Online Structural Change-point Detection of High-dimensional Streaming Data via Dynamic Sparse Subspace Learning

Ruiyu Xu xuruiyu@pku.edu.cn
Jianguo Wu j.wu@pku.edu.cn
Department of Industrial Engineering and Management
Peking University
Beijing, 100871, China

Xiaowei Yue xwy@vt.edu
Department of Industrial & Systems Engineering
Virginia Tech
1145 Perry Street, Blacksburg, VA 24061, USA

Yongxiang Li yongxiangli@sjtu.edu.cn
Department of Industrial Engineering and Management
Shanghai Jiao Tong University
Shanghai, 200240, China

Editor: XXXXX

Abstract

High-dimensional streaming data are becoming increasingly ubiquitous in many fields. They often lie in multiple low-dimensional subspaces, and the manifold structures may change abruptly on the time scale due to pattern shift or occurrence of anomalies. However, the problem of detecting the structural changes in a real-time manner has not been well studied. To fill this gap, we propose a dynamic sparse subspace learning (DSSL) approach for online structural change-point detection of high-dimensional streaming data. A novel multiple structural change-point model is proposed and it is shown to be equivalent to maximizing a posterior under certain conditions. The asymptotic properties of the estimators are investigated. The penalty coefficients in our model can be selected by AMDL criterion based on some historical data. An efficient Pruned Exact Linear Time (PELT) based method is proposed for online optimization and change-point detection. The effectiveness of the proposed method is demonstrated through a simulation study and a real case study using gesture data for motion tracking.

Keywords: Streaming data, Structural change-point detection, Sparse subspace learning, Asymptotic consistency, Dynamic correlation

1. Introduction

High-dimensional streaming data are ubiquitous in many fields such as bioinformatics, engineering, finance and social sciences. For example, in biological studies, neurons being monitored could generate hundreds or thousands time series signals (Qiu et al., 2016). In digital image correlation, each dynamic image in a video with high resolution could consist of more than one hundred thousand pixels. In semiconductor manufacturing, hundreds

of sensors are installed in the production system for real time monitoring of the manufacturing condition (Zhang et al., 2020). In gesture tracking, tens of sensors are mounted to dynamically capture the positions of body joints (Jiao et al., 2018). The relationship or correlation among these dimensions is of great value for research, as it provides insights into regularities and inter-dependencies between observed variables (Kolar and Xing, 2012). Usually, the correlation or dependence structure is sparse, i.e., a variable is only correlated with a small proportion of other variables. Besides, the correlation structure may change over time and the change-points often imply events or anomalies occurring at that moment. For instance, changes in the correlation between brain nerves may represent shifts in thinking content or patterns (Haslbeck and Waldorp, 2020). Changes in the correlation between image pixels may indicate transitions in the subject of the video (Tierney et al., 2014). Therefore, online detection of change in correlation or inter-dependence is of great importance to determine whether an event or anomaly has recently occurred in the system.

In this paper, we refer to the cross-correlation change as a structural change. The structural change- points separate the multivariate streaming data into multiple segments with different relationships. Multiple change-point problems have been actively studied in many fields, e.g., economics, climatic time series (Aminikhanghahi and Cook, 2017; Wu et al., 2016, 2019). However, these problems often refer to detection of breaks in trend or distributional parameters, e.g., a shift in mean or variance, while structural change detection focuses on detecting the changes of the underlying relationships among different dimensions. The structural change-point detection problem, especially the online one, has not been well explored compared with the traditional multiple change-point detection problems. Due to the curse of dimensionality?(Bellman and Corporation., 1957), it is often challenging to detect these change-points accurately and timely. Too many variables constitute an extremely complex correlation structure that is hard to estimate. Noise contamination and insufficient sample size further increase the difficulty of estimation.

Gaussian graphic model (GGM, Dempster, 1972) is a widely used method and continues to attract much attention to study the inter-dependence structure of multiple variables. A common assumption is that the sample $X \sim N_d(0, \Sigma)$ is a d-dimension Gaussian vector. Let $\Omega := \Sigma^{-1}$ denote the precision matrix, with entries $(\omega_{ij}), 1 \leq i, j \leq d$. It can be easily shown that the precision matrix Ω encodes the conditional independence structure among the variables. Variable i and j is conditionally independent given all other coordinates of X if and only if the entry ω_{ij} of the precision matrix is zero. Meinshausen and Buhlmann (2006) was the first to combine GGM with LASSO to get a sparse precision matrix Ω, and later a more systematic approach named Graphic LASSO was proposed by Friedman et al. (2008). Since then, there has been much similar work on estimating a single precision matrix Ω based on n independent samples (Drton and Perlman, 2008; Foygel and Drton, 2010; Rothman et al., 2008; Yuan and Lin, 2006; Zhao et al., 2015). However, these methods cannot track the evolution of the dependency graphs over time. To this end, several research groups (Haslbeck and Waldorp, 2020; Kolar and Xing; Qiu et al., 2016; Zhou et al., 2010) assumed that the dependency graph evolves continuously over time and proposed kernel smoothing methods for estimating time-varying graphical models. Kolar and Xing (2012) assumed that the graph changes abruptly at some time instants and proposed a penalized neighborhood selection method with a fused-type penalty for estimating a piece-wise constant graphical model. Considering that there may be prior knowledge of potential groups,

Gibberd and Nelson (2017) proposed a group-fused graphical lasso estimator for grouped estimation of change-points. However, all of the methods above are limited to the strong assumption that the samples follow a Gaussian distribution with a constant mean, which may contradict the fact that the mean of streaming data often varies over time.

Sparse subspace clustering (SSC) is another type of methods that can be used to capture the sparse dependencies or correlations across different variables (Elhamifar and Vidal, 2013). Subspace clustering is an extension of traditional clustering that seeks to find clusters in different subspaces. It is based on the fact that high-dimensional data often lie in multiple subspaces of significantly lower dimension instead of being uniformly distributed across the full space (Parsons et al., 2004). The key idea of SSC is the self-expressive property with sparse representation, i.e., each data point in a union of subspaces can be sparsely represented as a linear or affine combination of other points from its own subspace. SSC method builds a similarity graph by these sparse coefficients, and obtains data segmentation using spectral clustering. Later several structured SSC were developed by integrating the two separate stages of computing a sparse representation matrix and applying spectral clustering into a unified optimization framework (Li and Vidal; Li et al., 2017; Zhang et al.). Tierney et al. (2014) proposed an ordered subspace clustering method by including a new penalty term to handle data from a sequentially ordered union of subspaces. Guo et al. (2013) proposed a spatial subspace clustering (SpatSC) by combining subspace learning with the fused lasso for 1D hyperspectral data segmentation. However, all of these methods focus on static data of fixed length, and thus are not applicable to dynamic streaming data with increasing length. Besides, they are not able to detect the dynamic change of cross-correlation structure among variables. Recently, Zhang et al. (2020) proposed a dynamic multivariate functional data modeling approach to capture the change of cross-correlation structure over time. By formulating the problem as a sparse regression with fused LASSO penalty, the correlation structure among different variables as well as the change-points can be efficiently estimated using the Fast Iterative Shrinkage-thresholding Algorithm (FISTA). Nevertheless, this method is offline and cannot sequentially estimate the cross-correlation structure and detect the change-points. Jiao et al. (2018) proposed an online cumulative sum (CUSUM)-based control chart for subspace change-point detection. This method first learns the pre-change subspace from historical data, and then conducts online detection via a CUSUM statistic. However, this method is only applicable when there is one change-point and one subspace. Besides, it requires sufficient historical pre-change data to get the basis of the subspace.

To fill the research gap, we propose a novel dynamic sparse subspace learning approach for online detecting the change of sparse correlation structure of high-dimensional streaming data. Specifically, we follow the self-expressive assumption in (Elhamifar and Vidal, 2013) and formulate a novel multiple structural change-point model with two penalty terms in the loss function for encouraging sparse representation and avoiding excessive change-points respectively. It is then shown that the model formulation is fundamentally equivalent to maximizing a posterior under certain conditions. The asymptotic properties of the model estimators are further investigated, showing that with the number of change-points fixed, the positions of the change-points and the cross-correlation between variables within each segment converge to the true values as the length of segments increases. The model is

sequentially optimized through a customized PELT algorithm (Killick et al., 2012) to enable online change-point detection.

The rest of this paper is organized as follows. In Section 2, the dynamic sparse subspace learning based multiple structural change-point model is formulated, and is interpreted by showing its equivalence to maximizing a posterior with certain priors. The asymptotic properties of the proposed model are investigated in Section 3. In Section 4, we show how to solve the optimization problem sequentially via PELT algorithm, and propose some strategies to determine the penalty coefficients and proper hyperparameter to improve the computational efficiency. Numerical experiments with synthetic and real gesture data are conducted to demonstrate and validate the effectiveness of the proposed algorithm in Section 5. Section 6 presents the discussions and conclusions.

2. Multiple Structural Change-point Modeling via Dynamic Sparse Subspace Learning

In this section, we first introduce the subspace assumption and the self-expressive property or assumption, which lay the foundation for SSC (Elhamifar and Vidal, 2013). Then, the multiple structural change-point modeling approach is formulated. We further show that the model formulation is mathematically equivalent to maximizing a posterior (MAP) in a Bayesian framework.

2.1 Notations and Basic Assumptions

Consider a p-dimensional (e.g., p-channels) streaming data $[\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_p]$, where each dimension is of length N on the time scale, e.g., $\boldsymbol{Y}_i = [Y_{i1}, Y_{i2}, \ldots, Y_{iN}]$. We assume

$$Y_{ij} = X_{ij} + \epsilon_{ij}, i = 1, \ldots, p, j = 1, \ldots, N, \tag{1}$$

where X_{ij} is the true value and ϵ_{ij} is the independent noise with mean $E[\epsilon_{ij}] = 0$ and variance $\mathrm{Var}[\epsilon_{ij}] = \sigma^2$. We assume there is no autocorrelation in the noise. To facilitate understanding, we could treat a discrete time series $\boldsymbol{X}_i$ as a functional sample of $X_i(t)$, and assume that these functions can be partitioned into L different subspaces $S_l, l = 1, \ldots, L$. Functions in the same subspace have strong cross-correlations, while functions in different subspaces have no cross-correlations.

Assumption A1. (Subspace Assumption) *It is assumed that the multivariate streaming data can be partitioned into different subspaces. Each subspace or translated subspace S_l is defined as the set of all functions formed by linearly combining the d_l basis functions $\boldsymbol{\Phi}_l = [\phi_{l1}(t), \ldots, \phi_{ld_l}(t)]^T$ with a translation or shift function $\phi_{l0}(t)$*

$$\boldsymbol{S}_l \triangleq \left\{ X(t) \mid X(t) = \sum_{q=1}^{d_l} \alpha_q \phi_{lq}(t) + \phi_{l0}(t), \alpha_q \in \mathcal{R} \right\}. \tag{2}$$

If orthogonal basis functions are considered, then,

$$\int \phi_{li}(t)\phi_{lj}(t)dt = 0, \forall i, j = 1, \ldots, d_l, i \neq j.$$

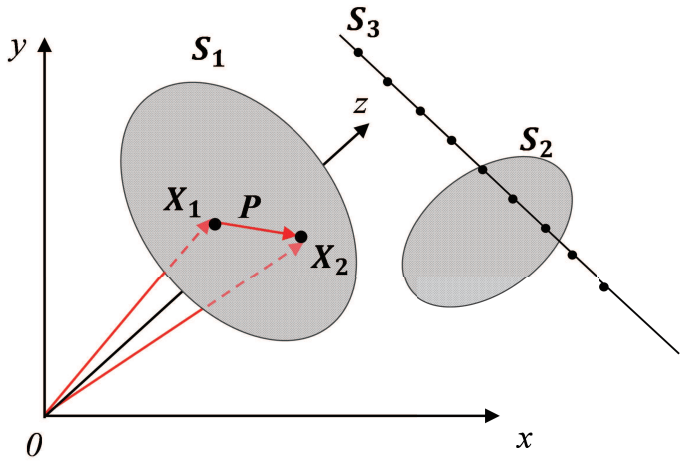

Figure 1: Illustration of one-dimensional (S_3) and two-dimensional (S_1 and S_2) subspaces in $\mathcal{R}^3$.

Note that if $\phi_{l0}(t) \neq 0$, it is a translated subspace, as it does not contain the origin (Nowinski, 1981). From the subspace assumption we can see that the time series within the same subspace share common basis functions, therefore they are expected to have strong correlations or linear relationship. In practical applications, the collected streaming data are discrete. In such a case, the basis vectors, e.g., $\boldsymbol{\phi}_{li} = (\phi_{li}(t_1), \ldots, \phi_{li}(t_N))'$ instead of the basis functions can be used to represent the subspace assumption. The formed vector subspace is actually an affine space.

Figure 1 is an illustrative example showing three subspaces in $\mathcal{R}^3$. Intuitively, given sufficient samples, each data point can be efficiently represented as a linear combination of other points in the same subspace. For example, in subspace S_1, point $\boldsymbol{X}_2 = \boldsymbol{X}_1 + \boldsymbol{P}$ where $\boldsymbol{P} = \alpha_1(\boldsymbol{X}_4 - \boldsymbol{X}_3) + \alpha_2(\boldsymbol{X}_6 - \boldsymbol{X}_5)$ for certain points $\boldsymbol{X}_3, \boldsymbol{X}_4, \boldsymbol{X}_5$ and $\boldsymbol{X}_6$ in S_1. This self-expressive property is summarized in Assumption A2 as follows.

