Detecting Change Points in Covariance Matrix from High Dimensional Sequence Data

Young Hyun Cho[1] · **Seonghun Cho** [1] · **Taehyun Koo**[2] · **Johan Lim**[1]

[1]Department of Statistics, Seoul National University

[2]Department of Statistics, Rutgers University

Introduction

This research proposes a method to test and estimate change points in the covariance structure of high-dimensional multivariate series data. Our method uses the trace of the beta matrix, known as the Pillai's statistics, at each time point. We prove the asymptotic normality of the Pillai's statistics for testing the equality of two covariance matrices when both n (sample size) and p (dimension) increase at the same rate. We compute the Pillai's statistics and its p-value for each time point. We then test the existence of single change point by combining individual p-values by using Cauchy combination test by Yaowu Lin and Jun Xie (2018) and estimate the change point as the point whose statistic is the greatest. To test and estimate multiple change points, we use the idea of the wild binary segmentation by Fryzlewicz (2014). We apply the above procedure to each segmented series until no significant change point exists. We numerically provide the size and power of our method. We finally apply our procedure to finding abnormal behavior in the investment of a private equity.

Methods

Assumption

For a sequence of p-dimensional vector valued data $\mathbf{y}_1, \ldots, \mathbf{y}_T \sim (\mathbf{0}, \Sigma_i)$, we consider the null hypothesis

$$\mathbf{H}_0 : \Sigma_1 = \Sigma_2 = \cdots = \Sigma_T$$

To find a change point(CP) k with $[\epsilon \cdot T] + 1 \leq k \leq [(1-\epsilon)T]$ for some $\epsilon > 0$, consider the following alternative hypothesis

$$\mathbf{H}_1 : \Sigma_1 = \cdots = \Sigma_{k-1} = \Sigma_k \neq \Sigma_{k+1} = \cdots = \Sigma_T \text{ for some } k$$

Model

Asymptotic normality

- Let $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_T]^\top$
- For each $k = [\epsilon \cdot T] + 1, \ldots, [(1-\epsilon) \cdot T]$, with predetermined window length wl, let

$$Y_{k,1} = [\mathbf{y}_{a_k}, \ldots, \mathbf{y}_k]^\top \text{ and } Y_{k,2} = [\mathbf{y}_{k+1}, \ldots, \mathbf{y}_{b_k}]^\top,$$

where $a_k = \max(k - wl + 1, 1)$, $b_k = \min(k + wl, T)$

- Compute the Beta-matrix $B_k = n_{k,1}S_{k,1}(n_{k,1}S_{k,1} + n_{k,2}S_{k,2})^{-1}$
  - $n_{k,1}$, $n_{k,2}$: the number of columns of $Y_{k,1}$, $Y_{k,2}$
  - $S_{k,1}$, $S_{k,2}$: covariance matrices of $Y_{k,1}$, $Y_{k,2}$
- Statistic $\mathcal{K}_k$ is defined by

$$\mathcal{K}_k = \frac{\sum \lambda_i^{\mathbf{B}_k} - pl_k - \mu_k}{\sigma_k},$$

  - $\lambda_i^{\mathbf{B}_k}$ denotes the i-th smallest eigenvalue of $\mathbf{B}_k$
  - $l_k = \frac{h^2\delta_{y_2>1} + y_2^2\delta_{y_2<1}}{y_2(y_1+y_2)}$, $\mu_k = -\frac{\Delta_1 h^2 y_1^2 y_2^2}{(y_1+y_2)^4} + \frac{\Delta_2 h^2 y_1^2 y_2^2}{(y_1+y_2)^4}$
  - $\sigma_k = \frac{2h^2 y_1^2 y_2^2}{(y_1+y_2)^4} + (\Delta_1 y_1 + \Delta_2 y_2)\frac{h^4 y_1^2 y_2^2}{(y_1+y_2)^6}$
  - $y_1 = \frac{p}{n_{k,1}}$, $y_2 = \frac{p}{n_{k,2}}$ and Δ_i stands for skewness
- Koo(2019) proved asymptotic normality under suitable conditions on moments and dimensionality,

$$\mathcal{K}_k \xrightarrow{D} N(0,1)$$

Test statistic and p-value

- The proposed statistic for detecting CP is based on the series of Beta-matrices as

$$k^* = \underset{[\epsilon \cdot n]+1, \ldots, [(1-\epsilon)\cdot n]}{\operatorname{argmax}} \mathcal{K}_k, \text{ and } T_{max} = \mathcal{K}_{k^*}$$

- For the observed value t_{max} of $\mathcal{T}_{max}$, the p-value for testing the existence of the CP is bounded by

$$P(\mathcal{T}_{max} > t_{max}) \leq \sum_{k=[\epsilon \cdot T]+1}^{[(1-\epsilon)\cdot T]} P(\mathcal{K}_k > t_{max}).$$

- However, Bonferroni type bound tends to be too conservative so we adopted Cauchy combination test by Yaowu Liu and Jun Xie (2018):

$$T = \sum_k \omega_k \tan\{(0.5 - P(\mathcal{K}_k > t_{max}))\pi\}$$

- The p-value is approximated by

$$p - \text{value} = \frac{1}{2} - (\arctan t_0)/\pi,$$

where t_0 is the observed value of T.

