A cumulative sum-based change-point detection method for high-dimensional data

Kento Egashira[a], **Kazuyoshi Yata**[b], **Makoto Aoshima**[b]

[a]**Department of Information Sciences, Tokyo University of Science**
[b]**Institute of Mathematics, University of Tsukuba**

1 Introduction

Change-point analysis can generally be classified into two broad approaches. The first deals with a given dataset to investigate the existence, number, and locations of change-points, while the second monitors data that are observed sequentially in order to detect a change-point as quickly as possible. The latter is commonly referred to as online change-point detection, whereas the former is known as offline change-point analysis. In this talk, we focus on the offline setting. Even within this framework, two distinct situations can be considered. In the first, the existence of change-points is assumed, and the objective is to estimate both their number and locations. In the second, the existence of change-points is itself uncertain, requiring both a statistical test for their presence and, if present, estimation of their number and locations. When a change-point is assumed to exist and there is exactly one change-point, the problem is referred to as single change-point detection. When multiple change-points are allowed, the problem is typically known as multiple change-point detection or time-series segmentation.

In recent years, the scope of change-point analysis has been extended to high-dimensional data, motivated by the increasing availability of large-scale observations in fields such as genomics, finance, neuroscience, and network analysis. Classical change-point detection methods are often designed under the assumption that the data dimension is fixed and relatively small compared to the sample size. However, such methods may fail or become unreliable when the number of variables is comparable to or even exceeds the number of observations. This high-dimensional regime introduces substantial challenges, including the need to handle strong dependence among variables, to ensure statistical consistency under sparsity or low-rank structures, and to maintain computational feasibility despite the explosive dimensionality. Accordingly, a growing body of research has been devoted to developing high-dimensional change-point detection methods that leverage modern statistical and computational tools such as principal component analysis (PCA), factor models, estimating covariance structure, and sparse regularization techniques. These approaches aim to detect structural changes not only in the mean vector but also in the covariance or latent factor structure, providing a more comprehensive understanding of change-point detection in high-dimensional settings.

A number of studies have been conducted on change-point detection in high-dimension, low-sample-size (HDLSS) settings. Among them, the cumulative sum (CUSUM) statistic,

a representative approach for detecting mean shifts, has been widely extended to the high-dimensional framework. For instance, Liu et al. [9] developed a CUSUM-type procedure based on U-statistics for high-dimensional data. Similarly, Yu and Chen [16] proposed a bootstrap-based CUSUM method tailored for HDLSS data to improve finite-sample performance. Li [7] proposed a nonparametric change-point detection procedure for high-dimensional data and established its asymptotic properties without imposing restrictive assumptions on the underlying population distribution. Dimension-reduction-based approaches have also attracted attention. Wang and Samworth [12] proposed a method that employs random projection to reduce dimensionality, assuming that mean changes occur only within a sparse subset of variables and under a normality assumption. Many of these CUSUM- and projection-based approaches derive their asymptotic properties under various sparsity conditions. For change-point detection using PCA, one may refer to Xiao et al. [13]. Relatedly, Yata and Aoshima [15] and Nakayama et al. [10] proposed clustering methods based on PCA that can be applied to change-point detection. More recently, Drikvandi and Modarres [2] introduced a change-point detection framework that first identifies candidate change-points and then tests whether these candidates are true change-points, providing theoretical properties when employing a unique distance function. In addition, Liu et al. [8] provided a comprehensive survey of high-dimensional change-point detection and conducted a comparative simulation studies across existing approaches.

