

Robust Estimation of Multiple Change Points in Multivariate Processes



Yana Melnykov, Marcus Perry, and Volodymyr Melnykov

Abstract Change point inference is important in various fields of science. Many different procedures have been proposed in the 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 for identifying multiple change points in multivariate processes. 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 its competitors. The application of the methodology to real-life datasets highlights its usefulness and broad applicability.

1 Introduction

The change point estimation in sequential data has 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 [1], ecology [2], pharmacy [3], engineering [4], finance [5, 6], 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

Y. Melnykov · M. Perry · V. Melnykov (✉)
The University of Alabama, Tuscaloosa AL, 35487, USA
e-mail: vmelnykov@cba.ua.edu

Y. Melnykov
e-mail: ymelnykov@cba.ua.edu

M. Perry
e-mail: mperry@cba.ua.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
A. Bekker et al. (eds.), *Innovations in Multivariate Statistical Modeling*,
Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-13971-0_3

devoted to the estimation of a change point in means of univariate normal distributions [7]. The problem with a constant mean but possible shift in variance parameters was considered by [8–11]. A generalization of both ideas was considered by [12] who developed a test capable of detecting a change in mean and variance parameters simultaneously.

Attention has been paid to multivariate settings as well. [13] and [14] 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 [15] and [16]. In the same setting of multivariate normal distribution, [17] proposed a procedure for estimating a change in covariance matrices under the assumption of a constant mean vector. Recently, [18] 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 [19, 20], nonparametric methods [21] including probabilistic pruning based on various goodness-of-fit measures [22], and some others.

In this paper, we consider the problem of estimating multiple change points in the framework with multivariate processes. The importance of this problem is rather substantial but the number of existing methods is very limited (e.g., see discussion on this topic in [22]). The most traditional approach taken by the majority of researchers assumes the independence of observations over time as well as their multivariate normality. Unfortunately, both assumptions are often inadequate or unrealistic. Among other alternatives, there are two nonparametric procedures employing probabilistic pruning with Energy statistic [23] and Kolmogorov-Smirnov statistic [24] that are available through the R package ECP [22]. It is worth mentioning that this R package is currently the only one that aims at identifying multiple change points in the multivariate setting. The lack of developments in this important area of change point inference motivates our methodology. Our proposed technique is based on 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 presents the proposed methodology. Section 3 investigates the performance of our procedure and three competitors in various settings. Section 4 applies the developed methods to the analysis of real-life data. The paper concludes with a discussion provided in Sect. 5.

2 Methodology

Matrix Normal Distribution

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

$$\mathbf{Y} = \begin{pmatrix} y_{11} & y_{12} & \dots & y_{1T} \\ y_{21} & y_{22} & \dots & y_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ y_{p1} & y_{p2} & \dots & y_{pT} \end{pmatrix}. \quad (1)$$

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 \mathbf{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 [25] that has the following probability density function (pdf):

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

where \mathbf{Y} is the $p \times T$ matrix argument defined in (1) and \mathbf{M} is a $p \times T$ mean matrix. The $p \times p$ matrix $\boldsymbol{\Sigma}$ and $T \times T$ matrix $\boldsymbol{\Psi}$ are covariance matrices that model variability associated with rows and columns, respectively. Also, $\text{tr}\{\cdot\}$ denotes the trace operator. It can be shown that $\text{vec}(\mathbf{Y}) \sim \mathcal{N}_{pT}(\text{vec}(\mathbf{M}), \boldsymbol{\Psi} \otimes \boldsymbol{\Sigma})$, where $\text{vec}(\cdot)$ 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 $\text{vec}(\mathbf{M})$ and covariance matrix $\boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}$. There is a minor non-identifiability issue caused by the properties of the Kronecker product since $a\boldsymbol{\Psi} \otimes \boldsymbol{\Sigma} = \boldsymbol{\Psi} \otimes a\boldsymbol{\Sigma}$ for any multiplier $a \in \mathbb{R}^+$. One simple restriction on $\boldsymbol{\Psi}$ or $\boldsymbol{\Sigma}$ 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 $\text{Cov}(y_{jt}, y_{j't'})$ 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 $\boldsymbol{\Psi}$ can be conveniently parameterized in terms of a desired time series process. In this paper, we illustrate the methodology based on the autoregressive

process of order 1 (AR(1)). Incorporating moving average or higher order autoregressive processes is very similar as it affects just the covariance matrix Ψ . In fact, the AR(1) model has been chosen as an illustration simply because it yields the best results for the application considered in Sect. 4. Under AR(1), the covariance matrix Ψ is given by