Assumption A2. (Self-Expressive Assumption) *If there are sufficient data points (time series) from each subspace, e.g., $p_l > d_l$ for $l = 1, \ldots L$, where p_l is the number of points in subspace S_l, we have and d_l is the subspace dimension, then $\boldsymbol{X}_i$ is self-expressive, i.e., $\forall i \in \boldsymbol{P}_l$ where $\boldsymbol{P}_l$ is the set of data point indices of subspace S_l, we have*

$$\boldsymbol{X}_i = \sum_{j \in \boldsymbol{P}_l, j \neq i} \beta_{ij} \boldsymbol{X}_j, \forall i = 1, \ldots, p. \tag{3}$$

With this assumption, we have $\boldsymbol{X} = \boldsymbol{X}\boldsymbol{B}$ where $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$ is the true-value time series data in $\mathcal{R}^{N \times p}$, $\boldsymbol{B}$ is a sparse coefficient matrix in $\mathcal{R}^{N \times p}$ with entry $\beta_{ii} = 0$ and $\beta_{ij} = 0$ if $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ belong to different subspaces.

2.2 Model Formulation

The self-expressive property states that each data point or time series can be efficiently reconstructed by a linear combination of other points in the same subspace. To be more precise, each data point can be represented by other points that are not necessarily from the same subspace, and this representation is not unique in general. Ideally, there is a sparse representation where the nonzero elements correspond to the points from the same subspace and the number of nonzero elements corresponds to the dimension of the underlying subspace. Considering the existence of measurement noise, the sparse representation can be obtained by minimizing the following objective function with an l_1 penalty

$$\min_{\beta_{ij}, j\neq i} \sum_{i=1}^{p} \left\{ \frac{1}{2} \left\| Y_i - \sum_{j\neq i} \beta_{ij} Y_j \right\|^2 + \lambda_1 \sum_{j\neq i} |\beta_{ij}| \right\} \text{ s.t. } \beta_{ii} = 0, \tag{4}$$

where λ_1 is the penalty weight to control the sparsity in the representation. As the representation of each point is independent of those of other points, the above optimization problem is equivalent to the following one that can be solved efficiently using the LASSO algorithm

$$\min_{\beta_{ij}, j\neq i} \frac{1}{2} \left\| Y_i - \sum_{j\neq i} \beta_{ij} Y_j \right\|^2 + \lambda_1 \sum_{j\neq i} |\beta_{ij}| \text{ s.t. } \beta_{ii} = 0 \text{ for } i = 1, \ldots, p. \tag{5}$$

In the traditional sparse subspace clustering, the dimension of the data points and the correlation structure are fixed. However, in our case, the length of the streaming data dynamically increases and there may be abrupt structural changes, i.e., changes in $\beta_{ij}, j \neq i$ at some unknown change-points.

For a p-channel streaming data of length N, suppose there are in total C change-points $\{\tau_1, \ldots, \tau_C\}$ with $0 < \tau_1 < \cdots < \tau_C < N$, which partitions the streaming data into $C+1$ segments. For notational convenience, we define $\tau_0 = 0$ and $\tau_{C+1} = N$. In the estimation of change-point models, the fused lasso is one of the most popular techniques, which penalizes the l_1-norm of both the coefficients and their successive differences (Tibshirani et al., 2005; Tierney et al., 2014; Zhang et al., 2020). Using the fused lasso, the problem can be formulated as

$$\min_{\substack{\beta_{ijt}, j\neq i, \\ t=1,\ldots,N}} \sum_{t=1}^{N} \left[\frac{1}{2} \left(Y_{it} - \sum_{j\neq i} \beta_{ijt} Y_{jt} \right)^2 + \lambda_1 \sum_{j\neq i} |\beta_{ijt}| \right] + \lambda_2 \sum_{t=2}^{N} \sum_{j\neq i} |\beta_{ijt} - \beta_{ijt-1}|, i = 1, \ldots, p. \tag{6}$$

The above optimization can be achieved by the Fast Iterative Shrinkage-thresholding Algorithm (FISTA, Beck and Teboulle, 2009). However, this formulation is inherently offline and cannot be efficiently solved in a real-time or sequential manner for online applications. To overcome this problem, we propose a new model formulation that can be sequentially optimized

$$\min_{\substack{C, \tau_1, \ldots, \tau_C \\ \boldsymbol{B}^{(c)}, c=1,\ldots,C+1}} \sum_{c=1}^{C+1} \left\{ \sum_{i=1}^{p} \left[\sum_{t=\tau_{c-1}+1}^{\tau_c} \frac{1}{2} \left(Y_{it} - \sum_{j\neq i} \beta_{ij}^{(c)} Y_{jt} \right)^2 + \lambda_1 \sum_{j\neq i} \left| \beta_{ij}^{(c)} \right| \right] + \lambda_2 \right\}, \tag{7}$$

or

$$\min_{\substack{C,\tau_1,\ldots,\tau_C \\ \boldsymbol{B}^{(c)},c=1,\ldots,C+1}} \sum_{c=1}^{C+1}\sum_{i=1}^{p}\left(\frac{1}{2}\left\|\boldsymbol{Y}_i^{(c)}-\boldsymbol{Y}^{(c)}\boldsymbol{\beta}_i^{(c)}\right\|^2+\lambda_1\left\|\boldsymbol{\beta}_i^{(c)}\right\|_1\right)+\lambda_2(C+1), \tag{8}$$

where $\boldsymbol{Y}^{(c)}=\left(\boldsymbol{Y}_1^{(c)},\ldots,\boldsymbol{Y}_p^{(c)}\right)$ is the streaming data in the cth segment, $\boldsymbol{Y}_i^{(c)}=\left(Y_{i\tau_{c-1}+1},\ldots,Y_{i\tau_c}\right)'$, $\boldsymbol{\beta}_i^{(c)}=\left(\beta_{i1}^{(c)},\beta_{i2}^{(c)},\ldots,\beta_{ip}^{(c)}\right)'$, $\beta_{ii}^{(c)}=0$ are the representation coefficients for channel i in the cth segments, $\boldsymbol{B}^{(c)}=\left(\boldsymbol{\beta}_1^{(c)},\ldots,\boldsymbol{\beta}_p^{(c)}\right)$ and λ_2 is a penalty weight penalizing the number of segments to avoid overfitting.

The formulation of Equation 7 has several advantages compared with Equation 6. First of all, the number of parameters significantly decreases by directly setting constant representation coefficients in each segment, which greatly reduces the problem complexity. Secondly, the smoothness penalty term $\lambda_2\sum_{t=2}^{N}\sum_{j\neq i}|\beta_{ijt}-\beta_{ijt-1}|$ in Equation 6 tends to reduce the differences between two successive segments, while Equation 7 does not have such an issue. Thirdly and most importantly, the Equation 7 can be sequentially solved, e.g., without the need to restart the optimization process from the very beginning once a new observation arrives, which will be shown in detail in Section 4. This property is very desirable for online applications. We name this formulation along with the sequential optimization algorithm as dynamic sparse subspace learning (DSSL).

2.3 The Mathematical Equivalence between DSSL and MAP

In this subsection, we will interpret the formulation of DSSL from the Bayesian perspective by showing its equivalence to maximizing a posterior (MAP) of the change-points and representation coefficients. In the Bayesian formulation of a multiple change-point model, the priors for the number and locations of the change-points, and the parameters of each segment need to be specified. We assume that all the self-expressive residuals follow *i.i.d.* Gaussian distribution with zero mean and a known variance σ^2. Denote the multiple structural change-point model as $\mathcal{M}=\left(C,\left\{\delta^{(c)}\right\}_{c=1}^{C+1},\left\{\boldsymbol{B}^{(c)}\right\}_{c=1}^{C+1}\right)$ where $\delta^{(c)}=\tau_c-\tau_{c-1}$ is the duration of the cth segment. In the existing literature, the number and locations of the change-points are often jointly modeled by a Markov process, where the occurrence of the next change-point only depends on the location of the previous one or the duration of the current segment (Wen et al., 2017, 2019). For the discrete process where the change-points only occur at the observation indices, a Bernoulli process is commonly applied, or equivalently a geometric distribution is applied to the durations. Here we also assume a geometric distribution with parameter p_0 for the change-point locations

$$\pi\left(C,\left\{\delta^{(c)}\right\}_{c=1}^{C+1}\right)=p_0^C\left(1-p_0\right)^{N-C}.$$

For the regression coefficients $\boldsymbol{B}^{(c)}$, a double exponential or Laplace prior $La(0,\lambda)$ can be specified for each coefficient $\beta_{ij}^{(c)}$ independently to induce sparsity, i.e.,

$$\pi\left(\beta_{ij}^{(c)}\right)=\frac{1}{2\lambda}\exp\left(-\frac{\left|\beta_{ij}^{(c)}\right|}{\lambda}\right).$$

Then the posterior of $\mathcal{M}$ can be expressed as

$$f(\mathcal{M} \mid \mathbf{Y}) \propto f(\mathbf{Y} \mid \mathcal{M})\pi(\mathcal{M}), \tag{9}$$

where

$$\begin{aligned}\pi(\mathcal{M}) &= \pi\left(C, \left\{\delta^{(c)}\right\}_{c=1}^{C+1}\right)\pi\left(\left\{\boldsymbol{B}^{(c)}\right\}_{c=1}^{C+1} \mid C, \left\{\delta^{(c)}\right\}_{c=1}^{C+1}\right) \\ &= p_0^C\left(1-p_0\right)^{N-C}\prod_{c=1}^{C+1}\prod_{i=1}^{p}\prod_{j\neq i}^{p}\frac{1}{2\lambda}\exp\left(-\frac{\left|\beta_{ij}^{(c)}\right|}{\lambda}\right),\end{aligned}$$

and

$$f(\mathbf{Y} \mid \mathcal{M}) = \prod_{c=1}^{C+1}\prod_{i=1}^{p}\prod_{t=\tau_{c-1}+1}^{\tau_c}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\left(Y_{it}-\sum_{j\neq i}\beta_{ij}^{(c)}Y_{jt}\right)^2}{2\sigma^2}\right).$$

The negative log-posterior can thus be calculated as

$$\begin{aligned}-\log f(\mathcal{M} \mid \mathbf{Y}) = &\sum_{c=1}^{C+1}\sum_{i=1}^{p}\left[\sum_{t=\tau_{c-1}+1}^{\tau_c}\frac{1}{2\sigma^2}\left(Y_{it}-\sum_{j\neq i}\beta_{ij}^{(c)}Y_{jt}\right)^2+\frac{1}{\lambda}\sum_{j\neq i}\beta_{ij}^{(c)}\right] \\ &+C\left[-p(p-1)\log\frac{1}{2\lambda}-\log\frac{p_0}{1-p_0}\right]+C_0,\end{aligned} \tag{10}$$

where C_0 is a constant not related with $\mathcal{M}$. The MAP is equivalent to minimizing Equation 10. Comparing Equation 7 and Equation 10 we can clearly see that this two formulations are identical when

$$\lambda_1 = \frac{\sigma^2}{\lambda}, \lambda_2 = \sigma^2\left[-p(p-1)\log\frac{1}{2\lambda}-\log\frac{p_0}{1-p_0}\right]. \tag{11}$$

3. Asymptotic Properties

In this section, the asymptotic properties of the estimators of DSSL are established. Given a fixed but arbitrary number of change-points, we show that the positions of the change-points and the sparse regression coefficients within each segment converge to the true values as the number of samples increases. For simplicity yet without loss of generality, we only consider one sparse regression in the following analysis, e.g., $\boldsymbol{Y}_i$ as the response variable. The problem of including all sparse regressions can follow the same approach.

Let $\gamma_c^0 = \tau_c^0/N$ for $c = 1, \ldots, C$, $\boldsymbol{\gamma}^0 = \left(\gamma_1^0, \gamma_2^0, \ldots, \gamma_C^0\right)$ where the superscript 0 refers to the true value. Similarly, define $\gamma_c = \tau_c/N$, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_C)$, $\mathcal{B}^0 = \left(\boldsymbol{B}^{(1)^0}, \ldots, \boldsymbol{B}^{(C+1)^0}\right)$ and $\mathcal{B} = \left(\boldsymbol{B}^{(1)}, \ldots, \boldsymbol{B}^{(C+1)}\right)$. Note that the number of the change-points is fixed and $\boldsymbol{\gamma}^0$ is set to be a constant vector as N goes to infinity. Now we detail the assumptions of the proposed method, under which its asymptotic properties can be better established.

Assumption A3. *It is assumed that for* $c = 1, 2, \ldots, C, i = 1, 2, \ldots, p$, $E_{Y_{it}}\left[f\left(Y_{it} \mid \boldsymbol{Y}_{(-i)t}, \boldsymbol{B} = \boldsymbol{B}^{(c+1)^0}\right)\right] \neq E_{Y_{it}}\left[f\left(Y_{it} \mid \boldsymbol{Y}_{(-i)t}, \boldsymbol{B} = \boldsymbol{B}^{(c)^0}\right)\right]$ *on a set of non-zero measure, where* $f\left(Y_{it} \mid \boldsymbol{Y}_{(-i)t}, \boldsymbol{B} = \boldsymbol{B}^{(c)^0}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(Y_{it} - \sum_{j \neq i} \beta_{ij}^{(c)^0} Y_{jt}\right)^2}{2\sigma^2}\right)$.

This assumption guarantees that the expectation of the distributions in two successive segments are different, which is consistent with the definition that change-points are the time points when the correlation structure changes.

Assumption A4. *It is assumed that:*
(1). For $c = 1, 2, \ldots, C+1$, $\boldsymbol{B}^{(c)}$ *and* $\boldsymbol{B}^{(c)^0}$ *are contained in* $\mathbb{B}_c$, *where* $\mathbb{B}_c$ *is a compact subset of* $\mathcal{R}^{p \times p}$.
(2). For $t = 1, 2, \ldots, N$, $\boldsymbol{y}_t = (Y_{1t}, \ldots, Y_{pt})$ *is contained in* $\mathbb{Y}$, *where* $\mathbb{Y}$ *is a compact subset of* $\mathcal{R}^p$. *Besides, there exists a finite value for* $E\left[\boldsymbol{y}_t^T \boldsymbol{y}_t\right]$ *and* $\|\boldsymbol{y}_t\|_\infty$ *has an upper bound* $\|\boldsymbol{y}_t\|_\infty \leq M$.

This assumption limits the sample space of $\boldsymbol{B}^{(c)}$ and $\boldsymbol{y}_t$, which is suitable for most applications.

