Statistica Sinica Preprint No: SS-2023-0216	
Title	Simultaneous Jump Detection for Multiple Sequences via Screening and Multiple Testing
Manuscript ID	SS-2023-0216
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0216
Complete List of Authors	Shengji Jia and Chunming Zhang
Corresponding Authors	Chunming Zhang
E-mails	cmzhang@stat.wisc.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

SIMULTANEOUS JUMP DETECTION FOR MULTIPLE SEQUENCES VIA SCREENING AND MULTIPLE TESTING

Shengji Jia and Chunming Zhang

Shanghai Lixin University of Accounting and Finance

University of Wisconsin-Madison

Abstract: The estimation of nonparametric discontinuous regression function is fundamental in many applied fields, but challenges arise when the number of jumps (or discontinuities) is large and unknown. We propose a new jump detection method, via the consecutive screening and multiple testing (SaMT) algorithm, for simultaneously estimating the unknown number of jump points and detecting their locations in the flexible nonparametric regression model, guaranteeing the desired accuracy. The initial jump candidates are obtained in the *consecutive screening* procedure combined with locally-linear smoothing method. To further assess the significance of an individual jump candidate, we develop a novel test based on profile likelihood inference. The ultimate selection of relevant jump points is conducted in a *multiple testing* procedure, which eliminates irrelevant jump points with large variations, due to heteroscedastic errors, from jump candidates. Moreover, we generalize the SaMT algorithm to detect common jump points shared across multiple aligned sequences. The proposed method is easy

Corresponding author: Chunming Zhang. E-mail: *cmzhang@stat.wisc.edu*.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

to implement, flexible in bandwidth and threshold selection, and outperforms existing approaches in simulations and real-data applications.

Key words and phrases: Copy number variation (CNV), False discovery rate (FDR), Jump regression analysis, Locally-linear smoothing, Wald statistic.

1. Introduction

In many areas, such as environmental statistics, genetics, finance and engineering, very long and noisy sequences of data will arise. Examples include high-throughput sequencing data in genetics and high-frequency financial data in econometrics. Such data usually have discontinuities or jumps which are important data structures, and the goal is to identify and understand structural variations, e.g., from a normal number to an excessive number of chromosome copies in genetics, or structural changes in trends, e.g., from a bull market to a bear market in finance. Thus, identifying multiple jump points in the regression function is a fundamental and challenging problem.

In the literature, the issue of jump detection was investigated using different approaches, including locally-linear smoothing (Grégoire and Hamrouni (2002); Xia and Qiu (2015)), splines (Ma and Yang (2011)), and wavelets (Wang (1995); Fan and Wang (2007)). Qiu and Yandell (1998); Joo and Qiu (2009) investigated the derivatives of a regression function

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

to detect jumps, while Eubank and Speckman (1994) developed a different method based on semiparametric model. Besides, Gijbels et al. (2004) proposed a two-step procedure to improve the efficiency of jump points estimators. Comprehensive reviews and comparisons of these methods are given by Qiu (2005).

In the context of change-point detection for the parametric piecewise-constant model, some particular algorithms have also been developed, including the binary segmentation algorithm (Olshen et al. (2004); Korkas and Fryzlewicz (2017)), moving sum (MOSUM) techniques (Eichinger and Kirch (2018)), and cumulative segmented model (Muggeo and Adelfio (2011)). Niu and Zhang (2012) proposed a screening and ranking algorithm (SaRa) based on locally-constant smoothing. Besides, the error rate control methods have been studied by Hao et al. (2013) and Li et al. (2016). Other approaches rely on the penalized least squares regression, using the L_1 penalty (Harchaoui and Lévy-Leduc (2010)), or combined penalties as in the fused LASSO approach (Tibshirani and Wang (2008)).

Recently, the problem of detecting common change points that occur at the same location in multiple noisy sequences has garnered significant attention, as pooling data across samples can enhance the power to detect simultaneously occurring signals. Many authors have addressed this

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

problem using different approaches. For example, Zhang et al. (2010) proposed the multi-sample scan algorithm, Bleakley and Vert (2011) applied group LASSO techniques, and Song et al. (2016) generalized the SaRa algorithm to accommodate multiple samples to detect the common change points shared across multiple aligned sequences.

Very little work, however, has been published on the detection of jump points for multiple sequences in the nonparametric regression. Suppose that we observe two-dimensional array $\{Y_k(T_i) : k = 1, \ldots, m;\ i = 1, \ldots, n\}$

$$Y_k(T_i) = \mu_k(T_i) + \varepsilon_k(T_i), \quad k = 1, \ldots, m;\ i = 1, \ldots, n, \tag{1.1}$$

where m is the number of sequences, n is the number of observations, $\mu_k(\cdot)$'s are smooth functions except at shared jump points $\{\tau_1, \ldots, \tau_J\}$, and $\varepsilon_k(\cdot)$'s are heteroscedastic error processes with means zero and conditional variance functions $\sigma_k^2(\cdot)$'s. Most existing methods failed to detect jump points in the presence of heteroscedastic errors, since small jumps are more likely to be contaminated by continuity points with large variations, which makes this problem more challenging.

The nonparametric model (1.1) not only relaxes the assumptions of parametric piecewise-constant model and reduces the risk of modeling biases theoretically, but also comes from practical needs in real data analysis. For example, as Olshen et al. (2004) mentioned, copy number data tend

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

to exhibit local trends consisting of oscillations that even elaborate preprocessing fails to remove completely. Marioni et al. (2007) utilized Lowess regression to break the waves and to improve the calling of copy number variants (CNVs) in whole-genome tiling path arrays. They found that the wavy patterns appear to be a general feature of array comparative genomic hybridization (aCGH) data sets, and may prevent accurate change-point detection if fitting is done using the conventional parametric piecewise-constant model. Besides, model (1.1) can also be applied to estimate both integrated volatility and jump variation of the high-frequency financial data from several stocks simultaneously, see Fan and Wang (2007). In this paper:

(i) We show the optimal convergence rate $O(n^{-1})$ of jump estimators in the classic and popular screening procedure (Xia and Qiu (2015)).

(ii) We improve the classic screening procedure via multiple testing, which is called screening and multiple testing (SaMT) algorithm, to estimate the number and the locations of jump points. The proposed algorithm is shown to be less sensitive to the choices of tuning parameters, thus alleviating the need to precisely estimate the jumps in the screening procedure.

(iii) We propose a more powerful local test of significance of an individual candidate jump point based on the profile likelihood inference.

(iv) We generalize the SaMT algorithm to accommodate the presence

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

of multiple sequences to detect common jump points shared across multiple aligned sequences by aggregating the test statistics or p-values.

(v) We develop a new change-point detection method for the array-based DNA copy number data after considering the impacts of genomic waves, which is quite different from the existing methods.

The rest of the paper is organized as follows. Section 2 presents the classic screening procedure (Xia and Qiu (2015)) and our improved theoretical results. Section 3 describes the proposed SaMT algorithm to detect jump points for both single sequence and multiple aligned sequences. Section 4 illustrates the proposed methods via simulation studies and real data examples respectively. Section 5 is devoted to discussion and suggestions for further work. All the technical proofs and an additional simulation study are relegated to the Supplementary Material. Both R codes and data are available at `https://github.com/ShengjiJia/SAMT`.

2. Screening procedure

Let the response variable $Y(t)$ be collected at points $\{T_i : i = 1, \ldots, n\} \subset \mathcal{T}$, where $\mathcal{T} \subset \mathbb{R}$ is a fixed interval (e.g., $\mathcal{T} = [0, 1]$), and n is the total number of observations. T_i's are allowed to be either random or fixed. Consider the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

model

$$Y(t) = \mu(t) + \varepsilon(t) = \alpha(t) + \sum_{j=1}^{J} \beta_j \mathrm{I}(t > \tau_j) + \varepsilon(t), \tag{2.1}$$

where $\alpha(\cdot)$ is a smooth function, J and $\{\tau_1 < \cdots < \tau_J\}$ are the number and the locations of jump points. Let $\beta_j \neq 0$ denote the jump size at point τ_j, and $\mathrm{I}(\cdot)$ be the indicator function which is equal to 1 when its argument is true. Suppose $\varepsilon(t)$ is a Gaussian random process with mean 0 and

$$\mathrm{cov}\{\varepsilon(T_i), \varepsilon(T_j) \mid T_i = s, T_j = t\} = \sigma^2(t)\mathrm{I}(t = s),$$

where $\sigma^2(\cdot)$ is a nonparametric smooth function. The Gaussian assumption for the error process $\varepsilon(t)$ is not essential in our framework and can be relaxed. Denote the left- and right-limits of $\mu(\cdot)$ at point t by

$$\mu_-(t) = \lim_{x \uparrow t} \mu(x), \quad \mu_+(t) = \lim_{x \downarrow t} \mu(x).$$