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

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 $\Psi \otimes \Sigma$ is to set $\delta^2 = 1 - \phi^2$. This restriction immediately leads to $\Psi \equiv \mathbf{R}_\phi$, where \mathbf{R}_ϕ denotes the corresponding correlation matrix that relies on a single parameter ϕ . It can be shown that

$$|\Psi| \equiv |\mathbf{R}_\phi| = (1 - \phi^2)^{T-1} \quad \text{and} \quad \Psi^{-1} \equiv \mathbf{R}_\phi^{-1} = \frac{1}{1 - \phi^2} (\mathbf{I}_T - \phi \mathbf{J}_1 + \phi^2 \mathbf{J}_2), \tag{3}$$

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

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

Expressions in (3) are helpful for speedier maximum likelihood estimation as the potentially time consuming inversion of the $T \times T$ covariance matrix Ψ can be completely avoided.

Change Point Estimation

Consider the problem of estimating change points in the given framework. Let μ_0 be the p -variate mean vector associated with the main process. Suppose, there are K alternative processes with means $\mu_1, \mu_2, \dots, \mu_K$. Then, the mean matrix \mathbf{M} can be written as $\mathbf{M} = \sum_{k=0}^K \mu_k \mathbf{m}_k^\top$, where \mathbf{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 $\sum_{k=0}^K \mathbf{m}_k = \mathbf{1}_T$, where $\mathbf{1}_T$ is the vector of length T with all elements equal to 1. It can be noted that vectors \mathbf{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, \dots, t_K , the mean matrix is given by

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

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

The log-likelihood function corresponding to Eq. (2) has the following form:

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

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 [26], alternative families of power transformations as in [27], or the exponential transformation proposed by [28]. Let \mathcal{T} be an invertible and differentiable mapping representing 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 [29–31], i.e., the p -variate transformation is given by $\mathcal{T}(\mathbf{y}; \boldsymbol{\lambda}) = (\mathcal{T}(y_1; \lambda_1), \mathcal{T}(y_2; \lambda_2), \dots, \mathcal{T}(y_p; \lambda_p))^\top$, where the transformation parameter vector is given by $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_p)^\top$. This idea can be readily generalized to the matrix framework with $\mathcal{T}(\mathbf{Y}; \boldsymbol{\lambda})$ 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 \mathbf{M} and implementing the transformation idea, the log-likelihood function can be further written as

$$\begin{aligned} \log \mathcal{L}(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}, \phi, \boldsymbol{\lambda}) &= -\frac{pT}{2} \log(2\pi) - \frac{T}{2} \log |\boldsymbol{\Sigma}| - \frac{p(T-1)}{2} \log(1 - \phi^2) \\ &\quad - \frac{1}{2(1 - \phi^2)} \text{tr} \left\{ \boldsymbol{\Sigma}^{-1} (\mathcal{T}(\mathbf{Y}; \boldsymbol{\lambda}) - \sum_{k=0}^K \boldsymbol{\mu}_k \mathbf{m}_k^\top) (\mathbf{I}_T - \phi \mathbf{J}_1 + \phi^2 \mathbf{J}_2) \right. \\ &\quad \left. \times (\mathcal{T}(\mathbf{Y}; \boldsymbol{\lambda}) - \sum_{k=0}^K \boldsymbol{\mu}_k \mathbf{m}_k^\top)^\top \right\} + \log \left| \frac{\partial \mathcal{T}(\mathbf{Y}; \boldsymbol{\lambda})}{\partial \mathbf{Y}} \right|, \end{aligned} \quad (4)$$

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

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

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

where \mathbf{R}_ϕ^{-1} is as in (3). Solving a system of $K + 1$ equations leads to the expressions for $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K$. Maximum likelihood estimation for $\boldsymbol{\Sigma}_k$ yields the following expression:

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

Substituting expressions for $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K$ and $\boldsymbol{\Sigma}$ into the log-likelihood function (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 $\mathcal{T}(y; \lambda) = y^{I(\lambda=0)} (\exp\{\lambda y - 1\} \lambda^{-1})^{I(\lambda \neq 0)}$, where $I(\cdot)$ is the indicator function. In this setting, the log of Jacobian in (4) is given by $\boldsymbol{\lambda}^\top \mathbf{Y} \mathbf{1}_T$, where $\mathbf{1}_T = (1, 1, \dots, 1)^\top$ with cardinality $|\boldsymbol{\lambda}| = T$.

