Journal of Applied Mathematics and Physics, 2024, 12, 2634-2651
https://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

Changepoint Detection with Outliers Based on RWPCA

Xin Zhang, Sanzhi Shi, Yuting Guo

School of Mathematics and Statistics, Changchun University of Science and Technology, Changchun, China
Email: shisz@cust.edu.cn

How to cite this paper: Zhang, X., Shi, S.Z. and Guo, Y.T. (2024) Changepoint Detection with Outliers Based on RWPCA. *Journal of Applied Mathematics and Physics*, **12**, 2634-2651.
https://doi.org/10.4236/jamp.2024.127156

Received: June 20, 2024
Accepted: July 27, 2024
Published: July 30, 2024

Abstract

Changepoint detection faces challenges when outlier data are present. This paper proposes a multivariate changepoint detection method which is based on the robust WPCA projection direction and the robust RFPOP method, RWPCA-RFPOP method. Our method is double robust which is suitable for detecting mean changepoints in multivariate normal data with high correlations between variables that include outliers. Simulation results demonstrate that our method provides strong guarantees on both the number and location of changepoints in the presence of outliers. Finally, our method is well applied in an ACGH dataset.

Keywords

RWPCA-RFPOP, Double Robust, Outlier Detection, Biweight Loss

1. Introduction

There has been extensive research on the topic of changepoints. After Page (1954) [1] first applied changepoint detection to industrial quality control, changepoint detection method is received widespread attention and in-depth research from scholars. The related literature encompasses a variety of methods for detection both univariate and multivariate data, as well as single and multiple changepoints. For the univariate changepoint detection methods, Scott (1974) [2] introduced the Binary Segmentation (BS) algorithm for approximating multiple changepoint detections by segmenting the series into non-overlapping subsections and minimizing a cost function. Fryzlewicz's (2014) [3] proposed the Wild Binary Segmentation (WBS) method, which can consistently estimate the number and locations of multiple changepoints. However, changepoint detection faces challenges when outliers or abnormal data are present, as these outliers may be falsely identified as changepoints. Therefore, ensuring the robustness of

changepoint detections is extremely important. Fearnhead and Rigaill (2019) [4] introduced a robust mean changepoint detection method that can handle outliers effectively. Dehling (2020) [5] demonstrated that a method performs well under heavy-tailed noise distributions, utilizing the two-sample Hodges-Lehmann test statistic. Anastasiou and Fryzlewicz (2022) [6] presented a changepoint detection procedure based on an isolation technique. Kovacs *et al.* (2023) [7] introduced the Seeded Binary Segmentation algorithm, with a high probability. Each true changepoint is well covered by at least one interval which does not contain any other true changepoints. Fryzlewicz (2024) [8] presented a method for detecting localized regions in data sequences that contain a changepoint in the median. This method uses a novel sign-multiresolution sup-norm-type loss and greedily identifies the shortest intervals where constancy is significantly violated. Changepoint detection methodologies have been applied to a wide range of research fields, including medical diagnosis, finance, and network security.

The multivariable changepoint problem has garnered widespread attention since Horváth *et al.* (1999) [9]. Matteson *et al.* (2014) [10] introduced the E-divisive method, which is a non-parametric multiple changepoint detection for multivariate independent sequences. Jirak (2015) [11] considered an l_∞ -aggregation of the CUSUM statistics that works well for sparse changepoint. Cho and Fryzlewicz (2015) [12] proposed Sparse Binary Segmentation, which also takes sparsity into account and can be viewed as a hard thresholding of the CUSUM matrix followed by an l_1-aggregation. Knoblauch *et al.* (2018) [13] proposed a robust Bayesian online changepoint detection method based on β-divergence, offering double robustness in terms of parameters and changepoint posteriors. Wang (2020) [14] studied high-dimensional mean changepoint detection problems. Grundy *et al.* (2020) [15] proposed to project a multivariate dataset to two dimensions instead of one, allowing the detection of a change in mean and variance by applying univariate changepoint detection methods to the two projected series. Wendelberger *et al.* (2021) [16] employed a multiple linear regression framework for Bayesian online changepoint detection, accommodating seasonal trends and enhancing robustness against occasional outliers.

In changepoint analysis, the outliers affect the changepoint detection process, sometimes leading to the misinterpretation of outliers as changepoints. Therefore, it is important to consider the influence of outliers on the changepoint detection. Inspired by the RFPOP method proposed by Fearnhead and Rigaill (2019), this paper develops a changepoint detection method for multivariate data with outliers. Our method is suitable for detecting mean changes in multivariate data that contain outliers and noise, and exhibit high correlations between variables. We proposed two-step changepoint detection method: First, search for a robust projection direction; second, perform changepoint detection using the RFPOP method. In the step of searching for projection directions, outliers are replaced. Combining principal component analysis (PCA) with weighted projection, a robust projection direction is derived. This projection direction is then used to reduce the dimensionality of the original data. In the second step, this

paper employs the RFPOP method proposed by Fearnhead and Rigaill, RFPOP is robust to outliers in univariate data.

This paper is organized as follows: Section 2 introduces the double robust multivariate data changepoint detection RWPCA-RFPOP method. Section 3 evaluates the changepoint detection performance of the proposed method through Monte Carlo numerical simulations. Section 4 applies the RWPCA-RFPOP method to the analysis of real data. Section 5 summarizes the entire paper.

2. Multivariate Changepoint Detection Method

This section proposes the RWPCA-RFPOP method for multiple changepoint detection in multivariate data with outliers. The combination of weighted principal direction and the univariate changepoint detection RFPOP method is used to detect changepoints in multivariate data containing outliers.

2.1. Model Assumptions

In the multiple changepoints problem with multivariate data, let $X_1, X_2, \cdots, X_n$ denote the n p-dimensional random vectors. We consider the mean-change model:

$$X_i = \mu_k + \Sigma^{1/2}\varepsilon_i, \tau_{k-1}^* \le i \le \tau_k^* - 1, k = 1, \cdots, K+1; \tag{1}$$

where K is the true number of changepoints, μ_k is the mean vector between τ_{k-1}^* and $\tau_k^* - 1$, Σ indicates a positive definite matrix, τ_k^* is the locations of the changepoint in the sequence, ε_i is a p-dimensional random vector with mean zero mean and an identity covariance matrix. For convenience, the symbols are used as $\tau_0^* = 1$ and $\tau_{K+1}^* = n+1$. And $\mu_k \ne \mu_{k+1}$ (as long as one component of the μ vectors is not equal). Let J_k be the set of variables for which changes occur at τ_k^*, say $J_k = \left\{ j : \mu_{jk} \ne \mu_{j,k+1} \right\}$, and $J = \bigcup_{k=1}^{K} J_k$, where μ_{jk} is the jth component of μ_k. Our goal is to test whether there is at least one changepoint in the data, with $H_0 : K = 0$ versus $H_1 : K > 0$, and to further estimate the τ_k^* if the null hypothesis is rejected.

Assume that there is an outlier at position $m_j \left(j = 1, 2, \cdots, s \right)$, and the magnitude of the outlier is a constant $\Delta_j \left(j = 1, 2, \cdots, s \right)$, the data containing outliers is recorded as $Y = \left(Y_1, Y_2, \cdots, Y_n \right)$, then

$$\begin{cases} Y_{m_1} = X_{m_1} + \Delta_1 & i = m_1 \\ Y_{m_2} = X_{m_2} + \Delta_2 & i = m_2 \\ \vdots & \\ Y_{m_s} = X_{m_s} + \Delta_s & i = m_s \\ Y = X & \text{other} \end{cases} \tag{2}$$

If the observed value Y deviates from the confidence interval $\left[\mu_Y - \left(P * \sigma_Y \right), \mu_Y + \left(P * \sigma_Y \right) \right]$, which is considered an outlier. In the absence of a changepoint, a common value for the coefficient P is $P = 3$. When changepoints are present, the coefficient is relaxed to be equal to the changepoint jump magnitude.