Assumption A5. *It is assumed that:*
For any $c = 1, 2, \ldots, C+1$, $i = 1, 2, \ldots, p$, *and any integers* m, n *satisfying* $0 \leq m \leq n \leq N$,

$$E\left\{\max_{\boldsymbol{B}^{(c)} \in \mathbb{B}_c}\left(\sum_{t=m+1}^{n}\left\{\log f\left(Y_{it} \mid \boldsymbol{Y}_{(-i)t}, \boldsymbol{B}^{(c)}\right) - E_{Y_{it}}\left[\log f\left(Y_{it} \mid \boldsymbol{Y}_{(-i)t}, \boldsymbol{B}^{(c)}\right)\right]\right\}\right)^2\right\}$$

$\leq C_0(n-m)^r$, *where* $r < 2$ *and* C_0 *is a constant.*

The Assumption A5 is a technical requirement on the behavior of the log-likelihood function between or within the segment. This condition is relatively weak and it is easy to check that it is satisfied by at least all distributions in the exponential family (He and Severini, 2010).

To prove the consistency, we first provide two lemmas.

Lemma 1.
Under (A1-A5), there exist two positive constants $C_1 > 0$ *and* $C_2 > 0$ *such that, for any* $\boldsymbol{\gamma}$ *and* $\mathcal{B}$,

$$\lim_{N \to \infty} J_1 \leq -\max\left\{C_1\left\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^0\right\|_\infty, C_2\rho\left(\mathcal{B}, \mathcal{B}^0\right)\right\},$$

where $\left\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^0\right\|_\infty = \max_c\left|\gamma_c - \gamma_c^0\right|$, $\rho\left(\mathcal{B}, \mathcal{B}^0\right) = \max_c\left|v\left(\boldsymbol{B}^{(c)}, \boldsymbol{B}^{(c)^0}\right)\right|$,

$$v\left(\boldsymbol{B}^{(c)}, \boldsymbol{B}^{(c')^0}\right) = E_{Y_{(-i)t}}\left(\int\left[\log f\left(Y_{it} \mid \boldsymbol{Y}_{(-i)t}, \boldsymbol{B}^{(c)}\right) - \log f\left(Y_{it} \mid \boldsymbol{Y}_{(-i)t}, \boldsymbol{B}^{(c')^0}\right)\right] f\left(Y_{it} \mid \boldsymbol{Y}_{(-i)t}, \boldsymbol{B}^{(c')^0}\right) dY_{it}\right).$$

and J_1 is defined in Appendix 3.

The proof of Lemma 1 is given in Appendix 1.

Lemma 2.
Under (A1-A5), for any $c = 1, 2, \ldots, C+1$, any $0 \leq m_1 \leq m_2 \leq N$ and any positive number $\varepsilon > 0$, there exist a constant A, independent of ε, and a constant $r < 2$, such that

$$\Pr\left(\max_{\substack{m_1 \leq s < t \leq m_2, \\ \boldsymbol{B}^{(\mathbf{c})} \in \mathbb{B}_c}} \left| \sum_{i=s+1}^{t} \left\{ \log f\left(Y_{it} \mid \boldsymbol{Y}_{(-i)t}, \boldsymbol{B}^{(c)}\right) - E_{Y_{it}}\left[\log f\left(Y_{it} \mid \boldsymbol{Y}_{(-i)t}, \boldsymbol{B}^{(c)}\right)\right]\right\} \right| > \varepsilon \right) \leq A\frac{(m_2 - m_1)^r}{\varepsilon^2}.$$

The proof of Lemma 2 is given in Appendix 2.

Theorem 1 (Consistency).
Under (A1-A5) and Equation 7, we have $\hat{\gamma}_i \xrightarrow{p} \gamma_i^0, \widehat{\boldsymbol{B}}^{(c)} \xrightarrow{p} \boldsymbol{B}^{(c)^0}$ as $N \to \infty$, that is, $\hat{\gamma}_i - \gamma_i^0 = o_p(1), \widehat{\boldsymbol{B}}^{(c)} - \boldsymbol{B}^{(c)^0} = o_p(1)$, where $\hat{\gamma}_i = \hat{\tau}_i / N$ for $i = 1, 2, \ldots, C$ and $c = 1, 2, \ldots, C+1$.

This property tells us that if the segments are sufficiently long, the change-points and representation coefficients can be estimated accurately. The proof of Theorem 1 is given in Appendix 3.

4. Online Optimization via PELT Algorithm

In this section, we first introduce the PELT algorithm, an efficient sequential optimization algorithm, that can be applied to solve our optimization problem 7. Then, we discuss how to determine the penalty coefficients and proper parameters for the PELT algorithm.

4.1 The PELT Algorithm

Various algorithms have been proposed to solve the optimization problem of the following form for multiple change-points models:

$$\min_{m, \tau_1, \ldots, \tau_m} \sum_{c=1}^{m+1} \left\{ \text{Cost}\left(\mathbf{Y}_{\tau_{c-1}+1:\tau_c}\right)\right\} + f(m), \tag{12}$$

where $\text{Cost}(\cdot)$ is a cost function for a segment, m is the number of change-points and $f(m)$, e.g., $f(m) = \beta(m+1)$, is a penalty term to guard against overfitting. Binary Segmentation (BS) algorithm proposed by Scott and Knott (1974) is one of the most established search method with an $\mathcal{O}(n \log n)$ computational cost for n samples. It begins by initially applying the single change-point method to the entire data set. The data set is split into two segments and the single change-point detection method is carried out again for these two segments independently. This procedure is repeated until no further change-points are detected. The

advantage of the BS method is that it is computationally efficient. But it does not guarantee to find the global optimal solution.

The optimal partitioning (OP) algorithm proposed by Jackson et al. (2005) focuses on finding the latest change-point (LCP) at each time step. It relates the optimal value of the cost function to the cost for the optimal partition of the data prior to the latest change-point plus the cost for the segment from the latest change-point to the end of the data. Let $F(n)$ denote the optimal value of the objective function for data $\boldsymbol{Y}_{1:n}$ and $\boldsymbol{\tau}_n = \{\boldsymbol{\tau} : 0 = \tau_0 < \tau_1 < \cdots < \tau_m < \tau_{m+1} = n\}$ be the set of all possible vectors of the change-points for this dataset. Set $f(m) = \beta(m+1)$ and $F(0) = -\beta$. Then,

$$\begin{aligned} F(n) &= \min_{\boldsymbol{\tau} \in \boldsymbol{\tau}_n} \left\{ \sum_{c=1}^{m+1} \left[\text{Cost}\left(\boldsymbol{Y}_{\tau_{c-1}+1:\tau_c} \right) + \beta \right] \right\} \\ &= \min_{t \in \{0,\ldots,n-1\}} \left\{ \min_{\boldsymbol{\tau} \in \boldsymbol{\tau}_t} \sum_{c=1}^{m} \left[\text{Cost}\left(\boldsymbol{Y}_{\tau_{c-1}+1:\tau_c} \right) + \beta \right] + \text{Cost}\left(\boldsymbol{Y}_{t+1:n} \right) + \beta \right\} \\ &= \min_{t \in \{0,\ldots,n-1\}} \left\{ F(t) + \text{Cost}\left(\boldsymbol{Y}_{t+1:n} \right) + \beta \right\}. \end{aligned} \tag{13}$$

As $F(t)$ only needs to be calculated once and can be used repeatedly in the following steps, this recursion can be solved sequentially or dynamically for $n = 1, \ldots, N$, and the computational cost is $\mathcal{O}\left(N^2\right)$. Although the OP algorithm can find the global optimal solution, it is still far from being computationally competitive with the BS method. To this end, Killick et al. (2012) proposed the PELT (Pruned Exact Linear Time) by adding a pruning step for the OP algorithm to reduce the computational cost. In the OP algorithm, to solve $F(n) = \min_{t \in \{0,\ldots,n-1\}} \{F(t) + \text{Cost}\left(\boldsymbol{Y}_{t+1:n}\right) + \beta\}$, we need to consider all time points prior to time n. But in the PELT algorithm, we remove the time points that can never be the optimal LCP. Specifically, we optimize $F(n) = \min_{t \in R(t)} \{F(t) + \text{Cost}\left(\boldsymbol{Y}_{t+1:n}\right) + \beta\}$ where $R(n)$ is the set of all time points that could be the possible LCP in terms of optimality at time n. The following theorem (Killick et al., 2012) provides a simple condition under which such pruning can be performed.

Theorem 2. *We assume there exists a constant K such that for all $s < n < T$,*

$$\text{Cost}\left(\boldsymbol{Y}_{s+1:n}\right) + \text{Cost}\left(\boldsymbol{Y}_{n+1:T}\right) + K \leq \text{Cost}\left(\boldsymbol{Y}_{s+1:T}\right). \tag{14}$$

Then if

$$F(s) + \text{Cost}\left(\boldsymbol{Y}_{s+1:n}\right) + K \geq F(n) \tag{15}$$

holds, at a future time $T > n$, s can never be the optimal last change-point prior to T.

The proof can be found in Section 5 of Supplementary Material in (Killick et al., 2012). This result states that if Equation 15 holds, then for any $T > n$, the best segmentation with the LCP prior to T being at n will be better than any with the LCP at s. Note that there exists a proper constant K satisfying Equation 14 for almost all cost functions used in practice. For example, if the cost function is the negative log-likelihood, then the constant can be selected as $K = 0$. Therefore, we have $R(n+1) = \{\tau \mid \tau \in R(n) \cup \{n\}, F(\tau) + \text{Cost}\left(\boldsymbol{Y}_{\tau+1:n}\right) + K < F(n)\}$. This pruning process makes the

Algorithm 1 The PELT based sequential optimization algorithm for DSSL

Input: Data set $\boldsymbol{Y}$, cost function Cost, penalty constant λ_2, constant K
Initialize: N: the length of data, $F(0) = -\lambda_2,\ cp(0) = NULL,\ R(1) = \{0\}$
Iterate for $n = 1, \ldots, N$

1. Calculate $F(n) = \min_{\tau \in R(n)} \{F(\tau) + \text{Cost}\left(\boldsymbol{Y}_{\tau+1:n}\right) + \lambda_2\}$.
2. Let $\hat{\tau} = \underset{\tau \in R(n)}{\text{argmin}} \{F(\tau) + \text{Cost}\left(\boldsymbol{Y}_{\tau+1:n}\right) + \lambda_2\}$ and $\widehat{\boldsymbol{B}}_n = \underset{B}{\text{argmin}} \{\text{Cost}\left(\boldsymbol{Y}_{\hat{\tau}+1:n}\right)\}$.
3. Set $cp(n) = \{cp(\hat{\tau}), \hat{\tau}\}$ and $\mathcal{B}(n) = \left(\mathcal{B}(\hat{t}), \widehat{\boldsymbol{B}}_n\right)$.
4. Set $R(n+1) = \{\tau \mid \tau \in R(n) \cup \{n\}, F(\tau) + \text{Cost}\left(\boldsymbol{Y}_{\tau+1:n}\right) + K < F(n)\}$.

End

optimization process very efficient under mild conditions with approximately linear computational cost with n, or on average a constant computational cost at each time step, which is highly desirable for online change-point detection.

In our optimization problem 7, the cost function and the penalty term can be expressed as

$$\begin{gathered}\text{Cost}\left(\boldsymbol{Y}_{\tau_{c-1}+1:\tau_c}\right) = \sum_{i=1}^{p}\left[\sum_{t=\tau_{c-1}+1}^{\tau_c}\frac{1}{2}\left(Y_{it} - \sum_{j\neq i}\widehat{\beta_{ij}^{(c)}}Y_{jt}\right)^2 + \lambda_1\sum_{j\neq i}\left|\widehat{\beta_{ij}^{(c)}}\right|\right], \\ f(C) = \lambda_2(C+1),\end{gathered} \tag{16}$$

where $\widehat{\beta_{ij}^{(c)}}$ is the estimated coefficients via LASSO algorithm for the following optimization problem

$$\min_{\beta_{ij}^{(c)}, j\neq i}\left\{\sum_{t=\tau_{c-1}+1}^{\tau_c}\frac{1}{2}\left(Y_{it} - \sum_{j\neq i}\beta_{ij}^{(c)}Y_{jt}\right)^2 + \lambda_1\sum_{j\neq i}\left|\beta_{ij}^{(c)}\right|\right\}. \tag{17}$$

The detailed PELT based algorithm to solve our optimization problem 7 is shown in Algorithm 1.

4.2 Parameter Selection

The parameter K is very critical in the pruning of the LCPs to achieve an efficient computation. Based on Equation 14 and 15, we can find that the least upper bound (LUB) of K satisfying Equation 14 is the optimal value for K. However, it is not realistic to obtain the LUB in most cases. Here we propose a more conservative upper bound UB such that $K \leq UB \leq LUB$ for the selection of K, which is give in Lemma 3.

Lemma 3.
For all $s < n < T$, define the least upper bound LUB as

$$LUB = \min_{s<n<T}\left\{\text{Cost}\left(\boldsymbol{Y}_{s+1:T}\right) - \text{Cost}\left(\boldsymbol{Y}_{s+1:n}\right) - \text{Cost}\left(\boldsymbol{Y}_{n+1:T}\right)\right\}.$$

Define $UB_{|\beta|}$ as the upper bound of the absolute value of all the regressive coefficients within any time interval, i.e.,

$$UB_{|\beta|} = \max\left\{\left|\hat{\beta}_{ij}\right|, i, j = 1, \ldots, p\right\}.$$

Then

$$LUB \geq -2p(p-1)\lambda_1 UB_{|\beta|}.$$

The proof is provided in Appendix 4. Lemma 3 tells us that we can select $K = -2p(p-1)\lambda_1 UB_{|\beta|}$. In practice, $UB_{|\beta|}$ can be roughly obtained based on empirical knowledge or historical data. Besides, as observed in our case studies, the detection is not very sensitive to this parameter. Note that this selection is very conservative. In most cases, $\text{Cost}\left(\boldsymbol{Y}_{s+1:T}\right) - \text{Cost}\left(\boldsymbol{Y}_{s+1:n}\right) - \text{Cost}\left(\boldsymbol{Y}_{n+1:T}\right) \geq 0$ and we can simply set $K = 0$.

