ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2017)
79, *Part* 1, *pp.* 5–27

Classification of non-parametric regression functions in longitudinal data models

Michael Vogt

University of Bonn, Germany

and Oliver Linton

University of Cambridge, UK

[Received February 2015. Revised October 2015]

Summary. We investigate a longitudinal data model with non-parametric regression functions that may vary across the observed individuals. In a variety of applications, it is natural to impose a group structure on the regression curves. Specifically, we may suppose that the observed individuals can be grouped into a number of classes whose members all share the same regression function. We develop a statistical procedure to estimate the unknown group structure from the data. Moreover, we derive the asymptotic properties of the procedure and investigate its finite sample performance by means of a simulation study and a real data example.

Keywords: Classification of regression curves; Kernel estimation; Longitudinal or panel data; Non-parametric regression

1. Introduction

Non-parametric and semiparametric regression models are a flexible framework to analyse longitudinal data from various fields such as economics, finance, biology and climatology. Most of the literature is based on the assumption that the regression function is the same across the observed individuals; see Ruckstuhl *et al.* (2000), Henderson *et al.* (2008) and Mammen *et al.* (2009) among many others. This assumption, however, is very unrealistic in many applications. In particular, when the number of observed individuals is large, it is quite unlikely that all individuals have the same regression function. In a wide range of cases, it is much more plausible to suppose that there are groups of individuals who share the same regression function (or at least have very similar regression curves). As a modelling approach, we may thus assume that the observed individuals can be grouped into a number of classes whose members all share the same regression function. The aim of this paper is to develop a statistical procedure to infer the unknown group structure from the data.

Throughout the paper, we work with the following model set-up. We observe a sample of longitudinal or panel data $\{(Y_{it}, X_{it}) : 1 \leqslant i \leqslant n, 1 \leqslant t \leqslant T\}$, where i denotes the ith individual and t is the time point of observation. The time series dimension T is assumed to be large or, more precisely, to tend to ∞. The cross-section dimension n, in contrast, may either be fixed or diverging. The data are supposed to come from the non-parametric regression model

Address for correspondence: Michael Vogt, Department of Economics and Hausdorff Center for Mathematics, University of Bonn, 53113 Bonn, Germany.
E-mail: michael.vogt@uni-bonn.de

 1369–7412/17/79005

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

where m_i are unknown non-parametric functions which may differ across individuals i and u_{it} denotes the error term. We impose the following group structure on the model: let $G_1, \ldots, G_K$ be a fixed number of disjoint sets which partition the index set $\{1, \ldots, n\}$, i.e. $G_1 \dot\cup \ldots \dot\cup G_K = \{1, \ldots, n\}$. We suppose that, for each $k \in \{1, \ldots, K\}$,

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

Hence, the members of the class G_k all have the same regression function, which we denote by g_k. The classes $G_k = G_{k,n}$ depend on the cross-section dimension n in general. To keep the exposition simple, we, however, suppress this dependence in the notation throughout the paper. Our aim is to estimate the groups $G_1, \ldots, G_K$, their number K and the group-specific regression functions $g_1, \ldots, g_K$ in model (1.1)–(1.2).

The error terms u_{it} in model (1.1) are supposed to have the form $u_{it} = \alpha_i + \gamma_t + \varepsilon_{it}$, which is similar to the structure of a two-way classification model as discussed in Rao (1997). The components ε_{it} are standard regression errors that satisfy $\mathbb{E}[\varepsilon_{it}|X_{it}] = 0$. The terms α_i are individual-specific errors: they control for individual-specific characteristics like intelligence or genetic make-up that are unobserved and stable over time. In a similar vein, the terms γ_t capture unobserved time-specific effects like calendar effects or trends that are common across individuals. In many applications, the regressors may be correlated with unobserved individual- or time-specific characteristics. To take this into account, we allow the errors α_i and γ_t to be correlated with the regressors in an arbitrary way. Specifically, defining $\mathcal{X}_{n,T} = \{X_{it} : 1 \leqslant i \leqslant n, 1 \leqslant t \leqslant T\}$, we allow that $\mathbb{E}[\alpha_i|\mathcal{X}_{n,T}] \neq 0$ and $\mathbb{E}[\gamma_t|\mathcal{X}_{n,T}] \neq 0$. Moreover, whereas the errors ε_{it} are assumed to be independent across i later on, the terms α_i may be correlated across i. 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}. The way that we model the errors u_{it} is motivated by the econometrics literature; see for example Hsiao (2003) and Baltagi (2013). Following the terminology from there, we call α_i and γ_t fixed effects. To identify the functions m_i in the presence of the fixed effects α_i and γ_t, we normalize them to satisfy $\mathbb{E}[m_i(X_{it})] = 0$ for all i and t. This normalization amounts to a harmless rescaling under our technical conditions in Section 3.

The group structure that is imposed in model (1.1)–(1.2) is an attractive working hypothesis in a wide number of applications. In Section 6, we illustrate this by an example from finance. Up to 2007, primary European stock exchanges such as the London Stock Exchange were essentially the only venues where stocks could be traded in Europe. This monopoly was ended by the so-called 'Markets in financial instruments directive' in 2007. Since then, various new trading platforms have emerged. Nowadays, the European equity market is strongly fragmented with stocks being traded simultaneously at a variety of venues. This restructuring of the European stock market has raised the question how competition between trading venues, i.e. trading venue fragmentation, affects the quality of the market. Obviously, the effect of fragmentation on market quality can be expected to differ across stocks. Moreover, it is plausible to suppose that there are different groups of stocks for which the effect is the same (or at least quite similar). Our modelling approach thus appears to be a suitable framework to investigate empirically the effect of fragmentation on market quality. In Section 6, we apply it to a sample of data for the Financial Times Stock Exchange (FTSE) 100 and FTSE 250 stocks.

To the best of our knowledge, the problem of classifying non-parametric regression functions in the longitudinal data framework (1.1) has not been considered so far in the literature. Recently, however, there have been some studies on a parametric version of this problem: consider the linear panel regression model $Y_{it} = \beta_i X_{it} + u_{it}$, where the coefficients β_i are allowed to vary across

individuals. As in our non-parametric model, we may suppose that the coefficients β_i can be grouped into a number of classes. Specifically, we may assume that there are classes $G_1, \ldots, G_K$ such that $\beta_i = \beta_j$ for all $i, j \in G_k$ and all $1 \leqslant k \leqslant K$. The problem of estimating the unknown classes in this parametric framework has been considered in Su *et al*. (2014) among others.

Our modelling approach is related to classification problems in functional data analysis. There, the observed data $X_1, \ldots, X_n$ are curves or, more specifically, sample paths of a stochastic process $X = \{X(t) : t \in \mathcal{T}\}$, where $\mathcal{T}$ is some index set and most commonly represents an interval of time. In some cases, the curves $X_1, \ldots, X_n$ are observed without noise; in others, they are observed with noise. In both the noiseless and the noisy case, the aim is to cluster the curves $X_1, \ldots, X_n$ into a number of groups. There is a vast amount of references which deal with this problem in different model set-ups; see for example Abraham *et al*. (2003) and Tarpey and Kinateder (2003) for procedures based on k-means clustering, James and Sugar (2003) and Chiou and Li (2007) for so-called model-based clustering approaches, Ray and Mallick (2006) for a Bayesian approach and Jacques and Preda (2014) for a recent survey.

Even though there is a natural link between our estimation problem and the issue of classifying curves in functional data analysis, these two problems substantially differ from each other. In functional data analysis, the objects to be clustered are realizations of random curves that depend on a deterministic index $t \in \mathcal{T}$. In our longitudinal model, in contrast, we aim to cluster deterministic curves that depend on random regressors. Hence, the objects to be clustered are of a very different nature. Moreover, the error structure in our model is much more involved than in functional data analysis, where the noise is most commonly independently and identically distributed across observations (if there is noise at all). Finally, whereas the number of observed curves n should diverge to ∞ in functional data models, we provide theory for both fixed and diverging n. For these reasons, substantially different theoretical arguments are required to analyse clustering algorithms in our framework and in functional data analysis.

Our estimation methods are introduced in Section 2. There, we develop a thresholding algorithm to estimate the classes $G_1, \ldots, G_K$. The algorithm has the very nice feature that it simultaneously estimates the classes along with their number K. Hence, we do not need a separate procedure to estimate K. This distinguishes our procedure from most other classification algorithms such as k-means clustering which presuppose knowledge of the true number of classes. Once we have constructed our estimators of the classes $G_1, \ldots, G_K$, we use these to come up with kernel-type estimators of the associated regression functions $g_1, \ldots, g_K$.

The asymptotic properties of our methods are investigated in Section 3. There, we show that our estimators of the classes $G_1, \ldots, G_K$ and of their number K are consistent. Moreover, we derive the limit distribution of the estimators of the group-specific regression functions $g_1, \ldots, g_K$. In Section 4, we discuss how to implement our methods in practice. Most importantly, our algorithm to estimate the classes $G_1, \ldots, G_K$ depends on a threshold parameter which needs to be tuned appropriately. We provide a detailed discussion on how to achieve this. We finally complement the theoretical analysis of the paper by a simulation study in Section 5 and by our empirical investigation of the effect of fragmentation on market quality in Section 6. The computer code to calculate our estimators in the simulation set-up of Section 5 may be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Estimation

In this section, we describe how to estimate the groups $G_1, \ldots, G_K$, their number K and the group-specific regression functions $g_1, \ldots, g_K$ in model (1.1)–(1.2). For simplicity of exposition,

we restrict attention to real-valued regressors X_{it}; the theory carries over to the multivariate case in a completely straightforward way. To set up our estimation method, we proceed in several steps: in a first step, we construct kernel-type smoothers of the individual functions m_i. With the help of these smoothers, we then set up estimators of the classes $G_1, \dots, G_K$ and of their number K. These are finally used to come up with estimators of the functions $g_1, \dots, g_K$.

2.1. Estimation of the regression functions m_i

To construct an estimator $\hat{m}_i$ of the regression function m_i of the ith individual, we proceed as follows: let $Y_{it}^{\text{fe}} = Y_{it} - \alpha_i - \gamma_t$ be the Y-observations purged of the individual and time fixed effects. If the fixed effects were observed, we could directly work with the model equation $Y_{it}^{\text{fe}} = m_i(X_{it}) + \varepsilon_{it}$, from which the function m_i can be estimated by standard non-parametric methods. In particular, we could employ a Nadaraya–Watson smoother of the form

$$\hat{m}_i^*(x) = \frac{\sum_{t=1}^{T} W_h(X_{it} - x) Y_{it}^{\text{fe}}}{\sum_{t=1}^{T} W_h(X_{it} - x)},$$

where h is the bandwidth and W denotes a kernel function with $W_h(x) = h^{-1} W(x/h)$. To obtain a feasible estimator of m_i, we replace the unobserved variables Y_{it}^{fe} in the above formula by the approximations $\hat{Y}_{it}^{\text{fe}} = Y_{it} - \bar{Y}_i - \bar{Y}_t^{(i)} + \bar{\bar{Y}}^{(i)}$, where $\bar{Y}_i = T^{-1} \Sigma_{t=1}^{T} Y_{it}$, $\bar{Y}_t^{(i)} = (n-1)^{-1} \Sigma_{j=1, j \neq i}^{n} Y_{jt}$ and $\bar{\bar{Y}}^{(i)} = \{(n-1)T\}^{-1} \Sigma_{j=1, j \neq i}^{n} \Sigma_{t=1}^{T} Y_{jt}$. In the definition of $\bar{Y}_t^{(i)}$ and $\bar{\bar{Y}}^{(i)}$, we leave out the data of the ith individual to avoid some bias terms that are particularly problematic when n is fixed. With this notation at hand, we define the feasible estimator

$$\hat{m}_i(x) = \frac{\sum_{t=1}^{T} W_h(X_{it} - x) \hat{Y}_{it}^{\text{fe}}}{\sum_{t=1}^{T} W_h(X_{it} - x)}$$

of the function m_i. For simplicity, we use the same bandwidth h for all estimators $\hat{m}_i$. It is, however, no problem at all to let the estimators depend on different bandwidths h_i. In particular, our theoretical results in Section 3 go through essentially unchanged for varying bandwidths h_i (as long as these fulfil the conditions on the common bandwidth h that are summarized in condition 4 of Section 3.1). It is also important to note that our theory is not restricted to Nadaraya–Watson estimators. Alternatively, we could work with local linear or more generally local polynomial estimators: our methods to estimate the groups G_k and the functions g_k for $1 \leqslant k \leqslant K$ remain the same no matter which type of kernel smoother we employ. Moreover, our theoretical results of Section 3 can be verified for local polynomial estimators of an arbitrary order at the cost of a more involved notation. Specifically, theorems 1 and 2 can be shown to hold true unchanged; only the limit distribution of theorem 3 must be adjusted.

