



Sequential Analysis

Design Methods and Applications

ISSN: (Print) (Online) Journal homepage: <https://www.tandfonline.com/loi/lsqa20>

Real-time change point detection in linear models using the ranking selection procedure

Chao Gu & Suthakaran Ratnasingam

To cite this article: Chao Gu & Suthakaran Ratnasingam (2023): Real-time change point detection in linear models using the ranking selection procedure, Sequential Analysis, DOI: [10.1080/07474946.2023.2187416](https://doi.org/10.1080/07474946.2023.2187416)

To link to this article: <https://doi.org/10.1080/07474946.2023.2187416>



Published online: 23 May 2023.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)



Real-time change point detection in linear models using the ranking selection procedure

Chao Gu^a and Suthakaran Ratnasingam^b 

^aDepartment of Mathematical Sciences, The Citadel–The Military College of South Carolina, Charleston, South Carolina, USA; ^bDepartment of Mathematics, California State University, San Bernardino, San Bernardino, California, USA

ABSTRACT

We propose a novel sequential change point detection method in linear models. Our method uses a given historical data set to determine the prechange model. Significant features are selected using the ranking procedure, which is an innovative approach aimed at revealing the rank of all features in terms of their effects on the model. We establish the asymptotic properties of the test statistic under the null and alternative hypotheses. Simulations are conducted to illustrate the performance of the proposed method. We conclude with a real data application to illustrate the detection procedure.

ARTICLE HISTORY

Received 4 October 2022

Revised 5 January 2023

Accepted 18 January 2023



KEYWORDS

Change point analysis;
linear models; ranking
selection; sequential
analysis

1. INTRODUCTION

Sequential change point analysis plays an important role in statistical quality control and reliability, bioinformatics, signal processing, and medical imaging. In sequential analysis, the goal is to detect the structural change as quickly as possible while controlling the false alarm rate. This decision must be taken in real time based on prior information. There is a broad range of literature discussing the sequential change point detection procedure for low-dimensional data; see, for example, Page (1954), Shiryaev (1963), Roberts (1966), Lorden (1971), Siegmund (1985), Horváth et al. (2004), Horváth, Kokoszka, and Steinebach (2007), and Tartakovsky, Nikiforov, and Basseville (2014).

The recent advances in technological development have enabled massive volumes of data to be collected in a short amount of time. For example, medical, genomics, traffic monitoring, and health care data contain an endless number of observations and have a large number of explanatory variables for each observation. Horváth et al. (2004) proposed the sequential change point detection method in linear regression. Unfortunately, their method cannot be applied to many practical problems because it only works with univariate data. The high-dimensional linear model, in which the dimension p is significantly greater than the sample size n , has garnered substantial attention recently, driven by a variety of applications. Penalization or regularization techniques are proven to be

CONTACT Suthakaran Ratnasingam  suthakaran.ratnasingam@csusb.edu  Department of Mathematics, California State University, San Bernardino, 5500 University Parkway, San Bernardino, CA 92407, USA.
Both authors contributed equally to this work as co-first authors.

© 2023 Taylor & Francis Group, LLC

effective where the number of covariates is larger than the sample size. Several penalization methods have been proposed, including the least absolute shrinkage and selection operator (LASSO; Tibshirani 1996), smoothly clipped absolute deviation (SCAD; Fan and Li 2001), elastic net (Zou and Hastie 2005), adaptive LASSO (Zou 2006), grouped LASSO (Yuan and Lin 2006), Dantzig (Candes and Tao 2007), and minimax concave penalty (MCP; Zhang 2010). In particular, regularized methods such as LASSO, adaptive LASSO, SCAD, and MCP can select significant variables and estimate the coefficient simultaneously. The penalty function remains crucial for high-dimensional data analysis.

Extensive work has been done on sequential change point analysis. Chen (2019) proposed an online detection framework that utilizes nearest-neighbor information for high-dimensional data. This method makes use of the similarity structure represented by nearest neighbors. L. Chu and Chen (2019) studied sequential change point detection for high-dimensional and non-Euclidean data that uses graph-based test statistics under k -nearest neighbors. Ratnasingam and Ning (2021b) studied the sequential change point detection method to monitor structural changes in penalized quantile regression models. Ratnasingam and Ning (2021a) proposed a method to detect the structural changes in the SCAD penalized regression model for high-dimensional data sequentially. They established the asymptotic properties of the test statistics under the null and alternative hypotheses. Following Gu (2021), we adopted the Bechhofer, Dunnett, and Sobel (1954) single-sample multiple decision procedures for low- and medium-dimensional variable selection in linear models. In other words, we are interested in selecting the best covariates that have the greatest effect on the regression model. In particular, when compared to other features, the best features contribute the most to the predictions. To the best of our knowledge, this is the first study that utilizes the ranking selection procedure in change point analysis.

The remainder of the article is organized as follows. In Section 2, we describe notations and a framework for the variable selection and estimation based on the ranking procedure. In Section 3, we briefly describe the sequential change point detection method and provide the corresponding asymptotic results. In Section 4, an extensive simulation study is conducted under different settings to investigate the finite sample performances of the proposed method. A real data application is given in Section 5 to illustrate the detecting process. In Section 6, we discuss our results and make conclusions. The proofs are deferred to the appendix.

2. METHODOLOGY

Throughout this article, we adopt notations similar to those of Ratnasingam and Ning (2021a). Suppose we have a random sample $\{Y_i, x_{i1}, \dots, x_{ip}\}, i = 1, \dots, m$. Consider the model

$$Y = X\beta + \mathcal{E}, \quad (2.1)$$

where $Y = (Y_1, \dots, Y_m)$ is a vector of responses, X is an $m \times p$ matrix of predictors with i th row $X_i^\top = (x_{i1}, \dots, x_{ip})$, where $i = 1, \dots, m$; and j th column $X_j = (x_{1j}, \dots, x_{mj})^\top$, where $j = 1, \dots, p$. In this model, X is considered fixed data. We also

assume that there is no random component, including no measurement error. The $\beta = (\beta_1, \dots, \beta_p)^\top$ is a p -vector of unknown parameters and $\mathcal{E} = (\mathcal{E}_1, \dots, \mathcal{E}_m)^\top$ represents an m -vector of independent and identically distributed (i.i.d.) random variables with mean 0 and variance σ^2 .

We consider a sparse model in which most regression coefficients are exactly zero and only certain predictors have regression coefficients that are nonzero. Without loss of generality, we assume that the first q regression coefficients are nonzero, whereas the rest of the $(p - q)$ coefficients are zero. Let $X = (X^{(1)}, X^{(2)})$, where $X^{(1)}$ is the first $m \times q$ submatrix and $X^{(2)}$ is the last $m \times (p - q)$ submatrix of X . Similarly, we denote $\beta = (\beta^{(1)}, \beta^{(2)})$. Let $C_m = \frac{1}{m} X^\top X$ and $C_m^{(u,v)} = \frac{1}{m} X^{(u)\top} X^{(v)}$, for $u, v = 1, 2$. Let $\beta_0 = (\beta_{01}, \dots, \beta_{0p})$ be the true unknown parameter vector. Let $\mathcal{A} = \{j \in \{1, \dots, p\} : \beta_{0j} \neq 0\}$ be the index set of the nonzero coefficients for the true parameter, where β_{0j} is the j th component of the true parameter vector β_0 . We denote the regression estimate based on the ranking procedure by $\hat{\beta}^w$. Let $\mathcal{A}^* = \{j \in \{1, \dots, p\} : \hat{\beta}_j^w \neq 0\}$ be the index set of the regression estimator based on the ranking procedure obtained using the historical sample size m , where $\hat{\beta}_j^w$ is the j th element of the regression estimator $\hat{\beta}^w$ that is obtained by the ranking procedure.

2.1. Rank-Based Variable Selection Method in Linear Models

Now we are in a position to discuss the rank-based variable selection method in linear models. For simplicity, let the columns of X be standardized and Y be centered (we no longer need a constant column in X). The true model is defined in (2.1). The $|\beta_j|, j = 1, \dots, p$ is defined as its effect size. The random error vector $\mathcal{E} \in R^{m \times 1}$ follows $N(0, \sigma^2)$, which is assumed known to us. It should be clear that the effect size reveals the impact of each predictor X_i on Y . We then define X_1, \dots, X_k with the largest effect sizes as the “best” k predictors, which are of interest in variable selection. In particular, the perfect case would be (X_1, \dots, X_k) exactly coincides with $X^{(1)}$, i.e. $k=q$.