2.2. Changepoint Detection Method

When there are outliers or abnormal data in the multivariate data, sometimes leading to the misinterpretation of outliers as changepoints. Considering multivariate data $Y=\left(Y_1,Y_2,\cdots,Y_n\right)$ with outliers, in this section, we combine the weighted principal component analysis WPCA and RFPOP algorithms to propose a multivariate data changepoint detection method that is doubly robust to outliers, termed RWPCA-RFPOP.

2.2.1. Finding Robust Projection Directions and Reducing Dimensionality

First, in order to find projection directions that are not influenced by outliers for dimensionality reduction of the original data $Y=\left(Y_1,Y_2,\cdots,Y_n\right)$, we preprocess the data. This paper uses Z-scores for outlier detection. A Z-score measures the degree of deviation of an observation from the mean, expressed as the number of standard deviations the observation is away from the mean:

$$Z=\frac{Y-\mu}{\sigma} \tag{3}$$

where Y represents the observed value, μ represents the mean, and σ represents the standard deviation. By calculating the Z-score, we can determine the degree of deviation of the observed value from the mean. When the Z-score is large, it indicates that the observed value deviates significantly from the mean; when the Z-score is small, it suggests that the observed value is close to the mean.

Secondly, we remove the outliers and replace them with the mean of each respective dimension at the original outlier positions $m_j\left(j=1,2,\cdots,s\right)$. For the multivariate data $Y'=\left(Y_1',Y_2',\cdots,Y_n'\right)$, after replacing outliers, the principal component directions are obtained through the singular value decomposition $Y'=U'S'V'^{\mathrm{T}}$ of the matrix. The first h principal components are selected to capture the majority of the information, and these principal components are multiplied by their corresponding weights ω_j to construct a matrix of weighted principal directions. The weights ω_j reflect the proportion of variance explained by a specific principal component. By aggregating the row of the weighted matrix, we obtain the robust projection direction $\hat{\omega}$. When the number of samples n is sufficiently large, the robust projection direction can also be obtained by directly removing outliers. The robust projection direction $\hat{\omega}$ can be utilized to perform dimensionality reduction on the original data $Y=\left(Y_1,Y_2,\cdots,Y_n\right)$.

Theorem 2.1. Let $Y_1,Y_2,\cdots,Y_n$ be iid random variable following a Gaussian distribution with mean μ and covariance matrix Σ, Let $\hat{\omega}$ be a linear projection direction. Then, $y_t=\hat{\omega}^{\mathrm{T}}Y_t,t=1,\cdots,n$ follows a normal distribution with mean $\hat{\omega}^{\mathrm{T}}\mu$ and variance $\hat{\omega}^{\mathrm{T}}\Sigma\hat{\omega}$.

The presence of outliers does not affect the distribution of the data; the projected data remains normally properties.

Theorem 2.2 states that the one-dimensional data obtained after linear projection satisfies the conditions given by Fearnhead and Rigail (2019).

Theorem 2.2. After linear dimensionality reduction $y_t = \hat{\omega}^{\mathrm{T}} Y_t, t = 1, \cdots, n$, vector $y = (y_1, y_2, \cdots, y_n)^{\mathrm{T}}$ satisfies the following conditions:

1) Exist a fixed number of changepoints k, and fixed constants $0 < \tau_1 < \cdots < \tau_k < 1$ so that for a dataset of size n, the ith changepoint at $\tau_i^* = \lfloor n\tau_i \rfloor$, for $i = 1, \cdots, k$, let $\tau_0^* = 0$ and $\tau_{k+1}^* = n$.

2) Exist a fixed segment-specific location parameters $\vec{\mu}_0, \cdots, \vec{\mu}_k$, with the obvious constraint that $\vec{\mu}_i \neq \vec{\mu}_{i-1}$ for $i = 1, \cdots, k$.

Let $Z_1, Z_2, \cdots, Z_n$ be iid noise random variables, so that for $t = 1, \cdots, n$ the observations are realizations of

$$y = \vec{\mu}_i + Z_t$$

where i is such that $\tau_i^* < t \leq \tau_{i+1}^*$.

2.2.2. Changepoint Detection

For the univariate data $y_{1:t}(t = 1, \cdots, n)$ after dimensionality reduction, this paper considers using the RFPOP method proposed by Fearnhead and Rigaill (2019) for robust changepoint detection. This method detects changepoints by minimizing a penalty cost function, thereby reduce the impact of outliers on the changepoint detection. With the Biweight loss function, RFPOP method can lead to the consistent estimation of the number of changepoints and accurate estimation of their location under weak conditions on the noise distribution. Fearnhead and Rigaill (2019) employ robust σ estimation of the median absolute deviation based on time series differences (Fryzlewicz, 2014) to handle univariate data with outliers, rendering the inference results more robust.

For the univariate data $y_{1:t}(t = 1, \cdots, n)$, Fearnhead and Rigail (2019) define the cost function associated with the data segment $(s = 1, \cdots, n; t = s, \cdots, n)$ as

$$C(y_{s:t}) = \min_{\theta} \sum_{i=s}^{t} \gamma(y_i; \theta), \tag{4}$$

in $\gamma(y_i; \theta) = \begin{cases} (y-\theta)^2 & |y-\theta| < L \\ L^2 & \text{other} \end{cases}$, L is a constant.

Fearnhead and Rigail (2019) introduce the minimum penalized cost of segmentin $y_{1:t}$ conditional on the most recent segment having parameter θ,

$$Q_t(\theta) = \min_{\tau \in J_t} \left\{ \sum_{i=0}^{k-1} \left[C(y_{\tau_i+1:\tau_{i+1}}) + \beta \right] + \sum_{j=\tau_k+1}^{t} \gamma(y_j; \theta) + \beta \right\}. \tag{5}$$

Fearnhead and Rigail (2019) considered recursively calculate $Q_t(\theta)$ for increasing values of t, thus, $Q_t = \min_{\theta} Q_t(\theta)$, $Q_1(\theta) = \gamma(y_1; \theta) + \beta$, thereby obtaining a set of candidate changepoints $\hat{J}_k = \{\hat{\tau} = \hat{\tau}_{1:k} : 0 < \hat{\tau}_1 < \cdots < \hat{\tau}_k < t\}$.

Which is Theorem 3 from the paper by Fearnhead and Rigail (2019), theorem 2.3 provides the asymptotic properties for estimating the number and positions of changepoints.

Theorem 2.3. For a given n, let $\hat{k}$ be the estimate of the number of changepoints, and $\hat{\tau}_1, \cdots, \hat{\tau}_{\hat{k}}$ their estimated locations, obtained by minimizing the pe-

nalized cost using the biweight loss function and a penalty β_n. Then, there exists constants $c_1 > 0$ and $c_2 > 0$ such that suppose conditions 1 and 2 hold

Conditions 1: $M(\theta) = E\left[\min\left\{(Z_i - \theta)^2, L^2\right\}\right] \geq M(0) + \min\left\{c_1\theta^2, c_2\right\}$, the mean of the loss function is $M(\theta) = E\left\{\gamma(Z_i;\theta)\right\}$ and will hold if $M(\theta)$ has a positive second derivative for all θ in a neighborhood around 0 and that $M(\theta) - M(0) \geq c_2 > 0$ for all θ outside this region.

Conditions 2: Let $q = \Pr(|Z_i| > L)$ and $\sigma^2 = E\left(Z_i^2 \middle| |Z_i| \leq L\right)$, then we need

$$L^2(1-2q) - (1-q)\sigma^2 > 0$$

such that

$$\Pr\left[\hat{k} = k \text{ and } \max_{i=1,\cdots,k}\left\{\min_{j=1,\cdots,\hat{k}}\left|\tau_i - \hat{\tau}_j\right|\right\} \leq C_2 \log(n)\right] \to 0, \text{ as } n \to \infty$$

provided that $C_1 \log(n) < \beta_n = o(n)$.

