Journal of Econometrics 216 (2020) 305–325

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Multiscale clustering of nonparametric regression curves

Michael Vogt [a,*,1], Oliver Linton [b,2,3]

[a] University of Bonn, Germany
[b] University of Cambridge, UK

ARTICLE INFO

Article history:
Available online 31 January 2020

JEL classification:
C14
C38
C55

Keywords:
Clustering of nonparametric curves
Nonparametric regression
Multiscale statistics
Multiple time series

ABSTRACT

In a wide range of modern applications, one observes 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 modeled 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 simulated and real-data examples.

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 $\mathcal{T}_i = \{(Y_{it}, X_{it}) : 1 \leq t \leq T\}$ for $1 \leq i \leq n$. For simplicity, we synonymously speak of the ith 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 $t = 1, \ldots, T$, where m_i is an unknown smooth function, X_{it} are random or deterministic regressors and u_{it} is the error term. The n time series in our sample 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 $k = 1, \ldots, K_0$,

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

According to (1.2), the time series of a given group G_k all have the same regression function. Model (1.1)–(1.2) provides a parsimonious way to deal with a potentially very large number of time series n. It thus stands in the tradition of multiple

* Correspondence to: Department of Economics and Hausdorff Center for Mathematics, University of Bonn, 53113 Bonn, Germany.
E-mail addresses: michael.vogt@uni-bonn.de (M. Vogt), obl20@cam.ac.uk (O. Linton).

[1] Supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Germany under Germany's Excellence Strategy – GZ 2047/1, Project-ID 390685813.

[2] Faculty of Economics, Austin Robinson Building, Sidgwick Avenue, Cambridge, CB3 9DD, UK.

[3] Financial support by the Cambridge-INET Institute, UK is gratefully acknowledged.

https://doi.org/10.1016/j.jeconom.2020.01.020
0304-4076/

time series analysis, an area which greatly benefited from the pioneering work of George Tiao. A detailed description of model (1.1)–(1.2) can be found in Section 2.

In many applications, it is quite natural to suppose that there is a group structure of the form (1.2) in the data. We give some examples to illustrate this. A first example comes from environmental statistics, a field to which George Tiao has contributed immensely with numerous articles including Box et al. (1975), Reinsel et al. (1989) and Niu and Tiao (1995). Suppose we observe time series $\mathcal{T}_i = \{Y_{it} : 1 \leq t \leq T\}$ of temperature, precipitation or ozone measurements at various spatial locations i. A simple model for the measurements at location i is given by $Y_{it} = m_i(t/T) + u_{it}$, where $X_{it} = t/T$ is (rescaled) time and m_i is a nonparametric time trend function. It is natural to suppose that the locations i in the observed sample can be grouped into geographical regions where the trend m_i is the same (or at least very similar). We come back to this example in Section 8. Another example which was analyzed in Vogt and Linton (2017) comes from finance. A recent question of policy interest is how competition between trading venues affects market quality in stock markets; cp. O'Hara and Ye (2009), Degryse et al. (2014) and Boneva et al. (2015, 2016) among others. To tackle this question, one may consider the model $Y_{it} = m_i(X_{it}) + u_{it}$, where Y_{it} is a measure of market quality for stock i at time t such as volatility and X_{it} is a measure of trading-venue fragmentation which gives information on whether stock i is traded simultaneously at many different venues at time t. 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 the same (or at least very similar). Hence, it makes sense to assume a group structure of the form (1.2) in this situation.

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 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 depend on a number of bandwidth or smoothing parameters required to estimate the nonparametric functions m_i. In general, nonparametric curve estimators are strongly affected by the chosen bandwidth parameters. A clustering algorithm which is based on such estimators can be expected to be strongly influenced by the choice of bandwidths as well. In particular, the clusters produced by the algorithm can be expected to vary considerably with the chosen bandwidths.

The main aim of this paper is to develop 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 that need to be selected. To achieve this, we make use of multiscale techniques from statistical hypothesis testing. In recent years, a number of multiscale methods have been developed in the context of different test problems. 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.

Our strategy 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) can be outlined as follows: To start with, we construct statistics $\widehat{d}_{ij}$ which measure the distance between pairs of functions m_i and m_j. Building on multiscale techniques, we design the statistics $\widehat{d}_{ij}$ in such a way that they do not depend on a specific bandwidth or smoothing parameter. To estimate the unknown classes $G_1, \ldots, G_{K_0}$, the multiscale distance statistics $\widehat{d}_{ij}$ are combined 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 $\widehat{d}_{ij}$ 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 S.2 of the Supplementary Material.

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 that 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 $1 \leq t \leq T$ and $1 \leq i \leq n$, 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 $1 \leq k \leq K_0$. The problem of estimating the unknown groups and their unknown number has been studied in different versions of this modeling framework in Bonhomme and Manresa (2015), Su et al. (2016), Wang et al. (2018) and Su and Ju (2018) among others. Notably, 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 S.2 of the Supplementary Material, 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 $G_1, \dots, G_{K_0}$. 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 behavior of the multiscale statistics and forms the basis to derive the theoretical properties of our clustering methods. To investigate 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 and explore a real-data example from environmental statistics in Section 8.

2. The model

We now introduce the model framework in detail which underlies our analysis. As already mentioned in the Introduction, we observe n time series $\mathcal{T}_i = \{(Y_{it}, X_{it}) : 1 \leq t \leq T\}$ of length T for $1 \leq i \leq n$. 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 modeled by the nonparametric regression equation

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

for $1 \leq t \leq T$, where m_i is an unknown smooth function and u_{it} denotes the error term. To keep the exposition as simple as possible, we assume that the regressors X_{it} are real-valued. As discussed in Section S.2 of the Supplementary Material, 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.

We consider both a random and a fixed design version of model (2.1), which we denote by (RD) and (FD), respectively. The regressors X_{it} are assumed to have the following properties in these two versions of the model:

(RD) For each i, the regressors X_{it} are real-valued random variables that are distributed according to some density f_i.
(FD) For each i, the regressors X_{it} are deterministic points on the unit interval with $0 \leq X_{i1} < X_{i2} < \cdots < X_{iT} \leq 1$. They are generated by a design density in the sense of Sacks and Ylvisaker (1970): for each i, there exists a density f_i such that $\int_{X_{i,t-1}}^{X_{it}} f_i(w)dw = 1/T$ for $1 \leq t \leq T$, where $X_{i0} = 0$.

Note that by setting $f_i \equiv 1$ in (FD), we obtain the important special case of equidistant design points $X_{it} = t/T$, which represent (rescaled) time in many applications. The error terms u_{it} are assumed to have an additive component structure both in the random and the fixed design case:

(RD) It holds that $u_{it} = \alpha_i + \gamma_t + \varepsilon_{it}$, where α_i and γ_t are fixed effects that may be correlated with the regressors X_{it} in an arbitrary way and ε_{it} are standard regression errors that satisfy $\mathbb{E}[\varepsilon_{it}|X_{it}] = 0$.
(FD) It holds that $u_{it} = \alpha_i + \varepsilon_{it}$, where α_i are fixed effects and ε_{it} are standard regression errors with $\mathbb{E}[\varepsilon_{it}] = 0$.

As discussed in more detail in Section 2.3, we do not include the time fixed effects γ_t in the (FD) case for identifiability reasons: Whereas the functions m_i can be identified in the presence of the time fixed effects γ_t in the (RD) model, this is in general not possible in the (FD) model. Both in the (RD) and the (FD) case, the time series $\mathcal{E}_i = \{\varepsilon_{it} : 1 \leq t \leq T\}$ are supposed to be weakly dependent stationary processes that are independent across i. The error terms ε_{it} are thus 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 $G_1, \dots, G_{K_0}$ with $\dot{\bigcup}_{k=1}^{K_0} G_k = \{1, \dots, n\}$ such that for each $1 \leq k \leq K_0$,

$$m_i = m_j \quad \text{for all } i, j \in G_k. \tag{2.2}$$

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

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

where g_k is the group-specific regression function associated with the class G_k. According to (2.3), 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 S.2 of the Supplementary Material. 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

We first discuss the (RD) case. If we drop the fixed effects α_i and γ_t from the error terms u_{it}, we obtain the standard regression equation $Y_{it} = m_i(X_{it}) + \varepsilon_{it}$. Obviously, m_i is identified in this case since $m_i(\cdot) = \mathbb{E}[Y_{it}|X_{it} = \cdot\,]$ almost surely. In the full model $Y_{it} = m_i(X_{it}) + \alpha_i + \gamma_t + \varepsilon_{it}$, by contrast, m_i is not identified. Specifically, we can rewrite the model 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 $\mathbb{E}[m_i(X_{it})] = \int m_i(w)f_i(w)dw$ for any i. A common choice is the identification constraint

$$\int m_i(w)f_i(w)dw = 0 \quad \text{for } 1 \leq i \leq n, \tag{2.4}$$

which is implicitly assumed to be fulfilled throughout the paper. Given this constraint, it is straightforward to show that the functions m_i are identified under our regularity conditions from Section 6. A formal identification result is provided in Section S.2 of the Supplementary Material for completeness.

