A Semiparametric Model for Heterogeneous Panel Data with Fixed Effects

Lena Körber[*]
London School of Economics

Oliver Linton[†]
University of Cambridge

Michael Vogt[‡]
University of Konstanz

December 19, 2013

Abstract

This paper develops methodology for semiparametric panel data models in a setting where both the time series and the cross section are large. Such settings are common in finance and other areas of economics. Our model allows for heterogeneous nonparametric covariate effects as well as unobserved time and individual specific effects that may depend on the covariates in an arbitrary way. To model the covariate effects parsimoniously, we impose a dimensionality reducing common component structure on them. In the theoretical part of the paper, we derive the asymptotic theory of the proposed procedure. In particular, we provide the convergence rates and the asymptotic distribution of our estimators. In the empirical part, we apply our methodology to a specific application that has been the subject of recent policy interest, that is, the effect of trading venue fragmentation on market quality. We use a unique dataset that reports the location and volume of trading on the FTSE350 companies from 2008 to 2011 at the weekly frequency. We find that the effect of fragmentation on market quality is nonlinear and non-monotonic. The implied quality of the market under perfect competition is superior to that under monopoly provision, but the transition between the two is complicated.

[*] Address: Department of Economics, London School of Economics, Houghton Street, London WC2A 2AE, UK. Email: `l.m.koerber@lse.ac.uk`. Financial support from the ESRC and Cusanuswerk is gratefully acknowledged. I would also like to thank participants at the UCL-LSE-cemmap PhD student conference for their helpful comments.

[†] Address: Faculty of Economics, Austin Robinson Building, Sidgwick Avenue, Cambridge, CB3 9DD, UK. Email: `obl20@cam.ac.uk`. I would like to thank the European Research Council for financial support.

[‡] Address: Department of Mathematics and Statistics, 78457 Konstanz, Germany. Email: `michael.vogt@uni-konstanz.de`. Financial support by the ERC is gratefully acknowledged.

1 Introduction

In this paper, we develop estimation methodology for semiparametric panel models in a setting where both the time series and the cross section dimension are large. Such settings have become increasingly common over the last couple of years. In particular, they are frequently encountered in finance and various areas of economics such as industrial organization or labour economics. Cheng Hsiao has been a pioneer in the development of panel data techniques and his monograph (1986, 2003) contains the main methodological background for our work.

We investigate a regression model which has a nonparametric covariate effect along with individual and time specific fixed effects. The covariate effect is allowed to be heterogeneous across individuals, which is feasible given the long time series we are assuming. To restrict the heterogeneity to be of low dimension, we propose a common component structure on the model. In particular, we assume the individual covariate effects to be composed of a finite number of unknown functions that are the same across individuals but loaded up differently for each cross-sectional unit. The covariate effects are thus modelled as linear combinations of a small number of common functions. The individual and time specific effects of the model are allowed to be related to the covariate in quite a general way. This allows a potential channel for endogeneity, which is important in many applications. We recognize that the endogeneity that is permitted is rather limited, but we remark that this type of restriction is extremely widely exploited in empirical microeconomics, see Angrist and Pischke (2009, Chapter 5). A rigorous formulation of the model together with a detailed description of its components is given in Section 2. The issue of identifying the various model components is discussed in Section 3.

Our model can be regarded as an intermediate case between two extremes. The one extreme is the homogeneous model, where the covariate effect is the same for each cross-sectional unit. This is a very common framework which has been investigated in various parametric and semiparametric studies, see for example Hsiao (1986). In a wide range of applications, it is however rather unrealistic to assume that the covariate effect is the same for all individuals. On the other extreme end, there is the fully flexible model without any restrictions on the covariate effects. One example is the classical SURE model. More recently, Chen, Gao, and Li (2012) among others have studied a semiparametric version of this very general framework. Even though it is highly flexible, it is however not well suited to some applications. In particular, if the number of individuals is in the hundreds or thousands, the estimation output consists of a huge number of individual functions. This makes the model hardly interpretable. Furthermore, the estimation precision may be very low. Our model lies between these two extremes and allows the user to select the degree of flexibility appropriate for the given application.

Our setting falls in the class of semiparametric panel data models for large cross-section and long time series. Most of the models proposed in the literature for this type of panel data are essentially parametric. Some important papers include Phillips and Moon (1999), Bai and Ng (2002), Bai (2003, 2004), and Pesaran (2006). These authors have addressed a variety of issues including nonstationarity, estimation of unobserved factors, and model selection. Most of the work on semiparametric panel models is in the context of short time series, see for example Kyriazidou (1997). Nonparametric additive models have been considered for instance in Porter (1996). More recent articles include Mammen, Støve, and Tjøstheim (2009), Qian and Wang (2011), and Hoderlein, Mammen, and Yu (2011).

Only recently, there have been a number of contributions to the non- and semiparametric literature on panels with large cross-section and time series dimension. Linton, Nielsen, and Nielsen (2009) consider estimation of a fixed effect time series. Atak, Linton, and Xiao (2011) are concerned with seasonality and trends in a panel setting; see also Li, Chen, and Gao (2013a). Connor, Hagmann, and Linton (2012) consider a semiparametric additive panel model for stock returns driven by observable covariates and unobservable "factor returns". They allow weak dependence in both time and cross-section direction, but the covariates are not time-varying and there is no individual effect. This model is suited for their application but does not allow a channel for endogeneity. The estimation method is made simpler by the fact that each additive term has a different covariate, whereas the common functions in our model all have the same covariate. Kneip, Sickles, and Song (2012) consider a model similar to ours except that they focus on time as the key nonparametric covariate. Moreover, they do not allow individual effects to be related to included covariates, that is, there is no endogeneity in their model.

In Section 8, we apply our methods to an empirical question of recent interest for policy makers and in academic research, that is, the effect of trading venue fragmentation on market quality. In 2007, the monopoly of primary European exchanges such as the London stock exchange was ended by the "Markets in Financial Instruments Directive". Since then, various new trading platforms have emerged and competed for trading volume. We investigate whether this competition has led to improved market quality for participants. It has been argued that High Frequency Trading has been a major beneficiary of the market fragmentation, and that this affects both the amount of fragmentation as well as the quality of the market outcomes.[1] Our model allows for this endogeneity channel by treating this unobservable as part of the individual and time effects. It also allows for heterogeneous nonlinear covariate effects of fragmentation on market quality, which we think are important for capturing the relationship of interest in an adequate way. We use a unique weekly dataset on the location and volume of trading for FTSE 100 and FTSE 250 companies over the period from 2008 to 2011, as well as publicly available measures of market quality. To summarize the results, we find that the effect of fragmentation on market quality is nonlinear and non-monotonic. The implied quality of the market under perfect competition is superior to that under monopoly provision, but the transition between the two regimes is complicated. Our model and procedures may also be applied

[1] See the UK government project "The future of computer based trading in financial markets" for a full description of High Frequency Trading and related concepts. www.bis.gov.uk/foresight/our-work/projects/current-projects/computer-trading.

in many other contexts in economics and finance.

Our method to estimate the common functions and the parameter vectors which constitute the individual covariate effects is introduced in Section 4. The asymptotic properties of the estimators are described in Section 5. In Subsection 5.2, we derive the uniform convergence rates as well as an asymptotic normality result for our estimators of the common functions. Importantly, the estimators can be shown to converge to the true functions at the uniform rate $\sqrt{\log nT/nTh}$ which is based on the pooled number of data points nT with n being the cross-section dimension and T the length of the time series. Intuitively, this fast rate is possible to achieve because the functions are the same for all individuals. This allows us to base our estimation procedure on information from the whole panel rather than on a single time series corresponding to a specific individual. In Subsection 5.3, we investigate the asymptotic behaviour of our parameter estimators. In particular, we show that they are asymptotically normal. As will turn out, the parameters are estimated with the same precision as in the case where the common functions are known. In particular, our estimators have the same asymptotic distribution as the oracle estimators constructed under the assumption that the functions are observed. To investigate the small sample performance of our estimation procedures, we conduct a series of simulation experiments. Overall, our procedures work well even for quite small sample sizes. For reasons of brevity, the detailed results are reported in the supplementary material.

To keep the arguments and discussion as simple as possible, we derive our estimation procedure as well as the asymptotic results under the simplifying assumption that the number of common functions is known. In Sections 6 and 7, we explain how to dispense with this assumption. In particular, we provide a simple rule to select the number of unknown common functions. This complements our estimation procedure and makes it ready to apply to real data.

2 The model

In this section, we provide a detailed description of our model framework. We observe a sample of panel data $\{(Y_{it}, X_{it}) : i = 1, \dots, n, \ t = 1, \dots, T\}$, where i denotes the i-th individual and t is the time point of observation. To keep the notation as simple as possible, we assume that both the variables Y_{it} and X_{it} are real-valued and focus on the case of a balanced panel.

The data are assumed to come from the model

$$Y_{it} = \mu_0 + \alpha_i + \gamma_t + m_i(X_{it}) + \varepsilon_{it}, \tag{1}$$

where $\mathbb{E}[\varepsilon_{it}|X_{it}] = 0$. Here, m_i are nonparametric functions which capture the covariate effect, μ_0 is the model constant and the variables ε_{it} are idiosyncratic error terms. The expressions α_i and γ_t are unobserved individual and time specific effects, respectively, which may depend on the regressors in an arbitrary way, e.g., $\alpha_i = G_i(X_{i1}, \ldots, X_{iT}; \eta_i)$ and $\gamma_t = H_t(X_{1t}, \ldots, X_{nt}; \delta_t)$ for some deterministic functions G_i, H_t and random errors η_i, δ_t that are independent of the covariates. As usual there is an identification shortfall here, and to identify the components of the model, we assume that $\mathbb{E}[m_i(X_{it})] = 0$ along with $\sum_{i=1}^{n} \alpha_i = \sum_{t=1}^{T} \gamma_t = 0$.

As the functions m_i may differ across individuals, the covariate effect in our model is allowed to be heterogeneous. However, rather than allowing the effect to vary completely freely, we impose some structure on it. In particular, we assume the functions m_i to have the common component structure

$$m_i(x) = \sum_{k=1}^{K} \beta_{ik} \mu_k(x), \tag{2}$$

where $\mu = (\mu_1, \ldots, \mu_K)^\intercal$ is a vector of nonparametric component functions and $\beta_i = (\beta_{i1}, \ldots, \beta_{iK})^\intercal$ are parameter vectors. Like the functions μ and the coefficient vectors β_i, the number of components K is unobserved. Identifying the functions μ together with the coefficients β_i in our setting is not completely straightforward and requires some care. We thus devote a separate section to this issue. In particular, we provide a detailed discussion in Section 3.

The model defined by (1) and (2) takes into account several issues which are important in a panel data context. To start with, it captures nonlinearities and heterogeneity in the covariate effect in a flexible but parsimonious way. Moreover, since $\mathbb{E}[\alpha_i + \gamma_t | \{X_{it}\}] \neq 0$ in general, the unobserved effects α_i and γ_t introduce a simultaneity between the covariates and the dependent variable. This allows a certain type of endogeneity. Our model and the estimation techniques we develop may thus be applied to a number of different empirical problems where heterogeneity and endogeneity are potential issues. In the applied part of the paper, we focus on a particular problem from the area of finance.[2]

The type of endogeneity allowed for by the unobserved effects α_i and γ_t is rather limited, but we remark that this type of restriction is extremely widely exploited in empirical microeconomics, see Angrist and Pischke (2009, Chapter 5). An alternative approach to dealing with endogeneity is to introduce instrumental variables, but there are advantages and disadvantages with that approach also. Our model has the benefit of simplicity and is in line with the simple approach to identifying empirical effects espoused both in Angrist and Pischke (2009) and Manski (2008), for example. It is a generalization of standard heterogeneous linear regression panel data models that are widely discussed in Hsiao (2003) and is part of a large developing literature on semiparametric panel models including Atak, Linton, and Xiao (2011), Chen, Gao, and Li (2012), Connor, Hagmann, and Linton (2012), Chen, Gao, and Li (2013a), and Chen, Gao, and Li (2013b) that explore different weakenings of these models.

[2] We note that a symmetric type of model where the heterogeneity in the covariate effect is driven by time rather than individual (i.e., $m_t(\cdot)$ instead of $m_i(\cdot)$) may be of interest in some cases.

The elements $\theta = \{\mu_0, \alpha_i, \gamma_t : i = 1, \ldots, n,\ t = 1, \ldots, T\}$ play the role of nuisance parameters in our framework. There is a large number of them which is increasing with the sample size. Nevertheless, we have an even larger number of observations, which enable us to estimate consistently all the unknown quantities of interest. We thus do not face the "incidental parameters problem" (Neyman and Scott (1948)) that is of wide concern in other panel data settings; see Hsiao (2003) for some discussion of this issue.

We take a pragmatic approach to estimation based on first eliminating the nuisance parameters. To achieve this we make use of a fixed effect transformation. Denote the time, cross sectional, and global averages by:

$$\overline{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it}, \qquad \overline{Y}_t = \frac{1}{n}\sum_{i=1}^{n} Y_{it}, \qquad \overline{\overline{Y}} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} Y_{it},$$

and define $Y_{it}^{\mathrm{fe}} = Y_{it} - \overline{Y}_i - \overline{Y}_t + \overline{\overline{Y}}$. Now note that

$$\begin{aligned} Y_{it}^{\mathrm{fe}} &= m_i(X_{it}) + \varepsilon_{it} - \frac{1}{T}\sum_{t=1}^{T} m_i(X_{it}) - \frac{1}{T}\sum_{t=1}^{T}\varepsilon_{it} - \frac{1}{n}\sum_{i=1}^{n} m_i(X_{it}) - \frac{1}{n}\sum_{i=1}^{n}\varepsilon_{it} \\ &\qquad + \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} m_i(X_{it}) + \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\varepsilon_{it} \\ &= m_i(X_{it}) + \varepsilon_{it} + O_p(T^{-1/2}) + O_p(n^{-1/2}), \end{aligned} \tag{3}$$

where we require the sample averages to converge to their population means at standard rates. (3) shows that the nuisance parameters θ can be eliminated by subtracting sample means from the data, although this method introduces some additional small error terms.

An alternative procedure is based on differencing, which is the most common method in linear models, see Angrist and Pischke (2009). Specifically, let $Y_{ijt}^{\mathrm{did}} = (Y_{it} - Y_{it-1}) - (Y_{jt} - Y_{jt-1})$ denote the difference-in-difference transformation. Then we have

$$Y_{ijt}^{\mathrm{did}} = (m_i(X_{it}) - m_i(X_{it-1})) - (m_j(X_{jt}) - m_j(X_{jt-1})) + u_{ijt}, \tag{4}$$

where $u_{ijt} = (\varepsilon_{it} - \varepsilon_{it-1}) - (\varepsilon_{jt} - \varepsilon_{jt-1})$ is a serially dependent error term. This approach also eliminates the nuisance parameters θ, but also not completely without cost. First of all, the right-hand side of (4) is an additive regression function of the covariates $X_{it}, X_{it-1}, X_{jt}, X_{jt-1}$. To estimate this function, either higher dimensional smoothing must be employed, see Linton and Nielsen (1995), or iterative smoothing techniques like backfitting, see Mammen, Linton, and Nielsen (1999). Second, the error term u_{ijt} is dependent across time and cross-section, in particular it has a four term "dyadic" (Fafchamps and Gubert (2007)) structure that needs to be accounted for. Finally, one needs stronger conditional moment restrictions on the original error terms to be able to consistently estimate this model. Specifically, we require $\mathbb{E}[\varepsilon_{it}|X_{it}, X_{it-1}, X_{jt}, X_{jt-1}] = 0$ rather than just the assumption $\mathbb{E}[\varepsilon_{it}|X_{it}] = 0$ that will be needed for the fixed effect method. Henderson, Carroll, and Li (2008) propose this method (with just time differencing) in the homogeneous one way model, i.e., $Y_{it} = \mu_0 + \alpha_i + m(X_{it}) + \varepsilon_{it}$.

3 Identification

The individual regression functions m_i in our model are identified through the normalizations $\mathbb{E}[m_i(X_{it})] = 0$ along with $\sum_{i=1}^n \alpha_i = \sum_{t=1}^T \gamma_t = 0$. We now describe how to identify the vector of common component functions $\mu = (\mu_1, \dots, \mu_K)^\intercal$ and the parameter vectors $\beta_i = (\beta_{i1}, \dots, \beta_{iK})^\intercal$ which constitute the functions m_i. Roughly speaking, the idea is to characterize μ and the parameter vectors β_i via an eigenvalue decomposition of a matrix related to the functions m_i. Exploiting the uniqueness properties of this decomposition, we are able to identify μ and the parameter vectors up to sign. Our strategy is thus very similar to the arguments usually used in factor analysis which can for example be found in Connor and Korajczyk (1988) and Bai (2003).

To lay out our strategy, we denote the vector of individual functions by $m = (m_1, \dots, m_n)^\intercal$ and define B to be a $n \times K$ matrix with the entries β_{ik} for $i = 1, \dots, n$ and $k = 1, \dots, K$. With this notation at hand, we can represent the vector of functions m as

$$m = B\mu. \tag{5}$$

We now put some slight regularity conditions on B and μ. In particular, the functions μ are assumed to be orthonormal with respect to a weighting function w, i.e., $\int \mu(x)\mu(x)^\intercal w(x)dx = I_K$. Moreover, the coefficient matrix B is supposed to have full rank K. These assumptions are rather harmless. In particular, the rank condition on B just makes sure that there is enough variation in the coefficients, i.e., in the linear combinations of the μ-functions across individuals.

The above two assumptions on μ and B can be replaced by a condition which parallels the set of assumptions usually used in factor analysis. In particular, they are equivalent to the following condition:

(I1) The matrix B is orthonormal, i.e. $B^\intercal B = I_K$, and $\int \mu(x)\mu(x)^\intercal w(x)dx$ is a diagonal matrix with non-zero diagonal entries.

To see this equivalence, assume that we start off with a matrix $B^{(1)}$ of rank K and a vector of common component functions $\mu^{(1)}$ which are orthonormal in the sense specified above. Then consider the symmetric, positive definite $K \times K$ matrix $(B^{(1)})^\intercal B^{(1)} = ODO^\intercal$, where $OO^\intercal = O^\intercal O = I_K$ and D is a diagonal matrix with positive entries. Let

$$B^{(2)} = B^{(1)} O D^{-1/2} \tag{6}$$

$$\mu^{(2)}(x) = D^{1/2} O^\intercal \mu^{(1)}(x). \tag{7}$$

Then

$$(B^{(2)})^\intercal B^{(2)} = D^{-1/2} O^\intercal (B^{(1)})^\intercal B^{(1)} O D^{-1/2} = I_K$$

and

$$\int \mu^{(2)}(x) \mu^{(2)}(x)^\intercal w(x) dx = D^{1/2} O^\intercal O D^{1/2} = D.$$

Hence, the normalized versions $B^{(2)}$ and $\mu^{(2)}$ satisfy (I1).

Let us now assume that the matrix B and the component functions μ are normalized according to (I1). In addition, suppose that the functions μ satisfy the following constraint:

(I2) The diagonal entries of the matrix $\int \mu(x)\mu(x)^\intercal w(x)dx$ are all distinct.