2.1.1. Estimation of β

The ordinary least squares method finds the coefficients β by solving the convex problem

$$\hat{\beta}^{LS} = \underset{\beta \in R^{p \times 1}}{\text{minimize}} \quad \|Y - X\beta\|_2^2,$$

which yields the unique solution $\hat{\beta}^{LS} = (X^\top X)^{-1} X^\top Y$ if $X^\top X$ is nonsingular. It can be shown that $\hat{\beta}^{LS}$ follows $N_p(\beta, \sigma^2(X^\top X)^{-1})$.

2.1.2. Decorrelation of estimators

The ordinary least squares method yields $\hat{\beta}^{LS}$ following $N_p(\beta, \sigma^2(X^\top X)^{-1})$. But the one-sample ranking approach relies on the assumption of independence among populations, which indicates that the covariance matrix needs to be a diagonal matrix. Thus, we shall

decorrelate $\hat{\beta}_j^{LS}$ ($j = 1, \dots, p$). According to Gu (2021), we apply a principal component analysis (PCA) whitening transformation on X as follows:

$$X^* = XV\Lambda^{-1/2},$$

where Λ is a diagonal matrix containing the eigenvalues of $X^T X$ (assume that they are all positive). V is a matrix containing the orthonormal eigenvectors of $X^T X$, which gives a rotation needed to decorrelate X . And the factor of $\Lambda^{-1/2}$ makes variances equal to 1. Then the least squares estimator in terms of X^* is given by

$$\hat{\beta}^w = (X^{*\top} X^*)^{-1} X^{*\top} Y.$$

As a result, $\hat{\beta}^w$ follows $N_p(\beta^w, \sigma^2 I_p)$, where $\beta^w = \Lambda^{\frac{1}{2}} V^\top \beta$. After decorrelation, our goal is switched to rank and select $|\beta^w|$. It is worth mentioning that this proposed method can be effective when there is multicollinearity.

2.1.3. Ranking procedure-based variable selection

Let $\beta_j^{w+} = |\beta_j^w|$ and $\bar{\beta}_j^{w+} = |\bar{\beta}_j^w| = \left| \frac{\sum_{l=1}^B \hat{\beta}_{jl}^w}{B} \right|$ ($j = 1, \dots, p; l = 1, \dots, B$), where B indicates the number of resamplings. The ranked $|\bar{\beta}_j^w|$ and $|\beta_j^w|$ are denoted as $\bar{\beta}_{[j]}^{w+}$ and $\beta_{[j]}^{w+}$, respectively. In addition, let $\bar{\beta}_{(j)}^w$ be the mean of resampling estimators related to $\beta_{[j]}^{w+}$. And $|\bar{\beta}_{(j)}^w|$ is denoted as $\bar{\beta}_{(j)}^{w+}$. We consider Bechhofer (1954) the least favorable configuration as

$$\begin{cases} \beta_{[p]}^{w+} - \beta_{[p-k+1]}^{w+} &= 0 \\ \beta_{[p-k+1]}^{w+} - \beta_{[p-k]}^{w+} &= \delta^* \\ \beta_{[p-k]}^{w+} - \beta_{[1]}^{w+} &= 0 \end{cases}.$$

Given one group of resampling estimators, let $\omega = |\bar{\beta}^w|$.

Theorem 2.1. *Under the foregoing assumptions about random error terms and the least favorable configuration, the probability of a correct ranking of $|\beta^w|$ can be expressed as*

$$\begin{aligned} \eta &= \Pr \left[\max \left\{ \bar{\beta}_{(1)}^{w+}, \dots, \bar{\beta}_{(p-k)}^{w+} \right\} < \min \left\{ \bar{\beta}_{(p-k+1)}^{w+}, \dots, \bar{\beta}_{(p)}^{w+} \right\} \right] \\ &= 2(p-k) \int_0^{+\infty} [2\Phi(z) - 1]^{p-k-1} [2 - \Phi(z-d) - \Phi(z+d)]^k \phi(z) dz, \end{aligned}$$

where $z = \frac{\omega}{\sigma/\sqrt{B}}$, $d = \frac{\delta^*}{\sigma/\sqrt{B}}$, and $\phi(\cdot)$, $\Phi(\cdot)$ stand for the probability density function and cumulative distribution function of the standard normal distribution, respectively.

Given fixed B , we prespecify δ^* as “worth detecting” for each possible value of k ($k < p$). Theorem 2.1 would produce the corresponding probability of a correct ranking of $|\beta^w|$. Gu (2021) proposed a ranking approach-based variable selection (RPVS) as follows:

The RPVS algorithm:

- Step 1: Generate a PCA whitening transformation on X and center Y , denoted as X^* and Y^* , respectively.
- Step 2: Given the known σ^2 , we prespecify η , B , and δ^* . Then we set k to start from 1 to $p - 1$ and decide the value(s) of k based on the criterion such that [Theorem 2.1](#) is ensured to be at least η , denoted as k_s , for $s = 1, \dots, t$ and $t \in [1, p - 1]$.
- Step 3: Regress Y^* on X^* and apply residuals bootstrap B times to generate Y^{B_l} , $l = 1, \dots, B$.
- Step 4: For each pair of (Y^{B_l}, X^*) , we apply ordinary least squares to generate $\hat{\beta}_{jl}^w$, for $j = 1, \dots, p$ and $l = 1, \dots, B$.
- Step 5: Compute $\bar{\beta}^{w+}$:

$$\bar{\beta}_j^{w+} = \left| \frac{1}{B} \sum_{l=1}^B \hat{\beta}_{jl}^w \right|, \text{ for } j = 1, \dots, p; l = 1, \dots, B;$$

- Step 6: Rank all $\bar{\beta}_j^{w+}$, denoted as $\bar{\beta}_{[j]}^{w+}$. For each k_s from step 2, we choose the k_s covariates according to the observed $\{\bar{\beta}_{[p-k_s+1]}^{w+}, \dots, \bar{\beta}_{[p]}^{w+}\}$ as the “best” k_s predictors at least at the confidence level of τ . Then we drop the rest of the “poor” $(p - k_s)$ covariate(s) out of the input matrix X . In other words, we construct a submatrix of X^* that has n rows and k_s columns, denoted as $X_{k_s}^*$, for $s = 1, \dots, t$.
- Step 7: Regress Y^* on $X_{k_s}^*$, and denote the corresponding model as M_s , for $s = 1, \dots, t$. Regressing Y^* on X^* , the full-sized model is denoted as M_F .
- Step 8: For each M_s and M_F , calculate the expected predictor error in terms of the test data set, denoted as $test.err_s$ and $test.err_F$, respectively. Then the optimum choice of k is defined as

$$k^{\text{opt}} = \begin{cases} \operatorname{argmin}_{s \in [1, p-1]} \{test.err_s\}, & \min\{test.err_s\} \leq test.err_F \\ p, & \min\{test.err_s\} > test.err_F. \end{cases}$$

Note that the value of δ^* “worth detecting” is assigned based on our experience with Y^* . In general, if one wants to include more predictors in the model at the start of the variable selection process, a smaller δ^* value is preferred, because it allows for more predictors to be included, and vice versa. However, a smaller δ^* will increase the computational time. In this approach, k is treated as a tuning parameter. We will choose k to minimize the expected prediction error according to the test data or incoming observations.

3. SEQUENTIAL CHANGE POINT PROBLEM

Let m be the size of the historical sample. We assume that there is no change in the historical sample. This assumption was considered in C.-S. Chu, Stinchcombe, and White (1996), Horváth et al. (2004), Zhou, Wang, and Tang (2015), and Ratnasingam and Ning (2021b). The historical sample is used to build the prechange regression model. After the historical sample size m , we continue to monitor the future incoming

observations $\{Y_i, x_{i1}, \dots, x_{ip}\}$, $i = m + 1, m + 2, \dots$ sequentially. Let T_m be the monitoring horizon. The linear model after historical observations m is

$$Y_i = X_i^\top \beta_i + \mathcal{E}_i, \quad i = m + 1, m + 2, \dots \quad (3.1)$$

Our objective is, at each time point i , to test whether our model is the same as the model with the historic sample m . Under the null hypothesis, if there is no change in the coefficients,

$$H_0 : \beta_i = \beta_0 \quad \text{for } i = m + 1, m + 2, \dots \quad (3.2)$$

Under the alternative hypothesis, at an unknown time point τ , the coefficients change from β_0 to β_1 . There exists $\tau \geq 1$ such that

$$H_1 : \begin{cases} \beta_i = \beta_0; & i = m + 1, \dots, m + \tau \\ \beta_i = \beta_1; & i = m + \tau + 1, \dots, m + T_m \end{cases} \quad \text{and} \quad \beta_0 \neq \beta_1. \quad (3.3)$$

