Statistics
Innovation

ARTICLE

Open Access

https://doi.org/10.48130/stati-0026-0002
Statistics Innovation **2026**, 3: e002

A synthetic data approach for FDR control in change-point detection

Ao Sun[1], Jianxin Bi[2*] and Jingyuan Liu[3]

[1] *Department of Data Sciences and Operations, Marshall School of Business, University of Southern California, Los Angeles, CA 90089, USA*
[2] *School of Statistics and Mathematics, Shandong University of Finance and Economics, Jinan 250014, China*
[3] *MOE Key Laboratory of Econometrics, Department of Statistics and Data Science in School of Economics, Laboratory of Digital Finance, and Fujian Key Laboratory of Statistical Science, Xiamen University, Xiamen 361005, China*
* Correspondence: bjxstatistics@stu.xmu.edu.cn (Bi J)

Abstract

In multiple change-point analysis, the resulting detection sets are typically conservative, often identifying more change points than actually exist, due to the issues of 'unreliability of assumptions' and 'unreliability of algorithms'. Therefore, controlling the false discovery rate is of vital importance to multiple change-point detection. Data-splitting-based methods have gained widespread attention for false discovery rate control. However, relying solely on a part of the dataset during the validation stage typically suffers from power loss. Instead, the study introduces a novel synthetic data framework and proposes the Synthetic Data Filter to control the false discovery rate in multiple change-point detection. Here, the study demonstrates that the proposed method effectively controls the false discovery rate and achieves asymptotic power approaching one under mild conditions. Numerical comparisons with existing methods provide evidence for the superiority of the approach in terms of both false discovery rate control and statistical power. The proposed method is further applied to a bladder tumor microarray dataset, and potential loci are identified with structural changes.

Citation: Sun A, Bi J, Liu JY. 2026. A synthetic data approach for FDR control in change-point detection. *Statistics Innovation* 3: e002 https://doi.org/10.48130/stati-0026-0002

Introduction

Change-point analysis, a process of detecting structural changes in a data sequence, has been an active area of research and attracted increasing attention with the growing availability of temporal data. It has applications across a wide range of fields, including but not limited to environmental sciences, econometrics, biology, geosciences, and linguistics. In this context, the accurate and efficient detection of multiple change-points (MCP) is undoubtedly one of the most crucial issues. For example, Liu et al.[1] introduced a general framework for high-dimensional change-point detection by constructing a U-statistic-based cumulative sum matrix C and aggregating it based on the adjusted L_p-norm, while Liu et al.[2] focused on high-dimensional linear models, providing asymptotic validity and an extension to multiple change points via binary segmentation. For more comprehensive reviews of various existing approaches to MCP inference, see Aue & Horváth[3], and Niu et al.[4].

However, obtaining consistent estimators for the number and locations of MCP typically requires stringent conditions on the magnitude of changes, as extensively documented in prior studies[5−13]. Unfortunately, such requirements are often unrealistic, as small change magnitudes tend to cause underfitting. Consequently, an overfitted selection set is often obtained via some conservative algorithms. Furthermore, the empirical performance of certain detection methods is intrinsically linked to the choice of tuning parameters, which requires access to unavailable population-level information. These two issues, referred to as the unreliability of assumptions and the unreliability of algorithms, can introduce false discoveries, potentially leading to the reproducibility crisis if an excessive number of false detections occur.

To tackle this problem, a natural solution is to detect the active set while controlling the false discovery rate (FDR) at a pre-specified level. A widely adopted strategy is to treat change point detection as a multiple hypothesis testing problem, utilizing classical p-value-based methods [14−18] to control the FDR. Notable works include Hao et al.[19], Li et al.[20], and Cheng et al.[21]. These methods work properly for the univariate mean change problem, but extending them to a multi-dimensional setting is challenging, since the model's complexity renders the derivation of p-values intractable. Leveraging the knockoff framework[22−26], a related study is Liu et al.[27], which proposed a generalized knockoff procedure (GKnockoff) to control the FDR for detecting structural changes in the coefficients of a linear regression model. More recently, Du et al.[28], and Dai et al.[29] proposed a data-splitting-based FDR control framework, which outperforms knockoff methods in power under moderate to strong dependence and is more robust than asymptotic p-value-based methods. Chen et al.[30] adopted this data-splitting philosophy and proposed a data-driven selection procedure for MCP detection while controlling the error rate. This approach is quite general and can handle complex MCP scenarios. However, on the one hand, the symmetry property of the proposed statistic depends on the sample size. If the sample size is too small, the comparison statistic may become asymmetric, distorting the FDR control. On the other hand, data-splitting inevitably reduces the power of change-point detection as only half of the information is utilized.

The limitations of data-splitting motivate us to develop a novel framework for controlling the error rates in high-dimensional MCP detection. This framework, termed the Synthetic Data Filter (SD filter), is designed to enhance the accuracy of multiple change-point identification. Basically, the SD filter procedure consists of three steps. First, the dataset is divided according to the temporal order's parity, as inspired by Chen et al.[30]. The change-point detection is performed on the odd-part and can be carried out using the adaptive ℓ_q aggregated CUSUM-type statistic introduced by Liu et al.[1]. Next, information from the odd part is leveraged to generate a synthetic dataset. In the final step, the synthetic dataset is merged with the reserved even-part dataset to construct symmetric statistics and control the false discovery rate (FDR). Figure 1 presents a flow chart summarizing the procedure of the proposed SD filter. The SD filter exhibits the capability to asymptotically control the FDR at the desired level while achieving a power that approaches one under mild conditions. Moreover, this work demonstrates its competence in addressing high-dimensional MCP problems and represents, to the best of current knowledge, the first attempt at controlling the error rate in this context.

The synthetic data framework clearly departs from the aforementioned error rate control procedures. Unlike conventional methods that quantify the distribution of the test statistic, the Gaussian multiplier bootstrap is employed [31,32] to construct a mirror statistic for FDR control. This framework is highly general, easily extendable to various model settings, and applicable to any scenario requiring data splitting for simultaneous hypothesis formulation and testing. By integrating the synthetic random sample with the second dataset in the testing phase, the original sample size is maintained, thereby boosting statistical power.

The synthetic data framework for MCP

Suppose that a sequence of independent data have been observed, $\mathcal{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_{2n}\}$ collecting from

$$\mathbf{z}_i \sim F(\cdot|\boldsymbol{\beta}_k), \tau_k < i \leqslant \tau_{k+1}, k = 1, \ldots, K_n; i = 1, \ldots, 2n \tag{1}$$

where, K_n is the number of change-points which could diverge with sample size n and $\tau_k's$ are the change locations with the convention that $\tau_0 = 0$ and $\tau_{K_n+1} = 2n$. $F(\cdot|\boldsymbol{\beta}_k)$ represents the model structure of segment k, where $\boldsymbol{\beta}_k$ is a d-dimensional parameter vector of interest, satisfying $\boldsymbol{\beta}_k \neq \boldsymbol{\beta}_{k+1}$. The setting of MCP in Eq. (1) is quite general and encompasses many classical models, such as the multivariate mean change model and the regression model with structural breaks [33].

The objective of this study is to detect the active set $\mathcal{S} = \{\tau_k, k = 1, \ldots, K_n\}$ while controlling the FDR. However, the definition of FDR in this setting is context-specific, as the optimal rate of change-point estimation is characterized by $O_p(1)$[34]. Therefore, it is essential to first establish the corresponding concepts in this context. Given a candidate change-point set $\hat{\mathcal{S}} = \{\hat{\tau}_k, k = 1, \ldots, \hat{K}_n\}$, the definition of false discovery in MCP detection is as follows:

Definition 2.1 (False discovery). *The candidate change-point $\hat{\tau}_k \in \hat{\mathcal{S}}$ is a* ***False Discovery*** *if there is no true change-point falls in*

$$G_k := [\lceil(\hat{\tau}_{k-1} + \hat{\tau}_k)/2\rceil, \lceil(\hat{\tau}_k + \hat{\tau}_{k+1})/2\rceil) \tag{2}$$

where, $\hat{\tau}_0 = 0$ and $\hat{\tau}_{\hat{K}_n} = 2n$.

Further, let the set $\mathcal{I}_0$ encompass all the false discoveries. Then, the set $\mathcal{I}_1 = \hat{\mathcal{S}} \cap \mathcal{I}_0^c$ contains all the true discoveries in the selection set $\hat{\mathcal{S}}$. Throughout this paper, $\mathcal{I}_1$ and $\mathcal{I}_0$ are referred to as the informative and uninformative sets, respectively. This definition is well-defined, as each candidate change point is unambiguously classified as either a true or false discovery, with no overlap between the two categories. Based on this, the corresponding FDR is defined as follows:

$$\mathrm{FDR}(\mathcal{T}) = \mathrm{E}\left[\frac{|\mathcal{T} \cap \mathcal{I}_0|}{|\mathcal{T}|}\right] \tag{3}$$

where $\mathcal{T}$ represents a subset of $\hat{\mathcal{S}}$ yielded by the selection procedure, and $|A|$ denotes the cardinality of set A. The FDR is the expected value of the False Discovery Proportion (FDP), which represents the ratio of false discoveries to the total number of discoveries.

Synthetic data generating process

This subsection first introduces a synthetic data generating procedure. Following the order-preserving splitting procedure in Zou et al.[33], the data $\mathcal{Z}$ is partitioned into odd and even parts

$$\mathcal{Z}^O := \{\mathbf{z}_{2i-1}, i = 1, \ldots, n\} \text{ and } \mathcal{Z}^E := \{\mathbf{z}_{2i}, i = 1, \ldots, n\}$$

On the subset $\mathcal{Z}^O$, a candidate set of change points is estimated $\hat{\mathcal{S}} = \{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{K}_n}\}$ using a suitable detection algorithm, such as the aggregated CUSUM method. This training phase allows for the possibility that $\hat{\mathcal{S}}$ may overestimate the true set of change-points $\mathcal{S}$. Based on the candidate set $\hat{\mathcal{S}}$, sets G_k are defined according to Definition 2.2. Then the odd sample $\mathcal{Z}^O$ and the even sample $\mathcal{Z}^E$ are partitioned into segments $\mathcal{Z}^O_{G_k} := \{\mathbf{z}_{2i-1} : i \in G_k\}$ and $\mathcal{Z}^E_{G_k} := \{\mathbf{z}_{2i} : i \in G_k\}$, respectively. In the validation stage, $\mathcal{Z}^O$ is treated as given to avoid dealing with intractable post-selection inference.