This assumption is needed to ensure that the eigenspaces in the spectral decomposition below are one-dimensional, which in turn makes sure that the eigenvectors of the decomposition are uniquely identified up to sign.

Given (I1) and (I2), the matrix B can be characterized via the "covariance" structure of the functions m. In particular, we have that

$$\Omega := \int m(x)m(x)^\intercal w(x)dx = B \int \mu(x)\mu(x)^\intercal w(x)dx \ B^\intercal = BDB^\intercal,$$

where D is a diagonal matrix with the diagonal entries $\int \mu_k^2(x)w(x)dx$ for $k = 1, \ldots, K$. These entries are the non-zero distinct eigenvalues of the matrix Ω. Moreover, the columns of the matrix B are the corresponding orthonormal eigenvectors. This spectral decomposition is unique up to the sign of the eigenvectors, i.e., up to the sign of the columns of the matrix B. Thus, the coefficients contained in the matrix B are identified up to sign as well.

Exploiting the fact that the columns of B are orthonormal, we can moreover represent the vector of functions μ by writing

$$\mu = B^\intercal m.$$

This equation almost surely identifies the functions μ up to sign: The functions m_i contained in the vector m are identified almost surely by our normalizing assumptions. Moreover, as seen above the columns of the matrix B are identified up to sign. As a result, the functions μ are almost surely identified up to sign as well.

Rather than working with the system (5) of dimension n directly, we transform it into a system of dimension K. Let $W = (\omega_{ki})$ be a $K \times n$ weighting matrix of rank K. Then we can write $Wm = WB\mu$. Introducing the shorthands $S = WB$ and $g = Wm$, we obtain that

$$g = S\mu. \tag{8}$$

Here, $g = (g_1, \ldots, g_K)^\intercal$ are weighted averages of the individual functions m_i given by

$g_k = \sum_{i=1}^n \omega_{ki} m_i$. Moreover, the $K \times K$ matrix S contains weighted averages of the model parameters as its elements, in particular $S = (s_{kl})$ with $s_{kl} = \sum_{i=1}^n \omega_{ki}\beta_{il}$ for $k, l = 1, \ldots, K$. Note that the vectors m and g as well as the matrices B, W, and S depend on the cross-section dimension n. To keep the notation readable, this dependence is suppressed throughout the paper.

Premultiplying the n-dimensional system (5) with the matrix W, we form K different weighted averages of the individual functions m. We thus replace the system (5) which characterizes the individual functions m as linear combinations of the common components μ by a system which represents weighted averages of these functions as linear combinations of μ. The reason for this is twofold: Firstly, the system (8) has a fixed dimension K rather than a growing dimension n, which is technically more convenient. Secondly, the functions g being averages of the individual functions m, they can be estimated much more precisely than the latter. In particular, g can be estimated with a much faster convergence rate than the individual functions. This will help us to achieve a fast convergence rate for our estimator of μ as well.

The elements of the system (8) can be normalized in an analogous way as those of the system (5): To start with, we assume that the matrix S has full rank K and that the functions μ are orthonormal, i.e. $\int \mu(x)\mu(x)^\intercal w(x)dx = I_K$. By the same arguments as before, this is equivalent to the following assumption:

($\mathrm{I_W}$1) The matrix S is orthonormal, i.e. $S^\intercal S = I_K$, and $\int \mu(x)\mu(x)^\intercal w(x)dx$ is a diagonal matrix with non-zero diagonal entries.

Note that the normalization of the functions μ in ($\mathrm{I_W}$1) depends on the matrix S and thus on the chosen weighting matrix W. This becomes visible from equation (7) which shows how the normalized version of μ is constructed. As before, we additionally suppose that the normalized vector of functions μ has the following property:

($\mathrm{I_W}$2) The diagonal entries of the matrix $\int \mu(x)\mu(x)^\intercal w(x)dx$ are all distinct.

We finally put a slight regularity condition on the weighting scheme W:

($\mathrm{I_W}$3) The weights ω_{ki} are of the form $\omega_{ki} = v_{ki}/n$ with non-negative parameters $v_{ki} \le C < \infty$ for some sufficiently large constant C. For each k, the number n_k of nonzero weights is such that $n_k/n \to c_k$ for some positive constant c_k.

The above condition is satisfied by a wide range of weighting schemes, for example by the simple choice

$$W = \begin{pmatrix} \overbrace{\frac{1}{n} \cdots \frac{1}{n}}^{[n/K] \text{ times}} & & & 0 \\ & \frac{1}{n} \cdots \frac{1}{n} & & \\ & & \ddots & \\ 0 & & & \frac{1}{n} \cdots \frac{1}{n} \end{pmatrix}. \tag{9}$$

Note that by assuming n_k/n to converge to a positive limit, we just make sure that the averages which result from applying the weighting matrix W are composed of $O(n)$ terms. This allows us to apply asymptotic arguments to them later on.

Given the normalization conditions ($\mathrm{I_W}1$) and ($\mathrm{I_W}2$) together with the assumption on the weights ($\mathrm{I_W}3$), the functions μ can be represented as follows: As the columns of the matrix S are orthonormal, we can write

$$\mu = S^\intercal g. \tag{10}$$

The matrix S in this equation can be characterized by a spectral decomposition of the matrix $\Sigma = \int g(x)g(x)^\intercal w(x)dx$. In particular, it holds that

$$\Sigma = S \int \mu(x)\mu(x)^\intercal w(x)dx\ S^\intercal = SDS^\intercal,$$

where $D = \mathrm{diag}(\lambda_1, \dots, \lambda_K)$ with $\lambda_k = \int \mu_k^2(x)w(x)dx$. The constants $\lambda_1, \dots, \lambda_K$ are the non-zero distinct eigenvalues of Σ. Moreover, the columns of S are the corresponding orthonormal eigenvectors, denoted by $s_1, \dots, s_K$ in what follows.

In the sequel, we shall assume throughout that the functions μ and the matrix S are normalized to fulfill ($\mathrm{I_W}1$) and ($\mathrm{I_W}2$). Moreover, we suppose that the matrix Σ converges to a full-rank matrix Σ^*. These seem like reasonable and innocuous assumptions. Finally, note that given the existence of a limit Σ^*, the matrix S converges to a limit S^* as well. This is due to the fact that the eigenvectors $s_1, \dots, s_K$ depend continuously on the entries of the matrix Σ.

4 Estimation

We now describe our procedure to estimate the functions $\mu_1, \dots, \mu_K$ and the coefficient vectors $\beta_i = (\beta_{i1}, \dots, \beta_{iK})^\intercal$ based on kernel methods. Of course, alternative methods can be used, including the iterative algorithms developed in Chen, Gao, and Li (2013a) or the sieve methods described in Chen (2010). One advantage of our procedures is that they are "in closed form" meaning that one does not have to rely on nonlinear optimization and that they can be computed very fast and accurately even with very large datasets.

For simplicity of exposition, we assume throughout the section that the number K of common components is known. In Sections 6 and 7, we will dispense with this assumption and provide a procedure to estimate K. Our approach splits up into four steps, each of which is described in a separate subsection. To start with, we construct preliminary estimators of the individual regression functions m_i. These are used to obtain estimators of the μ-functions and the coefficient vectors β_i in a second and third step, respectively. In a final step, we exploit the model structure to obtain improved estimators of the individual regression functions m_i.

4.1 Preliminary estimators of the individual functions

We estimate the individual functions m_i by applying nonparametric kernel techniques to the time series data $\{(Y_{it}^{\text{fe}}, X_{it}) : t = 1, \dots, T\}$. More specifically, Nadaraya-Watson or local linear smoothers may be used. The Nadaraya-Watson estimator of the function m_i is defined as

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

where h is a scalar bandwidth and $K(\cdot)$ denotes a kernel satisfying $\int K(u) du = 1$ and $K_h(\cdot) = h^{-1} K(h^{-1} \cdot)$. The local linear estimator of m_i is given by the formula

$$\widehat{m}_i^{\text{LL}}(x) = \frac{\sum_{t=1}^T w_{i,T}(x, X_{it}) Y_{it}^{\text{fe}}}{\sum_{t=1}^T w_{i,T}(x, X_{it})},$$

with

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

and

$$S_{i,T,k}(x) = \frac{1}{T} \sum_{t=1}^T K_h(x - X_{it}) \Big(\frac{x - X_{it}}{h} \Big)^k$$

for $k = 1, 2$; see Fan and Gijbels (1995) for a detailed account of the local linear smoothing method. The procedure to estimate the functions μ and the parameter vectors β_i is the same no matter whether we work with Nadaraya-Watson or local linear smoothers. In what follows, we thus use the symbol $\widehat{m}_i$ to denote either the local constant estimator $\widehat{m}_i^{\text{NW}}$ or the local linear smoother $\widehat{m}_i^{\text{LL}}$.

4.2 Estimating the common component functions $\boldsymbol{\mu}$

We now use the characterization (10) of the functions μ to construct an estimator of them. We proceed as follows:

Step 1: Construct estimators $\widehat{g} = (\widehat{g}_1, \dots, \widehat{g}_K)^\intercal$ of the functions $g = (g_1, \dots, g_K)^\intercal$ according to

$$\widehat{g}_k(x) = \sum_{i=1}^n \omega_{ki} \widehat{m}_i(x).$$

Step 2: Estimate the matrix Σ by

$$\widehat{\Sigma} = \int \widehat{g}(x) \widehat{g}(x)^\intercal w(x) dx.$$

Step 3: Estimate the eigenvalues and eigenvectors by

$$\widehat{\Sigma} = \widehat{S} \widehat{D} \widehat{S}^\intercal,$$

i.e., by performing an eigenvalue decomposition of the matrix $\widehat{\Sigma}$. Let $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_K$ be the eigenvalues of $\widehat{\Sigma}$ (i.e. the diagonal entries of the matrix $\widehat{D}$), and $\widehat{s}_1, \ldots, \widehat{s}_K$ the corresponding orthonormal eigenvectors (i.e. the columns of the matrix $\widehat{S}$).

Step 4: Define the estimator of μ by replacing S and g in (10) with their respective estimators, i.e.,

$$\widehat{\mu} = \widehat{S}^\intercal \widehat{g}.$$

4.3 Estimating the coefficients $\boldsymbol{\beta_i}$

Consider the time series data $\{(Y_{it}, X_{it}) : t = 1, \ldots, T\}$ of the i-th individual. These are assumed to come from the model

$$Y_{it} = \mu_0 + \alpha_i + \gamma_t + \sum_{k=1}^{K} \beta_{ik} \mu_k(X_{it}) + \varepsilon_{it}$$

for $t = 1, \ldots, T$, which is linear in the parameters $\beta_i = (\beta_{i1}, \ldots, \beta_{iK})^\intercal$. If the functions $\mu_1, \ldots, \mu_K$ were known, the coefficients β_i could be estimated by standard least squares methods from the time series data $\{(Y_{it}^{\text{fe}}, X_{it}) : t = 1, \ldots, T\}$. In particular, we could use a weighted least squares estimator given by

$$\widetilde{\beta}_i = \Big(\frac{1}{T} \sum_{t=1}^{T} \pi(X_{it}) \mu(X_{it}) \mu(X_{it})^\intercal \Big)^{-1} \frac{1}{T} \sum_{t=1}^{T} \pi(X_{it}) \mu(X_{it}) Y_{it}^{\text{fe}} \tag{11}$$

with a weighting function π. As the functions μ are not known, we replace them by the estimates $\widehat{\mu}$, thus yielding the estimator

$$\widehat{\beta}_i = \Big(\frac{1}{T} \sum_{t=1}^{T} \pi(X_{it}) \widehat{\mu}(X_{it}) \widehat{\mu}(X_{it})^\intercal \Big)^{-1} \frac{1}{T} \sum_{t=1}^{T} \pi(X_{it}) \widehat{\mu}(X_{it}) Y_{it}^{\text{fe}}. \tag{12}$$

4.4 Re-estimating the functions $\boldsymbol{m_i}$ and iterating the estimation procedure

Exploiting the model structure, we can now define new estimators of the individual functions m_i which have better asymptotic properties than the preliminary estimators $\widehat{m}_i$. Specifically, we let

$$\widehat{m}_i^e(x) = \widehat{\beta}_i^\intercal \widehat{\mu}(x).$$

As we will see later on, the estimators $\widehat{m}_i^e$ have a faster convergence rate than the preliminary smoothers $\widehat{m}_i$.

A possible extension of our estimation procedure is to iterate it. To do so, we first re-estimate the component functions μ and the parameters β_i by using $\widehat{m}_i^e$ instead of the preliminary smoothers $\widehat{m}_i$. This yields updated estimators of μ and β_i. In addition, we may update the estimated individual effects whose first round estimates were implicitly given by $\widehat{\alpha}_i = \overline{Y}_i - \overline{\overline{Y}}$, $\widehat{\gamma}_t = \overline{Y}_t - \overline{\overline{Y}}$, and $\widehat{\mu}_0 = \overline{\overline{Y}}$. Specifically, these may be replaced by:

$$\widehat{\alpha}_i^e = \frac{1}{T}\sum_{t=1}^{T}\left\{Y_{it} - \widehat{\mu}_0 - \widehat{m}_i^e(X_{it})\right\}; \quad \widehat{\gamma}_t^e = \frac{1}{n}\sum_{i=1}^{n}\left\{Y_{it} - \widehat{\mu}_0 - \widehat{m}_i^e(X_{it})\right\};$$

$$\widehat{\mu}_0^e = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left\{Y_{it} - \widehat{m}_i^e(X_{it})\right\}.$$

This process can be continued until some convergence criterion is satisfied, which is likely to be achieved in practice quite quickly. Note that we can view this iterative algorithm as a procedure to find the minimum of a least squares objective function along the lines of Connor, Linton, and Hagmann (2012).

5 Asymptotics

In what follows, we derive the asymptotic properties of our estimators. To start with, we list the assumptions needed for our analysis. We then present the results on the limiting behaviour of the estimators $\widehat{\mu}$, $\widehat{\beta}_i$, and $\widehat{m}_i^e$. The proofs of our theoretical results can be found in Appendix A.

5.1 Assumptions

We impose the following regularity conditions, which as usual are sufficient but not necessary for our results. The expression $T^a \ll n \ll T^b$ is used to mean that $CT^{a+\delta} \le n \le CT^{b-\delta}$ for some positive constant C, a small $\delta > 0$ and $0 < a < b$. The symbol $\gg$ is used analogously.

(A1) The data $\{(X_{it}, \varepsilon_{it}) : i = 1, \dots, n,\ t = 1, \dots, T\}$ are independent across i. Moreover, they are strictly stationary and strongly mixing (Rosenblatt, 1956) in the time direction. Let $\alpha_i(k)$ for $k = 1, 2, \dots$ be the mixing coefficients of the time series $\{(X_{it}, \varepsilon_{it}), t = 1, \dots, T\}$ of the i-th individual. It holds that $\alpha_i(k) \le \alpha(k)$ for all $i = 1, \dots, n$, where the coefficients $\alpha(k)$ decay exponentially fast to zero as $k \to \infty$.

(A2) The densities f_i of the variables X_{it} exist and have bounded support, $[0, 1]$ say. Moreover, they are uniformly bounded away from zero and from above, i.e., $0 < c \le \min_{1 \le i \le n} \inf_{x \in [0,1]} f_i(x)$ as well as $\max_i \sup_x f_i(x) \le C < \infty$ for some pair of constants $0 < c \le C < \infty$. Finally, the joint densities $f_{i;l}$ of (X_{it}, X_{it+l}) exist and are also uniformly bounded from above.

(A3) The functions $\mu_1, \dots, \mu_K$ are twice continuously differentiable on $[0, 1]$. Moreover, the densities f_i are twice continuously differentiable on $[0, 1]$ as well with uniformly bounded first and second derivatives f_i' and f_i''. Finally, the coefficients β_{ik} are bounded by some constant $\overline{\beta} < \infty$, i.e., $|\beta_{ik}| \le \overline{\beta}$ for all $i = 1, \dots, n$ and $k = 1, \dots, K$, which ensures that the functions m_i as well as the derivatives m_i' and m_i'' are uniformly bounded on $[0,1]$ as well.

(A4) It holds that $\mathbb{E}[\varepsilon_{it}|X_{it}] = 0$. Moreover, for some $\theta > 5$ and for all $l \in \mathbb{Z}$,

$$\max_{1 \le i \le n} \sup_{x \in [0,1]} \mathbb{E}\big[|\varepsilon_{it}|^{\theta}\big|X_{it} = x\big] \le C < \infty \tag{13}$$

$$\max_{1 \le i \le n} \sup_{x,x' \in [0,1]} \mathbb{E}\big[|\varepsilon_{it}|\big|X_{it} = x, X_{it+l} = x'\big] \le C < \infty \tag{14}$$

$$\max_{1 \le i \le n} \sup_{x,x' \in [0,1]} \mathbb{E}\big[|\varepsilon_{it}\varepsilon_{it+l}|\big|X_{it} = x, X_{it+l} = x'\big] \le C < \infty, \tag{15}$$

where C is a sufficiently large constant independent of l.

(A5) The cross-section dimension $n = n(T)$ depends on T and satisfies $T^{2/3} \ll n \ll T^{3/2}$.

(A6) The bandwidth h is of the order $(nT)^{-(1/5+\delta)}$ for some small $\delta > 0$.

(A7) The kernel K is bounded, symmetric about zero and has compact support ($[-C_1, C_1]$, say). Moreover, it fulfills the Lipschitz condition that there exists a positive constant L with $|K(u) - K(v)| \le L|u - v|$. Let $\mu_2(K) = \int K(\varphi)\varphi^2 d\varphi$ and $\|K\|_2^2 = \int K^2(\varphi) d\varphi$.

Note that we do not necessarily require exponentially decaying mixing rates as assumed in (A1). These could alternatively be replaced by sufficiently high polynomial rates. We nevertheless make the stronger assumption (A1) to keep the notation and structure of the proofs as clear as possible.

The cross-sectional independence of the data is maintained for simplicity, one could however allow some forms of dependence in the cross-section. For example, one could allow the type of clustering structure used in Connor, Hagmann, and Linton (2012). Our results would go through with minimal changes in this case. An alternative approach is to follow Connor and Koraczyk (1993) and to assume that there exists some ordering of the observations with respect to which the data $\{(X_{it}, \varepsilon_{it})\}$ are mixing across i. Jenish (2012) derives pointwise limit theorems for nonparametric regression with near-epoch dependent mixing processes defined on a general lattice dimension d, which includes that setting as a special case. Robinson (2011) has proposed an alternative approach based on linear processes that does not need a measure of cross-sectional distance. His framework allows for strongly dependent and nonstationary regression disturbances. These types of cross-sectional dependence are much harder to deal with in our framework and would involve a great deal of technical and notational effort to cope with. Heuristically speaking, however, we expect these dependence structures to have no effect on the asymptotic behaviour of our estimators provided the dependence is weak. Specifically, the cross-sectional dependence should wash out of the distribution for the nonparametric estimates and should not affect the univariate asymptotics for the loading coefficients.

