Statistica Sinica Preprint No: SS-2024-0355	
Title	Change-Point Detection with Local Trend Adjustment
Manuscript ID	SS-2024-0355
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0355
Complete List of Authors	Shengji Jia, Chunming Zhang and Yiming Tang
Corresponding Authors	Yiming Tang
E-mails	jstangyiming@163.com
Notice: Accepted author version.	

Statistica Sinica: Newly accepted Paper

Statistica Sinica

Change-Point Detection with Local Trend Adjustment

Shengji Jia[1], Chunming Zhang[2*] and Yiming Tang[1*]

[1] *Shanghai Lixin University of Accounting and Finance*

[2] *University of Wisconsin-Madison*

Abstract: Identifying the number and precise locations of multiple change points in long sequences is a critical issue in statistics and machine learning. However, accurate change point detection can be compromised by the presence of local trends in the sequence when using the conventional parametric piecewise-constant model. In this paper, we introduce an adaptive Neyman test to assess the presence of local trends. Subsequently, we develop a novel change point detection procedure based on a partially linear model that incorporates these local trends. Furthermore, we extend the proposed testing and estimation methods to multidimensional cases, facilitating the identification of common change points in array-based data. Our methods are straightforward to implement, and we evaluate their numerical performance through simulations and the analysis of SNP genotyping data.

Key words and phrases: Adaptive Neyman test, Change points, Copy number variation, Fourier transformation, Lasso, Penalization.

* Corresponding author.

Statistica Sinica: Newly accepted Paper

1. Introduction

Change-point model has played a vital role in statistics and machine learning, finding successful applications in various fields, including genetics, engineering, signal processing, climatology and econometrics. A compelling example is the detection of DNA copy number variation in bioinformatics. DNA copy number refers to the number of copies of a genomic DNA region. Chromosome copy number variation (CNV) is the deviation of genomic regions from their normal copy number states, representing a significant genetic structural variation that may be associated with various human diseases, including cancers (Fearnhead and Liu, 2007; Zhang et al., 2024). Therefore, identifying the number and precise locations of change points is fundamentally crucial for analyzing DNA copy number data.

However, due to the advancements and increasing applications of high-resolution CNV detection technologies, a genome-wide local trend interfering with accurate CNV detection in signal intensity data has been observed. As originally mentioned by Olshen et al. (2004), copy number data tend to display local trends in the form of wave patterns that even comprehensive preprocessing fails to completely eliminate. Marioni et al. (2007) employed Loess regression to mitigate these wave patterns and enhance CNV calling in whole-genome tiling path arrays. The wavy patterns they observed seem

Statistica Sinica: Newly accepted Paper

to be a common characteristic of array comparative genomic hybridization (aCGH) data. Additionally, Diskin et al. (2008) discovered that signal intensities measured by high-density single nucleotide polymorphism (SNP) arrays are also susceptible to genomic waves. Hence, the classical change-point (piecewise-constant) model may not be suitable, and it is advantageous for the CNV calling algorithm to be robust to genomic waves.

Before more advanced techniques for observing local trends (or wave patterns) in the copy number data became available, statisticians had already developed some nonparametric methods to identify abrupt jumps in a smooth curve. These methods are primarily based on local linear smoothing (Grégoire and Hamrouni, 2002) or wavelets (Wang, 1995). Müller and Song (1997) introduced a two-step estimation procedure to enhance the efficiency of change points estimators. M*ü*ller and Stadtm*ü*ller (1999) and Gijbels and Goderniaux (2004) proposed different methods for testing whether the regression function is smooth or not. However, very little work has been published examining the adequacy of parametric fits (i.e., change-point models) in comparison to nonparametric alternatives (i.e., smooth curves with discontinuities). Furthermore, to investigate the impact of local trends, especially wave patterns in DNA copy number data, specific estimation and inference procedures are required instead of relying on existing nonparamet-

Statistica Sinica: Newly accepted Paper

ric or semiparametric methods that are generally applicable to any type of smooth function.

Recently, the issue of detecting change points occurring at the same location in multiple sequences has garnered significant attention. The objective is to identify these shared change points and leverage multiple samples to enhance the accuracy of change point detection. For example, Bleakley and Vert (2011) introduced group fused Lasso techniques, while Song et al. (2016) proposed the multiple sample SaRa algorithm for detecting simultaneous change points in array-based DNA copy number data. Therefore, it is desirable to develop new methods that consolidate statistical evidence across samples to detect common change points, while considering the simultaneous impacts of local trends from multiple sequences.

The main contributions of this paper can be summarized as follows:

(i) We formulate a partially linear model that accommodates local trends and introduce an adaptive Neyman test to confirm the local trends.

(ii) We develop a partial penalized least squares (PPLS) algorithm for estimating the change points, if local trends exist.

(iii) We extend the testing procedure and estimation algorithm to multiple sequences, specifically the multiple-sample adaptive Neyman test and

Statistica Sinica: Newly accepted Paper

group partial penalized least squares algorithm (GPPLS), for detecting common change points.

(iv) We apply the PPLS and GPPLS algorithms to detect change points in DNA copy number data. For single sequence, PPLS demonstrates noteworthy performance compared to classical models. For multiple sequences, GPPLS enables the detection of common change points even with weak signals.

The rest of the paper is organized as follows: In Section 2, we provide an overview of existing related works. Section 3 introduces the test for examining the presence of local trends and estimation algorithm for detecting change points. Sections 4 and 5 offer illustrative examples of the proposed methods using simulation studies and real data. Finally, Section 6 is dedicated to discussions and suggestions for further research. Technical conditions and proofs of main results, together with additional simulation and real data analysis are collected in the Supplementary Material. The algorithms are implemented in the R program, and both the source code and data can be accessed at `https://github.com/ShengjiJia/wave_test`.

Statistica Sinica: Newly accepted Paper

2. Review of existing works

The change-point detection problem is typically formulated as a piecewise-constant model, focusing on the high-dimensional Normal means model:

$$y_j = \mu_j + \varepsilon_j, \qquad j = 1, \ldots, n, \tag{2.1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$ is a sequence of responses (e.g., $\log_2$ fluorescence ratios), with a mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ represented as a piecewise-constant vector, and errors $\varepsilon_j \overset{i.i.d.}{\sim} N(0, \sigma^2)$. A change point is a position τ such that $\mu_\tau \neq \mu_{\tau+1}$. Assuming there are K change points $0 < \tau_1 < \cdots < \tau_K < n$, we are particularly interested in situations where n is large, and K is small. The objective is to estimate both the number K of change points and the location vector $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K)^T$.

In the literature, many tools for detecting multiple change points have been developed and extensively implemented. For instance, Olshen et al. (2004) introduced the circular binary segmentation (CBS) algorithm, which recursively identifies segments with changes. Niu and Zhang (2012) introduced the screening and ranking algorithm (SaRa), a powerful change point detection tool with a computational complexity of $O(n)$. Muggeo and Adelfio (2011) and Jia and Shi (2022) proposed a fast and robust al-

Statistica Sinica: Newly accepted Paper

gorithm based on a simple data transformation. Other approaches utilize penalized least squares regression with methods like L_1 penalty (Huang *et al.*, 2005; Harchaoui and Lévy, 2010), or combined penalties as seen in the fused-Lasso (Tibshirani and Wang, 2008). Bayesian approaches have been adopted when the focus is on mean levels rather than change points, see Erdman and Emerson (2008). Additionally, hidden Markov models (HMM) can also be applied for change-points detection, as demonstrated by Fridlyand et al. (2004). Comprehensive reviews and comparisons of some of these methods are provided by Liu et al. (2022) and Zhao et al. (2023).