Let $\ell(\boldsymbol{\beta}; \mathbf{z}_i)$ be a suitable loss function evaluated at data point $\mathbf{z}_i$, with its derivative denoted as $\mathbf{s}_{\beta}(\mathbf{z}_i) = \partial\ell(\boldsymbol{\beta}; \mathbf{z}_i)/\partial\boldsymbol{\beta}$. Ideally, for a given d-dimensional reference vector $\boldsymbol{\gamma}$, $*E\{\mathbf{s}_{\gamma}(\mathbf{z}_i)\} \neq *E\{\mathbf{s}_{\gamma}(\mathbf{z}_{i'})\}$ is expected when there is a change between i and i', since the score function remains constant in regions without change. This motivates the decomposition of the score as

$$\mathbf{s}_{\gamma}(\mathbf{z}_i) = \boldsymbol{\mu}_i + \boldsymbol{\zeta}_i, i = 1, \ldots, 2n. \tag{4}$$

where, $\boldsymbol{\mu}_i = *E[\mathbf{s}_{\gamma}(\mathbf{z}_i)]$ is the expected score at $\mathbf{z}_i$, and $\boldsymbol{\zeta}_i = \mathbf{s}_{\gamma}(\mathbf{z}_i) - *E[\mathbf{s}_{\gamma}(\mathbf{z}_i)]$ represents the residual. It is further assumed that $*Cov(\boldsymbol{\zeta}_i) = \boldsymbol{\Sigma}^{(k)}$ for all i such that $\tau_k + 1 \leqslant i \leqslant \tau_{k+1}$. This score-type framework was first introduced by Zou et al.[33] for selecting the number of change-points. The procedure is often invariant to the choice of $\boldsymbol{\gamma}$, which can therefore be set as $\boldsymbol{\gamma} := \arg\min_{\beta} \sum_{\mathbf{z}_i \in \mathcal{Z}^O} \ell(\boldsymbol{\beta}; \mathbf{z}_i)$ when no prior information is available. Given a specific $\boldsymbol{\gamma}$, the score function $\mathbf{s}_{\gamma}(\mathbf{z}_i)$ is denoted as $\mathbf{s}_i$ for simplicity.

To monitor the change magnitude in the dataset$\mathcal{Z}^E_{G_k}$, the CUSUM statistic is employed for score $\mathbf{s}_i^E$ in $\mathcal{Z}^E_{G_k}$, which is defined as,

$$\mathbf{c}_k(s) = \sqrt{\frac{s(n_k - s)}{n_k}}\left(\frac{1}{s}\sum_{i \leqslant s, i \in G_k} \mathbf{s}_i^E - \frac{1}{n_k - s}\sum_{i > s, i \in G_k} \mathbf{s}_i^E\right), \tag{5}$$

where, $n_k = |G_k|$ and s varies in G_k with the exception of points that are too close to both ends. Substitute $\mathbf{s}_i^E$ by $\boldsymbol{\mu}_i^E + \boldsymbol{\zeta}_i^E$, it becomes

$$\mathbf{c}_k(s) = \sqrt{\frac{s(n_k - s)}{n_k}}\left(\frac{1}{s}\sum_{i \leqslant s, i \in G_k} \boldsymbol{\zeta}_i^E - \frac{1}{n_k - s}\sum_{i > s, i \in G_k} \boldsymbol{\zeta}_i^E\right) + \Delta_k(s), \tag{6}$$

where, $\Delta_k(s) = \sqrt{s(n_k - s)/n_k}\left(s^{-1}\sum_{i \leqslant s, i \in G_k} \boldsymbol{\mu}_i^E - (n_k - s)^{-1}\sum_{i > s, i \in G_k} \boldsymbol{\mu}_i^E\right)$. If $\hat{\tau}_k$, located at the s-th position of G_k, is a false discovery, then

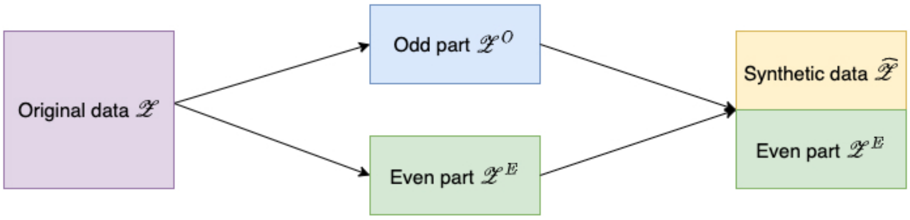
Odd part $\mathscr{X}^O$
Original data $\mathscr{X}$
Synthetic data $\widetilde{\mathscr{X}}$
Even part $\mathscr{X}^E$
Even part $\mathscr{X}^E$

Fig. 1 Flow chart of SD filter.

$\Delta_k(s) = 0$. Hence the CUSUM statistic $\mathbf{c}_k$ is just a combination of the errors,

$$\mathbf{c}_k(s) = \sqrt{\frac{s(n_k - s)}{n_k}} \left(\frac{1}{s} \sum_{i \leqslant s, i \in G_k} \zeta_i^E - \frac{1}{n_k - s} \sum_{i > s, i \in G_k} \zeta_i^E \right), \text{ if } \hat{\tau}_k \in \mathcal{I}_0. \tag{7}$$

Observing that for a certain false discovery $\hat{\tau}_k$ lies in the s-th position of G_k, $\mathbf{c}_k(s)$ consists of two parts: $s^{-1} \sum_{i \leqslant s, i \in G_k} \zeta_i^E$ and $(n_k - s)^{-1} \sum_{i > s, i \in G_k} \zeta_i^E$. According to the central limit theorem, it can be inferred that the two scaled summations $s^{-1/2} \sum_{i \leqslant s, i \in G_k} \zeta_i^E$ and $(n_k - s)^{-1/2} \sum_{i > s, i \in G_k} \zeta_i^E$ converge to a d-dimensional normal distribution with mean zero and covariance matrix $\mathbf{\Sigma}^{(k)}$, provided that both s and $n_k - s$ are sufficiently large.

Based on this intuition, i.i.d random variables $\xi_1, \ldots, \xi_{n_k}$ are introduced, each of which follows a standard normal distribution $N(0,1)$. These variables are also independent of the original dataset $\mathcal{Z}$ and define two partial sums based on $\mathcal{Z}^O$ as $\bar{\mathbf{s}}_k^{O,-}(s) = s^{-1} \sum_{i \leqslant s, i \in G_k} \mathbf{s}_i^O$ and $\bar{\mathbf{s}}_k^{O,+}(s) = (n_k - s)^{-1} \sum_{i > s, i \in G_k} \mathbf{s}_i^O$. Then a synthetic dataset $\tilde{\mathcal{Z}}_{G_k}$ is generated, consisting of the following two parts.

Definition 2.2. *Define the synthetic data based on the training sample as:*

$$\left\{ \xi_i \left(\mathbf{s}_i^O - \bar{\mathbf{s}}_k^{O,-}(s) \right), i \leqslant s, i \in G_k \right\} \text{ and } \left\{ \xi_i \left(\mathbf{s}_i^O - \bar{\mathbf{s}}_k^{O,+}(s) \right), i > s, i \in G_k \right\}. \tag{8}$$

By using the synthetic dataset $\tilde{\mathcal{Z}}_{G_k}$, a synthetic CUSUM can further be constructed

$$\tilde{\mathbf{c}}_k(s) = \sqrt{\frac{s(n_k - s)}{n_k}} \left(\frac{1}{s} \sum_{i \leqslant s, i \in G_k} \xi_i \left(\mathbf{s}_i^O - \bar{\mathbf{s}}_k^{O,-}(s) \right) - \frac{1}{n_k - s} \sum_{i > s, i \in G_k} \xi_i \left(\mathbf{s}_i^O - \bar{\mathbf{s}}_k^{O,+}(s) \right) \right). \tag{9}$$

Given the dataset $\mathcal{Z}_{G_k}^O$, two summations in the $\tilde{\mathbf{c}}_k(s)$ are also normally distributed, that is

$$\frac{1}{\sqrt{s}} \sum_{i \leqslant s, i \in G_k} \xi_i \left(\mathbf{s}_i^O - \bar{\mathbf{s}}_k^{O,-}(s) \right) \sim N\left(0, \widehat{\mathbf{\Sigma}}^{(k)-}\right)$$

$$\text{and } \frac{1}{\sqrt{n_k - s}} \sum_{i > s, i \in G_k} \xi_i \left(\mathbf{s}_i^O - \bar{\mathbf{s}}_k^{O,+}(s) \right) \sim N\left(0, \widehat{\mathbf{\Sigma}}^{(k)+}\right),$$

where,

$$\widehat{\mathbf{\Sigma}}^{(k)-}(s) = \frac{1}{s} \sum_{i \leqslant s, i \in G_k} \left(\mathbf{s}_i^O - \bar{\mathbf{s}}_k^{O,-}(s) \right) \left(\mathbf{s}_i^O - \bar{\mathbf{s}}_k^{O,-}(s) \right)^\top$$

$$\widehat{\mathbf{\Sigma}}^{(k)+}(s) = \frac{1}{n_k - s} \sum_{i > s, i \in G_k} \left(\mathbf{s}_i^O - \bar{\mathbf{s}}_k^{O,+}(s) \right) \left(\mathbf{s}_i^O - \bar{\mathbf{s}}_k^{O,+}(s) \right)^\top$$

are two plausible estimates of $\mathbf{\Sigma}^{(k)}$. Hence, the synthetic CUSUM statistic $\tilde{\mathbf{c}}_k(s)$ mimics the distributional behavior of the original CUSUM statistic $\mathbf{c}_k(s)$ when there are no change points, and it is independent of $\mathbf{c}_k(s)$ given the odd samples $\mathcal{Z}^O$.