The selection of optimal penalty weights λ_1 and λ_2 is nontrivial yet important for online change-points detection. In subsection 2.3, we have shown that the parameters λ_1 and λ_2 can be derived from hyperparameters in the Bayesian framework. However, these hyperparameters are also unknown and are still not easy to obtain. In this subsection, a new tuning method combining searching grid and approximate Minimum Description Length (AMDL, Saito, 1994) criterion are proposed for the PELT algorithm.

Cross validation methods and information criterion methods are two widely used methods in model selection or parameter selection. An advantage of information criterion methods is that they have considerably less computational expense than CV methods (Kirkland et al., 2015). Classical criterions, such as AIC and BIC criteria, have been applied for many fields and have proven their efficiency and accuracy. However, they require that the degrees of freedom and error variance must be known or well estimated, which is difficult in high-dimensional data with $p \gg n$. Modified criteria are then proposed without the estimation of error variance. Specifically, Saito introduced AMDL as an information-theoretic based criterion with

$$\text{AMDL} = N\log(\overline{err}) + 3D\log(N),$$

where $\overline{err}$ is the average residual and D is the dimension of the model.

Here we propose the following criterion for our model which is similar to the AMDL criterion:

$$\text{AMDL}(\lambda) = (Np)\log(\overline{err}) + 3\widehat{df}(\lambda)\log(Np),$$

where $D = \widehat{df}(\lambda)$ is estimated as the number of nonzero elements in regression coefficients. The first term favors complex models, while the second term is a penalty term balancing the bias-variance tradeoff.

Two-dimensional searching grid method is used to determine λ_1 and λ_2 using some historical data. We first find the minimum value of λ_1 that results in each estimated variable to be 0, and denote this value as $\lambda_1^{\max}$. By fixing λ_1, $\lambda_2^{\max}(\lambda_1)$ can be found as the minimum value that results in zero change-point as detection result. The optimal values of λ_1 and λ_2 are selected by searching grid in $\{(0, \lambda_1^{\max}), (0, \lambda_2^{\max}(\lambda_1))\}$ for the value that minimizes the proposed criterion. In the case studies, the detection results show that the proposed method is efficient and effective for the change-point detection algorithm.

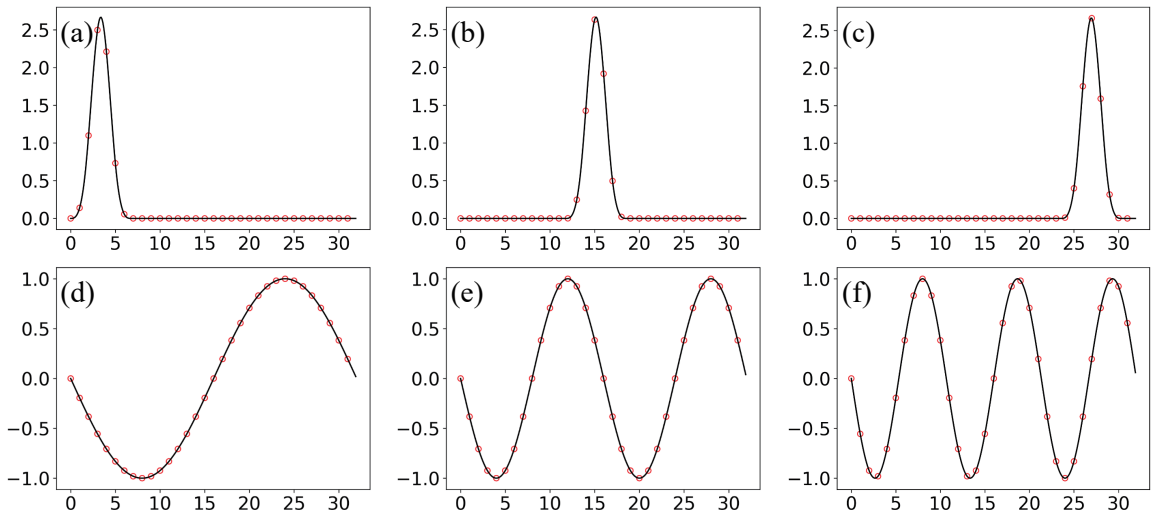

Figure 2: Basis functions of the two subspaces. (a-c): B-spline basis functions; (d-f): Fourier basis functions.

5. Case Studies

In this section, the proposed algorithm is evaluated through numerical experiments with synthetic and real gesture data. The results show that the algorithm has high detection accuracy, low detection delay and high analysis speed.

5.1 Simulation Study

In this subsection, we generate synthetic data from the assumed subspace model described in Section 2.1 to evaluate the proposed method. The length of the time series is set to $N = 128$ and suppose there are two change-points at $\tau_1 = N/4$ and $\tau_2 = N/2$. The dimension is set to $p = 40$. Two subspaces are simulated, where the first half of these time series are generated by linear representation of three B-spline basis functions, while the other half is generated by Fourier basis functions. These basis functions are shown in Figure 2.

The coefficients are randomly generated in [-0.5,0.5] and they change at the change-points. The noise variance is set to $\sigma^2 = 0.0025$.

To apply our detection algorithm, we set the tuning parameters based on the discussion in Section 4.2. The parameters are set as $\lambda_1 = 0.01, \lambda_2 = 2.0$, and $K = 0$. Figure 3 shows the 40 simulated time series in one run and the fitted values using the proposed method. Clearly, due to sufficient samples in each subspace, the self-expression is very accurate in all three segments.

Figure 4 shows the estimated coefficients at the final time step for the first three time series of each subspace for illustration, i.e., $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \widehat{\boldsymbol{\beta}}_3, \widehat{\boldsymbol{\beta}}_{21}, \widehat{\boldsymbol{\beta}}_{22}$ and $\widehat{\boldsymbol{\beta}}_{23}$. Clearly, only the time series within the same subspace have nonzero coefficients, and each representation is sparse. These coefficients are constant within each segment.

Figure 5 shows the selected candidates for the LCP at each time step using $K = 0$. Only a very small number of points including the true LCP are selected as the candidates, which can effectively control the computational cost at each step.

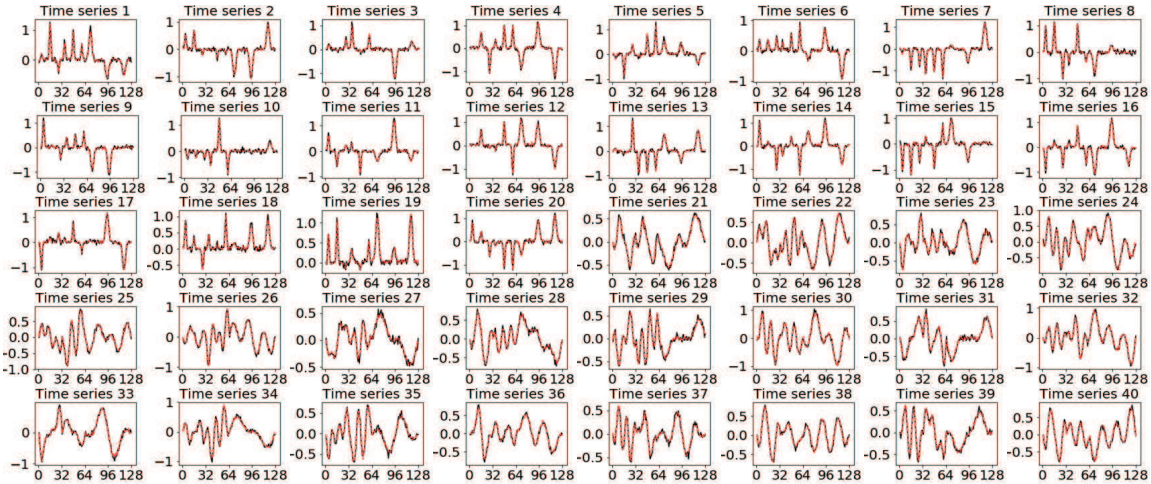

Figure 3: The original and fitted traces of 40 simulated time series in one run. The black solid lines are the raw curves, and the red dashed lines are estimated values using the proposed method.

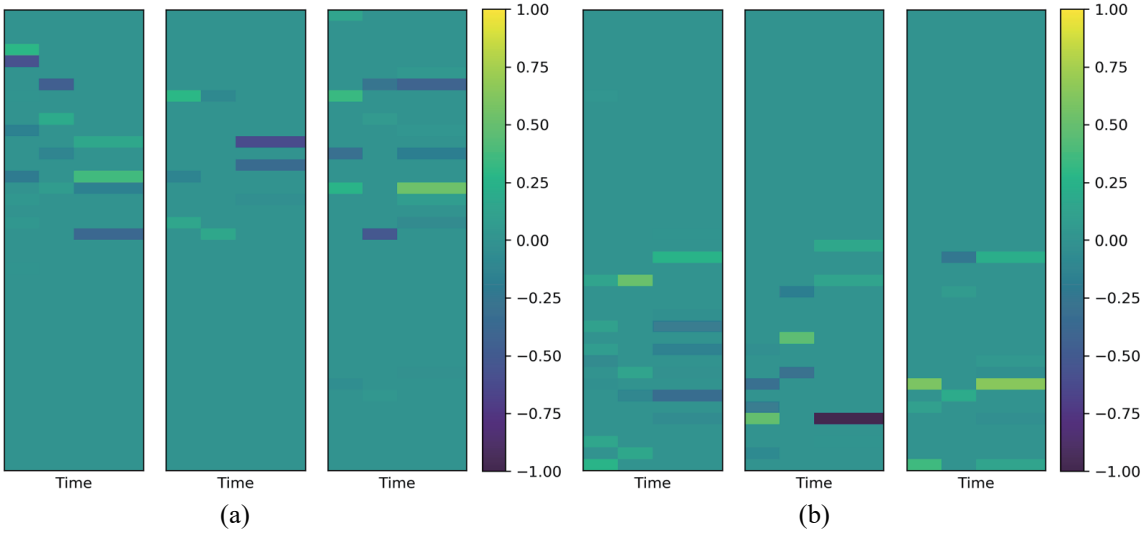

Figure 4: The estimated coefficients for the first three time series of each subspace as response variables: (a) subspace I and (b) subspace II.

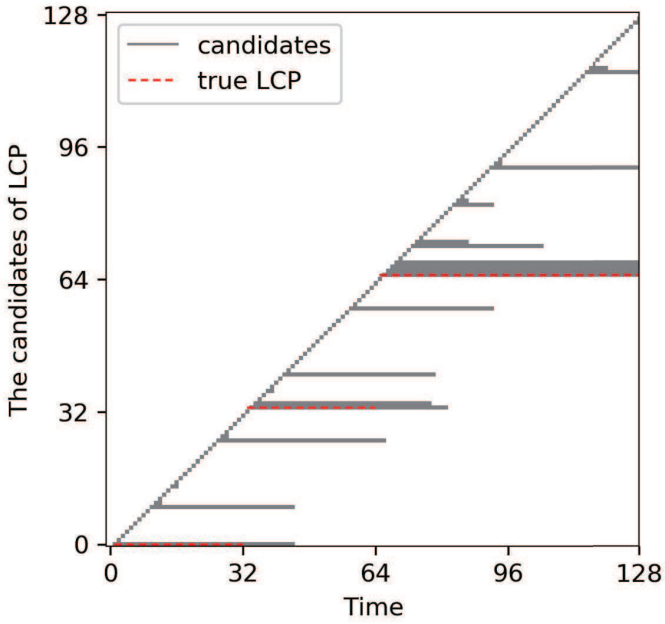

Figure 5: The selected candidates for the LCP at each time step.

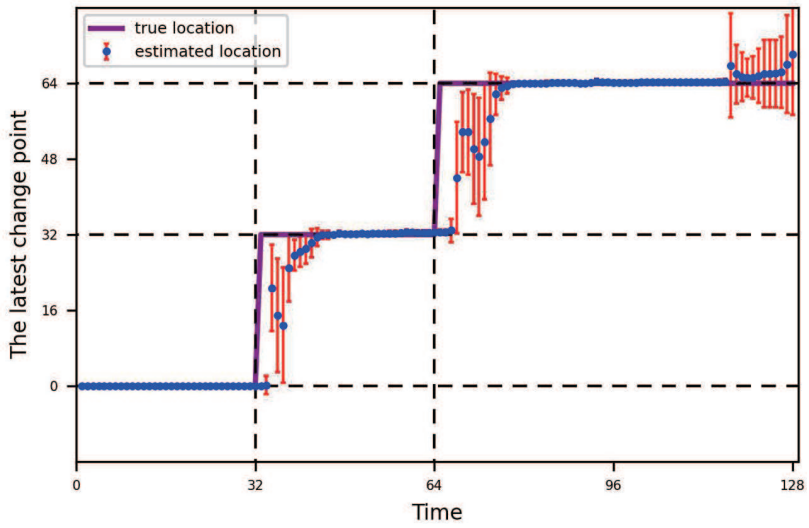

Figure 6: The online detected latest change-point.

To evaluate the detection timeliness and accuracy of the detected LCP, we repeat the experiment 100 times. Figure 6 shows the mean and confidence interval of the detected LCP at each time step. It can be clearly seen that in most of the cases, the change can be timely detected with only about 10 time steps after the change occurs. Besides, as the observations since the latest change accumulates, the detected location of the LCP becomes more and more accurate. Note that in each replication, the detected LCP often abruptly jumps to the true values shortly after the change occurs. As the detection delay may vary from run to run, the change of the mean detection is not as abrupt as a single run.

5.2 Human Motion Tracking

In this subsection, we apply the proposed DSSL to the MSRC-12 Gesture Dataset for gesture tracking (Fothergill et al., 2012). This dataset consists of sequences of human skeletal body part movements and the associated gesture that needs to be recognized by the system. Each sample of the sequences contains 60 variables, which are the three dimensional coordinates of 20 human joints. The body pose is captured at a sample rate of 30Hz with $\pm$2cm accuracy in joint positions. In the MSRC-12 Gesture Dataset, there are 30 subjects and they perform 12 gestures each for ten times. The position of these 20 joints and some snapshots of shoot gesture and throw gesture are shown in Figure 7.

