cemmap
centre for microdata methods and practice

Classification of nonparametric regression functions in heterogeneous panels

Michael Vogt
Oliver Linton

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP06/15

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Classification of Nonparametric Regression Functions in Heterogeneous Panels

Michael Vogt[1] University of Konstanz

Oliver Linton[2] University of Cambridge

February 20, 2015

We investigate a nonparametric panel model with heterogeneous regression functions. 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 observed 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; k-means clustering; kernel estimation; nonparametric regression; panel data.
AMS 2010 subject classifications: 62G08; 62G20; 62H30.

1 Introduction

Most of the literature on non- and semiparametric panel models is based on the assumption that the regression function is the same across individuals; see Henderson et al. (2008), Mammen et al. (2009) and Qian and Wang (2012) 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.

[1]Corresponding author. Address: Department of Mathematics and Statistics, University of Konstanz, 78457 Konstanz, Germany. Email: mv346@cam.ac.uk.
[2]Address: Faculty of Economics, Austin Robinson Building, Sidgwick Avenue, Cambridge, CB3 9DD, UK. Email: obl20@cam.ac.uk.

Throughout the paper, we work with the following model setup. We observe a sample of panel data $\{(Y_{it}, X_{it}) : 1 \leq i \leq n, 1 \leq t \leq T\}$, where i denotes the i-th individual and t is the time point of observation. 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 $G_1, \ldots, G_{K_0}$ be a fixed number of disjoint sets which partition the index set $\{1, \ldots, n\}$, that is, $G_1 \,\dot\cup \ldots \dot\cup\, G_{K_0} = \{1, \ldots, n\}$. Moreover, let $g_1, \ldots, g_{K_0}$ be functions associated with these sets. We suppose that

$$m_i = g_k \quad \text{for all } i \in G_k \text{ and } 1 \leq k \leq K_0. \tag{1.2}$$

Hence, the observed individuals can be grouped into a finite number of classes G_k whose members share the same regression curve g_k. Our aim is to estimate the groups $G_1, \ldots, G_{K_0}$, their number K_0 and the associated functions $g_1, \ldots, g_{K_0}$.

A great concern in many panel data applications is the issue of endogeneity, that is, the issue that the error term may be correlated with the regressors. Such a correlation may be produced, for example, by unobserved variables that are not controlled for and thus induce some sort of omitted variable bias. To take into account the issue of endogeneity in our model, we suppose the error terms u_{it} in (1.1) to have the structure $u_{it} = \alpha_i + \gamma_t + \varepsilon_{it}$, where ε_{it} are idiosyncratic error terms with $\mathbb{E}[\varepsilon_{it}|X_{it}] = 0$. Moreover, α_i and γ_t are unobserved individual and time specific error terms which may be correlated with the regressors in an arbitrary way. Specifically, defining $\mathcal{X}_{n,T} = \{X_{it} : 1 \leq i \leq n, 1 \leq t \leq T\}$, we allow that $\mathbb{E}[\alpha_i|\mathcal{X}_{n,T}] \neq 0$ and $\mathbb{E}[\gamma_t|\mathcal{X}_{n,T}] \neq 0$ in general. In the panel literature, α_i and γ_t are commonly termed individual and time specific fixed effects, respectively. The time series dimension T of the observed panel 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. To identify the functions m_i in (1.1), 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. Finally, 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.

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 and competed for trading volume. 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 fixed effects panel 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 $G_1, \dots, G_{K_0}$ along with associated coefficients $\theta_1, \dots, \theta_{K_0}$ such that $\beta_i = \theta_k$ for all $i \in G_k$ and all $1 \leq k \leq K_0$. The problem of estimating the unknown groups $G_1, \dots, G_{K_0}$ in this parametric framework has been considered, for example, in Sarafidis and Weber (2014) and Su et al. (2014) who work with penalization techniques, and in Lin and Ng (2012) who employ thresholding and k-means clustering methods.

Our modelling approach is related to classification problems in functional data analysis. There, the observed data $X_1, \dots, X_n$ 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 $X_1, \dots, X_n$ are observed without noise; in others, they are observed with noise. In the latter case, they have to be estimated from noisy observations $Y_1, \dots, Y_n$ 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 $X_1, \dots, X_n$ 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 panel 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 panel framework and in functional data analysis.

Our procedure to estimate the classes $G_1, \ldots, G_{K_0}$ along with the associated functions $g_1, \ldots, g_{K_0}$ in model (1.1)–(1.2) is presented in Section 2. There, we construct two algorithms to estimate the classes. In both cases, we compute the pairwise L_2-distances between kernel estimates of the regression curves m_i and use the information contained in these distances to infer the unknown class structure. The first algorithm exploits a particular pattern in the ordered L_2-distances, whereas the second one is a k-means clustering type of approach applied to the estimated distances. To obtain an estimation method with good theoretical and practical properties, we combine these two algorithms to form a two-step procedure: the first algorithm provides us with initial estimators that are used as starting values for the k-means type algorithm. With these estimators of the classes G_k at our disposal, it is straightforward to construct estimators of the functions g_k. Specifically, since $g_k = |G_k|^{-1} \sum_{i \in G_k} m_i$ with $|G_k|$ denoting the cardinality of G_k, we may simply estimate g_k by averaging the kernel estimates of the functions m_i whose index i belongs to the estimated class G_k.

The asymptotic properties of our estimation approach are investigated in Section 3. There, we show that our estimators of the classes $G_1, \ldots, G_{K_0}$ are consistent and we derive the limit distribution of the estimators of the associated functions $g_1, \ldots, g_{K_0}$. Our estimation approach can be used both when the number of classes K_0 is known and when K_0 is replaced by a consistent estimator. In Section 4, we describe how to construct such an estimator of K_0 and how to implement it in practice to obtain a good finite sample performance. 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 $G_1, \ldots, G_{K_0}$ and the corresponding functions $g_1, \ldots, g_{K_0}$ 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. Throughout the section, we assume that the number K_0 of groups is known. In Section 4, we drop this simplifying assumption and replace K_0 by an estimator. 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_k and finally use these to come up with estimators of the functions g_k for $1 \leq k \leq K_0$.

2.1 Estimation of the regression functions m_i

To construct an estimator of the regression function m_i of the i-th individual, we proceed as follows: Let $Y_{it}^{\mathrm{fe}} = Y_{it} - \alpha_i - \gamma_t$ 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 $Y_{it}^{\mathrm{fe}} = m_i(X_{it}) + \varepsilon_{it}$, 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

$$\widehat{m}_i^*(w) = \frac{\sum_{t=1}^T K_h(X_{it} - w) Y_{it}^{\mathrm{fe}}}{\sum_{t=1}^T K_h(X_{it} - w)},$$

where h is the bandwidth and K denotes a kernel function with $K_h(w) = h^{-1}K(w/h)$. To obtain a feasible estimator of m_i, we replace the unobserved variables Y_{it}^{fe} in the above formula by the approximations $\widehat{Y}_{it}^{\mathrm{fe}} = Y_{it} - \overline{Y}_{\,i} - \overline{Y}_{\,t}^{(i)} + \overline{\overline{Y}}^{(i)}$, where

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

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

$$\widehat{m}_i(w) = \frac{\sum_{t=1}^T K_h(X_{it} - w) \widehat{Y}_{it}^{\mathrm{fe}}}{\sum_{t=1}^T K_h(X_{it} - w)}$$

of the regression function m_i. Alternatively to the Nadaraya-Watson smoother $\widehat{m}_i$, we could work with a local linear or more generally with a local polynomial estimator. Indeed, our procedure to estimate the groups G_k and the corresponding functions g_k for $1 \leq k \leq K_0$ is the same no matter which type of kernel smoother we employ.

2.2 Estimation of the groups $G_1, \ldots, G_{K_0}$

We now present two methods to estimate the classes $G_1, \ldots, G_{K_0}$. The first method provides us with preliminary estimators of the classes. These serve as starting values for the second method, which yields improved estimators of $G_1, \ldots, G_{K_0}$. To formulate the two procedures, we introduce some notation: For two functions $q_1 : \mathbb{R} \to \mathbb{R}$ and

$q_2 : \mathbb{R} \to \mathbb{R}$, let

$$\Delta(q_1, q_2) = \int \big(q_1(w) - q_2(w)\big)^2 \pi(w) dw \tag{2.1}$$

be their weighted squared L_2-distance, where π is some weight function. To shorten notation, we write $\Delta_{ij} = \Delta(m_i, m_j)$ in what follows. Moreover, we let $\overline{m}_S = \frac{1}{|S|} \sum_{i \in S} m_i$ be the average of the functions m_i in the group S, where $|S|$ is the cardinality of S.

A preliminary estimation algorithm. Our first estimation approach is based on the following observation: Let $S \subseteq \{1, \ldots, n\}$ be an index set which contains at least two different classes G_k and $G_{k'}$. For each $i \in S$, the distances $\{\Delta_{ij} : j \in S\}$ exhibit a particular pattern when sorted in increasing order: Denoting the ordered distances by $\Delta_{i(1)} \leq \Delta_{i(2)} \leq \ldots \leq \Delta_{i(n_S)}$ with $n_S = |S|$ being the cardinality of S, there are κ points $j_1 < \ldots < j_\kappa$ with $1 \leq \kappa < K_0$ such that

$$\begin{aligned} \Delta_{i(1)} = \ldots = \Delta_{i(j_1 - 1)} &< \Delta_{i(j_1)} = \ldots = \Delta_{i(j_2 - 1)} \\ &< \Delta_{i(j_2)} = \ldots = \Delta_{i(j_3 - 1)} \\ &\vdots \\ &< \Delta_{i(j_\kappa)} = \ldots = \Delta_{i(n_S)}. \end{aligned}$$

The indices $j_1, \ldots, j_\kappa$ mark the positions where the L_2-distance jumps to another value. These jump points are informative on the group structure $\{G_k : 1 \leq k \leq K_0\}$: Two indices (j) and (j') can only belong to the same class G_k if $\Delta_{i(j)} = \Delta_{i(j')}$.

We now exploit the step structure of the ordered L_2-distances to partition the index set $\{1, \ldots, n\}$ into the classes $G_1, \ldots, G_{K_0}$ by an iterative procedure. In each iteration step, we split an index set S (which contains at least two different classes G_k and $G_{k'}$) into two subsets as follows:

(AL1) Pick some index $i \in S$, sort the distances $\{\Delta_{ij} : j \in S\}$ in increasing order and denote the ordered distances by $\Delta_{i(1)} \leq \Delta_{i(2)} \leq \ldots \leq \Delta_{i(n_S)}$.

(AL2) Determine the position of the largest jump,

$$j_{\max} = \arg \max_{2 \leq j \leq n_S} \big|\Delta_{i(j)} - \Delta_{i(j-1)}\big|.$$

(AL3) Partition S into two subgroups as follows: $S = S_< \dot\cup S_>$ with

$$S_< = \big\{(1), \ldots, (j_{\max} - 1)\big\} \quad \text{and} \quad S_> = \big\{(j_{\max}), \ldots, (n_S)\big\}.$$

The sets $S_<$ and $S_>$ have the following property: each class $G_k \subset S$ is either contained in $S_<$ or in $S_>$. Hence, the algorithm (AL1)–(AL3) separates the classes G_k contained in S into two groups. We now iterate this algorithm as follows:

1^{st} Step: Set $S = \{1, \ldots, n\}$ and split it up into two subgroups $S_1 = S_<$ and $S_2 = S_>$ by applying (AL1)–(AL3).

r^{th} Step: Let $\{S_1, \ldots, S_r\}$ be the partition of $\{1, \ldots, n\}$ from the previous iteration step. Pick some group S_{ℓ^*} from this partition for which $\max_{i,j \in S_{\ell^*}} \Delta_{ij} > 0$. This condition ensures that S_{ℓ^*} contains at least two different classes G_k and $G_{k'}$. Now split S_{ℓ^*} into two subgroups $S_{\ell^*,<}$ and $S_{\ell^*,>}$ by applying (AL1)–(AL3). This yields a refined partition with the $(r+1)$ elements $S_{\ell^*,<}$, $S_{\ell^*,>}$ and S_ℓ for $1 \leq \ell \leq r$, $\ell \neq \ell^*$.

Repeating this algorithm $(K_0 - 1)$ times partitions the index set $\{1, \ldots, n\}$ into K_0 groups. By construction, these groups are identical to the classes $G_1, \ldots, G_{K_0}$.

To obtain estimators of the classes $G_1, \ldots, G_{K_0}$, we apply the iterative procedure from above to estimated versions of the distances Δ_{ij}. In particular, we estimate Δ_{ij} by $\widehat{\Delta}_{ij} = \Delta(\widehat{m}_i, \widehat{m}_j)$ and perform the following iterative algorithm:

1^{st} Step: Set $S = \{1, \ldots, n\}$ and split it up into two subgroups $S_1 = S_<$ and $S_2 = S_>$ by applying (AL1)–(AL3) to the estimated distances $\{\widehat{\Delta}_{ij} : i, j \in S\}$.

r^{th} Step: Let $\{S_1, \ldots, S_r\}$ be the partition of $\{1, \ldots, n\}$ from the previous iteration step. Calculate the maximum L_2-distance $\max_{i,j \in S_\ell} \widehat{\Delta}_{ij}$ for each ℓ with $1 \leq \ell \leq r$. Pick the group S_ℓ with the largest maximum distance, say S_{ℓ^*}, and split it up into two subgroups $S_{\ell^*,<}$ and $S_{\ell^*,>}$ by applying (AL1)–(AL3) to the estimated distances $\{\widehat{\Delta}_{ij} : i, j \in S_{\ell^*}\}$. This yields a refined partition with the $(r+1)$ elements $S_{\ell^*,<}$, $S_{\ell^*,>}$ and S_ℓ for $1 \leq \ell \leq r$, $\ell \neq \ell^*$.