We may also allow for nonstationarity in $\{(X_{it}, \varepsilon_{it})\}$ of the type proposed in Dahlhaus (1997). This so-called local stationarity may arise in the time direction, that is, densities change smoothly in the argument t/T. In addition, it may arise in the cross-section, that is, densities change smoothly in the argument i/n with respect to an unknown ordering of the individuals. Vogt (2012) establishes a number of results for nonparametric regression with locally stationary processes, and we anticipate that his results can be extended to this case, although the technical effort to accomplish this would be considerable.

It is worth mentioning that our assumptions do not only allow for time series dependence but also for heteroskedasticity in the error terms ε_{it}. The errors may for example have the form $\varepsilon_{it} = \sigma(X_{it})\eta_{it}$, where η_{it} are i.i.d. variables independent of X_{it} and σ is an unknown volatility function. The moment bounds (13)–(15) on the error terms are needed to derive a couple of uniform convergence results later on. They are modifications of standard assumptions required to derive uniform convergence rates for kernel estimators; cp. for example Assumption 2 in Hansen (2008). They are for instance satisfied when the error terms take the form $\varepsilon_{it} = \sigma(X_{it})\eta_{it}$, where η_{it} are i.i.d. with $\mathbb{E}|\eta_{it}|^\theta < \infty$ and σ is a continuous function.

Finally, note that there is a trade-off between the moment condition (13) in (A4) and the conditions on the relative sample sizes in (A5). For example, if we restrict attention to the case $n = O(T)$, we can do with $\theta > 4$ in condition (A4). The restrictions in (A5) reflect two constraints on the relative sample sizes: Firstly, T needs to be large enough relative to n such that the preliminary estimators are sufficiently precisely estimated. Secondly, n needs to be large enough such that the error terms stemming from the fixed effect transformation can be ignored.

5.2 Asymptotics for the estimator $\widehat{\mu}$

Our first result characterizes the asymptotic behaviour of the estimator $\widehat{\mu}$. In particular, it shows that $\widehat{\mu}$ uniformly converges to μ and is asymptotically normal. To formulate it, we define $V(x)$ to be a $K \times K$ matrix with the entries

$$V_{k,l}(x) = \|K\|_2^2 \lim_{n\to\infty} \Big(n \sum_{i=1}^n \omega_{ki}\omega_{li} \frac{\sigma_i^2(x)}{f_i(x)} \Big),$$

where $\sigma_i^2(x) = \mathbb{E}[\varepsilon_{it}^2 | X_{it} = x]$.

Theorem 5.1. *Let (A1)–(A7) together with (I$_W$1)–(I$_W$3) be satisfied. Then*

$$\sup_{x\in I_h} \|\widehat{\mu}(x) - \mu(x)\| = O_p\Big(\sqrt{\frac{\log nT}{nTh}}\Big). \tag{16}$$

Here, $I_h = [C_1 h, 1 - C_1 h]$ if our procedure is based on the Nadaraya-Watson smoothers $\widehat{m}_i^{NW}$ and $I_h = [0, 1]$ if it is based on the local linear smoothers $\widehat{m}_i^{LL}$. Moreover, for any *fixed point* $x \in (0,1)$,

$$\sqrt{nTh}(\widehat{\mu}(x) - \mu(x)) \xrightarrow{d} N(0, \nu(x)) \tag{17}$$

with $\nu(x) = (S^*)^\intercal V(x) S^*$ *and* S^* *being the limit of* S.

The first part of the theorem shows that $\widehat{\mu}$ converges to μ at a fast rate based on the pooled number of observations nT. If we set up our estimation procedure with the local linear smoothers $\widehat{m}_i^{\mathrm{LL}}$, the rate is uniform over the whole support $[0,1]$. For the Nadaraya-Watson based procedure in contrast, the rate is only uniform on the subinterval $[C_1 h, 1 - C_1 h]$ which converges to the support $[0,1]$ as the sample size increases. This is due to the fact that the Nadaraya-Watson estimators $\widehat{m}_i^{\mathrm{NW}}$ suffer from slow convergence rates at the boundary of the support.

The second part of the theorem specifies the asymptotic distribution of $\widehat{\mu}$. The asymptotic covariance matrix $\nu(x)$ can be seen to depend on the weights ω_{ki}. The reason for this is as follows: The normalization of the functions μ depends on the choice of the weighting matrix W. In particular, different choices of W generally result in different eigenvalues $\lambda_k = \int \mu_k^2(x) w(x) dx$, i.e., in different values of the L_2-norm of the functions μ_k. This becomes reflected in the covariance matrix $\nu(x)$ through its dependence on the weights ω_{ki}. Moreover, note that $\nu(x)$ need not be diagonal in general: If the weighting matrix W is diagonal, then $V(x)$ is a diagonal matrix as well. However, even then the matrix S^* may have a more complicated non-diagonal structure. Hence, the components of $\widehat{\mu}$ are asymptotically mutually correlated in general.

Regarding inference, we propose a simple plug-in method. Let $\widehat{\varepsilon}_{it} = Y_{it}^{\mathrm{fe}} - \widehat{m}_i(X_{it})$ and

$$\widehat{V}_{k,l}(x) = \|K\|_2^2 \; n \sum_{i=1}^{n} \omega_{ki} \omega_{li} \frac{\widehat{\sigma}_i^2(x)}{\widehat{f}_i(x)},$$

where $\widehat{\sigma}_i^2(x)$ is a local constant or local linear time series regression smoother of $\widehat{\varepsilon}_{it}^2$ on X_{it} and $\widehat{f}_i(x) = T^{-1} \sum_{t=1}^{T} K_h(X_{it} - x)$ is the time series kernel density estimator of $f_i(x)$. Then, $\widehat{\nu}(x) = \widehat{S}^\intercal \widehat{V}(x) \widehat{S}$ consistently estimates $\nu(x)$, and pointwise confidence intervals based on this are consistent under our assumptions, see Härdle (1991).

To derive the results of Theorem 5.1, we work with the undersmoothing assumption (A6) on the bandwidth h. Moreover, we use the same bandwidth both to estimate the average functions g and the matrix Σ. It is however also possible to employ different bandwidths. In particular, one may use a slightly undersmoothed bandwidth h_Σ of the order $(nT)^{-(1/5+\delta)}$ to construct the estimate $\widehat{\Sigma}$ and a bandwidth h_g of the optimal order $(nT)^{-1/5}$ to set up the estimator $\widehat{g}$. Inspecting the proof of Theorem 5.1, it is easily seen that in this case

$$\sqrt{nTh_g}(\widehat{\mu}(x) - \mu(x)) = S^\intercal \left[\sqrt{nTh_g}\big(\widehat{g}(x) - g(x)\big)\right] + o_p(1)$$

with

$$\sqrt{nTh_g}\big(\widehat{g}(x) - g(x)\big) \xrightarrow{d} N(B(x), V(x)),$$

where the variance $V(x)$ has already been defined above and the bias term $B(x)$ is given by $B^{\mathrm{NW}}(x)$ and $B^{\mathrm{LL}}(x)$ in the Nadaraya-Watson and the local linear based case, respectively. The latter two expressions are defined by

$$B_k^{\mathrm{NW}}(x) = \frac{c_0\mu_2(K)}{2} \lim_{n\to\infty} \sum_{i=1}^{n} \omega_{ki}\big(2m_i'(x)f_i'(x) + m_i''(x)f_i(x)\big)/f_i(x)$$
$$B_k^{\mathrm{LL}}(x) = \frac{c_0\mu_2(K)}{2} \lim_{n\to\infty} \sum_{i=1}^{n} \omega_{ki}m_i''(x)$$

for $k = 1,\dots,K$, where c_0 is the limit of the sequence values $\sqrt{nTh_g^5}$.

Given the above remarks, we suggest a straightforward rule of thumb for bandwidth selection. In particular, we first select the bandwidth h_g and then choose the bandwidth h_Σ simply by picking a value slightly smaller than the choice of h_g. To select the bandwidth h_g (or rather $h_{g,k}$ if we allow a different bandwidth for each function g_k), we optimize the integrated mean-squared error criterion

$$\mathrm{IMSE}(h_{g,k}) = h_{g,k}^4 \int B_k^2(x)dx + \frac{1}{nTh_{g,k}} \int V_{k,k}(x)dx$$

for $k = 1,\dots,K$. Minimizing with respect to $h_{g,k}$, the optimal bandwidth turns out to be given by

$$h_{g,k}^* = \left(\frac{\int V_{k,k}(x)dx}{4\int B_k^2(x)dx}\right)^{\frac{1}{5}} (nT)^{-1/5}.$$

This expression still depends on some unknown quantities which have to be replaced by estimators. To do so, we apply a simple plug-in rule similar to the methods discussed in Fan and Gijbels (1994).

5.3 Asymptotics for the parameter estimators $\widehat{\boldsymbol{\beta}}_i$

The next theorem describes the asymptotic properties of the parameter estimates $\widehat{\beta}_i$ for a fixed individual i. To state the asymptotic distribution of $\widehat{\beta}_i$, we introduce the shorthands

$$\Gamma_i = \mathbb{E}[\pi(X_{i0})\mu(X_{i0})\mu(X_{i0})^\intercal] \quad \text{and} \quad \Psi_i = \sum_{l=-\infty}^{\infty} \mathrm{Cov}(\chi_{i0}, \chi_{il}),$$

where $\chi_{it} = \{\pi(X_{it})\mu(X_{it}) - \mathbb{E}[\pi(X_{it})\mu(X_{it})]\}\varepsilon_{it} - \mathbb{E}[\pi(X_{it})\mu(X_{it})]m_i(X_{it})$ and π is a bounded weighting function.

Theorem 5.2. *Suppose that all the assumptions of Theorem 5.1 are fulfilled and let Γ_i have full rank. Then for any fixed i,*

$$\sqrt{T}(\widehat{\beta}_i - \beta_i) \xrightarrow{d} N\big(0, \Gamma_i^{-1}\Psi_i(\Gamma_i^{-1})^\intercal\big).$$

If our procedure is based on Nadaraya-Watson smoothers, we have to restrict the weighting function π to equal zero within the boundary region $[0, C_1 h) \cup (1 - C_1 h, 1]$. This is necessary because the convergence rate of $\widehat{\mu}$ is only uniform over the interval $[C_1 h, 1 - C_1 h]$ in this case. If the local linear based procedure is applied, we do not have to impose any restrictions on π.

From the proof of Theorem 5.2, we can see that our parameter estimators $\widehat{\beta}_i$ have some type of oracle property. In particular, it holds that $\sqrt{T}(\widehat{\beta}_i - \widetilde{\beta}_i) = o_p(1)$. Our estimators $\widehat{\beta}_i$ thus have the same asymptotic distribution as the oracle estimators $\widetilde{\beta}_i$ which are constructed under the assumption that the functions $\mu_1, \ldots, \mu_K$ are known. To estimate the asymptotic variance Ψ_i, we may apply standard long-run variance estimation procedures to the residuals $\widehat{\chi}_{it}$ given by

$$\widehat{\chi}_{it} = \{\pi(X_{it})\widehat{\mu}(X_{it}) - \widehat{\pi\mu}\}\widehat{\varepsilon}_{it} - \widehat{\pi\mu}\widehat{m}_i^e(X_{it}),$$

where we define $\widehat{\pi\mu} = T^{-1}\sum_{t=1}^T \pi(X_{it})\widehat{\mu}(X_{it})$, $\widehat{\varepsilon}_{it} = Y_{it}^{\text{fe}} - \widehat{m}_i^e(X_{it})$ and $\widehat{m}_i^e(x) = \widehat{\beta}_i^\intercal \widehat{\mu}(x)$.

5.4 Asymptotics for the estimators $\widehat{m}_i^e$ and a parameter of interest

We finally discuss the asymptotic properties of the estimator $\widehat{m}_i^e(x) = \widehat{\beta}_i^\intercal \widehat{\mu}(x)$. It holds that

$$\widehat{m}_i^e(x) - m_i(x) = (\widehat{\beta}_i - \beta_i)^\intercal \mu(x) + \beta_i^\intercal(\widehat{\mu}(x) - \mu(x)) + o_p\Big(\frac{1}{\sqrt{nT}}\Big). \tag{18}$$

The first term on the right-hand side is of the order $T^{-1/2}$, while the second one has the (pointwise) order $(nTh)^{-1/2}$ under our conditions. Given assumption (A5) on the relationship between the dimensions n and T, the leading term is the first one of order $T^{-1/2}$. It follows that $\widehat{m}_i^e(x)$ is asymptotically normal at the rate $T^{-1/2}$, i.e., at a faster rate than the preliminary estimator $\widehat{m}_i(x)$ which converges at the (pointwise) rate $(Th)^{-1/2}$.

In our application below, we are interested in the parameter $c_i = m_i(1) - m_i(0)$, which measures the difference between monopoly and competition. Defining $\widehat{c}_i = \widehat{m}_i^e(1) - \widehat{m}_i^e(0)$, we obtain that

$$\widehat{c}_i - c_i = (\widehat{\beta}_i - \beta_i)^\intercal(\mu(1) - \mu(0)) + \beta_i^\intercal(\widehat{\mu}(1) - \mu(1)) - \beta_i^\intercal(\widehat{\mu}(0) - \mu(0)) + o_p\Big(\frac{1}{\sqrt{nT}}\Big).$$

Under the null hypothesis that $c_i = 0$, we should observe that

$$\sqrt{T}\widehat{c}_i \xrightarrow{d} N(0, \tau_i) \quad \text{with} \quad \tau_i = (\mu(1) - \mu(0))^\intercal \Gamma_i^{-1}\Psi_i(\Gamma_i^{-1})^\intercal(\mu(1) - \mu(0)),$$

which could form the basis of a test. Specifically, we can use the strategy to estimate the covariance matrix $\Gamma_i^{-1}\Psi_i(\Gamma_i^{-1})^\intercal$ from the previous subsection together with the estimators $\widehat{\mu}$ to obtain a consistent estimator $\widehat{\tau}_i$ of the asymptotic variance τ_i and let

$$t_i = \frac{\widehat{c}_i}{\sqrt{\widehat{\tau}_i/T}},$$

which is asymptotically standard normal.

6 Robustness of the estimation method

So far, we have worked under the simplifying assumption that the number K of common component functions $\mu_1, \ldots, \mu_K$ is known. We now drop this assumption and take into account that K is usually not observed in applications. We only suppose that there is some known upper bound $\overline{K}$ of the number of component functions. In what follows, we investigate how our procedure behaves if we work with this upper bound instead of the true number of components.

To do so, let $\overline{W} = (\overline{\omega}_{ki})$ be a $\overline{K} \times n$ weighting matrix satisfying ($\mathrm{I_W}$3). Writing $\overline{g} = \overline{W}m$ and $\overline{S} = \overline{W}B$, we obtain that

$$\overline{g} = \overline{S}\,\mu.$$

Using an analogous normalization as in Section 3, we can assume that (i) the matrix $\int \mu(x)\mu(x)^\intercal w(x)dx$ is diagonal with positive and distinct diagonal entries and that (ii) $\overline{S}$ is a $\overline{K} \times K$ matrix with orthonormal columns. Note that this normalization is somewhat different from that used in the previous sections as we have replaced the weighting scheme W by $\overline{W}$. For simplicity, we suppress this difference in the notation in what follows and again denote the normalized component functions by μ. We thus obtain that

$$\mu = \overline{S}^\intercal \overline{g}.$$

As in the case with known K, the matrix $\overline{S}$ can be characterized by an eigenvalue decomposition of the $\overline{K} \times \overline{K}$ matrix

$$\overline{\Sigma} = \int \overline{g}(x)\overline{g}(x)^\intercal w(x)dx.$$

In particular, it holds that $\overline{\Sigma} = \overline{S}D\overline{S}^\intercal$ with $D = \int \mu(x)\mu(x)^\intercal w(x)dx$. Note that this way of writing the spectral decomposition implicitly presupposes that K is known. For this reason, it is more appropriate to rewrite the decomposition as $\overline{\Sigma} = \overline{U}\,\overline{D}\,\overline{U}^\intercal$. Here, $\overline{U}$ is an orthonormal $\overline{K} \times \overline{K}$ matrix with the first K columns being equal to $\overline{S}$. Moreover, $\overline{D} = \int \overline{\mu}(x)\overline{\mu}(x)^\intercal w(x)dx$ is a diagonal $\overline{K} \times \overline{K}$ matrix with $\overline{\mu} = (\mu, 0, \ldots, 0)$ being a vector of length $\overline{K}$. Similarly to the case with known K, we assume that $\overline{\Sigma}$ converges to a matrix $\overline{\Sigma}^*$ of rank K.

To estimate the vector of functions $\overline{\mu} = (\mu, 0, \ldots, 0)$, we mimic the estimation procedure from Subsection 4.2. In particular, we proceed as follows:

Step 1: Estimate the function $\overline{g}_k(x)$ by $\widetilde{g}_k(x) = \sum_{i=1}^n \overline{\omega}_{ki} \widehat{m}_i(x)$ for $k = 1 \ldots, \overline{K}$.

Step 2: Estimate the matrix $\overline{\Sigma}$ by $\widetilde{\Sigma} = \int \widetilde{g}(x) \widetilde{g}(x)^\intercal w(x) dx$.

Step 3: Perform an eigenvalue decomposition of $\widetilde{\Sigma}$ to obtain estimators of $\overline{U}$ and $\overline{D}$. In particular, write $\widetilde{\Sigma} = \widetilde{U} \widetilde{D} \widetilde{U}^\intercal$ with $\widetilde{D}$ being diagonal and $\widetilde{U}$ being orthonormal.

Step 4: Estimate the vector of functions $\overline{\mu} = (\mu, 0, \ldots, 0)$ by

$$\widetilde{\mu} = \widetilde{U}^\intercal \widetilde{g}.$$

Inspecting the proof of Theorem 5.1, it is straightforward to see that for $k = 1, \ldots, K$, the estimator $\widetilde{\mu}_k$ has analogous asymptotic properties as $\widehat{\mu}_k$. In particular, it uniformly converges to μ_k and is asymptotically normal. The next theorem summarizes the properties of $\widetilde{\mu}_k$ for $k = 1, \ldots, K$. To formulate it, we let $\overline{V}(x)$ be a $\overline{K} \times \overline{K}$ matrix with the entries

$$\overline{V}_{k,l}(x) = \|K\|_2^2 \lim_{n \to \infty} \Big(n \sum_{i=1}^n \overline{\omega}_{ki} \overline{\omega}_{li} \frac{\sigma_i^2(x)}{f_i(x)} \Big),$$

where $\overline{\omega}_{ki}$ are the elements of the weighting matrix $\overline{W}$.

Theorem 6.1. *Let (A1)–(A7) be fulfilled. Then it holds that*

$$\sup_{x \in I_h} \big| \widetilde{\mu}_k(x) - \mu_k(x) \big| = O_p\Big(\sqrt{\frac{\log nT}{nTh}} \Big) \tag{19}$$

for all $k = 1, \ldots, K$. As before, $I_h = [C_1 h, 1 - C_1 h]$ for the Nadaraya-Watson based case and $I_h = [0, 1]$ for the local linear based procedure. Moreover, for any fixed point $x \in (0, 1)$,

$$\sqrt{nTh} \, [\widetilde{\mu}(x) - \mu(x)] \xrightarrow{d} N(0, \overline{\nu}(x)), \tag{20}$$

where $\overline{\nu}(x) = (\overline{S}^*)^\intercal \overline{V}(x) \, \overline{S}^*$ and $\overline{S}^*$ is the limit of $\overline{S}$.

