Statistics and Computing (2020) 30:1155–1166
https://doi.org/10.1007/s11222-020-09940-y

High-dimensional changepoint detection via a geometrically inspired mapping

Thomas Grundy[1] · **Rebecca Killick**[2] · **Gueorgui Mihaylov**[3]

Received: 25 September 2019 / Accepted: 13 March 2020 / Published online: 28 March 2020

Abstract
High-dimensional changepoint analysis is a growing area of research and has applications in a wide range of fields. The aim is to accurately and efficiently detect changepoints in time series data when both the number of time points and dimensions grow large. Existing methods typically aggregate or project the data to a smaller number of dimensions, usually one. We present a high-dimensional changepoint detection method that takes inspiration from geometry to map a high-dimensional time series to two dimensions. We show theoretically and through simulation that if the input series is Gaussian, then the mappings preserve the Gaussianity of the data. Applying univariate changepoint detection methods to both mapped series allows the detection of changepoints that correspond to changes in the mean and variance of the original time series. We demonstrate that this approach outperforms the current state-of-the-art multivariate changepoint methods in terms of accuracy of detected changepoints and computational efficiency. We conclude with applications from genetics and finance.

Keywords Changepoint · Time series · High-dimensional · PELT

1 Introduction

Time series data often have abrupt structural changes occurring at certain time points, known as changepoints. To appropriately analyze, model or forecast time series data that contain changes, we need to be able to accurately detect where changepoints occur. High-dimensional changepoint analysis aims to accurately and efficiently detect the location of changepoints as both the number of dimensions and time points increase. High-dimensional changepoint analysis is an ever-growing research area and has multiple applications including finance and economics (Modisett and Maboudou-Tchao 2010); longitudinal studies (Terrera et al. 2011); and genetics (Bleakley and Vert 2011).

Changepoint analysis in the univariate setting is a well-studied area of research with early work by Page (1954) and overviews can be found in Eckley et al. (2011) and Brodsky and Darkhovsky (2013). The multivariate extension has received less attention, see Truong et al. (2020) for a recent review. One major challenge with high-dimensional changepoint analysis is the computational burden of an increasing number of dimensions. To partially reduce this computational burden, a common assumption is that changepoints are assumed to occur in all series simultaneously (Maboudou-Tchao and Hawkins 2013), a sparse set of series (Wang and Samworth 2018); or a dense set of series (Zhang et al. 2010). Within these settings, a common approach is to first project the time series to a single dimension and then use a univariate changepoint method on the projected time series. For example, Zhang et al. (2010), Horváth and Hušková (2012) and Enikeeva and Harchaoui (2019) consider an l_2-aggregation of the CUSUM statistic, while Jirak (2015) considers an l_∞-aggregation that works well for sparse changepoints. A recent advancement was the Inspect method

Grundy is grateful for the support of the Engineering and Physical Sciences Research Council (Grant Number EP/L015692/1). The authors also acknowledge Royal Mail Group Ltd for financial support and are grateful to Jeremy Bradley in Royal Mail GBI Data Science for helpful discussions.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11222-020-09940-y) contains supplementary material, which is available to authorized users.

✉ Thomas Grundy
t.grundy1@lancaster.ac.uk

[1] STOR-i Centre for Doctoral Training, Lancaster University, Lancaster, UK

[2] Department of Mathematics and Statistics, Lancaster University, Lancaster, UK

[3] Royal Mail GBI Data Science, London, UK

proposed by Wang and Samworth (2018) who aim to find an optimal projection direction of the CUSUM statistic to maximize a change in mean.

Current projection methods are generally limited to detecting changes in a single parameter, usually the mean. Therefore, these methods cannot be used in many practical scenarios where multiple features of the time series change. An alternative, nonparametric approach was taken in Matteson and James (2014) where U-Statistics were used to segment the time series. As this is a nonparametric method, it can detect different types of changes in distribution but becomes computationally infeasible as the number of time points increases. The methods above almost exclusively use a Binary Segmentation approach (Scott and Knott 1974; Vostrikova 1981), or derivations thereof (Fryzlewicz 2014), to detect multiple changepoints. This can lead to poor detection rates as conditional identification of changes can lead to missing or poor placement of changepoints due to factors such as masking. This occurs when a large change is masked by two smaller changes on either side acting in opposite directions; this idea is explained further in Fryzlewicz (2014).

A key novelty in this paper is to map a given high-dimensional time series onto two dimensions instead of one. Inspired by a geometric representation of data, we map each high-dimensional time vector to its distance and angle from a fixed pre-defined reference vector based upon the standard scalar product. These mappings show shift and shape changes in the original data corresponding to mean and variance changes. Given the geometric inspiration, we denote the method GeomCP throughout.

In Sect. 2, we set up the high-dimensional changepoint problem before defining the geometric mappings used in GeomCP. Also, we discuss an alternative approach to Binary Segmentation that can be applied to the univariate mapped series. An extensive simulation study is performed in Sect. 3, which compares GeomCP to competing available multivariate changepoint methods, Inspect (Wang and Samworth 2018) and E-Divisive (Matteson and James 2014). Section 4 presents two applications from genetics and finance. Section 5 gives concluding remarks.

2 Methodology

In this section, we set up the high-dimensional changepoint problem for our scenario. We define our new method, GeomCP, and discuss how changes in high-dimensional time series manifest themselves in the mapped time series. We then suggest an appropriate univariate changepoint detection method for detecting changes in the mapped time series—although practically others could be used.

Before proceeding, we define some notation used throughout the paper. We define the $\mathbb{1}_p$ vector as a p-dimensional vector where each entry is 1 and the number of dimensions, p, is inferred from context. For a vector, $\boldsymbol{y} = (y_1, \ldots, y_p)^T$, we define the l_q-norm as $\|\boldsymbol{y}\|_q := \left(\sum_{j=1}^{p} |y_j|^q\right)^{\frac{1}{q}}$ for $q \in [1, \infty)$. We define $\langle \cdot, \cdot \rangle$ as the standard scalar product such that for vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ we have $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{j=1}^{p} x_j y_j$. Finally, the terms variables, series and dimensions shall be used interchangeably to indicate the multivariate nature of the problem.

2.1 Problem setup

We study the time series model where $\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n}$ are independent, p-dimensional time vectors that follow a multivariate Normal distribution where

$$\boldsymbol{Y_i} \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2 \boldsymbol{I}_p), \qquad 1 \leq i \leq n \ .$$

We assume there are an unknown number of changepoints, m, which occur at locations $\tau_{1:m} = (\tau_1, \ldots, \tau_m)$. These changepoints split the data into $m + 1$ segments, indexed k, that contain piecewise constant mean and variance vectors, $\boldsymbol{\mu}_k$ and $\boldsymbol{\sigma}_k^2$. Note we assume a diagonal covariance matrix, so the covariance matrix can be described by the variance vector and the identity matrix. We define $\tau_0 = 0$ and $\tau_{m+1} = n$ and assume the changepoints are ordered so, $\tau_0 = 0 < \tau_1 < \ldots < \tau_m < \tau_{m+1} = n$.