To conclude, for each candidate change point $\hat{\tau}_k$, if it is a false discovery, the distributions of $\mathbf{c}_k(s)$ and $\tilde{\mathbf{c}}_k(s)$ are highly similar. In contrast, if $\hat{\tau}_k$ corresponds to a true change point, the two distributions differ significantly. This distinction is visually demonstrated in Fig. 2, where the distributions align closely in the absence of a change point but diverge otherwise. The synthetic data thus serves as a diagnostic tool for detecting false discoveries, thereby enhancing selection power in the validation step.

Moreover, it is worth noting that solely comparing the difference at a fixed point $\hat{\tau}_k$ may not yield the optimal result because $\hat{\tau}_k$ maximizes the difference in the odd part dataset $\mathcal{Z}^O$ instead of the even part dataset $\mathcal{Z}^E$. When the candidate change-points in $\mathcal{Z}^E$ are validated, it is more reasonable to iterate through all possible locations in the interval G_k and select the point with the largest absolute difference. Furthermore, the CUSUM statistics are aggregated via a ℓ_q norm, where q takes value in {1, 2, ..., ∞ } to adapt the change structure under the alternatives. Hence, the comparison statistics for the original dataset $\mathcal{Z}_{G_k}^E$ and the synthetic dataset $\tilde{\mathcal{Z}}_{G_k}$ are defined as:

$$T_k^q = \max_{s \in G_k^*} \|\mathbf{c}_k(s)\|_q \text{ and } \tilde{T}_k^q = \max_{s \in G_k^*} \|\tilde{\mathbf{c}}_k(s)\|_q, \text{ for } q \in \{1, 2, \ldots, \infty\}. \tag{10}$$

where, for a vector $\mathbf{x} \in \mathbb{R}^d$, $\|\mathbf{x}\|_q := \left(\sum_{j=1}^d |x_j|^q \right)^{1/q}$ for $1 \leqslant q < \infty$, $\|\mathbf{x}\|_\infty := \max_{1 \leqslant j \leqslant d} |x_j|$ and G_k^* is a truncated version of G_k, obtained by removing the first $\underline{s}$ and last $\underline{s}$ indices. This truncation is necessary since $\widehat{\mathbf{\Sigma}}^{(k)-}(s)$ and $\widehat{\mathbf{\Sigma}}^{(k)+}(s)$ may not be reliable estimates of $\mathbf{\Sigma}^{(k)}$ when sample sizes within the left or right segments are too small. Therefore, if $\hat{\tau}_k$ is a false discovery, the distribution of $\tilde{T}_k^q$ is expected to closely approximate that of T_k^q, given the similarity between the synthetic and validation data.

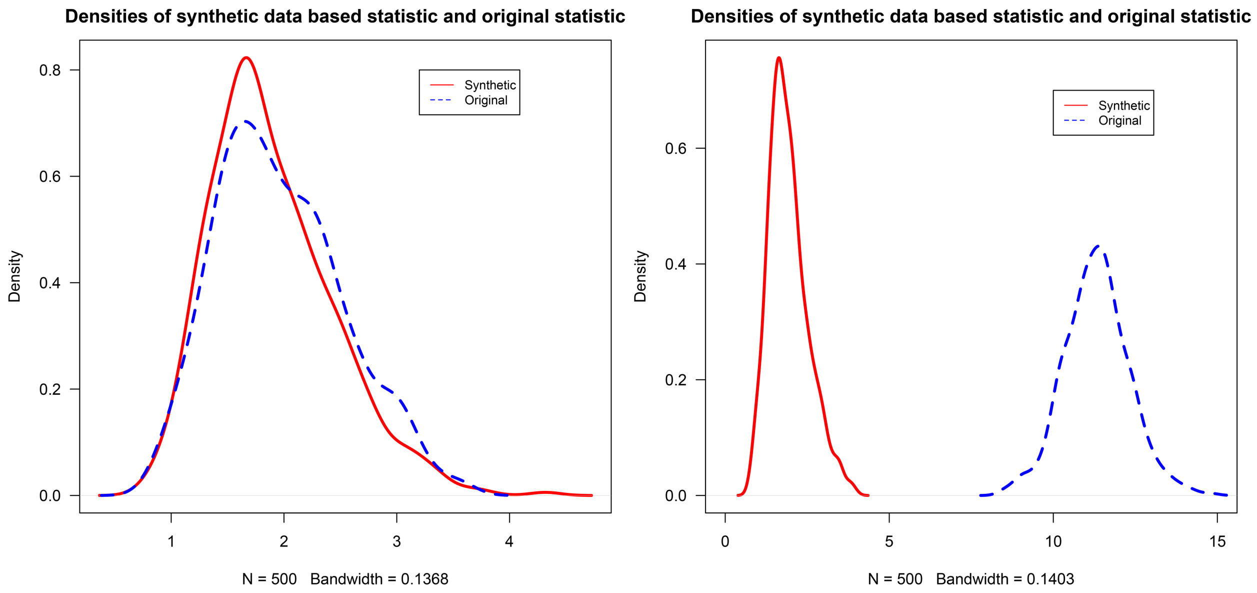

Fig. 2 Densities of the synthetic-data-based statistic and the original statistic under stationary (left) and non-stationary (right) settings. Detailed settings are described in the simulation study.

However, if the candidate change-point $\hat{\tau}_k$ belongs to the informative set, i.e., $\hat{\tau}_k \in \mathcal{I}_1$, then $\Delta_k(s)$ in Eq. (6) should be the leading term in the $\mathbf{c}_k(s)$. Denote

$$\mathbf{r}_k(s) = \sqrt{\frac{s(n_k - s)}{n_k}} \left(\frac{1}{s} \sum_{i \leqslant s, i \in G_k} \boldsymbol{\zeta}_i^E - \frac{1}{n_k - s} \sum_{i > s, i \in G_k} \boldsymbol{\zeta}_i^E \right)$$

then in this case,

$$T_k^q = \max_{s \in G_k^*} \|\mathbf{c}_k(s)\|_q \geqslant \max_{s \in G_k^*} \|\Delta_k(s)\|_q - \max_{s \in G_k^*} \|\mathbf{r}_k(s)\|_q$$

where $\mathbf{r}_k(s), s \in G_k^*$ are all zero mean vectors. Furthermore, the CUSUM statistics defined on the synthetic dataset $\tilde{\mathcal{Z}}_{G_k}$ also have zero mean, as they are transformations of zero-mean random vectors. Hence, the original comparison statistic T_k^q should be sufficiently larger than the synthetic comparison statistic $\tilde{T}_k^q$. That is,

$$T_k^q - \tilde{T}_k^q \geqslant \max_{s \in G_k^*} \|\Delta_k(s)\|_q - \max_{s \in G_k^*} \|\mathbf{c}_k(s)\|_q - \max\nolimits_{s \in G_k^*} \|\tilde{\mathbf{c}}_k(s)\|_q \gg 0$$

provided that the change magnitudes within G_k are sufficiently large and well-separated.

FDR control via SD filter

Since the synthetic data in Eq. (8) mimic the distributional behavior under the null, the distributions of T_k^q and $\tilde{T}_k^q$ are close when $\hat{\tau}_k \in \mathcal{I}_0$. Therefore, a comparison statistic is defined for each candidate change-point $\hat{\tau}_k$ as follows:

$$W_k^q = T_k^q - \tilde{T}_k^q, \text{ for } q \in \{1, 2, \ldots, \infty\} \tag{11}$$

discussed previously.

Moreover, the odd dataset $\mathcal{Z}^O$, used to detect the candidate change-points, provides valuable side information. For each $\hat{\tau}_k$, the CUSUM statistic can also be computed using $\mathcal{Z}_{G_k}^O$, denoted by $T_k^q(\mathcal{Z}_{G_k}^O)$. If the candidate set $\hat{\mathcal{S}}$ is a plausible estimate of the true set $\mathcal{S}$, it is typically observed that $T_k^q(\mathcal{Z}_{G_k}^O) > T_{k'}^q(\mathcal{Z}_{G_{k'}}^O)$ for $\hat{\tau}_k \in \mathcal{I}_1$ and $\hat{\tau}_{k'} \in \mathcal{I}_0$. By incorporating this information, a new statistic $W_k^{side,q}$ is introduced. This blends the original comparison statistic with the side information:

$$W_k^{side,q} = (T_k^q - \tilde{T}_k^q) T_k^q(\mathcal{Z}_{G_k}^O), \tag{12}$$

After incorporating the odd dataset $\mathcal{Z}^O$, the statistic still exhibits symmetry around zero in the case of a false discovery. The advantage of $W_k^{side,q}$ lies in its ability to enhance the separation between the comparison statistics corresponding to informative and uninformative sets. Specifically, when $\hat{\tau}_k \in \mathcal{I}_1$ and $\hat{\tau}_{k'} \in \mathcal{I}_0$, the ratio

$$\frac{W_k^{side,q}}{W_{k'}^{side,q}} = \left(\frac{W_k^q}{W_{k'}^q} \right) \cdot \left(\frac{T_k^q(\mathcal{Z}_{G_k}^O)}{T_k^q(\mathcal{Z}_{G_{k'}}^O)} \right) \geqslant \frac{W_k^q}{W_{k'}^q}, \tag{13}$$

since the second factor is typically greater than or equal to one.

Building on the definitions of W_k^q and $W_k^{side,q}$, the candidate set is refined by selecting positions with large values of the comparison statistics. Specifically, it is defined as:

$$\mathcal{T}(t) = \{\hat{\tau}_k \in \hat{\mathcal{S}} : W_k^q \geqslant t\} \text{ or } \mathcal{T}^{side}(t) = \{\hat{\tau}_k \in \hat{\mathcal{S}} : W_k^{side,q} \geqslant t\} \tag{14}$$

where, $\mathcal{T}(t)$ refered to as the selection set. Given such a set, the number of false discoveries can be estimated by exploiting the symmetric of W_k^q or $W_k^{side,q}$ around zero when $\hat{\tau}_k \in \mathcal{I}_0$. This leads to the following relationship:

$$\#\{\hat{\tau}_k \in \mathcal{I}_0 : W_k^{side,q} \geqslant t\} \approx \#\{\hat{\tau}_k \in \mathcal{I}_0 : W_k^{side,q} \leqslant -t\} \leqslant \#\{\hat{\tau}_k : W_k^{side,q} \leqslant -t\} \tag{15}$$

where, $W_k^{side,q}$ can be replaced by W_k^q. Based on this property, the FDR can be approximated by

$$\mathrm{FDR}(t) \approx \frac{|\mathcal{T}(t) \cap \mathcal{I}_0|}{|\mathcal{T}(t)|} \leqslant \frac{\#\{k : W_k \leqslant -t\}}{\#\{k : W_k \geqslant t\}}$$

To control the FDR at a target level α, the knockoff+ procedure [22] is followed and a data-dependent threshold $T(\alpha)$ is computed as follows:

$$T(\alpha) = \min\nolimits_t \left\{ t \in \mathcal{W} : \frac{1 + \#\{k : W_k \leqslant -t\}}{\#\{k : W_k \geqslant t\} \vee 1} \leqslant \alpha \right\} \tag{16}$$

where, $\mathcal{W} = \{W_1, \ldots W_{\hat{K}_n}\} \backslash \{0\}$ and the extra term 1 in the numerator makes the choice of $T(\alpha)$ slightly conservative.

