Multiscale Clustering of Nonparametric Regression Curves

Michael Vogt[1] Oliver Linton[2]
University of Bonn University of Cambridge

November 21, 2018

In a wide range of modern applications, we observe a large number of time series rather than only a single one. It is often natural to suppose that there is some group structure in the observed time series. When each time series is modelled by a nonparametric regression equation, one may in particular assume that the observed time series can be partitioned into a small number of groups whose members share the same nonparametric regression function. We develop a bandwidth-free clustering method to estimate the unknown group structure from the data. More precisely speaking, we construct multiscale estimators of the unknown groups and their unknown number which are free of classical bandwidth or smoothing parameters. In the theoretical part of the paper, we analyze the statistical properties of our estimators. Our theoretical results are derived under general conditions which allow the data to be dependent both in time series direction and across different time series. The technical analysis of the paper is complemented by a simulation study and a real-data application.

Key words: Clustering of nonparametric curves; nonparametric regression; multiscale statistics; multiple time series.
JEL classifications: C14; C38; C55.

1 Introduction

In this paper, we are concerned with the problem of clustering nonparametric regression curves. We consider the following model setup: We observe a large number of time series $\mathcal{T}_i = \{(Y_{it}, X_{it}) : 1 \le t \le T\}$ for $1 \le i \le n$. For simplicity, we synonymously speak of the i-th time series, the time series i and the time series $\mathcal{T}_i$ in what follows. Each time series $\mathcal{T}_i$ satisfies the nonparametric regression equation

$$Y_{it} = m_i(X_{it}) + u_{it} \tag{1.1}$$

for $t = 1, \dots, T$, where m_i is an unknown smooth function which is evaluated at the design points X_{it} and u_{it} denotes the error term. The n time series in our sample

[1] Corresponding author. Address: Department of Economics and Hausdorff Center for Mathematics, University of Bonn, 53113 Bonn, Germany. Email: michael.vogt@uni-bonn.de.

[2] Address: Faculty of Economics, Austin Robinson Building, Sidgwick Avenue, Cambridge, CB3 9DD, UK. Email: obl20@cam.ac.uk.

are supposed to belong to K_0 different groups. More specifically, the set of time series $\{1,\ldots,n\}$ can be partitioned into K_0 groups $G_1,\ldots,G_{K_0}$ such that for each $k = 1,\ldots,K_0$,

$$m_i = m_j \quad \text{for all } i,j \in G_k. \tag{1.2}$$

Hence, the members of each group G_k all have the same regression function. A detailed description of model (1.1)–(1.2) can be found in Section 2. Our modelling approach provides a parsimonious way to deal with a potentially very large number of time series n. It thus stands in the tradition of multiple time series analysis, an area which greatly benefited from the pioneering work of George Tiao.

An interesting statistical problem is how to construct estimators of the unknown groups $G_1,\ldots,G_{K_0}$ and their unknown number K_0 in model (1.1)–(1.2). For the special case that the design points $X_{it} = t/T$ represent (rescaled) time and the functions m_i are nonparametric time trends, this problem has been analyzed for example in Luan and Li (2003) and Degras et al. (2012). For the case that X_{it} are general random design points which may differ across time series i, Vogt and Linton (2017) have developed a thresholding method to estimate the unknown groups and their number. Notably, their approach can also be adapted to the case of deterministic regressors X_{it}, in particular to the case that $X_{it} = t/T$. The model (1.1)–(1.2) with the fixed design points $X_{it} = t/T$ is closely related to models from functional data analysis. There, the aim is to cluster smooth random curves that are functions of (rescaled) time and that are observed with or without noise. A number of different clustering approaches have been proposed in the context of functional data models; see for example Abraham et al. (2003), Tarpey and Kinateder (2003) and Tarpey (2007) for procedures based on k-means clustering, James and Sugar (2003) and Chiou and Li (2007) for model-based clustering approaches and Jacques and Preda (2014) for a recent survey.

Virtually all of the proposed procedures to cluster nonparametric curves in model (1.1)–(1.2) and in related functional data settings have the following drawback: they depend on a number of smoothing parameters required to estimate the nonparametric functions m_i. A common approach is to approximate the functions m_i by a series expansion $m_i(x) \approx \sum_{j=1}^{L} \beta_{ij}\phi_j(x)$, where $\{\phi_j : j = 1,2,\ldots\}$ is a function basis and L is the number of basis elements taken into account for the estimation of m_i. Here, L plays the role of the smoothing parameter and may vary across i, that is, $L = L_i$. To estimate the classes $G_1,\ldots,G_{K_0}$, estimators $\widehat{\boldsymbol{\beta}}_i$ of the coefficient vectors $\boldsymbol{\beta}_i = (\beta_{i1},\ldots,\beta_{iL})^\top$ are clustered into groups by a standard clustering algorithm. Variants of this approach have for example been investigated in Abraham et al. (2003), Luan and Li (2003), Chiou and Li (2007) and Tarpey (2007). Another approach is to compute nonparametric estimators $\widehat{m}_i = \widehat{m}_{i,h}$ of the functions m_i for some smoothing parameter h (which may differ across i) and to calculate distances $\widehat{\rho}_{ij} = \rho(\widehat{m}_i,\widehat{m}_j)$ between the estimates $\widehat{m}_i$ and $\widehat{m}_j$, where $\rho(\cdot,\cdot)$ is a distance measure such as a supremum or an

L_2-distance. A distance-based clustering algorithm is then applied to the distances $\widehat{\rho}_{ij}$. This strategy has for example been used in Vogt and Linton (2017).

In general, nonparametric curve estimators strongly depend on the chosen smoothing or bandwidth parameters. A clustering procedure which is based on such estimators can be expected to be strongly influenced by the choice of smoothing parameters as well. To see this issue more clearly, consider two time series i and j from two different groups. The corresponding regression functions m_i and m_j may differ on different scales. In particular, they may differ on a local/global scale, that is, they may have certain local/global features which distinguish them from each other. For example, they may be identical except for a sharp local spike, or they may have a slightly different curvature globally all over their support. Whether nonparametric estimators are able to pick up local/global features of m_i and m_j depends on the chosen bandwidth. When the bandwidth is large, the estimators capture global features of m_i and m_j but smooth out local ones. When the bandwidth is small, they pick up local features, whereas more global ones are poorly captured. As a consequence, a clustering algorithm which is based on nonparametric estimators of m_i and m_j will reliably detect local/global differences between the functions m_i and m_j only if the bandwidths are chosen appropriately. The clustering results produced by such an algorithm can thus be expected to vary considerably with the chosen bandwidths.

The main aim of this paper is to construct estimators of the unknown groups $G_1, \ldots, G_{K_0}$ and of their unknown number K_0 in model (1.1)–(1.2) which are free of classical smoothing or bandwidth parameters. To achieve this, we construct a clustering algorithm which is based on statistical multiscale methods. In recent years, a number of multiscale techniques have been developed in the context of statistical hypothesis testing. Early examples are the SiZer approach of Chaudhuri and Marron (1999, 2000) and the multiscale tests of Horowitz and Spokoiny (2001) and Dümbgen and Spokoiny (2001). More recent references include the tests in Schmidt-Hieber et al. (2013), Armstrong and Chan (2016), Eckle et al. (2017) and Proksch et al. (2018) among others. In this paper, we develop multiscale techniques for clustering rather than testing purposes. Roughly speaking, we proceed as follows: To start with, we construct statistics which measure the distances between pairs of functions m_i and m_j. To do so, we estimate the functions m_i and m_j at different resolution levels, that is, with the help of different bandwidths h. The resulting estimators are aggregated in supremum-type statistics which simultaneously take into account multiple bandwidth levels. We thereby obtain multiscale statistics which avoid the need to pick a specific bandwidth. To estimate the unknown classes $G_1, \ldots, G_{K_0}$, we combine the constructed multiscale statistics with a hierarchical clustering algorithm. To estimate the unknown number of classes K_0, we develop a thresholding rule that is applied to the dendrogram produced by the clustering algorithm. Alternatively, the multiscale statistics may be combined with other distance-based clustering algorithms. In particular, they can be used to turn the estimation strategy of Vogt and Linton (2017) into a bandwidth-free procedure. We comment on this in more detail in Section 9 of the paper.

By construction, our multiscale clustering methods allow to detect differences between the functions m_i at different scales or resolution levels. An alternative way to achieve this is to employ Wavelet methods. A Bayesian Wavelet-based method to cluster nonparametric curves has been developed in Ray and Mallick (2006). There, the model $Y_{it} = m_i(t/T) + u_{it}$ is considered, where m_i are smooth functions of rescaled time t/T and the error terms u_{it} are restricted to be i.i.d. Gaussian noise. To the best of our knowledge, there are no Wavelet-based clustering methods available in the literature which allow to deal with the model setting (1.1)–(1.2) under general conditions on the design points X_{it} and the error terms u_{it}. Our methods and theory, in contrast, allow to do so. In particular, we do not restrict attention to the special case that $X_{it} = t/T$ but allow for general design points X_{it} that may differ across i. Moreover, we do not restrict the error terms to be Gaussian but only impose some moderate moment conditions on them. In addition, we allow them to be dependent both across t and i.

The problem of estimating the unknown groups and their unknown number in model (1.1)–(1.2) is closely related to a developing literature in econometrics which aims to identify the unknown group structure in parametric panel regression models. The clustering problem considered in this literature can be regarded as a parametric version of our problem. In its simplest form, the panel regression model under consideration is given by the equation $Y_{it} = \boldsymbol{\beta}_i^\top X_{it} + u_{it}$ for $1 \le t \le T$ and $1 \le i \le n$, where the coefficient vectors $\boldsymbol{\beta}_i$ are allowed to vary across individuals i. Similarly as in our nonparametric model, the coefficients $\boldsymbol{\beta}_i$ are assumed to belong to a number of groups: there are K_0 groups $G_1, \ldots, G_{K_0}$ such that $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j$ for all $i, j \in G_k$ and all $1 \le k \le K_0$. The problem of estimating the unknown groups and their unknown unknown number has been studied in different versions of this modelling framework in Bonhomme and Manresa (2015), Su et al. (2016), Wang et al. (2018) and Su and Ju (2018) among others. Note that our clustering methods can be adapted in a straightforward way to a number of semiparametric models which are middle ground between the fully parametric panel models just discussed and our nonparametric framework. In Section 9, we discuss in more detail how to achieve this.

Our estimation methods are described in detail in Sections 3–5. In Section 3, we construct the multiscale statistics that form the basis of our clustering methods. Section 4 introduces the hierarchical clustering algorithm to estimate the unknown classes $G_1, \ldots, G_{K_0}$. In Section 5, we finally describe the procedure to estimate the unknown number of classes K_0. The main theoretical result of the paper is laid out in Section 6. This result characterizes the asymptotic convergence behaviour of the multiscale statistics and forms the basis to derive the theoretical properties of our clustering methods. To explore the finite sample properties of our approach and to illustrate its advantages over bandwidth-dependent clustering algorithms, we conduct a simulation study in Section 7. Moreover, we illustrate the procedure by an application from finance in Section 8.

2 The model

As already mentioned in the Introduction, we observe n different time series $\mathcal{T}_i = \{(Y_{it}, X_{it}) : 1 \le t \le T\}$ of length T for $1 \le i \le n$. In what follows, we describe in detail how the observed data $\{\mathcal{T}_i : 1 \le i \le n\}$ are modelled. For our theoretical analysis, we regard the number of time series n as a function of T, that is, $n = n(T)$. The time series length T is assumed to tend to infinity, whereas the number of time series n may be either bounded or diverging. The exact technical conditions on T and n are laid out in Section 6. Throughout the paper, asymptotic statements are to be understood in the sense that $T \to \infty$.

2.1 The model for time series $\mathcal{T}_i$

Each time series $\mathcal{T}_i$ in our sample is modelled by the nonparametric regression equation

$$Y_{it} = m_i(X_{it}) + u_{it} \tag{2.1}$$

for $1 \le t \le T$, where m_i is an unknown smooth function and u_{it} denotes the error term. We focus attention on the case that the design points X_{it} are random as this is the technically more involved case. Our methods can be adapted to deterministic design points X_{it} with some minor modifications. To keep the exposition as simple as possible, we assume that the regressors X_{it} are real-valued. As discussed in Section 9, our methods and theory carry over to the multivariate case in a straightforward way. We further suppose that the regressors X_{it} have compact support, which w.l.o.g. is equal to $[0, 1]$ for each i. The error terms u_{it} in (2.1) are assumed to have the additive component structure

$$u_{it} = \alpha_i + \gamma_t + \varepsilon_{it}, \tag{2.2}$$

where ε_{it} are standard regression errors that satisfy $\mathbb{E}[\varepsilon_{it}|X_{it}] = 0$ and the terms α_i and γ_t are so-called fixed effects. The expression α_i is an error component which is specific to the i-th time series $\mathcal{T}_i$. It can be interpreted as capturing unobserved characteristics of the time series $\mathcal{T}_i$ which are stable over time. Suppose for instance that the observations of $\mathcal{T}_i$ are sampled from some subject i. In this case, α_i can be regarded as controlling for time-invariant unobserved characteristics of subject i, such as intelligence or certain unknown genetic factors. Similarly, the term γ_t captures unobserved time-specific effects like calendar effects or trends that are common across time series i. In many applications, the regressors may be correlated with unobserved subject- or time-specific characteristics. To take this into account, we allow the errors α_i and γ_t to be correlated with the regressors in an arbitrary way. Specifically, defining $\mathcal{X}_{n,T} = \{X_{it} : 1 \le i \le n,\, 1 \le t \le T\}$, we permit that $\mathbb{E}[\alpha_i | \mathcal{X}_{n,T}] \neq 0$ and $\mathbb{E}[\gamma_t | \mathcal{X}_{n,T}] \neq 0$. The error terms ε_{it} are allowed to be dependent across t but are assumed to be independent across i. The fixed effects α_i, in contrast, may be correlated across i in an arbitrary way. Hence, by including α_i and γ_t in the error structure, we allow for some restricted types of cross-sectional dependence in the errors u_{it}. As a result, we accommodate for both time series dependence and certain forms of cross-sectional dependence in the error terms of our model. The exact conditions on the dependence structure are stated in (C1) in Section 6.

2.2 The group structure

We impose the following group structure on the time series $\mathcal{T}_i$ in our sample: There are K_0 groups of time series $G_1, \ldots, G_{K_0}$ with $\dot{\bigcup}_{k=1}^{K_0} G_k = \{1, \ldots, n\}$ such that for each $1 \le k \le K_0$,

$$m_i = m_j \quad \text{for all } i, j \in G_k. \tag{2.3}$$

Put differently, for each $1 \le k \le K_0$,

$$m_i = g_k \quad \text{for all } i \in G_k, \tag{2.4}$$

where g_k is the group-specific regression function associated with the class G_k. According to (2.4), the time series of a given class G_k all have the same regression curve g_k. To make sure that time series which belong to different classes have different regression curves, we suppose that $g_k \neq g_{k'}$ for $k \neq k'$. The exact technical conditions on the functions g_k are summarized in (C6) in Section 6. For simplicity, we assume that the number of groups K_0 is fixed. It is however straightforward to allow K_0 to grow with the number of time series n. We comment on this in more detail in Section 9. The groups $G_k = G_{k,n}$ depend on the cross-section dimension n in general. For ease of notation, we however suppress this dependence on n throughout the paper.

