High-Dimensional Changepoint Detection via a Geometrically Inspired Mapping

Thomas Grundy, Rebecca Killick, and Gueorgui Mihaylov
Presenter: JaeHoon Kim

December 2, 2024

Outline

- Introduction
- Methodologies
    - Geometric Mapping
    - Analyzing Mapped Time Series
    - GeomCP Algorithm
    - Non-Normal and Dependent Data

Problem Setting

- Suppose we observe time vectors $Y_1, \ldots, Y_n$ independently from p-dimensional Gaussian distributions with diagonal covariance:

$$Y_i \overset{\text{ind.}}{\sim} N_p(\mu_i, \sigma_i^2 I_p), \quad i = 1, \ldots, n.$$

- We define m change points $\tau_{1:m} = \{\tau_1, \ldots, \tau_m\}$ with $0 = \tau_0 < \tau_1 < \cdots < \tau_m < \tau_{m+1} = n$, such that:

$$(\mu_{\tau_k}, \sigma^2_{\tau_k}) = \cdots = (\mu_{\tau_{k+1}-1}, \sigma^2_{\tau_{k+1}-1}),$$
$$(\mu_{\tau_k}, \sigma^2_{\tau_k}) \neq (\mu_{\tau_{k+1}}, \sigma^2_{\tau_{k+1}}), \quad k = 0, \ldots, m.$$

Geometric mapping

- We aim to detect changepoints in the mean and variance vectors utilizing geometric properties that capture these changes.
  - Mean changes can be detected by observing variations in distances.
  - Variance changes can be identified by tracking changes in angles.
- Using this mapping, we can transform a *p*-dimensional time series into a two-dimensional representation.
- A pre-specified reference vector is required to calculate distances and angles.

Geometric mapping

- We propose a data-driven reference vector as follows:
  1. Set the reference vector $y_0 = \mathbf{1}$.
  2. Translate all points based on this vector:

$$y'_{i,j} = y_{i,j} - \left(\min_i y_{i,j} - y_{0,j}\right), \quad i \in [1, \dots, n], \quad j \in [1, \dots, p].$$

- This approach has several desirable properties:
  - It bounds the angle measure between 0 and $\pi/4$.
  - It ensures that changes in individual series are reflected in the angle measure.
  - It does not affect the distance measure.

Geometric mapping

- The distance and angle measures are defined using the standard scalar product.
- To compute the distance measure, d_i, we apply the mapping $\delta : \mathbb{R}^p \to \mathbb{R}_{>0}$:

$$d_i = \delta(\mathbf{y_i}) = \sqrt{\langle (\mathbf{y_i'} - \mathbf{1}), (\mathbf{y_i'} - \mathbf{1}) \rangle},$$

  which is equivalent to $\|\mathbf{y_i'} - \mathbf{1}_p\|_2$.
- To compute the angle measure, a_i, we use the mapping $\alpha : \mathbb{R}^p \to \left[0, \frac{\pi}{4}\right]$:

$$a_i = \alpha(\mathbf{y_i}) = \cos^{-1}\left(\frac{\langle \mathbf{y_i'}, \mathbf{1} \rangle}{\sqrt{\langle \mathbf{y_i'}, \mathbf{y_i'} \rangle}\sqrt{\langle \mathbf{1}, \mathbf{1} \rangle}}\right),$$

  which represents the principal angle between $\mathbf{y_i'}$ and $\mathbf{1}$.

Analyzing Mapped Time Series

Theorem 2.3.1

Suppose we have independent random variables, $Y_i \sim N(\mu_i, \sigma_i^2)$. Let

$$X = \sqrt{\sum_{i=1}^{p} Y_i^2},$$

then as $p \to \infty$,

$$\frac{X - \sqrt{\sum_{i=1}^{p}(\mu_i^2 + \sigma_i^2)}}{\sqrt{\frac{2\sum_{i=1}^{p}(\mu_i\sigma_i)^2 + \sum_{i=1}^{p}\sigma_i^4 + 2\rho\sqrt{\sum_{i=1}^{p}\sum_{j=1}^{p}\mu_i^2\sigma_i^2\sigma_j^4}}{2\sum_{i=1}^{p}(\mu_i^2+\sigma_i^2)}}} \xrightarrow{\mathcal{D}} N(0,1),$$

where ρ is an unknown correlation parameter.

Analyzing Mapped Time Series

- In the literature, it is commonly assumed that angles also follow a Normal distribution, as shown in Fearnhead et al. (2018).
- To detect changepoints in the mapped series, we use the PELT algorithm by Killick et al. (2012).
- When the Normal approximation holds for distance and angle measures, we use the Normal likelihood as the test statistic.
- If the Normal approximation is unsuitable, we recommend a non-parametric test statistic, such as the empirical distribution from Zou et al. (2014).

GeomCP Algorithm

Algorithm GeomCP

Require: $\mathbf{Y} \in \mathbb{R}^{n \times p}$, threshold ξ, *Univariate Cpt Method*.

Step 1: Centralize data by $y'_{i,j} = y_{i,j} - (\min_i y_{i,j} - 1)$.

Step 2: Perform distance mapping: $y_i \xrightarrow{\delta} d_i, \ \forall i$.

Step 3: Perform *Cpt Method* on d to recover changepoints, $\hat{\tau}^{(d)}$.

Step 4: Perform angle mapping: $y_i \xrightarrow{\alpha} a_i, \ \forall i$.

Step 5: Perform *Cpt Method* on a to recover changepoints, $\hat{\tau}^{(a)}$.

Step 6: $\forall k$, if $\min \left| \hat{\tau}^{(a)} - \hat{\tau}_k^{(d)} \right| < \xi$, then remove $\hat{\tau}_k^{(d)}$ from $\hat{\tau}^{(d)}$.

return $\hat{\tau} = \mathrm{sort}(\hat{\tau}^{(a)}, \hat{\tau}^{(d)})$.

Non-Normal and Dependent Data

- We may allow for an arbitrary covariance matrix:
  - $Y_{\text{pre}} \overset{\text{ind.}}{\sim} N_p(0, \Sigma)$
  - $Y_{\text{post}} \overset{\text{ind.}}{\sim} N_p(0, \sigma\Sigma)$
- We still expect the angles between the time vectors and the reference vector to change, indicating changes in covariance.
- Alternatively, other inner products, such as the Mahalanobis Distance (Mahalanobis, 1936), could be considered in the distance and angle mappings.

Non-Normal and Dependent Data

- If the data originates from a non-Normal distribution, changes in the first and second moments would still likely appear in the distance and angle mappings.
- However, understanding the distribution of the mapped series becomes more challenging.
- Allowing for temporal dependence between time points introduces temporal dependence in the mapped series as well.
- Understanding how temporal dependence in the multivariate series transfers to the mapped series is non-trivial.

References I

Fearnhead, P., Maidstone, R., and Letchford, A. (2018). Detecting changes in slope with an l0 penalty. *Journal of Computational and Graphical Statistics*, pages 1–11.

Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *National Institute of Science of India*.

Zou, C., Yin, G., Feng, L., and Wang, Z. (2014). Nonparametric maximum likelihood approach to multiple change-point problems. *The Annals of Statistics*, 42(3):970–1002.