In addition, we can show that for $k = K + 1, \ldots, \overline{K}$, the estimators $\widetilde{\mu}_k$ converge in an L_2-sense to zero.

Theorem 6.2. *Let (A1)–(A7) be fulfilled. Then it holds that*

$$\int \widetilde{\mu}_k^2(x) w(x) dx = o_p\Big(\frac{1}{\sqrt{nTh}} \Big) \tag{21}$$

for all $k = K + 1, \ldots, \overline{K}$.

The proof of Theorem 6.2 is given in Appendix A. Taken together, Theorems 6.1 and 6.2 show that our procedure is robust to overestimating the number of component functions K. In particular, applying it with the upper bound $\overline{K}$ instead of K, the first K components of the estimator $\widetilde{\mu}$ still uniformly converge to the vector of functions μ. Moreover, the remaining components converge to zero in an L_2-sense and thus become negligible as the sample size grows.

7 Selecting the number of components K

In this section, we propose a simple method to estimate the unknown number of components K. To define our estimator, let $\overline{\lambda} = (\overline{\lambda}_1, \ldots, \overline{\lambda}_{\overline{K}})^\intercal$ be the vector of eigenvalues of the matrix $\overline{\Sigma}$ arranged in descending order. Analogously, let $\widetilde{\lambda}$ be the eigenvalues of the estimator $\widetilde{\Sigma}$. Finally, let $\{\delta_{n,T}\}$ be any null sequence which converges to zero at the order $O(1/\sqrt{nTh})$ or at a slower rate. With this notation at hand, our estimator of K is defined as

$$\widehat{K} = \min \left\{ k \in \{1, \ldots, \overline{K}\} \;\middle|\; \frac{\widetilde{\lambda}_1 + \ldots + \widetilde{\lambda}_k}{\widetilde{\lambda}_1 + \ldots + \widetilde{\lambda}_{\overline{K}}} \geq 1 - \delta_{n,T} \right\}.$$

The intuition behind this estimator is simple: Under our assumptions, the matrix $\overline{\Sigma}$ has K non-zero eigenvalues, i.e., the first K entries of $\overline{\lambda}$ are non-zero. The first K entries of the estimator $\widetilde{\lambda}$ thus converge to some positive values, whereas the other ones approach zero as the sample size increases. Hence, the ratio

$$\frac{\widetilde{\lambda}_1 + \ldots + \widetilde{\lambda}_k}{\widetilde{\lambda}_1 + \ldots + \widetilde{\lambda}_{\overline{K}}}$$

should converge to a number strictly smaller than 1 for $k < K$ and to 1 for $k \geq K$. This suggests that $\widehat{K}$ consistently estimates the true number of components K.

This intuition can easily be turned into a formal argument: First of all, it can be shown that the convergence rate of $\widetilde{\lambda}$ is at least $o_p(1/\sqrt{nTh})$, i.e., $\|\widetilde{\lambda} - \overline{\lambda}\| = o_p(1/\sqrt{nTh})$. As a consequence, it holds that

$$\frac{\widetilde{\lambda}_1 + \ldots + \widetilde{\lambda}_k}{\widetilde{\lambda}_1 + \ldots + \widetilde{\lambda}_{\overline{K}}} = \frac{\lambda_1 + \ldots + \lambda_k}{\lambda_1 + \ldots + \lambda_{\overline{K}}} + o_p\Big(\frac{1}{\sqrt{nTh}}\Big).$$

for any $k \in \{1, \ldots, \overline{K}\}$. In particular,

$$\frac{\widetilde{\lambda}_1 + \ldots + \widetilde{\lambda}_K}{\widetilde{\lambda}_1 + \ldots + \widetilde{\lambda}_{\overline{K}}} = 1 + o_p\Big(\frac{1}{\sqrt{nTh}}\Big).$$

Using these two equations together with some straightforward arguments, it is easily seen that $\widehat{K}$ is indeed a consistent estimator of the true number of components K, i.e. $\widehat{K} = K + o_p(1)$.

When implementing the estimator $\widehat{K}$ in practice, an important question is how to choose the constant $\delta_{n,T}$. We suggest to pick it by a rule of thumb which is similar to the procedure usually used in principal component analysis for selecting the number of factors. To understand the intuitive idea behind the rule, first note that $\lambda_k = \int \mu_k^2(x)w(x)dx$ for $k = 1, \dots, K$ and $\lambda_k = 0$ for $k = K+1, \dots, \overline{K}$. The eigenvalues λ_k are thus equal to (the square of) a weighted L_2-norm of the component functions $\overline{\mu} = (\mu, 0, \dots, 0)$. Put differently, they measure the variation of these functions. As a result, the ratio

$$\frac{\lambda_1 + \ldots + \lambda_k}{\lambda_1 + \ldots + \lambda_{\overline{K}}}$$

can be interpreted to capture the percentage of the overall variation in the functions $\overline{\mu}$ that stems from the first k components. Hence, by picking a certain value of $\delta_{n,T}$, we select the number of component functions such that at least a certain percentage of the overall variation is explained by the chosen number of components. For instance, if we let $\delta_{n,T} = 0.05$, we pick the number of components to capture at least 95% of the total variation. Keeping in mind that our estimation procedure is robust to picking the number of components too large, we propose to choose the constant $\delta_{n,T}$ rather small (e.g. $\delta_{n,T} = 0.01$ or $\delta_{n,T} = 0.05$). This results in a conservative rule which tends to overestimate the true number K rather than to underestimate it. As already noted above, this way of selecting the number of components is very similar to the usual approach in factor analysis (see e.g. Zhu & Ghodsi (2006) or Chapter 6 of Jolliffe (2002)).

8 Application

The implementation of the “Markets in Financial Instruments Directive (MiFID)” ended the monopoly of primary security exchanges in Europe and served as a catalyst for the soaring of competition between marketplaces we observe today. The first round of MiFID was implemented on November 1st, 2007, but fragmentation of the UK equity market began sometime before that, and by 13th July, 2007, Chi-X was actively trading all of the FTSE 100 stocks. In October 2012, the volume of the FTSE 100 stocks traded via the London Stock Exchange had declined to 64%.[3]

There are theoretical reasons why the current trend towards fragmentation of order flow can improve market quality. Higher competition generally promotes technological innovation, improves efficiency and reduces the fees that have to be paid by investors. On the other hand, there are reasons to think that security exchanges are natural monopolies. Consolidated exchanges enjoy economies of scale because establishing a new exchange requires the payment of high fixed costs. Every additional trade lowers the average cost of the exchange. In addition, a single, consolidated exchange market creates network externalities. The larger the market, the more trading opportunities exist that attract

[3] `www.batstrading.co.uk/market_data/market_share/index`, assessed on October 20, 2012

even more traders.[4]

In view of these ambiguous theoretical predictions about the effect of order flow fragmentation on market quality, many researchers have approached this question with empirical methods. Gresse (2011) finds that increased competition between trading venues creates more liquidity – measured by spreads and best-quote depth – in a sample of stocks listed on the LSE and Euronext exchanges in Amsterdam, Paris and Brussels. The results of Degryse et al. (2011) suggest that fragmentation on trading venues with a visible order book improves global liquidity, but has a negative effect on local liquidity. On "dark" platforms with an invisible order book, liquidity is lower in more fragmented markets. O'Hara and Ye (2011) study the effect of market fragmentation on market quality in US equity markets and find that more fragmented stocks are associated with lower transaction costs and higher volatility.

The previous literature is subject to several methodological caveats. First, both Gresse (2011) and Degryse et al. (2011) assume that the conditional expectation of market quality on fragmentation is homogenous across all stocks. However, O'Hara and Ye (2011) provide evidence that the effect of fragmentation on market quality varies significantly across stocks. If there is indeed heterogeneity in the conditional expectation of market quality on fragmentation, the estimates are biased and policy implications can be misleading (Pesaran and Smith, 1995). In addition, previous studies use a parametric econometric model that presupposes a functional form for the effect of fragmentation on market quality. Gresse (2011) and O'Hara and Ye (2011) assume a linear functional form, while Degryse et al. (2011) specify a quadratic relationship. If the true regression model has a different functional form, then these studies suffer from misspecification which questions the validity of the results. The semiparametric model for heterogenous panel data we develop in this paper can address these limitations of previous work.

8.1 Data

Data on the volume of the individual FTSE 100 and FTSE 250 stocks traded on each equity venue was supplied to us by Fidessa. The data is recorded on a weekly basis and covers the period from May 2008 to June 2011. In total, we have $n = 350$ and $T = 152$ observations, which is broadly consistent with our assumptions. We use the volume traded on different venues to compute the Herfindahl index as a measure of market fragmentation.[5] In May 2008, equity trading in the UK was consolidated at the LSE as reflected by a Herfindahl index of 0.6 (Figure 1). By June 2011, the entry of new trading venues has changed the structure of the UK equity market dramatically: The Herfindahl index has fallen by about half over the sample period.

[4] These network externalities, however, are weakened as traders can now simultaneously access multiple markets via Smart Order Routing Technologies.

[5] The Herfindahl index of a stock is calculated as the sum of the squared market shares of the exchanges where the stock was traded. A value of 1 indicates a perfectly monopolistic market.

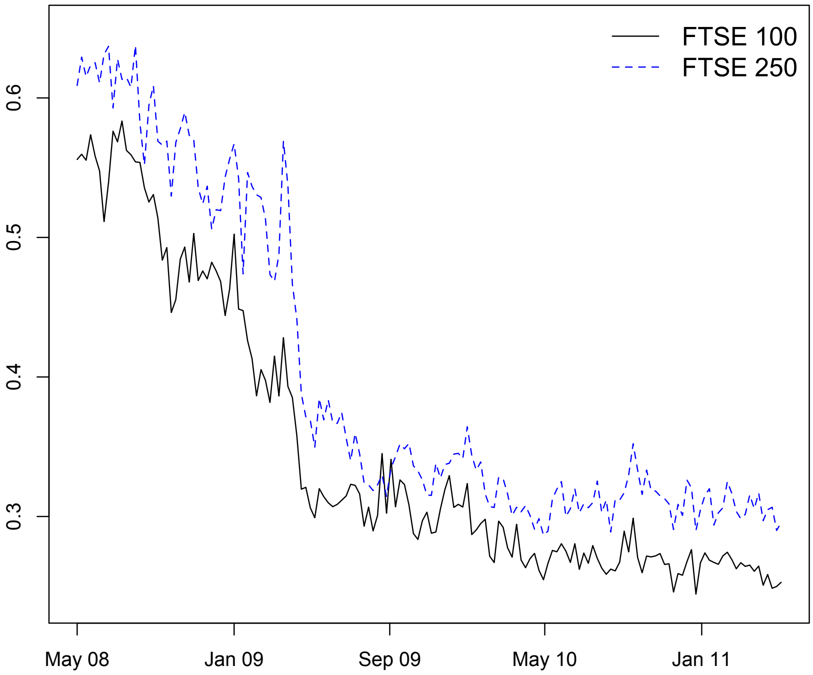

Figure 1: The Herfindahl index for the FTSE 100 and FTSE 250 stock indices. Data source: Fidessa.

The data allows us to distinguish between public exchanges with a visible order book ("lit venues"), venues with an invisible order book ("dark pools"), over the counter ("OTC") venues, and systematic internalizers ("SI venues").[6] It is interesting to inspect the evolution of volume traded at the different venue categories (Figure 2). The share of volume traded at dark, OTC and SI venues increased over the sample period, while the share of volume traded at lit venues has fallen considerably. For all categories, the observed changes are largest in the year 2009. In the period after 2009, volumes have approximately stabilized with the exception of dark venues. Quantitatively, the majority of trades are executed on lit and OTC venues while dark and SI venues attract only about 1% of the order flow.

We measure market quality by volatility and bid-ask spreads of the FTSE 100 and 250 stocks. Both measures of market quality are constructed as weekly medians of the daily measures. Volatility is calculated as the difference between price high and price low, scaled by price low. Bid-ask spreads are constructed as the difference between ask and bid price scaled by the midpoint. The evolution of volatility over the sample period clearly shows the effect of the global financial crisis in 2008/2009 (Figure 3).[7]

[6] The list of lit venues includes: Bats Europe, Chi-X, Equiduct, LSE, Nasdaq Europe, Nyse Arca, and Turquoise. The list of dark pools includes: BlockCross, Instinet BlockMatch, Liquidnet, Nomura NX, Nyfix, Posit, Smartpool, and UBS MTF. The list of OTC venues includes: Boat xoff, Chi-X OTC, Euronext OTC, LSE xoff, Plus, XOFF, and xplu/o. The list of SI venues includes: Boat SI and London SI.

[7] We do not show the evolution of the bid-ask spread as it does not exist for the index.

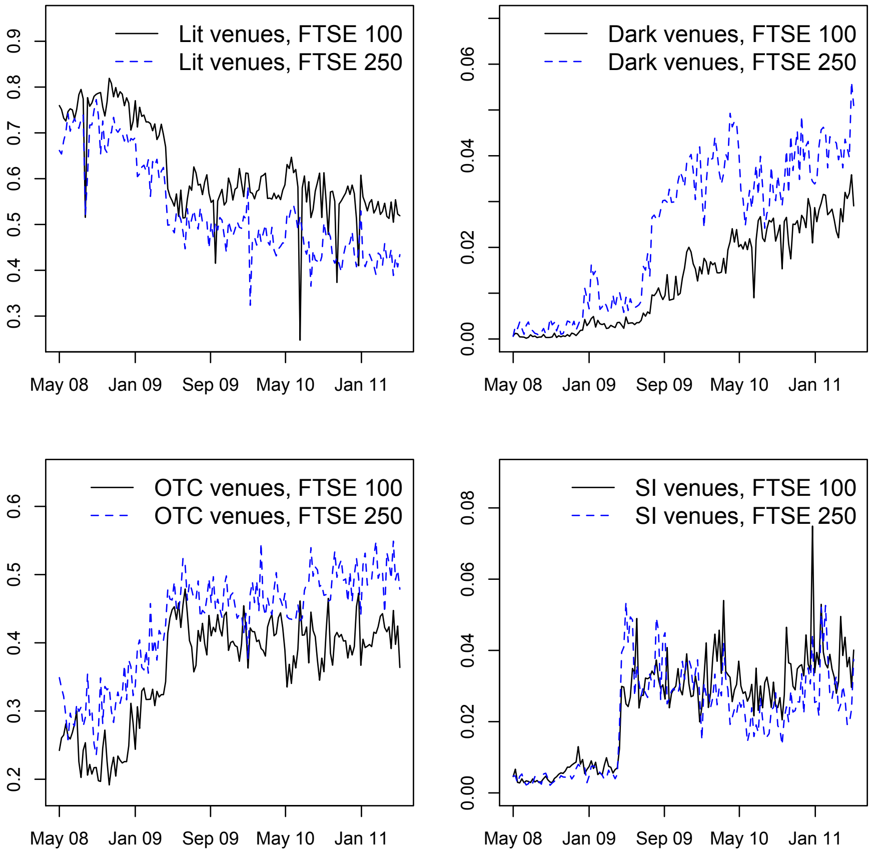

Figure 2: Share of volume traded by venue category. Data source: Fidessa.

8.2 The effect of market structure on market quality

The descriptive analysis documents a profound change in the organization of the UK equity market. In this section, we apply our model to assess the consequences of it for market quality. To do so, let $\{(Y_{it}, X_{it})\}$ be the data sample at hand, where Y_{it} denotes market quality and X_{it} is a measure of market structure, namely the Herfindahl index or the share of volume traded on lit venues. The effect of X_{it} on Y_{it} for firm i is captured by the individual regression function m_i. The functions μ can be interpreted as the common components of this effect, which for each firm i are weighted differently by the coefficients β_i. The common components are interesting because they measure the degree of heterogeneity that is hidden in the average effect, which is defined as $n^{-1} \sum_{i=1}^{n} m_i(x)$. The fixed effects γ_t and α_i capture the time trends of and cross-sectional exposure to High Frequency Trading, for example. As argued in Gresse (2011) among others, High Frequency Trading affects both the amount of fragmentation as well as the quality of the market outcomes and thus introduces a simultaneity in the data.

To estimate the parameters and functions of interest, we use our methods based on the local linear smoothers $\widehat{m}_i^{LL}$. Prior to estimation, we eliminate stocks with a very small time series dimension, in particular with less than 50 observations. In addition, we exclude stocks whose support of the observations X_{it} is particularly small, specifically

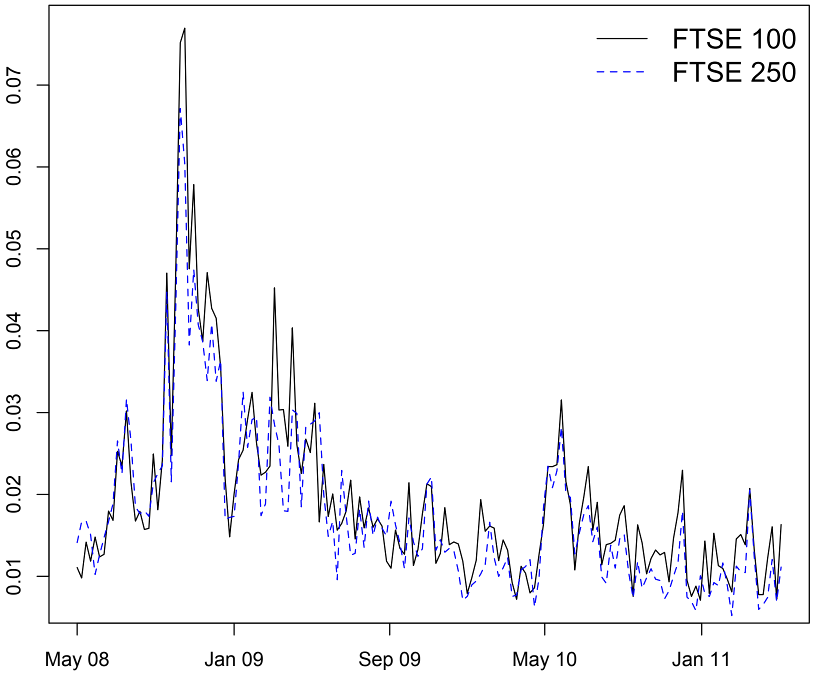

Figure 3: Volatility of the FTSE 100 and FTSE 250 stock indices. Data source: Datastream.

whose support of X_{it} does not span the interquartile range of the pooled distribution. The number of common components K is chosen according to the rule of thumb described in Section 7, where we pick $\delta_{n,T} = 0.05$ and $\overline{K} = 100$. The bandwidth is determined by a plug-in method as discussed in Subsection 5.2. Finally, the weighting matrix W is specified as in equation (9). As a robustness check, we have repeated the estimation for alternative matrices W. The results suggest that our procedure is not very sensitive to the choice of W.

The average effect of market fragmentation and of the volume share traded on lit venues on volatility is shown in Figure 4. We find that volatility is lower when equity venues compete for volume as compared to a monopolistic market, see Panel a) in Figure 4. However, the transition between these extreme forms of market organization is complicated: When new trading venues enter a monopolistic market, volatility first increases until the Herfindahl index reaches a value of 0.4 and then falls. Figure 5 decomposes the average effect into the common components μ_k. We find that the initial increase in volatility when competition increases – or when the value of the Herfindahl index falls – can be attributed to the second component while the decline in volatility at low values of the Herfindahl index is driven by the first component.