The situation is somewhat different in the (FD) case. There, it is in general not possible to identify the functions m_i in the presence of the time fixed effects γ_t. To see the issue, consider the special case $X_{it} = t/T$ and suppose for simplicity that $\alpha_i = 0$ and $\gamma_t = \gamma(t/T)$ with some deterministic function γ. In this case, $Y_{it} = m_i(t/T) + \gamma(t/T) + \varepsilon_{it}$, where $\tau_i(t/T) = m_i(t/T) + \gamma(t/T)$ is the trend function of time series i. Obviously, we cannot identify the trend components m_i and γ in this situation without imposing severe assumptions on their functional form. However, if we restrict attention to the model $Y_{it} = m_i(X_{it}) + \alpha_i + \varepsilon_{it}$ without the time fixed effects γ_t, we can proceed analogously as in the (RD) case. In particular, we get identification of the functions m_i under the constraint (2.4).

It is important to note that the identification constraint (2.4) and thus the fixed effects error structure of our model implicitly imposes certain restrictions on the design densities f_i. The identification constraint (2.4) requires that $\int g_k(w)f_i(w)dw = 0$ for all $i \in G_k$, where g_k is the regression function associated with the group G_k. It is in general not possible to satisfy the constraint $\int g_k(w)f_i(w)dw = 0$ simultaneously for all i when the densities f_i are arbitrarily different across i. However, if we suppose that for any $1 \leq k \leq K_0$,

$$f_i = f_j \quad \text{for all } i, j \in G_k, \tag{2.5}$$

then this constraint is satisfied quite naturally. In the remainder of the paper, we take for granted that the property (2.5) is fulfilled, that is, we assume the design density f_i to be the same for all time series i in a given group G_k.

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 $1 \leq k \leq K_0$. Using multiscale techniques, we design the statistic $\widehat{d}_{ij}$ in such a way that it is free of specific bandwidth parameters. 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. We work with the same local linear kernel smoother as in Vogt and Linton (2017). This estimator is given by

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

where $\widehat{Y}_{it}^* = Y_{it} - \overline{Y}_i - \overline{Y}_t^{(i)} + \overline{\overline{Y}}^{(i)}$ in the (RD) case and $\widehat{Y}_{it}^* = Y_{it} - \overline{Y}_i$ in the (FD) case with

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

Moreover, $W_{it}(x, h)$ are kernel weights of the form

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

where $S_{i,\ell}(x, h) = T^{-1}\sum_{t=1}^{T} K_h(X_{it} - x)(\{X_{it} - x\}/h)^\ell$ 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 $[-C_K, C_K]$. For ease of notation, we set $C_K = 1$ and take the kernel K to be the same for each i.

Step 2. As an intermediate step, we construct a bandwidth-dependent test statistic of the hypothesis H_0. Specifically, we consider the statistic

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

where

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

is a rescaled version of the difference between the curve estimators $\widehat{m}_{i,h}(x)$ and $\widehat{m}_{j,h}(x)$ at location x with bandwidth h. The term $\widehat{v}_{ij}(x, h)$ is a scaling factor which normalizes the asymptotic variance of $\widehat{\psi}_{ij}(x, h)$. Importantly, our theory does not require the statistics $\widehat{\psi}_{ij}(x, h)$ to have asymptotic variance exactly equal to 1. Nevertheless, for the multiscale methods we are about to develop, it is desirable to normalize the statistics $\widehat{\psi}_{ij}(x, h)$ such that their variance is approximately equal to 1 and thus comparable in size across locations x and bandwidths h. In order to achieve this, we set

$$\widehat{v}_{ij}(x, h) = \left\{\frac{\widehat{\sigma}_{i,h}^2}{\widehat{f}_{i,h}(x)} + \frac{\widehat{\sigma}_{j,h}^2}{\widehat{f}_{j,h}(x)}\right\} s(x, h), \tag{3.2}$$

where $s(x, h) = \{\int_{-x/h}^{(1-x)/h} K^2(u)[\kappa_2(x, h) - \kappa_1(x, h)u]^2 du\}/\{\kappa_0(x, h)\kappa_2(x, h) - \kappa_1(x, h)^2\}^2$ is a kernel constant with $\kappa_\ell(x, h) = \int_{-x/h}^{(1-x)/h} u^\ell K(u)du$, $\widehat{f}_{i,h}(x) = \{\kappa_0(x, h)T\}^{-1}\sum_{t=1}^{T} K_h(X_{it} - x)$ is a boundary-corrected kernel density estimator of the design density f_i and $\widehat{\sigma}_{i,h}^2 = T^{-1}\sum_{t=1}^{T}\{\widehat{Y}_{it}^* - \widehat{m}_{i,h}(X_{it})\}^2$ is an estimator of the error variance $\sigma_i^2 = \mathbb{E}[\varepsilon_{it}^2]$. In the (RD) case, the definition of $\widehat{\sigma}_{i,h}^2$ implicitly presupposes that the error terms ε_{it} are homoskedastic. When they are heteroskedastic, $\widehat{\sigma}_{i,h}^2$ can be replaced by an estimator of the conditional error variance $\sigma_i^2(x) = \mathbb{E}[\varepsilon_{it}^2|X_{it} = x]$, in particular, by $\widehat{\sigma}_{i,h}^2(x) = \{\sum_{t=1}^{T} K_h(X_{it} - x)[\widehat{Y}_{it}^* - \widehat{m}_{i,h}(X_{it})]^2\}/\{\sum_{t=1}^{T} K_h(X_{it} - x)\}$. In the (FD) case, the term $\widehat{\sigma}_{i,h}^2$ must give a reasonable approximation to the long-run error variance $\Gamma_i = \sum_{\ell=-\infty}^{\infty}\mathrm{Cov}(\varepsilon_{i0}, \varepsilon_{i\ell})$ in order to produce a correct normalization of the statistic $\widehat{\psi}_{ij}(x, h)$. As long as the time series dependence in the errors ε_{it} is not too strong, $\sigma_i^2 = \mathrm{Var}(\varepsilon_{i0}^2)$ will be the dominant term in the long-run variance Γ_i, implying that $\widehat{\sigma}_{i,h}^2$ should approximate Γ_i reasonably well. However, if the dependence in the errors is expected to be strong, $\widehat{\sigma}_{i,h}^2$ should be replaced by an estimator of the long-run variance Γ_i, for example, by a HAC-type estimator as discussed in Andrews (1991) or de Jong and Davidson (2000).

To motivate the following steps, it is instructive to examine the statistic $\widehat{d}_{ij}(h)$ in a simplified version of our model. We in particular consider the setting

$$Y_{it} = m_i(X_{it}) + \varepsilon_{it}, \tag{3.3}$$

where (a) the design density $f = f_i$ is the same for all i, (b) the fixed effects α_i and γ_t are dropped from the model and (c) the errors ε_{it} are i.i.d. both across i and t. In this simplified setting, the statistic $\widehat{\psi}_{ij}(x, h)$ can be decomposed into a bias part $\widehat{\psi}_{ij}^B(x, h)$ and a variance part $\widehat{\psi}_{ij}^V(x, h)$ according to

$$\widehat{\psi}_{ij}(x, h) = \widehat{\psi}_{ij}^B(x, h) + \widehat{\psi}_{ij}^V(x, h) + \text{lower order terms}, \tag{3.4}$$

where for any $x \in [h, 1 - h]$,

$$\widehat{\psi}_{ij}^B(x, h) = \sqrt{Th}\,\frac{\int\{m_i(x + hu) - m_j(x + hu)\}w(u, x, h)du}{\sqrt{\widehat{v}_{ij}(x, h)}} \tag{3.5}$$

$$\widehat{\psi}_{ij}^V(x, h) = \sqrt{Th}\,\frac{\big(\widehat{m}_{i,h}^V(x) - \widehat{m}_{j,h}^V(x)\big)}{\sqrt{\widehat{v}_{ij}(x, h)}} \tag{3.6}$$

with $\widehat{m}^V_{i,h}(x) = \{\sum_{t=1}^T W_{it}(x,h)\varepsilon_{it}\}/\{\sum_{t=1}^T W_{it}(x,h)\}$ and $w(u,x,h) = K(u)f(x+hu)/\int K(v)f(x+hv)dv$. Under standard conditions, $\widehat{\psi}^V_{ij}(x,h) \xrightarrow{d} N(0,1)$. Moreover, 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 $1 \leq k \leq K_0$.

The variance part $\widehat{\psi}^V_{ij}(x,h)$ captures the stochastic fluctuations of the statistic $\widehat{\psi}_{ij}(x,h)$. Inspecting (3.5) and recalling that the kernel K has support $[-1,1]$, the bias part $\widehat{\psi}^B_{ij}(x,h)$ can be seen to be a weighted integrated difference between m_i and m_j on the interval $[x-h, x+h]$. It can thus be regarded as a signal which indicates a deviation from H_0 locally around x. The strength of the signal $\widehat{\psi}^B_{ij}(x,h)$ depends on the choice of the bandwidth h. To see this more clearly, consider two regression functions m_i and m_j from two different groups. The 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. To fix ideas, suppose that m_i and m_j differ on the interval $I^* = [x-h^*, x+h^*]$ but are the same outside I^*. The parameter h^* can be regarded as the scale on which m_i and m_j differ: For small/large values of h^*, the interval I^* 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 strong for bandwidths h close to h^* and becomes weak for bandwidths h substantially smaller or larger than h^*. Very roughly speaking, the reason for this is as follows: Depending on the choice of h, the integration region in (3.5) changes. If h is much larger than h^*, we integrate over a much larger interval than I^* in (3.5) and thus smooth out the differences between m_i and m_j. If h is much smaller than h^*, in contrast, we only integrate over a small part of the region I^* where the two functions m_i and m_j differ and thus do not use all of the information available on the differences between m_i and m_j.