2.2. A thresholding procedure to estimate the groups G_k

We first consider the following estimation problem: let $S \subseteq \{1, \dots, n\}$ be some index set and pick an index $i \in S$. Moreover, let $G \in \{G_1, \dots, G_K\}$ be the class to which i belongs and suppose that $G \subseteq S$. We would like to infer which indices in S belong to the group G.

To tackle this estimation problem, we measure the distances between pairs of functions m_i and m_j. Specifically, we work with squared L_2-distances of the form $\Delta_{ij} = \int \{m_i(x) - m_j(x)\}^2 \pi(x) \mathrm{d}x$,

where π is some weight function. These are estimated by $\hat{\Delta}_{ij} = \int \{\hat{m}_i(x) - \hat{m}_j(x)\}^2 \pi(x)\,\mathrm{d}x$, where $\hat{m}_i$ and $\hat{m}_j$ are the kernel smoothers that were introduced in Section 2.1. We now sort the distances $\{\Delta_{ij} : j \in S\}$ along with their estimates $\{\hat{\Delta}_{ij} : j \in S\}$ in increasing order. Denote the ordered distances by

$$\Delta_{i(1)} \leqslant \Delta_{i(2)} \leqslant \ldots \leqslant \Delta_{i(n_S)},$$

$$\hat{\Delta}_{i[1]} \leqslant \hat{\Delta}_{i[2]} \leqslant \ldots \leqslant \hat{\Delta}_{i[n_S]},$$

where $n_S = |S|$ is the cardinality of S and the symbols $(\cdot)$ and $[\cdot]$ are used to distinguish between the orderings of the true and the estimated distances. The ordered distances $\Delta_{i(j)}$ have the following property: there is a point $p = p_{i,S}$ such that

$$\Delta_{i(j)} \begin{cases} = 0 & \text{for } j \leqslant p, \\ \geqslant c & \text{for } j > p \end{cases}$$

with some constant $c > 0$. From this, it immediately follows that $G = \{(1), \ldots, (p)\}$. The ordered estimates $\hat{\Delta}_{i[j]}$ exhibit a similar pattern: since $\max_{1 \leqslant i,j \leqslant n} |\hat{\Delta}_{ij} - \Delta_{ij}| = o_p(1)$ under appropriate regularity conditions, it holds that

$$\hat{\Delta}_{i[j]} \begin{cases} = o_p(1) & \text{for } j \leqslant p, \\ \geqslant c + o_p(1) & \text{for } j > p. \end{cases} \tag{2.1}$$

This in particular says that the first p order statistics $\hat{\Delta}_{i[1]}, \ldots, \hat{\Delta}_{i[p]}$ approximate the distances $\Delta_{i(1)}, \ldots, \Delta_{i(p)}$, which in turn implies that the two sets of indices $\{[1], \ldots, [p]\}$ and $\{(1), \ldots, (p)\}$ should coincide with probability tending to 1. Hence, if we knew the size $p = |G|$ of the class G, we could simply estimate $G = \{(1), \ldots, (p)\}$ by $\tilde{G} = \{[1], \ldots, [p]\}$.

As p is not observed in practice, we must estimate it. This can be achieved by a thresholding approach: let $\{\tau_{n,T}\}$ be a null sequence of threshold levels that converge to 0 sufficiently slowly. In particular, suppose that

$$\max_{j \in G} \hat{\Delta}_{ij} \leqslant \tau_{n,T} \qquad \text{with probability approaching 1,} \tag{2.2}$$

which says that the threshold parameter $\tau_{n,T}$ is not allowed to converge to 0 faster than $\max_{j \in G} \hat{\Delta}_{ij}$. By the above considerations, $\max_{j \in G} \hat{\Delta}_{ij} = \hat{\Delta}_{i[p]}$ with probability tending to 1. Hence, statement (2.1) immediately yields that

$$\hat{\Delta}_{i[j]} \begin{cases} \leqslant \tau_{n,T} & \text{for } j \leqslant p, \\ > \tau_{n,T} & \text{for } j > p \end{cases}$$

with probability approaching 1. This suggests that we estimate $p = p_{i,S}$ by

$$\hat{p} = \hat{p}_{i,S} = \max\{j \in \{1, \ldots, n_S\} : \hat{\Delta}_{i[j]} \leqslant \tau_{n,T}\} \tag{2.3}$$

and define our estimator of G as $\hat{G} = \{[1], \ldots, [\hat{p}]\}$. Fig. 1 provides a graphical illustration of this estimation approach.

We now set up an algorithm which iteratively applies the thresholding procedure from above to estimate the class structure $\{G_k : 1 \leqslant k \leqslant K\}$.

Step 1: set $S_1 = \{1, \ldots, n\}$, pick some index $i_1 \in S_1$ and denote the ordered estimated distances by $\hat{\Delta}_{i_1[1]} \leqslant \ldots \leqslant \hat{\Delta}_{i_1[n_{S_1}]}$. Compute $\hat{p} = \hat{p}_{i_1,S_1}$ as defined in equation (2.3) and estimate the group to which i_1 belongs by $\hat{G}_1 = \{[1], \ldots, [\hat{p}]\}$.

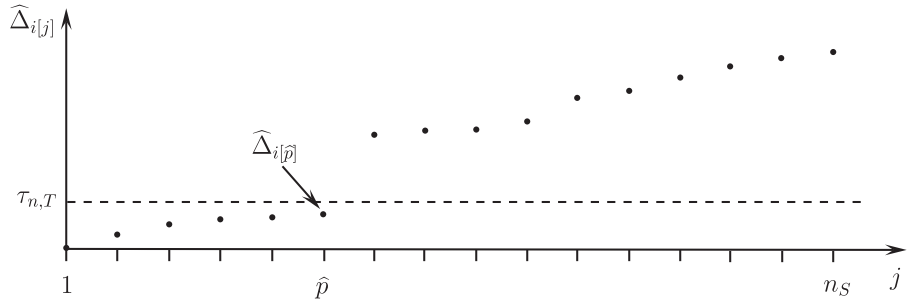

Fig. 1. Graphical illustration of the procedure underlying the estimator $\hat{G} = \{[1], \ldots, [\hat{p}]\}$: •, ordered estimated distances $\hat{\Delta}_{i[1]}, \ldots, \hat{\Delta}_{i[n_S]}$; – – –, threshold level $\tau_{n,T}$

Step k: let $\hat{G}_1, \ldots, \hat{G}_{k-1}$ be the class estimates from the previous iteration steps. Set $S_k = \{1, \ldots, n\} \setminus \cup_{l=1}^{k-1} \hat{G}_l$, pick some index $i_k \in S_k$ and denote the ordered estimated distances by $\hat{\Delta}_{i_k[1]} \leqslant \ldots \leqslant \hat{\Delta}_{i_k[n_{S_k}]}$. Compute $\hat{p} = \hat{p}_{i_k, S_k}$ as defined in equation (2.3) and estimate the group to which i_k belongs by $\hat{G}_k = \{[1], \ldots, [\hat{p}]\}$.

We iterate this algorithm $\hat{K}$ times until $\hat{\Delta}_{i_{\hat{K}}[j]} \leqslant \tau_{n,T}$ for all $1 \leqslant j \leqslant n_{S_{\hat{K}}}$, i.e. until our thresholding rule suggests that all indices in $S_{\hat{K}}$ belong to the same class. In this case, $S_{\hat{K}}$ is not split into two parts any more and $\hat{G}_{\hat{K}} = S_{\hat{K}}$. Our algorithm thus produces the partition $\{\hat{G}_k : 1 \leqslant k \leqslant \hat{K}\}$, which serves as our estimator of the class structure $\{G_k : 1 \leqslant k \leqslant K\}$. Importantly, the algorithm not only estimates the classes $G_1, \ldots, G_K$ but also their number K. In particular, K is implicitly estimated by the number of iterations $\hat{K}$. This is a very nice feature of the method, distinguishing it from most other classification algorithms which commonly presuppose knowledge of the true number of classes.

In Section 4, we discuss how to implement the estimators $\hat{G}_1, \ldots, \hat{G}_{\hat{K}}$ in practice. In particular, we explain how to choose the threshold parameter $\tau_{n,T}$ in an appropriate way. Besides the threshold $\tau_{n,T}$, we also need to pick an index $i_k \in S_k$ in each iteration step of the procedure. In principle, there is no restriction on how to do so. In particular, our theoretical results in Section 3 hold true no matter which indices i_k we pick. Nevertheless, we may try to improve the finite sample behaviour of our estimators by a good choice of the indices i_k. In Section 4, we discuss how to achieve this.

2.3. A k-means procedure to estimate the groups G_k

Overall, our thresholding method performs well in small samples as illustrated by the simulations in Section 5. However, when the noise level in the data is high, the estimates $\hat{m}_i$ tend to be poor, which in turn may lead to frequent classification errors. In such cases, we may improve on the performance of the thresholding method by an additional k-means clustering step. In particular, we may use the threshold estimators $\hat{G}_1, \ldots, \hat{G}_{\hat{K}}$ as the starting values of a k-means algorithm. As shown in the simulations, the resulting estimators tend to be quite precise even when the noise level in the data is high.

The k-means algorithm has a long tradition in the classification literature. Since its introduction in Cox (1957) and Fisher (1958), many people have worked on it; see for example Pollard (1981, 1982) for consistency and weak convergence results and Garcia-Escudero and Gordaliza (1999), Tarpey and Kinateder (2003), Sun *et al.* (2012) and Ieva *et al.* (2013) for more recent extensions and applications of the algorithm. For the k-means algorithm to work well, two conditions need to be satisfied.

(a) The algorithm presupposes knowledge of the number of classes K. Hence, if we want to apply it, we first must estimate K.
(b) Its performance heavily depends on the starting values. When these are not chosen appropriately, it tends to produce poor results.

Our thresholding method is a neat way to satisfy (a) and (b) simultaneously: it estimates the number of classes K and at the same time produces accurate starting values. It thus provides an appropriate basis for the k-means algorithm to work well.

Our version of the k-means algorithm proceeds as follows: to start with, we compute the mean functions $\hat{g}_k^{[1]}(x) = |\hat{G}_k|^{-1} \sum_{i \in \hat{G}_k} \hat{m}_i(x)$ for each class estimate $\hat{G}_k$ with $1 \leqslant k \leqslant \hat{K}$. Defining $\Delta(q_1, q_2) = \int \{q_1(x) - q_2(x)\}^2 \pi(x) \mathrm{d}x$ to be the squared L_2-distance between two functions $q_l : [0,1] \to \mathbb{R}$ with $l = 1, 2$, we then proceed as follows.