In addition to fragmentation of order flow, it is interesting to investigate how the share of volume traded on lit venues affects market quality. Interestingly, we find that volatility is higher if a larger share of volume is traded on lit venues as in Linton (2012), cp. Panel b) in Figure 4. While the average effect is linear, Figure 6 reveals that the second common

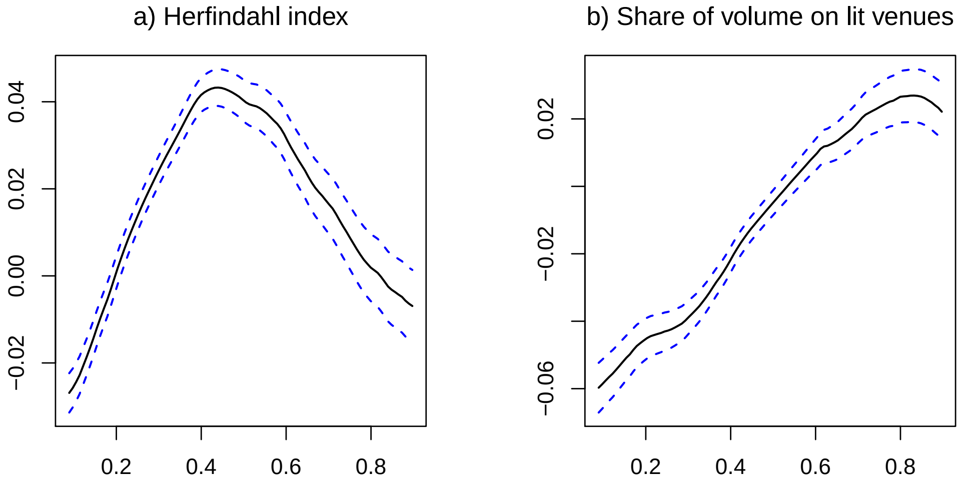

Figure 4: The average effect $n^{-1}\sum_{i=1}^{n} m_i(x)$ of changes in market structure on volatility.

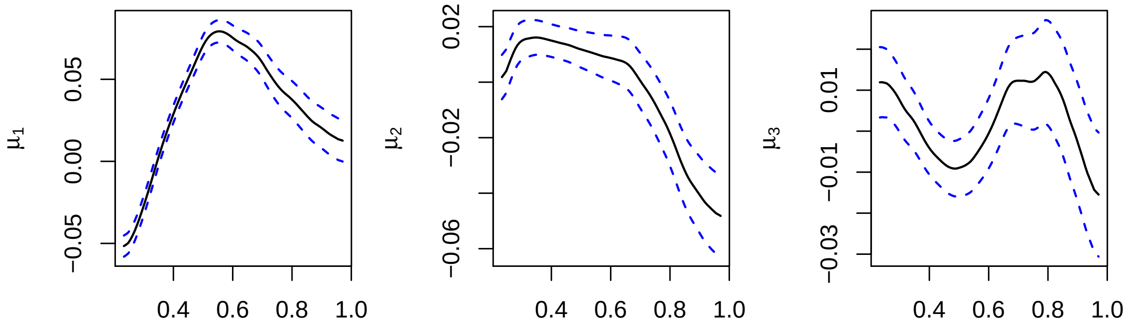

Figure 5: Component functions for the effect of market fragmentation on volatility.

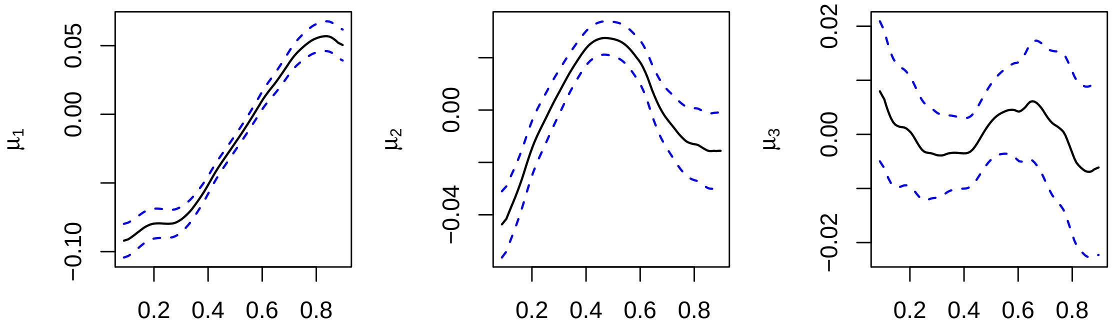

Figure 6: Component functions for the effect of the share of volume traded on lit venues on volatility.

component function has a quadratic shape.

Besides volatility, bid-ask spreads provide a good proxy for market quality. We find that bid-ask spreads are lower in a competitive market in comparison with a monopolistic market. During the transition to a competitive market structure, bid-ask spreads increase initially by a small magnitude before falling rapidly for values of the Herfindahl index below 0.6, see Panel a) in Figure 7. A disaggregation of this effect into its common components is provided in Figure 8.

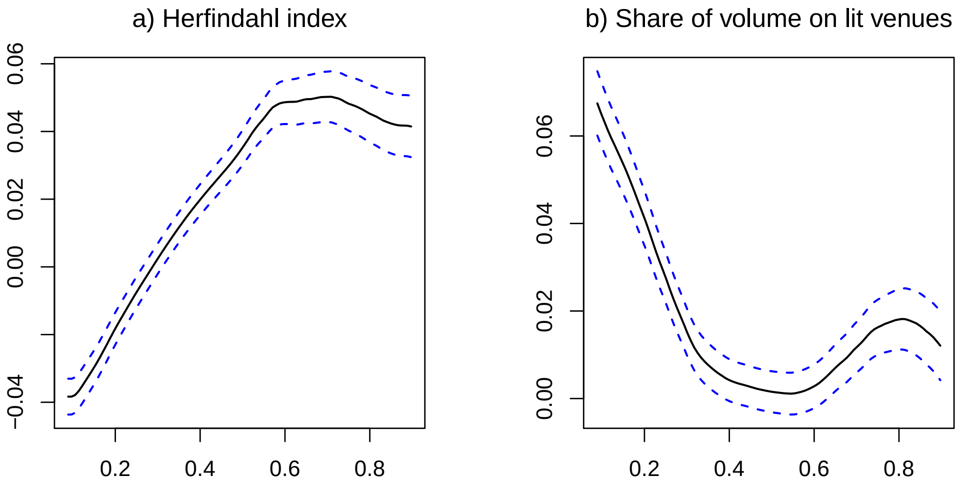

Figure 7: The average effect $n^{-1}\sum_{i=1}^{n} m_i(x)$ of changes in market structure on bid-ask spreads.

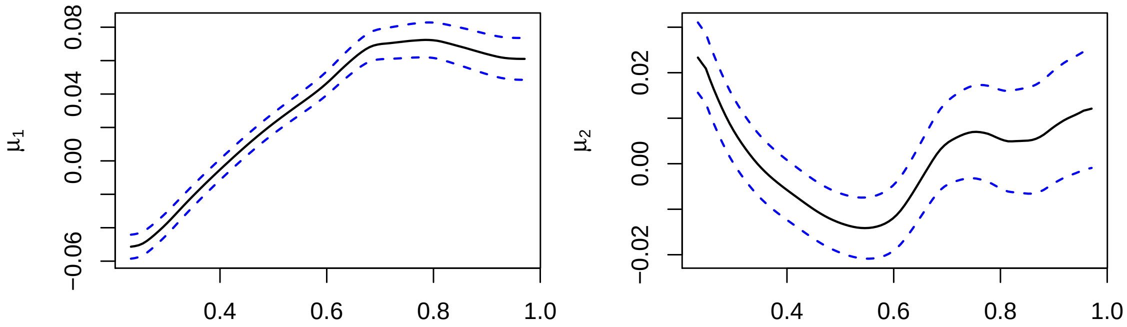

Figure 8: Component functions for the effect of market fragmentation on bid-ask spreads.

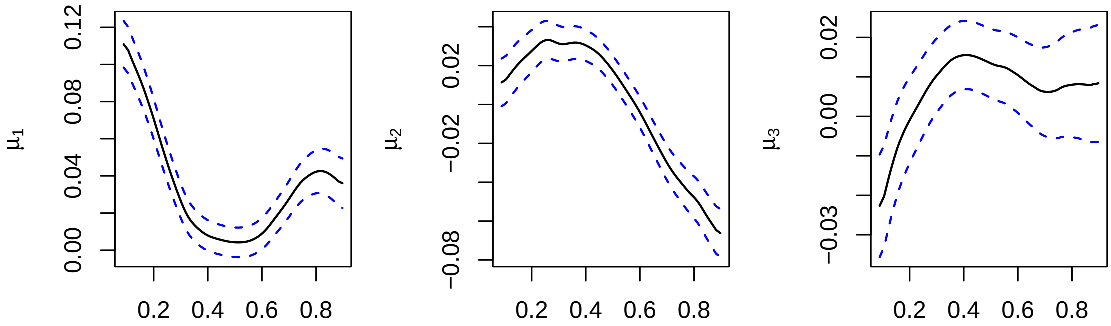

Figure 9: Component functions for the effect of the share of volume traded on lit venues on bid-ask spreads.

As shown in Panel b) in Figure 7, an increase in the share of volume traded at lit venues lowers bid-ask spreads, but not monotonically. As the share of volume traded at lit venues increases, bid-ask spreads fall until 60% of all shares are traded on lit venues and increase thereafter. The decline is primarily driven by the first component function, which can be seen from Figure 9.

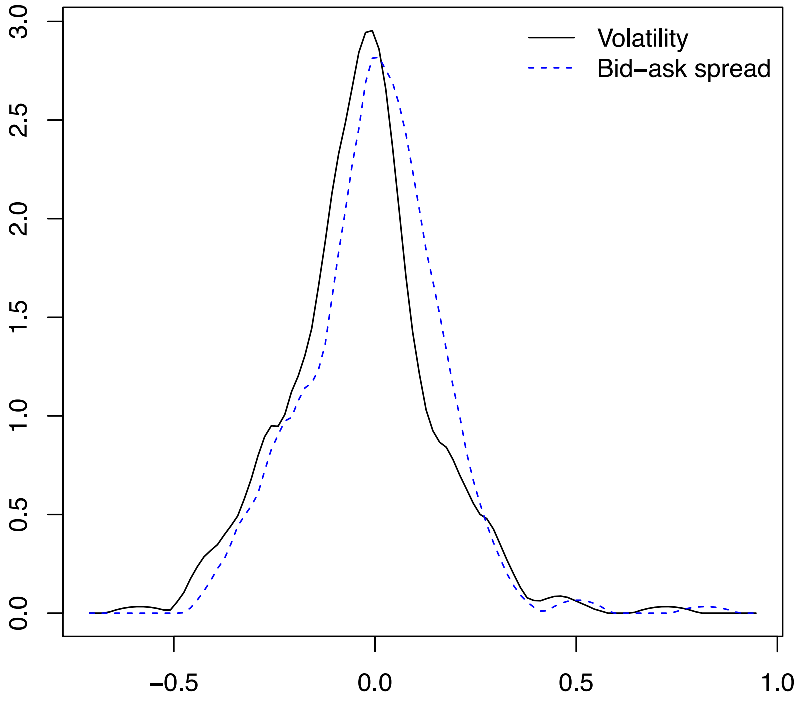

Figure 10: Kernel density estimate of the difference between monopoly and competition $\widehat{c}_i$.

8.3 Is there a difference between monopoly and competition?

One interesting question is whether market quality is significantly different under competition when compared to a monopolistic market. To answer that question, we calculate the statistic $\widehat{c}_i/\sqrt{\widehat{\tau}_i/T}$, where $\widehat{c}_i = \widehat{m}_i(1) - \widehat{m}_i(0)$ measures the difference between monopoly and competition (see Section 5.4). Here, we only consider the Herfindahl index as an independent variable, but we use both measures of market quality, volatility and bid-ask spreads, as a dependent variable. Recall that the Herfindahl index is 1 for a monopolistic market and 0 under perfect competition. To estimate τ_i, one requires an estimate of the long-run variance of the residuals $\widehat{\chi}_{it}$. We estimate the long-run variance by the HAC method with a quadratic spectral kernel (Andrews, 1991) where the bandwidth is chosen optimally. For α-mixing random variables as assumed in this paper, the HAC estimator based on the quadratic spectral kernel with an optimally chosen growth rate of the bandwidth parameter is consistent if the 2 1/2th moment is finite (Hansen, 1992). In our application, $\widehat{c}_i/\sqrt{\widehat{\tau}_i/T}$ is below the critical value even at a significance level of 10% suggesting that there is no statistically significant difference between monopoly and competition (Figure 10).

In addition to a stock-by-stock analysis, we also investigate whether on average, market quality is different under competition when compared to a monopolistic market. In this

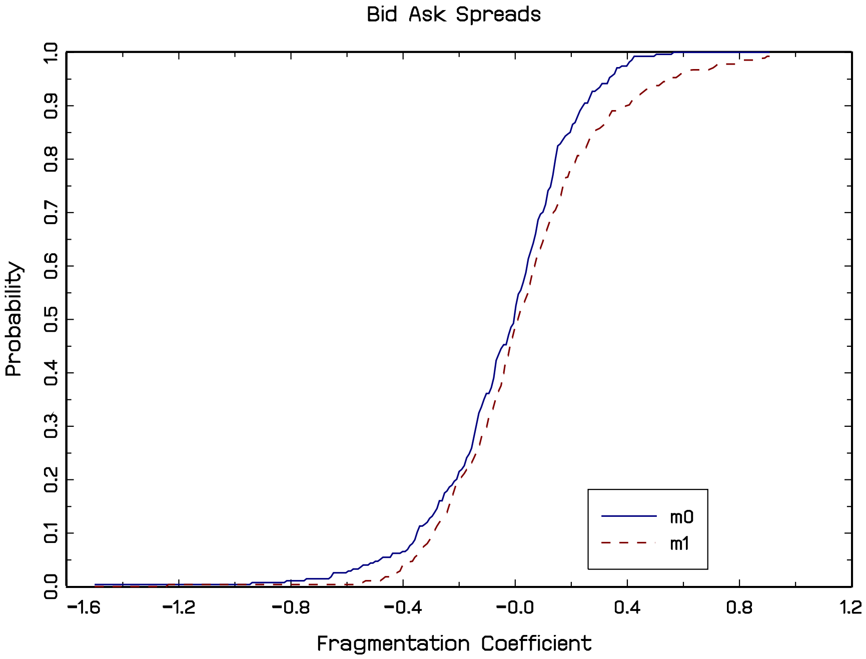

Figure 11: Comparison of cdfs for $m_i(1)$ and $m_i(0)$ coefficients in the bid-ask spread case.

case, the test statistic is given by

$$\widehat{t} = \frac{\sqrt{nTh}[\widehat{m}_{\text{av}}(1) - \widehat{m}_{\text{av}}(0)]}{\sqrt{\widehat{V}(1) + \widehat{V}(0)}},$$

which is asymptotically standard normal. Here, $\widehat{m}_{\text{av}}(x) = n^{-1}\sum_{i=1}^{n} \widehat{m}_i(x)$ is an estimator of the average regression function and $\widehat{V}(x)$ is the sample analogue of $V(x) = \|K\|_2^2 \lim_{n\to\infty}(\frac{1}{n}\sum_{i=1}^{n} \frac{\sigma_i^2(x)}{f_i(x)})$ with $\sigma_i^2(x) = \mathbb{E}[\varepsilon_{it}^2 | X_{it} = x]$. In our data, $\widehat{t}$ is 4.15 if volatility is used as a measure of market quality and 16.09 if market quality is measured by bid-ask spreads. These results are consistent with the findings in Figures 4a) and 7a) but counter to the evidence from individual stocks. When compared to a stock-by-stock analysis, pooling the data increases the power of the test in detecting a difference between competitive and monopolistic markets.

An alternative approach to assess the difference between competitive and monopolistic market structures is to compare the marginal distributions of the coefficients $m_i(1)$ and $m_i(0)$ according to stochastic dominance orderings, cp. Linton, Maasoumi, and Whang (2005). We find that the comparison between cdfs and integrated cdfs for volatility is inconclusive (meaning the two curves cross at least once), whereas the bid-ask spread case is clearer. In particular, the distribution of bid-ask spreads under the monopoly case dominates to first order the distribution under competition, which since bid-ask spread is a bad, means that competition would be preferred to monopoly by any non-satiated utility maximizer. We do not provide a formal test of this hypothesis, since the derivation of appropriate critical values would appear to be a substantial project in itself.

9 Conclusion

Our model captures in a general way two important features in many applications: heterogeneity and nonlinearity. We also allow for a limited type of endogeneity through the unobserved time and cross-section fixed effects. Nevertheless, our estimation procedures are particularly simple, and are in fact closed form at each step. We have provided the tools to conduct inference and to select tuning and order parameters. We applied our method to a question of recent policy interest and our results revealed substantial nonlinearity in the relationship between fragmentation of order flow and market quality, which is not unexpected. Overall, we may find weak evidence that competition between trading venues reduces bid-ask spreads and lowers volatility for traders. Additionally, we find that a higher share of volume traded on lit venues is associated with higher volatility and lower bid-ask spreads. We believe that these results will be of interest for policy makers to evaluate MiFID I and to stimulate further debate on MiFID II.

We close the paper by commenting on some extensions of our model. In our analysis, we have focused on the case of univariate regressors X_{it}. If the regressors are multivariate, the usual curse of dimensionality problem arises, cp. Stone (1980). One way to circumvent this problem is to assume that the regression functions m_i split up into additive components according to

$$m_i(x) = m_i^{(1)}(x_1) + \ldots + m_i^{(d)}(x_d),$$

where d is the dimension of the regressors. Analogously to the univariate case, we may suppose that for each j, the individual functions $m_i^{(j)}$ have the common component structure

$$m_i^{(j)}(x_j) = \sum_{k=1}^{K} \beta_{ik}^{(j)} \mu_k^{(j)}(x_j),$$

where K could also be allowed to differ across j. The additive functions $m_i^{(1)}, \ldots, m_i^{(d)}$ can be estimated by time series backfitting for each individual i, see Mammen et al. (1999). These backfitting estimators may be used as preliminary estimators in our procedure. In particular, the common functions $\mu^{(j)} = (\mu_1^{(j)}, \ldots, \mu_K^{(j)})$ may be estimated separately for each j by repeating the estimation steps of Section 4 based on the backfitting estimators.

Perhaps one is also concerned that we do not allow for sufficiently general time effects, since we have assumed homogeneous such effects. A more general model which allows for additional interactive (exogenous) time effects is given by

$$Y_{it} = \mu_0 + \alpha_i + \gamma_t + g_i(t/T) + m_i(X_{it}) + \epsilon_{it},$$

where $g_i(\cdot)$ is a smooth function of rescaled time. In practice, a number of authors adopt parametric specifications for $g_i(t/T)$ such as $g_i(t/T) = \zeta_i t + \eta_i t^2$, see for example Brogaard et al. (2013). In this case, we obtain

$$Y_{it}^{\text{fe}} = g_i(t/T) + m_i(X_{it}) + \epsilon_{it} + O_p(T^{-1/2}) + O_p(n^{-1/2}),$$

where we have assumed that $\sum_{t=1}^{T} g_i(t/T) = 0$. Similarly to the multivariate case discussed above, we here have an additive regression model that could be estimated by time series backfitting. Moreover, one could restrict $g_i(\cdot)$ to rely on a small number of principal components as we do for $m_i(\cdot)$, and do parallel analysis for both functions.

