}}}{N\left(1-\sum_{k=1}^{p}\hat{\phi}_k\right)}\left(\frac{\widehat{\mathbf{D}}'\left(\mathbf{I}_N - \mathcal{P}_{\widehat{\mathbf{A}}}\right)\widehat{\mathbf{D}}}{N\left(1-\sum_{k=1}^{p}\hat{\phi}_k\right)^2}\right)^{-1}\frac{\widehat{\mathbf{D}}'\left(\mathbf{I}_N - \mathcal{P}_{\widehat{\mathbf{A}}}\right)\boldsymbol{\beta}}{N\left(1-\sum_{k=1}^{p}\hat{\phi}_k\right)} \\ &= \frac{1}{N}\boldsymbol{\alpha}'\left[\left(\mathbf{I}_N - \mathcal{P}_{\widehat{\mathbf{A}}}\right)\mathbf{D}_j\left(\mathbf{D}_j'\left(\mathbf{I}_N - \mathcal{P}_{\widehat{\mathbf{A}}}\right)\mathbf{D}_j\right)^{-1}\mathbf{D}_j'\left(\mathbf{I}_N - \mathcal{P}_{\widehat{\mathbf{A}}}\right)\right]\boldsymbol{\beta} + O_P\left(\frac{1}{N}\right) \\ &= \frac{1}{N}\boldsymbol{\alpha}'\mathcal{P}_{\left(\mathbf{I}_N - \mathcal{P}_{\widehat{\mathbf{A}}}\right)\mathbf{D}_j}\boldsymbol{\beta} + O_P\left(\frac{1}{N}\right).\end{aligned}$$

Hence, from (8),

$$\frac{1}{N}\boldsymbol{\alpha}'\mathcal{P}_{[\widehat{\mathbf{A}}\ \widehat{\mathbf{D}}]}\boldsymbol{\beta} = \frac{1}{N}\boldsymbol{\alpha}'\mathcal{P}_{[\mathbf{A}_j\ \mathbf{D}_j]}\boldsymbol{\beta} + O_P\left(\frac{1}{N}\right). \tag{24}$$

Since $\widehat{\mathbf{X}} = \mathbf{X}_{(p+1):N} - \sum_{j=1}^{p}\hat{\phi}_j\mathbf{X}_{(p+1-j):(N-j)}$, for any $j,k \in \{0,1,\ldots,p\}$, (24) shows that

$$\begin{aligned}
&\frac{1}{N}\mathbf{X}'_{(p+1-j):(N-j)}\left(\mathbf{I}_N - \mathcal{P}_{[\widehat{\mathbf{A}}\ \widehat{\mathbf{D}}]}\right)\mathbf{X}_{(p+1-k):(N-k)} \\
&= \frac{1}{N}\left[\left(\mathbf{I}_N - \mathcal{P}_{[\mathbf{A}_j\ \mathbf{D}_j]}\right)\mathbf{X}_{(p+1-j):(N-j)}\right]'\left[\left(\mathbf{I}_N - \mathcal{P}_{[\mathbf{A}_k\ \mathbf{D}_k]}\right)\mathbf{X}_{(p+1-k):(N-k)}\right] \\
&+ O_P\left(\frac{1}{N}\right) \\
&= \frac{1}{N}\left(\boldsymbol{\epsilon}^{\text{ols}}_{(p+1-j):(N-j)}\right)'\boldsymbol{\epsilon}^{\text{ols}}_{(p+1-k):(N-k)} + O_P\left(\frac{1}{N}\right).
\end{aligned}$$

