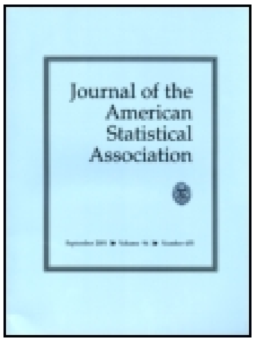

Journal of the
American
Statistical
Association

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Clustering High-Dimensional Time Series Based on Parallelism

Ting Zhang

To cite this article: Ting Zhang (2013) Clustering High-Dimensional Time Series Based on Parallelism, Journal of the American Statistical Association, 108:502, 577-588, DOI: 10.1080/01621459.2012.760458

To link to this article: https://doi.org/10.1080/01621459.2012.760458

Accepted author version posted online: 10 Jan 2013.
Published online: 10 Jan 2013.

Clustering High-Dimensional Time Series Based on Parallelism

Ting ZHANG

This article considers the problem of clustering high-dimensional time series based on trend parallelism. The underlying process is modeled as a nonparametric trend function contaminated by locally stationary errors, a special class of nonstationary processes. For each group where the parallelism holds, I semiparametrically estimate its representative trend function and vertical shifts of group members, and establish their central limit theorems. An information criterion, consisting of in-group similarities and number of groups, is then proposed for the purpose of clustering. I prove its theoretical consistency and propose a splitting-coalescence algorithm to reduce the computational burden in practice. The method is illustrated by both simulation and a real-data example.

KEY WORDS: Information criterion; Local linear estimation; Locally stationary processes; Low- and high-dimensional clustering; Semiparametric methods.

1. INTRODUCTION

Because objects of interest in scientific areas (e.g., weather stations in climate science, patients in clinical trials, fields in agriculture, stocks in finance, and countries in economics) are usually too numerous to process as individual entities, they are often grouped into categories based on their feature similarities. Cluster analysis is the act of searching homogenous subgroups in a dataset. Conventional methods such as the k-means clustering (MacQueen 1967), the EM algorithm (Dempster, Laird, and Rubin 1977) on Gaussian mixture models and the hierarchical clustering (Ward 1963) have been widely used in many applications. When objects of interest are functions, which are intrinsically infinite-dimensional, these conventional methods are usually coupled with dimension reduction techniques, for example, the principal component analysis. Hall, Lee, and Park (2007) proposed to focus on some summary statistics, quantities that are nonlinear functions of the data curves, to avoid the difficulty on interpreting higher-order principle components. Abraham et al. (2003) proposed to first fit the functional data by B-splines and then apply the k-means algorithm on the fitted coefficients to find clusters. Antoniadis, Bigot, and von Sachs (2009) used wavelet thresholding and Neyman truncation for dimension reduction and the EM algorithm on Gaussian mixture models for clustering. Other contributions can be found in Tarpey and Kinateder (2003), García-Escudero and Gordaliza (2005), Serban and Wasserman (2005), Chiou and Li (2007), and references therein.

The article is motivated by a compiled dataset containing daily cell phone download activities (applications, audio, images, ringtones, and wall papers) for 129 area codes in the United States from July 9, 2005 to May 31, 2006. Hence, the dimension $p = 129$, which is comparable to the sample size $n = 327$. These data were analyzed by Degras et al. (2012) by means of hypothesis testing, and the goal was to determine whether the download trends after a logarithmic transformation in different area codes were identical up to vertical shifts. These individual shifts are caused by the difference in numbers of cell phone users among different area codes. Hence, a problem of considerable interest is to find homogeneous subgroups based on trend parallelism. The phone company and its commercial partners can thus place the same advertising efforts and commercial incentives for area codes within the same subgroup. The results can also be used for bandwidth allocation in phone networks.

Although the problem of clustering has been widely studied in the literature, the majority of existing results focused on clustering based on equality, namely, two objects (usually vectors) are put in the same subgroup if their means are considered to be equal. Functional clustering based on shape similarities, however, has been a much less explored area. Heckman and Zamar (2000) considered clustering based on the rank correlation coefficient, where two functions are in the same subgroup if one is a strictly increasing function of the other. Hence, it is not directly applicable to the current problem concerning parallelism. Chiou and Li (2008) proposed a correlation-based functional clustering method, where the similarity measure was taken to be the functional inner product. Their method was based on the truncated Karhunen–Loève representation, and thus the corresponding theoretical results including the clustering consistency cannot be trivially generalized to allow time series data with discrete measurement errors; see Hitchcock, Casella, and Booth (2006) for their conceptual difference. In addition, their method requires that the number of curves in each subgroup goes to infinity, which can be quite restrictive in practice because, in some applications, a few objects can exhibit inhomogeneity from the remaining majority, as can be seen from our data analysis in Section 4.3 that there are six subgroups with size one in the estimated partition. This inevitably excludes their method from applications involving small subgroups. Furthermore, both the aforementioned methods require that the total number of clusters is known a priori, the same as the conventional k-means clustering. The major goal of this article is to develop a consistent clustering method for both low- and high-dimensional time series data based on trend parallelism, a category of shape similarity, which is also capable of automatically identifying the total number of clusters.

Ting Zhang is Assistant Professor, Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, IA 52242-1409 (E-mail: *ting-zhang@uiowa.edu*). The author is grateful to the Co-Editor, the Associate Editor, and an anonymous referee for their helpful comments and suggestions.

Journal of the American Statistical Association
June 2013, Vol. 108, No. 502, Theory and Methods
DOI: 10.1080/01621459.2012.760458

Recently, Degras et al. (2012) considered the problem of testing the parallelism hypothesis, and the test was applied to devise a clustering algorithm analogous to the backward variable selection for linear regression models. Compared with their results, the current article makes several substantial improvements. First, the clustering algorithm of Degras et al. (2012) is based on stepwise hypothesis testings, thus its theoretical property is somewhat hard to understand. In contrast, I propose an information criterion and prove its clustering consistency, which refers to selecting the true underlying partition with probability tending to one. Second, the current method does not require estimating the long-run variance function as in Degras et al. (2012), which itself can be a very challenging task. Although there are estimators proposed in the literature, for example, by Zhou and Wu (2010), they usually depend on (and are very sensitive to) other user-chosen parameters. Hence, it seems desirable to develop alternative methods, as in the current article, that do not involve the estimation of nuisance parameters and still lead to asymptotically valid inference. In addition, I establish central limit theorems of the semiparametric estimates and prove that estimates of the parametric components achieve the $n^{1/2}$-convergence rate. Hence, the current article shed new light on high-dimensional semiparametric inference.

Suppose we observe p (which can grow to infinity) time series $(y_{k,i})_{i=1}^n$, $k = 1, \ldots, p$, according to the model

$$y_{k,i} = \mu_k(i/n) + e_{k,i}, \quad i = 1, \ldots, n, \tag{1}$$

where $\mu_k : [0, 1] \to \mathbb{R}$ are unknown smooth trend functions, and $(e_{k,i})_{i=1}^n$ are zero mean error processes. The scaling device i/n in (1) is useful in characterizing the smoothness and is necessary for providing asymptotic justification for any nonparametric smoothing estimators as demonstrated by Robinson (1989); see Robinson (1991), Dahlhaus (1996, 1997), and Cai (2007) for more discussions. We want to find a minimal number of nonoverlapping subgroups $\mathcal{S}_1 \cup \cdots \cup \mathcal{S}_Q = \{1, \ldots, p\}$ such that for each $q = 1, \ldots, Q$ and $k \in \mathcal{S}_q$,

$$\mu_k(t) = \mu_{\mathcal{S}_q}(t) + c_k, \quad t \in [0, 1], \tag{2}$$

for some common trend function $\mu_{\mathcal{S}_q} : [0, 1] \to \mathbb{R}$ and individual shifts $c_k \in \mathbb{R}$. We assume that

$$\sum_{k \in \mathcal{S}_q} c_k = 0 \tag{3}$$

to ensure the identifiability. Compared with existing results on clustering, the present article has at least three distinctive features. First, the objects are clustered based on the parallelism of their trend functions, which are intrinsically infinite-dimensional, and the number of objects being clustered is allowed to grow to infinity. The high dimensionality of both kinds can make the clustering become much more challenging than in conventional settings. Second, I consider the situation with measurement errors that are modeled as locally stationary processes (see Section 2.1), a special class of nonstationary processes. Hence, the objects of interest (trend functions) need to be estimated nonparametrically from the contaminated data before calling the clustering algorithm. This nonparametric estimation step poses an extra difficulty in proving the clustering consistency. Third, unlike the conventional k-means clustering and the results in Chiou and Li (2008), I consider the situation that the total number of clusters, Q, is unknown (unsupervised learning) and can also grow to infinity. Hence, the p objects can belong to p distinguishable subgroups if this is indeed the case.

The rest of the article is organized as follows. Section 2.1 introduces the locally stationary framework and the corresponding dependence measure. Section 2.2 provides estimators of the underlying trend functions and individual shifts and presents their asymptotic properties. Section 2.3 proposes an information criterion, consisting of in-group similarities and number of groups, and provides its clustering consistency. Its detailed implementation is described in Section 3. Section 4 contains simulation results and a real-data analysis. Technical proofs are deferred to the Appendix.

2. MAIN RESULTS

2.1 Model Assumptions

To make our theory be widely applicable, we allow locally stationary error processes in (1). Locally stationary processes appear frequently in practice where the underlying data generating mechanism bears a smooth temporal change. Such processes have been studied in the literature by means of spectral representations (Dahlhaus 1996, 1997), pseudo-differential operators (Mallat, Papanicolaou, and Zhang 1998), discrete nondecimated wavelets (Nason, von Sachs, and Kroisandt 2000), and smooth localized complex exponentials (Ombao, von Sachs, and Guo 2005) among others. We shall here follow the framework of Draghicescu, Guillas, and Wu (2009) and assume that

$$e_{k,i} = G(i/n; \mathcal{F}_{k,i}), \quad \mathcal{F}_{k,i} = (\ldots, \epsilon_{k,i-1}, \epsilon_{k,i}), \tag{4}$$

where $\epsilon_{l,j}, l, j \in \mathbb{Z}$, are independent and identically distributed (iid) random variables, and G is a measurable function. For a random variable X, we write $X \in \mathcal{L}^r$, $r > 0$, if $\|X\|_r = \{E(|X|^r)\}^{1/r} < \infty$, and denote $\|\cdot\| = \|\cdot\|_2$. The process (4) is locally stationary if G is stochastically Lipschitz continuous ($G \in$ Lip in short), namely, there exists a constant $c_0 < \infty$ such that

$$\|G(t_1; \mathcal{F}_{k,i}) - G(t_2; \mathcal{F}_{k,i})\| \le c_0 |t_1 - t_2|$$

holds uniformly for all $t_1, t_2 \in [0, 1]$. Hence, the underlying data generating mechanism can only change smoothly over time. Let $(\epsilon^\star_{l,j})_{l,j\in\mathbb{Z}}$ be an iid copy of $(\epsilon_{l,j})_{l,j\in\mathbb{Z}}$, and $\mathcal{F}^\star_{k,i} = (\mathcal{F}_{k,-1}, \epsilon^\star_{k,0}, \epsilon_{k,1}, \ldots, \epsilon_{k,i})$ be the coupled shift process, we define the functional dependence measure

$$\theta_{i,r} = \sup_{t\in[0,1]} \|G(t; \mathcal{F}_{k,i}) - G(t; \mathcal{F}^\star_{k,i})\|_r.$$

According to the idea of coupling, the quantity $\theta_{i,r}$ measures the ith step ahead impact of the current innovation. If the short range dependence condition $\Theta_{0,r} = \sum_{i=0}^{\infty} \theta_{i,r} < \infty$ holds for some $r \ge 2$, then the long-run variance function $g(t) = \sum_{l\in\mathbb{Z}} E\{G(t; \mathcal{F}_{k,0}) G(t; \mathcal{F}_{k,l})\} < \infty$ uniformly over $t \in [0, 1]$. We need the following technical assumptions in establishing the main results.

(A1) $G \in$ Lip and $\sup_{t\in[0,1]} \|G(t; \mathcal{F}_{k,i})\|_r < \infty$ for some $r > 2$.
(A2) $\mu_k \in \mathcal{C}^3[0, 1]$, $k = 1, \ldots, p$, and $\max_{1\le k\le p} \sup_{t\in[0,1]} |\mu_k''(t)| \le c$ for some $c < \infty$.
(A3) $g \in \mathcal{C}^2[0, 1]$ and $(np)^{-1} \sum_{k=1}^p \sum_{i=1}^n E(e_{k,i}^2) \ge \varepsilon$ for some $\varepsilon > 0$.