According to these heuristic considerations, it strongly depends on the chosen bandwidth whether the test statistic $\widehat{d}_{ij}(h) = \sup_{x\in[0,1]} |\widehat{\psi}_{ij}(x,h)|$ is able to detect a deviation from the null H_0. In particular, if the bandwidth h is much smaller/larger than the scale h^* on which m_i and m_j mainly differ, the statistic $\widehat{d}_{ij}(h)$ is not able to pick up the differences between m_i and m_j.

Step 3. We now construct a test statistic of H_0 which does not depend on a specific bandwidth h but takes into account a wide range of different bandwidths simultaneously. By construction, such a statistic should be able to detect differences between the functions m_i and m_j on multiple scales simultaneously. To obtain such a statistic, we proceed as follows: We compute the bandwidth-dependent statistic $\widehat{d}_{ij}(h)$ for all bandwidths h in a large set $\mathcal{H} = \{h : h_{\min} \leq h \leq h_{\max}\}$, where $h_{\min}$ and $h_{\max}$ denote some minimal and maximal bandwidth values that are specified later on. We then combine the statistics $\widehat{d}_{ij}(h)$ for all $h \in \mathcal{H}$ to obtain a single test statistic.

A simple way of combining the statistics $\widehat{d}_{ij}(h)$ for all $h \in \mathcal{H}$ is to take their supremum, which leads to the definition

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

On first sight, the statistic $\widetilde{d}_{ij}$ seems to be a reasonable multiscale statistic which takes into account multiple bandwidths simultaneously. However, inspecting it more closely, it turns out to have the following defect: It does not take into account all bandwidths $h \in \mathcal{H}$ in an equal fashion. Its stochastic behavior is rather dominated by the statistics $\widehat{d}_{ij}(h)$ that correspond to small bandwidths h. To see this, we examine the statistic $\widetilde{d}_{ij}$ in the simplified model setting (3.3) introduced in Step 2 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, $\widehat{\psi}_{ij}(x,h) = \widehat{\psi}^V_{ij}(x,h) +$ lower order terms, since the bias term $\widehat{\psi}^B_{ij}(x,h)$ in (3.4) is equal to 0 for all x and h as already noted in Step 2. Hence, the statistic $\widehat{\psi}_{ij}(x,h)$ is approximately equal to the variance term $\widehat{\psi}^V_{ij}(x,h)$, which captures its stochastic fluctuations. Neglecting terms of lower order, we obtain that under H_0, $\widehat{\psi}_{ij}(x,h) = \widehat{\psi}^V_{ij}(x,h)$ and thus

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

For a given bandwidth h, the statistics $\widehat{\psi}^V_{ij}((2\ell-1)h,h)$ for $\ell = 1,\dots,\lfloor 1/2h\rfloor$ can be shown to be (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 $\max_\ell \widehat{\psi}^V_{ij}((2\ell-1)h,h)$ is approximately of size $\lambda(2h)$ for small bandwidths h. Moreover, since the statistics $\widehat{\psi}^V_{ij}(x,h)$ with $(2\ell-1)h < x < (2\ell+1)h$ are correlated with $\widehat{\psi}^V_{ij}((2\ell-1)h,h)$ and $\widehat{\psi}^V_{ij}((2\ell+1)h,h)$, the supremum $\sup_x \widehat{\psi}^V_{ij}(x,h)$ approximately behaves as the maximum $\max_\ell \widehat{\psi}^V_{ij}((2\ell-1)h,h)$. Taken together, these heuristic considerations suggest that

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

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 behavior of $\widetilde{d}_{ij}$ tends to be dominated by the statistics $\widehat{d}_{ij}(h)$ which correspond to small bandwidths h.

Step 4. To fix this bias issue, we follow Dümbgen and Spokoiny (2001) and replace the statistic $\widetilde{d}_{ij}$ by the modified version

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

where $\lambda(r) = \sqrt{2\log(1/r)}$. For each given bandwidth h, we thus subtract the additive correction term $\lambda(2h)$ from the statistic $\widehat{d}_{ij}(h)$. The idea behind this additive correction is as follows: 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 according to the heuristic considerations from above. 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 more comparable across bandwidths h. We thus put them on a more equal footing and prevent small bandwidths from dominating the stochastic behavior of the multiscale statistic. As a result, $\widehat{d}_{ij}$ should be a reliable test statistic of the null hypothesis H_0 which is able to detect differences between the functions m_i and m_j on multiple scales simultaneously.

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.

Remark 3.1. Alternatively to (3.10), one may define

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

where the multiplicative constant $\omega(r) = \sqrt{\log(e/r)}/\log\log(e^e/r)$ is motivated by Theorem 2.1 in Dümbgen and Spokoiny (2001). In simple special cases, the limit distribution of $\widehat{d}_{ij}$ can be shown to be degenerate for $i, j \in G_k$ as the largest bandwidth $h_{\max}$ converges to zero. For this reason, one may prefer the statistic $\widehat{d}^{\omega}_{ij}$ over $\widehat{d}_{ij}$ in the context of statistical testing. For our clustering purposes, however, both statistics are appropriate. In particular, it does not matter whether $\widehat{d}_{ij}$ has a degenerate limit. The main theoretical results on our clustering methods hold true no matter whether we work with $\widehat{d}_{ij}$ or $\widehat{d}^{\omega}_{ij}$. Moreover, from a practical point of view, the performance of $\widehat{d}^{\omega}_{ij}$ appears to be very similar to that of $\widehat{d}_{ij}$. In particular, the simulation results of Section 7 are almost identical when $\widehat{d}_{ij}$ is replaced by $\widehat{d}^{\omega}_{ij}$. For these reasons, we stick to the somewhat simpler statistic $\widehat{d}_{ij}$ throughout the paper.

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 $h_{\max}$ in practice. These guidelines are in particular used to implement our methods in the simulated and real-data examples of Sections 7 and 8.

Ideally, we would like to make the interval $[h_{\min}, h_{\max}]$ as large as possible, thus taking into account as many bandwidths h as possible. From a technical perspective, we can pick any bandwidths $h_{\min}$ and $h_{\max}$ with $h_{\min} \ge cT^{-(1-\delta)}$ and $h_{\max} \le CT^{-\delta}$ 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 $Th_{\min} \le 10$. 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 behavior of the multiscale statistic $\widehat{d}_{ij}$. To see why, write $\widehat{\psi}_{ij}(x, h) = \widehat{\psi}^{B}_{ij}(x, h) + \widehat{\psi}^{V}_{ij}(x, h) +$ lower order terms as in (3.4), where the variance term $\widehat{\psi}^{V}_{ij}(x, h)$ captures the stochastic fluctuations

of $\widehat{\psi}_{ij}(x,h)$ and the bias term $\widehat{\psi}^B_{ij}(x,h)$ 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

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

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 $\widehat{\psi}^B_{ij}(x,h)$ 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

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

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 roughly 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 formally prove 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 $\widehat{d}_{ij} \leq c$ 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:

Algorithm 4.1 (*HAC Algorithm*)**.**
Step 0 (Initialization): Let $\widehat{G}_i^{[0]} = \{i\}$ denote the ith singleton cluster for $1 \leq i \leq n$ 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

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

and merge them into a new cluster.

Iterating this procedure for $r = 1, \dots, n-1$ yields a tree of nested partitions $\{\widehat{G}_1^{[r]}, \dots, \widehat{G}_{n-r}^{[r]}\}$, 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:

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

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 $\mathcal{A} = \{A_1, \dots, A_r\}$ and $\mathcal{B} = \{B_1, \dots, B_{r'}\}$ be two partitions of the set $\{1, \dots, n\}$, that is, $\dot{\cup}_{\ell=1}^{r} A_\ell = \{1, \dots, n\}$ and $\dot{\cup}_{\ell=1}^{r'} B_\ell = \{1, \dots, n\}$. 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) $\mathbb{P}\Big(\big\{\widehat{G}_1^{[n-K_0]}, \dots, \widehat{G}_{K_0}^{[n-K_0]}\big\} = \big\{G_1, \dots, G_{K_0}\big\}\Big) \to 1$,
(b) $\mathbb{P}\Big(\big\{\widehat{G}_1^{[n-K]}, \dots, \widehat{G}_K^{[n-K]}\big\}$ *is a refinement of* $\big\{G_1, \dots, G_{K_0}\big\}\Big) \to 1$ *for any* $K > K_0$,
(c) $\mathbb{P}\Big(\big\{G_1, \dots, G_{K_0}\big\}$ *is a refinement of* $\big\{\widehat{G}_1^{[n-K]}, \dots, \widehat{G}_K^{[n-K]}\big\}\Big) \to 1$ *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 $\{\widehat{G}_1, \dots, \widehat{G}_{K_0}\}$ with $\widehat{G}_k = \widehat{G}_k^{[n-K_0]}$ for $1 \leq k \leq K_0$ is a consistent estimator of the true class structure $\{G_1, \dots, G_{K_0}\}$ in the following sense: $\{\widehat{G}_1, \dots, \widehat{G}_{K_0}\}$ coincides with $\{G_1, \dots, G_{K_0}\}$ with probability tending to 1. Hence, if the number of classes K_0 were known, we could consistently estimate the true class structure by $\{\widehat{G}_1, \dots, \widehat{G}_{K_0}\}$. The partitions $\{\widehat{G}_1^{[n-K]}, \dots, \widehat{G}_K^{[n-K]}\}$ 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 $\widehat{G}_1^{[n-K]}, \dots, \widehat{G}_K^{[n-K]}$ 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