Here, we outline the flowchart of the outlier and changepoint detection algorithm as shown in Figure 1, and the changepoint detection algorithm is detailed in Table 1.

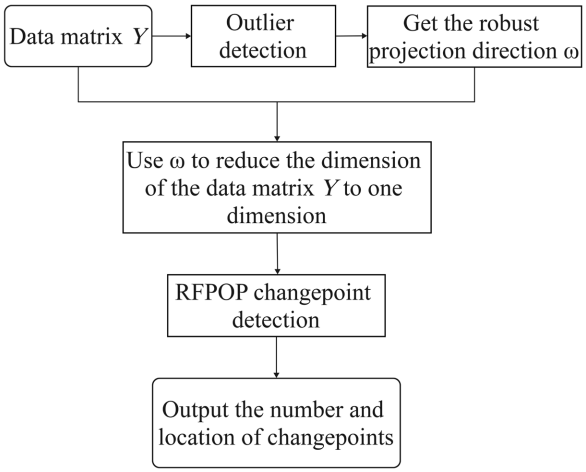

Figure 1. RWPCA-RFPOP algorithm flow chart.

Table 1. RWPCA-RFPOP changepoint detection algorithm.

Input: $Y \in \mathbb{R}^{p \times n}$.
Step 1: Outlier detection.
Step 2: Multivariate data $Y' = (Y_1', Y_2', \cdots, Y_n')$ after removing and replacing outliers.
Step 3: Use WPCA to solve for the projection direction $\hat{\omega}$ of $Y' = (Y_1', Y_2', \cdots, Y_n')$.
Step 4: Dimensionality reduction to univariate data $Y \xrightarrow{\hat{\omega}} y_{1:t}$.
Step 5: Perform Univariate RFPOP method on $y_{1:t}$ to recover cpts $\hat{\tau}_k$.
output: $\hat{K}, \hat{J}_k = \left\{\hat{\tau} = \hat{\tau}_{1:k} : 0 < \hat{\tau}_1 < \cdots < \hat{\tau}_k < t\right\}$.

3. Numerical Simulation

This section validates the robustness of the RWPCA-RFPOP method against out-

liers through simulation studies. It also compares the algorithm with other existing methods, including the Inspect method proposed by Wang and Samworth (2018), the GeomCP method by Grundy *et al.* (2020), and the E-divisive method by Matteson *et al.* (2014). Our RWPCA-RFPOP method estimates the standard deviation using the median absolute deviation of the differenced time series. With regard to the Biweight loss function, an appropriate value of L should be chosen. A reasonable default is 2 to 3 times the estimated standard deviation of the noise. This choice ensures that most observations fall within the segment-specific parameter L, thereby enhancing robustness against extreme outliers. Typically, the absolute median deviation of differenced time series is used as an estimate $\hat{\sigma}$ of the noise standard deviation. In practical applications using biweight loss, adjustments are often made based on extreme outlier behavior under a Gaussian model and BIC penalties, we choose $L = 3\hat{\sigma}$, $\beta_1 = 2\hat{\sigma}^2 \log(n)$. We implemented the Inspect method using the InspectChangepoint package (Wang and Samworth). For the global spatial dependence parameter in the Inspect method, we choose $\lambda = 2\sqrt{2\log(p\log n)}$. For the GeomCP and E-divisive methods, we also used the default settings.

3.1. Simulation Settings

First, we are given the sparsity of a p-dimensional data Y, where the sparsity is defined by $sp = \max\{|J_k|\}/p, j = 1,\cdots,k$, where $|J_k|$ is the cardinality of the set J_k, $sp \in [0,1]$. We ran simulations with different sample sizes $n \in \{500,1000,1500\}$, dimensions $p \in \{200,400,600\}$, and sparsity $sp \in \{0.4,0.6,0.8\}$, the number of changepoints $K \in \{2,3,5\}$, the variance of the noise $\sigma^2 = 1$, and the jump magnitude $\vartheta^{(i)} = \|\theta^{(i)}\|_2$ at the i-th changepoint, taking $\vartheta^{(i)} = \vartheta_1$, setting $\vartheta_1 \in \{2.2,2.4,2.6\}$ to observe the performance of the algorithm. All parameter settings are shown in **Table 2**.

Table 2. Parameter setting list.

Parameter	Related options
Sample size	$n \in \{500,1000,1500\}$
Dimension	$p \in \{200,400,600\}$
Sparsity	$sp \in \{0.4,0.6,0.8\}$
Number of changepoints	$K \in \{2,3,5\}$
Noise variance	$\sigma^2 = 1$
Jump magnitude	$\vartheta_1 \in \{2.2,2.4,2.6\}$
Number of outliers	$s = 6$
Outlier location	$m \in \{50,120,175,360,450,800\}$

Let $W = \Sigma^{1/2}\varepsilon_i$, for some positive definite matrix $\Sigma \in \mathbb{R}^{p\times p}$, assume that the noise vectors satisfy $W_1,\cdots,W_n \overset{IID}{\sim} N_p(0,\Sigma)$. Consider the problem of performing a changepoint detection in the case of global dependence, suppose that

$\Sigma = \mathrm{I}_p + \frac{\rho}{p} 1_p 1_p^{\mathrm{T}}$ for some $-1 \le \rho \le p$. This paper generates a multivariate normal distribution with s outliers, where the size of the outliers is set to be outside 3 standard deviations from the mean.

To evaluate the performance of the changepoint estimation, we performed 100 simulations under each scenario and provided a frequency distribution of $\hat{K} - K$. The Rand Index (ARI) was used to assess the consistency between the estimated changepoint locations and the true changepoint locations (Hubert and Arabie 1985), and the scaled *Hausdorff* distance was used for further evaluation.

3.2. Simulation Results and Analysis

The experimental data was generated according to the settings in Section 3.1, and 100 repeated simulations were performed under 11different parameter settings $M_1 \sim M_{11}$ to obtain the average results. The numerical simulation experiments were all implemented using the R language. Table 3 displays the parameter settings for the simulation experiments:

Table 3. Parameter settings in simulation experiments.

Setting	n	p	sp	$\vartheta = (\vartheta_1, \vartheta_1, \vartheta_1)$	Correlation
M_1	1000	600	0.6	$\vartheta_1 = 2.4$	0.5 - 0.7
M_2	500	600	0.6	$\vartheta_1 = 2.4$	0.5 - 0.7
M_3	1500	600	0.6	$\vartheta_1 = 2.4$	0.5 - 0.7
M_4	1000	200	0.6	$\vartheta_1 = 2.4$	0.5 - 0.7
M_5	1000	400	0.6	$\vartheta_1 = 2.4$	0.5 - 0.7
M_6	1000	600	0.4	$\vartheta_1 = 2.4$	0.5 - 0.7
M_7	1000	600	0.8	$\vartheta_1 = 2.4$	0.5 - 0.7
M_8	1000	600	0.6	$\vartheta_1 = 2.2$	0.5 - 0.7
M_9	1000	600	0.6	$\vartheta_1 = 2.6$	0.5 - 0.7
M_{10}	1000	600	0.6	$\vartheta_1 = 2.4$	0.3 - 0.5
M_{11}	1000	600	0.6	$\vartheta_1 = 2.4$	0.7 - 0.9

Compare M_1, M_2 and M_3, to examine the impact of sample size on algorithm performance, compare M_1, M_4 and M_5, to examine the impact of dimension on algorithm performance, compare M_1, M_6 and M_7, to examine the impact of sparsity on algorithm performance, compare M_1, M_8 and M_9, to examine the impact of the jump magnitude at the changepoint on algorithm performance, compare M_1, M_{10} and M_{11}, to examine the impact of data correlation on algorithm performance.