We will focus on the fused-Lasso approach, which serves as the motivation for our proposed methods. It's worth noting that the change-point model (2.1) can be reformulated as a linear regression model. Let $\beta_0 = \mu_1$ and $\beta_i = \mu_{i+1} - \mu_i$, for $i = 1, \ldots, n-1$. Then, (2.1) is equivalent to:

$$y_j = \sum_{i=0}^{j-1} \beta_i + \varepsilon_j, \quad j = 1, \ldots, n, \tag{2.2}$$

and the constraints for the signals, such as:

$$\mu_1 = \mu_2 = \ldots = \mu_{\tau_1} \neq \mu_{\tau_1+1} = \ldots = \mu_{\tau_K} \neq \mu_{\tau_K+1} = \ldots = \mu_n$$

are transformed into constraints for the parameters $\{\beta_i : \ i = 1, \ldots, n-1\}$:

Statistica Sinica: Newly accepted Paper

$$\begin{aligned} \beta_i &\neq 0 \quad \text{if } i \in S, \\ \beta_i &= 0 \quad \text{if } i \notin S. \end{aligned} \tag{2.3}$$

Here, $S = \{\tau_1, \ldots, \tau_K\} \subseteq \{1, \ldots, n-1\}$ represents the set of true change points. The assumption that $K = |S|$ is much smaller than the sample size n implies that the transformed parameters $\{\beta_1, \ldots, \beta_{n-1}\}$ are sparse. As a result, the primary task shifts to identifying significant covariates or the support set $S = \{1 \leq i \leq n-1 : \ \beta_i \neq 0\}$ in the linear model (2.2). This leads us to reformulate the multiple change-point detection problem as a Lasso-type problem (Tibshirani, 1996), enabling efficient variable selection. Huang et al. (2005) and Harchaoui and Lévy (2010) proposed to minimize the following penalized sum of squares:

$$\min_{\beta_0, \ldots, \beta_{n-1}} \frac{1}{2n} \sum_{j=1}^{n} \Big(y_j - \sum_{i=0}^{j-1} \beta_i \Big)^2 + \lambda \sum_{i=1}^{n-1} |\beta_i|. \tag{2.4}$$

This is expected to yield a sparse vector $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{n-1})^T$ with non-zero components corresponding to the change point locations. Notably, β_0 corresponds to the intercept in model (2.2) and is not penalized. Let's denote the $n \times n$ lower triangular matrix with non-zero elements equal to 1 as $\mathbf{X}$. With this, the problem (2.4) can be represented as:

$$\min_{\boldsymbol{\beta}} \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}_1\|_1, \tag{2.5}$$

Statistica Sinica: Newly accepted Paper

where $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_{n-1})^T$, and $\|\mathbf{u}\|_1 = \sum_{j=1}^{n} |u_j|$ and $\|\mathbf{u}\|_2 = \sqrt{\sum_{j=1}^{n} u_j^2}$ are the l_1 and l_2 norms, respectively, of a vector $\mathbf{u} = (u_1, \ldots, u_n)^T \in \mathbb{R}^n$. Equivalently, the convex programming problem (2.5) can be reformulated as the following l_1-constrained quadratic programming problem:

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2, \quad \text{subject to} \quad \|\boldsymbol{\beta}_1\|_1 < s. \tag{2.6}$$

The minimization problems (2.5) and (2.6) can be efficiently solved using the least angle regression (LAR) algorithm, as proposed by Efron et al. (2004). The computational complexity of the LAR algorithm in this particular case is $O(K_{\max} n \log(n))$, where $K_{\max}$ represents a known upper bound on the true number of change points. For more details, see Harchaoui and Lévy (2010). This substantial reduction in computational complexity stands in contrast to the $O(K_{\max} n^2)$ complexity of the dynamic programming (DP) algorithm, as discussed by Huber et al. (2006).

In practice, it is essential to determine the values of the tuning parameters λ in (2.5) and s in (2.6). Various methods for selecting tuning parameters have been employed in the literature. For example, Huang et al. (2005) suggested empirically choosing s. They examined the solutions of (2.6) for an increasing sequence of s. As they added more change points to the model with larger values of s, they stopped increasing s when the resulting mean difference at the new change point was not sufficiently

Statistica Sinica: Newly accepted Paper

large. Harchaoui and Lévy-Leduc (2010) provided an asymptotic order of $O(\sqrt{(\log(n))/n})$ for the tuning parameter λ in the optimization problem (2.5), and Muggeo and Adelfio (2011) proposed selecting the tuning parameter by minimizing the generalized Bayesian information criterion (gBIC). As discussed in Rinaldo (2009) and Niu et al. (2016), finding the optimal value of the tuning parameter remains an open problem.

3. Proposed methods

3.1 Change-point detection for single sequence

Rather than fitting the piecewise-constant model, we integrate local trends of signals into model (2.2) and formulate it as a partially linear model:

$$y_j = f(x_j) + \sum_{i=0}^{j-1} \beta_i + \varepsilon_j, \quad j = 1, \ldots, n, \tag{3.7}$$

where $x_j = j$ for $j = 1, \ldots, n$, and $f(\cdot)$ represents the nonlinear component describing the local trends (or wave patterns). A natural choice is to assume that $f(\cdot)$ can be represented by the following Fourier series expansion:

$$f(x) = \sum_{i=1}^{m} \left\{ a_i \cos\left(\frac{2i\pi x}{n}\right) + b_i \sin\left(\frac{2i\pi x}{n}\right) \right\}. \tag{3.8}$$

In comparison with the standard Fourier series expansion, there is no intercept term in (3.8) since it has been absorbed by the baseline β_0 in (3.7) to

Statistica Sinica: Newly accepted Paper

ensure model identifiability. Typically, the number (dozens) of basis functions in (3.8) is assumed to be much smaller than the number (thousands) of data points. Depending on the specific research objectives, alternative expansions of $f(\cdot)$ may include orthogonal polynomial expansion, spline basis expansion, and wavelet basis expansion (Vidakovic, 1999).

Testing for the presence of local trends (or wave patterns) involves conducting the hypothesis testing problem:

$$H_0: \ f(\cdot) = 0 \quad \text{versus} \quad H_1: \ f(\cdot) \neq 0. \tag{3.9}$$

If the null hypothesis H_0 is rejected, we must then detect the change points by estimating the support set $S = \{1 \leq i \leq n-1: \ \beta_i \neq 0\}$ based on the partially linear model (3.7) instead of the parametric linear model (2.2).

Next, we outline the testing procedure. We adapt the adaptive Neyman test (Fan and Huang, 2001) to the current high-dimensional regression setting. Assuming a parametric linear model under the null hypothesis in (3.9), i.e.,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$. After deriving the initial minimizer $\widehat{\boldsymbol{\beta}}^{\text{ini}} = (\widehat{\beta}_0^{\text{ini}}, \widehat{\beta}_1^{\text{ini}}, \ldots, \widehat{\beta}_{n-1}^{\text{ini}})^T$ in (2.5) or (2.6), we refit the following submodel to obtain the residuals $\widehat{\boldsymbol{\varepsilon}}$:

$$\mathbf{y} = \beta_0 \mathbf{x}_0 + \sum_{i \in \widehat{S}^{\text{ini}}} \beta_i \mathbf{x}_i + \boldsymbol{\varepsilon}, \tag{3.10}$$

Statistica Sinica: Newly accepted Paper

where $\mathbf{X} = (\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{n-1})$ and $\widehat{S}^{\text{ini}} = \{1 \le i \le n-1 : \ \widehat{\beta}_i^{\text{ini}} \neq 0\}$. Let $\widehat{\boldsymbol{\varepsilon}}^* = (\widehat{\varepsilon}_1^*, \ldots, \widehat{\varepsilon}_n^*)^T$ be the discrete Fourier transform (Vidakovic, 1999) of the residual vector $\widehat{\boldsymbol{\varepsilon}} = (\widehat{\varepsilon}_1, \ldots, \widehat{\varepsilon}_n)^T$ from (3.10). More precisely, we define

$$\widehat{\varepsilon}_{2i-1}^* = \left(\frac{2}{n}\right)^{1/2} \sum_{j=1}^{n} \cos\left(\frac{2\pi ij}{n}\right) \widehat{\varepsilon}_j,$$

$$\widehat{\varepsilon}_{2i}^* = \left(\frac{2}{n}\right)^{1/2} \sum_{j=1}^{n} \sin\left(\frac{2\pi ij}{n}\right) \widehat{\varepsilon}_j, \qquad i = 1, \ldots, [n/2].$$