Iterating this procedure $(K_0 - 1)$ times partitions the index set $\{1, \ldots, n\}$ into K_0 groups $\widetilde{G}_1, \ldots, \widetilde{G}_{K_0}$ which serve as our estimators of the classes $G_1, \ldots, G_{K_0}$.

When implementing the estimators $\widetilde{G}_1, \ldots, \widetilde{G}_{K_0}$, there is in principle no restriction on how to pick the index $i \in S$ in the first step (AL1) of the iterative algorithm (AL1)–(AL3). Indeed, the asymptotic properties of $\widetilde{G}_1, \ldots, \widetilde{G}_{K_0}$ derived in Section 3 hold true no matter how we choose the index i in (AL1). In practice, we suggest to pick the index i according to the following rule:

(R) For each $i \in S$, let $\widehat{J}_{ij} = |\widehat{\Delta}_{i(j+1)} - \widehat{\Delta}_{i(j)}|$ with $1 \leq j \leq n_S - 1$ be the jumps between the estimated L_2-distances and denote the ordered jumps by $\widehat{J}_{i(1)} \leq \widehat{J}_{i(2)} \leq \ldots \leq \widehat{J}_{i(n_S-1)}$. Define $\widehat{\Sigma}(i) = \sum_{j=1}^{n_S - K_0} \widehat{J}_{i(j)}$ for $i \in S$ and choose $i = i^*$ with $\widehat{\Sigma}(i^*) = \min_{i \in S} \widehat{\Sigma}(i)$.

The heuristic idea behind this rule is as follows: The algorithm (AL1)–(AL3) exploits the fact that for any $i \in S$, the ordered L_2-distances $\Delta_{i(1)} \leq \Delta_{i(2)} \leq \ldots \leq \Delta_{i(n_S)}$ have a step structure. When applied to the estimated distances $\widehat{\Delta}_{i(1)} \leq \widehat{\Delta}_{i(2)} \leq \ldots \leq \widehat{\Delta}_{i(n_S)}$, the algorithm can thus be expected to perform well only when the step structure is captured accurately by the estimates $\{\widehat{\Delta}_{i(j)} : j \in S\}$. This suggests to choose an index

$i \in S$ for which the step structure is well approximated. The rule (R) is designed to achieve such a choice: As the total number of classes is K_0, there are at most $(K_0 - 1)$ non-zero steps in the ordered L_2-distances $\{\Delta_{i(j)} : j \in S\}$, implying that the jumps $\widehat{J}_{i(j)}$ must converge to zero for $1 \leq j \leq n_S - K_0$. The smaller these jumps are, the better the step structure is captured by the estimates $\{\widehat{\Delta}_{i(j)} : j \in S\}$ corresponding to the index i. The expression $\widehat{\Sigma}(i)$ can thus be regarded as a measure of how well the step structure is approximated for the index i. In particular, the smaller $\widehat{\Sigma}(i)$, the better the approximation.

As we will see later on, the proposed estimation procedure consistently estimates the classes $G_1, \ldots, G_{K_0}$. Hence, from an asymptotic perspective, it works as desired. Moreover, from a computational point of view, it is a quite fast algorithm because roughly speaking, it only requires to determine the maximum of $O(n)$ terms and to repeat this maximum search for $(K_0 - 1)$ times. Its small sample performance, however, is not fully satisfactory in some situations. The reason is as follows: As already noted, the algorithm (AL1)–(AL3) can only be expected to perform well when the step structure is captured in a reasonable way by the estimated L_2-distances $\widehat{\Delta}_{ij}$. When the noise level of the error terms in model (1.1) is low, the estimates $\widehat{\Delta}_{ij}$ are quite precise and tend to approximate the step structure fairly accurately. In this case, the algorithm works well and produces estimates $\widetilde{G}_1, \ldots, \widetilde{G}_{K_0}$ with a good small sample performance. When the noise level is high in contrast, the step structure may not be captured appropriately by the estimates $\widehat{\Delta}_{ij}$ any more, which may lead to poor estimates of the classes $G_1, \ldots, G_{K_0}$. For this reason, we use $\widetilde{G}_1, \ldots, \widetilde{G}_{K_0}$ only as the starting values of a second estimation algorithm which produces more stable and accurate results.

A *k*-means estimation algorithm. The second estimation method makes use of the fact that the classes $G_1, \ldots, G_{K_0}$ can be characterized as follows: Let $P = \{S_1, \ldots, S_{K_0}\}$ be a partition of $\{1, \ldots, n\}$ into K_0 sets and let $\mathcal{P}$ be the set of all possible partitions with K_0 elements. Then the partition $\{G_1, \ldots, G_{K_0}\}$ can be characterized as

$$\{G_1, \ldots, G_{K_0}\} = \operatorname*{argmin}_{P=\{S_1,\ldots,S_{K_0}\}\in\mathcal{P}} \sum_{k=1}^{K_0} \sum_{i\in S_k} \Delta(m_i, \overline{m}_{S_k}). \tag{2.2}$$

Here, $\Delta(m_i, \overline{m}_{S_k})$ measures the squared L_2-distance between the function m_i and the mean function or centroid $\overline{m}_{S_k}$ of the cluster S_k. Moreover, $\sum_{i\in S_k} \Delta(m_i, \overline{m}_{S_k})$ specifies the sum of squared distances within the cluster S_k. The partition $\{G_1, \ldots, G_{K_0}\}$ thus minimizes the within-cluster sums of squared distances.

Formula (2.2) suggests to estimate $\{G_1, \ldots, G_{K_0}\}$ by minimizing a sample analogue of the within-cluster sums of squared distances, in particular by minimizing the criterion function $\sum_{k=1}^{K_0} \sum_{i\in S_k} \Delta(\widehat{m}_i, \widehat{\overline{m}}_{S_k})$, where $\widehat{\overline{m}}_{S_k} = |S_k|^{-1} \sum_{i\in S_k} \widehat{m}_i$. In practice, this minimizer can be approximated by applying a k-means clustering algorithm to the

estimated functions $\{\widehat{m}_i : 1 \leq i \leq n\}$. This type of algorithm is very popular and 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. Our version of the k-means clustering algorithm proceeds as follows:

1st Step: Choose starting values $\widehat{\overline{m}}_1^{[0]}, \ldots, \widehat{\overline{m}}_{K_0}^{[0]}$ for the cluster means and calculate the distances $\widehat{d}_k(i) = \Delta(\widehat{m}_i, \widehat{\overline{m}}_k^{[0]})$ for each i and k. Define the partition $\{S_1^{[0]}, \ldots, S_{K_0}^{[0]}\}$ by assigning the index i to the k-th group $S_k^{[0]}$ if $\widehat{d}_k(i) = \min_{1 \leq k' \leq K_0} \widehat{d}_{k'}(i)$.

r^{th} Step: Let $\{S_1^{[r-1]}, \ldots, S_{K_0}^{[r-1]}\}$ be the partition of $\{1, \ldots, n\}$ from the previous iteration step. Calculate mean functions

$$\widehat{\overline{m}}_k^{[r]} = \frac{1}{|S_k^{[r-1]}|} \sum_{i \in S_k^{[r-1]}} \widehat{m}_i \quad \text{for } 1 \leq k \leq K_0$$

based on this partition and compute the distances $\widehat{d}_k(i) = \Delta(\widehat{m}_i, \widehat{\overline{m}}_k^{[r]})$ for each i and k. Define the new partition $\{S_1^{[r]}, \ldots, S_{K_0}^{[r]}\}$ by assigning the index i to the k-th group $S_k^{[r]}$ if $\widehat{d}_k(i) = \min_{1 \leq k' \leq K_0} \widehat{d}_{k'}(i)$.

This algorithm is iterated until the computed partition does 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 $\{G_k : 1 \leq k \leq K_0\}$ which are denoted by $\{\widehat{G}_k : 1 \leq k \leq K_0\}$ in what follows.

The performance of our k-means clustering procedure obviously depends on the choice of the starting values $\widehat{\overline{m}}_1^{[0]}, \ldots, \widehat{\overline{m}}_{K_0}^{[0]}$. In particular, when these are not picked appropriately, the procedure may not converge to the partition which minimizes the within-cluster sums of squares but may be stuck in a local minimum. For this reason, it does not yield a consistent estimator of the partition $\{G_k : 1 \leq k \leq K_0\}$ in general. To ensure consistency, we have to make sure that we start off with appropriate mean functions $\widehat{\overline{m}}_1^{[0]}, \ldots, \widehat{\overline{m}}_{K_0}^{[0]}$. To achieve this, we make use of our first estimation method which provides us with preliminary consistent estimates $\widetilde{G}_1, \ldots, \widetilde{G}_{K_0}$ of the groups $G_1, \ldots, G_{K_0}$. From these, we can calculate starting values

$$\widehat{\overline{m}}_k^{[0]} = \frac{1}{|\widetilde{G}_k|} \sum_{i \in \widetilde{G}_k} \widehat{m}_i \quad \text{for } 1 \leq k \leq K_0. \tag{2.3}$$

As these converge to the true centroids $\overline{m}_k = |G_k|^{-1} \sum_{i \in G_k} m_i$, our k-means clustering algorithm can be proven to consistently estimate the partition $\{G_k : 1 \leq k \leq K_0\}$ when applied with the starting values defined in (2.3). In the sequel, we implicitly take for granted that the algorithm is always started with these values.

2.3 Estimation of the functions $g_1, \ldots, g_{K_0}$

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

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

where $|\widehat{G}_k|$ denotes the cardinality of the set $\widehat{G}_k$. Hence, we simply average the estimators $\widehat{m}_i$ with indices in the estimated group $\widehat{G}_k$.

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

We make the following assumptions.

(C1) The time series $\mathcal{Z}_i = \{(X_{it}, \varepsilon_{it}) : 1 \leq t \leq T\}$ 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 $\alpha_i(\ell) \leq \alpha(\ell)$ for all $1 \leq i \leq n$, where the coefficients $\alpha(\ell)$ decay exponentially fast to zero as $\ell \to \infty$.

(C2) The functions g_k $(1 \leq k \leq K_0)$ 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, $0 < c \leq \min_{1 \leq i \leq n} \inf_{w \in [0,1]} f_i(w)$ and $\max_{1 \leq i \leq n} \sup_{w \in [0,1]} f_i(w) \leq C < \infty$ for some constants $0 < c \leq C < \infty$. 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

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

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

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

for any $\ell \geq \ell^*$ and a fixed constant $C < \infty$.

(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 $n/T \leq C$ for some constant $C < \infty$. The bandwidth h converges to zero such that $T^{1/2}h \to \infty$ and $T^{\delta}h \to 0$ for some small $\delta > 0$.

(C5) The kernel K 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 $|K(w) - K(w')| \leq L|w - w'|$.

We finally suppose that the weight function π in the definition of the L_2-distance in (2.1) is bounded and that its support is contained in the support of the regressors, that is, $\text{supp}(\pi) \subseteq [0, 1]$.

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 $n/T^{3/2} \leq C$, 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 $1 \leq k \leq K_0$. They only need to fulfill the trivial conditions that $n_k \leq n$ for $1 \leq k \leq K_0$ and $\sum_{k=1}^{K_0} n_k = n$. 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

To start with, we examine the asymptotic properties of our estimators of the classes $\{G_k : 1 \leq k \leq K_0\}$. According to the first theorem, the preliminary estimators $\{\widetilde{G}_k : 1 \leq k \leq K_0\}$ are consistent in the following sense: they coincide with the true classes $\{G_k : 1 \leq k \leq K_0\}$ with probability tending to one as the sample size grows.

Theorem 3.1. *Let (C1)–(C5) be satisfied. Then*

$$\mathbb{P}\Big(\big\{\widetilde{G}_k : 1 \leq k \leq K_0\big\} \neq \big\{G_k : 1 \leq k \leq K_0\big\}\Big) = o(1).$$

The second-step estimators $\{\widehat{G}_k : 1 \leq k \leq K_0\}$ can be shown to inherit this consistency property from the preliminary estimators $\{\widetilde{G}_k : 1 \leq k \leq K_0\}$.

Theorem 3.2. *Let (C1)–(C5) be satisfied. Then*

$$\mathbb{P}\Big(\big\{\widehat{G}_k : 1 \leq k \leq K_0\big\} \neq \big\{G_k : 1 \leq k \leq K_0\big\}\Big) = o(1).$$

Note that the indexing of the estimators $\widetilde{G}_1, \ldots, \widetilde{G}_{K_0}$ and $\widehat{G}_1, \ldots, \widehat{G}_{K_0}$ is completely arbitrary. We could, for example, change the indexing according to the rule $k \mapsto K_0 - k + 1$. In the sequel, we suppose that the estimated classes are indexed such that $\mathbb{P}(\widetilde{G}_k = G_k) \to 1$ and $\mathbb{P}(\widehat{G}_k = G_k) \to 1$ for all k with $1 \leq k \leq K_0$. Theorems 3.1 and 3.2 imply that this is possible without loss of generality.