Supplementary Material

In the supplementary material, we investigate the small sample performance of our estimation procedures in a series of simulation experiments. Moreover, we provide the technical details which are omitted in the appendix.

Appendix A

In this appendix, we derive the main results of our theory. In particular, we provide a detailed proof of Theorems 5.1 and 5.2, which characterize the asymptotic behaviour of our estimators. For the proof, we require a series of uniform convergence results which are derived in Appendix B. Throughout the appendix, the symbol C is used to denote a universal real constant which may take a different value on each occurrence. Moreover, we let $I_h = [C_1 h, 1 - C_1 h]$ denote the interior of the support of the regressors X_{it} and use $I_h^c = [0,1] \setminus I_h$ to denote the boundary region. Finally, we frequently make use of the shorthand $\kappa_0(x) = \int_{-x/h}^{(1-x)/h} K(\varphi) d\varphi$.

Proof of Theorem 5.1

We restrict attention to the proof for the Nadaraya-Watson based estimators. The local linear case can be handled by similar arguments.

To start with, we list some auxiliary results needed to derive the statements (16) and (17) of Theorem 5.1. The proof of these results is postponed until the arguments for Theorem 5.1 are completed. The following uniform expansion of $\widehat{g}_k(x) - g_k(x)$ forms the basis of our arguments.

Proposition A1. *It holds that*

$$\widehat{g}_k(x) - g_k(x) = \sum_{i=1}^{n} \frac{\omega_{ki}}{\kappa_0(x) f_i(x)} \frac{1}{T} \sum_{t=1}^{T} K_h(X_{it} - x)\varepsilon_{it} + R_k(x), \tag{22}$$

where the remainder satisfies $\sup_{x \in I_h} |R_k(x)| = o_p(1/\sqrt{nTh})$ *and* $\sup_{x \in I_h^c} |R_k(x)| = O_p(h)$.

Using the uniform expansion of Proposition A1, we are able to derive the asymptotic properties of $\widehat{g}$. These are summarized in the next proposition.

Proposition A2. *It holds that*

$$\sup_{x \in I_h} \big\| \widehat{g}(x) - g(x) \big\| = O_p\Big(\sqrt{\frac{\log nT}{nTh}} \Big) \tag{23}$$

$$\sup_{x \in I_h^c} \big\| \widehat{g}(x) - g(x) \big\| = O_p(h). \tag{24}$$

Moreover, for any fixed $x \in (0,1)$,

$$\sqrt{nTh}(\widehat{g}(x) - g(x)) \xrightarrow{d} N(0, V(x)), \tag{25}$$

where $V(x) = (V_{k,l}(x))_{k,l=1,\ldots,K}$ and $V_{k,l}(x) = \|K\|_2^2 \lim_{n\to\infty}(n \sum_{i=1}^n \omega_{ki}\omega_{li} \frac{\sigma_i^2(x)}{f_i(x)})$ with $\sigma_i^2(x)$ $= \mathbb{E}[\varepsilon_{it}^2 | X_{it} = x]$.

Proposition A1 can further be used to characterize the convergence behaviour of the matrices $\widehat{\Sigma}$.

Proposition A3. *It holds that*

$$\|\widehat{\Sigma} - \Sigma\| = o_p\Big(\frac{1}{\sqrt{nTh}} \Big). \tag{26}$$

Finally, Proposition A3 together with a Taylor expansion argument yields the following result.

Proposition A4. *It holds that*

$$\|\widehat{S} - S\| = o_p\Big(\frac{1}{\sqrt{nTh}} \Big) \tag{27}$$

$$\|\widehat{\lambda} - \lambda\| = o_p\Big(\frac{1}{\sqrt{nTh}} \Big) \tag{28}$$

with $\lambda = (\lambda_1, \ldots, \lambda_K)^\intercal$ and $\widehat{\lambda} = (\widehat{\lambda}_1, \ldots, \widehat{\lambda}_K)^\intercal$.

With the help of the above propositions, it is now straightforward to prove the statements (16) and (17) of Theorem 5.1. We start with the proof of (16): Recalling that the matrix of eigenvectors S converges to a limit S^* and using (23) together with (27), we arrive at

$$\begin{aligned} \sup_{x \in I_h} \|\widehat{\mu}(x) - \mu(x)\| &\leq \|\widehat{S}^\intercal - S^\intercal\| \sup_{x \in I_h} \|\widehat{g}(x)\| \\ &\quad + \|S^\intercal\| \sup_{x \in I_h} \|\widehat{g}(x) - g(x)\| = O_p\Big(\sqrt{\frac{\log nT}{nTh}} \Big). \end{aligned}$$

Similarly, we obtain that

$$\begin{aligned}\sqrt{nTh}(\widehat{\mu}(x)-\mu(x)) &= \sqrt{nTh}(\widehat{S}^\intercal - S^\intercal)\widehat{g}(x) + S^\intercal\sqrt{nTh}(\widehat{g}(x)-g(x)) \\ &= S^\intercal\sqrt{nTh}(\widehat{g}(x)-g(x)) + o_p(1).\end{aligned}$$

Since S converges to S^*, the normality result (25) implies that

$$S^\intercal\sqrt{nTh}(\widehat{g}(x)-g(x)) \xrightarrow{d} N(0,(S^*)^\intercal V(x)S^*),$$

which yields (17). □

Proof of Proposition A1

Let $\widehat{f}_i(x) = T^{-1}\sum_{t=1}^T K_h(X_{it}-x)$, $Y_{it}^{\mathrm{fe}} = Y_{it} - \overline{Y}_i - \overline{Y}_t + \overline{\overline{Y}}$ and write

$$\widehat{g}_k(x) - g_k(x) = Q_{k,V}(x) + Q_{k,B}(x) + Q_{k,\gamma}(x) + Q_{k,\alpha} + Q_{k,\mu_0},$$

where

$$\begin{aligned}Q_{k,V}(x) &= \sum_{i=1}^n \omega_{ki}\frac{1}{T}\sum_{t=1}^T K_h(X_{it}-x)\varepsilon_{it}/\widehat{f}_i(x) \\ Q_{k,B}(x) &= \sum_{i=1}^n \omega_{ki}\frac{1}{T}\sum_{t=1}^T K_h(X_{it}-x)\big\{m_i(X_{it})-m_i(x)\big\}/\widehat{f}_i(x) \\ Q_{k,\gamma}(x) &= \sum_{i=1}^n \omega_{ki}\frac{1}{T}\sum_{t=1}^T K_h(X_{it}-x)\big\{\mu_0+\gamma_t-\overline{Y}_t\big\}/\widehat{f}_i(x) \\ Q_{k,\alpha} &= \sum_{i=1}^n \omega_{ki}\big\{\mu_0+\alpha_i-\overline{Y}_i\big\} \\ Q_{k,\mu_0} &= \Big(\sum_{i=1}^n \omega_{ki}\Big)\big\{\overline{\overline{Y}}-\mu_0\big\}.\end{aligned}$$

In what follows, we analyze these five terms one after the other.

(i) It holds that

$$Q_{k,V}(x) = \sum_{i=1}^n \omega_{ki}\frac{1}{T}\sum_{t=1}^T K_h(X_{it}-x)\varepsilon_{it}/\kappa_0(x)f_i(x) + R_{k,V}(x),$$

where the remainder term is given by

$$R_{k,V}(x) = \sum_{m=1}^M R_{k,V}^{(m)}(x) + R_{k,V}^{(M+1)}(x) + R_{k,V}^{(B)}(x)$$

with

$$R_{k,V}^{(m)}(x) = \sum_{i=1}^n \omega_{ki} \Big(\frac{(\mathbb{E}[\widehat{f}_i(x)] - \widehat{f}_i(x))^m}{\mathbb{E}[\widehat{f}_i(x)]^{m+1}} \Big) \Big(\frac{1}{T} \sum_{t=1}^T K_h(X_{it} - x) \varepsilon_{it} \Big)$$

for $m = 1, \ldots, M$,

$$R_{k,V}^{(M+1)}(x) = \sum_{i=1}^n \omega_{ki} \Big(\frac{(\mathbb{E}[\widehat{f}_i(x)] - \widehat{f}_i(x))^{M+1}}{\mathbb{E}[\widehat{f}_i(x)]^{M+1} \widehat{f}_i(x)} \Big) \Big(\frac{1}{T} \sum_{t=1}^T K_h(X_{it} - x) \varepsilon_{it} \Big)$$

and

$$R_{k,V}^{(B)}(x) = \sum_{i=1}^n \omega_{ki} \Big(\frac{\kappa_0(x) f_i(x) - \mathbb{E}[\widehat{f}_i(x)]}{\kappa_0(x) f_i(x) \mathbb{E}[\widehat{f}_i(x)]} \Big) \Big(\frac{1}{T} \sum_{t=1}^T K_h(X_{it} - x) \varepsilon_{it} \Big).$$

The remainder term has the property that

$$\sup_{x \in I_h} \big| R_{k,V}(x) \big| = o_p\Big(\frac{1}{\sqrt{nTh}} \Big) \tag{29}$$

$$\sup_{x \in I_h^c} \big| R_{k,V}(x) \big| = O_p(h). \tag{30}$$

We first derive (29): To start with, straightforward calculations yield that $\max_{1 \le i \le n} \sup_{x \in I_h} |\kappa_0(x) f_i(x) - \mathbb{E}[\widehat{f}_i(x)]| = O_p(h^2)$. Together with Lemma B1 in Appendix B, this directly implies that $\sup_{x \in I_h} |R_{k,V}^{(B)}(x)| = o_p(1/\sqrt{nTh})$. Moreover, by Lemma B3, it holds that $\sup_{x \in I_h} |R_{k,V}^{(m)}(x)| = o_p(1/\sqrt{nTh})$ for $m = 1, \ldots, M$. Finally, if M is chosen sufficiently large, then an application of Lemma B1 immediately shows that $\sup_{x \in I_h} |R_{k,V}^{(M+1)}(x)| = o_p(1/\sqrt{nTh})$ as well. (30) follows by analogous arguments.

(ii) We next show that

$$\sup_{x \in I_h} |Q_{k,B}(x)| = o_p\Big(\frac{1}{\sqrt{nTh}} \Big)$$

$$\sup_{x \in I_h^c} |Q_{k,B}(x)| = O_p(h).$$

To see this, decompose $Q_{k,B}(x)$ into the following two components:

$$Q_{k,B}(x) = Q_{k,B}^{(1)}(x) + Q_{k,B}^{(2)}(x)$$

with

$$Q_{k,B}^{(1)}(x) = \sum_{i=1}^n \omega_{ki} \frac{1}{T} \sum_{t=1}^T \Big(K_h(X_{it} - x) \big\{ m_i(X_{it}) - m_i(x) \big\} \\ - \mathbb{E}\big[K_h(X_{it} - x) \big\{ m_i(X_{it}) - m_i(x) \big\} \big] \Big) \big/ \widehat{f}_i(x)$$

$$Q_{k,B}^{(2)}(x) = \sum_{i=1}^n \omega_{ki} \frac{1}{T} \sum_{t=1}^T \mathbb{E}\big[K_h(X_{it} - x) \big\{ m_i(X_{it}) - m_i(x) \big\} \big] \big/ \widehat{f}_i(x).$$

Exploiting the smoothness conditions on the functions m_i and f_i in a standard way, the term $Q^{(2)}_{k,B}(x)$ can be shown to satisfy $\sup_{x \in I_h} |Q^{(2)}_{k,B}(x)| = O_p(h^2) = o_p(1/\sqrt{nTh})$ and $\sup_{x \in I_h^c} |Q^{(2)}_{k,B}(x)| = O_p(h)$. Moreover, $Q^{(1)}_{k,B}(x) = Q^{(1,a)}_{k,B}(x) + Q^{(1,b)}_{k,B}(x)$ with

$$
\begin{aligned}
Q^{(1,a)}_{k,B}(x) = \sum_{i=1}^n \omega_{ki} \frac{1}{T} \sum_{t=1}^T \Big(& K_h(X_{it} - x)\big\{m_i(X_{it}) - m_i(x)\big\} \\
& - \mathbb{E}\big[K_h(X_{it} - x)\big\{m_i(X_{it}) - m_i(x)\big\}\big]\Big) / \kappa_0(x) f_i(x) \\
Q^{(1,b)}_{k,B}(x) = \sum_{i=1}^n \omega_{ki} \Big(\frac{\kappa_0(x) f_i(x) - \widehat{f}_i(x)}{\kappa_0(x) f_i(x) \widehat{f}_i(x)}\Big) \frac{1}{T} \sum_{t=1}^T \Big(& K_h(X_{it} - x)\big\{m_i(X_{it}) - m_i(x)\big\} \\
& - \mathbb{E}\big[K_h(X_{it} - x)\big\{m_i(X_{it}) - m_i(x)\big\}\big]\Big).
\end{aligned}
$$

Using the proving strategy of Lemma B2, the term $Q^{(1,a)}_{k,B}(x)$ can be shown to be of the order $O_p(h\sqrt{\log nT / nTh}) = o_p(1/\sqrt{nTh})$ uniformly for $x \in [0,1]$. Moreover, applying Lemma B1, it is straightforward to see that $\sup_{x \in [0,1]} |Q^{(1,b)}_{k,B}(x)| = o_p(1/\sqrt{nTh})$ as well.

(iii) We now turn to the analysis of $Q_{k,\gamma}(x)$. In particular, we show that

$$
\sup_{x \in [0,1]} |Q_{k,\gamma}(x)| = o_p\Big(\frac{1}{\sqrt{nTh}}\Big).
$$

To do so, first note that

$$
Q_{k,\gamma}(x) = -\sum_{i=1}^n \omega_{ki} \frac{1}{T} \sum_{t=1}^T K_h(X_{it} - x)\Big\{\frac{1}{n} \sum_{j=1}^n (m_j(X_{jt}) + \varepsilon_{jt})\Big\} / \widehat{f}_i(x).
$$

This expression can be decomposed as follows: $Q_{k,\gamma}(x) = Q^{(1)}_{k,\gamma}(x) + Q^{(2)}_{k,\gamma}(x) + Q^{(3)}_{k,\gamma}(x)$, where

$$
\begin{aligned}
Q^{(1)}_{k,\gamma}(x) &= -\sum_{i=1}^n \omega_{ki} \frac{1}{T} \sum_{t=1}^T K_h(X_{it} - x)\Big\{\frac{1}{n} \sum_{j=1}^n (m_j(X_{jt}) + \varepsilon_{jt})\Big\} / \kappa_0(x) f_i(x) \\
Q^{(2)}_{k,\gamma}(x) &= -\sum_{i=1}^n \omega_{ki} \frac{1}{T} \sum_{t=1}^T K_h(X_{it} - x)\Big\{\frac{1}{n} (m_i(X_{it}) + \varepsilon_{it})\Big\} \Big(\frac{1}{\widehat{f}_i(x)} - \frac{1}{\kappa_0(x) f_i(x)}\Big) \\
Q^{(3)}_{k,\gamma}(x) &= -\sum_{i=1}^n \omega_{ki} \frac{1}{T} \sum_{t=1}^T K_h(X_{it} - x)\Big\{\frac{1}{n} \sum_{\substack{j=1 \\ j \neq i}}^n (m_j(X_{jt}) + \varepsilon_{jt})\Big\} \Big(\frac{1}{\widehat{f}_i(x)} - \frac{1}{\kappa_0(x) f_i(x)}\Big).
\end{aligned}
$$

To analyze the term $Q^{(1)}_{k,\gamma}(x)$, we further split it up into two components: $Q^{(1)}_{k,\gamma}(x) =$

$Q_{k,\gamma}^{(1,a)}(x) + Q_{k,\gamma}^{(1,b)}(x)$, where

$$Q_{k,\gamma}^{(1,a)}(x) = -\frac{1}{T}\sum_{t=1}^{T}\Big(\sum_{i=1}^{n}\frac{\omega_{ki}}{\kappa_0(x)f_i(x)}(K_h(X_{it}-x)-\mathbb{E}[K_h(X_{it}-x)])\Big) \times \Big\{\frac{1}{n}\sum_{j=1}^{n}(m_j(X_{jt})+\varepsilon_{jt})\Big\}$$

$$Q_{k,\gamma}^{(1,b)}(x) = -\sum_{i=1}^{n}\frac{\omega_{ki}}{\kappa_0(x)f_i(x)}\Big(\frac{1}{nT}\sum_{j=1}^{n}\sum_{t=1}^{T}\mathbb{E}[K_h(X_{it}-x)](m_j(X_{jt})+\varepsilon_{jt})\Big).$$

The term $Q_{k,\gamma}^{(1,a)}(x)$ can be handled by similar techniques as applied in Lemma B3. The details are summarized in Lemma B4 which yields that $\sup_{x\in[0,1]}|Q_{k,\gamma}^{(1,a)}(x)| = o_p(1/\sqrt{nTh})$. Moreover, it is straightforward to verify that $\sup_{x\in[0,1]}|Q_{k,\gamma}^{(1,b)}(x)| = O_p(1/\sqrt{nT})$. Turning to the expression $Q_{k,\gamma}^{(2)}(x)$, we can easily see with the help of Lemma B1 that $\sup_{x\in[0,1]}|Q_{k,\gamma}^{(2)}(x)| = o_p(1/\sqrt{nTh})$. To prove that $\sup_{x\in[0,1]}|Q_{k,\gamma}^{(3)}(x)| = o_p(1/\sqrt{nTh})$, some rather involved arguments are needed which are presented in Lemma B5. Setting $\widehat{\phi}_i(x) = (\widehat{f}_i(x))^{-1} - (\kappa_0(x)f_i(x))^{-1}$ in this lemma yields the result.