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

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 $K < K_0$,

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

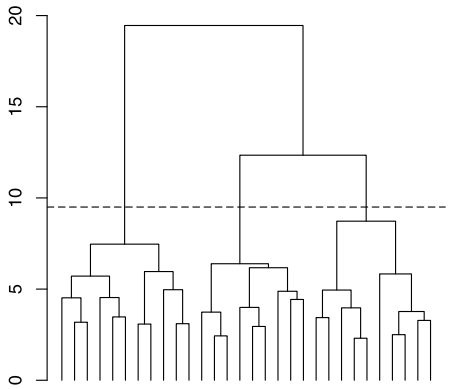


Fig. 1. Example of a dendrogram produced by the HAC algorithm. The dashed horizontal line indicates the dissimilarity level $\pi_{n,T}$. The estimator $\widehat{K}_0$ can be computed by counting the vertical lines that intersect the dashed horizontal threshold. In the above example, $\widehat{K}_0$ is equal to 3.

whereas for $K = K_0$,

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

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

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

is satisfied. Formally speaking, we estimate K_0 by

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

The estimator $\widehat{K}_0$ depends on the threshold parameter $\pi_{n,T}$ whose choice is discussed in detail below. Note that the clustering algorithm of Vogt and Linton (2017) also depends on a threshold parameter, which however plays a quite different role than $\pi_{n,T}$. We comment on the relationship between our clustering approach and the method in Vogt and Linton (2017) in more detail in Section S.2 of the Supplement.

Provided that $\pi_{n,T}$ satisfies (5.1), $\widehat{K}_0$ 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 speaking, 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

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

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 Fig. 1 for an illustration.

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 describe a data-driven procedure to choose the threshold level $\pi_{n,T}$. We first introduce the algorithm and then give a heuristic explanation why it should yield a suitable choice of $\pi_{n,T}$ in practice. To formulate the algorithm, we make use of the following notation: We write the set $\mathcal{G}_T$ of location-bandwidth points (x, h) as

$$G_T = \big\{z_{\nu,\ell} = (x_{\nu,\ell}, h_\nu) : 1 \le \nu \le p,\ 1 \le \ell \le N_\nu\big\},$$

where $x_{\nu,\ell}$ $(1 \le \ell \le N_\nu)$ are the locations corresponding to the bandwidth h_ν and p different bandwidths h_ν $(1 \le \nu \le p)$ are considered. Moreover, we let $\boldsymbol{\zeta}_i = (\zeta_{i,\nu,\ell} : 1 \le \nu \le p,\ 1 \le \ell \le N_\nu) = (\zeta_{i,1,1}, \dots, \zeta_{i,1,N_1}, \dots, \zeta_{i,p,1}, \dots, \zeta_{i,p,N_p})^\top$ be independent Gaussian random vectors of length $\sum_{\nu=1}^p N_\nu$ for $1 \le i \le n$. Each random vector $\boldsymbol{\zeta}_i$ has the covariance structure

$$\begin{aligned}
&\mathrm{Cov}\big(\zeta_{i,\nu,\ell},\ \zeta_{i,\nu',\ell'}\big) \\
&\quad = \big\{4\rho_{\nu,\ell}\,\rho_{\nu',\ell'}\big\}^{-1/2}\sqrt{\frac{h_\nu}{h_{\nu'}}}\bigg\{\int_{-x_{\nu,\ell}/h_\nu}^{(1-x_{\nu,\ell})/h_\nu} K(u)\big[\kappa_{2,\nu,\ell} - \kappa_{1,\nu,\ell}\,u\big] \\
&\qquad \times K\Big(\frac{h_\nu u + x_{\nu,\ell} - x_{\nu',\ell'}}{h_{\nu'}}\Big)\Big[\kappa_{2,\nu',\ell'} - \kappa_{1,\nu',\ell'}\Big(\frac{h_\nu u + x_{\nu,\ell} - x_{\nu',\ell'}}{h_{\nu'}}\Big)\Big]du\bigg\},
\end{aligned} \tag{5.4}$$

where we use the shorthands $\kappa_{j,\nu,\ell} = \kappa_j(x_{\nu,\ell}, h_\nu)$ with $\kappa_j(x,h) = \int_{-x/h}^{(1-x)/h} u^j K(u)du$ and $\rho_{\nu,\ell} = \rho(x_{\nu,\ell}, h_\nu)$ with $\rho(x,h) = \int_{-x/h}^{(1-x)/h} K^2(u)[\kappa_2(x,h) - \kappa_1(x,h)u]^2 du$. We further define the random variable

$$B_n = \max_{1\le i,j\le n}\big(\,|\boldsymbol{\zeta}_i - \boldsymbol{\zeta}_j| - \boldsymbol{\lambda}\,\big)_\infty,$$

where we employ the notation $|z| = (|z_1|, \dots, |z_q|)^\top$ and $(z)_\infty = \max_{1\le\ell\le q} z_\ell$ for vectors $z \in \mathbb{R}^q$ and $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_p)^\top$ with $\boldsymbol{\lambda}_\nu = (\lambda(2h_\nu), \dots, \lambda(2h_\nu)) \in \mathbb{R}^{N_\nu}$ for each ν. With this notation at hand, we compute $\pi_{n,T}$ as follows.

Algorithm 5.1 (*Choice of the Threshold Level* $\pi_{n,T}$)**.** For some pre-specified $\alpha \in (0, 1)$, compute the empirical $(1-\alpha)$-quantile $\widehat{q}_n(\alpha)$ of B_n by simulation. In particular, simulate a large number of realizations of $(\boldsymbol{\zeta}_1, \dots, \boldsymbol{\zeta}_n)$, compute the corresponding realizations of B_n and calculate the empirical $(1-\alpha)$-quantile $\widehat{q}_n(\alpha)$ from these. Set $\pi_{n,T} = \widehat{q}_n(\alpha)$, where we suggest to pick $\alpha \in \{0.01, 0.05, 0.1\}$, thus mimicking the usual significance levels of a statistical test in practice.

We now give some heuristic arguments why Algorithm 5.1 should yield an appropriate choice of $\pi_{n,T}$. 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 in the (RD) case 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 design densities f_i 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 design density in the kth class G_k. We can now make the following heuristic observations, where we use the notation introduced above:

(a) Consider any pair of time series i and j that belong to the same class G_k. As in (3.4), we can 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, where $\widehat{\psi}_{ij}^B(x,h) = 0$ for $i, j \in G_k$ and thus

$$\widehat{\psi}_{ij}(x,h) \approx \widehat{\psi}_{ij}^V(x,h) = \sqrt{Th}\big\{\widehat{m}_{i,h}^V(x) - \widehat{m}_{j,h}^V(x)\big\}\big/\{\widehat{v}_{ij}(x,h)\}^{1/2} \tag{5.5}$$

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

$$\begin{aligned}
\widehat{m}_{i,h}^V(x) &\approx \big\{f_k(x)\big[\kappa_0(x,h)\kappa_2(x,h) - \kappa_1(x,h)^2\big]\big\}^{-1} \\
&\quad \times \frac{1}{T}\sum_{t=1}^T K_h(X_{it} - x)\Big[\kappa_2(x,h) - \kappa_1(x,h)\Big(\frac{X_{it} - x}{h}\Big)\Big]\varepsilon_{it}.
\end{aligned} \tag{5.6}$$

Since by construction, $\widehat{v}_{ij}(x,h)$ is an estimator of $v_{ij}(x,h) = 2\{\sigma_k^2/f_k(x)\}s(x,h)$ with $s(x,h)$ introduced in (3.2), we can combine (5.5) and (5.6) to obtain the approximation $\widehat{\psi}_{ij}(x,h) \approx \widehat{\psi}_i(x,h) - \widehat{\psi}_j(x,h)$ with

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

For each i, we stack the random variables $\widehat{\psi}_i(x, h)$ with $(x, h) \in \mathcal{G}_T$ in the vector $\widehat{\boldsymbol{\psi}}_i = (\widehat{\psi}_i(x_{\nu,\ell}, h_\nu) : 1 \le \nu \le p,\ 1 \le \ell \le N_\nu) = (\widehat{\psi}_i(x_{1,1}, h_1), \ldots, \widehat{\psi}_i(x_{1,N_1}, h_1), \ldots, \widehat{\psi}_i(x_{p,1}, h_p), \ldots, \widehat{\psi}_i(x_{p,N_p}, h_p))^\top$. 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 $\mathrm{Cov}(\widehat{\psi}_i(x_{\nu,\ell}, h_\nu), \widehat{\psi}_i(x_{\nu',\ell'}, h_{\nu'})) \approx \mathrm{Cov}(\zeta_{i,\nu,\ell}, \zeta_{i,\nu',\ell'})$, we can approximate the random vector $\widehat{\boldsymbol{\psi}}_i$ by the Gaussian vector $\boldsymbol{\zeta}_i$. 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 $S \subseteq \{1, \ldots, n\}$.