In summary, compared with pure data-splitting methods such as MOPS and M-MOPS, the core methodological distinction of this approach lies in how the test statistics are constructed. Instead of performing inference on only half of the data, the method used generates a synthetic dataset via a Gaussian multiplier bootstrap, which integrates information from both the odd part and the reserved even part. This design increases the effective sample size and thus improves statistical power. Moreover, although the construction of W_k^q in Eq. (13) differs from that in MOPS and M-MOPS, it serves a conceptually similar role as the comparison statistic used in data-splitting–based procedures. As shown in Eq. (13), the formulation effectively amplifies the contrast between informative and uninformative subsets, which in turn facilitates better control of the false discovery rate.

Theoretical results

Asymptotic theory for FDR control

The error rate control results rely on the symmetry property of the comparison statistics W_k^q and $W_k^{side,q}$ when $\hat{\tau}_k$ is a false discovery. To lay the groundwork for FDR control, it is essential to systematically examine this symmetry property. Before presenting the main theorem, some regular conditions are first imposed.

Condition 3.1 (Moments and tails). *Let $\underline{b}, \overline{b}$, and any vector $\boldsymbol{\vartheta} \in \mathbb{S}^d$, and for all $\ell = 1, 2,\ i = 1, \ldots, 2n$ and $j = 1, \ldots, d$. (1)*$E[(\boldsymbol{\vartheta}^\top \boldsymbol{\zeta}_i)^2] \geqslant \underline{b}$; (2)*$E[(\boldsymbol{\vartheta}^\top \boldsymbol{\zeta}_i)^{\ell+2}] \leqslant \overline{b}^{\ell}$; (3) $\|\boldsymbol{\vartheta}^\top \boldsymbol{\zeta}_i\|_{\psi_1} \leqslant \overline{b}$, where for $\beta \in [1, \infty)$, $\|\cdot\|_{\psi_\beta}$ represents an Orlicz norm.*

Condition 3.2 (Detection ability). *Assume $\hat{K}_n \geqslant K_n$. There exist $\hat{\tau}_{j_1} < \ldots < \hat{\tau}_{j_{K_n}}$ belonging to $\mathcal{T}$ such that $\max_{1<k<K_n} |\hat{\tau}_{j_k} - \tau_k| \leqslant \delta_n$ holds with probability approaching one as $n \to \infty$, where δ_n is some positive sequence.*

Condition 3.3 (Minimum distance). *Assume that $\mathcal{T} \subseteq \mathcal{T}(\omega_n) = \{\mathcal{T} : \min_j(\tau_{j+1} - \tau_j) \geqslant \underline{\lambda}_n\}$, where $\underline{\lambda}_n$ is a positive sequence such that $\underline{\lambda}_n \geqslant n^\eta$ for some $0 < \eta < 1$ and $\underline{\lambda}_n \geqslant 2\delta_n$.*

A sufficient condition for Condition 3.1 (1) is that the minimum eigenvalue of $\boldsymbol{\Sigma}^{(k)}$ is uniformly bounded below for all $k = 1, \ldots, K_n$. This is a common assumption when the dimension d is fixed. Condition 3.1 (2) is a mild moment condition of the linear transformation of $\boldsymbol{\zeta}_i$. Condition 3.1 (3) assumes a sub-exponential tail for the transformation of the residuals. Note that $\|\boldsymbol{\vartheta}^\top \boldsymbol{\zeta}_i\|_{\psi_1} \leqslant \sum_{j=1}^d \|\vartheta_j \zeta_{ij}\|_{\psi_1} \leqslant \sum_{j=1}^d \|\zeta_{ij}\|_{\psi_1}$. Therefore, if $\sum_{j=1}^d \|\zeta_{ij}\|_{\psi_1} \leqslant \overline{b}$, then Condition 3.1 (3) is satisfied. Conditions similar to Condition 3.1 can be found in Liu et al.[1] and Yu & Chen[13] for change-point inference. Condition 3.2 requires that the set of candidate change-points $\hat{\mathcal{S}}$ is sufficiently accurate. Condition 3.3 imposes sparsity on the true change-points. Both Condition 3.2 and Condition 3.3 are also considered in Chen et al.[30].

Let $\overline{n} = \max_{k=1,\ldots,K_n} n_k$ and $\underline{n} = \min_{k=1,\ldots,K_n} n_k$ denote the maximum and minimum sample sizes of the intervals G_k, respectively. The following lemma is now established.

Lemma 3.1. *Assume Condition 3.1, 3.2 and 3.3 holds, then*

$$\Pr\left\{\max_{k\in\mathcal{I}_0}\rho(T_k^q,\tilde{T}_k^q)\leqslant c\left(\frac{\log^7(\overline{n})}{\underline{s}}\right)^{1/6}\mid\mathcal{Z}^O\right\}\geqslant 1-C/(\underline{n})^{\kappa} \qquad (17)$$

where, $\rho(T_1,T_2)=\sup_{t\in(0,\infty)}|\Pr(T_1\leqslant t)-\Pr(T_2\leqslant t)|$ denotes the Kolmogorov distance between T_1 and T_2. Here, $\underline{s}$ is defined as $\min_{k=1,\dots,K_n}\tau_{k+1}-\tau_k$, and κ, C, and c are positive constants.

Lemma 3.1 demonstrates that the distribution of T_k^q approximates that of its original CUSUM counterpart under mild conditions. Based on Lemma 3.1, the main results on FDR control are therefore presented.

Theorem 3.1. *Under Condition 3.1, 3.2 and 3.3, and $\log^7(\hat{K}_n\bar{n})/\underline{s}\to 0$, the SD filter satisfies*

$$\limsup_{n\to\infty}{}^*E\left[\frac{|\mathcal{T}\cap\mathcal{I}_0|}{|\mathcal{T}|}\Big|\mathcal{Z}^O\right]\leqslant\alpha,$$

for any $\alpha\in(0,1)$.

Theorem 3.1 establishes the asymptotic FDR control property of the SD filter. The condition $\log^7(\hat{K}_n\bar{n}d^2)/\underline{s}\to 0$ implies that the bound parameter $\underline{s}$ should not approach the endpoints too closely. This requirement is reasonable, since accurate covariance estimation requires sufficiently large samples on both sides of s. Unlike Theorem 1 in Chen et al.[30], Theorem 3.1 does not require $\delta_n/\underline{\lambda_n}\to 0$, indicating that the SD filter does not rely on highly accurate candidate localization.

The FDR control for the general ℓ_q-norm cannot be extended to high-dimensional MCP settings due to the constraints of high-dimensional central limit theorem for simple and sparse convex sets, as discussed in Chernozhukov et al.[32]. To address the high-dimensional MCP challenges, a practical approach is to set $q=\infty$. In this case, the high-dimensional central limit theorem for hyperrectangles can be employed to justify Lemma 3.1 even in high-dimensional MCP scenarios.

Theorem 3.2 (FDR control for high-dimensional MCP). *When $d\to\infty$, under Condition 3.1, 3.2 and 3.3, and $\log^7(\hat{K}_n\bar{n}d^2)/\underline{s}\to 0$. Let $q=\infty$, the SD filter satisfies*

$$\limsup_{n\to\infty}{}^*E\left[\frac{|\mathcal{T}\cap\mathcal{I}_0|}{|\mathcal{T}|}\Big|\mathcal{Z}^O\right]\leqslant\alpha$$

for any $\alpha\in(0,1)$.

Power analysis

Next, the power of the SD filter is analyzed under the following signal condition.

Condition 3.4 (Minimum signal). *Let $\boldsymbol{\delta}_k=\boldsymbol{\mu}_{k+1}-\boldsymbol{\mu}_k$ be the change magnitude, which satisfies*

$$\min_{k\in\mathcal{I}_1}\|\boldsymbol{\delta}^{(k)}\|_q\gg C\bar{\sigma}^2\sqrt{\frac{\log(\alpha_n\hat{K}_n\bar{n}d)}{t_k(1-t_k)\underline{n}}}$$

where, $t_k=\tau_k/n_k$, α_n is a sequence that converges to infinity with a slow rate and C is a positive constant.

Condition 3.4 imposes a minimum signal separation between any two true change-points, ensuring their asymptotic identifiability. Similar conditions can be found in Harchaoui & Lévy-Leduc[5], Fryzlewicz[6], and Yu & Chen[13].

Theorem 3.3. *Under Condtion 3.1, 3.2, 3.3 and 3.4 and $\log^7(\hat{K}_n\bar{n})/\underline{s}\to 0$. The power of SD filter satisfies*

$$\lim_{n\to\infty}{}^*E\left[\frac{|\mathcal{T}\cap\mathcal{I}_1|}{|\mathcal{I}_1|}\Big|\mathcal{Z}^O\right]=1$$

Theorem 3.3 states that the power of SD filter approaches 1 asymptotically. Furthermore, the selection consistency property can be established.

Corollary 3.1. *Under Conditions in Theorem 3.3, there is*

$$\lim_{n\to\infty}\Pr\left\{\mathcal{S}=\mathcal{T}\mid\mathcal{Z}^O\right\}=1$$

Compared with the condition for selection consistency in Chen et al.[30], which requires $\min_{k\in\mathcal{I}_1}\|\boldsymbol{\delta}^{(k)}\|_2\gg\sqrt{\log n/\underline{\lambda}_n}$, Condition 3.4 is of approximately the same order, noting that $\underline{\lambda}_n\approx\underline{n}$.