2.3 Identification of the functions m_i

Plugging (2.2) into (2.1), we obtain the model equation

$$Y_{it} = m_i(X_{it}) + \alpha_i + \gamma_t + \varepsilon_{it}, \tag{2.5}$$

where $\mathbb{E}[\varepsilon_{it}|X_{it}] = 0$. If we drop the fixed effects α_i and γ_t from (2.5), we are left with the standard regression equation $Y_{it} = m_i(X_{it}) + \varepsilon_{it}$. Obviously, m_i is identified in this case since $m_i(\,\cdot\,) = \mathbb{E}[Y_{it}|X_{it} = \,\cdot\,]$. In the full model (2.5), in contrast, m_i is not identified. In particular, we can rewrite (2.5) as $Y_{it} = \{m_i(X_{it})+a_i\}+\{\alpha_i-a_i\}+\gamma_t+\varepsilon_{it}$, where a_i is an arbitrary real constant. In order to get identification, we need to impose certain constraints which pin down the expectation $\mathbb{E}[m_i(X_{it})]$ for any i and t. We in particular work with the identification constraint that

$$\mathbb{E}[m_i(X_{it})] = 0 \quad \text{for } 1 \le t \le T \text{ and } 1 \le i \le n. \tag{2.6}$$

Under this constraint, it is straightforward to show that the functions m_i are identified. In particular, we can derive the following formal result whose proof is given in the Supplementary Material for completeness.

Proposition 2.1. *Let the constraint* (2.6) *be satisfied and suppose that the regularity conditions (C1)–(C6) from Section 6 are fulfilled. Then the functions m_i in model* (2.5) *are identified. More precisely, let m_i and $\widetilde{m}_i$ be two functions for some $i \in \{1, \ldots, n\}$ which satisfy the model equation* (2.5) *for any t and which are normalized such that $\mathbb{E}[m_i(X_{it})] = \mathbb{E}[\widetilde{m}_i(X_{it})] = 0$ for any t. Then $m_i(x) = \widetilde{m}_i(x)$ must hold for all $x \in [0, 1]$.*

Apart from a couple of technicalities, conditions (C1)–(C6) contain the following two assumptions which are essential for the identification result of Proposition 2.1:

(a) The time series $\{X_{it} : t = 1, 2, \ldots\}$ is strictly stationary with $X_{it} \sim f_i$ for each i.

(b) The density f_i is the same for all time series i in a given group G_k, that is, $f_i = f_j$ for all $i, j \in G_k$ and any k.

Under (a) and (b), the identification constraint (2.6) amounts to a harmless normalization of the functions m_i. On the other hand, it is in general not possible to satisfy (2.6) without the assumptions (a) and (b): Suppose that (a) is violated and that for some i, $X_{it} \sim f_{it}$ with a density f_{it} that differs across t. In this case, the constraint (2.6) requires that $\int m_i(x)f_{it}(x)dx = 0$ for all t. In general, it is however not possible to satisfy the equation $\int m_i(x)f_{it}(x)dx = 0$ simultaneously for all t if the density f_{it} differs across t. An analogous problem arises when (b) is violated and the density f_i varies across $i \in G_k$. According to these considerations, the normalization constraint (2.6) requires us to impose assumptions (a) and (b). Hence, in order to identify the functions m_i in the presence of a general fixed effects error structure, we need the regressors to satisfy (a) and (b). If we dropped the fixed effects from the model, we could of course do without these assumptions. There is thus a certain trade-off between a general fixed effects error structure and weaker conditions on the regressors.

3 The multiscale distance statistic

Let i and j be two time series from our sample. In what follows, we construct a test statistic $\widehat{d}_{ij}$ for the null hypothesis $H_0 : m_i(x) = m_j(x)$ for all $x \in [0,1]$, that is, for the null hypothesis that i and j belong to the same group G_k for some $1 \le k \le K_0$. We design the statistic $\widehat{d}_{ij}$ in such a way that it does not depend on a specific bandwidth or smoothing parameter. The statistic $\widehat{d}_{ij}$ will serve as a distance measure between the functions m_i and m_j in our clustering algorithm later on.

3.1 Construction of the multiscale statistic

Step 1. As a first preliminary step, we define a nonparametric estimator $\widehat{m}_{i,h}$ of the function m_i, where h denotes the bandwidth. To do so, suppose for a moment that the fixed effects α_i and γ_t are known, which implies that the variables $Y_{it}^* = Y_{it} - \alpha_i - \gamma_t$ are known as well. In this case, we can work with the model equation $Y_{it}^* = m_i(X_{it}) + \varepsilon_{it}$ and estimate the function m_i by applying standard nonparametric regression techniques to the sample $\{(Y_{it}^*, X_{it}) : 1 \le t \le T\}$. Since α_i and γ_t are unobserved in practice, we replace the unknown variables Y_{it}^* by the approximations $\widehat{Y}_{it}^* = Y_{it} - \overline{Y}_i - \overline{Y}_t^{(i)} + \overline{\overline{Y}}^{(i)}$, where

$$\overline{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it}, \quad \overline{Y}_t^{(i)} = \frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^{n} Y_{jt} \quad \text{and} \quad \overline{\overline{Y}}^{(i)} = \frac{1}{(n-1)T}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{t=1}^{T} Y_{jt}. \tag{3.1}$$

With these approximations at hand, we can estimate m_i by applying kernel regression techniques to the constructed sample $\{(\widehat{Y}_{it}^*, X_{it}) : 1 \le t \le T\}$. In particular, we define a local linear kernel estimator of m_i by

$$\widehat{m}_{i,h}(x) = \frac{\sum_{t=1}^{T} W_{it}(x,h)\widehat{Y}_{it}^*}{\sum_{t=1}^{T} W_{it}(x,h)}, \tag{3.2}$$

where the weights $W_{it}(x,h)$ have the form

$$W_{it}(x,h) = K_h(X_{it} - x)\Big\{S_{i,2}(x,h) - \Big(\frac{X_{it} - x}{h}\Big)S_{i,1}(x,h)\Big\} \tag{3.3}$$

with $S_{i,\ell}(x,h) = T^{-1}\sum_{t=1}^{T} K_h(X_{it} - x)(\frac{X_{it}-x}{h})^{\ell}$ for $\ell = 0,1,2$ and K is a kernel function with $K_h(\varphi) = h^{-1}K(\varphi/h)$. Throughout the paper, we assume that the kernel K has compact support $[-C_K, C_K]$ and we set $C_K = 1$ for ease of notation.

Step 2. As an intermediate step in our construction, we set up a bandwidth-dependent test statistic for a somewhat simpler hypothesis than H_0. Specifically, we consider the hypothesis $H_{0,x} : m_i(x) = m_j(x)$ for a fixed point $x \in [0,1]$. A test statistic for this problem is given by

$$\widehat{\psi}_{ij}(x,h) = \sqrt{Th}\,\frac{\big(\widehat{m}_{i,h}(x) - \widehat{m}_{j,h}(x)\big)}{\sqrt{\widehat{\nu}_{ij}(x,h)}}, \tag{3.4}$$

where

$$\widehat{\nu}_{ij}(x,h) = \left\{ \frac{\widehat{\sigma}^2_{i,h}(x)}{\widehat{f}_{i,h}(x)} + \frac{\widehat{\sigma}^2_{j,h}(x)}{\widehat{f}_{j,h}(x)} \right\} s(x,h) \tag{3.5}$$

is a scaling factor which normalizes the variance of $\widehat{\psi}_{ij}(x,h)$ to be approximately equal to 1 for sufficiently large T. In formula (3.5), $s(x,h) = \{\int_{-x/h}^{(1-x)/h} K^2(u)[\kappa_2(x,h) - \kappa_1(x,h)u]^2 du\}/\{\kappa_0(x,h)\kappa_2(x,h) - \kappa_1(x,h)^2\}^2$ is a kernel constant with $\kappa_\ell(x,h) = \int_{-x/h}^{(1-x)/h} u^\ell K(u) du$ for $0 \le \ell \le 2$. Moreover, $\widehat{f}_{i,h}(x) = \{\kappa_0(x,h)T\}^{-1} \sum_{t=1}^T K_h(X_{it} - x)$ is a boundary-corrected kernel density estimator of f_i, where f_i denotes the density of the regressor X_{it} as in Section 2.3, and $\widehat{\sigma}^2_{i,h}(x) = \{\sum_{t=1}^T K_h(X_{it} - x)[\widehat{Y}^*_{it} - \widehat{m}_{i,h}(X_{it})]^2\}/\{\sum_{t=1}^T K_h(X_{it} - x)\}$ is an estimator of the conditional error variance $\sigma_i^2(x) = \mathbb{E}[\varepsilon_{it}^2 | X_{it} = x]$. If the error terms ε_{it} are homoskedastic, that is, if $\sigma_i^2(x) \equiv \sigma_i^2 = \mathbb{E}[\varepsilon_{it}^2]$ for any x, we can replace $\widehat{\sigma}^2_{i,h}(x)$ by the simpler estimator $\widehat{\sigma}^2_{i,h} = T^{-1} \sum_{t=1}^T \{\widehat{Y}^*_{it} - \widehat{m}_{i,h}(X_{it})\}^2$.

For some of the discussion later on, it is convenient to decompose the statistic $\widehat{\psi}_{ij}(x,h)$ into a bias part $\widehat{\psi}^B_{ij}(x,h)$ and a variance part $\widehat{\psi}^V_{ij}(x,h)$. Standard calculations for kernel estimators yield that

$$\widehat{\psi}_{ij}(x,h) = \widehat{\psi}^B_{ij}(x,h) + \widehat{\psi}^V_{ij}(x,h) + \text{lower order terms}, \tag{3.6}$$

where

$$\widehat{\psi}^B_{ij}(x,h) = \sqrt{Th}\frac{\int_{-x/h}^{(1-x)/h} \{w_i(u,x,h) m_i(x+hu) - w_j(u,x,h) m_j(x+hu)\} K(u) du}{\sqrt{\widehat{\nu}_{ij}(x,h)}}$$

with $w_i(u,x,h) = \{\mathbb{E}[S_{i,2}(x,h)] - \mathbb{E}[S_{i,1}(x,h)]u\} f_i(x+hu)/\{\mathbb{E}[S_{i,0}(x,h)]\mathbb{E}[S_{i,2}(x,h)] - \mathbb{E}[S_{i,1}(x,h)]^2\}$ and

$$\widehat{\psi}^V_{ij}(x,h) = \sqrt{Th}\,\frac{\big(\widehat{m}^V_{i,h}(x) - \widehat{m}^V_{j,h}(x)\big)}{\sqrt{\widehat{\nu}_{ij}(x,h)}}$$

with $\widehat{m}^V_{i,h}(x) = \{\sum_{t=1}^T W_{it}(x,h)(\varepsilon_{it} - \overline{\varepsilon}^{(i)}_t - \overline{m}^{(i)}_t)\}\{\sum_{t=1}^T W_{it}(x,h)\}$ as well as $\overline{\varepsilon}^{(i)}_t = (n-1)^{-1} \sum_{j=1, j\ne i}^n \varepsilon_{jt}$ and $\overline{m}^{(i)}_t = (n-1)^{-1} \sum_{j=1, j\ne i}^n m_j(X_{jt})$. Under the regularity conditions from Section 6, it can be shown that $\widehat{\psi}^V_{ij}(x,h) \xrightarrow{d} N(0, V_{ij})$, where the asymptotic variance V_{ij} is exactly equal to 1 in the case that $n \to \infty$ and is approximately equal to 1 if n is large but bounded. Moreover, under these conditions, the bias term $\widehat{\psi}^B_{ij}(x,h)$ vanishes for any pair of time series i and j that belong to the same class G_k, that is, $\widehat{\psi}^B_{ij}(x,h) = 0$ for any $i, j \in G_k$ and $1 \le k \le K_0$.

The variance part $\widehat{\psi}^V_{ij}(x,h)$ captures the stochastic fluctuations of the statistic $\widehat{\psi}_{ij}(x,h)$, whereas $\widehat{\psi}^B_{ij}(x,h)$ can be regarded as a signal which indicates a deviation from the null $H_{0,x}$. The strength of the signal $\widehat{\psi}^B_{ij}(x,h)$ depends on the choice of the bandwidth h. To better understand how the signal varies with the bandwidth h, suppose that the two functions m_i and m_j differ on the interval $I(x,h_0) = [x-h_0, x+h_0]$ but are the same outside $I(x,h_0)$. The parameter h_0 specifies how local the differences between m_i and m_j are. Put differently, it specifies the scale on which m_i and m_j differ: For small/large values of h_0, the interval $I(x,h_0)$ is small/large compared to the overall support $[0,1]$, which means that m_i and m_j differ on a local/global scale. Usually, the signal $\widehat{\psi}^B_{ij}(x,h)$ is strongest for bandwidths h close to h_0 and becomes weak for bandwidths h that are substantially smaller or larger than h_0. The heuristic reason for this is as follows: If h is much larger than h_0, the differences between m_i and m_j get smoothed out by the kernel methods that underlie the statistic $\widehat{\psi}_{ij}(x,h)$. If h is much smaller than h_0, in contrast, we do not take into account all data points which convey information on the difference between m_i and m_j. As a result, the signal $\widehat{\psi}^B_{ij}(x,h)$ gets rather weak. Hence, if the bandwidth h is much smaller/larger than the scale h_0 on which m_i and m_j mainly differ, the statistic $\widehat{\psi}_{ij}(x,h)$ is not able to pick up the differences between m_i and m_j and thus to detect a deviation from the null $H_{0,x}$.

Step 3. Let us now turn to the problem of testing the hypothesis $H_0 : m_i(x) = m_j(x)$ for all $x \in [0,1]$. A simple bandwidth-dependent test statistic for H_0 is the supremum statistic

$$\widehat{d}_{ij}(h) = \sup_{x \in [0,1]} \big| \widehat{\psi}_{ij}(x,h) \big|.$$

Obviously, this statistic suffers from the same problem as the statistic $\widehat{\psi}_{ij}(x,h)$: It is not able to pick up local/global differences between the functions m_i and m_j in a reliable way if the bandwidth h is chosen too large/small. Its performance can thus be expected to strongly depend on the chosen bandwidth.

