Multiscale clustering of nonparametric regression curves

Michael Vogt
Oliver Linton

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP08/18

Multiscale Clustering of Nonparametric Regression Curves

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

We study a longitudinal data model with nonparametric regression functions that may vary across the observed subjects. In a wide range of applications, it is natural to assume that not every subject has a completely different regression function. We may rather suppose that the observed subjects can be grouped into a small number of classes whose members share the same regression curve. We develop a bandwidth-free clustering method to estimate the unknown group structure from the data. More specifically, we construct estimators of the unknown classes 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. The technical analysis is complemented by a simulation study and an application to temperature anomaly data.

Key words: Clustering of nonparametric curves; nonparametric regression; multiscale statistics; longitudinal/panel data.
AMS 2010 subject classifications: 62G08; 62G20; 62H30.

1 Introduction

In this paper, we are concerned with the problem of clustering nonparametric regression curves in a longitudinal data framework. We consider the following model setup: We observe data $\{(Y_{it}, x_{it}) : 1 \le t \le T_i\}$ for n different subjects $i = 1, \dots, n$. The data of subject i satisfy the nonparametric regression equation

$$Y_{it} = m_i(x_{it}) + \varepsilon_{it} \tag{1.1}$$

for $t = 1, \dots, T_i$, where m_i is an unknown smooth function, x_{it} are deterministic or random design points and ε_{it} denotes the error term. The subjects in our sample are supposed to belong to K_0 different classes. More specifically, the set of subjects $\{1, \dots, n\}$ can be partitioned into K_0 groups $G_1, \dots, G_{K_0}$ such that for each $k = 1, \dots, K_0$,

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

Hence, the members of each group G_k all have the same regression function. In Section 2, we introduce model (1.1)–(1.2) in detail.

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

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

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). A number of estimation methods have been proposed in the context of functional data models related to (1.1)–(1.2); 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, Ray and Mallick (2006) for a Bayesian approach and Jacques and Preda (2014) for a recent survey. In these functional data models, the design points usually represent time and are thus deterministic. Hence, fixed design settings are analyzed. In the random design case where the regressors x_{it} are stochastic, the literature is much more sparse. Abraham et al. (2003) allow for random design points, though only under the very strong restriction that the design points are independent of the error terms. An estimation method for the random design case under much more general conditions has recently been developed in Vogt and Linton (2017).

Most of the proposed procedures have the following drawback: they depend on a number of smoothing parameters required to estimate the nonparametric functions m_i. A common approach is to approximate the functions m_i by a series expansion $m_i(x) \approx \sum_{j=1}^{N} \beta_{ij} \phi_j(x)$, where $\{\phi_j : j = 1, 2, \ldots\}$ is a function basis and N is the number of basis elements taken into account for the estimation of m_i. Here, N plays the role of the smoothing parameter and may vary across i, that is, $N = N_i$. To estimate the classes $G_1, \ldots, G_{K_0}$, estimators $\widehat{\boldsymbol{\beta}}_i$ of the coefficient vectors $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{iN})^\top$ are clustered into groups by a standard clustering algorithm. Variants of this approach have for example been investigated in Abraham et al. (2003), Luan and Li (2003), Chiou and Li (2007) and Tarpey (2007). Another approach is to compute nonparametric estimators $\widehat{m}_i = \widehat{m}_{i,h}$ of the functions m_i for some smoothing parameter h (which may differ across i) and to calculate distances $\widehat{\rho}_{ij} = \rho(\widehat{m}_i, \widehat{m}_j)$ between the estimates $\widehat{m}_i$ and $\widehat{m}_j$, where $\rho(\cdot, \cdot)$ is a distance measure such as a supremum or an L_2-distance. A distance-based clustering algorithm is then applied to the distances $\widehat{\rho}_{ij}$. This strategy has for example been used in Vogt and Linton (2017). As is well-known, smoothing parameter selection is a quite delicate matter. In general, nonparametric curve estimates strongly depend on the chosen smoothing parameters. A clustering procedure which is based on nonparametric estimates of the curves m_i can thus be expected to be markedly influenced by the choice of smoothing parameters as well.

The main aim of this paper is to construct estimators of the unknown groups $G_1, \ldots, G_{K_0}$ and of their unknown number K_0 in model (1.1)–(1.2) which are free of classical smoothing or bandwidth parameters. To achieve this, we make use of techniques from statistical multiscale testing studied e.g. in Chaudhuri and Marron (1999), Dümbgen and Spokoiny (2001), Hannig and Marron (2006) and Schmidt-Hieber et al. (2013). More specifically, we develop multiscale statistics which measure the distances between pairs of functions m_i and m_j. To construct these statistics, we estimate the functions m_i and m_j at different resolution levels, that is, with the help of different bandwidths h. The resulting estimators are aggregated in supremum-type statistics. We thereby obtain multiscale statistics which simultaneously take into account multiple bandwidth levels and thus avoid the need to pick a specific bandwidth. To estimate the unknown classes $G_1, \ldots, G_{K_0}$, we combine the constructed multiscale statistics with a hierarchical clustering algorithm. To estimate the unknown number of classes K_0, we develop a thresholding rule that is applied to the dendrogram produced by the clustering algorithm.

Apart from being free of classical bandwidth parameters, our estimation methods have the following features:

(a) They can be applied in both the fixed and the random design case. We thus allow the design points x_{it} to be either deterministic or random. Section 2 provides a detailed description of the fixed and the random design model we work with.

(b) The multiscale statistics on which our methods are based need not be combined with a hierarchical clustering algorithm as proposed in Section 4. Alternatively, they 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 detail in Section 9.

(c) By construction, our methods allow to detect differences between the functions m_i at different scales or resolution levels. Local differences can be detected by inspecting the functions on a high resolution level, that is, by means of small bandwidths. Global differences can be spotted by examining the functions on a low resolution level, that is, by means of large bandwidths. Another (very different) way to construct clustering methods which allow to detect differences between the functions m_i at multiple scales is to employ Wavelet methods. In a fixed design setting, a Wavelet-based approach has for example been suggested in Ray and Mallick (2006). In many applications, it is not clear at all on which resolution levels the functions m_i mainly differ. Being able to detect differences on multiple scales is thus crucial to obtain a reliable clustering algorithm.

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 methods. Section 4 explains how to combine them with a hierarchical clustering algorithm to estimate the unknown classes $G_1, \ldots, G_{K_0}$ in model (1.1)–(1.2). In Section 5, we finally introduce our procedure to estimate the unknown number of classes K_0. The main theoretical result of the paper is laid out in Section 6. This result characterizes the asymptotic convergence behaviour of the multiscale statistics and forms the basis to derive the theoretical properties of our clustering methods. To explore the finite sample properties of our methods, we conduct a simulation study in Section 7. Moreover, we apply our procedure to a sample of temperature anomaly data from the Berkeley Earth project in Section 8. The aim of the application is to cluster the spatial locations in our sample into geographical regions which are characterized by distinct temperature anomaly profiles, or put differently, by distinct climate change patterns.

2 The model

We now introduce the model framework in detail which underlies our analysis. We develop estimation methods for both a fixed and a random design setting.

Fixed design model. The data $\{(Y_{it}, x_{it}) : 1 \le t \le T_i\}$ of each subject $i = 1, \ldots, n$ satisfy the model equation

$$Y_{it} = m_i(x_{it}) + \varepsilon_{it} \tag{2.1}$$

with $\mathbb{E}[\varepsilon_{it}] = 0$ for $1 \le t \le T_i$, where m_i is an unknown nonparametric function and x_{it} are deterministic design points. For simplicity, we restrict attention to real-valued design points, the theory carrying over to the multivariate case in a straightforward way. The points x_{it} are normalized to lie in the unit interval, that is, $0 \le x_{i1} < \ldots < x_{iT_i} \le 1$. An important special case is the uniform design with $x_{it} = t/T_i$. We do not only consider this special case but allow for a wide range of non-uniform designs. Technically speaking, we assume that the design points 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

$$\int_{x_{i,t-1}}^{x_{it}} f_i(w)dw = \frac{1}{T_i} \quad \text{for } t = 1, \ldots, T_i,$$

where we set $x_{i0} = 0$. The densities f_i are supposed to fulfill certain regularity conditions specified in Section 6. Roughly speaking, we require them to be sufficiently smooth and to be bounded away from zero on their support $[0, 1]$. Note that by setting $f_i \equiv 1$ for all i, we obtain the special case of a uniform design with $x_{it} = t/T_i$. The error terms ε_{it} have the property that $\mathbb{E}[\varepsilon_{it}] = 0$ for all i and t, implying that $\mathbb{E}[Y_{it}] = m_i(x_{it})$. They are allowed to be correlated both across i and t. The exact conditions on their dependence structure are summarized in assumptions ($\text{C}_{\text{FD}}1$) and (C4) in Section 6 and are briefly discussed in comments (a) and (b) in Section 6.

Random design model. We have data $\{(Y_{it}, X_{it}) : 1 \le t \le T_i\}$ for n different subjects $i = 1, \ldots, n$. The data of subject i follow the model

$$Y_{it} = m_i(X_{it}) + \varepsilon_{it} \tag{2.2}$$

with $\mathbb{E}[\varepsilon_{it}|X_{it}] = 0$ for $1 \le t \le T_i$, where X_{it} are random design points. Here and in what follows, we use the upper case letter X_{it} to distinguish the random from the fixed design points. As in the fixed design case, we restrict attention to real-valued regressors X_{it}, the theory easily extending to the vector-valued case. We suppose that the variables X_{it} have compact support, which w.l.o.g. is equal to $[0,1]$. As before, the errors ε_{it} are allowed to be correlated both across i and t. The exact conditions on their dependence structure can be found in (C_{RD}1) and (C4) in Section 6.

Group structure. We impose the following group structure on both the fixed and the random design model: there are K_0 groups of subjects $G_1, \ldots, G_{K_0}$ with $\dot{\bigcup}_{k=1}^{K_0} G_k = \{1, \ldots, n\}$ such that for each $1 \le k \le K_0$,

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

Put differently, $m_i = g_k$ for all $i \in G_k$, where g_k is the group-specific regression function associated with the class G_k. Hence, the subjects of a given class G_k all have the same regression curve g_k. To make sure that subjects of different classes have different regression curves, we suppose that $g_k \ne g_{k'}$ for $k \ne k'$. The exact technical conditions on the functions g_k are summarized in (C6) in Section 6. To keep the exposition simple, 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 subjects n. We comment on this in more detail in Section 9. The groups $G_k = G_{k,n}$ depend on the cross-section dimension n in general. For ease of notation, we however suppress this dependence on n throughout the paper.