Step 1: for each $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, \hat{K}\}$, calculate the distance $\hat{d}_k(i) = \Delta(\hat{m}_i, \hat{g}_k^{[1]})$ between the function $\hat{m}_i$ and the cluster mean $\hat{g}_k^{[1]}$. Define the classes $\{G_k^{[1]} : 1 \leqslant k \leqslant \hat{K}\}$ by assigning the index i to the kth class $G_k^{[1]}$ if $\hat{d}_k(i) = \min_{1 \leqslant k' \leqslant \hat{K}} \hat{d}_{k'}(i)$.
Step r: let $\{G_k^{[r-1]} : 1 \leqslant k \leqslant \hat{K}\}$ be the classes from the previous iteration step. Calculate mean functions $\hat{g}_k^{[r]} = |G_k^{[r-1]}|^{-1} \sum_{i \in G_k^{[r-1]}} \hat{m}_i$ and compute the distances $\hat{d}_k(i) = \Delta(\hat{m}_i, \hat{g}_k^{[r]})$ for each i and k. Define the new classes $\{G_k^{[r]} : 1 \leqslant k \leqslant \hat{K}\}$ by assigning the index i to the kth group $G_k^{[r]}$ if $\hat{d}_k(i) = \min_{1 \leqslant k' \leqslant \hat{K}} \hat{d}_{k'}(i)$.

This algorithm is iterated until the computed classes do not change any more. For a given sample of data, this is guaranteed to happen after finitely many steps. We thus obtain estimators of the classes $\{G_k : 1 \leqslant k \leqslant K\}$ which are denoted by $\{\hat{G}_k^{\mathrm{KM}} : 1 \leqslant k \leqslant \hat{K}\}$ in what follows.

2.4. Estimation of the functions g_k

Once we have constructed estimators of the groups G_k, it is straightforward to come up with good estimators of the functions g_k. In particular, we define

$$\hat{g}_k(x) = \frac{1}{|\hat{G}_k|} \sum_{i \in \hat{G}_k} \hat{m}_i(x),$$

where $|\hat{G}_k|$ denotes the cardinality of the set $\hat{G}_k$. Hence, we simply average the kernel smoothers $\hat{m}_i$ with indices in the estimated group $\hat{G}_k$. When we additionally perform the k-means algorithm from the previous section, the threshold estimators $\hat{G}_k$ should of course be replaced by the refined versions $\hat{G}_k^{\mathrm{KM}}$ in the definition of $\hat{g}_k$.

3. Asymptotics

In this section, we investigate the asymptotic properties of our estimators. We first list the assumptions that are needed for the analysis and then summarize the main results. The proofs can be found in the on-line supplementary material.

3.1. Assumptions

Condition 1. The time series processes $\mathcal{Z}_i = \{(X_{it}, \varepsilon_{it}) : 1 \leqslant t \leqslant T\}$ are independent across i. Moreover, they are strictly stationary and strongly mixing for each i. Let $\alpha_i(l)$ for $l = 1, 2, \ldots$ be the mixing coefficients corresponding to the ith time series $\mathcal{Z}_i$. It holds that $\alpha_i(l) \leqslant \alpha(l)$ for all $1 \leqslant i \leqslant n$, where the coefficients $\alpha(l)$ decay exponentially fast to 0 as $l \to \infty$.

Condition 2. The functions g_k $(1 \leqslant k \leqslant K)$ are twice continuously differentiable. The densities f_i of the variables X_{it} exist and have bounded support, which without loss of generality equals $[0, 1]$. They are uniformly bounded away from 0 and ∞, i.e. $0 < c \leqslant \min_{1 \leqslant i \leqslant n} \inf_{x \in [0,1]} f_i(x)$ and $\max_{1 \leqslant i \leqslant n} \sup_{x \in [0,1]} f_i(x) \leqslant C < \infty$ for some constants $0 < c \leqslant C < \infty$. Moreover, they are twice continuously differentiable on $[0, 1]$ with first and second derivatives that are uniformly bounded away from ∞. Finally, the joint densities $f_{i,l}$ of (X_{it}, X_{it+l}) exist and are uniformly bounded away from ∞ as well.

Condition 3. There are a real number $\theta > 4$ and a natural number l^* such that, for any $l \in \mathbb{Z}$ with $|l| \geqslant l^*$ and a fixed constant $C < \infty$,

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

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

Condition 4. The time series dimension T tends to ∞, whereas the cross-section dimension n may be either fixed or diverging. Their relative growth is such that $n/T \leqslant C$ for some constant $C < \infty$. The bandwidth h has the property that $cT^{-2/5} \leqslant h \leqslant CT^{-\delta}$ for some small $\delta > 0$ and positive constants c and C.

Condition 5. The kernel W is non-negative and bounded. Moreover, it is symmetric about zero, has compact support (say $[-C_1, C_1]$) and fulfils the Lipschitz condition that there is a positive constant L with $|W(x) - W(x')| \leqslant L|x - x'|$ for all $x, x' \in \mathbb{R}$. We use the notation $\|W\|^2 = \int W^2(x)\,\mathrm{d}x$ and $\|W * W\|^2 = \int \{\int W(x)\, W(x + y)\,\mathrm{d}x\}^2\,\mathrm{d}y$.

We finally suppose that the weight function π in the definition of the distances Δ_{ij} is bounded and that its support is contained in that of the regressors, i.e. $\mathrm{supp}(\pi) \subseteq [0, 1]$.

We briefly comment on these assumptions. First, we do not necessarily require exponentially decaying mixing rates as assumed in condition 1. 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. Conditions 2 and 3 are standard-type smoothness and moment conditions that are needed to derive uniform convergence results for the kernel estimators on which our methods are based; see Hansen (2008) for similar assumptions. Condition 4 imposes restrictions on the relative growth of the two dimensions n and T. There is a trade-off between these restrictions and the moment condition that $\theta > 4$ in condition 3. In particular, it is possible to relax condition 4 at the cost of a stronger moment condition. For example, we can weaken condition 4 to allow for $n/T^{3/2} \leqslant C$, if we strengthen the moment condition to $\theta > 5$. Importantly, we do not impose any restrictions on the class sizes $n_k = |G_k|$ for $1 \leqslant k \leqslant K$. They only need to fulfil the trivial conditions that $n_k \leqslant n$ for $1 \leqslant k \leqslant K$ and $\Sigma_{k=1}^{K} n_k = n$. The sizes n_k may thus be very different across the classes G_k. In particular, they may be fixed for some classes and grow to ∞ at different rates for others.

3.2. Main results

We first investigate the asymptotic properties of the threshold estimators $\{\hat{G}_k : 1 \leqslant k \leqslant \hat{K}\}$. To do so, we require the threshold parameter $\tau_{n,T}$ to fulfil the following condition.

Condition 6.

$$\tau_{n,T} \to 0 \qquad \text{such that} \quad \mathbb{P}\Big(\max_{i,j \in G_k} \hat{\Delta}_{ij} \leqslant \tau_{n,T}\Big) \to 1 \quad \text{for } 1 \leqslant k \leqslant K.$$

This condition is in particular satisfied by any threshold $\tau_{n,T}$ which converges to 0 more slowly than $\max_{i,j\in G_k}\hat{\Delta}_{ij}$ for $1\leqslant k\leqslant K$. More formally, suppose that $\max_{i,j\in G_k}\hat{\Delta}_{ij}=O_p(c_{n,T})$ for some null sequence $\{c_{n,T}\}$ and any k. Then any null sequence $\{\tau_{n,T}\}$ with $\tau_{n,T}/c_{n,T}\to\infty$ satisfies condition 6. In the on-line supplementary material, we show that $\max_{i,\,j\in G_k}\hat{\Delta}_{ij}=O_p(c_{n,T})$ with $c_{n,T}=T^{-1/5}+h^3$ under conditions 1–5 and $c_{n,T}=\log(T)/(Th)+h^3$ provided that the moment assumptions in condition 3 are strengthened to hold for some $\theta>20/3$. Notably, these are only upper bounds on the rate of $\max_{i,j\in G_k}\hat{\Delta}_{ij}$. In lemma S.2 in the supplementary material, we derive the sharp rate $\max_{i,j\in G_k}\hat{\Delta}_{ij}=O_p\{1/(Th)\}$ under more restrictive conditions than conditions 1–5. This lemma also provides us with a more concise characterization of the threshold sequences that satisfy condition 6. It shows that $\max_{i,j\in G_k}\hat{\Delta}_{ij}\leqslant b_{n,T}+\rho_{n,T}$, where the leading term $b_{n,T}$ has the form $b_{n,T}=\|W\|^2\max_{1\leqslant i<j\leqslant n}(b_i+b_j)/(Th)$ with $b_i=\int\sigma_i^2(x)\pi(x)/f_i(x)\mathrm{d}x$ and $\sigma_i^2(x)=\mathbb{E}[\varepsilon_{it}^2|X_{it}=x]$. The lower order terms are summarized by the expression $\rho_{n,T}=O_p\{\log(T)/(Th^{1/2})\}$. From this, it immediately follows that any null sequence $\{\tau_{n,T}\}$ with $\tau_{n,T}\geqslant b_{n,T}+\bar{\rho}_{n,T}$ fulfils condition 6, where $\bar{\rho}_{n,T}$ is an upper bound on the lower order terms $\rho_{n,T}$ satisfying $\bar{\rho}_{n,T}/\rho_{n,T}\to\infty$.

Our first result shows that the threshold estimators $\{\hat{G}_k:1\leqslant k\leqslant\hat{K}\}$ are consistent in the following sense: they coincide with the true classes $\{G_k:1\leqslant k\leqslant K\}$ with probability tending to 1, provided that the threshold parameter $\tau_{n,T}$ fulfils condition 6.

Theorem 1. Let conditions 1–5 be satisfied and suppose that $\tau_{n,T}$ fulfils condition 6. Then $\mathbb{P}(\hat{K}\neq K)=o(1)$ and

$$\mathbb{P}(\{\hat{G}_k:1\leqslant k\leqslant\hat{K}\}\neq\{G_k:1\leqslant k\leqslant K\})=o(1).$$

Note that the indexing of the estimators $\hat{G}_1,\ldots,\hat{G}_{\hat{K}}$ is completely arbitrary. We could, for example, change the indexing according to the rule $k\mapsto\hat{K}-k+1$. In what follows, we suppose that the estimated classes are indexed such that $\mathbb{P}(\hat{G}_k=G_k)\to 1$ for all k. Theorem 1 implies that this is possible without loss of generality. The next theorem shows that the k-means estimators $\{\hat{G}_k^{\mathrm{KM}}:1\leqslant k\leqslant\hat{K}\}$ inherit the consistency property of theorem 1 from the threshold estimators $\{\hat{G}_k:1\leqslant k\leqslant\hat{K}\}$.

Theorem 2. Under the conditions of theorem 1, it holds that

$$\mathbb{P}(\{\hat{G}_k^{\mathrm{KM}}:1\leqslant k\leqslant\hat{K}\}\neq\{G_k:1\leqslant k\leqslant K\})=o(1).$$

As above, the k-means estimators $\hat{G}_1^{\mathrm{KM}},\ldots,\hat{G}_{\hat{K}}^{\mathrm{KM}}$ are indexed such that $\mathbb{P}(\hat{G}_k^{\mathrm{KM}}=G_k)\to 1$ for all k.