Following Horváth et al. (2004), we define the detector based on the cumulative sum (CUSUM) of residuals. That is, the CUSUM of $\hat{\mathcal{E}}_i$, $i = m + 1, \dots, m + \tau$ is defined as

$$\Gamma(m, \tau) = \frac{1}{\hat{\sigma}_m} \left| \sum_{i=m+1}^{m+\tau} \hat{\mathcal{E}}_i \right|, \quad (3.4)$$

where $\hat{\mathcal{E}}_i = Y_i^* - X_i^{*\top} \hat{\beta}^w$ for $i = m + 1, m + 2, \dots$ and $\hat{\sigma}_m^2$ is the error variance, defined as

$$\hat{\sigma}_m^2 = \frac{1}{(m - k^{\text{opt}})} \sum_{i=1}^m (Y_i^* - X_i^{*\top} \hat{\beta}^w)^2, \quad (3.5)$$

where k^{opt} is the number of nonzero coefficients in the selected model based on the ranking procedure, and this is the estimated value of q . For a given constant $\gamma \in [0, 1/2)$, the $g(m, \tau, \gamma)$ is called the normalizing function, defined as

$$g(m, \tau, \gamma) = m^{1/2} \left(1 + \frac{\tau}{m} \right) \left(\frac{\tau}{\tau + m} \right)^\gamma, \quad (3.6)$$

where γ is called the control parameter. Following C.-S. Chu, Stinchcombe, and White (1996) and Horváth et al. (2004), we propose the test statistic for monitoring structural change:

$$\Omega = \sup_{1 \leq \tau \leq T_m} \frac{\Gamma(m, \tau)}{g(m, \tau, \gamma)}. \quad (3.7)$$

Now, let $N > 0$. Suppose $T_m < \infty$ with $\lim_{m \rightarrow \infty} T_m/m = N$. The stopping time for the monitoring process is defined as

$$\Lambda(m) = \begin{cases} \inf\{\tau \geq 1 : & \text{if } \Gamma(m, \tau) \geq g(m, \tau, \gamma) c_\alpha(\gamma)\}, \\ T_m & \text{for all } \tau = 1, \dots, T_m, \end{cases} \quad (3.8)$$

where $c_\alpha(\gamma)$ is the $(1 - \alpha)$ th quantile of the asymptotic distribution obtained in Theorem 3.1.

Under the null hypothesis,

$$\lim_{m \rightarrow \infty} P(\Lambda(m) < \infty) = \alpha, \quad (3.9)$$

and under the alternative hypothesis,

$$\lim_{m \rightarrow \infty} P(\Lambda(m) < \infty) = 1. \quad (3.10)$$

Our simulation study revealed that the monitoring process stops promptly for large γ values. Thus, the large value of γ is preferred when the change in the regression coefficients happens shortly after m . We assume that the following conditions hold:

- A1. The model errors $\mathcal{E}_1, \dots, \mathcal{E}_m, \mathcal{E}_{m+1}, \dots$ are i.i.d. random variables. $E(\mathcal{E}_i) = 0$, $\text{Var}(\mathcal{E}_i) = \sigma^2 < \infty$ and $E(|\mathcal{E}_1|^\rho) < \infty$ for some $\rho \geq 2$.
 A2. There is a positive definite matrix C and a constant $\zeta > 0$ such that

$$\left| \frac{1}{m} \sum_{1 \leq i \leq m} \mathbf{X}_i \mathbf{X}_i^\top - C \right| = O(m^{-\zeta}) \quad \text{a.s.}$$

- A3. $N = O(m^\lambda)$ with some $1 \leq \lambda < \infty$ and $\lim_{m \rightarrow \infty} \inf N/m > 0$.

Assumption A1 is standard in a regression model. Assumption A2 is used in Horváth et al. (2004), Horváth, Kokoszka, and Steinebach (2007), and Ratnasingam and Ning (2021a). Assumption A3 is considered in the linear regression model with a change point; see Horváth, Kokoszka, and Steinebach (2007) and Ciuperca (2015).

Theorem 3.1. *Under Assumptions A1–A3, if the null hypothesis holds,*

$$\lim_{m \rightarrow \infty} P(\Omega \leq c_\alpha(\gamma)) = P\left(\sup_{0 \leq t \leq N/(N+1)} \frac{\|W(t)\|_\infty}{t^\gamma} \leq c_\alpha(\gamma)\right),$$

where $\{W(t), 0 \leq t < \infty\}$ denotes the l -dimensional Wiener process, where l is the number of significant features in the model based on historical data.

The asymptotic distribution of test statistics may be obtained using Theorem 3.1. The asymptotic critical value $c_\alpha(\gamma)$ can be obtained from

$$P\left(\sup_{0 \leq t \leq N/(N+1)} \frac{\|W(t)\|_\infty}{t^\gamma} \geq c_\alpha(\gamma)\right) = \alpha,$$

where $\alpha \in (0, 1)$ and the tuning parameter $0 \leq \gamma < 1/2$. We obtain asymptotic critical values through simulation. First, we generate a sequence of i.i.d. l -dimensional random vectors $e_i = (e_{i1}, e_{i2}, \dots, e_{il})$, where $e_{ij} \sim N(0, 1)$. Define $W^*(t) = M^{-1/2} \sum_{i=1}^{tM} e_i$, where M is a grid of 10,000. In each iteration, we calculate the test statistic $\max \|W^*(t)/t^\gamma\|_\infty$ for the proposed method over $t \in \{1/M, 2/M, \dots, N/(N+1)\}$. The critical value for a level- α test can be estimated by the $(1 - \alpha)$ th quantile of the test statistics. The asymptotic critical values for various γ and N values are given in Table 1 (Ratnasingam and Ning 2021a).

Theorem 3.2. *Under Assumptions A1–A3, if the alternative hypothesis holds, we have*

$$\sup_{1 \leq \tau \leq T_m} \frac{\Gamma(m, \tau)}{g(m, \tau, \gamma)} \rightarrow \infty \quad \text{as } m \rightarrow \infty.$$

The proofs of Theorems 3.1 and 3.2 are given in the Appendix.

Table 1. Asymptotic critical values for various values of N and control parameter γ .

α	N	γ		
		0.00	0.25	0.49
0.010	2	2.4471	2.8169	3.7326
	4	2.6865	2.9540	3.7450
	6	2.7858	3.0044	3.7472
	9	2.8471	3.0451	3.7499
0.025	2	2.2118	2.5620	3.4637
	4	2.4306	2.6815	3.4744
	6	2.5148	2.7352	3.4786
	9	2.5721	2.7675	3.4799
0.050	2	2.0209	2.3590	3.2573
	4	2.2224	2.4698	3.2682
	6	2.2974	2.5104	3.2717
	9	2.3515	2.5393	3.2741

4. SIMULATION STUDIES

In this section, we conduct Monte Carlo simulations to evaluate the performance of the sequential change point detection procedure in linear models using the ranking selection procedure. To evaluate the effectiveness, we examine three criteria typically used for the determination of the quality of a sequential change point detection approach:

1. Type I error rate: Close to the nominal level
2. Power of the test: Preferably close to 1
3. Detection time under the alternative hypothesis: Stop as soon as possible after a change is noticed.

We generate data from the following model:

$$Y_i = X_i^\top \beta_0 + \mathcal{E}_i, \quad i = 1, \dots, m + T_m.$$

4.1. Low Dimension

To study the performance of the monitoring process in a medium-dimensional setting, we conducted another simulation study. We generated data sets with (p, m) , considering $(10, 75)$ and $(10, 100)$. We adopted the same settings used in Ratnasingam and Ning (2021a), as given below.

Setting I (type I error calculations)

- Under H_0 , the true parameter vectors $\beta_0 \in \{-2, 0, 2, 0, 10, 1, 0, 0, 8, -5\}$.
- The predictor variables X_i for all $i \in \{1, \dots, p\} \setminus \{3, 4, 5\}$ have a standard normal distribution $N(0, 1)$ and $X_3 \sim N(2, 1)$, $X_4 \sim N(4, 1)$ and $X_5 \sim N(5, 1)$.
- The model errors \mathcal{E}_i are i.i.d. $N(0, 1)$.

Setting II (stopping time and power analysis)

- Under H_0 , the true parameter vectors $\beta_0 \in \{0, 0, 2, 0, 0, 1, 0, 0, 1, 0\}$.
- Under H_1 , we consider the parameter vector $\beta_1 \in \{0, 0, 0, 3, 0, 0, 1, 0, 0, -1\}$.