Therefore, (21) can be further simplified as

$$\begin{aligned}
&\frac{1}{N}\widehat{\mathbf{X}}'\left(\mathbf{I}_N - \mathcal{P}_{[\widehat{\mathbf{A}}\ \widehat{\mathbf{D}}]}\right)\widehat{\mathbf{X}} \\
&= \frac{1}{N}\left[\boldsymbol{\epsilon}^{\text{ols}}_{(p+1):N} - \sum_{j=1}^{p}\hat{\phi}_j\boldsymbol{\epsilon}^{\text{ols}}_{(p+1-j):(N-j)}\right]'\left[\boldsymbol{\epsilon}^{\text{ols}}_{(p+1):N} - \sum_{k=1}^{p}\hat{\phi}_k\boldsymbol{\epsilon}^{\text{ols}}_{(p+1-k):(N-k)}\right] \\
&+ O_P\left(\frac{1}{N}\right) \\
&= \hat{\gamma}(0) - 2\sum_{j=1}^{p}\hat{\phi}_j\hat{\gamma}(j) + \sum_{j=1}^{p}\sum_{k=1}^{p}\hat{\phi}_j\hat{\phi}_k\hat{\gamma}(|j-k|) + O_P\left(\frac{1}{N}\right) \\
&= \hat{\gamma}(0) - 2\hat{\boldsymbol{\gamma}}'_p\hat{\boldsymbol{\phi}} + \hat{\boldsymbol{\phi}}'\hat{\boldsymbol{\Gamma}}_p\hat{\boldsymbol{\phi}} + O_P\left(\frac{1}{N}\right) \\
&= \hat{\gamma}(0) - \hat{\boldsymbol{\gamma}}'_p\hat{\boldsymbol{\Gamma}}_p^{-1}\hat{\boldsymbol{\gamma}}_p + O_P\left(\frac{1}{N}\right).
\end{aligned}$$

It is not hard to show that under the model $\boldsymbol{\lambda}_{\emptyset}$ ($m = 0$), because $C(\widehat{\mathbf{D}})$ is the null

space, (21) also holds.

Lastly, we show that $f(\delta^2)$ in (22) satisfies

$$\begin{aligned} f(\delta^2) &= \hat{\gamma}(0) - \hat{\boldsymbol{\gamma}}_p' \hat{\boldsymbol{\Gamma}}_p^{-1} \hat{\boldsymbol{\gamma}}_p \\ &= \gamma(0) + \delta^2 - \left(\boldsymbol{\gamma}_p + \delta^2 \mathbf{1}_p\right)' \left(\boldsymbol{\Gamma}_p + \delta^2 \mathbf{J}_p\right)^{-1} \left(\boldsymbol{\gamma}_p + \delta^2 \mathbf{1}_p\right) \end{aligned}$$

and is strictly increasing in δ^2.

If $\delta^2 > 0$, according to (2.22) in Harville (2008, pp. 428), for any matrices $\mathbf{R} \in \mathbb{R}^{r\times r}, \mathbf{S} \in \mathbb{R}^{r\times l}, \mathbf{T} \in \mathbb{R}^{l\times l}, \mathbf{U} \in \mathbb{R}^{l\times r}$ with $\mathbf{R}, \mathbf{U}$ non-singular,

$$(\mathbf{R} + \mathbf{STU})^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{S}(\mathbf{T}^{-1} + \mathbf{UR}^{-1}\mathbf{S})^{-1}\mathbf{UR}^{-1},$$

Hence,

$$\begin{aligned} \left(\boldsymbol{\Gamma}_p + \delta^2 \mathbf{J}_p\right)^{-1} &= \left(\boldsymbol{\Gamma}_p + \mathbf{1}_p \delta^2 \mathbf{1}_p'\right)^{-1} \\ &= \boldsymbol{\Gamma}_p^{-1} - \boldsymbol{\Gamma}_p^{-1}\mathbf{1}_p \left(\frac{1}{\delta^2} + \mathbf{1}_p'\boldsymbol{\Gamma}_p^{-1}\mathbf{1}_p\right)^{-1} \mathbf{1}_p'\boldsymbol{\Gamma}_p^{-1}. \end{aligned} \tag{25}$$

