ON ROBUST ESTIMATION OF MULTIPLE CHANGE POINTS IN MULTIVARIATE AND MATRIX PROCESSES

by

YANA MELNYKOV

MARCUS B. PERRY, COMMITTEE CHAIR
OLEKSANDRA BEZNOSOVA
JAMES J. COCHRAN
BRUCE BARRETT
CALI M. DAVIS

A DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Information Systems, Statistics and Management Science
in the Graduate School of
The University of Alabama

TUSCALOOSA, ALABAMA

2017

ABSTRACT

There are numerous areas of human activities where various processes are observed over time. If the conditions of the process change, it can be reflected through the shift in observed response values. The detection and estimation of such shifts is commonly known as change point inference. While the estimation helps us learn about the process nature, assess its parameters, and analyze identified change points, the detection focuses on finding shifts in the real-time process flow. There is a vast variety of methods proposed in the literature to target change point detections in both settings. Unfortunately, the majority of procedures impose very restrictive assumptions. Some of them include the normality of data, independence of observations, or independence of subjects in multisubject studies. In this dissertation, a new methodology, relying on more realistic assumptions, is developed. This dissertation report includes three chapters. The summary of each chapter is provided below. In the first chapter, we develop methodology capable of estimating and detecting multiple change points in a multisubject single variable process observed over time. In the second chapter, we introduce methodology for the robust estimation of change points in multivariate processes observed over time. In the third chapter, we generalize the ideas presented in the first two chapters by developing methodology capable of identifying multiple change points in multisubject matrix processes observed over time.

DEDICATION

I dedicate this dissertation to my grandparents Shahida Niyazymbetova and Seyfulla Orynbaev. I also dedicate it to my beloved husband Volodymyr Melnykov.

LIST OF ABBREVIATIONS AND SYMBOLS

N	number of subjects
p	number of variables
d	number of variables in two-factor data
q	number of explanatory variables
T	number of time points
K	number of change points
M	number of blocks
n_m	size of m^{th} block
$\boldsymbol{\lambda}$	transformation (skewness) parameter vector
$\boldsymbol{\Lambda}$	skewness parameter matrix
$\boldsymbol{\Theta}$	overall parameter vector
$\boldsymbol{\mu}_k$	k^{th} mean vector
σ^2	variance parameter responsible for modeling between-block variability
σ_b^2	variance parameter responsible for modeling within-block variability
η	ratio of between- and within-block variabilities
AR_1	autoregressive time series process of order one
$\boldsymbol{R}_\phi$	correlation matrix of AR_1 time series
ϕ	correlation parameter of AR_1 time series

δ^2	parameter responsible for modeling variability in a covariance matrix $\boldsymbol{\Psi}$
$\boldsymbol{I}_p$	$p \times p$ identity matrix
$\boldsymbol{V}_m$	covariance matrix associated with m^{th} block of $\boldsymbol{\Sigma}$
$\mathcal{D}_M(\boldsymbol{V})$	block-diagonal matrix consisting of M identical blocks $\boldsymbol{V}$
$\boldsymbol{M}$	mean matrix
$\boldsymbol{\Sigma}$	covariance matrix
$\boldsymbol{\Delta}$	covariance matrix
$\boldsymbol{\Psi}$	covariance matrix
$\mathbf{1}_p$	vector of length p consisting of ones
$\boldsymbol{B}$	matrix of linear model coefficients
$\boldsymbol{\beta}_k$	vector of coefficients associated with the k^{th} process
$\boldsymbol{b}_k$	vector consisting of ones and zeros with ones in positions of the k^{th} process
t_k	k^{th} change point
$\mathcal{M}$	mean tensor
$\mathcal{Y}$	data tensor
$\tilde{\mathcal{Y}}$	tensor matricization form
$\ddot{\mathcal{Y}}$	tensor matricization form
$\breve{\mathcal{Y}}$	tensor matricization form
$\mathcal{T}$	transformation operator
BIC	Bayesian information criterion

ACKNOWLEDGMENTS

First of all, I would like to thank my advisor, Professor Perry, for his research guidance, patience, and friendly support provided over the entire period of my dissertation research. I highly appreciate his trust in me and hope to contunue our research collaboration.

I would also like to thank my dissertation committee for their constructive criticism and constant encouragement on my way to the final defense.

I was lucky to have such classmates as Xuwen Zhu, Rong Zheng, Semhar Michael, and Shuchismita Sarkar who became not just my friends but rather a part of my family.

I would like to thank my beloved husband Volodymyr for supporting me during all these difficult years of my studies. He has been very patient teaching me repeatedly. It would be impossible to become who I am now without his love and support.

I am thankful to my sweet kids who sacrificed our time together on many nights and weekends that I spent working in the office.

I am grateful to my parents-in-law for being so supportive and playing a crucial role in taking care of my little ones.

I would like to thank my mother Banu Lebedeva who always believes in me and supports my ideas.

I am extremely lucky to have such a great and loving family that I have been always wishing for.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

ON MATRIX MANLY DISTRIBUTION FOR ROBUST CHANGE POINT DETECTION AND ESTIMATION

A variety of change point estimation and detection algorithms have been developed for random variables observed over time. The acquisition of data in current practice often results in multiple observation units studied. The traditional treatment of such observations involves the assumption of their independence. In practice, however, this assumption is often inadequate or unrealistic. We propose an effective and modern computerized approach to estimating and detecting change points in time series processes in the situation when the assumption of independent observations is not feasible. The developed methodology relies on the multivariate transformation and matrix normal distribution. The latter is used for separating the sources of variability. The application of the back-transform of the exponential transformation leads to a flexible distribution that effectively accounts for deviations from normality. The developed procedure has been successfully tested in various settings and applied to a crime rate data set.

1.1 Introduction

We live in a dynamic world where various changes occur or may happen at every moment. The ability to identify changes that occurred in the past helps explain the nature of observed processes. The detection of changes in timely manner is of vital importance in many areas of science, industry, and human activity. Applications of related methods can be found in finance [5], medicine [12], ecology [8], biology [41], history [10], engineering [31], and many other ar-

eas.

One of the first change point problems was considered more than a half of a century ago by [28] who monitored the mean to find a single change point in the case of a random sample from a univariate normal distribution. Different variations of this setting assuming known or unknown variances were considered by [13] and [42], respectively. In both cases, the variance was assumed to be common for both processes. Another group of change point estimation methods was devoted to identifying the variance shift under the assumption of a constant mean parameter [15, 9, 16, 5]. Later, [14] proposed an approach based on the assymptotic distribution of the likelihood ratio statistic for testing the change in mean and variance simultaneously.

In multivariate setting, the most common assumption is that data are distributed according to a multivariate normal distribution. A single change point estimation for mean vectors was investigated by [38] and [39]. Around the same time, scholars became interested in identifying multiple change points in mean vectors. [44] and [45] developed corresponding theory for various structures of covariance matrices. [6] developed a testing approach for estimating the change in covariance matrices. The proposed procedure was further enhanced by [7] who developed an approach for estimating the shift is mean vector and covariance matrix simultaneously. The development of change point methods employeeing distributions other than normal one were investigated by [32], [33], and etc. For the purpose of identifying the optimal model, it is common among researchers to apply a formal statistical testing procedure or employ popular information criteria such as the one proposed by [1] (known as AIC) or [37] (known as the Bayesian information criterion or just BIC).

In this paper, we consider a problem of estimating multiple change points in a univariate process monitored for multiple subjects. Here, no independence assumption is made with regard to the subjects. Moreover, the corresponding covariance structure can be of any form. We pro-

pose a novel model formulation by building on a matrix normal distribution. The advantage of this distribution is in easy handling of the variation associated with data rows and columns, which can be modeled separately. To improve the robustness of the developed methodology to deviations from normality, we employ the ideas related to the exponential transformation of [24]. This leads us to the so-called matrix Manly distribution which contains a skewness parameter that improves the robustness characteristics of the model.

The rest of the paper is constructed as follows. Section 2.2 is devoted to developing the proposed methodology. Section 2.3 considers simulations experiments conducted in various settings. Section 3.3 applies the developed methods to the analysis of crime rates for 125 cities representing five regions in the United States. The paper concludes with a summary in Section 3.4.

1.2 Methodology

1.2.1 Exponential transformation and Manly distribution

As per our discussion in Introduction, the normality assumption is oftentime violated. One of possible remedies is to employ a transformation to near-normality. Perhaps, the most famous one in this class is the power transformation proposed by [4]. Some criticism of the Box-Cox transformation is related to its restricted support and ability to handle right-skewed data exclusively. The exponentital transformation of [24] is somewhat less popular but free of the above-mentioned constraints, *i.e.*, it can be applied to left- and right-skewed data in $\mathbb{R}$. The Manly transformation is given by

$$\mathcal{M}(y;\lambda) = \lambda^{-1}(\exp(\lambda y) - 1)I(\lambda \neq 0) + yI(\lambda = 0), \tag{1.2.1}$$

where y is the original observation and $\mathcal{M}(\cdot;\lambda)$ is the transformation operator with parameter λ. $I(\mathcal{A})$ is the indicator function that returns 1 if $\mathcal{A}$ is true and yields 0 otherwise.

For vector-valued $\boldsymbol{y}$, it is commonly assumed that coordinatewise transformations lead to the joint normality of the transformed vector [2, 40, 23, 46]. In this case, the Manly transformation takes the form

$$\mathcal{M}(\boldsymbol{y};\boldsymbol{\lambda}) = (\mathcal{M}(y_1;\lambda_1), \mathcal{M}(y_2;\lambda_2), \ldots, \mathcal{M}(y_p;\lambda_p))^\top \tag{1.2.2}$$

with $\boldsymbol{y} = (y_1, y_2, \ldots, y_p)^\top$ and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_p)^\top$. Assuming that it is successful, *i.e.*, $\mathcal{M}(\boldsymbol{y};\boldsymbol{\lambda})$ roughly follows normal distribution, the Manly back transformation yields the so called Manly distribution [46] with probability density function (pdf) given by

$$g(\boldsymbol{y};\boldsymbol{\Theta}) = \phi_p(\mathcal{M}(\boldsymbol{y};\boldsymbol{\lambda});\boldsymbol{\mu},\boldsymbol{\Sigma})e^{\boldsymbol{\lambda}^\top \boldsymbol{y}}, \tag{1.2.3}$$

where $\phi_p(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$ denotes the p-variate Gaussian pdf with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. This distribution can be used for modeling the original data without transforming them. Under this setting, the transformation parameter $\boldsymbol{\lambda}$ can be more conveniently thought of as a skewness parameter.

1.2.2 Matrix normal and matrix Manly distributions

Suppose $\boldsymbol{Y}$ is a matrix-variate observation of dimensions $N \times T$ that follows matrix normal distribution [11] with pdf given by

$$\phi_{N\times T}(\boldsymbol{Y};\boldsymbol{M},\boldsymbol{\Sigma},\boldsymbol{\Psi}) = (2\pi)^{-\frac{NT}{2}}|\boldsymbol{\Sigma}|^{-\frac{T}{2}}|\boldsymbol{\Psi}|^{-\frac{N}{2}}\exp\left\{-\frac{1}{2}\text{tr}\left\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{M})\boldsymbol{\Psi}^{-1}(\boldsymbol{Y}-\boldsymbol{M})^\top\right\}\right\}, \tag{1.2.4}$$

where $\text{tr}\{\cdot\}$ represents the trace operator, $\boldsymbol{M}$ is the $N \times T$ mean matrix, and $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$ are covariance matrices of dimensions $N \times N$ and $T \times T$ describing variability associated with rows and

columns, respectively. It is known that the matrix normal distribution can be equivalently represented as multivariate normal distribution. In other words, if $\boldsymbol{Y}$ has the pdf $\phi_{N\times T}(\boldsymbol{Y};\boldsymbol{M},\boldsymbol{\Sigma},\boldsymbol{\Psi})$, it can be shown that $\text{vec}(\boldsymbol{Y})$ follows $\phi_{NT}(\text{vec}(\boldsymbol{Y});\text{vec}(\boldsymbol{M}),\boldsymbol{\Psi}\otimes\boldsymbol{\Sigma})$, where $\otimes$ represents the Kronecker product and $\text{vec}(\cdot)$ is an operator that stacks the columns of a $N\times T$ matrix on top of each other producing a NT-variate vector. It can be noted that $a\boldsymbol{\Psi}\otimes a^{-1}\boldsymbol{\Sigma}=\boldsymbol{\Psi}\otimes\boldsymbol{\Sigma}$ implying that the matrix normal distribution is non-identifiable. For this reason, a common practice is to incorporate a restriction on $\boldsymbol{\Sigma}$ or $\boldsymbol{\Psi}$.

As in the vector-valued case, the shortcoming of the matrix normal distribution is its incapability to model skewed data. One can generalize the idea outlined in Section 1.2.1 to produce the transformation

$$\mathcal{M}(\boldsymbol{Y};\boldsymbol{\Lambda})=\begin{pmatrix}\mathcal{M}(Y_{11};\Lambda_{11}) & \mathcal{M}(Y_{12};\Lambda_{12}) & \dots & \mathcal{M}(Y_{1T};\Lambda_{1T})\\ \mathcal{M}(Y_{21};\Lambda_{21}) & \mathcal{M}(Y_{22};\Lambda_{22}) & \dots & \mathcal{M}(Y_{2T};\Lambda_{2T})\\ \vdots & \vdots & \ddots & \vdots\\ \mathcal{M}(Y_{N1};\Lambda_{N1}) & \mathcal{M}(Y_{N2};\Lambda_{N2}) & \dots & \mathcal{M}(Y_{NT};\Lambda_{NT})\end{pmatrix}, \tag{1.2.5}$$

where $\boldsymbol{\Lambda}=(\Lambda_{ij})_{N\times T}$ is the $N\times T$ transformation parameter matrix. Then, the matrix analogue of the Manly distribution can be obtained based on (1.2.5) and its pdf is given by

$$g(\boldsymbol{Y};\boldsymbol{M},\boldsymbol{\Sigma},\boldsymbol{\Psi},\boldsymbol{\Lambda})=\phi_{N\times T}(\mathcal{M}(\boldsymbol{Y};\boldsymbol{\Lambda});\boldsymbol{M},\boldsymbol{\Sigma},\boldsymbol{\Psi})e^{\text{tr}\{\boldsymbol{\Lambda}^\top\boldsymbol{Y}\}}. \tag{1.2.6}$$

The obtained matrix Manly pdf inherits the non-identifiability problem from the matrix normal distribution. Fortunately, a variety of constraints can be employed to alleviate the problem.

1.2.3 Modeling mean matrix and shift change points

Suppose there is a sample consisting of N T-variate observations $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N$ that share some temporal structure. Then, the matrix $\boldsymbol{Y}$ can be formed in such a way that its rows represent the observed sample, *i.e.*, $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N)^\top$. Since all observations have to be measured on the same scale at all time points, this leads to the condition $\boldsymbol{\Lambda} = \lambda \mathbf{1}_N \mathbf{1}_T^\top$, where λ is the common skewness parameter. Then, denoting $\mathcal{M}(\boldsymbol{Y}; \lambda) \equiv \mathcal{M}(\boldsymbol{Y}; \lambda \mathbf{1}_N \mathbf{1}_T^\top)$ for notational simplicity, the model (1.2.6) will be $g(\boldsymbol{Y}; \boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}, \lambda) = \phi_{N \times T}(\mathcal{M}(\boldsymbol{Y}; \lambda); \boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}) e^{\lambda \mathbf{1}_N^\top \boldsymbol{Y} \mathbf{1}_T}$. In this setting, $\boldsymbol{\Sigma}$ represents the covariance matrix associated with N observations while $\boldsymbol{\Psi}$ is responsible for modeling the underlying temporal structure. Mean matrix $\boldsymbol{M}$ has NT parameters. In many problems, this number can be large. In case when explanatory variables are available, $\boldsymbol{M}$ can be modeled as $\boldsymbol{M} = \boldsymbol{X}\boldsymbol{B}$, where $\boldsymbol{X}$ is the $N \times q$ design matrix and $\boldsymbol{B}$ is the $q \times T$ matrix of linear model coefficients. Assuming there are no changes in mean over all T time points, $\boldsymbol{B}$ can be written as $\boldsymbol{B} = \boldsymbol{\beta} \mathbf{1}_T^\top$, where $\boldsymbol{\beta}$ is the common vector of coefficients with q elements.

Suppose now that there are K shift change points at times $t_1, t_2, \ldots, t_K$. Then, $\boldsymbol{B}$ can be written as

$$\begin{aligned} \boldsymbol{B} &= \left(\underbrace{\boldsymbol{\beta}_0, \boldsymbol{\beta}_0 \ldots, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1}_{t_1}, \underbrace{\boldsymbol{\beta}_1 \ldots, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2}_{t_2 - t_1}, \ldots, \underbrace{\boldsymbol{\beta}_{K-1}, \ldots, \boldsymbol{\beta}_{K-1}, \boldsymbol{\beta}_K}_{t_K - t_{K-1}}, \underbrace{\boldsymbol{\beta}_K, \ldots, \boldsymbol{\beta}_K}_{T - t_K} \right) \\ &= \sum_{k=0}^{K} \boldsymbol{\beta}_k \boldsymbol{b}^\top_{<t_k, t_{k+1}-1>}, \quad t_0 = 1 \quad \text{and} \quad t_{K+1} = T + 1, \end{aligned} \tag{1.2.7}$$

where $\boldsymbol{\beta}_k$ is a vector of length q and $\boldsymbol{b}_{<t_k, t_{k+1}-1>}$ is a vector of length T that consists of zeros and ones, with ones located between positions t_k and $t_{k+1} - 1$ inclusively. Thus, until the first change point is observed at time t_1, the process is unchanged and the vector $\boldsymbol{\beta}_0$ remains the same. Starting from time t_1 and up to the next change point observed at time t_2, the process is stable

with the parameter vector $\boldsymbol{\beta}_1$. Continuing this process, the entire matrix of coefficients $\boldsymbol{B}$ can be formed as shown in (1.2.7). To simplify further notation, we define $\boldsymbol{b}_k \equiv \boldsymbol{b}_{<t_k,t_{k+1}-1>}$. Hence, $\boldsymbol{B} = \sum_{k=0}^{K} \boldsymbol{\beta}_k \boldsymbol{b}_k^\top$ and the corresponding pdf can now be written as follows below:

$$g(\boldsymbol{Y}; \boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}, \lambda) = \phi_{N\times T}(\mathcal{M}(\boldsymbol{Y}; \lambda); \boldsymbol{X}\boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}) e^{\lambda \mathbf{1}_N^\top \boldsymbol{Y} \mathbf{1}_T}. \tag{1.2.8}$$