Despite the extensive literature on high-dimensional change-point analysis, most existing methods rely on specific structural assumptions—such as sparsity, independence, or normality—that may not hold in practice. Meanwhile, in the broader development of high-dimensional statistical analysis, the importance of the eigenvalue structure of the covariance matrix has been increasingly recognized. As exemplified by Johnstone [6] and Paul [11], spiked covariance models have become a central framework in high-dimensional asymptotics, capturing scenarios where a few dominant eigenvalues represent major sources of variation. Empirically, Yata and Aoshima [14] reported that the leading eigenvalues of covariance matrices often grow as power functions of the dimension. In response to these findings, Aoshima and Yata [1] proposed a classification of high-dimensional covariance structures based on the contribution of the leading eigenvalues to introduce the strongly spiked eigenvalue (SSE) and the non-SSE (NSSE) model for high dimensional covariance matrix. For a d-dimensional positive definite covariance matrix $\mathbf{\Sigma}$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$, the SSE and NSSE model in Aoshima and Yata [1] are defined as the conditions

$$\liminf_{d\to\infty} \frac{\lambda_1^2}{\mathrm{tr}(\mathbf{\Sigma}^2)} > 0 \tag{1}$$

and

$$\frac{\lambda_1^2}{\mathrm{tr}(\mathbf{\Sigma}^2)} \to 0 \text{ as } d \to \infty, \tag{2}$$

respectively. This dichotomy provides a fundamental classification for analyzing high-dimensional data, since the covariance matrix and the inference procedures differ drastically between these two regimes. Most notably, in spite of the extensive literature on high-dimensional change-point analysis, many existing methods implicitly assume a covariance structure corresponding to the NSSE model. Among the various forms of SSE models, Ishii et al. [5] introduced a uni-SSE (USSE) model defined as

$$\frac{\sum_{i=2}^{d} \lambda_i^2}{\lambda_1^2} \to 0 \text{ as } d \to \infty. \tag{3}$$

This model assumes that a single leading eigenvalue dominates the covariance structure, while the contributions of the remaining eigenvalues become asymptotically negligible. Thus, the USSE model provides a mathematically tractable and conceptually clear framework within the broader class of SSE models. Accordingly, the USSE model serves as a practical and theoretically sound basis for developing statistical methodologies. Under the USSE model, Ishii et al. [3, 4] studied the estimation of eigenspaces.

Building upon these developments, it is natural to consider change-point analysis within the framework of the USSE model. In addition, since many real datasets fall within the scope of the SSE model, developing theoretical frameworks that explicitly account for the USSE model, a specific case of the SSE model, can be expected to enhance methodological performance. Consequently, the importance of developing change-point detection procedures under the USSE model has become increasingly evident.

In this talk, we introduce a high-dimensional multivariate CUSUM procedure presented in Liu et al. [8], and derive its asymptotic properties in the framework of HDLSS, including the SSE model. Based on the asymptotic properties obtained, we propose a modification. Furthermore, we introduce a change-point detection procedure that captures not only shifts in mean vectors but also changes in covariance structures.

2 Change-point detection

Suppose that there are two independent d-dimensional populations. Each population π_i is assumed to have an unknown mean vector $\boldsymbol{\mu}_i$ and an unknown positive-definite covariance matrix $\boldsymbol{\Sigma}_i$. We do not assume that the population distributions are normal. Suppose that n_1 independent observations $\{x_{11}, \ldots, x_{1n_1}\}$ are obtained from population π_1, and subsequently n_2 independent observations $\{x_{21}, \ldots, x_{2n_2}\}$ are obtained from population π_2. Let $n = n_1 + n_2$ and $n_{\min} = \min\{n_1, n_2\}$. For simplicity, we define $y_j = x_{1j}$ for $j \in \{1, \ldots, n_1\}$, and $y_{n_1+k} = x_{2k}$ for $k \in \{1, \ldots, n_2\}$.