Table 2. Type I errors for various values of γ and the nominal significance level $\alpha = 0.05$ and $p = 10$.

m	N	γ		
		0.00	0.25	0.45
75	2	0.023	0.038	0.043
	4	0.036	0.039	0.045
	6	0.039	0.044	0.049
	8	0.045	0.047	0.050
100	2	0.017	0.032	0.039
	4	0.028	0.037	0.042
	6	0.034	0.041	0.046
	8	0.038	0.039	0.041

- Under H_0 , X_i for all $i \in \{1, \dots, 10\} \setminus \{3, 4, 5\}$ have a standard normal distribution $N(0, 1)$ and $X_3 \sim N(2, 1)$, $X_4 \sim N(4, 1)$, and $X_5 \sim N(5, 1)$.
- For the second distribution, under H_1 , the i th explanatory variable is $X_i + 0.8$, where $X_i \sim N(0, 1)$ for all $i \in \{1, \dots, 10\}$.
- The model errors \mathcal{E}_i are i.i.d. $N(0, 1)$.

Monte Carlo simulation was used to analyze the finite sample performance of the detection procedure. First, we evaluated the empirical type I errors. The various values of control parameter γ and the different sizes of the historical observations m were considered. The $\gamma \in \{0, 0.25, 0.45\}$ and $m \in \{75, 100\}$. The results are based on 1,000 iterations. The empirical type I error probabilities are summarized in Table 2. We observed slightly deflated type I errors for small N ; however, this improved as N increased. In addition, the empirical type I errors appeared to increase when the control parameter increased. Figure 1 compares the empirical type I errors for the proposed method. In terms of the results, utilizing a large N and a large control parameter γ could produce empirical type I errors that are above the nominal level. As a result, for big N , a small control parameter is preferable and vice versa. In addition, we observed a pattern in which the empirical type I errors tended to decrease as the historical sample size m increased.

Next, we conducted a power analysis using Setting II to assess the effectiveness of the proposed method. Various control parameter values were considered, including $\gamma \in \{0, 0.25, 0.45\}$. Simulations were conducted at different true change point locations $\tau^* \in \{1, 25, 50\}$ with various historical sample sizes, $m \in \{75, 100\}$. The results are based on 2,500 iterations and are summarized in Table 3. It can be clearly seen that as the historical sample size increased, the power tended to increase. In addition, the power tended to decrease as the control parameter increased. This was due to the fact that the empirical type I error increased as the control parameter increased. Moreover, as the change location moved further from the historical sample, the power tended to decrease. In the stopping time calculations below, larger τ^* values resulted in more delayed detection, as shown in Figure 2. This agrees with our foregoing conclusion that larger γ values are recommended if we want the monitoring process to stop promptly.

We monitored the process from $(m + 1)$ to $9m$ observations with various historical sample sizes, such as $m \in \{75, 100\}$. In addition, we changed the true change point location, considering $\tau^* \in \{1, 25, 50\}$ and level $\alpha = 0.05$. Simulation results are

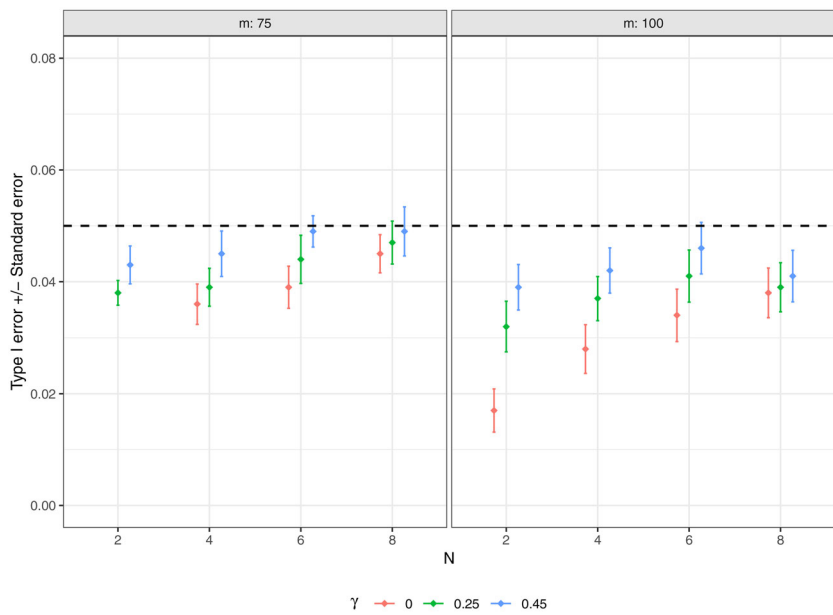


Figure 1. Empirical type I error comparison for the low dimension.

Table 3. Power analysis for various values of γ and the nominal significance level $\alpha = 0.05$ and $p = 10$.

m	τ^*	γ		
		0.00	0.25	0.45
75	1	0.969	0.943	0.902
	25	0.955	0.948	0.936
	50	0.942	0.913	0.902
100	1	0.979	0.959	0.945
	25	0.972	0.960	0.948
	50	0.965	0.950	0.943

summarized in Table 4. It is evident that the choice of the control parameter γ affected the stopping time. As noted in Horváth et al. (2004), our simulation results confirmed that smaller γ values result in a longer period in which structural changes can be detected, whereas larger γ values lead to quicker detection rates. Table 4 indicates that when the true change point location is far away from the historical sample, a small γ is preferable and vice versa. Moreover, we estimated the densities of the stopping time at various change point locations with different historical sample sizes (m) and various γ values. The historical sample size m had a significant influence on the stopping time determination. We observed a considerable variation in the estimated densities as m changed from 75 to 100. There was a small variation between the estimated densities for a fixed control value γ irrespective of the historical sample size. See the graph in Figure 2.

4.2. Medium Dimension

To study the performance of the monitoring process in a medium-dimensional setting, we conducted another simulation study. We generated data sets with (p, m) , considering $(100, 200)$ and $(100, 300)$. We considered the following two settings:

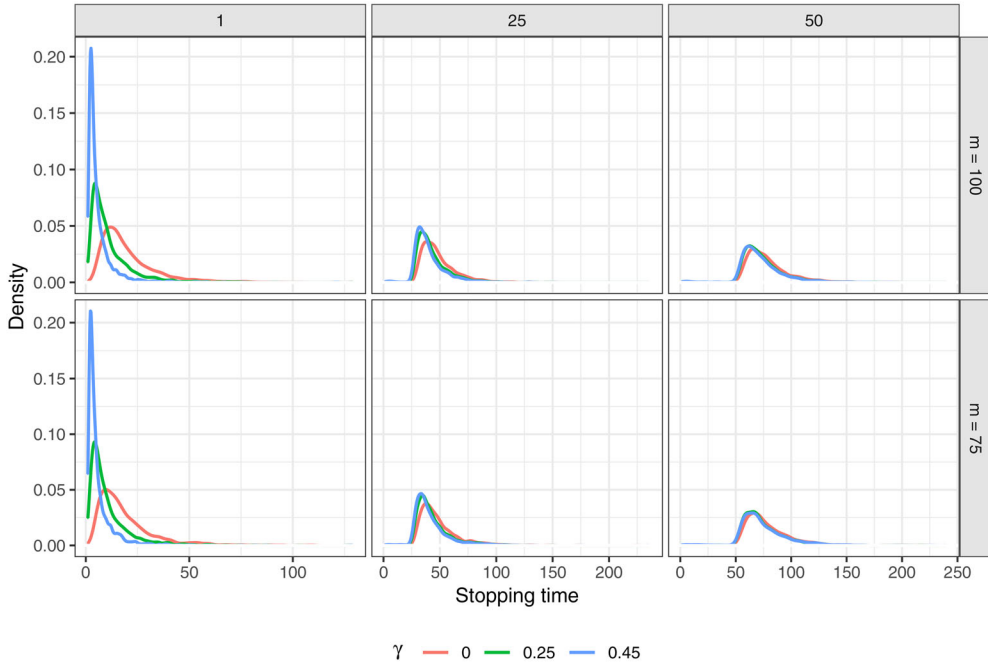


Figure 2. Estimated density of the stopping time for $\tau^* = \{1, 25, 50\}$ with low dimension.

Table 4. Summary statistics for the detection time for low dimension with $\tau^* \in \{1, 25, 50\}$, $\gamma \in \{0, 0.25, 0.45\}$ and $\alpha = 0.05$.