1.2.4 Modeling covariance matrices

The specific choice of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$ depends on a particular application. Without loss of generality, we illustrate the further model development in the random effect setting with time modeled by means of the autoregressive process of order 1 (AR_1). Then, covariance matrix $\boldsymbol{\Sigma}$ is given by

$$\boldsymbol{\Sigma} = \sigma^2 \text{diag}\left\{\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_M\right\} \quad \text{with} \quad \boldsymbol{V}_m = \eta \mathbf{1}_{n_m} \mathbf{1}_{n_m}^\top + \boldsymbol{I}_{n_m}, \quad \eta = \frac{\sigma_b^2}{\sigma^2},$$

where σ^2 and σ_b^2 are parameters responsible for modeling between- and within-block variability, respectively. M represents the number of blocks, each of size n_m, and $\boldsymbol{I}_{n_m}$ is the $n_m \times n_m$ identity matrix. Assuming equal block sizes, *i.e.*, $n_1 = \ldots = n_M \equiv n$, we obtain $\boldsymbol{V}_1 = \ldots = \boldsymbol{V}_M \equiv \boldsymbol{V}$. Then, $\boldsymbol{\Sigma}$ can be written as $\boldsymbol{\Sigma} = \sigma^2 \mathcal{D}_M(\boldsymbol{V})$, where $\mathcal{D}_M(\boldsymbol{V})$ represents a block-diagonal matrix consisting of M identical blocks $\boldsymbol{V}$. It can be shown that

$$\boldsymbol{V}^{-1} = \frac{1}{\eta n + 1} \left(\eta(n\boldsymbol{I}_n - \mathbf{1}_n \mathbf{1}_n^\top) + \boldsymbol{I}_n\right)$$

and, as a result,

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma^2} \mathcal{D}_M^{-1}(\boldsymbol{V}) = \frac{1}{\sigma^2} \mathcal{D}_M(\boldsymbol{V}^{-1}) = \frac{1}{\sigma^2(\eta n + 1)} \left(\eta \left(n\boldsymbol{I}_N - \mathcal{D}_M(\mathbf{1}_n \mathbf{1}_n^\top)\right) + \boldsymbol{I}_N\right). \tag{1.2.9}$$

It can be also shown that $|\boldsymbol{V}| = \eta n + 1$ and then,

$$|\boldsymbol{\Sigma}| = \sigma^{2N}|\mathcal{D}_M(\boldsymbol{V})| = \sigma^{2N}(n\eta + 1)^M. \tag{1.2.10}$$

The matrix $\boldsymbol{\Psi}$ corresponding to AR_1 process is given by

$$\boldsymbol{\Psi} = \frac{\delta^2}{1-\phi^2}\boldsymbol{R}_\phi \quad \text{with} \quad \boldsymbol{R}_\phi = \begin{pmatrix} 1 & \phi & \phi^2 & \dots & \phi^{T-1} \\ \phi & 1 & \phi & \dots & \phi^{T-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{T-1} & \phi^{T-2} & \phi^{T-3} & \dots & 1 \end{pmatrix},$$

where δ^2 and ϕ are corresponding variance and AR_1 parameters. $\boldsymbol{R}_\phi$ represents the correlation matrix associated with AR_1. Recall that a restriction on covariance matrices has to be introduced to ensure model identifiability. One convenient constraint in the considered setting is $\delta^2 = 1 - \phi^2$. Then, the covariance matrix $\boldsymbol{\Psi}$ can be readily reduced to the correlation matrix $\boldsymbol{R}_\phi$. It can be shown that

$$|\boldsymbol{\Psi}| = |\boldsymbol{R}_\phi| = (1-\phi^2)^{T-1} \tag{1.2.11}$$

and

$$\boldsymbol{\Psi}^{-1} = \boldsymbol{R}_\phi^{-1} = \frac{1}{1-\phi^2}(\boldsymbol{I}_T - \phi\boldsymbol{J}_1 + \phi^2\boldsymbol{J}_2), \tag{1.2.12}$$

where $\boldsymbol{J}_1$ and $\boldsymbol{J}_2$ are $T \times T$ matrices defined as follows below:

$$\boldsymbol{J}_1 = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 1 & 0 & 1 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & 0 & \dots & 1 & 0 \end{pmatrix} \quad \text{and} \quad \boldsymbol{J}_2 = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 \end{pmatrix}.$$

1.2.5 Estimation of model parameters

Equation (2.2.2) can be updated based on expressions (1.2.9)-(1.2.12) and maximum likelihood estimates (MLEs) of parameters σ^2, η, ϕ, $\boldsymbol{\beta}_0, \dots, \boldsymbol{\beta}_K$, and λ can be found. The total number of parameters in the model is $(K+1)q + 4$. The log-likelihood function corresponding to the pdf given in (2.2.2) has the following form:

$$\begin{aligned} \log \mathcal{L}(\boldsymbol{Y}; \boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}, \lambda) = &-\frac{NT}{2}\log(2\pi) - \frac{T}{2}\log|\boldsymbol{\Sigma}| - \frac{N}{2}\log|\boldsymbol{\Psi}| \\ &- \frac{1}{2}\text{tr}\left\{\boldsymbol{\Sigma}^{-1}(\mathcal{M}(\boldsymbol{Y};\lambda) - \boldsymbol{X}\boldsymbol{B})\boldsymbol{\Psi}^{-1}(\mathcal{M}(\boldsymbol{Y};\lambda) - \boldsymbol{X}\boldsymbol{B})^\top\right\} + \lambda \mathbf{1}_N^\top \boldsymbol{Y} \mathbf{1}_T. \end{aligned} \tag{1.2.13}$$

Taking partial derivatives of this log-likelihood function with respect to $\boldsymbol{\beta}_k$, leads to the expression

$$\boldsymbol{\beta}_k = \left(\boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1} \left(\mathcal{M}(\boldsymbol{Y};\lambda) - \boldsymbol{X} \sum_{\substack{k'=0 \\ k' \neq k}}^{K} \boldsymbol{\beta}_{k'} \boldsymbol{b}_{k'}^\top \right) \boldsymbol{\Psi}^{-1} \boldsymbol{b}_k \left(\boldsymbol{b}_k^\top \boldsymbol{\Psi}^{-1} \boldsymbol{b}_k\right)^{-1}.$$

Recalling that $\boldsymbol{\Sigma} = \sigma^2 \mathcal{D}_M(\boldsymbol{V})$ and $\boldsymbol{\Psi} = \boldsymbol{R}_\phi$ under constraint $\delta^2 = 1 - \phi^2$, this equation can be written as

$$\boldsymbol{\beta}_k = \left(\boldsymbol{X}^\top \mathcal{D}_M^{-1}(\boldsymbol{V}) \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \mathcal{D}_M^{-1}(\boldsymbol{V}) \left(\mathcal{M}(\boldsymbol{Y}; \lambda) - \boldsymbol{X} \sum_{\substack{k'=0 \\ k' \neq k}}^{K} \boldsymbol{\beta}_{k'} \boldsymbol{b}_{k'}^\top \right) \boldsymbol{R}_\phi^{-1} \boldsymbol{b}_k \left(\boldsymbol{b}_k^\top \boldsymbol{R}_\phi^{-1} \boldsymbol{b}_k \right)^{-1}. \tag{1.2.14}$$

Then, a system of $K+1$ equations (1.2.14) for $k = 0, 1, \ldots, K$ can be solved for $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$. For example, denoting $b_{kk'} \equiv \boldsymbol{b}_k^\top \boldsymbol{\Psi}^{-1} \boldsymbol{b}_{k'}$, $\boldsymbol{\beta}_0$ can be found for $K = 1$ as

$$\boldsymbol{\beta}_0 = \frac{\left(\boldsymbol{X}^\top \mathcal{D}_M^{-1}(\boldsymbol{V}) \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \mathcal{D}_M^{-1}(\boldsymbol{V}) \mathcal{M}(\boldsymbol{Y}; \lambda) \boldsymbol{R}_\phi^{-1} \left(b_{11} \boldsymbol{b}_0 - b_{01} \boldsymbol{b}_1 \right)}{b_{00} b_{11} - b_{01}^2},$$

and for $K = 2$ as

$$\begin{aligned} \boldsymbol{\beta}_0 = &\left(\boldsymbol{X}^\top \mathcal{D}_M^{-1}(\boldsymbol{V}) \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \mathcal{D}_M^{-1}(\boldsymbol{V}) \mathcal{M}(\boldsymbol{Y}; \lambda) \boldsymbol{R}_\phi^{-1} \times \\ &\frac{b_{12}(b_{01} \boldsymbol{b}_2 + b_{02} \boldsymbol{b}_1 - b_{12} \boldsymbol{b}_0) - b_{02} b_{11} \boldsymbol{b}_2 - b_{01} b_{22} \boldsymbol{b}_1 + b_{11} b_{22} \boldsymbol{b}_0}{b_{00} b_{11} b_{22} - b_{01}^2 b_{22} - b_{02}^2 b_{11} - b_{12}^2 b_{00} + 2 b_{01} b_{02} b_{12}}. \end{aligned}$$

Expressions for the rest of coefficients are symmetric. Taking the partial derivative with respect to σ^2, it can be shown that

$$\sigma^2 = \frac{\operatorname{tr}\left\{ \mathcal{D}_M^{-1}(\boldsymbol{V}) (\mathcal{M}(\boldsymbol{Y}; \lambda) - \boldsymbol{X}\boldsymbol{B}) \boldsymbol{R}_\phi^{-1} (\mathcal{M}(\boldsymbol{Y}; \lambda) - \boldsymbol{X}\boldsymbol{B})^\top \right\}}{TN}. \tag{1.2.15}$$

It can be noted that expressions for coefficients $\boldsymbol{\beta}_k$'s and variance σ^2 depend on the remaining parameters η, ϕ, and λ only. Analytical expressions for the latter parameters are not available and numerical optimization of the log-likelihood function should be carried out. Substituting expres-

sions (1.2.9)-(1.2.12) into (1.2.13), the log-likelihood can be rewritten as

$$\log \mathcal{L}(\boldsymbol{Y}; \eta, \phi, \lambda) = \frac{NT}{2} \log(2\pi\sigma^2_{\eta,\phi,\lambda}) - \frac{MT}{2} \log(n\eta + 1) - \frac{N(T-1)}{2} \log(1-\phi^2)$$
$$- \frac{1}{2\sigma^2_{\eta,\phi,\lambda}(\eta n + 1)(1-\phi^2)} \text{tr} \left\{ \left(\eta \left(n\boldsymbol{I}_N - \mathcal{D}_N(\mathbf{1}_n \mathbf{1}_n^\top)\right) + \boldsymbol{I}_N\right) \left(\mathcal{M}(\boldsymbol{Y}; \lambda) - \boldsymbol{X} \sum_{k=0}^{K} \boldsymbol{\beta}_{k;\eta,\phi,\lambda} \boldsymbol{b}_k^\top\right) \right.$$
$$\left. \times \left(\boldsymbol{I}_T - \phi \boldsymbol{J}_1 + \phi^2 \boldsymbol{J}_2\right) \left(\mathcal{M}(\boldsymbol{Y}; \lambda) - \boldsymbol{X} \sum_{k=0}^{K} \boldsymbol{\beta}_{k;\eta,\phi,\lambda} \boldsymbol{b}_k^\top\right)^\top \right\} + \lambda \mathbf{1}_N^\top \boldsymbol{Y} \mathbf{1}_T,$$

where $\sigma^2_{\eta,\phi,\lambda}$ and $\boldsymbol{\beta}_{k;\eta,\phi,\lambda}$ represent σ^2 and $\boldsymbol{\beta}_k$ expressed as functions of η, ϕ, and λ. Numerical maximization of $\log \mathcal{L}(\boldsymbol{Y}; \eta, \phi, \lambda)$ is a relatively simple optimization problem with numerous algorithms available. In this paper, a simplex method proposed by [25] has been employed.

1.2.6 Change point estimation algorithm

In this section, we outline the change point estimation algorithm that is devoted to identifying shift change points $t_1, t_2, \ldots, t_K$. The procedure starts with the assumption of no change point (*i.e.*, $K = 0$) and proceeds with comparisons to all possible one-change-point models (*i.e.*, $K = 1$ and $t = 2, 3, \ldots, T$). A formal testing procedure, such as a likelihood ratio test, can be developed for this purpose. In this paper, however, we employ an alternative approach based on the Bayesian Information Criterion (BIC) [37] to avoid potential issues with multiple comparisons and the choice of confidence level. If there are models with $K = 1$ such that their BIC values are lower than that of the model with no change point, we choose the one with the lowest BIC as the currently best model and proceed in a similar fashion to the case $K = 2$, *etc*. If no improvement can be made, the currently best model is selected as declared the final model. The pseudocode description of the proposed model selection technique is provided in Algorithm 1.

It can be noted that the outlined methodology and algorithm can be immediately generalized to the estimation of change points representing a much broader class than that associated

Data: $\boldsymbol{Y} = (y_{ij})_{N \times T}$
Initialization: $K \leftarrow 0$, $\boldsymbol{\tau}_{best} \leftarrow NA$, $BIC_{best} \leftarrow BIC^{<0>}$
repeat
  $K \leftarrow K + 1$;
  $\mathcal{T} \leftarrow \{(t_1, t_2, \ldots, t_K) : 2 \leq t_1 \leq t_2 \leq \ldots \leq t_K \leq T\}$;
  $\boldsymbol{\tau}^{\star} \leftarrow \underset{\boldsymbol{\tau} \in \mathcal{T}}{\operatorname{argmin}} BIC_{\boldsymbol{\tau}}^{<K>}$;
  if $BIC_{\boldsymbol{\tau}^{\star}}^{<K>} < BIC_{best}$ **then**
    $K_{best} \leftarrow K$;
    $\boldsymbol{\tau}_{best} \leftarrow \boldsymbol{\tau}^{\star}$;
    $BIC_{best} \leftarrow BIC_{\boldsymbol{\tau}^{\star}}^{<K>}$
  else
    $K_{best} \leftarrow K - 1$;
    break;
  end
until $K = T - 1$;
Result: K_{best}, $\boldsymbol{\tau}_{best}$, BIC_{best}

Algorithm 1: Change point estimation algorithm.

with shifts in means. The only modification needed is associated with identifying the structure of applicable $\boldsymbol{b}$ vectors. In the most general case, one can consider all possible permutations of processes at T time points. Indeed, such a procedure can be time consuming or even infeasible for high T and multiple processes considered. However, in those cases when there are just two processes and T is moderate, this idea is entirely practical.

1.2.7 Change point detection algorithm

A simple change point detection algorithm can also be developed as illustrated in Algorithm 2. The process starts with a single data vector observed at time point 1. Then, new data vectors come from the examined process one by one. Based on the available data, BIC values $BIC_T^{<0>}$ (assuming no change points) and $BIC_T^{<1>}$ (assuming a change point at time T) are calculated. As soon as $BIC_T^{<1>}$ becomes smaller than $BIC_T^{<0>}$, the process is terminated since a change point is detected at time T.

Data: $\boldsymbol{Y} = (y_{i1})_{N\times 1}$
Initialization: $T \leftarrow 1$
repeat
  $T \leftarrow T + 1$;
  Obtain new data point $(y_{iT})_{N\times 1}$;
  Update data $\boldsymbol{Y} \leftarrow \{\boldsymbol{Y}, (y_{iT})_{N\times 1}\}$;
until $BIC_T^{<1>} < BIC_T^{<0>}$;
Result: T, $BIC_T^{<1>}$

Algorithm 2: Change point detection algorithm.

1.3 Experiments

In this section, we consider several simulation studies devoted to the rigorous evaluation of the proposed methodology. In Section 1.3.1, we investigate the performance of the change point estimation Algorithm 1. In Section 1.3.2, there is a study concerned with the performance of the methodology under misspecified model. Finally, Section 1.3.3 provides details of change point detection experiments.

1.3.1 Change point estimation experiments

In this section, we investigate the performance of the methodology for change points estimation in various settings. For illustrative purposes, we consider a 2^2 factorial design (*i.e.*, two bilevel factors). Simulated data sets consist of 200 realizations (50 in each treatment group) observed over 10 time points. Thus, $N = 200$, $M = 4$, and $T = 10$. Table 1.3.1 contains the parameters of the model considered. As we can see from the table, the vectors of coefficient $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_2$ are rather similar. We vary the magnitude of σ^2 to see the effect of increasing variability on the change point estimation. In our experiments, we consider $\sigma^2 = 0.1, 0.5, 1.0$. Also, we study the effect of the correlation coefficient ϕ. For this purpose, we choose $\phi = 0.1, 0.5, 0.9$. The skewness parameter λ is equal to 0.5. Finally, the ratio of variance components η is chosen to be 0.2.

$\boldsymbol{\beta}_0$	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	η	λ	σ^2	ϕ
$(10, 1.5, -2.3, 1.7)^\top$	$(10, 1.7, -2.2, 2.1)^\top$	$(11, 1.5, -2.0, 1.9)^\top$	0.2	0.5	$0.1, 0.5, 1.0$	$0.1, 0.5, 0.9$

Table 1.3.1: Parameters of the model considered in experiments of Section 1.3.1.

Six different settings that differ in the number and location of change points are studied under varying complexity conditions as reflected by parameters σ^2 and ϕ. We study the following cases: (a) $K = 0$ (*i.e.*, no change point), (b) $K = 1$, $t_1 = 2$, (c) $K = 1$, $t_1 = 6$, (d) $K = 2$, $t_1 = 2$, $t_2 = 3$, (e) $K = 2$, $t_1 = 2$, $t_2 = 6$, (f) $K = 2$, $t_1 = 4$, $t_2 = 7$. As we can see, even some cases with the same number of change points K are substantially different. For example, (b) is more complicated than (c) since there is just one time point corresponding to the first process (and 9 to the second), while in the latter case both processes are equally represented with 5 time points. Cases (d), (e), and (f) have similar distinctions, with (d) being the most complicated and (f) being the easiest.

Table 1.3.2 contains the results of experiments presented in the form of change point sampling distributions constructed based on 500 simulated data sets in each case. Values provided in the bold font, illustrate the numbers of correct estimation cases. The first case with $K = 0$ shows that the procedure does not tend to detect false positives. In fact, in all six settings considered, the methodology does not overestimate K. Overall, the smaller σ^2 is, the better results are. This trend is well expected and observed for all five setting (b)-(f), where change points are present. It can be noted that higher correlation values ϕ also lead to better performance of the procedure. In some cases, even when σ^2 is low, small values of ϕ can pose a considerable challenge. For example, in case (d) with $\sigma^2 = 0.1$, there are just 148 correct cases of change point estimation when $\phi = 0.1$, while it increases up to 500 when $\phi = 0.9$. Another important remark can be made with regard to the estimation process in various settings. As we can see, case (c) is considerably easier than (b), especially for $\sigma^2 = 0.1$. This happens because data are more evenly distributed among

(a) $K = 0$		$\sigma^2 = 1$			$\sigma^2 = 0.5$			$\sigma^2 = 0.1$		
t_1	t_2	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$
–	**–**	**500**	**500**	**500**	**500**	**500**	**500**	**500**	**500**	**499**

(b) $K = 1$		$\sigma^2 = 1$			$\sigma^2 = 0.5$			$\sigma^2 = 0.1$		
t_1	t_2	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$
–	–	499	498	433	494	490	180	220	88	–
2	**–**	**1**	**2**	**67**	**6**	**10**	**320**	**279**	**410**	**500**

(c) $K = 1$		$\sigma^2 = 1$			$\sigma^2 = 0.5$			$\sigma^2 = 0.1$		
t_1	t_2	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$
–	–	492	496	414	440	473	181	1	12	–
5	–	1	–	–	4	–	–	**2**	–	–
6	**–**	**5**	**4**	**86**	**53**	**26**	**318**	**495**	**485**	**500**

(d) $K = 2$		$\sigma^2 = 1$			$\sigma^2 = 0.5$			$\sigma^2 = 0.1$		
t_1	t_2	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$
–	–	341	378	7	66	111	119	–	–	–
3	–	152	122	388	428	381	–	352	89	–
4	–	4	–	–	–	–	–	–	–	–
2	**3**	**–**	**–**	**105**	**4**	**8**	**381**	**148**	**411**	**500**

(e) $K = 2$		$\sigma^2 = 1$			$\sigma^2 = 0.5$			$\sigma^2 = 0.1$		
t_1	t_2	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$
–	–	161	300	1	5	46	–	–	–	–
6	–	331	197	387	479	440	113	203	64	–
7	–	3	1	–	–	–	–	–	–	–
2	**6**	**2**	**2**	**112**	**11**	**14**	**387**	**296**	**436**	**500**

(f) $K = 2$		$\sigma^2 = 1$			$\sigma^2 = 0.5$			$\sigma^2 = 0.1$		
t_1	t_2	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$
–	–	215	335	7	17	68	–	–	–	–
6	–	4	–	–	–	–	–	–	–	–
7	–	273	164	359	451	399	119	36	29	–
4	**7**	**5**	**1**	**134**	**26**	**33**	**381**	**462**	**471**	**500**

Table 1.3.2: Change point sampling distributions in six different settings described in Section 1.3.1.

the processes in case (c) than in case (b), which leads to better estimation results. Similarly, we can observe that case (e) is better than (d) and (f) outperforms both of them. Interestingly, there are numerous cases of detecting correctly at least one change point in cases (d)-(f). Overall, we can conclude that the developed procedure is capable of estimating change points in various settings but can be affected by specific values of variance-related parameters as well as the allocation of change points among T time points.