The following section introduces the geometric intuition and mappings used within GeomCP. These mappings reduce the dimension of the problem to make the problem computationally feasible as n and p grow large.

2.2 Geometric mapping

When analyzing multivariate time series from a geometric viewpoint, we seek to exploit relevant geometric structures defined in the multi-dimensional space. Here, we aim to detect changepoints in the mean and variance vectors of multivariate Normal random variables; therefore, we wish to utilize geometric properties that capture these changes.

A change in the mean vector of our data generating process will cause a location shift of the data points in the multi-dimensional space. Consider a distance between each data point and some fixed reference point, if the data points are shifted in the multi-dimensional space, then their distance to the reference point would be expected to change. Hence, we can detect when the mean vector of the data generating process changes by observing a change in the distances. For a change in distance not to occur after an underlying mean change, the new mean vector must remain exactly on the same $(p - 1)$-sphere (centered in the reference point) that the old mean vector lay on. Given that the computation of the mean vector is a linear operator on the multivariate time

series, the requirement to lie on the same sphere (a quadric in $\mathbb{R}^p$) is highly non-generic from a geometric prospective. As a result, these scenarios are rare especially in high dimensions.

A change in the covariance of our data generating process will cause a change in the shape of the data points. More specifically in our setup, a change in the variance would cause the shape of the data points to expand or contract. Consider the angle between each data vector and a reference vector; as the shape of the data points expands (contracts), the angles will become more (less) varied. Hence, we can detect changes in the variance of the data generating process by detecting changes in the angles.

By using distances and angles, we can map a p-dimensional time series to two dimensions. To calculate these mappings, we need a pre-specified reference vector to calculate a distance and angle from. Naturally, one may think to use the mean of the data points. However, this requires a rolling window to estimate the mean of data points prior to the point being mapped. Not only does this introduce tuning parameters, such as the size of the rolling window, but will result in spikes in the distance and angle measures at changepoints. To detect changepoints, we would need a threshold for these spikes and calculating such a threshold is a non-trivial task; hence, we seek an alternative.

We propose setting the reference vector to be a fixed vector, $\boldsymbol{y_0}$. We then translate all the points based upon this fixed reference vector,

$$y'_{i,j} = y_{i,j} - (\min_i y_{i,j} - y_{0,j}),$$
$$i \in [1, \ldots, n]\ ,\ \ j \in [1, \ldots, p]\ . \tag{2.1}$$

This results in a data-driven reference vector. We choose to set $\boldsymbol{y_0} = \mathbb{1}$ as this bounds the angle measure between 0 and $\pi/4$, meaning we do not get vectors close to the origin facing in opposite directions causing non-standard behavior within a segment. Moreover, having a nonzero element in every entry of $\boldsymbol{y_0}$ ensures changes in the individual series will manifest in the angle measure. Note due to the translation in (2.1), the choice of $\boldsymbol{y_0}$ does not affect the distance measure. Throughout we assume the reference vector is set as $\boldsymbol{y_0} = \mathbb{1}$.

For data points in the same segment, we would expect their distances and angles to the reference vector to have the same distribution. When a mean (variance) change occurs in the data, this leads to a shift (spread) in the data; hence, the distances (angles) will change. Therefore, by detecting changes in the distances and angles, using an appropriate univariate changepoint method, we recover changepoints in the p-dimensional series.

We define our distance and angle measures based upon the standard scalar product. To obtain our distance measure, d_i, we perform a mapping, $\delta : \mathbb{R}^p \to \mathbb{R}^1_{>0}$,

$$d_i = \delta(\boldsymbol{y_i}) = \sqrt{\langle (\boldsymbol{y'_i} - \mathbb{1}), (\boldsymbol{y'_i} - \mathbb{1}) \rangle}\ , \tag{2.2}$$

which is equivalent to $\left\| \boldsymbol{y'_i} - \mathbb{1}_p \right\|_2$.

To obtain our angle measure, a_i, we perform a mapping $\alpha : \mathbb{R}^p \to [0, \frac{\pi}{4}]$,

$$a_i = \alpha(\boldsymbol{y_i}) = \cos^{-1}\left(\frac{\langle \boldsymbol{y_i}', \mathbb{1} \rangle}{\sqrt{\langle \boldsymbol{y_i}', \boldsymbol{y'_i} \rangle} \sqrt{\langle \mathbb{1}, \mathbb{1} \rangle}} \right)\ , \tag{2.3}$$

which is the principal angle between $\boldsymbol{y'_i}$ and $\mathbb{1}$.

By using the standard scalar product, we are incorporating information from each series in the distance and angle measures. As such, we would expect GeomCP to perform well in scenarios where a dense set of the series change at each changepoint. This idea will be explored further and verified in Sect. 3.

2.3 Analyzing mapped time series

Understanding the distributional form of the distance and angle mappings will aid in the choice of univariate changepoint methods. Under our problem setup, Theorem 1 shows that the distance measure, asymptotically in p, follows a Normal distribution.

Theorem 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 $p \to \infty$,*

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

where ρ is an unknown correlation parameter (see proof).

Proof See the Supplementary Material. □

Theorem 1 shows that, asymptotically in p, the distance between each time vector and a pre-specified fixed vector follows a Normal distribution. Hence, for piecewise constant time vectors, the resulting distance measure will follow a piecewise constant Normal distribution. It is common in the literature to assume that angles also follow a Normal distribution, as in Fearnhead et al. (2018). We found by simulation, for large enough p, the angle measure defined in (2.3) is well approximated by a Normal distribution with piecewise constant mean and variance.

While any theoretically valid univariate method could be used to detect changepoints in the mapped series, we use the PELT algorithm of Killick et al. (2012) as this is an exact

and computationally efficient search. For $n \to \infty$, PELT is consistent in detecting the number and location of changes in mean and variance (Tickle et al. 2019; Fisch et al. 2018); hence, using Theorem 1, we gain consistency of our distance measure as $p \to \infty$ also. When the Normal approximation of the distance and angle measures holds, we use the Normal likelihood as our test statistic within PELT and allow for changes in mean and variance. If p is small, we may not want to make the Normal assumption. In this case, we recommend using a nonparametric test statistic, such as the empirical distribution from Zou et al. (2014) (where consistency has also been shown) as embedded within PELT in Haynes et al. (2017b).