The purpose of Fourier transform is to condense useful signals into low frequencies, thereby enhancing the power of our proposed test. Testing all components of the vector $\widehat{\boldsymbol{\varepsilon}}^*$ is not advisable. If there is a vague prior suggesting that large absolute values of $\widehat{\boldsymbol{\varepsilon}}^*$ are primarily concentrated in the first k components, then one would focus on testing only the first k-dimensional subproblem. This leads to $\sum_{j=1}^{k} \widehat{\varepsilon}_j^{*2}$ or, equivalently, its standardized form:

$$\frac{1}{\sqrt{2k\widehat{\sigma}^4}} \sum_{j=1}^{k} (\widehat{\varepsilon}_j^{*2} - \widehat{\sigma}^2).$$

Here $\widehat{\sigma}^2$ is the sample variance of $\{\widehat{\varepsilon}_j^* : j = I_n + 1, \ldots, n\}$ for some given I_n. Fan and Huang (2001) suggested choosing $I_n = [n/4]$, and they showed that, under certain mild conditions, this estimator is $\sqrt{n}$-consistent with σ^2 even under the alternative hypothesis. The parameter k must be determined, and based on power considerations, we will employ the following adaptive

Statistica Sinica: Newly accepted Paper

Neyman test statistic:

$$T_n^* = \max_{1\le k\le n} \frac{1}{\sqrt{2k\widehat{\sigma}^4}} \sum_{j=1}^{k} (\widehat{\varepsilon}_j^{*2} - \widehat{\sigma}^2). \tag{3.11}$$

The null hypothesis is rejected when T_n^* is large. Fan and Huang (2001) derived the asymptotic null distribution of (3.11) in the low-dimensional case. We demonstrate in the following theorem that the asymptotic result remains valid for the high-dimensional model (3.7).

Theorem 1. *Suppose that conditions* (C1)–(C6) *in the Supplementary hold. Then, under the null hypothesis in* (3.9)*, the normalized test statistic*

$$T_n = \sqrt{2\log(\log(n))}T_n^* - \{2\log(\log(n)) + 0.5\log(\log(\log(n))) - 0.5\log(4\pi)\} \tag{3.12}$$

with T_n^* *defined in* (3.11)*, asymptotically follows*

$$\mathrm{P}(T_n \le x) \to \exp(-\exp(-x)) \qquad \text{as } n \to \infty. \tag{3.13}$$

The proof is provided in Supplementary Material. Conditions (C1) and (C2) serve as fundamental assumptions for adaptive Neyman test, paralleling those in Fan and Huang (2001). The literature on change-point analysis necessitates some standard assumptions regarding the signal jumps and minimum spacing between consecutive change-points. For instance, Niu et al. (2016) summarises the sufficient conditions for $\min_{2\le j\le K}(\tau_j - \tau_{j-1})$ and

Statistica Sinica: Newly accepted Paper

$\min_{1\le j\le K}|\mu_{\tau_j+1}-\mu_{\tau_j}|$. Given our reformulation of the change-point model into a Lasso-type framework, these conditions translate into assumptions for $\beta_i=\mu_{i+1}-\mu_i$ and the support set $S=\{1\le i\le n-1:\ \beta_i\neq 0\}$. Specifically, $\min_{i\in S}|\beta_i|$ in condition (C5) is essentially the lower bound for signal jumps, and conditions (C3) and (C4) imply that the neighboring change-points cannot be close to each other.

As a consequence of Theorem 1, the critical region $T_n>-\log(-\log(1-\alpha))$ has an asymptotic significance level α. If the null hypothesis H_0 in (3.9) is rejected, and the wave patterns can be described by the Fourier series expansion (3.8), then the estimation and inference procedures based on the linear model (2.2) may be inaccurate because some waves (or local trends) are incorrectly recognized as change points. Therefore, we should consider how to improve change point detection by fitting the partially linear model (3.7). Note that model (3.7) can be expressed in the matrix form using the following notations. Let $\boldsymbol{\gamma}=(\beta_0,a_1,b_1,\ldots,a_m,b_m)^T$ and

$$\mathbf{Z}=\begin{pmatrix} 1 & \cos(\frac{2\pi x_1}{n}) & \sin(\frac{2\pi x_1}{n}) & \ldots & \cos(\frac{2m\pi x_1}{n}) & \sin(\frac{2m\pi x_1}{n}) \\ 1 & \cos(\frac{2\pi x_2}{n}) & \sin(\frac{2\pi x_2}{n}) & \ldots & \cos(\frac{2m\pi x_2}{n}) & \sin(\frac{2m\pi x_2}{n}) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & \cos(\frac{2\pi x_n}{n}) & \sin(\frac{2\pi x_n}{n}) & \ldots & \cos(\frac{2m\pi x_n}{n}) & \sin(\frac{2m\pi x_n}{n}) \end{pmatrix}_{n\times(2m+1)}.$$

Statistica Sinica: Newly accepted Paper

Then, (3.7) can be written as

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

with $\mathbf{X}_1 = (\mathbf{x}_1, \ldots, \mathbf{x}_{n-1})$ excluding the intercept. The elements of $\boldsymbol{\gamma}$ describe the magnitude of the waves, which can be small compared with $\boldsymbol{\beta}_1$. We consider the following minimization problem, which only penalizes $\boldsymbol{\beta}_1$, and is called the partial penalized least squares (PPLS):

$$\min_{(\boldsymbol{\beta}_1,\boldsymbol{\gamma})} \frac{1}{2n} \parallel \mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1 - \mathbf{Z}\boldsymbol{\gamma} \parallel_2^2 + \lambda\|\boldsymbol{\beta}_1\|_1. \tag{3.14}$$

The tuning parameter λ in (3.14) may differ from that in (2.5). For any given $\boldsymbol{\beta}_1$, the $\boldsymbol{\gamma}$ that minimizes (3.14) necessarily satisfies

$$\mathbf{Z}^T\mathbf{Z}\boldsymbol{\gamma} = \mathbf{Z}^T(\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1).$$

Let $\mathbf{P} = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$ be the projection matrix of the column space of $\mathbf{Z}$. Then the profile objective function of $\boldsymbol{\beta}_1$ becomes

$$\min_{\boldsymbol{\beta}_1} \frac{1}{2n} \parallel (\mathbf{I}_n - \mathbf{P})(\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1) \parallel_2^2 + \lambda\|\boldsymbol{\beta}_1\|_1, \tag{3.15}$$

where $\mathbf{I}_n$ is an identity matrix of order n. Because the profile objective function (3.15) does not involve $\boldsymbol{\gamma}$ and has an explicit form, it is useful for both theoretical investigation and computation. In practice, we can regress $\mathbf{y}$ and each column of $\mathbf{X}_1$ on $\mathbf{Z}$ separately, and we denote the residuals by $\widetilde{\mathbf{y}}$

Statistica Sinica: Newly accepted Paper

and $\widetilde{\mathbf{X}}_1$ respectively. The profile objective function (3.15) is equivalent to

$$\min_{\boldsymbol{\beta}_1} \frac{1}{2n} \parallel \widetilde{\mathbf{y}} - \widetilde{\mathbf{X}}_1 \boldsymbol{\beta}_1 \parallel_2^2 + \lambda \|\boldsymbol{\beta}_1\|_1, \tag{3.16}$$

which becomes a standard Lasso-type problem. Let $\widehat{\boldsymbol{\beta}}_1 = (\widehat{\beta}_1, \ldots, \widehat{\beta}_{n-1})^T$ be the minimizer of (3.16), then $\widehat{S} = \{1 \leq i \leq n-1 : \ \widehat{\beta}_i \neq 0\}$ is the refined estimator of the support set S. Note that when $m = 0$, the PPLS estimator (3.14) becomes the standard Lasso estimator (2.5).