In the final experiment of this section, we consider the most general change point estimation setting assuming that two processes can appear in any order at any time within T time points. As we discussed in Section 1.2.6, all possible permutations of these processes need to be considered. In our experiment, we assume that the first process is observed at all time points except times $t_1 = 2$ and $t_2 = 6$. This leads to vectors $\boldsymbol{b}_0 = (1, 0, 1, 1, 1, 0, 1, 1, 1, 1)^\top$ and $\boldsymbol{b}_1 = \mathbf{1}_{10} - \boldsymbol{b}_0$.

$K=2$		$\sigma^2=1$			$\sigma^2=0.5$			$\sigma^2=0.1$		
t_1	t_2	$\phi=0.1$	$\phi=0.5$	$\phi=0.9$	$\phi=0.1$	$\phi=0.5$	$\phi=0.9$	$\phi=0.1$	$\phi=0.5$	$\phi=0.9$
2	**6**	**0.014**	**0.048**	**0.842**	**0.064**	**0.204**	**0.992**	**0.904**	**0.992**	**1.000**

Table 1.3.3: Proportion of times the correct combination of change points is found in Section 1.3.1.

Table 1.3.3 provides the results of the simulation study based on 500 simulated datasets. Each table cell represents the proportion of times the correct model have been identified. As we can see, the procedure is very efficient for $\sigma^2 = 0.1$ or $\phi = 0.9$. This finding is consistent with that corresponding to Table 1.3.2. The increase in variability leads to considerable reductions in the number of correct models detected.

1.3.2 Change point estimation under model misspecification

In this section, we investigate the performance of the proposed algorithm under model misspecification. It can be noted that due to the presence of the skewness parameter λ, the proposed model is rather robust for model deviations associated with data rows. This happens because the exponential transformation automatically leads to near-normality consequently providing great modeling flexibility of Manly distributions. On the other hand, the effect of deviations from AR_1 model assumed for modeling columns needs to be studied.

In the following set of experiments, we simulate data assuming the same set of parameters as in Table 1.3.1 with the exception that instead of AR_1, the first order moving average (MA_1) dependence structure is associated with data columns. The moving average coefficient ψ is chosen to be $\psi = 0.1, 0.5, 0.9$. The experiments, however, are conducted assuming the original AR_1 relationship for columns.

Table 1.3.4 presents results of the outlined simulation study. As we can see, the impact of the variance σ^2 has become more severe. Reasonably good results are observed only for $\sigma^2 = 0.1$. However, in the most challenging situation (c), just one change point is found in the majority

(a) $K=0$		$\sigma^2=1$			$\sigma^2=0.5$			$\sigma^2=0.1$		
t_1	t_2	$\psi=0.1$	$\psi=0.5$	$\psi=0.9$	$\psi=0.1$	$\psi=0.5$	$\psi=0.9$	$\psi=0.1$	$\psi=0.5$	$\psi=0.9$
–	**–**	**500**	**500**	**499**	**500**	**500**	**500**	**500**	**500**	**500**

(b) $K=1$		$\sigma^2=1$			$\sigma^2=0.5$			$\sigma^2=0.1$		
t_1	t_2	$\psi=0.1$	$\psi=0.5$	$\psi=0.9$	$\psi=0.1$	$\psi=0.5$	$\psi=0.9$	$\psi=0.1$	$\psi=0.5$	$\psi=0.9$
–	–	499	498	499	491	495	495	226	195	203
2	**–**	**1**	**2**	**1**	**8**	**5**	**4**	**273**	**298**	**290**

(c) $K=1$		$\sigma^2=1$			$\sigma^2=0.5$			$\sigma^2=0.1$		
t_1	t_2	$\psi=0.1$	$\psi=0.5$	$\psi=0.9$	$\psi=0.1$	$\psi=0.5$	$\psi=0.9$	$\psi=0.1$	$\psi=0.5$	$\psi=0.9$
–	–	478	459	454	375	274	299	–	–	–
4	–	–	–	3	–	–	–	–	–	–
5	–	1	5	10	7	15	13	1	–	–
6	**–**	**20**	**34**	**28**	**112**	**194**	**171**	**499**	**500**	**500**
7	–	1	2	4	5	14	15	–	–	–

(d) $K=2$		$\sigma^2=1$			$\sigma^2=0.5$			$\sigma^2=0.1$		
t_1	t_2	$\psi=0.1$	$\psi=0.5$	$\psi=0.9$	$\psi=0.1$	$\psi=0.5$	$\psi=0.9$	$\psi=0.1$	$\psi=0.5$	$\psi=0.9$
–	–	242	110	145	23	1	4	–	–	–
2	–	1	6	2	1	–	1	–	–	–
3	–	255	372	344	472	496	492	419	438	438
4	–	2	10	9	1	1	1	–	–	–
2	**3**	**–**	**1**	–	**3**	**1**	**1**	**81**	**62**	**62**

(e) $K=2$		$\sigma^2=1$			$\sigma^2=0.5$			$\sigma^2=0.1$		
t_1	t_2	$\psi=0.1$	$\psi=0.5$	$\psi=0.9$	$\psi=0.1$	$\psi=0.5$	$\psi=0.9$	$\psi=0.1$	$\psi=0.5$	$\psi=0.9$
–	–	64	1	1	–	–	–	–	–	–
5	–	12	4	7	–	–	–	–	–	–
6	–	414	479	478	490	490	490	207	196	200
7	–	7	12	8	1	–	–	–	–	–
2	**6**	**1**	**3**	**–**	**8**	**9**	**9**	**292**	**303**	**299**

(f) $K=2$		$\sigma^2=1$			$\sigma^2=0.5$			$\sigma^2=0.1$		
t_1	t_2	$\psi=0.1$	$\psi=0.5$	$\psi=0.9$	$\psi=0.1$	$\psi=0.5$	$\psi=0.9$	$\psi=0.1$	$\psi=0.5$	$\psi=0.9$
–	–	113	21	29	2	–	–	–	–	–
5	–	3	–	1	–	–	–	–	–	–
6	–	8	12	11	–	–	–	–	–	–
7	–	367	439	434	440	396	409	–	–	1
8	–	5	4	4	–	–	–	–	–	–
3	7	–	3	2	5	10	9	2	3	3
4	**7**	**1**	**17**	**15**	**51**	**88**	**77**	**492**	**496**	**495**
5	7	1	3	3	2	6	5	–	1	1

Table 1.3.4: Change point sampling distributions in six different settings from Section 1.3.2.

of cases even for the lowest σ^2. The effect of a specific value of ψ can be seen only in cases (c) and (f), *i.e.*, when change points are distributed evenly among T time points. In these situations, the results are slightly better for higher ψ values. Overall, we can conclude that the considered model misspecification does not pose serious problems if the variance is small and change points are not distributed very unevenly over T time points.

1.3.3 Change point detection experiments

In our final series of experiments, we study the change point detection algorithm outlined in Section 1.2.7. We consider cases with (a) no change point and one change point at (b) $t_1 =$

2 as well as (c) $t_1 = 6$. Table 1.3.5 summarizes the obtained results. In case (a), we can note that the specific variance level has the minimal effect on observed sampling distributions. The same statement is true with regard to the parameter ϕ. The case (b) is considerably easier than (c) for the detection of change points. In particular, when $\sigma^2 = 1$, results in (a) and (c) are very similar due to very high variability in data. For smaller values of σ^2, results tend to improve for (c). At the same time, case (b) does not present such a severe challenge as (c). As we can see, even when $\sigma^2 = 1$, the procedure detects change points rather satisfactorily, especially for highly correlated matrix columns, *i.e.*, when $\phi = 0.9$. Thus, we can conclude that change points are easier to detect if they happen sooner. In this case, they can be successfully detected even for data with high variability.

(a) $K = 0$	$\sigma^2 = 1$			$\sigma^2 = 0.5$			$\sigma^2 = 0.1$		
t_1	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$
–	**351**	**391**	**374**	**343**	**381**	**376**	**365**	**358**	**366**
2	127	98	109	135	106	117	117	127	117
3	20	4	13	19	9	6	15	15	16
4	2	5	4	2	2	1	1	–	1
5	–	1	–	1	1	–	2	–	–
6	–	1	–	–	1	–	–	–	–
(b) $K = 1$	$\sigma^2 = 1$			$\sigma^2 = 0.5$			$\sigma^2 = 0.1$		
t_1	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$
–	232	193	7	171	72	–	2	1	–
2	**262**	**304**	**493**	**325**	**426**	**500**	**498**	**499**	**500**
3	6	3	–	4	2	–	–	–	–
(c) $K = 1$	$\sigma^2 = 1$			$\sigma^2 = 0.5$			$\sigma^2 = 0.1$		
t_1	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$
–	363	374	270	368	357	106	147	59	–
2	120	97	126	106	111	112	132	124	126
3	11	18	18	12	11	11	5	10	15
4	3	1	5	2	1	2	4	8	2
5	1	–	1	3	–	–	–	1	1
6	**2**	**10**	**80**	**9**	**20**	**269**	**212**	**298**	**356**

Table 1.3.5: Change point detection results for experiments from Section 1.3.3.

1.4 Application

In this section, we illustrate the proposed methodology on crime rates data obtained at the US Department of Justice, FBI web-site (http://www.ucrdatatool.gov/Search/Crime/Crime.cfm). In particular, we study two types of crime: burglary and motor vehicle theft. We study the behav-

	$\hat{\boldsymbol{\beta}}_0$	$\hat{\boldsymbol{\beta}}_1$	$\hat{\eta}$	$\hat{\lambda}$	$\hat{\sigma}^2$	$\hat{\phi}$
(a)	$(4.67, 4.11, 4.06, 3.64, 4.30)^\top$	–	0.379	-1	0.032	0.970
(b)	$(5.70, 7.19, 5.77, 6.93, 8.28)^\top$	$(5.97, 6.99, 5.48, 6.52, 7.81)^\top$	0.010	-0.778	0.034	0.967

Table 1.4.1: Parameter estimates for models analyzing (a) Motor Vehicle Theft and (b) Burglary.

ior of these characteristics over the 13-year time period (2000-2012) in five commonly referred regions in the United States: *West*, *MidWest*, *NorthEast*, *SouthWest*, and *SouthEast*. Figure 1.4.1 illustrates the division of states among the five regions. In each region, we obtained crime rates for 25 most populated cities with data available. Thus, in our study, $N = 125$, $M = 5$, and $T = 13$.

Due to the relatively small number of time points, we can apply the estimation procedure that considers all possible permutations of two processes. For the variable *Motor Vehicle Theft*, no change point has been found. Parameter estimates are presented in Table 1.4.1 row (a). As we can see, the lowest mean rate is observed for *SouthWest* (3.64 per 1000 people), while the highest one is at *West* (4.67). For the variable *Burglary*, there was a change point observed in the last year considered, 2012. The results can be found in row (b) of Table 1.4.1. As we can see, there was a considerable rate change in all regions. In particular, in all regions except *West*, we observe the decrease in *Burglary* rates. The largest change is detected for *SouthWest* (-0.41) and *SouthEast* (-47). The two safest regions in terms of *Burglary* are *West* and *NorthEast*. The highest rates are associated with *SouthEast*: 8.28 before 2012 and 7.81 after that.

As a final remark, we can notice that $\hat{\phi}$ values are close to 1 in both situations. At the same time, $\hat{\sigma}^2$ estimates are very small. According to our findings in the simulation studies, the change point estimation procedure in these settings is remarkably accurate.

1.5 Discussion

In this paper, we proposed a novel approach to estimating and detecting various change points in processes monitored for multiple subjects. For this purpose, a matrix normal distribu-

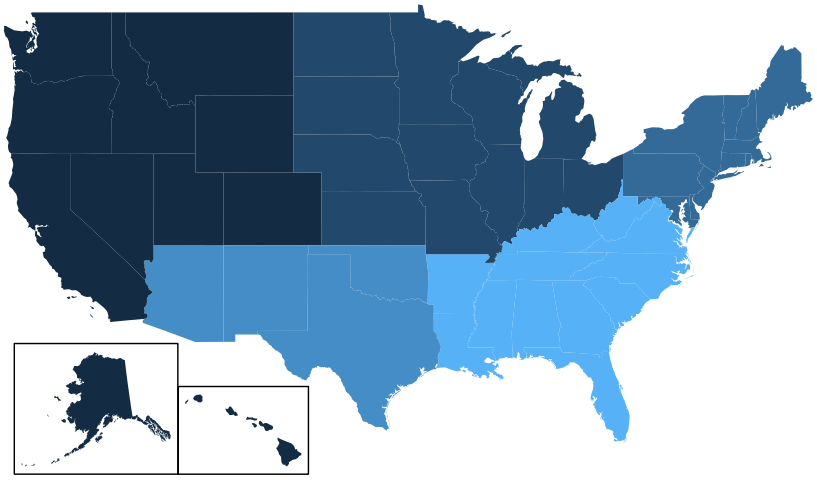


Figure 1.4.1: Five regions (*West*, *MidWest*, *NorthEast*, *SouthWest*, and *SouthEast*) considered in Section 3.3.

tion was employed. The convenience of this setting is justified by the structure of data in the matrix form with columns and rows representing time points and subjects, respectively. To make the developed procedure robust to deviations from normality, the skewness parameter originating from the exponential transformation is implemented into the model. The developed procedure is tested in various settings for change point estimation and detection, including cases with misspecified covariance structure. The application of the proposed methodology to the analysis of crime rate data in five US regions is also considered.

CHAPTER 2

ROBUST ESTIMATION OF MULTIPLE CHANGE POINTS IN MULTIVARIATE PROCESSES

Change point inference is important in various fields of science. Many different procedures have been proposed in literature but most of them rely on some restrictive assumptions such as the normality of underlying processes or independence of observations. In this paper, a novel likelihood-based technique is proposed. It provides a way to model various covariance patterns and is robust to skewness observed in data. Through simulation studies, we demonstrate that the proposed procedure is superior over some of its competitors. The application of the methodology to real-life datasets highlights its usefulness and broad applicability.

2.1 Introduction

The change point estimation in sequential data have become an important task in many areas of active research. It assumes the existence of at least two different processes observed over some time interval. Since the specific times associated with each process are typically unknown, they have to be estimated along with the processes themselves. The applications of change point estimation procedures can be found in medicine [18], ecology [29], pharmacy [3], engineering [26], finance [22, 30], and many other fields. The problem of process and change point estimation is also known as phase I in statistical process control. Then, phase II would deal with the detection of changes in a process flow based on the already estimated processes.

Researchers have been exploring change point problems for decades but there are still

many questions that remain open. One of the earliest papers on the subject was devoted to the estimation of a change point in means of univariate normal distributions [28]. The problem with a constant mean but possible shift in variance parameters was considered by [15, 9, 16, 5]. A generalization of both ideas was considered by [14] who developed a test capable of detecting a change in mean and variance parameters simultaneously.

Attention have been paid to multivariate settings as well. [38] and [39] considered the framework with a single change point in mean vectors of multivariate normal distributions. Soon after that, the estimation of multiple change points in mean vectors was studied by [44] and [45]. In the same setting of multivariate normal distribution, [6] proposed a procedure for estimating a change in covariance matrices under the assumption of a constant mean vector. Recently, [7] developed a test for estimating change points in mean vectors and covariance matrices simultaneously, thus generalizing the above-listed ideas. Other directions of research in the area of change point estimation include inference for the general exponential family [32, 27], nonparametrics methods [33] including probabilistic pruning based on various goodness-of-fit measures [17], and some others.

In this paper, we consider the problem of estimating multiple change points in the framework with multivariate processes. For this problem, we employ a matrix normal distribution. Due to its form, one can model the covariance structure associated not just with variables (given by matrix rows) or time points (provided by matrix columns), but also the overall covariance structure associated with variables and times. This effectively eliminates some of the common restrictive assumptions such as the independence of observations at different time points. To make the proposed procedure more robust to deviations from normality, we propose incorporating one of several available transformations to near-normality. As a result, the proposed procedure gains robustness features while being capable of accommodating various covariance structures in data.

The rest of the paper is organized as follows below. Section 2.2 presents the proposed methodology. Section 2.3 investigates the performance of our procedure and three competitors in various settings. Section 3.3 applies the developed methods to the analysis of real-life data. The paper concludes with a discussion provided in Section 3.4.

2.2 Methodology

2.2.1 Matrix normal distribution

Let $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T$ be a process observed over T time points with each $\boldsymbol{y}_i$ following a p-variate normal distribution. The entire dataset can be conveniently summarized in the matrix form as shown below

$$\boldsymbol{Y} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1T} \\ y_{21} & y_{22} & \cdots & y_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ y_{p1} & y_{p2} & \cdots & y_{pT} \end{pmatrix}. \tag{2.2.1}$$

Here, each row represents a particular variable observed over time, while every column stands for a p-variate measurement at a specific time point. The overall variability associated with $\boldsymbol{Y}$ can often be explained by the variation observed in rows and columns. This leads to the idea of modeling the variability corresponding to p variables separately from that associated with T time points.