2.4 GeomCP algorithm

Algorithm 1 details the pseudo-code for GeomCP. As changepoints can manifest in both the distance and angle measure, we post-process the two sets of changepoints to obtain the final set of changes. We introduce a threshold, ξ, and say that a changepoint in the distance measure, $\hat{\tau}^{(d)}$, and a changepoint in the angle measure, $\hat{\tau}^{(a)}$, are deemed the same if $\left|\hat{\tau}^{(d)} - \hat{\tau}^{(a)}\right| \leq \xi$. If we determine two changepoints to be the same, we set the changepoint location to be the one given by the angle measure as Sect. 3.2 demonstrates, this results in more accurate changepoint locations. The choice of ξ should be set based upon the minimum distance expected between changepoints. Alternatively, ξ could be set to zero and then an alternative post-processing step would be required to determine whether similar changepoint estimates correspond to the same change.

Algorithm 1 GeomCP

Input: $\boldsymbol{Y} \in \mathbb{R}^{n \times p}$, threshold $= \xi$, *Univariate Cpt Method*.

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

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

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

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

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

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

Return: $\hat{\boldsymbol{\tau}} = \text{sort}(\hat{\boldsymbol{\tau}}^{(a)}, \hat{\boldsymbol{\tau}}^{(d)})$

One of the major downfalls of many multivariate changepoint methods is they are computationally infeasible for large n and p. Within GeomCP, the computational cost to calculate both the distance and angle measures in (2.2) and (2.3) is $\mathcal{O}(np)$. If we implement the PELT algorithm for our univariate changepoint detection, this has expected computational cost $\mathcal{O}(n)$ under certain conditions. The main condition requires the number of changepoints to increase linearly with the number of time points, and further details are given in Killick et al. (2012). If these conditions are not satisfied, PELT has an at worst computational cost of $\mathcal{O}(n^2)$. Hence, the expected computational cost of GeomCP is $\mathcal{O}(np+n) = \mathcal{O}(np)$ (under the conditions in Killick et al. (2012)) and has at worst computational cost $\mathcal{O}\left(np + n^2\right) = \mathcal{O}\left(n(p+n)\right)$.

2.5 Non-Normal and dependent data

The current problem setup assumes multivariate Normal distributed data with a diagonal covariance matrix. These assumptions are made to facilitate our theoretical analysis and result in the Normality of the mapped series. If these assumptions are broken, the geometric intuition described in Sect. 2.2 still holds, but we can say less about the theoretical properties of the mapped series.

Firstly, if we allow for an arbitrary covariance matrix, this describes the shape and spread of the data points. Suppose our data undergoes a change from $X_{\text{pre}} \sim N(\mathbf{0}, \Sigma)$ to $X_{\text{post}} \sim (\mathbf{0}, \boldsymbol{\sigma}\Sigma)$ this will cause the data points to spread out in the directions of the principal components. Hence, we would still expect the angles between the time vectors and the reference vector to change, revealing the change in covariance. We investigate this further in Sect. 3.5. In fact, a Normal distributed data set with a known covariance matrix could be transformed into a Normal distributed data set with a diagonal covariance matrix (satisfying our initial problem setup) by an orthogonal transformation that aligns the axes with the principal components. Such a transformation would preserve the distances and angles by definition but requires knowledge of the true covariance structure.

Alternatively, we could consider other inner products in our distance and angle mappings defined in (2.2) and (2.3); here the geometric motivation of the method would remain valid. In this case, for an underlying mean change to occur without the distance measure changing, the new mean vector must remain exactly on the more general $(p-1)$-quadric in $\mathbb{R}^p$. This is still a highly non-generic requirement from a geometric prospective. In particular, we could use scalar products directly derived from the covariance matrix, such as the Mahalanobis Distance (Mahalanobis 1936). In such cases, the direct relation between angles and the correlation coefficients is well known (Wickens 1995). However, such inner products require an estimate of the covariance in each segment, which is non-trivial and therefore left as future work.

If we allow the data to be distributed from a non-Normal distribution, then we would expect changes to the first and second moment of these distributions to still manifest in the distance and angle mappings. However, being able to understand the distribution of the mapped series would be more challenging. In practice, the empirical cost function could be used within PELT (Haynes et al. 2017b), yet this would lead

to less power in the detection of changes in the univariate series.

Finally, if we allowed temporal dependence between the time points, this would lead to temporal dependence in the mapped series and an appropriate, cost function for PELT could be used. Understanding how the temporal dependence in the multivariate series manifests in the mapped series is non-trivial and is left as further work.

In the next section, we provide an extensive simulation study exploring the effectiveness of GeomCP at detecting multivariate changes in mean and variance and demonstrate an improved detection rate on current state-of-the-art multivariate changepoint methods. Furthermore, we illustrate the improved computational speed of GeomCP over current methods, especially as n and p grow large.

3 Simulation study

In this section, we provide a comparison of GeomCP; the Inspect method of Wang and Samworth (2018); and the E-Divisive method of Matteson and James (2014) using the statistical software R (R Core Team 2019). First, we investigate how changes in mean and variance of time series manifest themselves in the distance and angle measures within GeomCP. We then compare GeomCP to Inspect and E-Divisive in a wide range of scenarios including dense changepoints, where the change occurs in all or a large number of dimensions, and sparse changepoints, where the change occurs in a small number of dimensions. Changes in both mean, variance and a combination of the two will be considered.

Inspect is only designed for detecting changes in mean, therefore, it will only be included in such scenarios. In addition, Inspect is designed for detecting sparse changepoints, however, Inspect 'can be applied in non-sparse settings as well' (Wang and Samworth 2018) so we also include it in the dense change in mean scenarios. Like GeomCP, E-Divisive is designed for dense changepoints, but we will also include it in the sparse changepoint scenarios to assess performance.

For GeomCP, we perform the mappings in (2.2) and (2.3) before applying the PELT algorithm using the *changepoint* package (Killick and Eckley 2014). Unless otherwise stated, we use the default settings; namely, the MBIC penalty (Zhang and Siegmund 2007), Normal distribution and allow for changes in mean and variance. We implement the Inspect method using the *InspectChangepoint* package (Wang and Samworth 2016). The thresholds used to identify significant changepoints are calculated before timing the simulations using the data-driven approach suggested in Wang and Samworth (2018). For the remaining user-defined parameters, we use the default settings with $Q = 0$. Setting $Q = 0$ implements a Binary Segmentation approach (Scott and Knott 1974; Vostrikova 1981) for identifying multiple changepoints. When using $Q = 1000$, as suggested in Wang and Samworth (2018), a Wild Binary Segmentation (Fryzlewicz 2014) approach is implemented to detect multiple changes. However, this becomes computational infeasible even at moderate levels of n and p while only resulting in minor improvements in detection rate at the expense of higher false discovery rates. For $p > 1000$, the data-driven calculation of the thresholds was computationally infeasible; hence, the theoretical threshold derived in Wang and Samworth (2018) was originally implemented. However, this led to an excessive number of false positives and, as such, is not included. For the implementation of the E-Divisive method, we use the *ecp* package (James and Matteson 2014) with $\alpha = 1$; minimum segment size of 30; a significance level of 0.05; and $R = 499$ as suggested by Matteson and James (2014).