We next turn to the asymptotic properties of the estimators $\hat{g}_k$. To formulate them, we introduce some notation: let $\hat{n}_k=|\hat{G}_k|$ be the cardinality of $\hat{G}_k$ and let the constant c_k be implicitly defined by the formula $h/(\hat{n}_kT)^{-1/5}\to^{\mathrm{P}}c_k$. Noting that the group size $n_k=|G_k|$ depends on n in general, i.e. $n_k=n_k(n)$, we define the terms

$$B_k(x)=\frac{c_k^{5/2}}{2}\int W(\varphi)\varphi^2\mathrm{d}\varphi\lim_{n\to\infty}\left\{\frac{1}{n_k}\sum_{i\in G_k}\frac{g_k''(x)f_i(x)+2g_k'(x)f_i'(x)}{f_i(x)}\right\},$$

$$V_k(x)=\int W^2(\varphi)\mathrm{d}\varphi\lim_{n\to\infty}\left\{\frac{1}{n_k}\sum_{i\in G_k}\frac{\sigma_i^2(x)}{f_i(x)}\right\},$$

where we implicitly suppose that the limit expressions exist. The terms $B_k(x)$ and $V_k(x)$ play the role of the asymptotic bias and variance in what follows. The next theorem specifies the rate of convergence and the limit distribution of $\hat{g}_k$.

Theorem 3. Let the conditions of theorem 1 be satisfied and additionally suppose that $\max_{1\leqslant i\leqslant n}\sup_{x,x'\in[0,1]}\mathbb{E}[|\varepsilon_{it}\|X_{it}=x, X_{it+l}=x']\leqslant C<\infty$ for all $l=\pm1,\pm2,\ldots$ and some constant C. Then, for any fixed $x\in(0,1)$,

$$\hat{g}_k(x)-g_k(x)=O_p\bigg\{\frac{1}{\sqrt{(n_kTh)}}+h^2\bigg\}. \tag{3.1}$$

Moreover, if $n\to\infty$ and the bandwidth h is such that $h/(\hat{n}_kT)^{-1/5}\to^{\mathrm{P}}c_k$ for some constant $c_k>0$, then, for any fixed $x\in(0,1)$,

$$\sqrt{(\hat{n}_kTh)}\{\hat{g}_k(x)-g_k(x)\}\xrightarrow{\mathrm{d}}N\{B_k(x),V_k(x)\}. \tag{3.2}$$

When deriving the limit distribution in result (3.2), we restrict attention to the case that $n\to\infty$ for the following reason: if n is finite, the estimation error in $\hat{Y}_{it}^{\mathrm{fe}}$ that is induced by subtracting the sample averages $\bar{Y}_i$, $\bar{Y}_t^{(i)}$ and $\bar{\bar{Y}}^{(i)}$ is asymptotically not negligible but contributes to the limit distribution. If $n\to\infty$, this error is negligible in contrast, allowing us to derive clean expressions for the asymptotic bias and variance.

4. Implementation

Our thresholding approach to estimate the class structure $\{G_k:1\leqslant k\leqslant K\}$ depends on two tuning parameters: the threshold level $\tau_{n,T}$ and the bandwidth h of the kernel smoothers $\hat{m}_i$. In addition, we need to pick an index i_k in each iteration step of the algorithm. In what follows, we give some heuristic arguments on how to choose the threshold $\tau_{n,T}$ in an appropriate way. Moreover, we derive a selection rule for the bandwidth h and discuss the choice of the indices i_k. In addition, we outline some modifications of our estimation methods.

4.1. Choice of the threshold level $\tau_{n,T}$

Suppose that we are given some index $i\in G$ and want to estimate the unknown class G by our thresholding procedure. As suggested by the discussion in Section 2.2, in particular by formula (2.2), we would ideally like to choose the threshold $\tau_{n,T}$ to be slightly larger than $\max_{j\in G}\hat{\Delta}_{ij}$. We now explain how to achieve this.

To keep the derivations as clear as possible, we drop the fixed effects α_i and γ_t from the model. Modifying the arguments from Härdle and Mammen (1993), we can show that, for any $j\in G$ with $j\neq i$,

$$Th^{1/2}\hat{\Delta}_{ij}-h^{-1/2}\mathcal{B}_{ij}\xrightarrow{\mathrm{d}}N(0,\mathcal{V}_{ij}) \tag{4.1}$$

under slightly strengthened versions of conditions 1–5. The bias and variance expressions in formula (4.1) are of the form $\mathcal{B}_{ij}=\|W\|^2(b_i+b_j)$ and $\mathcal{V}_{ij}=\|W*W\|^2(2v_{ii}+4v_{ij}+2v_{jj})$, where $\|W\|^2$ and $\|W*W\|^2$ are defined in condition 5, $b_i=\int\sigma_i^2(x)\pi(x)/f_i(x)\mathrm{d}x$ and

$$v_{ij}=\int\frac{\sigma_i^2(x)\,\sigma_j^2(x)\,\pi^2(x)}{f_i(x)\,f_j(x)}\mathrm{d}x$$

with $\sigma_i^2(x)=\mathbb{E}[\varepsilon_{it}^2|X_{it}=x]$. Roughly speaking, expression (4.1) says that

$$\hat{\Delta}_{ij}\approx\Delta_{ij}^*:=\frac{\mathcal{B}_{ij}}{Th}+\frac{\sqrt{\mathcal{V}_{ij}}}{Th^{1/2}}Z_{ij}, \tag{4.2}$$

i.e. $\hat{\Delta}_{ij}$ is approximately distributed as Δ^*_{ij}, where Z_{ij} is a standard normal random variable. Since $\hat{\Delta}_{ii}=0$ by construction, we additionally set $\Delta^*_{ii}=0$. Neglecting the estimation error in expression (4.2), we replace the variables $\hat{\Delta}_{ij}$ by Δ^*_{ij} and aim to choose the threshold $\tau_{n,T}$ slightly larger than $\max_{j\in G}\Delta^*_{ij}$: to start with, it holds that

$$\max_{j\in G}\Delta^*_{ij} \leqslant \max_{j\in G_{-i}}\frac{\mathcal{B}_{ij}}{Th}+\max_{j\in G_{-i}}\left(\frac{\sqrt{\mathcal{V}_{ij}}}{Th^{1/2}}\max_{j\in G_{-i}}Z_{ij}\right),$$

where $G_{-i}=G\setminus\{i\}$. Moreover, since a standard normal random variable Z has the property that $\mathbb{P}(Z\geqslant z)\leqslant(2\pi z^2)^{-1/2}\exp(-z^2/2)$ for $z>0$, we obtain that

$$\mathbb{P}\Big\{\max_{j\in G_{-i}}Z_{ij}\geqslant(2\log|G|)^{1/2}\Big\}\leqslant\sum_{j\in G_{-i}}\mathbb{P}\{Z_{ij}\geqslant(2\log|G|)^{1/2}\}\leqslant\frac{1}{\sqrt{(4\pi\log|G|)}}.$$

Hence, if the class size $|G|$ is sufficiently large, the maximum $\max_{j\in G_{-i}}Z_{ij}$ will be rarely larger than $(2\log|G|)^{1/2}$. We can thus infer that

$$\max_{j\in G}\Delta^*_{ij}\leqslant\max_{j\in G_{-i}}\frac{\mathcal{B}_{ij}}{Th}+\max_{j\in G_{-i}}\left(\frac{\sqrt{\mathcal{V}_{ij}}}{Th^{1/2}}\right)(2\log|G|)^{1/2}\leqslant b_{n,T}+v_{n,T}(2\log|G|)^{1/2}$$

with probability tending to 1 as $|G|\to\infty$, where $b_{n,T}=\max_{1\leqslant i<j\leqslant n}\mathcal{B}_{ij}/(Th)$ and $v_{n,T}=\max_{1\leqslant i<j\leqslant n}\sqrt{\mathcal{V}_{ij}}/(Th^{1/2})$. These considerations suggest that an appropriate threshold level is given by

$$\tau_{n,T}=b_{n,T}+v_{n,T}(2\log|G|)^{1/2}. \tag{4.3}$$

Importantly, this heuristically motivated choice of the threshold is essentially in line with our theoretical results from Section 3.2. As discussed there, under the conditions of lemma S.2 from the on-line supplementary material, we can work with threshold sequences of the form $\tau_{n,T}\geqslant b_{n,T}+$ lower order terms. The threshold that is defined in equation (4.3) has such a form: its leading term is $b_{n,T}$ and the expression $v_{n,T}(2\log|G|)^{1/2}$ is a heuristically motivated bound on the lower order terms.

Of course, the threshold level in equation (4.3) is not a feasible choice as

(a) it depends on the unknown class G and
(b) the expressions $b_{n,T}$ and $v_{n,T}$ are not known.

To remove the dependence on G, we may replace the unknown class size $|G|$ by the trivial bound n. This leads to the threshold level

$$\tau_{n,T}(p)=b_{n,T}+v_{n,T}\{2\log(p)\}^{1/2} \tag{4.4}$$

with $p=n$. As n is quite a rough bound on the class size $|G|$, we refine this choice as follows: in the first step of our thresholding algorithm, we set the threshold level to $\tau_{n,T}(n)$. Next suppose that we are in the kth iteration step and let $\hat{G}_1,\ldots,\hat{G}_{k-1}$ be the estimated classes from the previous steps. Defining $\hat{n}_l=|\hat{G}_l|$, we set $p=n-\sum_{l=1}^{k-1}\hat{n}_l$ and use the threshold $\tau_{n,T}(p)$ to estimate G_k. We thus exploit the information from the previous iteration steps to obtain a better bound on the class size $|G_k|$.

To compute the threshold (4.4) in practice, we finally need to estimate the terms $b_{n,T}$ and $v_{n,T}$. The only unknown expressions in $b_{n,T}$ and $v_{n,T}$ are the conditional variances σ_i^2 and the densities f_i, which can be estimated by standard kernel smoothers. In particular, we may approximate σ_i^2 by $\hat{\sigma}_i^2(x)=(Th)^{-1}\sum_{t=1}^T W_h(X_{it}-x)\hat{\varepsilon}_{it}^2/\hat{f}_i(x)$, where $\hat{f}_i(x)=(Th)^{-1}\sum_{t=1}^T W_h(X_{it}-x)$ and $\hat{\varepsilon}_{it}=Y_{it}-\hat{m}_i(X_{it})$ are the estimated residuals. Moreover, we may estimate f_i by the modified

kernel density $\hat{f}_i^{\mathrm{bc}}(x) = \{\int_{-x/h}^{(1-x)/h} W(\varphi)\,\mathrm{d}\varphi\}^{-1} \hat{f}_i(x)$, where the correction $\{\int_{-x/h}^{(1-x)/h} W(\varphi)\,\mathrm{d}\varphi\}^{-1}$ prevents the estimator from becoming inconsistent at the boundary.

To make the estimates of $b_{n,T}$ and $v_{n,T}$ more robust, we recommend the following two modifications.

(a) The terms $b_{n,T}$ and $v_{n,T}$ are essentially maxima over the bias and variance expressions $\mathcal{B}_{ij}$ and $\mathcal{V}_{ij}$ that depend on the unknown functions σ_i^2 and f_i. It goes without saying that a poor estimate of σ_i^2 or f_i for some i may strongly influence our approximations of these maxima. To make our estimates of $b_{n,T}$ and $v_{n,T}$ more robust to such poor estimates, we suggest replacing the maxima over $\mathcal{B}_{ij}$ and $\mathcal{V}_{ij}$ by a high quantile, say the 95% quantile.

(b) As is well known from other studies, the conditional variances σ_i^2 are quite difficult to estimate accurately. We may thus expect to obtain poor estimates $\hat{\sigma}_i^2$ at least for some indices i. These few poor estimates may strongly affect our approximations of $b_{n,T}$ and $v_{n,T}$. To avoid this issue, we recommend replacing the estimates $\hat{\sigma}_i^2$ by the simple averages $\bar{\varepsilon}_i^2 := T^{-1} \sum_{t=1}^T \hat{\varepsilon}_{it}^2$, which estimate the unconditional variances $\mathbb{E}[\varepsilon_{it}^2]$. Strictly speaking, this is of course only allowed when the error terms ε_{it} are homoscedastic and thus $\mathbb{E}[\varepsilon_{it}^2 | X_{it} = x] = \mathbb{E}[\varepsilon_{it}^2]$. However, the error resulting from replacing $\hat{\sigma}_i^2$ with $\bar{\varepsilon}_i^2$ can be expected to be much smaller than the error stemming from the instabilities in the estimates $\hat{\sigma}_i^2$.