τ^*	Summary/ γ	$m = 75$			$m = 100$		
		0	0.25	0.45	0	0.25	0.45
1	Min	3	2	1	4	2	1
	Q1	10	4	2	11	5	2
	Med	15	7	4	16	8	4
	Q3	23	13	7	25	14	7
	Max	128	127	128	129	129	121
25	Min	25	13	1	27	22	1
	Q1	37	34	32	38	34	32
	Med	44	39	38	45	39	37
	Q3	54	49	47	55	49	46
	Max	225	150	236	152	150	152
50	Min	25	13	1	48	22	1
	Q1	64	62	61	65	62	61
	Med	73	70	70	74	70	69
	Q3	87	83	83	86	82	81
	Max	219	219	240	195	195	195

Setting I (type I error calculations)

- The nonzero components of the true parameters are $\beta_{0,1} = -5$, $\beta_{0,2} = 2$, $\beta_{0,3} = 5$, $\beta_{0,4} = 1$, $\beta_{0,5} = -3$, $\beta_{0,61} = -10$, and $\beta_{0,91} = 8$.
- The predictor variables X_i for all $i \in \{1, \dots, p\} \setminus \{3, 4, 5\}$ have a standard normal distribution $N(0, 1)$ and $X_3 \sim N(2, 1)$, $X_4 \sim N(4, 1)$ and $X_5 \sim N(5, 1)$.
- The model errors \mathcal{E}_i are i.i.d. $N(0, 1)$.

Table 5. Type I errors for various values of γ and the nominal significance level $\alpha = 0.05$ and $p = 100$.

m	N	γ		
		0.00	0.25	0.45
200	2	0.030	0.033	0.035
	4	0.032	0.038	0.038
	6	0.036	0.038	0.040
	8	0.038	0.041	0.044
300	2	0.020	0.024	0.033
	4	0.022	0.026	0.035
	6	0.025	0.027	0.036
	8	0.027	0.033	0.038

Setting II (stopping time and power analysis)

- Under H_0 , the true parameter vectors $\beta_{0,1} = -1$, $\beta_{0,2} = 1$, $\beta_{0,3} = -1$, $\beta_{0,4} = 4$, $\beta_{0,5} = -2$, $\beta_{0,58} = -3$, and $\beta_{0,86} = 2$.
- Under H_1 , we consider the parameter vectors $\beta_{1,1} = 3$, $\beta_{1,2} = 2$, $\beta_{1,45} = -2$, and $\beta_{1,93} = 2$.
- Under H_0 , X_i for all $i \in \{1, \dots, p\} \setminus \{3, 4, 5\}$ have a normal distribution $N(0, 1)$ and $X_3 \sim N(2, 1)$, $X_4 \sim N(4, 1)$ and $X_5 \sim N(5, 1)$
- For the second distribution, under H_1 , the i th explanatory variable is $X_i + 0.8$, where $X_i \sim N(0, 1)$ for all $i \in \{1, \dots, p\}$.
- The model errors \mathcal{E}_i are i.i.d. $N(0, 1)$.

Table 5 summarizes the empirical type I errors for the medium dimension. The various control parameter values $\gamma \in \{0, 0.25, 0.45\}$ and the different sizes of the historical observations $m \in \{200, 300\}$ were considered. The results are based on 2,500 iterations. Figure 3 compares the type I error for the proposed procedure. The empirical type I errors based on the historical sample size $m = 200$ were always larger than those based on $m = 300$. This suggests that the effect of m is more obvious in the medium dimension than in the low dimension, as shown in Figure 3. In the proposed method, it is vital to properly select the value of the control parameter. Type I errors were comparatively low for small γ values. In addition, smaller N provided slightly deflated type I errors, which improved for large N . Thus, for small N , we recommend higher control parameter values close to 0.5. A large historical sample, for example, $m = 300$, produces deflated type I errors.

Table 6 compares the power of the proposed method. The process was monitored from $(m + 1)$ to $9m$ observations. The power was roughly equal to 1 for a large historical sample size m independent of the level of significance. Next, we obtained the stopping time for the monitoring process. The true change point locations were $\tau^* \in \{1, 25, 50\}$ and level $\alpha = 0.05$. The results are summarized in Table 7. As we discussed previously, a larger γ value is preferable when the change occurs quickly after the historical sample size m . We observed variation in density plots but decreases due to the large historical sample size. The estimated density curves are graphed in Figure 4. Not surprisingly, all of the points highlighted under the low dimension apply to the medium dimension as well. Furthermore, when comparing the numerical results to those obtained in the low dimension, we observed that as the dimensionality increased, the power and accuracy increased, whereas the empirical type I error decreased. We

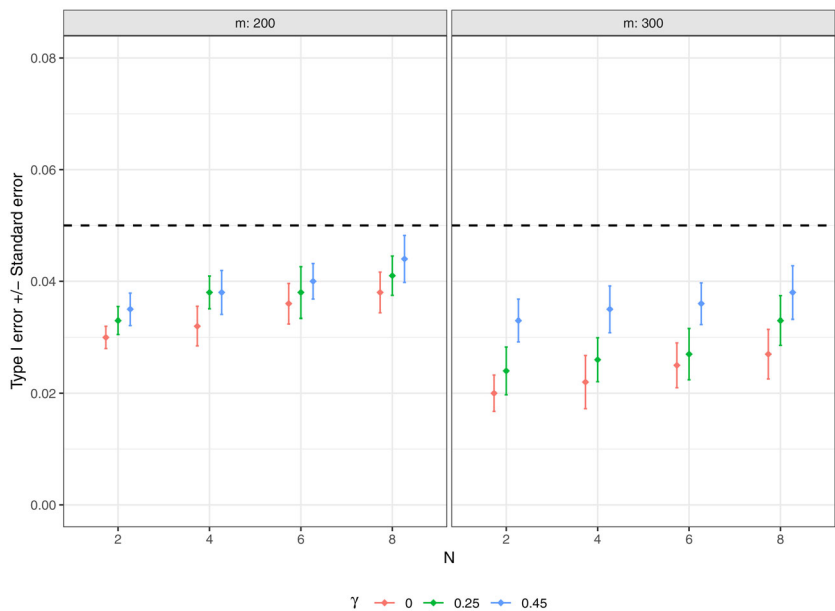


Figure 3. Empirical type I error comparison for the medium dimension.

Table 6. Power analysis for various values of γ and the nominal significance level $\alpha = 0.05$ and $p = 100$.

		γ		
m	τ^*	0.00	0.25	0.49
200	1	0.988	0.979	0.960
	25	0.975	0.968	0.956
	50	0.962	0.953	0.944
300	1	0.993	0.985	0.977
	25	0.982	0.974	0.968
	50	0.971	0.967	0.949

Table 7. Summary statistics for the detection time for medium dimension with $\tau^* \in \{1, 25, 50\}$, $\gamma \in \{0, 0.25, 0.45\}$ and $\alpha = 0.05$.

		$m = 200$			$m = 300$		
τ^*	Summary/ γ	0	0.25	0.45	0	0.25	0.45
1	Min	3	2	1	4	2	1
	Q1	13	3	2	18	4	2
	Med	23	6	3	33	7	3
	Q3	42	10	4	59	13	4
	Max	842	170	68	669	354	28
25	Min	26	26	1	28	26	1
	Q1	39	31	29	43	32	29
	Med	52	37	32	61	38	32
	Q3	78	49	40	93	51	38
	Max	817	467	405	1,395	923	197
50	Min	51	30	1	53	31	1
	Q1	66	59	56	72	60	56
	Med	82	67	63	92	70	63
	Q3	115	85	76	130	88	74
	Max	1,490	1,489	1,520	2,077	986	997

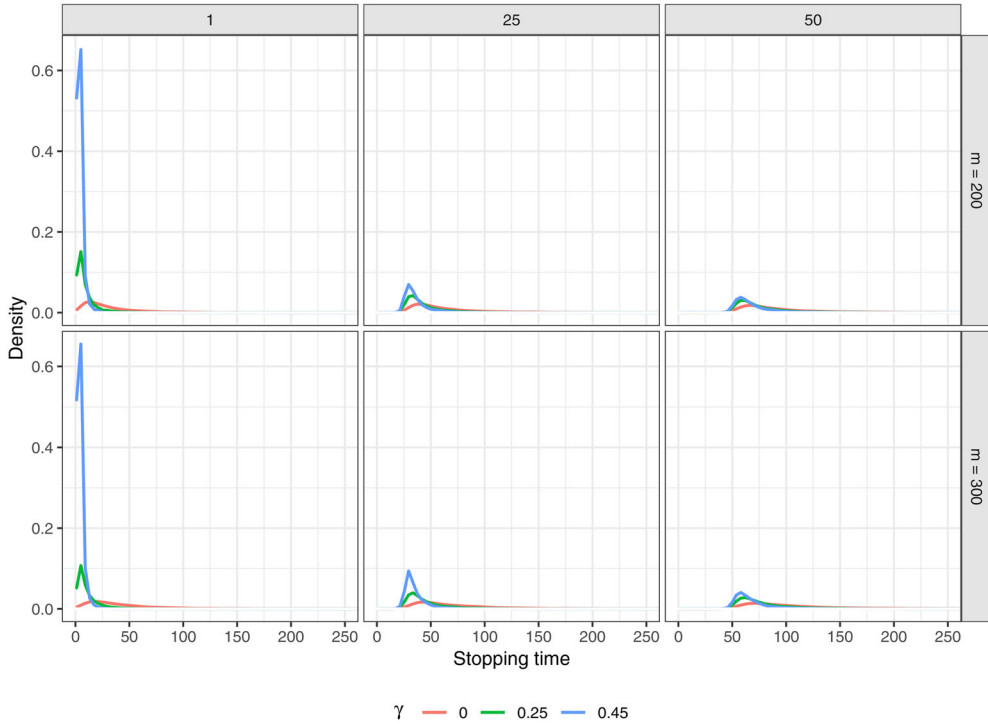


Figure 4. Estimated density of the stopping time for $\tau^* = \{1, 25, 50\}$ with medium dimension.