Unless indicated otherwise, we simulate data from a Normal distribution with changes in mean and variance given in each scenario. Additionally, the number of changepoints is set as $m = \lceil \frac{n}{200} \rceil$ and we distribute the changepoints uniformly at random throughout the time series with the condition that they are at least 30 time points apart. Where computationally feasible, we perform 500 repetitions of each scenario and display the true detection rate (TDR) and false detection rate (FDR) along with their confidence intervals given by two standard errors. For scenarios with $n \geq 1000$, E-Divisive was only run on 30 replications due to the high computational cost. Changepoint estimates are deemed correct if they are the closest to, and within 10 time points of, the true changepoint and contribute to the TDR. Changepoint estimates more than 10 time points from the true changepoints or where another estimated changepoint is closer to the true changepoint are deemed false and contribute to the FDR. We seek a TDR as close to 1 as possible and an FDR as close to 0 as possible. As GeomCP estimates changepoints in both the distance and angle measures, we apply the reconciling method from Sect. 2.4 with the threshold, $\xi = 10$. Then we apply the same TDR/FDR method to the reconciled changes.

3.1 Size of changepoints

As we are interested in multivariate changepoints, we need to decide upon the size of a change in each series. If we fixed a specific change size in each series, then as p increases, the change becomes easier to identify due to multivariate power. If we fixed a total change size across all series, then as p increases, the change becomes considerably harder to detect. Hence, we set our simulated change sizes so that GeomCP has an approximately constant performance across p, in terms of TDR and FDR.

To achieve a constant performance in the change in mean scenario, we require the difference in the expected distance measure pre- and post-change to be constant across p. If we

assume unit variance and a set mean across all series before the change, $\tilde{\mu}_{\text{pre}}$ and after the change, $\tilde{\mu}_{\text{post}}$, using Theorem 1 the expected difference in the distance measure before and after a changepoint is,

$$\mathbb{E}(d_{\text{post}} - d_{\text{pre}}) = \sqrt{p}\left(\sqrt{\tilde{\mu}_{\text{post}}^2 + 1} - \sqrt{\tilde{\mu}_{\text{pre}}^2 + 1}\right) .$$

If we set the total mean change size in our simulated data as,

$$\sum_{j=1}^{p} \mu_{j,\text{post}} - \mu_{j,\text{pre}} = \sqrt{p}\Theta , \tag{3.1}$$

for some constant Θ and, again, assume the mean of each series is the same, we gain,

$$\begin{aligned}\Theta &= \sqrt{p}\left(\tilde{\mu}_{\text{post}} - \tilde{\mu}_{\text{pre}}\right) \\ &\approx \sqrt{p}\left(\sqrt{\tilde{\mu}_{\text{post}}^2 + 1} - \sqrt{\tilde{\mu}_{\text{pre}}^2 + 1}\right) \\ &= \mathbb{E}(d_{\text{post}} - d_{\text{pre}}) .\end{aligned}$$

Hence, for a constant Θ, using a total mean change size scaling as in (3.1) will result in the expected difference of the distance pre- and post-change, and therefore the performance of GeomCP, being approximately constant across p. As we re-scale our data before applying our two mappings, the pre- and post-change means will be large enough that this approximation is reasonable.

Similarly, to gain an approximately constant performance of GeomCP across p for a change in variance, we set the total variance change size in our simulated data as

$$\prod_{j=1}^{p} \frac{\sigma_{j,\text{post}}}{\sigma_{j,\text{pre}}} = \Phi^{\sqrt{p}} , \tag{3.2}$$

for some constant Φ. When comparing methods, we shall use (3.1) and (3.2) to define the total change size for each scenario, with the change size being the same in all series that undergo a change.

3.2 GeomCP investigation

First, we investigate how changes in mean, variance and a combination of the two manifest themselves in the distance and angle measure within GeomCP. We set $n = 1000$ and $p = 200$ and simulate data with changepoints $\boldsymbol{\tau} = (250, 500, 750)$. At τ_1 we have a mean change of $+0.1$ in all series; at τ_2 we have a variance change of $\times 1.2$ in all series; and at τ_3 we have a mean change of -0.1 and a variance changes of $\times 1.2^{-1}$ in all series. Figure 1 shows 4 of the 200 series and shows the changepoints are undetectable by eye in the individual series. Applying the mappings within

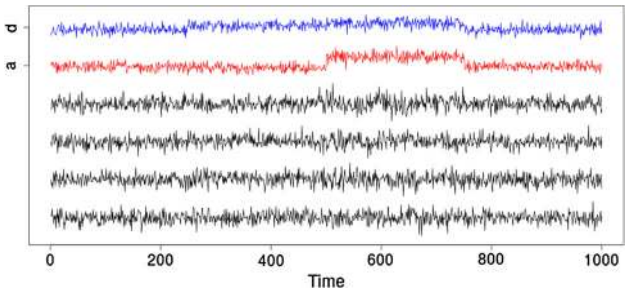


Fig. 1 4 series from the simulated data set with the distance (d) and angle (a) mappings showing 3 changepoints that are not obvious in the individual series

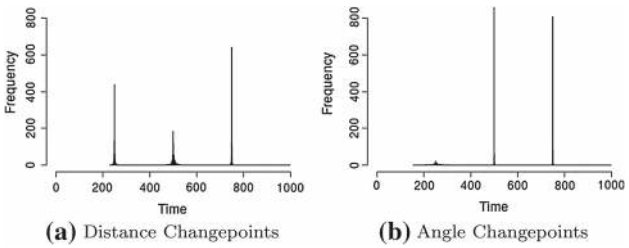


Fig. 2 Locations of detected changepoints in 1000 repetitions of simulated data set with changepoints at 250, 500 and 750, in mean, variance and, mean and variance, respectively

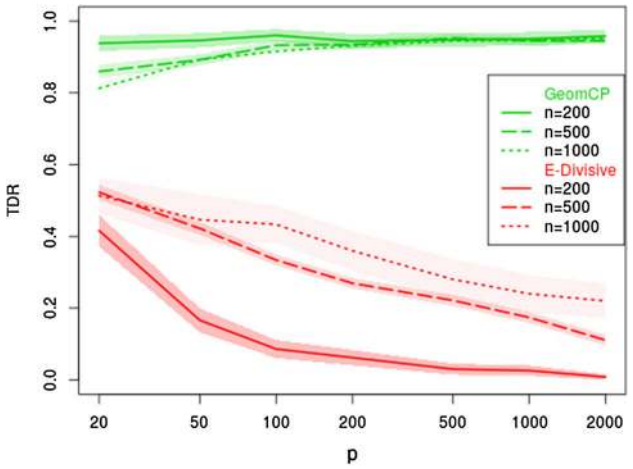
GeomCP results in the mapped series seen in Fig. 1 where the changes are clearly identifiable in at least one of the distance or angle measure.