A simple strategy to get rid of the dependence on the bandwidth h is as follows: We compute the statistic $\widehat{d}_{ij}(h)$ not only for a single bandwidth h but for a wide range of different bandwidths. We in particular consider all bandwidths h in the set $\mathcal{H} = \{h : h_{\min} \le h \le h_{\max}\}$, where $h_{\min}$ and $h_{\max}$ denote some minimal and maximal bandwidth values that are specified later on. This leaves us with a whole family of statistics $\{\widehat{d}_{ij}(h) : h \in \mathcal{H}\}$. By taking the supremum over all these statistics, we obtain the rudimentary multiscale statistic

$$\widetilde{d}_{ij} = \sup_{h \in \mathcal{H}} \widehat{d}_{ij}(h) = \sup_{h \in \mathcal{H}} \sup_{x \in [0,1]} \big| \widehat{\psi}_{ij}(x,h) \big|. \tag{3.7}$$

This statistic does not depend on a specific bandwidth h that needs to be selected. It rather takes into account a wide range of different bandwidths $h \in \mathcal{H}$ simultaneously.

It should thus be able to detect differences between the functions m_i and m_j on multiple scales simultaneously. Put differently, it should be able to pick up both local and global differences between m_i and m_j.

Inspecting the statistic $\widetilde{d}_{ij}$ more closely, it can be seen to have the following drawback: It does not take into account all scales $h \in \mathcal{H}$ in an equal fashion. Its stochastic behaviour is rather dominated by the statistics $\widehat{\psi}_{ij}(x,h)$ that correspond to small scales h. To see this, let us examine the statistic $\widetilde{d}_{ij}$ under the null hypothesis H_0, that is, in the case that i and j belong to the same group G_k. In this case, $\widehat{\psi}_{ij}(x,h) = \widehat{\psi}_{ij}^V(x,h) +$ lower order terms, since the bias term $\widehat{\psi}_{ij}^B(x,h)$ in (3.6) is equal to 0 for all x and h as already noted in Step 2 above. Hence, the statistic $\widehat{\psi}_{ij}(x,h)$ is approximately equal to the variance term $\widehat{\psi}_{ij}^V(x,h)$, which captures its stochastic fluctuations. Neglecting terms of lower order, we obtain that under H_0, $\widehat{\psi}_{ij}(x,h) = \widehat{\psi}_{ij}^V(x,h)$ and thus

$$\widetilde{d}_{ij} = \sup_{h \in \mathcal{H}} \widehat{d}_{ij}(h) \qquad \text{with} \qquad \widehat{d}_{ij}(h) = \sup_{x \in [0,1]} |\widehat{\psi}_{ij}^V(x,h)|.$$

For a given bandwidth h, the statistics $\widehat{\psi}_{ij}^V((2\ell-1)h,h)$ for $\ell = 1,\ldots,\lfloor 1/2h \rfloor$ are (approximately) standard normal and independent (for sufficiently large T). Since the maximum over $\lfloor 1/2h \rfloor$ independent standard normal random variables is $\lambda(2h) + o_p(1)$ as $h \to 0$ with $\lambda(r) = \sqrt{2\log(1/r)}$, it holds that $\max_\ell \widehat{\psi}_{ij}^V((2\ell-1)h,h)$ is approximately of size $\lambda(2h)$ for small bandwidths h. Moreover, since the statistics $\widehat{\psi}_{ij}^V(x,h)$ with $(2\ell-1)h < x < (2\ell+1)h$ are correlated with $\widehat{\psi}_{ij}^V((2\ell-1)h,h)$ and $\widehat{\psi}_{ij}^V((2\ell+1)h,h)$, the supremum $\sup_x \widehat{\psi}_{ij}^V(x,h)$ approximately behaves as the maximum $\max_\ell \widehat{\psi}_{ij}^V((2\ell-1)h,h)$. Taken together, these considerations suggest that

$$\widehat{d}_{ij}(h) \approx \max_{1 \le \ell \le \lfloor 1/2h \rfloor} \big|\widehat{\psi}_{ij}^V((2\ell-1)h,h)\big| \approx \lambda(2h) \tag{3.8}$$

for small bandwidth values h. According to (3.8), the statistic $\widehat{d}_{ij}(h)$ tends to be much larger in size for small than for large bandwidths h. As a consequence, the stochastic behaviour of $\widetilde{d}_{ij}$ tends to be dominated by the statistics $\widehat{d}_{ij}(h)$ which correspond to small bandwidths h.

To fix this problem, we follow Dümbgen and Spokoiny (2001) and replace the statistic $\widetilde{d}_{ij}$ by the modified version

$$\widehat{d}_{ij} = \sup_{h \in \mathcal{H}} \sup_{x \in [0,1]} \big\{ |\widehat{\psi}_{ij}(x,h)| - \lambda(2h) \big\}, \tag{3.9}$$

where $\lambda(r) = \sqrt{2\log(1/r)}$. For each given bandwidth h, we thus subtract the additive correction term $\lambda(2h)$ from the statistics $\widehat{\psi}_{ij}(x,h)$. The idea behind this additive correction is as follows: We can write $\widehat{d}_{ij} = \sup_{h \in \mathcal{H}} \{\widehat{d}_{ij}(h) - \lambda(2h)\}$ with $\widehat{d}_{ij}(h) = \sup_{x \in [0,1]} |\widehat{\psi}_{ij}(x,h)|$. According to the heuristic considerations from above, when i and

j belong to the same class, the statistic $\widehat{d}_{ij}(h)$ is approximately of size $\lambda(2h)$ for small values of h. Hence, we correct $\widehat{d}_{ij}(h)$ by subtracting its approximate size under the null hypothesis H_0. This calibrates the statistics $\widehat{d}_{ij}(h)$ in such a way that their stochastic fluctuations are comparable across scales h. We thus put them on a more equal footing and prevent small scales from dominating the stochastic behaviour of the multiscale statistic. As a result, the statistic $\widehat{d}_{ij}$ should be able to detect differences between the functions m_i and m_j on multiple scales simultaneously without being dominated by a particular scale. It should thus be a reliable test statistic for H_0, no matter whether the differences between m_i and m_j are on local or global scales.

To make the statistic $\widehat{d}_{ij}$ defined in (3.9) computable in practice, we replace the supremum over $x \in [0,1]$ and $h \in \mathcal{H}$ by the maximum over all points (x,h) in a suitable grid $\mathcal{G}_T$. The final version of the multiscale statistic is thus defined as

$$\widehat{d}_{ij} = \max_{(x,h)\in\mathcal{G}_T} \big\{ |\widehat{\psi}_{ij}(x,h)| - \lambda(2h) \big\}. \tag{3.10}$$

In this definition, $\mathcal{G}_T$ may be any subset of $\mathcal{G} = \{(x,h) \,|\, h_{\min} \le h \le h_{\max} \text{ and } x \in [0,1]\}$ with the following properties: (a) $\mathcal{G}_T$ becomes dense in $\mathcal{G}$ as $T \to \infty$, (b) $|\mathcal{G}_T| \le CT^{\beta}$ for some arbitrarily large but fixed constants $C, \beta > 0$, where $|\mathcal{G}_T|$ denotes the cardinality of $\mathcal{G}_T$, and (c) $h_{\min} \ge cT^{-(1-\delta)}$ and $h_{\max} \le CT^{-\delta}$ for some arbitrarily small but fixed $\delta > 0$ and some positive constants c and C. According to conditions (a) and (b), the number of points (x,h) in $\mathcal{G}_T$ should grow to infinity as $T \to \infty$, however it should not grow faster than CT^{β} for some arbitrarily large constants $C, \beta > 0$. This is a fairly weak restriction as it allows the set $\mathcal{G}_T$ to be extremely large as compared to the sample size T. As an example, we may use the Wavelet multiresolution grid $\mathcal{G}_T = \{(x,h) = (2^{-\nu}r, 2^{-\nu}) \,|\, 1 \le r \le 2^{\nu} - 1 \text{ and } h_{\min} \le 2^{-\nu} \le h_{\max}\}$. Condition (c) is quite weak as well, allowing us to choose the bandwidth window $[h_{\min}, h_{\max}]$ extremely large. In particular, we can choose the minimal bandwidth $h_{\min}$ to converge to zero almost as quickly as the time series length T and thus to be extremely small. Moreover, the maximal bandwidth $h_{\max}$ is allowed to converge to zero very slowly, in particular much more slowly than the optimal bandwidths for estimating the functions m_i, which are of the order $T^{-1/5}$ for all i under our technical conditions from Section 6. Hence, $h_{\max}$ can be chosen very large.

3.2 Tuning parameter choice

The multiscale statistic $\widehat{d}_{ij}$ does not depend on a specific bandwidth h that needs to be selected. It is thus free of a classical bandwidth or smoothing parameter. However, it is of course not completely free of tuning parameters. It obviously depends on the minimal and maximal bandwidths $h_{\min}$ and $h_{\max}$. Importantly, $h_{\min}$ and $h_{\max}$ are much more harmless tuning parameters than a classical bandwidth h. In particular, (a) they are much simpler to choose and (b) the multiscale methods are much less sensitive to their exact choice than conventional methods are to the choice of bandwidth. In what follows, we discuss the reasons for (a) and (b) in detail and give some guidelines how to choose $h_{\min}$ and $h_{\max}$ appropriately in practice. These guidelines are in particular used to implement our methods in the simulations of Section 7 and the empirical application of Section 8.

Ideally, we would like to make the interval $[h_{\min}, h_{\max}]$ as large as possible, thus taking into account as many scales h as possible. From a technical perspective, we can pick any bandwidths $h_{\min}$ and $h_{\max}$ with $h_{\min} \geq cT^{-(1-\delta)}$ and $h_{\max} \leq CT^{-\delta}$ for some small $\delta > 0$. Hence, our theory allows us to choose $h_{\min}$ and $h_{\max}$ extremely small and large, respectively. Heuristically speaking, the bandwidth $h_{\min}$ can be considered very small if the effective sample size $Th_{\min}$ for estimating the functions m_i is very small, say $Th_{\min} \leq 10$. Likewise, $h_{\max}$ can be regarded as extremely large if the effective sample size $Th_{\max}$ is very large compared to the full sample size T, say $Th_{\max} \approx T/4$ or $Th_{\max} \approx T/3$. Hence, in practice, we have a pretty good idea of what it means for $h_{\min}$ and $h_{\max}$ to be very small and large, respectively. It is thus clear in which range we need to pick the bandwidths $h_{\min}$ and $h_{\max}$ in practice.

As long as the bandwidth window $[h_{\min}, h_{\max}]$ is chosen reasonably large, the exact choice of $h_{\min}$ and $h_{\max}$ can be expected to have little effect on the overall behaviour of the multiscale statistic $\widehat{d}_{ij}$. To see why, write $\widehat{\psi}_{ij}(x,h) = \widehat{\psi}_{ij}^B(x,h) + \widehat{\psi}_{ij}^V(x,h) +$ lower order terms as in (3.6), where the variance term $\widehat{\psi}_{ij}^V(x,h)$ captures the stochastic fluctuations of $\widehat{\psi}_{ij}(x,h)$ and the bias term $\widehat{\psi}_{ij}^B(x,h)$ is a signal which picks up differences between the functions m_i and m_j locally around x. Neglecting terms of lower order, the multiscale statistic $\widehat{d}_{ij}$ from (3.9) can be written as

$$\widehat{d}_{ij} = \sup_{h \in [h_{\min}, h_{\max}]} \sup_{x \in [0,1]} \big\{ |\widehat{\psi}_{ij}^B(x,h) + \widehat{\psi}_{ij}^V(x,h)| - \lambda(2h) \big\}.$$

If the bandwidth window $[h_{\min}, h_{\max}]$ is chosen sufficiently large, it will contain all the scales h^* on which the two functions m_i and m_j mainly differ. As discussed in Section 3.1, the signals $\widehat{\psi}_{ij}^B(x,h)$ should be strongest for bandwidths h which are close to the scales h^*. Hence, as long as the window $[h_{\min}, h_{\max}]$ is chosen large enough to contain all the scales h^*, the size of the overall signal of the multiscale statistic $\widehat{d}_{ij}$ should be hardly affected by the exact choice of $h_{\min}$ and $h_{\max}$. Moreover, the size of the stochastic fluctuations of $\widehat{d}_{ij}$ should not be strongly influenced either: The stochastic part of $\widehat{d}_{ij}$ can be expressed as

$$\sup_{h \in [h_{\min}, h_{\max}]} \widehat{V}_{ij}(h) \quad \text{with} \quad \widehat{V}_{ij}(h) = \sup_{x \in [0,1]} \big\{ |\widehat{\psi}_{ij}^V(x,h)| - \lambda(2h) \big\},$$

where $\widehat{V}_{ij}(h)$ captures the stochastic fluctuations corresponding to bandwidth h. According to our heuristic considerations from Section 3.1, the variables $\widehat{V}_{ij}(h)$ are comparable in size across bandwidths h. Moreover, for h and h' close to each other, $\widehat{V}_{ij}(h)$ and $\widehat{V}_{ij}(h')$ are strongly correlated. For these reasons, the size of the stochastic part $\sup_{h \in [h_{\min}, h_{\max}]} \widehat{V}_{ij}(h)$ should not change much when we make the very large bandwidth window $[h_{\min}, h_{\max}]$ somewhat larger or smaller.

In view of these heuristic considerations, we suggest to choose $h_{\min}$ in practice such that the effective sample size $Th_{\min}$ is small, say ≤ 10, and $h_{\max}$ such that the effective sample size $Th_{\max}$ is large compared to T, say $Th_{\max} \geq T/4$.

3.3 Properties of the multiscale statistic

We now discuss some theoretical properties of the multiscale statistic $\widehat{d}_{ij}$ which are needed to derive the formal properties of the clustering methods developed in the following sections. Specifically, we compare the maximal multiscale distance between two time series i and j from the same class,

$$\max_{1 \leq k \leq K_0} \max_{i,j \in G_k} \widehat{d}_{ij},$$

with the minimal distance between two time series i and j from two different classes,

$$\min_{1 \leq k < k' \leq K_0} \min_{\substack{i \in G_k, \\ j \in G_{k'}}} \widehat{d}_{ij}.$$

In Section 6, we show that under appropriate regularity conditions,

$$\max_{1 \leq k \leq K_0} \max_{i,j \in G_k} \widehat{d}_{ij} = O_p\big(\sqrt{\log n + \log T}\big) \tag{3.11}$$

$$\min_{1 \leq k < k' \leq K_0} \min_{\substack{i \in G_k, \\ j \in G_{k'}}} \widehat{d}_{ij} \geq c_0 \sqrt{Th_{\max}} + o_p\big(\sqrt{Th_{\max}}\big), \tag{3.12}$$

where c_0 is a sufficiently small positive constant. These two statements imply that

$$\max_{1 \leq k \leq K_0} \max_{i,j \in G_k} \widehat{d}_{ij} \big/ \sqrt{Th_{\max}} = o_p(1) \tag{3.13}$$

$$\min_{1 \leq k < k' \leq K_0} \min_{\substack{i \in G_k, \\ j \in G_{k'}}} \widehat{d}_{ij} \big/ \sqrt{Th_{\max}} \geq c_0 + o_p(1). \tag{3.14}$$

According to (3.13) and (3.14), the maximal distance between time series of the same class converges to zero when normalized by $\sqrt{Th_{\max}}$, whereas the minimal distance between time series of two different classes remains bounded away from zero. Asymptotically, the distance measures $\widehat{d}_{ij}$ thus contain enough information to detect which time series belong to the same class. Technically speaking, we can make the following statement for any fixed positive constant $c < c_0$: with probability tending to 1, any time series i and j with $\widehat{d}_{ij} \le c$ belong to the same class, whereas those with $\widehat{d}_{ij} > c$ belong to two different classes. The hierarchical clustering algorithm introduced in the next section exploits this information in the distances $\widehat{d}_{ij}$.