The problem of change point estimation requires assessing the number of processes. To avoid potential problems with the adjustment for multiple comparisons, simplify calculations, and avoid testing procedures in general, we employ the variant of the Bayesian Information Criterion (BIC) [32] proposed by [33] specifically for the change point framework. BIC is also an appealing option due to its connection to the Bayes factor commonly used in Bayesian inference for comparing competing models.

As a final note in this section, we would like to remark that the proposed procedure focuses on processes with mean vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_K$. In real-life applications, it is possible that just some parts of these vectors will be different while the remaining variables exhibit no change point behavior. The task of detecting changes in specific variables is a challenging standalone problem that is beyond the scope of this work. One practical approach can be to search for such variables after detecting differences in mean vectors first. Such a scenario is considered in Sect. 4.

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

Table 1 Parameter values used in the simulation study of Sect. 3

j	μ_0	μ_1	μ_2	Σ			λ	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	

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 1.

Various levels of correlation and scaling as reflected by parameters ϕ and Σ , respectively, are studied. In particular, we consider $\phi = 0.1, 0.5, 0.9$ and $\Sigma, \Sigma/2, \Sigma/4$. 250 datasets were simulated for each combination of the covariance matrix and correlation parameter in both considered setting, thus, yielding 4,500 simulated datasets in total. The proposed technique assumes that the exact location of change points is known. The quality of the model fit is assessed by means of BIC. It can be noticed that in the search for the optimal model with K change points, $(T - 1)!/(T - 1 - K)!$ alternatives should be considered. As K is usually rather low, the approach is computationally feasible even for moderate T values. In our experiments, each model could be fitted in under one second. In addition, parallel computing can be readily implemented if the number of models becomes restrictively high.

The illustration of some simulated datasets can be found in Fig. 1. Here, plots (a) and (b) show datasets simulated with $\phi = 0.1$ but with different covariance matrices Σ and $\Sigma/4$, respectively. Plots (c) and (d) correspond to the same covariance matrices Σ and $\Sigma/4$ 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 μ_0, μ_1 , and μ_2 .