Figure 2a, b shows the position of identified changepoints in the distance and angle measure in 1000 replications of the current scenario using PELT. The relatively small change in mean at time point 250 is only reliably picked up by the distance measure. The change in variance is picked up by the angle measure in almost all cases and is also seen in the distance measure, however, with less accuracy and less often. The change in mean and variance at time point 750 is reliably detected in both the distance and angle measures. These findings were similar for varying mean and variance changes. As such, this justifies setting the location of changepoints that occur in both series to be given by the angle changepoint location as stated in Sect. 2.4.

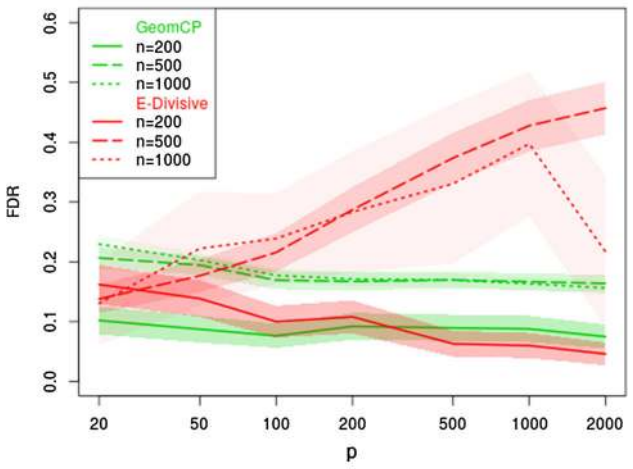
3.3 Dense changepoints

Now we compare GeomCP's performance with E-Divisive and Inspect. We investigate dense variance changes here, with mean, and mean and variance changes given in the Supplementary Material.

We simulate data with variance changes that occur in all series for a wide range of n and p and show a subset of the results here. We keep the mean vector constant, and we split the total change size defined in (3.2) evenly across all series. We display results with $\Phi = 3$ as this is shown to give a high



(a) TDR



(b) FDR

Fig. 3 TDR and FDR for GeomCP and E-Divisive for simulated data sets containing variance changes that occur in all series for multiple n and p

TDR while maintaining a low FDR in Eckley et al. (2011) for $p = 1$. Similar findings occur with varying values of Φ; see the Supplementary Material. We apply the GeomCP and E-Divisive methods to these simulated data sets, and the TDR and FDR are shown in Fig. 3.

Figure 3a shows the TDR across different numbers of dimensions and time points. It is clear that GeomCP outperforms E-Divisive in terms of TDR and the gap between the methods widens as the number of dimensions increases. Figure 3b shows that the improved TDR of GeomCP does not come at the expense of a higher FDR, which has similar rates across n and p.

In the mean, and mean and variance change scenarios, GeomCP similarly outperforms both E-Divisive and Inspect in terms of TDR while maintaining a low-level FDR across n and p. Results can be found in the Supplementary Material.

3.4 Sparsity investigation

Thus far we assumed that all series undergo a change at each changepoint. We now explore the effect of the sparsity of the changepoint. We define $\kappa \in (0, 1]$ to be the probability that a series undergoes a change. We explore sparse mean changes here, with sparse variance changes included in the Supplementary Material.

For the sparse changepoint scenarios, we set $n = 500$, $p = 200$ and vary κ; we note that there were similar findings for different n and p. We keep the variance vector constant and the change size in each series that undergoes a change, is the total change size defined in (3.1), split between the expected number of series to undergo a change. This means the expected total change size is the same as when all series undergo a change. We display results with $\Theta = 1.2$ and similar findings occur with varying values of Θ; see the Supplementary Material. We apply the GeomCP, Inspect and E-Divisive methods to these scenarios, and the TDR and FDR are shown in Fig. 4.

Figure 4a shows that GeomCP maintains a constant TDR across κ as expected. This reflects the setup of the scenario where the expected total change size is constant across κ. For dense changepoints, GeomCP compares well as we might expect. Interestingly, E-Divisive also assumes dense changepoints but performs poorly in this scenario. Inspect is designed for sparse changes and as expected, for very sparse changes the method performs the best. For sparse changepoints, the improved performance of Inspect and E-Divisive may be due to the size of change in each affected series increasing as κ decreases.

3.5 Between-series dependence

Now we will relax the assumption of a diagonal covariance matrix and investigate how this affects the performance of GeomCP. We will investigate how two different covariance matrix structures compare to the independent, diagonal covariance case. Here we will investigate variance changes in these covariance structures with mean changes explored in the Supplementary Material.

For these scenarios, we set $n = 200$, $p = 100$ and have one changepoint at $\tau = 100$. The pre-changepoint data will be distributed from a $N(\mathbf{0}, \Sigma)$, while the post-changepoint data distributed from a $N(\mathbf{0}, \boldsymbol{\sigma}\Sigma)$. We will vary the change size, $\boldsymbol{\sigma}$, while each entry of $\boldsymbol{\sigma}$ will be identical for each change size. We will compare three structures for Σ:

1. Independent case: $\Sigma = I$.
2. Block-diagonal case: Here Σ will be a block-diagonal matrix with block size of 2. The off-diagonal entries will be randomly sampled from a $U(-0.6, -0.3) \cup$

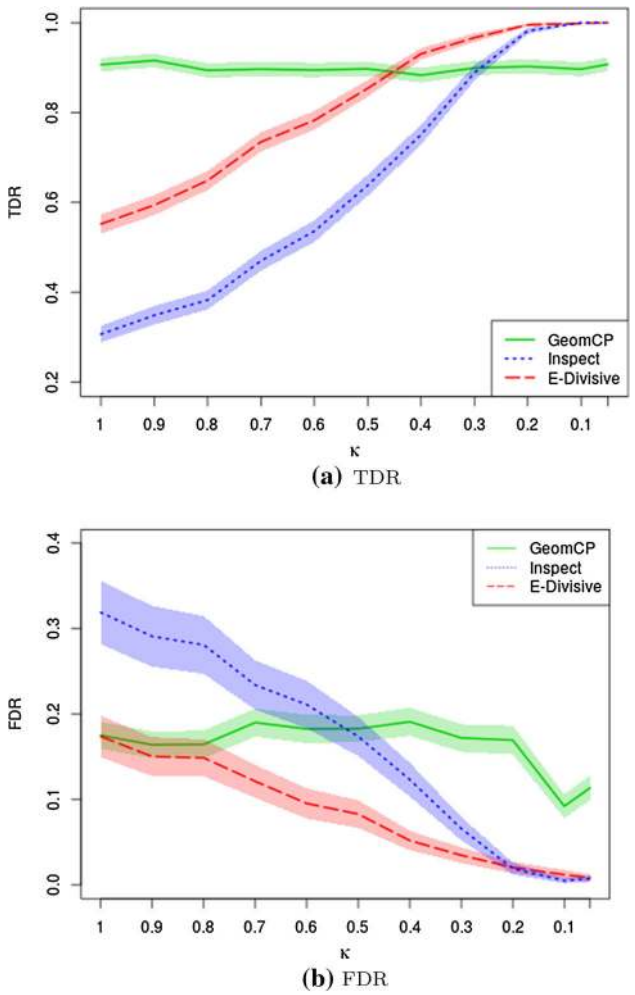


Fig. 4 TDR and FDR for GeomCP, Inspect and E-Divisive for simulated data sets with sparse mean changes for $n = 500$ and $p = 200$