Simulation study

In this section, aseries of change-point detection experiments is conducted to evaluate the empirical performance of the SD filter. Before presenting the results, the competing methods, the Mirror with Order-Preserved Splitting (MOPS) method and its variant, the Modified-MOPS (M-MOPS) are briefly summarized, both introduced in Chen et al.[30].

The M-MOPS method controls the FDR via a mirror statistic

$$W_k^{\text{M-MOPS}}=\frac{n_kn_{k+1}}{n_k+n_{k+1}}\left(\overline{\mathbf{S}}_k^{O,-}-\overline{\mathbf{S}}_k^{O,+}\right)^{\top}\Omega_n\left(\overline{\mathbf{S}}_k^{E,-}-\overline{\mathbf{S}}_k^{E,+}\right),\quad k=1,\dots,\hat{K}_n$$

where, $\overline{\mathbf{S}}_k^{O,-}$, $\overline{\mathbf{S}}_k^{O,+}$, $\overline{\mathbf{S}}_k^{E,-}$, $\overline{\mathbf{S}}_k^{E,+}$ are previously defined, and Ω_n is a positive matrix. Since the performance of M-MOPS is not sensitive to Ω_n, the study sets $\Omega_n=\mathbf{I}_d$, the *d*-dimensional identity matrix, in the simulations. Compared with M-MOPS, the MOPS statistic differs only in that it uses a larger sample to compute the sample means of the score functions:

$$W_k^{\text{MOPS}}=\frac{n_kn_{k+1}}{n_k+n_{k+1}}\left(\tilde{\mathbf{S}}_k^{O,-}-\tilde{\mathbf{S}}_k^{O,+}\right)^{\top}\Omega_n\left(\tilde{\mathbf{S}}_k^{E,-}-\tilde{\mathbf{S}}_k^{E,+}\right),\quad k=1,\dots,\hat{K}_n$$

where, $\tilde{\mathbf{S}}_k^{O,-}$ and $\tilde{\mathbf{S}}_k^{O,+}$ denote the sample means of $\{\mathbf{s}_i^O,\hat{\tau}_{k-1}<i\leqslant\hat{\tau}_k\}$ and $\{\mathbf{s}_i^O,\hat{\tau}_k<i\leqslant\hat{\tau}_{k+1}\}$, respectively; $\tilde{\mathbf{S}}_k^{E,-}$ and $\tilde{\mathbf{S}}_k^{E,+}$ are defined analogously using the even subsample.

Then the computational complexity of the methods is compared. Treating basic arithmetic operations as *O*(1), computing $W_k^{\text{M−MOPS}}$ requires *O*(*dn*) opreations, where *d* is the data dimension. Similarly, the quantities $\mathbf{c}_k(s)$ and $\tilde{\mathbf{c}}_k(s)$ also involve *O*(*dn*) operations, implying that both T_k^q and $\tilde{T}_k^q$, and therefore $W_k^{side,q}$, have the same complexity. This method generates only a single set of multiplier bootstrap samples and therefore adds no extra computational overhead. In practice, one may run B bootstrap repetitions, yielding a cost of *O*(*Bdn*); however, the procedure can be parallelized with ease, so the runtime can approach that of a single iteration. Overall, the SD filter does not incur high computational cost.

Beyond computational considerations, an important issue is statistical reliability. In particular, MOPS may fail to control the FDR when the discrepancy between the candidate and true change-point sets is large. To assess the performance of all methods, 200 simulation replications were conducted and each method was evaluated using the empirical FDR and power:

$$\widehat{\text{FDR}}=\frac{1}{200}\sum_{i=1}^{200}\frac{|\mathcal{T}_i\cap\mathcal{I}_0|}{|\mathcal{T}_i|}\text{ and }\widehat{\text{Power}}=\frac{1}{200}\sum_{i=1}^{200}\frac{|\mathcal{T}_i\cap\mathcal{I}_1|}{|\mathcal{I}_1|}$$

where, $\mathcal{T}_i$ denotes the estimated selection set in the ith replication.

The detailed pseudocode of this approach is as the Algorithm 1.

Simulation for multiple mean changes model

Consider a sequence of d-dimensional mean vectors $\boldsymbol{\mu}_i, i=1,\dots,n$, where the means are piecewise constant with change-points at positions $\{\tau_k,k=1,\dots,K\}$. Specifically,

$$\boldsymbol{\mu}_i=\boldsymbol{\mu}_{\tau_k},\text{ for }\tau_k+1\leqslant i\leqslant\tau_{k+1}$$

The sequence is initialized with $\boldsymbol{\mu}_{\tau_1}=(A/2)\mathbf{1}_d$, where $\mathbf{1}_d$ denotes the d-dimensional vector of ones. To define $\boldsymbol{\mu}_{\tau_2}$, randomly selecting

Algorithm 1. Synthetic data filter (SD filter) for MCP detection.

```
Input : Observed data sequence $\mathcal{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_{2n}\}$, target FDR level $\alpha$, suitable candidate change-point detection algorithm $\mathcal{A}(\cdot)$
Output : Selected change-point set $\mathcal{T}$
Split data into odd and even parts: $\mathcal{Z}^O = \{\mathbf{z}_1, \mathbf{z}_3, \dots, \mathbf{z}_{2n-1}\}$, $\mathcal{Z}^E = \{\mathbf{z}_2, \mathbf{z}_4, \dots, \mathbf{z}_{2n}\}$
Detect candidate change-points $\hat{\mathcal{S}} = \{\hat{\tau}_1, \dots, \hat{\tau}_{\hat{K}_n}\}$ through $\mathcal{A}(\mathcal{Z}^O)$
for $k \in 1, \cdots, \hat{K}_n$ do
    Define $G_k := [\lceil(\hat{\tau}_{k-1} + \hat{\tau}_k)/2\rceil, \lceil(\hat{\tau}_k + \hat{\tau}_{k+1})/2\rceil)$, where $\hat{\tau}_0 = 0$, $\hat{\tau}_{\hat{K}_n} = 2n$ $\mathcal{Z}^O_{G_k} := \{\mathbf{z}_{2i-1} : i \in G_k\}$, $\mathcal{Z}^E_{G_k} := \{\mathbf{z}_{2i} : i \in G_k\}$
    Compute score function $\mathbf{s}_i^E$ and $\mathbf{c}_k(s) = \sqrt{\frac{s(n_k - s)}{n_k}} \left( \frac{1}{s} \sum_{i \leqslant s, i \in G_k} \mathbf{s}_i^E - \frac{1}{n_k - s} \sum_{i > s, i \in G_k} \mathbf{s}_i^E \right)$
    Generate synthetic data $\tilde{\mathcal{Z}}_{G_k}$ through $\left\{ \xi_i \left( \mathbf{s}_i^O - \bar{\mathbf{s}}_k^{O,-}(s) \right), i \leqslant s, i \in G_k \right\}$ and $\left\{ \xi_i \left( \mathbf{s}_i^O - \bar{\mathbf{s}}_k^{O,+}(s) \right), i > s, i \in G_k \right\}$, where $\xi_i \sim N(0,1)$
    Compute synthetic CUSUM $\tilde{\mathbf{c}}_k(s) = \sqrt{\frac{s(n_k - s)}{n_k}} \left( \frac{1}{s} \sum_{i \leqslant s, i \in G_k} \xi_i \left( \mathbf{s}_i^O - \bar{\mathbf{s}}_k^{O,-}(s) \right) - \frac{1}{n_k - s} \sum_{i > s, i \in G_k} \xi_i \left( \mathbf{s}_k^O - \bar{\mathbf{s}}_k^{O,+}(s) \right) \right)$
    Calculate: $T_k^q = \max_{s \in G_k^*} \|\mathbf{c}_k(s)\|_q$, $\tilde{T}_k^q = \max_{s \in G_k^*} \|\tilde{\mathbf{c}}_k(s)\|_q$ and $W_k^{side,q} = (T_k^q - \tilde{T}_k^q) \cdot T_k^q(\mathcal{Z}^O_{G_k})$
end for
Obtain threshold $T(\alpha)$:
    $T(\alpha) = \min \left\{ t : \frac{1 + \#\{k : W_k^{side,q} \leqslant -t\}}{\#\{k : W_k^{side,q} \geqslant t\} \vee 1} \leqslant \alpha \right\}$
Select final change-point set: $\mathcal{T}^{side} = \{\hat{\tau}_k \in \hat{\mathcal{S}} : W_k^{side,q} \geqslant T(\alpha)\}$
return $\mathcal{T}^{side}$
```

r positions in $\boldsymbol{\mu}_{\tau_1}$ and flip the sign of those entries. Subsequent vectors $\boldsymbol{\mu}_{\tau_k}$ are constructed recursively from their predecessors using the same procedure. Under this setup, the study has

$$\|\boldsymbol{\mu}_{\tau_k} - \boldsymbol{\mu}_{\tau_{k+1}}\|_\infty = A, \quad k = 1, \dots, K$$

The data points are generated as $\mathbf{z}_i = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_i, i = 1, \dots, n$, and the signal strength is controlled by the value of A. Throughout this subsection, the dimensionality is set to $d = 50$, the sample size to $n = 4{,}000$. The true change-points are set to $\tau_k = 200\,k$, $k = 1, \dots, 19$, in Example 1, and to $\tau_k = 400\,k$, $k = 1, \dots, 9$, in Example 2. To mitigate the undesired bias from detection algorithms, a sequence of candidate change-points are manually constructed as $\mathcal{T} = \{150k + (-1)^{\mathrm{B}_k}\mathrm{P}_k \mid k = 1, \dots 26\}$, where B_k, P_k are independently drawn from $\mathrm{Bernoulli}\,(1/2)$ and $\mathrm{Poisson}\,(5)$, respectively. Theoretically, $q = \infty$ is required in high-dimensional settings, while any $q \geqslant 1$ is admissible in low-dimensional cases. Since the optimal choice of q depends on unknown signal conditions, $q = \infty$ is adopted as a practical default in both simulations and empirical analysis. For the truncation size, $s \in [10, 30]$ is chosen to ensure that both s and $n_k - s$ are sufficiently large, thereby mitigating boundary effects and improving the stability of the algorithm. In addition, the study sets $r = 1$ and FDR level $\alpha = 0.15$.