Both in the simulations and in the application, we work with the modifications (a) and (b).

We finally note that the estimation of $b_{n,T}$ and $v_{n,T}$ strongly simplifies if it is possible to impose some additional restrictions on the functions σ_i^2 and f_i. Suppose for example that the conditional error variance and the distribution of the covariates are (virtually) the same across individuals i. In this case, $\sigma_i^2 = \sigma^2$ and $f_i = f$ for all i and some functions σ^2 and f. The terms $b_{n,T}$ and $v_{n,T}$ simplify to $b_{n,T} = 2(Th)^{-1} \|W\|^2 \int \sigma^2(x)\,\pi(x)/f(x)\,\mathrm{d}x$ and $v_{n,T} = (Th^{1/2})^{-1} \{8\,\|W * W\|^2 \int \sigma^4(x)\,\pi^2(x)/f^2(x)\,\mathrm{d}x\}^{1/2}$. To estimate them, we do not have to compute any maxima any more. Moreover, the common functions σ^2 and f can be estimated much more precisely than σ_i^2 and f_i.

4.2. Choice of the indices i_k

To compute the threshold estimators $\{\hat{G}_k : 1 \leqslant k \leqslant \hat{K}\}$, we need to pick an index i_k from the index set $S_k = \{1, \ldots, n\} \setminus \cup_{l=1}^{k-1} \hat{G}_l$ in each iteration step of the algorithm. As already mentioned in Section 2.2, there is in principle no restriction on how to choose the indices i_k. Nevertheless, there are ways of selecting i_k which can be expected to improve the finite sample performance of the estimators. We now describe such a selection rule.

Rule 1. For each $i \in S_k$, compute $\hat{p}_{i,S_k}$ as defined in equation (2.3) and calculate the jump size $\hat{J}_{i,S_k} = \hat{\Delta}_{i[\hat{p}_{i,S_k}+1]} - \hat{\Delta}_{i[\hat{p}_{i,S_k}]}$, where we set $\hat{\Delta}_{i[n_{S_k}+1]} = (2+\delta)\max_{i,j \in S_k} \hat{\Delta}_{ij}$ with some $\delta > 0$ and $n_{S_k} = |S_k|$. Pick the index $i \in S_k$ for which $\hat{J}_{i,S_k}$ is maximal, i.e. define $i_k = \arg\max_{i \in S_k} \hat{J}_{i,S_k}$.

The heuristic idea behind this rule is as follows: $\hat{p}_{i,S_k}$ is the position where the ordered estimates $\hat{\Delta}_{i[1]}, \ldots, \hat{\Delta}_{i[n_{S_k}]}$ exceed the threshold value $\tau_{n,T}$. Put differently, $\hat{p}_{i,S_k}$ estimates the position where the ordered distances $\Delta_{i(1)}, \ldots, \Delta_{i(n_{S_k})}$ jump from 0 to a positive value. Rule 1 suggests that we pick the index i for which the estimated jump size is largest, i.e. for which the jump is most clearly visible in the data. Moreover, the rule is constructed such that we pick an index i with $\hat{p}_{i,S_k} = n_{S_k}$ as soon as such an index occurs. The rationale behind this is as follows: if $\hat{p}_{i,S_k} = n_{S_k}$, then all distances $\hat{\Delta}_{i[1]}, \ldots, \hat{\Delta}_{i[n_{S_k}]}$ are smaller than the threshold $\tau_{n,T}$. This indicates that all indices in S_k should belong to the same class. We thus stop the algorithm as soon as we

encounter such an index. This in particular prevents our estimator $\hat{K}$ from strongly overshooting the true number of classes K.

Rule 1 requires us to compute the positions $\hat{p}_{i,S_k}$ for each $i \in S_k$. This is of course computationally burdensome when the cross-section dimension n is very large. We thus recommend the use of rule 1 only for data samples with a moderately large dimension n. For very large n, more rudimentary rules are needed. For example, one may simply select the indices i_k as random draws from the sets S_k.

4.3. Bandwidth choice for $\hat{m}_i$

When deriving our methods, we have implicitly assumed that the smoothers $\hat{m}_i$ depend on a common bandwidth h. We now drop this assumption and allow for different bandwidths h_i. From a practical point of view, it is, however, not very desirable to select a different bandwidth for each individual i. The computational cost is simply too high, in particular when n is large. For this reason, we suggest choosing group-specific bandwidths: for each group G_k, we select a bandwidth h_k which is used to compute the estimators $\hat{m}_i = \hat{m}_{i,h_k}$ with $i \in G_k$. We derive our group-specific bandwidth selection rule under the assumption that the stochastic behaviour of the time series processes $\mathcal{Z}_i = \{(Y_{it}, X_{it}) : 1 \leqslant t \leqslant T\}$ does not differ too much within groups. Technically speaking, we suppose that not only are the functions m_i the same within groups but also the densities f_i and the conditional variances $\sigma_i^2(\cdot) = \mathbb{E}[\varepsilon_{it}^2 | X_{it} = \cdot]$. To keep the derivations as clear as possible, we additionally make the following simplifications: we drop the fixed effects α_i and γ_t from the model, we ignore the time series dependence in the data and we suppose that the errors ε_{it} are independent from the covariates X_{it}. We now derive our bandwidth selector step by step.

First suppose that we want to optimize the bandwidth h of $\hat{m}_i = \hat{m}_{i,h}$ for a fixed individual i. This can be achieved by standard methods: following Härdle *et al.* (1988), we take the optimal bandwidth to be the minimizer h_i^{opt} of the average squared error

$$\text{ASE}_i(h) = \frac{1}{T} \sum_{t=1}^{T} \{\hat{m}_{i,h}(X_{it}) - m_i(X_{it})\}^2 \, w(X_{it}),$$

where w is some weight function, and approximate h_i^{opt} by minimizing some estimate of $\text{ASE}_i(h)$ with respect to h. The estimates of $\text{ASE}_i(h)$ that are commonly considered in the literature are closely related to the residual sum of squares

$$\text{RSS}_i(h) = \frac{1}{T} \sum_{t=1}^{T} \{Y_{it} - \hat{m}_{i,h}(X_{it})\}^2 \, w(X_{it}),$$

but they are not identical with it. Indeed, we cannot minimize $\text{RSS}_i(h)$ directly but must modify it. The heuristic reason is as follows: the residual sum of squares $\text{RSS}_i(h)$ can be interpreted as a prediction error. More specifically, it measures the error which results from predicting the observations Y_{it} by the estimates $\hat{m}_{i,h}(X_{it})$ for $t = 1, \ldots, T$. Since the observation Y_{it} is contained in the estimate $\hat{m}_{i,h}(X_{it})$, it is used to predict itself. This creates a bias term which prevents the minimizer of $\text{RSS}_i(h)$ from being a reasonable approximation of h_i^{opt}. Formally speaking, it holds that $\mathbb{E}[\text{RSS}_i(h)] = \mathbb{E}[\text{ASE}_i(h)] + B_{i,1}(h) + B_{i,2}(h)$, where $B_{i,1}(h) = T^{-1} \sum_{t=1}^{T} \sigma_i^2 \, \mathbb{E}[w(X_{it})]$ and

$$B_{i,2}(h) = -\frac{2W(0)}{T^2 h} \sum_{t=1}^{T} \sigma_i^2 \, \mathbb{E}\left[\frac{w(X_{it})}{\hat{f}_i(X_{it})}\right]$$

with $\sigma_i^2 = \mathbb{E}[\varepsilon_{it}^2]$ and $\hat{f}_i(x) = T^{-1} \sum_{t=1}^{T} W_h(X_{it} - x)$. The first bias term $B_{i,1}(h)$ is harmless as it is

independent of h. The second bias $B_{i,2}(h)$, however, is very problematic. As one can show, it has the effect that minimizing the residual sum of squares leads to bandwidths which are too small.

To correct for the bias $B_{i,2}(h)$, cross-validation or penalization techniques are commonly used; see for example Härdle *et al.* (1988). In our panel set-up, we can circumvent this bias issue in a simpler way, in particular by borrowing information from other individuals j: suppose that we know that i and j belong to the same class G_k. In this situation, we may replace the residual sum of squares $\mathrm{RSS}_i(h)$ by

$$\mathrm{RSS}_i^{(j)}(h) = \frac{1}{T}\sum_{t=1}^{T}\big\{Y_{jt} - \hat{m}_{i,h}(X_{jt})\big\}^2 w(X_{jt}),$$

i.e. we may use the estimator $\hat{m}_{i,h}$ to predict the Y-observations of the jth rather than the ith individual. This avoids the bias problem since the data of the jth individual are independent from those of the ith subject. Formally speaking, we obtain that $\mathbb{E}[\mathrm{RSS}_i^{(j)}(h)] = \mathbb{E}[\mathrm{ASE}_i^{(j)}(h)] + T^{-1}\sum_{t=1}^{T}\sigma_j^2\,\mathbb{E}[w(X_{jt})]$, where

$$\mathrm{ASE}_i^{(j)}(h) = \frac{1}{T}\sum_{t=1}^{T}\big\{\hat{m}_{i,h}(X_{jt}) - m_i(X_{jt})\big\}^2 w(X_{jt}).$$

This shows that we can remove the problematic bias component. We may thus choose the bandwidth of the ith individual by simply minimizing the residual sum of squares criterion $\mathrm{RSS}_i^{(j)}(h)$. Since

$$\begin{aligned}\mathbb{E}[\mathrm{ASE}_i^{(j)}(h)] &= \mathbb{E}[\mathbb{E}[\mathrm{ASE}_i^{(j)}(h)|\{(Y_{it}, X_{it}) : 1 \leqslant t \leqslant T\}]] \\ &= \mathbb{E}\bigg[\int \{\hat{m}_{i,h}(x) - m_i(x)\}^2 f_i(x)\, w(x)\,\mathrm{d}x\bigg] \\ &=: \mathrm{MISE}_i(h), \end{aligned} \tag{4.5}$$

i.e., since the expectation of $\mathrm{ASE}_i^{(j)}(h)$ is nothing other than the mean integrated squared error $\mathrm{MISE}_i(h)$, the chosen bandwidth can be regarded as an approximation of the optimal bandwidth in an MISE-sense.

So far, we have discussed the choice of the bandwidth for a fixed individual i. We now use the ideas from above to set up a group-specific bandwidth selector. To start with, suppose that the class G_k is known and write $G_k = \{i_1, i_2, \ldots, i_{n_k}\}$ with $n_k = |G_k|$. Moreover, pick pairs of indices (i_{2l-1}, i_{2l}) for $1 \leqslant l \leqslant L$ and some $L \leqslant \lfloor n_k/2 \rfloor$. We compute the bandwidth estimate $\hat{h}_{i_{2l-1}}^{(i_{2l})} = \arg\min_h \mathrm{RSS}_{i_{2l-1}}^{(i_{2l})}(h)$ for each $1 \leqslant l \leqslant L$ and define our group-specific bandwidth selector by

$$\hat{h}_k = \frac{1}{L}\sum_{1 \leqslant l \leqslant L} \hat{h}_{i_{2l-1}}^{(i_{2l})}.$$

Since the mean integrated squared error $\mathrm{MISE}_i(h)$ is the same for all $i \in G_k$ under our conditions, the bandwidth estimate $\hat{h}_k$ can be interpreted as an approximation to the groupwide optimal bandwidth in an MISE-sense. It is worth noting that we need not take into account all pairs of indices (i_{2l-1}, i_{2l}) to compute $\hat{h}_k$; we may rather pick a small number L of them to keep the computational burden of the selection procedure to a minimum.