4 Estimation of the unknown groups

Let $S \subseteq \{1, \dots, n\}$ and $S' \subseteq \{1, \dots, n\}$ be two sets of time series from our sample. We define a dissimilarity measure between S and S' by setting

$$\widehat{\Delta}(S, S') = \max_{\substack{i \in S, \\ j \in S'}} \widehat{d}_{ij}. \tag{4.1}$$

This is commonly called a complete linkage measure of dissimilarity. Alternatively, we may work with an average or a single linkage measure. To partition the set of time series $\{1, \dots, n\}$ into groups, we combine the multiscale dissimilarity measure $\widehat{\Delta}$ with a hierarchical agglomerative clustering (HAC) algorithm which proceeds as follows:

Step 0 (Initialization): Let $\widehat{G}_i^{[0]} = \{i\}$ denote the i-th singleton cluster for $1 \le i \le n$ and define $\{\widehat{G}_1^{[0]}, \dots, \widehat{G}_n^{[0]}\}$ to be the initial partition of time series into clusters.

Step r (Iteration): Let $\widehat{G}_1^{[r-1]}, \dots, \widehat{G}_{n-(r-1)}^{[r-1]}$ be the $n - (r-1)$ clusters from the previous step. Determine the pair of clusters $\widehat{G}_k^{[r-1]}$ and $\widehat{G}_{k'}^{[r-1]}$ for which

$$\widehat{\Delta}(\widehat{G}_k^{[r-1]}, \widehat{G}_{k'}^{[r-1]}) = \min_{1 \le \ell < \ell' \le n-(r-1)} \widehat{\Delta}(\widehat{G}_\ell^{[r-1]}, \widehat{G}_{\ell'}^{[r-1]})$$

and merge them into a new cluster.

Iterating this procedure for $r = 1, \dots, n-1$ yields a tree of nested partitions $\{\widehat{G}_1^{[r]}, \dots$ $\dots, \widehat{G}_{n-r}^{[r]}\}$, which can be graphically represented by a dendrogram. Roughly speaking, the HAC algorithm merges the n singleton clusters $\widehat{G}_i^{[0]} = \{i\}$ step by step until we end up with the cluster $\{1, \dots, n\}$. In each step of the algorithm, the closest two clusters are merged, where the distance between clusters is measured in terms of the dissimilarity $\widehat{\Delta}$. We refer the reader to Ward (1963) for an early reference on HAC clustering and to Section 14.3.12 in Hastie et al. (2009) for an overview of hierarchical clustering methods.

We now examine the properties of our HAC algorithm. In particular, we investigate how the partitions $\{\widehat{G}_1^{[r]}, \dots, \widehat{G}_{n-r}^{[r]}\}$ for $r = 1, \dots, n-1$ are related to the true class structure $\{G_1, \dots, G_{K_0}\}$. From (3.13) and (3.14), it immediately follows that the multiscale statistics $\widehat{d}_{ij}$ have the following property:

$$\mathbb{P}\Big(\max_{1 \le k \le K_0} \max_{i,j \in G_k} \widehat{d}_{ij} < \min_{1 \le k < k' \le K_0} \min_{\substack{i \in G_k, \\ j \in G_{k'}}} \widehat{d}_{ij} \Big) \to 1. \tag{4.2}$$

To formulate the results on the HAC algorithm, we do not restrict attention to the multiscale statistics $\widehat{d}_{ij}$ from (3.10) but let $\widehat{d}_{ij}$ denote any statistics with the high-level property (4.2). We further make use of the following notation: Let $\mathcal{A} = \{A_1, \ldots, A_r\}$ and $\mathcal{B} = \{B_1, \ldots, B_{r'}\}$ be two partitions of the set $\{1, \ldots, n\}$, that is, $\dot{\bigcup}_{\ell=1}^{r} A_\ell = \{1, \ldots, n\}$ and $\dot{\bigcup}_{\ell=1}^{r'} B_\ell = \{1, \ldots, n\}$. We say that $\mathcal{A}$ is a refinement of $\mathcal{B}$ if each $A_\ell \in \mathcal{A}$ is a subset of some $B_{\ell'} \in \mathcal{B}$. With this notation at hand, the properties of the HAC algorithm can be summarized as follows:

Theorem 4.1. *Suppose that the statistics $\widehat{d}_{ij}$ satisfy condition* (4.2). *Then*

(a) $\mathbb{P}\Big(\big\{\widehat{G}_1^{[n-K_0]}, \ldots, \widehat{G}_{K_0}^{[n-K_0]}\big\} = \big\{G_1, \ldots, G_{K_0}\big\}\Big) \to 1,$

(b) $\mathbb{P}\Big(\big\{\widehat{G}_1^{[n-K]}, \ldots, \widehat{G}_K^{[n-K]}\big\}$ *is a refinement of* $\big\{G_1, \ldots, G_{K_0}\big\}\Big) \to 1$ *for any* $K > K_0$,

(c) $\mathbb{P}\Big(\big\{G_1, \ldots, G_{K_0}\big\}$ *is a refinement of* $\big\{\widehat{G}_1^{[n-K]}, \ldots, \widehat{G}_K^{[n-K]}\big\}\Big) \to 1$ *for any* $K < K_0$.

The proof of Theorem 4.1 is trivial and thus omitted, the statements (a)–(c) being immediate consequences of condition (4.2). By (a), the partition $\{\widehat{G}_1, \ldots, \widehat{G}_{K_0}\}$ with $\widehat{G}_k = \widehat{G}_k^{[n-K_0]}$ for $1 \le k \le K_0$ is a consistent estimator of the true class structure $\{G_1, \ldots, G_{K_0}\}$ in the following sense: $\{\widehat{G}_1, \ldots, \widehat{G}_{K_0}\}$ coincides with $\{G_1, \ldots, G_{K_0}\}$ with probability tending to 1. Hence, if the number of classes K_0 were known, we could consistently estimate the true class structure by $\{\widehat{G}_1, \ldots, \widehat{G}_{K_0}\}$. The partitions $\{\widehat{G}_1^{[n-K]}, \ldots, \widehat{G}_K^{[n-K]}\}$ with $K \neq K_0$ can of course not serve as consistent estimators of the true class structure. According to (b) and (c), there is nevertheless a close link between these partitions and the unknown class structure. In particular, by (b), for any $K > K_0$, the estimated clusters $\widehat{G}_1^{[n-K]}, \ldots, \widehat{G}_K^{[n-K]}$ are subsets of the unknown classes with probability tending to 1. Conversely, by (c), for any $K < K_0$, the unknown classes are subsets of the estimated clusters with probability tending to 1.

5 Estimation of the unknown number of groups

5.1 The estimation method

Let $\widehat{\Delta}(S, S')$ be the dissimilarity measure from (4.1) and define the shorthand $\widehat{\Delta}(S) = \widehat{\Delta}(S, S)$. Moreover, let $\{\pi_{n,T}\}$ be any sequence with the property that

$$\sqrt{\log n + \log T} \ll \pi_{n,T} \ll \sqrt{T h_{\max}}, \tag{5.1}$$

where the notation $a_{n,T} \ll b_{n,T}$ means that $a_{n,T} = o(b_{n,T})$. Combining properties (3.11) and (3.12) of the multiscale distance statistics $\widehat{d}_{ij}$ with the statements of Theorem 4.1, we immediately obtain the following: For any $K < K_0$,

$$\mathbb{P}\Big(\max_{1 \le k \le K} \widehat{\Delta}\big(\widehat{G}_k^{[n-K]}\big) \le \pi_{n,T} \Big) \to 0, \tag{5.2}$$

whereas for $K = K_0$,

$$\mathbb{P}\Big(\max_{1 \le k \le K_0} \widehat{\Delta}\big(\widehat{G}_k^{[n-K_0]}\big) \le \pi_{n,T} \Big) \to 1. \tag{5.3}$$

Taken together, (5.2) and (5.3) motivate to estimate the unknown number of classes K_0 by the smallest number K for which the criterion

$$\max_{1 \le k \le K} \widehat{\Delta}\big(\widehat{G}_k^{[n-K]}\big) \le \pi_{n,T}$$

is satisfied. Formally speaking, we estimate K_0 by

$$\widehat{K}_0 = \min \Big\{ K = 1, 2, \ldots \,\Big|\, \max_{1 \le k \le K} \widehat{\Delta}\big(\widehat{G}_k^{[n-K]}\big) \le \pi_{n,T} \Big\}.$$

$\widehat{K}_0$ can be shown to be a consistent estimator of K_0 in the sense that $\mathbb{P}(\widehat{K}_0 = K_0) \to 1$. More precisely, we can prove the following result.

Theorem 5.1. *Suppose that the multiscale statistics $\widehat{d}_{ij}$ defined in* (3.10) *have the properties* (3.11) *and* (3.12)*. Moreover, let $\{\pi_{n,T}\}$ be any threshold sequence with the property* (5.1)*. Then it holds that $\mathbb{P}(\widehat{K}_0 = K_0) \to 1$.*

The proof of Theorem 5.1 is straightforward: As already noted, the properties (3.11) and (3.12) of the multiscale distance statistics and the statements of Theorem 4.1 immediately imply (5.2) and (5.3). From (5.2), it further follows that $\mathbb{P}(\widehat{K}_0 < K_0) = o(1)$, whereas (5.3) yields that $\mathbb{P}(\widehat{K}_0 > K_0) = o(1)$. As a consequence, we obtain that $\mathbb{P}(\widehat{K}_0 = K_0) \to 1$.

The estimator $\widehat{K}_0$ can be interpreted in terms of the dendrogram produced by the HAC algorithm. It specifies a simple cutoff rule for the dendrogram: The value

$$\max_{1 \le k \le K} \widehat{\Delta}\big(\widehat{G}_k^{[n-K]}\big) = \min_{1 \le k < k' \le K+1} \widehat{\Delta}\big(\widehat{G}_k^{[n-(K+1)]}, \widehat{G}_{k'}^{[n-(K+1)]}\big)$$

is the dissimilarity level at which two clusters are merged to obtain a partition with K clusters. In the dendrogram, the clusters are usually indicated by vertical lines and the dissimilarity level at which two clusters are merged is marked by a horizontal line which connects the two vertical lines representing the clusters. To compute the estimator $\widehat{K}_0$, we simply have to cut the dendrogram at the dissimilarity level $\pi_{n,T}$ and count the vertical lines that intersect the horizontal cut at the level $\pi_{n,T}$. See Figure 1 for an illustration.

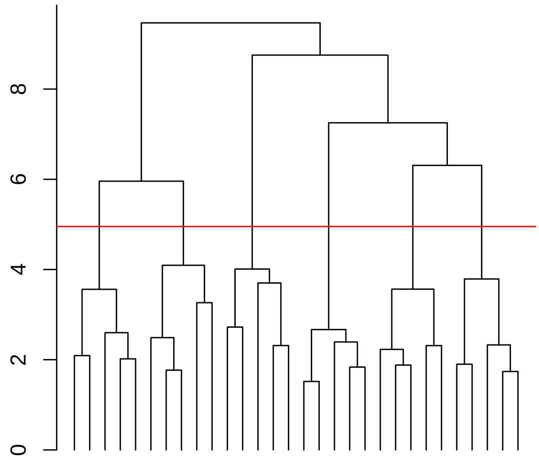


Figure 1: Example of a dendrogram produced by the HAC algorithm. The red horizontal line indicates the dissimilarity level $\pi_{n,T}$. The estimator $\widehat{K}_0$ can be computed by counting the vertical lines that intersect the red horizontal threshold. In the above example, $\widehat{K}_0$ is equal to 6.

5.2 Choice of the threshold level $\pi_{n,T}$

As shown in Theorem 5.1, $\widehat{K}_0$ is a consistent estimator of K_0 for any threshold sequence $\{\pi_{n,T}\}$ with the property that $\sqrt{\log n + \log T} \ll \pi_{n,T} \ll \sqrt{Th_{\max}}$. From an asymptotic perspective, we thus have a lot of freedom to choose the threshold $\pi_{n,T}$. In finite samples, a totally different picture arises. There, different choices of $\pi_{n,T}$ may result in markedly different estimates of K_0. Selecting the threshold level $\pi_{n,T}$ in a suitable way is thus a crucial issue in finite samples.