One distribution that can be effectively applied in the considered framework is a so-called matrix normal one [20] that has the following probability density function (pdf):

$$\phi_{p\times T}(\boldsymbol{Y}; \boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}) = (2\pi)^{-\frac{pT}{2}} |\boldsymbol{\Sigma}|^{-\frac{T}{2}} |\boldsymbol{\Psi}|^{-\frac{p}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{M})\boldsymbol{\Psi}^{-1}(\boldsymbol{Y}-\boldsymbol{M})^\top\right\}\right\}, \tag{2.2.2}$$

where $\boldsymbol{Y}$ is the $p \times T$ matrix argument defined in (2.2.1) and $\boldsymbol{M}$ is a $p \times T$ mean matrix. The

$p \times p$ matrix $\mathbf{\Sigma}$ and $T \times T$ matrix $\mathbf{\Psi}$ are covariance matrices that model variability associated with rows and columns, respectively. Also, tr$\{\cdot\}$ denotes the trace operator. It can be shown that vec$(\boldsymbol{Y}) \sim \mathcal{N}_{pT}(\text{vec}(\boldsymbol{M}), \mathbf{\Psi} \otimes \mathbf{\Sigma})$, where vec$(\cdot)$ denotes the vectorization operator that stacks matrix columns on top of each other, $\otimes$ is the Kronecker product, and $\mathcal{N}_{pT}$ is the pT-variate normal distribution with mean vector vec$(\boldsymbol{M})$ and covariance matrix $\mathbf{\Psi} \otimes \mathbf{\Sigma}$. There is a minor non-identifiability issue caused by the properties of the Kronecker product since $a\mathbf{\Psi} \otimes \mathbf{\Sigma} = \mathbf{\Psi} \otimes a\mathbf{\Sigma}$ for any multiplier $a \in I\!R^{+}$. One simple restriction on $\mathbf{\Psi}$ or $\mathbf{\Sigma}$ can effectively resolve this problem. The main advantage of taking into account the matrix data structure is the ability to reduce the number of parameters to $T(T+1)/2 + p(p+1)/2 - 1$ from $pT(pT+1)/2$ in the case of the most general covariance matrix. Hence, the proposed model effectively addresses a potential overparameterization issue while still allowing non-zero covariances $Cov(y_{jt}, y_{j't'})$ for any variables j and j' at time points t and t'.

As the specific problem considered in our setting deals with vectors observed over time, matrix $\mathbf{\Psi}$ can be conveniently parameterized in terms of a desired time series process. In this paper, we assume the autoregressive process of order 1 (AR(1)) in all experiments and applications. Under this setting, the covariance matrix $\mathbf{\Psi}$ is given by

$$\mathbf{\Psi} = \frac{\delta^2}{1-\phi^2} \begin{pmatrix} 1 & \phi & \phi^2 & \dots & \phi^{T-1} \\ \phi & 1 & \phi & \dots & \phi^{T-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{T-1} & \phi^{T-2} & \phi^{T-3} & \dots & 1 \end{pmatrix},$$

where ϕ is the correlation coefficient and δ^2 is the variance parameter. Then, one convenient constraint to avoid the non-identifiability issue associated with $\mathbf{\Psi} \otimes \mathbf{\Sigma}$ is to set $\delta^2 = 1 - \phi^2$. This restriction immediately leads to $\mathbf{\Psi} \equiv \boldsymbol{R}_\phi$, where $\boldsymbol{R}_\phi$ denotes the corresponding correlation ma-

trix that relies on a single parameter ϕ. It can be shown that

$$|\boldsymbol{\Psi}| \equiv |\boldsymbol{R}_\phi| = (1-\phi^2)^{T-1} \quad \text{and} \quad \boldsymbol{\Psi}^{-1} \equiv \boldsymbol{R}_\phi^{-1} = \frac{1}{1-\phi^2}(\boldsymbol{I}_T - \phi\boldsymbol{J}_1 + \phi^2\boldsymbol{J}_2), \tag{2.2.3}$$

where $\boldsymbol{J}_1$ and $\boldsymbol{J}_2$ are $T \times T$ matrices defined as follows below:

$$\boldsymbol{J}_1 = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 1 & 0 & 1 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & 0 & \dots & 1 & 0 \end{pmatrix} \quad \text{and} \quad \boldsymbol{J}_2 = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 \end{pmatrix}.$$

Expressions in (3.2.7) are helpful for speedier maximum likelihood estimation as the potentially time consuming inversion of the $T \times T$ covariance matrix $\boldsymbol{\Psi}$ can be completely avoided.

2.2.2 Change point estimation

Consider the problem of estimating change points in the given framework. Let $\boldsymbol{\mu}_0$ be the p-variate mean vector associated with the main process. Suppose, there are K alternative processes with means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_K$. The mean matrix $\boldsymbol{M}$ can be written as $\boldsymbol{M} = \sum_{k=0}^{K} \boldsymbol{\mu}_k \boldsymbol{m}_k^\top$, where $\boldsymbol{m}_k$ ($k = 0, 1, \dots, K$) is the vector of length T consisting of zeros and ones, with ones being located in those positions where the k^{th} process is observed. From the definition, it follows that $\sum_{k=0}^{K} \boldsymbol{m}_k = \mathbf{1}_T$, where $\mathbf{1}_T$ is the vector of length T with all elements equal to 1. It can be noted that vectors $\boldsymbol{m}_k$ can present various permutations of zeros and ones. However, in the case

of K shift change points at times $t_1, t_2, \ldots, t_K$, the mean matrix is given by

$$\boldsymbol{M} = \left(\underbrace{\boldsymbol{\mu}_0, \ldots, \boldsymbol{\mu}_0}_{t_1-1}, \underbrace{\boldsymbol{\mu}_1 \ldots, \boldsymbol{\mu}_1}_{t_2-t_1}, \ldots, \underbrace{\boldsymbol{\mu}_{K-1}, \ldots, \boldsymbol{\mu}_{K-1}}_{t_K-t_{K-1}}, \underbrace{\boldsymbol{\mu}_K, \ldots, \boldsymbol{\mu}_K}_{T-t_K+1} \right).$$

Also, $\boldsymbol{m}_k = \left(\underbrace{\mathbf{0}, \ldots, \mathbf{0}}_{t_k-1}, \underbrace{\mathbf{1}, \ldots, \mathbf{1}}_{t_{k+1}-t_k}, \underbrace{\mathbf{0}, \ldots, \mathbf{0}}_{T-t_{k+1}+1} \right)$ with boundary conditions $t_0 = 1$ and $t_{K+1} = T + 1$. As a result of such parameterization, the mean matrix $\boldsymbol{M}$ involves $p(K + 1)$ parameters.

The log-likelihood function corresponding to Equation (2.2.2) has the following form:

$$\begin{aligned} \log \mathcal{L}(\boldsymbol{Y}; \boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}) = &- \frac{pT}{2} \log(2\pi) - \frac{T}{2} \log |\boldsymbol{\Sigma}| - \frac{p}{2} \log |\boldsymbol{\Psi}| \\ &- \frac{1}{2} \text{tr} \left\{ \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y} - \boldsymbol{M}) \boldsymbol{\Psi}^{-1} (\boldsymbol{Y} - \boldsymbol{M})^\top \right\}. \end{aligned}$$

Oftentimes, the normality assumption is not adequate and inference based on such a model may be incorrect or misleading. One possible treatment of such a situation is to employ a transformation to near-normality. Incorporating a transformation into the model makes it considerably more robust to possible violations of the normality assumption. Several immediate candidates include the famous power transformation proposed by [4], alternative families of power transformations as in [43], or the the exponential transformation proposed by [24]. Let $\mathcal{T}$ represent the transformation operator such that $\mathcal{T}(y; \lambda)$ is approximately normally distributed upon the appropriate choice of the transformation parameter λ. In the p-variate setting, the traditional assumption is that the coordinatewise transformation leads to the joint near-normality [2, 23, 46], *i.e.*, the p-variate transformation is given by $\mathcal{T}(\boldsymbol{y}; \boldsymbol{\lambda}) = (\mathcal{T}(y_1; \lambda_1), \mathcal{T}(y_2; \lambda_2), \ldots, \mathcal{T}(y_p; \lambda_p))^\top$, where the transformation parameter vector is given by $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_p)^\top$. This idea can be readily generalized to the matrix framework with $\mathcal{T}(\boldsymbol{Y}; \boldsymbol{\lambda})$ representing data transformed to matrix near-normality based on the p-variate vector $\boldsymbol{\lambda}$.

Taking into account the special forms of $\boldsymbol{\Psi}$ and $\boldsymbol{M}$ and implementing the transformation idea, the log-likelihood function can be further written as

$$\log \mathcal{L}(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}, \phi, \boldsymbol{\lambda}) = -\frac{pT}{2}\log(2\pi) - \frac{T}{2}\log|\boldsymbol{\Sigma}| - \frac{p(T-1)}{2}\log(1-\phi^2)$$
$$- \frac{1}{2(1-\phi^2)}\text{tr}\left\{\boldsymbol{\Sigma}^{-1}\big(\mathcal{T}(\boldsymbol{Y};\boldsymbol{\lambda}) - \sum_{k=0}^{K}\boldsymbol{\mu}_k\boldsymbol{m}_k^\top\big)\left(\boldsymbol{I}_T - \phi\boldsymbol{J}_1 + \phi^2\boldsymbol{J}_2\right)\big(\mathcal{T}(\boldsymbol{Y};\boldsymbol{\lambda}) - \sum_{k=0}^{K}\boldsymbol{\mu}_k\boldsymbol{m}_k^\top\big)^\top\right\}$$
$$+ \log\left|\frac{\partial\mathcal{T}(\boldsymbol{Y};\boldsymbol{\lambda})}{\partial\boldsymbol{Y}}\right|, \tag{2.2.4}$$

where the term $\log\left|\frac{\partial\mathcal{T}(\boldsymbol{Y};\boldsymbol{\lambda})}{\partial\boldsymbol{Y}}\right|$ represents the log of Jacobian associated with the transformation.

Maximum likelihood estimation leads to the following expressions for $\boldsymbol{\mu}_k$'s:

$$\boldsymbol{\mu}_k = \Big(\mathcal{T}(\boldsymbol{Y};\boldsymbol{\lambda}) - \sum_{\substack{k'=0\\k'\neq k}}^{K}\boldsymbol{\mu}_{k'}\boldsymbol{m}_{k'}^\top\Big)\boldsymbol{R}_\phi^{-1}\boldsymbol{m}_k\left(\boldsymbol{m}_k^\top\boldsymbol{R}_\phi^{-1}\boldsymbol{m}_k\right)^{-1},$$

where $\boldsymbol{R}_\phi^{-1}$ is as in (3.2.7). Solving a system of $K+1$ equations leads to expressions for each $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$. Maximum likelihood estimation for $\boldsymbol{\Sigma}_k$ yields the following expression:

$$\boldsymbol{\Sigma} = \frac{(\mathcal{T}(\boldsymbol{Y};\boldsymbol{\lambda}) - \sum_{k=0}^{K}\boldsymbol{\mu}_k\boldsymbol{m}_k^\top)\boldsymbol{R}_\phi^{-1}(\mathcal{T}(\boldsymbol{Y};\boldsymbol{\lambda}) - \sum_{k=0}^{K}\boldsymbol{\mu}_k\boldsymbol{m}_k^\top)^\top}{T}.$$

Substituting expressions for $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$ and $\boldsymbol{\Sigma}$ into the log-likelihood function (2.2.4) makes the log-likelihood a function of the parameters ϕ and $\boldsymbol{\lambda}$. The maximization with respect to these parameters can be done numerically using one of many available optimization algorithms.

For the purpose of illustration, in this paper we focus on the exponential transformation of Manly given by $\mathcal{T}(y;\lambda) = y^{I(\lambda=0)}\left(\exp\{\lambda y - 1\}\lambda^{-1}\right)^{I(\lambda\neq 0)}$, where $I(\cdot)$ is the indicator function. In this setting, the log of Jacobian in (2.2.4) is given by $\boldsymbol{\lambda}^\top\boldsymbol{Y}\boldsymbol{1}_T$, where $\boldsymbol{1}_T = (1, 1, \ldots, 1)^\top$ with

cardinality $|\boldsymbol{\lambda}| = T$.

2.2.3 Model selection

In the previous section, we discussed how the maximum likelihood estimates of model parameters can be obtained efficiently. However, the problem of change point estimation requires assessing a number of models assuming change points at different times. To avoid potential problems with the adjustment for multiple comparisons, simplify calculations, and avoid testing procedures in general, we employ Bayesian Information Criterion (BIC) [37]. BIC is also an appealing option due to its connection to the Bayes factor commonly used in Bayesian inference for comparing competing models.

2.3 Experiments

In this section, we consider simulation studies devoted to the rigorous evaluation of the proposed methodology. We investigate the performance of the change point estimation procedure in two general settings. In both cases, we assume the existence of three processes observed over 100 time points. In the first case, the first process is observed until the change point at $t_1 = 10$, when the second process starts. Then, the second process runs until the next change point at $t_2 = 20$, when the third process starts and runs for the remaining time. In the second setting, the change points are set to be at times $t_1 = 10$ and $t_2 = 50$. The difference between these two settings is that in the first situation, the first two processes are observed for a relatively short period of time, while the third process is observed for much longer. On the contrary, in the second experiment setting, just the first process is observed for a short period of time as opposed to the other two processes. The parameters used in the simulation study are provided in Table 2.3.1. Various levels of correlation and scaling as reflected by parameters ϕ and $\boldsymbol{\Sigma}$, respectively, are studied. In particular, we consider $\phi = 0.1, 0.5, 0.9$ and $\boldsymbol{\Sigma}, \boldsymbol{\Sigma}/2, \boldsymbol{\Sigma}/4$. 250 datasets were simu-

Table 2.3.1: Parameter values used in the simulation study of Section 2.3.

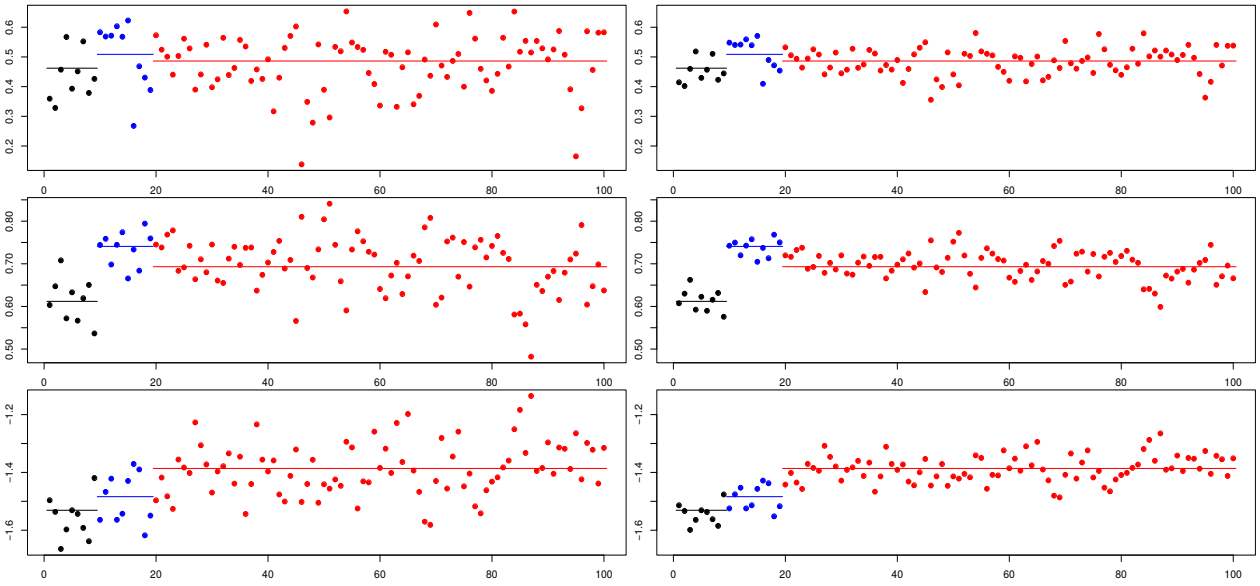
j	$\boldsymbol{\mu}_0$	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}$			$\boldsymbol{\lambda}$	ϕ
1	1	1.2	1.1	0.133	-0.033	0	3	$\{0.1, 0.5, 0.9\}$
2	1.2	1.7	1.5	-0.033	0.067	-0.033	2	
3	-2.3	-2.2	-2.0	0	-0.033	0.033	-0.5	

lated for each combination of the covariance matrix and correlation parameter in both considered setting, thus, yielding 4,500 simulated datasets in total.

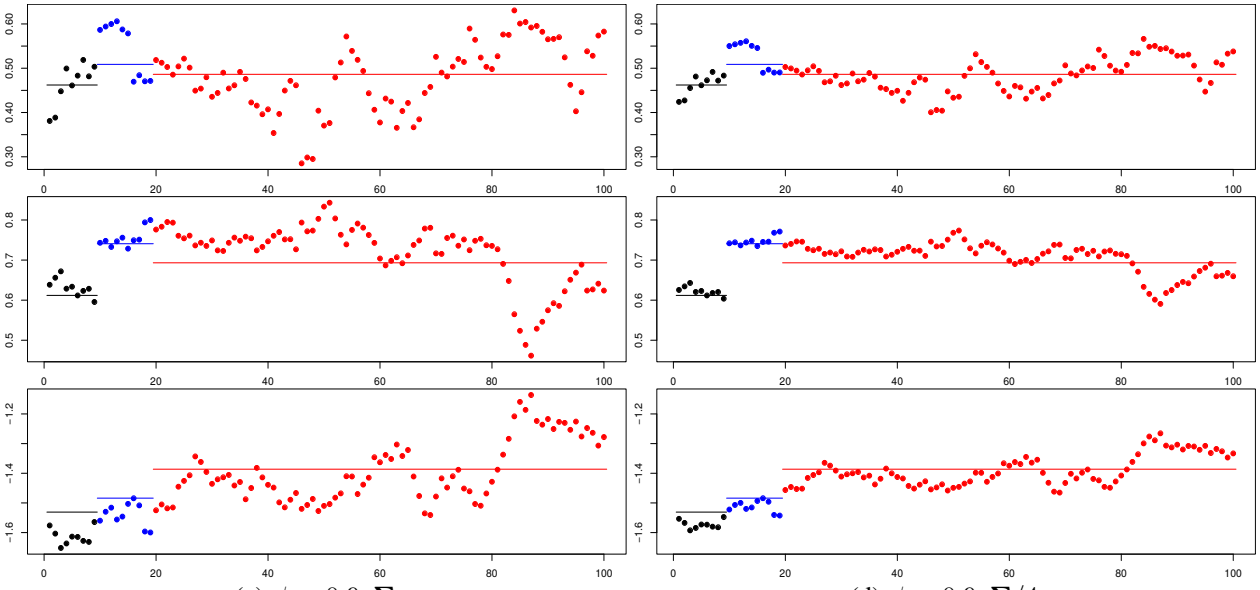
The illustration of some simulated datasets can be found in Figure 2.3.1. Here, plots (a) and (b) show datasets simulated with $\phi = 0.1$ but with different covariance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}/4$, respectively. Plots (c) and (d) correspond to the same covariance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}/4$ but with high correlation of $\phi = 0.9$. The four considered datasets represent the first setting with change points at $t_1 = 10$ and $t_2 = 20$. Within each of the four plots, there are three subplots representing the coordinatewise behavior of the processes reflected by means of the black, blue, and red colors. The top subplot corresponds to the first coordinate, the middle stands for the second one, and the bottom plot represents the third coordinate. Horizontal lines show the true back-transformed values of the corresponding coordinates of vectors $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, and $\boldsymbol{\mu}_2$.