According to (2.1), $\mu_+(\tau_j) - \mu_-(\tau_j) = \beta_j$ is the jump size at point τ_j, and $\mu_+(t) - \mu_-(t) = 0$ when t is a continuity point. It is anticipated that estimators, say $\widehat{\mu}_+(t)$ and $\widehat{\mu}_-(t)$, of $\mu_+(t)$ and $\mu_-(t)$ respectively, will satisfy

$$|\widehat{\mu}_+(t) - \widehat{\mu}_-(t)| \approx |\beta_j|, \quad \text{if } t \in \{\tau_1, \ldots, \tau_J\},$$

$$|\widehat{\mu}_+(t) - \widehat{\mu}_-(t)| \approx 0, \qquad \text{if } t \text{ is a continuity point with } \min_{1 \le j \le J} |t - \tau_j| > h_1,$$

where the definition of h_1 is given below. Therefore the local maximizers of $|\widehat{\mu}_+(t) - \widehat{\mu}_-(t)|$ may be considered as the jump candidates. Xia and Qiu

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

(2015) applied the locally-linear smoothing (Fan and Gijbels (1996)), which is known to reduce estimation bias at boundary points and perform better than the kernel estimators (e.g., Nadaraya-Watson and Gasser-Müller), to estimate $\mu_+(t)$ and $\mu_-(t)$ respectively. Specifically, let $K^+(\cdot)$ be a kernel function which is continuous within its support $[0,1]$, and $K^-(t) = K^+(-t)$ for $t \in [-1,0]$. Then we can replace a symmetric kernel function $K(\cdot)$ in the conventional locally-linear regression by the one-sided kernels $K^-(\cdot)$ and $K^+(\cdot)$ respectively to derive the estimators $\widehat{\mu}_+(t) := \widehat{a}_+$ and $\widehat{\mu}_-(t) := \widehat{a}_-$ by minimizing the following sum of squares:

$$\min_{(a_\pm, b_\pm)} \sum_{i=1}^{n} \Big\{ Y(T_i) - a_\pm - b_\pm (T_i - t) \Big\}^2 K^\pm\Big(\frac{T_i - t}{h_1}\Big), \tag{2.2}$$

where $h_1 > 0$ is the bandwidth. The estimators $\widehat{\mu}_+(t)$ and $\widehat{\mu}_-(t)$ can be written as the weighted sum of the responses (Fan and Gijbels (1996)):

$$\widehat{\mu}_\pm(t) = \frac{\sum_{i=1}^n w_i^\pm(t) Y(T_i)}{\sum_{i=1}^n w_i^\pm(t)}, \quad t \in \mathcal{T}, \tag{2.3}$$

$$\begin{aligned} w_i^\pm(t) &= K^\pm\Big(\frac{T_i - t}{h_1}\Big)\big\{S_2^\pm(t) - (T_i - t) S_1^\pm(t)\big\}, \quad i = 1, \ldots, n, \\ S_l^\pm(t) &= \sum_{i=1}^{n} (T_i - t)^l K^\pm\Big(\frac{T_i - t}{h_1}\Big), \quad l = 0, 1, 2. \end{aligned}$$

Now the estimation and inference of the jump points $\{\tau_j : j = 1, \ldots, J\}$ and jump sizes $\{\beta_j : j = 1, \ldots, J\}$ are based on the difference process:

$$L(t) := |\widehat{\mu}_+(t) - \widehat{\mu}_-(t)|, \quad t \in \mathcal{T}. \tag{2.4}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

If there is a unique discontinuity point in the regression function, i.e., $J = 1$, then the global maximizer of $L(t)$ serves to estimate the single jump point. Grégoire and Hamrouni (2002) derived the optimal convergence rate $O_{\mathrm{P}}(n^{-1})$ and the asymptotic distribution for this single jump estimator. As to multiple jumps, i.e., $J > 1$, the difference process $L(t)$ should have several local maximizers. We need a threshold λ such that any local maximum of $L(t)$ that exceeds λ can be regarded as a jump candidate. Assume $\mathcal{T} = [0, 1]$ without loss of generality. Denote by $\mathcal{S}_\lambda$ the candidate set of the initial jump estimators for a given threshold λ in the screening procedure, which can be updated in the following way:

(i) Extract local maximizers $\{\omega_1, \omega_2, \ldots\}$ of $L(t)$ over $[h_1, 1 - h_1]$ s.t.

$$L(\omega_1) \geq L(\omega_2) \geq \cdots \geq L(\omega_q) \geq \ \lambda \ > L(\omega_{q+1}) \geq \cdots ; \quad (2.5)$$

(ii) Let $\mathcal{S}_\lambda = \{\omega_1\}$ and the index set $\mathfrak{S} = \{1\}$;

(iii) For the candidate ω_i, with $i = 2, \ldots, q$ and q in (2.5), if

$$\omega_i \notin \bigcup_{j \in \mathfrak{S}} (\omega_j - h_1, \omega_j + h_1), \quad (2.6)$$

then we add ω_i into the candidate set: $\mathcal{S}_\lambda \leftarrow \mathcal{S}_\lambda \cup \{\omega_i\}$, and $\mathfrak{S} \leftarrow \mathfrak{S} \cup \{i\}$; otherwise, ω_i is not included in the candidate set $\mathcal{S}_\lambda$;

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

(iv) Finally, select $\widetilde{J} := |\mathfrak{S}|$ initial jump points in the candidate set

$$\mathcal{S}_\lambda := \{\widetilde{\tau}_1 < \widetilde{\tau}_2 < \cdots < \widetilde{\tau}_{\widetilde{J}}\} = \{\omega_i : \ i \in \mathfrak{S}\}. \tag{2.7}$$

Proposition 1. *If λ' is another threshold such that $\lambda > \lambda' > 0$, then we have $\mathcal{S}_\lambda \subseteq \mathcal{S}_{\lambda'}$.*

Proposition 1 shows the compatibility of the candidate set $\mathcal{S}_\lambda$ for different threshold λ: if one jump candidate is included in $\mathcal{S}_\lambda$ for some λ, then it will also be included in $\mathcal{S}_{\lambda'}$ as long as $\lambda' < \lambda$, which motivates the following sure screening property (Theorem 1(a)): when λ is small enough, for any true jump point τ_j, we can find an estimator in the candidate set $\mathcal{S}_\lambda$ that is very close to τ_j.

Theorem 1 (estimators for the number and locations of jumps)**.** *Suppose that conditions* A1–A7 *in the Supplementary Material hold.*

(a) (*Sure screening property*). *If the threshold $\lambda \to 0$ as $n \to \infty$, then*

$$\lim_{n\to\infty} \mathrm{P}(\widetilde{J} \geq J) = 1.$$

Besides, for each τ_j, $j = 1, \ldots, J$, there exists $\zeta \in \mathcal{S}_\lambda$ such that

$$|\zeta - \tau_j| = O_{\mathrm{P}}(n^{-1}), \quad j = 1, \ldots, J.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

(b) *Assume conditions in part* (a)*, and further assume* $\lambda^{-1}h_1^2 = o(1)$ *and* $\lambda^{-2}\log(h_1^{-1})/(nh_1) = o(1)$*. Then we have*

$$\lim_{n\to\infty} \mathrm{P}(\widetilde{J} = J) = 1. \tag{2.8}$$

Moreover, conditioning on the event $\{\widetilde{J} = J\}$*, we have*

$$|\widetilde{\tau}_j - \tau_j| = O_{\mathrm{P}}(n^{-1}), \quad j = 1, \ldots, J, \tag{2.9}$$

where $\widetilde{\tau}_j$ *are in* (2.7)*.*

Actually we have shown in the proof of Theorem 1 that if we choose λ appropriately, there exists an integer N such that $\widetilde{J} = J$ a.s. when $n > N$. So when the sample size is large enough, with probability one, we can estimate the number of jumps without error. Besides, we have shown the optimal convergence rate n^{-1} (Grégoire and Hamrouni (2002)) of the jump estimator in Theorem 1, compared with the convergence rate $\{h_1 \log(n)/n\}^{1/2}$ derived by Xia and Qiu (2015). Criteria other than the difference process $L(t)$ may be utilized in the screening procedure to select the jump candidates. For example, the wavelets method (Fan and Wang (2007)) for detecting jumps has the rate of convergence $n^{-1}\log^2(n)$, a little bit slower than that of the screening procedure.

Next, we will derive the asymptotic distribution of the jump size estimator $\widehat{\beta}_j = \widehat{\mu}_+(\widetilde{\tau}_j) - \widehat{\mu}_-(\widetilde{\tau}_j)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

Theorem 2 (jump size estimator). *Assume conditions of Theorem* 1(b), *and* $nh_1^5 = O(1)$. *Then conditioning on the event* $\{\widetilde{J} = J\}$, *we have*

$$\sqrt{nh_1}(\widehat{\beta}_j - \beta_j) \xrightarrow{\mathcal{D}} N\Big(0, \frac{2V\sigma^2(\tau_j)}{f(\tau_j)}\Big), \quad j = 1, \ldots, J, \tag{2.10}$$

where $f(\cdot)$ *is the density function of i.i.d. points* $\{T_i : i = 1, \ldots, n\}$, *and*

$$V = \frac{\mu_2^2\nu_0 - 2\mu_1\mu_2\nu_1 + \mu_1^2\nu_2}{(\mu_0\mu_2 - \mu_1^2)^2},$$

in which $\mu_j = \int u^j K^+(u)\mathrm{d}u$ *and* $\nu_j = \int u^j (K^+)^2(u)\mathrm{d}u$.