also compared the results presented in Tables 3, 4, and 6 in Ratnasingam and Ning (2021a), and the results showed that the proposed method performs at least as well as the SCAD-based sequential change point detection procedure. This demonstrates that our proposed method is competitive compared to the existing methods for low- and medium-dimensional data.

5. REAL DATA APPLICATION

In this section, we apply the proposed sequential change point detection method to the HIV Drug Resistance Mutations data set. The data set was originally described in Rhee et al. (2006). This data set includes the outcomes for one specific drug, nel-finavir, a protease inhibitor, as well as the existence of protease gene mutations, which may lead to drug resistance. In Rhee et al. (2006), HIV isolates from infected people were extracted, sequenced, and evaluated for resistance to several medications used in HIV therapy. The goal of their research was to identify the mutations associated with treatment resistance, which will aid in the development of novel anti-retroviral medications as well as the efficient use of those that are already available. The data contain expression measurements of 361 genes from 842 patients. The explanatory variables indicate the position and mutation. The response variable y is the outcome of the drug susceptibility assay. Higher numbers suggest greater resistance to the drug.

To find any structural changes in the data set, we first apply the standard log-likelihood method while assuming normality of the data. There is no change point detected in the first 150 observations. Thus, we can use any number of observations between 0 and 150. In our case, the first 125 observations are considered historical data. Additionally, we set δ^* equal to the standard deviation of the response variable y . Because we standardize the input matrix X , a change of one unit in the standard deviation of the response variable is considered a reasonably significant change in drug resistance testing.

We applied the method sequentially with the control parameter value $\gamma = \{0, 0.45\}$ with $\alpha = 0.05$. When $\gamma = 0$, we found seven change points including $\{317, 383, 448, 542, 606, 671, 737\}$, and for $\gamma = 0.45$ we detected eight change points including $\{314, 377, 441, 505, 567, 629, 692, 755\}$.

The change point locations corresponding to $\gamma = 0$ and $\gamma = 0.45$ are graphed in Figure 5. We observed an early detection when γ is close to 0.5. In particular, the change points that correspond to $\gamma = 0$ are completely different from those that correspond to $\gamma = 0.45$, as we expected. When $\gamma = 0.45$, we detected an additional change point in the data. This is consistent with our simulation results, because the larger γ s are typically more sensitive to change points. In this particular example, $\gamma = 0.45$ is preferable. Figure 5(b) clearly shows a change at 505; however, the choice of $\gamma = 0$ fails to identify the change due to the delay detection, as shown in Figure 5(a).

The advantage of the sequential change point detection approach is that fewer samples are required for decision making in comparison with change point detection with a

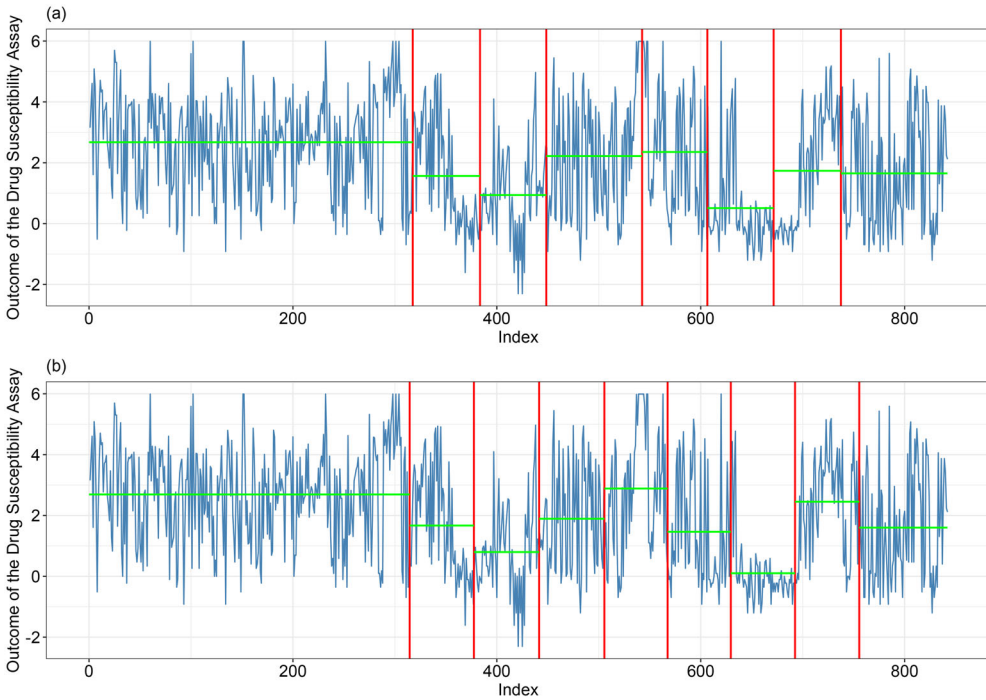


Figure 5. Change point detection for the drug susceptibility assay when (a): $\gamma = 0$ and (b) $\gamma = 0.45$.

fixed sample size. In this application, for example, our method only requires 125 observations and the monitoring process stops after 317 samples after the first change point is detected. The traditional log-likelihood method, however, requires all observations ($n = 842$) to estimate the change location.

6. DISCUSSION

In this article, we proposed a new sequential change point detection procedure for high-dimensional data using the ranking selection procedure. Simulation studies were conducted under two different scenarios, including low and medium dimensions. When N increases, the type I error probability improves significantly. The proposed method is superior for $N \geq 6$, whereas a large control parameter is preferable for small N ($N < 6$). Under the medium dimension, we observe that the type I error decreases as the historical sample size m increases, whereas under the low dimension, the effect of m is not outstanding. Additionally, both the historical sample size and the control parameter have a significant impact on test power in low and medium dimensions. Specifically, the test power increases as the historical sample size increases, whereas it decreases as the control parameter increases. Next, we determine the stopping time under various settings. From our results, the larger control parameter values tend to detect structural change much faster, whereas smaller control parameter values contribute to detection delays. Similar to Ratnasingam and Ning's (2021a) findings, if a change happens soon after a historical sample size, we recommend a larger value of γ close to 0.5. Smaller γ values are preferable if the structural change happens far away from the historical sample. Application to real data sets illustrates the sequential detection procedure. It should be noted that the choice of γ and m has an effect on the detection process and can be determined in the context of the study.

Because our proposed method relies on PCA whitening transformation on X , there is a trade-off between information loss and dimensionality reduction. However, this article focuses on sequential change point detection rather than the interpretability of predictors, and this approach has shown its competitiveness compared to the existing methods, including Ratnasingam and Ning (2021a). In the future, we would like to investigate a sequential change point detection procedure using the ranking selection procedure for high-dimensional data. Moreover, the simulation study indicates a reduction in the rate of type I errors for a large historical sample size. This has adverse side effects such as reduction in power and detection delays. Therefore, a modified test statistic is of interest.

APPENDIX PROOFS OF THEOREMS

Proof of Theorem 2.1. Let X_i be independently derived from normal distributions with unknown mean μ_i ($i = 1, \dots, p$) and a common known variance σ^2 . And $|\mu_i|$ and the ranked $|\mu_i|$ are denoted by μ_i^{abs} and $\mu_{[i]}^{abs}$, respectively, where

$$0 \leq \mu_{[1]}^{abs} \leq \dots \leq \mu_{[p]}^{abs}.$$

Let \bar{X}_i be the sample mean based on N independent observations from the i th population. $|\bar{X}_i|$ and the ranked $|\bar{X}_i|$ are denoted by \bar{X}_i^{abs} and $\bar{X}_{[i]}^{abs}$, respectively. Additionally, let $\bar{X}_{(i)}^s$ be the sample mean related to the population with $\mu_{[i]}^{abs}$, and $|\bar{X}_{(i)}^s|$ is denoted by $\bar{X}_{(i)}^{abs}$.