Dimensions n and T_i. We impose the following conditions on the sample sizes T_i and the number of subjects n:

(a) The sample sizes T_i all tend to infinity. Technically speaking, we regard $T_i = \tau_i(T)$ as a function $\tau_i : \mathbb{N} \to \mathbb{N}$ of some underlying sample size parameter T and suppose that $T_i = \tau_i(T) \to \infty$ as $T \to \infty$ for any i.

(b) The sample sizes T_i may differ across i. However, they are not allowed to differ too much in the sense that they all grow at the same rate T. Technically speaking, we assume the following: there exist constants c_i with $0 < \underline{c} \le c_i \le \overline{c} < \infty$ such that

$$\left| \frac{T_i}{T} - c_i \right| \le \rho(T) \to 0 \quad \text{as} \quad T \to \infty \tag{2.4}$$

for all $1 \le i \le n$, where $\underline{c}$, $\overline{c}$ and $\rho(\cdot)$ do not depend on i. (2.4) in particular implies that $T_i / T \to c_i$ for each i.

(c) The number of subjects n may either be fixed or diverging. We only impose the condition that n does not grow too quickly as compared to the sample sizes T_i. Technically speaking, we regard n as a function of T, that is, $n = n(T)$, and suppose that $n \le CT^{\rho}$, where $C > 0$ is a fixed constant and the parameter $\rho > 0$

is specified by conditions (C8) and (C9) in Section 6.

W.l.o.g. we assume that (i) $\max_{1 \le i \le n(T)} c_i \to \overline{c}$ as $T \to \infty$ and (ii) $\overline{c} = 1$. (i) is no restriction at all since we can set $\overline{c} = \lim_{T\to\infty} \max_{1\le i \le n(T)} c_i$. Moreover, (ii) can always be satisfied by replacing T with $\widetilde{T} = \overline{c}T$ and by writing $T_i = \widetilde{\tau}_i(\widetilde{T})$ along with $\widetilde{\tau}_i(\cdot) = \tau_i(\cdot/\overline{c})$. Under (i) and (ii), (2.4) implies that $\max_{1\le i\le n(T)} T_i/T \to 1$ as $T \to \infty$. Hence, the constants c_i can be approximated by $\widehat{c}_i = T_i / \max_{1 \le i \le n} T_i$. To simplify notation, we use the shorthand $T_{\max} = \max_{1 \le i \le n} T_i$ in what follows. Throughout the paper, asymptotic statements are to be understood in the sense that $T \to \infty$.

3 The multiscale distance statistics

3.1 Construction of the statistics

In what follows, we construct multiscale statistics $\widehat{d}_{ij}$ which estimate the distance between any two functions m_i and m_j. To do so, let d_{ij} be a measure of distance between m_i and m_j. In particular, consider the supremum distance $d_{ij} = \sup_{x \in [0,1]} |m_i(x) - m_j(x)|$. A straightforward estimator of d_{ij} is

$$\widetilde{d}_{ij,h} = \sup_{x \in [h,1-h]} |\widehat{m}_{i,h}(x) - \widehat{m}_{j,h}(x)|,$$

where $\widehat{m}_{i,h}$ denotes some kernel estimator of the function m_i and h is the bandwidth. For simplicity, we take the supremum over all $x \in [h, 1-h]$ rather than over all $x \in [0,1]$ in the definition of $\widetilde{d}_{ij,h}$ to avoid boundary effects. In Section 9, we outline some technical modifications which allow the supremum to run over the whole unit interval. The estimator $\widetilde{d}_{ij,h}$ obviously depends on the chosen bandwidth h. To get rid of this dependence, we compute $\widetilde{d}_{ij,h}$ not only for a single bandwidth h but for a wide range of different bandwidths. This leaves us with a whole family of statistics $\{\widetilde{d}_{ij,h} : h_{\min} \le h \le h_{\max}\}$, where $h_{\min}$ and $h_{\max}$ are the minimal and maximal bandwidths that are taken into account, respectively. We now define an estimator of d_{ij} by taking the supremum over all the statistics in this family. Specifically, we set

$$\widetilde{d}_{ij} = \sup_{h \in [h_{\min}, h_{\max}]} |\widetilde{d}_{ij,h}| = \sup_{h \in [h_{\min}, h_{\max}]} \sup_{x \in [h,1-h]} |\widehat{m}_{i,h}(x) - \widehat{m}_{j,h}(x)|. \tag{3.1}$$

This is a rudimentary multiscale statistic which serves as a starting point for the construction of our multiscale distance statistic $\widehat{d}_{ij}$. The statistic $\widetilde{d}_{ij}$ does not depend on a specific bandwidth h but takes into account a wide range of different bandwidths $h \in [h_{\min}, h_{\max}]$ simultaneously. It is thus free of a classical bandwidth parameter that needs to be selected. To compute it, we only have to choose the minimal and maximal bandwidth levels $h_{\min}$ and $h_{\max}$. In Section 3.2, we explain in detail how to pick $h_{\min}$ and $h_{\max}$ in practice.

The multiscale statistic $\widetilde{d}_{ij}$ has the following drawback: By definition, it is the supremum over the statistics

$$\widetilde{\psi}_{ij}(x,h) := \widehat{m}_{i,h}(x) - \widehat{m}_{j,h}(x),$$

where the supremum runs over all $x \in [h, 1-h]$ and all $h \in [h_{\min}, h_{\max}]$. In general, the statistics $\widetilde{\psi}_{ij}(x,h)$ may have a very different stochastic behaviour across x and h as well as across i and j. In particular, their variance may strongly differ across x, h, i and j. Hence, the supremum $\widetilde{d}_{ij} = \sup_{x,h} |\widetilde{\psi}_{ij}(x,h)|$ may be dominated by only a small number of random variables $\widetilde{\psi}_{ij}(x,h)$ with a very large variance. To avoid this issue and to put the statistics $\widetilde{\psi}_{ij}(x,h)$ on an equal footing, we replace them by normalized versions. In particular, we normalize them such that their variances are approximately equal to 1 for all x, h, i and j. We now explain how to obtain such a normalization. We first consider the fixed and then the random design case.

Fixed design model. For technical reasons concerning the smoothing bias, we estimate the functions m_i by the local linear smoothers

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

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

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

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

Suppose that the technical conditions from Section 6 are satisfied and assume for a moment that the errors ε_{it} are independent across i and t. In this case, standard calculations yield that for any given point $x \in (0,1)$,

$$\sqrt{T_{\max} h}\, \Big[\big(\widehat{m}_{i,h}(x) - \widehat{m}_{j,h}(x)\big) - B_{ij,h}(x) \Big] \xrightarrow{d} N\big(0, \nu_{ij}(x)\big), \tag{3.4}$$

where $T_{\max} = \max_{1 \le i \le n} T_i$ and

$$B_{ij,h}(x) = \big\{m_i(x) - m_j(x)\big\} + \frac{h^2}{2}\kappa_2 \big\{m_i''(x) - m_j''(x)\big\} + O_p(h^3) \tag{3.5}$$

$$\nu_{ij}(x) = \Big\{ \frac{\sigma_i^2}{c_i f_i(x)} + \frac{\sigma_j^2}{c_j f_j(x)} \Big\} \|K\|_2^2. \tag{3.6}$$

Here, $\sigma_i^2 = \mathbb{E}[\varepsilon_{it}^2]$ denotes the error variance and we make use of the shorthands

$\kappa_2 = \int \varphi^2 K(\varphi) d\varphi$ and $\|K\|_2^2 = \int K^2(\varphi) d\varphi$. If the subjects i and j belong to the same class, that is, if $m_i = m_j$, we in particular obtain that $B_{ij,h}(x) = O_p(h^3)$ and

$$\sqrt{T_{\max} h}\,\big(\widehat{m}_{i,h}(x) - \widehat{m}_{j,h}(x)\big) \xrightarrow{d} N\big(0, \nu_{ij}(x)\big) \tag{3.7}$$

for any h with $Th \to \infty$ and $h = o(T^{-1/7})$.

The normality results (3.4) and (3.7) motivate to replace the variables $\widetilde{\psi}_{ij}(x,h)$ by the normalized versions

$$\widehat{\psi}_{ij}(x,h) = \sqrt{T_{\max} h}\big(\widehat{m}_{i,h}(x) - \widehat{m}_{j,h}(x)\big) \Big/ \sqrt{\widehat{\nu}_{ij,h}(x)}, \tag{3.8}$$

where

$$\widehat{\nu}_{ij,h}(x) = \left\{ \frac{\widehat{\sigma}_{i,h}^2}{\widehat{c}_i \widehat{f}_{i,h}(x)} + \frac{\widehat{\sigma}_{j,h}^2}{\widehat{c}_j \widehat{f}_{j,h}(x)} \right\} \|K\|_2^2 \tag{3.9}$$

is an estimator of $\nu_{ij}(x)$. Here, $\widehat{f}_{i,h}(x) = T_i^{-1} \sum_{t=1}^{T_i} K_h(x_{it} - x)$ is a kernel estimator of the design density $f_i(x)$ and $\widehat{\sigma}_{i,h}^2 = T_i^{-1} \sum_{t=1}^{T_i} \{Y_{it} - \widehat{m}_{i,h}(x_{it})\}^2$ is an estimator of the error variance σ_i^2. Standard arguments show that the normalized random variables $\widehat{\psi}_{ij}(x,h)$ are asymptotically normal with unit variance for any pair of subjects i and j from the same class, for any $x \in (0,1)$ and for any h with $Th \to \infty$ and $h = o(T^{-1/7})$. Hence, their variances can be expected to be approximately equal to 1 in finite samples as well, provided that the sample sizes T_i are sufficiently large.

To construct the normalized statistics $\widehat{\psi}_{ij}(x,h)$, we have made the simplifying assumption that the errors ε_{it} are independent across i and t. In principle, it is straightforward to take into account dependencies in the errors when setting up the statistics $\widehat{\psi}_{ij}(x,h)$. Suppose for example that the variables ε_{it} are (weakly) dependent across t but independent across i. The normality results (3.4) and (3.7) remain to hold true in this case if we substitute σ_i^2 in the asymptotic variance $\nu_{ij}(x)$ by the long-run error variance $\Gamma_i = \sum_{\ell=-\infty}^{\infty} \gamma_i(\ell)$ with $\gamma_i(\ell) = \mathbb{E}[\varepsilon_{it}\varepsilon_{i,t+\ell}]$. Hence, we can simply replace the estimator $\widehat{\sigma}_{i,h}^2$ in (3.9) by a suitable estimator of Γ_i to obtain statistics $\widehat{\psi}_{ij}(x,h)$ which are asymptotically normal with unit variance. From a practical point of view, it is however not recommendable to proceed in this way. The long-run variances Γ_i are quite difficult to estimate, in particular much more difficult than σ_i^2. It is thus quite likely that the estimates of Γ_i are fairly poor at least for some indices i, implying that the statistics $\widehat{\psi}_{ij}(x,h)$ are normalized inappropriately for these indices.