Motion segmentation is often a very critical step for gesture recognition. Here we combine the sequences of two gestures of the same subject together to demonstrate the effectiveness of the proposed method. Specifically, we choose shoot?and throw?as the two gestures. In the first segment, the subject stretches his arms out in front of him, holds a pistol in both hands, makes a recoil movement and then returns to the original position. In the second segment, the subject uses his right arm to make an overarm throwing movement and then returns to the original position. For the sake of simplicity, we convert coordinates of the same joint into one distance variable, which represents the distance of the joint from the reference point.

Clearly, some joints share similar trajectories with each other, as these joints move in similar ways, such as the joints on the same arm or the same leg (Figure 8(a)). Some joints have totally different trajectories because they have no correlations. With this regard, we can infer that these joints lie in some subspaces and they can be naturally clustered into different groups. Besides, some joints share similar trajectories in the first segment but have different trajectories in the second segment, such as the joints of the two hands in the combined data. As shown in Figure 8(b), in the first segment, they increase or decrease synchronously, while in the second segment, they increase or decrease in the opposite direction. Therefore, we can apply the proposed method to detect when the gesture changes.We delete the data of the left and right wrist because these data are generally the same with the data of the left and right hand. So in total $p = 18$ variables are considered. The tuning parameters are set to $\lambda_1 = 0.0004,\ \lambda_2 = 0.27$ and $K = 0$.

Figure 9 shows the original and fitted variables and the sequentially detected LCPs. The change of the gesture is timely detected only after about 10 time steps. The positions of the joints of head, trunk, legs and feet do not change much in the whole motion sequence. The joints of left arm and right arm play a key role to detect the change-point. The subspaces

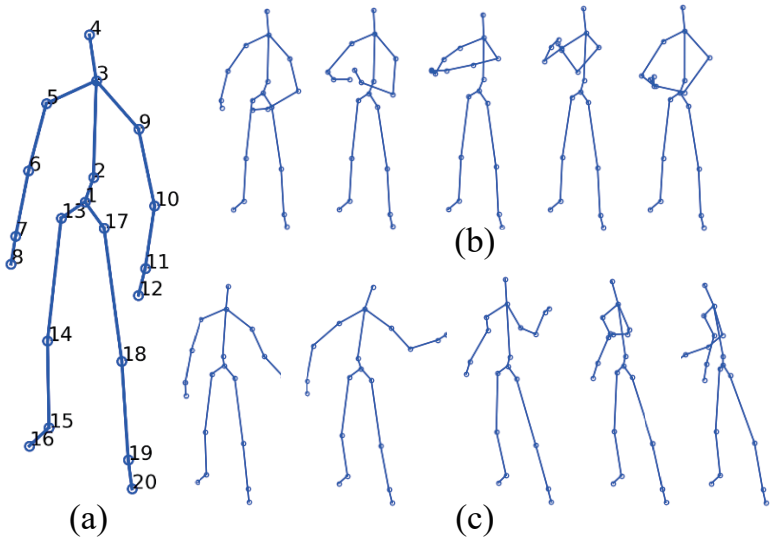

Figure 7: (a) The position of the 20 joints; (b) five snapshots of the shoot gesture; (c) five snapshots of the throw gesture.

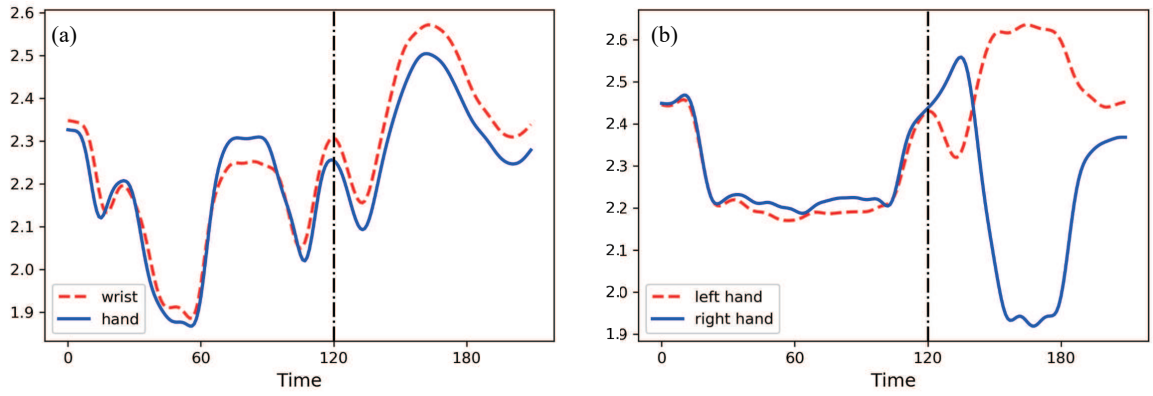

Figure 8: The representative correlations between the joints: (a) the trace of two joints of the left arm and (b) the trace of the joints of the left and right hand.

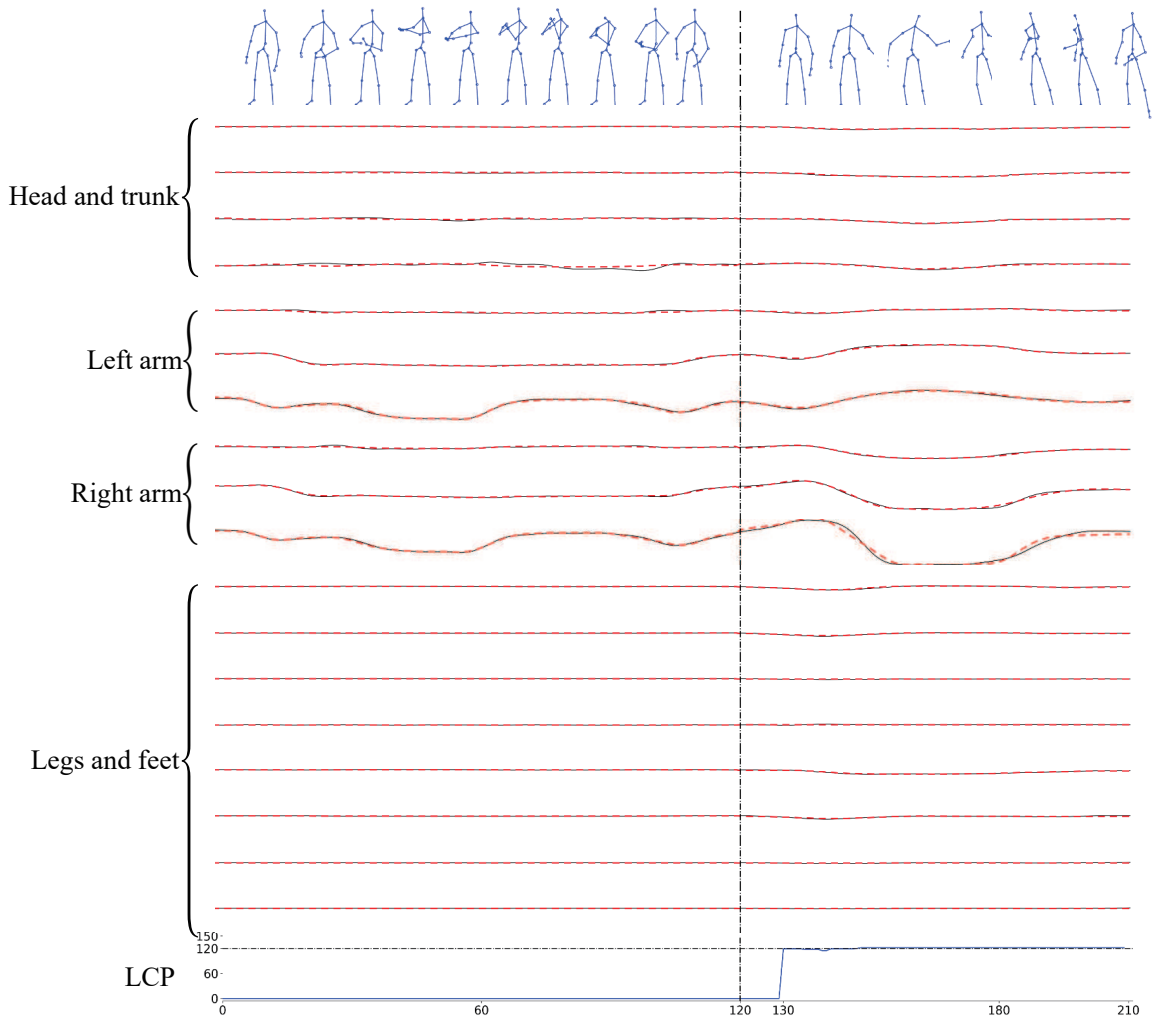

Figure 9: The original and fitted traces of the 18 variables and the sequentially detected LCP. The black solid lines are the original curves, and the red dashed lines are the estimated ones. The vertical dashed line denote the true change-point.

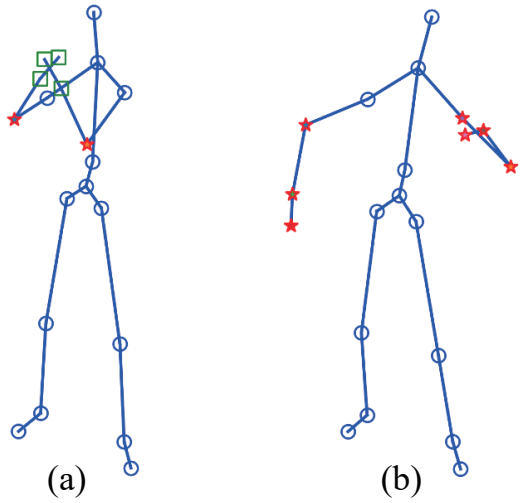

Figure 10: The clustering of the joints: (a) the first segment and (b) the second segment. The joints of the same color and shape are grouped into one cluster.

for each gesture are further identified via spectral clustering, which are shown in Figure 10. This result is consistent with the actual movement of each joint in these two gestures.

6. Discussion and Conclusions

In this paper, we proposed a dynamic sparse subspace learning (DSSL) approach for online structural change-point detection of high-dimensional streaming data. Specifically, it is assumed that the high-dimensional data lie in multiple low-dimensional subspaces and the subspace structures may abruptly change over time. Only the variables from the same subspace correlate with each other and each variables can be sparsely represented by others. Based on the self-expressive property, we proposed a novel multiple structural change-point model with two penalty terms in the loss function to encourage sparse representation and avoid excessive change-points. The model formulation was then shown to be equivalent to maximizing a posterior under a Bayesian framework. The consistency of the estimators was further established, which shows that with the number of change-points fixed or known, the positions of the change-points and the representation coefficients within each segment converge to true values as the length of the streaming data increases. A PELT based algorithm was proposed for online optimization and change-point detection. The parameter K of PELT algorithm could be conservatively estimated by $K = -2p(p-1)\lambda_1 UB_{|\beta|}$ and we can simply set $K = 0$ without loss of accuracy in most cases. Based on some historical data, the penalty coefficients in our model were selected by a tuning method combining searching grid and AMDL criterion. The effectiveness of the proposed method was demonstrated on synthetic data and gesture data for motion tracking.

There are several issues that are worthy of further investigation. Firstly, the proposed method assumes that the measurement noises of all variables are independent and identically distributed. However, the variance heterogeneity, cross-correlation and even auto-

correlation may exist in practice. In addition, all the subspaces are assumed to be linear manifolds in the current work. To make it more flexible, nonlinear manifolds can be considered by using nonlinear regressions, such as kernel based methods. Last but not least, when there are not sufficient samples in the subspace, the self-expressive assumption may not hold. How to learn the subspace sequentially with insufficient samples is a challenging problem that needs to be solved.

Acknowledgments

The authors would like to thank the editor, associate editor, and anonymous reviewers for many constructive comments which greatly improved the paper. This work was partially supported by National Natural Science Foundation of China grant NSFC-51875003 and key program NSFC-71932006.

