Classification of Nonparametric Regression Functions in Longitudinal Data Models

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

August 20, 2015

We investigate a longitudinal data model with nonparametric 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.

Key words: Classification of regression curves; kernel estimation; nonparametric regression; longitudinal/panel data.
AMS 2010 subject classifications: 62G08; 62G20; 62H30.

1 Introduction

Non- and semiparametric regression models are a flexible framework to analyze 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 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 func-

[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.

tion. 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 setup. We observe a sample of longitudinal or panel data $\{(Y_{it}, X_{it}) : 1 \le i \le n, 1 \le t \le T\}$, where i denotes the i-th 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 infinity. The cross-section dimension n, in contrast, may either be fixed or diverging. The data are supposed to come from the nonparametric regression model

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

where m_i are unknown nonparametric 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 $\{1, \ldots, n\}$, that is, $G_1 \,\dot{\cup} \ldots \dot{\cup}\, G_K = \{1, \ldots, n\}$. We suppose that for each $k \in \{1, \ldots, K\}$,

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

Hence, the members of the class G_k all have the same regression function, which we denote by g_k in what follows. Note that 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 (1.1) are supposed to have the structure $u_{it} = \alpha_i + \gamma_t + \varepsilon_{it}$. 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 makeup 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 $\mathcal{X}_{n,T} = \{X_{it} : 1 \le i \le n, 1 \le t \le T\}$, we allow that $\mathbb{E}[\alpha_i|\mathcal{X}_{n,T}] \ne 0$ and $\mathbb{E}[\gamma_t|\mathcal{X}_{n,T}] \ne 0$. 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}. In the econometrics literature, the error structure $u_{it} = \alpha_i + \gamma_t + \varepsilon_{it}$ is very common; see e.g. the books by 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 imposed in (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 different venues. This restructuring of the European stock market has raised the question how competition between trading venues, that is, trading venue fragmentation affects the quality of the market from the point of view of the typical trader. 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 empirically investigate the effect of fragmentation on market quality. In Section 6, we apply it to a sample of data for the FTSE 100 and FTSE 250 stocks.

To the best of our knowledge, the problem of classifying nonparametric 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. Similarly as in our nonparametric 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 $1 \le k \le K$. The problem of estimating the unknown groups $G_1, \ldots, G_K$ 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 the latter case, they have to be estimated from noisy observations $Y_1, \ldots, Y_n$ which are realizations of a process $Y = \{Y(t) = X(t) + \varepsilon(t) : t \in \mathcal{T}\}$ with ε being the noise process. 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 papers which deal with this problem in different model setups; 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 i.i.d. across observations (if there is noise at all). Finally, whereas the number of observed curves n should diverge to infinity in functional data models, we provide theory both for fixed and diverging n. For these reasons, substantially different theoretical arguments are required to analyze 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 of 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.

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 carrying 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 $G_1, \ldots, G_K$ and of their number K. These are finally used to come up with estimators of the functions $g_1, \ldots, g_K$.

2.1 Estimation of the regression functions m_i

To construct an estimator $\widehat{m}_i$ of the regression function m_i of the i-th 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 nonparametric methods. In particular, we could employ a Nadaraya-Watson smoother of the form

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

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 $\widehat{Y}_{it}^{\text{fe}} = Y_{it} - \overline{Y}_i - \overline{Y}_t^{(i)} + \overline{\overline{Y}}^{(i)}$, where

$$\overline{Y}_i = \frac{1}{T} \sum_{t=1}^T Y_{it}, \qquad \overline{Y}_t^{(i)} = \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^n Y_{jt}, \qquad \overline{\overline{Y}}^{(i)} = \frac{1}{(n-1)T} \sum_{\substack{j=1 \\ j \neq i}}^n \sum_{t=1}^T Y_{jt}$$

are sample averages of the Y-observations. In the definition of $\overline{Y}_t^{(i)}$ and $\overline{\overline{Y}}^{(i)}$, we leave out the i-th observation to avoid some bias terms that are particularly problematic when n is fixed. With this notation at hand, we define the feasible estimator

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

of the function m_i. For simplicity, we use the same bandwidth h for all estimators $\widehat{m}_i$. 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 fulfill the conditions on the common bandwidth h summarized in (C4) of Section 3.1). Alternatively to the Nadaraya-Watson smoothers $\widehat{m}_i$, we could work with local linear or more generally local polynomial estimators. Indeed, our procedure to estimate the groups G_k and the functions g_k for $1 \leq k \leq K$ is the same no matter which type of kernel smoother we employ.

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 $\Delta_{ij} =$

$\int(m_i(x) - m_j(x))^2\pi(x)dx$, where π is some weight function. These are estimated by

$$\widehat{\Delta}_{ij} = \int \big(\widehat{m}_i(x) - \widehat{m}_j(x)\big)^2\pi(x)dx,$$

where $\widehat{m}_i$ and $\widehat{m}_j$ are the kernel smoothers introduced in the previous section. We now sort the distances $\{\Delta_{ij} : j \in S\}$ along with their estimates $\{\widehat{\Delta}_{ij} : j \in S\}$ in increasing order. Denote the ordered distances by

$$\Delta_{i(1)} \le \Delta_{i(2)} \le \ldots \le \Delta_{i(n_S)} \quad \text{and} \quad \widehat{\Delta}_{i[1]} \le \widehat{\Delta}_{i[2]} \le \ldots \le \widehat{\Delta}_{i[n_S]},$$

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 exists a point $p = p_{i,S}$ such that

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

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

$$\widehat{\Delta}_{i[j]} \begin{cases} = o_p(1) & \text{for } j \le p \\ \ge c + o_p(1) & \text{for } j > p \end{cases} \tag{2.1}$$

with $c = \Delta_{i(p+1)} > 0$. This in particular says that the first p order statistics $\widehat{\Delta}_{i[1]}, \ldots$ $\ldots, \widehat{\Delta}_{i[p]}$ 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 one. Hence, if we knew the size $p = |G|$ of the class G, we could simply estimate $G = \{(1), \ldots, (p)\}$ by $\widetilde{G} = \{[1], \ldots, [p]\}$.

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

$$\max_{j \in G} \widehat{\Delta}_{ij} \le \tau_{n,T} \text{ with prob. approaching 1,} \tag{2.2}$$

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

$$\widehat{\Delta}_{i[j]} \begin{cases} \le \tau_{n,T} & \text{for } j \le p \\ > \tau_{n,T} & \text{for } j > p \end{cases} \quad \text{with prob. approaching 1.}$$

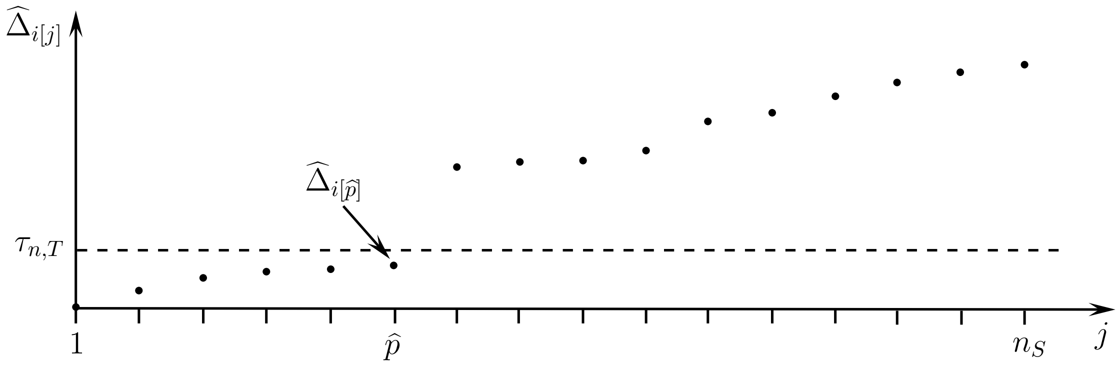

Figure 1: Graphical illustration of the procedure underlying the estimator $\widehat{G} = \{[1], \ldots, [\widehat{p}]\}$. The black dots indicate the ordered estimated distances $\widehat{\Delta}_{i[1]}, \ldots, \widehat{\Delta}_{i[n_S]}$, the dashed line represents the threshold level $\tau_{n,T}$.