From examining Figure 2.3.1, it is easy to conclude that the task of change point estimation is far from trivial in these cases. Especially in those cases when the variability is higher (left column of plots), we can observe a number of points that can be mistakenly thought of as change points. Thus, it is fully expected that false change points will be found oftentimes. Moreover, we can observe that the first change point should be considerably easier to find than the second one due to the substantial gap in the second coordinate of means related to the first two processes (*i.e.*, between black and blue horizontal lines).

As pointed out by [17], the number of procedures capable of estimating multiple change points in multivariate processes is rather limited. In this section, the developed methodology is



(a) $\phi = 0.1$, $\mathbf{\Sigma}$ (b) $\phi = 0.1$, $\mathbf{\Sigma}/4$



(c) $\phi = 0.9$, $\mathbf{\Sigma}$ (d) $\phi = 0.9$, $\mathbf{\Sigma}/4$

Figure 2.3.1: Datasets generated in the course of the simulation study in Section 2.3 with different scaling (reflected by $\mathbf{\Sigma}$ and $\mathbf{\Sigma}/4$) and correlation ($\phi = 0.1, 0.9$). Horizontal lines represent true back-transformed values of the corresponding coordinates of parameters $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, and $\boldsymbol{\mu}_2$.

compared with one parametric approach that we call naive and two nonparametric procedures available for practitioners through the R package ECP [17]. The naive method is mimicking the most common practical approach with all observations assumed independent and following multivariate normal processes. The two nonparametric procedures are based on probabilistic prun-

Table 2.3.2: Interpretation of notation used in Tables 2.3.3 and 2.3.4.

Notation	Interpretation
$\{t_1, t_2\}$	both change points are correctly found
$\{t_1, t_2, x\}$	both change poitns are correctly identified, but there are also false change points found
$\{t_1, \tilde{t}_2\}/\{\tilde{t}_1, t_2\}$	one change point is identified correctly, the other one is close by, *i.e.*, $0 < \lvert t_k - \hat{t}_k \rvert \leq 3$
$\{t_1\}/\{t_2\}$	one change point is identified correctly and it is the only one found
$\{t_1, !t_2\}/\{!t_1, t_2\}$	one change point is identified correctly, the others are not close, *i.e.*, $\lvert t_k - \hat{t}_k \rvert > 3$

ing with Energy statistic [35, 36] and Kolmogorov-Smirnov statistic [19] used as goodness-of-fit measures. Tables 2.3.3 and 2.3.4 provide the results of the simulation study in the first ($t_1 = 10$, $t_2 = 20$) and second ($t_1 = 10$, $t_2 = 50$) settings, respectively. The tables include proportions of times various solutions, as per description in Table 2.3.2, were found.

As we can observe from Table 2.3.3, the proposed method can rather effectively identify change points. Expectedly, the performance of the procedure improves considerably when the variabilty decreases. For example, in the case with $\phi = 0.9$ and $\mathbf{\Sigma}$, we are able to correctly identify the combination of change points in 14.8% of all cases. The percentage improves to 49.2% and 93.2% for $\mathbf{\Sigma}/2$ and $\mathbf{\Sigma}/4$, respectively. The performance of the procedure somewhat degrades for lower values of parameter ϕ. In particular, the correct setting was found in 63.2% and 55.6% of cases for $\mathbf{\Sigma}/4$ with $\phi = 0.1$ and $\phi = 0.5$, respectively. In the settings with higher variability, the task of estimating both change points correctly is considerably more difficult. It is worth mentioning that in these settings our procedure is capable of identifying at least one change point effectively. In particular, we can notice that there is a relatively low proportion of times when our method identified one point correctly and the other change point estimate was considerably off. Another observation can be made with regard to a low number of false change point detections made by our procedure. In addition, due to a strong penalty carried out by BIC, there is no tendency to overestimate the number of change points as we can see from the line $\{t_1, t_2, x\}$.

From examining Table 2.3.3, we can conclude that the closest competitor is the naive pro-

cedure. In particular, it demonstrates quite similar results in terms of the proportion of correct solutions for the majority of cases unless $\phi = 0.9$. When ϕ is high, the naive procedure is substantially outperformed by the proposed method in all settings. This observation is not surprising since the cases with lower correlations are more similar to the naive model assuming the independence of observations. Our developed method dramatically outperforms the two nonparametric methods. In the easiest case considered with $\phi = 0.9$ and $\mathbf{\Sigma}/4$, the probabilistic pruning with Energy statistic is capable of finding the correct combination of change points in 35.6% of cases. In all other cases, both procedures face considerable challenges. One can also notice that nonparametric methods struggle to find even one of the two change points correctly. In the case of $\mathbf{\Sigma}/4$, the Kolmogorov-Smirnov statistic shows some improvement for $\phi = 0.1$. It is able to estimate one change point correctly and the other one in close proximity to the true change point in 22.4% of all cases.

The inference drawn from Table 2.3.4 is mostly similar. In the meantime, we can notice that our method improves the performance in all cases. This happens due to the fact that the number of time points is more evenly distributed among the processes and thus more accurate estimation of parameters is possible. As a result, the difference between the proposed and naive approaches can now be observed for the case with $\mathbf{\Sigma}/4$ and $\phi = 0.9$. To conclude this section, we can remark that the proposed procedure proves to be a powerful tool for identifying multiple change points.

2.4 Applications

2.4.1 Illustration of crime rates in US cities

First, we apply the proposed methodology to the US cities crime data publicly available at the US Department of Justice, FBI Web-site (http://www.ucrdatatool.gov/Search/Crime/Crime.cfm).

Table 2.3.3: Simulation study from Section 2.3 assuming two change points at times $t_1 = 10$ and $t_2 = 20$. The four methods considered are our proposed procedure, naive procedure, and probabilistic pruning with Energy statistic and Kolmogorov-Smirnov statistic used as the goodness-of-fit measure. The notation interpretation is provided in Table 2.3.2. The bold font highlights the proportion of times the correct combination was found.

$K = 2$		$\mathbf{\Sigma}$			$\mathbf{\Sigma}/2$			$\mathbf{\Sigma}/4$		
$t_1 = 10, t_2 = 20$		$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$
Method	$\{10, 20\}$	**0.060**	**0.032**	**0.148**	**0.332**	**0.168**	**0.492**	**0.632**	**0.556**	**0.932**
	$\{10, 20, x\}$	0	0	0	0	0	0	0	0	0
	$\{10, \tilde{20}\}/\{\tilde{10}, 20\}$	0.200	0.084	0.012	0.336	0.168	0.016	0.304	0.160	0
	$\{10\}/\{20\}$	0.576	0.736	0.692	0.212	0.516	0.424	0.012	0.140	0.040
	$\{10, !20\}/\{!10, 20\}$	0.104	0.112	0.136	0.120	0.148	0.068	0.052	0.144	0.028
Naive	$\{10, 20\}$	**0.060**	**0.044**	**0.048**	**0.344**	**0.232**	**0.116**	**0.628**	**0.536**	**0.308**
	$\{10, 20, x\}$	0	0	0	0	0	0	0	0	0
	$\{10, \tilde{20}\}/\{\tilde{10}, 20\}$	0.188	0.192	0.028	0.362	0.224	0.048	0.308	0.252	0.056
	$\{10\}/\{20\}$	0.488	0.108	0	0.136	0.036	0	0.004	0.080	0
	$\{10, !20\}/\{!10, 20\}$	0.212	0.604	0.880	0.152	0.504	0.828	0.060	0.142	0.636
Energy	$\{10, 20\}$	**0**	**0**	**0.004**	**0**	**0**	**0.028**	**0.036**	**0.020**	**0.356**
	$\{10, 20, x\}$	0	0	0.004	0	0	0.008	0.016	0.008	0.044
	$\{10, \tilde{20}\}/\{\tilde{10}, 20\}$	0	0	0	0.004	0	0.004	0.028	0.012	0.016
	$\{10\}/\{20\}$	0	0	0	0	0	0.004	0.012	0.004	0.068
	$\{10, !20\}/\{!10, 20\}$	0.024	0.020	0.120	0.080	0.060	0.188	0.192	0.176	0.148
KS	$\{10, 20\}$	**0.024**	**0**	**0.004**	**0.020**	**0.016**	**0.012**	**0.044**	**0.028**	**0.024**
	$\{10, 20, x\}$	0	0	0	0	0	0	0	0.004	0.004
	$\{10, \tilde{20}\}/\{\tilde{10}, 20\}$	0.116	0.076	0.044	0.148	0.092	0.052	0.224	0.132	0.076
	$\{10\}/\{20\}$	0.040	0.032	0.020	0.056	0.092	0.040	0.064	0.132	0.060
	$\{10, !20\}/\{!10, 20\}$	0.024	0.016	0.032	0.016	0.024	0.048	0.028	0.020	0.060

Table 2.3.4: Simulation study from Section 2.3 assuming two change points at times $t_1 = 10$ and $t_2 = 50$. The description of the table is similar to that of Table 2.3.3.

$K = 2$		$\mathbf{\Sigma}$			$\mathbf{\Sigma}/2$			$\mathbf{\Sigma}/4$		
$t_1 = 10, t_2 = 50$		$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$
Method	$\{10, 50\}$	**0.232**	**0.116**	**0.216**	**0.384**	**0.324**	**0.576**	**0.632**	**0.624**	**0.948**
	$\{10, 50, x\}$	0	0	0	0	0	0	0	0	0
	$\{10, \tilde{50}\}/\{\tilde{10}, 50\}$	0.368	0.156	0.008	0.460	0.248	0.008	0.336	0.220	0
	$\{10\}/\{50\}$	0.068	0.376	0.600	0	0.096	0.316	0	0.004	0.044
	$\{10, !50\}/\{!10, 50\}$	0.276	0.332	0.168	0.156	0.332	0.100	0.032	0.152	0.008
Naive	$\{10, 50\}$	**0.228**	**0.152**	**0.100**	**0.404**	**0.320**	**0.240**	**0.632**	**0.556**	**0.520**
	$\{10, 50, x\}$	0	0	0	0	0	0	0	0	0
	$\{10, \tilde{50}\}/\{\tilde{10}, 50\}$	0.372	0.256	0.132	0.432	0.284	0.168	0.336	0.284	0.128
	$\{10\}/\{50\}$	0.036	0.008	0	0	0	0	0	0	0
	$\{10, !50\}/\{!10, 50\}$	0.288	0.548	0.696	0.152	0.388	0.588	0.032	0.160	0.352
Energy	$\{10, 50\}$	**0**	**0**	**0**	**0**	**0**	**0.008**	**0.008**	**0.004**	**0.128**
	$\{10, 50, x\}$	0	0	0	0.004	0.004	0	0.008	0	0.064
	$\{10, \tilde{50}\}/\{\tilde{10}, 50\}$	0	0	0.004	0	0	0.012	0	0	0.012
	$\{10\}/\{50\}$	0.068	0.036	0.156	0.128	0.116	0.412	0.296	0.284	0.580
	$\{10, !50\}/\{!10, 50\}$	0.012	0.024	0.076	0.052	0.052	0.088	0.084	0.088	0.152
KS	$\{10, 50\}$	**0**	**0**	**0.004**	**0**	**0.004**	**0.004**	**0.008**	**0.004**	**0**
	$\{10, 50, x\}$	0	0	0	0	0	0	0	0	0
	$\{10, \tilde{50}\}/\{\tilde{10}, 50\}$	0	0	0.004	0.004	0.008	0.008	0.012	0.004	0.016
	$\{10\}/\{50\}$	0.036	0.028	0.016	0.056	0.024	0.044	0.060	0.068	0.076
	$\{10, !50\}/\{!10, 50\}$	0.104	0.056	0.064	0.112	0.096	0.096	0.176	0.104	0.136

There are seven crime types grouped into two general categories: violent and property crimes.

The former includes *Murder*, *Rape*, *Robbery*, and *Aggravated Assault*. The property crimes are

Burglary, *Larceny Theft*, and *Motor Vehicle Theft*. We focus on crime rates observed between

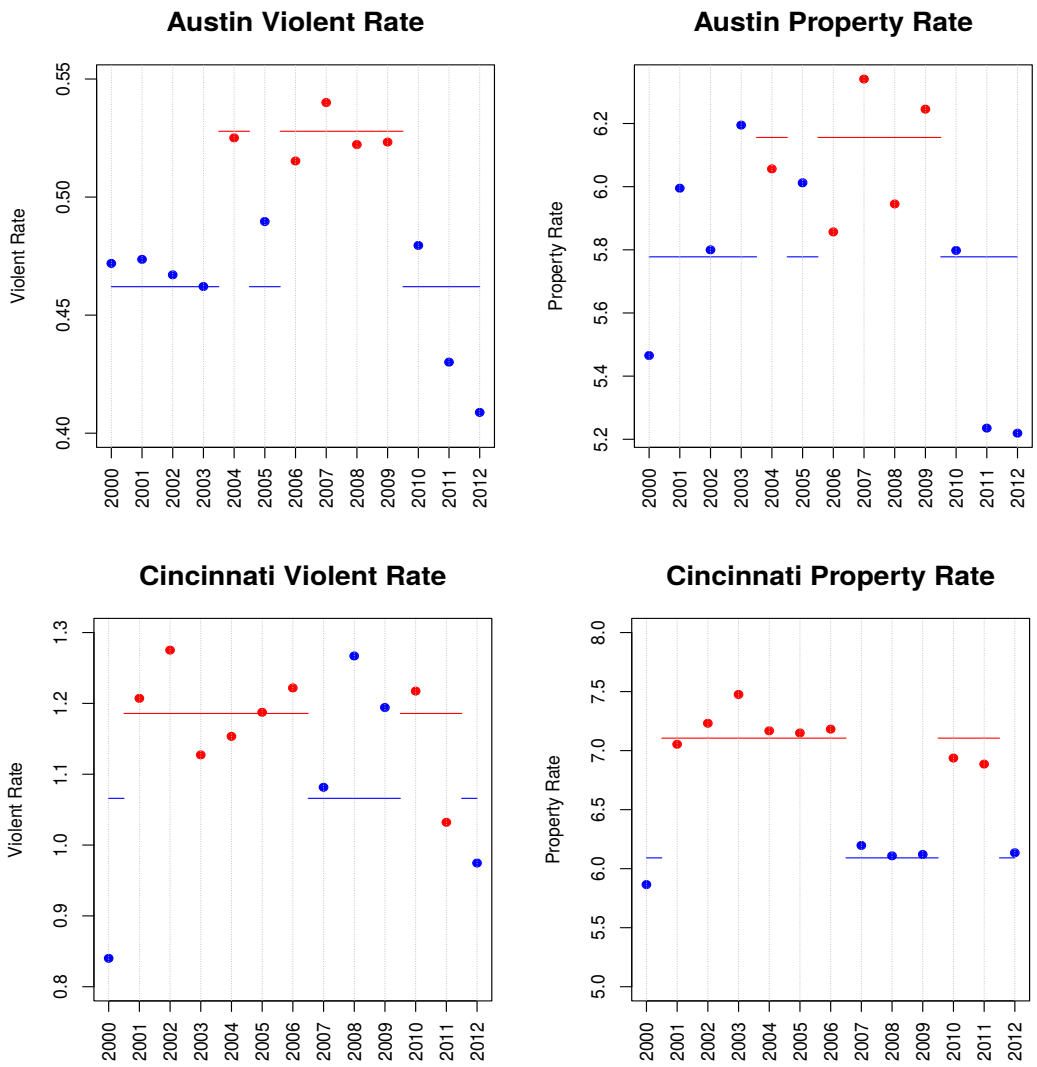


Figure 2.4.1: Violent and Property crime rates in Austin and Cincinnati over the 13-year time period (2000-2012). The blue and red colors represent two processes detected. Horizontal lines stand for the means of the processes.

2000 and 2012. As an example, we choose the data reported by Austin and Cincinnati Police Departments. Figure 2.4.1 illustrates violent (left column) and property (right column) crime rates.

In the case of Austin, the BIC value associated with a single process (*i.e.*, no change points) is equal to -9.933. After running the developed procedure over all possible permutations of processes, the lowest BIC of -40.564 was found. The parameter estimates associated with the model can be found in Table 2.4.1. A corresponding illustration is provided in the first row of plots in

Table 2.4.1: Parameter estimates, log-likelihood and BIC values for Austin and Cincinnati.

City	$\hat{\boldsymbol{\mu}}_0$	$\hat{\boldsymbol{\mu}}_1$	$\hat{\boldsymbol{\Sigma}}$		$\hat{\boldsymbol{\lambda}}$	$\hat{\phi}$	$\log \mathcal{L}$	BIC
Austin	168.234	524.023	4,422.5	105,136.9	17.258	−0.402	39.831	−40.564
	4,941.351	8,870.934	105,136.9	5,810,522	1.548			
Cincinnati	4.130	5.478	1.372	0.004	2.148	0.315	19.693	−0.288
	2.394	2.480	0.004	0.0001	−0.375			

Figure 2.4.1. Here, the years 2004, 2006, 2007, 2008, and 2009 are associated with the second process (provided in the red color), while the rest of the years represent the first process (given in the blue color). The horizontal lines reflect back-transformed parameters $\hat{\boldsymbol{\mu}}_0$ and $\hat{\boldsymbol{\mu}}_1$ detected by our methodology. As we can clearly see, the separation into two processes is strongly driven by the variable *Violent Crime*. In the meantime, the variable *Property Crime* demonstrates considerable variability associated with both processes. As we can see, the proposed methodology is efficient not only for detecting shifts in processes but also for separating arbitrary processes.

The opposite situation is observed for Cincinnati (second row in Figure 2.4.1). Here, the variable *Property Crime* contributes to the separation of the processes more than *Violent Crime*. Model parameters are also provided in Table 2.4.1. The BIC value of the best model detected is equal to -0.288 which is considerably better than that of the model with a single process, 19.568. The years 2000, 2007, 2008, 2009, and 2012 are associated with the first process (presented in the blue color), while the rest of the years represent the other process (given in the red color).

2.4.2 Effect of Colorado Amendment 64

In this section, we demonstrate how our proposed methodology can be applied to the analysis of the effects of public policies. As an example, we focus on studying the effects of the Colorado Amendment 64 that makes the private consumption, production, and possession of marijuana legal. Amendment 64 has been added to the constitution of Colorado in December, 2012 but the stores officially opened in January, 2014.

The crime rate data have been obtained from the Colorado Bureau of Investigation De-

partment of Public Safety Web-site (https://www.colorado.gov/pacific/cbi/crime-colorado1) for the last 10 years: from 2007 to 2016. The same seven variables as described in Section 2.4.1 have been explored without combining them into the two categories. The goal of our analysis was to check whether the last three years, when the use of marijuana was legal, were any different from the previous seven years. The value of BIC corresponding to the model with no change points is equal to -996.2, while that related to the model with the change point at 2014 yields BIC equal to -1,006.1. The likelihood ratio test conducted to varify the significance of the change yields p-value 1.47×10^{-6}. As we can see, there is very strong evidence in favor of the change point model based on both BIC and likelihood ratio test.