3. SaMT algorithm

According to Theorem 1(b) and 2, the choice of threshold λ is crucial, since the estimator for the number of jumps is sensitive to λ. In the literature, various threshold selection methods have been adopted. Grégoire and Hamrouni (2002) briefly described a method, but it is less practical since the choice of λ depends on $\min_{1\le j\le J}|\beta_j|$, which is unknown in advance and cannot be estimated before we know the number of jumps. Niu and Zhang (2012) applied locally-constant smoothing to detect change points in the parametric model, and suggested using a conservative threshold $\lambda = C\widehat{\sigma}\sqrt{2/(nh_1)}$, with $C = 2$ or 3, and $\widehat{\sigma}$ is the estimated standard deviation of homoscedastic errors. Xia and Qiu (2015) proposed a jump information criterion and Wang et al. (2022) proposed sample-splitting strat-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

egy to choose the optimal threshold. Besides, when the errors are heteroscedastic, the "global" threshold λ is not appropriate to detect jumps since some continuity points with large variations can be incorrectly recognized as jump candidates. Thus we suggest to choose small threshold λ and small bandwidth h_1 such that all the true jumps can be detected in the screening procedure (Theorem 1(a)) while allowing some continuity points to be included, and we then conduct multiple testing procedure, which will be introduced below, to rule out these spurious continuity points.

3.1 Local test: significance of individual jump candidate

In the literature, there are two types of tests in the jump detection problem: the local test and the global test. The local test focuses on testing whether the regression function has a discontinuity point at a certain fixed (prespecified) point t^*, and thus the local test relies on available information of the location of a possible jump point. The global test examines the null hypothesis of a smooth curve versus the alternative hypothesis of a curve with at least one discontinuity point (at an unknown position). There are a lot of references dealing with these tests, based on either the asymptotic laws (Müller and Stadtmüller (1999); Grégoire and Hamrouni (2002)), or the bootstrap procedure (Gijbels and Goderniaux (2004); Antoch et al.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

(2007)).

We focus on the local test since the jump candidates are available in Section 2, and first consider the following hypothesis testing problem:

$$\begin{aligned} H_0: &\quad \mu(\cdot) \text{ is continuous on } \mathcal{T}, \\ \text{versus } H_1: &\quad \mu(\cdot) \text{ is discontinuous at a prespecified time point } t^*. \end{aligned} \tag{3.1}$$

The asymptotic distribution of $\widehat{\mu}_+(t^*) - \widehat{\mu}_-(t^*)$ in (2.10) may be applied directly to construct a test statistic (Müller and Stadtmüller (1999); Grégoire and Hamrouni (2002)), but only the observations in the neighborhood of t^* are utilized, and the slower convergence rate $\sqrt{nh_1}$ will result in the less powerful test statistic. To enhance the efficacy, we will modify the profile likelihood estimation (Fan and Huang (2005)) to construct a new local test statistic which is more powerful than that of directly utilizing the asymptotic distribution (2.10). Note that model (2.1) can be regarded as a partially linear model if the jumps τ_j's are known. Consider the following model with a possible jump point t^*:

$$Y(T_i) = \mu(T_i) + \varepsilon(T_i) = \alpha(T_i) + \beta \mathrm{I}(T_i > t^*) + \varepsilon(T_i), \quad i = 1, \ldots, n, \tag{3.2}$$

and then the original hypothesis testing problem is transformed to

$$H_0 : \beta = 0 \quad \text{versus} \quad H_1 : \beta \neq 0. \tag{3.3}$$

We apply the profile locally-linear smoothing method to estimate $\{\alpha(\cdot),\, \beta\}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

Specifically, given β, model (3.2) reduces to a nonparametric model:

$$Y(T_i) - Z_i\beta = \alpha(T_i) + \varepsilon(T_i), \quad i = 1, \ldots, n, \tag{3.4}$$

where $Z_i = \mathrm{I}(T_i > t^*)$. The profile least squares estimator of β will enjoy a closed form if the following notations are used. Let $\boldsymbol{Y} = (Y(T_1), \ldots, Y(T_n))^T$, $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^T$, $\boldsymbol{m} = (\alpha(T_1), \ldots, \alpha(T_n))^T$, and $\boldsymbol{\varepsilon} = (\varepsilon(T_1), \ldots, \varepsilon(T_n))^T$, then model (3.4) can be written as

$$\boldsymbol{Y} - \boldsymbol{Z}\beta = \boldsymbol{m} + \boldsymbol{\varepsilon}. \tag{3.5}$$

Let $K(\cdot)$ be a symmetric kernel function, and $K_h(\cdot) = h^{-1}K(\cdot/h)$. Then using a bandwidth $h_2 > 0$ which differs from h_1 in (2.2), the locally-linear estimator $\widehat{\alpha}_\beta(t_0) := \widehat{a}_1$ for $\alpha(\cdot)$ at time point t_0 can be derived by:

$$\min_{a_1, b_1} \sum_{i=1}^{n} \big\{Y(T_i) - Z_i\beta - a_1 - b_1(T_i - t_0)\big\}^2 K_{h_2}(T_i - t_0), \tag{3.6}$$

which implies that $\widehat{\boldsymbol{m}} = \mathbf{S}(\boldsymbol{Y} - \boldsymbol{Z}\beta)$, where $\mathbf{S}$ is called a smoothing matrix of the locally-linear smoother, see Fan and Huang (2005). Substituting $\widehat{\boldsymbol{m}}$ into (3.5) results in the synthetic linear model,

$$(\mathbf{I} - \mathbf{S})\boldsymbol{Y} = (\mathbf{I} - \mathbf{S})\boldsymbol{Z}\beta + \boldsymbol{\varepsilon},$$

where $\mathbf{I} \in \mathbb{R}^{n\times n}$ is an identity matrix. So the estimator of β is

$$\widehat{\beta} = \{\boldsymbol{Z}^T(\mathbf{I} - \mathbf{S})^T(\mathbf{I} - \mathbf{S})\boldsymbol{Z}\}^{-1}\boldsymbol{Z}^T(\mathbf{I} - \mathbf{S})^T(\mathbf{I} - \mathbf{S})\boldsymbol{Y}. \tag{3.7}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

According to the sandwich formula (Fan and Huang (2005)), we construct the Wald statistic for testing the hypothesis (3.3):

$$W = \frac{\widehat{\beta}^2}{(\widetilde{\boldsymbol{Z}}^T\widetilde{\boldsymbol{Z}})^{-1}(\widetilde{\boldsymbol{Z}}^T\widehat{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{Z}})(\widetilde{\boldsymbol{Z}}^T\widetilde{\boldsymbol{Z}})^{-1}}, \tag{3.8}$$

where $\widetilde{\boldsymbol{Z}} = (\mathbf{I} - \mathbf{S})\boldsymbol{Z}$, and $\widehat{\boldsymbol{\Sigma}} = \mathrm{diag}\{\widehat{\sigma}^2(T_1), \ldots, \widehat{\sigma}^2(T_n)\}$ with $\widehat{\sigma}^2(\cdot)$ being an estimator of $\sigma^2(\cdot)$. In the literature, various methods can be applied to estimate the conditional variance function $\sigma^2(\cdot)$, e.g., the locally-linear smoothing estimator of the squared residuals (Li (2011)), which is defined by $\widehat{\sigma}^2(t_0) := \widehat{a}_2$ at time point t_0, where $(\widehat{a}_2, \widehat{b}_2)$ are the minimizers of

$$\min_{a_2, b_2} \sum_{i=1}^{n} \left\{\widehat{\varepsilon}_i^2 - a_2 - b_2(T_i - t_0)\right\}^2 K_{h_3}(T_i - t_0), \tag{3.9}$$

in which $\widehat{\varepsilon}_i = Y(T_i) - \widehat{\alpha}_{\widehat{\beta}}(T_i) - Z_i\widehat{\beta}$ is the residual, and $h_3 > 0$ is a new bandwidth differs from h_1 in (2.2) and h_2 in (3.6). In practice, we can also use the methods proposed by Jia et al. (2019) to guarantee the positivity of $\widehat{\sigma}^2(\cdot)$. Note that our proposed method reduces to Theorem 4.1 of Fan and Huang (2005) when the errors are homoscedastic, i.e., $\sigma^2(t) \equiv \sigma^2$. Moreover, compared with (2.10), the profile least squares estimator $\widehat{\beta}$ in (3.7) not only enjoys the faster convergence rate $\sqrt{n}$ (see the proof of Theorem 3), which will result in the more powerful test statistic W, but also avoids estimating the density function $f(\cdot)$. Thus we can test the hypothesis (3.3) based on the following asymptotic distribution of the Wald statistic W in (3.8).