The change-point detection problem can be formulated as the following hypothesis test:

$$H_0 : \pi_2 = \pi_1 \quad \text{v.s.} \quad H_1 : \pi_2 \neq \pi_1.$$

We conduct the test using the given dataset $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$. For a test statistics, we consider the multivariate CUSUM method shown in Liu et al. [8] for the given dataset $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$. For $k \in \{1, \ldots, n-1\}$, define

$$C(k) = \frac{k(n-k)}{n} \left\| \frac{1}{k} \sum_{i=1}^{k} \boldsymbol{y}_i - \frac{1}{n-k} \sum_{i=k+1}^{n} \boldsymbol{y}_i \right\|^2 .$$

The change-point location is estimated as

$$\hat{\tau}_C = 1 + \underset{1 \le k \le n-1}{\operatorname{argmax}} C(k),$$

where $C(k)$ measures the squared difference between the mean of the sample before and after time k.

For $i \in \{1, 2\}$, the eigen-decomposition of $\boldsymbol{\Sigma}_i$ is expressed as $\boldsymbol{\Sigma}_i = \boldsymbol{H}_i \boldsymbol{\Lambda}_i \boldsymbol{H}_i^T = \sum_{j=1}^{d} \lambda_{i(j)} \boldsymbol{h}_{i(j)} \boldsymbol{h}_{i(j)}^T$, where $\boldsymbol{\Lambda}_i = \operatorname{diag}(\lambda_{i(1)}, ..., \lambda_{i(d)})$ is a diagonal matrix of eigenvalues with $\lambda_{i(1)} \ge \cdots \ge \lambda_{i(d)} \ge 0$, and $\boldsymbol{H}_i = (\boldsymbol{h}_{i(1)}, ..., \boldsymbol{h}_{i(d)})$ is an orthogonal matrix whose columns are the corresponding eigenvectors. It should be noted that $\lambda_{i(1)}$ represents the largest eigenvalue of $\boldsymbol{\Sigma}_i$ for $i \in \{1, 2\}$.

We assume that

$$\limsup_{d \to \infty} \frac{\|\boldsymbol{\mu}_i\|^2}{d} < \infty, \quad \liminf_{d \to \infty} \frac{\operatorname{tr}(\boldsymbol{\Sigma}_i)}{d} > 0, \quad \text{and} \quad \limsup_{d \to \infty} \frac{\operatorname{tr}(\boldsymbol{\Sigma}_i)}{d} < \infty$$

for $i \in \{1, 2\}$. We assume that

$$\boldsymbol{y}_i = \boldsymbol{H}_1 \boldsymbol{\Lambda}_1^{1/2} \boldsymbol{z}_i + \boldsymbol{\mu}_1 \ \text{ for } i \in \{1, ..., n_1\}, \quad \boldsymbol{y}_i = \boldsymbol{H}_2 \boldsymbol{\Lambda}_2^{1/2} \boldsymbol{z}_i + \boldsymbol{\mu}_2 \ \text{ for } i \in \{n_1 + 1, ..., n\}$$

where $\boldsymbol{z}_i, \ i \in \{1, ..., n\}$, are i.i.d. random vectors having $\mathrm{E}[\boldsymbol{z}_i] = \boldsymbol{0}$ and $\mathrm{Var}[\boldsymbol{z}_i] = \boldsymbol{I}_d$. We introduce the following notations to state the theoretical results: Let $\Delta_\mu = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$ and $\Delta_\Sigma = |\operatorname{tr}(\boldsymbol{\Sigma}_1) - \operatorname{tr}(\boldsymbol{\Sigma}_2)|$. Additionally, let $L_1 = \mathrm{Var}[\|\boldsymbol{y}_j - \boldsymbol{\mu}_1\|^2]$ for $j \in \{1, ..., n_1\}$ and $L_2 = \mathrm{Var}[\|\boldsymbol{y}_j - \boldsymbol{\mu}_2\|^2]$ for $j \in \{n_1 + 1, ..., n\}$.

3 Asymptotic property

The following assumptions were considered.

(A-i): $nL_i/\Delta_\mu^2 \to 0$ and $n\operatorname{tr}(\boldsymbol{\Sigma}_i^2)/\Delta_\mu^2 \to 0$, $i \in \{1, 2\}$ as $d, n \to \infty$

Note that $L_i = 2\operatorname{tr}(\boldsymbol{\Sigma}_i^2)$ when Π_i is Gaussian.