Figure 2.4.2 illustrates the obtained results. The first column consisting of four plots represents violent crimes, while the second column with three plots shows property crimes. The description of individual plots is similar to that of Figure 2.4.1. As we can see, some variables such as *Rape* or *Burglary* seem to contribute substantially to the difference between the two models analyzed. To formalize the analysis, we employed a variable selection procedure. As the number of variables in our experiment is relatively low, we decided to test the model with no change point against the model with the change point at 2014 over all possible combinations of involved variables. The lowest p-value of 1.36×10^{-6} was observed for the combination of variables *Murder*, *Rape*, and *Burglary*. Thus, the most dramatic change in 2014 has been observed for these three variables considered jointly. The corresponding p-value is just marginally lower than the p-value observed for the full model when all seven variables are included, but it gives a good idea about the combination of variables that contribute the most to separating the processes. By examining the contributions of the three variables, we can notice that the crime rate of *Burglary* dropped considerably, while *Rape* and to some extent *Murder* are on the grow in the last three years. Indeed, the proposed analysis does not assume any cause-and-effect conclusions. In fact,

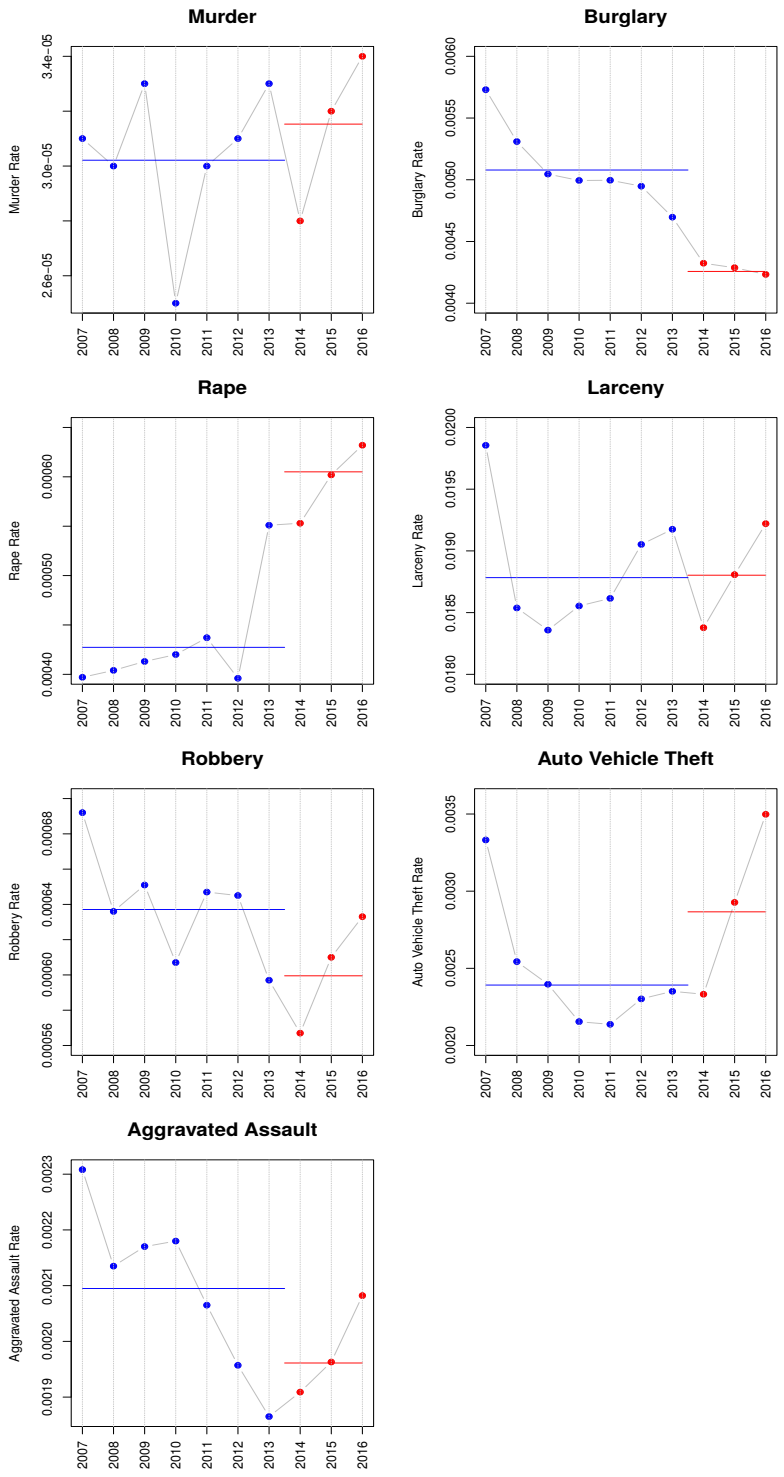


Figure 2.4.2: Crime rates in Colorado over the 10-year time period. The blue and red colors represent two processes. Horizontal lines stand for the back-transformed means of the processes.

we can notice a considerable descrease in *Murder* rates in 2014 and we can also observe that the increase in *Rape* rates began in 2013, *i.e.*, one year earlier than when Amendment 64 became effective. Nevertheless, it is obvious that the proposed methodology presents a powerful exploratory tool for studying effects of public policies.

2.5 Discussion

In this paper, we developed an efficient method capable of estimating multiple change points in multivariate processes. The proposed technique relies on the matrix normal distribution adjusted by the exponential Manly transformation. Such an adjustment makes the proposed methodology robust to violations of the normality assumption. The matrix setting has an appealing form as rows can represent variables and columns can be associated with time points. Based on the results of challenging simulation studies, we can conclude that the proposed technique is very promising. It outperforms the two non-parametric competitors in all settings dramatically. Two applications to crime data considered in the paper demonstrate the usefulness of our method.

CHAPTER 3

ROBUST ESTIMATION OF MULTIPLE CHANGE POINTS IN THREE-DIMENSIONAL DATA

Processes observed over time occur in all areas of human activity. One of the most important problems in this setting is the task of identifying change points. In this paper, we propose novel likelihood-based methodology capable of identifying multiple chagne points in matrix-valued processes effectively. The proposed methodology is a flexible tool that does not make restrictive assumptions such as independence of observations, independence of subjects, or normality. To improve the robustness of the model to deviations from symmetry and normality, a power or exponential transformation can be built into the model. The developed methodology is illustrated on two real-life datasets, with good and easily interpretable results.

3.1 Introduction

The problem of change point inference have found multiple applications in engineering [26, 31], ecology [8], finance [30], biology [41], pharmacy [3], and many other areas. The change point estimation problem has been recognized in the end of 1950's with a paper by [28]. The paper proposed an approach for identifying a change in the mean of a univariate process. Since then, numerous advances have been proposed in the change point literature. For univariate processes, methods for finding single and multiple change points in means [13, 42], variances [15, 9, 16, 5], or means and variances simultaneously [14] have been developed. Many real-life problems assume multivariate processes. Upon recognizing this fact, researches started tackling

change point problems in multivariate settings. Similarly to the cases of univariate processes, the changes in mean vectors [38, 39, 45] or covariances matrices [6] have been paid attention. While the focus of the majority of papers have been made on the treatment of univariate or multivariate Gaussian processes, other families such as Poisson [27] or general exponential family [32] are also considered. Some developments in nonparametric change point detection and estimation can also be found in literature [33, 17].

Despite all the attention to the change point inference, there are still many open problems. In particular, many procedures make often unrealistic assumptions about the independence of observations or subjects or are sensitive to violations of the normality assumption. Moreover, there have been no methods addressing the estimation of change points in matrix-valued processes. In this paper, we aim at relaxing overly restrictive assumptions and introducing novel likelihood-based inference for identifying multiple change points in two-dimensional processes. We also illustrate how change points can be treated in multisubject multivariate processes.

The paper is presented in the following way. Section 3.2 discusses some needed preliminaries and introduces the proposed methodology. Section 3.3 applies the developed procedure to the analysis of Alabama university professor salary data as well as crime rates in 125 major American cities. The paper concludes with a discussion provided in Section 3.4.

3.2 Methodology

3.2.1 Matrix normal distribution

Suppose $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T$ represent a p-variate process observed over T time points. Under the assumption that the process follows a p-variate Gaussian distribution, the entire dataset can conveniently modeled using a matrix normal distribution [20]. The matrix normal probability

density function (pdf) is given by the following expression

$$\phi_{p\times T}(\boldsymbol{Y};\boldsymbol{M},\boldsymbol{\Sigma},\boldsymbol{\Psi}) = (2\pi)^{-\frac{pT}{2}}|\boldsymbol{\Sigma}|^{-\frac{T}{2}}|\boldsymbol{\Psi}|^{-\frac{p}{2}}\exp\left\{-\frac{1}{2}\text{tr}\left\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{M})\boldsymbol{\Psi}^{-1}(\boldsymbol{Y}-\boldsymbol{M})^{\top}\right\}\right\}, \tag{3.2.1}$$

where $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T)$ is the $p \times T$ data matrix, $\boldsymbol{M}$ is the $p \times T$ mean matrix, and $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$ are $p\times p$ and $T\times T$ covariance matrices, respectively. The matrix $\boldsymbol{\Sigma}$ describes the variability associated with the rows of $\boldsymbol{Y}$, while $\boldsymbol{\Psi}$ models the variability related to the columns of $\boldsymbol{Y}$. The formulation provided in expression (3.2.1) is convenient due to the fact that it takes into consideration the matrix structure of the data and splits the sources of variability. It can be noted, however, that the $p \times T$ matrix normal distribution can be seen as a pT-variate Gaussian one with the mean produced by the vectorization of matrix $\boldsymbol{M}$ and overall covariance matrix $\boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}$. In other words, $\text{vec}(\boldsymbol{Y}) \sim \mathcal{N}_{pT}(\text{vec}(\boldsymbol{M}), \boldsymbol{\Psi} \otimes \boldsymbol{\Sigma})$, where $\text{vec}(\cdot)$ represents the vectorization operator stacking matrix columns on top of each other and $\otimes$ stands for the Kronecker product. As the covariance matrix is given in a Kronecker product form, the matrix normal distribution can be thought of as a special case of the pT-variate Gaussian distribution. This special form can be justified by the fact that the original data are provided as a matrix, with rows and columns representing some common characteristics of matrix elements. The reduction in the number of unique parameters associated with the covariance matrix is rather substantial as just $T(T+1)/2 + p(p+1)/2$ parameters are needed for matrix data as opposed to $pT(pT+1)/2$ parameters in the case of a pT-variate normal distribution with the unrestricted covariance matrix. One can notice that the product $\boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}$ yields a non-identifiability issue related to the property of the Kronecker product indicating that for any multiplier $a \in I\!R^{+}$, $a\boldsymbol{\Psi} \otimes \boldsymbol{\Sigma} = \boldsymbol{\Psi} \otimes a\boldsymbol{\Sigma}$. This minor issue can be easily resolved in practice by setting a restriction on one or both covariance matrices.

3.2.2 Modeling change points in matrix processes

Now, let $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_T$ represent a two-dimensional process observed over T time points. Here, each $\boldsymbol{Y}_i$ stands for a $d \times p$ matrix. Within this matrix, we observe d variables presented in rows as well as p variables given in matrix columns. Thus, the entire dataset can be seen as a three-way tensor $\mathcal{Y}$ summarized as shown below:

$$\mathcal{Y} \equiv \left\{ \begin{pmatrix} y_{11t} & y_{12t} & \cdots & y_{1pt} \\ y_{21t} & y_{22t} & \cdots & y_{2pt} \\ \vdots & \vdots & \ddots & \vdots \\ y_{d1t} & y_{d2t} & \cdots & y_{dpt} \end{pmatrix}, \; t = 1, 2, \ldots, T \right\}. \tag{3.2.2}$$

Suppose $d \times p \times T$ tensor $\mathcal{Y}$ follows a tensor normal distribution with mean tensor $\mathcal{M}$ and covariance matrices $\boldsymbol{\Delta}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\Psi}$ with dimensions $d \times d$, $p \times p$, and $T \times T$, respectively. The probability density function in this case as well as operations with tensors are not trivial. One popular way of treating tensors lies in the dimensionality reduction that leads to matrices. In particular, one of the possible vectorizations of matrices $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_T$ within $\mathcal{Y}$ produces a matrix object $\tilde{\mathcal{Y}}$ with dimensions $dp \times T$ that follows a matrix normal distribution $\phi_{dp \times T}(\tilde{\mathcal{Y}}; \tilde{\mathcal{M}}, \boldsymbol{\Sigma} \otimes \boldsymbol{\Delta}, \boldsymbol{\Psi})$. Here, $\tilde{\mathcal{Y}} = (\text{vec}(\boldsymbol{Y}_1), \text{vec}(\boldsymbol{Y}_2), \ldots, \text{vec}(\boldsymbol{Y}_T))$ and $\tilde{\mathcal{M}}$ is a similarly constructed $dp \times T$ mean matrix, *i.e.*, $\tilde{\mathcal{M}} = (\text{vec}(\boldsymbol{M}_1), \text{vec}(\boldsymbol{M}_2), \ldots, \text{vec}(\boldsymbol{M}_T))$ with each $\boldsymbol{M}_t$ being a $d \times p$ matrix. The covariance matrix corresponding to the rows is given by $\boldsymbol{\Sigma} \otimes \boldsymbol{\Delta}$. Under this considered vectorization, the matrix normal pdf is given by

$$\begin{aligned} \phi_{dp \times T}(\tilde{\mathcal{Y}}; \tilde{\mathcal{M}}, \boldsymbol{\Sigma} \otimes \boldsymbol{\Delta}, \boldsymbol{\Psi}) = (2\pi)^{-\frac{dpT}{2}} |\boldsymbol{\Sigma} \otimes \boldsymbol{\Delta}|^{-\frac{T}{2}} |\boldsymbol{\Psi}|^{-\frac{dp}{2}} \\ \times \exp\left\{ -\frac{1}{2} \text{tr}\left\{ (\boldsymbol{\Sigma} \otimes \boldsymbol{\Delta})^{-1} \left(\tilde{\mathcal{Y}} - \tilde{\mathcal{M}} \right) \boldsymbol{\Psi}^{-1} \left(\tilde{\mathcal{Y}} - \tilde{\mathcal{M}} \right)^{\top} \right\} \right\}. \end{aligned}$$

Using the properties of the Kronecker product, the density can be further written as

$$\phi_{dp\times T}(\tilde{\mathcal{Y}};\tilde{\mathcal{M}},\boldsymbol{\Sigma}\otimes\boldsymbol{\Delta},\boldsymbol{\Psi}) = (2\pi)^{-\frac{dpT}{2}}|\boldsymbol{\Delta}|^{-\frac{pT}{2}}|\boldsymbol{\Sigma}|^{-\frac{dT}{2}}|\boldsymbol{\Psi}|^{-\frac{dp}{2}} \times \exp\left\{-\frac{1}{2}\text{tr}\left\{\left(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Delta}^{-1}\right)\left(\tilde{\mathcal{Y}}-\tilde{\mathcal{M}}\right)\boldsymbol{\Psi}^{-1}\left(\tilde{\mathcal{Y}}-\tilde{\mathcal{M}}\right)^{\top}\right\}\right\}.$$

Suppose there are K change points and hence $K+1$ matrix processes. Each process has the same mean parameters until the next process starts. Then, the mean matrix $\tilde{\mathcal{M}}$ can be written in the following way:

$$\begin{aligned}\tilde{\mathcal{M}} &= \Bigg(\underbrace{\text{vec}(\boldsymbol{M}_0),\ldots,\text{vec}(\boldsymbol{M}_0)}_{t_1-1},\underbrace{\text{vec}(\boldsymbol{M}_1),\ldots,\text{vec}(\boldsymbol{M}_1)}_{t_2-t_1},\ldots, \\ &\qquad \underbrace{\text{vec}(\boldsymbol{M}_{K-1}),\ldots,\text{vec}(\boldsymbol{M}_{K-1})}_{t_K-t_{K-1}},\underbrace{\text{vec}(\boldsymbol{M}_K),\ldots,\text{vec}(\boldsymbol{M}_K)}_{T-t_K+1}\Bigg) \\ &= \sum_{k=0}^{K}\text{vec}(\boldsymbol{M}_k)\boldsymbol{m}_k^{\top},\end{aligned} \tag{3.2.3}$$

where $\boldsymbol{m}_k$ is a T-variate vector given by $\boldsymbol{m}_k^{\top} = \left(\underbrace{\mathbf{0},\ldots,\mathbf{0}}_{t_k-1},\underbrace{\mathbf{1},\ldots,\mathbf{1}}_{t_{k+1}-t_k},\underbrace{\mathbf{0},\ldots,\mathbf{0}}_{T-t_{k+1}+1}\right)$ with boundary conditions $t_0 = 1$ and $t_{K+1} = T+1$. As we can see, $\boldsymbol{m}_k$ represents a vector consisting of ones and zeros with the former located at those time points when the kth process is observed.

Violations of the normality assumption can severely impact the performance of the proposed procedure. To improve its robustness to deviations from normality, we can employ one of several available transformations to near-normality. Perhaps, the most well-known transformation in this class is the power transformation proposed by [4]. Some criticism of this procedure is related to its incapability to transform negative values and lack of flexibility in handling left-skewed

data. One flexible modification of the power transformation was proposed by [43] and is given by

$$\mathcal{T}_{pow}(\lambda; y) = \left\{ \left[\frac{(y+1)^{\lambda} - 1}{\lambda} \right]^{I(\lambda \neq 0)} [\log(y+1)]^{I(\lambda=0)} \right\}^{I(y \geq 0)} \left\{ \left[-\frac{(1-y)^{2-\lambda} - 1}{2-\lambda} \right]^{I(\lambda \neq 2)} [-\log(1-y)]^{I(\lambda=2)} \right\}^{I(y<0)}. \tag{3.2.4}$$

This transformation can be applied to $y \in I\!R$ and is equally efficient with left- and right-skewed data. One more popular alternative is the exponential transformation developed by [24] and given by

$$\mathcal{T}_{exp}(\lambda; y) = \left[\frac{e^{\lambda y} - 1}{\lambda} \right]^{I(\lambda \neq 0)} y^{I(\lambda=0)}. \tag{3.2.5}$$

The application of these univariate transformations in the multivariate framework is commonly undertaken based on the assumption that coordinatewise transformations can effectively lead to joint near-normality. There is extensive history of successful applications of transformations in multivariate settings [2, 34, 21, 46]. We denote the matrix transformation to near-normality as $\mathcal{T}(\boldsymbol{Y}; \boldsymbol{\Lambda})$, where $\boldsymbol{\Lambda}$ represents the $d \times p$ matrix of transformation parameters. For the matter of model interpretability as well as for reducing the number of parameters that have to be estimated by the numerical optimization of the log-likelihood function, we assume the additive effect of row and column transformation parameters $\boldsymbol{\nu}$ and $\boldsymbol{\lambda}$, *i.e.*, $\boldsymbol{\Lambda} = \boldsymbol{\nu}\mathbf{1}_p^{\top} + \mathbf{1}_d\boldsymbol{\lambda}^{\top}$, where $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_d)^{\top}$ and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_p)^{\top}$. Such parameterization leads to one more non-identifiability issue that can be resolved by setting $\lambda_p = 0$.