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 depends on the cross-section dimension n in general, i.e., $n_k = n_k(n)$, we define the terms

$$B_k(w) = \frac{c_k^{5/2}}{2}\Big(\int K(\varphi)\varphi^2 d\varphi\Big) \lim_{n\to\infty}\Big(\frac{1}{n_k}\sum_{i\in G_k} \frac{g_k''(w)f_i(w) + 2g_k'(w)f_i'(w)}{f_i(w)}\Big)$$

$$V_k(w) = \Big(\int K^2(\varphi)d\varphi\Big) \lim_{n\to\infty}\Big(\frac{1}{n_k}\sum_{i\in G_k} \frac{\sigma_i^2(w)}{f_i(w)}\Big),$$

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

Theorem 3.3. *Let (C1)–(C5) be satisfied. Moreover, write $\widehat{n}_k = |\widehat{G}_k|$ and choose the bandwidth h such that $h/(\widehat{n}_k T)^{-1/5} \xrightarrow{P} c_k$ for some fixed constant $c_k > 0$. Then*

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

for any fixed $w \in (0,1)$.

Theorem 3.3 implies that the pointwise convergence rate of $\widehat{g}_k$ is $O_p(1/\sqrt{n_k T h})$, or put differently,

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

for any $w \in (0,1)$. This rate depends on the class size n_k. Specifically, the faster n_k grows to infinity, the faster the convergence rate of $\widehat{g}_k$. The reason for this is simple: By construction, $\widehat{g}_k$ essentially is an average of the individual smoothers $\widehat{m}_i$. If n_k is bounded, we only average over finitely many smoothers, implying that the convergence rate of $\widehat{g}_k$ is identical to that of the individual time series smoothers $\widehat{m}_i$. In particular, the rate equals $O_p(1/\sqrt{Th})$ in this case. If n_k tends to infinity in contrast, we average over infinitely many smoothers. As is well known from other nonparametric estimation

problems (see e.g. Linton (1997) or Wang and Yang (2007)), the averaging leads to a variance reduction in this case and thus helps to speed up the rate of $\widehat{g}_k$.

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 B.1 from the Appendix directly implies that under (C1)–(C5),

$$\sup_{w\in[0,1]} \big|\widehat{g}_k(w) - g_k(w)\big| = o_p(1).$$

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 $n_k = O(n)$ and that the bandwidth h is of the order $(nT)^{-(1/5+\delta)}$ for some small $\delta > 0$, this has been done in Körber et al. (2014b). Their results immediately imply that in this case,

$$\sup_{w\in I_h} \big|\widehat{g}_k(w) - g_k(w)\big| = O_p\Big(\sqrt{\frac{\log(n_k T)}{n_k T h}}\Big), \tag{3.2}$$

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 verify (3.2) under more general conditions on the size of n_k.

4 Estimating the Number of Classes K_0

So far, we have worked under the simplifying assumption that the number of classes K_0 is known. We now drop this assumption and take into account that K_0 is unknown in many applications. We only suppose that there is some known upper bound $\overline{K}$ on the number of classes, i.e., we take for granted that $K_0 \leq \overline{K}$. Importantly, our theoretical results of Section 3 remain to hold true when K_0 gets replaced by a consistent estimator. We now explain how to construct such an estimator.

Step 1: For each candidate number of classes K with $1 \leq K \leq \overline{K}$, construct a partition $\{\widehat{G}_k^{(K)} : 1 \leq k \leq K\}$ of the index set with K elements as described below. Moreover, define associated function estimates $\widehat{g}_k^{(K)}$ by $\widehat{g}_k^{(K)}(w) = |\widehat{G}_k^{(K)}|^{-1} \sum_{i\in\widehat{G}_k^{(K)}} \widehat{m}_i(w)$.

Step 2: Let $\Delta(\widehat{m}_i, \widehat{g}_k^{(K)})$ be the squared L_2-distance between the function $\widehat{m}_i$ and the centre $\widehat{g}_k^{(K)}$ of the k-th cluster. For each K with $1 \leq K \leq \overline{K}$, compute the average L_2-distance

$$\widehat{\Psi}(K) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i\in\widehat{G}_k^{(K)}} \Delta(\widehat{m}_i, \widehat{g}_k^{(K)}).$$

Step 3: Estimate K_0 by

$$\widehat{K}_0 = \min\big\{K \in \{1,\dots,\overline{K}\} : \widehat{\Psi}(K) \leq \rho_{n,T}\big\},$$

where $\rho_{n,T} \searrow 0$ at a rate specified in Theorem 4.1 below.

To construct the partitions in Step 1, we may simply run the estimation algorithm from Subsection 2.2 for each K with $1 \leq K \leq \overline{K}$. Computationally, this is however not very efficient. We thus proceed as follows: To start with, we run the algorithm for $\overline{K}$, which yields a partition $\{\widehat{G}_k^{(\overline{K})} : 1 \leq k \leq \overline{K}\}$ with $\overline{K}$ elements. Next, we calculate the distances $\Delta(\widehat{g}_k^{(\overline{K})}, \widehat{g}_{k'}^{(\overline{K})}) = \int [\widehat{g}_k^{(\overline{K})}(w) - \widehat{g}_{k'}^{(\overline{K})}(w)]^2 \pi(w) dw$ for each pair of indices $1 \leq k < k' \leq \overline{K}$ and take the pair corresponding to the minimal distance, say (k_1, k_2). We then replace the two groups $\widehat{G}_{k_1}^{(\overline{K})}$ and $\widehat{G}_{k_2}^{(\overline{K})}$ by their union, which yields a partition $\{\widehat{G}_k^{(\overline{K}-1)} : 1 \leq k \leq \overline{K}-1\}$ with $(\overline{K}-1)$ elements. We thus construct a partition with $(\overline{K}-1)$ elements by taking the union of the two classes whose centres are closest to each other. Iterating this construction principle, we obtain a partition $\{\widehat{G}_k^{(K)} : 1 \leq k \leq K\}$ for each K.

The heuristic idea behind the estimator $\widehat{K}_0$ is as follows: When $K \geq K_0$, the partition $\{\widehat{G}_k^{(K)} : 1 \leq k \leq K\}$ can be expected to estimate a refinement of the true class structure $\{G_k : 1 \leq k \leq K_0\}$. In particular, it should converge to a partition $\{G_k^{(K)} : 1 \leq k \leq K\}$ with the property that for any $\ell \in \{1\dots,K\}$ there exists $k \in \{1,\dots,K_0\}$ with $G_\ell^{(K)} \subseteq G_k$. This suggests that $\widehat{\Psi}(K)$ converges to zero for $K \geq K_0$. When $K < K_0$ in contrast, we are not able to consistently estimate the true class structure. In particular, at least one of the estimates $\widehat{G}_\ell^{(K)}$ should approximate the union of two or more classes G_k. Hence, $\widehat{\Psi}(K)$ can be expected to converge to some positive number rather than zero for $K < K_0$. Taken together, these considerations suggest that $\widehat{\Psi}(K)$ is bounded away from zero for $K < K_0$ but converges to zero for $K \geq K_0$. The estimator $\widehat{K}_0$, that is, the smallest K for which $\widehat{\Psi}(K)$ is close to zero should thus give a good approximation to K_0. This intuition is confirmed by the next result which shows that $\widehat{K}_0$ consistently estimates the true number of groups K_0.

Theorem 4.1. *Let (C1)–(C5) be satisfied and suppose that $n_k/n \to c_k$ for some constants $c_k > 0$ and all $1 \leq k \leq K_0$. Moreover, let $\rho_{n,T} \searrow 0$ with $\rho_{n,T} \geq c(nT)^{-1/(20+\eta)}$ for some $c > 0$ and a small $\eta > 0$, and choose the bandwidth h such that $h/\rho_{n,T} \to 0$. Then*

$$\mathbb{P}\big(\widehat{K}_0 \neq K_0\big) = o(1).$$

The proof of Theorem 4.1 easily follows with the help of the arguments from the Appendix. We are thus content with providing a brief outline: As stated in Theorem 4.1, we impose the additional condition that $n_k/n \to c_k > 0$ for $1 \leq k \leq K_0$, i.e., the class sizes n_k are all of the order $O(n)$ and thus do not differ too much. Under this condition, it can be shown that for each $K < K_0$, the average distance $\widehat{\Psi}(K)$ converges

in probability to a fixed positive number. As $\rho_{n,T} \searrow 0$, we can thus infer that

$$\begin{aligned}\mathbb{P}\big(\widehat{K}_0 < K_0\big) &= \mathbb{P}\Big(\widehat{\Psi}(K) \leq \rho_{n,T} \text{ for some } K < K_0\Big) \\ &\leq \sum_{k=1}^{K_0-1} \mathbb{P}\big(\widehat{\Psi}(K) \leq \rho_{n,T}\big) = o(1) \end{aligned} \tag{4.1}$$

as well as

$$\begin{aligned}\mathbb{P}\big(\widehat{K}_0 > K_0\big) &= \mathbb{P}\big(\widehat{\Psi}(K) > \rho_{n,T} \text{ for all } K \leq K_0\big) \\ &= \mathbb{P}\big(\widehat{\Psi}(K_0) > \rho_{n,T}\big) + o(1). \end{aligned} \tag{4.2}$$

Combining (4.1) and (4.2), we arrive at

$$\begin{aligned}\mathbb{P}(\widehat{K}_0 \neq K_0) &= \mathbb{P}(\widehat{K}_0 < K_0) + \mathbb{P}(\widehat{K}_0 > K_0) \\ &= \mathbb{P}\big(\widehat{\Psi}(K_0) > \rho_{n,T}\big) + o(1). \end{aligned} \tag{4.3}$$

Inspecting the proof of Theorems 3.1 and 3.2, it is not difficult to see that the partition $\{\widehat{G}_k^{(K_0)} : 1 \leq k \leq K_0\}$ consistently estimates the classes $\{G_k : 1 \leq k \leq K_0\}$. This allows us to infer that $\widehat{\Psi}(K_0) = o_p(1)$. More specifically, with the help of the arguments for Lemma B.1 in the Appendix, we can show that $\widehat{\Psi}(K_0) = O_p((nT)^{-1/(20+\delta)} + h)$ for some arbitrarily small $\delta > 0$. Setting $\delta = \eta/2$ without loss of generality, we thus obtain that $\mathbb{P}(\widehat{\Psi}(K_0) > \rho_{n,T}) = o(1)$. Combining this statement with (4.3) completes the proof of Theorem 4.1.

It goes without saying that the small sample performance of the estimator $\widehat{K}_0$ strongly hinges on the choice of the threshold parameter $\rho_{n,T}$. It is thus essential to pick $\rho_{n,T}$ in an appropriate way. In what follows, we provide some heuristic arguments on how to achieve this. To start with, we replace the average distance $\widehat{\Psi}(K_0)$ by the term

$$\Psi(K_0) = \frac{1}{n} \sum_{k=1}^{K_0} \sum_{i \in G_k} \Delta(\widehat{m}_i, g_k),$$

thus ignoring the estimation error in $\widehat{G}_k$ and $\widehat{g}_k$ for $1 \leq k \leq K_0$. Letting the bandwidth h converge to zero slightly faster than $T^{-2/9}$ and neglecting the time series dependence in the data, the arguments in Härdle and Mammen (1993) show that

$$Th^{1/2}\Delta(\widehat{m}_i, g_k) \xrightarrow{d} N(B_{i,h}, V_i) \tag{4.4}$$

with

$$B_{i,h} = h^{-1/2} C_B(K) \int \frac{\sigma_i^2(w)\pi(w)}{f_i(w)} dw$$

$$V_i = 2C_V(K) \int \frac{(\sigma_i^2(w))^2 \pi^2(w)}{f_i^2(w)} dw,$$

where $C_B(K) = \int K^2(w)dw$, $C_V(K) = \int(\int K(v)K(v+w)dv)^2 dw$ and $\sigma_i^2(w) = \mathbb{E}[\varepsilon_{it}^2 | X_{it} = w]$. Let us now suppose that the fixed effects α_i and γ_t are observed. In this case, we can replace the estimators $\widehat{m}_i$ by the infeasible versions $\widehat{m}_i^*$ defined in Subsection 2.1, which are independent across i. Moreover, assume that the cross-section dimension n is fixed. Under these conditions, (4.4) immediately implies that

$$Th^{1/2}\Psi(K_0) \xrightarrow{d} N(B_h, V), \tag{4.5}$$

where $B_h = n^{-1}\sum_{i=1}^n B_{i,h}$ and $V = n^{-2}\sum_{i=1}^n V_i$. Letting q_α be the $(1-\alpha)$-quantile of a normal distribution with mean zero and variance V and setting $\rho_{n,T} = (q_\alpha + B_h)/(Th^{1/2})$, we further obtain that

$$\mathbb{P}\big(\Psi(K_0) \leq \rho_{n,T}\big) = \mathbb{P}\big(Th^{1/2}\Psi(K_0) - B_h \leq q_\alpha\big) \to 1 - \alpha \tag{4.6}$$

by (4.5). With the help of (4.3), we may finally conclude that

$$\mathbb{P}(\widehat{K}_0 \neq K_0) \approx \mathbb{P}(\widehat{\Psi}(K_0) > \rho_{n,T}) \approx \mathbb{P}(\Psi(K_0) > \rho_{n,T}) \approx \alpha. \tag{4.7}$$

With the choice $\rho_{n,T} = (q_\alpha + B_h)/(Th^{1/2})$, we should thus be able to approximately control the estimation error in $\widehat{K}_0$. In particular, the probability that $\widehat{K}_0 \neq K_0$ should not be much more than α.