In practice, our group-specific bandwidth selector is implemented as follows.

Step 1: as the classes $G_1, \ldots, G_K$ are not known in practice, we replace them by preliminary estimators. To do so, we choose a preliminary bandwidth h_0 which is the same for all $i \in \{1, \ldots, n\}$. This is done as follows: pick a small number N of indices $i_1, \ldots, i_N \in \{1, \ldots, n\}$

and apply a standard bandwidth selection rule to each index separately. For example, we may minimize a penalized version of the residual sum of squares criterion $\mathrm{RSS}_i(h)$ for each of the indices or apply a plug-in type of selection rule as described in Fan and Gijbels (1996). We finally set h_0 to be the average of the computed bandwidths. On the basis of the bandwidth h_0, we can compute preliminary estimators $\tilde{G}_1, \ldots, \tilde{G}_{\tilde{K}}$ of the classes.

Step 2: for each estimated class $\tilde{G}_k$, we calculate the bandwidth $\hat{h}_k$ as described above.

On the basis of the bandwidths $\hat{h}_k$, we can re-estimate the classes $G_1, \ldots, G_K$ by our thresholding procedure. To do so, we work with a slightly modified threshold parameter $\tau_{n,T}(p)$, which exploits the information that is contained in the preliminary estimates $\tilde{G}_1, \ldots, \tilde{G}_{\tilde{K}}$. In particular, we let $\tau_{n,T}(p) = \max_{1 \leqslant k \leqslant \tilde{K}} [b_{n,T}(\tilde{G}_k) + v_{n,T}(\tilde{G}_k)\{2\log(p)\}^{1/2}]$, where $b_{n,T}(\tilde{G}_k) = \max_{i,j \in \tilde{G}_k, i<j} \mathcal{B}_{ij}/(T\hat{h}_k)$ and $v_{n,T}(\tilde{G}_k) = \max_{i,j \in \tilde{G}_k, i<j} \sqrt{\mathcal{V}_{ij}}/(T\hat{h}_k^{1/2})$ with $\mathcal{B}_{ij}$ and $\mathcal{V}_{ij}$ defined in Section 4.1. We thus obtain updated estimates $\hat{G}_1, \ldots, \hat{G}_{\hat{K}}$. We finally calculate group-specific bandwidths based on the updated estimates $\hat{G}_k$, which we again denote by $\hat{h}_k$. These are used in the next section to come up with a good bandwidth selection rule for the estimators $\hat{g}_k$.

4.4. Bandwidth choice for $\hat{g}_k$

Suppose that the conditions of Section 4.3 are fulfilled. In particular, assume that the densities f_i and the conditional variances σ_i^2 are the same for all $i \in G_k$. In this situation, the individual smoothers $\hat{m}_i(x) = \hat{m}_{i,h}(x)$ have the same asymptotic bias $b_{i,h}(x)$ and variance $v_{i,h}(x)$ for all $i \in G_k$. Specifically, $b_{i,h}(x) = (h^2/2)\,\beta_k(x)$ and $v_{i,h}(x) = (Th)^{-1}\,\nu_k(x)$ with

$$\beta_k(x) = \int W(\varphi)\varphi^2 \,\mathrm{d}\varphi \frac{g_k''(x)\,f_k(x) + 2g_k'(x)\,f_k'(x)}{f_k(x)}$$

and $\nu_k(x) = \int W^2(\varphi)\,\mathrm{d}\varphi \sigma_k^2(x)/f_k(x)$, where, by a slight abuse of notation, we denote the group-specific density and conditional variance by f_k and σ_k^2 respectively. By theorem 3, the asymptotic bias and variance expressions of $\hat{g}_k(x) = \hat{g}_{k,h}(x)$ have a very similar form: they are equal respectively to $B_{k,h}(x) = (h^2/2)\,\beta_k(x)$ and $V_{k,h}(x) = (n_k Th)^{-1}\,\nu_k(x)$ with $n_k = |G_k|$. With these expressions at hand, we define the criterion functions $\xi_i(h) = \int \{b_{i,h}^2(x) + v_{i,h}(x)\}\, f_k(x) w(x)\,\mathrm{d}x$ and $\Xi_k(h) = \int \{B_{k,h}^2(x) + V_{k,h}(x)\}\, f_k(x) w(x)\,\mathrm{d}x$. Optimizing the bandwidth of the smoother $\hat{m}_{i,h}$ with respect to $\xi_i(h)$ leads to

$$h_k^* = \left\{ \frac{\int \nu_k(x)\, f_k(x)\, w(x)\,\mathrm{d}x}{\int \beta_k^2(x)\, f_k(x) w(x)\,\mathrm{d}x} \right\}^{1/5} T^{-1/5}$$

for all $i \in G_k$. Analogously, optimizing the bandwidth of $\hat{g}_{k,h}$ with respect to $\Xi_k(h)$ yields $H_k^* = n_k^{-1/5} h_k^*$.

As $b_{i,h}(x)$ and $v_{i,h}(x)$ are the leading terms in an asymptotic expansion of $\mathrm{bias}\{\hat{m}_{i,h}(x)\} = \mathbb{E}[\hat{m}_{i,h}(x)] - m_i(x)$ and $\mathrm{var}\{\hat{m}_{i,h}(x)\}$, the criterion function $\xi_i(h)$ is closely related to the mean integrated squared error

$$\begin{aligned}\mathrm{MISE}_i(h) &= \mathbb{E}\Big[\int \{\hat{m}_{i,h}(x) - m_i(x)\}^2 f_i(x) w(x)\,\mathrm{d}x \Big] \\ &= \int \mathrm{bias}\{\hat{m}_{i,h}(x)\}^2 f_i(x) w(x)\,\mathrm{d}x + \int \mathrm{var}\{\hat{m}_{i,h}(x)\} f_i(x) w(x)\,\mathrm{d}x.\end{aligned}$$

Moreover, $\mathrm{MISE}_i(h)$ relates to the criterion function $\mathrm{ASE}_i^{(j)}(h)$ from the previous subsection by equation (4.5). These relationships imply that our group-specific bandwidth selector $\hat{h}_k$ can

be regarded as an approximation of h_k^*. This suggests estimating H_k^* by $\hat{H}_k = \hat{n}_k^{-1/5}\hat{h}_k$, where $\hat{n}_k = |\hat{G}_k|$ is the size of the estimated class $\hat{G}_k$. We thus do not need to run a separate bandwidth selection routine for $\hat{g}_{k,h}$ but can make use of our group-specific bandwidth selector $\hat{h}_k$.

4.5. Rescaling

In many applications, the noise level of the time series data $\mathcal{Z}_i = \{(Y_{it}, X_{it}) : 1 \leqslant t \leqslant T\}$ can be expected to vary across individuals i. As a result, the quality of the estimates $\hat{\Delta}_{ij}$ can be expected to vary as well. To take into account different noise levels in the data, we may replace the estimators $\hat{\Delta}_{ij}$ by suitably scaled versions. This can be achieved as follows: let i and j be two indices that belong to the same class G_k. Equation (4.1) implies that $\hat{\Delta}_{ij} = \mathcal{B}_{ij}/(Th) +$ lower order terms. We can thus infer that $\hat{\Delta}_{ij}^{\text{sc}} := \hat{\Delta}_{ij}/\mathcal{B}_{ij} = (Th)^{-1} +$ lower order terms. The leading term of this expansion is independent of the indices i and j. Hence, the scaled estimators $\hat{\Delta}_{ij}^{\text{sc}}$ should be of comparable size for any pair of indices i and j that belong to the same group.

To account for different noise levels in the data, we may thus base our methods on the scaled estimators $\hat{\Delta}_{ij}^{\text{sc}}$ rather than $\hat{\Delta}_{ij}$. Of course, we cannot take the expressions $\hat{\Delta}_{ij}^{\text{sc}}$ at face value but must estimate the scaling factors $\mathcal{B}_{ij} = \|W\|^2(b_i + b_j)$, which can be achieved by the methods that were described at the end of Section 4.1. Moreover, we need to adjust the threshold level $\tau_{n,T}$. Applying the heuristic arguments from Section 4.1 to the scaled estimators $\hat{\Delta}_{ij}^{\text{sc}}$, the threshold parameter $\tau_{n,T}(p) = b_{n,T} + v_{n,T}\{2\log(p)\}^{1/2}$ from equation (4.4) must be replaced by $\tau_{n,T}^{\text{sc}}(p) = b_{n,T}^{\text{sc}} + v_{n,T}^{\text{sc}}\{2\log(p)\}^{1/2}$. Here, $b_{n,T}^{\text{sc}}$ and $v_{n,T}^{\text{sc}}$ have exactly the same form as $b_{n,T}$ and $v_{n,T}$ with $\mathcal{B}_{ij}$ and $\mathcal{V}_{ij}$ being replaced by $\mathcal{B}_{ij}^{\text{sc}} = 1$ and $\mathcal{V}_{ij}^{\text{sc}} = \|W * W\|^2(2v_{ii} + 4v_{ij} + 2v_{jj})/\mathcal{B}_{ij}^2$ respectively.

5. Simulations

We now investigate the small sample behaviour of our methods by means of a Monte Carlo experiment. The simulation design is set up to mimic the situation in the application of Section 6: we consider the panel model

$$Y_{it} = m_i(X_{it}) + \varepsilon_{it} \qquad 1 \leqslant i \leqslant n, \quad 1 \leqslant t \leqslant T \tag{5.1}$$

with $n = 120$ and $T \in \{100, 150, 200\}$, where $(n, T) = (120, 150)$ approximately corresponds to the sample size in the application. The individuals i are supposed to split into the five groups $G_1 = \{1, \ldots, 50\}, G_2 = \{51, \ldots, 80\}, G_3 = \{81, \ldots, 100\}, G_4 = \{101, \ldots, 110\}$ and $G_5 = \{111, \ldots, 120\}$. The functions that are associated with these groups are $g_1(x) = 0$, $g_2(x) = 1 - 2x$, $g_3(x) = 0.75\tan^{-1}\{10(x - 0.6)\}$, $g_4(x) = 2.5\vartheta\{(x - 0.75)/0.8\} - 0.75$ with $\vartheta(x) = (1 - x^2)^4\mathbf{1}(|x| \leqslant 1)$ and $g_5(x) = 1.75\tan^{-1}\{5(x - 0.6)\} + 0.75$. Fig. 2 provides a plot of these functions, which are chosen to approximate roughly the shapes of the estimates $\hat{g}_1, \ldots, \hat{g}_5$ in the application later.

The model errors ε_{it} are independent draws from a normal distribution with mean 0 and standard deviation 1.3, which matches the average standard deviation of the estimated residuals in the application. Moreover, the regressors X_{it} are drawn independently from a uniform distribution with support $[0, 1]$, taking into account that the regressors in the application are supported on $[0, 1]$ as well. As can be seen, there is no time series dependence in the error terms and the regressors, and we do not include fixed effects α_i and γ_t in the error structure. We do not take into account these complications in our simulation design because their effect on the results is obvious: the stronger the time series dependence in the model variables and the more noise we add in terms of the fixed effects, the more difficult it becomes to estimate the curves m_i and thus to infer the unknown group structure from the simulated data.