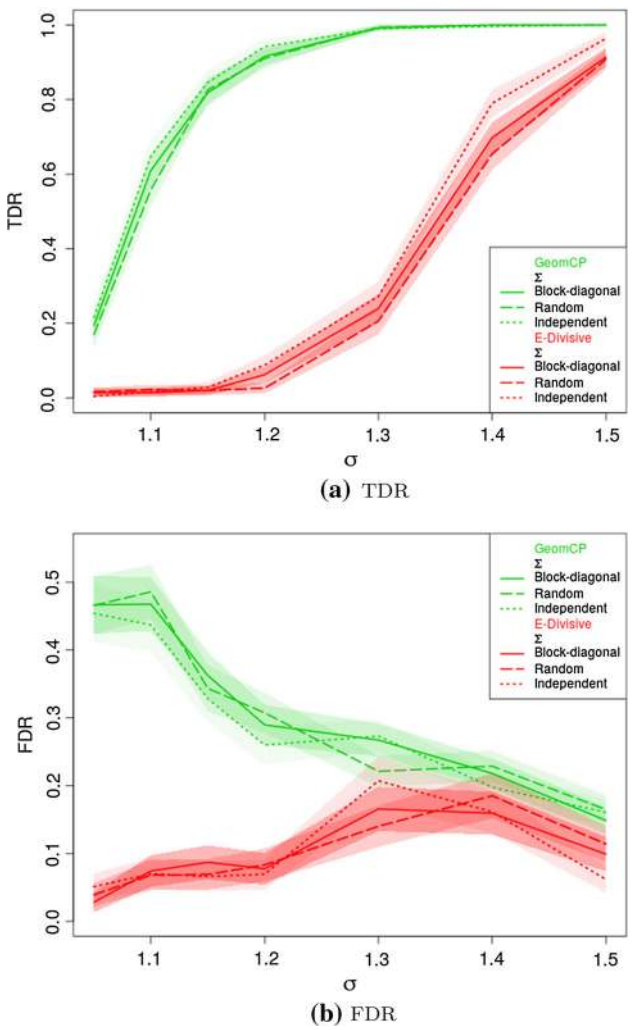


Fig. 5 TDR and FDR for GeomCP and E-Divisive for simulated data with a change in covariance for $n = 200$ and $p = 100$

$U(0.3, 0.6)$ distribution with the diagonal entries equal to 1.

3. Random case: Here we let $\Sigma = PDP'$ where P is an orthogonalized matrix of standard Normal random variables and D is a diagonal matrix with entries decreasing from 30 to 1.

As we no longer have independence between series, we cannot assume Normality of the distance and angle measures within GeomCP. Hence, we use the empirical cost function (Haynes et al. 2017b) within PELT to detect changes in the distance and angle measures. We similarly use the empirical cost function in the independent case for comparability.

Figure 5a shows the TDR of GeomCP and E-Divisive for varying change sizes, σ, and the different covariance structures. GeomCP clearly has a greater TDR than E-Divisive for smaller change sizes. However, Fig. 5b shows this comes at the expense of a higher FDR. This is to be expected when using the empirical cost function within PELT as this generally produces a higher FDR. By altering the penalty used within PELT, this FDR could be reduced at the cost of some power in detecting changes. Yet for $\sigma \geq 1.3$ GeomCP has a competitive FDR with E-Divisive while having a superior TDR. Interestingly, the covariance structure has very little impact on the performance of GeomCP; this follows our intuition from Sect. 2.5.

3.6 Computational speed

A major issue with high-dimensional changepoint detection is: as n and p grow large, many multivariate changepoint methods become computationally infeasible. Here, we compare the computational speeds of GeomCP, Inspect and E-Divisive for a range of n, p and m. We will compare the speeds in three scenarios:

1. n increasing while $p = 200$ and $m = \lceil \frac{n}{200} \rceil$.
2. n increasing while $p = 200$ and $m = 2$.
3. p increasing while $n = 500$ and $m = \lceil \frac{n}{200} \rceil = 3$.

The second scenario breaks PELT's assumption of a linearly increasing number of changepoints as the number of time points increases. This means the speed of detecting changepoints using GeomCP will no longer be linear in time. We performed simulations using the three scenarios defined above and only included mean changes, so we can compare with Inspect. We set the mean change size to be $\theta_j = 0.8$ in all series so that the changes are obvious. For scenario 1 and 2, E-Divisive was computationally infeasible for $n \geq 1000$. For scenario 3, Inspect's speed is only shown for $p < 1000$ due to the excessive computational time of generating a data-driven threshold. Note that the data-driven thresholds needed for Inspect were calculated outside of the recorded times. In practice, if a threshold was required, then Inspect would take considerably longer to run especially as p increases. Within GeomCP, we run the algorithm in serial, performing the mapping and changepoint identification for the distance and then for the angle. These could be processed in parallel, leading to a further reduction in computational time.

Figure 6 shows the computational speed of each method in the three scenarios. We can see from Fig. 6a that, in scenario 1, GeomCP is the fastest of the three methods for all n. As n increases the difference between the speeds of GeomCP and Inspect increases (note the log scale on both axes). We can also see, E-Divisive is substantially slower than GeomCP and Inspect for all n and its run time increases rapidly as n gets large. Scenario 1 supports our claim that GeomCP has linear run time in n when the required assumptions of PELT are met.

In scenario 2, shown in Fig. 6b, we break the assumption within PELT that the number of changepoints is increasing linearly in time. This results in a comparatively slower performance of GeomCP, although it remains computationally faster than Inspect for all n shown. Similarly to scenario 1, E-Divisive has a much longer run time than both GeomCP and Inspect.

Finally, for scenario 3 Fig. 6c shows that, for small p, Inspect is the fastest of the methods but as p increases above 50, GeomCP is computationally faster. While Inspect is faster for $p < 50$, recall that this does not include the time for the calculation of the threshold. Interestingly, the run time of E-Divisive appears unaffected by p until $p \geq 1000$. This is likely due to its computational cost being mainly affected by the number of changepoints and time points, which remain constant. Scenario 3 also supports our claim that GeomCP has linear run time as p increases, note the log scale that distorts the linearity of the plot.