In what follows, we give some heuristic discussion on how to pick the threshold level $\pi_{n,T}$ appropriately in practice. To do so, we suppose that the technical conditions from Section 6 are fulfilled. In addition, we make the simplifying assumption that $\alpha_i = \gamma_t = 0$ for all i and t, that is, we drop the fixed effects from the model. Moreover, we suppose that the errors ε_{it} are homoskedastic and that the error variances $\sigma_i^2 = \mathbb{E}[\varepsilon_{it}^2]$ are the same within groups. As already discussed in Section 2.3, the densities f_i of the regressors X_{it} are supposed to be the same within groups as well. Slightly abusing notation, we write σ_k^2 and f_k to denote the group-specific error variance and regressor density in the k-th class G_k. We can now make the following heuristic observations:

(a) Consider any pair of time series i and j that belong to the same class G_k. As in (3.6), we decompose $\widehat{\psi}_{ij}(x,h)$ into a bias and a variance part according to $\widehat{\psi}_{ij}(x,h) = \widehat{\psi}_{ij}^B(x,h) + \widehat{\psi}_{ij}^V(x,h) +$ lower order terms. As already noted in Section 3.1, $\widehat{\psi}_{ij}^B(x,h) = 0$ for $i,j \in G_k$, which implies that

$$\widehat{\psi}_{ij}(x,h) \approx \widehat{\psi}_{ij}^V(x,h) = \sqrt{Th}\{\widehat{m}_{i,h}^V(x) - \widehat{m}_{j,h}^V(x)\}/\{\widehat{\nu}_{ij}(x,h)\}^{1/2}, \tag{5.4}$$

where $\widehat{m}_{i,h}^V(x) = \{\sum_{t=1}^T W_{it}(x,h)\varepsilon_{it}\}/\{\sum_{t=1}^T W_{it}(x,h)\}$ under our simplifying assumptions. Standard arguments for kernel smoothers suggest that

$$\widehat{m}^V_{i,h}(x) \approx \big\{ f_k(x) \big[\kappa_0(x,h)\kappa_2(x,h) - \kappa_1(x,h)^2 \big] \big\}^{-1} \\ \times \frac{1}{T} \sum_{t=1}^T K_h(X_{it} - x) \Big[\kappa_2(x,h) - \kappa_1(x,h) \Big(\frac{X_{it} - x}{h} \Big) \Big] \varepsilon_{it}, \qquad (5.5)$$

where $\kappa_\ell(x,h) = \int_{-x/h}^{(1-x)/h} u^\ell K(u) du$ for $0 \le \ell \le 2$. Since by construction, $\widehat{\nu}_{ij}(x,h)$ is an estimator of $\nu_{ij}(x,h) = 2\{\sigma_k^2/f_k(x)\} s(x,h)$ with $s(x,h)$ introduced in (3.5), we can combine (5.4) and (5.5) to obtain the approximation $\widehat{\psi}_{ij}(x,h) \approx \widehat{\psi}_i(x,h) - \widehat{\psi}_j(x,h)$ with

$$\widehat{\psi}_i(x,h) = \big\{ 2\rho(x,h)\sigma_k^2 f_k(x) \big\}^{-1/2} \\ \times \frac{1}{\sqrt{Th}} \sum_{t=1}^T K\Big(\frac{X_{it} - x}{h} \Big) \Big[\kappa_2(x,h) - \kappa_1(x,h) \Big(\frac{X_{it} - x}{h} \Big) \Big] \varepsilon_{it},$$

where we use the shorthand $\rho(x,h) = \int_{-x/h}^{(1-x)/h} K^2(u)[\kappa_2(x,h) - \kappa_1(x,h)u]^2 du$. For each i, we stack the random variables $\widehat{\psi}_i(x,h)$ with $(x,h) \in \mathcal{G}_T$ in the vector

$$\widehat{\boldsymbol{\psi}}_i = \Big(\widehat{\psi}_i\big(x_1^1, h_1\big), \ldots, \widehat{\psi}_i\big(x_1^{N_1}, h_1\big), \ldots\ldots, \widehat{\psi}_i\big(x_p^1, h_p\big), \ldots, \widehat{\psi}_i\big(x_p^{N_p}, h_p\big) \Big)^\top,$$

where $\mathcal{G}_T = \bigcup_{\nu=1}^p \mathcal{G}_{T,\nu}$ and $\mathcal{G}_{T,\nu} = \{(x_\nu^\ell, h_\nu) : 1 \le \ell \le N_\nu\}$ is the set of points corresponding to the bandwidth level h_ν. Moreover, we write $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_p)^\top$ with $\boldsymbol{\lambda}_\nu = (\lambda(2h_\nu), \ldots, \lambda(2h_\nu))$ being a vector of length N_ν for each ν and we introduce the notation $|z| = (|z_1|, \ldots, |z_q|)^\top$ and $(z)_\infty = \max_{1 \le \ell \le q} z_\ell$ for $z \in \mathbb{R}^q$. With this notation at hand, we obtain that

$$\widehat{d}_{ij} \approx \big(|\widehat{\boldsymbol{\psi}}_i - \widehat{\boldsymbol{\psi}}_j| - \boldsymbol{\lambda} \big)_\infty$$

for any pair of time series i and j that belong to the same class.

(b) For any fixed number of points $z_1, \ldots, z_q \in (0,1)$ and related bandwidths h_{z_ℓ} with $h_{\min} \le h_{z_\ell} \le h_{\max}$ for $1 \le \ell \le q$, the random vector $[\, \widehat{\psi}_i(z_1, h_{z_1}), \ldots, \widehat{\psi}_i(z_q, h_{z_q}) \,]^\top$ is asymptotically normal. Hence, the random vector $\widehat{\boldsymbol{\psi}}_i$ can be treated as approximately Gaussian for sufficiently large sample sizes. More specifically, since

$$\mathrm{Cov}\big(\widehat{\psi}_i(x,h), \widehat{\psi}_i(x',h')\big) \\ \approx \big\{ 2\sqrt{\rho(x,h)\rho(x',h')} \big\}^{-1} \sqrt{\frac{h}{h'}} \Big\{ \int_{-x/h}^{(1-x)/h} K(u) \big[\kappa_2(x,h) - \kappa_1(x,h)u \big] \\ \times K\Big(\frac{hu + x - x'}{h'} \Big) \Big[\kappa_2(x',h') - \kappa_1(x',h') \Big(\frac{hu + x - x'}{h'} \Big) \Big] du \Big\}, \qquad (5.6)$$

we can approximate the random vector $\widehat{\boldsymbol{\psi}}_i$ by a Gaussian vector with the covariance structure specified on the right-hand side of (5.6). Moreover, since the vectors $\widehat{\boldsymbol{\psi}}_i$ are independent across i under our assumptions, we can approximate the distribution of

$$\max_{i,j \in S} \big(|\widehat{\boldsymbol{\psi}}_i - \widehat{\boldsymbol{\psi}}_j| - \boldsymbol{\lambda} \big)_\infty$$

by that of

$$\max_{i,j \in S} \big(|\boldsymbol{\zeta}_i - \boldsymbol{\zeta}_j| - \boldsymbol{\lambda} \big)_\infty$$

for any $S \subseteq \{1, \ldots, n\}$, where $\boldsymbol{\zeta}_i$ are independent Gaussian random vectors with the covariance structure from (5.6).

Ideally, we would like to tune the threshold level $\pi_{n,T}$ such that $\widehat{K}_0 = K_0$ with high probability. Put differently, we would like to choose $\pi_{n,T}$ such that it is slightly larger than $\max_{1 \le k \le K_0} \widehat{\Delta}(\widehat{G}_k^{[n-K_0]})$ with high probability. With the help of the observations (a) and (b) as well as some further heuristic arguments, this can be achieved as follows: Since the partition $\{\widehat{G}_1^{[n-K_0]}, \ldots, \widehat{G}_{K_0}^{[n-K_0]}\}$ consistently estimates the class structure $\{G_1, \ldots, G_{K_0}\}$, we have that

$$\max_{1 \le k \le K_0} \widehat{\Delta}(\widehat{G}_k^{[n-K_0]}) \approx \max_{1 \le k \le K_0} \widehat{\Delta}(G_k). \tag{5.7}$$

By observation (a), we further obtain that

$$\begin{aligned} \max_{1 \le k \le K_0} \widehat{\Delta}(G_k) &= \max_{1 \le k \le K_0} \Big\{ \max_{i,j \in G_k} \widehat{d}_{ij} \Big\} \\ &\approx \max_{1 \le k \le K_0} \Big\{ \max_{i,j \in G_k} \big(|\widehat{\boldsymbol{\psi}}_i - \widehat{\boldsymbol{\psi}}_j| - \boldsymbol{\lambda} \big)_\infty \Big\}, \end{aligned} \tag{5.8}$$

and by (b),

$$\max_{1 \le k \le K_0} \Big\{ \max_{i,j \in G_k} \big(|\widehat{\boldsymbol{\psi}}_i - \widehat{\boldsymbol{\psi}}_j| - \boldsymbol{\lambda} \big)_\infty \Big\} \overset{d}{\approx} \max_{1 \le k \le K_0} \Big\{ \max_{i,j \in G_k} \big(|\boldsymbol{\zeta}_i - \boldsymbol{\zeta}_j| - \boldsymbol{\lambda} \big)_\infty \Big\}, \tag{5.9}$$

where $Z \overset{d}{\approx} Z'$ means that Z is approximately distributed as Z'. Since the right-hand side of (5.9) depends on the unknown groups $G_1, \ldots, G_{K_0}$, we apply the trivial bound

$$\begin{aligned} &\max_{1 \le k \le K_0} \Big\{ \max_{i,j \in G_k} \big(|\boldsymbol{\zeta}_i - \boldsymbol{\zeta}_j| - \boldsymbol{\lambda} \big)_\infty \Big\} \\ &\qquad \le B_n := \max_{1 \le i,j \le n} \big(|\boldsymbol{\zeta}_i - \boldsymbol{\zeta}_j| - \boldsymbol{\lambda} \big)_\infty \end{aligned} \tag{5.10}$$

and define $q_n(\alpha)$ to be the α-quantile of B_n. Taken together, (5.7)–(5.10) suggest that

$$\max_{1 \le k \le K_0} \widehat{\Delta}(\widehat{G}_k^{[n-K_0]}) \le q_n(\alpha)$$

holds with high probability if we pick α close to 1. In particular, if the random variable $\max_{1 \le k \le K_0} \widehat{\Delta}(\widehat{G}_k^{[n-K_0]})$ is not only approximately but exactly distributed as $\max_{1 \le k \le K_0} \max_{i,j \in G_k} (|\boldsymbol{\zeta}_i - \boldsymbol{\zeta}_j| - \boldsymbol{\lambda})_\infty$, then

$$\mathbb{P}\Big(\max_{1 \le k \le K_0} \widehat{\Delta}(\widehat{G}_k^{[n-K_0]}) \le q_n(\alpha) \Big) \ge \alpha.$$

According to these considerations, $\pi_{n,T} = q_n(\alpha)$ with α close to 1 should be an appropriate threshold level. Throughout the simulations and applications, we set $\alpha = 0.95$.

6 Theoretical results

In this section, we derive the statements (3.11) and (3.12) under appropriate regularity conditions. These statements characterize the convergence behaviour of the multiscale statistics $\widehat{d}_{ij}$ and underlie Theorems 4.1 and 5.1 which describe the theoretical properties of our clustering methods. To prove (3.11) and (3.12), we impose the following conditions.

(C1) The time series processes $\mathcal{P}_i = \{(X_{it}, \varepsilon_{it}) : t = 1, 2, \ldots\}$ are independent across i. Moreover, they are strictly stationary and strongly mixing for each i. Let $\alpha_i(\ell)$ for $\ell = 1, 2, \ldots$ be the mixing coefficients corresponding to the i-th time series $\mathcal{P}_i$. It holds that $\alpha_i(\ell) \le \alpha(\ell)$ for all i, where the coefficients $\alpha(\ell)$ decay exponentially fast to zero as $\ell \to \infty$.

(C2) For each $1 \le i \le n$, the random variables X_{it} have a density f_i with the following properties: (a) f_i has bounded support, which w.l.o.g. equals $[0, 1]$ for all i, (b) f_i is bounded away from zero and infinity on $[0, 1]$ uniformly over i, that is, $0 < c \le f_i(x) \le C < \infty$ for all $x \in [0, 1]$ with some constants c and C that neither depend on x nor on i, (c) f_i is twice continuously differentiable on $[0, 1]$ with first and second derivatives that are bounded away from infinity in absolute value uniformly over i. Moreover, the variables $(X_{it}, X_{it+\ell})$ have a joint density $f_{i,\ell}$ which is bounded away from infinity uniformly over i, that is, $f_{i,\ell}(x, x') \le C < \infty$ for all i, x, x' and ℓ, where the constant C neither depends on i, x, x' nor on ℓ.

(C3) The error terms ε_{it} are homoskedastic, that is, $\sigma_i^2 = \mathbb{E}[\varepsilon_{it}^2] = \mathbb{E}[\varepsilon_{it}^2 | X_{it} = x]$ for all $x \in [0, 1]$. The error variances σ_i^2 are uniformly bounded away from zero and infinity, that is, $0 < c \le \sigma_i^2 \le C < \infty$ for all i, where the constants c and C do not depend on i.

(C4) The densities f_i and the error variances σ_i^2 are the same within groups. That is, for any k with $1 \le k \le K_0$, it holds that $f_i = f_j$ and $\sigma_i^2 = \sigma_j^2$ for all $i, j \in G_k$.

(C5) There exist a real number $\theta > 4$ and a natural number ℓ^* such that for any $\ell \in \mathbb{Z}$ with $|\ell| \geq \ell^*$ and some constant $C < \infty$,

$$\max_{1 \le i \le n} \sup_{x \in [0,1]} \mathbb{E}\big[|\varepsilon_{it}|^{\theta} \big| X_{it} = x\big] \le C < \infty$$
$$\max_{1 \le i \le n} \sup_{x,x' \in [0,1]} \mathbb{E}\big[|\varepsilon_{it}\varepsilon_{it+\ell}| \big| X_{it} = x, X_{it+\ell} = x'\big] \le C < \infty.$$

(C6) The group-specific regression functions g_k are twice continuously differentiable on $[0, 1]$ for $1 \le k \le K_0$ with Lipschitz continuous second derivatives g_k'', that is, $|g_k''(v) - g_k''(w)| \le L|v - w|$ for any $v, w \in [0, 1]$ and some constant L. Moreover, for any pair of indices (k, k') with $1 \le k < k' \le K_0$, the functions g_k and $g_{k'}$ are different in the sense that $g_k(x) \neq g_{k'}(x)$ for some point $x \in [0, 1]$.

(C7) It holds that

$$n = n(T) \le C \frac{(T^{1/2} \wedge T h_{\min})^{\frac{\theta - \delta}{2}}}{T^{1+\delta}} \tag{6.1}$$

for some small $\delta > 0$ and a sufficiently large constant $C > 0$, where we use the notation $a \wedge b = \min\{a, b\}$ and θ is defined in (C5).

(C8) The minimal and maximal bandwidths have the form $h_{\min} = aT^{-B}$ and $h_{\max} = AT^{-b}$ with some positive constants a, A, b and B, where $0 < b \le B < 1$.

(C9) The kernel K is non-negative, bounded and integrates to one. Moreover, it is symmetric about zero, has compact support $[-1, 1]$ and fulfills the Lipschitz condition that $|K(v) - K(w)| \le L|v - w|$ for some L and all $v, w \in \mathbb{R}$.

Remark 6.1. We briefly comment on the above assumptions.

(i) (C1) imposes some weak dependence conditions on the variables $(X_{it}, \varepsilon_{it})$ across t in the form of mixing assumptions. Note that we do not necessarily require exponentially decaying mixing rates as assumed in (C1). These could alternatively be replaced by sufficiently high polynomial rates. We nevertheless make the stronger assumption of exponential mixing to keep the proofs as clear as possible. (C1) further restricts the regressors X_{it} and the errors ε_{it} to be independent across i. Some restricted types of cross-sectional dependence in the data are however possible via the fixed effect error terms α_i and γ_t.

(ii) The homoskedasticity assumption in (C3) as well as the condition in (C4) that the error variances σ_i^2 are the same within groups are not necessarily needed but are imposed for simplicity. The restriction in (C4) that the densities f_i are the same within groups, in contrast, is required for identification purposes as already discussed in Section 2.3.

(iii) (C2), (C5) and (C6) are standard moment, boundedness and smoothness conditions to derive uniform convergence results for the kernel estimators on which the multiscale statistics $\widehat{d}_{ij}$ are based; see Hansen (2008) for similar assumptions.