This suggests to estimate $p = p_{i,S}$ by

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

and to define our estimator of G by $\widehat{G} = \{[1], \ldots, [\widehat{p}]\}$. Figure 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 $\{G_k : 1 \le k \le K\}$:

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

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

We iterate this algorithm $\widehat{K}$ times until $\widehat{\Delta}_{i_{\widehat{K}}[j]} \le \tau_{n,T}$ for all $j \in S_{\widehat{K}}$, that is, until our thresholding rule suggests that all indices in $S_{\widehat{K}}$ belong to the same class. In this case, $S_{\widehat{K}}$ is not split into two parts any more and $\widehat{G}_{\widehat{K}} = S_{\widehat{K}}$. Our algorithm thus produces the partition $\{\widehat{G}_k : 1 \le k \le \widehat{K}\}$, which serves as our estimator of the class structure $\{G_k : 1 \le k \le K\}$. Importantly, the algorithm does not only estimate the classes $G_1, \ldots, G_K$ but also their number K. In particular, K is implicitly estimated by the number of iterations $\widehat{K}$. 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 $\widehat{G}_1, \ldots, \widehat{G}_{\widehat{K}}$ 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 $\widehat{m}_i$ 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 $\widehat{G}_1, \ldots, \widehat{G}_{\widehat{K}}$ 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 e.g. 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 have to 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 simultaneously satisfy (a) and (b): 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 $\widehat{g}_k^{[1]}(x) = |\widehat{G}_k|^{-1} \sum_{i \in \widehat{G}_k} \widehat{m}_i(x)$ for each class estimate $\widehat{G}_k$ with $1 \le k \le \widehat{K}$. Defining $\Delta(q_1, q_2) = \int (q_1(x) - q_2(x))^2 \pi(x) dx$ to be the squared L_2-distance between two functions $q_\ell : [0,1] \to \mathbb{R}$ with $\ell = 1, 2$, we then proceed as follows:

1st Step: For each $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, \widehat{K}\}$, calculate the distance $\widehat{d}_k(i) = \Delta(\widehat{m}_i, \widehat{g}_k^{[1]})$ between the function $\widehat{m}_i$ and the cluster mean $\widehat{g}_k^{[1]}$. Define the classes $\{G_k^{[1]} : 1 \le k \le \widehat{K}\}$ by assigning the index i to the k-th class $G_k^{[1]}$ if $\widehat{d}_k(i) = \min_{1 \le k' \le \widehat{K}} \widehat{d}_{k'}(i)$.

rth Step: Let $\{G_k^{[r-1]} : 1 \le k \le \widehat{K}\}$ be the classes from the previous iteration step. Calculate mean functions $\widehat{g}_k^{[r]} = |G_k^{[r-1]}|^{-1} \sum_{i \in G_k^{[r-1]}} \widehat{m}_i$ and compute the distances $\widehat{d}_k(i) = \Delta(\widehat{m}_i, \widehat{g}_k^{[r]})$ for each i and k. Define the new classes $\{G_k^{[r]} : 1 \le k \le \widehat{K}\}$ by assigning the index i to the k-th group $G_k^{[r]}$ if $\widehat{d}_k(i) = \min_{1 \le k' \le \widehat{K}} \widehat{d}_{k'}(i)$.

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 $\{G_k : 1 \le k \le K\}$ which are denoted by $\{\widehat{G}_k^{\text{KM}} : 1 \le k \le \widehat{K}\}$ 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

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

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

3 Asymptotics

In this section, we investigate the asymptotic properties of our estimators. We first list the assumptions needed for the analysis and then summarize the main results. The proofs can be found in the Appendix.

3.1 Assumptions

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

(C2) The functions g_k $(1 \le k \le K)$ are twice continuously differentiable. The densities f_i of the variables X_{it} exist and have bounded support, which w.l.o.g. equals $[0,1]$. They are uniformly bounded away from zero and infinity, that is, $0 < c \le \min_{1 \le i \le n} \inf_{x \in [0,1]} f_i(x)$ and $\max_{1 \le i \le n} \sup_{x \in [0,1]} f_i(x) \le C < \infty$ for some constants $0 < c \le C < \infty$. Moreover, they are twice continuously differentiable on $[0,1]$ with uniformly bounded first and second derivatives. Finally, the joint densities $f_{i,\ell}$ of $(X_{it}, X_{it+\ell})$ exist and are uniformly bounded away from infinity.

(C3) There exist a real number $\theta > 4$ and a natural number ℓ^* such that for any $\ell \in \mathbb{Z}$ with $|\ell| \ge \ell^*$ and a fixed constant $C < \infty$,

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

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

(C4) The time series dimension T tends to infinity, while the cross-section dimension n may either be fixed or diverging. Their relative growth is such that $n/T \le C$ for some constant $C < \infty$. The bandwidth h has the property that $cT^{-2/5+\delta} \le h \le CT^{-\delta}$ for some small $\delta > 0$ and positive constants c, C.

(C5) The kernel W is non-negative and bounded. Moreover, it is symmetric about zero, has compact support (say $[-C_1, C_1]$), and fulfills the Lipschitz condition that there exists a positive constant L with $|W(x) - W(x')| \le L|x - x'|$. We use the notation $\|W\|^2 = \int W^2(x)dx$ and $\|W * W\|^2 = \int (\int W(x)W(x+y)dx)^2 dy$.

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., $\text{supp}(\pi) \subseteq [0,1]$.

We briefly comment on the above assumptions. First of all, note that we do not necessarily require exponentially decaying mixing rates as assumed in (C1). These could alternatively be replaced by sufficiently high polynomial rates. We nevertheless make the stronger assumption of exponential mixing to keep the notation and structure of the proofs as clear as possible. (C2) and (C3) 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; cp. for example Hansen (2008) for similar assumptions. (C4) 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 (C3). In particular, it is possible to relax (C4) at the cost of a stronger moment condition. For example, we can weaken (C4) to allow for $n/T^{3/2} \le C$, 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 $1 \le k \le K$. They only need to fulfill the trivial conditions that $n_k \le n$ for $1 \le k \le K$ and $\sum_{k=1}^K n_k = n$. 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 infinity at different rates for others.

3.2 Main results

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

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

This condition is in particular satisfied by any threshold $\tau_{n,T}$ which converges to zero more slowly than $\max_{i,j \in G_k} \widehat{\Delta}_{ij}$ for $1 \le k \le K$. More formally, suppose that $\max_{i,j \in G_k}$

$\widehat{\Delta}_{ij} = O_p(c_{n,T})$ 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 (C_τ). In the Appendix, we show that $\max_{i,j \in G_k} \widehat{\Delta}_{ij} = O_p(c_{n,T})$ with $c_{n,T} = T^{-1/5} + h^3$ under the conditions (C1)–(C5) and $c_{n,T} = \log T/(Th) + h^3$ provided that the moment assumptions in (C3) are strengthened to hold for some $\theta > 20/3$. Notably, these are only upper bounds on the rate of $\max_{i,j \in G_k} \widehat{\Delta}_{ij}$. In Lemma A.2 in the Appendix, we derive the sharp rate $\max_{i,j \in G_k} \widehat{\Delta}_{ij} = O_p(1/(Th))$ under more restrictive conditions than (C1)–(C5). This lemma also provides us with a more concise characterization of the threshold sequences that satisfy (C_τ). It shows that $\max_{i,j \in G_k} \widehat{\Delta}_{ij} \le b_{n,T} + \rho_{n,T}$, where the leading term $b_{n,T}$ has the form

$$b_{n,T} = \frac{\|W\|^2 \max_{1 \le i < j \le n}(b_i + b_j)}{Th}$$

with $b_i = \int \sigma_i^2(x)\pi(x)/f_i(x)dx$ 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 $\tau_{n,T} \ge b_{n,T} + \overline{\rho}_{n,T}$ fulfills (C_τ), where $\overline{\rho}_{n,T}$ is an upper bound on the lower order terms $\rho_{n,T}$ satisfying $\overline{\rho}_{n,T}/\rho_{n,T} \to \infty$.