Employing a multivariate transformation $\mathcal{T}(\boldsymbol{Y}; \boldsymbol{\nu}, \boldsymbol{\lambda})$ and taking into consideration both Equations (3.2.1) and (3.2.3), the corresponding log-likelihood function can be written as follows

below:

$$\log \mathcal{L}(\tilde{\mathcal{Y}};\, \boldsymbol{M}_1, \ldots, \boldsymbol{M}_K, \boldsymbol{\Delta}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}) = -\frac{pdT}{2}\log(2\pi) - \frac{pT}{2}\log|\boldsymbol{\Delta}| - \frac{dT}{2}\log|\boldsymbol{\Sigma}| - \frac{pd}{2}\log|\boldsymbol{\Psi}|$$
$$- \frac{1}{2}\text{tr}\Bigg\{ \left(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Delta}^{-1}\right) \left(\mathcal{T}(\tilde{\mathcal{Y}}; \boldsymbol{\nu}, \boldsymbol{\lambda}) - \sum_{k=0}^{K} \text{vec}(\boldsymbol{M}_k)\boldsymbol{m}_k^\top\right) \boldsymbol{\Psi}^{-1}$$
$$\times \left(\mathcal{T}(\tilde{\mathcal{Y}}; \boldsymbol{\nu}, \boldsymbol{\lambda}) - \sum_{k=0}^{K} \text{vec}(\boldsymbol{M}_k)\boldsymbol{m}_k^\top\right)^\top \Bigg\} - \log\left|\frac{\partial \mathcal{T}(\tilde{\mathcal{Y}}; \boldsymbol{\nu}, \boldsymbol{\lambda})}{\partial \tilde{\mathcal{Y}}}\right|. \tag{3.2.6}$$

Here, $\left|\frac{\partial \mathcal{T}(\mathcal{Y}; \boldsymbol{\nu}, \boldsymbol{\lambda})}{\partial \mathcal{Y}}\right|$ represents the Jacobian associated with the transformation. By straightforward differentiation of the log-likelihood function with respect to vec$(\boldsymbol{M}_k)$, we obtain

$$\text{vec}(\boldsymbol{M}_k) = \left(\mathcal{T}(\tilde{\mathcal{Y}}; \boldsymbol{\nu}, \boldsymbol{\lambda}) - \sum_{\substack{k'=0 \\ k' \neq k}}^{K} \text{vec}(\boldsymbol{M}_{k'})\boldsymbol{m}_{k'}^\top\right) \boldsymbol{\Psi}^{-1}\boldsymbol{m}_k \left(\boldsymbol{m}_k^\top \boldsymbol{\Psi}^{-1}\boldsymbol{m}_k\right)^{-1}.$$

This system of $K+1$ linear equations can be solved for vec$(\boldsymbol{M}_k)$, $k = 0, 1, \ldots, K$. It is easy to show that for $K = 1$,

$$\text{vec}(\boldsymbol{M}_0) = \mathcal{T}(\tilde{\mathcal{Y}}; \boldsymbol{\nu}, \boldsymbol{\lambda})\boldsymbol{\Psi}^{-1}(m_{11}\boldsymbol{m}_0 - m_{01}\boldsymbol{m}_1)/(m_{00}m_{11} - m_{01}^2),$$

$$\text{vec}(\boldsymbol{M}_1) = \mathcal{T}(\tilde{\mathcal{Y}}; \boldsymbol{\nu}, \boldsymbol{\lambda})\boldsymbol{\Psi}^{-1}(m_{00}\boldsymbol{m}_1 - m_{01}\boldsymbol{m}_0)/(m_{00}m_{11} - m_{01}^2),$$

where $m_{kk'} = \boldsymbol{m}_k^\top \boldsymbol{\Psi}^{-1}\boldsymbol{m}_{k'}$. For $K = 2$, the following expression can be obtained for vec$(\boldsymbol{M}_0)$:

$$\text{vec}(\boldsymbol{M}_0) = (m_{00}m_{11}m_{22} - m_{01}^2 m_{22} - m_{02}^2 m_{11} - m_{12}^2 m_{00} + 2m_{01}m_{02}m_{12})^{-1}\mathcal{T}(\tilde{\mathcal{Y}}; \boldsymbol{\nu}, \boldsymbol{\lambda})\boldsymbol{\Psi}^{-1}$$
$$\times (m_{12}(m_{01}\boldsymbol{m}_2 + m_{02}\boldsymbol{m}_1 - m_{12}\boldsymbol{m}_0) - m_{02}m_{11}\boldsymbol{m}_2 - m_{01}m_{22}\boldsymbol{m}_1 + m_{11}m_{22}\boldsymbol{m}_0)$$

and expressions for vec$(\boldsymbol{M}_1)$ and vec$(\boldsymbol{M}_2)$ can be found similarly. Upon the substitution of the

expressions derived for the process means into the log-likelihood function provided in (3.2.6), we need to focus on estimating covariance matrices $\boldsymbol{\Psi}$, $\boldsymbol{\Delta}$, and $\boldsymbol{\Sigma}$.

While matrices $\boldsymbol{\Delta}$ and $\boldsymbol{\Psi}$ can have a general unrestricted form, in the considered framework the user can employ a desired time series relationship through the specification of the covariance matrix $\boldsymbol{\Psi}$. The search for the optimal temporal model is beyond the scope of this paper, although it can be easily implemented by considering various parameterizations of $\boldsymbol{\Psi}$. Without loss of generality, we assume the autoregressive order one time series denoted by AR(1). The corresponding covariance matrix is given by $\boldsymbol{\Psi} = \frac{\delta^2}{1-\phi^2}\boldsymbol{R}_\phi$, where

$$\boldsymbol{R}_\phi = \begin{pmatrix} 1 & \phi & \phi^2 & \dots & \phi^{T-1} \\ \phi & 1 & \phi & \dots & \phi^{T-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{T-1} & \phi^{T-2} & \phi^{T-3} & \dots & 1 \end{pmatrix}$$

is the correlation matrix of the AR(1) process and ϕ and δ^2 are corresponding correlation and variance parameters. Recall that the covariance matrix of the vectorized tensor is given by $\boldsymbol{\Psi} \otimes \boldsymbol{\Sigma} \otimes \boldsymbol{\Delta}$. This relationship implies that two restrictions must be implemented. One convenient restriction is $\delta^2 = 1 - \phi^2$. Then, the covariance matrix $\boldsymbol{\Psi}$ reduces to the correlation matrix $\boldsymbol{R}_\phi$. To avoid potentially time consuming operations with determinants and inverses of $T \times T$ matrix $\boldsymbol{\Psi}$, the following expressions can be employed:

$$|\boldsymbol{\Psi}| \equiv |\boldsymbol{R}_\phi| = (1-\phi^2)^{T-1} \quad \text{and} \quad \boldsymbol{\Psi}^{-1} \equiv \boldsymbol{R}_\phi^{-1} = \frac{1}{1-\phi^2}(\boldsymbol{I}_T - \phi\boldsymbol{J}_1 + \phi^2\boldsymbol{J}_2), \tag{3.2.7}$$

where $\boldsymbol{J}_1$ and $\boldsymbol{J}_2$ are $T \times T$ matrices given below:

$$\boldsymbol{J}_1 = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 1 & 0 & 1 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & 0 & \dots & 1 & 0 \end{pmatrix} \quad \text{and} \quad \boldsymbol{J}_2 = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 \end{pmatrix}.$$

Being the only parameter of $\boldsymbol{\Psi}$, ϕ cannot be estimated analytically.

Consider now the problem of estimating matrix $\boldsymbol{\Sigma}$. It is easy to see that $\boldsymbol{\Sigma}$ is not readily available from the log-likelihood expression in (3.2.6) due to its involvement into the Kronecker product. However, a different way of constructing a matrix from tensor $\mathcal{Y}$ can easily resolve this problem. Consider a $dT \times p$ matrix $\ddot{\mathcal{Y}}$ constructed by the vectorization of $d \times T$ matrices at each $j = 1, 2, \dots, p$, *i.e.*,

$$\ddot{\mathcal{Y}} \equiv \left\{ \text{vec} \begin{pmatrix} y_{1j1} & y_{1j2} & \dots & y_{1jT} \\ y_{2j1} & y_{2j2} & \dots & y_{2jT} \\ \vdots & \vdots & \ddots & \vdots \\ y_{dj1} & y_{dj2} & \dots & y_{djT} \end{pmatrix}, \; j = 1, 2, \dots, p \right\}.$$

Applying the same matricization approach to the mean tensor $\mathcal{M}$, a $dT \times p$ mean matrix $\ddot{\mathcal{M}}$ can be obtained and a corresponding log-likelihood function constructed. Note that this log-likelihood is the same as the one given in (3.2.6), but it is written in an alternative form. Now, the

estimate of $\boldsymbol{\Sigma}$, which is not involved into the Kronecker product any more, can be readily found:

$$\boldsymbol{\Sigma} = \frac{(\mathcal{T}(\ddot{\mathcal{Y}}; \boldsymbol{\nu}, \boldsymbol{\lambda}) - \ddot{\mathcal{M}})^\top (\boldsymbol{\Psi}^{-1} \otimes \boldsymbol{\Delta}^{-1})(\mathcal{T}(\ddot{\mathcal{Y}}; \boldsymbol{\nu}, \boldsymbol{\lambda}) - \ddot{\mathcal{M}})}{dT}.$$

The last covariance matrix, $\boldsymbol{\Delta}$, can be obtained in a similar way that yields the following expression

$$\boldsymbol{\Delta} = \frac{(\mathcal{T}(\breve{\mathcal{Y}}; \boldsymbol{\nu}, \boldsymbol{\lambda}) - \breve{\mathcal{M}})^\top (\boldsymbol{\Psi}^{-1} \otimes \boldsymbol{\Sigma}^{-1})(\mathcal{T}(\breve{\mathcal{Y}}; \boldsymbol{\nu}, \boldsymbol{\lambda}) - \breve{\mathcal{M}})}{pT},$$

where $\breve{\mathcal{Y}}$ is a $pT \times d$ matrix given by

$$\breve{\mathcal{Y}} \equiv \left\{ \text{vec} \begin{pmatrix} y_{i11} & y_{i12} & \cdots & y_{i1T} \\ y_{i21} & y_{i22} & \cdots & y_{i2T} \\ \vdots & \vdots & \ddots & \vdots \\ y_{ip1} & y_{ip2} & \cdots & y_{ipT} \end{pmatrix}, \; i = 1, 2, \ldots, d \right\}$$

and $\breve{\mathcal{M}}$ is a $pT \times d$ mean matrix obtained in the same way from the mean tensor $\mathcal{M}$. As per our discussion on constraints necessary to avoid the non-identifiability issues, we can impose an additional restriction on $\boldsymbol{\Delta}$, *e.g.*, $|\boldsymbol{\Delta}| = 1$.

Estimators of the parameters $\boldsymbol{\nu}$ and $\boldsymbol{\lambda}$ cannot be expressed as closed form solutions. Based on the obtained expressions for means and covariance matrices, an iterative numerical optimization procedure can be developed to estimate all parameters. The total number of parameters in the considered model is $dp(K+1) + d(d+1)/2 + p(p+1)/2 + d + p - 1$.

3.2.3 Change point detection in multisubject multivariate processes

Suppose now that there are p variables observed for N subjects over T time points. The corresponding data can be represented as a tensor with dimensions $N \times p \times T$. Now, covariance

matrix $\boldsymbol{\Delta}$ will model the variability associated with N subjects. In such a setting, a special form of $\boldsymbol{\Delta}$ can be assumed. Without loss of generality, suppose that subjects are observed within M blocks, n_m subjects per block, where $m = 1, 2, \ldots, M$. Let σ^2 and σ_b^2 represent parameters modeling between- and within-block variability in subjects. Then, $\boldsymbol{\Delta}$ is a block-diagonal matrix given by

$$\boldsymbol{\Delta} = \sigma^2 \text{diag}\left\{\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_M\right\} \quad \text{with} \quad \boldsymbol{V}_m = \eta \mathbf{1}_{n_m}\mathbf{1}_{n_m}^\top + \boldsymbol{I}_{n_m} \quad \text{and} \quad \eta = \frac{\sigma_b^2}{\sigma^2}.$$

For the matter of notational simplicity, assume equal block sizes $n \equiv n_1 = \ldots = n_M = N/M$. The determinant and inverse of $\boldsymbol{\Delta}$ are given by the following expressions:

$$|\boldsymbol{\Delta}| = (\sigma^2)^N(\eta n + 1)^M \quad \text{and} \quad \boldsymbol{\Delta}^{-1} = \frac{1}{\sigma^2(\eta n + 1)}(\eta(n\boldsymbol{I}_N - \mathcal{D}_M(\mathbf{1}_n\mathbf{1}_n^\top)) + \boldsymbol{I}_N).$$

As per our discussion in Section 3.2.2 about non-identifiability issues, one restriction needs to be imposed on matrix $\boldsymbol{\Delta}$. One convenient constraint in the considered framework is $\sigma^2\eta = 1$. Under this restriction, $\boldsymbol{\Delta} = \text{diag}\left\{\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_M\right\}$ with $\boldsymbol{V}_m = \mathbf{1}_n\mathbf{1}_n^\top + \sigma^2\boldsymbol{I}_n$, or simply $\boldsymbol{\Delta} = \sigma^2\boldsymbol{I}_N + \mathcal{D}_M(\mathbf{1}_n\mathbf{1}_n^\top)$, where $\mathcal{D}_M(\mathbf{1}_n\mathbf{1}_n^\top)$ denotes the block-diagonal matrix consisting of M blocks $\mathbf{1}_n\mathbf{1}_n^\top$. Then, the determinant and inverse of $\boldsymbol{\Delta}$ can be simplified to

$$|\boldsymbol{\Delta}| = \left(\sigma^2\right)^{N-M}\left(n + \sigma^2\right)^M \quad \text{and} \quad \boldsymbol{\Delta}^{-1} = \frac{1}{\sigma^2(n + \sigma^2)}\left((n + \sigma^2)\boldsymbol{I}_N - \mathcal{D}_M(\mathbf{1}_n\mathbf{1}_n^\top)\right).$$

Mean tensor $\mathcal{M}$ has dimensions $N \times p \times T$. Then, similarly to the notation introduced in Section 3.2.2, $\tilde{\mathcal{M}} = (\text{vec}(\boldsymbol{M}_1), \text{vec}(\boldsymbol{M}_2), \ldots, \text{vec}(\boldsymbol{M}_T))$ is an $Np \times T$ matrix with each $\boldsymbol{M}_t$ being an $N \times p$ matrix. In the multisubject study, the assumption that all subjects have different means is hardly realistic. To reduce the number of parameters, we consider a linear model $\boldsymbol{M}_t =$

$\boldsymbol{X}\boldsymbol{B}_t$, where $\boldsymbol{X}$ is an $N \times q$ design matrix and $\boldsymbol{B}_t$ is a $q \times p$ matrix of coefficients. The design matrix $\boldsymbol{X}$ is specified by the user based on the problem considered. For example, in our setting with M blocks, without loss of generality, it can be assumed that subjects from the same block have a common mean.

In the framework with K change points, there are $K + 1$ processes with corresponding matrices of coefficients $\boldsymbol{B}_0, \boldsymbol{B}_1, \ldots, \boldsymbol{B}_K$. Since $\text{vec}(\boldsymbol{ACD}) = (\boldsymbol{D}^\top \otimes \boldsymbol{A})\text{vec}(\boldsymbol{C})$ for any conforming matrices $\boldsymbol{A}$, $\boldsymbol{C}$, and $\boldsymbol{D}$, we obtain

$$\text{vec}(\boldsymbol{M}_k) = \text{vec}(\boldsymbol{X}\boldsymbol{B}_k) = \text{vec}(\boldsymbol{X}\boldsymbol{B}_k\boldsymbol{I}_p) = (\boldsymbol{I}_p \otimes \boldsymbol{X})\text{vec}(\boldsymbol{B}_k).$$

Then, the log-likelihood can be written as

$$\begin{aligned}\log \mathcal{L}(\tilde{\mathcal{Y}};\, \boldsymbol{B}_0, \ldots, \boldsymbol{B}_K, \boldsymbol{\Sigma}, \boldsymbol{\Delta}, \boldsymbol{\Psi}) = &-\frac{NpT}{2}\log(2\pi) - \frac{NT}{2}\log|\boldsymbol{\Sigma}| - \frac{pT}{2}\log|\boldsymbol{\Delta}| - \frac{Np}{2}\log|\boldsymbol{\Psi}| \\ &- \frac{1}{2}\text{tr}\left\{\left(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Delta}^{-1}\right)\left(\mathcal{T}(\tilde{\mathcal{Y}};\boldsymbol{\lambda}) - (\boldsymbol{I}_p \otimes \boldsymbol{X})\sum_{k=0}^{K}\text{vec}(\boldsymbol{B}_k)\boldsymbol{m}_k^\top\right)\boldsymbol{\Psi}^{-1}\right. \\ &\left.\times\left(\mathcal{T}(\tilde{\mathcal{Y}};\boldsymbol{\lambda}) - (\boldsymbol{I}_p \otimes \boldsymbol{X})\sum_{k=0}^{K}\text{vec}(\boldsymbol{B}_k)\boldsymbol{m}_k^\top\right)^\top\right\} + \log\left|\frac{\partial\mathcal{T}(\tilde{\mathcal{Y}};\boldsymbol{\lambda})}{\partial\tilde{\mathcal{Y}}}\right|.\end{aligned}$$

As we can see, the p-variate vector $\boldsymbol{\lambda}$ is the only transformation parameter in the considered framework. By straightforward differentiation of the log-likelihood function over $\text{vec}(\boldsymbol{B}_k)$, we obtain the system of $K + 1$ equations. In the simplest case with $K = 1$, the solution is given by

$$\begin{aligned}\text{vec}(\boldsymbol{B}_0) = (m_{00}m_{11} - m_{01}^2)^{-1}&\left((\boldsymbol{I}_p \otimes \boldsymbol{X})^\top(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Delta}^{-1})(\boldsymbol{I}_p \otimes \boldsymbol{X})\right)^{-1}(\boldsymbol{I}_p \otimes \boldsymbol{X})^\top \\ &\times (\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Delta}^{-1})\mathcal{T}(\tilde{\mathcal{Y}};\boldsymbol{\lambda})\boldsymbol{\Psi}^{-1}(m_{11}\boldsymbol{m}_0 - m_{01}\boldsymbol{m}_1),\end{aligned}$$

$$\text{vec}(\boldsymbol{B}_1) = (m_{00}m_{11} - m_{01}^2)^{-1}\big((\boldsymbol{I}_p \otimes \boldsymbol{X})^\top(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Delta}^{-1})(\boldsymbol{I}_p \otimes \boldsymbol{X})\big)^{-1}(\boldsymbol{I}_p \otimes \boldsymbol{X})^\top$$
$$\times (\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Delta}^{-1})\mathcal{T}(\tilde{\mathcal{Y}}; \boldsymbol{\lambda})\boldsymbol{\Psi}^{-1}(m_{00}\boldsymbol{m}_1 - m_{01}\boldsymbol{m}_0).$$

The optimization of the log-likelihood function is now readily available through an iterative numerical optimization procedure.

3.3 Applications

In this section, we consider two real-life applications that illustrate the range of possible applications of the developed methodology.