The statistics $\widehat{\psi}_{ij}(x,h)$ defined in (3.8), in contrast, are much simpler to estimate and have a more robust behaviour. Moreover, they are not only useful when the errors ε_{it} are independent across i and t. As long as the dependence in the errors is not too strong, the normalization (3.9) can be expected to make sure that the statistics $\widehat{\psi}_{ij}(x,h)$ have variances of comparable size across x, h, i and j, even though the asymptotic variances are not exactly equal to 1 in general. Hence, from a practical point of view, it makes sense to work with the statistics $\widehat{\psi}_{ij}(x,h)$ defined in (3.8) even when the errors ε_{it} are dependent across i and t. These statistics are also a valid choice from a theoretical point of view. Notably, our theory does not require us to work with statistics that have unit variance asymptotically. We can thus base our methods on the statistics $\widehat{\psi}_{ij}(x,h)$ from (3.8) while allowing the errors to be correlated across i and t. The exact conditions on the dependence structure of the error terms ε_{it} are specified in ($\mathrm{C}_{\mathrm{FD}}1$), ($\mathrm{C}_{\mathrm{RD}}1$) and (C4) in Section 6. In addition, they are briefly discussed in comments (a) and (b) in Section 6.

Random design model. The estimators of the functions m_i are constructed in exactly the same way as in the fixed design case. Specifically, we define the estimator $\widehat{m}_{i,h}$ of m_i as described in equations (3.2) and (3.3), with the fixed design points x_{it} replaced by the random points X_{it}. Under the technical conditions from Section 6, it holds that for any given $x \in (0,1)$,

$$\sqrt{T_{\max}h}\,\Big[\big(\widehat{m}_{i,h}(x) - \widehat{m}_{j,h}(x)\big) - B_{ij,h}(x)\Big] \xrightarrow{d} N\big(0, \nu_{ij}(x)\big), \tag{3.10}$$

where the bias $B_{ij,h}(x)$ has the same structure (3.5) as in the fixed design case. The asymptotic variance has the form

$$\nu_{ij}(x) = \Big\{\frac{\sigma_i^2(x)}{c_i f_i(x)} + \frac{\sigma_j^2(x)}{c_j f_j(x)}\Big\}\|K\|_2^2, \tag{3.11}$$

where $\sigma_i^2(x) = \mathbb{E}[\varepsilon_{it}^2|X_{it} = x]$ is the conditional error variance. If the errors are homoskedastic, that is, if $\sigma_i^2(x) \equiv \sigma_i^2 = \mathbb{E}[\varepsilon_{it}^2]$ for all x, we obtain the following result which is completely analogous to that from the fixed design case: For any pair of subjects i and j from the same class, for any $x \in (0,1)$ and for any h with $Th \to \infty$ and $h = o(T^{-1/7})$, the statistic

$$\widehat{\psi}_{ij}(x,h) = \sqrt{T_{\max}h}\big(\widehat{m}_{i,h}(x) - \widehat{m}_{j,h}(x)\big)\Big/\sqrt{\widehat{\nu}_{ij,h}(x)} \tag{3.12}$$

is asymptotically standard normal, where we define

$$\widehat{\nu}_{ij,h}(x) = \left\{\frac{\widehat{\sigma}_{i,h}^2}{\widehat{c}_i\widehat{f}_{i,h}(x)} + \frac{\widehat{\sigma}_{j,h}^2}{\widehat{c}_j\widehat{f}_{j,h}(x)}\right\}\|K\|_2^2 \tag{3.13}$$

with $\widehat{f}_{i,h}(x) = T_i^{-1}\sum_{t=1}^{T_i} K_h(X_{it} - x)$ and $\widehat{\sigma}_{i,h}^2 = T_i^{-1}\sum_{t=1}^{T_i}\{Y_{it} - \widehat{m}_{i,h}(X_{it})\}^2$. If the errors are heteroskedastic, we replace $\widehat{\sigma}_{i,h}^2$ in (3.13) by an estimator $\widehat{\sigma}_{i,h}^2(x)$ of the conditional error variance $\sigma_i^2(x) = \mathbb{E}[\varepsilon_{it}^2|X_{it} = x]$, for example by $\widehat{\sigma}_{i,h}^2(x) = T_i^{-1}\sum_{t=1}^{T_i} K_h(X_{it} - x)\{Y_{it} - \widehat{m}_{i,h}(X_{it})\}^2/\widehat{f}_{i,h}(x)$. For simplicity, we assume throughout the paper that the errors ε_{it} are homoskedastic. However, our theoretical arguments easily carry over to the case of heteroskedastic error terms.

Our discussion so far suggests to replace the multiscale statistics $\widetilde{d}_{ij}$ from (3.1) by the normalized versions

$$\widehat{d}_{ij} = \sup_{h \in [h_{\min}, h_{\max}]} \sup_{x \in [h, 1-h]} |\widehat{\psi}_{ij}(x,h)|, \tag{3.14}$$

where the statistics $\widehat{\psi}_{ij}(x,h)$ are defined in (3.8) and (3.12) for the fixed and the random design case, respectively. The asymptotic normality results (3.4) and (3.10) which motivate this normalization include the bias term $B_{ij,h}(x)$. To make sure that this bias is asymptotically negligible, we have assumed that $Th \to \infty$ with $h = o(T^{-1/7})$. We take this restriction into account by supposing that $h_{\min} \ge cT^{-(1-\delta)}$ and $h_{\max} \le CT^{-(1/7+\delta)}$ for some small $\delta > 0$ and positive constants c and C. These conditions on $h_{\min}$ and $h_{\max}$ are fairly moderate: Since the optimal bandwidth for estimating m_i is of the order $T^{-1/5}$ for any i under our technical conditions from Section 6, we can choose the interval $[h_{\min}, h_{\max}]$ to contain a wide variety of bandwidths, thus allowing for both substantial under- and oversmoothing.

Even though the statistics $\widehat{d}_{ij}$ are an improvement on the initial versions $\widetilde{d}_{ij}$, they still suffer from the following drawback: they do not take into account all bandwidths $h \in [h_{\min}, h_{\max}]$ in an equal fashion but tend to be dominated by small bandwidths h. The reason for this is as follows: Let i and j be two subjects that belong to the same class and suppose for simplicity that the variables (Y_{it}, x_{it}) and (Y_{it}, X_{it}) are independent across i and t in the fixed and the random design case, respectively. By construction, the statistics $\widehat{\psi}_{ij}((2\ell-1)h, h)$ for $\ell = 1, \ldots, \lfloor 1/2h \rfloor$ are approximately standard normal and independent for any given bandwidth h. Since the maximum over $\lfloor 1/2h \rfloor$ independent standard normal random variables is $C(2h) + o_p(1)$ as $h \to 0$ with $C(r) = \sqrt{2\log(1/r)}$, we obtain that $\max_\ell \widehat{\psi}_{ij}((2\ell-1)h,h)$ is approximately of size $C(2h)$ for small bandwidths h. Moreover, since the statistics $\widehat{\psi}_{ij}(x,h)$ with $(2\ell-1)h < x < (2\ell+1)h$ are correlated with $\widehat{\psi}_{ij}((2\ell-1)h,h)$ and $\widehat{\psi}_{ij}((2\ell+1)h,h)$, the supremum $\sup_x \widehat{\psi}_{ij}(x,h)$ approximately behaves as the maximum $\max_\ell \widehat{\psi}_{ij}((2\ell-1)h,h)$. As a result, we obtain that

$$\widehat{d}_{ij} \approx \sup_{h \in [h_{\min}, h_{\max}]} \max_{1 \le \ell \le \lfloor 1/2h \rfloor} |\widehat{\psi}_{ij}((2\ell-1)h,h)|,$$

where $\max_\ell |\widehat{\psi}_{ij}((2\ell-1)h,h)| \approx C(2h)$ for small values of h. In particular, the maximum $\max_\ell |\widehat{\psi}_{ij}((2\ell-1)h,h)|$ tends to have a much larger size for small than for large bandwidths h. This suggests that the stochastic behaviour of $\widehat{d}_{ij}$ is dominated by the statistics $\widehat{\psi}_{ij}(x,h)$ corresponding to small bandwidths h.

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

$$\widehat{d}_{ij} = \sup_{h \in [h_{\min}, h_{\max}]} \sup_{x \in [h, 1-h]} \big\{ |\widehat{\psi}_{ij}(x,h)| - C(2h) \big\}, \tag{3.15}$$

where $C(r) = \sqrt{2\log(1/r)}$. For each given bandwidth h, we thus subtract the additive correction term $C(2h)$ from the statistics $\widehat{\psi}_{ij}(x,h)$. According to the heuristic considerations from above, when i and j belong to the same class, the maximum of the statistics $\widehat{\psi}_{ij}(x,h)$ with the bandwidth h is approximately of size $C(2h)$ for small values of h. Hence, we correct the statistics $\widehat{\psi}_{ij}(x,h)$ corresponding to a small bandwidth h by subtracting the approximate size of their maximum. This puts the statistics $\widehat{\psi}_{ij}(x,h)$ with different bandwidth values h on a more equal footing and prevents small bandwidths from dominating the behaviour of the multiscale statistics.

To make the statistics $\widehat{d}_{ij}$ defined in (3.15) computable in practice, we finally replace the supremum over $x \in [h, 1-h]$ and $h \in [h_{\min}, h_{\max}]$ by the maximum over all points x and h in a suitable grid $\mathcal{G}_T$. We may work with any grid which has the following properties: $\mathcal{G}_T$ is a subset of

$$\mathcal{G} = \big\{(x,h) \,\big|\, h_{\min} \le h \le h_{\max} \text{ and } h \le x \le 1-h\big\}, \tag{3.16}$$

$\mathcal{G}_T$ becomes dense in $\mathcal{G}$ as $T \to \infty$, and $|\mathcal{G}_T| \le CT^{\beta}$ for some arbitrarily large but fixed constants $C, \beta > 0$, where $|\mathcal{G}_T|$ denotes the cardinality of $\mathcal{G}_T$. For example, we may use the Wavelet multiresolution grid $\mathcal{G}_T = \{(x,h) = (2^{-\nu}r, 2^{-\nu}) \,|\, 1 \le r \le 2^{\nu}-1 \text{ and } h_{\min} \le 2^{-\nu} \le h_{\max}\}$. With this notation at hand, we can make the following formal definition:

Definition 3.1. *For any pair of subjects i and j with $1 \le i, j \le n$, we call*

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

a multiscale distance statistic or, synonymously, a multiscale distance measure. Here, $C(r) = \sqrt{2\log(1/r)}$ and the expressions $\widehat{\psi}_{ij}(x,h)$ are defined in (3.8) *and* (3.12) *for the fixed and the random design, respectively. Moreover, $\mathcal{G}_T$ is any grid with the properties specified above.*

3.2 Tuning parameter choice

The multiscale statistics $\widehat{d}_{ij}$ do not depend on a specific bandwidth h that needs to be selected. They rather take into account a wide range of different bandwidths $h \in [h_{\min}, h_{\max}]$ simultaneously. They are thus free of classical bandwidth or smoothing parameters. However, they are of course not completely free of tuning parameters. They obviously depend 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, (i) they are much simpler to choose and (ii) their exact choice influences the estimation results much less. In what follows, we discuss the reasons for (i) and (ii) in detail and give some guidelines how to choose $h_{\min}$ and $h_{\max}$ appropriately in practice. These guidelines are used in particular to implement our methods in the simulations of Section 7 and the empirical application of Section 8. We first have a closer look at $h_{\min}$ and then turn to $h_{\max}$.