Our first result shows that the threshold estimators $\{\widehat{G}_k : 1 \le k \le \widehat{K}\}$ are consistent in the following sense: they coincide with the true classes $\{G_k : 1 \le k \le K\}$ with probability tending to one, provided that the threshold parameter $\tau_{n,T}$ fulfills (C_τ).

Theorem 3.1. *Let (C1)–(C5) be satisfied and suppose that $\tau_{n,T}$ fulfills (C_τ). Then $\mathbb{P}(\widehat{K} \ne K) = o(1)$ and*

$$\mathbb{P}\Big(\big\{\widehat{G}_k : 1 \le k \le \widehat{K}\big\} \ne \big\{G_k : 1 \le k \le K\big\}\Big) = o(1).$$

Note that the indexing of the estimators $\widehat{G}_1, \ldots, \widehat{G}_{\widehat{K}}$ is completely arbitrary. We could, for example, change the indexing according to the rule $k \mapsto \widehat{K} - k + 1$. In the sequel, we suppose that the estimated classes are indexed such that $\mathbb{P}(\widehat{G}_k = G_k) \to 1$ for all k. Theorem 3.1 implies that this is possible without loss of generality. The next theorem shows that the k-means estimators $\{\widehat{G}_k^{\text{KM}} : 1 \le k \le \widehat{K}\}$ inherit the consistency property of Theorem 3.1 from the threshold estimators $\{\widehat{G}_k : 1 \le k \le \widehat{K}\}$.

Theorem 3.2. *Under the conditions of Theorem 3.1, it holds that*

$$\mathbb{P}\Big(\big\{\widehat{G}_k^{\text{KM}} : 1 \le k \le \widehat{K}\big\} \ne \big\{G_k : 1 \le k \le K\big\}\Big) = o(1).$$

As above, we suppose without loss of generality that the k-means estimators $\widehat{G}_1^{\text{KM}}, \ldots$ $\ldots, \widehat{G}_K^{\text{KM}}$ are indexed such that $\mathbb{P}(\widehat{G}_k^{\text{KM}} = G_k) \to 1$ for all k.

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

$$\begin{aligned} B_k(x) &= \frac{c_k^{5/2}}{2}\big(\textstyle\int W(\varphi)\varphi^2 d\varphi\big) \lim_{n\to\infty} \Big(\frac{1}{n_k}\sum_{i\in G_k} \frac{g_k''(x)f_i(x)+2g_k'(x)f_i'(x)}{f_i(x)}\Big) \\ V_k(x) &= \big(\textstyle\int W^2(\varphi)d\varphi\big) \lim_{n\to\infty} \Big(\frac{1}{n_k}\sum_{i\in G_k} \frac{\sigma_i^2(x)}{f_i(x)}\Big), \end{aligned}$$

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 convergence rate and the limit distribution of $\widehat{g}_k$.

Theorem 3.3. *Let the conditions of Theorem 3.1 be satisfied. Then for any fixed $x \in (0,1)$,*

$$\widehat{g}_k(x) - g_k(x) = O_p\Big(\frac{1}{\sqrt{n_k T h}}\Big). \tag{3.1}$$

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

$$\sqrt{\widehat{n}_k T h}\big(\widehat{g}_k(x) - g_k(x)\big) \xrightarrow{d} N\big(B_k(x), V_k(x)\big). \tag{3.2}$$

When deriving the limit distribution in (3.2), we restrict attention to the case that $n \to \infty$ for the following reason: If n is finite, the estimation error in $\widehat{Y}_{it}^{\text{fe}}$ induced by subtracting the sample averages $\overline{Y}_i$, $\overline{Y}_t^{(i)}$ and $\overline{\overline{Y}}^{(i)}$ 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.

In addition to the pointwise rate in (3.1), it is possible to derive results on the uniform convergence behaviour of $\widehat{g}_k$: Lemma A.1 from the Appendix directly implies that under (C1)–(C5), $\sup_{x\in[0,1]} |\widehat{g}_k(x) - g_k(x)| = o_p(1)$. To derive the exact rate at which $\widehat{g}_k$ uniformly converges to g_k, we essentially have to compute the uniform rate of an average of kernel smoothers. This can be achieved by following the usual strategy to derive uniform convergence rates for kernel estimators; see for example Masry (1996), Bosq (1998) or Hansen (2008). For the case that $n_k/n \to c > 0$ and that the bandwidth h is of the order $(nT)^{-(1/5+\delta)}$ for some small $\delta > 0$, this has been done in Boneva et al. (2014b). Their results immediately imply that in this case, $\sup_{x\in I_h} |\widehat{g}_k(x) - g_k(x)| = O_p(\sqrt{\log(n_k T)/(n_k T h)})$, where $I_h = [C_1 h, 1 - C_1 h]$ is the interior of the support of the regressors. By fairly straightforward modifications of these results, it is possible to derive this uniform rate under more general conditions on the size of n_k.

4 Implementation

Our thresholding approach to estimate the class structure $\{G_k : 1 \le k \le K\}$ depends on two tuning parameters: the threshold level $\tau_{n,T}$ and the bandwidth h of the kernel smoothers $\widehat{m}_i$. 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 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 $\tau_{n,T}$ slightly larger than $\max_{j \in G} \widehat{\Delta}_{ij}$. 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. Writing $G_{-i} = G \setminus \{i\}$, we can modify the arguments from Härdle and Mammen (1993) to show that for any $j \in G_{-i}$,

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

under slightly strengthened versions of the conditions (C1)–(C5). The bias and variance expressions in (4.1) are of the form

$$\mathcal{B}_{ij} = \|W\|^2(b_i + b_j) \qquad \text{and} \qquad \mathcal{V}_{ij} = \|W * W\|^2(2v_{ii} + 4v_{ij} + 2v_{jj}),$$

where $\|W\|^2$ and $\|W * W\|^2$ are defined in (C5), $b_i = \int \sigma_i^2(x)\pi(x)/f_i(x)dx$ and $v_{ij} = \int \sigma_i^2(x)\sigma_j^2(x)\pi^2(x)/(f_i(x)f_j(x))dx$ with $\sigma_i^2(x) = \mathbb{E}[\varepsilon_{it}^2|X_{it} = x]$. Roughly speaking, (4.1) says that

$$\widehat{\Delta}_{ij} \approx \Delta_{ij}^* := \frac{\mathcal{B}_{ij}}{Th} + \frac{1}{Th^{1/2}}Z_{ij},$$

where Z_{ij} is Gaussian with mean zero and variance $\mathcal{V}_{ij}$. As a consequence, it holds that $\max_{j \in G} \widehat{\Delta}_{ij} = \max_{j \in G_{-i}} \widehat{\Delta}_{ij} \approx \max_{j \in G_{-i}} \Delta_{ij}^*$ and

$$\max_{j \in G_{-i}} \Delta_{ij}^* \le \max_{j \in G_{-i}} \frac{\mathcal{B}_{ij}}{Th} + \max_{j \in G_{-i}} \frac{Z_{ij}}{Th^{1/2}} = \max_{j \in G_{-i}} \frac{\mathcal{B}_{ij}}{Th} + \max_{j \in G_{-i}} \frac{\sqrt{\mathcal{V}_{ij}}}{Th^{1/2}} Z_{ij}^\circ$$

with $Z_{ij}^\circ = Z_{ij}/\sqrt{\mathcal{V}_{ij}}$. Since a standard normal random variable Z has the property that $\mathbb{P}(Z \ge z) \le (2\pi z^2)^{-1/2}\exp(-z^2/2)$ for $z > 0$, we obtain that

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

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

$$\begin{aligned}\max_{j \in G} \widehat{\Delta}_{ij} = \max_{j \in G_{-i}} \widehat{\Delta}_{ij} \approx \max_{j \in G_{-i}} \Delta^*_{ij} &\leq \max_{j \in G_{-i}} \frac{\mathcal{B}_{ij}}{Th} + \max_{j \in G_{-i}} \frac{\sqrt{\mathcal{V}_{ij}}}{Th^{1/2}} (2 \log |G|)^{1/2} \\ &\leq b_{n,T} + v_{n,T} (2 \log |G|)^{1/2},\end{aligned}$$

where $b_{n,T} = \max_{1 \leq i < j \leq n} \mathcal{B}_{ij}/(Th)$, $v_{n,T} = \max_{1 \leq i < j \leq n} \sqrt{\mathcal{V}_{ij}}/(Th^{1/2})$ and the first inequality holds with probability approaching one as $|G|$ tends to infinity. These considerations suggest that an appropriate threshold level is given by

$$\tau_{n,T} = b_{n,T} + v_{n,T} (2 \log |G|)^{1/2}. \tag{4.2}$$

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 A.2, we can work with threshold sequences of the form $\tau_{n,T} \geq b_{n,T} + \overline{\rho}_{n,T}$, where $b_{n,T}$ is the leading term. The threshold defined in (4.2) 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 choice of the bound $\overline{\rho}_{n,T}$ on the lower order terms.

