Statistica Sinica Preprint No: SS-2021-0309	
Title	A Unified Framework for Change Point Detection in High-Dimensional Linear Models
Manuscript ID	SS-2021-0309
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0309
Complete List of Authors	Yue Bai and Abolfazl Safikhani
Corresponding Authors	Abolfazl Safikhani
E-mails	a.safikhani@ufl.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

A UNIFIED FRAMEWORK FOR CHANGE POINT DETECTION IN HIGH-DIMENSIONAL LINEAR MODELS

Yue Bai Abolfazl Safikhani

Department of Statistics, University of Florida

Abstract: Although change-point detection for high-dimensional data has become increasingly important in many scientific fields, most existing methods are designed for specific models (e.g., mean shift model, vector auto-regressive model, graphical model). Here, we provide a unified framework for structural break detection that is suitable for a large class of models. Moreover, we propose a three-step algorithm that automatically achieves consistent parameter estimates during the change-point detection process, without needing to refit the model. The first step combines the block segmentation strategy and a fused lasso-based estimation criterion, leading to significant computational gains, without compromising the statistical accuracy of identifying the number and location of the structural breaks. Then, we use hard-thresholding and exhaustive search steps to consistently estimate the number and location of the break points. We prove strong guarantees on both the number of estimated change points and the rates of convergence of their locations, and provide consistent estimates of the model parameters. The findings of our numerical studies support the theory and validate the competitive performance of the algorithm for a wide range of models. The proposed algorithm is implemented in the **R** package **LinearDetect**.

Key words and phrases: High-dimensional data; Piecewise stationarity; Structural breaks; Fused lasso; Block segmentation; Linear model.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

1. Introduction

Methods for detecting change points (break points) in dynamic systems have become increasingly important in areas such as quality control (Qiu, 2013), neuroscience (Ombao et al., 2005), economics and finance (Frisén, 2008), and social network analysis (Savage et al., 2014). A change point represents a discontinuity in the parameters of the data-generating process. Previous studies have investigated both *offline* and *online* versions of the problem (Basseville and Nikiforov, 1993; Csörgö and Horváth, 1997). In the former case, we have a sequence of observations, and we wish to determine, for example, whether change (break) points exists, and, if so, their locations, as well as estimating the parameters of the data-generating process. In the online case, we sequentially obtain new observations, with the goal of finding the change point as quickly as possible (Wang and Mei, 2015; Chan et al., 2021).

The fused lasso (Rinaldo, 2009) is a computationally attractive offline change-point detection method, owing to its linear computation time with respect to the sample size (Bleakley and Vert, 2011). In this method, we first expand the parameter space to allow the model parameters to change at all time points. Here, the consecutive differences of the parameters are fused (forced to zero) to reduce the parameter space dimension. It is known that the fused lasso method over estimates the number of change points, that is, it has a nonvanishing false positive rate (Harchaoui and Lévy-Leduc, 2010). In addition, there is no unified result for deriving upper bounds for the total positive rate of the fused lasso, which means we typically need additional steps to consistently estimate the number of change points;

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

see, for example, the screening step in Safikhani and Shojaie (2020). These additional steps usually include several hyperparameters, and the finite-sample detection performance can be sensitive to small changes in these hyperparameters. Furthermore, the theoretical rates of these hyperparameters depend on the model, and need to be derived separately for each statistical model under consideration. Note that despite these issues, the fused lasso is a popular detection algorithm, because it has a faster computational speed than that of more exhaustive search methods, such as dynamic programming, which has at least a quadratic computation time with respect to the sample size, and thus is not scalable to large-scale (and high-dimensional) data sets.

We propose a new detection algorithm called the threshold block-fused lasso (TBFL). Although it is motivated by the fused lasso, the problems with the latter method are mitigated in the proposed algorithm. Unlike the fused lasso, the TBFL consistently estimates the number of change points in a single step, with computational complexity similar to that of the fused lasso (or better; see Remark 2). Furthermore, we estimate the locations of the change points consistently by developing a local exhaustive search step. The proposed algorithm is flexible and can detect breaks in a wide range of statistical models. Here, we focus on detecting break points and estimating the model parameters for general sparse multivariate regression models with high-dimensional covariates (Rothman et al., 2010). In such models (model 2.1), both the response variable and the covariates are multivariate, and their dimensions can potentially be much larger than the sample size. Moreover, unlike in the case

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

of typical regression models, we do not assume independence among the covariates in different samples (see Sections 2 and 4). This makes the model flexible enough to include a wide range of models (with possible temporal and/or spatial correlations), including the mean shift models (Harchaoui and Lévy-Leduc, 2010), multiple linear regression model (Leonardi and Bühlmann, 2016), vector auto-regressive models (Lütkepohl, 2005), Gaussian graphical models (Yuan and Lin, 2007), and network auto-regressive model (Zhu et al., 2017).

The TBFL first partitions the time domain into blocks, assuming the model parameters remain fixed within each block and change among neighboring blocks. The block sizes (b_n, with n as the sample size) are selected carefully to control the false positive rates, while not missing any true break point. Then, we estimate the model parameters among all blocks simultaneously using regularized estimation procedures motivated by the fused lasso. Furthermore, we calculate the Frobenius norm of the differences between the estimated model parameters in consecutive blocks, called "jumps." Intuitively, a large magnitude of jump implies that a true break point exists inside the neighboring blocks, whereas a small jump may be caused by a finite-sample estimation error. Thus, jumps are thresholded using a certain data-driven threshold, and only block ends corresponding to jumps above the threshold are regarded as "candidate" change points. Note that the hard-thresholding technique has been used in lasso regularization to reduce the false positive rate (van de Geer et al., 2011), but thresholding has not been investigated fully for the fused lasso. We verify (Theorem 1) that, under certain conditions, this procedure leads to a set of "clusters" of

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

candidate change points, and the number of clusters matches the true number of break points in the model (denoted by m_0) with high probability. As a byproduct of this result, the total number of candidate change points is at most $2m_0$, with high probability, converging to one as the sample size diverges. This can be interpreted as an upper bound that controls the false positive rate, a result that is not available for the fused lasso for such a general linear model. Moreover, a simple exhaustive search within each estimated cluster gives the final estimation for the locations of the break points. We derive the nonasymptotic consistency rates of these estimates in Theorem 2, where we show that the change-point estimates are optimal up to a logarithmic factor (see Section 4). Few works have examined model parameter estimates after break detection. In addition, having consistent estimators for the model parameters before and after break points can reveal the main drivers of breaks in the system, providing valuable insights into the main features that contributes to a shock/break in a system (e.g., see the application of the TBFL to an EEG data set in Section 8). Interestingly, we can use the estimated parameters within the TBFL to develop model parameter estimates between any two consecutive break points, *without* refitting, and derive their consistency as well (see Theorem 3). The steps of the TBFL algorithm are illustrated in Figure 1. A random realization from model 2.1 is generated with sample size $n = 1000$, $p_x = 20$, $p_y = 1$, two true change points at 333 and 666 (solid red lines), and a block size of $b_n = 30$. In Figure 1, we plot the square of the jump sizes (i.e., the square of the Frobenius norm of the differences between the estimated model parameters in consecutive blocks) at block ends $(30, 60, 90, \ldots)$ in all

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

panels (for the model settings, see the Supplementary Material S7). The left panel shows that there are large jumps close to two break points, and some small jumps far from any true break point that can be removed using thresholding (green horizontal dashed line). The middle panel shows clusters of candidate change points in neighborhoods of the true break points. Three candidate break points remain after thresholding, supporting Theorem 1, which states that there should be at most $2m_0 = 4$ candidate break points. Finally, the right panel shows the final estimated break points as blue vertical dashed lines from using a local search within each cluster.

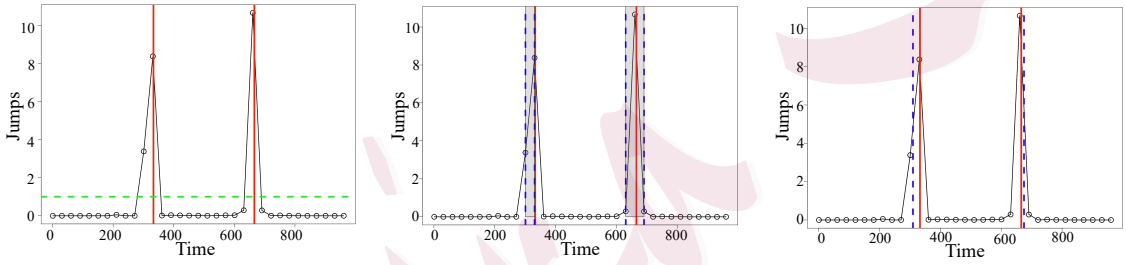

Figure 1: Illustration of the TBFL algorithm.

In summary, this study contributes to the literature in four ways. First, we propose a detection algorithm that can handle a wide range of linear models, including the change-in-mean model, multiple linear regression model, vector autoregressive (VAR) model, and Gaussian graphical model, in both high-dimensional and fixed-dimensional cases. Second, we provide theoretical guarantees in terms of the consistency rate of change-point detection and parameter estimation. Third, we provide consistent model parameter estimates during the change-point detection process, *without* needing to refit the model. Lastly, we provide

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

data-driven methods for selecting all hyperparameters in the algorithm. The algorithm is implemented in the **R** package **LinearDetect** (Bai and Safikhani, 2021).