Clearly, the estimation procedure (3.16) also depends on the choice of m in $\mathbf{Z}$, which is related to the number of basis functions in the Fourier series expansion (3.8). Simulation results in Section 4 demonstrate that even for a complicated wave pattern, approximating the Fourier series with a small m (e.g., $m = 5$ or 10) yields quite good results. This enhances the performance of the standard Lasso approach, especially in the presence of local trends. In real data applications, we select m and tuning parameter λ in (3.16) concurrently using the generalized Bayesian information criterion (Muggeo and Adelfio, 2011), which, based on our experience, works effectively.

3.2 Common change-point detection for array-based data

In this subsection, we extend the proposed testing and estimation procedures from Section 3.1 to accommodate multiple samples for the detection of common change points. Suppose the observed two-dimensional array

Statistica Sinica: Newly accepted Paper

$\{y_{i,j} : i = 1, \ldots, d,\ j = 1, \ldots, n\}$ satisfies:

$$y_{i,j} = f_i(x_j) + \sum_{k=0}^{j-1} \beta_{i,k} + \varepsilon_{i,j}, \quad i = 1, \ldots, d;\ j = 1, \ldots, n,$$

where d is the number of independent sequences (or samples), and n is the total number of observations for each sequence. Once again, the support set $S = \{1 \le k \le n-1 :\ \sum_{i=1}^{d} \beta_{i,k}^2 > 0\}$ is the set of shared change points. The wave patterns are defined as:

$$f_i(x_j) = \sum_{k=1}^{m} \Big\{ a_{i,k} \cos\Big(\frac{2k\pi x_j}{n}\Big) + b_{i,k} \sin\Big(\frac{2k\pi x_j}{n}\Big) \Big\}, \quad i = 1, \ldots, d, \quad (3.17)$$

and the variances $\sigma_i^2 = \mathrm{var}(\varepsilon_{i,j})$ may vary depending on i to be sample-specific. We first propose a multiple-sample adaptive Neyman test for:

$$H_0 :\ f_i(\cdot) = 0,\ i = 1, \ldots, d, \quad H_1 :\ f_i(\cdot) \neq 0,\ \exists\, i \in \{1, \ldots, d\}. \quad (3.18)$$

Under the null hypothesis H_0, the initial estimated support set is defined as $\widehat{S}^{\mathrm{ini}} = \{1 \le k \le n-1 :\ \sum_{i=1}^{d} (\widehat{\beta}_{i,k}^{\mathrm{ini}})^2 > 0\}$, where $\widehat{\beta}_{i,k}^{\mathrm{ini}}$'s are derived by minimizing the following group Lasso-type problem (Yuan and Lin, 2006):

$$\min_{\beta_{i,k}} \frac{1}{2n} \sum_{i=1}^{d} \sum_{j=1}^{n} \Big(y_{i,j} - \sum_{k=0}^{j-1} \beta_{i,k} \Big)^2 + \lambda \sum_{k=1}^{n-1} \Big(\sum_{i=1}^{d} \beta_{i,k}^2 \Big)^{1/2}. \quad (3.19)$$

Once we have obtained the residuals $\widehat{\boldsymbol{\varepsilon}}_i = (\widehat{\varepsilon}_{i,1}, \ldots, \widehat{\varepsilon}_{i,n})^T$ and the corresponding discrete Fourier transform $\widehat{\boldsymbol{\varepsilon}}_i^* = (\widehat{\varepsilon}_{i,1}^*, \ldots, \widehat{\varepsilon}_{i,n}^*)^T$ for the ith se-

Statistica Sinica: Newly accepted Paper

quence by refitting the submodel:

$$\mathbf{y}_i = \beta_{i,0}\mathbf{x}_0 + \sum_{k\in\widehat{S}^{\text{ini}}} \beta_{i,k}\mathbf{x}_k + \boldsymbol{\varepsilon}_i,$$

where $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,n})^T$ and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i,1}, \ldots, \varepsilon_{i,n})^T$, we generalize the adaptive Neyman test statistic in Section 3.1 to the multivariate case:

$$T_n^* = \max_{1\le k\le n} \Big\{ \sum_{i=1}^{d} \Big(\sum_{j=1}^{k} \frac{\widehat{\varepsilon}_{i,j}^{*2} - \widehat{\sigma}_i^2}{\sqrt{2k\widehat{\sigma}_i^4}} \Big)^2 \Big\}^{1/2}, \tag{3.20}$$

where $\widehat{\sigma}_i^2$ is the sample variance of $\{\widehat{\varepsilon}_{i,j}^* : j = [n/4]+1, \ldots, n\}$ for the ith sequence. The adaptive Neyman test statistic (3.11) for a single sequence is a special case of (3.20) with $d = 1$. The following theorem shows the asymptotic null distribution of the multiple-sample adaptive Neyman test statistic T_n^* (3.20), but a different standardization form (3.21) is required.

Theorem 2. *Suppose that conditions* (C1) *and* (C2′)–(C6′) *in the Supplementary hold. Then, under the null hypothesis* H_0 *in* (3.18)*, the normalized test statistic is given by:*

$$T_n = \sqrt{2\log(\log(n))}T_n^* - \Big\{ 2\log(\log(n)) + \frac{d}{2}\log(\log(\log(n))) - \log\big(\Gamma\big(\frac{d}{2}\big)\big) \Big\}, \tag{3.21}$$

with T_n^* *defined in* (3.20)*, and* $\Gamma(\cdot)$ *being the Gamma function. It asymptotically follows:*

$$\mathrm{P}(T_n \le x) \to \exp(-\exp(-x)), \quad \text{as } n \to \infty.$$

Statistica Sinica: Newly accepted Paper

If H_0 is rejected with a large value of T_n, then we should estimate the common change points after considering the impacts of the wave patterns. Let $\boldsymbol{\gamma}_i = (\beta_{i,0}, a_{i,1}, b_{i,1}, \ldots, a_{i,m}, b_{i,m})^T$, $\boldsymbol{\beta}_{i,\bullet} = (\beta_{i,1}, \ldots, \beta_{i,n-1})^T$ and $\boldsymbol{\beta}_{\bullet,j} = (\beta_{1,j}, \ldots, \beta_{d,j})^T$. Consider the following group partial penalized least squares (GPPLS):

$$\frac{1}{2n}\sum_{i=1}^{d} \| \mathbf{y}_i - \mathbf{X}_1\boldsymbol{\beta}_{i,\bullet} - \mathbf{Z}\boldsymbol{\gamma}_i \|_2^2 + \lambda \sum_{j=1}^{n-1} \|\boldsymbol{\beta}_{\bullet,j}\|_2, \tag{3.22}$$

where $\mathbf{X}_1$ and $\mathbf{Z}$ are defined in Section 3.1. Again, (3.22) can be reduced to the following standard group Lasso-type problem using the profile least squares technique:

$$\frac{1}{2n}\sum_{i=1}^{d} \| \widetilde{\mathbf{y}}_i - \widetilde{\mathbf{X}}_1\boldsymbol{\beta}_{i,\bullet} \|_2^2 + \lambda \sum_{j=1}^{n-1} \|\boldsymbol{\beta}_{\bullet,j}\|_2,$$

where $\widetilde{\mathbf{y}}_i$ and $\widetilde{\mathbf{X}}_1$ are the residuals after regressing $\mathbf{y}_i$ and each column of $\mathbf{X}_1$ on $\mathbf{Z}$ separately. The refined estimated set of common change points is given by $\widehat{S} = \{1 \le j \le n-1 : \ \|\widehat{\boldsymbol{\beta}}_{\bullet,j}\|_2 > 0\}$.