Example 1: Normal distribution

Consider the error term $\boldsymbol{\epsilon}_i$ to be drawn from multivariate normal distribution with covariance matrix $\boldsymbol{\Sigma} = \{\rho^{|i-j|}\}_{(i,j)}$. The study chose the change magnitude A and the correlation coefficient ρ as follows:

- Fix $\rho = 0$, and let A vary in {1.5, 1.7, 1.9, 2.1, 2.3, 2.5}.
- Fix $A = 1.5$, and let ρ vary in {0, 0.2, 0.4, 0.6, 0.8}.

Example 2: Beyond normal distribution

Consider the error term $\boldsymbol{\epsilon}_i$ to be drawn from either a multivariate t distribution or a multivariate chi-square distribution, each having a covariance matrix $\boldsymbol{\Sigma} = \mathbf{I}_d$. We choose the change magnitude A and the degrees of freedom *df* as follows:

- Fix $A = 2$, and let *df* vary in {8, 9, 10, 11, 12}.
- Fix $A = 3$, and let *df* vary in {3, 4, 5, 6, 7}.

The simulation results summarized in Figs 3 and 4 demonstrate the superior performance of the SD filter across various signal strengths and dependence structures. The results indicate that both the SD filter and M-MOPS exhibit the capability to maintain control over the false discovery rate (FDR) at the predetermined level. In contrast, MOPS struggles to maintain FDR control due to the deliberate reduction in the quality of the candidate change-point set. When the candidate change positions are significantly distant from the actual change-points, MOPS fails to perform effectively. In terms of empirical power, the SD filter consistently outperforms the M-MOPS method across all settings.

Structural breaks in linear regression model

Consider a linear regression model with structural breaks, defined as $\mathbf{y}_i = \mathbf{x}_i^\top \boldsymbol{\beta}_{\tau_k} + \epsilon_i$, for $\tau_{k-1} \leqslant i \leqslant \hat{\tau}_k$. Initially, let $\boldsymbol{\beta}_{\tau_1} = (A/2)\mathbf{1}_d$. Then, define $\boldsymbol{\beta}_{\tau_2}$ by randomly selecting s positions in $\boldsymbol{\beta}_{\tau_1}$ and flipping the signs of those entries. Each subsequent vector $\boldsymbol{\beta}_{\tau_k}$ is generated from $\boldsymbol{\beta}_{\tau_{k-1}}$ using the same procedure. Under this setup, the study has

$$\|\boldsymbol{\beta}_{\tau_k} - \boldsymbol{\beta}_{\tau_{k+1}}\|_\infty = A, \quad k = 1, \dots, K.$$

The covariates $\mathbf{x}_i$ are drawn from a multivariate normal distribution $N(\mathbf{0}_d, \boldsymbol{\Sigma})$, where $\mathbf{0}_d$ is the *d*-dimensional vector of zeros and the covariance matrix $\boldsymbol{\Sigma} = \{\rho^{|i-j|}\}_{(i,j)}$. The error terms ϵ_i are independently drawn from $N(0,1)$. In this model, the sample size is set as $n =$ 8,000 and the number of covariates d = 10. The true change-point set is defined as $\mathcal{S} = \{1000k, k = 1, \dots, 7\}$, and the candidate change-point set is given by $\hat{\mathcal{S}} = \{450k + (-1)^{\mathrm{B}}_{\mathrm{k}}\mathrm{P}_k \mid k = 1, \dots, 16\}$, where $\mathrm{B}_k \sim \mathrm{Bernoulli}(1/2)$ and $\mathrm{P}_k \sim \mathrm{Poisson}\,(5)$. Except that the FDR level is set to $\alpha = 0.2$, all other parameters remain unchanged. The change magnitude A and correlation coefficient ρ were chosen as follows:

- Fix $\rho = 0$, and let A vary in {0.20, 0.22, 0.24, 0.26, 0.28, 0.30}.
- Fix $A = 0.25$, and let ρ vary in {0, 0.1, 0.2, 0.3, 0.4, 0.5}.

The simulation results presented in Figs 5 and 6 display the estimated FDR and empirical power results for the linear regression

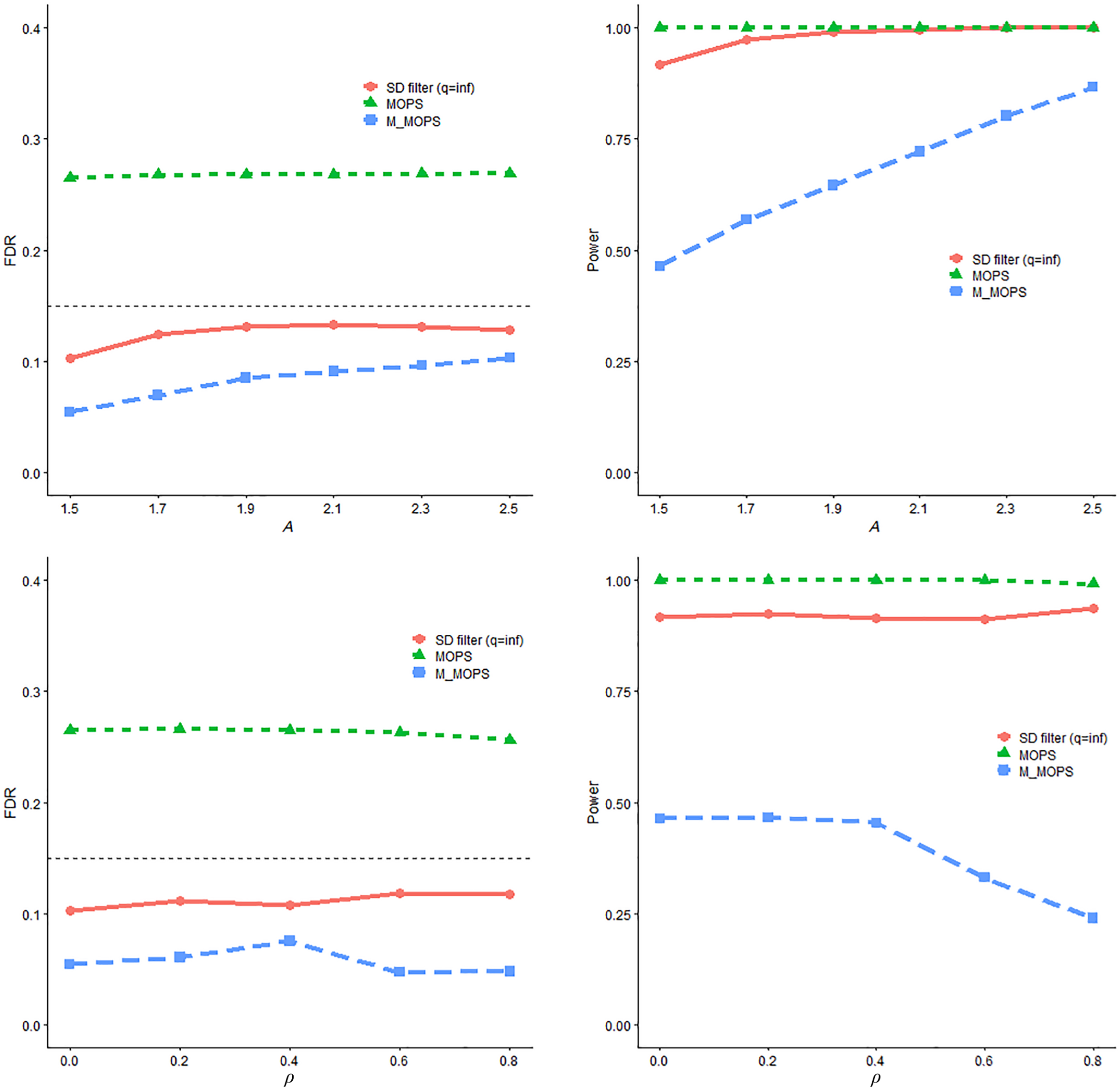

Fig. 3 FDR and power trends with respect to A and ρ for SD filter, MOPS, and M-MOPS under the mean change model, where $n = 4{,}000$, d = 50, $\alpha = 0.15$.

model with structural breaks under various scenarios. As before, it was observed that both the SD filter and M-MOPS successfully control the FDR at the pre-specified level. Again, MOPS fails to do so due to the deliberately reduced quality of the candidate change-point set. In terms of empirical power, the SD filter still consistently outperforms M-MOPS in this setting.

Bladder tumor micro-array analysis

In this section, the proposed SD filter is applied to analyze the bladder tumor micro-array dataset sourced from Bleakley & Vert[35], which is conveniently available in the *ecp* R package. The dataset consists of log–intensity-ratio measurements for 2,215 genetic loci obtained from 43 individuals diagnosed with bladder tumors. The primary objective is to identify change-points within the genetic loci, enabling the study to pinpoint potential influential genes related to bladder tumors. This dataset has been widely used as a benchmark in several prior studies on change-point detection[35–37], making it a representative and well-established dataset for evaluating the empirical performance of new methods. The analysis was conducted on the full dataset. However, for visualization purposes and to provide a clearer and more interpretable presentation of the results, the findings for the first ten individuals were reported, specifically individuals 3, 4, 5, 6, 7, 8, 9, 10, 14, and 15.