1.1 Related Works

Numerous works have examined the problem of change-point detection in the offline version, focusing mostly on fix-dimensional regimes. These studies can be categorized into the following three groups according to the dimension of the coefficient parameters considered by the model: univariate, multivariate, and high-dimensional. Several works focus on different types of models in the univariate case. For example, Davis et al. (2006) use the minimum description length principle to locate change points in piecewise univariate auto-regressive models, and Killick et al. (2012) propose a pruned exact linear time (PELT) method using the optimal partitioning approach of Jackson et al. (2005) and a pruning step within the dynamic program to detect the structural breaks. Fryzlewicz (2017) applies a tail-greedy Haar transformation to consistently estimate the number and locations of multiple change points in the univariate piecewise-constant model, and Aue et al. (2017) develop a method based on the (scaled) functional cumulative sum (CUSUM) statistic for detecting shifts in the mean of a functional data model. In the multivariate case, with the number of model parameters p fixed, Ombao et al. (2005) develop a spectral representation to locate the break points, Zhang and Lavitas (2018) propose a self-normalized technique for testing change points, and Matteson and James (2014) propose a nonparametric approach based on the Euclidean distances between sample observations. There is increasing interest in the high-dimensional case, in

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

which the number of model parameters p is much larger than the number of observations n (Hastie et al., 2009). Cho and Fryzlewicz (2015) and Cho (2016) use binary segmentation to locate break points in high-dimensional data, and Wang and Samworth (2016) propose a high-dimensional change-point detection method that uses a sparse projection to project the high-dimensional data into a univariate case. The algorithm for estimating a single change point can be combined with the wild binary segmentation scheme of Frick et al. (2014) to locate multiple change points sequentially in high-dimensional time series. Wang et al. (2019) develop an l_0-optimization for change point detection in VAR models, and Roy et al. (2017) propose a likelihood-based method for locating a single break point in high-dimensional Markov random fields, and provide the rate of estimating the change point and the model parameters. In addition, a U-statistic-based cumulative sum statistic is developed in Liu et al. (2020) to test for the existence of a single change point, and Safikhani and Shojaie (2020), Bai et al. (2020), and Safikhani et al. (2021) use a fused lasso (Tibshirani et al., 2005) and a screening step to estimate multiple break points in a VAR model and establish consistency results for both the break points and the model parameters. Moreover, Kolar and Xing (2012) consider a fused lasso regularization together with a neighborhood selection approach to detect the change points in the Gaussian graphical model, and Bybee and Atchadé (2018) introduce a majorize-minimize algorithm plus a simulated annealing (SA) algorithm for computing change points in large graphical models. Finally, Gibberd and Roy (2017) use a group-fused graphical lasso (GFGL) to detect multiple change points

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

in a high-dimensional setting. See Aue and Horváth (2013); Yu (2020) for a comprehensive review.

The remainder of the paper is organized as follows. In Section 2, we introduce the general model formulation, and in Section 3, we describe the proposed TBFL algorithm. We establish the asymptotic properties, including the consistency of the number of change points and their locations, in Section 4, and provide examples of models in Sections 5 and S3 (Supplementary Material). In Section 6, we discuss the optimal block size selection method, and in Section 7, we compared the numerical performance of the proposed TBFL in various simulation settings with that of other methods in Sections 7, S9, and S10 (Supplementary Material). In Section 8, we present a real-data application of electroencephalograms (EEGs) recorded during eyes-closed and eyes-open resting conditions. Section 9 concludes the paper.

Notation: Denote the indicator function of a subset S as $\mathbb{1}_S$. For any vector $v \in \mathbb{R}^p$, we use $\|v\|_\infty$ to denote $\max_{1\le i\le p}\{|v_i|\}$. For any matrix $\boldsymbol{A}$, the ℓ_1, ℓ_2, and ℓ_∞ norms of the vectorized form of $\boldsymbol{A}$ are denoted by $\|\boldsymbol{A}\|_1 = \|\mathrm{vec}(\boldsymbol{A})\|_1$, $\|\boldsymbol{A}\|_F = \|\mathrm{vec}(\boldsymbol{A})\|_2$ and $\|\boldsymbol{A}\|_\infty = \|\mathrm{vec}(\boldsymbol{A})\|_\infty$, respectively. The transpose of a matrix $\boldsymbol{A}$ is denoted by $\boldsymbol{A}'$. Let $\Lambda_{\max}(\Sigma)$ and $\Lambda_{\min}(\Sigma)$ denote the maximum and minimum eigenvalues of the symmetric matrix Σ. Denote the tensor product of two matrices as $\otimes$. For functions $f(n)$ and $g(n)$, we write $f(n) = \Omega(g(n))$ if and only if for some constants $c \in (0, \infty)$ and $n_0 > 0$, $f(n) \ge cg(n)$ for all $n \ge n_0$; we write $f(n) = O(g(n))$ if and only if for some constants $c \in (0, \infty)$ and $n_0 > 0$, $f(n) \le cg(n)$, for all $n \ge n_0$. We define the Hausdorff distance between two countable sets on

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

the real line as $d_H(A, B) = \max\left\{\max_{b\in B}\min_{a\in A}|b-a|, \max_{a\in A}\min_{b\in B}|b-a|\right\}$. For scalars a and b, define $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$.

2. Model Formulation

We consider a multivariate regression model (Rothman et al., 2010) with a structural break such that the values of the coefficient matrix change over time in a piecewise constant manner. Specifically, suppose there exist m_0 change points $\{t_1, \ldots, t_{m_0}\}$ such that $1 = t_0 < t_1 < \cdots < t_{m_0} < t_{m_0+1} = n + 1$. Then, the structural break multivariate regression model is given by

$$\boldsymbol{y}_t = \sum_{j=1}^{m_0+1} \left(\boldsymbol{B}_j^\star \boldsymbol{x}_t + \boldsymbol{\varepsilon}_{j,t}\right) \mathbb{1}_{\{t_{j-1}\le t<t_j\}},\ t = 1, \ldots, n, \tag{2.1}$$

where $\boldsymbol{y}_t \in \mathbb{R}^{p_y}$ is the response vector at time t, $\boldsymbol{B}_j^\star \in \mathbb{R}^{p_y\times p_x}$ is the true coefficient matrix during the jth segment, $\boldsymbol{x}_t \in \mathbb{R}^{p_x}$ is the predictor vector at time t, and $\boldsymbol{\varepsilon}_{j,t} \in \mathbb{R}^{p_y}$ is a multivariate white noise during the jth segment at time t. All parameters in the model are considered fixed during each segment, whereas the coefficient matrices $\boldsymbol{B}_j^\star$ are allowed to vary over segments. The multivariate regression model requires that we estimate $p_x p_y$ parameters within each segment, which is challenging when either the number of predictors p_x or the number of responses p_y becomes large. We work under the high-dimensional setting in which we allow the number of predictors p_x and the number of response p_y to grow with the sample size, and possibly exceed the sample size n, that is, $p_x \gg n$ and/or $p_y \gg n$. As a result, we assume the coefficient matrices $\boldsymbol{B}_j^\star$ are sparse. Specifically, denote the number of nonzero elements in $\boldsymbol{B}_j^\star$ by d_j, for $j = 1, 2, \ldots, m_0 + 1$. Let $d_n^\star = \max_{1\le j\le m_0+1} d_j$ be the

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

maximum sparsity of the model. We assume that $d_n^\star$ is much smaller than p_x and p_y; see Section 4.

3. The TBFL

In this section, we introduce the proposed three-step TBFL estimation procedure. The first step selects candidate change points from among blocks and estimates each segment's coefficient matrix by solving a block-fused lasso problem. A hard-thresholding step is then added to reduce the over-selection problem from the fused lasso step. In the third step, a local exhaustive search examines every time point inside a neighborhood region based on the cluster of candidate change points estimated in the previous step. Moreover, we obtain a consistent model parameter estimate during the block-fused lasso step.

(Step I) Block-Fused Lasso. Define a sequence of time points $1 = r_0 < r_1 < \cdots < r_{k_n} = n + 1$ for block segmentation, such that $r_i - r_{i-1} \approx b_n$, for $i = 1, \ldots, k_n - 1$, where $k_n = \lfloor n/b_n \rfloor$ is the total number of blocks. To simplify the notation and without loss of generality, throughout the rest of the paper, we assume that n is divisible by b_n such that $r_i - r_{i-1} = b_n$, for all $i = 1, \ldots, k_n$. By partitioning the observations into blocks of size b_n and fixing the model parameters within each block, we set $\boldsymbol{\Theta}_1 = \boldsymbol{B}_1^\star$ and $\boldsymbol{\Theta}_i = \boldsymbol{B}_{j+1}^\star - \boldsymbol{B}_j^\star$ when $t_j \in [r_{i-1}, r_i)$, for some j, and $\boldsymbol{\Theta}_i = 0$ otherwise, for $i = 2, 3, \ldots, k_n$. Note that $\boldsymbol{\Theta}_i \neq \mathbf{0}$ for $i \geq 2$ means that $\boldsymbol{\Theta}_i$ has at least one nonzero entry, and implies a change in the coefficients.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

We now formulate the following linear regression model in terms of $\boldsymbol{\Theta}(k_n) = (\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_{k_n})'$:

$$\underbrace{\begin{pmatrix} \boldsymbol{y}_{(1:r_1-1)} \\ \boldsymbol{y}_{(r_1:r_2-1)} \\ \vdots \\ \boldsymbol{y}_{(r_{k_n-1}:r_{k_n}-1)} \end{pmatrix}}_{\mathcal{Y}} = \underbrace{\begin{pmatrix} \boldsymbol{x}_{(1:r_1-1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \boldsymbol{x}_{(r_1:r_2-1)} & \boldsymbol{x}_{(r_1:r_2-1)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_{(r_{k_n-1}:r_{k_n}-1)} & \boldsymbol{x}_{(r_{k_n-1}:r_{k_n}-1)} & \cdots & \boldsymbol{x}_{(r_{k_n-1}:r_{k_n}-1)} \end{pmatrix}}_{\mathcal{X}} \underbrace{\begin{pmatrix} \boldsymbol{\Theta}_1{}' \\ \boldsymbol{\Theta}_2{}' \\ \vdots \\ \boldsymbol{\Theta}_{k_n}{}' \end{pmatrix}}_{\boldsymbol{\Theta}(k_n)} + \underbrace{\begin{pmatrix} \boldsymbol{\zeta}_{(1:r_1-1)} \\ \boldsymbol{\zeta}_{(r_1:r_2-1)} \\ \vdots \\ \boldsymbol{\zeta}_{(r_{k_n-1}:r_{k_n}-1)} \end{pmatrix}}_{E}, \tag{3.1}$$