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

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

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

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

and define $q_n(\alpha)$ to be the $(1 - \alpha)$-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 0. 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

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

According to these considerations, $\pi_{n,T} = \widehat{q}_n(\alpha)$ with α close to 0 (in particular with $\alpha \in \{0.01, 0.05, 0.1\}$) should be an appropriate threshold level.

6. Theoretical results

In this section, we derive the statements (3.11) and (3.12) under appropriate regularity conditions. These statements characterize the convergence behavior 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) Define $\mathcal{P}_i = \{(X_{it}, \varepsilon_{it}) : t = 1, 2, \ldots\}$ in the (RD) case and $\mathcal{P}_i = \{\varepsilon_{it} : t = 1, 2, \ldots\}$ in the (FD) case. The time series processes $\mathcal{P}_i$ 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 ith 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 $1 \leq i \leq n$, the design density f_i has 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, $0 < c \leq f_i(x) \leq C < \infty$ for all $x \in [0, 1]$ with some constants c and C that neither depend on x nor on i, and (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, in the (RD) case, 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, $f_{i,\ell}(x, x') \leq C < \infty$ for all i, x, x' and ℓ, where the constant C neither depends on i, x, x' nor on ℓ.

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

(C4) The densities f_i and the error variances σ_i^2 are the same within groups. That is, for any k with $1 \leq k \leq K_0$, 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 $\theta > 4$, a natural number ℓ^* and a positive constant C such that the following holds: In the (RD) case,

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

for any $\ell \in \mathbb{Z}$ with $|\ell| \geq \ell^*$, and in the (FD) case,

$$\max_{1 \leq i \leq n} \mathbb{E}\big[|\varepsilon_{it}|^\theta\big] \leq C < \infty.$$

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

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

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 $0 < b \leq B < 1$.

(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 $|K(v) - K(w)| \leq L|v - w|$ 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 time series $\mathcal{P}_i$ 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.

(ii) (C1) restricts the error components ε_{it} to be independent across i. Nevertheless, some restricted types of cross-sectional dependence in the error terms u_{it} are possible via the fixed effects α_i and γ_t.

(iii) 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 made 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.

(iv) (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.

(v) (C6) requires the functions g_k to be different across groups. However, it does not impose any quantitative restrictions on the size of their differences. From an asymptotic perspective, such statements are not needed. Asymptotically, the clustering algorithm developed in Sections 3–5 is able to detect the true group structure, no matter how small the differences between the functions g_k are in comparison to the noise level in the data, that is, in comparison to the error variances σ_i^2. The situation in practice is of course very different: In finite samples, the algorithm is only able to distinguish between two groups G_k and $G_{k'}$ if the difference between the functions g_k and $g_{k'}$ is sufficiently large compared to the noise level in the data. Otherwise, the multiscale statistics $\widehat{d}_{ij}$ will not pick up this difference, implying that the algorithm treats $G_k \cup G_{k'}$ as one group.

(vi) (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.

(vii) (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.

(viii) 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*

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

$$\min_{1\leq k<k'\leq K_0} \min_{\substack{i\in G_k,\\ j\in G_{k'}}} \widehat{d}_{ij} \geq c_0\sqrt{Th_{\max}} + o_p\big(\sqrt{Th_{\max}}\big), \tag{6.3}$$

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

7. Simulation study

The simulation study splits up into two main parts. In the first, we carry out some simulations to illustrate the advantages of our multiscale approach over clustering methods that depend on a specific bandwidth. In the second, we compare our estimator $\widehat{K}_0$ of the number of groups with alternative methods. Due to space constraints, the second part of the simulation study is presented in Section S.1 of the Supplement.

7.1. Comparison with bandwidth-dependent alternatives

When the grid $\mathcal{G}_T$ of location-bandwidth 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 and h the bandwidth under consideration.[4] In what follows, we compare our multiscale approach with this bandwidth-dependent procedure.

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

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

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

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

where $b(x, x_0, h) = 1(|x - x_0|/h \leq 1)\{1 - ((x - x_0)/h)^2\}^2$. Fig. 2 provides a graphical illustration of the functions g_k. The error process $\mathcal{E}_i = \{\varepsilon_{it} : 1 \leq t \leq T\}$ has an autoregressive (AR) structure for each i, in particular $\varepsilon_{it} = a\varepsilon_{it-1} + \eta_{it}$ for $1 \leq t \leq T$, 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

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

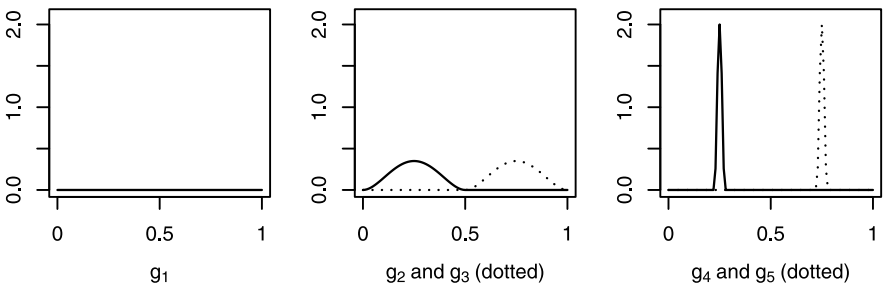


Fig. 2. Plot of the functions g_k for $1 \leq k \leq 5$.

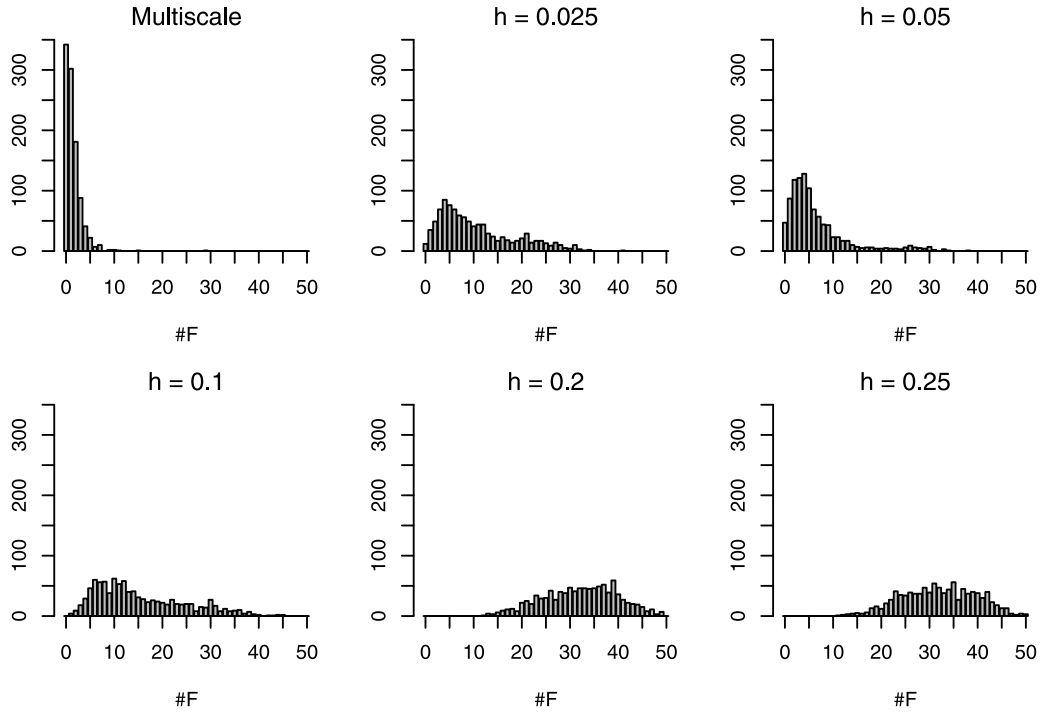
is chosen as $v^2 = 1 - a^2$, which implies that $\text{Var}(\varepsilon_{it}) = 1$. The regressors X_{it} are drawn independently from a uniform distribution on $[0, 1]$ for each i. As can be seen, there is no time series dependence in the regressors, and we do not include fixed effects α_i and γ_t 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-bandwidth 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 locations and $\mathcal{H} = \{h : 0.025 \leq h \leq 0.25 \text{ with } h = 0.025k \text{ for } k = 1, 2, \ldots\}$ is the set of bandwidths. The bandwidth-dependent algorithm is implemented with the same set of locations $\mathcal{X}$ and five different bandwidths $h \in \{0.025, 0.05, 0.1, 0.2, 0.25\}$. The local linear smoothers $\widehat{m}_{i,h}$ which underlie the clustering algorithms are computed with an Epanechnikov kernel K. 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 $\pi_{n,T} = \widehat{q}_n(\alpha)$ with $\alpha = 0.05$. 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.