2.2 Semiparametric Estimation

For each individual time series $(y_{k,i})_{i=1}^n$, its trend function (along with its derivative) can be estimated nonparametrically by the local linear estimate (Fan and Gijbels 1996)

$$\{\hat{\mu}_k(t), \hat{\mu}'_k(t)\} = \operatorname*{argmin}_{\eta_0,\eta_1\in\mathbb{R}} \sum_{i=1}^n \{y_{k,i} - \eta_0 - (i/n - t)\eta_1\}^2 K\left(\frac{i/n - t}{b_n}\right), \quad (5)$$

where $K(\cdot)$ is the kernel function and b_n is a bandwidth sequence satisfying $b_n \to 0$ and $nb_n \to \infty$. Throughout the article, we assume that the kernel function $K(\cdot)$ is a symmetric and bounded function in $\mathcal{C}^1[-1, 1]$ with $\int_{-1}^1 K(v)dv = 1$. A popular choice is the Epanechnikov kernel $K(v) = (3/4)\max(0, 1 - v^2)$. Observe that (5) has the closed form solution

$$\hat{\mu}_k(t) = \sum_{i=1}^n y_{k,i} w_i(t), \quad (6)$$

where $w_i(t) = K\{(i/n - t)/b_n\}\{S_2(t) - (t - i/n)S_1(t)\}/\{S_2(t) S_0(t) - S_1^2(t)\}$ are the local linear weights and $S_j(t) = \sum_{i=1}^n (t - i/n)^j K\{(i/n - t)/b_n\}$.

Suppose, toward the end of this subsection, that the parallelism assumption holds for the subgroup $\mathcal{S}_q$. Then its common trend function $\mu_{\mathcal{S}_q}(\cdot)$ as in (2) can be naturally estimated by

$$\hat{\mu}_{\mathcal{S}_q}(t) = |\mathcal{S}_q|^{-1} \sum_{k\in\mathcal{S}_q} \hat{\mu}_k(t), \quad t \in [0, 1], \quad (7)$$

where $|\mathcal{S}_q|$ is the cardinality, number of elements, of $\mathcal{S}_q$. By (2), the individual shifts c_k, $k \in \mathcal{S}_q$, can be estimated by

$$\hat{c}_k = n^{-1} \sum_{i=1}^n \{\hat{\mu}_k(i/n) - \hat{\mu}_{\mathcal{S}_q}(i/n)\}. \quad (8)$$

For the constraint (3) on identifiability, by (7), it is not hard to see that $\sum_{k\in\mathcal{S}_q} \hat{c}_k = 0$. Hence, $\hat{c}_k = c_k = 0$ if $|\mathcal{S}_q| = 1$. Theorem 2.1 provides the central limit theorems for estimators (7) and (8).

Theorem 2.1. Assume (A1)–(A3), $\Theta_{0,2} < \infty$, $b_n \to 0$, and $nb_n \to \infty$. Let $\phi_2 = \int_{-1}^1 K(v)^2 dv$, $\kappa_2 = \int_{-1}^1 v^2 K(v)dv$, and $g_1 = \int_0^1 g(t)dt$. If the parallelism holds for the subgroup $\mathcal{S}_q$, then

$$(nb_n|\mathcal{S}_q|)^{1/2}[\hat{\mu}_{\mathcal{S}_q}(t) - E\{\hat{\mu}_{\mathcal{S}_q}(t)\}] \Rightarrow N\{0, \phi_2 g(t)\}, \quad (9)$$

where $E\{\hat{\mu}_{\mathcal{S}_q}(t)\} = \mu_{\mathcal{S}_q}(t) + b_n^2 \mu''_{\mathcal{S}_q}(t)\kappa_2/2 + O(b_n^3)$. If in addition $|\mathcal{S}_q| > 1$, then

$$\{n|\mathcal{S}_q|/(|\mathcal{S}_q| - 1)\}^{1/2}(\hat{c}_k - c_k) \Rightarrow N(0, g_1). \quad (10)$$

2.3 High-Dimensional Clustering

Existing methods on clustering usually rely on a properly chosen pairwise distance, which could be the widely used Euclidean distance, the functional inner product, and the rank correlation (Heckman and Zamar 2000) among others. However, this may not be advantageous because a homogeneous subgroup may contain more than two objects. Hence, I define the in-group (instead of pairwise) similarity measure

$$\text{RSS}(\mathcal{S}_q) = \sum_{k\in\mathcal{S}_q} \sum_{i=1}^n \{y_{k,i} - \hat{\mu}_{\mathcal{S}_q}(i/n) - \hat{c}_k\}^2. \quad (11)$$

Since $\hat{\mu}_{\mathcal{S}_q}(i/n) + \hat{c}_k$ is the semiparametric estimate of $E(y_{k,i})$, $k \in \mathcal{S}_q$, $i = 1, \ldots, n$, the statistic (11) provides the residual sum of squares for the subgroup $\mathcal{S}_q$ under the parallelism (2). Let $\mathcal{P} = \{\mathcal{S}_1, \ldots, \mathcal{S}_Q\}$ be the implied partition with cardinality $|\mathcal{P}| = Q$, and $\text{RSS}(\mathcal{P}) = \sum_{q=1}^Q \text{RSS}(\mathcal{S}_q)$ be the residual sum of squares across all subgroups. I consider an extended Bayesian information criterion (EBIC) for clustering high-dimensional time series data, which takes the form

$$\text{EBIC}(\mathcal{P}) = (np)\log\{\text{RSS}(\mathcal{P})/(np)\} + \tau_n|\mathcal{P}|. \quad (12)$$

The criterion (12) depends on a tuning parameter τ_n. If $\tau_n = \log n$, then (12) becomes the traditional BIC. Larger τ_n leads to stronger penalization on the number of clusters, and vice versa. We estimate $\mathcal{P}_0$, the true underlying partition defined by (14), by minimizing (12).

Throughout the article, $\mathscr{P}$ denotes the space of set partitions of $\{1, \ldots, p\}$. For partitions $\mathcal{P}_1, \mathcal{P}_2 \in \mathscr{P}$, we write $\mathcal{P}_1 \preceq \mathcal{P}_2$ if each cluster of $\mathcal{P}_1$ is a subset of some cluster of $\mathcal{P}_2$ or, equivalently, if two individuals belong to the same cluster in $\mathcal{P}_1$, they belong to the same cluster in $\mathcal{P}_2$. Recall that $\mathcal{P} = \{\mathcal{S}_1, \ldots, \mathcal{S}_Q\} \in \mathscr{P}$ is the partition with cardinality $|\mathcal{P}| = Q$. For $q = 1, \ldots, Q$, let

$$\Delta_n(\mathcal{S}_q) = \min_{(\nu_i)\in\mathbb{R}^n,\ (\omega_k)\in\mathbb{R}^{|\mathcal{S}_q|}} \sum_{k\in\mathcal{S}_q} \sum_{i=1}^n \{\mu_k(i/n) - \nu_i - \omega_k\}^2. \quad (13)$$

Then $\Delta_n(\mathcal{P}) = \sum_{q=1}^Q \Delta_n(\mathcal{S}_q)$ provides the minimal squared distance of $\{E(y_{k,i})\}_{k,i} \in \mathbb{R}^{p\times n}$ to the subspace in which the parallelism (2) holds for each subgroup $\mathcal{S}_1, \ldots, \mathcal{S}_Q$. We define the true underlying partition

$$\mathcal{P}_{0,n} = \operatorname*{argmin}_{\mathcal{P}\in\mathscr{P}}\{|\mathcal{P}| : \Delta_n(\mathcal{P}) = 0\}, \quad (14)$$

namely, the partition with the smallest cardinality that satisfies $\Delta_n(\mathcal{P}) = 0$. We shall in the following suppress the subscript n and write $\Delta(\mathcal{P}) = \Delta_n(\mathcal{P})$ and $\mathcal{P}_0 = \mathcal{P}_{0,n}$ for notational ease. Hence, $\Delta(\mathcal{P}) = 0$ if $\mathcal{P} \preceq \mathcal{P}_0$. If $\mathcal{P} \npreceq \mathcal{P}_0$, then at least one individual is incorrectly clustered and, by the construction (13), $n^{-1}\Delta(\mathcal{P})$ is bounded away from zero. If in addition $|\mathcal{P}| < |\mathcal{P}_0|$, then at least $|\mathcal{P}_0| - |\mathcal{P}|$ individuals are incorrectly clustered and we expect that $\{(|\mathcal{P}_0| - |\mathcal{P}|)n\}^{-1}\Delta(\mathcal{P})$ is bounded away from zero. To summarize, we assume that

(A4) there exists a constant $\epsilon' > 0$ such that

$$\inf_{\mathcal{P}\npreceq\mathcal{P}_0} \frac{\Delta(\mathcal{P})}{n\max(1, |\mathcal{P}| - |\mathcal{P}_0|)} \geq \epsilon'.$$

For either fixed or large p, Theorem 2.2 provides the selection consistency of (12), which refers to the selection of the true underlying partition $\mathcal{P}_0$.

Theorem 2.2. Assume (A1)–(A4), $\Theta_{0,4} < \infty$ and $b_n = cn^{-1/5}$ for some $0 < c < \infty$. Suppose the tuning parameter satisfies

(i) $b_n^{-1} = o(\tau_n)$ and $\tau_n = o(n)$ if $p < \infty$ is fixed; or

(ii) $b_n^{-1}(nb_n)^{1/2} = O(\tau_n)$ and $\tau_n = o(n)$ if $p = O(n^\iota)$ for some $\iota < 1/4$.

Then for any $\mathcal{P} \neq \mathcal{P}_0$,

$$\text{pr}\{\text{EBIC}(\mathcal{P}) < \text{EBIC}(\mathcal{P}_0)\} = O[\{(nb_n)^{-1/2} + b_n^2\}^2 p^2]. \tag{15}$$

In addition, with probability tending to 1 as $n \to \infty$,

$$\mathcal{P}_0 = \underset{\mathcal{P}\in\mathscr{P}}{\text{argmin}}\, \text{EBIC}(\mathcal{P}). \tag{16}$$

Remark 2.1. Given a partition $\mathcal{P} = \{\mathcal{S}_1, \ldots, \mathcal{S}_{|\mathcal{P}|}\}$, we estimate semiparametrically the common trend functions $\mu_{\mathcal{S}_q}(\cdot)$, $q = 1, \ldots, |\mathcal{P}|$, and the individual shifts c_k, $k = 1, \ldots, p$. For each nonparametric component, the effective number of parameters used in kernel smoothing is b_n^{-1} (Hurvich, Simonoff, and Tsai 1998). Hence, by the constraint (3), the total effective number of parameters under $\mathcal{P}$ is $b_n^{-1}|\mathcal{P}| + (p - |\mathcal{P}|)$. If the dimension $p < \infty$, then Theorem 2.2 (i) allows the nonparametric BIC,

$$\text{BIC}(\mathcal{P}) = (np)\log\{\text{RSS}(\mathcal{P})/(np)\} + (b_n^{-1} - 1)|\mathcal{P}|\log(nb_n), \tag{17}$$

namely, $\tau_n = (b_n^{-1} - 1)\log(nb_n)$. Note that the effective sample size nb_n is used in (17). If the dimension $p \to \infty$, then Theorem 2.2(ii) indicates a stronger penalization on the model complexity. Chen and Chen (2008) considered variable selection for linear models with iid Gaussian errors, and demonstrated that a heavier penalization than traditional BIC is necessary when the number of covariates grows to infinity with the sample size.