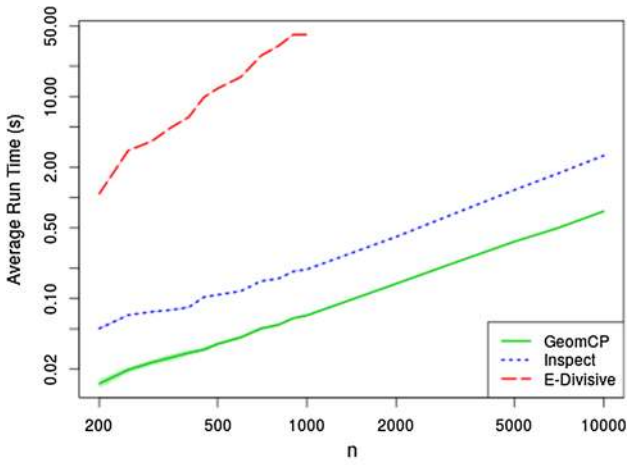


Fig. 6 Average run time for each method when: **a** n is increasing, $p = 200$ and m is increasing; **b** n is increasing, $p = 200$ and $m = 1$; **c** $n = 500$, p is increasing and $m = 3$ by default

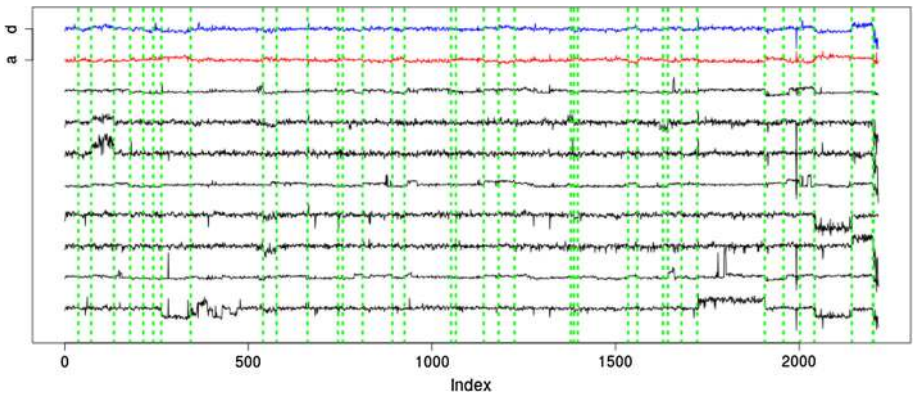


Fig. 7 Log-intensity-ratio measurements of microarray data from 6 out of 43 individuals and distance (d) and angle (a) mappings with vertical lines showing the identified changepoints

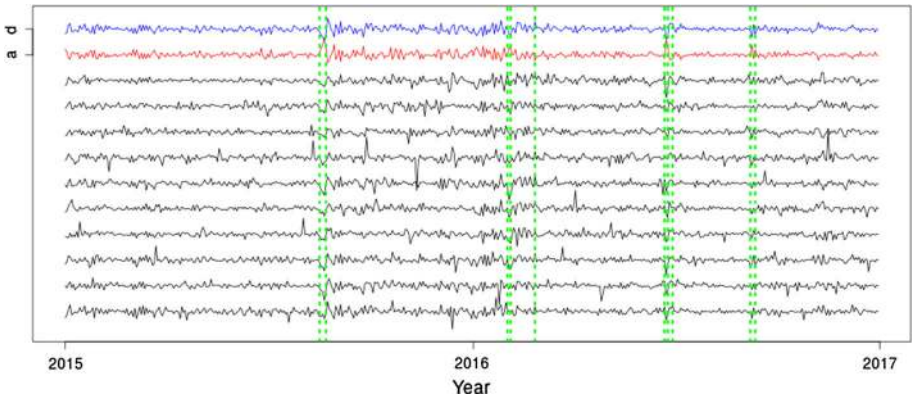


Fig. 8 Log-returns of 10 out of 447 companies within the S&P500 and the distance (d) and angle (a) mappings with vertical lines showing the identified changepoints

4 Applications

4.1 Comparative genomic hybridization

We study the comparative genomic hybridization microarray data set from Bleakley and Vert (2011). Comparative genomic hybridization allows the detection of copy number abnormalities in chromosomes by comparing the fluorescent intensity levels of DNA fragments between a test and reference sample. The data set contains log-intensity-ratio measurements from 43 individuals with bladder tumors with measurements taken at 2215 different positions on the genome. This data set is available in the *ecp* R package (James and Matteson 2014).

Copy number abnormalities come in regions on the genome and can either be specific to the individual or can be shared across several individuals. It is the latter that are of more interest as these are more likely to be disease-related. E-Divisive and Inspect have both been used to segment this data set with their results shown in Matteson and James (2014) and Wang and Samworth (2018), respectively. Under the default settings, these two methods fitted a large number of changepoints, 93 and 254, respectively, which may not be representative of changes occurring across multiple individuals. Wang and Samworth (2018) suggest selecting the 30 most significant changepoints to counter this; however, the justification for choosing 30 is unknown.

To perform our analysis, we first scale each series, similarly to Inspect, using the median absolute deviation to allow a better comparison. We then use the two mappings within GeomCP and apply the PELT algorithm, using the R package *changepoint.np* (Haynes and Killick 2016), to the resulting mapped series. The mappings do not appear Normal for this application; hence, we use the empirical cost function and set the number of quantiles as $4\log(n)$, as suggested in Haynes et al. (2017b). We use the CROPS algorithm of Haynes et al. (2017a) to identify an appropriate penalty value with diagnostic plots shown in the Supplementary Material. This leads to 37 changepoints being identified, and these are shown in Fig. 7 with the signal for 8 individuals from the study and the distance and angle mappings. Approximately 67.5% of the changepoints identified by GeomCP corresponded to those identified by E-Divisive (within 3 time points), with the majority of the rest corresponding to where E-Divisive fitted two changepoints. Also, the changepoints identified seem to be common across multiple individuals, while changes specific to a series are not detected. It is promising that our proposed segmentation identifies similar changepoints as other methods, while only identifying those that seem common across multiple individuals. Other GeomCP segmentations, using different potential penalty values identified

in CROPS, resulted in more or less of the individual features from specific series being detected.

4.2 S&P500 Stock prices

We now investigate the daily log-returns of the closing stock prices for 447 companies included in the S&P500 from January 2015 through to December 2016. This data set was created by Nugent (2018) and was loaded using the R package *SP500R* (Foret 2019). The aim is to identify changes in log-returns that affect a large number of companies rather than changes that are specific to individual companies. First we scale each series using the median absolute deviation. Next we apply the mappings within GeomCP, before using the PELT algorithm from the *changepoint* package (Killick et al. 2016) to both mapped series using the Normal cost function. We used the CROPS algorithm of Haynes et al. (2017a) to identify an appropriate penalty value for both series with diagnostic plots shown in the Supplementary Material.