From examining Fig. 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

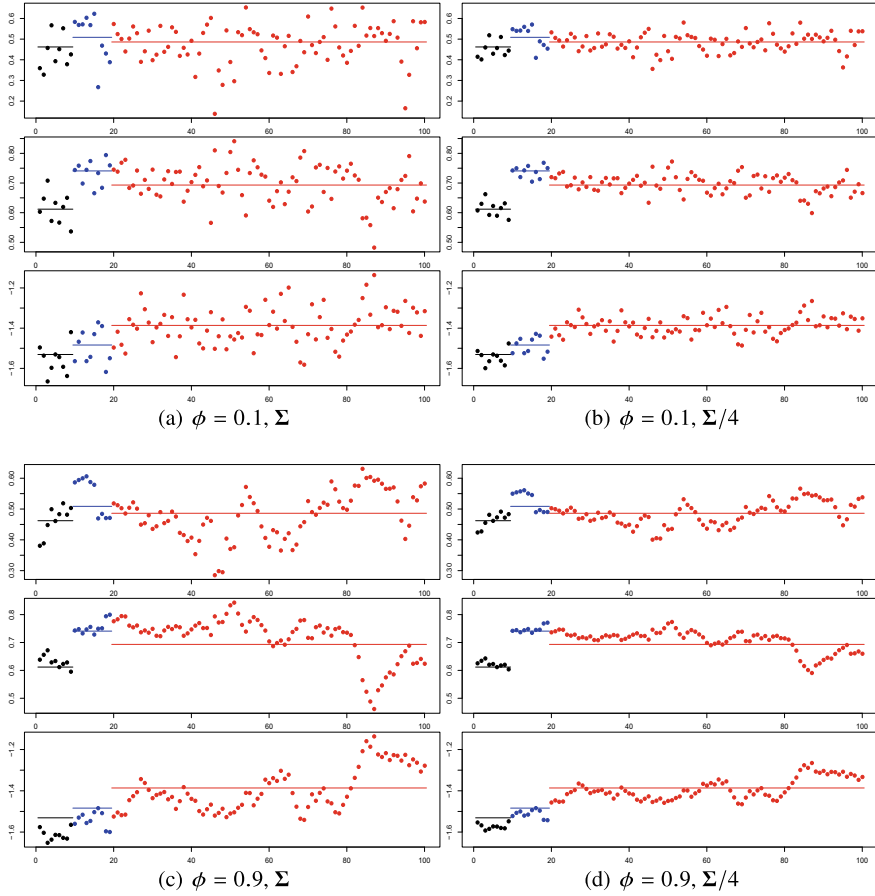


Fig. 1 Datasets generated in the course of the simulation study in Sect. 3 with different scaling (reflected by Σ and $\Sigma/4$) and correlation ($\phi = 0.1, 0.9$). Horizontal lines represent true back-transformed values of the corresponding coordinates of parameters μ_0 , μ_1 , and μ_2

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 [22], the number of procedures capable of estimating multiple change points in multivariate processes is rather limited. In this section, the developed methodology is compared with one parametric approach that we call naive and two nonparametric procedures available for practitioners through the R package ECP [22]. 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 pruning with

Table 2 Interpretation of notation used in Tables 3 and 4

Notation	Interpretation
$\{t_1, t_2\}$	Both change points are correctly found
$\{t_1, t_2, x\}$	Both change points are correctly identified, but there are false change points found as well
$\{t_1, \tilde{t}_2\}/\{\tilde{t}_1, t_2\}$	One change point is identified correctly, the other one is close by, i.e. $ t_k - \hat{t}_k \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., $ t_k - \hat{t}_k > 3$

Energy statistic [23, 34] and Kolmogorov-Smirnov statistic [24] used as goodness-of-fit measures. Tables 3 and 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, were found.

As we can observe from Table 3, the proposed method can rather effectively identify change points. Expectedly, the performance of the procedure improves considerably when the variability decreases. For example, in the case with $\phi = 0.9$ and Σ , 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 $\Sigma/2$ and $\Sigma/4$, 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 $\Sigma/4$ 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 3, we can conclude that the closest competitor is the naive procedure. 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

Table 3 Simulation study from Sect. 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 (KS) used as the goodness-of-fit measure. The notation interpretation is provided in Table 2. The bold font highlights the proportion of times the correct combination was found

$K = 2$		Σ			$\Sigma/2$			$\Sigma/4$		
$t_1 = 10, t_2 = 20$		$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$
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	0.004	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

easiest case considered with $\phi = 0.9$ and $\Sigma/4$, 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 $\Sigma/4$, the Kolmogorov-Smirnov statistic (denoted as KS) 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 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 $\Sigma/4$ and $\phi = 0.9$. It is worth mentioning that similar analysis has been repeated for negative parameters $\phi = -0.9, -0.5, -0.1$. The results and findings of

Table 4 Simulation study from Sect. 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 3

$K = 2$		Σ			$\Sigma/2$			$\Sigma/4$		
$t_1 = 10, t_2 = 50$		$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$
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

these experiments were similar and consistent with those presented in this section. To conclude this section, we can remark that the proposed procedure proves to be a powerful tool for identifying change points.

4 Applications

Illustration on Crime Rates in US Cities

First, we apply the proposed methodology to the US cities crime data obtained from the US Department of Justice, Federal Bureau of Investigation Website (<http://www.ucrdatatool.gov/Search/Crime/Crime.cfm>). 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 2000 and 2012. As an example, we choose the data reported by Austin and Cincinnati Police Departments. Figure 2 illustrates violent (left column) and property (right