According to the results in Table 4, the RWPCA-RFPOP method performs well in the given settings. For multivariate normal distributions with outliers, our RWPCA-RFPOP method is comparable to the E-divisive method in accurately detecting the number and locations of changepoints. The RWPCA-RFPOP

method performs well in scenarios with moderate sparsity as well as dense cases. The Inspect and GeomCP methods perform poorly, often overestimating the number of changepoints and misidentifying outliers as changepoints when outliers are present. The RWPCA-RFPOP method shows strong performance in accurately estimating the number and positions of changepoints. The error in the number of estimated changepoints relative to the actual number is within 5%, and the method achieves high Rand Index (ARI) values and low scaled *Hausdorff* distances. As the jump magnitude increases, the changepoint detection capability of the RWPCA-RFPOP method improves. Furthermore, as the sample size and dimensionality increase, the performance of the RWPCA-RFPOP method also improves significantly. We also conducted simulations for multivariate normal distributions without outliers, where the RWPCA-RFPOP method also performed excellently. However, this paper mainly presents scenarios with outliers in multivariate normal distributions.

Table 4. Data simulation results in a multivariate normal distribution scenario with outliers.

Setting	Method	Frequency of $\hat{K}-K$						*ARI*	d_H	Time (s)
		−1	0	1	2	3	4			
M_1	RWPCA-RFPOP	0	**99**	1	0	0	0	**0.9877**	**0.0131**	540.64
	Inspect	0	0	0	0	0	100	0.8503	0.5241	1612.98
	GeomCP	0	0	0	0	0	100	0.8176	0.5629	225.98
	E-divisive	0	94	5	1	0	0	0.9871	0.0248	1597.62
M_2	RWPCA-RFPOP	0	**96**	4	0	0	0	0.9781	**0.0272**	671.07
	Inspect	0	0	0	0	0	100	0.7233	0.6812	859.75
	GeomCP	0	0	0	0	7	93	0.6495	0.8146	209.19
	E-divisive	0	95	5	0	0	0	**0.9809**	0.0302	571.62
M_3	RWPCA-RFPOP	0	**97**	3	0	0	0	**0.9914**	**0.0178**	707.75
	Inspect	0	0	0	0	0	100	0.8482	0.9522	2233.88
	GeomCP	0	0	0	0	0	100	0.8250	0.9736	241.69
	E-divisive	0	96	3	1	0	0	0.9902	0.0186	5292.05
M_4	RWPCA-RFPOP	0	**97**	2	1	0	0	0.9872	**0.0217**	75.89
	Inspect	0	0	0	0	0	100	0.8499	0.5239	429.89
	GeomCP	0	0	0	0	2	98	0.8179	0.5572	18.94
	E-divisive	0	**97**	3	0	0	0	**0.9896**	0.0230	1335.66
M_5	RWPCA-RFPOP	0	**96**	4	0	0	0	0.9882	0.0224	294.45
	Inspect	0	0	0	0	0	100	0.8501	0.5239	912.56
	GeomCP	0	0	0	0	2	98	0.8164	0.5562	82.40
	E-divisive	0	**96**	3	1	0	0	**0.9885**	**0.0221**	1409.37

Continued

M_6	RWPCA-RFPOP	0	**99**	0	1	0	0	0.9714	0.0293	593.64
	Inspect	0	0	0	0	0	100	0.8500	0.5239	1490.17
	GeomCP	0	0	0	0	51	49	0.7913	0.5316	223.88
	E-divisive	0	97	2	1	0	0	**0.9890**	**0.0198**	1584.28
M_7	RWPCA-RFPOP	0	**100**	0	0	0	0	**0.9941**	**0.0064**	579.17
	Inspect	0	0	0	0	0	100	0.8505	0.5242	1540.21
	GeomCP	0	0	0	0	0	100	0.8215	0.5648	232.54
	E-divisive	0	94	6	0	0	0	0.9879	0.0239	1584.19
M_8	RWPCA-RFPOP	0	**99**	1	0	0	0	0.9863	0.0146	530.31
	Inspect	0	0	0	0	0	100	0.8505	0.5242	1592.94
	GeomCP	0	0	0	0	0	100	0.8195	0.5639	240.56
	E-divisive	0	98	2	0	0	0	**0.9901**	**0.0133**	1590.20
M_9	RWPCA-RFPOP	0	**99**	1	0	0	0	**0.9898**	**0.0112**	523.14
	Inspect	0	0	0	0	0	100	0.8505	0.5242	1616.61
	GeomCP	0	0	0	0	0	100	0.8207	0.5644	230.11
	E-divisive	0	97	3	0	0	0	0.9897	0.0142	1583.39
M_{10}	RWPCA-RFPOP	0	**100**	0	0	0	0	**0.9929**	**0.0078**	545.00
	Inspect	0	0	0	0	0	100	0.8506	0.5242	2545.47
	GeomCP	0	0	0	0	0	100	0.8211	0.5652	240.70
	E-divisive	0	93	7	0	0	0	0.9871	0.0321	1592.63
M_{11}	RWPCA-RFPOP	0	**99**	1	0	0	0	0.9842	**0.0168**	528.68
	Inspect	0	0	0	0	0	100	0.8501	0.5239	1589.39
	GeomCP	0	0	0	0	7	93	0.8128	0.5587	231.80
	E-divisive	0	95	3	2	0	0	**0.9878**	0.0257	1598.02

The following is an analysis of the accuracy and robustness of our proposed method, along with a comparison to existing methods. $\hat{K}-K$ represents the difference between the estimated number of changepoints and the true number of changepoints. Figure 2 depicts the results of the RWPCA-RFPOP method, the Inspect method, the GeomCP method, and the E-divisive method under $\hat{K}-K$ for multivariate normal distribution data with outliers. The settings are as follows: $n=1000$, $sp=0.6$, $\vartheta_1=2.2$, and the simulations include 2, 3, and 5 changepoints. Each box plot represents a different combination of simulation parameters. For instance, the first box plot illustrates the empirical distribution of a simulated 200-dimensional multivariate normal distribution data with outliers and 2 changepoints. The empirical distribution is calculated based on 100 repeated simulations conducted under each combination of simulation parameters.