where $\boldsymbol{y}_{(a:b)} := (\boldsymbol{y}_a, \ldots, \boldsymbol{y}_b)'$, $\boldsymbol{x}_{(a:b)} := (\boldsymbol{x}_a, \ldots, \boldsymbol{x}_b)'$, $\boldsymbol{\zeta}_{(a:b)} := (\boldsymbol{\zeta}_a, \ldots, \boldsymbol{\zeta}_b)'$; $\mathcal{Y} \in \mathbb{R}^{n \times p_y}$, $\mathcal{X} \in \mathbb{R}^{n \times k_n p_x}$, $\boldsymbol{\Theta}(k_n) \in \mathbb{R}^{k_n p_x \times p_y}$, and $E \in \mathbb{R}^{n \times p_y}$. Letting $\pi_n = k_n p_x p_y$, $\mathbf{y} = \mathrm{vec}(\mathcal{Y})$, $\mathbf{Z} = I_{p_y} \otimes \mathcal{X}$, and $\boldsymbol{\theta} = \mathrm{vec}(\boldsymbol{\Theta}(k_n))$, the regression model (3.1) can be written int vector form as $\mathbf{y} = \mathbf{Z}\boldsymbol{\theta} + \mathrm{vec}(E)$, where $\mathbf{y} \in \mathbb{R}^{np_y}$, $\mathbf{Z} \in \mathbb{R}^{np_y \times \pi_n}$, $\boldsymbol{\theta} \in \mathbb{R}^{\pi_n}$, and $\mathrm{vec}(E) \in \mathbb{R}^{np_y}$. Owing to the sparsity of the parameter $\boldsymbol{\theta}$, we can estimate it using an ℓ_1-penalized least squares regression of the form:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^{\pi_n}}{\arg\min} \left\{ \frac{1}{n} \|\mathbf{y} - \mathbf{Z}\boldsymbol{\theta}\|_2^2 + \lambda_{1,n} \|\boldsymbol{\theta}\|_1 + \lambda_{2,n} \sum_{i=1}^{k_n} \left\| \sum_{j=1}^{i} \boldsymbol{\Theta}_j \right\|_1 \right\}, \tag{3.2}$$

which uses a fused lasso penalty to control the number of change points, and a lasso penalty to control the sparsity of the coefficient parameter in the model. Denote the sets of indices of blocks with nonzero jumps and estimated change points obtained from solving (3.2) by

$$\widehat{I}_n = \left\{ \widehat{i}_1, \widehat{i}_2, \ldots, \widehat{i}_{\widehat{m}} \right\} = \left\{ i : \left\| \widehat{\boldsymbol{\Theta}}_i \right\|_F \neq 0,\ i = 2, \ldots, k_n \right\} \text{ and } \widehat{\mathcal{A}}_n = \{ \widehat{t}_1, \ldots, \widehat{t}_{\widehat{m}} \} = \left\{ r_{i-1} : i \in \widehat{I}_n \right\},$$

where $\widehat{m} = \left| \widehat{\mathcal{A}}_n \right|$ and $\widehat{\mathcal{A}}_n \subset \{r_1, \ldots, r_{k_n - 1}\}$. A data-driven method for selecting the optimal block size is provided in Section 6.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

(Step II) Hard-thresholding Procedure. The estimated change points estimated from (3.2) in the block-fused lasso step include all block-end points with nonzero $\widehat{\boldsymbol{\Theta}}$, leading to an over-estimation of the number of true change points in the model. To remedy this, we include a hard-thresholding step to "thin out" redundant change points with small changes in the estimated coefficients. Intuitively, we keep estimated change points from the first step that have jumps that are sufficiently large (above a threshold). Specifically, denote the sets of indices of candidate blocks and estimated change points after hard-thresholding by

$$\widetilde{I}_n = \left\{ i : \left\| \widehat{\boldsymbol{\Theta}}_i \right\|_F > \omega_n,\ i = 2, \ldots, k_n \right\} \text{ and } \widetilde{A}_n = \{\widetilde{t}_1, \ldots, \widetilde{t}_{\widetilde{m}}\} = \left\{ r_{i-1} : i \in \widetilde{I}_n \right\},$$

where ω_n is proportional to the minimum jump sizes $\nu_n = \min_{1 \le j \le m_0} \left\| \boldsymbol{B}^\star_{j+1} - \boldsymbol{B}^\star_j \right\|_F$. Given that ν_n is unknown, we introduce a data-driven procedure to select a threshold value ω_n (see the Supplementary Material S2.1).

(Step III) Exhaustive Search Procedure. After hard thresholding, the candidate change points located far from any true change points have been removed. However, there may be more than one change point remaining in the set $\widetilde{\mathcal{A}}_n$ in the neighborhood of each true change point. Thus, we cluster the remaining estimated change points based on how close they are to each other, assuming that the number of clusters is a reasonable estimate for m_0, the number of true change points. We consider a block clustering step based on a data-driven procedure to partition the $\widetilde{m}$ candidate change points into $\widetilde{m}^f$ clusters. In particular, we select the optimal number of clusters using gap statistics (Tibshirani et al., 2001) (see Supplementary Material S2.2). For a set A, define the cluster (A) as the partition of A

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

based on the clustering algorithm. Denote the subset in cluster $\left(\widetilde{\mathcal{A}}_n\right)$ by cluster $\left(\widetilde{\mathcal{A}}_n\right) = \{R_1, \ldots, R_{\widetilde{m}^f}\}$, where $\widetilde{m}^f = \left|\text{cluster}\left(\widetilde{\mathcal{A}}_n\right)\right|$. Denote the set of corresponding indices by cluster $\left(\widetilde{I}_n\right) = \{J_1, J_2 \ldots, J_{\widetilde{m}^f}\}$.

Next, we describe the local exhaustive search procedure for estimating the location of the change points. First, define the following local coefficient parameter estimates for each segment:

$$\widehat{\boldsymbol{B}}_j = \sum_{i=1}^{\lfloor \frac{1}{2}(\max(J_{j-1})+\min(J_j))\rfloor} \widehat{\boldsymbol{\Theta}}_i, \text{ for } j = 1, \ldots, \widetilde{m}^f + 1, \tag{3.3}$$

where $J_0 = \{1\}$, $J_{\widetilde{m}^f+1} = \{k_n\}$, and $\left\{\widehat{\boldsymbol{\Theta}}_i, i = 1, \ldots, k_n\right\}$ are matrix-form parameters estimated from (3.2). Define $l_j = (\min(R_j) - b_n)\mathbb{1}_{\{|R_j|=1\}} + \min(R_j)\mathbb{1}_{\{|R_j|>1\}}$ and $u_j = (\max(R_j) + b_n)\mathbb{1}_{\{|R_j|=1\}} + \max(R_j)\mathbb{1}_{\{|R_j|>1\}}$. Now, given a subset R_j, we apply the exhaustive search method for each time point s in the interval (l_j, u_j) to the data set truncated by the two end points in time, that is, we consider only data within the interval $[\min(R_j) - b_n, \max(R_j) + b_n)$. Specifically, define the final estimated change point $\widetilde{t}_j$ as

$$\widetilde{t}_j^f = \underset{s\in(l_j,u_j)}{\arg\min} \left\{ \sum_{t=\min(R_j)-b_n}^{s-1} \left\| \boldsymbol{y}_t - \widehat{\boldsymbol{B}}_j \boldsymbol{x}_t \right\|_2^2 + \sum_{t=s}^{\max(R_j)+b_n-1} \left\| \boldsymbol{y}_t - \widehat{\boldsymbol{B}}_{j+1} \boldsymbol{x}_t \right\|_2^2 \right\}, \tag{3.4}$$

for $j = 1, \ldots, \widetilde{m}^f$. Denote the set of final estimated change points from (3.4) by $\widetilde{\mathcal{A}}_n^f = \left\{\widetilde{t}_1^f, \ldots, \widetilde{t}_{\widetilde{m}^f}^f\right\}$. Note that the local model parameter estimates $\widehat{\boldsymbol{B}}_j$ defined in (3.3) can serve as estimations for the parameters $\boldsymbol{B}_j^\star$. Thus, as mentioned in Section 1, the TBFL can estimate the model parameters together with the change-point detection, without any refitting. To enhance the variable selection properties of the model parameter estimates, we propose hard-

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

thresholding $\widehat{\boldsymbol{B}}_j$. Specifically, define the thresholded estimate $\widetilde{\boldsymbol{B}}_j$ as

$$\widetilde{\boldsymbol{B}}_j = \widehat{\boldsymbol{B}}_j \mathbb{1}_{\{|\widehat{\boldsymbol{B}}_j| > \eta_{n,j}\}}, \text{ for } j = 1, \ldots, \widetilde{m}^f + 1, \tag{3.5}$$

which is element-wise thresholding such that $\widehat{\boldsymbol{B}}_{j,hl} = 0$ if $\left|\widehat{\boldsymbol{B}}_{j,hl}\right| \leq \eta_{n,j}$, and is unchanged otherwise, for all $j = 1, \ldots, \widetilde{m}^f + 1, h = 1, \ldots, p_y, l = 1, \ldots, p_x$. The thresholding parameter $\eta_{n,j}$ is selected using the BIC (see the Supplementary Material S2.4).

Remark 1. Note that the hard-thresholding (Step II) is used only to select potential change-point locations with large changes in their estimated coefficients. To guarantee a consistent estimation of the segment-specific model parameters $\boldsymbol{B}_j$, those $\widehat{\Theta}_i$ with smaller norm values are still kept in the local coefficient parameter estimates (3.3); see the Supplementary Material S2.3.