Finally, it is trivial to see that $Q_{k,\alpha} = O_p(1/\sqrt{nT})$ as well as $Q_{k,\mu_0} = O_p(1/\sqrt{nT})$. Together with (i)–(iii), this yields the expansion (22). □

Proof of Proposition A2

The proof easily follows with the help of the uniform expansion from Proposition A1. The latter says that

$$\widehat{g}_k(x) - g_k(x) = W_{k,V}(x) + R_k(x),$$

where

$$W_{k,V}(x) = \sum_{i=1}^{n}\omega_{ki}\frac{1}{T}\sum_{t=1}^{T}K_h(X_{it}-x)\varepsilon_{it}/\kappa_0(x)f_i(x)$$

and the remainder term $R_k(x)$ satisfies $\sup_{x\in I_h}|R_k(x)| = o_p(1/\sqrt{nTh})$ as well as $\sup_{x\in I_h^c}|R_k(x)| = O_p(h)$. Applying Lemma B2 to $W_{k,V}(x)$, we immediately obtain that $\sup_{x\in[0,1]}|W_{k,V}(x)| = O_p(\sqrt{\log nT/nTh})$. This yields the uniform convergence results (23) and (24). Furthermore, standard arguments show that

$$\sqrt{nTh}W_{k,V}(x) \xrightarrow{d} N\Big(0, \|K\|_2^2 \lim_{n\to\infty} n\sum_{i=1}^{n}\omega_{ki}^2\frac{\sigma_i^2(x)}{f_i(x)}\Big).$$

From this, the normality result (25) easily follows. □

Proof of Proposition A3

It holds that

$$\begin{aligned}\widehat{\Sigma}_{kl} - \Sigma_{kl} &= \int \widehat{g}_k(x)\widehat{g}_l(x)w(x)dx - \int g_k(x)g_l(x)w(x)dx \\ &= \int \big[\widehat{g}_k(x) - g_k(x)\big]\widehat{g}_l(x)w(x)dx + \int g_k(x)\big[\widehat{g}_l(x) - g_l(x)\big]w(x)dx \\ &= \int \big[\widehat{g}_k(x) - g_k(x)\big]g_l(x)w(x)dx + \int g_k(x)\big[\widehat{g}_l(x) - g_l(x)\big]w(x)dx \\ &\quad + o_p\Big(\frac{1}{\sqrt{nTh}}\Big),\end{aligned}$$

where the last equality follows by Proposition A2. Using the uniform expansion of Proposition A1, we obtain

$$\int \big[\widehat{g}_k(x) - g_k(x)\big]g_l(x)w(x)dx = J_V + R$$

with

$$J_V = \sum_{i=1}^{n} \omega_{ki}\frac{1}{T}\sum_{t=1}^{T}\Big(\int K_h(X_{it}-x)g_l(x)(\kappa_0(x)f_i(x))^{-1}w(x)dx\Big)\varepsilon_{it}$$

and $R = \int g_l(x)R_k(x)w(x)dx$. As $\sup_{x\in I_h}|R_k(x)| = o_p(1/\sqrt{nTh})$ and $\sup_{x\in I_h^c}|R_k(x)| = O_p(h)$, we have that $R = o_p(1/\sqrt{nTh})$. Moreover, applying Chebychev's inequality and exploiting the mixing conditions on the data with the help of Davydov's inequality (see Corollary 1.1 in Bosq (1998)), it is not difficult to see that $J_V = o_p(1/\sqrt{nTh})$. This completes the proof. □

Proof of Proposition A4

Let $v(A) = \text{vec}(A)$ be the vectorized representation of a $K \times K$ matrix A. There are fixed vector-valued functions $f_k(\cdot)$ and scalar functions $\psi_k(\cdot)$ with first and second derivatives existing and being continuous in a neighbourhood of $v(\Sigma^*)$ such that

$$\begin{aligned} s_k &= f_k(v(\Sigma)) \quad \text{and} \quad \lambda_k = \psi_k(v(\Sigma)) \\ \widehat{s}_k &= f_k(v(\widehat{\Sigma})) \quad \text{and} \quad \widehat{\lambda}_k = \psi_k(v(\widehat{\Sigma}))\end{aligned}$$

(cp. Magnus (1985)). In what follows, we show that $\|\widehat{s}_k - s_k\| = o_p(1/\sqrt{nTh})$ for all $k = 1, \ldots, K$, which immediately yields (27). The result (28) for the estimates of the eigenvalues follows by exactly the same argument. From Proposition A3, we know that

$$\|v(\widehat{\Sigma}) - v(\Sigma)\| = o_p\Big(\frac{1}{\sqrt{nTh}}\Big).$$

As f_k is continuously differentiable in a neighbourhood of $v(\Sigma^*)$, a first-order Taylor expansion yields

$$\widehat{s}_k - s_k = f_k(v(\widehat{\Sigma})) - f_k(v(\Sigma)) = f_k'(\xi)\big[v(\widehat{\Sigma}) - v(\Sigma)\big]$$

with ξ being an intermediate point between $v(\widehat{\Sigma})$ and $v(\Sigma)$. Since $f_k'(\xi) - f_k'(v(\Sigma^*)) = o_p(1)$, we immediately arrive at

$$\|\widehat{s}_k - s_k\| = o_p\Big(\frac{1}{\sqrt{nTh}}\Big). \qquad \square$$

Proof of Theorem 5.2

We again restrict attention to the Nadaraya-Watson based case, the arguments for the local linear case being essentially the same. Write

$$\sqrt{T}(\widehat{\beta}_i - \beta_i) = \sqrt{T}(\widehat{\beta}_i - \widetilde{\beta}_i) + \sqrt{T}(\widetilde{\beta}_i - \beta_i),$$

where $\widetilde{\beta}_i$ is the infeasible parameter estimator defined in (11). In what follows, we analyze the two terms on the right-hand side separately.

(i) First consider the term $\sqrt{T}(\widehat{\beta}_i - \widetilde{\beta}_i)$. It holds that

$$\begin{aligned}
&\sqrt{T}(\widehat{\beta}_i - \widetilde{\beta}_i) \\
&= \Big(\frac{1}{T}\sum_{t=1}^{T} \pi(X_{it})\widehat{\mu}(X_{it})\widehat{\mu}(X_{it})^\intercal\Big)^{-1} \frac{1}{\sqrt{T}}\sum_{t=1}^{T} \pi(X_{it})\big\{\widehat{\mu}(X_{it}) - \mu(X_{it})\big\} Y_{it}^{\text{fe}} \\
&\quad + \Big\{\Big(\frac{1}{T}\sum_{t=1}^{T} \pi(X_{it})\widehat{\mu}(X_{it})\widehat{\mu}(X_{it})^\intercal\Big)^{-1} - \Big(\frac{1}{T}\sum_{t=1}^{T} \pi(X_{it})\mu(X_{it})\mu(X_{it})^\intercal\Big)^{-1}\Big\} \\
&\qquad \times \frac{1}{\sqrt{T}}\sum_{t=1}^{T} \pi(X_{it})\mu(X_{it}) Y_{it}^{\text{fe}}.
\end{aligned}$$

Here,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \pi(X_{it})\mu(X_{it}) Y_{it}^{\text{fe}} = L_1 + L_2 + L_3 + L_4$$

with

$$\begin{aligned}
L_1 &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T} \pi(X_{it})\mu(X_{it})\varepsilon_{it} \\
L_2 &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T} \pi(X_{it})\mu(X_{it}) m_i(X_{it}) \\
L_3 &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T} \pi(X_{it})\mu(X_{it})\big(\mu_0 + \gamma_t - \overline{Y}_t\big)
\end{aligned}$$

$$L_4 = \Big(\frac{1}{T}\sum_{t=1}^{T}\pi(X_{it})\mu(X_{it})\Big)\sqrt{T}\big(\alpha_i - \overline{Y}_i + \overline{\overline{Y}}\big).$$

It is straightforward to see that $L_1 = O_p(1)$, $L_2 = O_p(\sqrt{T})$, $L_3 = o_p(1)$ and $L_4 = O_p(1)$. Hence,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\pi(X_{it})\mu(X_{it})Y_{it}^{\text{fe}} = O_p(\sqrt{T}). \tag{31}$$

As $\sup_{x\in I_h}\|\widehat{\mu}(x) - \mu(x)\| = O_p(\sqrt{\log nT/nTh}) = o_p(1/\sqrt{T})$, we further obtain that

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\pi(X_{it})\widehat{\mu}(X_{it})\widehat{\mu}(X_{it})^\intercal - \frac{1}{T}\sum_{t=1}^{T}\pi(X_{it})\mu(X_{it})\mu(X_{it})^\intercal& \\ = O_p\Big(\sqrt{\frac{\log nT}{nTh}}\Big) = o_p\Big(\frac{1}{\sqrt{T}}\Big)&\end{aligned} \tag{32}$$

as well as

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\pi(X_{it})\big\{\widehat{\mu}(X_{it}) - \mu(X_{it})\big\}Y_{it}^{\text{fe}} = o_p(1). \tag{33}$$

Combining (31)–(33) yields $\sqrt{T}(\widehat{\beta}_i - \widetilde{\beta}_i) = o_p(1)$.

(ii) We next turn to $\sqrt{T}(\widetilde{\beta}_i - \beta_i)$. Write

$$\sqrt{T}(\widetilde{\beta}_i - \beta_i) = \Big(\frac{1}{T}\sum_{t=1}^{T}\pi(X_{it})\mu(X_{it})\mu(X_{it})^\intercal\Big)^{-1}(L_1 + L_3 + L_4)$$

with L_1, L_3 and L_4 introduced above. Since $L_3 = o_p(1)$ and $T^{-1}\sum_{t=1}^{T}\pi(X_{it})\mu(X_{it}) \xrightarrow{P} \mathbb{E}[\pi(X_{it})\mu(X_{it})]$, we can rewrite L_4 as

$$L_4 = -\mathbb{E}[\pi(X_{it})\mu(X_{it})]\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(m_i(X_{it}) + \varepsilon_{it}) + o_p(1).$$

This yields that

$$L_1 + L_3 + L_4 = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\chi_{it} + o_p(1),$$

where $\chi_{it} = (\pi(X_{it})\mu(X_{it}) - \mathbb{E}[\pi(X_{it})\mu(X_{it})])\varepsilon_{it} - \mathbb{E}[\pi(X_{it})\mu(X_{it})]m_i(X_{it})$. Applying a central limit theorem, we now arrive at

$$\sqrt{T}(\widetilde{\beta}_i - \beta_i) \xrightarrow{d} N(0, \Gamma_i^{-1}\Psi_i(\Gamma_i^{-1})^\intercal),$$

where the matrices Γ_i and Ψ_i are given by $\Gamma_i = \mathbb{E}[\pi(X_{it})\mu(X_{it})\mu(X_{it})^\intercal]$ and $\Psi_i = \sum_{l=-\infty}^{\infty}\text{Cov}(\chi_{i0}, \chi_{il})$. □

Proof of Theorem 6.2

The same arguments as for the proof of Proposition A3 show that

$$\|\widetilde{\Sigma} - \overline{\Sigma}\| = o_p\Big(\frac{1}{\sqrt{nTh}}\Big).$$

Moreover, letting $\overline{\lambda}_1 \ge \ldots \ge \overline{\lambda}_{\overline{K}}$ be the eigenvalues of the matrix $\overline{\Sigma}$ and $\widetilde{\lambda}_1 \ge \ldots \ge \widetilde{\lambda}_{\overline{K}}$ the eigenvalues of $\widetilde{\Sigma}$, we have that

$$\widetilde{\lambda}_k = \int \widetilde{\mu}_k^2(x) w(x) dx$$

and $\overline{\lambda}_k = 0$ for $k = K+1, \ldots, \overline{K}$. Finally, note that the mapping of symmetric matrices to their eigenvalues is Lipschitz continuous. In particular, let A and B be any real symmetric $\overline{K} \times \overline{K}$ matrices and let $\lambda_1(A) \ge \lambda_2(A) \ge \ldots \ge \lambda_{\overline{K}}(A)$ and $\lambda_1(B) \ge \lambda_2(B) \ge \ldots \ge \lambda_{\overline{K}}(B)$ be the corresponding eigenvalues. Then there exists a constant L independent of A and B such that

$$|\lambda_k(A) - \lambda_k(B)| \le L\|A - B\|.$$

Combining the above remarks, we arrive at

$$\int \widetilde{\mu}_k^2(x) w(x) dx = \widetilde{\lambda}_k = |\widetilde{\lambda}_k - \overline{\lambda}_k| \le L\|\widetilde{\Sigma} - \overline{\Sigma}\| = o_p\Big(\frac{1}{\sqrt{nTh}}\Big).$$

for all $k = K+1, \ldots, \overline{K}$. □

Appendix B

In this appendix, we list some results on uniform convergence which are needed to derive the main theorems. As the proofs are rather lengthy and involved, they are deferred to the Supplementary Material. We formulate the results for a general array $\{(X_{it}, Z_{it})\} = \{(X_{it}, Z_{it}),\ i = 1, \ldots, n,\ t = 1, \ldots, T\}$ which satisfies the following conditions:

(A1') The data $\{(X_{it}, Z_{it})\}$ are independent across i. Moreover, they are strictly stationary and strongly mixing in the time direction. Let $\alpha_i(k)$ for $k = 1, 2, \ldots$ be the mixing coefficients of the time series $\{(X_{it}, Z_{it}), t = 1, \ldots, T\}$ of the i-th individual. It holds that $\alpha_i(k) \le \alpha(k)$ for all $i = 1, \ldots, n$, where the coefficients $\alpha(k)$ decay exponentially fast to zero as $k \to \infty$.

(A4') For some $\theta > 5$ and for all $l \in \mathbb{Z}$,

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

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

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

where C is a sufficiently large constant independent of l.

In addition, we suppose that the variables X_{it} and (X_{it}, X_{it+l}) have densities f_i and $f_{i;l}$ which satisfy (A2) and that the kernel K and the dimensions n and T fulfill (A5)–(A7).

Throughout the appendix, we assume that the above conditions are satisfied. We now formulate the various results:

Lemma B1. *For kernel averages $\Psi_i(x)$ of the form*

$$\Psi_i(x) = \frac{1}{T} \sum_{t=1}^{T} K_h(X_{it} - x) Z_{it},$$

it holds that

$$\max_{1 \le i \le n} \sup_{x \in [0,1]} \big|\Psi_i(x) - \mathbb{E}[\Psi_i(x)]\big| = o_p(1). \tag{34}$$

If the variables Z_{it} are bounded, i.e., if $|Z_{it}| \le C$ for some constant C independent of i and t, then we even have that

$$\max_{1 \le i \le n} \sup_{x \in [0,1]} \big|\Psi_i(x) - \mathbb{E}[\Psi_i(x)]\big| = O_p\Big(\sqrt{\frac{\log T}{Th}}\Big). \tag{35}$$

Lemma B2. *Let $\Psi(x)$ be a kernel average of the form*

$$\Psi(x) = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} K_h(X_{it} - x) Z_{it}.$$

It holds that

$$\sup_{x \in [0,1]} \big|\Psi(x) - \mathbb{E}[\Psi(x)]\big| = O_p\Big(\sqrt{\frac{\log nT}{nTh}}\Big).$$

Lemma B3. *Let*

$$\Psi(x) = \frac{1}{n} \sum_{i=1}^{n} V_i(x) W_i(x)$$

with

$$\begin{aligned} V_i(x) &= \Big(\frac{1}{T} \sum_{t=1}^{T} \big(K_h(X_{it} - x) - \mathbb{E}[K_h(X_{it} - x)]\big)\Big)^{\nu} \\ W_i(x) &= \frac{1}{T} \sum_{t=1}^{T} K_h(X_{it} - x) Z_{it} \end{aligned}$$

for some fixed natural number ν and assume that the variables Z_{it} satisfy $\mathbb{E}[Z_{it}|X_{it}] = 0$.

Then

$$\sup_{x\in[0,1]} \big|\Psi(x)\big| = o_p\Big(\frac{1}{\sqrt{nTh}}\Big).$$

Lemma B4. *Let*

$$\Psi(x) = \frac{1}{T}\sum_{t=1}^{T} V_t(x) W_t,$$

where $W_t = \frac{1}{n}\sum_{i=1}^n Z_{it}$ *and*

$$V_t(x) = \frac{1}{n}\sum_{i=1}^{n} \big(K_h(X_{it}-x) - \mathbb{E}[K_h(X_{it}-x)]\big).$$

Assume that the variables Z_{it} *have mean zero. Then it holds that*

$$\sup_{x\in[0,1]} \big|\Psi(x)\big| = o_p\Big(\frac{1}{\sqrt{nTh}}\Big).$$

Lemma B5. *Let*

$$\Psi(x) = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{1}{nT}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{t=1}^{T}\varphi_{it}(x) Z_{jt}\Big)$$

with $\varphi_{it}(x) = K_h(X_{it}-x)\widehat{\phi}_i(x)$ *and* $\widehat{\phi}_i(x)$ *an estimator based on the data* $\{X_{it} : t = 1,\ldots,T\}$*. Assume that* $\widehat{\phi}_i(x)$ *has the following two properties:*

(a) $\mathbb{P}(\max_{1\le i\le n}\sup_{x\in[0,1]}|\widehat{\phi}_i(x)| > Cb_{n,T}) = o(1)$ *for a sufficiently large constant* C *and a null sequence* $\{b_{n,T}\}$ *which satisfies* $b_{n,T}^2/h \le C(nT)^{-\eta}$ *for some small* $\eta > 0$*.*

(b) $\max_{1\le i\le n}|\widehat{\phi}_i(x) - \widehat{\phi}_i(x')| \le c_{n,T}|x-x'|$ *with probability tending to one for some sequence* $\{c_{n,T}\}$ *which satisfies* $c_{n,T} \le (nT)^C$ *for some positive constant* C*.*

In addition, let the variables Z_{it} *have mean zero. Then it holds that*

$$\sup_{x\in[0,1]} \big|\Psi(x)\big| = o_p\Big(\frac{1}{\sqrt{nTh}}\Big).$$

To prove the above lemmas, we use a covering argument together with an exponential inequality, thus following the common strategy to be found for example in Bosq (1998), Masry (1996) or Hansen (2008). For the proof of Lemmas B1 and B2, these standard arguments have to be modified only slightly. For the proof of Lemmas B3–B5 in contrast, some rather intricate and non-standard arguments are needed to get the overall strategy to work.

References

[1] Arellano, M. (2003). *Panel data econometrics.* Oxford University Press.

[2] Atak, A., Linton, O. & Xiao, Z. (2011). A semiparametric model for unbalanced data with application to climate change in the United Kingdom. *Journal of Econometrics* **164** 92-115.

[3] Alizadeh, S., Brandt, M. W. & Diebold, F. X. (2002) Range-based estimation of stochastic volatility models. *Journal of Finance* **57** 1047-1091.

[4] Amihud, Y. (2002). Illiquidity and stock returns: cross-section and time series effects. *Journal of Financial Markets* **5** 31-56.

[5] Anderson, T., Bollerslev, T., Diebold, F. X. & Labys, P. (2003). Modelling and forecasting realized volatility. *Econometrica* **71** 579-625.

[6] Angrist, J. D. & Pischke, J. (2009). *Mostly harmless econometrics: an empiricist's companion.* Princeton University Press.

[7] Bai, J. (2003). Inferential theory for factor models of large dimension. *Econometrica* **71** 135-171.

[8] Bai, J. (2004). Estimating cross-section common stochastic trends in nonstationary panel data. *Journal of Econometrics* **122** 137-183.

[9] Bai, J. & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191-221.

[10] Baltagi, B. H., Hidalgo, J. & Li, Q. (1996). A nonparametric test for poolability using panel data. *Journal of Econometrics* **75** 345-367.

[11] Borak, S., Härdle, W., Mammen, E. & Park, B. U. (2009). Time series modelling with semiparametric factor dynamics. *Journal of the American Statistical Association* **104** 284-298.

[12] Bosq, D. (1998). *Nonparametric statistics for stochastic processes: estimation and prediction.* Springer, Berlin.

[13] Bradley, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys* **2** 107-144.

[14] Breiman, L. & Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlations (with discussion). *Journal of the American Statistical Association* **80** 580-619.