Theorem 3.1. *Assume some regularity conditions, (A-i), and*

$$\limsup_{d \to \infty} \Delta_\Sigma / (n_{\min} \Delta_\mu) < 1. \tag{4}$$

Under H_1, $\hat{\tau}_C = n_1 + 1 + o_p(1)$ as $d \to \infty$, either when n is fixed or when $n \to \infty$.

In this talk, we modify $C(k)$ to have consistency without condition (4) and propose a test statistic based on the modification and investigate its asymptotic properties under both the null hypothesis H_0 and the alternative hypothesis H_1. In addition, we introduce a test statistic designed to detect not only changes in mean vectors but also changes in covariance matrix.

Acknowledgments
This research of the first author was partially supported by Grants-in-Aid for Early-Career Scientists, JSPS, under Contract Number 24K20748. This research of the second author was partially supported by a Grant-in-Aid for Scientific Research (C), JSPS, under Contract Number 22K03412. This research of the third author was partially supported by Grant-in-Aid for Scientific Research (A), JSPS, under Contract Number 25H01107.

References

[1] Aoshima, M., Yata, K. (2018). Two-sample tests for high-dimension, strongly spiked eigenvalue models. *Statistica Sinica*, 28, 43–62.

[2] Drikvandi, R., Modarres, R. (2024). A distribution-free method for change point detection in non-sparse high dimensional data. *Journal of Computational and Graphical Statistics*, 34, 290–305.

[3] Ishii, A., Yata, K., Aoshima, M. (2016). Asymptotic properties of the first principal component and equality tests of covariance matrices in high-dimension, low-sample-size context. *Journal of Statistical Planning and Inference*, 170, 186–199.

[4] Ishii, A., Yata, K., Aoshima, M. (2019). Equality tests of high-dimensional covariance matrices under the strongly spiked eigenvalue model. *Journal of Statistical Planning and Inference*, 202, 99–111.

[5] Ishii, A., Yata, K. Aoshima, M. (2021). Hypothesis tests for high-dimensional covariance structures. *Annals of the Institute of Statistical Mathematics*, 73, 599–622.

[6] Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 29, 295–327.

[7] Li, J. (2020). Asymptotic distribution-free change-point detection based on interpoint distances for high-dimensional data. *Journal of Nonparametric Statistics* 32. 157–184

[8] Liu, B., Zhang, X., Liu, Y. (2022). High dimensional change point inference: recent developments and extensions. *Journal of Multivariate Analysis*, 188, 104833.

[9] Liu, B., Zhou, C., Zhang, X., Liu, Y. (2020). A unified data-adaptive framework for high dimensional change point detection. *Journal of the Royal Statistical Society*, Series B, **82**, 933–963.

[10] Nakayama, Y., Yata, K., Aoshima, M. (2021). Clustering by principal component analysis with Gaussian kernel in high-dimension, low-sample-size settings. *Journal of Multivariate Analysis*, 185, 104779.

[11] Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17, 1617–1642.

[12] Wang, T., Samworth, R. J. (2018). High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society*, Series B, 80, 57–83.

[13] Xiao, W., Huang, X., He, F., Silva, J., Emrani, S., Chaudhuri, A. (2019). Online robust principal component analysis with change point detection *IEEE Transactions on Multimedia*, 22, 59–68.

[14] Yata, K., Aoshima, M. (2013). PCA consistency for the power spiked model in high-dimensional settings. *Journal of Multivariate Analysis*, 122, 334–354.

[15] Yata, K., Aoshima, M. (2020). Geometric consistency of principal component scores for high-dimensional mixture models and its application. *Scandinavian Journal of Statistics*, 47, 899–921.

[16] Yu, M., Chen, X. (2021). Finite sample change point inference and identification for high-dimensional mean vectors. *Journal of the Royal Statistical Society*, Series B, 83, 247–270.