(iv) (C7) imposes restrictions on the growth of the number of time series n. Loosely speaking, it says that n is not allowed to grow too quickly in comparison to T. More specifically, let $h_{\min} = aT^{-B}$ with some $B \leq 1/2$ and $h_{\max} = AT^{-b}$ with some $b > 0$. In this case, (6.1) simplifies to $n \leq CT^{(\theta-4-5\delta)/4}$ with some small $\delta > 0$. This shows that the growth restriction (6.1) on n is closely related to the moment conditions on the error terms ε_{it} in (C5). In particular, the larger the value of θ, that is, the stronger the moment conditions on ε_{it}, the faster n may grow in comparison to T. If $\theta = 8$, for example, then n may grow (almost) as quickly as T. If θ can be picked arbitrarily large, that is, if all moments of ε_{it} exist, then n may grow as quickly as any polynomial of T, that is, $n \leq CT^{\rho}$ with $\rho > 0$ as large as desired.

(v) (C8) imposes some conditions on the minimal and maximal bandwidths $h_{\min}$ and $h_{\max}$. Specifically, it requires that $h_{\min} \geq cT^{-(1-\delta)}$ and $h_{\max} \leq CT^{-\delta}$ for some small $\delta > 0$ and positive constants c and C. These conditions are fairly weak as already discussed in Section 3: According to them, we can choose $h_{\min}$ to converge to zero extremely fast, in particular much faster than the optimal bandwidths for estimating the functions m_i, which are of the order $T^{-1/5}$ for any i under the smoothness conditions (C2) and (C6). Similarly, we can let $h_{\max}$ converge to zero much more slowly than the optimal bandwidths. Hence, we can choose the interval $[h_{\min}, h_{\max}]$ to be very large, allowing for both substantial under- and oversmoothing.

(vi) Finally, it is worth noting that our assumptions do not impose any restrictions on the class sizes $|G_k|$. The sizes $|G_k|$ may thus be very different across the classes G_k. In particular, they may be fixed for some classes and grow to infinity at different rates for others.

Under the regularity conditions just discussed, we can derive the following result whose proof is provided in the Supplementary Material.

Theorem 6.1. *Under (C1)–(C9), it holds that*

$$\max_{1 \leq k \leq K_0} \max_{i,j \in G_k} \widehat{d}_{ij} = O_p\big(\sqrt{\log n + \log T}\big) \tag{6.2}$$

$$\min_{1 \leq k < k' \leq K_0} \min_{\substack{i \in G_k, \\ j \in G_{k'}}} \widehat{d}_{ij} \geq c_0 \sqrt{T h_{\max}} + o_p\big(\sqrt{T h_{\max}}\big), \tag{6.3}$$

where c_0 is a fixed positive constant that does not depend on T (nor on $n = n(T)$).

7 Simulations

In this section, we carry out some simulations to illustrate the advantages of our multiscale approach over clustering methods that depend on a specific bandwidth. When the grid $\mathcal{G}_T$ of location-scale points (x,h) comprises only one bandwidth value h, our multiscale approach reduces to a bandwidth-dependent procedure. Specifically, the resulting procedure consists in applying a hierarchical clustering algorithm to the supremum distances $\widehat{d}_{ij}(h) = \max_{x \in \mathcal{X}} |\widehat{\psi}_{ij}(x,h)|$, where $\mathcal{X}$ is the set of locations under consideration and h is the chosen bandwidth.[3] In what follows, we compare our multiscale approach with this bandwidth-dependent procedure for several bandwidth values h.

We consider the following setup for the simulations: The data are drawn from the model

$$Y_{it} = m_i(X_{it}) + \varepsilon_{it} \quad (1 \le t \le T,\, 1 \le i \le n), \tag{7.1}$$

where $T = 1000$ and $n = 100$. The time series $i \in \{1,\ldots,n\}$ belong to $K_0 = 5$ different groups $G_1,\ldots,G_{K_0}$ of the same size. In particular, we set $G_k = \{(k-1)n/5+1,\ldots,kn/5\}$ for $1 \le k \le K_0 = 5$. The group-specific regression functions $g_k : [0,1] \to \mathbb{R}$ are given by $g_1(x) = 0$ and

$$\begin{aligned} g_2(x) &= 0.35\, b\big(x, \tfrac{1}{4}, \tfrac{1}{4}\big) \qquad & g_4(x) &= 2\, b\big(x, \tfrac{1}{4}, \tfrac{1}{40}\big) \\ g_3(x) &= 0.35\, b\big(x, \tfrac{3}{4}, \tfrac{1}{4}\big) \qquad & g_5(x) &= 2\, b\big(x, \tfrac{3}{4}, \tfrac{1}{40}\big), \end{aligned}$$

where $b(x,x_0,h) = 1(|x-x_0|/h \le 1)\,\{1-((x-x_0)/h)^2\}^2$. Figure 2 provides a graphical illustration of the functions g_k for $1 \le k \le 5$. The error process $\mathcal{E}_i = \{\varepsilon_{it} : 1 \le t \le T\}$ has an autoregressive (AR) structure for each i, in particular $\varepsilon_{it} = a\varepsilon_{it-1} + \eta_{it}$ for $1 \le t \le T$, where a is the AR parameter and the innovations η_{it} are i.i.d. normal with $\mathbb{E}[\eta_{it}] = 0$ and $\mathbb{E}[\eta_{it}^2] = \nu^2$. We consider two different values for the AR parameter a, in particular $a = -0.25$ and $a = 0.25$. The innovation variance ν^2 is chosen as $\nu^2 = 1-a^2$, which implies that $\mathrm{Var}(\varepsilon_{it}) = 1$. The regressors X_{it} are drawn independently from a uniform distribution on $[0,1]$ for each i. As can be seen, there is no time series dependence in the regressors, and we do not include fixed effects α_i and γ_t in the model. We do not take into account these complications because the main aim of the simulations is to display the advantages of our multiscale approach over bandwidth-dependent procedures. These advantages can be seen most clearly in a simple stylized simulation setup as the one under consideration.

To implement our multiscale approach, we use the location-scale grid $\mathcal{G}_T = \{(x,h) : x \in \mathcal{X} \text{ and } h \in \mathcal{H}\}$, where $\mathcal{X} = \{x : x = r/100 \text{ for } r = 5,\ldots,95\}$ is the set of

[3] Note that the additive correction term $\lambda(2h)$ can be dropped as it is a fixed constant when only one bandwidth value h is considered.

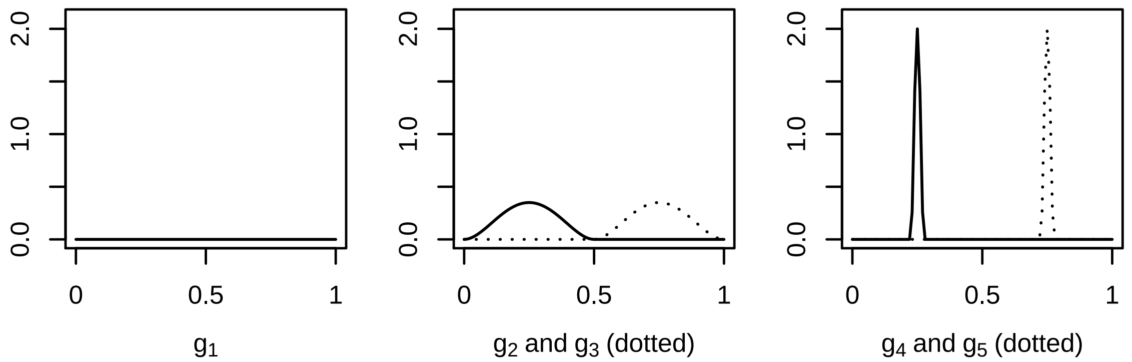
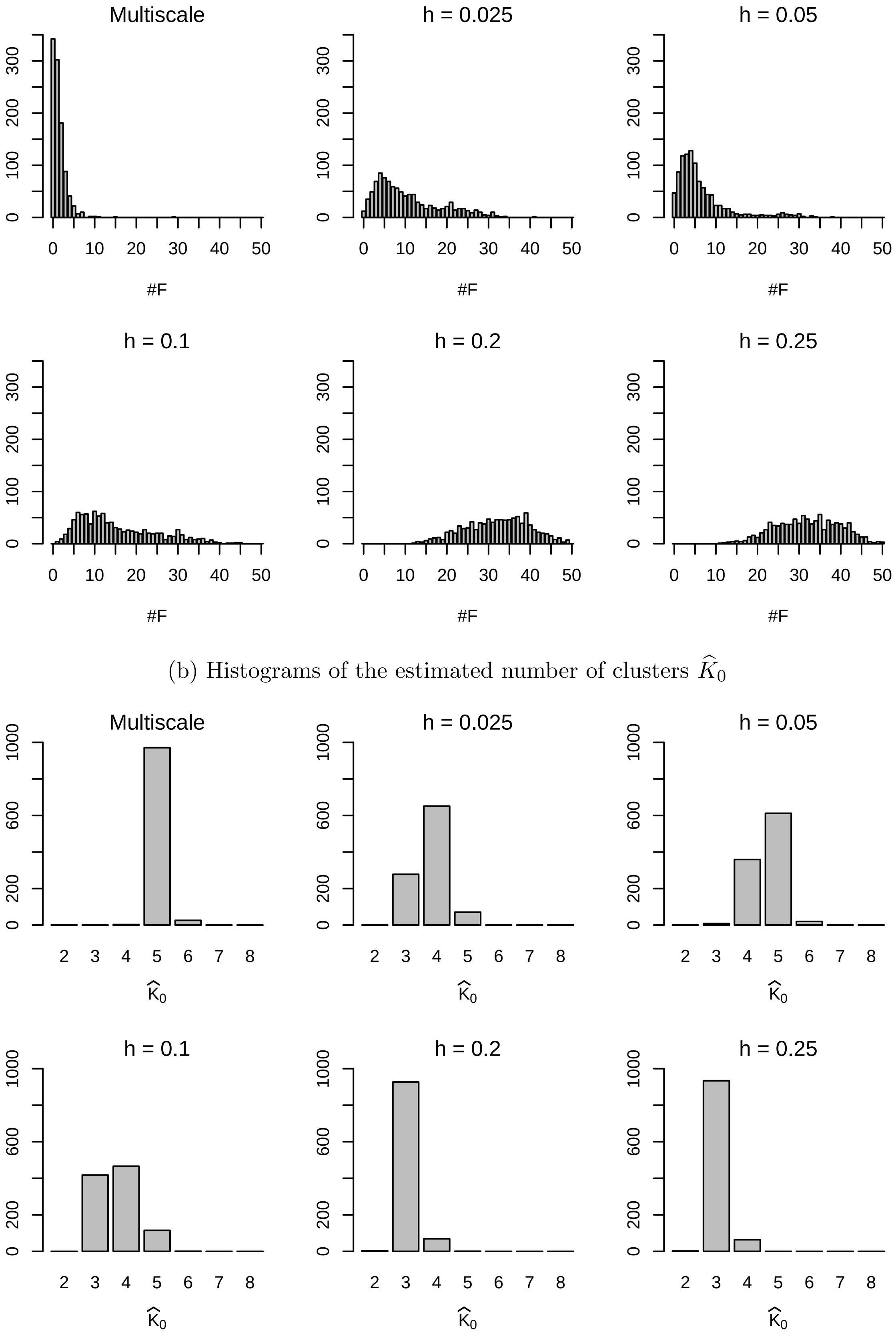


Figure 2: Plot of the functions g_k for $1 \le k \le 5$.

locations and $\mathcal{H} = \{h : 0.025 \le h \le 0.25 \text{ with } h = 0.025k \text{ for } k = 1, 2, \ldots\}$ is the set of bandwidths. The bandwidth-dependent algorithm is implemented with the same set of locations $\mathcal{X}$ and five different bandwidth values h, in particular $h \in \{0.025, 0.05, 0.1, 0.2, 0.25\}$. The number of classes $K_0 = 5$ is estimated as described in Section 5 both when the multiscale and the bandwidth-dependent algorithm is used. The threshold parameter $\pi_{n,T}$ is set to $\pi_{n,T} = q_n(\alpha)$ with $\alpha = 0.95$. To produce our simulation results, we draw $S = 1000$ samples from model (7.1) and compute the estimates of the classes $G_1, \ldots, G_{K_0}$ and their number K_0 for each simulated sample both for the multiscale and the bandwidth-dependent algorithm.