Choice of $\boldsymbol{h_{\min}}$. Ideally, we would like to make the interval $[h_{\min}, h_{\max}]$ as large as possible, thus taking into account as many different bandwidth values h as possible. In particular, we would like to choose $h_{\min}$ as small as possible. From a technical perspective, we can pick any value $h_{\min}$ such that $h_{\min} \geq cT^{-(1-\delta)}$ for some small $\delta > 0$ and some positive constant c. This allows $h_{\min}$ to converge to zero almost as quickly as the sample size parameter T and thus to be extremely small. Heuristically speaking, the bandwidth $h_{\min}$ can be considered very small if the effective sample size $T_i h_{\min}$ is very small, say $T_i h_{\min} \leq 10$ for all i (when an Epanechnikov kernel is used). Hence, from an applied point of view, it is clear in which range we have to pick the bandwidth $h_{\min}$. Moreover, the exact choice of $h_{\min}$ can be expected to have little effect on the estimation results. The reason is as follows: According to the heuristic discussion from the previous subsection, $\max_x |\widehat{\psi}_{ij}(x,h)| \approx C(2h) + o_p(1)$ as $h \to 0$, implying that $\max_x\{|\widehat{\psi}_{ij}(x,h)| - C(2h)\}$ tends to zero as $h \to 0$. This suggests that the multiscale statistic $\widehat{d}_{ij} = \max_{x,h}\{|\widehat{\psi}_{ij}(x,h)| - C(2h)\}$ should not attain its maximum at extremely small bandwidth values h. As a consequence, the precise value of $h_{\min}$ should barely influence the overall behaviour of the multiscale statistics. In view of these considerations, we propose to choose $h_{\min}$ in practice such that the effective sample size $T_i h_{\min}$ is small, say ≤ 10 for all i (when an Epanechnikov kernel is used).

Choice of $\boldsymbol{h_{\max}}$. Ideally, we would like to choose $h_{\max}$ as large as possible. From a technical point of view, we can pick any value $h_{\max}$ such that $h_{\max} \leq CT^{-(1/7+\delta)}$ for some small $\delta > 0$ and some positive constant C. This allows us to oversmooth substantially and to choose $h_{\max}$ much larger 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.

The condition that $h_{\max} \leq CT^{-(1/7+\delta)}$ is mainly needed because we allow the design densities f_i to be very different across i. In a wide range of applications, however, the design densities are fairly similar. In a fixed design context, for instance, the design points are often close to equidistant for all i. In these cases, the condition that $h_{\max} \leq CT^{-(1/7+\delta)}$ tends to be overly pessimistic and may be weakened. The reason is as follows: For i and j in the same class, we would like the difference $\widehat{m}_{i,h}(x) - \widehat{m}_{j,h}(x)$ to converge to zero sufficiently fast for any $h \in [h_{\min}, h_{\max}]$. To ensure this, the bias part of $\widehat{m}_{i,h}(x) - \widehat{m}_{j,h}(x)$ in particular needs to converge fast enough. The bias of the smoother $\widehat{m}_{i,h}(x)$ has the following two properties: it depends on the design density f_i and it tends to become larger when the bandwidth increases. Hence, if the design densities f_i and f_j strongly differ, $\widehat{m}_{i,h}(x) - \widehat{m}_{j,h}(x)$ can be expected to have a strong bias for large bandwidths h. To control this bias, we impose the condition that

$h_{\max} \le CT^{-(1/7+\delta)}$. If the design densities f_i and f_j do not differ so much, the bias of $\widehat{m}_{i,h}(x) - \widehat{m}_{j,h}(x)$ will commonly not be very pronounced for i and j in the same class in contrast. Hence, the condition that $h_{\max} \le CT^{-(1/7+\delta)}$ is overly restrictive in this case. As an example, consider the fixed design setting with $T_i = T$ and $f_i = f$ for all i. In this case, the bias terms of $\widehat{m}_{i,h}(x)$ and $\widehat{m}_{j,h}(x)$ are exactly the same for i and j in the same class. This allows us to drastically weaken the conditions on $h_{\max}$. In particular, we only require that $h_{\max} = o(1)$. (Indeed, we could even allow $h_{\max}$ not to converge to zero but to remain fixed.)

According to the above considerations, $h_{\max}$ can be chosen extremely large as long as the design densities do not differ strongly across i. From a heuristic perspective, the bandwidth value $h_{\max}$ can be regarded as extremely large if the effective sample size $T_i h_{\max}$ is very large as compared to the full sample size T_i for all i, say $T_i h_{\max} \approx T_i/4$ or $T_i h_{\max} \approx T_i/3$ (when an Epanechnikov kernel is used). Hence, it is clear in which range we need to pick the bandwidth $h_{\max}$ in practice. Moreover, the exact choice of $h_{\max}$ can be expected to have little influence on the estimation results: If we pick $h_{\max}$ very large, we smooth out virtually all features of the curves m_i and basically fit a straight line to the data. Whether we pick $h_{\max}$ a bit smaller or larger will not have a strong effect on the produced fits. In either case, we will end up with strongly oversmoothed, approximately linear estimates of the regression functions. In view of these points, we suggest to choose $h_{\max}$ in practice such that the effective sample size $T_i h_{\max}$ is large in comparison to the sample size T_i, say $T_i h_{\max} \ge T_i/4$ for all i (when an Epanechnikov kernel is used). This should yield a reasonable value for $h_{\max}$, at least as long as the design densities are not extremely different across i.

3.3 Properties of the multiscale statistics

We now discuss some theoretical properties of the multiscale statistics $\widehat{d}_{ij}$ 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 subjects i and j from the same class,

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

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

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

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

$$\max_{1 \le k \le K_0} \max_{i,j \in G_k} \widehat{d}_{ij} = O_p\big(\sqrt{\log n + \log T} + \sqrt{T h_{\max}^7}\big) \tag{3.17}$$

$$\min_{1 \le k < k' \le K_0} \min_{\substack{i \in G_k, \\ j \in G_{k'}}} \widehat{d}_{ij} \ge c_0 \sqrt{Th_{\max}} + o_p\big(\sqrt{Th_{\max}}\big), \tag{3.18}$$

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

$$\max_{1 \le k \le K_0} \max_{i,j \in G_k} \widehat{d}_{ij} \big/ \sqrt{Th_{\max}} = o_p(1) \tag{3.19}$$

$$\min_{1 \le k < k' \le K_0} \min_{\substack{i \in G_k, \\ j \in G_{k'}}} \widehat{d}_{ij} \big/ \sqrt{Th_{\max}} \ge c_0 + o_p(1). \tag{3.20}$$

According to (3.19) and (3.20), the maximal distance between subjects of the same class converges to zero when normalized by $\sqrt{Th_{\max}}$, whereas the minimal distance between subjects of two different classes remains bounded away from zero. Asymptotically, the distance measures $\widehat{d}_{ij}$ thus contain enough information to detect which subjects 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 subjects i and j with $\widehat{d}_{ij} \le c$ belong to the same class, whereas those with $\widehat{d}_{ij} > c$ belong to two different classes. The hierarchical clustering algorithm introduced in the next section exploits this information in the distances $\widehat{d}_{ij}$.

4 Estimation of the unknown classes

Let $S \subseteq \{1, \dots, n\}$ and $S' \subseteq \{1, \dots, n\}$ be two sets of subjects 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 subjects $\{1, \dots, n\}$ into groups, we combine the multiscale dissimilarity measure $\widehat{\Delta}$ with a hierarchical agglomerative clustering (HAC) algorithm which proceeds as follows:

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

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

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

and merge them into a new cluster.

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

We now examine the properties of our HAC algorithm. In particular, we investigate how the partitions $\{\widehat{G}_1^{[r]}, \dots, \widehat{G}_{n-r}^{[r]}\}$ for $r = 1, \dots, n-1$ are related to the true class structure $\{G_1, \dots, G_{K_0}\}$. From (3.19) and (3.20), it immediately follows that the multiscale statistics $\widehat{d}_{ij}$ have the following property:

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

To formulate the results on the HAC algorithm, we do not restrict attention to the multiscale statistics $\widehat{d}_{ij}$ from Definition 3.1 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, $\dot{\bigcup}_{\ell=1}^{r} A_\ell = \{1, \dots, n\}$ and $\dot{\bigcup}_{\ell=1}^{r'} B_\ell = \{1, \dots, n\}$. 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 $1 \le k \le K_0$ 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 $\widehat{G}_1^{[n-K]}, \ldots, \widehat{G}_K^{[n-K]}$ are subsets of the unknown classes with probability tending to 1. Conversely, by (c), for any $K < K_0$, the unknown classes are subsets of the estimated clusters with probability tending to 1.

5 Estimation of the unknown number of classes

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

$$\sqrt{\log n + \log T} + \sqrt{T h_{\max}^7} \ll \pi_{n,T} \ll \sqrt{T h_{\max}}, \tag{5.1}$$

where the notation $a_{n,T} \ll b_{n,T}$ means that $a_{n,T} = o(b_{n,T})$. Combining properties (3.17) and (3.18) of the multiscale distance statistics $\widehat{d}_{ij}$ with the statements of Theorem 4.1, we immediately obtain the following: For any $K < K_0$,

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

whereas for $K = K_0$,

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

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

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

is satisfied. Formally speaking, we estimate K_0 by

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

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

Theorem 5.1. *Suppose that the multiscale statistics $\widehat{d}_{ij}$ from Definition 3.1 have the properties* (3.17) *and* (3.18)*. 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 (3.17) and (3.18) 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 $\mathbb{P}(\widehat{K}_0 < K_0) =$

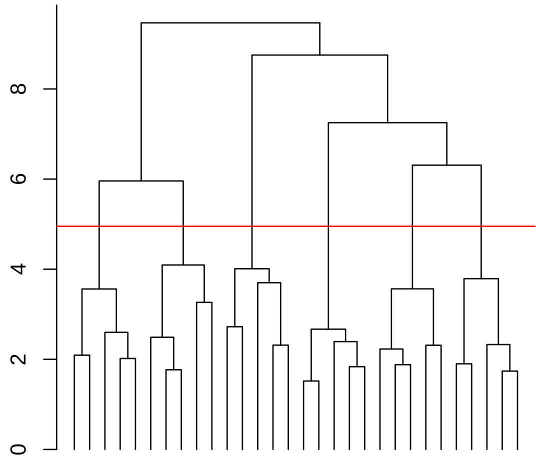

Figure 1: Example of a dendrogram produced by the HAC algorithm. The red horizontal line indicates the dissimilarity level $\pi_{n,T}$. The estimator $\widehat{K}_0$ can be computed by counting the vertical lines that intersect the red horizontal threshold. In the above example, $\widehat{K}_0$ is equal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

$o(1)$, whereas (5.3) yields that $\mathbb{P}(\widehat{K}_0 > K_0) = o(1)$. As a consequence, we obtain that $\mathbb{P}(\widehat{K}_0 = K_0) \to 1$.