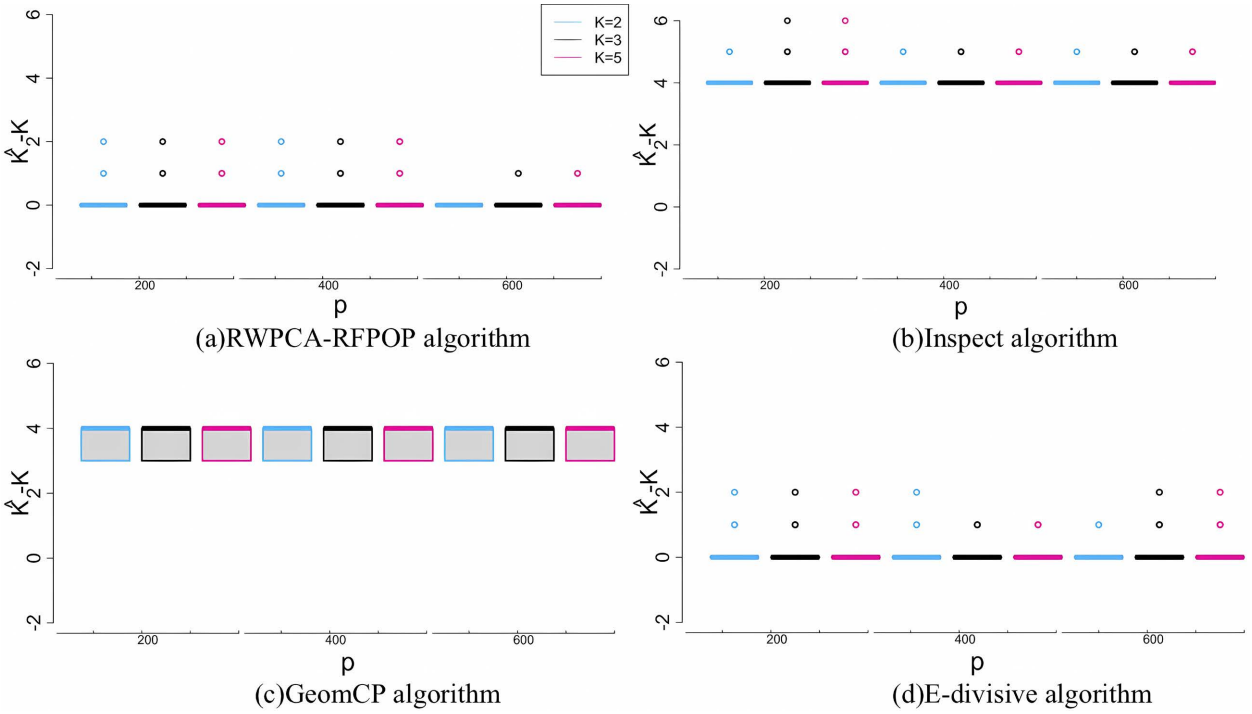

Figure 2. $\hat{K}-K$ (Empirical) distribution of multivariate normal distribution with outliers by different methods.

Figure 2 shows the results of the simulation scenarios where the data come from a multivariate normal distribution with outliers. Figure 2 shows that the RWPCA-RFPOP method and the E-divisive method perform exceptionally well. As seen in Figure 2, the performance of the RWPCA-RFPOP method significantly improves with an increase in dimensionality. When the data come from a multivariate normal distribution with outliers, the Inspect and GeomCP methods tend to overestimate the number of changepoints. This phenomenon is primarily due to the lack of robustness in these two methods, which are unable to effectively handle outlier data.

Figure 3 presents the boxplots for each changepoint when the data come from a multivariate normal distribution with outliers. Each boxplot in the above figure represents the ability to estimate the location of specific changepoints, without considering the method mistakenly identifying outliers as changepoints. In this context, the RWPCA-RFPOP, Inspect, GeomCP, and E-divisive methods all estimate the changepoint locations fairly accurately. When the Inspect method and the GeomCP method correctly identify changepoints, their accuracy is not significantly different from our proposed method. However, Figure 3 shows that the Inspect and GeomCP methods are more sensitive to outliers, often overestimating the number of changepoints. The RWPCA-RFPOP method demonstrates strong robustness and accuracy in handling data with outliers.

Figure 4 shows the scaled *Hausdorff* distance and ARI value of different methods under different sparsity, setting $n \in \{500,1000,1500\}$, $p=600$, $\vartheta_1=2.4$, the simulation included three changepoints, and both the RWPCA-RFPOP and E-divisive methods demonstrated advantages in terms of ARI values and scaled

Hausdorff distances. As the sample size increased, the performance of the RWPCA-RFPOP method significantly improved. The Inspect and GeomCP methods were more sensitive to outliers, leading to less satisfactory results.

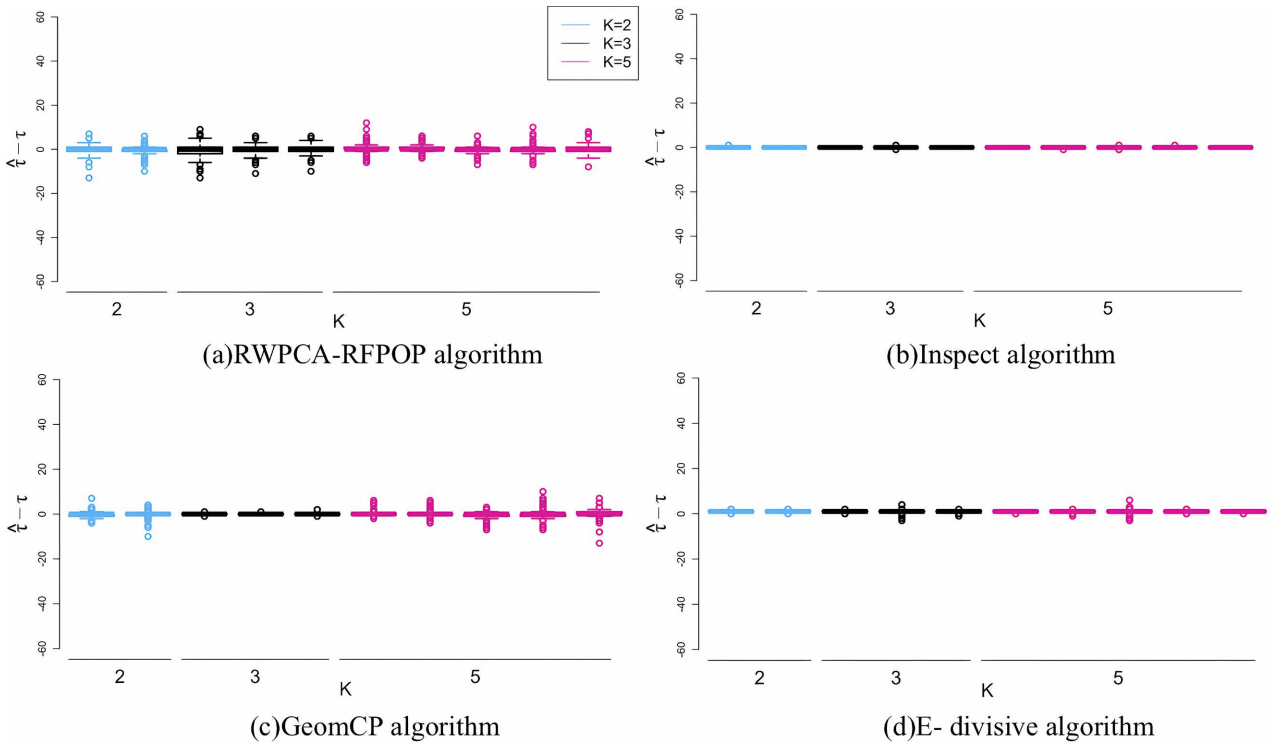

Figure 3. Estimation ability of different methods for $\hat{\tau}-\tau$ in a multivariate normal distribution with outliers.

One of the main issues with multivariate changepoint detection is that as the number of sample size n and dimension p increase, many multivariate changepoint methods become computationally infeasible. We compare the running time of the RWPCA-RFPOP, Inspect, GeomCP, and E-divisive methods, with the simulation settings including $n \in \{500,1000,1500\}$, $p \in \{200,400,600\}$, and 3 changepoints. We will compare the running time under two different scenarios.

We can see from the left graph of **Figure 5** that GeomCP is the fastest among the four methods, followed by the RWPCA-RFPOP method. We observe that E-Divisive and Inspect are much slower than GeomCP and RWPCA-RFPOP, with their running times increasing rapidly as n increases. In the right panel of **Figure 5**, the Inspect method performs well for small p, but its running time slows down when p increases beyond 500. The running time of E-Divisive does not seem to be affected by p, remaining relatively slow. This is likely because its computational cost is primarily influenced by the sample size, showing a slow increase. As p increases, both RWPCA-RFPOP and GeomCP exhibit linear running times and faster running times. Although the E-divisive method proposed by Matteson *et al.* (2014) performs well in terms of accuracy, its running time is relatively slow. Considering all factors, we recommend using the RWPCA-RFPOP method for changepoint detection in multivariate data with outliers.