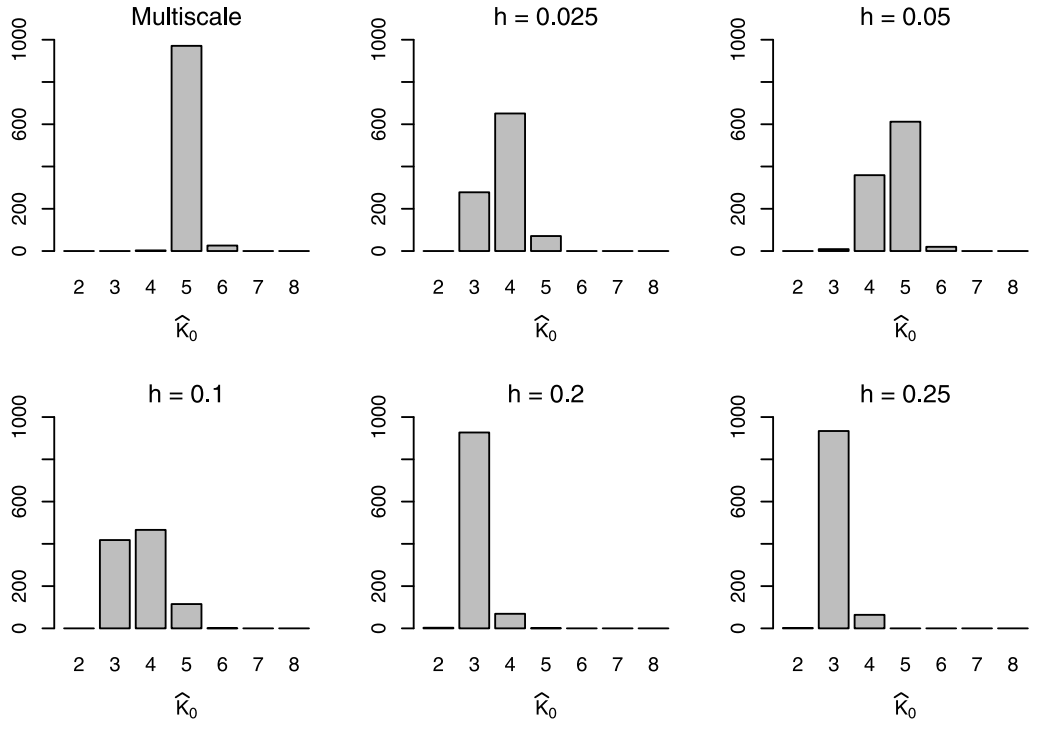
The simulation results for the scenario with the negative AR parameter $a = -0.25$ are reported in Fig. 3 and those for the scenario with the positive parameter $a = 0.25$ in Fig. 4. We first have a closer look at Fig. 3. To produce Fig. 3(a), we treat K_0 as known and compute the number of classification errors $\#F$, that is, the number of wrongly classified indices i for each of the $S = 1000$ simulated samples.[5] The upper left panel of Fig. 3(a) shows the histogram of these $S = 1000$ values for our multiscale approach. The other panels of Fig. 3(a) 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. Fig. 3(b) 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 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 Fig. 3(b). The results in Fig. 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 differences between the functions on a global scale. When implemented with a large bandwidth, in contrast, it is hardly able to capture differences on a local scale. Our multiscale approach, in contrast, is able to produce appropriate estimates since it analyzes the data on various scales simultaneously.

[5] Formally, $\#F$ 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 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 i, where the estimated classes are labeled according to the permutation π. Specifically, set $\widehat{\rho}_\pi(i) = \pi(k)$ if $i \in \widehat{G}_k = \widehat{G}_k^{[n-K_0]}$. With this notation at hand, we define $\#F = \min_{\pi \in \Pi} \sum_{i=1}^n 1(\rho(i) \neq \widehat{\rho}_\pi(i))$.

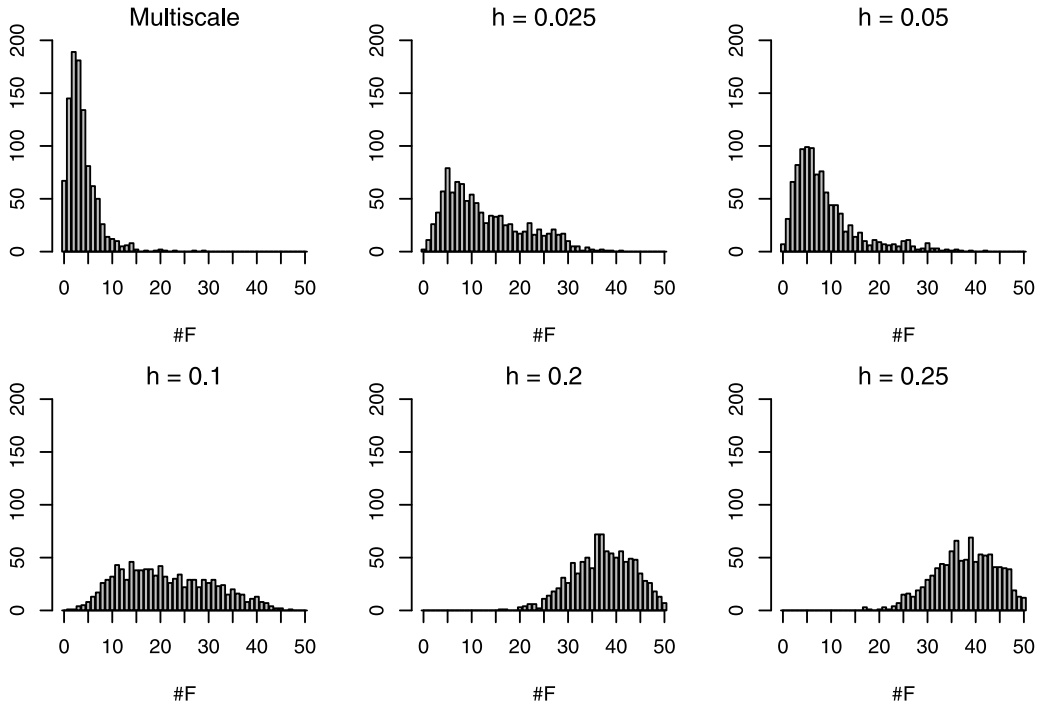


(a) Histograms of the number of classification errors $\#F$

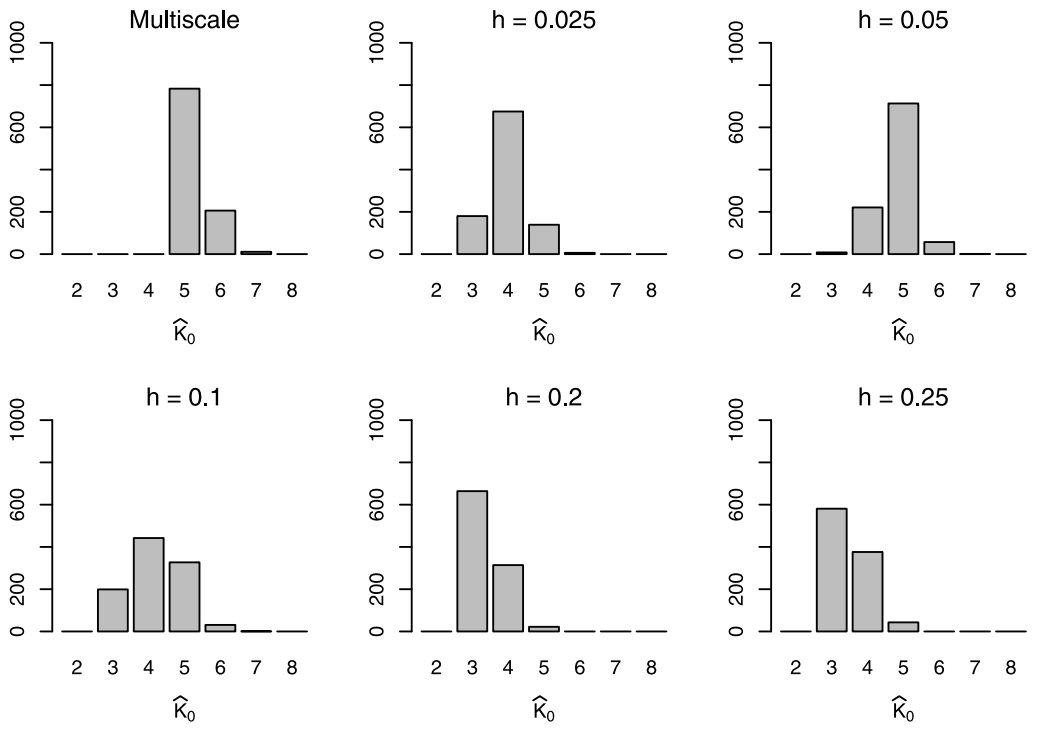


(b) Histograms of the estimated number of clusters $\widehat{K}_0$

Fig. 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 the 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 $\widehat{K}_0$

Fig. 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 the multiscale approach and the other panels those for the bandwidth-dependent competitor with different bandwidths h.