Remark 2. The approximate computational complexity of the TBFL method is $O\left(n/b_n + b_n\right)$ for fixed p_y, p_x, and finite m_0. The computational time is $O\left(n/b_n\right)$ in the first step (Bleakley and Vert, 2011), and $O\left(2b_n\right)$ in the exhaustive search step. Note that b_n can be selected as n^ϵ such that $0 \leq \epsilon < 1$. Setting $\epsilon = 0$ (i.e., selecting b_n as a constant) yields to linear computational complexity (which matches the complexity of the fused lasso). When $0 < \epsilon < 1$, the computational complexity is $O\left(n^{\max(\epsilon, 1-\epsilon)}\right)$, which is *sub-linear* with respect to the sample size.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

4. Theoretical Properties

In this section, we provide the asymptotic properties of the TBFL in terms of both detection accuracy and model parameter estimation consistency. The following assumptions are needed:

(A1.) Lower restricted eigenvalue condition (Lower-RE condition). There exist constants $c_1, c_2 > 0$, a sequence $\delta_n \to +\infty$, $a_n = \Omega(\log(p_x p_y \vee n))$, and parameters $\alpha_1 > 0$ and $\tau = c_0\alpha_1(a_n)^{-1}\log(p_x p_y \vee n) > 0$ such that, with probability at least $1 - c_1\exp(-c_2\delta_n)$, for all $v \in \mathbb{R}^{p_x p_y}$,

$$\inf_{1\le j\le m_0+1, t_j > u > l \ge t_{j-1}, |u-l|>a_n} v' I_{p_y} \otimes \left((l-u)^{-1} \sum_{t=l}^{u-1} \boldsymbol{x}_t \boldsymbol{x}_t' \right) v \ge \alpha_1 \|v\|_2^2 - \tau \|v\|_1^2. \quad (4.1)$$

Upper restricted eigenvalue condition (Upper-RE condition). There exist constants $c_1, c_2 > 0$, a sequence $\delta_n \to +\infty$, $a_n = \Omega(\log(p_x p_y \vee n))$, and parameters $\alpha_2 > 0$ and $\tau = c_0\alpha_2(a_n)^{-1}\log(p_x p_y \vee n) > 0$ such that, with probability at least $1 - c_1\exp(-c_2\delta_n)$, for all $v \in \mathbb{R}^{p_x p_y}$,

$$\sup_{1\le j\le m_0+1, t_j > u > l \ge t_{j-1}, |u-l|>a_n} v' I_{p_y} \otimes \left((l-u)^{-1} \sum_{t=l}^{u-1} \boldsymbol{x}_t \boldsymbol{x}_t' \right) v \le \alpha_2 \|v\|_2^2 + \tau \|v\|_1^2. \quad (4.2)$$

(A2.) Deviation bound condition. There exist constants $c_1, c_2 > 0$ and a sequence $\delta_n \to +\infty$ such that, with probability at least $1 - c_1\exp(-\delta_n)$, for any sequence a_n,

$$\sup_{1\le j\le m_0+1, t_j > u > l \ge t_{j-1}, |u-l|>a_n} \left\| (l-u)^{-1} \sum_{t=l}^{u-1} \boldsymbol{x}_t \boldsymbol{\varepsilon}_t' \right\|_\infty \le c_2 \sqrt{\frac{\log(p_x p_y \vee n)}{a_n}}. \quad (4.3)$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

(A3.) The matrices $\boldsymbol{B}_j^\star$ are d_j-sparse. More specifically, for all $j = 1, \ldots, m_0 + 1$, $d_j \ll p_x p_y$, that is, $d_j/(p_x p_y) = o(1)$. Moreover, there exists a positive constant $M_{\boldsymbol{B}} > 0$ such that

$$\max_{1 \le j \le m_0+1} \left\| \boldsymbol{B}_j^\star \right\|_\infty \le M_{\boldsymbol{B}}.$$

(A4.) Let $\nu_n = \min_{1 \le j \le m_0} \left\| \boldsymbol{B}_{j+1}^\star - \boldsymbol{B}_j^\star \right\|_F$ and $\Delta_n = \min_{1 \le j \le m_0} |t_{j+1} - t_j|$. There exists a positive sequence b_n such that, as $n \to \infty$,

$$\frac{\Delta_n}{b_n} \to +\infty, \; d_n^\star \frac{\log(p_x p_y \vee n)}{b_n} \to 0, \nu_n = \Omega\left(\sqrt{\frac{d_n^\star \log(p_x p_y \vee n)}{b_n}} \right).$$

(A5.) The regularization parameters $\lambda_{1,n}$ and $\lambda_{2,n}$ satisfy $\lambda_{1,n} = C_1 \sqrt{\log(p_x p_y \vee n)/n} \sqrt{b_n/n}$ and $\lambda_{2,n} = C_2 \sqrt{\log(p_x p_y \vee n)/n} \sqrt{b_n/n}$, for some large constants $C_1, C_2 > 0$.

Assumptions A1 and A2 are common in high-dimensional linear regression models (Loh and Wainwright, 2012) and hold for a wide range of models with possible temporal dependence (Basu and Michailidis, 2015). These assumptions should hold uniformly over all $(m_0 + 1)$ segments, owing to changes in the model parameters. Assumption A3 is related to the sparsity of the model which we need because of the high dimensionality of the model (i.e., $p_x \gg n$ and $p_y \gg n$). Furthermore, it puts an upper bound on the entries of the coefficient matrices, which is a common assumption in the change-point detection literature (e.g., see Assumption A2 in Safikhani and Shojaie (2020)). Assumption A4 connects several important quantities, including the minimum jump size required for the coefficient matrices to make the change point detectable, the block size used in the TBFL algorithm, the total

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

sparsity allowed in the model, and the minimum spacing between consecutive change points. Specifically, the block size should be selected to be significantly smaller than Δ_n so that the TBFL does not miss any true break points (i.e., to ensure there is at most one true change point in each block). The method can also handle the case of a diverging number of change points (i.e., $m_0 \to \infty$) as the sample size n diverges. On the other hand, the total sparsity allowed in the model, d_n^*, can increase proportionally with the block size b_n. Note that in the case of no change points, we can set $b_n = n$; thus, the constraint on the model sparsity becomes similar to that for high-dimensional linear regression models with no change points (Loh and Wainwright, 2012). Furthermore, a higher b_n allows the jump size ν_n to be smaller, while still allowing the TBFL to detect all change points in the model consistently. Finally, Assumption A5 specifies the rate of the tuning parameters $\lambda_{1,n}$ and $\lambda_{2,n}$ in the block-fused lasso problem in (3.2). Note that in the case of no change points, we can set $b_n = n$, in which case, the rates in Assumption A5 become the typical rates of the tuning parameters in high-dimensional regression models (Loh and Wainwright, 2012).

The first theorem is one of our main results about the false positive rate of the first step of the TBFL, and the consistency of the number of change points in the second step of the TBFL.

Theorem 1. *Suppose A1–A5 hold. Then, as* $n \to +\infty$,

$$\mathbb{P}\left(d_H\left(\widetilde{\mathcal{A}}_n, \mathcal{A}_n\right) < b_n, m_0 \leq \left|\widetilde{\mathcal{A}}_n\right| \leq 2m_0 \text{ and } \widetilde{m}^f = m_0\right) \to 1.$$

Theorem 1 states that the number of clusters obtained in the second step of the TBFL

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

is a consistent estimator for the number of true change points m_0, despite the fact that the total number of estimated change points in this step can be larger than m_0. Note that although the number of candidate change points in the second step of the TBFL can be larger than m_0, Theorem 1 states that it can be at most $2m_0$, with high probability. Moreover, all candidate change points in the second step of the TBFL are within a b_n-radius neighborhood of a true change point, with high probability. In other words, none of the candidate change points are far from true change points, which is not true for the fused lasso (Safikhani and Shojaie, 2020).

The exhaustive search procedure (third step of the TBFL) removes additional candidate break points from the clusters estimated in the second step of the TBFL. The next theorem states our main result on the accuracy of locating break points in the TBFL.

Theorem 2. *Suppose Assumptions A1–A5 hold. Then, as $n \to +\infty$, there exists a large enough constant $K > 0$ such that*

$$\mathbb{P}\left(\max_{1 \le j \le m_0} \left| \widetilde{t}_j^f - t_j \right| \le \frac{K d_n^\star \log(p_x p_y \vee n)}{\nu_n^2} \right) \to 1.$$

Theorem 2 states the localization error of the TBFL algorithm uniformly over all m_0 change points. It scales logarithmically with respect to the model dimensions p_x and p_y. Moreover, small jump sizes can potentially worsen the consistency rate for locating break points because the localization error scales proportionally with respect to the reciprocal of ν_n^2. Note that the rate stated in Theorem 2 is optimal up to a logarithm factor (Csörgö and Horváth, 1997).

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

Finally, we can achieve consistent estimations of segment-specific model parameters, as stated in the following theorem.

Theorem 3. *Suppose Assumptions A1–A5 hold. Then, solution* $\widehat{\boldsymbol{B}}_j$ *from* (3.3) *satisfies*

$$\max_{1\le j\le m_0+1}\left\|\widehat{\boldsymbol{B}}_j-\boldsymbol{B}_j^\star\right\|_F=O_p\left(\sqrt{\frac{d_n^\star\log(p_xp_y\vee n)}{b_n}}\right).$$

Further, if $\eta_{n,j}=C_j\sqrt{\frac{\log(p_xp_y\vee n)}{b_n}}$, *for some positive constant* C_j, *then the thresholded variant* $\widetilde{\boldsymbol{B}}_j$ *in* (3.5) *satisfies*

$$\max_{1\le j\le m_0+1}\left|supp(\widetilde{\boldsymbol{B}}_j)\backslash supp(\boldsymbol{B}_j^\star)\right|=O_p\left(d_n^\star\right).$$

Theorem 3 states that the estimator $\widehat{\boldsymbol{B}}_j$ of a model parameter exhibits proper consistency, whereas its thresholded version $\widetilde{\boldsymbol{B}}_j$ satisfies the variable selection property. Note that in the case of no break points, by selecting $b_n=n$, the rates stated in Theorem 3 match the typical consistency rates in high-dimensional regression models (Loh and Wainwright, 2012; Basu and Michailidis, 2015). Thus, the b_n in the denominator of the consistency rate in Theorem 3 serves as a proxy for the sample size in each segment.