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

Theorem 3. *Assume conditions* A1–A3 *and* B1–B4 *in the Supplementary Material. Then under the null hypothesis* H_0 *in* (3.3)*, the Wald statistic* W *in* (3.8) *asymptotically follows* χ^2_1 *distribution with one degree of freedom.*

We demonstrate the performance of the proposed Wald test statistic W through an additional simulation example in the Supplementary Material.

3.2 Multiple testing procedure: ultimate detection of jump points

According to the sure screening property (Theorem 1), with a small threshold λ, the candidate set $\mathcal{S}_\lambda$ will contain all of the true jumps. Of course, some continuity points with large variances may also be added into $\mathcal{S}_\lambda$. So after deriving the jump candidates in the screening procedure, we need to conduct multiple testing procedure to check whether each element in the candidate set $\mathcal{S}_\lambda = \{\widetilde{\tau}_1 < \widetilde{\tau}_2 < \cdots < \widetilde{\tau}_{\widetilde{J}}\}$ is the true jump point or not. This motivates us to consider testing the following multiple hypotheses:

$$\begin{aligned} H_{0,j}: \ & \mu_-(\widetilde{\tau}_j) = \mu_+(\widetilde{\tau}_j), \\ \text{versus } H_{1,j}: \ & \mu_-(\widetilde{\tau}_j) \neq \mu_+(\widetilde{\tau}_j), \quad j = 1, \ldots, \widetilde{J}. \end{aligned} \tag{3.10}$$

Here we assume the initial estimators $\widetilde{\tau}_j$ and $\widetilde{J}$ are prespecified, and apply the Benjamini-Hochberg (BH) multiple testing procedure (Benjamini and Hochberg (1995)) to control the false discovery rate (FDR):

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

(i) Define $\mathcal{N}_j$ as the neighborhood of a candidate $\widetilde{\tau}_j$ $(j = 1, \ldots, \widetilde{J})$:

$$\mathcal{N}_j := \Big\{T_i : \ T_i \in \Big(\frac{\widetilde{\tau}_{j-1} + \widetilde{\tau}_j}{2}, \frac{\widetilde{\tau}_j + \widetilde{\tau}_{j+1}}{2}\Big), i = 1, \ldots, n\Big\}, \qquad (3.11)$$

where $\widetilde{\tau}_0 := 2T_1 - \widetilde{\tau}_1$, and $\widetilde{\tau}_{\widetilde{J}+1} := 2T_n - \widetilde{\tau}_{\widetilde{J}}$, such that sets $\mathcal{N}_j$ constitute the partitions of all the points $\{T_1, \ldots, T_n\}$;

(ii) For $j = 1, \ldots, \widetilde{J}$, use the subsequence $\{Y(T_i) : \ T_i \in \mathcal{N}_j\}$ to fit the partially linear model, with the possible jump point $t^* = \widetilde{\tau}_j$ in (3.2),

$$Y(T_i) = \alpha(T_i) + \beta \mathrm{I}(T_i > \widetilde{\tau}_j) + \varepsilon(T_i), \quad T_i \in \mathcal{N}_j,$$

and derive the corresponding Wald test statistic W_j in (3.8) and the corresponding p-value p_j for testing the hypothesis $H_{0,j}$;

(iii) Apply the Benjamini-Hochberg (BH) multiple testing procedure to the p-values $\{p_1, \ldots, p_{\widetilde{J}}\}$ in step (ii). Let $p_{(1)} \le p_{(2)} \le \cdots \le p_{(\widetilde{J})}$ be the ordered p-values p_j's. For a nominal level α, say $\alpha = 0.05$ or 0.1, we reject the null hypotheses $H_{0,j}$ if $j \in \mathcal{A} := \{j : \ p_j \le p_{(\widehat{J})}\}$, where

$$\widehat{J} = \max\Big\{j : p_{(j)} \le \frac{\alpha j}{\widetilde{J}}\Big\}. \qquad (3.12)$$

Thus $\widehat{J}$ is the final estimator for the number of jumps, and $\{\widehat{\tau}_1 < \widehat{\tau}_2 < \cdots < \widehat{\tau}_{\widehat{J}}\} := \{\widetilde{\tau}_j : \ j \in \mathcal{A}\} \subseteq \mathcal{S}_\lambda$ are the corresponding estimated locations. For different subsequences $\{Y(T_i) : \ T_i \in \mathcal{N}_j\}_{j=1}^{\widetilde{J}}$, the bandwidths h_2 and h_3

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

that are utilized to construct the local tests may be different, and we will just use the same h_2 and h_3 for all these subsequences for simplicity.

It is easy to check that both the partitions $\mathcal{N}_j$'s and the subsequences $\{Y(T_i):\ T_i \in \mathcal{N}_j\}$'s are disjoint. Since the p-value p_j for testing the hypothesis $H_{0,j}$ depends on only the data points in the subsequence $\{Y(T_i):\ T_i \in \mathcal{N}_j\}$, the p-values $\{p_j\}_{j=1}^{\widetilde{J}}$ are independent, and the FDR can be controlled by the standard BH procedure (Benjamini and Hochberg (1995)).

3.3 Extension to multiple sequences

We now study the simultaneous jump detection problem for multiple aligned sequences, i.e., $m > 1$ in (1.1), and generalize the proposed SaMT algorithm to accommodate multiple sequences. The data consist of m independent sequences (samples), where all the observations are made on a fixed interval $\mathcal{T} \subset \mathbb{R}$ (e.g., $\mathcal{T} = [0, 1]$). For the kth sequence, $k = 1, \ldots, m$, the response $Y_k(t)$ is collected at points $\{T_i:\ i = 1, \ldots, n\}$, where n is the total number of observations for each sequence. Let us consider the following model:

$$\begin{aligned} Y_k(t) &= \mu_k(t) + \varepsilon_k(t) \\ &= \alpha_k(t) + \sum_{j=1}^{J} \beta_{k,j} \mathrm{I}(t > \tau_j) + \varepsilon_k(t), \quad k = 1, \ldots, m, \end{aligned} \tag{3.13}$$

where $\alpha_k(\cdot)$'s are smooth functions, J and $\{\tau_1, \ldots, \tau_J\}$ are the number and the locations of the shared jump points respectively, and $\beta_{k,j}$ denotes the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

jump size at τ_j for the kth sequence. Suppose $\varepsilon_k(\cdot)$'s are independent Gaussian random processes with means 0 and

$$\mathrm{cov}\{\varepsilon_k(T_i), \varepsilon_k(T_j) \mid T_i = s, T_j = t\} = \sigma_k^2(t)\mathrm{I}(t = s),$$

where $\sigma_k^2(\cdot)$'s are nonparametric smooth functions. We allow the baseline mean levels $\alpha_k(\cdot)$'s, jump sizes $\{\beta_{k,j}\}_{k=1}^m$, and variance functions $\sigma_k^2(\cdot)$'s to be subject-specified, which can differ substantially across samples. Besides, for each j, some of the jump sizes $\{\beta_{k,j}\}_{k=1}^m$ are allowed to be zero, which means that only a subset of the sequences, called the *carriers* (Zhang et al. (2010)), experience a shift in mean at the jump point τ_j.

Let $\widehat{\mu}_{k;\pm}(t)$ be the estimator of $\mu_{k;\pm}(t)$, where $\mu_{k;-}(t)$ and $\mu_{k;+}(t)$ are the left- and right-limits of $\mu_k(\cdot)$ at point t respectively. Let $L_k(t) = |\widehat{\mu}_{k;+}(t) - \widehat{\mu}_{k;-}(t)|$ be the difference process (2.4) for the kth sequence $\{Y_k(T_i) : \ i = 1, \ldots, n\}$, which measures the level of discontinuity at t for the kth sequence. To combine information across sequences to identify shared jump points, we need to combine the statistics $L_k(t)$'s. A natural choice is the *multiple sample difference process* defined as follows:

$$L^{\mathrm{multi}}(t) = \sum_{k=1}^{m} L_k^2(t). \tag{3.14}$$

$L^{\mathrm{multi}}(\cdot)$ is likely to be large around a jump point τ_j if $\sum_{k=1}^m \left\{\mu_{k;+}(\tau_j) - \mu_{k;-}(\tau_j)\right\}^2$ is large. Other metrics of the difference processes $L_k(\cdot)$'s may be

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

considered, such as $\max_{1\le k\le m} L_k(t)$ or $\sum_{k=1}^m L_k(t)$. Following Section 2, let

$$\mathcal{S}_\lambda^{\text{multi}} = \{\widetilde{\tau}_1 < \widetilde{\tau}_2 < \cdots < \widetilde{\tau}_{\widetilde{J}}\} \tag{3.15}$$

be the set of jump candidates for a threshold λ after we replace $L(t)$ by $L^{\text{multi}}(t)$ in the screening procedure. Then Theorem 1 can be generalized to the following Theorem 4 for the multi-sequence case, except that different conditions for the threshold λ are imposed in Theorem 4(b).