Of course, the threshold level in (4.2) 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 get rid of the dependence on G, we may replace the unknown class size $|G|$ by the trivial bound n. This leads to the threshold level

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

with $p = n$. As n is a quite 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 we are in the k-th iteration step and let $\widehat{G}_1, \ldots, \widehat{G}_{k-1}$ be the estimated classes from the previous steps. Defining $\widehat{n}_\ell = |\widehat{G}_\ell|$, we set $p = n - \sum_{\ell=1}^{k-1} \widehat{n}_\ell$ and use the threshold $\tau_{n,T}(p)$ to estimate G_k. We thus exploit the information from the previous iteration steps to get a better bound on the class size $|G_k|$. It is straightforward to show that our theoretical results from Section 3.2 remain to hold true when we proceed in this way.

To compute the threshold (4.3) 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 $\widehat{\sigma}_i^2(x) = (Th)^{-1} \sum_{t=1}^T W_h(X_{it} - x) \widehat{\varepsilon}_{it}^2 / \widehat{f}_i(x)$, where $\widehat{f}_i(x) = (Th)^{-1} \sum_{t=1}^T W_h(X_{it} - x)$ and $\widehat{\varepsilon}_{it} = Y_{it} - \widehat{m}_i(X_{it})$ are the estimated residuals. Moreover, we may estimate f_i by the modified kernel density $\widehat{f}_i^{\text{bc}}(x) = (\int_{-x/h}^{(1-x)/h} W(\varphi) d\varphi)^{-1} \widehat{f}_i(x)$, where the correction $(\int_{-x/h}^{(1-x)/h} W(\varphi) d\varphi)^{-1}$ 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: (i) 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 to replace the maxima over $\mathcal{B}_{ij}$ and $\mathcal{V}_{ij}$ by a high quantile, say the 95%-quantile. (ii) 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 $\widehat{\sigma}_i^2$ 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 to replace the estimates $\widehat{\sigma}_i^2$ by the simple averages $\overline{\varepsilon}_i^2 := T^{-1}\sum_{t=1}^T \widehat{\varepsilon}_{it}^2$, 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 homoskedastic and thus $\mathbb{E}[\varepsilon_{it}^2|X_{it} = x] = \mathbb{E}[\varepsilon_{it}^2]$. However, the error resulting from replacing $\widehat{\sigma}_i^2$ with $\overline{\varepsilon}_i^2$ can be expected to be much lower than the error stemming from the unstabilities of the estimates $\widehat{\sigma}_i^2$. Both in the simulations and the application, we work with the modifications (i) and (ii).

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 (more or less) 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 $b_{n,T} = 2(Th)^{-1}\|W\|^2 \int \sigma^2(x)\pi(x)/f(x)dx$ and $v_{n,T} = 8(Th^{1/2})^{-1}\|W * W\|^2 \int \sigma^4(x)\pi^2(x)/f^2(x)dx$. 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 $\{\widehat{G}_k : 1 \le k \le \widehat{K}\}$, we need to pick an index i_k from the index set $S_k = \{1,\dots,n\} \setminus \bigcup_{\ell=1}^{k-1} \widehat{G}_\ell$ 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:

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

The heuristic idea behind this rule is as follows: $\widehat{p}_{i,S_k}$ is the position where the ordered estimates $\widehat{\Delta}_{i[1]},\dots,\widehat{\Delta}_{i[n_k]}$ exceed the threshold value $\tau_{n,T}$. Put differently, $\widehat{p}_{i,S_k}$ estimates the position where the ordered distances $\Delta_{i(1)},\dots,\Delta_{i(n_k)}$ jump from zero to a positive value. The rule (R) suggests to pick the index i for which the estimated jump size is largest, that is, for which the jump is most clearly visible in the data. Moreover, the rule is constructed such that we pick an index i with $\widehat{p}_{i,S_k} = n_k$ as soon as such an index occurs. The rationale behind this is the following: If $\widehat{p}_{i,S_k} = n_k$, then all distances $\widehat{\Delta}_{i[1]}, \ldots, \widehat{\Delta}_{i[n_k]}$ are smaller than the threshold $\tau_{n,T}$, indicating 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 $\widehat{K}$ from strongly overshooting the true number of classes K.

The rule (R) requires us to compute the positions 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 to use the rule (R) 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 $\widehat{m}_i$

When deriving our estimation methods, we have implicitly assumed that the smoothers $\widehat{m}_i$ 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 the cross-section dimension n is large. For this reason, we suggest to choose group-specific bandwidths: For each group G_k, we select a bandwidth h_k which is used to compute the estimators $\widehat{m}_i = \widehat{m}_{i,h_k}$ 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 $\mathcal{Z}_i = \{(Y_{it}, X_{it}) : 1 \le t \le T\}$ does not differ too much within groups. Technically speaking, we suppose that not only the functions m_i are 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 we want to optimize the bandwidth h of the Nadaraya-Watson estimator $\widehat{m}_i = \widehat{m}_{i,h}$ 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

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

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

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

but they are not identical with it. Indeed, we cannot minimize $\text{RSS}_i(h)$ directly but have to 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 $\widehat{m}_{i,h}(X_{it})$ for $t = 1, \ldots, T$. Since the observation Y_{it} is contained in the estimate $\widehat{m}_{i,h}(X_{it})$, it is used to predict itself. This creates a bias term which prevents the minimizer of $\text{RSS}_i(h)$ to be a reasonable approximation of h_i^{opt}. Formally speaking, it holds that

$$\begin{aligned}\mathbb{E}\big[\text{RSS}_i(h)\big] &= \mathbb{E}\big[\text{ASE}_i(h)\big] + \frac{1}{T}\sum_{t=1}^{T}\sigma_i^2\mathbb{E}[w(X_{it})] - \frac{2}{T^2h}\sum_{t=1}^{T}\sigma_i^2 W(0)\mathbb{E}\Big[\frac{w(X_{it})}{\widehat{f}_i(X_{it})}\Big] \\ &=: \mathbb{E}\big[\text{ASE}_i(h)\big] + B_{i,1}(h) + B_{i,2}(h)\end{aligned}$$

with $\sigma_i^2 = \mathbb{E}[\varepsilon_{it}^2]$ and $\widehat{f}_i(x) = T^{-1}\sum_{t=1}^{T} W_h(X_{it} - x)$. 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 e.g. Härdle et al. (1988). In our panel setup, we can circumvent the above bias issue in a simpler way, in particular by borrowing information from other individuals j: Suppose we know that i and j belong to the same class G_k. In this situation, we may replace the residual sum of squares $\text{RSS}_i(h)$ by