4. Simulation studies

4.1 Simulation 1

We first investigate the finite sample properties of the adaptive Neyman test proposed in Section 3 through Monte Carlo simulations. Suppose the

Statistica Sinica: Newly accepted Paper

true data generating process is as follows:

$$y_j = f(x_j) + \sum_{k=1}^{K} \beta_k \mathrm{I}(x_j \geq \tau_k) + \varepsilon_j, \quad j = 1, \ldots, n, \tag{4.23}$$

where $\mathrm{I}(\cdot)$ is the indicator function (equal to 1 when true, 0 when false), the sample size $n = 500$, and $x_j = j$, $j = 1, \ldots, n$. We set the number of change points $K = 4$, with corresponding locations of change points $(\tau_1, \tau_2, \tau_3, \tau_4) = (150, 200, 400, 450)$, and the differences in mean levels $(\beta_1, \beta_2, \beta_3, \beta_4) = (1, -2, 2, -1)$. Assume $\varepsilon_j \overset{i.i.d.}{\sim} N(0, 0.5^2)$, and

$$f(x_j) = \theta\big\{ \sin(2\pi x_j/96 + \psi) + 2\sin(2\pi x_j/240 + \phi)\big\}, \quad j = 1, \ldots, n, \tag{4.24}$$

where $\psi \sim \mathrm{Unif}(0, 2\pi)$ and $\phi \sim \mathrm{Unif}(0, 2\pi)$. We set $\theta = 0.025 \times k$, $k = 0, 1, \ldots, 5$, and $\theta = 0$ corresponds to the null hypothesis H_0 in (3.9). We conduct the simulation 500 times. The performance of adaptive Neyman test depends on the accuracy of the initially estimated support set $\widehat{S}^{\mathrm{ini}}$ via the variable selection procedure (2.5), and we consider the following cases:

Case I (oracle) $\widehat{S}^{\mathrm{ini}} = S$.

Case II (overestimate) λ in (2.5) is selected such that $|\widehat{S}^{\mathrm{ini}}| = 10$.

Case III (underestimate) λ in (2.5) is selected such that $|\widehat{S}^{\mathrm{ini}}| = 5$.

It is worth mentioning that in Case III, although we have $|\widehat{S}^{\mathrm{ini}}| > K = 4$, $\widehat{S}^{\mathrm{ini}}$ is still an underestimate of S since the probability of detecting the true

Statistica Sinica: Newly accepted Paper

change point $\tau_4 = 450$ successfully ($\widehat{\tau}_4 \in [448, 452]$) over 500 simulations is just 0.22. Occasionally, two adjacent x_is may appear in $\widehat{S}^{\text{ini}}$ as the estimators of the same change point, and we will just discard one of them to avoid the problem of collinearity. For each simulated piece of data, we first get the initial estimated support set $\widehat{S}^{\text{ini}} = \{1 \leq i \leq n-1 : \ \widehat{\beta}_i^{\text{ini}} \neq 0\}$, and we then fit the submodel (3.10) to get the residuals $\widehat{\boldsymbol{\varepsilon}}$ and the corresponding adaptive Neyman test statistic T_n in (3.12).

Figure 1 shows the QQ plots and powers of the adaptive Neyman test over 500 simulations. From Figure 1, it can be seen that the testing procedure remains valid as long as all the true change points are contained in $\widehat{S}^{\text{ini}}$ (Case II), although it may lose some power when $\widehat{S}^{\text{ini}}$ is large. This loss of power is a result of the excessive biases in the estimation of S. Thus, there is a wide range for selecting the tuning parameter λ in the variable selection procedure (2.5). However, if we miss some true change points in $\widehat{S}^{\text{ini}}$ (Case III), the asymptotic distribution (3.13) no longer holds, and the null hypothesis H_0 is more likely to be rejected.

It is well known that the convergence rate of (3.13) is relatively slow, therefore the distribution under the null hypothesis does not approximate well even under the oracle setting. However, as Horváth (1993) pointed out, this approximation proves satisfactory for the upper tail of the dis-

Statistica Sinica: Newly accepted Paper

Table 1: *Average size and power of the adaptive Neyman test statistic for Cases I-III over 500 simulations (with $\alpha = 0.05$).*

	size		power			
θ	0	0.025	0.050	0.075	0.100	0.125
Case I	0.042	0.066	0.264	0.672	0.900	0.990
Case II	0.040	0.052	0.158	0.336	0.532	0.714
Case III	0.788	0.806	0.840	0.930	0.980	0.996

tribution, a property that suffices for the purpose of size control. Table 1 presents the average size and power of the adaptive Neyman test over 500 simulations. Notably, in Case I-II, the average size closely aligns with the predetermined significance level $\alpha = 0.05$, thereby affirming the suitability of our test for applications. Moreover, as Fan and Huang (2001) suggested, to further improve the performance of adaptive Neyman test, we advocate the utilization of simulations to empirically derive the finite sample null distribution of T_n instead of the asymptotic null distribution (3.13) to determine the more precise critical values. For illustration, leveraging 100 000 simulations with $n = 500$ and $\alpha = 0.05$, we establish a revised rejection criterion $T_n > 3.88$, as opposed to the asymptotic threshold of

Statistica Sinica: Newly accepted Paper

$T_n > -\log(-\log(1-0.05)) = 2.97$; see Fan and Huang (2001).

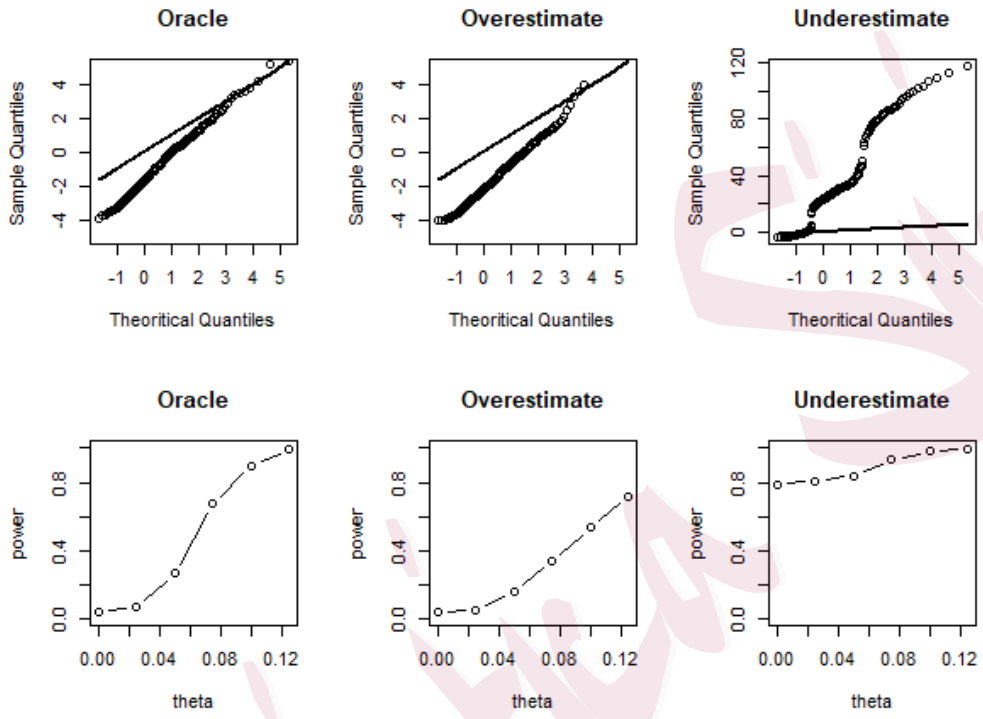

Figure 1: *QQ plots and powers of the adaptive Neyman test* (3.12) *with different estimated support set* $\widehat{S}^{\text{ini}}$. *Left: oracle* ($\widehat{S}^{\text{ini}} = S$); *middle: overestimate* ($|\widehat{S}^{\text{ini}}| = 10$); *right: underestimate* ($|\widehat{S}^{\text{ini}}| = 5$).

4.2 Simulation 2

Now we evaluate the performance of the proposed PPLS estimator (3.14). The data are generated under the same conditions as in Simulation 1, with

Statistica Sinica: Newly accepted Paper

variations in the mechanisms for generating the wave pattern $f(\cdot)$ in (4.23):

Scenario I Setting $m = 2$ and configuring all Fourier coefficients $\{a_1, a_2, b_1, b_2\}$ in (3.8) to be equal to 0.1.