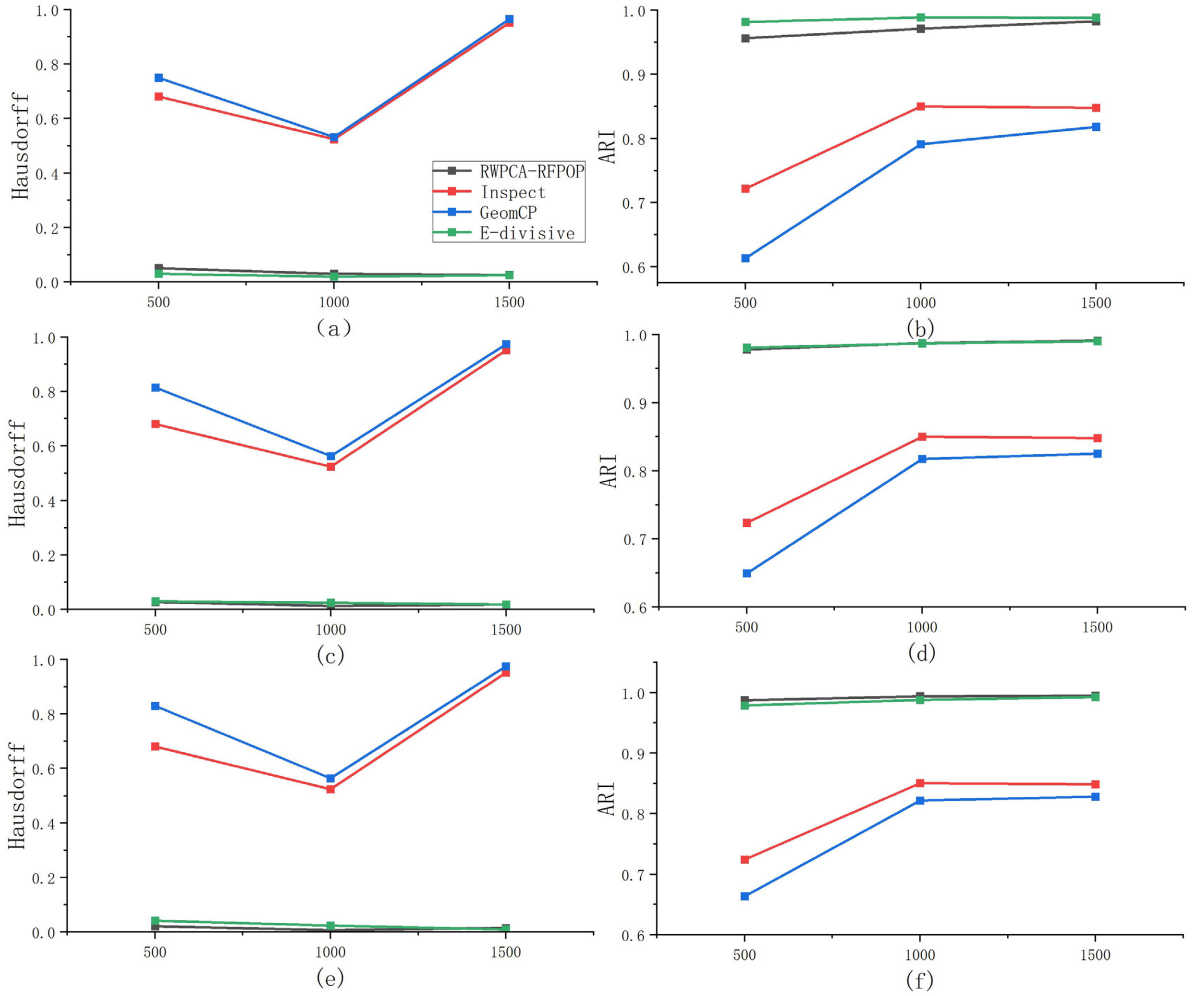

Figure 4. Scaled *Hausdorff* distances of different methods with different sparsity $sp = 0.4$ (a), $sp = 0.6$ (c), and $sp = 0.8$ (e), and ARI values of different methods as n changes with different sparsity $sp = 0.4$ (b), $sp = 0.6$ (d), and $sp = 0.8$ (f).

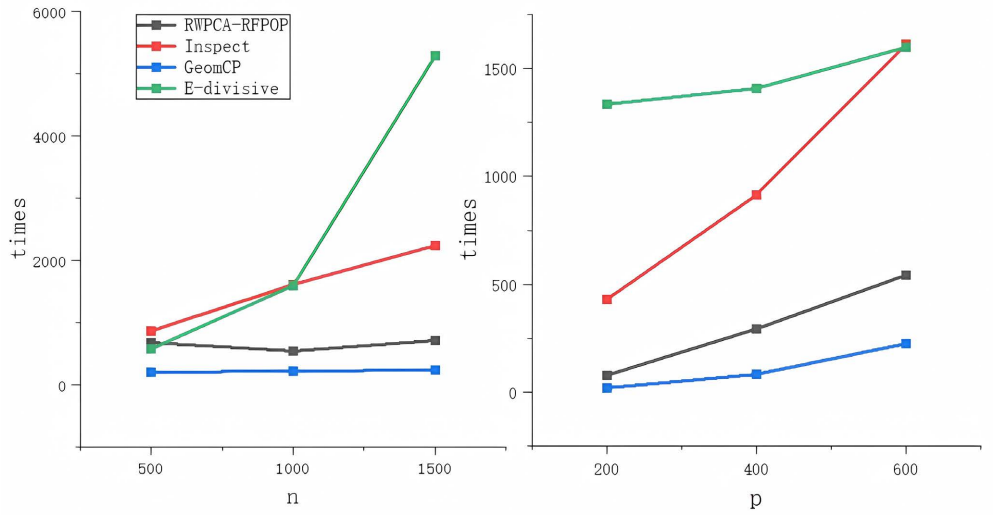

Figure 5. The running time of different methods as n changes (left) and the calculation speed as p changes (right).

In this simulation setting, we found that our proposed RWPCA-RFPOP method performs exceptionally well in handling outliers. The advantage of this method lies in its accurate estimation capability for changepoint locations, especially in the presence of outliers. The results in Figure 3 show that the RWPCA-RFPOP method can estimate the changepoint locations relatively accurately. Considering its performance and efficiency, we suggest using the RWPCA-RFPOP method for changepoint detection.

3.3. Outliers in Data

In the generated multivariate data, we set the sample size $n=1000$, dimension $p=600$, sparsity $sp=0.6$, and generated three types of outliers: small outliers (5 - 10 standard deviations away from the true mean), large outliers (25 - 30 standard deviations away), and very large outliers (40 or more standard deviations away). For each type of outlier setting, we randomly generated 5 - 10 outlier instances located at $m \in \{50,120,175,360,390,450,550,670,800,850\}$, with the dimensions of outlier variation randomly chosen. We specified the number of changepoints $K=3$, their positions $\tau_k \in \{250,500,750\}$, noise variance $\sigma^2=1$, and set the jump size at the i-th changepoint to $\vartheta_1=2.2$ to observe algorithm performance. For each outlier setting, we ran the RWPCA-RFPOP algorithm and reported the results of changepoint detection accuracy.

Table 5 reports the frequency distribution of $\hat{K}-K$, Adjusted *Rand Index* (ARI), and scaled *Hausdorff* distance for changepoint detection using the RWPCA-RFPOP method under different outlier scenarios. The results show that the RWPCA-RFPOP method exhibits high ARI and low scaled *Hausdorff* distance for all three types of outliers. Specifically, the RWPCA-RFPOP method consistently achieves an ARI of at least 0.97, indicating its accuracy in identifying true changepoints. Considering that multivariate data can contain a large number of outliers, which poses a significant challenge for other changepoint detection methods, the RWPCA-RFPOP algorithm's ability to handle noisy and outlier-affected changepoint detection makes it highly versatile and broadly applicable.

Table 5. RWPCA-RFPOP method changepoint detection results under various outlier conditions.