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

i.e., we may use the estimator $\widehat{m}_{i,h}$ to predict the Y-observations of the j-th rather than the i-th individual. This avoids the bias problem since the data of the j-th individual are independent from those of the i-th subject. Formally speaking, we obtain that

$$\mathbb{E}\big[\text{RSS}_i^{(j)}(h)\big] = \mathbb{E}\big[\text{ASE}_i^{(j)}(h)\big] + \frac{1}{T}\sum_{t=1}^{T}\sigma_j^2\mathbb{E}[w(X_{jt})],$$

where

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

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

$$\begin{aligned}\mathbb{E}\big[\mathrm{ASE}_i^{(j)}(h)\big] &= \mathbb{E}\Big[\mathbb{E}\big[\mathrm{ASE}_i^{(j)}(h)\big|\{(Y_{it},X_{it}):1\le t\le T\}\big]\Big] \\ &= \mathbb{E}\Big[\int \big(\widehat{m}_{i,h}(x)-m_i(x)\big)^2 f_i(x)w(x)dx\Big] =: \mathrm{MISE}_i(h),\end{aligned}$$

i.e., since the expectation of $\mathrm{ASE}_i^{(j)}(h)$ is nothing else than the mean integrated squared error $\mathrm{MISE}_i(h)$, the chosen bandwidth can be regarded as an approximation of the optimal bandwidth in a 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 $(i_{2\ell-1}, i_{2\ell})$ for $1 \le \ell \le L$ and some $L \le \lfloor n_k/2 \rfloor$. We compute the bandwidth estimate $\widehat{h}_{i_{2\ell-1}}^{(i_{2\ell})} = \mathrm{argmin}_h \mathrm{RSS}_{i_{2\ell-1}}^{(i_{2\ell})}(h)$ for each $1 \le \ell \le L$ and define our group-specific bandwidth selector by

$$\widehat{h}_k = \frac{1}{L}\sum_{1\le\ell\le L}\widehat{h}_{i_{2\ell-1}}^{(i_{2\ell})}.$$

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 $\widehat{h}_k$ can be interpreted as an approximation to the group-wide optimal bandwidth in a MISE sense. It is worth noting that we need not take into account all pairs of indices $(i_{2\ell-1}, i_{2\ell})$ to compute $\widehat{h}_k$; we may rather pick a small number L of them in order 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 selection rule as described in Fan and Gijbels (1996). We finally set h_0 to be the average of the computed bandwidths. Based on the bandwidth h_0, we can compute preliminary estimators $\widetilde{G}_1, \ldots, \widetilde{G}_{\widetilde{K}}$ of the classes.

Step 2: For each estimated class $\widetilde{G}_k$, we calculate the bandwidth $\widehat{h}_k$ as described above.

Based on the bandwidths $\widehat{h}_k$, 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 contained in the preliminary class estimates

$\widetilde{G}_1, \ldots, \widetilde{G}_{\widetilde{K}}$. In particular, we let $\tau_{n,T}(p) = \max_{1 \le k \le \widetilde{K}} \{ b_{n,T}(\widetilde{G}_k) + v_{n,T}(\widetilde{G}_k)(2 \log p)^{1/2} \}$, where $b_{n,T}(\widetilde{G}_k) = \max_{i,j \in \widetilde{G}_k, i<j} \mathcal{B}_{ij}/(T\widehat{h}_k)$ and $v_{n,T}(\widetilde{G}_k) = \max_{i,j \in \widetilde{G}_k, i<j} \sqrt{\mathcal{V}_{ij}}/(T\widehat{h}_k^{1/2})$ with $\mathcal{B}_{ij}$ and $\mathcal{V}_{ij}$ defined in Section 4.1. We thus obtain updated estimators $\widehat{G}_1, \ldots, \widehat{G}_{\widehat{K}}$ of the classes. We finally calculate group-specific bandwidths for the updated class estimates $\widehat{G}_k$, which we again denote by $\widehat{h}_k$. These are used in the next section to come up with a good bandwidth selection rule for the estimators $\widehat{g}_k$ of the group-specific regression functions.

4.4 Bandwidth choice for $\widehat{g}_k$

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 $\widehat{m}_i(x) = \widehat{m}_{i,h}(x)$ 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

$$\beta_k(x) = \big(\textstyle\int W(\varphi)\varphi^2 d\varphi\big) \dfrac{g_k''(x) f_k(x) + 2g_k'(x) f_k'(x)}{f_k(x)}$$

$$\nu_k(x) = \big(\textstyle\int W(\varphi^2) d\varphi\big) \dfrac{\sigma_k^2(x)}{f_k(x)},$$

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 3.3, the asymptotic bias and variance expressions of $\widehat{g}_k(x) = \widehat{g}_{k,h}(x)$ have a very similar form: they are equal to $B_{k,h}(x) = (h^2/2)\beta_k(x)$ and $V_{k,h}(x) = (n_k Th)^{-1}\nu_k(x)$, respectively. With these expressions at hand, we define the criterion functions $\xi_i(h) = \int [b_{i,h}^2(x) + v_{i,h}(x)] f_k(x) w(x) dx$ and $\Xi_k(h) = \int [B_{k,h}^2(x) + V_{k,h}(x)] f_k(x) w(x) dx$. Optimizing the bandwidth of the smoother $\widehat{m}_{i,h}$ with respect to $\xi_i(h)$ leads to

$$h_k^* = \Big(\frac{\int \nu_k(x) f_k(x) w(x) dx}{\int \beta_k^2(x) f_k(x) w(x) dx} \Big)^{1/5} T^{-1/5}$$

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

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

$$\begin{aligned} \text{MISE}_i(h) &= \mathbb{E}\Big[\int \big(\widehat{m}_{i,h}(x) - m_i(x) \big)^2 f_i(x) w(x) dx \Big] \\ &= \int \text{Bias}(\widehat{m}_{i,h}(x))^2 f_i(x) w(x) dx \\ &\qquad + \int \text{Var}(\widehat{m}_{i,h}(x)) f_i(x) w(x) dx. \end{aligned}$$

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

4.5 Rescaling

In many applications, the noise level of the time series data $\mathcal{Z}_i = \{(Y_{it}, X_{it}) : 1 \le t \le T\}$ can be expected to vary across individuals i. As a result, the quality of the estimates $\widehat{\Delta}_{ij}$ can be expected to vary as well. In order to take into account different noise levels in the data, we may replace the estimators $\widehat{\Delta}_{ij}$ 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 $\widehat{\Delta}_{ij} = \mathcal{B}_{ij}/(Th)$ + lower order terms. We can thus infer that

$$\widehat{\Delta}_{ij}^{\text{sc}} := \frac{\widehat{\Delta}_{ij}}{\mathcal{B}_{ij}} = \frac{1}{Th} + \text{ lower order terms.}$$

The leading term of this expansion is independent of the indices i and j. Hence, the scaled estimators $\widehat{\Delta}_{ij}^{\text{sc}}$ 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 estimates $\widehat{\Delta}_{ij}^{\text{sc}}$ rather than $\widehat{\Delta}_{ij}$. Of course, we cannot take the expressions $\widehat{\Delta}_{ij}^{\text{sc}}$ at face value but have to estimate the scaling factors $\mathcal{B}_{ij} = \|W\|^2(b_i + b_j)$, which can be achieved by the methods 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 estimates $\widehat{\Delta}_{ij}^{\text{sc}}$, the threshold parameter $\tau_{n,T}(p) = b_{n,T} + v_{n,T}(2\log p)^{1/2}$ from (4.3) has to 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

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

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 $G_1 = \{1, \dots, 50\}$, $G_2 = \{51, \dots, 80\}$, $G_3 = \{81, \dots, 100\}$, $G_4 = \{101, \dots, 110\}$ and $G_5 = \{111, \dots, 120\}$. The functions associated with these

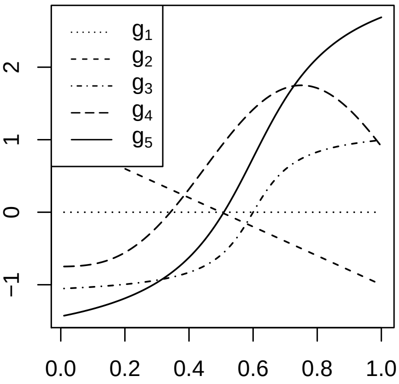

Figure 2: Plot of the functions g_k for $1 \le k \le 5$.

groups are $g_1(x) = 0$, $g_2(x) = 1 - 2x$, $g_3(x) = 0.75 \arctan(10(x - 0.6))$, $g_4(x) = 2.5\vartheta((x-0.75)/0.8)-0.75$ with $\vartheta(x) = (1-x^2)^4 1(|x| \le 1)$ and $g_5(x) = 1.75 \arctan(5(x-0.6)) + 0.75$. Figure 2 provides a plot of these functions, which are chosen to roughly approximate the shapes of the estimates $\widehat{g}_1, \ldots, \widehat{g}_5$ in the application later on.