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

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

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

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 $\sqrt{\log n + \log T} + \sqrt{T h_{\max}^7} \ll \pi_{n,T} \ll \sqrt{T h_{\max}}$. From an asymptotic perspective, we thus have a lot of freedom to choose the threshold $\pi_{n,T}$. In finite samples, a totally different picture arises. There, different choices of $\pi_{n,T}$ may result in markedly different estimates of K_0. Selecting the threshold level $\pi_{n,T}$ in a suitable way is thus a crucial issue in finite samples.

In what follows, we give some heuristic discussion on how to pick the threshold level $\pi_{n,T}$ appropriately in practice. To do so, we concentrate on the fixed design case.

The arguments for the random design are fully analogous. For the heuristic discussion, we make the following simplifications: (i) We assume that $T_i = T$ and $f_i \equiv f$ for all i, where f is some design density. Loosely speaking, we thus suppose that the sample sizes T_i and the design densities f_i are similar enough to neglect differences between them. It of course depends on the application at hand whether this assumption is justified or not. (ii) We suppose that not only the functions m_i are the same within groups but also the error variances σ_i^2. Slightly abusing notation, we employ the symbol σ_k^2 to denote the group-specific error variance of the class G_k. Under these simplifying assumptions, we can make the following heuristic observations:

(a) Consider any pair of subjects i and j that belong to the same class G_k. According to the normality result (3.7), the bias part of $\sqrt{Th}\{\widehat{m}_{i,h}(x) - \widehat{m}_{j,h}(x)\}$ is asymptotically negligible for any $h = o(T^{-1/7})$. This motivates the approximation

$$\begin{aligned} \widehat{\psi}_{ij}(x,h) &\approx \sqrt{Th}\big(\widehat{m}_{i,h}(x) - \widehat{m}_{j,h}(x)\big) \Big/ \sqrt{2\|K\|_2^2\sigma_k^2/f(x)} \\ &\approx \widehat{\psi}_i(x,h) - \widehat{\psi}_j(x,h) \end{aligned}$$

with

$$\widehat{\psi}_i(x,h) = \Big\{\frac{1}{\sqrt{Th}}\sum_{t=1}^{T} K\Big(\frac{x_{it}-x}{h}\Big)\varepsilon_{it}\Big\} \Big/ \sqrt{2\|K\|_2^2\sigma_k^2 f(x)}.$$

For each i, we stack the random variables $\widehat{\psi}_i(x,h)$ with $(x,h) \in \mathcal{G}_T$ in the vector

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

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

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

for any pair of subjects 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 $\ell = 1,\ldots,q$, 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

$$\operatorname{Cov}\big(\widehat{\psi}_i(x,h),\widehat{\psi}_i(x',h')\big) \approx \sqrt{\frac{h}{h'}}\Big\{\int K(\varphi)K\Big(\frac{h\varphi+(x-x')}{h'}\Big)d\varphi\Big\}/2\|K\|_2^2, \quad (5.4)$$

we can approximate the random vector $\widehat{\boldsymbol{\psi}}_i$ by a Gaussian vector with the covariance structure specified on the right-hand side of (5.4). Moreover, since the vectors $\widehat{\boldsymbol{\psi}}_i$ are independent across i under our simplifying assumptions, we can approximate the distribution of

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

by that of

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

for any $S \subseteq \{1, \ldots, n\}$, where $\boldsymbol{\zeta}_i$ are independent Gaussian vectors with the covariance structure from (5.4).

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

$$\max_{1 \le k \le K_0} \widehat{\Delta}(\widehat{G}_k^{[n-K_0]}) \approx \max_{1 \le k \le K_0} \widehat{\Delta}(G_k). \tag{5.5}$$

By observation (a), we further obtain that

$$\begin{aligned} \max_{1 \le k \le K_0} \widehat{\Delta}(G_k) &= \max_{1 \le k \le K_0} \Big\{ \max_{i,j \in G_k} \widehat{d}_{ij} \Big\} \\ &\approx \max_{1 \le k \le K_0} \Big\{ \max_{i,j \in G_k} \big(|\widehat{\boldsymbol{\psi}}_i - \widehat{\boldsymbol{\psi}}_j| - \boldsymbol{C} \big)_\infty \Big\}, \end{aligned} \tag{5.6}$$

and by (b),

$$\max_{1 \le k \le K_0} \Big\{ \max_{i,j \in G_k} \big(|\widehat{\boldsymbol{\psi}}_i - \widehat{\boldsymbol{\psi}}_j| - \boldsymbol{C} \big)_\infty \Big\} \overset{d}{\approx} \max_{1 \le k \le K_0} \Big\{ \max_{i,j \in G_k} \big(|\boldsymbol{\zeta}_i - \boldsymbol{\zeta}_j| - \boldsymbol{C} \big)_\infty \Big\}, \tag{5.7}$$

where $Z \overset{d}{\approx} Z'$ means that Z is approximately distributed as Z'. Since the right-hand side of (5.7) depends on the unknown groups $G_1, \ldots, G_{K_0}$, we apply the trivial bound

$$\begin{aligned} &\max_{1 \le k \le K_0} \Big\{ \max_{i,j \in G_k} \big(|\boldsymbol{\zeta}_i - \boldsymbol{\zeta}_j| - \boldsymbol{C} \big)_\infty \Big\} \\ &\qquad \le B_n := \max_{1 \le i,j \le n} \big(|\boldsymbol{\zeta}_i - \boldsymbol{\zeta}_j| - \boldsymbol{C} \big)_\infty \end{aligned} \tag{5.8}$$

and define $q_n(\alpha)$ to be the α-quantile of B_n. Taken together, (5.5)–(5.8) suggest that

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

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

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

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

6 Theoretical results

In this section, we examine the asymptotic properties of the multiscale statistics $\widehat{d}_{ij}$. We in particular derive statements (3.17) and (3.18) under appropriate regularity conditions. These statements characterize the convergence behaviour of the statistics $\widehat{d}_{ij}$ and underlie Theorems 4.1 and 5.1 which describe the theoretical properties of our clustering methods.

To prove (3.17) and (3.18), we impose the following regularity conditions. We first summarize the assumptions for the fixed design case:

(C_{FD}1) The error processes $\mathcal{P}_i = \{\varepsilon_{it} : 1 \le t \le T_i\}$ are strictly stationary and strongly mixing for all $1 \le i \le n$.

(C_{FD}2) For each $1 \le i \le n$, there exists a density f_i such that $\int_{x_{i,t-1}}^{x_{it}} f_i(w) dw = 1/T_i$ for $1 \le t \le T_i$, where $x_{i0} = 0$.

(C_{FD}3) There exist a real number $\theta > 4$ and a positive constant C such that

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

The conditions in the random design case are essentially analogous:

(C_{RD}1) The processes $\mathcal{P}_i = \{(X_{it}, \varepsilon_{it}) : 1 \le t \le T_i\}$ are strictly stationary and strongly mixing for all $1 \le i \le n$.

(C_{RD}2) For each $1 \le i \le n$, the random variables X_{it} have a density f_i and the variables $(X_{it}, X_{it+\ell})$ have a joint density $f_{i,\ell}$. The densities f_i have bounded support, which w.l.o.g. equals $[0,1]$ for all i. The joint densities are uniformly bounded away from infinity, that is, $f_{i,\ell}(x,x') \le C < \infty$ for all i, x, x' and ℓ, where the constant C neither depends on i, x, x' nor on ℓ.

(C$_{\mathrm{RD}}$3) There exist a real number $\theta > 4$ and a natural number ℓ^* such that for any $\ell \in \mathbb{Z}$ with $|\ell| \geq \ell^*$ and some constant $C < \infty$,

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

For simplicity, the error terms ε_{it} are homoskedastic, that is, $\sigma_i^2 = \mathbb{E}[\varepsilon_{it}^2] = \mathbb{E}[\varepsilon_{it}^2 | X_{it} = x]$ for all $x \in [0,1]$.

The following conditions are needed in both the fixed and the random design case:

(C4) Let $\alpha_i(\ell)$ for $\ell = 1, 2, \ldots$ be the mixing coefficients corresponding to the i-th process $\mathcal{P}_i$. It holds that $\alpha_i(\ell) \le \alpha(\ell)$ for all i, where the coefficients $\alpha(\ell)$ decay exponentially fast to zero as $\ell \to \infty$.

(C5) The densities f_i are uniformly bounded away from zero and infinity on $[0,1]$, that is, $0 < c \le f_i(x) \le C < \infty$ for all i and $x \in [0,1]$, where the constants c and C neither depend on i nor on x. Moreover, they are twice continuously differentiable on $[0,1]$ with first and second derivatives that are uniformly bounded away from infinity in absolute value.

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

(C7) The error variances σ_i^2 are uniformly bounded away from zero and infinity, that is, $0 < c \le \sigma_i^2 \le C < \infty$ for all i, where the constants c and C do not depend on i.

(C8) It holds that $T_i = \tau_i(T) \to \infty$ as $T \to \infty$ for all i. Moreover, there exist constants c_i with $0 < \underline{c} \le c_i \le \overline{c} < \infty$ for all i such that $|T_i/T - c_i| \le \rho(T) \to 0$ as $T \to \infty$, where $\underline{c}$, $\overline{c}$ and $\rho(\cdot)$ do not depend on i. Finally, $n = n(T)$ is such that

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

for some small $\delta > 0$ and a sufficiently large constant $C > 0$, where we use the notation $a \wedge b = \min\{a, b\}$ and θ is defined in (C$_{\mathrm{FD}}$3) and (C$_{\mathrm{RD}}$3).

(C9) 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 $1/7 < b \le B < 1$. Moreover, $h_{\max}^5 / h_{\min} = o(1)$.