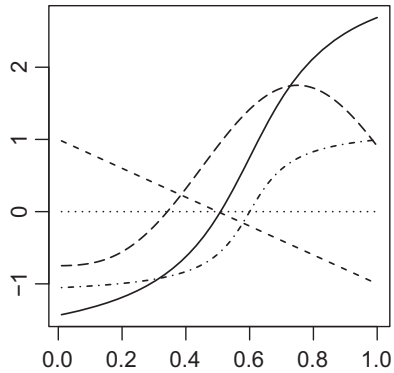

Fig. 2. Plot of the functions g_k for $1 \leqslant k \leqslant 5$: · · · · · ·, g_1; - - - - - -, g_2; - · - · - ·, g_3; – – –, g_4; ———, g_5

To implement our thresholding procedure, we compute the threshold level $\tau_{n,T}$ as described in Section 4.1, we pick the indices i_k according to rule 1 from Section 4.2 and work with scaled estimators of the L_2-distances Δ_{ij} as defined in Section 4.5. To compute the Nadaraya–Watson smoothers $\hat{m}_i$, we employ an Epanechnikov kernel and the bandwidth $h = 0.25$ throughout the simulations. As a check of robustness, we have repeated the simulations for various other bandwidths. As this yields very similar results, we, however, do not report them here. We do not use the bandwidth selection rule from Section 4.3 but work with the fixed bandwidth $h = 0.25$, since we focus on the performance of our classification methods and do not want our analysis to be influenced by effects of the bandwidth selection procedure. Additional simulations on the small sample behaviour of the bandwidth selection rule from Section 4.3 can be found in the on-line supplementary material.

For each sample size (n, T) with $n = 120$ and $T \in \{100, 150, 200\}$, we draw $N = 1000$ samples from setting (5.1) and compute the threshold estimates $\{\hat{G}_k : 1 \leqslant k \leqslant \hat{K}\}$ as well as the k-means estimates $\{\hat{G}_k^{\mathrm{KM}} : 1 \leqslant k \leqslant \hat{K}\}$. To measure how well these estimates fit the real class structure $\{G_k : 1 \leqslant k \leqslant K\}$, we compute the number $\#F$ of wrongly classified indices i, which is defined as follows: let π be some permutation of the class labels $\{1, \ldots, \hat{K}\}$ and denote the set of all possible permutations by Π. Moreover, denote the group membership of index i by $\rho(i)$, i.e. set $\rho(i) = k$ if $i \in G_k$. Similarly, let $\hat{\rho}_\pi(i)$ be the estimated group membership of index i, where the estimated classes are labelled according to the permutation π. More specifically, set $\hat{\rho}_\pi(i) = \pi(k)$ if $i \in \hat{G}_k$ (or if $i \in \hat{G}_k^{\mathrm{KM}}$ when considering the k-means estimators). With this notation at hand, we define $\#F = \min_{\pi \in \Pi} \sum_{i=1}^{n} \mathbf{1}\{\rho(i) \neq \hat{\rho}_\pi(i)\}$. For each sample size (n, T), we obtain $N = 1000$ values of $\#F$ for both the threshold and the k-means estimators. Fig. 3 shows the distribution of these values. In particular, the bars in the plots give the number of simulations (out of a total of 1000) in which a certain number of wrong classifications is obtained.

We now have a closer look at the simulation results in Fig. 3. Figs 3(a)–3(c) show the distribution of the number $\#F$ of wrong classifications for the threshold estimators $\{\hat{G}_k : 1 \leqslant k \leqslant \hat{K}\}$. Overall, the estimates can be seen to approximate the group structure reasonably well. Their precision improves quickly as the sample size grows. At a sample size of $T = 200$, all indices are correctly classified in about 80% of the cases and there is only one wrongly classified index in most other cases. For $T = 150$, which is approximately equal to the time series length in the application, our thresholding procedure also produces accurate results in most simulations with only a few indices being wrongly classified. Finally, for $T = 100$, the procedure yields good results with at most five wrongly classified indices in about 70% of the cases. There is, however, a

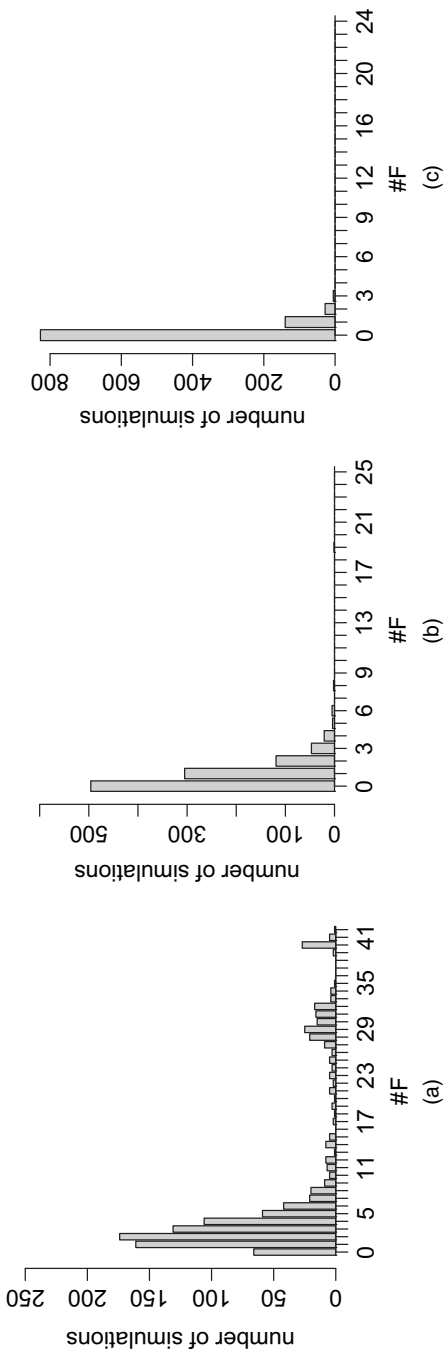

Fig. 3. Simulation results for the estimation of the classes $G_1, \ldots, G_5$: (a)–(c) distributions of the number $\#F$ of wrong classifications for the threshold estimators $\{\hat{G}_k : 1 \leqslant k \leqslant \hat{K}\}$ and the time series lengths (a) $T = 100$, (b) $T = 150$ and (c) $T = 200$; (d)–(f) corresponding distributions for the k-means estimators $\{\hat{G}_k^{\mathrm{KM}} : 1 \leqslant k \leqslant \hat{K}\}$ and the time series lengths (d) $T = 100$, (e) $T = 150$ and (f) $T = 200$

Table 1. Simulation results for the estimation of K†

$\hat{K}$	*Results for* $T = 100$	*Results for* $T = 150$	*Results for* $T = 200$
4	33	0	0
5	749	932	967
6	194	63	31
7	22	4	2
8	2	1	0

†The entries in the table specify the number of simulations in which a certain value of $\hat{K}$ is obtained.

substantial fraction of simulations in which many classification errors occur. This is not surprising since the time series length $T = 100$ is comparatively small given the noise level of the error terms. The fits $\hat{m}_i$ thus tend to be quite imprecise, which in turn leads to frequent classification errors.

Figs 3(d)–3(f) depict the distribution of $\#F$ for the k-means estimators $\{\hat{G}_k^{\mathrm{KM}} : 1 \leqslant k \leqslant \hat{K}\}$. As we can see, for the smallest sample size $T = 100$, i.e. when the signal-to-noise ratio is still quite low, the estimators $\{\hat{G}_k^{\mathrm{KM}} : 1 \leqslant k \leqslant \hat{K}\}$ strongly improve on the performance of the threshold estimators $\{\hat{G}_k : 1 \leqslant k \leqslant \hat{K}\}$. As already discussed in Section 2.3, we thus recommend refining our threshold estimators by an additional k-means clustering step when the noise level in the data is high. For $T = 150$, we still obtain quite a substantial improvement on the performance of the thresholding procedure, whereas, for the largest sample size $T = 200$, the additional gain from performing a k-means clustering step is comparatively small.

We finally turn to the finite sample performance of the estimator $\hat{K}$ which approximates the number of classes K. The simulation results are presented in Table 1. They suggest that the estimator $\hat{K}$ performs reasonably well in small samples. Already for the smallest time series length $T = 100$, it selects the true number of classes $K = 5$ in around 75% of the simulations. This value can be seen to improve to more than 95% as the sample size increases to $T = 200$.

6. Application

In 2007, the 'Markets in financial instruments directive' ended the monopoly of primary European stock exchanges. It paved the way for the emergence of various new trading platforms and brought about a strong fragmentation of the European stock market. Both policy makers and academic researchers aim to analyse and evaluate the effects of the directive; see for example UK Government Office for Science (2012). A particular interest lies in better understanding how trading venue fragmentation influences market quality. This question has been investigated with the help of parametric panel models in O'Hara and Ye (2009) and Degryse *et al.* (2014) among others. A semiparametric panel model with a factor structure has been employed in Boneva *et al.* (2015a).

In what follows, we use our modelling approach to gain further insights into the effect of fragmentation on market quality. We apply it to a large sample of volume and price data on the FTSE 100 and FTSE 250 stocks from May 2008 to June 2011. The volume data were supplied to us by Fidessa. The sample consists of weekly observations on the volume of all the FTSE stocks traded at various venues in the UK; see Boneva *et al.* (2015a, b) for a more detailed description of the data set. The price data were taken from Datastream and comprise the lowest

and the highest daily price of the various FTSE stocks. From these data, we calculate measures of fragmentation and market quality for all stocks in our sample on a weekly frequency. As a measure of fragmentation, we use the so-called Herfindahl index. The Herfindahl index of stock i is defined as the sum of the squared market shares of the venues where the stock is traded. It thus takes values between 0 and 1 or, more exactly, between $1/M$ and 1 with M being the number of trading venues. A value of $1/M$ indicates the perfect competition case where the stock is traded at equal shares at all existing venues. A value of 1 represents the monopoly case where the stock is traded at only one venue. As a measure of market quality, we employ volatility or, more specifically, the so-called high–low range, which is defined as the difference between the highest and the lowest price of the stock divided by the latter. To obtain volatility levels on a weekly frequency, we calculate the weekly median of the daily levels.

Denoting the Herfindahl index of stock i at time t by X_{it} and the corresponding logarithmic volatility level by Y_{it}, we model the relationship between Y_{it} and X_{it} by

$$Y_{it} = m_i(X_{it}) + u_{it}, \tag{6.1}$$

where the error term has the fixed effects structure $u_{it} = \alpha_i + \gamma_t + \varepsilon_{it}$. In this model, the function m_i captures the effect of fragmentation on market quality for stock i. This effect can be expected to differ across stocks. In particular, it is quite plausible to suppose that there are different groups of stocks for which the effect is fairly similar. We thus impose a group structure on the stocks in our sample: we suppose that there are K classes of stocks $G_1, \dots, G_K$ along with associated functions $g_1, \dots, g_K$ such that $m_i = g_k$ for all $i \in G_k$ and all $1 \leqslant k \leqslant K$. The effect of fragmentation on market quality is thus modelled to be the same within each group of stocks.