Theorem 4. *Assume conditions* A1′, A2′, A3–A7 *in the Supplementary Material hold.*

(a) (*Sure screening property*). *If the threshold* $\lambda \to 0$ *as* $n \to \infty$, *then*

$$\lim_{n\to\infty} \mathrm{P}(\widetilde{J} \ge J) = 1.$$

Besides, for each jump point τ_j, *there exists* $\zeta \in \mathcal{S}_\lambda^{\text{multi}}$ *such that*

$$|\zeta - \tau_j| = O_{\mathrm{P}}(n^{-1}), \quad j = 1, \ldots, J.$$

(b) *Assume conditions in part* (a), *and further assume* $\lambda^{-1}h_1^4 = o(1)$ *and* $\lambda^{-1}\log(h_1^{-1})/(nh_1) = o(1)$. *Then we have*

$$\lim_{n\to\infty} \mathrm{P}(\widetilde{J} = J) = 1.$$

Moreover, conditioning on the event $\{\widetilde{J} = J\}$, *we have*

$$|\widetilde{\tau}_j - \tau_j| = O_{\mathrm{P}}(n^{-1}), \quad j = 1, \ldots, J.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

After we derive the common jump candidates $\mathcal{S}_\lambda^{\text{multi}} = \{\widetilde{\tau}_1 < \widetilde{\tau}_2 < \cdots < \widetilde{\tau}_{\widetilde{J}}\}$ in (3.15), we conduct multiple testing procedure to check whether each element in $\mathcal{S}_\lambda^{\text{multi}}$ is the true shared jump or not, i.e., test the hypotheses,

$$\begin{aligned} H_{0,j}: &\ \textstyle\sum_{k=1}^m \left\{\mu_{k;+}(\widetilde{\tau}_j) - \mu_{k;-}(\widetilde{\tau}_j)\right\}^2 = 0, \\ \text{versus } H_{1,j}: &\ \textstyle\sum_{k=1}^m \left\{\mu_{k;+}(\widetilde{\tau}_j) - \mu_{k;-}(\widetilde{\tau}_j)\right\}^2 \neq 0, \quad j = 1, \ldots, \widetilde{J}. \end{aligned} \tag{3.16}$$

In the parametric paradigms (i.e., piecewise-constant model), various strategies including Fisher's method, Stouffer's method and the higher criticism method (Cai et al. (2011); Song et al. (2016)) have been proposed to detect the common change-points. In the nonparametric setting of this paper, to pool statistical evidence across samples, we consider to combine the Wald test statistics (Zhang et al. (2010)), or the p-values (Du and Zhang (2014)) from different sequences. Let $W_{k,j}$ be the Wald statistic for testing the jump candidate $\widetilde{\tau}_j$ based on the kth sequence, and define the following *multiple sample Wald statistic* for the hypothesis $H_{0,j}$:

$$W_j^{\text{multi}} = \sum_{k=1}^m W_{k,j}, \quad j = 1, \ldots, \widetilde{J}. \tag{3.17}$$

To combine the p-values, let $p_{k,j}$ be the p-value derived from the Wald test statistic $W_{k,j}$, for testing jump candidate $\widetilde{\tau}_j$ on the kth sequence, and define the *single-index modulated* (SIM) p-value for hypothesis $H_{0,j}$:

$$p_j^{\text{SIM}} = \Phi\Big(\sum_{k=1}^m w_k \Phi^{-1}(p_{k,j})\Big), \quad j = 1, \ldots, \widetilde{J}, \tag{3.18}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

where $\Phi(\cdot)$ is the c.d.f. of the standard normal distribution, and the weights w_k's satisfy $w_k \geq 0$ and $\sum_{k=1}^{m} w_k^2 = 1$. Du and Zhang (2014) suggested to choose the optimal weights w_k's based on the power consideration, and we will, in the absence of prior information, take $w_k = m^{-1/2}$, i.e., all the sequences are of the same importance. When the occurrence of jump points is sparse across sequences, a hybrid test statistic would be a better choice, combining the maximum-type and sum-type, see Song et al. (2016) for a review and comparison of these methods.

Proposition 2. *Suppose that conditions* A1′, A2′, A3, *and* B1′, B2–B4 *in the Supplementary Material hold.*

(a) *The multiple sample Wald statistics* $\{W_j^{\text{multi}}\}_{j=1}^{\widetilde{J}}$ *are independent. And under the null hypothesis* $H_{0,j}$ *in* (3.16), W_j^{multi} *asymptotically follows the* χ_m^2 *distribution with* m *degrees of freedom as* $n \to \infty$.

(b) *The single-index modulated p-values* $\{p_j^{\text{SIM}}\}_{j=1}^{\widetilde{J}}$ *are independent. And under the null hypothesis* $H_{0,j}$ *in* (3.16), p_j^{SIM} *asymptotically follows the uniform distribution over the interval* $[0, 1]$ *as* $n \to \infty$.

After deriving the *multiple sample p-values* $\{p_j^{\text{multi}}\}_{j=1}^{\widetilde{J}}$ based on the asymptotic null distributions of $\{W_j^{\text{multi}}\}_{j=1}^{\widetilde{J}}$ in (3.17), or the single-index modulated p-values $\{p_j^{\text{SIM}}\}_{j=1}^{\widetilde{J}}$ in (3.18), we conduct BH procedure to rule

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

out continuity points, in a way similar to Section 3.2.

4. Numerical studies

In this section, we investigate the performance of our proposed SaMT algorithm through Monte Carlo simulations and real data analysis.

4.1 Simulation 1: stability of SaMT algorithm

Suppose the true data generating process is model (2.1):

$$Y(T_i) = \alpha(T_i) + \sum_{j=1}^{J} \beta_j \mathrm{I}(T_i > \tau_j) + \varepsilon(T_i), \quad i = 1, \ldots, n,$$

with $n = 2000$ and $T_i = i/n$. We set $J = 20$, and the jump locations $\tau_j = 0.02 \times \varsigma_j$, where ς_j's are randomly selected from $\{1, 2, \ldots, 49\}$ without replacement. Let $\{\beta_j\}_{j=1}^{J}$ be independent, having discrete uniform distribution on $\{\pm 1, \pm 0.5\}$. Suppose $\varepsilon(t)$ is a Gaussian random process with mean 0 and $\sigma(t) = 0.1\{1 + \theta \sin(2\pi t)\}$, where $\theta \in \{0, 0.5\}$, and $\theta = 0$ corresponds to the homoscedastic errors. The nonparametric component $\alpha(\cdot)$ is

$$\alpha(t) = 0.1\big\{ \sin(20\pi t + \phi_1) + 2\sin(8\pi t + \phi_2)\big\}, \quad t \in [0, 1],$$

where $\{\phi_1, \phi_2\} \overset{\text{i.i.d.}}{\sim} \mathrm{unif}(0, 2\pi)$. We conduct the simulation 100 times, with the nominal level $\alpha = 0.1$. Throughout the paper, we use the Epanechnikov kernel $K(u) = 0.75(1-u^2)_+$, $u \in [-1, 1]$, and the one-sided kernel functions

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

$K^+(u) = K(u)\mathrm{I}(0 \le u \le 1)$ and $K^-(u) = K(u)\mathrm{I}(-1 \le u < 0)$. For each simulated data, we take the bandwidth $h_1 = 0.01 \times 0.85^k$, $k = 0, 1, 2$, and threshold $\lambda \in \{0.18, 0.20, 0.22, 0.24\}$ in the screening procedure. Besides, we assume the bandwidth $h_2 = 0.005$, and h_3 is selected automatically by R package "np" (Li and Racine (2007)) in the multiple testing procedure.

Table 1: *Means and standard errors (in parentheses) of* $\widehat{J}$ *and* FDP *over* 100 *simulations with different* λ *and* h_1.

		$h_1 = 0.01 \times 0.85^2$		$h_1 = 0.01 \times 0.85$		$h_1 = 0.01$	
θ	λ	$\widehat{J}$	FDP	$\widehat{J}$	FDP	$\widehat{J}$	FDP
0	0.24	$20.42_{(1.30)}$	$0.06_{(0.05)}$	$19.77_{(1.36)}$	$0.03_{(0.04)}$	$19.48_{(0.90)}$	$0.01_{(0.02)}$
	0.22	$20.73_{(1.73)}$	$0.09_{(0.06)}$	$20.21_{(1.55)}$	$0.06_{(0.05)}$	$19.64_{(1.13)}$	$0.03_{(0.03)}$
	0.20	$21.85_{(2.08)}$	$0.14_{(0.06)}$	$20.86_{(1.59)}$	$0.09_{(0.06)}$	$20.10_{(1.48)}$	$0.05_{(0.04)}$
	0.18	$21.81_{(2.95)}$	$0.16_{(0.08)}$	$21.61_{(1.90)}$	$0.13_{(0.07)}$	$20.59_{(1.48)}$	$0.08_{(0.05)}$
0.5	0.24	$20.68_{(1.88)}$	$0.10_{(0.06)}$	$20.01_{(1.52)}$	$0.07_{(0.05)}$	$19.62_{(1.32)}$	$0.04_{(0.04)}$
	0.22	$20.29_{(1.99)}$	$0.12_{(0.07)}$	$20.14_{(1.74)}$	$0.09_{(0.06)}$	$19.54_{(1.46)}$	$0.06_{(0.04)}$
	0.20	$20.90_{(2.36)}$	$0.14_{(0.07)}$	$20.65_{(2.05)}$	$0.10_{(0.06)}$	$20.19_{(1.43)}$	$0.07_{(0.05)}$
	0.18	$20.94_{(2.64)}$	$0.15_{(0.08)}$	$20.44_{(2.38)}$	$0.12_{(0.07)}$	$20.26_{(1.79)}$	$0.09_{(0.06)}$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