5. Examples of Models

In this section, we discuss two examples of well-known models that fit into the modeling framework (2.1). A third example based on a high-dimensional regression model is presented in the Supplementary Material, Section S3.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

5.1 Mean Shift Model

We consider a simple regression model in which the value of the mean changes over time. In this case, setting the parameters $\boldsymbol{x}_t = 1$, $\boldsymbol{B}_j^\star = \boldsymbol{\mu}_j^\star$, $p_x = 1$, and $p_y = p$ in the model representation in (2.1), the structural break mean shift model is given by

$$\boldsymbol{y}_t = \sum_{j=1}^{m_0+1} \left(\boldsymbol{\mu}_j^\star + \boldsymbol{\varepsilon}_{j,t}\right) \mathbb{1}_{\{t_{j-1} \le t < t_j\}}, \ t = 1, \ldots, n, \tag{5.1}$$

where $\boldsymbol{y}_t \in \mathbb{R}^p$ is the observation vector at time t, $\boldsymbol{\mu}_j^\star \in \mathbb{R}^p$ is the sparse mean vector during the jth segment, and $\boldsymbol{\varepsilon}_{j,t} \in \mathbb{R}^p$ is multivariate white noise during the jth segment at time t. Define $\boldsymbol{\Theta}_1 = \boldsymbol{\mu}_1^\star$, $\boldsymbol{\Theta}_i = \boldsymbol{\mu}_{j+1}^\star - \boldsymbol{\mu}_j^\star$ when $t_j \in [r_{i-1}, r_i)$, and $\boldsymbol{\Theta}_i = \boldsymbol{0}$ otherwise, for $i = 2, 3, \ldots, k_n$. In this case, the linear regression model in terms of Θ can be written as

$$\underbrace{\begin{pmatrix} \boldsymbol{y}_{(1:r_1-1)} \\ \boldsymbol{y}_{(r_1:r_2-1)} \\ \vdots \\ \boldsymbol{y}_{(r_{k_n-1}:r_{k_n}-1)} \end{pmatrix}}_{\mathcal{Y}} = \underbrace{\begin{pmatrix} \boldsymbol{1}_{(1:r_1-1)} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{1}_{(r_1:r_2-1)} & \boldsymbol{1}_{(r_1:r_2-1)} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{1}_{(r_{k_n-1}:r_{k_n}-1)} & \boldsymbol{1}_{(r_{k_n-1}:r_{k_n}-1)} & \cdots & \boldsymbol{1}_{(r_{k_n-1}:r_{k_n}-1)} \end{pmatrix}}_{\mathcal{X}} \underbrace{\begin{pmatrix} \boldsymbol{\Theta}_1{}' \\ \boldsymbol{\Theta}_2{}' \\ \vdots \\ \boldsymbol{\Theta}_{k_n}{}' \end{pmatrix}}_{\Theta} + \underbrace{\begin{pmatrix} \boldsymbol{\zeta}_{(1:r_1-1)} \\ \boldsymbol{\zeta}_{(r_1:r_2-1)} \\ \vdots \\ \boldsymbol{\zeta}_{(r_{k_n-1}:r_{k_n}-1)} \end{pmatrix}}_{E}, \tag{5.2}$$

where $\boldsymbol{y}_{(a:b)} := (\boldsymbol{y}_a, \ldots, \boldsymbol{y}_b)'$, $\boldsymbol{\zeta}_{(a:b)} := (\boldsymbol{\zeta}_a, \ldots, \boldsymbol{\zeta}_b)'$, $\boldsymbol{1}_{(a:b)} \in \mathbb{R}^{b-a+1}$ is an all-ones vector, $\mathcal{Y} \in \mathbb{R}^{n \times p}$, $\mathcal{X} \in \mathbb{R}^{n \times k_n}$, $\Theta \in \mathbb{R}^{k_n \times p}$, and $E \in \mathbb{R}^{n \times p}$. Applying the TFBL algorithm to this model, the estimated coefficient parameters are given by

$$\widehat{\boldsymbol{\mu}}_j = \sum_{i=1}^{\left\lfloor \frac{1}{2}(\max(J_{j-1}) + \min(J_j)) \right\rfloor} \widehat{\boldsymbol{\Theta}}_i, \ \text{for } j = 1, \ldots, \widetilde{m}^f + 1, \tag{5.3}$$

and its thresholded variant estimate $\widetilde{\boldsymbol{\mu}}_j$ is defined as

$$\widetilde{\boldsymbol{\mu}}_j = \widehat{\boldsymbol{\mu}}_j \mathbb{1}_{\{|\widehat{\boldsymbol{\mu}}_j| > \eta_{n,j}\}}, \ \text{for } j = 1, \ldots, \widetilde{m}^f + 1. \tag{5.4}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

To establish the consistency properties of the detection/estimation procedure, the following assumptions are needed:

(B1.) For the jth segment, where $j = 1, 2, \ldots, m_0 + 1$, the process can be written as $\boldsymbol{y}_{j,t} = \boldsymbol{\mu}_j^\star + \boldsymbol{\varepsilon}_{j,t}$, where the error $\{\boldsymbol{\varepsilon}_{j,t}\}$ is a subGaussian random vector with parameter (Σ_j, σ_j^2) (see the subGaussian definition in Appendix S1). Furthermore,

$$1/C_1 \leq \min_{1\leq j\leq m_0+1} \Lambda_{\min}(\Sigma_j) \leq \max_{1\leq j\leq m_0+1} \Lambda_{\max}(\Sigma_j) \leq C_1, \text{ and } 1/C_2 < \min_{1\leq j\leq m_0+1} \sigma_j^2 \leq \max_{1\leq j\leq m_0+1} \sigma_j^2 < C_2,$$

where C_1 and C_2 are positive constants.

(B2.) The mean vectors $\boldsymbol{\mu}_j^\star$ are sparse. More specifically, for all $j = 1, 2, \ldots, m_0 + 1$, $d_j \ll p$, that is, $d_j/p = o(1)$. Moreover, there exists a positive constant $M_{\boldsymbol{\mu}} > 0$ such that

$$\max_{1\leq j\leq m_0+1} \|\boldsymbol{\mu}_j^\star\|_\infty \leq M_{\boldsymbol{\mu}}.$$

(B3.) Let $\nu_n = \min_{1\leq j\leq m_0} \left\|\boldsymbol{\mu}_{j+1}^\star - \boldsymbol{\mu}_j^\star\right\|_2$. There exists a positive sequence b_n such that, as $n \to \infty$,

$$\frac{\min_{1\leq j\leq m_0+1} |t_j - t_{j-1}|}{b_n} \to +\infty, \; d_n^\star \frac{\log(p \vee n)}{b_n} \to 0 \text{ and } \nu_n = \Omega\left(\sqrt{\frac{d_n^\star \log(p \vee n)}{b_n}}\right).$$

(B4.) The regularization parameters $\lambda_{1,n}$ and $\lambda_{2,n}$ satisfy $\lambda_{1,n} = C_1\sqrt{\log(p \vee n)/n}\sqrt{b_n/n}$ and $\lambda_{2,n} = C_2\sqrt{\log(p \vee n)/n}\sqrt{b_n/n}$, for some large constants $C_1, C_2 > 0$.

Assumption B1 is a standard assumption in mean shift models that allows us to obtain the concentration inequalities needed in high dimensions, including the restricted eigenvalue

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

and deviation bound conditions (see Loh and Wainwright (2012)). Assumptions B2–B4 are special cases of Assumptions A3–A5 in Section 4. The next theorem states the detection and estimation consistency of the TBFL in a mean shift model.

Theorem 4 (Results for mean shift model)**.** *Suppose Assumptions B1–B4 hold. Then, there exists a large enough constant $K > 0$ such that, as $n \to +\infty$,*

$$\mathbb{P}\left(\widetilde{m}^f = m_0, \max_{1\le j\le m_0} \left|\widetilde{t}_j - t_j\right| \le \frac{K d_n^\star \log(p\vee n)}{\nu_n^2}\right) \to 1.$$

In addition, the solution $\widehat{\boldsymbol{\mu}}_j$ from (5.3) *satisfies*

$$\max_{1\le j\le m_0+1} \left\|\widehat{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_j^\star\right\|_F = O_p\left(\sqrt{\frac{d_n^\star \log(p\vee n)}{b_n}}\right).$$

Furthermore, if $\eta_{n,j} = C_j\sqrt{\frac{\log(p\vee n)}{b_n}}$ for some positive constant C_j, the thresholded variant $\widetilde{\boldsymbol{\mu}}_j$ from (5.4) *satisfies*

$$\max_{1\le j\le m_0+1} \left|supp(\widetilde{\boldsymbol{\mu}}_j)\backslash supp(\boldsymbol{\mu}_j^\star)\right| = O_p\left(d_n^\star\right).$$

The localization error rate obtained in Theorem 4 for the mean shift model is superior to those of the sparsified binary segmentation (SBS) algorithm of Cho and Fryzlewicz (2015) and the Inspect algorithm (Wang and Samworth, 2016). Note that our rate of consistency for estimating the break point locations is of order $d_n^\star \log(p\vee n)/\nu_n^2$, which could be as low as $(\log(p\vee n))^{1+v}$ if we set a constant ν_n and $d_n^\star = (\log(p\vee n))^\nu$. Cho and Fryzlewicz (2015) achieve a similar rate when Δ_n is of order n. However, when Δ_n is smaller and is of order n^ψ, for some $\psi \in (6/7, 1)$, their rate of consistency is of order $n^{2-2\psi}$, which is larger

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

than our logarithmic rate. Moreover, Wang and Samworth (2016) proposed a two-stage procedure called "Inspect" for estimating change points that guarantees the recovery of the correct number of change points, with high probability. Translating to our notation, their best localization error is at least of order $m_0^4(\log n + \log p)/\nu_n^2$ (see Theorem 5 in Wang and Samworth (2016)), where m_0 is the number of change points. This rate can be larger than the rate stated in Theorem 4, especially when m_0 is large. We also compared the performance of the three methods (TBFL, SBS, and Inspect) numerically; see Section 7.

5.2 Gaussian Graphical Model

In this section, we consider a Gaussian graphical model with possible changes in its covariance (precision) matrix. Specifically, suppose there exist m_0 change points $\{t_1, \ldots, t_{m_0}\}$ such that $1 = t_0 < t_1 < \cdots < t_{m_0} < t_{m_0+1} = n + 1$. Then,

$$\boldsymbol{x}_t \sim \sum_{j=1}^{m_0+1} \mathcal{N}_p\left(\mathbf{0}, \Sigma_j\right) \mathbb{1}_{\{t_{j-1} \le t < t_j\}}, \ t = 1, \ldots, n, \tag{5.5}$$

such that the observations $\boldsymbol{x}_t \in \mathbb{R}^p$ are p-dimensional realizations of a multivariate normal distribution with zero mean and covariance matrix Σ_j during the jth segment. Let $\Omega_j := \Sigma_j^{-1}$ denote the precision matrix during the jth segment, with elements $(\Omega_j(l,k))$, for $1 \le l, k \le p$. We estimate the change points and the nonzero elements of the precision matrices. Setting the parameters $\boldsymbol{x}_t = \boldsymbol{y}_t$ and $p_x = p_y = p$ in the model representation in (2.1), the model (5.5) can be expressed equivalently as the following regression equation (using the neighborhood