Scenario II Setting $m = 7$, with the Fourier coefficients $\{a_3, a_4, a_5, b_3, b_4, b_5\}$ in (3.8) equal to 0, while the other Fourier coefficients are set to 0.1.

Scenario III Defining $f(\cdot)$ as in (4.24) with $\theta = 0.1$.

We conduct the simulation 500 times. For each simulated dataset, we employ three methods to estimate the support set S:

Method I Standard Lasso estimator (2.5) [i.e., $m = 0$ in (3.14)].

Method II PPLS estimator (3.14) with $m = 5$.

Method III PPLS estimator (3.14) with $m = 10$.

The top panels of Figure 2 present the average number of true change-points correctly detected by the estimated support set $\widehat{S}$ using different methods as λ decreases (or equivalently, as the estimated number of change-points $\widehat{K} = |\widehat{S}|$ increases from 1 to 10) across 500 simulations. We say that a true change-point τ_k is correctly detected by $\widehat{S}$ if there exists $\widehat{\tau} \in \widehat{S}$ such that $\widehat{\tau} \in [\tau_k - 2, \tau_k + 2]$. Figure 2 illustrates that the PPLS estimator (with

Statistica Sinica: Newly accepted Paper

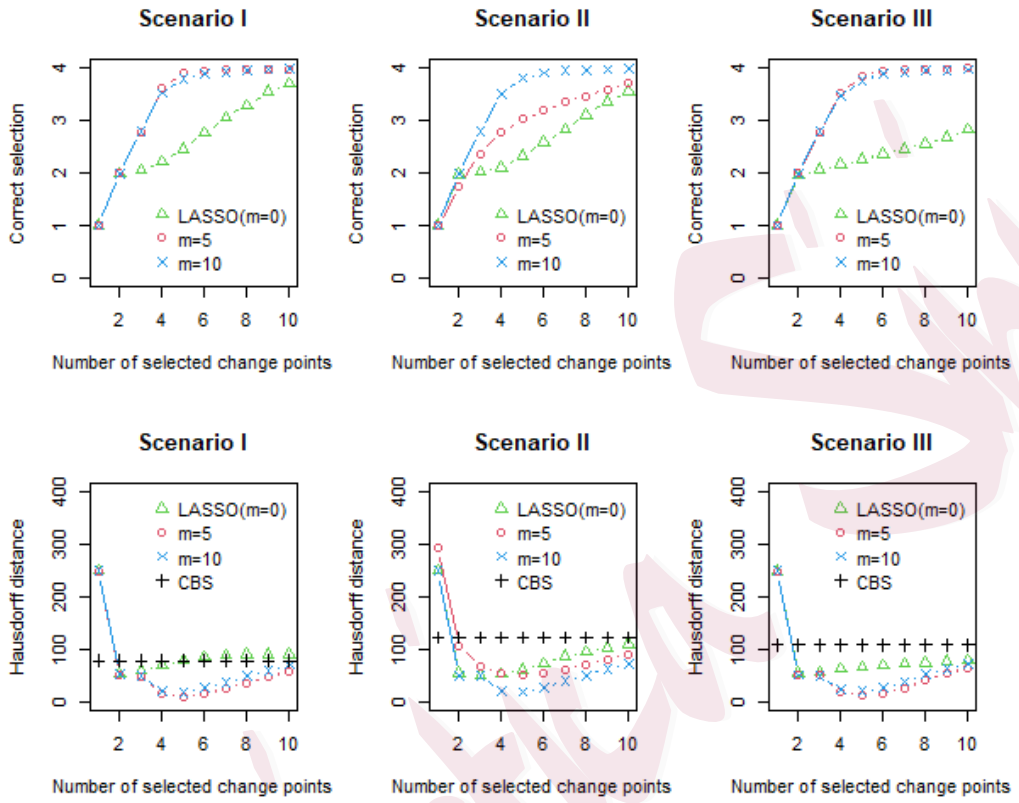

Figure 2: *Average true positives and Hausdorff distances by different methods over* 500 *simulations: Lasso* (△)*; PPLS with* $m = 5$ (∘) *and* $m = 10$ (×)*; CBS* (+)*. Left:* Scenario I*; middle:* Scenario II*; right:* Scenario III*.*

$m = 5$ or 10) outperforms the standard Lasso estimator (2.5) in all scenarios, because our proposed methods tend to detect all the true change-points much earlier than the standard Lasso estimator as $\widehat{K}$ increases. Particu-

Statistica Sinica: Newly accepted Paper

larly in Scenario III, the standard Lasso estimator fails to detect all the true change points, even when $\widehat{S}$ is very large. Furthermore, to demonstrate the efficacy in detecting change-point locations, we compute the mean Hausdorff distance (Harchaoui and Lévy, 2010) between the sets of estimated change-points and true change-points over 500 simulations. For the sake of comparative analysis, we incorporate the CBS algorithm (Olshen et al., 2004), which is a horizontal line in the plot because it is irrelevant with tuning parameter λ. The performance of Lasso and CBS algorithms are similar, and they are outperformed by the proposed PPLS estimators (with $m = 5$ or 10) when local trends are present. Method II (with $m = 5$) and Method III (with $m = 10$) are comparable, therefore in general, using $m = 5$ in (3.14) is sufficient to capture the local trends in the sequence, except in Scenario II, where the true number $m = 7$. Even for the most complex wave pattern (Scenario III) which does not adhere to the standard Fourier series expansion, our proposed method performs remarkably well.

4.3 Simulation 3

In this simulation, we assess the performance of the multiple-sample adaptive Neyman test statistic T_n (3.21) and the GPPLS estimator (3.22) as

Statistica Sinica: Newly accepted Paper

outlined in Section 3.2. The data generating process is described as follows:

$$y_{i,j} = f_i(x_j) + \sum_{k=1}^{K} \beta_{i,k} \mathrm{I}(x_j \geq \tau_k) + \varepsilon_{i,j}, \quad i = 1, \ldots, d; \ j = 1, \ldots, n,$$

where the number of sequences $d = 1$ or 4. The case with $d = 1$ corresponds to detecting change-points in a single sequence. The settings for the other parameters, including the sample size n, the number K and locations τ_k of change-points, the jump sizes $\beta_{i,k}$ and the errors terms $\varepsilon_{i,j}$, are identical to those defined in Simulation 1. The wave patterns $f_i(\cdot)$'s are sample-specific, and are expressed as follows:

$$f_i(x_j) = \theta\big[\sin(2\pi x_j/96 + \psi_i) + 2\sin(2\pi x_j/240 + \phi_i)\big], \tag{4.25}$$

where $\psi_i \overset{i.i.d.}{\sim} \mathrm{Unif}(0, 2\pi)$, and $\phi_i \overset{i.i.d.}{\sim} \mathrm{Unif}(0, 2\pi)$. The value θ is defined as $0.02 \times k$, with $k = 0, 1, \ldots, 5$, and $\theta = 0$ corresponds to the null hypothesis H_0 in (3.18). The significance level $\alpha = 0.05$, and the simulation is repeated 200 times for each case. In each simulated dataset, the initial estimated support set $\widehat{S}^{\mathrm{ini}}$ is obtained through group Lasso estimation (3.19). Subsequently, a submodel is fitted to acquire the residuals $\widehat{\boldsymbol{\varepsilon}}_i$'s and the corresponding multiple-sample adaptive Neyman test statistic T_n in (3.21). We examine three cases:

Case I (oracle) $\widehat{S}^{\mathrm{ini}} = S$.

Statistica Sinica: Newly accepted Paper

Case II The tuning parameter λ in (3.19) is chosen such that $|\widehat{S}^{\text{ini}}| = 10$.

Case III The tuning parameter λ in (3.19) is selected to achieve $|\widehat{S}^{\text{ini}}| = 5$.