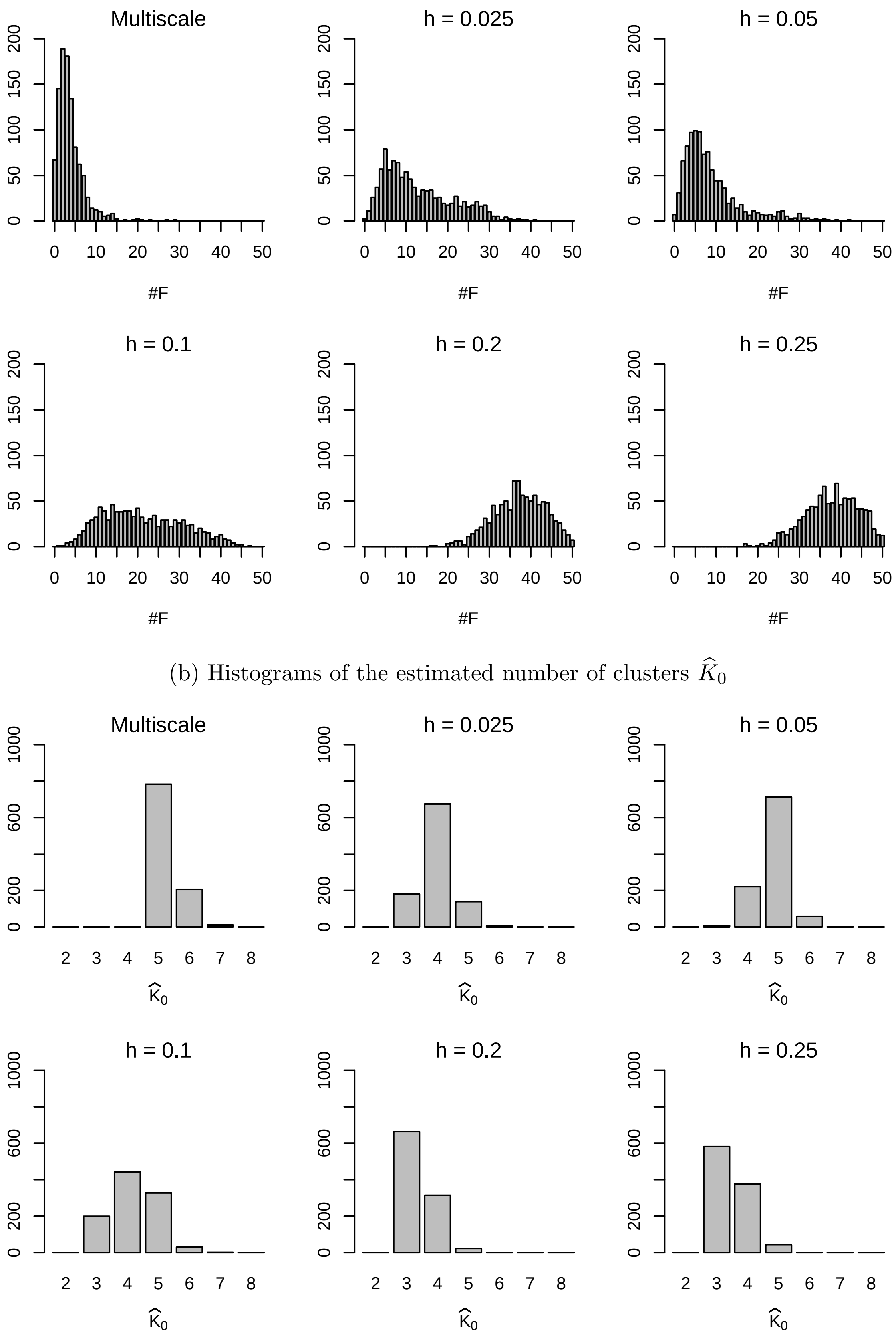
The simulation results for the scenario with the negative AR parameter $a = -0.25$ are reported in Figure 3 and those for the scenario with the positive parameter $a = 0.25$ in Figure 4. We first have a closer look at the results in Figure 3. To produce Figure 3a, we treat K_0 as known and compute the number of classification errors $\#F$, that is, the number of wrongly classified indices i for each of the $S = 1000$ simulated samples.[4] The upper left panel of Figure 3a shows the histogram of these $S = 1000$ values for our multiscale approach. The other panels of Figure 3a present the corresponding histograms for the bandwidth-dependent algorithm with the five different bandwidth values h under consideration. As can be seen very clearly, our multiscale approach performs much better than the bandwidth-dependent competitor for any of the considered bandwidths. Figure 3b shows the simulation results for the estimated number of classes $\widehat{K}_0$. The upper left panel depicts the histogram of the $S = 1000$ values of $\widehat{K}_0$ produced by the multiscale approach. As one can see, the estimate $\widehat{K}_0$ equals the true number of classes $K_0 = 5$ in about 95% of the

[4] Precisely speaking, $\#F$ is defined as follows: Let π be some permutation of the class labels $\{1, \ldots, K_0\}$ and denote the set of all possible permutations by Π. Moreover, denote the group membership of index i by $\rho(i)$, i.e. set $\rho(i) = k$ if $i \in G_k$. Similarly, let $\widehat{\rho}_\pi(i)$ be the estimated group membership of index i, where the estimated classes are labelled according to the permutation π. More specifically, set $\widehat{\rho}_\pi(i) = \pi(k)$ if $i \in \widehat{G}_k$. With this notation at hand, we define $\#F = \min_{\pi \in \Pi} \sum_{i=1}^n 1(\rho(i) \ne \widehat{\rho}_\pi(i))$.


(a) Histograms of the number of classification errors $\#F$

(b) Histograms of the estimated number of clusters $\widehat{K}_0$

Figure 3: Simulation results for the design with the negative AR parameter $a = -0.25$. In both subfigures (a) and (b), the upper left panel shows the results for our multiscale approach and the other panels those for the bandwidth-dependent competitor with different bandwidths h.



(a) Histograms of the number of classification errors $\#F$

(b) Histograms of the estimated number of clusters $\widehat{K}_0$

Figure 4: Simulation results for the design with the positive AR parameter $a = 0.25$. In both subfigures (a) and (b), the upper left panel shows the results for our multiscale approach and the other panels those for the bandwidth-dependent competitor with different bandwidths h.

cases (that is, in about 950 out of $S = 1000$ simulations). The performance of the bandwidth-dependent algorithm is considerably worse, which becomes apparent upon inspecting the other panels of Figure 3b. The results in Figure 4 for the scenario with the positive AR parameter $a = 0.25$ give a very similar picture. In particular, our multiscale approach shows a much better performance than the bandwidth-dependent competitor for any of the considered bandwidths. Comparing Figures 3 and 4, one can further see that the estimation precision is a bit better for the negative than the positive AR parameter (both for the multiscale and the bandwidth-dependent approach). This is not very surprising but simply reflects the fact that it is more difficult for the procedures to handle positive rather than negative correlation in the error terms.

Overall, our multiscale approach clearly outperforms the bandwidth-dependent algorithm in the simulation setup under consideration. Heuristically, this can be explained as follows: The setup comprises two very different types of signals. The signals g_4 and g_5 are very local in nature; they differ from a flat line only by a sharp, very local spike. The signals g_2 and g_3, in contrast, are much more global in nature; they differ from a flat line on a large part of the support $[0, 1]$, but they are much smaller in magnitude than g_4 and g_5. A bandwidth-dependent clustering algorithm is hardly able to distinguish these signals reliably from each other. When a small bandwidth value is used, local features of the functions (the spikes in g_4 and g_5) can be detected reliably, but more global features (the slight curvature in g_2 and g_3) are hard to see. Hence, when implemented with a small bandwidth, the algorithm is barely able to detect the global differences between the functions. When implemented with a large bandwidth, in contrast, it is hardly able to capture the local differences. Our multiscale approach, in contrast, is able to produce appropriate estimates since it analyzes the data on various scales simultaneously.

Even though we have considered a quite stylized setup in our simulations, the advantages of our multiscale approach that become visible in this setup can be expected to persist in real-data applications. In practice, it is usually not known whether the group-specific regression functions g_k $(1 \le k \le K_0)$ differ on a local or global scale. Hence, it is usually not clear at all which bandwidth is appropriate for implementing a bandwidth-dependent clustering algorithm. If the bandwidth is not picked suitably, the clustering results may not be very accurate. Moreover, when the functions g_k differ on multiple scales, a clustering approach which is based on a single bandwidth h can be expected to perform not very well, regardless of the specific value of h. Our multiscale approach, in contrast, can be expected to produce reliable clustering results, no matter whether the functions g_k differ on a local, global or multiple scales.

8 Application

In what follows, we revisit the application example from Vogt and Linton (2017). The aim of this example is to investigate the effect of trading venue fragmentation on market quality in the European stock market. For each stock i in the FTSE 100 and FTSE 250 index, we observe a time series $\mathcal{T}_i = \{(Y_{it}, X_{it}) : 1 \le t \le T\}$ of weekly data from May 2008 to June 2011, where Y_{it} is a measure of market quality and X_{it} a measure of fragmentation for stock i at time t. More specifically, Y_{it} denotes the logarithmic volatility level of stock i at time t, where volatility is measured by the so-called high-low range, which is defined as the difference between the highest and the lowest price of the stock at time t divided by the latter. As a measure of fragmentation, we use the so-called Herfindahl index. The Herfindahl index of stock i at time t is defined as the sum of the squared market shares of the venues where the stock is traded at time t. It thus takes values between 0 and 1. If X_{it} takes a value close to 0, there is strong fragmentation in stock i at time t, that is, stock i is traded at many different venues at time t. A value of X_{it} close to 1, in contrast, indicates little fragmentation, that is, stock i is traded only at a few venues at time t. The measures Y_{it} and X_{it} are constructed from data provided by Fidessa and Datastream. More details on the underlying data set and on variable construction can be found in Boneva et al. (2015) and Boneva et al. (2016).

For each stock i, we model the relationship between Y_{it} and X_{it} by the nonparametric regression equation

$$Y_{it} = m_i(X_{it}) + u_{it}, \tag{8.1}$$

where the error term has the fixed effects structure $u_{it} = \alpha_i + \gamma_t + \varepsilon_{it}$. The function m_i captures the effect of trading-venue fragmentation on market quality for stock i. It is quite plausible to suppose that there are groups of stocks for which this effect is fairly similar. We thus impose a formal group structure on the stocks in our sample. In particular, we suppose that there are K_0 groups of stocks $G_1, \ldots, G_{K_0}$ such that $m_i = g_k$ for all $i \in G_k$ and all $1 \le k \le K_0$, where g_k denotes the group-specific regression function associated with group G_k. Hence, we model the effect of fragmentation on market quality to be the same for all stocks in a given group.

We now use our multiscale clustering methods to estimate the unknown groups $G_1, \ldots, G_{K_0}$ along with their unknown number K_0 from the data sample at hand. As in Vogt and Linton (2017), we drop stocks from the sample for which data points are missing. Moreover, we eliminate stocks i with a very small empirical support $\mathcal{S}_i$ of the fragmentation data $\{X_{it} : 1 \le t \le T\}$. In particular, we only take into account stocks i for which the support $\mathcal{S}_i$ contains the interval $[0.275, 0.8]$. We thus use exactly the same data set as in Vogt and Linton (2017), which comprises $n = 125$ time series of length $T = 151$ weeks. To implement our multiscale methods, we employ the location-scale

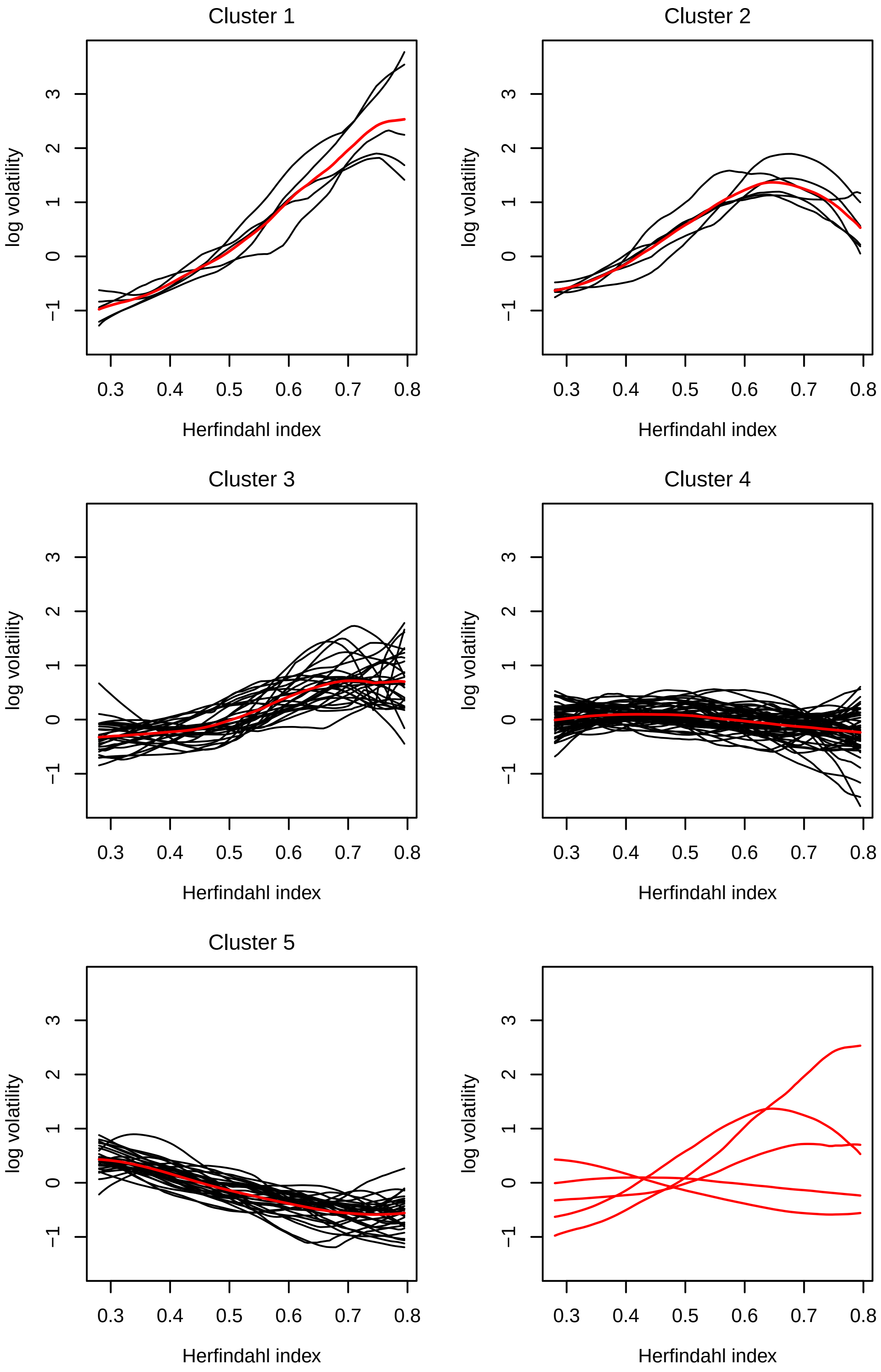


Figure 5: Estimated clusters in the application example of Section 8. Each panel corresponds to one cluster. The black lines are the estimated regression curves $\widehat{m}_{i,h}$ that belong to the respective cluster. The red lines are estimates of the group-specific regression functions. These are plotted once again together in the lower right panel of the figure.

grid $\mathcal{G}_T = \{(x,h) : x \in \mathcal{X} \text{ and } h \in \mathcal{H}\}$, where $\mathcal{X} = \{x : x = r/100 \text{ for } r = 1,\dots,99\}$ is the set of locations and $\mathcal{H} = \{h : 0.175 \le h \le 0.5 \text{ with } h = 0.025k \text{ for } k = 1,2,\dots\}$ is the set of bandwidths. Note that $h = 0.175$ is the smallest possible bandwidth we can use: If we pick h smaller than 0.175, we cannot compute the statistics $\widehat{\psi}_{ij}(x,h)$ for all stocks i and locations $x \in \mathcal{X}$ any more because for some i and x, there are less than two data points in the bandwidth window. The threshold parameter for the estimation of K_0 is set to $\pi_{n,T} = q_n(\alpha)$ with $\alpha = 0.95$.

The estimation results are presented in Figure 5. Each panel of the figure corresponds to one of the estimated groups $\widehat{G}_k^{[\widehat{K}_0]}$ for $1 \le k \le \widehat{K}_0$, where the estimated number of groups is $\widehat{K}_0 = 5$. In particular, each panel depicts the estimated curves $\widehat{m}_{i,h}$ that belong to some cluster $\widehat{G}_k^{[\widehat{K}_0]}$. The red curve in each panel is an estimate $\widehat{g}_{k,h}$ of the group-specific regression function g_k. More specifically, we define

$$\widehat{g}_{k,h}(x) = \frac{1}{\widehat{G}_k^{[\widehat{K}_0]}} \sum_{i \in \widehat{G}_k^{[\widehat{K}_0]}} \widehat{m}_{i,h}(x),$$

that is, we simply average the fits $\widehat{m}_{i,h}$ with $i \in \widehat{G}_k^{[\widehat{K}_0]}$. The estimates $\widehat{g}_{k,h}(x)$ are once again plotted together in the lower right panel of Figure 5. Whereas we do not need to specify a bandwidth h to compute the multiscale estimates $\widehat{G}_k^{[\widehat{K}_0]}$ of the unknown groups for $1 \le k \le \widehat{K}_0$, the kernel smoothers $\widehat{m}_{i,h}$ of course depend on a specific bandwidth h. As these smoothers are only computed for illustrative purposes, in particular for the graphical illustration of the results in Figure 5, we use the same bandwidth h for all stocks i. In particular, we choose the bandwidth adhoc as $h = 0.25$ for all i, which produces a good visual impression of the results.