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

selection method of Meinshausen and Bühlmann (2006)):

$$\boldsymbol{x}_t = \sum_{j=1}^{m_0+1} \left(\boldsymbol{A}_j^\star \boldsymbol{x}_t + \boldsymbol{\varepsilon}_{j,t}\right) \mathbb{1}_{\{t_{j-1} \le t < t_j\}},\ t = 1, \ldots, n, \tag{5.6}$$

where $\boldsymbol{x}_t$ is the p-vector of the observation at time t; $\boldsymbol{A}_j^\star \in \mathbb{R}^{p\times p}$ is a sparse coefficient matrix with a zero diagonal during the jth segment, such that the off-diagonal elements $\boldsymbol{A}_j^\star(l,-l) = \Sigma_j(l,-l)\left(\Sigma_j(-l,-l)\right)^{-1} = -\left(\Omega_j(l,l)\right)^{-1}\Omega_j(l,-l)$, where $\Sigma(-l,-k)$ is the sub-matrix of Σ with its lth row and kth column removed; $\Sigma(l,k)$ is the entry of matrix Σ that lies in the lth row and kth column; and $\boldsymbol{\varepsilon}_{j,t}$ is a multivariate Gaussian white noise, such that $\boldsymbol{\varepsilon}_{j,t} \sim \mathcal{N}\left(0, \left(I_p - \boldsymbol{A}_j^\star\right)\Sigma_j\left(I_p - \boldsymbol{A}_j^\star\right)'\right)$, where the variance of the lth component in the error term $\mathrm{Var}(\varepsilon_{j,t}(l)) = \Sigma_j(l,l) - \Sigma_j(l,-l)\left(\Sigma_j(-l,-l)\right)^{-1}\Sigma_j(-l,l)$. Therefore, we have $\Omega_j(l,l) = \left(\mathrm{Var}(\varepsilon_{j,t}(l))\right)^{-1}$ and $\Omega_j(l,-l) = -\left(\mathrm{Var}(\varepsilon_{j,t}(l))\right)^{-1}\boldsymbol{A}_j^\star(l,-l)$, where $l = 1,\ldots,p$, $j = 1,\ldots,m_0+1$, $t = 1,\ldots,n$. The sparsity in the entries of Ω_j can be translated into sparsity in the regression coefficient matrix $\boldsymbol{A}_j^\star$.

Define $\boldsymbol{\Theta}_1 = \boldsymbol{A}_j^\star$, $\boldsymbol{\Theta}_i = \boldsymbol{A}_{j+1}^\star - \boldsymbol{A}_j^\star$ when $t_j \in [r_{i-1}, r_i)$ for some j, and $\boldsymbol{\Theta}_i = \boldsymbol{0}$ otherwise, for $i = 2, 3, \ldots, k_n$. In this case, the linear regression model in terms of Θ can be written as

$$\underbrace{\begin{pmatrix} \boldsymbol{x}_{(1:r_1-1)} \\ \boldsymbol{x}_{(r_1:r_2-1)} \\ \vdots \\ \boldsymbol{x}_{(r_{k_n-1}:r_{k_n}-1)} \end{pmatrix}}_{\mathcal{Y}} = \underbrace{\begin{pmatrix} \boldsymbol{x}_{(1:r_1-1)} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{x}_{(r_1:r_2-1)} & \boldsymbol{x}_{(r_1:r_2-1)} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_{(r_{k_n-1}:r_{k_n}-1)} & \boldsymbol{x}_{(r_{k_n-1}:r_{k_n}-1)} & \cdots & \boldsymbol{x}_{(r_{k_n-1}:r_{k_n}-1)} \end{pmatrix}}_{\mathcal{X}} \underbrace{\begin{pmatrix} \boldsymbol{\Theta}_1{}' \\ \boldsymbol{\Theta}_2{}' \\ \vdots \\ \boldsymbol{\Theta}_{k_n}{}' \end{pmatrix}}_{\Theta} + \underbrace{\begin{pmatrix} \boldsymbol{\zeta}_{(1:r_1-1)} \\ \boldsymbol{\zeta}_{(r_1:r_2-1)} \\ \vdots \\ \boldsymbol{\zeta}_{(r_{k_n-1}:r_{k_n}-1)} \end{pmatrix}}_{E}, \tag{5.7}$$

where $\boldsymbol{x}_{(a:b)} := (\boldsymbol{x}_a, \ldots, \boldsymbol{x}_b)'$, $\boldsymbol{\zeta}_{(a:b)} := (\boldsymbol{\zeta}_a, \ldots, \boldsymbol{\zeta}_b)'$, $\mathcal{Y} \in \mathbb{R}^{n\times p}$, $\mathcal{X} \in \mathbb{R}^{n\times k_n p}$, $\Theta \in \mathbb{R}^{k_n p\times p}$, and $E \in \mathbb{R}^{n\times p}$. The TBFL algorithm can be applied to detect the change points, with the

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

estimated coefficient parameters given by

$$\widehat{\boldsymbol{A}}_j = \sum_{i=1}^{\left\lfloor \frac{1}{2}(\max(J_{j-1})+\min(J_j)) \right\rfloor} \widehat{\boldsymbol{\Theta}}_i, \text{ for } j = 1, \ldots, \widetilde{m}^f + 1, \tag{5.8}$$

and its thresholded variant estimate $\widetilde{\boldsymbol{A}}_i$ given by

$$\widetilde{\boldsymbol{A}}_j = \widehat{\boldsymbol{A}}_j \mathbb{1}_{\left\{|\widehat{\boldsymbol{A}}_j| > \eta_{n,j}\right\}}, \text{ for } j = 1, \ldots, \widetilde{m}^f + 1. \tag{5.9}$$

To establish the consistency properties of the detection/estimation procedure, the following assumptions are needed:

(D1.) For each $j = 1, 2, \ldots, m_0 + 1$, the process follows (5.6) such that $\boldsymbol{x}_{j,t} \sim \mathcal{N}(\boldsymbol{0}, \Sigma_j)$ and $\boldsymbol{\varepsilon}_{j,t} \sim \mathcal{N}\left(0, \left(I_p - \boldsymbol{A}_j^\star\right) \Sigma_j \left(I_p - \boldsymbol{A}_j^\star\right)'\right)$. Furthermore,

$$1/C_1 \leq \min_{1 \leq j \leq m_0+1} \Lambda_{\min}(\Sigma_j) \leq \max_{1 \leq j \leq m_0+1} \Lambda_{\max}(\Sigma_j) \leq C_1, \text{ and } 1/C_2 \leq \min_{1 \leq j \leq m_0+1, 1 \leq l \leq p} (\Omega_j(l,l))^{-1},$$

where C_1 and C_2 are positive constants.

(D2.) The coefficient vectors $\boldsymbol{A}_j^\star$ are sparse. More specifically, for all $j = 1, 2, \ldots, m_0 + 1$, $d_j \ll p^2$, that is, $d_j/p^2 = o(1)$. Moreover, there exists a positive constant $M_{\boldsymbol{A}} > 0$ such that

$$\max_{1 \leq j \leq m_0+1} \|\boldsymbol{A}_j^\star\|_\infty \leq M_{\boldsymbol{A}}.$$

(D3.) Let $\nu_n = \min_{1 \leq j \leq m_0} \|\boldsymbol{A}_{j+1}^\star - \boldsymbol{A}_j^\star\|_F$. There exists a positive sequence b_n such that, as $n \to \infty$,

$$\frac{\min_{1 \leq j \leq m_0+1} |t_j - t_{j-1}|}{b_n} \to +\infty, \; d_n^\star \frac{\log(p \vee n)}{b_n} \to 0 \text{ and } \nu_n = \Omega\left(\sqrt{\frac{d_n^\star \log(p \vee n)}{b_n}}\right).$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

(D4.) The regularization parameters $\lambda_{1,n}$ and $\lambda_{2,n}$ satisfy $\lambda_{1,n} = C_1\sqrt{\log(p \vee n)/n}\sqrt{b_n/n}$ and $\lambda_{2,n} = C_2\sqrt{\log(p \vee n)/n}\sqrt{b_n/n}$, for some large constants $C_1, C_2 > 0$.

We can exclude singular or nearly singular covariance matrices, based on Assumption D1, thus guaranteeing the uniqueness of $\boldsymbol{\Theta}$ (Wang et al., 2016; Meinshausen and Bühlmann, 2006). The RE condition (A1) and the deviation bound (A2) hold under Assumption D1 (see Section 4 in Bickel et al. (2009) and Lemma 12 in Zhou et al. (2011)). Assumptions D2–D4 are special cases of Assumptions A3–A5 in Section 4.

The next theorem is about the detection and estimation consistency of the TBFL when applied to Gaussian graphical model with breaks.

Theorem 5 (Results for Gaussian graphical model)**.** *Suppose Assumptions D1–D4 hold. Then, there exists a large enough constant $K > 0$ such that, as $n \to +\infty$,*

$$\mathbb{P}\left(\widetilde{m}^f = m_0, \max_{1\le j\le m_0}\left|\widetilde{t}_j - t_j\right| \le \frac{K d_n^\star \log(p \vee n)}{\nu_n^2}\right) \to 1.$$

In addition, the solution $\widehat{\boldsymbol{A}}_j$ from (5.8) satisfies

$$\max_{1\le j\le m_0+1}\left\|\widehat{\boldsymbol{A}}_j - \boldsymbol{A}_j^\star\right\|_F = O_p\left(\sqrt{\frac{d_n^\star \log(p \vee n)}{b_n}}\right).$$

Furthermore, if $\eta_{n,j} = C_j\sqrt{\frac{\log(p\vee n)}{b_n}}$, for some positive constant C_j, the thresholded variant $\widetilde{\boldsymbol{A}}_j$ from (5.9) satisfies

$$\max_{1\le j\le m_0+1}\left|supp\left(\widetilde{\boldsymbol{A}}_j\right) \setminus supp\left(\boldsymbol{A}_j^\star\right)\right| = O_p\left(d_n^\star\right).$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

The localization error stated in Theorem 5 is optimal up to a logarithmic factor. This rate is an improvement over the consistency rate of the group-fused graphical lasso (GFGL) method of Gibberd and Roy (2017), in which the localization error is of order $O(p^2 \log p / v_n^2)$ (as shown in Theorem 3.2 in Gibberd and Roy (2017)). Moreover, the TBFL achieves a better consistency rate than the $O(p \log n / v_n^2)$ localization error rate established in Kolar and Xing (2012). Finally, it achieves a similar consistency rate in terms of the localization error to that of the method of Bybee and Atchadé (2018) for a single change point, although they provide no theoretical results for the consistency of the number of change points detected using their method. We perform a numerical comparison between the TBFL and the method of Bybee and Atchadé (2018) finding that the TBFL outperforms the other method both in terms of the estimated number of change points and their location accuracy; see Section S10 of the Supplementary Material.