Types of outliers	s	Frequency of $\hat{K}-K$						*ARI*	d_H
		−2	−1	0	1	2	3		
Small outliers	5	0	0	96	4	0	0	0.9755	0.0330
Large outliers	5	0	0	95	5	0	0	0.9760	0.0448
Very large outliers	5	0	0	96	2	2	0	0.9740	0.0457
Small outliers	10	0	0	98	2	0	0	0.9764	0.0289
Large outliers	10	0	0	95	4	0	1	0.9791	0.0357
Very large outliers	10	0	0	98	2	0	0	0.9781	0.0334

4. Real Data Application

In this section, we apply RWPCA-RFPOP method to an Array Comparative Genomic Hybridization (ACGH) dataset, which is available in the ecp R package (James and Matteson, 2015), and perform a comparative analysis with results from the literature. Comparative genomic hybridization is a technique that allows detection of chromosomal copy number abnormality by comparing the fluorescence intensity levels of DNA fragments from a test sample and a reference sample. Chromosome copy number variations reflect DNA amplifications and deletions. From a medical perspective, amplified segments may contain oncogenes, while deleted segments may contain tumor suppressor genes. Whereas some of the copy number variations are specific to one individual, some copy number abnormality regions (e.g. between loci 2044 and 2143) are shared across several individuals and are more likely to be disease related. This dataset contains (test-to-reference) log-intensity-ratio measurements of 43 individuals with bladder tumours at 2215 different loci on their genome. The log-intensity-ratios for the first 10 individuals are plotted in Figure 6. We used the RWPCA-RFPOP method to estimate the positions of changepoints (red lines indicate detected changepoint locations, and blue lines indicate detected outlier positions).

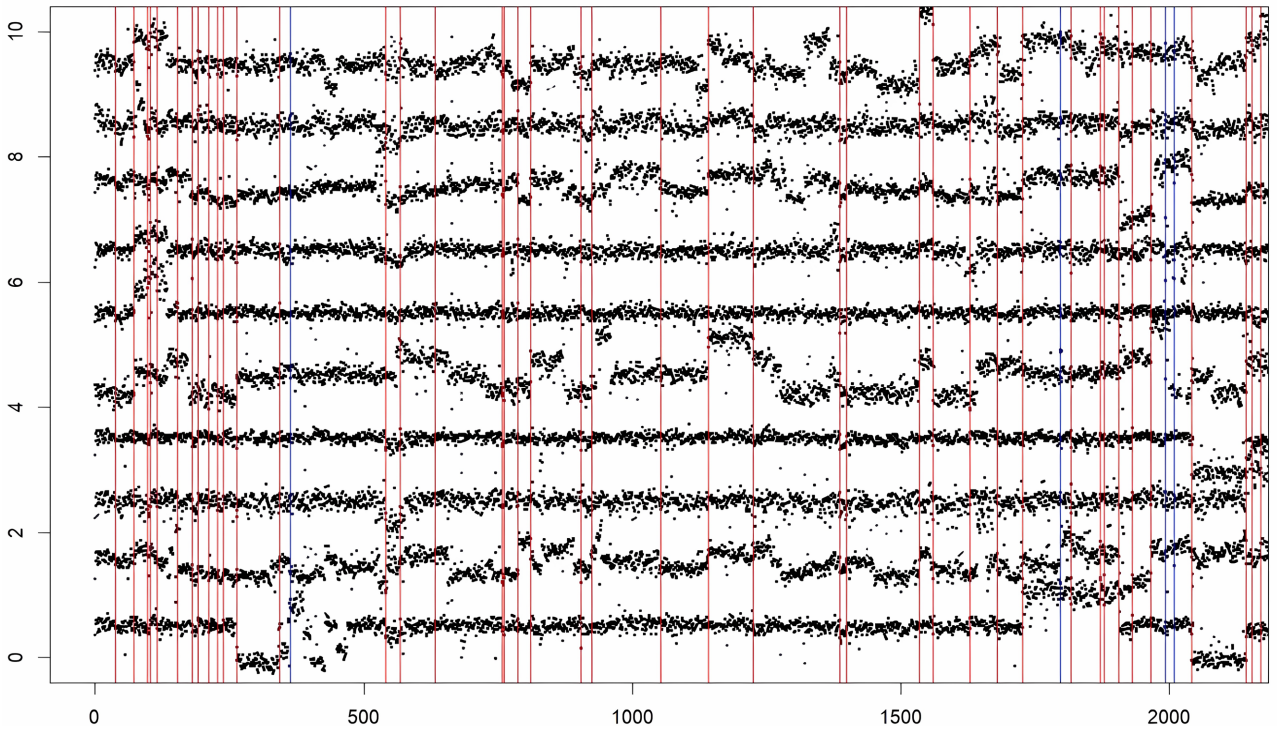

Figure 6. Changepoints of log intensity ratio in the entire ACGH data estimated by RWPCA-RFPOP (first 10 patients are shown).

The RWPCA-RFPOP method estimates the starting and ending points of copy number variations by aggregating changes across different individuals. Figure 7 displays the correlation among the 43 bladder tumor individuals. Given the presence of a substantial number of individual-specific copy number variations

and measurement outliers, we choose $L = 3\hat{\sigma}$, and the penalty threshold is accordingly set to $\beta_1 = 2\hat{\sigma}^2 \log(n)$, the RWPCA-RFPOP method directly applied the default threshold level to identify 49 changepoints, which are indicated by red vertical solid lines in Figure 8. Inspection revealed 140 changepoints, potentially attributed to the threshold setting. GeomCP estimated 27 changepoints. We compared the changepoint detection results of these four methods for the gene loci ranging from the 1700th to the 2100th in the bladder tumor microarray dataset. The results are presented in Table 6.

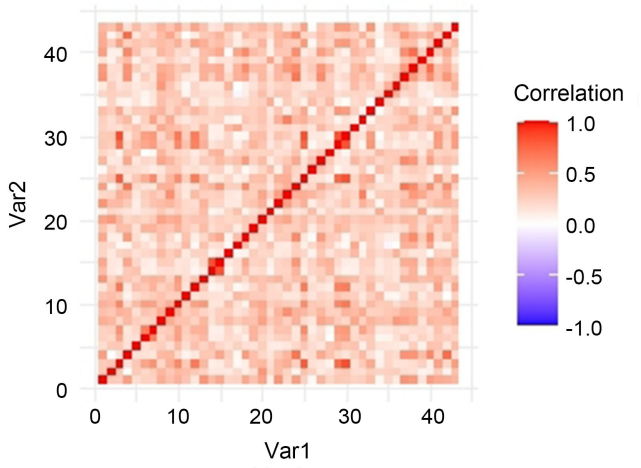

Figure 7. Correlation of 43 bladder tumor individuals.