For notational simplicity, denote the following scalars by

$$a \stackrel{\text{def}}{=} \mathbf{1}_p'\boldsymbol{\Gamma}_p^{-1}\mathbf{1}_p, \quad b \stackrel{\text{def}}{=} \mathbf{1}_p'\boldsymbol{\Gamma}_p^{-1}\boldsymbol{\gamma}_p. \tag{26}$$

Then $f(\delta^2)$ can be expanded as

$$f(\delta^2) = \gamma(0) + \delta^2 - \boldsymbol{\gamma}_p'\boldsymbol{\Gamma}_p^{-1}\boldsymbol{\gamma}_p - 2b\delta^2 - a(\delta^2)^2 + \frac{b^2}{\frac{1}{\delta^2} + a} + \frac{2ab\delta^2}{\frac{1}{\delta^2} + a} + \frac{a^2(\delta^2)^2}{\frac{1}{\delta^2} + a}.$$

Differentiation of $f(\delta^2)$ with respect to δ^2 gives

$$\begin{aligned} f'(\delta^2) &= 1 - 2b - 2a\delta^2 + \frac{b^2 \frac{1}{(\delta^2)^2}}{\left(\frac{1}{\delta^2} + a\right)^2} + \frac{2ab\left(\frac{2}{\delta^2} + a\right)}{\left(\frac{1}{\delta^2} + a\right)^2} + \frac{a^2\left(3 + 2a\delta^2\right)}{\left(\frac{1}{\delta^2} + a\right)^2} \\ &= \frac{(b-1)^2}{(1 + a\delta^2)^2} > 0. \end{aligned}$$

The last inequality follows since $\{\epsilon_t\}_{t=1}^N$ is causal, which implies that $b = \sum_{k=1}^p \phi_k > 1$. Therefore, $f(\delta^2)$ is strictly increasing in δ^2.

□

The asymptotic consistency of the BMDL can now be proven.

Proof of Theorem 2.4.1. For any changepoint model $\boldsymbol{\lambda}$ (including the true model $\boldsymbol{\lambda}^0$), the proof of Lemma A.2 shows that as $N \to \infty$,

$$\frac{\widehat{\mathbf{D}}'\widehat{\mathbf{D}}}{N} + \frac{\mathbf{I}_m}{N\nu} = \frac{\widehat{\mathbf{D}}'\widehat{\mathbf{D}}}{N} + O\left(\frac{1}{N}\right) = \frac{\mathbf{D}'\mathbf{D}}{N\left(1 - \sum_{k=1}^p \hat{\phi}_k\right)^2} + O_P\left(\frac{1}{N}\right).$$

As the matrices are $m \times m$ (of finite dimension), the determinant of the limit converges to a number c:

$$\left|\frac{\widehat{\mathbf{D}}'\widehat{\mathbf{D}}}{N} + \frac{\mathbf{I}_m}{N\nu}\right| = c + O_P\left(\frac{1}{N}\right).$$

Here, c may depend on the model $\boldsymbol{\lambda}$. The c for the true model $\boldsymbol{\lambda}^0$ is denoted by c^0. Therefore, the asymptotic BMDL in (2.37) for the changepoint configuration $\boldsymbol{\lambda}$ with $m > 0$ is

$$\begin{aligned}\mathrm{BMDL}(\boldsymbol{\lambda}) = \frac{N-p}{2}\left\{\log N + \log\left[\frac{1}{N}\widehat{\mathbf{X}}'\left(\mathbf{I}_N - \mathcal{P}_{[\widehat{\mathbf{A}}\ \widehat{\mathbf{D}}]}\right)\widehat{\mathbf{X}} + O_P\left(\frac{1}{N}\right)\right]\right\} \\ + \frac{m}{2}\log\nu + \frac{1}{2}\log\left\{N^m\left[c + O_P\left(\frac{1}{N}\right)\right]\right\} \\ - \sum_{k=1}^{2}\log\left[\Gamma\left(a + m^{(k)}\right)\Gamma\left(b^{(k)} + N^{(k)} - m^{(k)}\right)\right].\end{aligned}$$

This convergence also holds for the null model $\boldsymbol{\lambda}_{\emptyset}$ in (2.38) (here, $c = 1$).