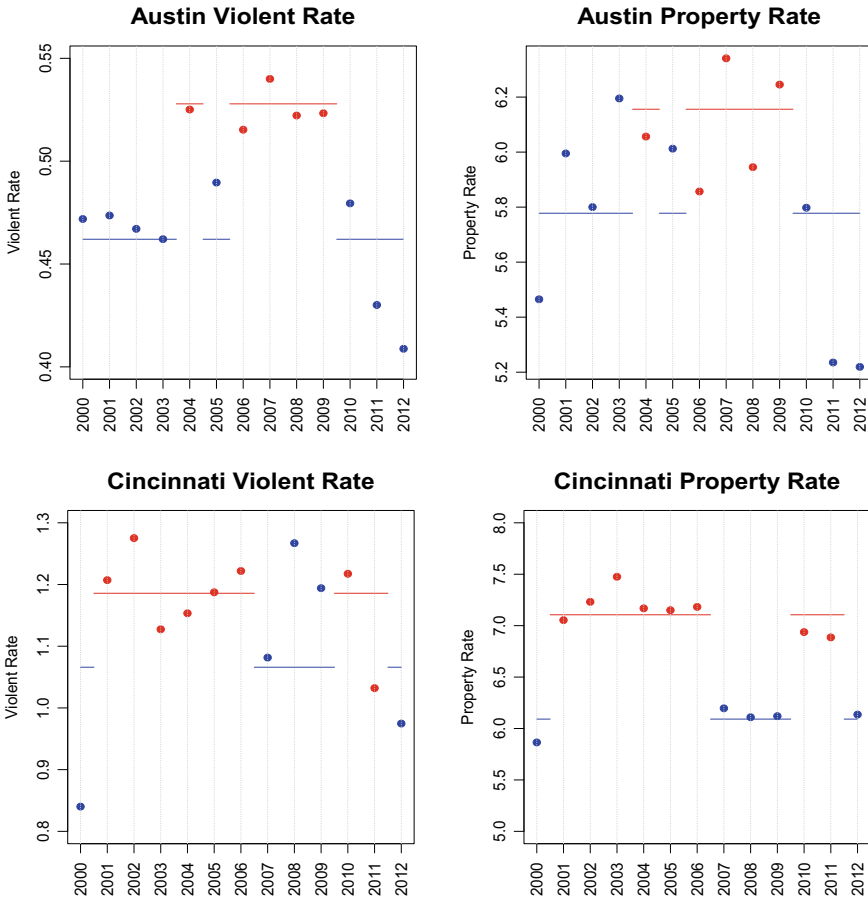


Fig. 2 Violent and Property crime rates in Austin and Cincinnati over the 13-year time period (2000-2012). The blue and red colors represent two processes detected. Horizontal lines stand for the means of the processes

column) crime rates. As the value $T = 13$ is quite low, instead of assuming models with shift-related change points only, we consider all possible orderings of processes.

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 orderings of processes, the lowest BIC of -47.081 was found. It is worth mentioning that the naive procedure outlined in Sect. 3 yields BIC -45.225 and the model with the AR(1) structure of Ψ but no transformation parameters produces BIC -44.099 . This suggests that even for so few data points as in the considered application, the proposed procedure can be useful. The parameter estimates associated with the model can be found in Table 5. A corresponding illustration is provided in the first row of plots in Fig. 2. Here, the years 2004, 2006, 2007, 2008, and 2009 are associated with

Table 5 Parameter estimates, log-likelihood, and BIC values for Austin and Cincinnati

City	$\hat{\mu}_0$	$\hat{\mu}_1$	$\hat{\Sigma}$		$\hat{\lambda}$	$\hat{\phi}$	$\log \mathcal{L}$	BIC
Austin	168.234	524.023	4422.5	105136.9	17.258	-0.402	39.831	-47.081
	4941.351	8870.934	105136.9	5810, 522	1.548			
Cincinnati	4.130	5.478	1.372	0.004	2.148	0.315	19.693	-6.804
	2.394	2.480	0.004	0.0001	-0.375			

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 $\hat{\mu}_0$ and $\hat{\mu}_1$ 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.

The opposite situation is observed for Cincinnati (second row in Fig. 2): the variable *Property Crime* contributes to the separation of the processes more than *Violent Crime*. Model parameters are also provided in Table 5. The BIC value of the best model detected is equal to -6.804 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). The BIC value associated with the naive approach is equal to -10.846 suggesting that AR(1) structure of Ψ as well as transformation-related parameters do not bring an improvement to the naive model in this case.

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 which 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 Department of Public Safety Website (<https://www.colorado.gov/pacific/cbi/crime-colorado1>) for 10 years: from 2007 to 2016. The same seven variables as described in Sect. 4 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 in 2014 yields BIC equal to $-1,006.1$. The likelihood ratio test conducted to verify 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 3 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 Fig. 2. 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 grown in the last 3 years. Indeed, the proposed analysis does not assume any cause-and-effect conclusions. In fact, we can notice a considerable decrease in *Murder* rates in 2014 and we can also observe that the increase in *Rape* rates began in 2013, i.e., 1 year earlier than when Amendment 64 became effective. Nevertheless, it is obvious that the proposed methodology presents a powerful exploratory tool for studying the effects of public policies.