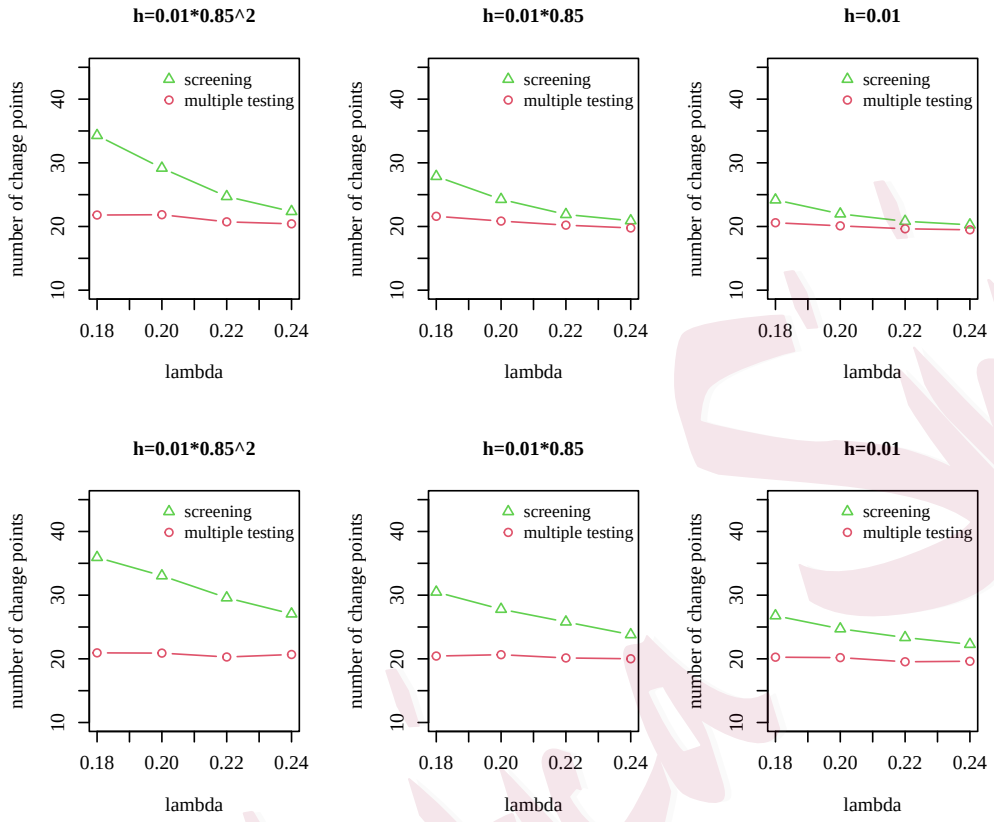

Figure 1: *Average number of* $\widetilde{J}$ *in screening procedure* ($\triangle$) *and* $\widehat{J}$ *in multiple testing procedure* ($\circ$) *over 100 simulations with different* λ *and* h_1. *Top panels:* $\theta = 0$*; bottom panels:* $\theta = 0.5$.

In Figure 1, we record the average number of jump candidates ($\widetilde{J}$) in the screening procedure, and the average number of final estimated jump points ($\widehat{J}$) after multiple testing over 100 simulations. We find SaMT works

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

well for both homoscedastic ($\theta = 0$) and heteroscedastic ($\theta = 0.5$) errors. In the screening procedure, as λ decreases, more points are included in the candidate set $\mathcal{S}_\lambda$. On the other hand, the final estimator $\widehat{J}$ remains quite stable and is close to the true number ($J = 20$) of jumps, implying that most of the spurious points are ruled out after multiple testing. Moreover, our proposed algorithm is not sensitive to the choices of λ and h_1, so there is a wide range for selecting the tuning parameters.

Table 1 shows the accuracy of the jump points estimators $\{\widehat{\tau}_1, \widehat{\tau}_2, \ldots, \widehat{\tau}_{\widehat{J}}\}$. We say a true jump τ_j is correctly detected if there exists $\widehat{\tau} \in \{\widehat{\tau}_1, \widehat{\tau}_2, \ldots, \widehat{\tau}_{\widehat{J}}\}$ such that $|\widehat{\tau} - \tau_j| \leq 2/n$. We record the average estimated number of jumps ($\widehat{J}$) and average false discovery proportion (FDP) over 100 simulations. From Table 1, FDP will increase as the threshold λ decreases, because we are adding just noises after all the true jumps being included in the candidate set $\mathcal{S}_\lambda$. Besides, FDP will be smaller as h_1 increases, since more data points are utilized in each neighborhood to derive the difference process $L(\cdot)$ in (2.4), which will result in the more precise jump candidates.

4.2 Simulation 2: comparison of different methods

Suppose the true data generating process follows model (3.13),

$$Y_k(T_i) = \alpha_k(T_i) + \sum_{j=1}^{J} \beta_{k,j} \mathrm{I}(T_i > \tau_j) + \varepsilon_k(T_i), \quad k = 1, \ldots, m; i = 1, \ldots, n,$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

with $n = 2048$ and $T_i = i/n$. We assume $J = 20$, and the common jump points $\tau_j = \{100 \times (j-1) + \varsigma_j\}/n$, where ς_j's are randomly selected from $\{1, 2, \ldots, 100\}$ with replacement. We take $m = 1$ or 4, where $m = 1$ corresponds to the jump detection for a single sequence. Suppose $\varepsilon_k(t)$ is a Gaussian random process with mean 0 and $\sigma_k(t) = 0.1\{1 + \theta_k \sin(2\pi t)\}$, with $\{\theta_k\}_{k=1}^m \overset{\text{i.i.d.}}{\sim} \text{unif}(0, 0.5)$. The nonparametric component $\alpha_k(\cdot)$ admits the following form:

$$\alpha_k(t) = 0.1\big\{\sin(20\pi t + \phi_{k,1}) + 2\sin(50\pi t + \phi_{k,2})\big\}, \quad t \in [0, 1],$$

where $\{\phi_{k,1}, \phi_{k,2}\}_{k=1}^m \overset{\text{i.i.d.}}{\sim} \text{unif}(0, 2\pi)$. The mechanism for generating $\beta_{k,j}$ is:

Case I. $\beta_{k,j}$'s are i.i.d. following the discrete uniform distribution on $\{-0.5, 0.5\}$;

Case II. $\beta_{k,j}$'s are i.i.d. following the discrete uniform distribution on $\{0, \pm 0.5, \pm 1\}$.

The following different methods are implemented to detect jumps:

1. CBS algorithm (Olshen et al. (2004)) for the single sequence;
2. cumSeg algorithm (Muggeo and Adelfio (2011)) for single sequence;
3. SaRa algorithm (Niu and Zhang (2012)) for the single sequence;
4. Wavelet: the wavelets method using Haar wavelets and universal threshold rule (Fan and Wang (2007)) for the single sequence;

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

5. SaMT$^{(\mathrm{I})}$: the proposed SaMT algorithm for the single sequence;

6. SaMT$^{(\mathrm{II})}$: the proposed SaMT algorithm for multiple sequences ($m = 4$) with the multiple sample Wald statistics $\{W_j^{\mathrm{multi}}\}_{j=1}^{\widetilde{J}}$ in (3.17);

7. SaMT$^{(\mathrm{III})}$: the proposed SaMT algorithm for multiple sequences ($m = 4$) with the single-index modulated p-values $\{p_j^{\mathrm{SIM}}\}_{j=1}^{\widetilde{J}}$ in (3.18).

We conduct the simulation 100 times with the nominal level $\alpha = 0.1$. For the SaMT algorithms (methods 5–7), we take $h_1 = 0.01$ and $\lambda = 0.2$ in the screening procedure. The choices of the (one-sided) kernel functions, the bandwidths h_2 and h_3 are the same as those in Section 4.1.