6. Optimal Block Size Selection

In this section, we develop a data-driven method to select the optimal block size. If the true number of change points m_0 is relatively small, the proposed TBFL algorithm is robust to changes in the block size b_n (see Section S9). However, for a large m_0, we propose selecting the optimal block size by minimizing the high-dimensional Bayesian information criterion (HBIC) of Wang and Zhu (2011) over a grid search domain. Specifically, we select the

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

optimal b_n as

$$\widehat{b}_n = \underset{b_n \in S}{\arg\min}\, \mathrm{HBIC}(b_n) = \underset{b_n \in S}{\arg\min} \left(n \log \left(\frac{1}{n} \mathrm{RSS}(b_n) \right) + 2\gamma \log(p_x p_y)\, |M(b_n)| \right),$$

where $\mathrm{RSS}(b_n) = \sum_{j=1}^{\widetilde{m}^f(b_n)+1} \sum_{t=\widetilde{t}^f_{j-1}(b_n)}^{\widetilde{t}^f_j(b_n)-1} \left\| \boldsymbol{y}_t - \widetilde{\boldsymbol{B}}_j(b_n) \boldsymbol{x}_t \right\|_2^2$ is the residual sum of squares; $\widetilde{\boldsymbol{B}}_j(b_n)$, $\widetilde{m}^f(b_n)$ and $\widetilde{t}^f_j(b_n)$ are the estimated parameters, number of change points, and locations of the change points using block size b_n, respectively; $|M(b_n)| = \sum_{j=1}^{\widetilde{m}^f(b_n)+1} \widetilde{d}_j(b_n)$, where $\widetilde{d}_j(b_n)$ is the number of nonzero elements in the coefficient parameter $\widetilde{B}_j$ in (3.5) while using the block size b_n. We follow the suggestion of Wang and Zhu (2011) for selecting γ. Note that the detection and estimation results are robust with respect to changes in γ, as investigated in Section S8 in the Supplementary Material. The details of the selection of the search domain S are provided in Section S2.5 in the Supplementary Material.

7. Numerical Performance Evaluation

In this section, we compare the empirical performance of our method (TBFL) with that of several competing methods. For the mean shift model, we compare our method with SBS (Cho and Fryzlewicz, 2015) and Inspect (Wang and Samworth, 2016). For the Gaussian graphical model, we compare our method with the SA algorithm (Bybee and Atchadé, 2018). We also evaluate the performance of the TBFL method with respect to both structural break detection and parameter estimation over several simulation scenarios. Owing to space limitations, we report only the comparison between the performance of the proposed method with that of SBS and Inspect; details on thecomparison with the method of (Bybee and

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

Atchadé, 2018) and empirical performance of the TBFL over several simulation scenarios are provided in the Supplementary Material, Sections S10 and S9, respectively.

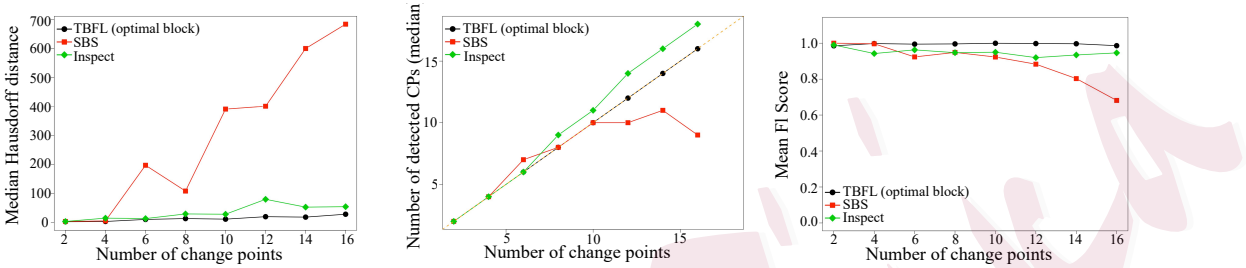

Figure 2: (a) Hausdorff distance $d_H\left(\widetilde{\mathcal{A}}_n^f, \mathcal{A}_n\right)$ for the TBFL, SBS, and Inspect methods; (b) median number of detected change points for the three methods; (c) F1 score.

Before describing the simulation settings, we need to define how we measure the detection performance of the methods. First, we use the Hausdorff distance $d_H\left(\widetilde{\mathcal{A}}_n^f, \mathcal{A}_n\right)$ to measure the estimation accuracy of the locations of the break points. Moreover, following Hushchyn et al. (2020), we define a set of correctly detected change points as true positive change points (TPCP):

$$\text{TPCP} = \left\{ t_j \middle| \exists \widetilde{t}_{j'}^f \text{ such that } \widetilde{t}_{j'}^f \in \left[t_j - \frac{t_j - t_{j-1}}{5}, t_j + \frac{t_{j+1} - t_j}{5} \right], j = 1, \ldots, m_0 \right\}.$$

In addition, the precision, recall, and F1-score are calculated as follows:

$$\text{Precision} = \frac{|\text{TPCP}|}{\widetilde{m}^f}, \quad \text{Recall} = \frac{|\text{TPCP}|}{m_0}, \quad \text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}},$$

where $|\text{TPCP}|$ is the cardinality of the set TPCP. The highest possible value of an F1 score is one, indicating perfect precision and recall, and the lowest possible value is zero, indicating

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

that either the precision or the recall is zero. We select the F1 score as another quantitative measurement to evaluate the detection performance. The simulation setting are as follows.

Setting A (Mean Shift Model). In setting A, $n = 5000$ and $p = 20$, with the number of nonzero elements in the jth segments $d_j = 2$, for all $j = 1, \ldots, m_0 + 1$. The mean coefficient μ is chosen to be multivariate with a random sparse structure and random entries sampled from $\text{Uniform}(-1, -0.5)\mathbb{1}_{\{j \text{ is odd}\}} + \text{Uniform}(0.5, 1)\mathbb{1}_{\{j \text{ is even}\}}$, for each $j = 1, \ldots, m_0 + 1$. We consider different setting of m_0, from 2 to 16.

Table 1: Results of the difference between $\widetilde{m}^f$ and m_0 for the TBFL, SBS, and Inspect methods in simulation scenario A.

method	$\lvert\widetilde{m}^f - m_0\rvert$	$m_0 = 2$	$m_0 = 4$	$m_0 = 6$	$m_0 = 8$	$m_0 = 10$	$m_0 = 12$	$m_0 = 14$	$m_0 = 16$
TBFL	0	94	**98**	**93**	**93**	**99**	**97**	**95**	**74**
	1	4	2	7	7	1	3	5	16
	2	2	0	0	0	0	0	0	6
	> 2	0	0	0	0	0	0	0	4
SBS	0	**100**	96	30	57	45	8	1	0
	1	0	3	60	40	28	25	11	2
	2	0	0	10	3	26	40	18	1
	> 2	0	1	0	0	1	27	70	97
Inspect	0	95	56	64	38	31	3	7	11
	1	5	35	24	37	41	28	22	32
	2	0	8	9	19	18	36	42	30
	> 2	0	1	3	6	10	33	29	27

The detection results of the TBFL, SBS, and Inspect methods are summarized in Figure 2 and Table 1. As shown in Figure 2 (left panel), the Hausdorff distance between the set of estimated change points and true change points increases significantly for the SBS method when m_0 increases, whereas the TBFL and Inspect seem to be more stable. The middle panel shows the median of the number of detected change points for all three methods,

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

indicating that Inspect (SBS) overestimates (underestimates) the true number of change points, whereas the TBFL correctly identifies m_0. The right panel of Figure 2 depicts the F1 score, showing that for small m_0, all models perform reasonably well, but that the TBFL outperforms SBS and Inspect for larger m_0. Overall, the TBFL outperforms these two competing methods in terms of estimating both the number of change points and their locations. Finally, as shown in Table 1, among 100 replicates, our method correctly estimates m_0 over 90% replicates when $m_0 = 2$ to 14, whereas SBS (Inspect) tends to underestimate (overestimate) m_0 starting from $m_0 = 6$. Note that for $m_0 = 16$, the TBFL selects the true number of change points in 74% replicates, which implies that with model specifications in this simulation setting, the TBFL has reached its detection limit.

8. An Application to EEG Data

In this section, we apply the TBFL method and SA method (Bybee and Atchadé, 2018) to an EEG data set analyzed in Trujillo (2019). In this database, EEG signals from active electrodes for 72 channels are recorded at a sampling frequency of 256 Hz, for approximately three minutes. The stimulus procedure tested on the selected subject comprised three one-min interleaved sessions with eyes open and closed. To speed up the computations, we construct a subset of the EEG data observations by selecting one record in every 16. After detrending and scaling the data, the total time points is reduced to $n = 2,922$. The data are also preprocessed to remove the temporal structure pattern (see the Supplementary

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

Material S11).

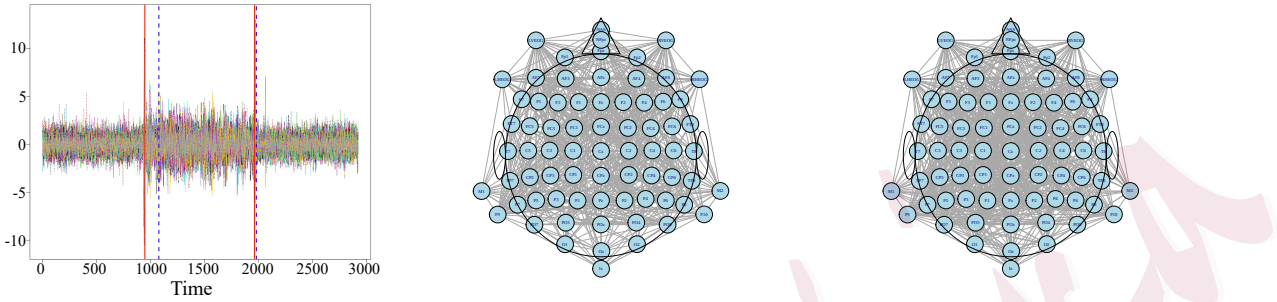

Figure 3: (left) The detrended EEG data with 72 channels. The blue dashed line indicates the detected change point, and the red vertical line indicates the true change point. The first and third parts correspond to eyes open status, and the second part corresponds to eyes closed status; (middle and right) Directed graph of EEG channels connectivity before (middle panel) and after (right panel) the first change point.