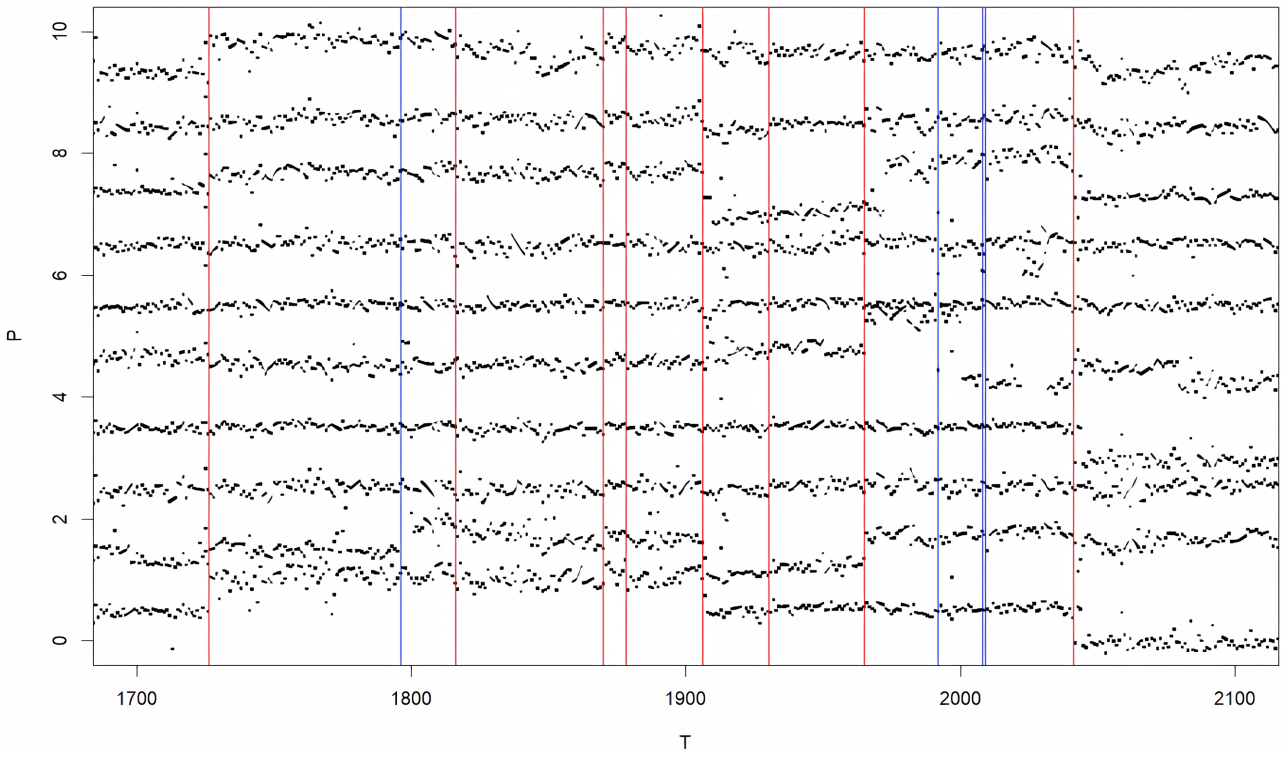

Figure 8. Changepoints of the logarithmic intensity ratios of the 1700th to 2100th gene loci in the ACGH dataset estimated by RWPCA-RFPOP.

Table 6. Changepoint detection results of four methods on the 1700th to 2100th gene loci in the bladder tumor microarray dataset.

Method	Result
RWPCA-RFPOP	1726 1816 1870 1878 1906 1930 1965 2041
Inspect	1724 1725 1748 1795 1799 1831 1850 1868 1906 1957 1965 1982 1991 1992 1996 1997 2005 2007 2009 2010 2022 2027 2031 2036 2041 2044 2072 2084 2086
GeomCP	1722 1906 1957 1991 2010 2041
E-divisive	1727 1757 1796 1832 1871 1907 1966 2001 2045 2085

The results indicate that the RWPCA-RFPOP method detected 8 changepoints in the data from loci 1700 to 2100: 1726, 1816, 1870, 1878, 1906, 1930, 1965, and 2041. These are comparable to the findings reported in the literature. However, some changepoints, such as those in the data from loci 1982 to 2010, were not detected due to the absence of apparent segmentation characteristics. Compared to the GeomCP method, the RWPCA-RFPOP method identified a slightly higher number of changepoints, Four changepoints are consistent or similar to the results from GeomCP. However, the RWPCA-RFPOP method failed to detect changepoints at loci 1991 and 2010, likely because the individual data in those regions were more dispersed, leading to a lack of distinct segmentation patterns in the projected data. In the presence of outliers, such as segments (1724, 1836) or (1965, 2044), the RWPCA-RFPOP method is capable of identifying the locations of these outliers, as depicted in **Figure 8**. This makes our method more reasonable and stable.

5. Summary

In this paper, we propose a novel RWPCA-RFPOP method for detecting changepoints in multivariate data. Our objective is to detect mean changes in multivariate data, particularly in the presence of outliers, noise, and high correlations between data variables. Simulation results demonstrate that the RWPCA-RFPOP method exhibits good robustness and outperforms state-of-the-art methods in detecting and identifying the number and location of multiple changepoints. RWPCA-RFPOP method we proposed applied in an ACGH dataset with outliers. The results show that our method performs well in detecting the number and locations of changepoints. Furthermore, we can investigate the detection of smaller outliers that may be missed by the current method.

Acknowledgements

The support from the Jilin Provincial Department of Education Project (JJKH20210809KJ), and the funding from the National Natural Science Foundation of China (11601039).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Page, E.S. (1954) A Test for a Change in a Parameter Occurring at an Unknown Point. *Biometrika*, **42**, 523-527. https://doi.org/10.1093/biomet/42.3-4.523

[2] Scott, A.J. and Knott, M. (1974) A Cluster Analysis Method for Grouping Means in the Analysis of Variance. *Biometrics*, **30**, 507-512. https://doi.org/10.2307/2529204

[3] Fryzlewicz, P. (2014) Wild Binary Segmentation for Multiple Change-Point Detection. *Annals of Statistics*, **42**, 2243-2281. https://doi.org/10.1214/14-AOS1245

[4] Fearnhead, P. and Rigaill, G. (2019) Changepoint Detection in the Presence of Outliers. *Journal of the American Statistical Association*, **114**, 169-183. https://doi.org/10.1080/01621459.2017.1385466

[5] Dehling, H., Fried, R. and Wendler, M. (2020) A Robust Method for Shift Detection in Time Series. *Biometrika*, **107**, 647-660. https://doi.org/10.1093/biomet/asaa004

[6] Anastasiou, A. and Fryzlewicz, P. (2022) Detecting Multiple Generalized Change-Points by Isolating Single Ones. *Metrika*, **85**, 141-174. https://doi.org/10.1007/s00184-021-00821-6

[7] Kovács, S., Li, H., Bühlmann, P. and Munk, A. (2023) Seeded Binary Segmentation: A General Methodology for Fast and Optimal Change Point Detection. *Biometrika*, **110**, 249-256. https://doi.org/10.1093/biomet/asac052

[8] Fryzlewicz, P. (2024) Robust Narrowest Significance Pursuit: Inference for Multiple Change-Points in the Median. *Journal of Business & Economic Statistics*. https://doi.org/10.1080/07350015.2024.2316103

[9] Horvath, L., Kokoszka, P. and Steinebach, J. (1999) Testing for Changes in Multivariate Dependent Observations with an Application to Temperature Changes. *Journal of Multivariate Analysis*, **68**, 96-199. https://doi.org/10.1006/jmva.1998.1780

[10] Matteson, D. and James, N.A. (2014) A Nonparametric Approach for Multiple Change Point Analysis of Multivariate Data. *Journal of the American Statistical Association*, **109**, 334-345. https://doi.org/10.1080/01621459.2013.849605

[11] Jirak, M. (2015) Uniform Change Point Tests in High Dimension. *Annals of Statistics*, **43**, 2451-2483. https://doi.org/10.1214/15-AOS1347

[12] Cho, H. and Fryzlewicz, P. (2015) Multiple-Change-Point Detection for High Dimensional Time Series via Sparsified Binary Segmentation. *Journal of the Royal Statistical Society*: *Series B*, **77**, 475-507. https://doi.org/10.1111/rssb.12079

[13] Knoblauch, J., Jewson, J.E. and Damoulas, T. (2018) Doubly Robust Bayesian Inference for Non-Stationary Streaming Data with β-Divergences. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Montréal, 3-8 December 2018, 64-75.

[14] Wang, T. and Samworth, R.J. (2020) High Dimensional Change Point Estimation via Sparse Projection. *Journal of the Royal Statistical Society*: *Series B*, **80**, 57-83. https://doi.org/10.1111/rssb.12243

[15] Grundy, T., Killick, R. and Mihaylov, G. (2020) High-Dimensional Changepoint Detection via a Geometrically Inspired Mapping. *Statistics and Computing*, **30**, 1155-1166. https://doi.org/10.1007/s11222-020-09940-y

[16] Wendelberger, L., Gary, J., Reich, B.J. and Wilson, A. (2021) Monitoring Deforestation Using Multivariate Bayesian Online Changepoint Detection with Outliers.