Appendix 1. The proof of Lemma 1

Proof. Let us define

$$g_i\left(\alpha, \mathcal{B}^0\right) = \sup_{1\le j\le C+1} \sup_{B} \left\{\alpha v\left(\boldsymbol{B}^{(j)}, \boldsymbol{B}^{(i+1)^0}\right) + (1-\alpha) v\left(Y, \boldsymbol{B}^{(j)}, \boldsymbol{B}^{(i)^0}\right)\right\},$$

where $0 \le \alpha \le 1$.

$$\begin{aligned} v\left(\boldsymbol{B}^{(c)}, \boldsymbol{B}^{(c')^0}\right) =& E_{\boldsymbol{Y}_{(-i)t}}\left(\int \left[\log f\left(Y_{it} \mid \boldsymbol{Y}_{(-i)t}, \boldsymbol{B}^{(c)}\right) - \log f\left(Y_{it} \mid \boldsymbol{Y}_{(-i)t}, \boldsymbol{B}^{(c')^0}\right)\right]\right. \\ & \left. f\left(Y_{it} \mid \boldsymbol{Y}_{(-i)t}, \boldsymbol{B}^{(c')^0}\right) dY_{it}\right). \end{aligned}$$

Therefore, $v\left(\boldsymbol{B}^{(c)}, \boldsymbol{B}^{(c)}\right) = 0$.
We then have that $g_i\left(0, \mathcal{B}^0\right) = g_i\left(1, \mathcal{B}^0\right) = 0$ for $i = 1, 2, \ldots, C$.
Obviously, $g_i\left(\alpha, \mathcal{B}^0\right)$ is a convex function with respect to α for any i.
Let $\mathrm{G_i}\left(\mathcal{B}^0\right) = 2g_i\left(1/2, \mathcal{B}^0\right)$. Because $\alpha = 2\alpha(1/2) + (1-2\alpha)0$, convexity of $g_i\left(\alpha, \boldsymbol{B}^0\right)$ gives that

$$g_i\left(\alpha, \mathcal{B}^0\right) \le 2\alpha g_i\left(1/2, \mathcal{B}^0\right) = \alpha \mathrm{G_i}\left(\mathcal{B}^0\right), \forall i.$$

Noting that

$$g_i\left(1/2, \mathcal{B}^0\right) = \frac{1}{2} \sup_{1\le j\le C+1} \sup_{\boldsymbol{B}^{(j)}\in\mathbb{B}_j} \left\{v\left(\boldsymbol{B}^{(j)}, \boldsymbol{B}^{(i+1)^0}\right) + v\left(\boldsymbol{B}^{(j)}, \boldsymbol{B}^{(i)^0}\right)\right\},$$

It follows from Assumption 1 that $\mathrm{G}_i\left(\mathcal{B}^0\right) < 0$. If we let $\bar{\mathrm{G}}\left(\mathcal{B}^0\right) = \max_{1\le i\le C} \mathrm{G}_i\left(\mathcal{B}^0\right)$, then $\bar{\mathrm{G}}\left(\mathcal{B}^0\right) < 0$.
Let $\Delta_\gamma^0 = \min_{1\le c\le C-1}\left|\gamma_{c+1}^0 - \gamma_c^0\right|$. Consider a change-point fraction configuration $\boldsymbol{\gamma}$ such that $\left\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^0\right\|_\infty \le \Delta_\gamma^0/4$. For any j, there are two cases: a candidate change-point fraction γ_j may be on the left or the right of the true change-point fraction γ_j^0.
For any j with γ_j on the right of γ_j^0, we have that $\gamma_{j-1} \le \gamma_j^0 \le \gamma_j$. then

$$\lim_{N\to\infty} J_1 \le \frac{n_{j,j+1}}{N} v\left(\boldsymbol{B}^{(j)}, \boldsymbol{B}^{(j+1)^0}\right) + \frac{n_{j,j}}{N} v\left(\boldsymbol{B}^{(j)}, \boldsymbol{B}^{(j)^0}\right),$$

where $n_{j,i}$ refers to the number of observations in the set $[\tau_{j-1}+1, \tau_j] \cap \left[\tau_{i-1}^0 + 1, \tau_i^0\right]$.
If we define $\alpha_{j,j+1} = \frac{n_{j,j+1}}{n_{j,j+1}+n_{j,j}}$, then the case $\left\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^0\right\|_\infty \le \Delta_\gamma^0/4$ gives that $\alpha_{j,j+1} \le \frac{1}{2}$ and

$$\begin{aligned} &\lim_{N\to\infty} J_1 \le \tfrac{n_{j,j+1}+n_{j,j}}{N} \left\{\alpha_{j,j+1} v\left(\boldsymbol{B}^{(j)}, \boldsymbol{B}^{(j+1)^0}\right) + (1-\alpha_{j,j+1}) v\left(\boldsymbol{B}^{(j)}, \boldsymbol{B}^{(j)^0}\right)\right\} \\ &\le \tfrac{n_{j,j+1}}{N} \mathrm{G}_j\left(\mathcal{B}^0\right) \le \left(\gamma_j - \gamma_j^0\right) \overline{\mathrm{G}}\left(\mathcal{B}^0\right). \end{aligned}$$

For any j with γ_j on the left of γ_j^0, we have that $\gamma_j \le \gamma_j^0 \le \gamma_{j+1}$. Similarly, we have

$$\lim_{N\to\infty} J_1 \le \left(\gamma_j^0 - \gamma_j\right) \overline{\mathrm{G}}\left(\mathcal{B}^0\right).$$

Therefore, if $\left\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^0\right\|_\infty \leq \Delta_\gamma^0/4$, then we have $\lim_{N\to\infty} J_1 \leq \left|\gamma_j^0-\gamma_j\right|_\infty \overline{\mathrm{G}}\left(\mathcal{B}^0\right)$. On the other hand,

$$\lim_{N\to\infty} J_1 \leq \min_{1\leq j\leq C+1} \frac{n_{j,j}}{N} v\left(\boldsymbol{B}^{(j)}, \boldsymbol{B}^{(j)^0}\right) = -\max_{1\leq j\leq C+1} \frac{n_{j,j}}{N}\left|v\left(\boldsymbol{B}^{(j)}, \boldsymbol{B}^{(j)^0}\right)\right|.$$

We have $\frac{n_{j,j}}{N} \geq \frac{\Delta_\gamma^0}{2}$ for any j, so

$$\lim_{N\to\infty} J_1 \leq -\frac{\Delta_\gamma^0}{2} \sup_{1\leq j\leq C+1}\left|v\left(\boldsymbol{B}^{(j)}, \boldsymbol{B}^{(j)^0}\right)\right| = -\frac{\Delta_\gamma^0}{2}\rho\left(\mathcal{B}, \mathcal{B}^0\right).$$

Now, consider the other case of a change-point fraction configuration $\boldsymbol{\gamma}$, where $\left\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^0\right\|_\infty > \Delta_\gamma^0/4$. It is clear that there exists a pair of integers (i,j) such that $n_{ij}/N \geq \Delta_\gamma^0/4,\ n_{i,j+1}/N \geq \Delta_\gamma^0/4$.

$$\begin{aligned}
\lim_{N\to\infty} J_1 &\leq \frac{n_{i,j+1}}{N} v\left(\boldsymbol{B}^{(i)}, \boldsymbol{B}^{(j+1)^0}\right) + \frac{n_{i,j}}{N} v\left(\boldsymbol{B}^{(i)}, \boldsymbol{B}^{(j)^0}\right) \\
&\leq \frac{n_{i,j+1}+n_{i,j}}{N}\left\{\alpha_{i,j+1} v\left(\boldsymbol{B}^{(i)}, \boldsymbol{B}^{(j+1)^0}\right) + \left(1-\alpha_{i,j+1}\right) v\left(\boldsymbol{B}^{(i)}, \boldsymbol{B}^{(j+1)^0}\right)\right\} \\
&\leq \frac{n_{i,j+1}+n_{i,j}}{N} \min\left(\alpha_{i,j+1}, 1-\alpha_{i,j+1}\right) \overline{\mathrm{G}}\left(\mathcal{B}^0\right) \\
&\leq \frac{\Delta_\gamma^0}{2} \min\left(\frac{n_{i,j+1}}{N}, \frac{n_{i,j}}{N}\right) \overline{\mathrm{G}}\left(\mathcal{B}^0\right) \\
&\leq \frac{1}{2}\left(\frac{\Delta_\gamma^0}{2}\right)^2 \overline{\mathrm{G}}\left(\mathcal{B}^0\right).
\end{aligned}$$

Combining the results from the two cases of $\left\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^0\right\|_\infty \leq \Delta_\gamma^0/4$ and $\left\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^0\right\|_\infty > \Delta_\gamma^0/4$, it follows that

$$\lim_{N\to\infty} J_1 \leq \overline{\mathrm{G}}\left(\mathcal{B}^0\right) \min\left(\frac{1}{2}\left(\frac{\Delta_\gamma^0}{2}\right)^2, \left|\gamma_j^0-\gamma_j\right|_\infty\right) \leq \frac{1}{2}\left(\frac{\Delta_\gamma^0}{2}\right)^2 \overline{\mathrm{G}}\left(\mathcal{B}^0\right)\left|\gamma_j^0-\gamma_j\right|_\infty$$

and

$$\lim_{n\to\infty} J_1 \leq \frac{\Delta_\gamma^0}{2} \max\left[-\rho\left(\mathcal{B}, \mathcal{B}^0\right), \frac{\Delta_\gamma^0}{4}\overline{\mathrm{G}}\left(\mathcal{B}^0\right)\right] \leq -\frac{\Delta_\gamma^0}{2} \min\left[\rho\left(\mathcal{B}, \mathcal{B}^0\right), -\frac{\Delta_\gamma^0}{4}\overline{\mathrm{G}}\left(\mathcal{B}^0\right)\right].$$

$\rho\left(\mathcal{B}, \mathcal{B}^0\right)$ have upper bound $\varrho\left(\mathcal{B}, \mathcal{B}^0\right) = \max_i \max_{1\leq j\leq C+1} \sup_{\boldsymbol{B}^{(\mathrm{j})}\in\mathbb{B}_{\boldsymbol{j}}}\left|v\left(\boldsymbol{B}^{(j)}, \boldsymbol{B}^{(i)^0}\right)\right|$. We have

$$\frac{\rho\left(\mathcal{B}, \mathcal{B}^0\right)}{\varrho\left(\mathcal{B}, \mathcal{B}^0\right)} \leq 1.$$

Therefore,

$$\lim_{N\to\infty} J_1 \leq -\frac{\Delta_\gamma^0}{2}\varrho\left(\mathcal{B}, \mathcal{B}^0\right) \min\left[\frac{\rho\left(\mathcal{B}, \mathcal{B}^0\right)}{\mathrm{Q}\left(\mathcal{B}, \mathcal{B}^0\right)}, -\frac{\frac{\Delta_\gamma^0}{4}\overline{\mathrm{G}}\left(\mathcal{B}^0\right)}{\varrho\left(\mathcal{B}, \mathcal{B}^0\right)}\right].$$

If $-\frac{\Delta_\gamma^0}{4}\mathrm{G}\left(\mathcal{B}^0\right)/\varrho\left(\mathcal{B},\mathcal{B}^0\right)\leq 1$, then we have

$$\lim_{N\to\infty} J_1 \leq \left(\frac{\Delta_\gamma^0}{2}\right)^2 \left(\frac{\rho\left(\mathcal{B},\mathcal{B}^0\right)}{\varrho\left(\mathcal{B},\mathcal{B}^0\right)}\right)\left(\frac{\overline{\mathrm{G}}\left(\mathcal{B}^0\right)}{2}\right).$$

If $-\frac{\Delta_\gamma^0}{4}\mathrm{G}\left(\mathcal{B}^0\right)/\varrho\left(\mathcal{B},\mathcal{B}^0\right) > 1$,, then $\lim_{N\to\infty} J_1 \leq -\frac{\Delta_\gamma^0}{2}\rho\left(\mathcal{B},\mathcal{B}^0\right)$.
Let $C_2 = \min\left\{-\left(\frac{\Delta_\gamma^0}{2}\right)^2\left(\frac{\overline{\mathrm{G}}(\mathcal{B}^0)}{2}\right), \frac{\Delta_\gamma^0}{2}\right\}$. We have

$$\lim_{N\to\infty} J_1 \leq -C_2\rho\left(\mathcal{B},\mathcal{B}^0\right).$$

Setting $C_1 = \frac{1}{2}\left(\frac{\Delta_\gamma^0}{2}\right)^2\overline{\mathrm{G}}\left(\mathcal{B}^0\right)$, we finally have that

$$\lim_{N\to\infty} J_1 \leq -\max\left\{C_1\left\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^0\right\|_\infty, C_2\rho\left(\mathcal{B},\mathcal{B}^0\right)\right\},$$

which concludes the proof.

■

Appendix 2. The proof of Lemma 2

Proof. Under A5, for $0\leq s\leq t\leq N$, we have

$$\begin{aligned}
&E\left\{\max_{\boldsymbol{B}^{(c)}\in\mathbb{B}_c}\left(\sum_{t=m+1}^{n}\left\{\log f\left(Y_{it}\mid \boldsymbol{Y}_{(-i)t},\boldsymbol{B}^{(c)}\right)-E_{Y_{it}}\left[\log f\left(Y_{it}\mid \boldsymbol{Y}_{(-i)t},\boldsymbol{B}^{(c)}\right)\right]\right\}\right)^2\right\}\\
&\leq C_0(n-m)^r.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&E\left\{\max_{\substack{m_1\leq s<t\leq m_2,\\ \boldsymbol{\beta}^{(c)}\in\mathbf{B_c}}}\left(\sum_{i=s+1}^{t}\left\{\log f\left(Y_{it}\mid \boldsymbol{Y}_{(-i)t},\boldsymbol{B}^{(c)}\right)-E_{Y_{it}}\left[\log f\left(Y_{it}\mid \boldsymbol{Y}_{(-i)t},\boldsymbol{B}^{(c)}\right)\right]\right\}\right)^2\right\}\\
&\leq C_0\left(m_2-m_1\right)^r.
\end{aligned}$$

Based on Markov inequality,

$$\begin{aligned}
&\Pr\left(\max_{\substack{m_1\leq s<t\leq m_2,\\ \boldsymbol{\beta}^{(c)}\in\mathbf{B_c}}}\left|\sum_{i=s+1}^{t}\left\{\log f\left(Y_{it}\mid \boldsymbol{Y}_{(-i)t},\boldsymbol{B}^{(c)}\right)-E_{Y_{it}}\left[\log f\left(Y_{it}\mid \boldsymbol{Y}_{(-i)t},\boldsymbol{B}^{(c)}\right)\right]\right\}\right|>\varepsilon\right)\\
&=\Pr\left(\max_{\substack{m_1\leq s<t\leq m_2,\\ \boldsymbol{\beta}^{(c)}\in\mathbf{B_c}}}\left(\sum_{i=s+1}^{t}\left\{\log f\left(Y_{it}\mid \boldsymbol{Y}_{(-i)t},\boldsymbol{B}^{(c)}\right)-E_{Y_{it}}\left[\log f\left(Y_{it}\mid \boldsymbol{Y}_{(-i)t},\boldsymbol{B}^{(c)}\right)\right]\right\}\right)^2\right.\\
&\left.>\varepsilon^2\right)\\
&\leq C_0\frac{(m_2-m_1)^r}{\varepsilon^2}.
\end{aligned}$$