3. IMPLEMENTATION

3.1 The Splitting-Coalescence (SC) Algorithm

In practice, it is computationally infeasible to calculate the EBIC for every possible partition, since the total number of partitions of a set with p individuals is the pth Bell number, which is very large even when p is relatively small. For high-dimensional data, it is typical that the selection criterion is applied only to a manageable number of candidate models (Chen and Chen 2008). I shall here propose an SC algorithm that can effectively reduce the computational burden. The algorithm searches the homogeneous subgroup by repeatedly splitting individuals having the largest residual sum of squares and applying the EBIC on this sequence of candidate models. After identifying a homogeneous subgroup, it coalesces the split individuals and repeats. Its detailed implementation is as follows, which involves iteratively a splitting step and a coalescence step.

```
(1) (Initialization) Set $\mathcal{U}_1 = \{1, \ldots, p\}$.
(2) (Splitting) Set $\mathcal{S}^{(1)} = \mathcal{U}_1$ and the partition $\mathcal{P}^{(1)} = \{\mathcal{S}^{(1)}\}$.
    (a) Treat $\mathcal{S}^{(1)}$ as a subgroup, and compute $\hat{\mu}_{\mathcal{S}^{(1)}}(\cdot)$ and
        $\hat{c}_k$, $k \in \mathcal{S}^{(1)}$, by (7) and (8);
    (b) Compute $\text{RSS}_k = \sum_{i=1}^n \{y_{k,i} - \hat{\mu}_{\mathcal{S}^{(1)}}(i/n) - \hat{c}_k\}^2$ for
        each individual $k \in \mathcal{S}^{(1)}$;
    (c) Remove the one with the largest $\text{RSS}_k$ from $\mathcal{S}^{(1)}$ and
        denote the remaining $\mathcal{S}^{(2)}$;
    (d) Let $\mathcal{P}^{(2)}$ be the subpartition of $\mathcal{P}^{(1)}$ by splitting $\mathcal{S}^{(1)}$
        into $\mathcal{S}^{(2)}$ and an individual;
    (e) Repeat (a)–(d) to get $\mathcal{S}^{(p)} \subset \cdots \subset \mathcal{S}^{(1)}$ and $\mathcal{P}^{(p)} \preceq$
        $\cdots \preceq \mathcal{P}^{(1)}$.
(3) (Coalescence)
    (a) Compute the EBIC for $\mathcal{P}^{(1)}, \ldots, \mathcal{P}^{(p)}$ and denote the
        minimizer $\mathcal{P}^{(l)}$;
    (b) Set $\mathcal{S}_1 = \mathcal{S}^{(l)}$, and stop the algorithm if $\mathcal{S}_1 = \mathcal{U}_1$;
    (c) Coalesce the remaining individuals in $\mathcal{P}^{(l)}$ to get
        $\mathcal{U}_2 = \mathcal{U}_1 \backslash \mathcal{S}_1$.
(4) Repeat Steps 2 and 3 to get the estimated partition $\hat{\mathcal{P}} =$
    $\{\mathcal{S}_1, \mathcal{S}_2, \ldots\}$.
```

3.2 Bandwidth Selection

The construction of semiparametric estimators in Section 2.2 and development of the clustering procedure in Sections 2.3 and 3.1 both rely on the local linear estimates (6), where the same bandwidth is used for $k = 1, \ldots, p$. This can help correct the bias in semiparametric estimation, simplify the implementation of the algorithm, and derive rigorous asymptotic properties. Specifically, if the parallelism holds for the subgroup $\mathcal{S}_q$ and the same bandwidth is used, then for any $k \in \mathcal{S}_q$,

$$E\{\hat{\mu}_k(t) - \hat{\mu}_{\mathcal{S}_q}(t)\} = \sum_{i=1}^n \{\mu_k(i/n) - \mu_{\mathcal{S}_q}(i/n)\}w_i(t) = c_k,$$

which does not have a bias. Also, the clustering algorithm requires the search over all the possible partitions (or a considerable but manageable amount of candidate partitions by using the SC algorithm). If the same bandwidth is used, then we only need to perform the nonparametric estimation (6) once, and simply use Equations (7) and (8) to obtain the corresponding semiparametric estimates for different partitions $\mathcal{P}$, which can significantly reduce the computational burden. In addition, since the number of time series being clustered is allowed to grow with the sample size, it will be quite nontrivial to derive a bound for the maximum as in Lemma A.1, a crucial part in proving the consistency of the EBIC, if the bandwidths are different.

Note that $\mathscr{P}$ contains partitions with subgroups that contain only one individual time series, where the semiparametric estimate reduces to the local linear estimate (6). The choice $b_n = cn^{-1/5}$ for some constant $0 < c < \infty$ achieves the optimal rate for the asymptotic mean integrated squared error. However, the problem of choosing an optimal constant c is quite nontrivial. For independent data, it was considered by Härdle and Marron (1985), Härdle, Hall, and Marron (1988), and Park and Marron (1990) for kernel density and regression estimation, and Kulasekera and Wang (1997) and Fan and Linton (2003) for nonparametric hypothesis testings. In our case, however, it is further complicated by the presence of dependence and nonstationarity. Following Zhang and Wu (2012), we consider the generalized cross-validation (GCV) selector by Craven and Wahba (1979) and estimate the covariance matrix $\boldsymbol{\Gamma}_n = \{E(e_{k,i}e_{k,j})\}_{1\le i,j\le n}$ to correct the dependence. In particular, let $\boldsymbol{Y}_k = (y_{k,1}, \ldots, y_{k,n})^\top$, then for any bandwidth $b \in (0, 1)$, the local linear fitted values can be written as $\hat{\boldsymbol{Y}}_k(b) = \boldsymbol{H}(b)\boldsymbol{Y}_k$, where $\boldsymbol{H}(b) = \{w_j(i/n)\}_{1\le i,j\le n}$ is the associated hat matrix. The bandwidth $\hat{b}_n$ is chosen by minimizing

$$\text{GCV}(b) = p^{-1}\sum_{k=1}^p \frac{n^{-1}\{\hat{\boldsymbol{Y}}_k(b) - \boldsymbol{Y}_k\}^\top \hat{\boldsymbol{\Gamma}}_n^{-1}\{\hat{\boldsymbol{Y}}_k(b) - \boldsymbol{Y}_k\}}{[1 - \text{tr}\{\boldsymbol{H}(b)\}/n]^2}. \tag{18}$$

The covariance matrix estimate $\hat{\boldsymbol{\Gamma}}_n$ can be obtained by using the banding technique as in Bickel and Levina (2008) and Wu and Pourahmadi (2009). The GCV selector (18) works reasonably well in our simulation studies.

We shall here discuss the choice of the tuning parameter τ_n. If $p < \infty$, then one can simply choose $\tau_n = (b_n^{-1} - 1)\log(nb_n)$ to form the nonparametric BIC as in Remark 2.1. If $p = p(n) \to \infty$ as $n \to \infty$, by Theorem 2.2(ii), it suffices to set $\tau_n = (b_n^{-1} - 1)(nb_n)^{1/2}$. Note that $(nb_n)^{1/2}$ grows to infinity at a faster rate than $\log(nb_n)$. In practice, however, it can be very ambiguous whether the dimension p should be treated as fixed or not. To avoid this practical inconvenience, following Chen and Chen (2008), let $\gamma \in [0, 1]$ and we consider

$$\begin{aligned}\text{EBIC}_\gamma(\mathcal{P}) &= (np)\log\{\text{RSS}(\mathcal{P})/(np)\} \\ &\quad + \left(b_n^{-1} - 1\right)|\mathcal{P}|\{(1-\gamma)\log(nb_n) + \gamma(nb_n)^{1/2}\}.\end{aligned} \tag{19}$$

Larger γ indicates stronger penalization on the number of clusters, and vice versa. If $\gamma = 0$, then (19) reduces to the nonparametric BIC (17) that is consistent if $p < \infty$ is fixed, while $\gamma > 0$ entails the consistency no matter whether the dimension is fixed or grows to infinity with the sample size. For high-dimensional problems, the simple choice $\gamma = 1$ seems to perform reasonably well as can be seen from our simulation studies.

4. NUMERICAL EXPERIMENTS

4.1 A Comparison With Testing-Based Method

Although functional clustering based on shape similarities has been studied by Heckman and Zamar (2000) and Chiou and Li (2008), their methods, as discussed in Section 1, cannot be trivially generalized to handle the current problem. Despite the different focus as in Heckman and Zamar (2000) and the different framework as in Chiou and Li (2008), both of them considered the problem of clustering in a supervised learning setting, namely, the total number of subgroups $|\mathcal{P}|$ is prespecified and known to the practitioner. In addition, the method of Chiou and Li (2008) requires that the number of curves in each subgroup grows to infinity for all the subgroups, which can be quite restrictive in practice and inevitably excludes their method from clustering a finite number of objects (not necessarily time series). In contrast, the current method provides consistent clustering results for both low- and high-dimensional locally stationary time series data, and can automatically identify the total number of clusters. Recently, Degras et al. (2012) developed a central limit theorem of the $\mathcal{L}^2$-distance between nonparametric and semiparametric estimates

$$T_{\mathcal{S}_q} = \sum_{k \in \mathcal{S}_q} \int_0^1 \{\hat{\mu}_k(t) - \hat{\mu}_{\mathcal{S}_q}(t) - \hat{c}_k\}^2 dt,$$

under the null hypothesis that the parallelism holds for the subgroup $\mathcal{S}_q$. Due to its poor finite sample performance by directly applying the central limit theorem to obtain the cutoff value, they suggested a simulation-based approximation to the null distribution. In particular, let $\hat{g}(t)$, $t \in [0, 1]$, be an estimate of the long-run variance function that can be obtained by using the results of Degras et al. (2012) and Zhang and Wu (2012). Let $y_{k,i}^\circ = \hat{g}(i/n)z_{k,i}$, where $z_{k,i}$, $k = 1, \ldots, p$, $i = 1, \ldots, n$, are iid standard normal random variables, and $T_{\mathcal{S}_q}^\circ$ be the corresponding test statistic. Then the parallelism (null) hypothesis on $\mathcal{S}_q$ is rejected at level α if $T_{\mathcal{S}_q} > \hat{q}_{1-\alpha}$, where $\hat{q}_{1-\alpha}$ is the $(1-\alpha)$th quantile of $T_{\mathcal{S}_q}^\circ$. The test can be combined with a backward algorithm to devise a clustering algorithm. However, due to the problem of multiple testings, its consistency is not guaranteed. In addition, it can be computationally very intensive if the dimension p is large because the simulation-based approximation needs to be performed at each step. I shall here compare the proposed method with this testing-based (TB) alternative.

Consider model (1) with $e_{k,i} = \vartheta_{k,i}(i/n)$, where for any $t \in [0, 1]$, $\{\vartheta_{k,i}(t)\}_{i \in \mathbb{Z}}$ follows the recursion

$$\vartheta_{k,i}(t) = \rho(t)\vartheta_{k,i-1}(t) + \epsilon_{k,i}. \tag{20}$$

Here, $\epsilon_{l,j}$, $l, j \in \mathbb{Z}$ are iid standard normal innovations and $\rho(t) = 0.5\cos(2\pi t)$. Let $n = 500$, $p = 6$, and the true partition $\mathcal{P}_0 = \{\mathcal{S}_1, \mathcal{S}_2\}$, where $\mathcal{S}_1 = \{1, 2, 3\}$ and $\mathcal{S}_2 = \{4, 5, 6\}$. Let $\mu_{\mathcal{S}_1}(t) = t^3 - t^2$, $\mu_{\mathcal{S}_2}(t) = 0.5\sin(2\pi t)$, and $\mu_k(t) = \mu_{\mathcal{S}_q}(t) + c_k$ for $k \in \mathcal{S}_q$ with individual shifts $c_k = k - 3$. Both the information criterion (19) and the TB method of Degras et al. (2012) are applied to find homogeneous subgroups based on trend parallelism. An exhaustive search is conducted to find the minimizer of EBIC_γ, while for the TB method, the significance level is chosen as $\alpha = 0.05$, and at each step, $\hat{q}_{0.95}$ is obtained with 1000 simulated $T_{\mathcal{S}_q}^\circ$. The Epanechnikov kernel is used, and the clustering quality is measured by the adjusted Rand index (Hubert and Arabie 1985), which measures the agreement between two partitions $\mathcal{P}_0$ and $\hat{\mathcal{P}}$. The index is bounded above by one, and equals to one if $\hat{\mathcal{P}} = \mathcal{P}_0$. Larger indices indicate higher clustering qualities. The results are reported in Table 1 for different choices of bandwidth b and penalization γ. Sample time series plots of the simulated data are provided in Figure 1.