Clearly, (4.7) is not a theoretically rigorous result but is based on heuristic arguments which are subject to a number of simplifications. Nevertheless, it suggests that our estimator $\widehat{K}_0$ should perform reasonably well when implemented with the choice $\rho_{n,T} = (q_\alpha + B_h)/(Th^{1/2})$. In practice, of course, we cannot take this choice at face value but have to replace the expressions q_α and B_h by estimators. This can be achieved by replacing the unknown functions σ_i and f_i in the bias and variance terms B_h and V by standard kernel estimators. Denoting the resulting estimators of q_α and B_h by $\widehat{q}_\alpha$ and $\widehat{B}_h$, respectively, we propose to choose $\rho_{n,T} = (\widehat{q}_\alpha + \widehat{B}_h)/(Th^{1/2})$. The simulations in the next section suggest that this choice of $\rho_{n,T}$ yields an estimator $\widehat{K}_0$ with good finite sample properties.

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

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

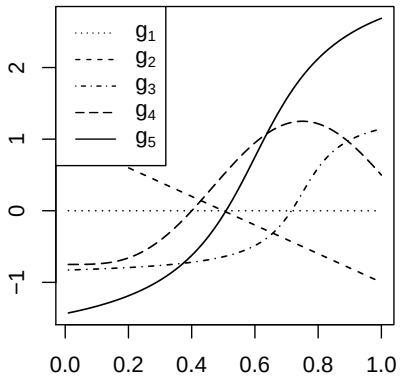

Figure 1: Plot of the functions g_k for $1 \leq k \leq 5$.

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 associated with these groups are $g_1(w) = 0$, $g_2(w) = 1 - 2w$, $g_3(w) = 0.75 \arctan(10(w - 0.75)) + 0.25$, $g_4(w) = 2\vartheta((w - 0.75)/0.75) - 0.75$ with $\vartheta(w) = (1 - w^2)^4 1(|w| \leq 1)$ and $g_5(w) = 1.75 \arctan(5(u - 0.6)) + 0.75$. Figure 1 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.

In our first simulation exercise, we treat the number of classes $K_0 = 5$ as known and focus on the estimation of the class structure $\{G_k : 1 \leq k \leq K_0\}$. To compute the estimates $\widehat{m}_i$, we work with an Epanechnikov kernel and the bandwidth $h = 0.25$, which is used 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. 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 estimates $\{\widehat{G}_k : 1 \leq k \leq K_0\}$. In order to measure how well these estimates fit the real class structure $\{G_k : 1 \leq k \leq K_0\}$, we calculate the number of wrongly classified

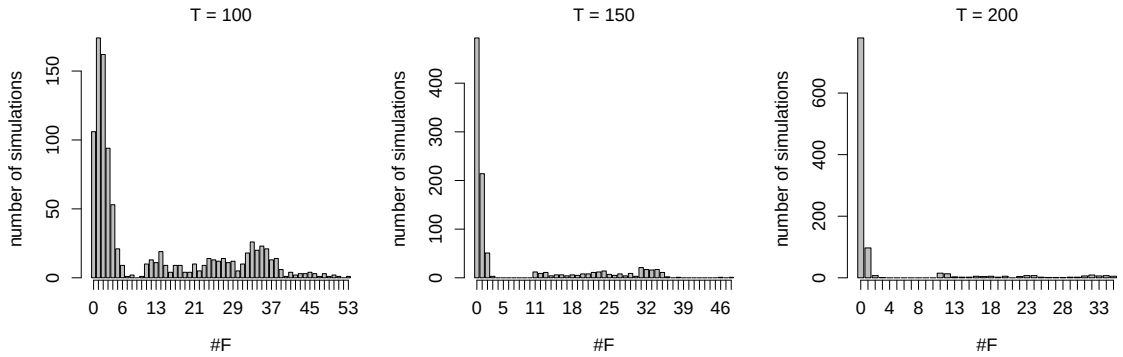

Figure 2: Simulation results for the estimation of the classes $G_1, \ldots, G_5$. The three plots show the distributions of the number $\#F$ of wrong classifications for the samples sizes (n, T) with $n = 120$ and $T \in \{100, 150, 200\}$.

indices i, which is denoted by $\#F$ in what follows. For each sample size (n, T), we thus obtain $N = 1000$ values for the number $\#F$ of wrong classifications. Figure 2 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.

Inspecting the plots of Figure 2, our estimators $\{\widehat{G}_k : 1 \leq k \leq K_0\}$ can be seen to approximate the group structure quite 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 classification 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 only a few wrong classifications in about 50–60% 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 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.

	$T = 100$	$T = 150$	$T = 200$
$\widehat{K}_0 = 5$	811	881	912
$\widehat{K}_0 = 6$	114	80	59
$\widehat{K}_0 = 7$	46	28	27
$\widehat{K}_0 = 8$	3	9	2
$\widehat{K}_0 = 9$	3	0	0
$\widehat{K}_0 = 10$	23	2	0

Table 1: Simulation results for the estimation of K_0. The entries in the table specify the number of simulations in which a certain value of $\widehat{K}_0$ is obtained.

Having investigated the performance of our estimators $\{\widehat{G}_k : 1 \leq k \leq K_0\}$ for a given number of classes K_0, we now examine the finite sample properties of the estimator $\widehat{K}_0$. To do so, we implement the estimator as described in Section 4, where we set $\overline{K} = 10$ and $\rho_{n,T} = (\widehat{q}_\alpha + \widehat{B}_h)/(Th^{1/2})$ with $\alpha = 0.05$. As before, we drawn $N = 1000$ samples for each model specification and calculate the estimate $\widehat{K}_0$ for each simulated sample. The simulation results are presented in Table 1. They suggest that the estimator $\widehat{K}_0$ performs reasonably well in small samples. Already for the smallest time series length $T = 100$, it selects the true number of classes $K_0 = 5$ in around 80% of the simulations. This value can be seen to improve as the sample size increases. For $T = 200$, it is around 91% and thus comes close to a level of 95%, which is the level predicted by our heuristic considerations in Section 4, in particular by (4.7).

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 Körber 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 Körber 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

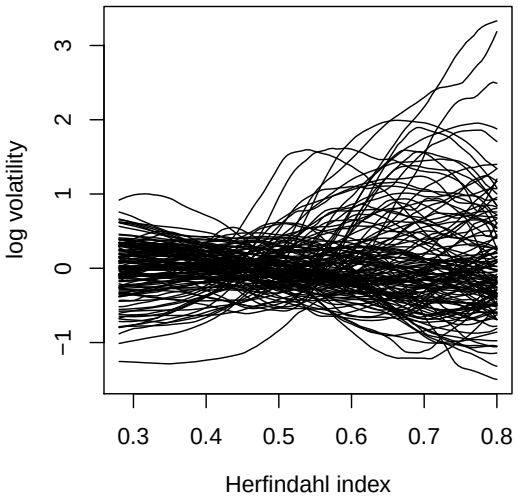

Figure 3: Estimates $\widehat{m}_i$ for the $n = 127$ stocks in our sample.

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 exist K_0 classes of stocks $G_1, \ldots, G_{K_0}$ along with associated functions $g_1, \ldots, g_{K_0}$ such that $m_i = g_k$ for all $i \in G_k$ and all $1 \leq k \leq K_0$. 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_0 and to estimate the groups G_k along with the functions g_k for $1 \leq k \leq K_0$, we employ the estimation techniques developed in the previous sections. As in the simulations, we use an Epanechnikov kernel to compute the Nadaraya-Watson smoothers $\widehat{m}_i$. Prior to estimation, we eliminate stocks i with a very small empirical support $\mathcal{S}_i$ of the fragmentation data $\{X_{it} : 1 \leq t \leq T\}$. In particular, we only take into account stocks i for which the support $\mathcal{S}_i$ contains the interval $[0.275, 0.8]$. This leaves us with $n = 127$ 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. To start with, we estimate the number of classes K_0 by means of the procedure from Section 4, where we set $\overline{K} = 10$ and $\rho_{n,T} = (\widehat{q}_\alpha + \widehat{B}_h)/(Th^{1/2})$ with $\alpha = 0.05$ as in the simulations. We compute the estimate $\widehat{K}_0$ for various bandwidths h, in particular for $h \in \{0.2, 0.225, 0.25, 0.275, 0.3\}$. For all

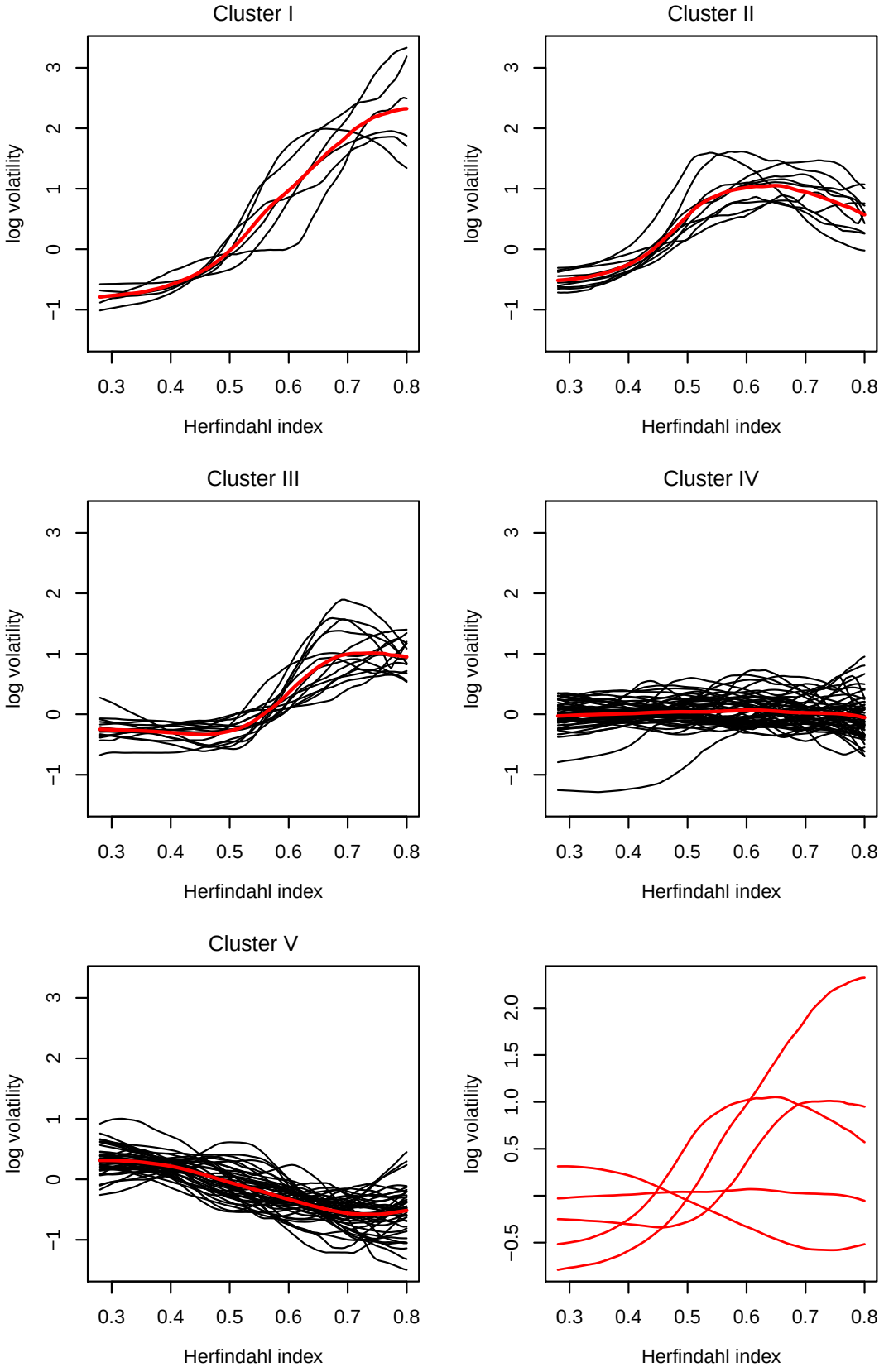

Figure 4: Clusters of the curve estimates $\widehat{m}_i$. The black lines are the estimates $\widehat{m}_i$, the red lines the estimates $\widehat{g}_k$. The latter are once again plotted together in the lower right panel.

these bandwidths except for $h = 0.2$, we obtain the estimate $\widehat{K}_0 = 5$, whereas $\widehat{K}_0 = 4$ for $h = 0.2$. This indicates that the group structure imposed by our model yields a good fit of the data when the number of classes equals 4 or 5. In what follows, we only present the results for 5 classes, those for 4 groups giving a similar picture and suggesting essentially the same interpretation.