■

Appendix 3. The proof of Theorem 1

Proof. Our optimization problem

$$\min_{\substack{C,\tau_1,\ldots,\tau_C \\ \boldsymbol{B}^{(c)},c=1,\ldots,C+1}} \sum_{c=1}^{C+1} \left\{ \sum_{t=\tau_{c-1}+1}^{\tau_c} \frac{1}{2} \left(Y_{it} - \sum_{j \neq i} \beta_{ij}^{(c)} Y_{jt} \right)^2 + \lambda_1 \sum_{j \neq i} \left| \beta_{ij}^{(c)} \right| + \lambda_2 \right\}$$

could be reformulated as

$$\min_{c,\tau_1,\ldots,\tau_C} \sum_{C=1}^{C+1} \left\{ \sum_{t=\tau_{c-1}+1}^{\tau_c} \log f\left(Y_{it} \mid \boldsymbol{B}^{(c)}\right) + \log f_{\boldsymbol{\beta}}\left(\boldsymbol{B}^{(c)}\right) + \lambda_3 \right\},$$

where $f_{\boldsymbol{\beta}}\left(\boldsymbol{B}^{(c)}\right) = \frac{1}{(2\lambda)^p} \exp\left(-\frac{\sum_{j \neq i} \left|\beta_{ij}^{(c)}\right|}{\lambda} \right)$ and λ_3 is a constant.
When the number of the change-points is known, λ_3 can be ignored.
Let $l = \sum_{c=1}^{C+1} \left\{ \sum_{t=\tau_{c-1}+1}^{\tau_c} \log f\left(Y_{it} \mid \boldsymbol{B}^{(c)}\right) + \log f_{\boldsymbol{\beta}}\left(\boldsymbol{B}^{(c)}\right) \right\}$.
Define a function J by $J = J_1 + J_2 + J_3$, where

$$\begin{aligned}
J_1 = \sum_{c=1}^{C+1} \sum_{c'=1}^{C+1} \sum_{i \in [\tau_{c-1}+1, \tau_c] \cap \left[\tau_{c'-1}^0+1, \tau_{c'}^0\right]} \frac{1}{N} \Big\{ \int \Big[& \log f\left(Y_{it} \mid \boldsymbol{B}^{(c)}\right) \\
& - \log f\left(Y_{it} \mid \boldsymbol{B}^{(c')^0}\right) \Big] f\left(Y_{it} \mid \boldsymbol{B}^{(c')^0}\right) dY_{it} \Big\}, \\
J_2 = \frac{1}{N} \sum_{c=1}^{C+1} \sum_{t=\tau_{c-1}+1}^{\tau_c} & \left\{ \log f\left(Y_{it} \mid \boldsymbol{B}^{(c)}\right) - E\left[\log f\left(Y_{it} \mid \boldsymbol{B}^{(c)}\right) \right] \right\} \\
- \frac{1}{N} \sum_{c'=1}^{c+1} \sum_{t=\tau_{c'-1}+1}^{\tau_{c'}} & \left\{ \log f\left(Y_{it} \mid \boldsymbol{B}^{(c')^0}\right) - E\left[\log f\left(Y_{it} \mid \boldsymbol{B}^{(c')^0}\right) \right] \right\}, \\
J_3 = \frac{1}{N} \sum_{c=1}^{C+1} & \left(\log f_{\boldsymbol{\beta}}\left(\boldsymbol{B}^{(c)}\right) - \log f_{\boldsymbol{\beta}}\left(\boldsymbol{B}^{(c)^0}\right) \right).
\end{aligned}$$

We obviously have that

$$\operatorname*{argmin}_{\substack{C,\tau_1,\ldots,\tau_C \\ \boldsymbol{B}^{(c)},c=1,\ldots,C+1}} l = \operatorname*{argmin}_{\substack{C,\tau_1,\ldots,\tau_C \\ \boldsymbol{B}^{(c)},c=1,\ldots,C+1}} J.$$

Denote that

$$\Lambda = \{(\gamma_1, \ldots, \gamma_C)\}, \ \Lambda_\delta = \left\{ \boldsymbol{\gamma} \in \Lambda : \left\| \boldsymbol{\gamma} - \boldsymbol{\gamma}^0 \right\|_\infty > \delta \right\},$$

$$\mathbb{B} = \mathbb{B}_1 \times \ldots \times \mathbb{B}_{C+1}, \ \mathbb{B}_\delta = \left\{ \mathcal{B} \in \mathbb{B} : \rho\left(\mathcal{B}, \mathcal{B}^0\right) > \delta \right\}.$$

Then, for any $\delta > 0$, it follows from Lemma 1 that

$$-\max_{\boldsymbol{\gamma} a \in \Lambda_\delta, \mathcal{B} \in \mathbb{B}} \lim_{N \to \infty} J_1 \geq C_1 \delta, \ -\max_{\boldsymbol{\gamma} \in \Lambda, \mathcal{B} \in \mathbb{B}_\delta} \lim_{N \to \infty} J_1 \geq C_2 \delta.$$

Therefore, we obtain that, when $N \to \infty$,

$$
\begin{aligned}
&\lim_{N\to\infty} \Pr\left(\left\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0\right\|_\infty > \delta\right) \\
&\leq \lim_{N\to\infty} \Pr\left(\max_{\boldsymbol{\gamma}\in\Lambda_\delta, \mathcal{B}\in\mathbb{B}} J > 0\right) = \Pr\left(\lim_{N\to\infty} \max_{\boldsymbol{\gamma}\in\Lambda_\delta, \mathcal{B}\in\mathbb{B}} J > 0\right) \\
&\leq \Pr\left(\max_{\boldsymbol{\gamma}\in\Lambda_\delta, \mathcal{B}\in\mathbb{B}} \lim_{N\to\infty} (J_2 + J_3) > -\max_{\boldsymbol{\gamma}\in\Lambda_\delta, \mathcal{B}\in\mathbb{B}} \lim_{N\to\infty} J_1\right) \\
&\leq \Pr\left(\max_{\boldsymbol{\gamma}\in\Lambda_\delta, \mathcal{B}\in\mathbb{B}} \lim_{N\to\infty} (J_2 + J_3) > \mathrm{C}_1\delta\right) \leq \Pr\left(\max_{\boldsymbol{\gamma}\in\Lambda_\delta, \mathcal{B}\in\mathbb{B}} \lim_{N\to\infty} (|J_2| + |J_3|) > \mathrm{C}_1\delta\right) \\
&\leq \Pr\left(\max_{\boldsymbol{\gamma}\in\Lambda_\delta, \mathcal{B}\in\mathbb{B}} \left\{\lim_{N\to\infty} \frac{1}{N} \sum_{c=1}^{C+1} \left| \sum_{t=\tau_{c-1}+1}^{\tau_c} \left\{\log f\left(Y_{it} \mid \boldsymbol{B}^{(c)}\right) - E\left[\log f\left(Y_{it} \mid \boldsymbol{B}^{(c)}\right)\right]\right\}\right|\right\} \right. \\
&\qquad \left. > \frac{C_1}{3}\delta\right) \\
&+ \Pr\left(\lim_{N\to\infty} \frac{1}{N} \sum_{c=1}^{C+1} \left| \sum_{t=\tau_{c-1}^0+1}^{\tau_c^0} \left\{\log f\left(Y_{it} \mid \boldsymbol{B}^{(c)^0}\right) - E\left[\log f\left(Y_{it} \mid \boldsymbol{B}^{(c)^0}\right)\right]\right\}\right| > \frac{C_1}{3}\delta\right) \\
&+ \Pr\left(\max_{\boldsymbol{\gamma}\in\Lambda_\delta, \mathcal{B}\in\mathbb{B}} \left\{\lim_{N\to\infty} \frac{1}{N} \sum_{c=1}^{C+1} \left|\log f_{\boldsymbol{\beta}}\left(\boldsymbol{B}^{(c)}\right) - \log f_{\boldsymbol{\beta}}\left(\boldsymbol{B}^{(c)^0}\right)\right|\right\} > \frac{C_1}{3}\delta\right).
\end{aligned}
$$

The first part

$$
\begin{aligned}
&\Pr\left(\max_{\boldsymbol{\gamma}\in\Lambda_\delta, \mathcal{B}\in\mathbb{B}} \left\{\lim_{N\to\infty} \frac{1}{N} \sum_{c=1}^{C+1} \left| \sum_{t=\tau_{c-1}+1}^{\tau_c} \left\{\log f\left(Y_{it} \mid \boldsymbol{B}^{(c)}\right) - E\left[\log f\left(Y_{it} \mid \boldsymbol{B}^{(c)}\right)\right]\right\}\right|\right\} \right. \\
&\qquad \left. > \frac{C_1}{3}\delta\right) \\
&= \lim_{N\to\infty} \Pr\left(\max_{\boldsymbol{\gamma}\in\Lambda_\delta, \mathcal{B}\in\mathbb{B}} \left\{\frac{1}{N} \sum_{c=1}^{c+1} \left| \sum_{t=\tau_{c-1}+1}^{\tau_c} \left\{\log f\left(Y_{it} \mid \boldsymbol{B}^{(c)}\right) - E\left[\log f\left(Y_{it} \mid \boldsymbol{B}^{(c)}\right)\right]\right\}\right|\right\} \right. \\
&\qquad \left. > \frac{C_1}{3}\delta\right) \\
&\leq \lim_{N\to\infty} \left[\frac{3(C+1)}{C_1\delta}\right]^2 \left(\sum_{j=1}^{c+1} A_j\right) N^{r-2} \to 0.
\end{aligned}
$$

Similarly,

$$
\Pr\left(\lim_{N\to\infty} \frac{1}{N} \sum_{c=1}^{C+1} \left| \sum_{t=\tau_{c-1}^0+1}^{\tau_c^0} \left\{\log f\left(Y_{it} \mid \boldsymbol{B}^{(c)^0}\right) - E\left[\log f\left(Y_{it} \mid \boldsymbol{B}^{(c)^0}\right)\right]\right\}\right| > \frac{C_1}{3}\delta\right) \to 0.
$$

Since

$$
\lim_{N\to\infty} \frac{1}{N} \sum_{c=1}^{C+1} \left|\log f_{\boldsymbol{\beta}}\left(\boldsymbol{B}^{(c)}\right) - \log f_{\boldsymbol{\beta}}\left(\boldsymbol{B}^{(c)^0}\right)\right| \leq \lim_{N\to\infty} \frac{C+1}{N} \times 2 \max_{\boldsymbol{B}^{(c)}\in\mathbb{B}_c} \left|\log f_{\boldsymbol{\beta}}\left(\boldsymbol{B}^{(c)}\right)\right| \to 0,
$$

we have

$$\Pr\left(\max_{\boldsymbol{\gamma}\in\Lambda_\delta,\mathcal{B}\in\mathbb{B}}\left\{\lim_{N\to\infty}\frac{1}{N}\sum_{c=1}^{C+1}\left|\log f_{\boldsymbol{\beta}}\left(\boldsymbol{B}^{(c)}\right)-\log f_{\boldsymbol{\beta}}\left(\boldsymbol{B}^{(c)^0}\right)\right|\right\}>\frac{C_1}{3}\delta\right)=0.$$

Therefore,

$$\lim_{N\to\infty}\Pr\left(\left\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^0\right\|_\infty>\delta\right)\to 0,\ \text{as } N\to\infty.$$

Under A3, $v\left(\boldsymbol{B}^{(c)},\boldsymbol{B}^{(c)^0}\right)=0$ iff $\boldsymbol{B}^{(c)}=\boldsymbol{B}^{(c)^0}$.
It follows that $\hat{\boldsymbol{\gamma}}\to_p\boldsymbol{\gamma}_0,\ \widehat{\mathcal{B}}\to_p\mathcal{B}^{\mathbf{0}}$ as $N\to\infty$.

■

Appendix 4. The proof of Lemma 3

Proof.
K must satisfy

$$\begin{aligned}&\sum_{i=1}^{p}\left\{\sum_{t=s+1}^{n}\frac{1}{2}\left(Y_{it}-\sum_{j\neq i}\hat{\beta}_{ij}^{(1)}Y_{jt}\right)^2+\lambda_1\sum_{j\neq i}\left|\hat{\beta}_{ij}^{(1)}\right|\right\}\\&+\sum_{i=1}^{p}\left\{\sum_{t=n+1}^{T}\frac{1}{2}\left(Y_{it}-\sum_{j\neq i}\hat{\beta}_{ij}^{(2)}Y_{jt}\right)^2+\lambda_1\sum_{j\neq i}\left|\hat{\beta}_{ij}^{(2)}\right|\right\}\\&+K\leq\sum_{i=1}^{p}\left\{\sum_{t=s+1}^{T}\frac{1}{2}\left(Y_{it}-\sum_{j\neq i}\hat{\beta}_{ij}^{(0)}Y_{jt}\right)^2+\lambda_1\sum_{j\neq i}\left|\hat{\beta}_{ij}^{(0)}\right|\right\}.\end{aligned}$$

for all case.
Therefore,

$$\begin{aligned}K\leq\sum_{i=1}^{p}\Bigg\{&\lambda_1\sum_{j\neq i}\left|\widehat{\beta}_{ij}^{(0)}\right|-\lambda_1\sum_{j\neq i}\left|\widehat{\beta}_{ij}^{(1)}\right|-\lambda_1\sum_{j\neq i}\left|\widehat{\beta}_{ij}^{(2)}\right|+\sum_{t=s+1}^{T}\frac{1}{2}\left(Y_{it}-\sum_{j\neq i}\widehat{\beta}_{ij}^{(0)}Y_{jt}\right)^2\\&-\sum_{t=s+1}^{n}\frac{1}{2}\left(Y_{it}-\sum_{j\neq i}\widehat{\beta}_{ij}^{(1)}Y_{jt}\right)^2-\sum_{t=n+1}^{T}\frac{1}{2}\left(Y_{it}-\sum_{j\neq i}\widehat{\beta}_{ij}^{(2)}Y_{jt}\right)\Bigg\}\end{aligned}$$