From Figure 1, we can see that it is very hard to find subgroups based on trend parallelism without using an appropriate

Table 1. Means and medians (in parentheses) of the adjusted Rand index for the information criterion (19) with $\gamma \in \{0, 0.5, 1\}$ and the TB method of Degras et al. (2012) across different bandwidths $b = 0.1 + 0.05l$, $l = 0, \ldots, 5$. For each configuration, the results are based on 1000 simulated realizations

	b					
Method	0.1	0.15	0.2	0.25	0.3	0.35
EBIC_0	$0.922_{(1.000)}$	$0.902_{(1.000)}$	$0.871_{(1.000)}$	$0.822_{(1.000)}$	$0.791_{(0.706)}$	$0.750_{(0.706)}$
$\text{EBIC}_{0.5}$	$0.948_{(1.000)}$	$0.971_{(1.000)}$	$0.969_{(1.000)}$	$0.955_{(1.000)}$	$0.937_{(1.000)}$	$0.919_{(1.000)}$
EBIC_1	$0.842_{(1.000)}$	$0.940_{(1.000)}$	$0.971_{(1.000)}$	$0.964_{(1.000)}$	$0.974_{(1.000)}$	$0.957_{(1.000)}$
TB	$0.687_{(0.706)}$	$0.650_{(0.706)}$	$0.601_{(0.706)}$	$0.466_{(0.324)}$	$0.390_{(0.324)}$	$0.342_{(0.324)}$

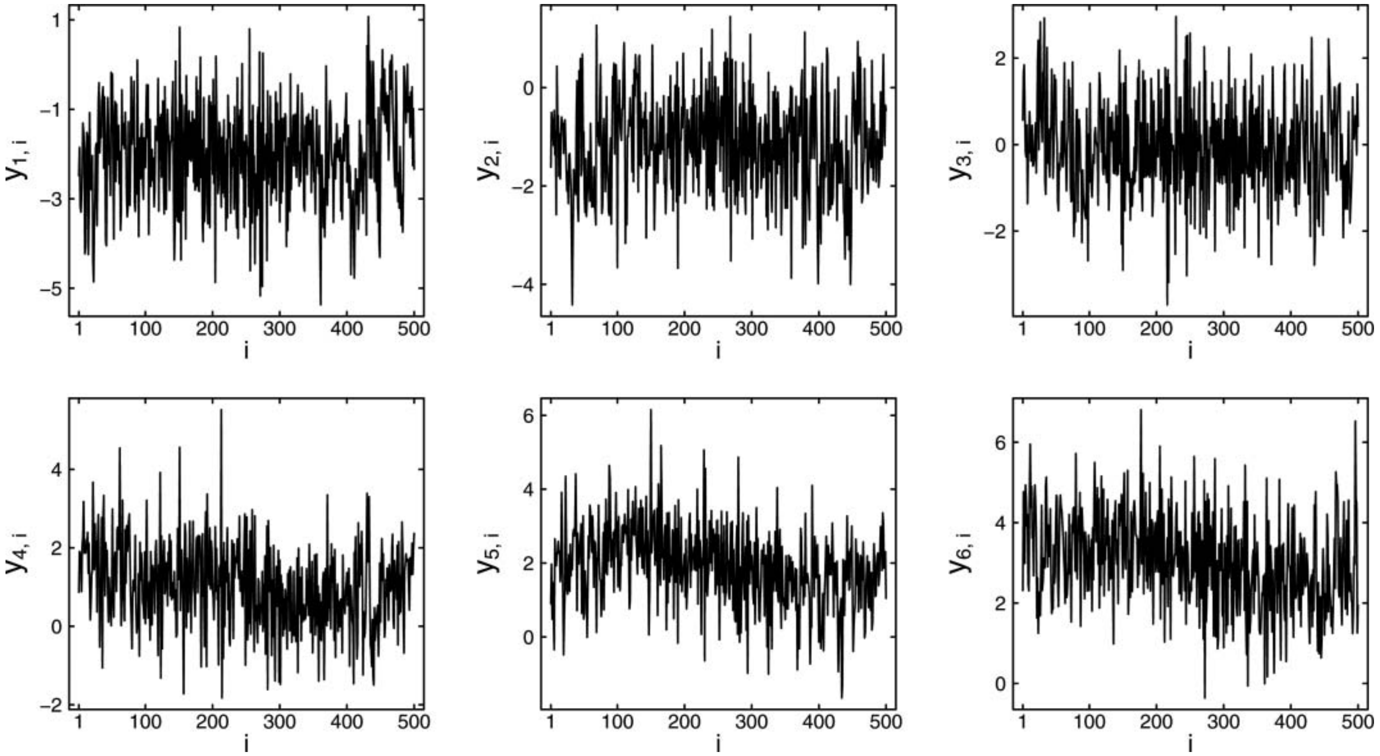

Figure 1. Sample time series plots of simulated data, where the two rows represent two different subgroups based on trend parallelism.

clustering algorithm. For example, the time series displayed in the top left and middle bottom both share an increasing trend at the beginning, then a decreasing trend in the middle and another increasing trend at the end, while they actually belong to two different subgroups. In contrast, the time series displayed in the left bottom has a decreasing trend at the beginning and seems to jiggling around a constant toward the end, while it actually belongs to the same subgroup as the one displayed in the middle bottom. Therefore, the problem of clustering is quite challenging for time series data because it can be very difficult to tell whether an observed pattern is structural or merely due to the dependence. Note that the adjusted Rand index is 0.706 when measuring the similarity between $\{\{1, 2, 3\}, \{4, 5\}, \{6\}\}$ and $\mathcal{P}_0 = \{\{1, 2, 3\}, \{4, 5, 6\}\}$, namely, only a single time series is misspecified to an additional subgroup. From Table 1, we can see that the proposed information criterion (19) does provide very satisfactory results on clustering, and is superior to the TB method of Degras et al. (2012).

4.2 Simulation Studies

I shall in this section examine the finite sample performance of the proposed information criterion (19) under moderate to large dimensions, and its sensitivity to the choice of bandwidth b_n and penalization γ. Consider model (1) with $e_{k,i} = \lambda(i/n)\vartheta_{k,i}(i/n)$, where for any $t \in [0, 1]$, $\lambda(t) = 2\sigma(t - 0.5)^2$ and $\{\vartheta_{k,i}(t)\}_{i\in\mathbb{Z}}$ follows the recursion (20) with iid Rademacher innovations and $\rho(t) = 0.5t - 0.2$. Thus, $(e_{k,i})_{i=1}^n$ is a first-order autoregressive process with time-varying coefficient and variance, and non-Gaussian innovations. I consider the following two situations:

(M1) $\mathcal{P}_0 = \{\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3\}$ with $\mathcal{S}_1 = \{k \mid 1 \leq k \leq 0.4p\}$, $\mathcal{S}_2 = \{k \mid 0.4p < k \leq 0.8p\}$, and $\mathcal{S}_3 = \{k \mid 0.8p < k \leq p\}$.

(M2) $\mathcal{P}_0 = \{\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4\}$ with $\mathcal{S}_1 = \{k \mid 1 \leq k \leq 0.4p\}$, $\mathcal{S}_2 = \{k \mid 0.4p < k \leq 0.7p\}$, $\mathcal{S}_3 = \{k \mid 0.7p < k \leq 0.9p\}$, and $\mathcal{S}_4 = \{k \mid 0.9p < k \leq p\}$.

Denote $\lfloor a \rfloor$ the integer part of $a \in \mathbb{R}$ and $l_j(t)$ the jth order Legendre polynomial. Set $\mu_k(t) = \mu_{\mathcal{S}_q}(t) + c_k$ for $k \in \mathcal{S}_q$, where the common trend function $\mu_{\mathcal{S}_j}(t) = l_{s+j-1}(2t - 1)$, and vertical shifts $c_k = k - 3 - 5\lfloor k/5 \rfloor$. Note that s indicates the level of smoothness. The information criterion (19) is applied to find homogeneous subgroups based on trend parallelism. Due to the large dimension p considered, the SC algorithm proposed in Section 3.1 is used, and the clustering quality is measured by the adjusted Rand index (Hubert and Arabie 1985). Let $n = 500$. The Epanechnikov kernel is used, and different choices of dimension p, noise-to-signal level σ^2, and level of smoothness s are considered. The results are reported in Tables 2 and 3 for cases (M1) and (M2), respectively.

It can be seen that the proposed clustering method has a very good performance as the adjusted Rand indices are all very close to one indicating high similarities between $\mathcal{P}_0$, the true partition, and $\hat{\mathcal{P}}$, the estimated partition. In addition, larger values of γ seems to yield higher clustering quality, which corroborates Theorem 2.2 that a heavier penalization is needed when considering high-dimensional problems, as $p \in \{50, 100\}$ here. If $\gamma = 1$ is chosen, the corresponding clustering quality is not sensitive to the choice of bandwidth. The bandwidth plays an important role in parameter estimation because it should be appropriately chosen to reflect the degree of smoothness of the underlying function. However, in terms of clustering, we only need to determine whether the given functions belong to the same subgroup, and thus the result is less sensitive to the choice of bandwidth. For all the cases, medians of the selected

Table 2. Means and medians (in parentheses) of the adjusted Rand index for the information criterion (19) applied on model (M1) with different combinations of dimension $p \in \{50, 100\}$, noise-to-signal level $\sigma^2 \in \{1, 2\}$, level of smoothness $s \in \{0, 1\}$, bandwidth $b = 0.1 + 0.05l$, $l = 0, \dots, 5$, and penalization $\gamma \in \{0.25, 0.5, 1\}$. For each configuration, the results are based on 1000 simulated realizations

			b					
σ^2	s	γ	0.1	0.15	0.2	0.25	0.3	0.35
					$p = 50$			
1	0	0.25	$0.979_{(1.000)}$	$0.966_{(0.984)}$	$0.955_{(0.965)}$	$0.946_{(0.965)}$	$0.932_{(0.949)}$	$0.922_{(0.949)}$
		0.5	$0.987_{(1.000)}$	$0.983_{(1.000)}$	$0.981_{(1.000)}$	$0.974_{(0.984)}$	$0.968_{(0.984)}$	$0.963_{(0.984)}$
		1	$0.996_{(1.000)}$	$0.994_{(1.000)}$	$0.992_{(1.000)}$	$0.991_{(1.000)}$	$0.989_{(1.000)}$	$0.988_{(1.000)}$
	1	0.25	$0.981_{(1.000)}$	$0.971_{(0.984)}$	$0.960_{(0.965)}$	$0.954_{(0.965)}$	$0.946_{(0.965)}$	$0.944_{(0.965)}$
		0.5	$0.988_{(1.000)}$	$0.985_{(1.000)}$	$0.980_{(1.000)}$	$0.976_{(1.000)}$	$0.974_{(0.984)}$	$0.975_{(0.984)}$
		1	$0.996_{(1.000)}$	$0.994_{(1.000)}$	$0.993_{(1.000)}$	$0.993_{(1.000)}$	$0.992_{(1.000)}$	$0.992_{(1.000)}$
2	0	0.25	$0.976_{(0.984)}$	$0.968_{(0.984)}$	$0.956_{(0.965)}$	$0.943_{(0.965)}$	$0.933_{(0.949)}$	$0.919_{(0.934)}$
		0.5	$0.986_{(1.000)}$	$0.981_{(1.000)}$	$0.977_{(1.000)}$	$0.972_{(0.984)}$	$0.964_{(0.971)}$	$0.961_{(0.971)}$
		1	$0.994_{(1.000)}$	$0.992_{(1.000)}$	$0.991_{(1.000)}$	$0.990_{(1.000)}$	$0.986_{(1.000)}$	$0.986_{(1.000)}$
	1	0.25	$0.978_{(1.000)}$	$0.970_{(0.984)}$	$0.956_{(0.965)}$	$0.951_{(0.965)}$	$0.938_{(0.955)}$	$0.924_{(0.934)}$
		0.5	$0.987_{(1.000)}$	$0.985_{(1.000)}$	$0.978_{(1.000)}$	$0.973_{(0.984)}$	$0.968_{(0.971)}$	$0.960_{(0.965)}$
		1	$0.996_{(1.000)}$	$0.993_{(1.000)}$	$0.993_{(1.000)}$	$0.990_{(1.000)}$	$0.989_{(1.000)}$	$0.987_{(1.000)}$
					$p = 100$			
1	0	0.25	$0.982_{(0.984)}$	$0.975_{(0.983)}$	$0.966_{(0.976)}$	$0.960_{(0.973)}$	$0.946_{(0.960)}$	$0.937_{(0.955)}$
		0.5	$0.989_{(1.000)}$	$0.985_{(0.992)}$	$0.984_{(0.984)}$	$0.979_{(0.983)}$	$0.973_{(0.983)}$	$0.973_{(0.983)}$
		1	$0.995_{(1.000)}$	$0.994_{(1.000)}$	$0.993_{(1.000)}$	$0.991_{(1.000)}$	$0.990_{(1.000)}$	$0.990_{(1.000)}$
	1	0.25	$0.983_{(0.984)}$	$0.977_{(0.983)}$	$0.970_{(0.983)}$	$0.960_{(0.966)}$	$0.958_{(0.966)}$	$0.954_{(0.966)}$
		0.5	$0.991_{(1.000)}$	$0.988_{(1.000)}$	$0.985_{(0.992)}$	$0.981_{(0.983)}$	$0.978_{(0.983)}$	$0.976_{(0.983)}$
		1	$0.997_{(1.000)}$	$0.995_{(1.000)}$	$0.995_{(1.000)}$	$0.994_{(1.000)}$	$0.994_{(1.000)}$	$0.994_{(1.000)}$
2	0	0.25	$0.979_{(0.983)}$	$0.974_{(0.983)}$	$0.965_{(0.974)}$	$0.956_{(0.966)}$	$0.946_{(0.960)}$	$0.932_{(0.948)}$
		0.5	$0.987_{(0.992)}$	$0.984_{(0.992)}$	$0.981_{(0.983)}$	$0.977_{(0.983)}$	$0.974_{(0.983)}$	$0.968_{(0.977)}$
		1	$0.993_{(1.000)}$	$0.993_{(1.000)}$	$0.991_{(1.000)}$	$0.990_{(1.000)}$	$0.988_{(0.992)}$	$0.987_{(0.992)}$
	1	0.25	$0.983_{(0.984)}$	$0.976_{(0.983)}$	$0.968_{(0.983)}$	$0.959_{(0.966)}$	$0.948_{(0.957)}$	$0.936_{(0.948)}$
		0.5	$0.990_{(1.000)}$	$0.987_{(1.000)}$	$0.984_{(0.992)}$	$0.979_{(0.983)}$	$0.974_{(0.983)}$	$0.969_{(0.983)}$
		1	$0.996_{(1.000)}$	$0.995_{(1.000)}$	$0.994_{(1.000)}$	$0.993_{(1.000)}$	$0.992_{(1.000)}$	$0.991_{(1.000)}$