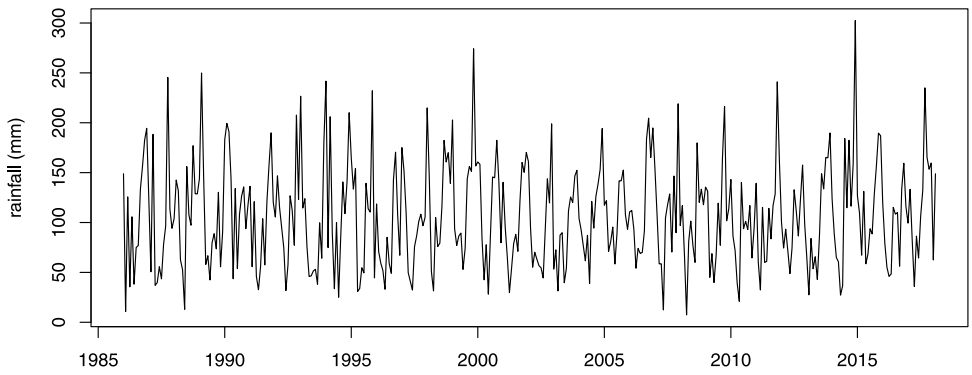


Fig. 5. Time series of total monthly rainfall (in mm) at Lerwick weather station.

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 $(1 \leq k \leq K_0)$ 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

We now illustrate the advantages of our multiscale clustering method by a real-data example. To do so, we compare it with the bandwidth-dependent algorithm introduced in Section 7.1. Unlike in simulations, the true groups are not known in real-data applications. Hence, we cannot simply evaluate the performance of the clustering algorithms by comparing the estimated groups with the true ones. Nevertheless, some clusterings may be more plausible than others in the light of the application context. In what follows, we use plausibility arguments to obtain a meaningful comparison of our multiscale method with the bandwidth-dependent alternative.

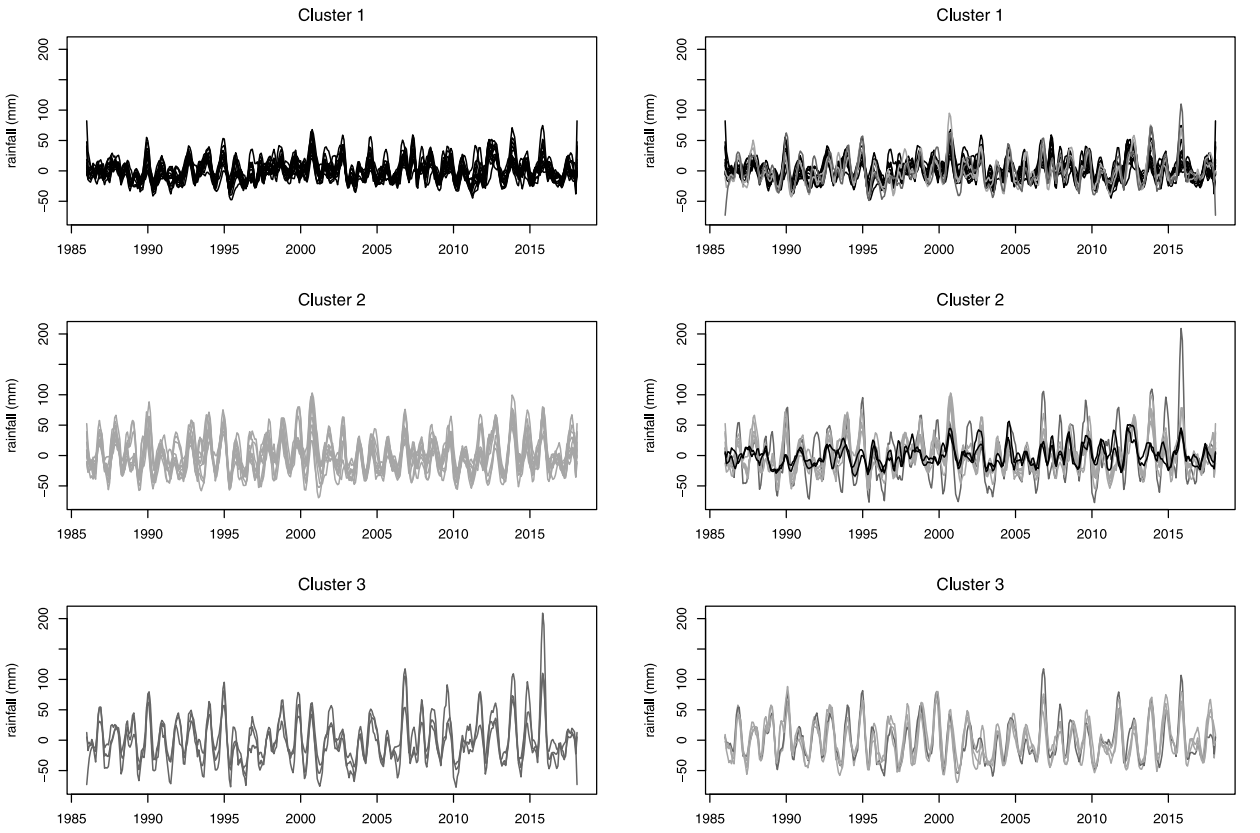
Our application example comes from environmental statistics. We examine a sample of monthly rainfall data from 34 UK weather stations. The data are publicly available on the webpage of the UK Met Office. We use a subset of 27 stations for which data are available over the time span from 1986 to 2018. We thus observe a time series $\mathcal{Y}_i = \{Y_{it} : 1 \leq t \leq T\}$ of length $T = 385$ for each station $i \in \{1, \ldots, 27\}$, where Y_{it} denotes total monthly rainfall (in mm) at station i at time t. Each of the $n = 27$ rainfall time series $\mathcal{Y}_i$ in our sample is assumed to follow the model

$$Y_{it} = m_i\Big(\frac{t}{T}\Big) + \alpha_i + \varepsilon_{it}, \tag{8.1}$$

where m_i is an unknown nonparametric time trend function which satisfies the normalization constraint $\int_0^1 m_i(u)du = 0$, α_i is a fixed effect error term and ε_{it} is an idiosyncratic error with $\mathbb{E}[\varepsilon_{it}] = 0$. As usual in nonparametric regression, we let m_i depend on rescaled time t/T rather than on real time t; see e.g. Robinson (1989), Dahlhaus (1997) and Vogt and Linton (2014) for the use and some discussion of the rescaled time argument. The trend function m_i describes the rainfall pattern at station i corrupted by noise $\alpha_i + \varepsilon_{it}$. Due to the normalization $\int_0^1 m_i(u)du = 0$, it holds that $T^{-1}\sum_{t=1}^T Y_{it} = \alpha_i + O_p(T^{-1/2})$, that is, the average rainfall level $T^{-1}\sum_{t=1}^T Y_{it}$ at station i is absorbed into the term α_i. As in the theoretical part of the paper, we assume that the stations i can be partitioned into K_0 groups $G_1, \ldots, G_{K_0}$ such that for each $1 \leq k \leq K_0$, $m_i = m_j$ for all $i, j \in G_k$. We thus suppose that the time trends m_i are the same (or at least very similar) at all stations i in a given group G_k.

To estimate the unknown group structure in the data, we implement our multiscale method as follows: We use an Epanechnikov kernel K to compute the local linear smoothers $\widehat{m}_{i,h}$ and consider the location-bandwidth grid $\mathcal{G}_T = \{(x, h) : x \in \mathcal{X} \text{ and } h \in \mathcal{H}\}$, where $\mathcal{X} = \{t/T : 1 \leq t \leq T\}$ is the set of locations and $\mathcal{H} = \{h : h = 3\ell/T \text{ with } 1 \leq \ell \leq 20\}$ is the set of bandwidths. The bandwidths $h = 3/T, 6/T, 9/T, \ldots$ correspond to effective sample sizes of $3, 6, 9, \ldots$ months of data. To implement the bandwidth-dependent algorithm, we use the same grid of locations $\mathcal{X}$ and consider different bandwidths h.

An example of the $n = 27$ rainfall time series in our sample is shown in Fig. 5. The plot depicts the time series of total monthly rainfall at Lerwick weather station. As can be clearly seen, the time series exhibits strong seasonal fluctuations. The underlying trend function m_i can thus be expected to strongly vary on local scales, in particular, over short time periods of only a few months. There may of course be variation in the function m_i on more global scales as well. However,



(a) Clusters produced by the algorithm with $h = 3/T$. Each panel shows the curve estimates $\widehat{m}_{i,h}$ that belong to a particular cluster.

(b) Clusters produced by the algorithm with $h = 48/T$. Each panel shows the curve estimates $\widehat{m}_{i,h}$ that belong to a particular cluster.

Fig. 6. Clusters produced by the bandwidth-dependent algorithm. The curve estimates $\widehat{m}_{i,h}$ are plotted with the bandwidth $h = 3/T$ in both subfigures for reasons of comparability. Their colors correspond to the clusters in subfigure (a).

the variation on local scales due to seasonal fluctuations in rainfall appears to be quite dominant. Hence, it seems crucial to take into account differences between the functions m_i on local scales when clustering the rainfall time series. As a consequence, a bandwidth-dependent algorithm can be expected to produce an appropriate clustering when implemented with a small bandwidth. When implemented with a large bandwidth, in contrast, it will presumably neglect important local differences between the functions m_i and thus produce inappropriate results.

This is illustrated in Fig. 6. Fig. 6(a) shows the clustering results produced by the bandwidth-dependent algorithm with the very small bandwidth $h = 3/T$, which corresponds to an effective sample size of 3 months of data. In order to estimate the number of clusters, we apply the procedure from Section 5 with $\alpha = 0.05$, which yields the estimate $\widehat{K}_0 = 3$. For reasons of comparability, we do not re-estimate the number of clusters when running the multiscale algorithm and the bandwidth-dependent method with other bandwidths h. The number of clusters is thus set to $\widehat{K}_0 = 3$ throughout the empirical analysis. Each panel of Fig. 6(a) represents one of the three estimated clusters and shows the curve estimates $\widehat{m}_{i,h}$ with $h = 3/T$ which belong to the respective cluster. The clusters appear to capture the local differences between the curves reasonably well. Curves with similar shape are sorted into the same group. In particular, the overall pattern of oscillations, their amplitudes and peaks are similar within each group.