(C10) 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 there exists a positive constant L with $|K(x)-K(x')| \le L|x-x'|$ for all $x,x' \in \mathbb{R}$. We use the notation $\|K\|^2 = \int K^2(\varphi)d\varphi$ and $\kappa_\ell = \int K(\varphi)\varphi^\ell d\varphi$ for $\ell = 0,1,2,\ldots$

We briefly comment on the above conditions.

(a) Assumptions ($\mathrm{C_{FD}}1$) and ($\mathrm{C_{RD}}1$) restrict the dependence structure of the model variables across t by imposing mixing conditions on them. (C4) requires the mixing coefficients to decay exponentially fast. This assumption is not necessarily needed but it is nevertheless imposed to keep the proofs as clear as possible. The exponential mixing rate may alternatively be replaced by a sufficiently high polynomial rate.

(b) Assumptions ($\mathrm{C_{FD}}1$) and ($\mathrm{C_{RD}}1$) do not impose any restrictions on the dependence structure of the model variables across i. Hence, our theory allows the model variables to be dependent across i in an arbitrary way.

(c) ($\mathrm{C_{FD}}3$), ($\mathrm{C_{RD}}3$), (C5) and (C6) are standard-type moment 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.

(d) (C8) imposes restrictions on the growth of the number of subjects n. Loosely speaking, it says that n is not allowed to grow too quickly as compared to T. More specifically, let $h_{\min} = aT^{-B}$ with some $B \le 1/2$ and $h_{\max} = AT^{-b}$ with some $b > 1/7$, which implies that (C9) is satisfied. In this case, (6.1) simplifies to

$$n \le CT^{\frac{\theta}{4}-1-\frac{5}{4}\delta},$$

which essentially says that n should not grow more quickly than $T^{\theta/4-1}$. According to this, the growth restriction (6.1) on n is closely connected with the moment conditions on the error terms ε_{it} in ($\mathrm{C_{FD}}3$) and ($\mathrm{C_{RD}}3$). In particular, the larger the value of θ, that is, the stronger the moment conditions on ε_{it}, the faster n may grow as compared 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, $n \le CT^\rho$ with $\rho > 0$ as large as desired.

(e) (C9) imposes some conditions on the minimal and maximal bandwidths $h_{\min}$ and $h_{\max}$. Specifically, it requires that $h_{\min} \ge cT^{-(1-\delta)}$ and $h_{\max} \le CT^{-(1/7+\delta)}$ for some small $\delta > 0$ and positive constants c and C. These conditions are fairly moderate: Since the optimal bandwidth for estimating m_i is of the order $T^{-1/5}$ for any i under the smoothness conditions (C5) and (C6), we can choose the interval $[h_{\min}, h_{\max}]$ to be quite large, allowing for both substantial under- and oversmoothing.

(f) 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 statements on the convergence behaviour of the multiscale statistics $\widehat{d}_{ij}$.

Theorem 6.1. *Let (C_{FD}1)–(C_{FD}3) and (C_{RD}1)–(C_{RD}3) be fulfilled in the fixed and the random design case, respectively. Moreover, suppose that (C4)–(C10) are satisfied. Then it holds that*

$$\max_{1 \le k \le K_0} \max_{i,j \in G_k} \widehat{d}_{ij} = O_p\big(\sqrt{\log n + \log T} + \sqrt{T h_{\max}^7}\big) \tag{6.2}$$

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

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

The proof of Theorem 6.1 is provided in the Supplementary Material.

7 Simulations

To explore the finite sample properties of our methods, we carry out some simulations. We consider the model

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

where $T = T_i = 200$ for all i and $n = 240$. The individuals i are supposed to belong to $K_0 = 6$ different groups of the same size. In particular, we set $G_k = \{(k-1)n/6+1, \dots, kn/6\}$ for $1 \le k \le K_0 = 6$. The group-specific regression functions $g_k : [0,1] \to \mathbb{R}$ are given by

$$\begin{aligned}
g_1(x) &= G(x, \tfrac{1}{2}, \tfrac{1}{2}) \\
g_2(x) &= G(x, \tfrac{1}{4}, \tfrac{1}{4}) + G(x, \tfrac{3}{4}, \tfrac{1}{4}) \\
g_3(x) &= G(x, \tfrac{1}{8}, \tfrac{1}{8}) + G(x, \tfrac{3}{8}, \tfrac{1}{8}) + G(x, \tfrac{3}{4}, \tfrac{1}{4}) \\
g_4(x) &= G(x, \tfrac{1}{4}, \tfrac{1}{4}) + G(x, \tfrac{5}{8}, \tfrac{1}{8}) + G(x, \tfrac{7}{8}, \tfrac{1}{8}) \\
g_5(x) &= G(x, \tfrac{1}{12}, \tfrac{1}{12}) + G(x, \tfrac{1}{4}, \tfrac{1}{12}) + G(x, \tfrac{5}{12}, \tfrac{1}{12}) + G(x, \tfrac{3}{4}, \tfrac{1}{4}) \\
g_6(x) &= G(x, \tfrac{1}{4}, \tfrac{1}{4}) + G(x, \tfrac{7}{12}, \tfrac{1}{12}) + G(x, \tfrac{3}{4}, \tfrac{1}{12}) + G(x, \tfrac{11}{12}, \tfrac{1}{12}),
\end{aligned}$$

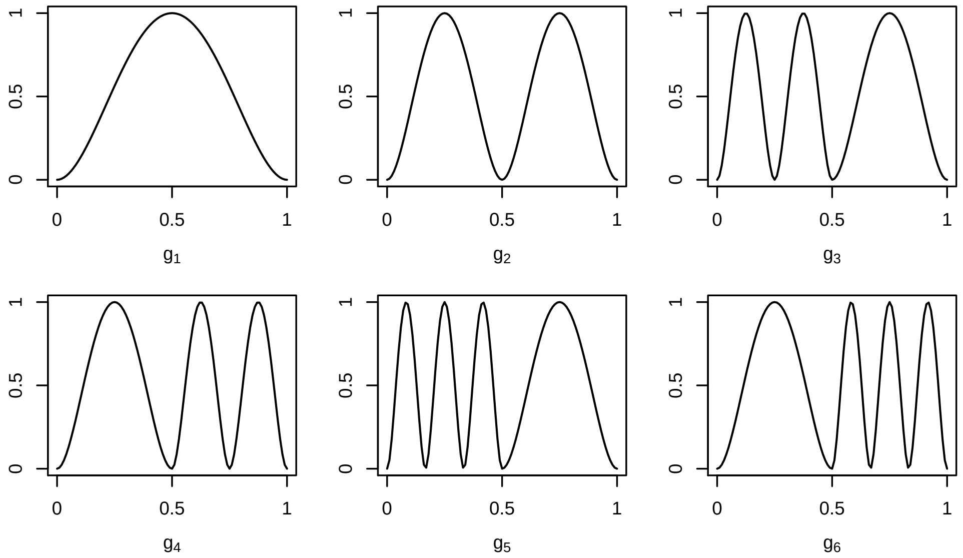

Figure 2: Plot of the functions g_k for $1 \leq k \leq 6$.

where $G(x, x_0, h) = \mathbb{1}(|(x - x_0)/h| \leq 1)\,(1 - ((x - x_0)/h)^2)^2$. Figure 2 gives a graphical illustration of the functions g_k for $1 \leq k \leq 6$. As can be seen, some of the functions are much smoother than others.[3] Moreover, the smoothness of some functions varies across the support $[0,1]$. The function g_6, for instance, is much smoother on the interval $[0, 0.5]$ than on $[0.5, 1]$. To deal with these varying degrees of smoothness, we need to inspect the functions at different resolution levels. Whereas an approach with a fixed bandwidth is barely able to do so, our multiscale approach easily accommodates functions of varying smoothness.

Both the regressors X_{it} and the error terms ε_{it} are supposed to be independent across i and t. We draw the regressors X_{it} from a uniform distribution on the unit interval $[0,1]$ and the errors ε_{it} from a normal distribution with mean 0 and different variance levels σ^2. We ignore dependence structures in the model variables X_{it} and ε_{it} across i and t because their effect is obvious: the stronger the dependence (in particular the dependence across t), the more difficult it gets to estimate the curves m_i and thus to infer the unknown group structure from the data. To assess the noise level in the simulated data, we define the noise-to-signal ratios $\mathrm{NSR}_k = \mathrm{Var}(\varepsilon_{it})/\mathrm{Var}(g_k(X_{it})) = \sigma^2/\mathrm{Var}(g_k(X_{it}))$ for $1 \leq k \leq 6$. Since NSR_k is the same for all k in our design, we simply write $\mathrm{NSR} = \mathrm{NSR}_k$ for all k. We consider three different noise-to-signal ratios $\mathrm{NSR} = 2$, 3 and 4, which correspond to the error variances $\sigma^2 \approx 0.49^2, 0.60^2$ and 0.70^2.

For each noise-to-signal ratio $\mathrm{NSR} = 2, 3$ and 4, we draw $S = 1000$ samples from model (7.1). For each sample, we compute the class estimates $\{\widehat{G}_1^{[n-K]}, \ldots, \widehat{G}_K^{[n-K]}\}$ for $K = 1, 2, \ldots$ as well as the estimate $\widehat{K}_0$ of the number of classes. To implement

[3] We here use the term "smoothness" in an informal way. By saying that a function is smoother than another one, we simply mean that it is more wiggly.

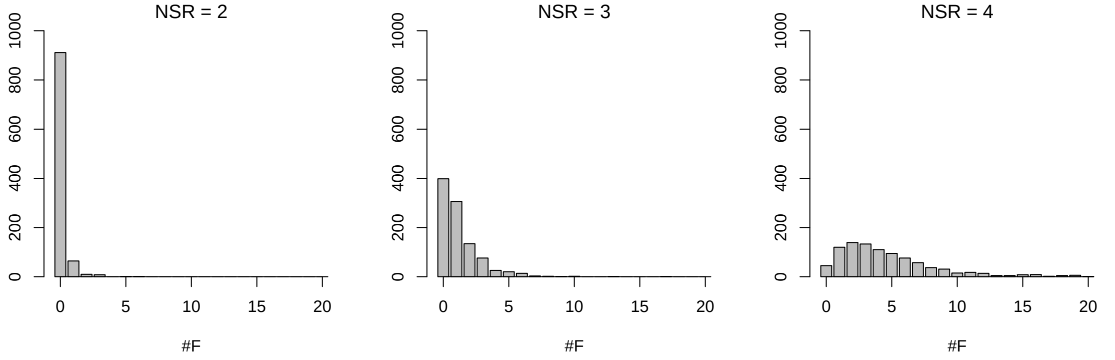

Figure 3: Histograms of the number $\#F$ of classification errors for the three simulation scenarios with the noise-to-signal ratios NSR $= 2$, 3 and 4. (To visualize the histograms, we concentrate on values of $\#F$ in the range $[0, 20]$. $\#F > 20$ in 5, 16 and 69 of the $S = 1000$ simulations for NSR $= 2$, 3 and 4, respectively.)

our clustering methods, we use the grid

$$\mathcal{G}_T = \big\{(x,h) \,\big|\, [x-h, x+h] \subseteq [0,1] \text{ with } x = r/100 \text{ for some } r = 1, \ldots, 100 \\ \text{and } h \in \{0.05, 0.1, 0.15, 0.2, 0.25\}\big\} \tag{7.2}$$

and the threshold parameter $\pi_{n,T} = q_n(\alpha)$ with $\alpha = 0.95$. The simulation study splits into two parts: In the first part, we treat the number of classes K_0 as known and investigate how well the estimated partition $\{\widehat{G}_1, \ldots, \widehat{G}_{K_0}\} = \{\widehat{G}_1^{[K_0]}, \ldots, \widehat{G}_{K_0}^{[K_0]}\}$ approximates the true class structure. In the second part, we examine how well the estimates $\widehat{K}_0$ approximate the true number of classes K_0.