3.3.1 Salaries in four major universities in Alabama

There are four major universities in Alabama that are recognized by Carnegie classification as universities with high research activity. These four schools include the University of Alabama at Tuscaloosa, University of Alabama at Birmingham, University of Alabama at Huntsville, and Auburn university. The goal of the study considered in this section is to investigate whether there were change points in salaries paid at each of these four universities in the recent years. While the information on salaries at these four schools can be found in public records, we obtain it from the Web-site http://data.chronicle.com. This site provides self-reported data, thus it just mimics the real salary data. For each university, the information is summarized by gender (males, females) and position rank (Assistant, Associate, and Full Professors) over 13 years (from 2003/2004 to 2015/2016). As a result, the data collected for each school represent a tensor with dimensions $2 \times 3 \times 13$. The methodology considered in Section 3.2.2 can be readily applied in this setting.

Table 3.3.1 presents the most important findings obtained in the course of running the proposed approach over all possible shift change points. To investigate the importance of incorporating transformations into the model, three models are studied. The first one considered is a

Table 3.3.1: Study of professor salaries at four universities in Alabama. The results are obtained without (None) and with transformation parameters (Exponential and Power).

	University	Transformation	$\hat{t}_1$	$\hat{t}_2$	$\log \mathcal{L}$	BIC
1	UA Tuscaloosa	None	–	–	354.26	−643.17
		Exponential	–	–	354.29	−625.79
		Power	–	–	354.43	−626.08
2	UA Birmingham	None	–	–	353.62	−641.90
		Exponential	2005	2012	385.42	−635.78
		Power	2005	2012	385.46	−635.86
3	UA Huntsville	None	2007	2012	374.39	−631.14
		Exponential	2007	2012	374.41	−613.75
		Power	2007	2012	374.40	−613.74
4	Auburn	None	2005	2006	392.68	−667.74
		Exponential	2005	2006	408.97	−682.89
		Power	2005	2006	409.08	−683.11

tensor Gaussian model, while the other two involve exponential and power transformations. Several interesting findings can be reported based on the results presented in Table 3.3.1. Also, illustrations are provided in Figure 3.3.1. The blue and red colors represent salaries earned by males and females, respectively. Circles, squares, and triangles illustrate salaries received by Full, Associate, and Assistant Professors.

For the UA Tuscaloosa, all three models found no change points. In addition, the reported log-likelihood values are very close to each other. This is an indication that there is no skewness in salary data for this university. For the UA Birmingham, different models are found by methods with and without transformations incorporated. In the first case, there are no change points detected and in the second case, there are two change points at times $\hat{t}_1 = 2005$ and $\hat{t}_2 = 2012$. While there is nearly 35-unit difference in log-likelihood values corresponding to the models, the BIC value of -641.90 suggests that the model without transformation parameters should be prefered. Thus, no change points have been detected for this university as well. The third case considered is for the UA Huntsville. The log-likelihood values are again very similar implying that there is no skewness in data. In the meantime, all three models identified two change points at times $\hat{t}_1 = 2007$ and $\hat{t}_2 = 2012$. As we can see from the corresponding plot in Figure 3.3.1,

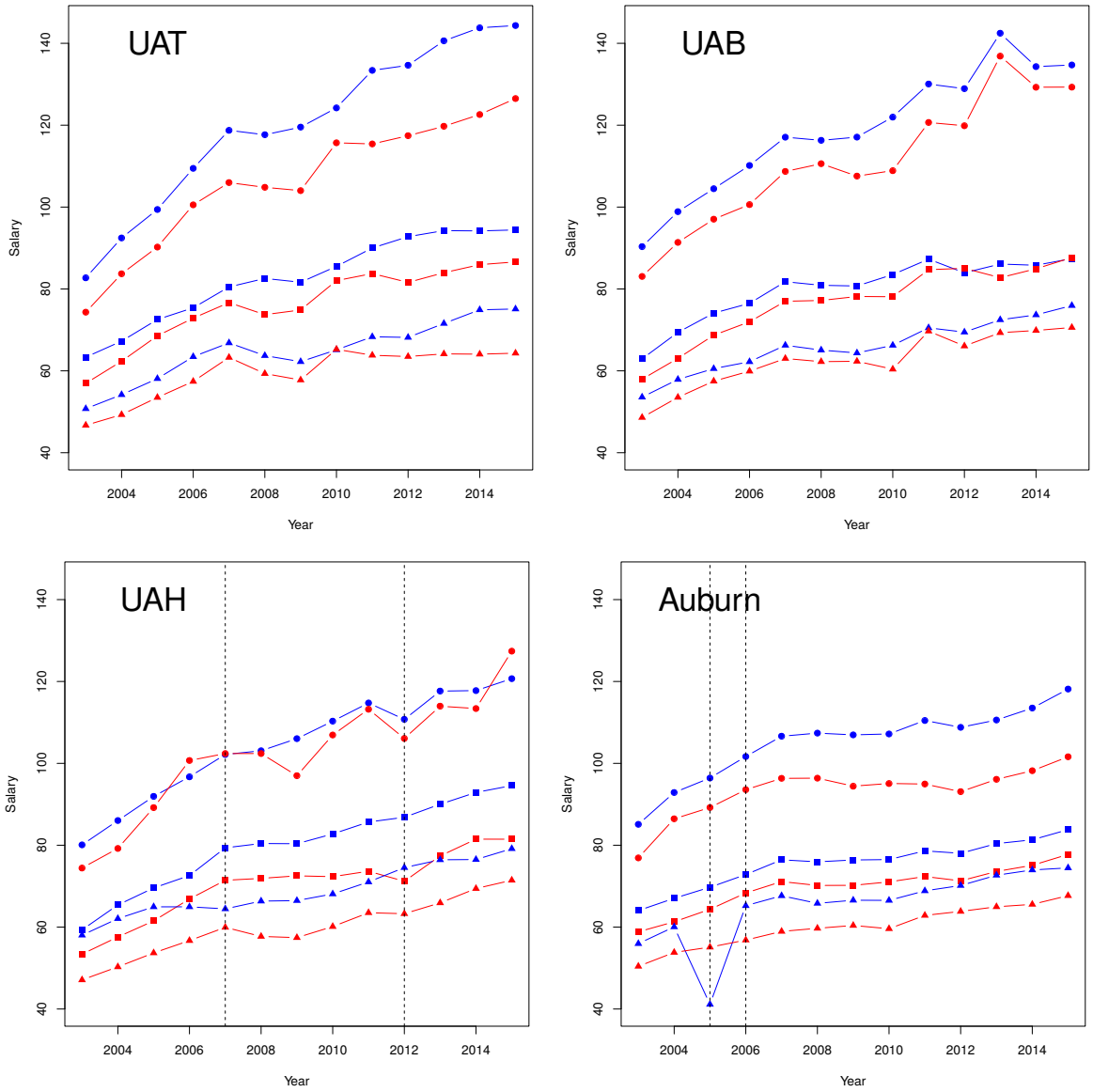


Figure 3.3.1: Salaries for the four major research universities in Alabama. Blue and red colors represent males and females. Circles, squares, and triangles illustrate Full, Associate, and Assistant Professors. Vertical dashed lines show estimated change point times.

both change points can be justified by visual inspection. In particular, a slight direction change is observed for several sequences in 2007. Also, Full Professor and female Associate Professor salaries decreased in 2012. The last case considered involves Auburn university. Despite the fact that the same change points have been detected by all three methods ($\hat{t}_1 = 2005$ and $\hat{t}_2 = 2006$), we can notice that the best models in terms of BIC are those corresponding to transformations.

This finding clearly suggests the presence of skewness in the salary data for Auburn. The plot illustrating Auburn salaries in Figure 3.3.1, suggests that there was a considerable drop in male Assistant Professor salaries at 2005. In 2006, the salaries rebounded to previously observed values causing the other change point. In all cases, models based on exponential and power transformations produced nearly the same results. This implies that both transformations are almost equally effective in reaching near-normality.

3.3.2 Crime rates in major US cities

Now, we study crime rates in 125 major American cities. The data were obtained from the US Department of Justice, FBI Web-site (http://www.ucrdatatool.gov/Search/Crime/Crime.cfm). There are seven crime types combined into two categories: violent and property crimes. The data are provided for 13 years between 2000 and 2012, inclusively. The 125 largest cities with populations over 100,000 people have been identified in the following five regions: West (WA, OR, MT, ID, WY, NV, UT, CO, CA, AK, HI), MidWest (ND, SD, MN, NE, IA, KS, MO, IL, WI, IN, MI, OH), NorthEast (PA, NY, VT, ME, NH, MA, RI, CT, DE, MD, NJ, DC), SouthWest (AZ, NM, TX, OK), and SouthEast (AR, LA, MS, AL, TN, GA, FL, SC, NC, KY, VA, WV). Thus, in our considered framework, there are 125 cities within 5 equally represented blocks, 2 crime categories, and 13 years. The data can be summarized in the form of a $125 \times 2 \times 13$ tensor and the methodology discussed in Section 3.2.3 can be readily applied.

Table 3.3.2 provides results obtained by the same three models as those employed in Section 3.3.1. As we can see, all three models cannot find change points in the crime data based on BIC. The worst performance is demonstrated by the model with no transformation incorporated. This observation suggests that data are severely skewed. The other two models are more similar, but the exponential transformation is better in all cases. This remark implies that the Manly

Table 3.3.2: Log-likelihood, BIC, and p-value results obtained without (None) and with transformation parameters (Exponential and Power).

	None			Power			Exponential		
$\hat{t}_1$	$\log \mathcal{L}$	BIC	p-value	$\log \mathcal{L}$	BIC	p-value	$\log \mathcal{L}$	BIC	p-value
–	415.25	-709.20		815.71	-1493.95		854.19	-1570.92	
2001	419.96	-637.75	0.493	820.75	-1423.16	0.434	859.20	-1500.06	0.439
2002	418.01	-633.86	0.854	818.05	-1417.76	0.912	856.66	-1494.98	0.895
2003	418.33	-634.50	0.802	818.00	-1417.67	0.917	856.42	-1494.51	0.924
2004	420.94	-639.73	0.329	820.10	-1421.87	0.553	858.55	-1498.77	0.559
2005	426.04	-649.91	0.017	824.14	-1429.94	0.078	862.91	-1507.49	0.065
2006	422.07	-641.98	0.190	822.21	-1426.08	0.224	860.52	-1502.72	0.243
2007	420.83	-639.50	0.345	820.64	-1422.95	0.452	859.13	-1499.92	0.451
2008	423.22	-644.28	0.101	822.74	-1427.15	0.170	861.13	-1503.93	0.179
2009	425.17	-648.19	0.031	827.03	-1435.72	0.012	865.40	-1512.46	0.013
2010	421.04	-639.91	0.314	822.82	-1427.30	0.163	861.09	-1503.84	0.182
2011	418.79	-635.42	0.718	819.50	-1420.66	0.670	857.96	-1497.59	0.674
2012	421.02	-639.88	0.317	824.00	-1429.66	0.084	862.66	-1506.98	0.076

transformation is more successful in reaching near-normality for these particular data. Among all models with change points, the one with $\hat{t}_1 = 2009$ produces the lowest BIC value equal to -1512.46. The corresponding p-value obtained from the likelihood ratio test suggests that the change point at 2009 is significant with either method applied. This finding makes good sense as the year of 2009 is the year of the world financial crisis.

3.4 Discussion

In this paper, we developed an efficient method capable of estimating multiple change points in multivariate processes. The proposed technique relies on the matrix normal distribution adjusted by the exponential Manly transformation. Such an adjustment makes the proposed methodology robust to violations of the normality assumption. The matrix setting has an appealing form as rows can represent variables and columns can be associated with time points. Based on the results of challenging simulation studies, we can conclude that the proposed technique is very promising. It outperforms the two non-parametric competitors in all settings dramatically. Two applications to crime data considered in the paper demonstrate the usefulness of our method.

REFERENCES

[1] H Akaike. Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory*, pages 267–281, 1973.

[2] D. F. Andrews, R. Gnanadesikan, and J. L. Warner. Transformations of multivariate data. *Biometrics*, 27(4):825–840, 1971.

[3] Y. Baddour, R. Tholmer, and P. Gavit. Use of change-point analysis for process monitoring and control. *BioPharm International*, 22, 2009.

[4] George E.P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26(2):211–252, 1964.

[5] J. Chen and A. K. Gupta. Testing and locating variance changepoints with application to stock prices. *Jornal of the American Statistical Association*, 92:739–747, 1997.

[6] J. Chen and A. K. Gupta. Statistical inference of covariance change points in gaussian model. *Journal of Theoretical and Applied Statistics*, 38:17–28, 2004.

[7] J. Chen and A. K. Gupta. *Parametric statistical change point analysis*. Springer, 2nd edition, 2011.

[8] P. Coppin, I. Jonckheere, B. Nackaerts, B. Muys, and E. Lambin. Review articlegigital change detection methods in ecosystem monitoring: a review. *International Journal of Remote Sensing*, 25:1565–1596, 2004.

[9] W. W. Davis. Robust methods for detection of shifts of the innovation variance of a time series. *Technometrics*, 21:313–320, 1979.

[10] L. S. Guild, W. B. Cohen, and J. B. Kauffman. Detection of deforestation and land conversion in rondania, brazil using change detection techniques. *International Journal of Remote Sensing*, 25:731–750, 2004.

[11] A. K. Gupta and D. K. Nagar. *Matrix variate distributions*. Chapman & Hall / CRC, 2000.

[12] C. B. Hall, J. Ying, L. Kuo, M. Sliwinski, H. Buschke, M. Katz, and R. B. Lipton. Estimation of bivariate measurements having different change point, with application to cognitive ageing. *Statistics in Medicine*, 20:3695–3714, 2001.

[13] D. M. Hawkins. Testing a sequence of observations for a shift in location. *Journal of the American Statistical Association*, 72:180–186, 1977.

[14] L. Horváth. The maximum likelihood method for testing changes in the parameters of normal observations. *Annals of Statistics*, 21:671–680, 1993.

[15] D. A. Hsu. Tests for variance shifts at an unknown time point. *Applied Statistics*, 26:279–284, 1977.

[16] C. Inclán. Detection of multiple changes of variance using posterior odds. *Journal of Business and Economics Statistics*, 11:189–300, 1993.

[17] James, N. A. and Matteson, D. S. ecp: An R package for nonparametric multiple change point analysis of multivariate data. *Journal of Statistical Software*, 62:1–25, 2014.

[18] A. T. Kass-Hout, Z. Xu, P. McMurray, S. Park, D. Buckeridge, J. S. Brownstein, L. Finelli, and S. L. Groseclose. Application of change point analysis to daily influenza-like illness emergency department visits. *Journal of the American Medical Informatics Association*, 19:1075–1081, 2012.

[19] D. Kifer, S. Ben-David, and J. Gehrke. Detecting change in data streams. *International Conference on Very Large Data Bases*, 30:180–191, 2004.

[20] W. J. Krzanowski and F. H. C. Marriott. *Multivariate Analysis, part 1: Distributions, Ordination and Inference*. John Wiley & Sons Inc, 1994.

[21] J. Lee, H Su, P. E Cheng, M. Liou, J. A. D Aston, A. C. Tsai, and C. Che. MR image segmentation using a power transformation approach. *IEEE transactions on medical imaging*, 28(6):894–905, 2009.

[22] M. J. Lenardon and A. Amirdjanove. Interaction between stock indices via changepoint analysis. *Applied stochastic models in business and industry*, 22:573–586, 2006.

[23] C. Lindsey and S Sheather. Power transformation via multivariate Box-Cox. *The Stata Journal*, 10(1):69–81, 2010.

[24] B. F. J. Manly. Exponential data transformations. *Journal of the Royal Statistical Society, Series D*, 25(1):37–42, 1976.

[25] J. A. Nelder and R. Mead. A simplex algorithm for function minimization. *Computer Journal*, 7(4):308 – 313, 1965.

[26] M. B. Nigro, S. N Pakzad, and S. Dorvash. Localized structural damage detction:a change point analysis. *Computer-Aided civil and infrastructure engineering*, 29:416–432, 2014.

[27] S. Nyambura, S. Mundai, and A. Waititu. Estimation of change point in poissin random variable using the maximum likelihood method. *American Journal of Theoretical and Applied Statistics*, 5:219–224, 2016.

[28] E. S. Page. On problem in which a change in parameter occurs at an unknown points. *Biometrika*, 42:248–252, 1957.

[29] S. H. Patel, S. J. Morreale, A. P. Panagopoulou, H. Bailey, N. J. Robinson, F. V. Paladino, D. Margaritoulis, and J. R. Spotila. Changepoint analysis: a new approach for revealing animal movements and behaviors from satellite telemetry data. *Ecosphere*, 6:1–13, 2015.

[30] A. Pepelyshev and A. S. Polunchenko. Real-time financial surveillance via quickest change-point detection methods. *Statistics and its interface*, 0:1–14, 2015.

[31] M. B. Perry. Identifying the time of polynomial drift in the mean of autocorrelated processes. *Quality and Reliability Engineering International*, 25:399–415, 2010.

[32] M. B. Perry and J. J. Pignatiello. A change point model for the location parameter of exponential family densities. *IIE Transactions*, 40:947–956, 2008.

[33] A. N. Pettitt. A non-parametric approach to the change point problem. *Journal of the American Statistical Association*, 28:126–135, 1979.

[34] A. J. Quiroz, M. Nakamura, and F. J. Perez. Estimation of a multivariate box-cox transforamtion to elliptical symmetry via the empirical characteristic function. *J. Instit. Statist. Math.*, 48:687–709, 1996.

[35] M.L. Rizzo and G.J. Szekely. Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method. *Journal of Classification*, 22:151–183, 2005.

[36] M.L. Rizzo and G.J. Szekely. Disco analysis: A nonparametric extension of analysis of variance. *The Annals of Applied Statistics*, 4:1034–1055, 2010.

[37] G. Schwarz. Estimating the dimensions of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[38] A. K. Sen and M. S. Srivastava. On multivariate tests for detecting change in mean. *Sankhyá*, A35:173–186, 1973.

[39] M. S. Srivastava and K. J. Worsley. Likelihood ratio tests for a chance in the multivariate normal mean. *Journal of the American Statistical Association*, 81:199–204, 1986.

[40] S. Velilla. A note on the multivariate Box-Cox transformation to normality. *Statistics & Probability Letters*, 17(4):259–263, 1993.

[41] S. Weiss. Fluorescence spectroscopy of single biomolecules. *Science*, 283:1676–1683, 1999.

[42] K. J. Worsley. On the likelihood ratio test for a shift in location of normal populations. *Journal of the American Statistical Association*, 74:365–367, 1979.

[43] I.-K. Yeo and R. A. Johnson. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87:954–959, 2000.

[44] L. C. Zhao, P. R. Krishnaiah, and Z. D. Bai. On detection of the number of signals in presence of white noise. *Journal of Multivariate Analysis*, 20:1–25, 1986a.

[45] L. C. Zhao, P. R. Krishnaiah, and Z. D. Bai. On detection of the number of signals when the noise covariance matrix is arbitrary. *Journal of Multivariate Analysis*, 20:26–49, 1986b.

[46] X. Zhu and V. Melnykov. Manly transformation in finite mixture modeling. *accepted by Computational Statistics and Data Analysis*, 2016.