Table 2 presents the average estimated number ($\widehat{J}$) and the coverage probabilities for some jump points τ_j's obtained by different methods. We record the relative frequency that τ_j is correctly detected (i.e., there exists $\widehat{\tau} \in \{\widehat{\tau}_1, \widehat{\tau}_2, \ldots, \widehat{\tau}_{\widehat{J}}\}$ such that $|\widehat{\tau} - \tau_j| \leq 2/n$) over 100 simulations. For the single sequence, the CBS, cumSeg and SaRa algorithms (methods 1–3), which target at parametric piecewise-constant model, detect too many jumps because the wave patterns are incorrectly recognized as jump points. Besides, the proposed SaMT$^{(\mathrm{I})}$ works better than the wavelets-based method, especially when the jump points are located in the interval $[0.5, 1]$ where the signal-to-noise ratio is high. We also find that the coverage prob-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

Table 2: *Average estimated number* $(\widehat{J})$ *and average coverage probabilities of jump points with different methods over* 100 *simulations.*

$\beta_{k,j}$	method	$\widehat{J}$	τ_2	τ_4	τ_6	τ_8	τ_{10}	τ_{12}	τ_{14}	τ_{16}	τ_{18}	τ_{20}
	CBS	101.26	0.98	1.00	0.98	0.97	0.99	1.00	1.00	1.00	1.00	1.00
	cumSeg	37.15	0.36	0.39	0.37	0.32	0.41	0.44	0.40	0.37	0.37	0.47
	SaRa	44.05	0.73	0.66	0.73	0.66	0.70	0.74	0.73	0.76	0.67	0.78
Case I	Wavelet	22.67	0.69	0.75	0.75	0.68	0.75	0.67	0.74	0.79	0.75	0.72
	SaMT$^{(I)}$	18.24	0.76	0.78	0.79	0.77	0.90	0.86	0.94	0.95	0.93	0.98
	SaMT$^{(II)}$	19.34	0.98	0.98	0.97	0.99	0.99	0.96	0.97	0.99	0.96	1.00
	SaMT$^{(III)}$	19.34	0.97	0.99	0.97	0.99	0.99	0.96	0.97	0.99	0.96	1.00
	CBS	100.11	0.80	0.85	0.88	0.75	0.82	0.85	0.89	0.88	0.85	0.83
	cumSeg	33.41	0.38	0.50	0.36	0.44	0.45	0.47	0.48	0.46	0.41	0.41
	SaRa	43.83	0.68	0.66	0.68	0.66	0.67	0.75	0.74	0.71	0.76	0.67
Case II	Wavelet	26.53	0.62	0.60	0.65	0.63	0.64	0.56	0.65	0.62	0.68	0.66
	SaMT$^{(I)}$	16.14	0.68	0.72	0.71	0.67	0.68	0.73	0.83	0.79	0.81	0.81
	SaMT$^{(II)}$	19.19	0.98	0.95	0.96	0.94	0.96	0.99	0.98	0.99	0.99	0.99
	SaMT$^{(III)}$	19.08	0.96	0.95	0.96	0.94	0.95	0.99	0.98	0.99	0.99	0.99

abilities of jump points for multiple sequences (SaMT$^{(II)}$ and SaMT$^{(III)}$) are larger than those for a single sequence (SaMT$^{(I)}$ and Wavelet), which

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

demonstrates the advantage of our proposed method in pooling statistical evidence across samples to detect common jumps. Finally, SaMT$^{(II)}$ and SaMT$^{(III)}$ are comparable in detecting the weak signals (Case I), but when the jump sizes differ substantially across samples or when common jumps are only shared within part of the sequences (Case II), the jump estimators detected by SaMT$^{(II)}$ (which combines the Wald statistics) are more precise.

4.3 bladder tumor aCGH data

We now illustrate the proposed algorithm using the bladder tumor aCGH data. The chromosome copy number is the number of copies of a DNA region, and DNA copy number variation (CNV) refers to deletion or duplication of a region of DNA sequences compared to a reference genome assembly, and is associated with many human diseases including cancers. Therefore a major goal of DNA copy number data analysis is to identify the number and exact locations of the copy number changes. The bladder tumor aCGH dataset is publicly available from UCSF Cancer Center Array CGH Core Facility (`http://microarrays.curie.fr/publications/oncologie_moleculaire/bladder_TCM/`). The bladder tumor samples are analyzed on CGH microarrays, which consist of more than 2000 bacterial artificial chromosome clones covering the human genome with an average

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

resolution of 1.3 Mb (HumArray 2.0). Spots located in zones of spatial bias (abnormally high $\log_2$ ratios measured in some areas of the array, generally due to an edge or corner effect) are ignored, see Stransky et al. (2006) for more details. For subsequent transcriptome and CGH correlations, the $\log_2$ ratios of positions 2171559–37334583 kilobases from 4 samples (X1333-4, X1533-1, X1533-10 and X1533-13) are included, giving us a final list of 2300 probes for each sample. The segments with concentrated high or low $\log_2$ ratios correspond to gains or losses of copy numbers.

We compare the CBS, cumSeg, SaRa, wavelet-based methods and our proposed SaMT algorithm to detect change points in the bladder tumor aCGH data. In our proposed algorithm, both h_1 and h_2 are selected by the data-driven (order-preserved sample-splitting) strategy proposed by Wang et al. (2022), and h_3 is selected automatically by R package "np" (Li and Racine (2007)). We take the threshold $\lambda = 4\text{mad}(L(T_1), \ldots, L(T_n))$, with $\text{mad}(\cdot)$ being the median absolute deviation operator. Besides, to implement the wavelets-based method, the data are augmented via a symmetric reflection in the boundary so that the length of new sequence is 4096.

Table 3 compares the estimated number of change points for each sequence using different methods. We also calculate the adaptive Neyman test statistic (Fan and Huang (2001)) T_n^{AN} for each sequence to examine

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

Table 3: *Estimated number of change points with different methods and the adaptive Neyman test statistic* T_n^{AN} *for each sample, with* $n = 2300$.

Sample	CBS	cumSeg	SaRa	Wavelet	SaMT	T_n^{AN}
X1333-4	13	2	16	14	14	21.90
X1533-1	46	16	23	26	18	1.12
X1533-10	32	13	18	25	18	14.39
X1533-13	37	16	19	24	16	13.16

whether the conventional piecewise-constant model is adequate. The test statistics for samples X1333-4, X1533-10 and X1533-13 are larger than the critical value ($-\log\{-\log(1-0.05)\} = 2.97$) with significance level 0.05, thus the nonparametric models (2.1) and (3.13) are necessary. The CBS, SaRa and Wavelet algorithms tend to detect more change points, most of which are likely to be false positives because of the existence of oscillations (wave patterns). Since the original sequences are too long for us to observe the oscillations, Figure 2 plots just part of the data and the corresponding fitted lines using the SaMT algorithm. Compared with the classic piecewise-constant model, the significant change points have been preserved, while the wave patterns are also fitted pretty well by our proposed method. We also

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

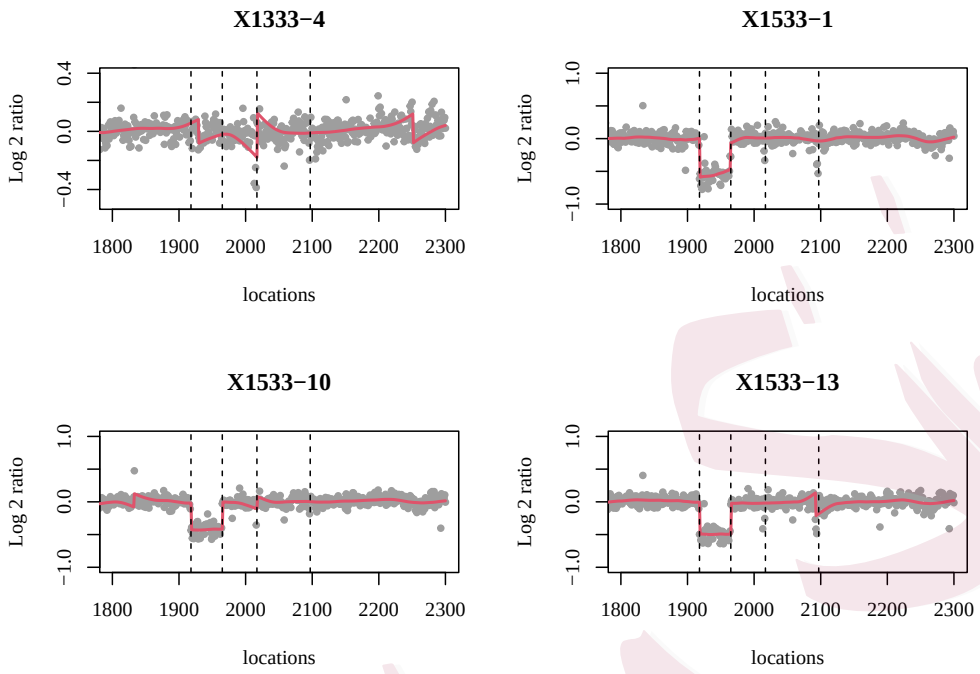

Figure 2: *Data (the last* 500 *probes), the fitted (solid) line by* SaMT *algorithm for each sequence and the common change points (vertical dashed lines) according to* (3.17) *for multiple sequences.*

show the estimators of the common change points (vertical dashed lines) according to (3.17) for multiple sequences. Some change points estimators (e.g., location 2251) that are significant for only one or two sequences (which may be correlated with other diseases instead of bladder tumor), are not detected as common change points by our proposed method.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

5. Discussion

In this paper, we propose the SaMT algorithm, which is an improvement of the classic screening procedure (Xia and Qiu (2015)), to estimate the number and locations of jump points in the nonparametric regression model with heteroscedastic errors. The jump candidates are first obtained by the screening techniques and we then conduct multiple testing procedure, which is based on the profile likelihood inference in the partially linear model, to rule out the continuity points with large variations. We also consider the detection of common jump points shared across multiple aligned sequences by combining the Wald test statistics or p-values to increase the statistical powers. The proposed method is stable and not sensitive to the choices of bandwidth h_1 and threshold λ, thus alleviating the need to precisely estimate the number of jumps in the screening procedure.