Firstly, the inspect method[9] is applied to narrow the scope and obtain a candidate set. To ensure optimal performance, a minimum difference of 50 was established between two change-points. The set of identified change-points is as follows:

$$\begin{aligned}\hat{S} = \{&73, 263, 428, 669, 811, 960,\\ &1050, 1378, 1436, 1559, 1724, 1831, 1906, 2084\}\end{aligned}$$

Subsequently, the SD filter, MOPS and M-MOPS are applied to further refine the results, controlling the FDR at a level of 0.1. The final sets of detected change-points from MOPS and M-MOPS are identical to $\hat{S}$, whereas the set of change-points detected by the SD filter is as follows:

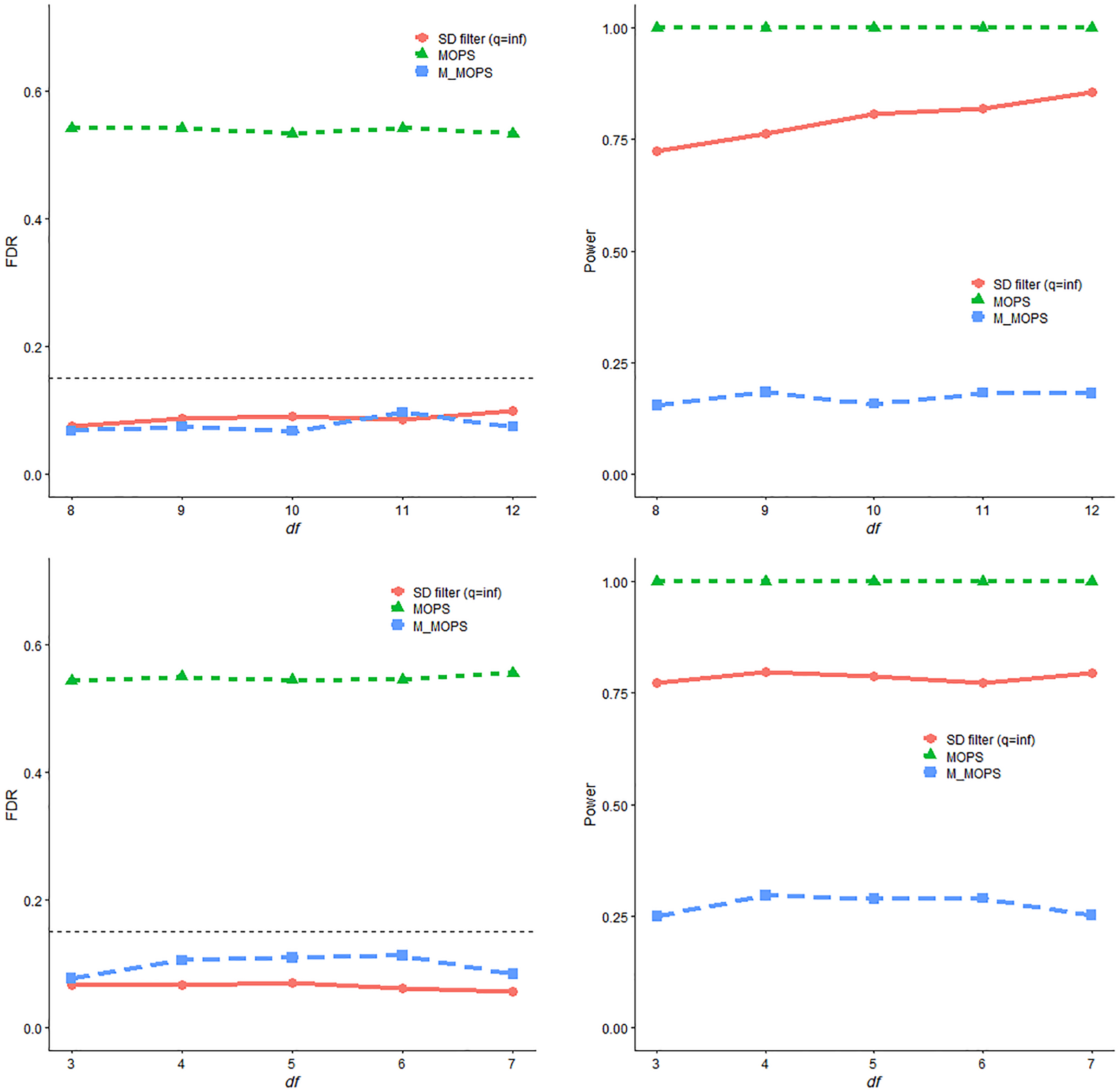

Fig. 4 FDR and power trends with respect to *df* of the t-distribution and chi-square distribution for SD filter, MOPS, and M-MOPS under the mean change model, where $n = 4{,}000$, d = 50, $\alpha = 0.15$.

$$\mathcal{T}_{SD} = \{73, 263, 428, 669, 811, 960, 1050, 1378, 1436, 1559, 1724, 1906, 2084\}$$

This result of SD filter excludes the position 1,831. Figure 7 visually demonstrates the change-points identified through the SD filter. In each plot, individual data points represent log-intensity ratios on a specific genetic locus, with each plot corresponding to a different test subject. Vertical lines are used to indicate the locations of detected change-points. The change-points identified by the SD filter are shown as dashed lines. Notably, the only solid line—positioned at 1,831—does not correspond to any apparent change across the ten individuals, highlighting the greater precision and accuracy of the SD filter in identifying true change-points.

Conclusions

To overcome the limitations of existing FDR control methods for multiple change-point detection-particularly the drawbacks associated with data-splitting approaches, the study proposes a synthetic data filter (SD filter) for change-point detection and FDR control. After identifying potential change-points, Gaussian multiplier bootstrap is applied to generate synthetic data based on information from the detection dataset. This synthetic data is then used to construct a mirror statistic that enables control of the FDR, offering the flexibility to leverage information from the entire dataset and improve statistical power under a variety of alternatives and dimensions. The symmetry property of the mirror statistic is then established andits ability to rigorously control FDR asymptotically is proven. The detection power is also demonstrated under mild conditions. Simulation studies empirically verified the outstanding performance of the SD filter in terms of FDR control and power. The study also applies the proposed method to analyze a micro array dataset that describes the change loci of bladder tumor patients. As mentioned above, the framework of the SD filter is quite general, and it would be interesting to extend it to a broader range of cases where data splitting is required for formulating and testing hypotheses within the same dataset.

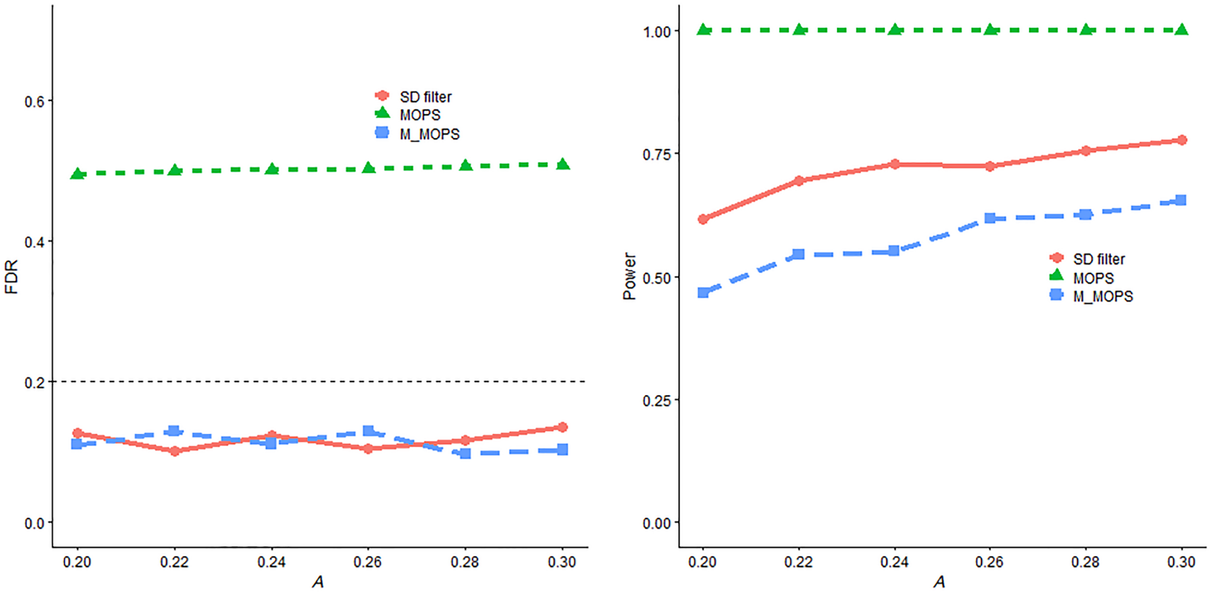

Fig. 5 FDR and power trends with respect to A for SD filter, MOPS, and M-MOPS under the structural breaks linear regression model, where n = 8,000, d = 10, α = 0.2.

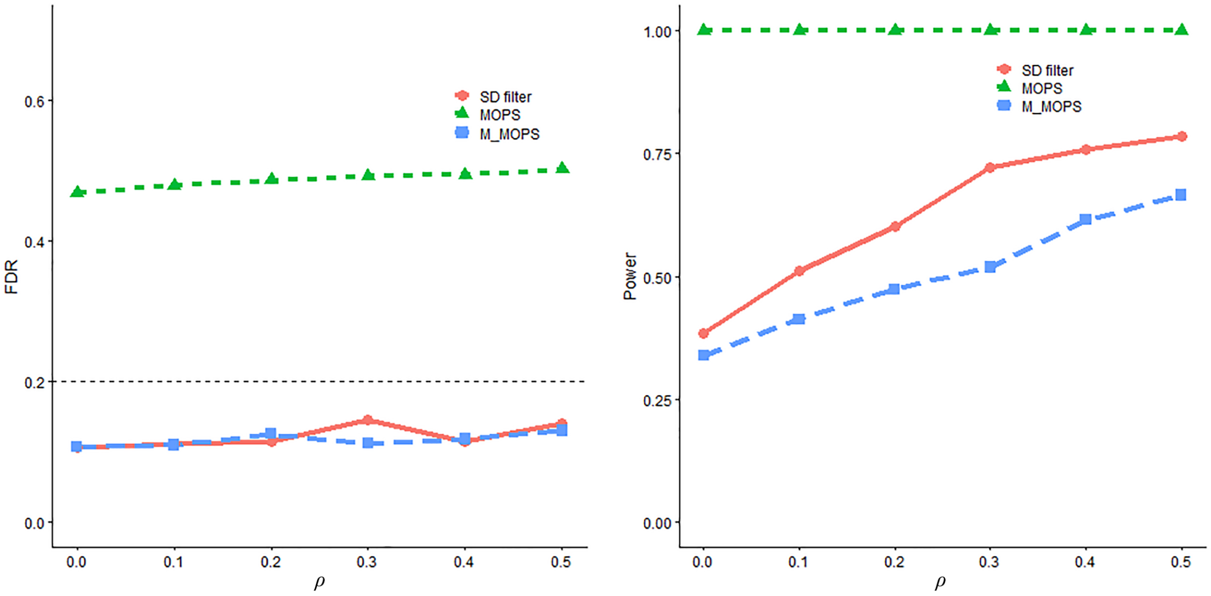

Fig. 6 FDR and power trends with respect to ρ for SD filter, MOPS and M-MOPS under the structural breaks linear regression model, where n = 8,000, d = 10, α = 0.2.