We considered the Gaussian graphical model with breaks for this data set, and applied the TBFL with an optimal block size procedure under the search domain $b_n = 80, 90, 100, 110, 120$ to detect the change points and estimate the model parameters. The selected optimal block size is $b_n = 90$. As shown in left panel of Figure 3, our method detects two break points at $\widetilde{t}_1^f = 1077$ and $\widetilde{t}_2^f = 1980$, which are close to the open-eye and closed-eye times identified by neurologists ($t_1 = 947$ and $t_2 = 1963$). We also applied the SA method (Bybee and Atchadé, 2018) to this EEG data set. The SA method detects only one change point close to the boundaries (30), for which there are no recorded stimuli, but no estimated change points close to the true change points. To demonstrate the changes between the eye-open and eye-closed segments, we focus on the first two segments, and estimate the model parameters in both segments using the thresholded estimator defined in (5.9), that is, $\widetilde{\boldsymbol{A}}_1$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

(segment 1, open-eye) and $\widetilde{\boldsymbol{A}}_2$ (segment 2, closed-eye). Network edges corresponding to nonzero coefficients in these two estimated parameters are depicted in the middle and right panels of Figure 3. During the second segment (the eyes-closed state), the overall network connectivity increased. Specifically, the total number of edges in the eyes-open (EO) state is 724, whereas the total number of edges in the eyes-closed (EC) state is 857. Among the channels with the most connectivity changes, that is, the degree (number of edges) between the two segments changed the most, there are six EEG channels, namely PO4, POz, PO3, Pz, P3, and CP2, which are located in the visual cortex in the brain (Nezamfar et al., 2011). This result confirms the satisfactory variable selection performance of the model parameter estimation, as stated in Theorem 3, after detecting break points in the TBFL procedure. Such estimations can produce insights into which channels have been most affected by the stimulus procedure.

9. Conclusion

We have introduced a novel unified framework that can consistently identify structural breaks and estimate model parameters for general sparse multivariate linear models with high-dimensional covariates. We have developed a regularized estimation procedure to simultaneously detect the structural break points and estimate the model parameters. Key technical developments include the calibration of the block size and the introduction of hard-thresholding for screening out redundant candidate change points. Note that our method

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

can also handle the VAR model. An extension of the current framework to nonlinear models would be an interesting future research direction.

Supplementary Material

In Section S1 of the online Supplementary Material, we define a sub-Gaussian random variable and a sub-Gaussian random vector, and in Section S2, we discuss the algorithm and the data-driven procedures. Section S3 presents an example of a high-dimensional multiple linear regression model, and additional details about the Gaussian graphical model are provided in Section S4. Section S5 contains the technical lemmas needed to prove the main results. The proofs of the main results are given in Section S6. Finally, Sections S8 and S9 and S10 present additional results for simulation scenario A, additional simulation results for several different settings, and a comparison with the simulated annealing method (Bybee and Atchadé, 2018), respectively.

References

Aue, A. and L. Horváth (2013). Structural breaks in time series. *Journal of Time Series Analysis 34*(1), 1–16.

Aue, A., G. Rice, and O. Sönmez (2017). Detecting and dating structural breaks in functional data without dimension reduction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Bai, P., A. Safikhani, and G. Michailidis (2020). Multiple change points detection in low rank and sparse high dimensional vector autoregressive models. *IEEE Transactions on Signal Processing 68*, 3074–3089.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

Bai, Y. and A. Safikhani (2021). *LinearDetect: Change Point Detection in High-Dimensional Linear Regression Models*. R package version 0.1.4.

Basseville, M. and I. V. Nikiforov (1993). *Detection of abrupt changes: theory and application*, Volume 104. Prentice Hall Englewood Cliffs.

Basu, S. and G. Michailidis (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics 43*(4), 1535–1567.

Bickel, P. J., Y. Ritov, A. B. Tsybakov, et al. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics 37*(4), 1705–1732.

Bleakley, K. and J.-P. Vert (2011). The group fused lasso for multiple change-point detection. *arXiv preprint arXiv:1106.4199*.

Bybee, L. and Y. Atchadé (2018). Change-point computation for large graphical models: a scalable algorithm for gaussian graphical models with change-points. *The Journal of Machine Learning Research 19*(1), 440–477.

Chan, N. H., W. L. Ng, and C. Y. Yau (2021). A self-normalized approach to sequential change-point detection for time series. *Statistica Sinica 31*(1), 491–517.

Cho, H. (2016). Change-point detection in panel data via double cusum statistic. *Electronic Journal of Statistics 10*(2), 2000–2038.

Cho, H. and P. Fryzlewicz (2015). Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 77*(2), 475–507.

Csörgö, M. and L. Horváth (1997). *Limit theorems in change-point analysis*, Volume 18. John Wiley & Sons Inc.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

Davis, R. A., T. C. M. Lee, and G. A. Rodriguez-Yam (2006). Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association 101*(473), 223–239.

Frick, K., A. Munk, and H. Sieling (2014). Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(3), 495–580.

Frisén, M. (2008). *Financial surveillance*, Volume 71. John Wiley & Sons.

Fryzlewicz, P. (2017). Tail-greedy bottom-up data decompositions and fast mulitple change-point detection. *Annals of Statistics*.

Gibberd, A. J. and S. Roy (2017). Multiple changepoint estimation in high-dimensional gaussian graphical models. *arXiv preprint arXiv:1712.05786*.

Harchaoui, Z. and C. Lévy-Leduc (2010). Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association 105*(492), 1480–1493.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hushchyn, M., K. Arzymatov, and D. Derkach (2020). Online neural networks for change-point detection. *arXiv preprint arXiv:2010.01388*.

Jackson, B., J. D. Scargle, D. Barnes, S. Arabhi, A. Alt, P. Gioumousis, E. Gwin, P. Sangtrakulcharoen, L. Tan, and T. T. Tsai (2005). An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters 12*(2), 105–108.

Killick, R., P. Fearnhead, and I. A. Eckley (2012). Optimal detection of changepoints with a linear computational

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

cost. *Journal of the American Statistical Association 107*(500), 1590–1598.

Kolar, M. and E. P. Xing (2012). Estimating networks with jumps. *Electronic journal of statistics 6*, 2069.

Leonardi, F. and P. Bühlmann (2016). Computationally efficient change point detection for high-dimensional regression. *arXiv preprint arXiv:1601.03704*.

Liu, B., C. Zhou, X. Zhang, and Y. Liu (2020). A unified data-adaptive framework for high dimensional change point detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82*(4), 933–963.

Loh, P.-L. and M. J. Wainwright (2012, 06). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Statist. 40*(3), 1637–1664.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.

Matteson, D. S. and N. A. James (2014). A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association 109*(505), 334–345.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *Annals of statistics 34*(3), 1436–1462.

Nezamfar, H., U. Orhan, S. Purwar, K. Hild, B. Oken, and D. Erdogmus (2011). Decoding of multichannel eeg activity from the visual cortex in response to pseudorandom binary sequences of visual stimuli. *International Journal of Imaging Systems and Technology 21*(2), 139–147.

Ombao, H., R. Von Sachs, and W. Guo (2005). Slex analysis of multivariate nonstationary time series. *Journal of the American Statistical Association 100*(470), 519–531.

Qiu, P. (2013). *Introduction to statistical process control*. CRC press.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

Rinaldo, A. (2009). Properties and refinements of the fused lasso. *Annals of Statistics 37*(5B), 2922–2952.

Rothman, A. J., E. Levina, and J. Zhu (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics 19*(4), 947–962.

Roy, S., Y. Atchadé, and G. Michailidis (2017). Change point estimation in high dimensional markov random-field models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*(4), 1187–1206.

Safikhani, A., Y. Bai, and G. Michailidis (2021). Fast and scalable algorithm for detection of structural breaks in big var models. *Journal of Computational and Graphical Statistics*, 1–14.

Safikhani, A. and A. Shojaie (2020). Joint structural break detection and parameter estimation in high-dimensional nonstationary var models. *Journal of the American Statistical Association*, 1–14.

Savage, D., X. Zhang, X. Yu, P. Chou, and Q. Wang (2014). Anomaly detection in online social networks. *Social Networks 39*, 62–70.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(1), 91–108.

Tibshirani, R., G. Walther, and T. Hastie (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*(2), 411–423.

Trujillo, L. (2019). Raw Empirical EEG Data.

van de Geer, S., P. Bühlmann, S. Zhou, et al. (2011). The adaptive and the thresholded lasso for potentially misspecified models (and a lower bound for the lasso). *Electronic Journal of Statistics 5*, 688–749.

Wang, D., Y. Yu, A. Rinaldo, and R. Willett (2019). Localizing changes in high-dimensional vector autoregressive

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

processes. *arXiv preprint arXiv:1909.06359*.

Wang, L., X. Ren, and Q. Gu (2016). Precision matrix estimation in high dimensional gaussian graphical models with faster rates. In *Artificial Intelligence and Statistics*, pp. 177–185.

Wang, T. and R. J. Samworth (2016). High-dimensional changepoint estimation via sparse projection. *arXiv preprint arXiv:1606.06246*.

Wang, T. and L. Zhu (2011). Consistent tuning parameter selection in high dimensional sparse linear regression. *Journal of Multivariate Analysis 102*(7), 1141–1151.

Wang, Y. and Y. Mei (2015). Large-scale multi-stream quickest change detection via shrinkage post-change estimation. *IEEE Transactions on Information Theory 61*(12), 6926–6938.

Yu, Y. (2020). A review on minimax rates in change point detection and localisation. *arXiv preprint arXiv:2011.01857*.

Yuan, M. and Y. Lin (2007). Model selection and estimation in the gaussian graphical model. *Biometrika 94*(1), 19–35.

Zhang, T. and L. Lavitas (2018). Unsupervised self-normalized change-point testing for time series. *Journal of the American Statistical Association 113*(522), 637–648.

Zhou, S., P. Rütimann, M. Xu, and P. Bühlmann (2011). High-dimensional covariance estimation based on gaussian graphical models. *The Journal of Machine Learning Research 12*, 2975–3026.

Zhu, X., R. Pan, G. Li, Y. Liu, and H. Wang (2017). Network vector autoregression. *The Annals of Statistics 45*(3), 1096–1123.

Department of Statistics, University of Florida

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0309

E-mail: (baiyue@ufl.edu)

Department of Statistics, University of Florida

E-mail: (a.safikhani@ufl.edu)