Denote the least upper bound

$$\begin{aligned}LUB=\min\sum_{i=1}^{p}\Bigg\{&\lambda_1\sum_{j\neq i}\left|\widehat{\beta}_{ij}^{(0)}\right|-\lambda_1\sum_{j\neq i}\left|\widehat{\beta}_{ij}^{(1)}\right|-\lambda_1\sum_{j\neq i}\left|\widehat{\beta}_{ij}^{(2)}\right|+\sum_{t=s+1}^{T}\frac{1}{2}\left(Y_{it}-\sum_{j\neq i}\widehat{\beta}_{ij}^{(0)}Y_{jt}\right)^2\\&-\sum_{t=s+1}^{n}\frac{1}{2}\left(Y_{it}-\sum_{j\neq i}\widehat{\beta}_{ij}^{(1)}Y_{jt}\right)^2-\sum_{t=n+1}^{T}\frac{1}{2}\left(Y_{it}-\sum_{j\neq i}\widehat{\beta}_{ij}^{(2)}Y_{jt}\right)\Bigg\}\end{aligned}$$

Since $\sum_{i=1}^{p}\left\{\lambda_1\sum_{j\neq i}\left|\widehat{\beta}_{ij}^{(0)}\right|\right\}\geq 0$ and

$$\sum_{t=s+1}^{T}\frac{1}{2}\left(Y_{it}-\sum_{j\neq i}\hat{\beta}_{ij}^{(0)}Y_{jt}\right)^2-\sum_{t=s+1}^{n}\frac{1}{2}\left(Y_{it}-\sum_{j\neq i}\hat{\beta}_{ij}^{(1)}Y_{jt}\right)^2$$
$$-\sum_{t=n+1}^{T}\frac{1}{2}\left(Y_{it}-\sum_{j\neq i}\hat{\beta}_{ij}^{(2)}Y_{jt}\right)^2\geq 0,$$

$\min\sum_{i=1}^{p}\left\{-\lambda_1\sum_{j\neq i}\left|\widehat{\beta}_{ij}^{(1)}\right|-\lambda_1\sum_{j\neq i}\left|\widehat{\beta}_{ij}^{(2)}\right|\right\}\leq LUB.$
We can also get

$$\min\sum_{i=1}^{p}\left\{-\lambda_1\sum_{j\neq i}\left|\widehat{\beta}_{ij}^{(1)}\right|-\lambda_1\sum_{j\neq i}\left|\hat{\beta}_{ij}^{(2)}\right|\right\}\geq -2p(p-1)\lambda_1\max\left\{\left|\hat{\beta}_{ij}\right|,i,j=1,\ldots,p\right\}.$$

Therefore,

$$-2p(p-1)\lambda_1\max\left\{\left|\hat{\beta}_{ij}\right|,i,j=1,\ldots,p\right\}\leq LUB.$$

$K\leq UB^*=-2p(p-1)\lambda_1\max\left\{\left|\hat{\beta}_{ij}\right|,i,j=1,\ldots,p\right\}$ satisfying Equation 14.

■

References

Samaneh Aminikhanghahi and Diane J Cook. A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51(2):339–367, 2017.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Richard Bellman and Rand Corporation. *Dynamic programming*. Princeton University Press, Princeton,, 1957.

A. P. Dempster. Covariance selection. *Biometrics*, 28(1):157, 1972.

Mathias Drton and Michael D Perlman. A sinful approach to gaussian graphical model selection. *Journal of Statistical Planning and Inference*, 138(4):1179–1200, 2008.

Ehsan Elhamifar and Rene Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11): 2765–2781, 2013.

Simon Fothergill, Helena M Mentis, Pushmeet Kohli, and Sebastian Nowozin. Instructing people for training gestural interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems(CHI)*, pages 1737–1746, 2012.

Rina Foygel and Mathias Drton. Extended bayesian information criteria for gaussian graphical models. *Advances in Neural Information Processing Systems*, pages 604–612, 2010.

Jerome H Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Alexander J Gibberd and James D B Nelson. Regularized estimation of piecewise constant gaussian graphical models: The group-fused graphical lasso. *Journal of Computational and Graphical Statistics*, 26(3):623–634, 2017.

Yi Guo, Junbin Gao, and Feng Li. Spatial subspace clustering for hyperspectral data segmentation. In *International Conference on Digital Information Processing and Communications(ICDIPC)*, pages 180–190, 2013.

Jonas M. B. Haslbeck and Lourens J. Waldorp. mgm: Estimating time-varying mixed graphical models in high-dimensional data. *Journal of Statal Software*, 93(8), 2020.

Heping He and Thomas A Severini. Asymptotic properties of maximum likelihood estimators in models with multiple change points. *Bernoulli*, 16(3):759–779, 2010.

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

Yuchen Jiao, Yanxi Chen, and Yuantao Gu. Subspace change-point detection: A new model and solution. *IEEE Journal of Selected Topics in Signal Processing*, 12(6):1224–1239, 2018.

Rebecca Killick, Paul Fearnhead, and Idris A Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107 (500):1590–1598, 2012.

Lisa Kirkland, Frans Kanfer, and Sollie Millard. *LASSO Tuning Parameter Selection*. 2015.

Mladen Kolar and Eric P Xing. On time varying undirected graphs. In *International Conference on Artificial Intelligence and Statistics(AISTATS)*, volume 15, pages 407–415.

Mladen Kolar and Eric P. Xing. Estimating networks with jumps. *Electronic Journal of Statistics*, 6:2069–2106, 2012.

Chunguang Li and Rene Vidal. Structured sparse subspace clustering: A unified optimization framework. In *Computer Vision and Pattern Recognition(CVPR)*, pages 277–286.

Chunguang Li, Chong You, and Rene Vidal. Structured sparse subspace clustering: A joint affinity learning and subspace clustering framework. *IEEE Transactions on Image Processing*, 26(6):2988–3001, 2017.

Nicolai Meinshausen and Peter Buhlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

J. L. Nowinski. Applications of functional analysis in engineering. *International Journal of Engineering Science*, 19(11):1377–1390, 1981.

Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: a review. *Sigkdd Explorations*, 6(1):90–105, 2004.

Huitong Qiu, Fang Han, Han Liu, and Brian S Caffo. Joint estimation of multiple graphical models from high dimensional time series. *Journal of The Royal Statistical Society Series B-statistical Methodology*, 78(2):487–504, 2016.

Adam J Rothman, Peter J Bickel, Elizaveta Levina, and J Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

Naoki Saito. Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum description length criterion. *Wavelet Analysis and Its Applications*, 4:299–324, 1994.

A J Scott and M Knott. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, 30(3):507, 1974.

Robert Tibshirani, Michael A Saunders, Saharon Rosset, J Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of The Royal Statistical Society Series B-statistical Methodology*, 67(1):91–108, 2005.

Stephen Tierney, Junbin Gao, and Yi Guo. Subspace clustering for sequential data. In *Computer Vision and Pattern Recognition(CVPR)*, pages 1019–1026, 2014.

Yuxin Wen, Jianguo Wu, and Yuan Yuan. Multiple-phase modeling of degradation signal for condition monitoring and remaining useful life prediction. *IEEE Transactions on Reliability*, 66(3):924–938, 2017.

Yuxin Wen, Jianguo Wu, Qiang Zhou, and Tzuliang Tseng. Multiple-change-point modeling and exact bayesian inference of degradation signal for prognostic improvement. *IEEE Transactions on Automation Science and Engineering*, 16(2):613–628, 2019.

Jianguo Wu, Yong Chen, and Shiyu Zhou. Online detection of steady-state operation using a multiple-change-point model and exact bayesian inference. *IIE Transactions*, 48(7): 599–613, 2016.

Jianguo Wu, Honglun Xu, Chen Zhang, and Yuan Yuan. A sequential bayesian partitioning approach for online steady-state detection of multivariate systems. *IEEE Transactions on Automation Science and Engineering*, 16(4):1882–1895, 2019.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society Series B-statistical Methodology*, 68(1):49–67, 2006.

Chen Zhang, Hao Yan, Seungho Lee, and Jianjun Shi. Dynamic multivariate functional data modeling via sparse subspace learning. *Technometrics*, pages 1–14, 2020.

Junjian Zhang, Chunguang Li, Honggang Zhang, and Jun Guo. Low-rank and structured sparse subspace clustering. In *Visual Communications and Image Processing(VCIP)*, pages 1–4.

Zhao, Ren, Tingni, Sun, Cun-Hui, Zhang, Harrison, H., and Zhou. Asymptotic normality and optimalities in estimation of large gaussian graphical models. *Annals of Statistics*, 2015.

Shuheng Zhou, John Lafferty, and Larry Wasserman. Time varying undirected graphs. *Machine Learning*, 80(2):295–319, 2010.

References

Samaneh Aminikhanghahi and Diane J Cook. A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51(2):339–367, 2017.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Richard Bellman and Rand Corporation. *Dynamic programming*. Princeton University Press, Princeton,, 1957.

A. P. Dempster. Covariance selection. *Biometrics*, 28(1):157, 1972.

Mathias Drton and Michael D Perlman. A sinful approach to gaussian graphical model selection. *Journal of Statistical Planning and Inference*, 138(4):1179–1200, 2008.

Ehsan Elhamifar and Rene Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11): 2765–2781, 2013.

Simon Fothergill, Helena M Mentis, Pushmeet Kohli, and Sebastian Nowozin. Instructing people for training gestural interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems(CHI)*, pages 1737–1746, 2012.

Rina Foygel and Mathias Drton. Extended bayesian information criteria for gaussian graphical models. *Advances in Neural Information Processing Systems*, pages 604–612, 2010.

Jerome H Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Alexander J Gibberd and James D B Nelson. Regularized estimation of piecewise constant gaussian graphical models: The group-fused graphical lasso. *Journal of Computational and Graphical Statistics*, 26(3):623–634, 2017.

Yi Guo, Junbin Gao, and Feng Li. Spatial subspace clustering for hyperspectral data segmentation. In *International Conference on Digital Information Processing and Communications(ICDIPC)*, pages 180–190, 2013.

Jonas M. B. Haslbeck and Lourens J. Waldorp. mgm: Estimating time-varying mixed graphical models in high-dimensional data. *Journal of Statal Software*, 93(8), 2020.

Heping He and Thomas A Severini. Asymptotic properties of maximum likelihood estimators in models with multiple change points. *Bernoulli*, 16(3):759–779, 2010.

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

Yuchen Jiao, Yanxi Chen, and Yuantao Gu. Subspace change-point detection: A new model and solution. *IEEE Journal of Selected Topics in Signal Processing*, 12(6):1224–1239, 2018.

Rebecca Killick, Paul Fearnhead, and Idris A Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107 (500):1590–1598, 2012.

Lisa Kirkland, Frans Kanfer, and Sollie Millard. *LASSO Tuning Parameter Selection*. 2015.

Mladen Kolar and Eric P Xing. On time varying undirected graphs. In *International Conference on Artificial Intelligence and Statistics(AISTATS)*, volume 15, pages 407–415.

Mladen Kolar and Eric P. Xing. Estimating networks with jumps. *Electronic Journal of Statistics*, 6:2069–2106, 2012.

Chunguang Li and Rene Vidal. Structured sparse subspace clustering: A unified optimization framework. In *Computer Vision and Pattern Recognition(CVPR)*, pages 277–286.

Chunguang Li, Chong You, and Rene Vidal. Structured sparse subspace clustering: A joint affinity learning and subspace clustering framework. *IEEE Transactions on Image Processing*, 26(6):2988–3001, 2017.

Nicolai Meinshausen and Peter Buhlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

J. L. Nowinski. Applications of functional analysis in engineering. *International Journal of Engineering Science*, 19(11):1377–1390, 1981.

Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: a review. *Sigkdd Explorations*, 6(1):90–105, 2004.

Huitong Qiu, Fang Han, Han Liu, and Brian S Caffo. Joint estimation of multiple graphical models from high dimensional time series. *Journal of The Royal Statistical Society Series B-statistical Methodology*, 78(2):487–504, 2016.

Adam J Rothman, Peter J Bickel, Elizaveta Levina, and J Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

Naoki Saito. Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum description length criterion. *Wavelet Analysis and Its Applications*, 4:299–324, 1994.

A J Scott and M Knott. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, 30(3):507, 1974.

Robert Tibshirani, Michael A Saunders, Saharon Rosset, J Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of The Royal Statistical Society Series B-statistical Methodology*, 67(1):91–108, 2005.

Stephen Tierney, Junbin Gao, and Yi Guo. Subspace clustering for sequential data. In *Computer Vision and Pattern Recognition(CVPR)*, pages 1019–1026, 2014.

Yuxin Wen, Jianguo Wu, and Yuan Yuan. Multiple-phase modeling of degradation signal for condition monitoring and remaining useful life prediction. *IEEE Transactions on Reliability*, 66(3):924–938, 2017.

Yuxin Wen, Jianguo Wu, Qiang Zhou, and Tzuliang Tseng. Multiple-change-point modeling and exact bayesian inference of degradation signal for prognostic improvement. *IEEE Transactions on Automation Science and Engineering*, 16(2):613–628, 2019.

Jianguo Wu, Yong Chen, and Shiyu Zhou. Online detection of steady-state operation using a multiple-change-point model and exact bayesian inference. *IIE Transactions*, 48(7): 599–613, 2016.

Jianguo Wu, Honglun Xu, Chen Zhang, and Yuan Yuan. A sequential bayesian partitioning approach for online steady-state detection of multivariate systems. *IEEE Transactions on Automation Science and Engineering*, 16(4):1882–1895, 2019.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society Series B-statistical Methodology*, 68(1):49–67, 2006.

Chen Zhang, Hao Yan, Seungho Lee, and Jianjun Shi. Dynamic multivariate functional data modeling via sparse subspace learning. *Technometrics*, pages 1–14, 2020.

Junjian Zhang, Chunguang Li, Honggang Zhang, and Jun Guo. Low-rank and structured sparse subspace clustering. In *Visual Communications and Image Processing(VCIP)*, pages 1–4.

Zhao, Ren, Tingni, Sun, Cun-Hui, Zhang, Harrison, H., and Zhou. Asymptotic normality and optimalities in estimation of large gaussian graphical models. *Annals of Statistics*, 2015.

Shuheng Zhou, John Lafferty, and Larry Wasserman. Time varying undirected graphs. *Machine Learning*, 80(2):295–319, 2010.