Ethical statements

This study uses publicly available and anonymized data from the *ecp* R package. As the data are de-identified and distributed for open research purposes, no ethical approval was required from the authors' institution. All analyses were conducted in accordance with standard ethical guidelines for statistical research and data use.

Author contributions

The authors confirm contribution to the paper as follows: study conception and design: Sun A, Liu J; data collection: Sun A; analysis and interpretation of results: Sun A, Bi J, Liu J; draft manuscript preparation: Sun A, Bi J, Liu J. All authors reviewed the results and approved the final version of the manuscript.

Data availability

The data that support the findings of this study are available through the *ecp* R package.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant Nos 12271456 and 71988101), the Ministry of Education Research in the Humanities and Social Sciences (Grant No. 22YJA910002). The authors sincerely thank the editor and the referees for their constructive comments and helpful suggestions.

Conflict of interest

The authors declare that they have no conflict of interest.

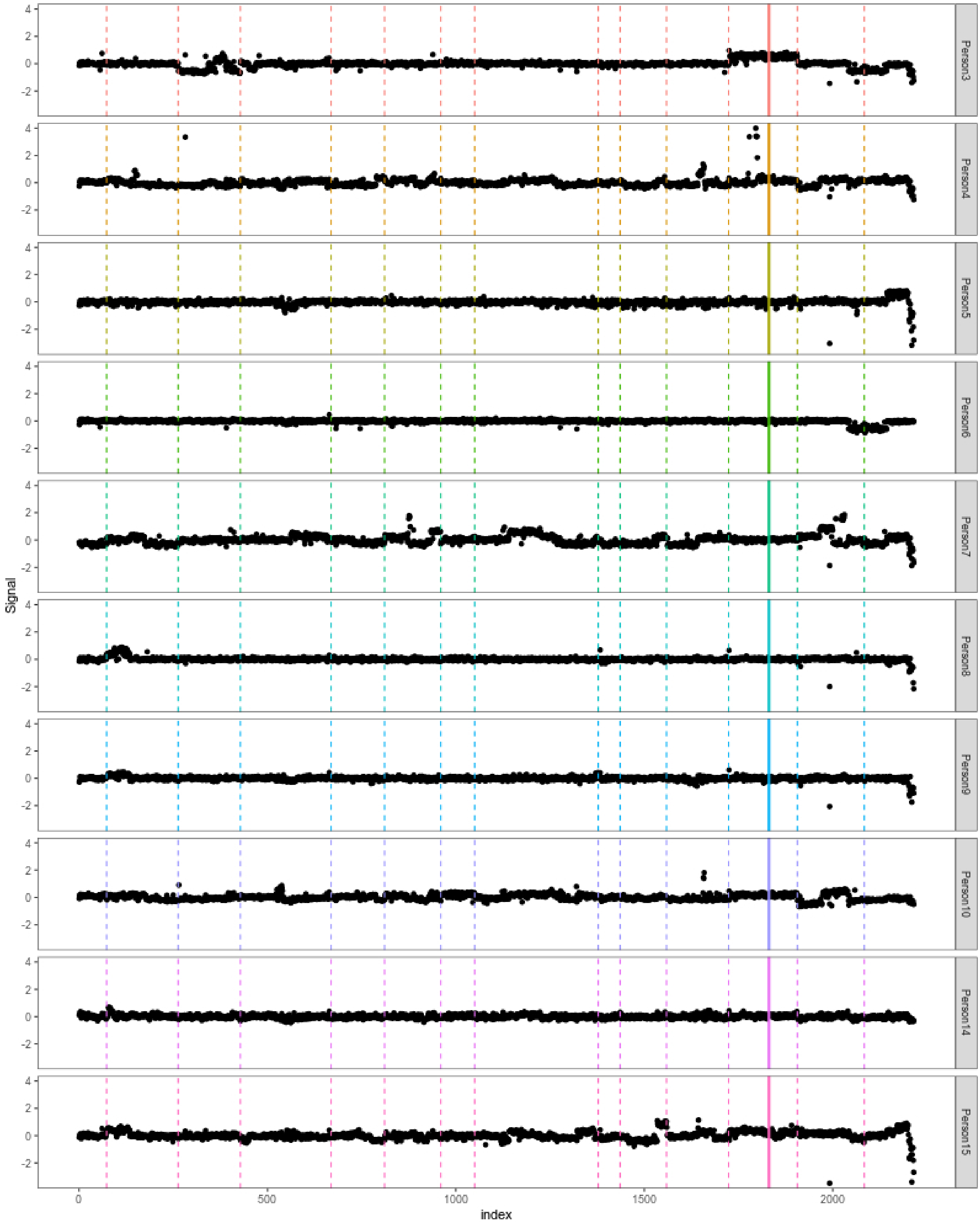

Fig. 7 Detected change-points on bladder tumor micro-array dataset (first ten persons are presented).

Dates

Received 22 May 2025; Revised 15 December 2025; Accepted 29 December 2025; Published online 28 February 2026

References

[1] 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 Statistical Methodology* 82:933–963

[2] Liu B, Zhang X, Liu Y. 2024. Simultaneous change point detection and identification for high dimensional linear models. *arXiv* 2401.08173

[3] Aue A, Horváth L. 2013. Structural breaks in time series. *Journal of Time Series Analysis* 34:1–16

[4] Niu YS, Hao N, Zhang H. 2016. Multiple change-point detection: a selective overview. *Statistical Science* 31(4):611–623

[5] Harchaoui Z, Lévy-Leduc C. 2010. Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association* 105:1480–1493

[6] Fryzlewicz P. 2014. Wild binary segmentation for multiple change-point detection. *The Annals of Statistics* 42:2243–2281

[7] Cho H, Fryzlewicz P. 2015. Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B Statistical Methodology* 77:475–507

[8] Lee S, Seo MH, Shin Y. 2016. The lasso for high dimensional regression with a possible change point. *Journal of the Royal Statistical Society: Series B Statistical Methodology* 78:193–210

[9] Wang T, Samworth RJ. 2018. High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society: Series B Statistical Methodology* 80:57–83

[10] Enikeeva F, Harchaoui Z. 2019. High-dimensional change-point detection under sparse alternatives. *The Annals of Statistics* 47:2051–2079

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

[12] Wang D, Zhao Z, Lin KZ, Willett R. 2021. Statistically and computationally efficient change point localization in regression settings. *Journal of Machine Learning Research* 22:1–46

[13] 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 Statistical Methodology* 83:247–270

[14] Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: Series B Statistical Methodology* 57:289–300

[15] Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* 29:1165–1180

[16] Storey JD. 2002. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B Statistical Methodology* 64:479–98

[17] Genovese C, Wasserman L. 2004. A stochastic process approach to false discovery control. *The Annals of Statistics* 32:1035–1061

[18] Storey JD, Taylor JE, Siegmund D. 2004. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B Statistical Methodology* 66:187–205

[19] Hao N, Niu YS, Zhang H. 2013. Multiple change-point detection via a screening and ranking algorithm. *Statistica Sinica* 23:1553–1572

[20] Li H, Munk A, Sieling H. 2016. FDR-control in multiscale change-point segmentation. *Electronic Journal of Statistics* 10:918–959

[21] Cheng D, He Z, Schwartzman A. 2020. Multiple testing of local extrema for detection of change points. *Electronic Journal of Statistics* 14:3705–3729

[22] Barber RF, Candès EJ. 2015. Controlling the false discovery rate via knockoffs. *The Annals of Statistics* 43:2055–85

[23] Candès E, Fan Y, Janson L, Lv J. 2018. Panning for gold: 'model-X'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B Statistical Methodology* 80:551–577

[24] Barber RF, Candès EJ. 2019. A knockoff filter for high-dimensional selective inference. *The Annals of Statistics* 47:2504–2537

[25] Fan Y, Demirkaya E, Li G, Lv J. 2020. RANK: large-scale inference with graphical nonlinear knockoffs. *Journal of the American Statistical Association* 115:362

[26] Barber RF, Candès EJ, Samworth RJ. 2020. Robust inference with knockoffs. *The Annals of Statistics* 48:1409–1431

[27] Liu J, Sun A, Ke Y. 2024. A Generalized knockoff procedure for FDR control in structural change detection. *Journal of Econometrics* 239:105331

[28] Du L, Guo X, Sun W, Zou C. 2023. False discovery rate control under general dependence by symmetrized data aggregation. *Journal of the American Statistical Association* 118:607–621

[29] Dai C, Lin B, Xing X, Liu JS. 2023. False discovery rate control via data splitting. *Journal of the American Statistical Association* 118:2503–2520

[30] Chen H, Ren H, Yao F, Zou C. 2023. Data-driven selection of the number of change-points via error rate control. *Journal of the American Statistical Association* 118:1415–1428

[31] Chernozhukov V, Chetverikov D, Kato K. 2013. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics* 41:2786–2819

[32] Chernozhukov V, Chetverikov D, Kato K. 2017. Central limit theorems and bootstrap in high dimensions. *The Annals of Probability* 45:2309–2352

[33] Zou C, Wang G, Li R. 2020. Consistent selection of the number of change-points via sample-splitting. *The Annals of Statistics* 48:413–439

[34] Yao YC, Au ST. 1989. Least-squares estimation of a step function Sankhyā. *The Indian Journal of Statistics (Series A)* 51:370–381

[35] Bleakley K, Vert JP. 2011. The group fused lasso for multiple change-point detection. *arXiv* 1106.4199

[36] James NA, Matteson DS. 2015. ecp: An R package for nonparametric multiple change point analysis of multivariate data. *Journal of Statistical Software* 62:1–25

[37] Matteson DS, James NA. 2014. A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association* 109:334