bandwidths based on the GCV criterion (18) range from 0.091 to 0.097.

4.3 A Real-Data Analysis

We shall here apply the proposed clustering method to the daily cell phone download data described in Section 1, for which $n = 327$ days, from July 9, 2005 to May 31, 2006, and $p = 129$ area codes in the United States. A logarithmic transformation is taken to turn the multiplicative effect caused by population into an additive effect. Degras et al. (2012) conducted a test on the parallelism hypothesis for the entire dataset, and rejected the null hypothesis at 1% significance level. The goal is to find homogeneous subgroups in terms of trend parallelism. In particular, we apply the information criterion (19) to estimate the underlying partition and the SC algorithm in Section 3.1 is used because of the high dimensionality. The bandwidth $\hat{b}_n = 0.081$ is selected by the GCV criterion (18), and $\gamma = 1$ is used due to its superior performance in our simulation studies on handling high-dimensional problems. The results are reported in Table 4 containing possible subgroup sizes and corresponding number of subgroups.

The largest four subgroups together constitute 74.4% of the entire 129 area codes. Degras et al. (2012) considered the same data and found homogeneous subgroups by means of hypothesis testings. There is a similarity between their results and ours. For example, the largest subgroup found by our approach, which we denote $\hat{\mathcal{S}}_1$, shares many common area codes with the largest one found by Degras et al. (2012), which we denote $\tilde{\mathcal{S}}_1$. However, they do differ from each other. Time series plots of their representatives are given in Figure 2(i) and 2(ii) with estimated common trends imposed. The two solid curves in Figure 2(i) are contained in $\hat{\mathcal{S}}_1$ but not in $\tilde{\mathcal{S}}_1$, and they seem to have trend functions that are parallel to the dashed curves that are contained in both $\hat{\mathcal{S}}_1$ and $\tilde{\mathcal{S}}_1$. This is supported by the current method. On the other hand, the two solid curves in Figure 2(ii) seem to have faster growth rates in the middle, while Degras et al. (2012) included them in the same subgroup as the dashed curves. Their inhomogeneity from the imposed estimated common trend can be identified. To have a close comparison between the two, let $\hat{\mu}_{\hat{\mathcal{S}}_1 \setminus \tilde{\mathcal{S}}_1}(t)$, $\hat{\mu}_{\hat{\mathcal{S}}_1 \cap \tilde{\mathcal{S}}_1}(t)$, and $\hat{\mu}_{\tilde{\mathcal{S}}_1 \setminus \hat{\mathcal{S}}_1}(t)$, $t \in [0, 1]$, be as in (7), namely, they are the estimated common trend functions of $\hat{\mathcal{S}}_1 \setminus \tilde{\mathcal{S}}_1$, $\hat{\mathcal{S}}_1 \cap \tilde{\mathcal{S}}_1$, and $\tilde{\mathcal{S}}_1 \setminus \hat{\mathcal{S}}_1$, respectively. If the parallelism holds for both $\hat{\mathcal{S}}_1$ and $\tilde{\mathcal{S}}_1$, then both $\hat{\mu}_{\hat{\mathcal{S}}_1 \setminus \tilde{\mathcal{S}}_1}(\cdot) - \hat{\mu}_{\hat{\mathcal{S}}_1 \cap \tilde{\mathcal{S}}_1}(\cdot)$ and $\hat{\mu}_{\tilde{\mathcal{S}}_1 \setminus \hat{\mathcal{S}}_1}(\cdot) - \hat{\mu}_{\hat{\mathcal{S}}_1 \cap \tilde{\mathcal{S}}_1}(\cdot)$ would be very close to some constants, namely, $d_{\hat{\mathcal{S}}_1 \setminus \tilde{\mathcal{S}}_1}(t) = \hat{\mu}_{\hat{\mathcal{S}}_1 \setminus \tilde{\mathcal{S}}_1}(t) - \hat{\mu}_{\hat{\mathcal{S}}_1 \cap \tilde{\mathcal{S}}_1}(t) - v_{\hat{\mathcal{S}}_1 \setminus \tilde{\mathcal{S}}_1}$ and $d_{\tilde{\mathcal{S}}_1 \setminus \hat{\mathcal{S}}_1}(t) = \hat{\mu}_{\tilde{\mathcal{S}}_1 \setminus \hat{\mathcal{S}}_1}(t) - \hat{\mu}_{\hat{\mathcal{S}}_1 \cap \tilde{\mathcal{S}}_1}(t) - v_{\tilde{\mathcal{S}}_1 \setminus \hat{\mathcal{S}}_1}$, $t \in [0, 1]$, would both be very close to zero, where $v_{\hat{\mathcal{S}}_1 \setminus \tilde{\mathcal{S}}_1} = n^{-1} \sum_{i=1}^{n} \{\hat{\mu}_{\hat{\mathcal{S}}_1 \setminus \tilde{\mathcal{S}}_1}(i/n) - \hat{\mu}_{\hat{\mathcal{S}}_1 \cap \tilde{\mathcal{S}}_1}(i/n)\}$ and $v_{\tilde{\mathcal{S}}_1 \setminus \hat{\mathcal{S}}_1} = n^{-1} \sum_{i=1}^{n} \{\hat{\mu}_{\tilde{\mathcal{S}}_1 \setminus \hat{\mathcal{S}}_1}(i/n) - \hat{\mu}_{\hat{\mathcal{S}}_1 \cap \tilde{\mathcal{S}}_1}(i/n)\}$. From Figure 2(iii), we can see that $d_{\hat{\mathcal{S}}_1 \setminus \tilde{\mathcal{S}}_1}(\cdot)$, represented by the dashed curve, is closer

Table 3. Means and medians (in parentheses) of the adjusted Rand index for the information criterion (19) applied on model (M2) with different combinations of dimension $p \in \{50, 100\}$, noise-to-signal level $\sigma^2 \in \{1, 2\}$, level of smoothness $s \in \{0, 1\}$, bandwidth $b = 0.1 + 0.05l$, $l = 0, \dots, 5$, and penalization $\gamma \in \{0.25, 0.5, 1\}$. For each configuration, the results are based on 1000 simulated realizations

			b					
σ^2	s	γ	0.1	0.15	0.2	0.25	0.3	0.35
					$p = 50$			
1	0	0.25	$0.981_{(1.000)}$	$0.971_{(0.982)}$	$0.962_{(0.972)}$	$0.953_{(0.968)}$	$0.946_{(0.961)}$	$0.938_{(0.957)}$
		0.5	$0.989_{(1.000)}$	$0.982_{(1.000)}$	$0.980_{(1.000)}$	$0.975_{(0.988)}$	$0.974_{(0.982)}$	$0.968_{(0.982)}$
		1	$0.996_{(1.000)}$	$0.994_{(1.000)}$	$0.993_{(1.000)}$	$0.992_{(1.000)}$	$0.991_{(1.000)}$	$0.990_{(1.000)}$
	1	0.25	$0.978_{(1.000)}$	$0.972_{(0.992)}$	$0.967_{(0.982)}$	$0.964_{(0.972)}$	$0.962_{(0.972)}$	$0.957_{(0.972)}$
		0.5	$0.989_{(1.000)}$	$0.985_{(1.000)}$	$0.984_{(1.000)}$	$0.982_{(1.000)}$	$0.983_{(1.000)}$	$0.981_{(1.000)}$
		1	$0.996_{(1.000)}$	$0.995_{(1.000)}$	$0.995_{(1.000)}$	$0.994_{(1.000)}$	$0.994_{(1.000)}$	$0.995_{(1.000)}$
2	0	0.25	$0.978_{(0.992)}$	$0.969_{(0.982)}$	$0.957_{(0.972)}$	$0.952_{(0.968)}$	$0.936_{(0.953)}$	$0.924_{(0.949)}$
		0.5	$0.988_{(1.000)}$	$0.984_{(1.000)}$	$0.977_{(1.000)}$	$0.975_{(0.988)}$	$0.971_{(0.982)}$	$0.962_{(0.972)}$
		1	$0.996_{(1.000)}$	$0.994_{(1.000)}$	$0.994_{(1.000)}$	$0.991_{(1.000)}$	$0.990_{(1.000)}$	$0.988_{(1.000)}$
	1	0.25	$0.978_{(1.000)}$	$0.970_{(0.982)}$	$0.961_{(0.972)}$	$0.956_{(0.972)}$	$0.946_{(0.963)}$	$0.941_{(0.958)}$
		0.5	$0.988_{(1.000)}$	$0.984_{(1.000)}$	$0.980_{(1.000)}$	$0.979_{(0.992)}$	$0.971_{(0.982)}$	$0.972_{(0.982)}$
		1	$0.996_{(1.000)}$	$0.994_{(1.000)}$	$0.993_{(1.000)}$	$0.993_{(1.000)}$	$0.990_{(1.000)}$	$0.990_{(1.000)}$
					$p = 100$			
1	0	0.25	$0.983_{(0.991)}$	$0.977_{(0.986)}$	$0.970_{(0.981)}$	$0.965_{(0.974)}$	$0.958_{(0.968)}$	$0.951_{(0.963)}$
		0.5	$0.991_{(1.000)}$	$0.988_{(0.996)}$	$0.985_{(0.991)}$	$0.983_{(0.991)}$	$0.981_{(0.986)}$	$0.979_{(0.986)}$
		1	$0.997_{(1.000)}$	$0.996_{(1.000)}$	$0.995_{(1.000)}$	$0.995_{(1.000)}$	$0.994_{(1.000)}$	$0.993_{(1.000)}$
	1	0.25	$0.984_{(0.991)}$	$0.980_{(0.986)}$	$0.976_{(0.982)}$	$0.972_{(0.977)}$	$0.969_{(0.976)}$	$0.968_{(0.974)}$
		0.5	$0.992_{(1.000)}$	$0.989_{(0.996)}$	$0.987_{(0.996)}$	$0.985_{(0.991)}$	$0.984_{(0.991)}$	$0.984_{(0.991)}$
		1	$0.997_{(1.000)}$	$0.996_{(1.000)}$	$0.996_{(1.000)}$	$0.995_{(1.000)}$	$0.996_{(1.000)}$	$0.996_{(1.000)}$
2	0	0.25	$0.984_{(0.991)}$	$0.978_{(0.986)}$	$0.969_{(0.981)}$	$0.961_{(0.972)}$	$0.952_{(0.963)}$	$0.942_{(0.958)}$
		0.5	$0.991_{(1.000)}$	$0.988_{(0.996)}$	$0.986_{(0.991)}$	$0.982_{(0.991)}$	$0.977_{(0.986)}$	$0.973_{(0.982)}$
		1	$0.997_{(1.000)}$	$0.996_{(1.000)}$	$0.995_{(1.000)}$	$0.994_{(1.000)}$	$0.993_{(1.000)}$	$0.991_{(1.000)}$
	1	0.25	$0.986_{(0.991)}$	$0.978_{(0.986)}$	$0.972_{(0.981)}$	$0.964_{(0.972)}$	$0.958_{(0.968)}$	$0.951_{(0.962)}$
		0.5	$0.992_{(1.000)}$	$0.988_{(0.996)}$	$0.986_{(0.991)}$	$0.982_{(0.986)}$	$0.980_{(0.986)}$	$0.977_{(0.982)}$
		1	$0.997_{(1.000)}$	$0.996_{(1.000)}$	$0.995_{(1.000)}$	$0.995_{(1.000)}$	$0.993_{(1.000)}$	$0.993_{(1.000)}$