Setting the number of groups to 5, we apply our procedure from Section 2 to cluster the estimated curves $\widehat{m}_i$ into groups. Figure 3 depicts the estimates $\widehat{m}_i$ for the $n = 127$ stocks in our sample. Figure 4 shows the clusters produced by our procedure. Each panel of Figure 4 depicts the estimates which belong to a particular class $\widehat{G}_k$. The corresponding estimates $\widehat{g}_k$ are indicated by the solid red curves and are once again plotted together in the lower right panel of the figure. The estimates $\widehat{m}_i$ in Figures 3 and 4 are computed with the bandwidth $h = 0.25$. As a robustness check, we have repeated the estimation with the bandwidths $h \in \{0.2, 0.225, 0.25, 0.275, 0.3\}$. As this yields similar clusters as in the case with $h = 0.25$, we do not report the results.

Inspecting Figure 4, 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 level 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 4 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. Cluster III) or deteriorate (cp. Cluster V) market quality.

7 Extensions

Our estimation approach may be extended in various directions. We close the paper by outlining some of them.

7.1 The fixed design case

So far, we have focused on the case of stochastic covariates X_{it}. In a variety of applications, however, we are interested in a design with deterministic regressors. A particularly important case arises when X_{it} is (rescaled) time, that is, $X_{it} = t/T$. In

this case, the model equation (1.1) becomes

$$Y_{it} = m_i\Big(\frac{t}{T}\Big) + u_{it}, \tag{7.1}$$

where m_i are unknown nonparametric time trend functions. In many applications, the data do not only exhibit a trending but also a seasonal behaviour over time. For this reason, we enrich (7.1) by a seasonal component, yielding the model

$$Y_{it} = s_i(t) + m_i\Big(\frac{t}{T}\Big) + u_{it}. \tag{7.2}$$

Here, s_i is the seasonal component of the time series of the i-th individual. In particular, $(s_i(1), \dots, s_i(T))$ is a periodic sequence with a known integer-valued period p, that is, $s_i(t) = s_i(t + \ell p)$ for all ℓ and t.

As in the random design case, we suppose that there is a finite number of groups $G_1, \dots, G_{K_0}$ such that $m_i = g_k$ for all $i \in G_k$ and $1 \leq k \leq K_0$. The error terms u_{it} are assumed to split into two components, $u_{it} = \alpha_i + \varepsilon_{it}$, where ε_{it} are idiosyncratic error terms satisfying $\mathbb{E}[\varepsilon_{it}] = 0$ and α_i are unobserved individual specific fixed effects. Unlike in the random design, we do not include a time fixed effect in the error structure, because the trending behaviour of the data over time is now explicitly modelled by the functions m_i. As before, the time series dimension T is assumed to tend to infinity, whereas the cross-section dimension n may either be fixed or diverging. To identify the functions m_i in (7.2), we normalize them to satisfy $\int_0^1 m_i(w)dw = 0$ for all i. This is sufficient for identification by Lemma A2 in Vogt and Linton (2014).

To estimate the groups G_k and the functions g_k for $1 \leq k \leq K_0$, we proceed in the same way as in the random design case. The only difference is that the estimators of the trend functions m_i are not exactly the same as those of the regression functions in the random design. To construct estimators of the trend functions, we define $Y_{it}^{\text{fe}} = Y_{it} - \nu_i(t)$ with $\nu_i(t) = s_i(t) + \alpha_i$. These variables can be approximated by $\widehat{Y}_{it}^{\text{fe}} = Y_{it} - \widehat{\nu}_i(t)$, where

$$\widehat{\nu}_i(t) = \frac{1}{L_t} \sum_{\ell=1}^{L_t} Y_{i,t_p+(\ell-1)p}$$

with $t_p = t - \lfloor \frac{t-1}{p} \rfloor p$ and $L_t = 1 + \lfloor \frac{T-t_p}{p} \rfloor$. With this notation at hand, we define Nadaraya-Watson type estimators of the functions m_i by

$$\widehat{m}_i(w) = \frac{\sum_{t=1}^T K_h(\frac{t}{T} - w)\widehat{Y}_{it}^{\text{fe}}}{\sum_{t=1}^T K_h(\frac{t}{T} - w)}.$$

Replacing the smoothers of Subsection 2.1 by those defined above, the groups G_k and the associated functions g_k can be estimated exactly as described in Subsections 2.2 and 2.3. As in the random design case, the resulting estimators are denoted by $\widetilde{G}_k$, $\widehat{G}_k$ and $\widehat{g}_k$.

To derive asymptotic results analogous to those from Section 3, we impose the following conditions.

(C'1) The time series $\mathcal{Z}_i = \{\varepsilon_{it} : 1 \leq t \leq T\}$ 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 $\alpha_i(\ell) \leq \alpha(\ell)$ for all $1 \leq i \leq n$, where the coefficients $\alpha(\ell)$ decay exponentially fast to zero as $\ell \to \infty$.

(C'2) The functions g_k $(1 \leq k \leq K_0)$ are twice continuously differentiable on $[0, 1]$.

(C'3) There exists a real number $\theta > 4$ such that $\mathbb{E}[|\varepsilon_{it}|^\theta] \leq C < \infty$.

Under conditions (C'1)–(C'3) along with (C4)–(C5), our estimators of the groups G_k can be shown to be consistent in the sense that

$$\mathbb{P}\Big(\big\{\widetilde{G}_k : 1 \leq k \leq K_0\big\} \neq \big\{G_k : 1 \leq k \leq K_0\big\}\Big) = o(1) \tag{7.3}$$

$$\mathbb{P}\Big(\big\{\widehat{G}_k : 1 \leq k \leq K_0\big\} \neq \big\{G_k : 1 \leq k \leq K_0\big\}\Big) = o(1). \tag{7.4}$$

The proof of these two results is completely analogous to that of Theorems 3.1 and 3.2 for the random design. It is worth mentioning that (7.3) and (7.4) remain to hold true when we drop the independence assumption on the time series $\mathcal{Z}_i$ and allow them to be dependent across i in an arbitrary way. This is possible because we do not include a time fixed effect in the error structure. The time fixed effect γ_t in the random design is essentially approximated by the cross-sectional average $\overline{Y}_t^{(i)} = \frac{1}{n-1}\sum_{j \neq i} Y_{jt}$ in our estimation approach. To control the behaviour of this average asymptotically, we have to impose some restrictions on the dependence of the time series $\mathcal{Z}_i$ across i. When dropping γ_t from the model, the averages $\overline{Y}_t^{(i)}$ are not needed any more and our technical arguments, in particular those for Lemma B.1 in the Appendix, go through without any such restrictions on the dependence structure.

We next derive the limit distribution of the estimators $\widehat{g}_k$. By arguments analogous to those for Theorem 3.3, we can prove the following result: Let (C'1)–(C'3) along with (C4)–(C5) be satisfied and suppose that the bandwidth h is such that $h/(\widehat{n}_k T)^{-1/5} \xrightarrow{P} c_k$ for some fixed constant $c_k > 0$. Then

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

for any fixed $w \in (0, 1)$, where the asymptotic bias and variance terms are given by

$$B_k(w) = \frac{c_k^{5/2}}{2}\Big(\int K(\varphi)\varphi^2 d\varphi\Big) g_k''(w)$$

$$V_k(w) = \Big(\int K^2(\varphi) d\varphi\Big) \lim_{n\to\infty} \Big(\frac{1}{n_k}\sum_{i \in G_k}\sum_{\ell=-\infty}^{\infty} \mathrm{Cov}(\varepsilon_{it}, \varepsilon_{it+\ell})\Big).$$

We finally mention that the number of classes K_0 can be estimated by techniques as discussed in Section 4. For reasons of brevity, we neglect the details.

7.2 Additive models

Our estimation techniques are not only useful in a nonparametric context but easily carry over to semi- and structured nonparametric settings. As an example, consider the additive model

$$Y_{it} = \sum_{j=1}^{d} m_{i,j}(X_{it,j}) + u_{it}, \tag{7.5}$$

where $X_{it} = (X_{it,1}, \ldots, X_{it,d})^\top$, $m_{i,j}$ are nonparametric functions for $1 \leq j \leq d$, and the error terms u_{it} have the fixed effects structure $u_{it} = \alpha_i + \gamma_t + \varepsilon_{it}$ with $\mathbb{E}[\varepsilon_{it}|X_{it}] = 0$. Suppose we are mainly interested in the effect of $X_{it,1}$ on the response Y_{it}, which is captured by the functions $m_{i,1}$. As in our nonparametric framework (1.1), we may model this effect by imposing a group structure on the curves $m_{i,1}$ in (7.5). In particular, we may suppose that there exist classes $G_1, \ldots, G_{K_0}$ and associated functions $g_1, \ldots, g_{K_0}$ such that $m_{i,1} = g_k$ for all $i \in G_k$ and $1 \leq k \leq K_0$. Our estimation procedures can be applied almost unchanged in this additive context. We only have to replace the Nadaraya-Watson smoothers $\widehat{m}_i$ of Section 2 by more complicated estimators of $m_{i,1}$ that take into account the additive structure of (7.5).

Appendix A

We now provide the proofs of Theorems 3.1–3.3. Throughout the Appendix, the symbol C denotes a universal real constant which may take a different value on each occurrence. For the proofs of the theorems, we require some auxiliary results on the uniform convergence of kernel estimators which are derived in Appendix B.

Proof of Theorem 3.1

Let S be a subset of $\{1, \ldots, n\}$ with cardinality $|S| = n_S$ and suppose that S contains elements from at least two different classes G_k and $G_{k'}$. For each $i \in S$, let

$$\Delta_{i(1)} \leq \ldots \leq \Delta_{i(n_S)}$$

be the ordered distances $\{\Delta_{ij} : j \in S\}$ and write

$$\widehat{\Delta}_{i[1]} \leq \ldots \leq \widehat{\Delta}_{i[n_S]}$$

to denote the ordered estimates $\{\widehat{\Delta}_{ij} : j \in S\}$. We here use the two different symbols $(\,\cdot\,)$ and $[\,\cdot\,]$ to distinguish between the orderings of the true and the estimated distances.

Next, let $j_{\max}(i), \widehat{j}_{\max}(i) \in S$ be indices with the property that

$$j_{\max}(i) = \arg \max_{2 \le j \le n_S} \big|\Delta_{i(j)} - \Delta_{i(j-1)}\big|$$
$$\widehat{j}_{\max}(i) = \arg \max_{2 \le j \le n_S} \big|\widehat{\Delta}_{i[j]} - \widehat{\Delta}_{i[j-1]}\big|.$$

For simplicity, we assume that the index $j_{\max}(i)$ is unique for each i. This helps us to keep the proof as clear as possible. In particular, it avoids some cumbersome case distinctions in what follows. Finally, we introduce the sets

$$\begin{aligned} S_<(i) &= \big\{(1), \ldots, (j_{\max}(i) - 1)\big\} \quad \text{and} \quad S_>(i) = \big\{(j_{\max}(i)), \ldots, (n_S)\big\}, \\ \widehat{S}_<(i) &= \big\{[1], \ldots, [\widehat{j}_{\max}(i) - 1]\big\} \quad \text{and} \quad \widehat{S}_>(i) = \big\{[\widehat{j}_{\max}(i)], \ldots, [n_S]\big\} \end{aligned} \tag{A.1}$$

for any $i \in S$.