The model errors ε_{it} are i.i.d. draws from a normal distribution with mean zero 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 into 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.

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 the rule (R) 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 $\widehat{m}_i$, we employ an Epanechnikov kernel and the bandwidth $h = 0.25$ throughout the simulations. As a robustness check, 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 Supplementary Material.

For each sample size (n, T) with $n = 120$ and $T \in \{100, 150, 200\}$, we drawn $N = 1000$ samples from the setting (5.1) and compute the threshold estimators $\{\widehat{G}_k : 1 \le$

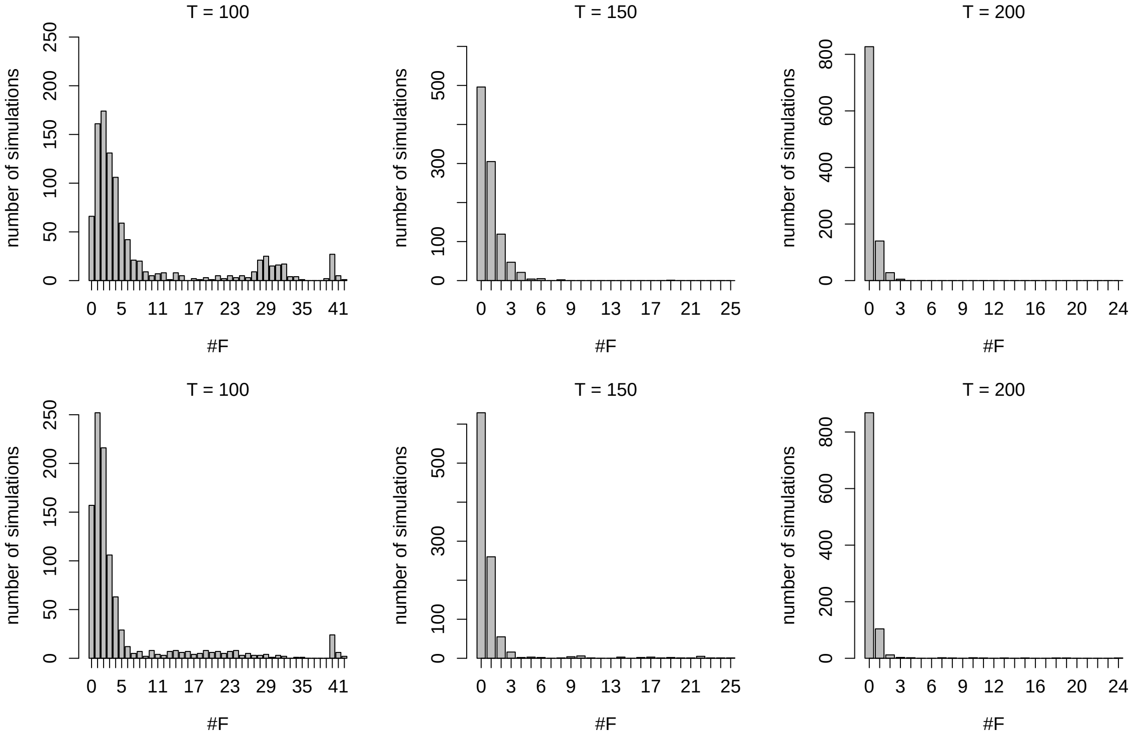

Figure 3: Simulation results for the estimation of the classes $G_1, \ldots, G_5$. The upper three panels show the distributions of the number $\#F$ of wrong classifications for the threshold estimators $\{\widehat{G}_k : 1 \le k \le K\}$ and the time series lengths $T = 100, 150, 200$. The lower three panels show the corresponding distributions for the k-means estimators $\{\widehat{G}_k^{\text{KM}} : 1 \le k \le K\}$.

$k \le \widehat{K}\}$ as well as the k-means estimators $\{\widehat{G}_k^{\text{KM}} : 1 \le k \le \widehat{K}\}$. In order to measure how well these estimates fit the real class structure $\{G_k : 1 \le k \le K\}$, we calculate the number of wrongly classified indices i, which is denoted by $\#F$ in what follows. For each sample size (n, T), we thus obtain $N = 1000$ values of $\#F$ both for the threshold and the k-means estimators. Figure 3 shows the distribution of these values. In particular, the bars in the plots give the number of simulations (out of total of 1000) in which a certain number of wrong classifications is obtained.

We now have a closer look at the simulation results in Figure 3. The upper three panels show the distribution of the number of wrong classifications $\#F$ for the threshold estimators $\{\widehat{G}_k : 1 \le k \le \widehat{K}\}$. Overall, the estimates can be seen to approximate the group structure reasonably well, their precision improving 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 5 wrongly classified indices in about 70% of the cases. There is however a substantial fraction of simulations in which many classification errors occur. This is not surprising since the time series length $T = 100$ is comparably small given

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

Table 1: Simulation results for the estimation of K. The entries in the table specify the number of simulations in which a certain value of $\widehat{K}$ is obtained.

the noise level of the error terms. The fits $\widehat{m}_i$ thus tend to be quite imprecise, which in turn leads to frequent classification errors.

The lower three panels of Figure 3 depict the distribution of $\#F$ for the k-means estimators $\{\widehat{G}_k^{\text{KM}} : 1 \le k \le \widehat{K}\}$. As one can see, for the smallest sample size $T = 100$, i.e., when the signal-to-noise ratio is still quite low, the estimators $\{\widehat{G}_k^{\text{KM}} : 1 \le k \le \widehat{K}\}$ strongly improve on the performance of the threshold estimators $\{\widehat{G}_k : 1 \le k \le \widehat{K}\}$. As already discussed in Section 2.3, we thus recommend to refine our threshold estimators by an additional k-means clustering step when the noise level in the data is high. For $T = 150$, we still get a quite substantial improvement on the performance of the thresholding procedure, while for the largest sample size $T = 200$, the additional gain from performing a k-means clustering step is comparably small.

We finally turn to the finite sample performance of the estimator $\widehat{K}$ which approximates the number of classes K. The simulation results are presented in Table 1. They suggest that the estimator $\widehat{K}$ 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 (MiFID)" 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 analyze and evaluate the effects of MiFID. 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. (2014b).

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 is supplied to us by Fidessa. The sample consists of weekly observations on the volume of all the FTSE stocks traded at a number of different venues in the UK; see Boneva et al. (2014a,b) for a more detailed description of the data set. The price data is taken from Datastream and comprises 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 the equation

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

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 $G_1, \ldots, G_K$ along with associated functions $g_1, \ldots, g_K$ such that $m_i = g_k$ for all $i \in G_k$ and all $1 \le k \le K$. 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 $1 \le k \le K$, we use the estimation techniques 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 the rule (R) 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 $\widehat{m}_i$ are based on an Epanechnikov kernel and their bandwidths are chosen as explained in Section 4.3. Prior to estimation, we eliminate stocks i with a very small empirical support $\mathcal{S}_i$ of the fragmentation data $\{X_{it} : 1 \le t \le T\}$. In particular, we

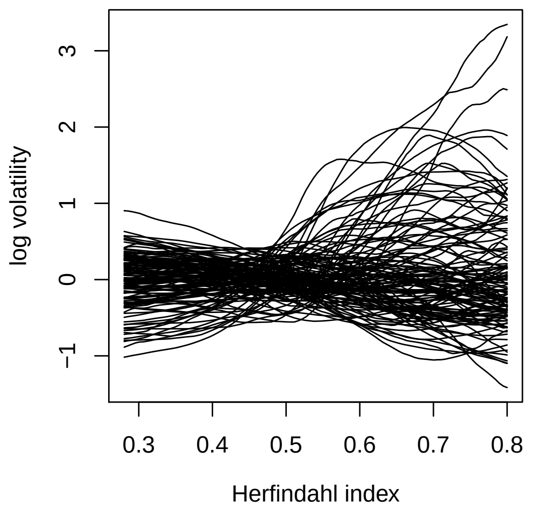

Figure 4: Estimates $\widehat{m}_i$ for the $n = 125$ stocks in our sample.

only take into account 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 dimension amounts to $T = 151$ weeks. These sizes of n and T are broadly consistent with our assumptions from Section 3.