Several issues are desirable for further research. First, in this paper we only considered estimating the shared jump points through the summary statistics. It is possible that the common jump points may only be shared within part of the sequences (carriers), while we do not know which individuals carry a particular jump. Zhang et al. (2010) and Cai et al. (2011) proposed some methods to post-process the candidates of shared jump points in the parametric setting (or piecewise-constant model). Thus

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

it's of interest to develop new methods to identify the carriers of a given jump point in the nonparametric paradigm.

Secondly, the profile likelihood inference will be affected by the estimation errors of the jump candidates $\widetilde{\tau}_j$ in the screening procedure, since they are derived from the same data used for testing. Recently, Jewell et al. (2022) and Chen et al. (2023) addressed the challenges associated with post-selection inference in the context of piecewise-constant models, while this issue is rarely studied in the nonparametric regression. In the Supplementary Material, we conduct an additional simulation to examine the impact of estimation error in $\widetilde{\tau}$ on the Wald test statistic, but a formal investigation into post-selection inference for jump detection is highly desirable.

Besides, in this study, our primary focus was on the detection of common jumps within a fixed number of sequences. However, as Bleakley and Vert (2011) mentioned, the length (n) of sequence in genomic studies is typically fixed by the underlying technology, while the number (m) of sequences can increase when we collect the data from a greater number of patients. The asymptotic distributions derived in Proposition 2 are no longer valid when m tends to infinity. From a statistical point of view, it is of interest to develop new methods for the cases with fixed n and large m.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

Finally, very little work has been published on examining the adequacy of parametric fits compared with nonparametric alternatives for change-point detection problem. In Section 4.3, we used the adaptive Neyman test (Fan and Huang (2001)) heuristically to verify the existence of (non-parametric) wave patterns. It is desirable to more formally investigate the adaptive Neyman test. This topic is beyond the scope of this article and we plan to address this issue in a separate paper.

Supplementary Material

The online Supplementary Material contains all the technical conditions, complete proofs of the main theoretical results, and an additional simulation study.

Acknowledgments

We thank the editor, associate editor, and two reviewers for their constructive comments. Jia was supported by the Shanghai Natural Science Foundation, grants 25ZR1402404. Zhang was supported by U.S. National Science Foundation grants DMS-2013486 and DMS-1712418, as well as funding from the University of Wisconsin-Madison Office of the Vice Chancellor for Research and Graduate Education provided by the Wisconsin Alumni Re-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

search Foundation.

References

Antoch, J., Grégoire, G. and Hušková, M. (2007). Tests for continuity of regression function. *Journal of Statistical Planning and Inference* **137**, 753–777.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**, 289–300.

Bleakley, K. and Vert, J. P. (2011). The group fused Lasso for multiple change-point detection. arXiv preprint arXiv:1106.4199. `https://doi.org/10.48550/arXiv.1106.4199`

Cai, T. T., Jeng, X. J. and Jin, J. S. (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. *J. R. Stat. Soc. Ser. B* **73**, 629–662.

Chen, H., Ren, H., Yao, F. and Zou, C. (2023). Data-driven selection of the number of change-points via error rate control. *Jour. Ameri. Statist. Assoc.* **118**, 1415–1428.

Du, L. L. and Zhang, C. M. (2014). Single-index modulated multiple testing. *Ann. Statist.* **42**, 1262–1311.

Eichinger, B. and Kirch, C. (2018). A MOSUM procedure for the estimation of multiple random change-points. *Bernoulli* **24**, 526–564.

Eubank, R. L. and Speckman, P. (1994). Nonparametric estimation of functions with jump discontinuities. *Lecture Notes-Monograph Series* **23**, 130–144.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

Hall, New York.

Fan, J. and Huang, L. (2001). Goodness-of-fit test for parametric regression models. *Jour. Ameri. Statist. Assoc.* **96**, 640–652.

Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying coefficient partially linear models. *Bernoulli* **11**, 1031–1059.

Fan, J. and Wang, Y. (2007). Multi-scale jump and volatility analysis for high-frequency financial data. *Jour. Ameri. Statist. Assoc.* **102**, 1349–1362.

Gijbels, I. and Goderniaux, A. C. (2004). Bootstrap test for change-points in nonparametric regression. *Nonparametric Statistics* **16**, 591–611.

Gijbels, I., Hall, P. and Kneip, A. (2004). Interval and band estimation for curves with jumps. *Journal of Applied Probability* **41**, 65–79.

Grégoire, G. and Hamrouni, Z. (2002). Change point estimation by local linear smoothing. *J. Multivariate Anal.* **83**, 56–83.

Hao, N., Niu, Y. S. and Zhang, H. (2013). Multiple change-point detection via a screening and ranking algorithm. *Statist. Sinica* **23**, 1553–1572.

Harchaoui, Z. and Lévy-Leduc, C. (2010). Multiple changepoint estimation with a total variation penalty. *J. Amer. Statist. Assoc.* **105**, 1480–1493.

Jewell, S., Fearnhead, P. and Witten, D. (2022). Testing for a change in mean after changepoint detection. *J. R. Stat. Soc. Ser. B* **84**, 1082–1104.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

REFERENCES

Jia, S., Zhang, C. and Wu, H. (2019). Efficient semiparametric regression for longitudinal data with regularised estimation of error covariance function. *Journal of Nonparametric Statistics* **31**, 867–886.

Joo, J. H. and Qiu, P. (2009). Jump detection in a regression curve and its derivative. *Technometrics* **51**, 289–305.

Korkas, K. K. and Fryzlewicz, P. (2017). Multiple change-point detection for non-stationary time series using wild binary segmentation. *Statist. Sinica* **27**, 287–311.

Li, H., Munk, A., and Sieling, H. (2016). FDR-control in multiscale change-point segmentation. *Electron. J. Stat.* **10**, 918–959.

Li, Q. and Racine, J. S. (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press, New Jersey.

Li, Y. (2011). Efficient semiparametric regression for longitudinal data with nonparametric covariance estimation. *Biometrika* **98**, 355–370.

Ma, S. and Yang, L. (2011). A jump-detecting procedure based on polynomial spline estimation. *Journal of Nonparametric Statistics* **23**, 67–81.

Marioni, J. C., Thorne, N. P., Valsesia, A. et al. (2007). Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome biology* **8**(10), R228.

Muggeo, V. M. R. and Adelfio, G. (2011). Efficient change point detection for genomic sequences

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

of continuous measurements. *Bioinformatics* **27**, 161–166.

Müller, H. G. and Stadtmüller, U. (1999). Discontinuous versus smooth regression. *Ann. Statist.* **27**, 299–337.

Niu, Y. S. and Zhang, H. (2012). The screening and ranking algorithm to detect DNA copy number variations. *Ann. Appl. Stat.* **6**, 1306–1326.

Olshen, A., Venkatraman, E., Lucito, R. and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572.

Qiu, P. (2005). *Image Processing and Jump Regression Analysis*. John Wiley and Sons, New Jersey.

Qiu, P. and Yandell, B. (1998). A local polynomial jump detection algorithm in nonparametric regression. *Technometrics* **40**, 141–152.

Song, C., Min, X. and Zhang, H. (2016). The screening and ranking algorithm for change-points detection in multiple samples. *Ann. Appl. Stat.* **10**, 2102–2129.

Stransky, N., Vallot, C., Reyal, F. et al. (2006). Regional copy number-independent deregulation of transcription in cancer. *Nat. Genet.* **38**, 1386–1396.

Tibshirani, R. and Wang, P. (2008). Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics* **9**, 18–29.

Wang, G., Zou, C. and Qiu, P. (2022). Data-driven determination of the number of jumps in regression curves. *Technometrics* **64**, 312–322.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0216

REFERENCES

Wang, Y. (1995). Jump and sharp cusp detection by wavelets. *Biometrika* **82**, 385–397.

Xia, Z. M. and Qiu, P. H. (2015). Jump information criterion for statistical inference in estimating discontinuous curves. *Biometrika* **102**, 397–408.

Zhang, N. R., Siegmund, D. O., Ji, H. and Li, J. Z. (2010). Detecting simultaneous changepoints in multiple sequences. *Biometrika* **97**, 631–645.

School of Statistics and Mathematics; Interdisciplinary Research Institute of Data Science, Shanghai Lixin University of Accounting and Finance, Shanghai 201209, China.

E-mail: 20200026@lixin.edu.cn

Department of Statistics, University of Wisconsin, Madison, WI 53706, USA.

E-mail: cmzhang@stat.wisc.edu