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 nonparametric competitors in all settings. Two applications to crime data considered in the paper demonstrate the usefulness of the developed method.

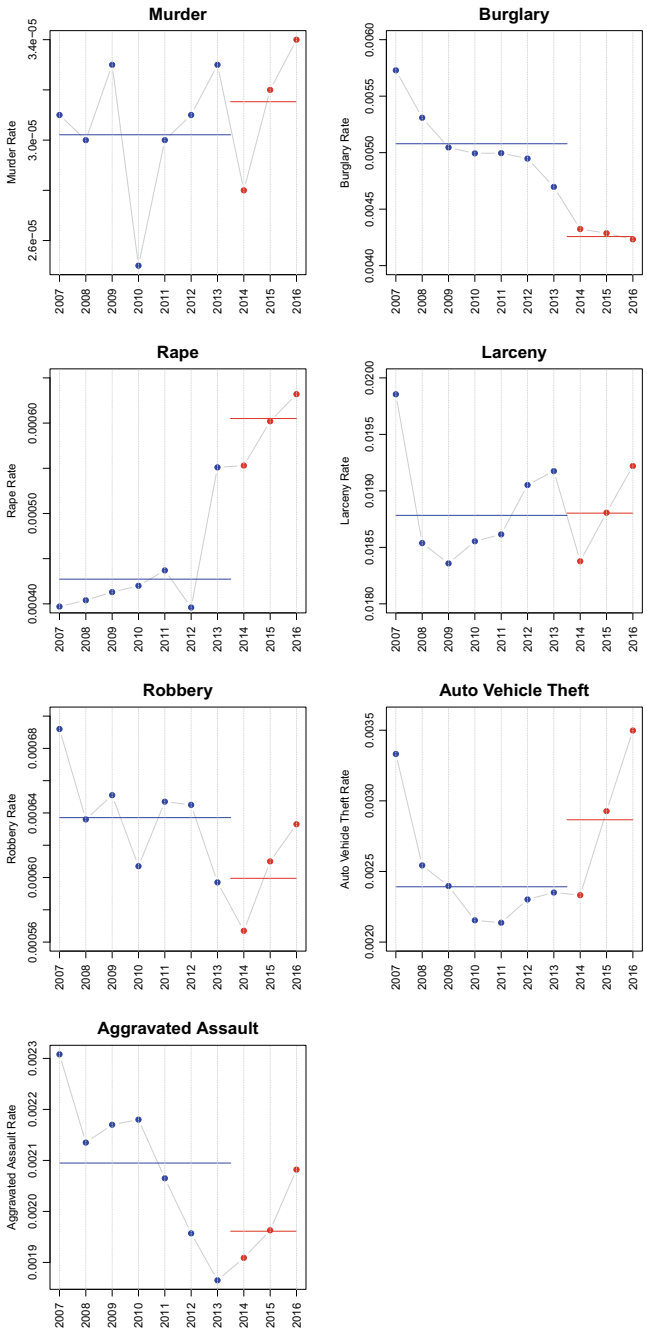


Fig. 3 Crime rates in Colorado over the 10-year time period. The blue and red colors represent two processes. Horizontal lines stand for the back-transformed means of the processes