When applied to the true distances $\{\Delta_{ij} : j \in S\}$, the algorithm (AL1)–(AL3) partitions S into two subsets $S_<$ and $S_>$. When applied to the estimated distances $\{\widehat{\Delta}_{ij} : j \in S\}$, it splits S into two subsets which we denote by $\widehat{S}_<$ and $\widehat{S}_>$ to distinguish them from $S_<$ and $S_>$. By construction, the subsets S_ℓ and $\widehat{S}_\ell$ ($\ell \in \{<, >\}$) are closely related to the sets defined in (A.1). In particular, it holds that $S_\ell = S_\ell(i)$ and $\widehat{S}_\ell = \widehat{S}_\ell(i)$ for some $i \in S$. In the sequel, we show that

$$\mathbb{P}\big(\widehat{S}_\ell(i) = S_\ell(i) \text{ for all } i \in S\big) \to 1 \tag{A.2}$$

for $\ell \in \{<, >\}$. Hence, with probability tending to one, the algorithm (AL1)–(AL3) partitions S in the same way when applied to the true and the estimated L_2-distances. As the algorithm (AL1)–(AL3) is repeated only finitely many times in the course of our estimation procedure, this immediately implies that with probability tending to one, the two partitions $\{\widetilde{G}_k : 1 \le k \le K_0\}$ and $\{G_k : 1 \le k \le K_0\}$ are identical. This completes the proof of Theorem 3.1. □

Proof of (A.2). With the help of Lemma B.1 from Appendix B, we can show that

$$\max_{i,j \in S} \big|\widehat{\Delta}_{ij} - \Delta_{ij}\big| = o_p(1), \tag{A.3}$$

or put differently,

$$\mathbb{P}\Big(\max_{i,j \in S} \big|\widehat{\Delta}_{ij} - \Delta_{ij}\big| > \frac{\delta}{4}\Big) = o(1) \tag{A.4}$$

for any $\delta > 0$, in particular for $\delta_0 = \min_{i \in S} |\Delta_{i(j_{\max}(i))} - \Delta_{i(j_{\max}(i)-1)}|$. By definition of δ_0, it holds that $\Delta_{i(j)} + \delta_0 \le \Delta_{i(j_{\max}(i))} \le \Delta_{i(j')}$ for any pair j, j' with $j < j_{\max}(i) \le j'$ and any $i \in S$. By applying (A.4) with $\delta = \delta_0$, we can thus infer that with probability tending to one, the following holds: If j is an index in $S_\ell(i)$ for some $i \in S$ and $\ell \in \{<, >\}$, then it must also be an index in $\widehat{S}_\ell(i)$. As a result, $\widehat{S}_\ell(i) = S_\ell(i)$ for any $i \in S$ and $\ell \in \{<, >\}$ with probability approaching one. □

Proof of Theorem 3.2

With the help of Lemma B.1 from Appendix B, it is straightforward to see that

$$\int \big(\widehat{m}_i(w) - \widehat{\overline{m}}_k^{[0]}\big)^2 \pi(w)dw = \int \big(m_i(w) - g_k(w)\big)^2 \pi(w)dw + o_p(1)$$

uniformly over i and k, or put differently,

$$\max_{1\le k\le K_0} \max_{1\le i\le n} \big|\Delta(\widehat{m}_i, \widehat{\overline{m}}_k^{[0]}) - \Delta(m_i, g_k)\big| = o_p(1). \tag{A.5}$$

By construction, the index i is assigned to the group $S_k^{[0]}$ in the first step of the k-means algorithm if $\widehat{d}_k(i) = \Delta(\widehat{m}_i, \widehat{\overline{m}}_k^{[0]})$ is minimal, i.e., if $\widehat{d}_k(i) = \min_{1\le k'\le K_0} \widehat{d}_{k'}(i)$. By (A.5), we know that

$$\widehat{d}_k(i) = \begin{cases} r_k(i) & \text{if } i \in G_k \\ \Delta(m_i, g_k) + r_k(i) & \text{if } i \notin G_k, \end{cases} \tag{A.6}$$

where the remainder term $r_k(i)$ has the property that $\max_{1\le k\le K_0} \max_{1\le i\le n} |r_k(i)| = o_p(1)$. Since $\min_{1\le k\le K_0} \min_{i\notin G_k} \Delta(m_i, g_k) \ge \Delta_{\min} > 0$ for some positive constant $\Delta_{\min}$, (A.6) implies that

$$\mathbb{P}\Big(\big\{S_k^{[0]} : 1 \le k \le K_0\big\} \ne \big\{G_k : 1 \le k \le K_0\big\}\Big) = o(1).$$

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 $1 \le k \le K_0$. □

Proof of Theorem 3.3

In a first step, we replace the estimator $\widehat{g}_k$ by the infeasible version

$$\widehat{g}_k^*(w) = \frac{1}{n_k} \sum_{i\in G_k} \widehat{m}_i(w)$$

and show that the difference between the two estimators is asymptotically negligible: For any null sequence $\{a_{n,T}\}$ of positive numbers, it holds that

$$\begin{aligned} &\mathbb{P}\Big(\big|\widehat{g}_k(w) - \widehat{g}_k^*(w)\big| > a_{n,T}\Big) \\ &\quad \le \mathbb{P}\Big(\big|\widehat{g}_k(w) - \widehat{g}_k^*(w)\big| > a_{n,T}, \widehat{G}_k = G_k\Big) + \mathbb{P}\big(\widehat{G}_k \ne G_k\big) = o(1), \end{aligned}$$

since the first probability on the right-hand side is equal to zero by definition of $\widehat{g}_k$ and $\widehat{g}_k^*$ and the second one is of the order $o(1)$ by Theorem 3.2. Hence, $|\widehat{g}_k(w) - \widehat{g}_k^*(w)| =$

$O_p(a_{n,T})$ for any null sequence $\{a_{n,T}\}$ of positive numbers, which in turn implies that

$$\sqrt{\widehat{n}_k Th}\big(\widehat{g}_k(w) - g_k(w)\big) = \sqrt{\widehat{n}_k Th}\big(\widehat{g}_k^*(w) - g_k(w)\big) + o_p(1).$$

The difference between $\widehat{g}_k$ and $\widehat{g}_k^*$ can thus be asymptotically ignored.

To complete the proof of Theorem 3.3, we derive the limit distribution of the term $\sqrt{\widehat{n}_k Th}(\widehat{g}_k^*(w) - g_k(w))$: Since $\mathbb{P}(\widehat{n}_k \neq n_k) = o(1)$ by Theorem 3.2, it holds that $\sqrt{\widehat{n}_k Th}(\widehat{g}_k^*(w) - g_k(w)) = \sqrt{n_k Th}(\widehat{g}_k^*(w) - g_k(w)) + o_p(1)$. It thus suffices to compute the limit distribution of $\sqrt{n_k Th}(\widehat{g}_k^*(w) - g_k(w))$. To do so, write

$$\widehat{m}_i(w) - m_i(w) = \big[Q_{i,V}(w) + Q_{i,B}(w) - Q_{i,\gamma}(w)\big]\big/\widehat{f}_i(w) - \overline{Q}_i + \overline{\overline{Q}}_i,$$

where

$$\begin{aligned}
Q_{i,V}(w) &= \frac{1}{T}\sum_{t=1}^{T} K_h(X_{it} - w)\varepsilon_{it} \\
Q_{i,B}(w) &= \frac{1}{T}\sum_{t=1}^{T} K_h(X_{it} - w)\big[m_i(X_{it}) - m_i(w)\big] \\
Q_{i,\gamma}(w) &= \frac{1}{T}\sum_{t=1}^{T} K_h(X_{it} - w)\Big(\frac{1}{n-1}\sum_{j\neq i}\big[m_j(X_{jt}) + \varepsilon_{jt}\big]\Big) \\
\overline{Q}_i &= \frac{1}{T}\sum_{t=1}^{T}\big[m_i(X_{it}) + \varepsilon_{it}\big] \\
\overline{\overline{Q}}_i &= \frac{1}{(n-1)T}\sum_{j\neq i}\sum_{t=1}^{T}\big[m_j(X_{jt}) + \varepsilon_{jt}\big]
\end{aligned}$$

and $\widehat{f}_i(w) = T^{-1}\sum_{t=1}^{T} K_h(X_{it} - w)$. With this notation at hand, we obtain that

$$\begin{aligned}
&\sqrt{n_k Th}\,\big(\widehat{g}_k^*(w) - g_k(w)\big) \\
&= \sqrt{n_k Th}\Big\{\frac{1}{n_k}\sum_{i\in G_k}\frac{Q_{i,V}(w)}{\widehat{f}_i(w)} + \frac{1}{n_k}\sum_{i\in G_k}\frac{Q_{i,B}(w)}{\widehat{f}_i(w)} - \frac{1}{n_k}\sum_{i\in G_k}\frac{Q_{i,\gamma}(w)}{\widehat{f}_i(w)} \\
&\qquad\qquad - \frac{1}{n_k}\sum_{i\in G_k}\overline{Q}_i + \frac{1}{n_k}\sum_{i\in G_k}\overline{\overline{Q}}_i\Big\} \\
&= \sqrt{n_k Th}\,\Big\{\frac{1}{n_k}\sum_{i\in G_k}\frac{Q_{i,V}(w)}{\widehat{f}_i(w)} + \frac{1}{n_k}\sum_{i\in G_k}\frac{Q_{i,B}(w)}{\widehat{f}_i(w)} - \frac{1}{n_k}\sum_{i\in G_k}\frac{Q_{i,\gamma}(w)}{\widehat{f}_i(w)}\Big\} + o_p(1),
\end{aligned}$$

the last line following by standard calculations. In the sequel, we show that

$$\frac{1}{n_k}\sum_{i\in G_k}\frac{Q_{i,\gamma}(w)}{\widehat{f}_i(w)} = o_p\Big(\frac{1}{\sqrt{n_k Th}}\Big) \tag{A.7}$$

$$\frac{1}{n_k}\sum_{i\in G_k}\frac{Q_{i,V}(w)}{\widehat{f}_i(w)} = \frac{1}{n_k}\sum_{i\in G_k}\frac{Q_{i,V}(w)}{f_i(w)} + o_p\Big(\frac{1}{\sqrt{n_kTh}}\Big) \tag{A.8}$$

$$\frac{1}{n_k}\sum_{i\in G_k}\frac{Q_{i,B}(w)}{\widehat{f}_i(w)} = \frac{1}{n_k}\sum_{i\in G_k}\frac{Q_{i,B}(w)}{f_i(w)} + o_p\Big(\frac{1}{\sqrt{n_kTh}}\Big). \tag{A.9}$$

(A.7)–(A.9) allow us to conclude that

$$\begin{aligned}
&\sqrt{n_kTh}\,\big(\widehat{g}_k^*(w) - g_k(w)\big)\\
&= \sqrt{n_kTh}\Big\{\frac{1}{n_k}\sum_{i\in G_k}\frac{Q_{i,V}(w)}{f_i(w)} + \frac{1}{n_k}\sum_{i\in G_k}\frac{Q_{i,B}(w)}{f_i(w)}\Big\} + o_p\Big(\frac{1}{\sqrt{n_kTh}}\Big)\\
&= \sqrt{n_kTh}\Big(\frac{1}{n_kT}\sum_{i\in G_k}\sum_{t=1}^{T}\frac{K_h(X_{it}-w)}{f_i(w)}\varepsilon_{it}\Big)\\
&\quad + \sqrt{n_kTh}\Big(\frac{1}{n_kT}\sum_{i\in G_k}\sum_{t=1}^{T}\frac{K_h(X_{it}-w)}{f_i(w)}\big[m_i(X_{it}) - m_i(w)\big]\Big) + o_p(1).
\end{aligned}$$

With the help of a standard central limit theorem, the first term on the right-hand side can be shown to weakly converge to a normal distribution with mean zero and variance $V_k(w)$. Moreover, standard bias calculations yield that the second term converges in probability to the bias expression $B_k(w)$. This completes the proof. □

Proof of (A.7). In a first step, we show that

$$R_\gamma := \frac{1}{n_k}\sum_{i\in G_k}\frac{Q_{i,\gamma}(w)}{\widehat{f}_i(w)} - \frac{1}{n_k}\sum_{i\in G_k}\frac{Q_{i,\gamma}(w)}{\mathbb{E}[\widehat{f}_i(w)]} = o_p\Big(\frac{1}{\sqrt{n_kTh}}\Big). \tag{A.10}$$

To do so, we write $R_\gamma = R_{\gamma,1} + R_{\gamma,2}$, where

$$\begin{aligned}
R_{\gamma,1} &= \frac{1}{n_k}\sum_{i\in G_k}\frac{\mathbb{E}[\widehat{f}_i(w)] - \widehat{f}_i(w)}{\mathbb{E}[\widehat{f}_i(w)]^2}Q_{i,\gamma}(w)\\
R_{\gamma,2} &= \frac{1}{n_k}\sum_{i\in G_k}\frac{(\mathbb{E}[\widehat{f}_i(w)] - \widehat{f}_i(w))^2}{\mathbb{E}[\widehat{f}_i(w)]^2\,\widehat{f}_i(w)}Q_{i,\gamma}(w).
\end{aligned}$$

Defining $Z_{it}(w) = \mathbb{E}[K_h(X_{it}-w)] - K_h(X_{it}-w)$, the first term $R_{\gamma,1}$ can be expressed as

$$\begin{aligned}
R_{\gamma,1} = \frac{1}{n_k}\sum_{i\in G_k}\frac{1}{\mathbb{E}[\widehat{f}_i(w)]^2}\Big\{\frac{1}{T}\sum_{t=1}^{T}Z_{it}(w)\Big\}\\
\times\Big\{\frac{1}{T}\sum_{t=1}^{T}K_h(X_{it}-w)\Big(\frac{1}{n-1}\sum_{j\neq i}\big[m_j(X_{jt})+\varepsilon_{jt}\big]\Big)\Big\}.
\end{aligned}$$

We thus obtain that

$$\mathbb{E}[R_{\gamma,1}^2] = \frac{1}{n_k^2(n-1)^2} \sum_{i,i' \in G_k} \sum_{\substack{j \neq i \\ j' \neq i'}} \frac{1}{\mathbb{E}[\widehat{f}_i(w)]^2} \frac{1}{\mathbb{E}[\widehat{f}_{i'}(w)]^2} \times \Big(\frac{1}{T^4} \sum_{t,t',s,s'=1}^{T} \Psi_{i,i',j,j',t,t',s,s'}(w) \Big), \tag{A.11}$$

where we use the shorthand

$$\Psi_{i,i',j,j',t,t',s,s'}(w) = \mathbb{E}\big[Z_{it}(w) K_h(X_{is} - w) \big\{ m_j(X_{js}) + \varepsilon_{js} \big\} \times Z_{i't'}(w) K_h(X_{i's'} - w) \big\{ m_{j'}(X_{j's'}) + \varepsilon_{j's'} \big\} \big].$$