Using GeomCP, we identified 10 changepoints. These are shown in Fig. 8 along with the log-returns of the first 10 companies from the S&P500 list and the mapped distance and angle measures. These changepoints correspond to key events that we would expect to impact the stocks of a large number of companies. For example, the changepoints in August 2015 correspond to large falls in the Chinese stock markets with the Dow Jones industrial average falling by 1300 points over 3 days. The changepoints in February and late June 2016 likely correspond to the announcement and subsequent result of the British referendum to leave the European Union. Applying the E-Divisive method (with the minimum segment length set to 2 and the rest of the user-defined parameters set as in Sect. 3) resulted in only 2 changepoints, both occurring in August 2015 similar to those detected in GeomCP.

5 Conclusion

We have presented a new high-dimensional changepoint detection method that can detect mean and variance changes in multivariate time series. This is achieved by implementing a univariate changepoint detection method on two related geometric mappings of the time series. We have shown that looking at the high-dimensional changepoint problem from a geometric viewpoint allows us to utilize relevant geometric structures to detect changepoints. We have displayed an improved performance in detecting and identifying the location of multiple changepoints over current state-of-the-art methods. Furthermore, we have shown an improved computational speed over competing methods when using the univariate changepoint method PELT. Finally, we have shown the effectiveness of GeomCP at detecting changepoints when applied to applications.

We have discussed how to extend this methodology to non-Gaussian data along with temporal and between-series dependence. However, a thorough investigation of how changes manifest in the distance and angle measure in the presence of these structures is left as future work.

All of the methods proposed here are implemented in the R package *changepoint.geo* available on CRAN https://cran.r-project.org/package=changepoint.geo.

References

Bleakley, K., Vert, J.P.: The group fused Lasso for multiple change-point detection. (2011) arXiv e-prints arXiv:1106.4199

Brodsky, E., Darkhovsky, B.: Nonparametric Methods in Change Point Problems. Mathematics and Its Applications. Springer, Berlin (2013)

Eckley, I.A., Fearnhead, P., Killick, R.: Analysis of changepoint models. In: Barber, D., Cemgil, A.T., Chiappa, S. (eds.) Bayesian Time Series Models chap 10, pp. 205–224. Cambridge University Press, Cambridge (2011)

Enikeeva, F., Harchaoui, Z.: High-dimensional change-point detection under sparse alternatives. Ann. Stat. **4**, 2051–2079 (2019)

Fearnhead, P., Maidstone, R., Letchford, A.: Detecting changes in slope with an l_0 penalty. J. Comput. Gr. Stat. **28**(2), 1–11 (2018)

Fisch, A.T.M., Eckley, I.A., Fearnhead, P.: A linear time method for the detection of point and collective anomalies. arXiv e-prints arXiv:1806.01947 (2018)

Foret, P.: SP500R: Easy loading of SP500 stocks data. Github R package version 0.1.0 (2019)

Fryzlewicz, P.: Wild binary segmentation for multiple change-point detection. Ann. Stat. **42**(6), 2243–2281 (2014)

Haynes, K., Killick, R.: changepoint.np: Methods for nonparametric changepoint detection. CRAN R package version 1.0.1 (2016)

Haynes, K., Eckley, I.A., Fearnhead, P.: Computationally efficient changepoint detection for a range of penalties. J. Comput. Gr. Stat. **26**(1), 134–143 (2017a)

Haynes, K., Fearnhead, P., Eckley, I.: A computationally efficient nonparametric approach for changepoint detection. Stat. Comput. **27**(5), 1293–1305 (2017b)

Horváth, L., Hušková, M.: Change-point detection in panel data. J. Time Ser. Anal. **33**(4), 631–648 (2012)

James, N.A., Matteson, D.S.: ecp: An R package for nonparametric multiple change point analysis of multivariate data. J. Stat. Softw. **62**(7), 1–25 (2014)

Jirak, M.: Uniform change point tests in high dimension. Ann. Stat. **43**(6), 2451–2483 (2015)
Killick, R., Eckley, I.A.: changepoint: An R package for changepoint analysis. J. Stat. Softw. **58**(3), 1–19 (2014)
Killick, R., Fearnhead, P., Eckley, I.A.: Optimal detection of changepoints with a linear computational cost. J. Am. Stat. Assoc. **107**(500), 1590–1598 (2012)
Killick, R., Haynes, K., Eckley, I.A.: changepoint: An R package for changepoint analysis. CRAN R package version 2.2.2 (2016)
Maboudou-Tchao, E.M., Hawkins, D.M.: Detection of multiple changepoints in multivariate data. J. Appl. Stat. **40**(9), 1979–1995 (2013)
Mahalanobis, P.C.: On the Generalized Distance in Statistics. National Institute of Science of India, India (1936)
Matteson, D.S., James, N.A.: A nonparametric approach for multiple change point analysis of multivariate data. J. Am. Stat. Assoc. **109**(505), 334–345 (2014)
Modisett, M.C., Maboudou-Tchao, E.M.: Significantly lower estimates of volatility arise from the use of open-high-low-close price data. N. Am. Actuar. J. **14**(1), 68–85 (2010)
Nugent, C.: S&P 500 stock data. (2018) https://www.kaggle.com/camnugent/sandp500, Kaggle dataset version 4
Page, E.S.: Continuous inspection schemes. Biometrika **41**(1/2), 100–115 (1954)
R Core Team: R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2019)
Scott, A.J., Knott, M.: A cluster analysis method for grouping means in the analysis of variance. Biometrics **30**(3), 507–512 (1974)
Terrera, G.M., van den Hout, A., Matthews, F.E.: Random change point models: investigating cognitive decline in the presence of missing data. J. Appl.Stat. **38**(4), 705–716 (2011)
Tickle, S.O., Eckley, I.A., Fearnhead, P., Haynes, K.: Parallelization of a common changepoint detection method. J. Comput. Gr. Stat. **0**, 1–13 (2019)
Truong, C., Oudre, L., Vayatis, N.: Selective review of offline change point detection methods. Signal Process. **167** (2020). https://doi.org/10.1016/j.sigpro.2019.107299
Vostrikova, L.: Detecting 'disorder' in multidimensional random processes. Sov. Math. Dokl. **24**, 55–59 (1981)
Wang, T., Samworth, R.: InspectChangepoint: high-dimensional changepoint estimation via sparse projection. CRAN R package version 1.0.1 (2016)
Wang, T., Samworth, R.J.: High dimensional change point estimation via sparse projection. J. R. Stat. Soc. Ser. B **80**(1), 57–83 (2018)
Wickens, T.: The Geometry of Multivariate Statistics. Lawrence Erlbaum Associates Inc., Hillsdale (1995)
Zhang, N.R., Siegmund, D.O.: A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. Biometrics **63**(1), 22–32 (2007)
Zhang, N.R., Siegmund, D.O., Ji, H., Li, J.Z.: Detecting simultaneous changepoints in multiple sequences. Biometrika **97**(3), 631–645 (2010)
Zou, C., Yin, G., Feng, L., Wang, Z.: Nonparametric maximum likelihood approach to multiple change-point problems. Ann. Stat. **42**(3), 970–1002 (2014)