Fig. 6(b) depicts the clusters produced by the bandwidth-dependent algorithm with the quite large bandwidth $h = 48/T$, which corresponds to an effective sample size of 4 years of data. As before, each panel shows the curve estimates $\widehat{m}_{i,h}$ of one cluster. For comparability reasons, the estimates $\widehat{m}_{i,h}$ are computed with the same bandwidth $h = 3/T$ as in Fig. 6(a) and their colors correspond to the clusters in Fig. 6(a). Fig. 6 illustrates two important points: First, the bandwidth-dependent algorithm yields quite different clusters depending on which bandwidth is used. Second, the algorithm with $h = 48/T$ is not able to detect the local differences between the curves appropriately. Cluster 2, for example, contains curves of quite different shapes, some having a strong oscillatory pattern whereas others have fluctuations with a much smaller amplitude.

Our multiscale approach produces exactly the same clusters as the bandwidth-dependent algorithm with the small bandwidth $h = 3/T$, which are depicted in Fig. 6(a). As argued above, it is quite plausible to suppose that the functions

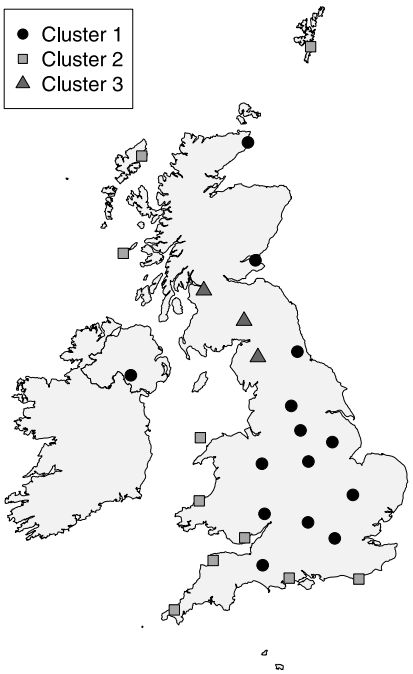


Fig. 7. Map of the UK with the locations of the $n = 27$ weather stations in our sample. The different symbols that indicate the stations correspond to the clusters produced by our multiscale approach.

m_i differ predominantly on local scales. Importantly, we do not feed this information into the multiscale algorithm. The method is rather designed to automatically select the important scales on which the functions m_i mainly differ. In the data example at hand, this appears to work quite well: The multiscale algorithm picks out very local scales, which are quite plausibly the most important ones. As a consequence, it produces clusters which reflect the seasonal fluctuations in the data quite well. The performance of the bandwidth-dependent algorithm, in contrast, strongly depends on the chosen bandwidth.

Fig. 7 presents a map of Great Britain which shows the locations of the $n = 27$ weather stations in our sample. The symbols that indicate the stations reflect the clustering produced by our multiscale approach. In particular, the stations that belong to a specific cluster are depicted by the same symbol. As can be seen, most stations of Cluster 1 are located in the eastern part of the UK, whereas those of Cluster 2 are mainly situated in the western part. The clusters thus show a clear division between the west and east of the UK. This makes sense as the UK weather is strongly influenced by winds from the Atlantic ocean that move from west to east across the UK, implying that the precipitation patterns in the west are different from those in the east.

Before we close this section, we should note that the real-data example we have considered here is of course not meant to be a full-blown empirical application. In serious environmental applications, data are collected on huge spatial grids, with rainfall, temperature or ozone measurements being available at hundreds or thousands of different locations. Our application example, in contrast, has a purely illustrative purpose. We have deliberately kept the example simple and the number of locations i small such that the main advantages of our multiscale method can be demonstrated in a clear and easy way.

Acknowledgments

We thank the Editor and two referees for their helpful and constructive comments on an earlier version of the paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2020.01.020.

References

Abraham, C., Cornillon, P.A., Matzner-Løber, E., Molinari, N., 2003. Unsupervised curve clustering using B-splines. Scand. J. Stat. 30, 581–595.
Andrews, D.W.K., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 59, 817–858.
Armstrong, T.B., Chan, H.P., 2016. Multiscale adaptive inference on conditional moment inequalities. J. Econometrics 194, 24–43.

Boneva, L., Linton, O., Vogt, M., 2015. A semiparametric model for heterogeneous panel data with fixed effects. J. Econometrics 188, 327–345.
Boneva, L., Linton, O., Vogt, M., 2016. The effect of fragmentation in trading on market quality in the UK equity market. J. Appl. Econometrics 31, 192–213.
Bonhomme, S., Manresa, E., 2015. Grouped patterns of heterogeneity in panel data. Econometrica 83, 1147–1184.
Box, G.E.P., Hamming, W.J., Tiao, G.C., 1975. A statistical analysis of the Los Angeles ambient carbon monoxide data. J. Air Pollut. Control Assoc. 25, 1129–1136.
Chaudhuri, P., Marron, J., 1999. SiZer for exploration of structures in curves. J. Amer. Statist. Assoc. 94, 807–823.
Chaudhuri, P., Marron, J., 2000. Scale space view of curve estimation. Ann. Statist. 28, 408–428.
Chiou, J.-M., Li, P.-L., 2007. Functional clustering and identifying substructures of longitudinal data. J. R. Stat. Soc. Ser. B Stat. Methodol. 69, 679–699.
Dahlhaus, R., 1997. Fitting time series models to nonstationary processes. Ann. Statist. 25, 1–37.
Degras, D., Xu, Z., Zhang, T., Wu, W.B., 2012. Testing for parallelism among trends in multiple time series. IEEE Trans. Signal Process. 60, 1087–1097.
Degryse, H., De Jong, F., Van Kervel, V., 2014. The impact of dark trading and visible fragmentation on market quality. Rev. Finance 1–36.
Dümbgen, L., Spokoiny, V.G., 2001. Multiscale testing of qualitative hypotheses. Ann. Statist. 29, 124–152.
Eckle, K., Bissantz, N., Dette, H., 2017. Multiscale inference for multivariate deconvolution. Electron. J. Stat. 11, 4179–4219.
Hansen, B., 2008. Uniform convergence rates for kernel estimation with dependent data. Econometric Theory 24, 726–748.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Springer.
Horowitz, J.L., 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., Preda, C., 2014. Functional data clustering: a survey. Adv. Data Anal. Classif. 8, 231–255.
James, M., Sugar, C.A., 2003. Clustering for sparsely sampled functional data. J. Amer. Statist. Assoc. 98, 397–408.
de Jong, R.M., Davidson, J., 2000. Consistency of kernel estimators of heteroscedastic and autocorrelated covariance matrices. Econometrica 68, 407–423.
Luan, Y., Li, H., 2003. Clustering of time-course gene expression data using a mixed-effects model with B-splines. Bioinformatics 19, 474–482.
Niu, X., Tiao, G.C., 1995. Modeling satellite ozone data. J. Amer. Statist. Assoc. 90, 969–983.
O'Hara, M., Ye, M., 2009. Is fragmentation harming market quality? J. Financ. Econ. 100, 459–474.
Proksch, K., Werner, F., Munk, A., 2018. Multiscale scanning in inverse problems. Ann. Statist. 46, 3569–3602.
Reinsel, G.C., Tiao, G.C., DeLuisi, J.J., Basu, S., Carriere, K., 1989. Trend analysis of aerosol-corrected Umkehr ozone profile data through 1987. J. Geophys. Res. Atmospheres 94, 16373–16386.
Robinson, P.M., 1989. Nonparametric estimation of time-varying parameters. In: Hackl, P. (Ed.), Statistical Analysis and Forecasting of Economic Structural Change. Springer, pp. 253–264.
Sacks, J., Ylvisaker, D., 1970. Designs for regression problems with correlated errors. III. Ann. Math. Stat. 41, 2057–2074.
Schmidt-Hieber, J., Munk, A., Dümbgen, L., 2013. Multiscale methods for shape constraints in deconvolution: confidence statements for qualitative features. Ann. Statist. 41, 1299–1328.
Su, L., Ju, G., 2018. Identifying latent grouped patterns in panel data models with interactive fixed effects. J. Econometrics 206, 554–573.
Su, L., Shi, Z., 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. Amer. Statist. 61, 34–40.
Tarpey, T., Kinateder, K.K.J., 2003. Clustering functional data. J. Classification 20, 93–114.
Vogt, M., Linton, O., 2014. Nonparametric estimation of a periodic sequence in the presence of a smooth trend. Biometrika 101, 121–140.
Vogt, M., Linton, O., 2017. Classification of non-parametric regression functions in longitudinal data models. J. R. Stat. Soc. Ser. B Stat. Methodol. 79, 5–27.
Wang, W., Phillips, P.C.B., Su, L., 2018. Homogeneity pursuit in panel data models: theory and application. J. Appl. Econometrics 33, 797–815.
Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. J. Amer. Statist. Assoc. 58, 236–244.