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

Inspecting Figure 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 $\widehat{g}_4$. The fits of Cluster V slightly slope downwards, indicating that the volatility level is a bit lower in the monopoly case than under competition. Most of the fits in Cluster III are moderately increasing, suggesting that the volatility is a bit lower under competition. In contrast to 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 a bad, the results of Figure 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 (cp.

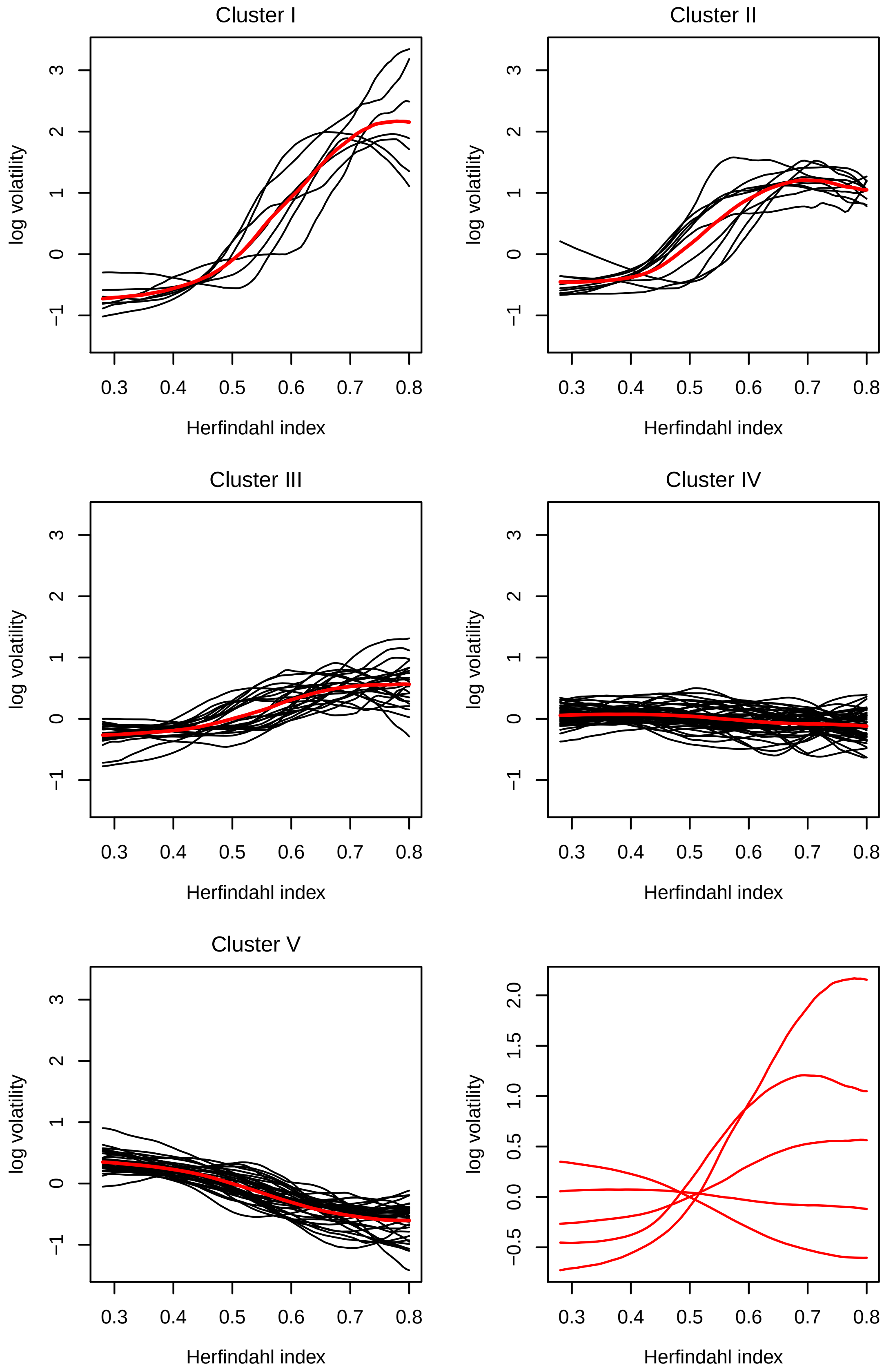

Figure 5: Clusters of the curve estimates $\widehat{m}_i$. The black lines are the estimates $\widehat{m}_i$, the red lines the estimates $\widehat{g}_k$. The latter are once again plotted together in the lower right panel.

Cluster III) or deteriorate (cp. Cluster V) market quality.

We shortly compare these findings to the empirical results in Boneva et al. (2014a). In contrast to our approach, they impose the factor structure $m_i(x) = \sum_{k=1}^{K} \beta_{ik}\mu_k(x)$ 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 $\beta_i = (\beta_{i1}, \dots \beta_{iK})^\top$ assign different individual-specific weights to these factors. Applying their model to the data at hand, Boneva et al. (2014a) find 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 the latter, 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. (2014a).

Appendix A

In what follows, we prove Theorems 3.1 and 3.2. The proof of Theorem 3.3 can be found in the Supplementary Material. To derive the theorems, we make use of the following uniform convergence result.

Lemma A.1. *Let (C1)–(C5) be satisfied, define $I_h = [C_1 h, 1 - C_1 h]$ and set $a_{n,T} = T^{-1/10}$. It holds that*

$$\max_{1 \le i \le n} \sup_{x \in I_h} \big| \widehat{m}_i(x) - m_i(x) \big| = O_p(a_{n,T} + h^2)$$

$$\max_{1 \le i \le n} \sup_{x \in [0,1] \setminus I_h} \big| \widehat{m}_i(x) - m_i(x) \big| = O_p(a_{n,T} + h).$$

If we strengthen the moment assumptions in (C3) to hold for some $\theta > 20/3$, we can improve this result to hold with $a_{n,T} = \sqrt{\log T/(Th)}$. The proof is deferred to the Supplementary Material. From Lemma A.1, it easily follows that

$$\max_{1 \le i,j \le n} \big| \widehat{\Delta}_{ij} - \Delta_{ij} \big| = o_p(1). \tag{A.1}$$

Moreover, we obtain that

$$\max_{i,j \in G_k} \widehat{\Delta}_{ij} = O_p(a_{n,T}^2 + h^3) \tag{A.2}$$

for any $1 \le k \le K$. Notably, (A.2) merely provides an upper bound on the rate of $\max_{i,j \in G_k} \widehat{\Delta}_{ij}$. The reason is as follows: Directly applying Lemma A.1 does not take into account that the argument x of the smoothers $\widehat{m}_i(x)$ and $\widehat{m}_j(x)$ is integrated out in $\widehat{\Delta}_{ij}$. We now derive the sharp rate of $\max_{i,j \in G_k} \widehat{\Delta}_{ij}$ under stronger conditions than (C1)–(C5).

Lemma A.2. *Let (C1)–(C5) be satisfied, let $h \le CT^{-(1/5+\delta)}$ for some constant C and some small $\delta > 0$, and choose the weight function π such that its support is contained in $I_h = [C_1 h, 1 - C_1 h]$. In addition, drop the fixed effects α_i and γ_t from the model and suppose that the following conditions hold:*

(i) The variables X_{it} and ε_{it} are independent both across i and t. Moreover, X_{it} and ε_{it} are independent of each other for any i and t.