Importantly, the expressions $\Psi_{i,i',j,j',t,t',s,s'}(w)$ in (A.11) have the following property: $\Psi_{i,i',j,j',t,t',s,s'}(w) \neq 0$ only if (a) $i = j'$ and $i' = j$ or (b) $j = j'$. Exploiting the mixing conditions of (C1) by means of Davydov's inequality (see Corollary 1.1 in Bosq (1998)), we can show that in case (a), $|T^{-4} \sum_{t,t',s,s'=1}^{T} \psi_{i,i',j,j',t,t',s,s'}(w)| \leq C(\log T)^2/(Th)^2$ and in case (b), $|T^{-4} \sum_{t,t',s,s'=1}^{T} \psi_{i,i',j,j',t,t',s,s'}(w)| \leq C(\log T)^3/(Th)^3$. Plugging these bounds into (A.11), we immediately arrive at $R_{\gamma,1} = o_p(1/\sqrt{n_k Th})$. Furthermore, with the help of Hölder's inequality and Lemma B.2, we obtain that

$$\begin{aligned} R_{\gamma,2} &\leq \Big\{ \max_{1 \leq i \leq n} \sup_{w \in [0,1]} \frac{(\mathbb{E}[\widehat{f}_i(w)] - \widehat{f}_i(w))^2}{\mathbb{E}[\widehat{f}_i(w)]^2 \widehat{f}_i(w)} \Big\} \Big\{ \frac{1}{n_k} \sum_{i \in G_k} \Big(\frac{1}{T} \sum_{t=1}^{T} K_h^{4/3}(X_{it} - w) \Big)^{3/4} \\ &\quad \times \Big(\frac{1}{T} \sum_{t=1}^{T} \Big(\frac{1}{n-1} \sum_{j \neq i} \big[m_j(X_{jt}) + \varepsilon_{jt} \big] \Big)^4 \Big)^{1/4} \Big\} \\ &= O_p\Big(\Big(\sqrt{\frac{\log T}{Th}} \Big)^2 \frac{1}{h^{1/4}(n-1)^{1/2}} \Big) = o_p\Big(\frac{1}{\sqrt{n_k Th}} \Big), \end{aligned}$$

which completes the proof of (A.10).

In the next step, we show that

$$\frac{1}{n_k} \sum_{i \in G_k} \frac{Q_{i,\gamma}(w)}{\mathbb{E}[\widehat{f}_i(w)]} = o_p\Big(\frac{1}{\sqrt{n_k Th}} \Big). \tag{A.12}$$

To do so, we derive the convergence rate of the second moment

$$\mathbb{E}\left[\Big\{ \frac{1}{n_k} \sum_{i \in G_k} \frac{Q_{i,\gamma}(w)}{\mathbb{E}[\widehat{f}_i(w)]} \Big\}^2 \right] = \frac{1}{n_k^2(n-1)^2} \sum_{i,i' \in G_k} \sum_{\substack{j \neq i \\ j' \neq i'}} \frac{1}{\mathbb{E}[\widehat{f}_i(w)]} \frac{1}{\mathbb{E}[\widehat{f}_{i'}(w)]} \times \Big(\frac{1}{T^2} \sum_{t,t'=1}^{T} \Psi_{i,i',j,j',t,t'}(w) \Big), \tag{A.13}$$

where $\Psi_{i,i',j,j',t,t'}(w) = \mathbb{E}[K_h(X_{it} - w)\{m_j(X_{jt}) + \varepsilon_{jt}\} K_h(X_{i't'} - w)\{m_{j'}(X_{j't'}) + \varepsilon_{j't'}\}]$.

Similarly as above, $\Psi_{i,i',j,j',t,t'}(w) \neq 0$ only if (a) $i = j'$ and $i' = j$ or (b) $j = j'$. Applying Davydov's inequality once again, we get that in case (a), $T^{-2}\sum_{t,t'=1}^{T}\Psi_{i,i',j,j',t,t'}(w) \leq C\log T/T$ and in case (b),

$$\frac{1}{T^2}\sum_{t,t'=1}^{T}\Psi_{i,i',j,j',t,t'}(w) \leq \begin{cases} \frac{C\log T}{Th} & \text{if } i = i' \\ \frac{C\log T}{T} & \text{if } i \neq i'. \end{cases}$$

Plugging these bounds into (A.13), we easily arrive at (A.12). The statement (A.7) now follows upon combining (A.10) with (A.12). □

Proof of (A.8) and (A.9). By arguments similar to those for (A.7),

$$\frac{1}{n_k}\sum_{i\in G_k}\frac{Q_{i,\ell}(w)}{\widehat{f}_i(w)} - \frac{1}{n_k}\sum_{i\in G_k}\frac{Q_{i,\ell}(w)}{\mathbb{E}[\widehat{f}_i(w)]} = o_p\Big(\frac{1}{\sqrt{n_kTh}}\Big) \tag{A.14}$$

for $\ell \in \{V, B\}$. With the help of standard bias calculations, we further obtain that

$$\frac{1}{n_k}\sum_{i\in G_k}\frac{Q_{i,\ell}(w)}{\mathbb{E}[\widehat{f}_i(w)]} - \frac{1}{n_k}\sum_{i\in G_k}\frac{Q_{i,\ell}(w)}{f_i(w)} = o_p\Big(\frac{1}{\sqrt{n_kTh}}\Big). \tag{A.15}$$

Combining (A.14) and (A.15) completes the proof. □

Appendix B

In the proof of Theorems 3.1–3.3, we repeatedly make use of the following uniform convergence result.

Lemma B.1. *Under (C1)–(C5), it holds that*

$$\max_{1\leq i\leq n}\sup_{w\in[0,1]}\big|\widehat{m}_i(w) - m_i(w)\big| = o_p(1).$$

To show this lemma, we modify standard arguments to derive uniform convergence rates for kernel estimators, which can be found e.g. in Masry (1996), Bosq (1998) or Hansen (2008). These arguments are designed to derive the rate of $\sup_{w\in[0,1]}|\widehat{m}_i(w) - m_i(w)|$ for a fixed individual i. They thus yield the rate which is uniform over w but pointwise in i. In contrast to this, we aim to derive the rate which is uniform both over w and i. The additional uniformity over i slows down the convergence rate in general, that is, the term $\max_{1\leq i\leq n}\sup_{w\in[0,1]}|\widehat{m}_i(w) - m_i(w)|$ converges more slowly than $\sup_{w\in[0,1]}|\widehat{m}_i(w) - m_i(w)|$ for a fixed i. Inspecting the proof of Lemma B.1, our arguments can be seen to imply that the rate is at least $O_p((nT)^{-1/(20+\delta)} + h)$ for some small $\delta > 0$, that is, $\max_{1\leq i\leq n}\sup_{w\in[0,1]}|\widehat{m}_i(w) - m_i(w)| = O_p((nT)^{-1/(20+\delta)} + h)$.

Proof of Lemma B.1. To start with, write

$$\widehat{m}_i(w) - m_i(w) = \big[Q_{i,V}(w) + Q_{i,B}(w) - Q_{i,\gamma}(w)\big]/\widehat{f}_i(w) - \overline{Q}_i + \overline{\overline{Q}}_i,$$

where $Q_{i,V}(w)$, $Q_{i,B}(w)$, $Q_{i,\gamma}(w)$ along with $\overline{Q}_i$, $\overline{\overline{Q}}_i$ and $\widehat{f}_i(w)$ are defined as in the proof of Theorem 3.3. In what follows, we show that

$$\max_{1\leq i\leq n} \sup_{w\in[0,1]} \big|Q_{i,V}(w)\big| = o_p(1) \tag{B.1}$$

$$\max_{1\leq i\leq n} \sup_{w\in[0,1]} \big|Q_{i,B}(w)\big| = o_p(1) \tag{B.2}$$

$$\max_{1\leq i\leq n} \sup_{w\in[0,1]} \big|Q_{i,\gamma}(w)\big| = o_p(1) \tag{B.3}$$

$$\max_{1\leq i\leq n} \sup_{w\in[0,1]} \big|\widehat{f}_i(w) - f_i(w)\big| = o_p(1). \tag{B.4}$$

Moreover, a simplified version of the arguments for (B.1) yields that $\max_{1\leq i\leq n} |\overline{Q}_i| = o_p(1)$ as well as $\max_{1\leq i\leq n} |\overline{\overline{Q}}_i| = o_p(1)$. Lemma B.1 immediately follows upon combining these statements. □

Proof of (B.1). Let $\{a_{n,T}\}$ be a sequence of positive numbers that slowly converges to zero. In particular, we set $a_{n,T} = (nT)^{-\xi}$ for some sufficiently small constant $\xi > 0$. In addition, define

$$\begin{aligned}\varepsilon_{it}^{\leq} &= \varepsilon_{it} 1\big(|\varepsilon_{it}| \leq \tau_{n,T}\big)\\ \varepsilon_{it}^{>} &= \varepsilon_{it} 1\big(|\varepsilon_{it}| > \tau_{n,T}\big),\end{aligned}$$

where $\tau_{n,T} = (nT)^{1/(\theta-\delta)}$, θ is introduced in (C3) and $\delta > 0$ is a small positive number. With this notation at hand, we can rewrite the term $Q_{i,V}(w)$ as

$$Q_{i,V}(w) = \sum_{t=1}^{T} Z_{it,T}^{\leq}(w) + \sum_{t=1}^{T} Z_{it,T}^{>}(w),$$

where

$$\begin{aligned}Z_{it,T}^{\leq}(w) &= \big(K_h(X_{it} - w)\varepsilon_{it}^{\leq} - \mathbb{E}\big[K_h(X_{it} - w)\varepsilon_{it}^{\leq}\big]\big)/T\\ Z_{it,T}^{>}(w) &= \big(K_h(X_{it} - w)\varepsilon_{it}^{>} - \mathbb{E}\big[K_h(X_{it} - w)\varepsilon_{it}^{>}\big]\big)/T.\end{aligned}$$

We thus split $Q_{i,V}(w)$ into the "interior part" $\sum_{t=1}^{T} Z_{it,T}^{\leq}(w)$ and the "tail part" $\sum_{t=1}^{T} Z_{it,T}^{>}(w)$. This parallels the standard arguments for deriving the convergence rate of $\sup_{w\in[0,1]} |Q_{i,V}(w)|$ for a fixed individual i. As we maximize over i, however, we have to choose the truncation sequence $\tau_{n,T}$ to go to infinity much faster than in the standard case with a fixed i. (When n is bounded, we can of course choose $\tau_{n,T}$ to diverge as quickly as in the standard case. When n goes to infinity in contrast, $\tau_{n,T}$

is required to diverge much faster.) This is the main reason why we obtain a slower convergence rate than in the standard case.

We now proceed in several steps. To start with, we show that

$$\max_{1\le i\le n}\sup_{w\in[0,1]}\Big|\sum_{t=1}^{T} Z^{>}_{it,T}(w)\Big| = o_p(1). \tag{B.5}$$

This can be achieved as follows:

$$\begin{aligned}
&\mathbb{P}\Big(\max_{1\le i\le n}\sup_{w\in[0,1]}\Big|\sum_{t=1}^{T} Z^{>}_{it,T}(w)\Big| > a_{n,T}\Big)\\
&\le \sum_{i=1}^{n}\mathbb{P}\Big(\sup_{w\in[0,1]}\Big|\frac{1}{T}\sum_{t=1}^{T} K_h(X_{it}-w)\varepsilon^{>}_{it}\Big| > \frac{a_{n,T}}{2}\Big)\\
&\quad+\sum_{i=1}^{n}\mathbb{P}\Big(\sup_{w\in[0,1]}\Big|\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[K_h(X_{it}-w)\varepsilon^{>}_{it}\big]\Big| > \frac{a_{n,T}}{2}\Big).
\end{aligned}$$

With the help of assumption (C3), we obtain that

$$\begin{aligned}
&\sum_{i=1}^{n}\mathbb{P}\Big(\sup_{w\in[0,1]}\Big|\frac{1}{T}\sum_{t=1}^{T} K_h(X_{it}-w)\varepsilon^{>}_{it}\Big| > \frac{a_{n,T}}{2}\Big)\\
&\le \sum_{i=1}^{n}\mathbb{P}\Big(|\varepsilon_{it}| > \tau_{n,T} \text{ for some } 1\le t\le T\Big)\\
&\le C(nT)^{1-\frac{\theta}{\theta-\delta}} = o(1).
\end{aligned}$$

Once more applying (C3), it can be seen that

$$\begin{aligned}
\big|\mathbb{E}\big[K_h(X_{it}-w)\varepsilon^{>}_{it}\big]\big| &\le \mathbb{E}\left[K_h(X_{it}-w)\mathbb{E}\Big[\frac{|\varepsilon_{it}|^{\theta}}{\tau_{n,T}^{\theta-1}}1(|\varepsilon_{it}|>\tau_{n,T})\Big|X_{it}\Big]\right]\\
&\le C(nT)^{-\frac{\theta-1}{\theta-\delta}}
\end{aligned}$$

with some constant C independent of w. If we choose the exponent $\xi > 0$ in the definition of $a_{n,T}$ small enough, we get that $C(nT)^{-\frac{\theta-1}{\theta-\delta}} < a_{n,T}/2$ as the sample size grows large, implying that

$$\sum_{i=1}^{n}\mathbb{P}\Big(\sup_{w\in[0,1]}\Big|\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[K_h(X_{it}-w)\varepsilon^{>}_{it}\big]\Big| > \frac{a_{n,T}}{2}\Big) = 0$$

for sufficiently large sample sizes. This yields (B.5).