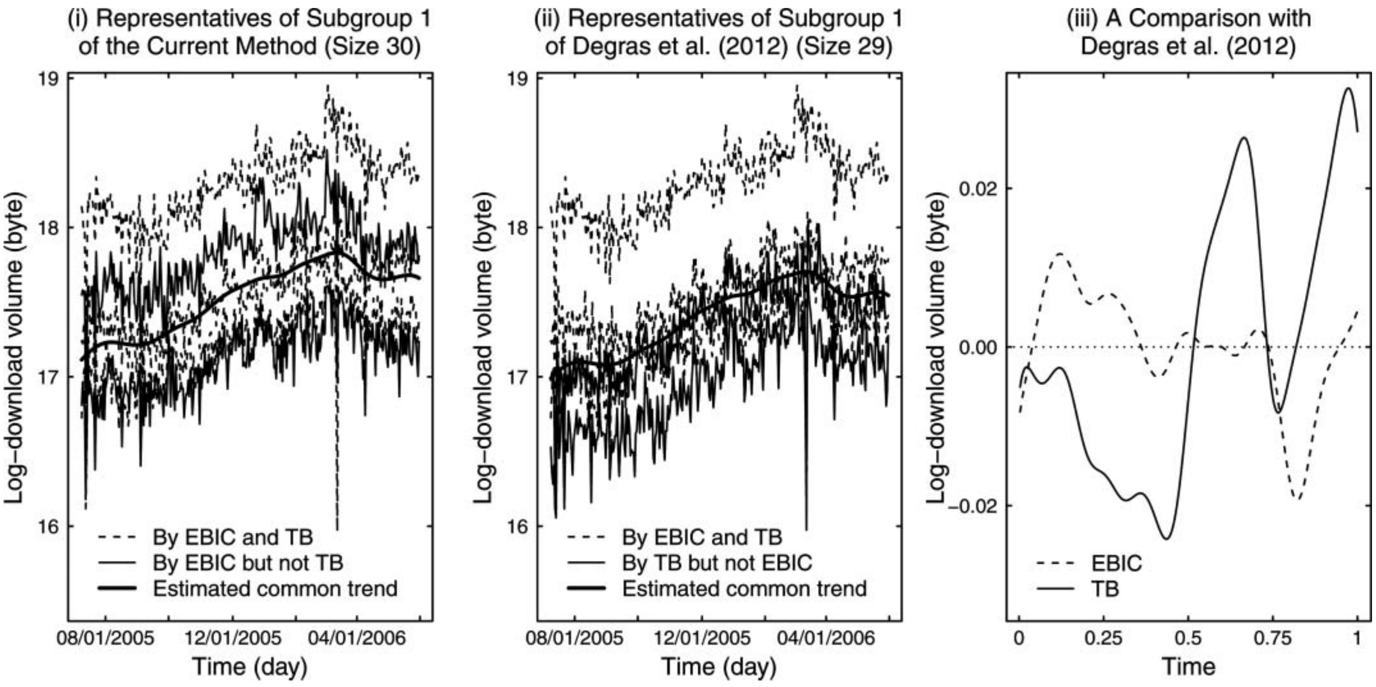

Figure 2. Time series plots of representatives from the largest estimated subgroups by using (i) the proposed EBIC, and (ii) the TB method of Degras et al. (2012). In both plots, the dashed curves are shared by the two methods while the solid curves represent the difference. The two thick curves represent the estimated common trends of $\hat{\mathcal{S}}_1$ and $\tilde{\mathcal{S}}_1$, respectively. In (iii), the dashed and solid curves represent $d_{\hat{\mathcal{S}}_1 \setminus \tilde{\mathcal{S}}_1}(\cdot)$ and $d_{\tilde{\mathcal{S}}_1 \setminus \hat{\mathcal{S}}_1}(\cdot)$, respectively.

Table 4. Possible subgroup sizes of the estimated partition $\hat{\mathcal{P}}$, and the corresponding number of subgroups with that particular size

Subgroup size	30	26	20	7	5	2	1
Number of subgroups with that size	1	1	2	1	2	5	6

to zero than $d_{\tilde{\mathcal{S}}_1\backslash\hat{\mathcal{S}}_1}(\cdot)$, represented by the solid curve. Therefore, the time series in $\hat{\mathcal{S}}_1$ are more homogeneous than those in $\tilde{\mathcal{S}}_1$ based on trend parallelism, and the current method seems to yield more reasonable cluster patterns than the one by Degras et al. (2012).

5. CONCLUSION

This article considers the problem of clustering functional data contaminated by locally stationary errors based on trend parallelism. The dimension of the observed data can either be fixed or grow to infinity with the sample size. Existing results on high-dimensional data are usually confined to parametric inference, while the estimation problem under the parallelism (2) is semiparametric (see Section 2.2). I shall here discuss some future topics to conclude the article. As discussed in Section 3.2, the same bandwidth is used for constructing the semiparametric estimates and developing the clustering algorithm. This is helpful in correcting the bias for semiparametric estimation, reducing the computational burden, and obtaining rigorous theoretical results. While in practice, different bandwidths can be used for different time series to obtain their local linear estimates, and the methodology developed in the current article can still be applied. Nevertheless, including p data-driven bandwidths can make the theoretical property of the resulting estimates very hard to understand. In addition, bandwidth selectors developed under independence can break down very often for dependent data as demonstrated in Opsomer, Wang, and Yang (2001). Furthermore, traditional bandwidth selector developed for estimation and hypothesis testing problems might not be advantageous in clustering which is quite different. The problem of finding optimal bandwidths and tuning parameters for clustering and their theoretical justifications can be an interesting topic for open discussion.

APPENDIX: TECHNICAL PROOFS

Proof. (Theorem 2.1) By (2) and (3), we have

$$\hat{\mu}_{\mathcal{S}_q}(t) = \sum_{i=1}^{n} \mu_{\mathcal{S}_q}(i/n) w_i(t) + \sum_{i=1}^{n} \left(|\mathcal{S}_q|^{-1} \sum_{k\in\mathcal{S}_q} e_{k,i} \right) w_i(t).$$

Let $\xi_{\mathcal{S}_q,i} = |\mathcal{S}_q|^{-1/2} \sum_{k\in\mathcal{S}_q} e_{k,i}$, $i = 1,\ldots,n$. Since the process $(\xi_{\mathcal{S}_q,i})_{i=1}^n$ has functional dependence measure

$$\sup_{t\in[0,1]} \left\| |\mathcal{S}_q|^{-1/2} \sum_{k\in\mathcal{S}_q} G(t;\mathcal{F}_{k,i}) - |\mathcal{S}_q|^{-1/2} \sum_{k\in\mathcal{S}_q} G(t;\mathcal{F}^{\star}_{k,i}) \right\| \le \theta_{i,2},$$

and long-run variance function $g(t)$, $t \in [0,1]$, (9) follows by the proof of Theorem 1 in Zhou and Wu (2010). For (10), by (2) and elementary calculation, we have

$$\begin{aligned}\hat{c}_k &= n^{-1} \sum_{i=1}^{n}\sum_{j=1}^{n} \left(y_{k,j} - |\mathcal{S}_q|^{-1} \sum_{l\in\mathcal{S}_q} y_{l,j} \right) w_j(i/n) \\ &= c_k + n^{-1} \sum_{j=1}^{n} \left\{ (1 - |\mathcal{S}_q|^{-1}) e_{k,j} - |\mathcal{S}_q|^{-1} \sum_{l\in\mathcal{S}_q\backslash\{k\}} e_{l,j} \right\}.\end{aligned}$$

Let $\zeta_{k,j} = e_{k,j} - |\mathcal{S}_q|^{-1} \sum_{l\in\mathcal{S}_q} e_{l,j}$, $j = 1,\ldots,n$. Then the process $(\zeta_{k,j})_{j=1}^n$ has functional dependence measure

$$\begin{aligned}&\sup_{t\in[0,1]} \Bigg\| \left\{ G(t;\mathcal{F}_{k,j}) - |\mathcal{S}_q|^{-1} \sum_{l\in\mathcal{S}_q} G(t;\mathcal{F}_{l,j}) \right\} \\ &\quad - \left\{ G(t;\mathcal{F}^{\star}_{k,j}) - |\mathcal{S}_q|^{-1} \sum_{l\in\mathcal{S}_q} G(t;\mathcal{F}^{\star}_{l,j}) \right\} \Bigg\| \\ &\le \left\{ (1 - |\mathcal{S}_q|^{-1})^2 \theta_{j,2}^2 + |\mathcal{S}_q|^{-2}(|\mathcal{S}_q| - 1)\theta_{j,2}^2 \right\}^{1/2} \le \theta_{j,2},\end{aligned}$$

and long-run variance function

$$\sum_{j=1}^{n} \{(1 - |\mathcal{S}_q|^{-1})^2 + |\mathcal{S}_q|^{-2}(|\mathcal{S}_q| - 1)\} g(t) = |\mathcal{S}_q|^{-1}(|\mathcal{S}_q| - 1) g(t), \qquad t \in [0,1].$$

By Lemma A1 in Zhang and Wu (2011) and the m-dependence approximation, we obtain the asymptotic normality (10). □

The following lemma provides approximations of the residual sum of squares, and is useful in proving Theorem 2.2. In the proof C denotes constants whose value may vary from place to place.

Lemma A.1. Assume (A1)–(A3), $\Theta_{0,4} < \infty$, $b_n \to 0$, and $nb_n^{3/2} \to \infty$. Let $\psi_n = b_n^{-1} + nb_n^4$, then (i) there exists a constant c_1 such that for any $\mathcal{P} \in \mathscr{P}$,

$$\begin{aligned}&\left\| \mathrm{RSS}(\mathcal{P}) - \sum_{k=1}^{p}\sum_{i=1}^{n} e_{k,i}^2 - \Delta(\mathcal{P}) \right\| \\ &\quad \le c_1 \left\{ b_n^{-1}|\mathcal{P}| + nb_n^4 p + p + \left(nb_n^4 p\right)^{1/2} + (np)^{1/2} \mathbb{1}_{\{\mathcal{P}\not\preceq\mathcal{P}_0\}} \right\};\end{aligned}$$

(ii)

$$\max_{\mathcal{P}\not\preceq\mathcal{P}_0} \left| \mathrm{RSS}(\mathcal{P}) - \sum_{k=1}^{p}\sum_{i=1}^{n} e_{k,i}^2 - \Delta(\mathcal{P}) \right| = O_p\left\{ \psi_n p + \psi_n^{1/2} p^2 + n^{1/2} p^2 \right\};$$

and (iii)

$$\max_{\mathcal{P}\preceq\mathcal{P}_0} \left| \mathrm{RSS}(\mathcal{P}) - \sum_{k=1}^{p}\sum_{i=1}^{n} e_{k,i}^2 \right| = O_p\left\{ \psi_n p + \psi_n^{1/2} p^2 \right\}.$$