References

1. Kass-Hout, A. T., Xu, Z., McMurray, P., Park, S., Buckenridge, D., Brownstein, J. S., et al. (2012). Application of change point analysis to daily influenza-like illness emergency department visits. *Journal of the American Medical Informatics Association*, *19*, 1075–1081.
2. Patel, S. H., Morreale, S. J., Panagopoulou, A. P., Bailey, H., Robinson, N. J., Paladino, F. V., et al. (2015). Change-point analysis: A new approach for revealing animal movements and behaviors from satellite telemetry data. *Ecosphere*, *6*, 1–13.
3. Baddour, Y., Tholmer, R., & Gavit, P. (2009). Use of change-point analysis for process monitoring and control. *BioPharm international* (Vol. 22).
4. Nigro, M. B., Pakzad, S. N., & Dorvash, S. (2014). Localized structural damage detection: A change point analysis. *Computer-Aided civil and infrastructure engineering*, *29*, 416–432.
5. Lenardon, M. J., & Amirdjanove, A. (2006). Interaction between stock indices via change-point analysis. *Applied Stochastic Models in Business and Industry*, *22*, 573–586.
6. Pepelyshev, A., & Polunchenko, A. S. (2015). Real-time financial surveillance via quickest change-point detection methods. *Statistics and its interface* (Vol. 0, pp. 1–14).
7. Page, E. S. (1957). On problem in which a change in parameter occurs at an unknown points. *Biometrika*, *42*, 248–252.
8. Hsu, D. A. (1977). Tests for variance shifts at an unknown time point. *Applied Statistics*, *26*, 279–284.
9. Davis, W. W. (1979). Robust methods for detection of shifts of the innovation variance of a time series. *Technometrics*, *21*, 313–320.
10. Inclán, C. (1993). Detection of multiple changes of variance using posterior odds. *Journal of Business and Economics Statistics*, *11*, 189–300.
11. Chen, J., & Gupta, A. K. (1997). Testing and locating variance change-points with application to stock prices. *Journal of the American Statistical Association*, *92*, 739–747.
12. Horváth, L. (1993). The maximum likelihood method for testing changes in the parameters of normal observations. *Annals of Statistics*, *21*, 671–680.
13. Sen, A. K., & Srivastava, M. S. (1973). On multivariate tests for detecting change in mean. *Sankhyá*, *A35*, 173–186.
14. Srivastava, M. S., & Worsley, K. J. (1986). Likelihood ratio tests for a change in the multivariate normal mean. *Journal of the American Statistical Association*, *81*, 199–204.
15. Zhao, L. C., Krishnaiah, P. R., & Bai, Z. D. (1986). On detection of the number of signals in presence of white noise. *Journal of Multivariate Analysis*, *20*, 1–25.
16. Zhao, L. C., Krishnaiah, P. R., & Bai, Z. D. (1986). On detection of the number of signals when the noise covariance matrix is arbitrary. *Journal of Multivariate Analysis*, *20*, 26–49.
17. Chen, J., & Gupta, A. K. (2004). Statistical inference of covariance change points in Gaussian model. *Journal of Theoretical and Applied Statistics*, *38*, 17–28.
18. Chen, J., & Gupta, A. K. (2011). *Parametric statistical change point analysis*, 2nd ed. Springer.
19. Perry, M. B., & Pignatiello, J. J. (2008). A change point model for the location parameter of exponential family densities. *IIE Transactions*, *40*, 947–956.
20. Nyambura, S., Mundai, S., & Waititu, A. (2016). Estimation of change point in Poisson random variable using the maximum likelihood method. *American Journal of Theoretical and Applied Statistics*, *5*, 219–224.
21. Pettitt, A. N. (1979). A non-parametric approach to the change point problem. *Journal of the American Statistical Association*, *28*, 126–135.
22. James, N. A., & Matteson, D. S. (2014). ECP: An R package for nonparametric multiple change point analysis of multivariate data. *Journal of Statistical Software*, *62*, 1–25.
23. Rizzo, M., & Szekely, G. (2005). Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method. *Journal of Classification*, *22*, 151–183.
24. Kifer, D., Ben-David, S., & Gehrke, J. (2004). Detecting change in data streams. *International Conference on Very Large Data Bases*, *30*, 180–191.
25. Krzanowski, W. J., & Marriott, F. H. C. (1994). *Multivariate analysis, part 1: Distributions, ordination and inference*. Wiley.

26. Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26(2), 211–252.
27. Yeo, I.-K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87, 954–959.
28. Manly, B. F. J. (1976). Exponential data transformations. *Journal of the Royal Statistical Society, Series D*, 25(1), 37–42.
29. Andrews, D. F., Gnanadesikan, R., & Warner, J. L. (1971). Transformations of multivariate data. *Biometrics*, 27(4), 825–840.
30. Lindsey, C., & Sheather, S. (2010). Power transformation via multivariate Box-Cox. *The Stata Journal*, 10(1), 69–81.
31. Zhu, X., & Melnykov, V. (2018). Manly transformation in finite mixture modeling. *Computational Statistics and Data Analysis*, 121, 190–208.
32. Schwarz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics*, 6(2), 461–464.
33. Shen, G., & Ghosh, J. (2011). Developing a new BIC for detecting change-points. *Journal of Statistical Planning & Inference*, 141, 1436–1447.
34. Rizzo, M., & Szekely, G. (2010). Disco analysis: A nonparametric extension of analysis of variance. *The Annals of Applied Statistics*, 4, 1034–1055.