The results for the first part of the study are presented in Figure 3. To measure how well the partition $\{\widehat{G}_1, \ldots, \widehat{G}_{K_0}\}$ approximates the class structure $\{G_1, \ldots, G_{K_0}\}$, we compute the number $\#F$ of wrongly classified indices i for each simulated sample.[4] The three panels of Figure 3 show histograms of the $S = 1000$ values of $\#F$ that are obtained for the three noise-to-signal ratios under consideration. For the lowest noise-to-signal ratio NSR $= 2$, our algorithm produces very accurate results: in almost all of the $S = 1000$ simulations, the number of classification errors $\#F$ is at most 3, and in more than 90% of the cases, $\#F$ is equal to 0. For the ratio level NSR $= 3$, the results are less precise but still quite accurate: $\#F \le 5$ in about 96% of the simulations and $\#F = 0$ in around 40% of the cases. When NSR $= 4$, the noise level in the data is very high, the error variance σ^2 being 4 times larger than the variance $\text{Var}(g_k(X_{it}))$ of the signal. In this case, the estimation results are much less precise than in the

[4] $\#F$ 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 subject 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 subject i, where the estimated classes are labelled according to the permutation π. More specifically, set $\widehat{\rho}_\pi(i) = \pi(k)$ if $i \in \widehat{G}_k$. With this notation at hand, we define $\#F = \min_{\pi \in \Pi} \sum_{i=1}^n \mathbb{1}(\rho(i) \neq \widehat{\rho}_\pi(i))$.

	NSR = 2	NSR = 3	NSR = 4
$\widehat{K}_0 = 5$	0	0	13
$\widehat{K}_0 = 6$	873	854	825
$\widehat{K}_0 = 7$	119	136	156
$\widehat{K}_0 = 8$	8	10	6

Table 1: Simulation results for the estimator $\widehat{K}_0$ of $K_0 = 6$. The three columns of the table correspond to the three simulation scenarios with NSR = 2, 3 and 4. The entries in each column give the number of simulations (out of a total of $S = 1000$) in which $\widehat{K}_0$ takes a certain value.

previous two cases. Nevertheless, we still get that $\#F \le 10$ in around 85% of the simulations and $\#F \le 20$ in around 93% of them, meaning that we misclassify at most 10 out of 240 subjects in 85% of the cases and not more than 20 subjects in 93% of them. Hence, even though the noise level in the data is very high, the partition $\{\widehat{G}_1, \dots, \widehat{G}_{K_0}\}$ gives a reasonable approximation to the true class structure in a large part of the simulations. All in all, the simulations suggest that our clustering algorithm produces appropriate estimates of the true class structure as long as the noise level in the data is not extremely high.

The simulation results for the second part of the study are summarized in Table 1. The entries of the table specify the number of simulations in which a certain value of $\widehat{K}_0$ is obtained. The results suggest that the estimator $\widehat{K}_0$ performs well in all three simulation scenarios with the noise-to-signal ratios NSR = 2, 3 and 4. As one can see, the estimation precision deteriorates as the noise-to-signal ratio increases. Nevertheless, even for the highest ratio level NSR = 4, the results are still fairly accurate: the estimated number of clusters $\widehat{K}_0$ is equal to the true number $K_0 = 6$ in around 82% of the cases. Moreover, $\widehat{K}_0$ lies between 5 and 8 in all of the simulations, thus being reasonably close to $K_0 = 6$ in all of the cases.

8 Application

Global warming is a very pressing issue which has received a lot of attention in both the scientific and the public debate over the last few decades. There is a huge number of climatological studies which aim to shed light on both the global warming trend and the regional variations thereof. Global temperature records are analyzed in Bloomfield (1992) and Hansen et al. (2010) among many others. Regional variations of the warming trend are investigated for example in Karoly and Wu (2005), Stott et al. (2010) and Knutson et al. (2013).

In this section, we use our clustering methods to estimate the regional patterns of global warming from a large data set on land surface temperature anomalies that was collected by the investigators of the Berkeley Earth project. The data are publicly available at `http://berkeleyearth.org/data`. A detailed description of them can

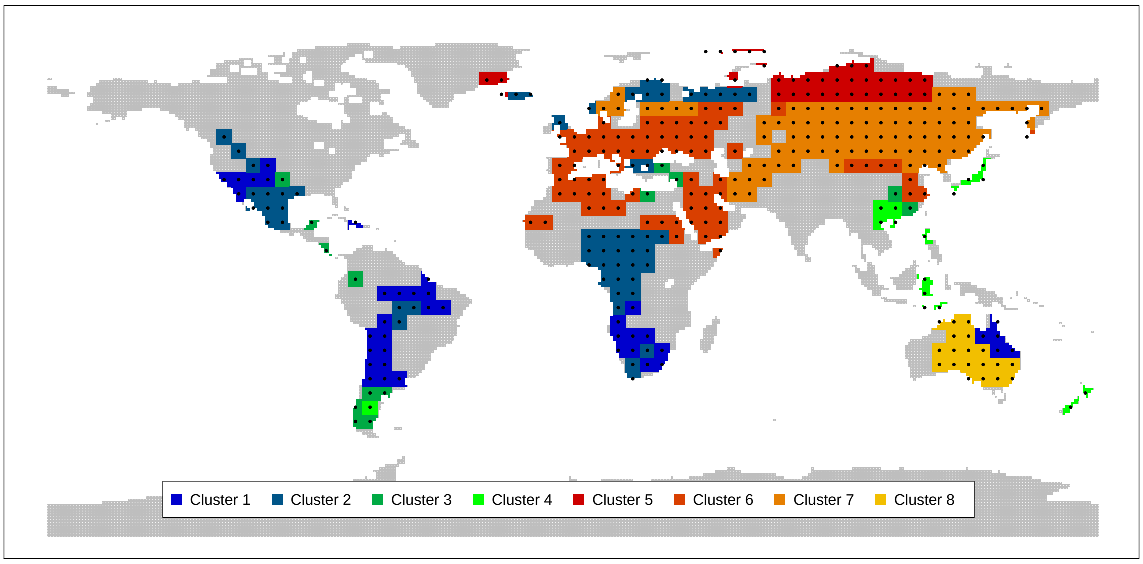

Figure 4: Estimated temperature anomaly clusters presented in a worldmap. The black dots are the exact locations i where we observe data. The colour around each location i indicates which cluster the location belongs to.

be found in Rohde et al. (2013). From the data set, we compute yearly temperature anomalies (measured in degree Celsius) for a wide range of spatial locations. The temperature anomaly at location i in year t is defined as the departure of the average temperature at location i in year t from a certain reference value, in particular from the average temperature at location i over the period from 1951 to 1980. For our analysis, we consider all spatial locations on land which lie on a 5 degree (longitude) by 5 degree (latitude) grid. We take into account all grid points where data are available for the time span from 1880 to 2016. This leaves us with a total of $n = 347$ spatial grid points. For each grid point, we observe a time series of length $T = 137$ which consists of the yearly temperature anomalies from 1880 to 2016.

Throughout the section, we use the symbol Y_{it} to denote the temperature anomaly at location i and time point t. The time series $\{Y_{it} : 1 \leq t \leq T\}$ at location i is supposed to follow the model

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

for $1 \leq t \leq T$, where $x_{it} = t/T$ are the design points and $\mathbb{E}[\varepsilon_{it}] = 0$. We thus model the temperature anomalies Y_{it} at location i as a nonparametric trend $m_i(t/T)$ corrupted by noise ε_{it}. As in the theoretical part of the paper, we assume that the locations i in our sample can be partitioned into K_0 groups $G_1, \ldots, G_{K_0}$ such that for each $1 \leq k \leq K_0$,