By (21) and (22), it now follows that the difference between BMDLs in a (non-true) model $\boldsymbol{\lambda}$ and the true model $\boldsymbol{\lambda}^0$ is asymptotically

$$\begin{aligned}&\mathrm{BMDL}\left(\boldsymbol{\lambda}\right) - \mathrm{BMDL}\left(\boldsymbol{\lambda}^0\right) \qquad (27)\\ &= \frac{N-p}{2}\log\left[\frac{f(\delta^2) + O_P\left(\frac{1}{N}\right)}{f(0) + O_P\left(\frac{1}{N}\right)}\right] + \frac{m-m^0}{2}(\log\nu + \log N) + \frac{1}{2}\log\left(\frac{c}{c^0}\right) \\ &\quad + \sum_{k=1}^{2}\log\left[\frac{\Gamma\left(a + m^{0(k)}\right)\Gamma\left(b^{(k)} + N^{(k)} - m^{0(k)}\right)}{\Gamma\left(a + m^{(k)}\right)\Gamma\left(b^{(k)} + N^{(k)} - m^{(k)}\right)}\right] + O_P\left(\frac{1}{N}\right).\end{aligned}$$

Since $\delta^2 = 0$ if and only if $\boldsymbol{\lambda} \supset \boldsymbol{\lambda}^0$, as $N \to \infty$, the first term of (27) is

$$\frac{N-p}{2}\log\left[\frac{f(\delta^2) + O_P\left(\frac{1}{N}\right)}{f(0) + O_P\left(\frac{1}{N}\right)}\right] = \begin{cases} O_P(N) > 0, & \text{if } \boldsymbol{\lambda} \not\supset \boldsymbol{\lambda}^0, \\ O_P(1), & \text{if } \boldsymbol{\lambda} \supset \boldsymbol{\lambda}^0. \end{cases}$$

Without loss of generality, the number of documented and undocumented changepoint times can be assumed to be increasing linearly with N — say $N^{(k)} = O(N)$, for $k \in \{1, 2\}$. Stirling's formula allows us to find the asymptotic limit of Gamma function ratios:

$$
\begin{aligned}
&\frac{\Gamma\left(b^{(k)}+N^{(k)}-m^{0(k)}\right)}{\Gamma\left(b^{(k)}+N^{(k)}-m^{(k)}\right)} \\
\approx\ & e^{m^{0(k)}-m^{(k)}}\ \frac{\left(b^{(k)}+N^{(k)}-m^{0(k)}-1\right)^{b^{(k)}+N^{(k)}-m^{0(k)}-1/2}}{\left(b^{(k)}+N^{(k)}-m^{(k)}-1\right)^{b^{(k)}+N^{(k)}-m^{(k)}-1/2}} \\
=\ & O\left(N^{m^{(k)}-m^{0(k)}}\right).
\end{aligned}
$$

Therefore, the last term of (27) is

$$
\sum_{k=1}^{2}\log\left[\frac{\Gamma\left(a+m^{0(k)}\right)\Gamma\left(b^{(k)}+N^{(k)}-m^{0(k)}\right)}{\Gamma\left(a+m^{(k)}\right)\Gamma\left(b^{(k)}+N^{(k)}-m^{(k)}\right)}\right]=(m-m^{0})\log N+\text{Const.}
$$

If $\boldsymbol{\lambda}\not\supset\boldsymbol{\lambda}^{0}$, the first term in (27) is asymptotically dominant:

$$
\text{BMDL}\left(\boldsymbol{\lambda}\right)-\text{BMDL}\left(\boldsymbol{\lambda}^{0}\right)=O_{P}(N)+(m-m^{0})\log N=O_{P}(N)>0.
$$

In contrast, if $\boldsymbol{\lambda}\supset\boldsymbol{\lambda}^{0}$, then since $m>m^{0}$,

$$
\text{BMDL}\left(\boldsymbol{\lambda}\right)-\text{BMDL}\left(\boldsymbol{\lambda}^{0}\right)=O_{P}(1)+(m-m^{0})\log N=O_{P}(\log N)>0.
$$

□

Table 1: Changepoints times and corresponding mean shifts.