Proof. With a little abuse of notation, let $\mu_{\mathcal{S}_q}(t) = |\mathcal{S}_q|^{-1} \sum_{k\in\mathcal{S}_q} \mu_k(t)$, $t \in [0,1]$, and $c_k = n^{-1} \sum_{i=1}^{n} \{\mu_k(i/n) - \mu_{\mathcal{S}_q}(i/n)\}$, $k \in \mathcal{S}_q$, then a careful check of the proof would reveal that Theorem 2.1 holds regardless of the parallelism (2). Recall that $\xi_{\mathcal{S}_q,i} = |\mathcal{S}_q|^{-1/2} \sum_{k\in\mathcal{S}_q} e_{k,i}$, $i = 1,\ldots,n$. By the proof of Theorem 2.1 and Lemmas A.1 and A.2. in Zhang and Wu (2012),

$$\begin{aligned}\|\mathrm{I}_n\| &= \left\| \sum_{q=1}^{Q}\sum_{k\in\mathcal{S}_q}\sum_{i=1}^{n} \{\mu_{\mathcal{S}_q}(i/n) - \hat{\mu}_{\mathcal{S}_q}(i/n)\}^2 \right\| \\ &\le 2 \sum_{q=1}^{Q}\sum_{k\in\mathcal{S}_q}\sum_{i=1}^{n} [\mu_{\mathcal{S}_q}(i/n) - E\{\hat{\mu}_{\mathcal{S}_q}(i/n)\}]^2 \\ &\quad + 2 \left\| \sum_{q=1}^{Q}\sum_{i=1}^{n} \left\{ \sum_{j=1}^{n} \xi_{\mathcal{S}_q,j} w_j(i/n) \right\}^2 \right\| \\ &\le C\left(nb_n^4 p + b_n^{-1}|\mathcal{P}| \right),\end{aligned}$$

and, by Schwarz's inequality,

$$\max_{\mathcal{P}\in\mathscr{P}} |\mathrm{I}_n| \le 2 \max_{\mathcal{P}\in\mathscr{P}} \sum_{q=1}^{Q}\sum_{i=1}^{n}\left\{|\mathcal{S}_q|^{-1/2}\sum_{l\in\mathcal{S}_q}\sum_{j=1}^{n} e_{l,j}w_j(i/n)\right\}^2 + O\left(nb_n^4 p\right) \le 2\sum_{q=1}^{Q}\sum_{l\in\mathcal{S}_q}\sum_{i=1}^{n}\left\{\sum_{j=1}^{n} e_{l,j}w_j(i/n)\right\}^2 + O\left(nb_n^4 p\right) = O_p\left(b_n^{-1}p + nb_n^4 p\right).$$

Similarly,

$$\|\mathrm{II}_n\| = \left\|\sum_{q=1}^{Q}\sum_{k\in\mathcal{S}_q}\sum_{i=1}^{n}(c_k-\hat{c}_k)^2\right\| \le 2\left\|\sum_{q=1}^{Q}\sum_{k\in\mathcal{S}_q}\left(n^{-1/2}\sum_{j=1}^{n}e_{k,j}\right)^2\right\| + 2\left\|\sum_{q=1}^{Q}\left(n^{-1/2}\sum_{j=1}^{n}\xi_{\mathcal{S}_q,j}\right)^2\right\| \le Cp,$$

and

$$\max_{\mathcal{P}\in\mathscr{P}} |\mathrm{II}_n| \le 2\max_{\mathcal{P}\in\mathscr{P}}\sum_{q=1}^{Q}\sum_{k\in\mathcal{S}_q}\sum_{i=1}^{n}\left\{\left(n^{-1}\sum_{j=1}^{n}e_{k,j}\right)^2 + \left(n^{-1}|\mathcal{S}_q|^{-1}\sum_{l\in\mathcal{S}_q}\sum_{j=1}^{n}e_{l,j}\right)^2\right\} \le 2\sum_{k=1}^{p}\sum_{i=1}^{n}\left(n^{-1}\sum_{j=1}^{n}e_{k,j}\right)^2 + 2\sum_{l=1}^{p}\sum_{i=1}^{n}\left(n^{-1}\sum_{j=1}^{n}e_{l,j}\right)^2 = O_p(p).$$

Since $\mu_k(i/n) - \mu_{\mathcal{S}_q}(i/n) - c_k = \mu_k(i/n) - \mu_{\mathcal{S}_q}(i/n) - n^{-1}\sum_{j=1}^{n}\{\mu_k(j/n) - \mu_{\mathcal{S}_q}(j/n)\}$, by Lemma A1 in Zhang and Wu (2011),

$$\|\mathrm{III}_n\| = \left\|\sum_{q=1}^{Q}\sum_{k\in\mathcal{S}_q}\sum_{i=1}^{n}\{\mu_k(i/n)-\mu_{\mathcal{S}_q}(i/n)-c_k\}e_{k,i}\right\| \le C(np)^{1/2},$$

and

$$\max_{\mathcal{P}\in\mathscr{P}}|\mathrm{III}_n| \le \left|\sum_{k=1}^{p}\sum_{i=1}^{n}\left\{\mu_k(i/n) - n^{-1}\sum_{j=1}^{n}\mu_k(j/n)\right\}e_{k,i}\right| + \max_{\mathcal{P}\in\mathscr{P}}\left|\sum_{q=1}^{Q}\sum_{k\in\mathcal{S}_q}\sum_{i=1}^{n}|\mathcal{S}_q|^{-1}\sum_{l\in\mathcal{S}_q}\left\{\mu_l(i/n) - n^{-1}\sum_{j=1}^{n}\mu_l(j/n)\right\}e_{k,i}\right| \le \sum_{k=1}^{p}\sum_{l=1}^{p}\left|\sum_{i=1}^{n}\left\{\mu_l(i/n) - n^{-1}\sum_{j=1}^{n}\mu_l(j/n)\right\}e_{k,i}\right| + O_p\{(np)^{1/2}\} = O_p\left(n^{1/2}p^2\right).$$

Since $\sum_{k\in\mathcal{S}_q}\{\mu_k(t)-\mu_{\mathcal{S}_q}(t)-c_k\}=0,\ t\in[0,1],\ \sum_{i=1}^{n}\{\mu_k(i/n) - \mu_{\mathcal{S}_q}(i/n) - c_k\} = 0, k\in\mathcal{S}_q$, and $\sum_{k\in\mathcal{S}_q}(c_k-\hat{c}_k)=0$, we have

$$\mathrm{IV}_n = \sum_{q=1}^{Q}\sum_{k\in\mathcal{S}_q}\sum_{i=1}^{n}\{\mu_k(i/n)-\mu_{\mathcal{S}_q}(i/n)-c_k\} \times [\{\mu_{\mathcal{S}_q}(i/n)-\hat{\mu}_{\mathcal{S}_q}(i/n)\} + (c_k-\hat{c}_k)] = 0,$$

and

$$\mathrm{V}_n = \sum_{q=1}^{Q}\sum_{k\in\mathcal{S}_q}\sum_{i=1}^{n}\{\mu_{\mathcal{S}_q}(i/n)-\hat{\mu}_{\mathcal{S}_q}(i/n)\}(c_k-\hat{c}_k) = 0.$$

By the proof of Theorem 1 in Zhang and Wu (2011), we have

$$\|\mathrm{VI}_n\| = \left\|\sum_{q=1}^{Q}\sum_{k\in\mathcal{S}_q}\sum_{i=1}^{n}\{\mu_{\mathcal{S}_q}(i/n)-\hat{\mu}_{\mathcal{S}_q}(i/n)\}e_{k,i}\right\| \le 2\left\|\sum_{q=1}^{Q}\sum_{k\in\mathcal{S}_q}\sum_{i=1}^{n}[\mu_{\mathcal{S}_q}(i/n)-E\{\hat{\mu}_{\mathcal{S}_q}(i/n)\}]e_{k,i}\right\| + 2\left\|\sum_{q=1}^{Q}\sum_{i=1}^{n}\sum_{j=1}^{n}\xi_{\mathcal{S}_q,j}\xi_{\mathcal{S}_q,i}w_j(i/n)\right\| \le C\{(nb_n^4p)^{1/2} + b_n^{-1}|\mathcal{P}|\},$$

and

$$\max_{\mathcal{P}\in\mathscr{P}}|\mathrm{VI}_n| \le \sum_{k=1}^{p}\sum_{l=1}^{p}\left|\sum_{i=1}^{n}\left\{\mu_l(i/n) - \sum_{j=1}^{n}\mu_l(j/n)w_j(i/n) - \sum_{j=1}^{n}e_{l,j}w_j(i/n)\right\}e_{k,i}\right| \le \sum_{k=1}^{p}\left(\sum_{l=k}+\sum_{l\ne k}\right)\left|\sum_{i=1}^{n}\sum_{j=1}^{n}e_{l,j}w_j(i/n)e_{k,i}\right| + O_p\left(n^{1/2}b_n^2p^2\right) = O_p\left(b_n^{-1}p + b_n^{-1/2}p^2 + n^{1/2}b_n^2p^2\right).$$

Similarly,

$$\|\mathrm{VII}_n\| = \left\|\sum_{q=1}^{Q}\sum_{k\in\mathcal{S}_q}\sum_{i=1}^{n}(c_k-\hat{c}_k)e_{k,i}\right\| = n^{-1}\left\|\sum_{q=1}^{Q}\sum_{k\in\mathcal{S}_q}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(e_{k,j} - |\mathcal{S}_q|^{-1}\sum_{l\in\mathcal{S}_q}e_{l,j}\right)e_{k,i}\right\| \le Cp,$$

and

$$\max_{\mathcal{P}\in\mathscr{P}}|\mathrm{VII}_n| \le n^{-1}\left|\sum_{k=1}^{p}\sum_{i=1}^{n}\sum_{j=1}^{n}e_{k,j}e_{k,i}\right| + n^{-1}\sum_{k=1}^{p}\sum_{l=1}^{p}\left|\sum_{i=1}^{n}\sum_{j=1}^{n}e_{l,j}e_{k,i}\right| = O_p(p^2).$$

Note that (13) has the close form solution

$$\Delta(\mathcal{S}_q) = \sum_{k\in\mathcal{S}_q}\sum_{i=1}^{n}\{\mu_k(i/n)-\mu_{\mathcal{S}_q}(i/n)-c_k\}^2,$$

we have

$$\mathrm{RSS}(\mathcal{P}) = \sum_{k=1}^{p}\sum_{i=1}^{n}e_{k,i}^2 + \Delta(\mathcal{P}) + \mathrm{I}_n + \mathrm{II}_n + 2(\mathrm{III}_n + \mathrm{IV}_n + \mathrm{V}_n + \mathrm{VI}_n + \mathrm{VII}_n).$$

Since $\mathrm{III}_n = \mathrm{IV}_n = 0$ if $\mathcal{P}\preceq\mathcal{P}_0$, Lemma A.1 follows. □

Proof. (Theorem 2.2) If $\mathcal{P} \npreceq \mathcal{P}_0$, then by assumptions (A3) and (A4),

$$\log\left\{\sum_{k=1}^{p}\sum_{i=1}^{n} E\left(e_{k,i}^2\right) + \frac{3\Delta(\mathcal{P})}{5}\right\} - \log\left\{\sum_{k=1}^{p}\sum_{i=1}^{n} E\left(e_{k,i}^2\right) + \frac{2\Delta(\mathcal{P})}{5}\right\} + \frac{\tau_n(|\mathcal{P}| - |\mathcal{P}_0|)}{np} > 0 \tag{21}$$

for all large n. For any two real numbers a and b, we denote $a \vee b = \max(a, b)$. By Lemma A.1(i), assumption (A4), and the Markov inequality,

$$\mathrm{pr}\left\{\left|\mathrm{RSS}(\mathcal{P}_0) - \sum_{k=1}^{p}\sum_{i=1}^{n} e_{k,i}^2\right| \vee \left|\mathrm{RSS}(\mathcal{P}) - \sum_{k=1}^{p}\sum_{i=1}^{n} e_{k,i}^2 - \Delta(\mathcal{P})\right| > \frac{\Delta(\mathcal{P})}{5}\right\} = O(p/n). \tag{22}$$

Similarly, by Lemma A1 in Zhang and Wu (2011),

$$\mathrm{pr}\left[\left|\sum_{k=1}^{p}\sum_{i=1}^{n}\left\{e_{k,i}^2 - E\left(e_{k,i}^2\right)\right\}\right| > \frac{\Delta(\mathcal{P})}{5}\right] = O(p/n). \tag{23}$$