(ii) The second derivatives m_i'' fulfill the Lipschitz condition that $|m_i''(x) - m_i''(x')| \le L|x - x'|$ for all x, x' and a constant L independent of i.

(iii) There exist constants $M, \gamma > 0$ such that for all indices i, t and for all $c \ge 0$, $\mathbb{P}(|\varepsilon_{it}| \ge c) \le M \int_c^\infty \exp(-\gamma r^2) dr$.

Then for any $1 \le k \le K$,

$$\max_{i,j \in G_k} \widehat{\Delta}_{ij} = \max_{\substack{i,j \in G_k \\ i<j}} \frac{\mathcal{B}_{ij}}{Th} + O_p\Big(\frac{\log T}{Th^{1/2}}\Big) = O_p\Big(\frac{1}{Th}\Big),$$

where $\mathcal{B}_{ij}$ is defined in Subsection 4.1.

Proof of Theorem 3.1

Let $S \subseteq \{1, \ldots, n\}$ be some index set with $n_S = |S|$, pick an index $i \in S$, and let $G \subseteq S$ be the class to which i belongs. As seen in Subsection 2.2, the group G has the form $G = \{(1), \ldots, (p)\}$, where $\Delta_{i(1)} = \ldots = \Delta_{i(p)} < \Delta_{i(p+1)} \le \ldots \le \Delta_{i(n_S)}$ are the ordered L_2-distances. Denoting the ordered estimated distances by $\widehat{\Delta}_{i[1]} \le \widehat{\Delta}_{i[2]} \le \ldots \le \widehat{\Delta}_{i[n_S]}$, we estimate G by $\widehat{G} = \{[1], \ldots, [\widehat{p}]\}$ with $\widehat{p}$ defined in (2.3). In what follows, we show that

$$\mathbb{P}\Big(\big\{[1], \ldots, [\widehat{p}]\big\} \ne \big\{(1), \ldots, (p)\big\}\Big) = o(1), \tag{A.3}$$

from which the statements of Theorem 3.1 can be easily inferred. For the proof of (A.3), it suffices to show that

$$\mathbb{P}\Big(\big\{[1], \ldots, [p]\big\} \ne \big\{(1), \ldots, (p)\big\}\Big) = o(1) \tag{A.4}$$

$$\mathbb{P}\big(\widehat{p} \ne p\big) = o(1). \tag{A.5}$$

These two statements can be verified as follows: By (A.1), it holds that $\widehat{\Delta}_{i(j)} - \Delta_{i(j)} = o_p(1)$ uniformly over $j \in S$. As $\Delta_{i(j)} = 0$ for all $j \le p$ and $\Delta_{i(j)} \ge c$ for all $j > p$ and some constant $c > 0$, we obtain that

$$\max_{j \le p} \widehat{\Delta}_{i(j)} = o_p(1) \quad \text{and} \quad \min_{j > p} \widehat{\Delta}_{i(j)} \ge c + o_p(1). \tag{A.6}$$

This immediately implies that the ordered estimates $\widehat{\Delta}_{i[j]}$ have the same property, i.e.,

$$\max_{j\le p} \widehat{\Delta}_{i[j]} = o_p(1) \quad \text{and} \quad \min_{j>p} \widehat{\Delta}_{i[j]} \ge c + o_p(1). \tag{A.7}$$

From (A.6) and (A.7), it is obvious that the index sets $\{[1],\dots,[p]\}$ and $\{(1),\dots,(p)\}$ coincide with probability tending to one, which is the statement of (A.4). From (A.4), it follows that $\max_{j\in G} \widehat{\Delta}_{ij} = \widehat{\Delta}_{i[p]}$ with probability tending to one. Moreover, as the threshold parameter $\tau_{n,T}$ satisfies (C_τ), $\widehat{\Delta}_{i[p]} \le \tau_{n,T}$ with probability approaching one. Finally, by (A.7), $\widehat{\Delta}_{i[p+1]} > \tau_{n,T}$ with probability approaching one as well. We thus arrive at

$$\mathbb{P}\big(\widehat{\Delta}_{i[p]} \le \tau_{n,T} \text{ and } \widehat{\Delta}_{i[p+1]} > \tau_{n,T}\big) \to 1,$$

which immediately implies that $\mathbb{P}(\widehat{p} = p) \to 1$. □

Proof of Theorem 3.2

As $\widehat{K} = K$ with probability tending to one, we can neglect the estimation error in $\widehat{K}$ and treat K as known. With the help of Lemma A.1, it is straightforward to see that

$$\int \big(\widehat{m}_i(x) - \widehat{g}_k^{[1]}(x)\big)^2 \pi(x)dx = \int \big(m_i(x) - g_k(x)\big)^2 \pi(x)dx + o_p(1)$$

uniformly over i and k, or put differently,

$$\max_{1\le k\le K} \max_{1\le i\le n} \big|\Delta(\widehat{m}_i, \widehat{g}_k^{[1]}) - \Delta(m_i, g_k)\big| = o_p(1). \tag{A.8}$$

By construction, the index i is assigned to the group $G_k^{[1]}$ in the first step of the k-means algorithm if $\widehat{d}_k(i) = \Delta(\widehat{m}_i, \widehat{g}_k^{[1]})$ is minimal, i.e., if $\widehat{d}_k(i) = \min_{1\le k'\le K} \widehat{d}_{k'}(i)$. By (A.8), we know that

$$\widehat{d}_k(i) = \begin{cases} \widehat{r}_k(i) & \text{if } i \in G_k \\ \Delta(m_i, g_k) + \widehat{r}_k(i) & \text{if } i \notin G_k, \end{cases} \tag{A.9}$$

where the remainder term $\widehat{r}_k(i)$ has the property that $\max_{1\le k\le K} \max_{1\le i\le n} |\widehat{r}_k(i)| = o_p(1)$. Since $\min_{1\le k\le K} \min_{i\notin G_k} \Delta(m_i, g_k) \ge \Delta_{\min} > 0$ for some positive constant $\Delta_{\min}$, (A.9) implies that

$$\mathbb{P}\Big(\big\{G_k^{[1]} : 1 \le k \le K\big\} \ne \big\{G_k : 1 \le k \le K\big\}\Big) = o(1).$$

Hence, with probability tending to one, our k-means clustering algorithm converges already after the first iteration step and produces estimates which coincide with the classes G_k for $1 \le k \le K$. □

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.

Baltagi, B. H. (2013). *Econometric analysis of panel data.* Wiley.

Boneva, L., Linton, O. and Vogt, M. (2014a). The effect of fragmentation in trading on market quality in the uk equity market. *Forthcoming in Journal of Applied Econometrics.*

Boneva, L., Linton, O. and Vogt, M. (2014b). A semiparametric model for heterogeneous panel data with fixed effects. *Forthcoming in Journal of Econometrics.*

Bosq, D. (1998). *Nonparametric statistics for stochastic processes.* New York, Springer.

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

Cox, D. R. (1957). Note on grouping. *Journal of the American Statistical Association*, **52** 543–547.

Degryse, H., De Jong, F. and Van Kervel, V. (2014). The impact of dark trading and visible fragmentation on market quality. *Review of Finance* 1–36.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications.* Chapman & Hall.

Fisher, D. (1958). On grouping for maximum homogeneity. *Journal of the American Statistical Association*, **53** 789–798.

Garcia-Escudero, L. A. and Gordaliza, A. (1999). Robustness of properties of k-means and trimmed k-means. *Journal of the American Statistical Association*, **94** 956–969.

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

Härdle, W., Hall, P. and Marron, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *Journal of the American Statistical Association*, **83** 86–95.

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

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

Hsiao, C. (2003). *Analysis of panel data.* Cambridge University Press.

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

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.

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

Masry, E. (1996). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *Journal of Time Series Analysis*, **17** 571–599.

O'Hara, M. and Ye, M. (2009). Is fragmentation harming market quality? *Journal of Financial Economics*, **100** 459–474.

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

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

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

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

Su, L., Shi, Z. and Phillips, P. C. B. (2014). Identifying latent structures in panel data. *Preprint*.

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

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