We next have a closer look at the expression $\sum_{t=1}^{T} Z^{\le}_{it,T}(w)$. Let $0 = w_0 < w_1 < \ldots < w_L = 1$ be an equidistant grid of points covering the unit interval and set $L = L_{n,T} = \tau_{n,T}/(a_{n,T}h^2)$. Exploiting the Lipschitz continuity of the kernel K,

straightforward calculations yield that

$$\max_{1\le i\le n}\sup_{w\in[0,1]}\Big|\sum_{t=1}^{T} Z^{\leq}_{it,T}(w)\Big| \le \max_{1\le i\le n}\max_{1\le \ell\le L}\Big|\sum_{t=1}^{T} Z^{\leq}_{it,T}(w_\ell)\Big| + Ca_{n,T}. \tag{B.6}$$

We can thus replace the supremum over w by a maximum over the grid points w_ℓ. Moreover, it holds that

$$\begin{aligned}\mathbb{P}\Big(\max_{1\le i\le n}\max_{1\le \ell\le L}\Big|\sum_{t=1}^{T} Z^{\leq}_{it,T}(w_\ell)\Big| > C_0 a_{n,T}\Big) \\ \le \sum_{i=1}^{n}\sum_{\ell=1}^{L}\mathbb{P}\Big(\Big|\sum_{t=1}^{T} Z^{\leq}_{it,T}(w_\ell)\Big| > C_0 a_{n,T}\Big),\end{aligned} \tag{B.7}$$

where C_0 is a sufficiently large constant to be specified later on. In what follows, we show that for each fixed w_ℓ,

$$\mathbb{P}\Big(\Big|\sum_{t=1}^{T} Z^{\leq}_{it,T}(w_\ell)\Big| > C_0 a_{n,T}\Big) \le CT^{-r}, \tag{B.8}$$

where the constants C and r are independent of w_ℓ and $r>0$ can be chosen arbitrarily large. Plugging (B.8) into (B.7) and combining the result with (B.6), we arrive at

$$\max_{1\le i\le n}\sup_{w\in[0,1]}\Big|\sum_{t=1}^{T} Z^{\leq}_{it,T}(w)\Big| = o_p(1), \tag{B.9}$$

which completes the proof.

It thus remains to prove (B.8). To do so, we split the term $\sum_{t=1}^{T} Z^{\leq}_{it,T}(w_\ell)$ into blocks as follows:

$$\sum_{t=1}^{T} Z^{\leq}_{it,T}(w_\ell) = \sum_{s=1}^{q_{n,T}} B_{2s-1} + \sum_{s=1}^{q_{n,T}} B_{2s}$$

with $B_s = \sum_{t=(s-1)r_{n,T}+1}^{sr_{n,T}} Z^{\leq}_{it,T}(w_\ell)$, where $2q_{n,T}$ is the number of blocks and $r_{n,T} = T/2q_{n,T}$ is the block length. In particular, we choose the block length such that $r_{n,T} = O(T^\eta)$ for some small $\eta>0$. With this notation at hand, we get

$$\begin{aligned}\mathbb{P}\Big(\Big|\sum_{t=1}^{T} Z^{\leq}_{it,T}(w_\ell)\Big| > C_0 a_{n,T}\Big) \le \mathbb{P}\Big(\Big|\sum_{s=1}^{q_{n,T}} B_{2s-1}\Big| > \frac{C_0}{2} a_{n,T}\Big) \\ + \mathbb{P}\Big(\Big|\sum_{s=1}^{q_{n,T}} B_{2s}\Big| > \frac{C_0}{2} a_{n,T}\Big).\end{aligned}$$

As the two terms on the right-hand side can be treated analogously, we focus attention to the first one. By Bradley's lemma (see Lemma 1.2 in Bosq (1998)), we can construct a sequence of random variables $B_1^*, B_3^*, \ldots$ such that (a) $B_1^*, B_3^*, \ldots$ are independent, (b)

B_{2s-1} and B^*_{2s-1} have the same distribution for each s, and (c) for $0 < \mu \leq \|B_{2s-1}\|_\infty$, $\mathbb{P}(|B^*_{2s-1} - B_{2s-1}| > \mu) \leq 18(\|B_{2s-1}\|_\infty/\mu)^{1/2}\alpha(r_{n,T})$. With these variables, we obtain the bound

$$\mathbb{P}\Big(\Big|\sum_{s=1}^{q_{n,T}} B_{2s-1}\Big| > \frac{C_0}{2}a_{n,T}\Big) \leq P_1 + P_2,$$

where

$$P_1 = \mathbb{P}\Big(\Big|\sum_{s=1}^{q_{n,T}} B^*_{2s-1}\Big| > \frac{C_0}{4}a_{n,T}\Big)$$
$$P_2 = \mathbb{P}\Big(\Big|\sum_{s=1}^{q_{n,T}} \big(B_{2s-1} - B^*_{2s-1}\big)\Big| > \frac{C_0}{4}a_{n,T}\Big).$$

Using (c) together with the fact that the mixing coefficients $\alpha(\cdot)$ decay to zero exponentially fast, it is not difficult to see that P_2 converges to zero at an arbitrarily fast polynomial rate. To deal with P_1, we make use of the following three facts:

(i) For a random variable B and $\lambda > 0$, Markov's inequality yields that

$$\mathbb{P}\big(\pm B > \delta\big) \leq \frac{\mathbb{E}\exp(\pm\lambda B)}{\exp(\lambda\delta)}.$$

(ii) We have that $|B_{2s-1}| \leq C_B r_{n,T}\tau_{n,T}/(Th)$ for some constant $C_B > 0$. Define $\lambda_{n,T} = Th/(2C_B r_{n,T}\tau_{n,T})$, which implies that $\lambda_{n,T}|B_{2s-1}| \leq 1/2$. As $\exp(x) \leq 1 + x + x^2$ for $|x| \leq 1/2$, we get that

$$\mathbb{E}\Big[\exp\big(\pm\lambda_{n,T}B_{2s-1}\big)\Big] \leq 1 + \lambda^2_{n,T}\mathbb{E}\big[(B_{2s-1})^2\big] \leq \exp\big(\lambda^2_{n,T}\mathbb{E}\big[(B_{2s-1})^2\big]\big)$$

along with

$$\mathbb{E}\Big[\exp\big(\pm\lambda_{n,T}B^*_{2s-1}\big)\Big] \leq \exp\big(\lambda^2_{n,T}\mathbb{E}\big[(B^*_{2s-1})^2\big]\big).$$

(iii) Standard calculations for kernel estimators imply that

$$\sum_{s=1}^{q_{n,T}} \mathbb{E}\big[(B^*_{2s-1})^2\big] \leq \frac{C}{Th}.$$

Using (i)–(iii), we arrive at

$$\mathbb{P}\Big(\Big|\sum_{s=1}^{q_{n,T}} B^*_{2s-1}\Big| > \frac{C_0}{4}a_{n,T}\Big)$$
$$\leq \mathbb{P}\Big(\sum_{s=1}^{q_{n,T}} B^*_{2s-1} > \frac{C_0}{4}a_{n,T}\Big) + \mathbb{P}\Big(-\sum_{s=1}^{q_{n,T}} B^*_{2s-1} > \frac{C_0}{4}a_{n,T}\Big)$$

$$
\begin{aligned}
&\leq \exp\Big(-\frac{C_0}{4}\lambda_{n,T}a_{n,T}\Big)\left\{\mathbb{E}\Big[\exp\Big(\lambda_{n,T}\sum_{s=1}^{q_{n,T}}B^*_{2s-1}\Big)\Big]+\mathbb{E}\Big[\exp\Big(-\lambda_{n,T}\sum_{s=1}^{q_{n,T}}B^*_{2s-1}\Big)\Big]\right\}\\
&\leq \exp\Big(-\frac{C_0}{4}\lambda_{n,T}a_{n,T}\Big)\left\{\prod_{s=1}^{q_{n,T}}\mathbb{E}\Big[\exp\big(\lambda_{n,T}B^*_{2s-1}\big)\Big]+\prod_{s=1}^{q_{n,T}}\mathbb{E}\Big[\exp\big(-\lambda_{n,T}B^*_{2s-1}\big)\Big]\right\}\\
&\leq 2\exp\Big(-\frac{C_0}{4}\lambda_{n,T}a_{n,T}\Big)\prod_{s=1}^{q_{n,T}}\exp\Big(\lambda^2_{n,T}\mathbb{E}\big[(B^*_{2s-1})^2\big]\Big)\\
&= 2\exp\Big(-\frac{C_0}{4}\lambda_{n,T}a_{n,T}\Big)\exp\Big(\lambda^2_{n,T}\sum_{s=1}^{q_{n,T}}\mathbb{E}\big[(B^*_{2s-1})^2\big]\Big)\\
&\leq 2\exp\Big(-\frac{C_0}{4}\lambda_{n,T}a_{n,T}+\lambda^2_{n,T}\frac{C}{Th}\Big).
\end{aligned}
$$

Moreover, taking into account that $n/T \leq C$ and $T^{1/2}h \to \infty$ by assumption, setting θ to a value slightly larger than 4 and supposing that $a_{n,T} = (nT)^{-\xi}$ with $\xi \leq 1/(20+\delta)$ for some small $\delta > 0$, it holds that

$$
\exp\Big(-\frac{C_0}{4}\lambda_{n,T}a_{n,T}+\lambda^2_{n,T}\frac{C}{Th}\Big) \leq T^{-r}
$$

for sufficiently large sample sizes, where the constant $r > 0$ can be chosen arbitrarily large. This implies that $P_1 \leq CT^{-r}$, which in turn completes the proof of (B.8). □

Proof of (B.3). Define $Z_{it} = (n-1)^{-1}\sum_{j\neq i}(m_j(X_{jt})+\varepsilon_{jt})$ and write

$$
Q_{i,\gamma}(w) = \frac{1}{T}\sum_{t=1}^{T}K_h(X_{it}-w)Z_{it}. \tag{B.10}
$$

By construction, the time series processes $\{X_{it} : 1 \leq t \leq T\}$ and $\{Z_{it} : 1 \leq t \leq T\}$ are independent of each other. Moreover, by Theorem 5.2 in Bradley (2005), the process $\{Z_{it} : 1 \leq t \leq T\}$ is strongly mixing with mixing coefficients that are bounded by $n\alpha(k)$. (B.3) can thus be shown by applying the arguments from the proof of (B.1) to (B.10). □

Proof of (B.2) and (B.4). The two statements follow by the arguments from the proof of (B.1) together with standard bias calculations. □

The proof of Theorem 3.3 makes use of an additional uniform convergence result which specifies the rate of the kernel density estimator $\widehat{f}_i(w) = T^{-1}\sum_{t=1}^{T}K_h(X_{it}-w)$.

Lemma B.2. *Under (C1)–(C5), it holds that*

$$
\max_{1\leq i\leq n}\sup_{w\in[0,1]}\big|\widehat{f}_i(w)-\mathbb{E}\big[\widehat{f}_i(w)\big]\big| = O_p\Big(\sqrt{\frac{\log T}{Th}}\Big).
$$

Proof of Lemma B.2. The overall strategy is the same as that for the proof of (B.1). There is however one important difference: In the proof of (B.1), we have examined a kernel average of the form $T^{-1}\sum_{t=1}^{T} K_h(X_{it}-w)Z_{it}$ with $Z_{it}=\varepsilon_{it}$. As the variables ε_{it} have unbounded support in general, we have introduced the truncation sequence $\tau_{n,T}$ and have split ε_{it} into the two parts $\varepsilon_{it}^{\leq}$ and $\varepsilon_{it}^{>}$. Here in contrast, we are concerned with the case $Z_{it}\equiv 1$. Importantly, the random variables $Z_{it}\equiv 1$ are bounded, implying that we do not have to truncate them at all. Keeping this in mind and going step by step along the proof of (B.1), we arrive at the statement of Lemma B.2. □

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.

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

Bradley, R. C. (2005). Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, **2** 107–144.

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.

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. 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.

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.

Körber, 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*.

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

Lin, C.-C. and Ng, S. (2012). Estimation of panel data models with parameter heterogeneity when group membership is unknown. *Journal of Econometric Methods*, **1** 42–55.

Linton, O. (1997). Efficient estimation of additive nonparametric regression models. *Biometrika*, **84** 469–473.

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.

Qian, J. and Wang, L. (2012). Estimating semiparametric panel data models by marginal integration. *Journal of Econometrics*, **167** 483–493.

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

Sarafidis, V. and Weber, N. (2014). A partially heterogeneous framework for analyzing panel data. *Oxford Bulletin of Economics and Statistics*.

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.

Vogt, M. and Linton, O. (2014). Nonparametric estimation of a periodic sequence in the presence of a smooth trend. *Biometrika*, **101** 121–140.

Wang, L. and Yang, L. (2007). Spline-backfitted kernel smoothing of nonlinear additive autoregression model. *Annals of Statistics*, **35** 2474–2503.