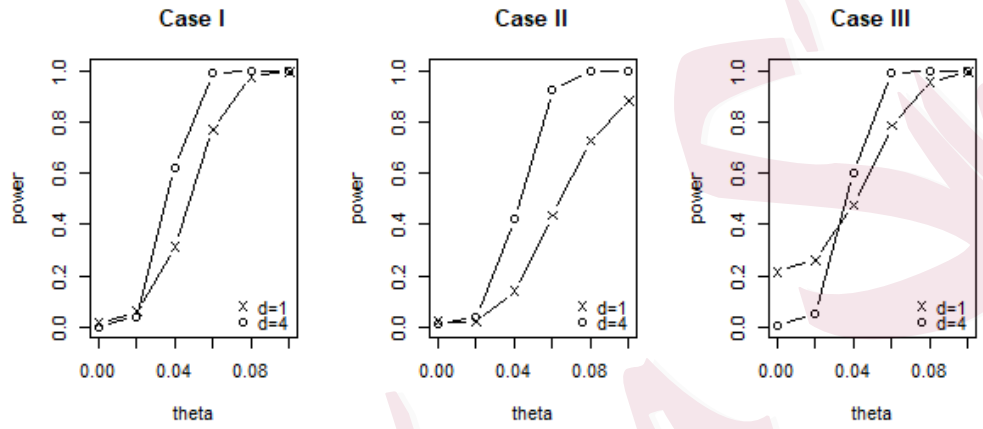

Figure 3: *Powers of the adaptive Neyman tests* (3.12) *for a single sequence* ($\times$) *and* (3.21) *for multiple sequences* ($\circ$) *with different initial sets* $\widehat{S}^{\text{ini}}$. *Left panel:* Case I*; middle panel:* Case II*; right panel:* Case III.

Figure 3 depicts the powers of the adaptive Neyman tests for single sequence ($d = 1$) and multiple sequences ($d = 4$), each with different initial estimated support set $\widehat{S}^{\text{ini}}$. The results for Case I-II align with those observed in Simulation 1: the testing procedure remains valid when all the true change points are contained in $\widehat{S}^{\text{ini}}$ (Case II), although there is a slight loss of power when $\widehat{S}^{\text{ini}}$ is large. Moreover, compared to the adaptive Neyman

Statistica Sinica: Newly accepted Paper

Table 2: *Coverage Probabilities (CP) for all change points* $(\tau_1, \tau_2, \tau_3, \tau_4)$ *with different methods over* 200 *simulations.*

	scenario I				scenario II			
(m, d)	τ_1	τ_2	τ_3	τ_4	τ_1	τ_2	τ_3	τ_4
$(0, 1)$	0.630	1.000	1.000	0.425	0.530	1.000	1.000	0.300
$(0, 4)$	0.845	1.000	1.000	0.940	0.510	1.000	1.000	0.605
$(5, 1)$	0.955	1.000	1.000	0.980	0.880	1.000	1.000	0.935
$(\mathbf{5}, \mathbf{4})$	**0.990**	**1.000**	**1.000**	**1.000**	**0.830**	**1.000**	**1.000**	**1.000**

test (3.12) for a single sequence, the multiple-sample adaptive Neyman test (3.21) benefits from the presence of multiple sequences, resulting in an increase in statistical power. The performance of these two tests differs significantly in Case III. Given that the group Lasso method yields a more precise estimator $\widehat{\mathcal{S}}^{\text{ini}}$ (e.g., the probabilities of successfully detecting the true change point $\tau_4 = 450$ ($\widehat{\tau}_4 \in [448, 452]$) are 0.8 and 1 for single and multiple sequences), the multiple-sample adaptive Neyman test ($d = 4$) remains effective, whereas the original test ($d = 1$) falls short.

Next, we evaluate the performance of the GPPLS estimator (3.22) under different settings. We consider the following two wave patterns:

Scenario I Setting $m = 2$ and configuring all Fourier coefficients $\{a_{i,1}, a_{i,2}, b_{i,1}, b_{i,2}\}$

Statistica Sinica: Newly accepted Paper

in (3.17) to be equal to θ_i, where θ_i is randomly selected from $\{-0.3, 0.3\}$.

Scenario II Defining $f_i(\cdot)$'s as in (4.25) with $\theta = 0.2$.

Table 2 presents the coverage probabilities (CP) for all the change points τ_k obtained with different estimated support sets $\widehat{S}$ in (3.22). Here, $m = 0$ corresponds to the standard (group) Lasso estimator. The table records the relative frequency with which τ_k is correctly detected by a given estimated support set $\widehat{S}$ (i.e., there exists $\widehat{\tau} \in \widehat{S}$ such that $\widehat{\tau} \in [\tau_k - 2, \tau_k + 2]$) over 200 simulations. To ensure a fair comparison, the tuning parameters are chosen such that $|\widehat{S}| = 5$ for all methods. The table shows that the coverage probabilities of the change points are higher for multiple sequences ($d = 4$) compared to a single sequence ($d = 1$), highlighting the advantage of the proposed estimator in leveraging statistical evidence across samples to detect common change points. Additionally, the standard Lasso estimators (with $m = 0$) successfully detect only the significant change-points τ_2 and τ_3 ($|\beta_{i,2}| = |\beta_{i,3}| = 2$), while the insignificant change-points τ_1 and τ_4 ($|\beta_{i,1}| = |\beta_{i,4}| = 1$) are more likely to be influenced by wave patterns. In contrast, the proposed estimators (with $m = 5$), which account for the impact of local trends, exhibit much better performance. Finally, even for the most complex wave pattern (Scenario II), which lacks the standard Fourier series expansion, our proposed method continues to perform effectively.

Statistica Sinica: Newly accepted Paper

5. Real data analysis

We now demonstrate the proposed methods using SNP genotyping data derived from an Illumina 550K platform, featuring a familial trio comprising a father, mother and offspring. This dataset, which is accessible in the PennCNV package (`https://penncnv.openbioinformatics.org/en/latest/`), comprises measurements of the Log R ratio, a normalized total signal intensity ratio computed as $\log_2(R_{obs}/R_{exp})$. Here R_{obs} represents the observed total intensity of the two alleles for a given SNP, while R_{exp} is the expected intensity, estimated through linear interpolation of the observed allelic ratio relative to canonical genotype clusters, as detailed in Niu and Zhang (2012). For our analysis, we focus on the Log R ratios along chromosomes 21 and 22, which contain 8251 and 8462 SNPs respectively, for each member of the trio. Thus, for each chromosome, our dataset consists of $d = 3$ sequences of length $n = 8251$ or $n = 8462$. The segments with concentrated high or low Log R ratios are regarded as gains or losses of copy numbers.

For this dataset, we employ the multiple-sample CBS algorithm (Zhang *et al.*, 2010), SaRa algorithm (Song *et al.*, 2016), group Lasso algorithm (Bleakley and Vert, 2011), and our proposed GPPLS to detect common change-points within the SNP genotyping data. For our proposed method,

Statistica Sinica: Newly accepted Paper

the tuning parameter is selected by minimizing the following generalized Bayesian information criterion (gBIC):

$$\text{gBIC} = \log(\hat{\sigma}^2) + \text{edf}\frac{\log(n)}{n}C_n,$$

where $\hat{\sigma}^2$ is the residual variance estimator, edf is the actual model dimension quantified by the number of estimated parameters, and $C_n = \log(\log(n))$, as described in Muggeo and Adelfio (2011).

Table 3: *Estimated number of change-points by different methods and multiple-sample adaptive Neyman test statistic* T_n.

data	CBS	SaRa	Lasso	Proposed	T_n
Chromosome 21	9	1	7	5	68.96
Chromosome 22	4	0	7	2	38.36

Table 3 presents the outcomes of SNP genotyping data analysis for chromosomes 21 and 22. Additionally, the multiple-sample adaptive Neyman test statistics T_n in (3.21) are provided, since both of them are large, we reject the null hypothesis H_0 in (3.18), therefore our proposed partially linear model is more appropriate. From Table 3, we find the CBS and group Lasso algorithms detected more change-points, largely attributed to genomic waves. Conversely, the SaRa algorithm detected less change-points,

Statistica Sinica: Newly accepted Paper

primarily stemming from the fact the SaRa algorithm is essentially a local method and only local information in the neighborhood of each probe is exploited. Based on simulation studies, we assert that the change-points estimators by our proposed GPPLS algorithm are more precise when local trends or genomic waves are present.