Recall that we are interested in selecting the “best” k populations among p populations in terms of the rank of μ_i^{abs} ($i = 1, \dots, p$), which is unknown to us. Now we define the probability of a correct ranking as

$$\eta = \Pr \left[\max \{ \bar{X}_{(1)}^{abs}, \dots, \bar{X}_{(p-k)}^{abs} \} < \min \{ \bar{X}_{(p-k+1)}^{abs}, \dots, \bar{X}_{(p)}^{abs} \} \right] \quad (\text{A.1})$$

where $\frac{1}{C(p,k)} \leq \eta \leq 1$.

We consider the following least favorable configuration:

$$\begin{cases} \mu_{[p]}^{abs} - \mu_{[p-k+1]}^{abs} &= 0 \\ \mu_{[p-k+1]}^{abs} - \mu_{[p-k]}^{abs} &= \delta^* \\ \mu_{[p-k]}^{abs} - \mu_{[1]}^{abs} &= 0 \end{cases}$$

Then we rewrite (A.1) as

$$\begin{aligned} \eta &= \Pr \left[\max \{ \bar{X}_{(1)}^{abs}, \dots, \bar{X}_{(p-k)}^{abs} \} < \min \{ \bar{X}_{(p-k+1)}^{abs}, \dots, \bar{X}_{(p)}^{abs} \} \right] \\ &= k \times \Pr \left[\max \{ \bar{X}_{(1)}^{abs}, \dots, \bar{X}_{(p-k)}^{abs} \} < \bar{X}_{(p-k+1)}^{abs} < \min \{ \bar{X}_{(p-k+2)}^{abs}, \dots, \bar{X}_{(p)}^{abs} \} \right] \\ &= k \int_0^{+\infty} \left[\Phi \left(\frac{y - \mu_{[p-k]}^{abs}}{\sigma/\sqrt{N}} \right) + \Phi \left(\frac{y + \mu_{[p-k]}^{abs}}{\sigma/\sqrt{N}} \right) - 1 \right]^{p-k} \\ &\quad \left[2 - \Phi \left(\frac{y - \mu_{[p-k]}^{abs} - \delta^*}{\sigma/\sqrt{N}} \right) - \Phi \left(\frac{y + \mu_{[p-k]}^{abs} + \delta^*}{\sigma/\sqrt{N}} \right) \right]^{k-1} \\ &\quad \frac{\sqrt{N}}{\sigma} \left[\phi \left(\frac{y - \mu_{[p-k]}^{abs} - \delta^*}{\sigma/\sqrt{N}} \right) + \phi \left(\frac{y + \mu_{[p-k]}^{abs} + \delta^*}{\sigma/\sqrt{N}} \right) \right] dy \quad , \quad (\text{A.2}) \\ &= (p-k) \int_0^{+\infty} \left[\Phi \left(\frac{y - \mu_{[p-k]}^{abs}}{\sigma/\sqrt{N}} \right) + \Phi \left(\frac{y + \mu_{[p-k]}^{abs}}{\sigma/\sqrt{N}} \right) - 1 \right]^{p-k-1} \\ &\quad \left[2 - \Phi \left(\frac{y - \mu_{[p-k]}^{abs} - \delta^*}{\sigma/\sqrt{N}} \right) - \Phi \left(\frac{y + \mu_{[p-k]}^{abs} + \delta^*}{\sigma/\sqrt{N}} \right) \right]^k \\ &\quad \frac{\sqrt{N}}{\sigma} \left[\phi \left(\frac{y - \mu_{[p-k]}^{abs}}{\sigma/\sqrt{N}} \right) + \phi \left(\frac{y + \mu_{[p-k]}^{abs}}{\sigma/\sqrt{N}} \right) \right] dy \end{aligned}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function and cumulative distribution function of the standard normal distribution, respectively.

To ensure that the probability of a correct ranking based on $\bar{X}_{[i]}^{abs}$ ($i = 1, \dots, p$) is at least η , Gu (2021) reformulated (A.2) as

$$\begin{aligned}
\eta &= (p-k) \int_0^{+\infty} \left[\Phi\left(\frac{y - \mu_{[p-k]}^{abs}}{\sigma/\sqrt{N}}\right) + \Phi\left(\frac{y + \mu_{[p-k]}^{abs}}{\sigma/\sqrt{N}}\right) - 1 \right]^{p-k-1} \\
&\quad \left[2 - \Phi\left(\frac{y - \mu_{[p-k]}^{abs} - \delta^*}{\sigma/\sqrt{N}}\right) - \Phi\left(\frac{y + \mu_{[p-k]}^{abs} + \delta^*}{\sigma/\sqrt{N}}\right) \right]^k \\
&\quad \frac{\sqrt{N}}{\sigma} \left[\phi\left(\frac{y - \mu_{[p-k]}^{abs}}{\sigma/\sqrt{N}}\right) + \phi\left(\frac{y + \mu_{[p-k]}^{abs}}{\sigma/\sqrt{N}}\right) \right] dy \\
&\geq (p-k) \int_0^{+\infty} \left[\Phi\left(\frac{y}{\sigma/\sqrt{N}}\right) + \Phi\left(\frac{y}{\sigma/\sqrt{N}}\right) - 1 \right]^{p-k-1} \\
&\quad \left[2 - \Phi\left(\frac{y - \delta^*}{\sigma/\sqrt{N}}\right) - \Phi\left(\frac{y + \delta^*}{\sigma/\sqrt{N}}\right) \right]^k \\
&\quad \frac{\sqrt{N}}{\sigma} \left[\phi\left(\frac{y}{\sigma/\sqrt{N}}\right) + \phi\left(\frac{y}{\sigma/\sqrt{N}}\right) \right] dy \\
&= 2(p-k) \int_0^{+\infty} \left[2\Phi(z) - 1^{[p-k-1]} [2 - \Phi(z-d) - \Phi(z+d)]^k \phi(z) \right] dz,
\end{aligned} \tag{A.3}$$

where $z = \frac{y}{\sigma/\sqrt{N}}$ and $d = \frac{\delta^*}{\sigma/\sqrt{N}}$.

Because $\hat{\beta}^{LS}$ follows $N_p(\beta, \sigma^2(X^\top X)^{-1})$ and the above ranking approach relies on the assumption of independence and constant variance among p populations, we shall decorrelate $\hat{\beta}^{LS}$ before (A.3) is applied. First, we apply a singular value decomposition on X :

$$X_{n \times p} = U_{n \times n} S_{n \times p} V_{p \times p}^\top,$$

where $U^\top U = I_n$, $V^\top V = I_p$, and S is a rectangular diagonal matrix with nonnegative real numbers on the diagonal.

More specifically, the columns of U are generated by the eigenvectors of XX^\top , the eigenvectors of $X^\top X$ contribute the columns of V , and $S^\top S = \Lambda_{p \times p}$ forms a diagonal matrix with the eigenvalues of $X^\top X$ (assume that they are all positive) on the diagonal. Then

$$\begin{aligned}
X^\top X &= VS^\top U^\top USV^\top \\
&= V\Lambda V^\top.
\end{aligned}$$

Secondly, we apply a PCA whitening transformation on X . Referring to Gu (2021), this process is defined as follows.

PCA whitening transformation:

$$X^* = XV\Lambda^{-1/2},$$

where matrix V gives a rotation needed to decorrelate X (maps the original features to principal components). The factor of $\Lambda^{-1/2}$ makes variances equal to 1.

Then the least squares estimator in terms of X^* is given by

$$\hat{\beta}^w = (X^{*\top} X^*)^{-1} X^{*\top} Y.$$

Now we have

$$\begin{aligned}
 E(\hat{\beta}^w) &= E\left[(X^{*\top} X^*)^{-1} X^{*\top} (X\beta + \varepsilon)\right] \\
 &= \Lambda^{\frac{1}{2}} V^\top \beta \\
 &= \beta^w \\
 \Sigma_{\hat{\beta}^w} &= (X^{*\top} X^*)^{-1} X^{*\top} \text{Var}(Y) X^* (X^{*\top} X^*)^{-1} \\
 &= \sigma^2 (X^{*\top} X^*)^{-1} \\
 &= \sigma^2 (\Lambda^{-\frac{1}{2}} V^\top V S^\top U^\top U S V^\top V \Lambda^{-\frac{1}{2}})^{-1} \\
 &= \sigma^2 (\Lambda^{-\frac{1}{2}} \Lambda \Lambda^{-\frac{1}{2}})^{-1} \\
 &= \sigma^2 I_p
 \end{aligned}$$

Thus, $\hat{\beta}^w$ follows $N_p(\beta^w, \sigma^2 I_p)$, where β^w is a transformation of β due to whitening X . Therefore, the probability of a correct ranking of $|\beta^w|$ is expressed as [Theorem 2.1](#). This completes the proof.