Changepoint(τ_j)	Mean Shift at τ_j
1954-11-11	-1.9756
1955-09-14	2.5941
1960-09-13	-0.81925
1963-03-10	1.2962
1965-09-11	0.95114
1966-08-06	-3.6765
1967-01-16	2.2279
1971-06-22	0.51685
1981-04-25	-1.6267
1987-07-24	-1.40932
1993-12-26	1.77702
1995-05-06	-1.54974
1997-04-18	-1.18564

Bibliography

Barry, D. and Hartigan, J. A. (1993). A Bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88:309–319.

Beasley, D., Bull, D. R., and Martin, R. R. (1993). An overview of genetic algorithms: Part 1, fundamentals. *University Computing*, 15:58–69.

Beaulieu, C., Ouarda, T. B., and Seidou, O. (2010). A Bayesian normal homogeneity test for the detection of artificial discontinuities in climatic series. *International Journal of Climatology*, 30:2342–2357.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*. Springer Series in Statistics, 2 edition.

Carlin, B. P., Gelfand, A. E., and Smith, A. F. (1992). Hierarchical Bayesian analysis of changepoint problems. *Applied Statistics*, 41:389–405.

Carlin, B. P. and Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall/CRC Boca Raton.

Caussinus, H. and Mestre, O. (2004). Detection and correction of artificial shifts in climate series. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53:405–425.

Chernoff, H. and Zacks, S. (1964). Estimating the current mean of a normal distribution which is subjected to changes in time. *The Annals of Mathematical Statistics*, 35:999–1018.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86:221–241.

Christensen, R. (2002). *Plane Answers to Complex Questions: the Theory of Linear Models*. Springer.

Davis, L. D. (1991). *Handbook Of Genetic Algorithms*. Van Nostrand Reinhold.

Davis, R. A., Lee, Thomas, C., and Rodrigues-Yam, G. A. (2006). Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, 101:223–239.

Davis, R. A., Lee, T., and Rodriguez-Yam, G. A. (2008). Break detection for a class of nonlinear time series models. *Journal of Time Series Analysis*, 29(5):834–867.

Della-Marta, P. M. and Wanner, H. (2006). A method of homogenizing the extremes and mean of daily temperature measurements. *Journal of Climate*, 19:4179–4197.

Du, C., Kao, C.-L. M., and Kou, S. C. (2015). Stepwise signal extraction via marginal likelihood. *Journal of the American Statistical Association*, page in press.

Eckley, I. A., Fearnhead, P., and Killick, R. (2011). *Analysis of Changepoint Models*, pages 205–224. Cambridge University Press.

Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems. *Statistical Computing*, 16:203–213.

Fearnhead, P. and Liu, Z. (2011). Efficient Bayesian analysis of multiple changepoint models with dependence across segments. *Stat Comput*, 21:217–229.

García-Donato, G. and Martínez-Beneito, M. A. (2013). On sampling strategies in Bayesian variable selection problems with large model spaces. *Journal of the American Statistical Association*, 108(501):340–352.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistics Sinica*, 7:339–373.

Giordani, P. and Kohn, R. (2008). Efficient Bayesian inference for multiple change-point and mixture innovation models. *Journal of Business and Economic Statistics*, 26(1):66–77.

Goldberg, D. E. and Holland, J. H. (1988). Genetic algorithms and machine learning. *Machine Learning*, 3:95–99.

Green, Peter, J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732.

Grünwald, P. D., Myung, I. J., and Pitt, M. A. (2005). *Advances in Minimum Description Length: Theory and Applications*. MIT press.

Hannart, A. and Naveau, P. (2009). Bayesian multiple change points and segmentation: Application to homogenization of climatic series. *Water Resources Research*, 45.