Since $\Delta(\mathcal{P}) \geq 0$, (15) follows by Equations (21)–(23). If $\mathcal{P} \preceq \mathcal{P}_0$ and $\mathcal{P} \neq \mathcal{P}_0$, then $\Delta(\mathcal{P}) = 0$ and $|\mathcal{P}| > |\mathcal{P}_0|$ by the definition of $\mathcal{P}_0$. Let $\varepsilon > 0$ be as in assumption (A3) and A denote the event that

$$\max\left[\left|\sum_{k=1}^{p}\sum_{i=1}^{n}\left\{e_{k,i}^2 - E\left(e_{k,i}^2\right)\right\}\right|, \left|\mathrm{RSS}(\mathcal{P}) - \sum_{k=1}^{p}\sum_{i=1}^{n} e_{k,i}^2\right|, \left|\mathrm{RSS}(\mathcal{P}_0) - \sum_{k=1}^{p}\sum_{i=1}^{n} e_{k,i}^2\right|\right] \leq \frac{np\varepsilon}{4}.$$

By Lemma A.1(i), Lemma A1 in Zhang and Wu (2011) and the Markov inequality, $\mathrm{pr}(A^c) = O\{(np)^{-1}\}$, where A^c is the complement of A. Under the event A, by condition (A3),

$$\left|\frac{\mathrm{RSS}(\mathcal{P}) - \sum_{k=1}^{p}\sum_{i=1}^{n} e_{k,i}^2}{\sum_{k=1}^{p}\sum_{i=1}^{n} e_{k,i}^2}\right| \leq \frac{1}{3}, \quad \left|\frac{\mathrm{RSS}(\mathcal{P}_0) - \sum_{k=1}^{p}\sum_{i=1}^{n} e_{k,i}^2}{\sum_{k=1}^{p}\sum_{i=1}^{n} e_{k,i}^2}\right| \leq \frac{1}{3},$$

and by the concavity of the logarithmic function,

$$|\log\{\mathrm{RSS}(\mathcal{P})\} - \log\{\mathrm{RSS}(\mathcal{P}_0)\}| \leq \frac{3}{2}\frac{|\mathrm{RSS}(\mathcal{P}) - \mathrm{RSS}(\mathcal{P}_0)|}{\sum_{k=1}^{p}\sum_{i=1}^{n} e_{k,i}^2} \leq 2\frac{|\mathrm{RSS}(\mathcal{P}) - \mathrm{RSS}(\mathcal{P}_0)|}{np\varepsilon}.$$

Since

$$\mathrm{pr}\{\mathrm{EBIC}(\mathcal{P}) < \mathrm{EBIC}(\mathcal{P}_0)\} \leq \mathrm{pr}[\{\mathrm{EBIC}(\mathcal{P}) - \mathrm{EBIC}(\mathcal{P}_0) < 0\} \cap A] + \mathrm{pr}(A^c),$$

(15) follows by Lemma A.1(i) and another application of the Markov inequality. We shall now prove (16). By Lemma A.1(iii),

$$\max_{\mathcal{P} \preceq \mathcal{P}_0}\left|\mathrm{RSS}(\mathcal{P}) - \sum_{k=1}^{p}\sum_{i=1}^{n} e_{k,i}^2\right| = O_p\left\{b_n^{-1}p + b_n^{-1/2}p^2\right\} = o_p(\tau_n).$$

Therefore, with probability tending to 1,

$$\inf_{\mathcal{P} \preceq \mathcal{P}_0,\ |\mathcal{P}| > |\mathcal{P}_0|}\{\mathrm{EBIC}(\mathcal{P}) - \mathrm{EBIC}(\mathcal{P}_0)\} = \inf_{\mathcal{P} \preceq \mathcal{P}_0,\ |\mathcal{P}| > |\mathcal{P}_0|}\{\tau_n(|\mathcal{P}| - |\mathcal{P}_0|)\} + o_p(\tau_n) > 0.$$

It remains to deal with the cases that $\mathcal{P} \npreceq \mathcal{P}_0$. By assumption (A4),

$$\text{VIII}_n = \inf_{\mathcal{P} \npreceq \mathcal{P}_0}\left\{(np)\log\frac{\sum_{k=1}^{p}\sum_{i=1}^{n} E\left(e_{k,i}^2\right) + \Delta(\mathcal{P})}{\sum_{k=1}^{p}\sum_{i=1}^{n} E\left(e_{k,i}^2\right)} + \tau_n(|\mathcal{P}| - |\mathcal{P}_0|)\right\} > 0$$

for all large n. By Lemma A.1(ii),

$$\max_{\mathcal{P} \npreceq \mathcal{P}_0}\left|\mathrm{RSS}(\mathcal{P}) - \sum_{k=1}^{p}\sum_{i=1}^{n} e_{k,i}^2 - \Delta(\mathcal{P})\right| = O_p\left\{b_n^{-1}p + b_n^{-1/2}p^2 + n^{1/2}p^2\right\} = o_p(n).$$

Since $p = o(n)$, by Lemma A1 in Zhang and Wu (2011), $\sum_{k=1}^{p}\sum_{i=1}^{n}\{e_{k,i}^2 - E(e_{k,i}^2)\} = o_p(n)$. Therefore, with probability tending to 1,

$$\inf_{\mathcal{P} \npreceq \mathcal{P}_0}\{\mathrm{EBIC}(\mathcal{P}) - \mathrm{EBIC}(\mathcal{P}_0)\} = \text{VIII}_n + o_p(n) > 0.$$

Hence, with probability tending to 1, the minimizer $\hat{\mathcal{P}} = \mathrm{argmin}_{\mathcal{P}}\,\mathrm{EBIC}(\mathcal{P})$ satisfies $\hat{\mathcal{P}} \preceq \mathcal{P}_0$ and $|\hat{\mathcal{P}}| \leq |\mathcal{P}_0|$. Theorem 2.2 follows by the definition of $\mathcal{P}_0$. □

[Received March 2012. Revised December 2012.]

REFERENCES

Abraham, C., Cornillon, P. A., Matzner-Løber, E., and Molinari, N. (2003), "Unsupervised Curve Clustering Using B-Splines," *Scandinavian Journal of Statistics*, 30, 581–595. [577]

Antoniadis, A., Bigot, J., and von Sachs, R. (2009), "A Multiscale Approach for Statistical Characterization of Functional Images," *Journal of Computational and Graphical Statistics*, 18, 216–237. [577]

Bickel, P. J., and Levina, E. (2008), "Regularized Estimation of Large Covariance Matrices," *The Annals of Statistics*, 36, 199–227. [581]

Cai, Z. (2007), "Trending Time-Varying Coefficient Time Series Models With Serially Correlated Errors," *Journal of Econometrics*, 136, 163–188. [578]

Chen, J., and Chen, Z. (2008), "Extended Bayesian Information Criteria for Model Selection With Large Model Spaces," *Biometrika*, 95, 759–771. [580,581]

Chiou, J.-M., and Li, P.-L. (2007), "Functional Clustering and Identifying Substructures of Longitudinal Data," *Journal of The Royal Statistical Society,* Series B, 69, 679–699. [577]

——— (2008), "Correlation-Based Functional Clustering via Subspace Projection," *Journal of the American Statistical Association*, 103, 1684–1692. [577,578,581]

Craven, P., and Wahba, G. (1979), "Smoothing Noisy Data With Spline Functions," *Numerical Mathematics*, 31, 377–403. [580]

Dahlhaus, R. (1996), "On the Kullback-Leibler Information Divergence of Locally Stationary Processes," *Stochastic Processes and Their Applications*, 62, 139–168. [578]

——— (1997), "Fitting Time Series Models to Nonstationary Processes," *The Annals of Statistics*, 25, 1–37. [578]

Degras, D., Xu, Z., Zhang, T., and Wu, W. B. (2012), "Testing for Parallelism Among Trends in Multiple Time Series," *IEEE Transactions on Signal Processing*, 60, 1087–1097. [577,581,583]

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of The Royal Statistical Society,* Series B, 39, 1–38. [577]

Draghicescu, D., Guillas, S., and Wu, W. B. (2009), "Quantile Curve Estimation and Visualization for Non-Stationary Time Series," *Journal of Computational and Graphical Statistics*, 18, 1–20. [578]

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modeling and Its Applications*, London: Chapman & Hall. [579]

Fan, Y., and Linton, O. (2003), "Some Higher-Order Theory for a Consistent Non-Parametric Model Specification Test," *Journal of Statistical Planning and Inference*, 109, 125–154. [580]

García-Escudero, L. A., and Gordaliza, A. (2005), "A Proposal for Robust Curve Clustering," *Journal of Classification*, 22, 185–201. [577]

Hall, P., Lee, Y. K., and Park, B. U. (2007), "A Method for Projecting Functional Data Onto a Low-Dimensional Space," *Journal of Computational and Graphical Statistics*, 16, 799–812. [577]

Härdle, W., Hall, P., and Marron, J. S. (1988), "How Far are Automatically Chosen Regression Smoothing Parameters From Their Optimum?" *Journal of the American Statistical Association*, 83, 86–95. [580]

Härdle, W., and Marron, J. S. (1985), "Optimal Bandwidth Selection in Nonparametric Regression Function Estimation," *The Annals of Statistics*, 13, 1465–1481. [580]

Heckman, N. E., and Zamar, R. H. (2000), "Comparing the Shapes of Regression Functions," *Biometrika*, 87, 135–144. [577,579,581]

Hitchcock, D. B., Casella, G., and Booth, J. G. (2006), "Improved Estimation of Dissimilarities by Presmoothing Functional Data," *Journal of the American Statistical Association*, 101, 211–222. [577]

Hubert, L., and Arabie, P. (1985), "Comparing Partitions," *Journal of Classification*, 2, 193–218. [581,582]

Hurvich, C. M., Simonoff, J. S., and Tsai, C. L. (1998), "Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion," *Journal of The Royal Statistical Society,* Series B, 60, 271–293. [580]

Kulasekera, K. B., and Wang, J. (1997), "Smoothing Parameter Selection for Power Optimality in Testing of Regression Curves," *Journal of the American Statistical Association*, 92, 500–511. [580]

MacQueen, J. (1967), "Some Methods for Classification and Analysis of Multivariate Observations," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281–297. [577]

Mallat, S., Papanicolaou, G., and Zhang, Z. (1998), "Adaptive Covariance Estimation of Locally Stationary Processes," *The Annals of Statistics*, 26, 1–47. [578]

Nason, G. P., von Sachs, R., and Kroisandt, G. (2000), "Wavelet Processes and Adaptive Estimation of the Evolutionary Wavelet Spectrum," *Journal of The Royal Statistical Society,* Series B, 62, 271–292. [578]

Ombao, H., von Sachs, R., and Guo, W. (2005), "SLEX Analysis of Multivariate Nonstationary Time Series," *Journal of the American Statistical Association*, 100, 519–531. [578]

Opsomer, J., Wang, Y. D., and Yang, Y. H. (2001), "Nonparametric Regression With Correlated Errors," *Statistical Science*, 16, 134–153. [585]

Park, B. U., and Marron, J. S. (1990), "Comparison of Data-Driven Bandwidth Selectors," *Journal of the American Statistical Association*, 85, 66–72. [580]

Robinson, P. M. (1989), "Nonparametric Estimation of Time-Varying Parameters," In *Statistical Analysis and Forecasting of Economic Structural Change*, ed. P. Hackl, pp. 253–264, Berlin: Springer. [578]

——— (1991), "Time-Varying Nonlinear Regression," In *Economic Structure Change Analysis and Forecasting*, eds. P. Hackl and A. H. Westland, pp. 179–190, Berlin: Springer. [578]

Serban, N., and Wasserman, L. (2005), "CATS: Clustering After Transformation and Smoothing," *Journal of the American Statistical Association*, 100, 990–999. [577]

Tarpey, T., and Kinateder, K. K. J. (2003), "Clustering Functional Data," *Journal of Classification*, 20, 93–114. [577]

Ward, J. H. (1963), "Hierarchical Groupings to Optimize an Objective Function," *Journal of the American Statistical Association*, 58, 236–244. [577]

Wu, W. B., and Pourahmadi, M. (2009), "Banding Sample Covariance Matrices of Stationary Processes," *Statistica Sinica*, 19, 1755–1768. [581]

Zhang, T., and Wu, W. B. (2011), "Testing Parametric Assumptions of Trends of a Nonstationary Time Series," *Biometrika*, 98, 599–614. [585,586,587]

——— (2012), "Inference of Time-Varying Regression Models," *The Annals of Statistics*, 40, 1376–1402. [580,581,585]

Zhou, Z., and Wu, W. B. (2010), "Simultaneous Inference of Linear Models With Time Varying Coefficients," *Journal of The Royal Statistical Society,* Series B, 72, 513–531. [578,585]