In order to interpret the results in Figure 5, we regard volatility as a bad, meaning that higher volatility implies lower market quality. As can be seen, the effect of fragmentation on volatility is quite moderate for most stocks: Most of the curve fits in Cluster 4 are close to a flat line, whereas those in Clusters 3 and 5 slightly slope upwards and downwards, respectively. In contrast to this, the fits in Cluster 1 and to a lesser extent also those in Cluster 2 exhibit a strong increase. This indicates that higher fragmentation is accompanied by lower volatility and thus higher market quality for these stocks. To summarize, fragmentation appears to substantially improve market quality only for a small share of stocks (in particular for those in Clusters 1 and 2), whereas the effect of fragmentation is quite moderate for the great bulk of stocks (in particular for those in Clusters 3, 4 and 5). These findings are in line with those in Vogt and Linton (2017). Indeed, the clusters produced by our multiscale method are fairly similar to those obtained there. Hence, our multiscale approach confirms the results of the bandwidth-dependent algorithm from Vogt and Linton (2017), but without the need to go through the complicated bandwidth-selection procedure from there which may very well perform less accurate in other applications.

9 Extensions and modifications

9.1 Extension to the multivariate case and to other model settings

Throughout the paper, we have restricted attention to real-valued regressors X_{it}. Our approach extends to $\mathbb{R}^d$-valued regressors $X_{it} = (X_{it,1}, \ldots, X_{it,d})^\top$ in a straightforward way. The clustering methods described in Sections 4 and 5 remain the same in the multivariate case, only the multiscale statistics $\widehat{d}_{ij}$ need to be adjusted. To do so, we simply need to (i) replace the involved kernel estimators by multivariate versions and (ii) modify the scaling factors $\widehat{\nu}_{ij}(x, h)$ appropriately to normalize the variance of the statistics $\widehat{\psi}_{ij}(x, h)$. We neglect the details as these modifications are very straightforward.

The kernel smoothers on which the multiscale statistics $\widehat{d}_{ij}$ are based suffer from the usual curse of dimensionality. Hence, our fully nonparametric approach is only useful in practice as long as the dimension d of the regressors is moderate. If d is large, it makes sense to resort to structured nonparametric or semiparametric approaches. As an example, consider the partially linear model

$$Y_{it} = m_i(X_{it}) + \boldsymbol{\beta}^\top Z_{it} + u_{it}, \tag{9.1}$$

where X_{it} is real-valued, $Z_{it} = (Z_{it,1}, \ldots, Z_{it,d})^\top$ is an $\mathbb{R}^d$-valued vector and the error terms u_{it} have the fixed effects structure $u_{it} = \alpha_i + \gamma_t + \varepsilon_{it}$ with $\mathbb{E}[\varepsilon_{it}|X_{it}, Z_{it}] = 0$. In this model, Z_{it} is a vector of controls which enters the equation (9.1) linearly for simplicity. In particular, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^\top$ is an unknown parameter vector which is assumed to be the same for all i. Suppose we are mainly interested in the effect of X_{it} on the response Y_{it}, which is captured by the functions m_i. As in Section 2, we may model this effect by imposing a group structure on the curves m_i: We may suppose that there exist classes $G_1, \ldots, G_{K_0}$ and associated functions $g_1, \ldots, g_{K_0}$ such that $m_i = g_k$ for all $i \in G_k$ and $1 \le k \le K_0$. In order to apply our estimation methods in this context, we merely need to adjust the multiscale statistics $\widehat{d}_{ij}$. In particular, we need to replace the local linear smoothers $\widehat{m}_{i,h}(x)$ by appropriate estimators of m_i and adjust the scaling factors $\widehat{\nu}_{ij}(x, h)$. The functions m_i may for example be estimated with the help of the methods developed in Robinson (1988). Once the multiscale statistics $\widehat{d}_{ij}$ have been adjusted to the partially linear model setting (9.1), estimators of the unknown classes and their unknown number can be obtained as described in Sections 4 and 5. We conjecture that the two main Theorems 4.1 and 5.1 on the multiscale clustering methods remain to hold true in the context of the partially linear model (9.1). However, extending our theoretical results to model (9.1) is by no means trivial but would require a substantial deal of additional work.

Another interesting model setting to which our methods can be extended is the following: Suppose that

$$Y_{it} = m_i(X_{it}) + \alpha_i + \gamma_t + \varepsilon_{it}, \tag{9.2}$$

where X_{it} is a continuous treatment effect that is applied to unit i during periods $\tau_i \subset \{1, \ldots, T\}$ and is not present during pre- and post-treatment periods. Moreover, suppose that there is a matched control group that never receives the treatment and so satisfies

$$Y_{j(i),t} = \alpha_{j(i)} + \gamma_t + \varepsilon_{j(i),t}, \tag{9.3}$$

where $j(i)$ denotes the unit in the control group which is matched with i. This setting is similar to that considered in Boneva et al. (2018) who evaluate the effects of the UK government's corporate bond purchase scheme on market quality measures such as liquidity. The treatment in this study is continuously distributed and applied during an 18-month-period to a subset of all UK listed corporate bonds. The authors assume a linear homogeneous treatment effect in the baseline model and apply difference-in-difference methods to estimate the effect. However, one could easily allow for more general nonlinear and heterogeneous effects m_i as in equation (9.2) and impose a group structure on them. Notice that for $t \in \tau_i$ and $s \in \tau_i^c$, we have

$$(Y_{it} - Y_{j(i),t}) - (Y_{is} - Y_{j(i),s}) = m_i(X_{it}) + \varepsilon_{it} - \varepsilon_{j(i),t} - \varepsilon_{is} + \varepsilon_{j(i),s}, \tag{9.4}$$

which is essentially a nonparametric regression equation for each i. We could thus apply our methods to the difference-in-difference equation (9.4).

9.2 Alternatives to hierarchical clustering

In order to estimate the unknown class structure in model (2.1)–(2.2), we have combined the multiscale statistics $\widehat{d}_{ij}$ with a hierarchical clustering algorithm. It is also possible to combine them with other distance-based clustering approaches. In particular, they can be employed as distance statistics in the thresholding algorithm of Vogt and Linton (2017). To do so, we replace the L_2-type distance statistics $\widehat{\Delta}_{ij}$ from Vogt and Linton (2017) by the multiscale statistics $\widehat{d}_{ij}$ and construct the threshold estimators of the unknown groups $G_1, \ldots, G_{K_0}$ and of their unknown number K_0 exactly as described in Section 2.2 of Vogt and Linton (2017). This leads to estimators $\widetilde{K}_0$ and $\widetilde{G}_1, \ldots, \widetilde{G}_{\widetilde{K}_0}$, which unlike those constructed in Vogt and Linton (2017) are free of classical bandwidth parameters.

Under regularity conditions very similar to those from Section 6, we can derive some basic theoretical properties of the estimators $\widetilde{K}_0$ and $\widetilde{G}_1, \ldots, \widetilde{G}_{\widetilde{K}_0}$: Suppose that the threshold parameter $\tau_{n,T}$ of the procedure fulfills Condition 6 from Section 3.2 of Vogt and Linton (2017), that is, $\tau_{n,T} \searrow 0$ such that $\max_{i,j \in G_k} \widehat{d}_{ij} \leq \tau_{n,T}$ with probability tending to 1 for all k. Then it can be shown that $\mathbb{P}(\widetilde{K}_0 = K_0) \to 1$ as well as $\mathbb{P}(\{\widetilde{G}_1, \ldots, \widetilde{G}_{\widetilde{K}_0}\} = \{G_1, \ldots, G_{K_0}\}) \to 1$.

To implement the estimators $\widetilde{K}_0$ and $\widetilde{G}_1, \ldots, \widetilde{G}_{\widetilde{K}_0}$ in practice, we need to choose the threshold level $\tau_{n,T}$. In view of Condition 6 from Vogt and Linton (2017), we would like to tune $\tau_{n,T}$ such that $\max_{i,j \in G_k} \widehat{d}_{ij} \le \tau_{n,T}$ holds with high probability for all k. According to our heuristic arguments from Section 5.2, this may be achieved by setting $\tau_{n,T} = q_n(\alpha)$ with α close to 1. We thus suggest to choose the threshold parameter $\tau_{n,T}$ in the same way as the dissimilarity level $\pi_{n,T}$ at which we cut the dendrogram to estimate K_0.

9.3 Letting K_0 grow with the sample size

Throughout the paper, we have assumed that the number of classes K_0 is fixed. We now allow K_0 to grow with the number of time series n, that is, we admit of $K_0 = K_{0,n} \to \infty$ as $n \to \infty$. To deal with this situation, we require the group-specific regression functions g_k to fulfill the following additional condition:

(C10) The functions g_k as well as their first and second derivatives are uniformly bounded in absolute value, that is, $|g_k^{(\ell)}(x)| \le C$ for all $x \in [0,1]$ and $\ell = 0, 1, 2$, where $g_k^{(\ell)}$ denotes the ℓ-th derivative of g_k and the constant $C < \infty$ does not depend on k. Moreover,

$$\min_{1 \le k < k' \le K_0} \max_{\{x \,:\, (x, h_{\max}) \in \mathcal{G}_T\}} |g_k(x) - g_{k'}(x)| \gg \frac{\sqrt{\log n + \log T} + \sqrt{T h_{\max}^5}}{\sqrt{T h_{\max}}}. \tag{9.5}$$

As before, the expression $a_{n,T} \gg b_{n,T}$ means that $b_{n,T} = o(a_{n,T})$ and the notation $a_{n,T} \ll b_{n,T}$ is used analogously. (9.5) essentially says that the regression functions g_k and $g_{k'}$ of two different classes do not approach each other too quickly as $n \to \infty$. If condition (C10) is fulfilled, a slightly modified version of Theorem 6.1 can be proven. In particular, with the help of the technical arguments from the Supplementary Material, it is not difficult to show that

$$\max_{1 \le k \le K_0} \max_{i,j \in G_k} \widehat{d}_{ij} = O_p\big(\sqrt{\log n + \log T}\big)$$

$$\min_{1 \le k < k' \le K_0} \min_{\substack{i \in G_k, \\ j \in G_{k'}}} \widehat{d}_{ij} \gg \sqrt{\log n + \log T} + \sqrt{T h_{\max}^5}.$$

These two statements immediately imply that Theorem 4.1 remains to hold true. Moreover, Theorem 5.1 remains valid as well if the threshold level $\pi_{n,T}$ satisfies a strengthened version of condition (5.1), namely the condition that $\sqrt{\log n + \log T} \ll \pi_{n,T} \ll \sqrt{T h_{\max}} \min_{1 \le k < k' \le K_0} \max_{\{x \,:\, (x, h_{\max}) \in \mathcal{G}_T\}} |g_k(x) - g_{k'}(x)|$.

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.

Armstrong, T. B. and Chan, H. P. (2016). Multiscale adaptive inference on conditional moment inequalities. *Journal of Econometrics*, **194** 24–43.

Boneva, L., Elliott, D., Kaminska, I., Linton, O., McLaren, N. and Morley, B. (2018). The impact of QE on liquidity: evidence from the UK corporate bond purchase scheme. *Preprint.*

Boneva, L., Linton, O. and Vogt, M. (2015). A semiparametric model for heterogeneous panel data with fixed effects. *Journal of Econometrics*, **188** 327–345.

Boneva, L., Linton, O. and Vogt, M. (2016). The effect of fragmentation in trading on market quality in the UK equity market. *Journal of Applied Econometrics*, **31** 192–213.

Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, **83** 1147–1184.

Chaudhuri, P. and Marron, J. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, **94** 807–823.

Chaudhuri, P. and Marron, J. (2000). Scale space view of curve estimation. *Annals of Statistics*, **28** 408–428.

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.

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.

Dümbgen, L. and Spokoiny, V. G. (2001). Multiscale testing of qualitative hypotheses. *Annals of Statistics*, **29** 124–152.

Eckle, K., Bissantz, N. and Dette, H. (2017). Multiscale inference for multivariate deconvolution. *Electronic Journal of Statistics*, **11** 4179–4219.

Hansen, B. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, **24** 726–748.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning.* New York, Springer.

Horowitz, J. L. and Spokoiny, V. G. (2001). An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. *Econometrica*, **69** 599–631.

Jacques, J. and Preda, C. (2014). Functional data clustering: a survey. *Advances in Data Analysis and Classification*, **8** 231–255.

James, M. and Sugar, C. A. (2003). Clustering for sparsely sampled functional data.

Journal of the American Statistical Association, **98** 397–408.

Luan, Y. and Li, H. (2003). Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics*, **19** 474–482.

Proksch, K., Werner, F. and Munk, A. (2018). Multiscale scanning in inverse problems. *Annals of Statistics*, **46** 3569–3602.

Ray, S. and Mallick, B. (2006). Functional clustering by Bayesian wavelet methods. *Journal of the Royal Statistical Society: Series B*, **68** 305–332.

Robinson, P. M. (1988). Root-n consistent semiparametric regression. *Econometrica*, **56** 931–954.

Schmidt-Hieber, J., Munk, A. and Dümbgen, L. (2013). Multiscale methods for shape constraints in deconvolution: confidence statements for qualitative features. *Annals of Statistics*, **41** 1299–1328.

Su, L. and Ju, G. (2018). Identifying latent grouped patterns in panel data models with interactive fixed effects. *Journal of Econometrics*, **206** 554–573.

Su, L., Shi, Z. and Phillips, P. C. B. (2016). Identifying latent structures in panel data. *Econometrica*, **84** 2215–2264.

Tarpey, T. (2007). Linear transformations and the k-means clustering algorithm. *The American Statistician*, **61** 34–40.

Tarpey, T. and Kinateder, K. K. J. (2003). Clustering functional data. *Journal of Classification*, **20** 93–114.

Vogt, M. and Linton, O. (2017). Classification of non-parametric regression functions in longitudinal data models. *Journal of the Royal Statistical Society: Series B*, **79** 5–27.

Wang, W., Phillips, P. C. B. and Su, L. (2018). Homogeneity pursuit in panel data models: theory and application. *Journal of Applied Econometrics*, **33** 797–815.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, **58** 236–244.