To determine the number of classes K and to estimate the groups G_k along with the functions g_k for $1 \leqslant k \leqslant K$, we use the estimation techniques that were developed in the previous sections. As the data are quite noisy, we refine our thresholding procedure by the additional k-means clustering step from Section 2.3. To implement the thresholding procedure, we compute the threshold parameter $\tau_{n,T}$ as explained in Section 4.1, we choose the indices i_k according to rule 1 from Section 4.2 and work with scaled estimators of the L_2-distances Δ_{ij} as described in Section 4.5. The Nadaraya–Watson smoothers $\hat{m}_i$ are based on an Epanechnikov kernel and their bandwidths are chosen as explained in Section 4.3. Before estimation, we drop stocks

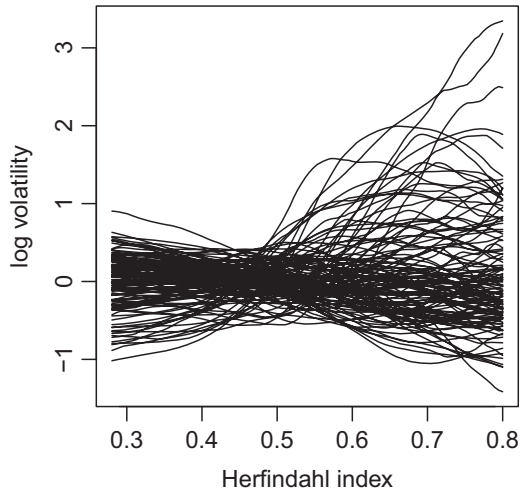

Fig. 4. Estimates $\hat{m}_i$ for the $n = 125$ stocks in our sample

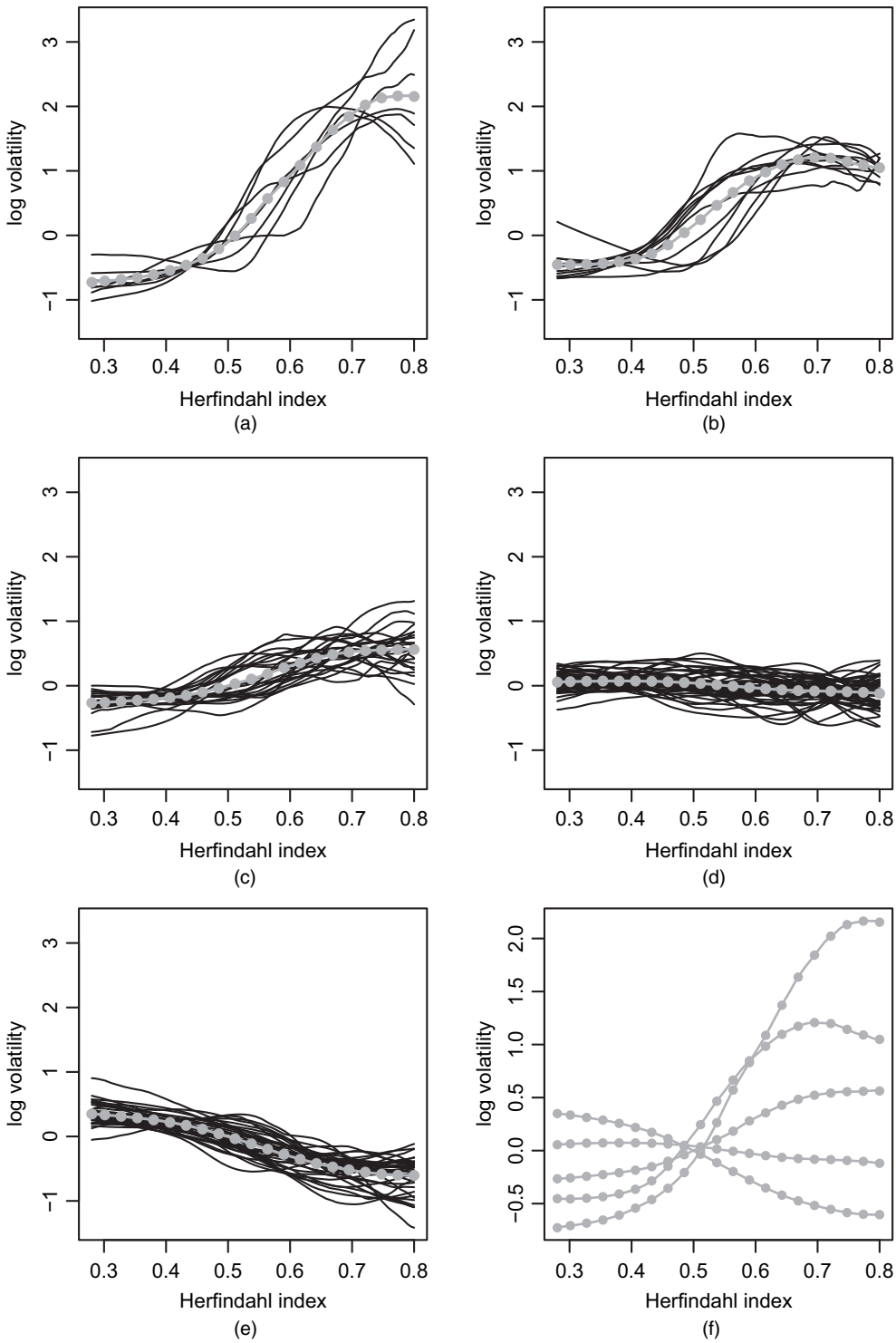

Fig. 5. Clusters of the curve estimates $\hat{m}_i$ (———, $\hat{m}_i$; ●, $\hat{g}_k$): (a) cluster I; (b) cluster II; (c) cluster III; (d) cluster IV; (e) cluster V; (f) estimates $\hat{g}_k$

from our sample for which we do not observe data over the whole time span from May 2008 to June 2011. Moreover, we eliminate stocks i with a very small empirical support $\mathcal{S}_i$ of the fragmentation data $\{X_{it} : 1 \leqslant t \leqslant T\}$. In particular, we take into account only stocks i for which the support $\mathcal{S}_i$ contains the interval $[0.275, 0.8]$. This leaves us with $n = 125$ stocks. The time series length amounts to $T = 151$ weeks.

We now turn to the estimation results. Fig. 4 depicts the smoothers $\hat{m}_i$ for the $n = 125$ stocks in our sample. Our thresholding procedure yields the estimate $\hat{K} = 5$, thus suggesting grouping the curves $\hat{m}_i$ into five clusters. The estimated clusters are shown in Fig. 5. In particular, each panel of Fig. 5 depicts the estimated curves which belong to a particular class $\hat{G}_k^{\mathrm{KM}}$. The corresponding estimates $\hat{g}_k$ of the group-specific regression functions are indicated by the grey dotted curves and are once again plotted together in Fig. 5(f).

Inspecting Fig. 5, the effect of fragmentation on (logarithmic) volatility appears to be quite moderate for a large number of stocks i: most of the curves in cluster IV are close to a flat line, which is reflected by the shape of the associated function $\hat{g}_4$. The fits of cluster V slightly slope downwards, indicating that the volatility level is a little lower in the monopoly case than under competition. Most of the fits in cluster III are moderately increasing, suggesting that the volatility is a little lower under competition. In contrast with the fits in clusters III, IV and V, those in clusters I and II exhibit a more pronounced effect of fragmentation on volatility: most of the fits substantially slope upwards, the increase being stronger in cluster I than in II. Regarding volatility as bad, the results of Fig. 5 can be interpreted as follows: for the stocks in clusters I, II and III, fragmentation leads to a decrease of volatility and thus to an improvement of market quality. For some stocks—specifically for those of cluster I—this improvement is quite substantial. For most of the stocks, however—in particular for those in clusters III, IV and V—the effect of fragmentation on volatility is fairly moderate and may go into both directions. In particular, fragmentation may either slightly improve (see cluster III) or deteriorate (see cluster V) market quality.

We briefly compare these findings with the empirical results in Boneva *et al.* (2015a). In contrast with our approach, they imposed the factor structure $m_i(x) = \Sigma_{k=1}^{K} \beta_{ik} \mu_k(x)$ on the regression curves. The functions μ_k in this model structure can be interpreted as common factors that are the same across individuals. The coefficient vectors $\beta_i = (\beta_{i1}, \ldots \beta_{iK})^{\mathrm{T}}$ assign different individual-specific weights to these factors. Applying their model to the data at hand, Boneva *et al.* (2015a) found evidence that market quality is better under competition than in the monopoly case. However, their results also reveal that the improvement is quite moderate. These findings are essentially in line with our own results. According to our results, the effect of fragmentation on market quality is quite moderate for the great bulk of stocks and competition substantially improves market quality only for a small fraction of stocks. This translates into a moderate positive effect of fragmentation on market quality when working with the factor structure of Boneva *et al.* (2015a).

7. Supplementary material

In the on-line supplement, we examine the bandwidth selection rule from Section 4.3 by means of a simulation study. Moreover, we provide the proofs of theorems 1–3.

Acknowledgements

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

References

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

Baltagi, B. H. (2013) *Econometric Analysis of Panel Data*. Hoboken: Wiley.

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

Boneva, L., Linton, O. and Vogt, M. (2015b) The effect of fragmentation in trading on market quality in the UK equity market. *J. Appl. Econmetr.*, to be published.

Chiou, J.-M. and Li, P.-L. (2007) Functional clustering and identifying substructures of longitudinal data. *J. R. Statist. Soc.* B, **69**, 679–699.

Cox, D. R. (1957) Note on grouping. *J. Am. Statist. Ass.*, **52**, 543–547.

Degryse, H., De Jong, F. and Van Kervel, V. (2014) The impact of dark trading and visible fragmentation on market quality. *Rev. Finan.*, **19**, 1587–1622.

Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and Its Applications*. London: Chapman and Hall.

Fisher, D. (1958) On grouping for maximum homogeneity. *J. Am. Statist. Ass.*, **53**, 789–798.

Garcia-Escudero, L. A. and Gordaliza, A. (1999) Robustness of properties of k-means and trimmed k-means. *J. Am. Statist. Ass.*, **94**, 956–969.

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

Härdle, W., Hall, P. and Marron, J. S. (1988) How far are automatically chosen regression smoothing parameters from their optimum? *J. Am. Statist. Ass.*, **83**, 86–95.

Härdle, W. and Mammen, E. (1993) Comparing nonparametric versus parametric regression fits. *Ann. Statist.*, **21**, 1926–1947.

Henderson, D. J., Carroll, R. J. and Li, Q. (2008) Nonparametric estimation and testing of fixed effects panel data models. *J. Econmetr.*, **144**, 257–275.

Hsiao, C. (2003) *Analysis of Panel Data*. Cambridge: Cambridge University Press.

Ieva, F., Paganoni, A. M., Pigoli, D. and Vitelli, V. (2013) Multivariate functional clustering for the morphological analysis of electrocardiograph curves. *Appl. Statist.*, **62**, 401–418.

Jacques, J. and Preda, C. (2014) Functional data clustering: a survey. *Adv. Data Anal. Classificn*, **8**, 231–255.

James, M. and Sugar, C. A. (2003) Clustering for sparsely sampled functional data. *J. Am. Statist. Ass.*, **98**, 397–408.

Mammen, E., Støve, B. and Tjøstheim, D. (2009) Nonparametric additive models for panels of time series. *Econmetr. Theor.*, **25**, 442–481.

O'Hara, M. and Ye, M. (2009) Is fragmentation harming market quality? *J. Finan. Econ.*, **100**, 459–474.

Pollard, D. (1981) Strong consistency of k-means clustering. *Ann. Statist.*, **9**, 135–140.

Pollard, D. (1982) A central limit theorem for k-means clustering. *Ann. Probab.*, **10**, 919–926.

Rao, P. S. R. S. (1997) *Variance Components Estimation: Mixed Models, Methodologies and Applications*. New York: Chapman and Hall.

Ray, S. and Mallick, B. (2006) Functional clustering by Bayesian wavelet methods. *J. R. Statist. Soc.* B, **68**, 305–332.

Ruckstuhl, A. F., Welsh, A. H. and Carroll, R. J. (2000) Nonparametric function estimation of the relationship between two repeatedly measured variables. *Statist. Sin.*, **10**, 51–71.

Su, L., Shi, Z. and Phillips, P. C. B. (2014) Identifying latent structures in panel data. *Preprint*. Singapore Management University. (Available from `http://dx.doi.org/10.2139/ssrn.2448189`.)

Sun, W., Wang, J. and Fang, Y. (2012) Regularized k-means clustering of high-dimensional data and its asymptotic consistency. *Electron. J. Statist.*, **6**, 148–167.

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

UK Government Office for Science (2012) Future of computer trading in financial markets. UK Government Office for Science, London. (Available from `www.gov.uk/government/publications/future-of-computer-trading-in-financial-markets-an-international-perspective`.)

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplement to "Classification of non-parametric regression functions in longitudinal data models"'.