A.1. Proofs in Section 3

We will show that [Theorems 3.1](#) and [3.2](#) hold under the historical sample size m .

Proof of Theorem 3.1. Consider

$$\sum_{i=m+1}^{m+\tau} \hat{\varepsilon}_i = \sum_{i=m+1}^{m+\tau} \varepsilon_i + \sum_{i=m+1}^{m+\tau} X_i^\top (\hat{\beta}^w - \beta_0). \quad (\text{A.4})$$

Using Assumptions A1 and A2, by the central limit theorem, we have

$$(X_{\mathcal{A}}^\top X_{\mathcal{A}})^{-1} X_{\mathcal{A}}^\top \mathcal{E} = O_P(m^{-1/2}). \quad (\text{A.5})$$

Under Assumption A2, we have that

$$(X_{\mathcal{A}}^\top X_{\mathcal{A}})^{-1} = \frac{1}{m} C_{\mathcal{A}}^{-1} (1 + o_P(1)), \quad (\text{A.6})$$

where matrix $C_{\mathcal{A}}$ contains the elements of matrix C with the index in set \mathcal{A} . Because $(\hat{\beta}^w - \beta_0)$ converges to zero with the rate $m^{-1/2}$, we have

$$(\hat{\beta}^w - \beta_0)_{\mathcal{A}} = (X_{\mathcal{A}}^\top X_{\mathcal{A}})^{-1} X_{\mathcal{A}}^\top \mathcal{E} (1 + o_P(1)). \quad (\text{A.7})$$

Similarly, we can show that for any $e > 0, \tau \geq 1$, we have

$$P\left(\sum_{i=m+1}^{m+\tau} X_i^\top (\hat{\beta}^w - \beta_0) = \sum_{i=m+1}^{m+\tau} X_{i,\mathcal{A}}^\top (\hat{\beta}^w - \beta_0)_{\mathcal{A}}\right) > 1 - 2e. \quad (\text{A.8})$$

Using [\(A.8\)](#) and [\(A.7\)](#), the relation [\(A.4\)](#) becomes

$$\sum_{i=m+1}^{m+\tau} \hat{\varepsilon}_i = \sum_{i=m+1}^{m+\tau} \varepsilon_i - \left(\sum_{i=m+1}^{m+\tau} X_{i,\mathcal{A}}^\top\right) (X_{\mathcal{A}}^\top X_{\mathcal{A}})^{-1} X_{\mathcal{A}}^\top \mathcal{E} (1 + o_P(1)). \quad (\text{A.9})$$

Thus, the CUSUM of residuals based on the ranking selection procedure on \mathcal{A} is given as

$$\sum_{i=m+1}^{m+\tau} \varepsilon_i - \left(\sum_{i=m+1}^{m+\tau} X_{i,\mathcal{A}}^\top\right) (X_{\mathcal{A}}^\top X_{\mathcal{A}})^{-1} X_{\mathcal{A}}^\top \mathcal{E}. \quad (\text{A.10})$$

Now, applying [Theorem 2.1](#) of Horváth et al. (2004), we can prove [Theorem \(A.4\)](#), which is therefore excluded. ■

Proof of Theorem 3.2. The proof is similar to Horváth et al. (2004). Thus, details are omitted here. ■

ACKNOWLEDGMENTS

The authors thank two anonymous referees and the Associate Editor for their constructive comments and suggestions, which helped to improve this article significantly.

DISCLOSURE

The authors have no conflicts of interest to report.

ORCID

Suthakaran Ratnasingam  <http://orcid.org/0000-0002-6802-192X>

REFERENCES

- Bechhofer, R. E. 1954. "A Single-Sample Multiple Decision Procedure." *Annals of Mathematical Statistics* 25 (1):16–39.
- Bechhofer, R. E., C. W. Dunnett, and M. Sobel. 1954. "A Two-Sample Multiple Decision Procedure for Ranking Means of Normal Populations with a Common Unknown Variance." *Biometrika* 41 (1–2):170–6.
- Candes, E., and T. Tao. 2007. "The Dantzig Selector: Statistical Estimation When p is Much Larger than n ." *Annals of Statistics* 35 (6):2313–51.
- Chen, H. 2019. "Sequential Change-Point Detection Based on Nearest Neighbors." *The Annals of Statistics* 47 (3):1381–407.
- Chu, C.-S., M. Stinchcombe, and H. White. 1996. "Monitoring Structural Change." *Econometrica* 64 (5):1045–65.
- Chu, L., and H. Chen. 2019. "Asymptotic Distribution-Free Change-Point Detection for Multivariate and Non-Euclidean Data." *The Annals of Statistics* 47 (1):382–414.
- Ciuperca, G. 2015. "Real Time Change-Point Detection in a Model by Adaptive Lasso and Cusum." *Journal de la Société Française de Statistique* 156 (4):113–32.
- Fan, J., and R. Li. 2001. "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties." *Journal of the American Statistical Association* 96:1348–60.
- Gu, C. 2021. "Advancing Bechhofer's Ranking Procedures to High-Dimensional Variable Selection." Mathematics Ph.D. Dissertations, Bowling Green State University, 81.
- Horváth, L., M. Hušková, P. Kokoszka, and J. Steinebach. 2004. "Monitoring Changes in Linear Models." *Journal of Statistical Planning and Inference* 126:225–51.
- Horváth, L., P. Kokoszka, and J. Steinebach. 2007. "On Sequential Detection of Parameter Changes in Linear Regression." *Statistics and Probability Letters* 77 (9):885–95.
- Lorden, G. 1971. "Procedures for Reacting to a Change in Distribution." *The Annals of Mathematical Statistics* 42 (6):1897–908.
- Page, E. S. 1954. "Continue Inspection Schemes." *Biometrika* 41:100–35.
- Ratnasingam, S., and W. Ning. 2021a. "Monitoring Sequential Structural Changes in Penalized High-Dimensional Linear Models." *Sequential Analysis* 40 (3):381–404.
- Ratnasingam, S., and W. Ning. 2021b. "Sequential Change Point Detection for High-Dimensional Data Using Nonconvex Penalized Quantile Regression." *Biometrical Journal* 63 (3):575–98.
- Rhee, S.-Y., J. Taylor, G. Wadhera, A. Ben-Hur, D. L. Brutlag, and R. W. Shafer. 2006. "Genotypic Predictors of Human Immunodeficiency Virus Type 1 Drug Resistance." *Proceedings of the National Academy of Sciences* 103:17355–60.

- Roberts, S. W. 1966. "A Comparison of Some Control Chart Procedures." *Technometrics* 8:411–30.
- Shiryaev, A. N. 1963. "On Optimum Methods in Quickest Detection Problems." *Theory of Probability and Its Applications* 8 (1):22–46.
- Siegmund, D. 1985. *Sequential Analysis: Tests and Confidence Intervals*. New York: Springer. doi: [10.1007/978-1-4757-1862-1](https://doi.org/10.1007/978-1-4757-1862-1).
- Tartakovsky, A., I. Nikiforov, and M. Basseville. 2014. *Sequential Analysis: Hypothesis Testing and Changepoint Detection*, 1st ed. Chapman and Hall/CRC. doi:[10.1201/b17279](https://doi.org/10.1201/b17279).
- Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society Series B* 58 (1):267–88.
- Yuan, M., and Y. Lin. 2006. "Model Selection and Estimation in Regression with Grouped Variables." *Journal of the Royal Statistical Society Series B* 68 (1):49–67.
- Zhang, C.-H. 2010. "Nearly Unbiased Variable Selection under Minimax Concave Penalty." *Annals of Statistics* 38 (2):894–942.
- Zhou, M., H. Wang, and Y. Tang. 2015. "Sequential Change Point Detection in Linear Quantile Regression Models." *Statistics and Probability Letters* 100:98–103.
- Zou, H. 2006. "The Adaptive Lasso and Its Oracle Properties." *Journal of the American Statistical Association* 101 (476):1418–28.
- Zou, H., and T. Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society* 67 (2):301–20.