Multiple points detection

- If an estimated change point from the entire data is statistically significant, split the data into two sub-data with the estimated point
- Detection is applied on both sub-data, which possibly result in further splits.
- The recursion on a given segment continues until there is no significant change point

Results

Simulation study: Single point detection

- We numerically provide powers and empirical size of the proposed test, while comparing to the existing method: Ian Barnett, Jukka-Pekka Onnela(2016)

Data

- For $T \in \{200, 500, 800, 1000\}$, prefix a change point at $k^* = \frac{T}{2} + 1$ and generated 1000 data sets as:
  - Case1: $\mathbf{y}_j \sim \mathbf{MVN}(\mathbf{0}, \Sigma_{0.5})$, for all $1 \leq j \leq T$
  - Case2: $\mathbf{y}_j \sim \mathbf{MVN}(\mathbf{0}, \Sigma_{0.4})$, for all $1 \leq j \leq k^*$, $\mathbf{y}_j \sim \mathbf{MVN}(\mathbf{0}, \Sigma_{0.6})$, for all $k^* + 1 \leq j \leq T$
- Powers and sizes are evaluated by counting the number of rejection under significant level 0.05.

Result

Table 1: Single detection results

Data length	Size			Power		
	Proposed method		Simulation method	Proposed method		Simulation method
	wl=30	wl=50		wl=30	wl=50	
200	0.129	0.067	0.035	0.870	**0.998**	0.274
500	0.062	**0.051**	**0.050**	0.705	**0.991**	0.646
800	**0.050**	**0.055**	**0.051**	0.699	**0.983**	0.84
1000	0.0630	**0.045**	**0.048**	0.630	**0.983**	0.911

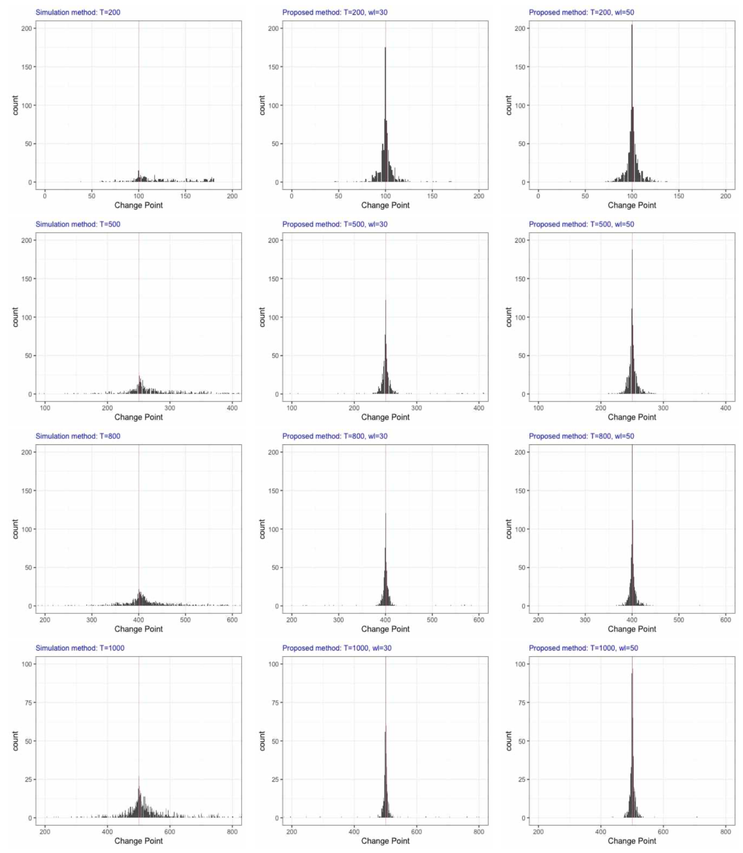

Figure 1: Simulation results for single point detection

Simulation study: Multiple points detection

Data

- Generate 1000 data sets as:

$$\mathbf{y}_j \sim \begin{cases} \mathbf{MVN}(\mathbf{0}, \Sigma_{0.3}) & \text{for all } 1 \leq j \leq k_1^* \\ \mathbf{MVN}(\mathbf{0}, \Sigma_{0.7}) & \text{for all } k_1^* + 1 \leq j \leq k_2^* \\ \mathbf{MVN}(\mathbf{0}, \Sigma_{0.5}) & \text{for all } k_2^* + 1 \leq j \leq T \end{cases}$$

,where $T \in \{200, 500, 800, 1000\}$, $k_1^* = \left\lfloor \frac{T}{3} \right\rfloor$ and $k_2^* = \left\lfloor \frac{2T}{3} \right\rfloor$

Results

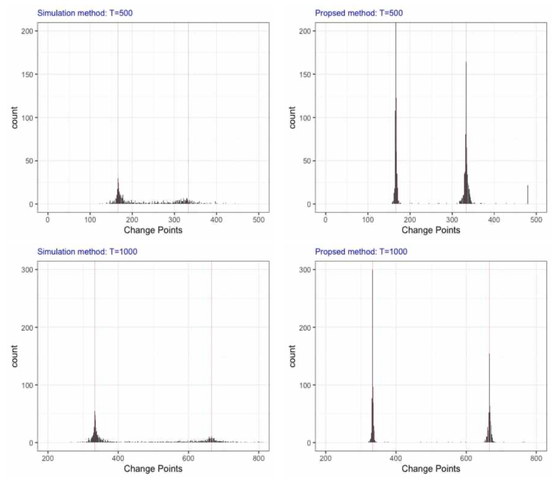

Figure 2: Simulation results for multiple points detection

Real data analysis

Data

- Daily records of the holdings of stocks in Korean stock market of a private fund from 2013-01-02 to 2018-01-25, $n = 1252$ working days
- We group the stocks by 11 sectors, followed by the Global Industry Classification Standard

Result

Table 2: Results of real data analysis

Window Length	CP_1	CP_2	CP_3
25	**904**	233	**1120**
45	**878**	468	**1055**
90	**870**	536	**1152**

Conclusions

- Taking wl at least larger than p is recommendable, and one might try with various level of window length.
- If window length is chosen, we recommend to drop the data as much as the window length
- According to our method, the fund managers seem to have changed the investment strategy, which deviates their initial public offerings