Hannart, A. and Naveau, P. (2012). An improved Bayesian information criterion for multiple change-point models. *Technometrics*, 54(3):256–268.

Hansen, M. H. and Yu, B. (2001). Model selection and the principle of minimum description lengths. *Journal of the American Statistical Association*, 96:746–774.

Harville, D. A. (2008). *Matrix Algebra From a Statistician's Perspective*. Springer-Verlag.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Ko, S. I. M., Chong, T. T. L., and Ghosh, P. (2015). Dirichlet process hidden Markov multiple change-point model. *Bayesian Analysis*, 10(2):275–296.

Kuglitsch, F. G., Toreti, A., Xoplaki, E., Della-Marta, P. M., Luterbacher, J., and Wanner, H. (2009). Homogenization of daily maximum temperature series in the Mediterranean. *Journal of Geophysical Research: Atmospheres (1984–2012)*, 114:D15108.

Lai, T. L. and Xing, H. (2011). A simple Bayesian approach to multiple change-points. *Statistica Sinica*, pages 539–569.

Li, F. and Zhang, N. R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association*, 105(491):1202–1214.

Li, S. and Lund, R. (2012). Multiple changepoint detection via genetic algorithms. *Journal of Climate*, 25:674–686.

Li, Y. and Lund, R. (2015). Multiple changepoint detection using metadata. *Journal of Climate*, 28:4199–4216.

Li, Y., Lund, R., and Priyadarshani, H. (2015). Bayesian minimal description lengths for multiple changepoint detection. *working paper*.

Liu, G., Shao, Q., Lund, R., and Woody, J. (2015). Testing seasonal means in time series data. *In revision for, Environmetrics*.

Lu, Q. Q. and Lund, R. (2007). Simple linear regression with multiple level shifts. *Canadian Journal of Statistics*, 37:447–458.

Lu, Q. Q., Lund, R., and Lee, T. C. (2010). An MDL approach to the climate segmentation problem. *The Annals of Applied Statistics*, 4:299–319.

Lund, R., Hurd, H., Bloomfield, P., and Smith, R. (1995). Climatological time series with periodic correlation. *Journal of Climate*, 8:2787–2809.

Lund, R. and Reeves, J. (2002). Detection of undocumented changepoints: A revision of the two-phase regression model. *Journal of Climate*, 15:2547–2554.

Lund, R., Seymour, L., and Kafadar, K. (2001). Temperature trends in the United States. *Environmetrics*, 12:673–690.

Menne, M. J. and Williams Jr., C. N. (2005). Detection of undocumented changepoints using multiple test statistics and composite reference series. *Journal of Climate*, 18:4271–4286.

Menne, M. J. and Williams Jr., C. N. (2009). Homogenization of temperature series via pairwise comparisons. *Journal of Climate*, 22:1700–1717.

Mitchell, J. M. (1953). On the causes of instrumentally observed secular temperature trends. *Journal of Meteorology*, 10:244–261.

Potter, K. W. (1981). Illustration of a new test for detecting a shift in mean in precipitation series. *Monthly Weather Review*, 109:2040–2045.

Ray, B. K. and Tsay, R. S. (2002). Bayesian methods for change-point detection in long-range dependent processes. *Journal of Time Series Analysis*, 23(6):687–705.

Reeves, J., Chen, J., Wang, X., Lund, R., and Lu, Q. Q. (2007). A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, 46:900–915.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific Publ. Co.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Scott, J. and Berger, J. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38:2587–2619.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(143-157):623.

Vincent, L. A. (1998). A technique for the identification of inhomogeneities in Canadian temperature series. *Journal of Climate*, 11:1094–1104.

Vincent, L. A. and Zhang, X. (2002). Homogenization of daily temperatures over Canada. *Journal of Climate*, 15:1322–1334.

Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences*. Academic Press.

Yao, Y.-C. (1984). Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches. *The Annals of Statistics*, 12:1434–1447.

Zhang, N. R. and Siegmund, D. O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63:22–32.