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

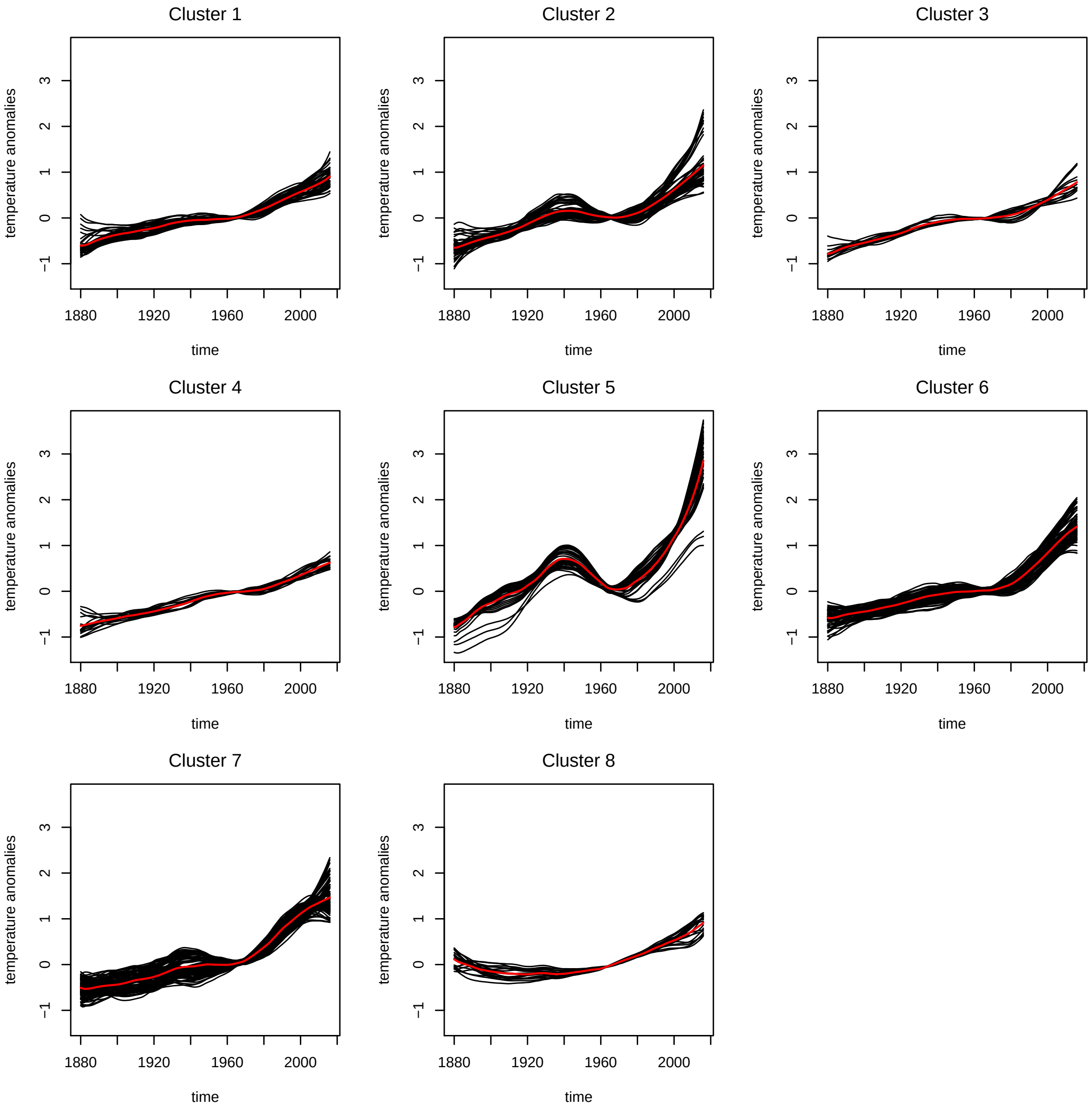

Figure 5: Estimated temperature anomaly clusters. Each panel corresponds to one cluster. The black lines are the estimated regression curves $\widehat{m}_{i,h}$ that belong to the respective cluster. The red lines are estimates of the group-specific regression functions.

We thus suppose that there is a certain number of spatial regions G_k where the temperature anomalies evolve in the same way over time (or at least very similarly). The aim of our analysis is to estimate the unknown regions $G_1, \ldots, G_{K_0}$ along with their unknown number K_0. Put differently, we want to estimate the regional patterns of global warming from the data.

To implement our clustering methods, we employ the grid $\mathcal{G}_T$ defined in (7.2) and the threshold parameter $\pi_{n,T} = q_n(\alpha)$ with $\alpha = 0.95$. Our estimation results are presented in Figures 4 and 5. The estimated number of regions is $\widehat{K}_0 = 8$. Figure 4 visualizes the estimated regions in a worldmap. Figure 5 presents them in an alternative way: Each panel of the figure depicts the estimated curves $\widehat{m}_{i,h}$ that belong to one of the $\widehat{K}_0 = 8$ estimated clusters. The red curve in each panel is an estimate $\widehat{g}_{k,h}$ of the group-specific regression function g_k. In particular, we define

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

that is, we simply average the fits $\widehat{m}_{i,h}$ with $i \in \widehat{G}_k^{[\widehat{K}_0]}$. To compute the local linear smoothers $\widehat{m}_{i,h}$, we of course need to select a bandwidth value h. As the smoothers are only computed for illustrative purposes, we use the same bandwidth h for all locations i. In particular, we choose the bandwidth adhoc as $h = 0.15$ for all i, which produces a good visual impression of the results.

Inspecting Figures 4 and 5, our clustering methods appear to give a reasonable picture of the regional patterns present in the temperature anomaly data. They produce clusters which mainly correspond to connected geographical regions. Cluster 5, for example, corresponds to Northern Russia (and a small part of Greenland for which data are available), whereas Cluster 8 covers most of Australia. As can be seen from Figure 5, the clusters are fairly homogeneous, consisting of curves with similar shapes. Moreover, the curves of some clusters have very different shapes from those of other clusters. This suggests that the pattern of climate change may be quite different across geographical regions. The curves of Cluster 5, for instance, which mainly represent locations in Northern Russia have a highly nonlinear shape. Moreover, they exhibit a strong increase from the 1970s onwards. This sharply contrasts with the curves in Cluster 8, for example, which cover most of Australia. These curves are fairly flat over the whole time range from 1880 to 2016. To summarize, our clustering methods appear to be a useful data mining tool. The produced clusters allow to get a first overview of the regional patterns of the anomaly data and may thus serve as a starting point for a more thorough data analysis.

9 Extensions and modifications

Before closing the paper, we discuss some possible extensions and modifications of our methods.

Extension 1. Throughout the paper, we have assumed that the number of classes K_0 is fixed. We now allow K_0 to grow with the number of subjects n, that is, we admit of $K_0 = K_{0,n} \to \infty$ as $n \to \infty$. To deal with this situation, we require the group-specific regression functions g_k to fulfill the following additional condition:

(C11) The functions g_k as well as their first and second derivatives are uniformly bounded in absolute value, that is, $|g_k^{(\ell)}(x)| \le C$ for all $x \in [0,1]$ and $\ell = 0, 1, 2,$

where $g_k^{(\ell)}$ denotes the ℓ-th derivative of g_k and the constant $C < \infty$ does not depend on k. Moreover,

$$\min_{1 \le k < k' \le K_0} \max_{\{x \,:\, (x, h_{\max}) \in \mathcal{G}_T\}} |g_k(x) - g_{k'}(x)| \gg \frac{\sqrt{\log n + \log T} + \sqrt{T h_{\max}^5}}{\sqrt{T h_{\max}}}. \qquad (9.1)$$

As before, the expression $a_{n,T} \gg b_{n,T}$ means that $b_{n,T} = o(a_{n,T})$ and the notation $a_{n,T} \ll b_{n,T}$ is used analogously. (9.1) essentially says that the regression functions g_k and $g_{k'}$ of two different classes do not approach each other too quickly. If condition (C11) is fulfilled, a slightly modified version of Theorem 6.1 can be proven. In particular, with the help of the technical arguments from the Supplementary Material, it is not difficult to show that

$$\max_{1 \le k \le K_0} \max_{i,j \in G_k} \widehat{d}_{ij} = O_p\big(\sqrt{\log n + \log T} + \sqrt{T h_{\max}^7}\big)$$

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

These two statements immediately imply that Theorem 4.1 remains to hold true. Moreover, Theorem 5.1 remains valid as well if the threshold level $\pi_{n,T}$ satisfies a strengthened version of condition (5.1), namely the condition that $\sqrt{\log n + \log T} + \sqrt{T h_{\max}^7} \ll \pi_{n,T} \ll \sqrt{T h_{\max}} \min_{1 \le k < k' \le K_0} \max_{\{x \,:\, (x, h_{\max}) \in \mathcal{G}_T\}} |g_k(x) - g_{k'}(x)|$.

Extension 2. By assumption, the grid $\mathcal{G}_T$ of location-bandwidth points (x, h) is a subset of $\mathcal{G} = \{(x, h) \,|\, h_{\min} \le h \le h_{\max} \text{ and } h \le x \le 1 - h\}$. Hence, the multiscale statistics $\widehat{d}_{ij}$ take into account locations $x \in [h, 1-h]$ but ignore points x in the boundary region $[0, h) \cup (1-h, 1]$. To include boundary points x, we need to modify the multiscale statistics $\widehat{d}_{ij}$. Specifically, we need to adjust the normalization term $\widehat{\nu}_{ij,h}(x)$ in two ways: (i) The normalization $\widehat{\nu}_{ij,h}(x)$ is an estimator of the asymptotic variance $\nu_{ij}(x)$ from the normality results (3.4) and (3.10). As the variance $\nu_{ij}(x)$ has a slightly different form for boundary points x, the normalization $\widehat{\nu}_{ij,h}(x)$ needs to be changed accordingly at the boundary. (ii) The kernel densities $\widehat{f}_{i,h}(x)$ and $\widehat{f}_{j,h}(x)$ in the definition of $\widehat{\nu}_{ij,h}(x)$ need to be replaced by boundary-corrected versions. Importantly, the other kernel estimators involved in the definition of $\widehat{d}_{ij}$ do not suffer from boundary effects and can thus remain unchanged. Modifying the multiscale statistics $\widehat{d}_{ij}$ according to (i) and (ii), we can enlarge the grid $\mathcal{G}_T$ to contain points (x, h) with x in the boundary region $[0, h) \cup (1-h, 1]$. The theoretical results of the paper remain unaffected by these changes, the proofs in the Supplementary Material being easily adjusted.

Extension 3. In order to estimate the unknown class structure in model (1.1)–(1.2), we have combined the multiscale statistics $\widehat{d}_{ij}$ with a hierarchical clustering algorithm.

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

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

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

References

Abraham, C., Cornillon, P. A., Matzner-Løber, E. and Molinari, N. (2003). Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics*, **30** 581–595.

Bloomfield, P. (1992). Trends in global temperature. *Climatic Change*, **21** 1–16.

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

Chiou, J.-M. and Li, P.-L. (2007). Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society: Series B*, **69** 679–699.

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

Hannig, J. and Marron, J. S. (2006). Advanced distribution theory for SiZer. *Journal of the American Statistical Association*, **101** 484–499.

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

Hansen, J., Ruedy, R., Sato, M. and Lo, K. (2010). Global surface temperature change. *Reviews of Geophysics*, **48**.

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

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

James, M. and Sugar, C. A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, **98** 397–408.

Karoly, D. J. and Wu, Q. (2005). Detection of regional surface temperature trends. *Journal of Climate*, **18** 4337–4343.

Knutson, T. R., Zeng, F. and Wittenberg, A. T. (2013). Multimodel assessment of regional surface temperature trends: CMIP3 and CMIP5 twentieth-century simulations. *Journal of Climate*, **26** 8709–8743.

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

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

Rohde, R., Muller, R., Jacobsen, R., Perlmutter, S., Rosenfeld, A., Wurtele, J., Curry, J., Wickham, C. and Mosher, S. (2013). Berkeley Earth Temperature averaging process. *Geoinformatics & Geostatistics: An Overview*, **1:2**.

Sacks, J. and Ylvisaker, D. (1970). Designs for regression problems with correlated errors. III. *Annals of Mathematical Statistics*, **41** 2057–2074.

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

Stott, P. A., Gillett, N. P., Hegerl, G. C., Karoly, D. J., Stone, D. A., Zhang, X. and Zwiers, F. (2010). Detection and attribution of climate change: a regional perspective. *Wiley Interdisciplinary Reviews: Climate Change*, **1** 192–211.

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

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

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

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