Due to space constraints, an additional application to the bladder tumor aCGH dataset is provided in the Supplementary Material.

6. Discussion

In this paper, we propose a testing procedure based on the adaptive Neyman test statistic (Fan and Huang, 2001) to verify the existence of genomic wave patterns that interfere with accurate CNV detection. Additionally, we propose a partial penalized least squares method to detect change points in the partially linear model that incorporates these wave patterns, substantially enhancing the performance of the standard Lasso estimator. Finally, we extend the proposed testing and estimation procedures to identify common change points shared in multiple sequences.

Several issues merit further investigation. First, in this study, our primary focus was on the detection of common change-points within a fixed number of independent sequences. Notably, our proposed method is flexible

Statistica Sinica: Newly accepted Paper

enough to accommodate scenarios when the sequences are correlated. This is accomplished by the utilization of weighted penalized least squares. On the other hand, as highlighted by Bleakley and Vert (2011), the length n of sequence in genomic studies typically remains fixed for a given technique, while the number d of sequences can increase as data are collected from a greater number of patients. It is crucial to note that the asymptotic distribution of the proposed adaptive Neyman test no longer holds as d tends to infinity. From a statistical perspective, it is, therefore, of interest to develop testing and estimation methods for cases with fixed n and large d.

Secondly, while our focus in this paper was on detecting change-points using Lasso-based techniques, other sparsity-inducing penalties, such as the smoothly clipped absolute deviation (SCAD) penalty (Fan, 1997) or minimax concave (MCP) penalty (Zhang, 2010), are possible options. Due to space limitations, we have included an additional simulation study exploring nonconvex penalties in the Supplementary Material. A more formal investigation into the theoretical properties and finite-sample performance of these penalties is warranted; however, this topic is beyond the scope of the current work and we plan to address this issue in a separate paper.

Furthermore, whereas theoretical results for estimating change-point positions in canonical Lasso problems like (2.5) and (3.19) have been well-

Statistica Sinica: Newly accepted Paper

established, the performance of the estimated change-points arising from formulation (3.15) or (3.22) remains unknown. Unfortunately, although numerical results in Section 4 reveal that the finite sample performance of estimated change-points in (3.15) and (3.22) are satisfactory, we have conducted additional simulations and found that the design matrix $\widetilde{\mathbf{X}}_1 = (\mathbf{I}_n - \mathbf{P})\mathbf{X}_1$ in (3.15) fails to satisfy the assumptions of standard LASSO theory. Consequently, the direct application of standard LASSO theory is hindered, necessitating the development of novel techniques to justify the theoretical behaviors of estimated change-point locations in (3.15) or (3.22).

Finally, in this paper we primarily focus on detecting common change-points across all sequences using group Lasso techniques. This approach overlooks the discrepancies introduced by heterogeneity in multiple sequences, where common change-points may be shared within only some of the sequences rather than all sequences. In this scenario, one might either post-process the results of the group Lasso, or consider sparse group selection techniques (Tian *et al.*, 2012; Simon *et al.*, 2013) to identify group-specific change-points in the presence of local trends. This topic falls beyond the scope of this paper, and further research is required.

Statistica Sinica: Newly accepted Paper

Supplementary Materials

The online Supplementary Material includes the conditions and proofs of the theoretical results, additional simulations, and an additional real data analysis.

Acknowledgements

The authors thank the Associate Editor and three reviewers for their careful review and helpful suggestions. The research of Zhang was supported by U.S. National Science Foundation grants DMS-2013486 and DMS-1712418, and by the University of Wisconsin-Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation.

References

[1] Bleakley, K. and Vert, J. P. (2011). The group fused Lasso for multiple change-point detection. Technical Report.

[2] Diskin, S. J., Li, M., Hou, C., Yang, S., Glessner, J., Hakonarson, H., Bucan, M., Maris, J. M. and Wang, K. (2008). Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res*, 36, e126.

Statistica Sinica: Newly accepted Paper

REFERENCES

[3] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.*, 32, 407–489.

[4] Erdman, C. and Emerson, J. (2008). A fast Bayesian change point analysis for the segmentation of microarray data. *Bioinformatics*, 24, 2143–2148.

[5] Fan, J. (1997). Comments on "Wavelets in statistics: A review" by A. Antoniadis. *Journal of the Italian Statistical Association*, 6, 131–138.

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

[7] Fearnhead, P. and Liu, Z. (2007). On-line inference for multiple changepoint problems. *J. R. Statist. Soc. B*, 69, 589–605.

[8] Fridlyand, J., Snijders, A. M., Pinkel, D., Albertson, D. G. and Jain, A. N. (2004). Hidden Markov models approach to the analysis of array CGH data. *Journal of Multivariate Analysis*, 90, 132–153.

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

[10] Grégoire, G. and Hamrouni, Z. (2002). Change point estimation by local linear smoothing. *Journal of Multivariate Analysis*, 83, 56–83.

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

Statistica Sinica: Newly accepted Paper

[12] Huang, T., Wu, B., Lizardi, P. and Zhao, H. (2005). Detection of DNA copy number alterations using penalized least squares regression. *Bioinformatics*, 21, 3811–3817.

[13] Huber, W., Toedling, J. and Steinmetz, L. M. (2006). Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics*, 22, 1963–1970.

[14] Jia, S. and Shi, L. (2022). Efficient change-points detection for genomic sequences via cumulative segmented regression. *Bioinformatics*, 38, 311–317.

[15] Liu, B., Zhang, X. and Liu, Y. (2022). High dimensional change point inference: Recent developments and extensions. *Journal of multivariate analysis*, 188, 104833.

[16] Marioni, J. C., Thorne, N. P., Valsesia, A., Fitzgerald, T., Redon, R., Fiegler, H., Andrews, T. D., Stranger, B. E., Lynch, A. G., Dermitzakis, E. T. et al. (2007). Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biol.*, 8, R228.

[17] Muggeo, V. M. R. and Adelfio, G. (2011). Efficient change point detection for genomic sequences of continuous measurements. *Bioinformatics*, 27, 161–166.

[18] Müller, H. G. and Song, K. S. (1997). Two-stage change-point estimators in smooth regression models. *Statistics & Probability Letters*, 34, 323–335.

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

[20] Niu, Y. S. and Zhang, H. (2012). The screening and ranking algorithm to detect DNA copy

Statistica Sinica: Newly accepted Paper

number variations. *Ann. Appl. Stat.*, 6, 1306–1326.

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

[22] 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.

[23] Rinaldo, A. (2009). Properties and refinements of the fused lasso. *Ann. Statist.*, 37, 2922–2952.

[24] Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2013). A sparse-group Lasso, *Journal of Computational and Graphical Statistics*, 22, 231-245.

[25] 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.

[26] Tian, Z., Zhang, H. and Kuang, R. (2012). Sparse group selection on fused lasso components for identifying group-specific DNA copy number variations. *IEEE International Conference on Data Mining*, 12, 665–674.

[27] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Statist. Soc. B*, 58, 267–288.

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

[29] Vidakovic, B. (1999). Statistical modeling by wavelets. Wiley, New York.

Statistica Sinica: Newly accepted Paper

REFERENCES

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

[31] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B*, 68, 49–67.

[32] Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38, 894–942.

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

[34] Zhang, Y., Liu, W. and Duan, J. (2024). On the core segmentation algorithms of copy number variations detection tools. *Briefings in Bioinformatics*, 25(2), 1–10.

[35] Zhao, W., Zhu, X. and Zhu, L. (2023). Detecting multiple change points: The pulse criterion. *Statistica Sinica*, 33, 431–451.

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

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

E-mail: tang.yiming@lixin.edu.cn