[15] Brogaard, J., Hendershott, T., Hunt, S., Latza, T., Pedace, L. & Ysusi, C. (2013). High-frequency trading and the execution costs of institutional investors. FSA Occasional Paper Series 43.

[16] Chen, X. (2010). Penalized sieve estimation and inference of semi-nonparametric dynamic models: a selective review. In: *Advances in Economics and Econometrics: Tenth World Congress, Volume III, Econometrics*, (Eds. D. Acemoglu, M. Arellano and E. Dekel). Cambridge University Press.

[17] Chen, J., Gao, J. & Li, D. (2012). Semiparametric trending panel data models with cross-sectional dependence. *Journal of Econometrics* **171** 81-85.

[18] Chen, J., Gao, J. & Li, D. (2013a). Estimation in partially linear single-index panel data models with fixed effects. *Journal of Business and Economic Statistics* **31** 315-330.

[19] Chen, J., Gao, J. & Li, D. (2013b). Estimation in single-index panel data models with heterogeneous link functions. *Econometric Reviews* **32** 928-955.

[20] Connor, G. & Korajczyk, R. A. (1988). Risk and return in an equilibrium APT: application of a new test methodology. *Journal of Financial Economics* **21** 255-289.

[21] Connor, G. & Korajczyk, R. A. (1993). A test for the number of factors in an approximate factor model. *Journal of Finance* **48** 1263-1288.

[22] Connor, G., Hagmann, M. & Linton, O. (2012). Efficient estimation of a semiparametric characteristic-based factor model for security returns. *Econometrica* **80** 713-754

[23] Cutler, D. M., Poterba, J. M. & Summers, L. H. (1989). What moves stock prices? *Journal of Portfolio Management* **15** 4-12.

[24] Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *Annals of Statistics* **25** 1-37.

[25] Fafchamps, M. & Gubert, F. (2007). The formation of risk sharing networks. *Journal of Development Economics* **83** 326-350.

[26] Fan, J. & Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association* **99** 710-723.

[27] Fan, J., Huang, T. & Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *Journal of the American Statistical Association* **102** 632-641.

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

[29] Fan, J. & Yao, Q. (2003). *Nonlinear time series analysis.* Springer, Berlin.

[30] Fengler, M., Härdle, W. & Mammen, E. (2007). A semiparametric factor model for implied volatility surface dynamics. *Journal of Financial Econometrics* **5** 189-218.

[31] Gresse, C. (2010). Effects of the competition between multiple trading platforms on market liquidity: evidence from the MiFID experience. Working Paper.

[32] Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory* **24** 726-748.

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

[34] Hoderlein, S., Klemelä, J. & Mammen, E. (2010). Analyzing the random coefficient model nonparametrically. *Econometric Theory* **26** 804-837.

[35] Hoderlein, S., Mammen, E. & Yu, K. (2011). Non-parametric models in binary choice fixed effects panel data. *The Econometrics Journal* **14** 351-367.

[36] Hsiao, C. (1986, 2003). *Analysis of panel data.* Cambridge University Press.

[37] Jenish, N. (2012). Nonparametric spatial regression under near-epoch dependence. *Journal of Econometrics* **167** 224-239.

[38] Jolliffe, I. T. (2002). *Principal component analysis.* Springer, Berlin.

[39] Kneip, A., Sickles, R. C. & Song, W. (2012). A new panel data treatment for heterogeneity in time trends. *Econometric Theory* **28** 590-628.

[40] Kyriazidou, E. (1997). Estimation of a panel data sample selection model. *Econometrica* **65** 1335-1364.

[41] Li, Q. & Sun, Y. (2011). A consistent nonparametric test of parametric regression functional form in fixed effects panel data models. Working Paper.

[42] Li, D., Chen, J. & Gao, J. (2011). Non-parametric time-varying coefficient panel data models with fixed effects. *The Econometrics Journal* **14** 387-408.

[43] Lin, X. & Carroll, R. J. (2006). Semiparametric estimation in general repeated measures problems. *Journal of the Royal Statistical Society B* **68** 69-88.

[44] Linton, O. & Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* **82** 93-100.

[45] Linton, O., Maasoumi, E.& Whang, Y. (2005). Testing for stochastic dominance under general conditions: a subsampling approach. *Review of Economic Studies* **72** 735-765. (Corrigendum, 2007, 75, 1-5.)

[46] Linton, O., Nielsen, J. P. & Nielsen, S. F. (2009) Non-parametric regression with a latent time series. *The Econometrics Journal* **12** 187-207.

[47] Magnus, J. R. (1985). On differentiating eigenvalues and eigenvectors. *Econometric Theory* **1** 179-191.

[48] Mammen, E., Linton, O. & Nielsen, J. P. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics* **27** 1443-1490.

[49] Mammen, E., Støve, B. & Tjøstheim, D. (2009). Nonparametric additive models for panels of time series. *Econometric Theory* **25** 442-481.

[50] Mandelbrot, B. (1963). The variation of certain speculative prices. *The Journal of Business* **36** 394-419.

[51] Manski, C. F. (2008). *Identification for prediction and decision.* Harvard University Press.

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

[53] O'Hara, M. & Ye, M. (2009). Is fragmentation harming market quality? *Journal of Financial Economics* **100** 459-474.

[54] Neyman, J. & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16** 1-32.

[55] Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica* **74** 967-1012.

[56] Phillips, P. C. B. & Moon, H. R. (1999). Linear regression limit theory for nonstationary panel data. *Econometrica* **67** 1057-1113.

[57] Porter, J. (1996). Nonparametric regression estimation for a flexible panel data model. Ph.D. Thesis, Department of Economics, MIT.

[58] Qian, J. & Wang, L. (2011). Estimating semiparametric panel data models by marginal integration. *Journal of Econometrics* **167** 483-493.

[59] Robinson, P. M. (2011). Asymptotic theory for nonparametric regression with spatial data. *Journal of Econometrics* **165** 5-19.

[60] Robinson, P. M. (2012). Nonparametric trending regression with cross-sectional dependence. *Journal of Econometrics* **169** 4-14.

[61] Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Nat. Acad. Sci. U.S.A.* **4** 43-47.

[62] Sarr, A. & Lybek, T. (2002). Measuring liquidity in financial markets. Working Paper.

[63] Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Annals of Statistics* **8** 1348-1360.

[64] Vogt, M. (2012). Nonparametric regression for locally stationary time series. *Annals of Statistics* **40** 2601-2633.

[65] Wang, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika* **90** 43-52.

[66] Wooldridge, J. M. (2006). Cluster-sample methods in applied econometrics: an extended analysis. Working Paper.

[67] Zhu, M. & Ghodsi, A. (2006). Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics and Data Analysis* **51** 918-930.

Supplementary Material for "A Semiparametric Model for Heterogeneous Panel Data with Fixed Effects"

Lena Körber
London School of Economics

Oliver Linton
University of Cambridge

Michael Vogt
University of Konstanz

December 19, 2013

Abstract

In this supplement, we investigate the finite sample performance of our estimators by means of a simulation study. In addition, we provide the technical details and proofs that are omitted in the paper.

1 Simulation study

To assess the small sample properties of our estimation methods, we simulate data from the following model setup: The regressors X_{it} are i.i.d. draws from a uniform distribution on the unit interval. Moreover, there are $K = 2$ common component functions defined by

$$\mu_1(x) = \sqrt{2}\sin(2\pi x) \qquad \text{and} \qquad \mu_2(x) = \sqrt{2}\cos(2\pi x).$$

These functions are orthonormal with respect to the standard scalar product on $[0, 1]$, i.e., $\int_0^1 \mu_1(x)\mu_2(x)dx = 0$ and $\int_0^1 \mu_k^2(x)dx = 1$ for $k = 1, 2$. We have chosen these functions as they indeed look similar to some of the estimated μ-functions from the application in Section 8 of the paper. As the regressors are uniformly distributed on $[0, 1]$, we obtain that $\mathbb{E}[\mu_k(X_{it})] = 0$ for $k = 1, 2$ and thus $\mathbb{E}[m_i(X_{it})] = 0$ with $m_i(x) = \beta_{i1}\mu_1(x) + \beta_{i2}\mu_2(x)$. Thus, the regression functions fulfill the normalization $\mathbb{E}[m_i(X_{it})] = 0$ that is assumed for identification.

The factor loadings β_{ik} $(i = 1, \dots, n,\ k = 1, \dots, K)$ are generated deterministically according to

$$\beta_{i1} = 1 + \frac{i-1}{n-1} \qquad \text{and} \qquad \beta_{i2} = 2 - \frac{i-1}{n-1}.$$

With this choice, the coefficient β_{i1} of the function μ_1 linearly increases from 1 to 2 as the index i grows larger. Similarly, the loading β_{i2} of μ_2 decreases from 2 to 1. Hence, the component function μ_1 becomes more and more important as the index i gets larger and vice versa for the second component μ_2. The weighting matrix W is given by

$$W = \begin{pmatrix} 2/n & \dots & 2/n & 0 & \dots & 0 \\ 0 & \dots & 0 & 2/n & \dots & 2/n \end{pmatrix}.$$

Note that the coefficient matrix B and the weighting matrix W are chosen such that $S = WB$ has full rank. In addition, the μ-functions are orthonormal. Hence, the normalization conditions of Section 3 in the paper are fulfilled. In the simulations, S and μ are renormalized such that they fulfill condition (I_W1) of the paper.

The individual and time fixed effects α_i and γ_t are i.i.d. standard normal random variables. The model constant μ_0 is set to zero, and the disturbances ε_{it} are i.i.d. normal random variables with zero mean and standard deviation σ_ε. To vary the signal-to-noise ratio in the model, we choose two different values for σ_ε, in particular $\sigma_\varepsilon \in \{1, 2\}$. As can be seen, there is no time series dependence in the error terms and the regressors, and we have only included a very limited form of fixed effects. These simplifications allow us to get a clear picture of the performance of our estimation methods. It goes without saying that they may be relaxed, i.e., we may allow for time series dependence in the model variables and add some more complicated forms of fixed effects.

In what follows, we examine the performance of our estimators $\widehat{\mu}$ and $\widehat{\beta}_i$. Moreover, we assess the small sample behaviour of two estimators of the average regression function $m_{\text{av}}(x) = n^{-1} \sum_{i=1}^n m_i(x)$ defined by $\widehat{m}_{\text{av}}(x) = n^{-1} \sum_{i=1}^n \widehat{m}_i(x)$ and $\widehat{m}^e_{\text{av}}(x) = n^{-1} \sum_{i=1}^n \widehat{m}^e_i(x)$, where $\widehat{m}^e_i(x) = \widehat{\beta}_i^\intercal \widehat{\mu}(x)$ are the reconstructed regression functions. As performance measures, we employ the mean squared errors

$$\text{MSE}(\widehat{\mu}_k) = \int_0^1 \big[\widehat{\mu}_k(x) - \mu_k(x)\big]^2 dx$$

for $k = 1, 2$ along with

$$\text{MSE}(\widehat{m}_{\text{av}}) = \int_0^1 \big[\widehat{m}_{\text{av}}(x) - m_{\text{av}}(x)\big]^2 dx$$
$$\text{MSE}(\widehat{m}^e_{\text{av}}) = \int_0^1 \big[\widehat{m}^e_{\text{av}}(x) - m_{\text{av}}(x)\big]^2 dx.$$

Table 1: Small sample properties of the estimators in the design with $\sigma_\varepsilon = 1$.

a) MSE of $\widehat{m}_{av}$

$T\backslash n$	50	100	150	200
50	0.0449	0.0362	0.0331	0.032
100	0.0425	0.0347	0.0324	0.0311
150	0.042	0.0345	0.0321	0.0309
200	0.0418	0.0343	0.0321	0.0308

b) MSE of $\widehat{m}_{av}^e$

$T\backslash n$	50	100	150	200
50	0.017	0.0124	0.0107	0.0101
100	0.0109	0.0077	0.0071	0.0066
150	0.0092	0.0067	0.0061	0.0058
200	0.0085	0.0062	0.0057	0.0054

c) MSE of $\widehat{\mu}_1$

$T\backslash n$	50	100	150	200
50	0.0159	0.0099	0.0064	0.0052
100	0.008	0.004	0.003	0.0024
150	0.0053	0.0029	0.0021	0.0016
200	0.0041	0.0022	0.0016	0.0012

d) MSE of $\widehat{\mu}_2$

$T\backslash n$	50	100	150	200
50	0.0159	0.0092	0.0063	0.005
100	0.009	0.0051	0.0039	0.0035
150	0.0065	0.0043	0.0035	0.003
200	0.0054	0.0035	0.003	0.0027

e) L_1*-norm of the coefficient estimates* $\widehat{\beta}_{i1}$

$T\backslash n$	50	100	150	200
50	0.129	0.125	0.124	0.123
100	0.089	0.0853	0.0841	0.0837
150	0.072	0.0684	0.0679	0.0675
200	0.0627	0.0591	0.0583	0.0581

f) L_1*-norm of the coefficient estimates* $\widehat{\beta}_{i2}$

$T\backslash n$	50	100	150	200
50	0.136	0.13	0.128	0.128
100	0.0973	0.0914	0.0895	0.0886
150	0.0822	0.0752	0.0732	0.0721
200	0.0732	0.0658	0.0641	0.0629

Table 2: Small sample properties of the estimators in the design with $\sigma_\varepsilon = 2$.

a) MSE of $\widehat{m}_{av}$

$T\backslash n$	50	100	150	200
50	0.0512	0.039	0.0352	0.034
100	0.0456	0.0362	0.0334	0.0318
150	0.0442	0.0354	0.0327	0.0314
200	0.0428	0.035	0.0326	0.0312

b) MSE of $\widehat{m}^e_{av}$

$T\backslash n$	50	100	150	200
50	0.0233	0.0153	0.0129	0.0118
100	0.0144	0.00936	0.008	0.00749
150	0.0115	0.00779	0.00682	0.00632
200	0.00993	0.0069	0.00634	0.00579

c) MSE of $\widehat{\mu}_1$

$T\backslash n$	50	100	150	200
50	0.0343	0.019	0.0129	0.0103
100	0.0169	0.0089	0.00604	0.00465
150	0.0106	0.0057	0.00402	0.00294
200	0.00804	0.00418	0.00292	0.00225

d) MSE of $\widehat{\mu}_2$

$T\backslash n$	50	100	150	200
50	0.0339	0.0183	0.0125	0.00993
100	0.0171	0.00942	0.0071	0.00568
150	0.0117	0.007	0.00542	0.00429
200	0.00955	0.00555	0.00433	0.00366

e) L_1*-norm of the coefficient estimates* $\widehat{\beta}_{i1}$

$T\backslash n$	50	100	150	200
50	0.233	0.231	0.231	0.231
100	0.162	0.162	0.161	0.161
150	0.134	0.131	0.131	0.131
200	0.115	0.113	0.114	0.114

f) L_1*-norm of the coefficient estimates* $\widehat{\beta}_{i2}$

$T\backslash n$	50	100	150	200
50	0.237	0.234	0.234	0.234
100	0.169	0.166	0.164	0.164
150	0.138	0.135	0.134	0.134
200	0.12	0.118	0.116	0.116

The small sample behavior of the coefficient estimates $\widehat{\beta}_i$ is evaluated by the L_1-norm

$$\frac{1}{n}\sum_{i=1}^{n}|\widehat{\beta}_{ik}-\beta_{ik}|$$

for $k=1,2$. Throughout, we assume the number of components $K=2$ to be known and use the version of our method which is based on local linear estimators. Moreover, the bandwidth is set to $h=0.15$ and we use an Epanechnikov kernel. As a robustness check, we have varied the bandwidth. As this produces very similar results, we have however not reported them here. Finally, the number of replications is set to $N=1000$.

Tables 1 and 2 report the simulation results. Overall, our estimators perform well even for the moderate sample sizes $n=T=50$. The accuracy of the estimators increases steadily as the dimensions n and T grow larger, the only exception being the estimates of the factor loadings which improve above all in T but not so much in n. This is a very natural phenomenon as the factor loadings are estimated from individual time series regressions. Hence, their quality should depend above all on the time series dimension and not so much on the length of the cross-section. It is also worth mentioning that the MSE of the reconstructed average $\widehat{m}^e_{\text{av}}$ is smaller and converges faster to zero than the MSE of $\widehat{m}_{\text{av}}$. This observation is consistent with the asymptotic properties of the estimators $\widehat{m}_i$ and $\widehat{m}^e_i$: While $\widehat{m}_i$ converges at the rate $(Th)^{-1/2}$, $\widehat{m}^e_i$ converges at the faster rate $T^{-1/2}$ (cp. Section 5.4 in the paper). Finally, when the standard deviation σ_ε of the disturbance terms is increased to 2, the signal-to-noise ratio in the model decreases. This makes it harder to estimate the functions and parameters of interest, which is reflected in higher values of the MSE and the L_1-norm as can be seen upon comparing Tables 1 and 2.

2 Technical details

In what follows, we prove the uniform convergence results which are stated in Appendix B of the paper.

Proof of Lemma B1. The proof proceeds by slightly modifying standard arguments to derive uniform convergence rates for kernel estimators. We are thus content with giving some remarks on the necessary modifications.

We start with the proof of (35). Write

$$\mathbb{P}\Big(\max_{1\le i\le n}\sup_{x\in[0,1]}\big|\Psi_i(x)-\mathbb{E}[\Psi_i(x)]\big|>Ca_T\Big)\le\sum_{i=1}^{n}\mathbb{P}\Big(\sup_{x\in[0,1]}\big|\Psi_i(x)-\mathbb{E}[\Psi_i(x)]\big|>Ca_T\Big)$$

with $a_T=\sqrt{\log T/Th}$. Going along the lines of the standard proving strategy, the probabilities on the right-hand side can be bounded by a null sequence $\{c_T\}$ which does not depend on i. Under our conditions, this sequence can be chosen such that $\{nc_T\}$ is a null sequence as well. This yields the result.

We now turn to (34). As the variables Z_{it} are not bounded, we have to replace them by truncated versions $Z_{it}^{\leq} = Z_{it} I(|Z_{it}| \leq \tau_{n,T})$ in a first step. Since we maximize over i, the truncation sequence $\tau_{n,T}$ must be chosen to go to infinity much faster than in the standard case where i is fixed. In particular, we take $\tau_{n,T} = (nT)^{1/(\theta-\delta)}$ for some small $\delta > 0$. Applying the same proving strategy as for (35) to the truncated version of $\Psi_i(x)$, one can see that the arguments still go through. However, as the truncation points $\tau_{n,T}$ diverge much faster than in the standard case with fixed i, the convergence rate turns out to be slower than the standard rate $\sqrt{\log T/Th}$. □

Proof of Lemma B2. As the proof closely follows standard arguments, we only provide a short sketch: Let $a_{n,T} = \sqrt{\log nT/nTh}$ and write $\Psi(x) = \Psi^{\leq}(x) + \Psi^{>}(x)$ with

$$\begin{aligned}\Psi^{\leq}(x) &= \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} K_h(X_{it}-x) Z_{it} I(|Z_{it}| \leq \tau_{n,T}) \\ \Psi^{>}(x) &= \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} K_h(X_{it}-x) Z_{it} I(|Z_{it}| > \tau_{n,T}),\end{aligned}$$

where the truncation sequence $\tau_{n,T}$ is given by $\tau_{n,T} = (nT)^{1/(\theta-\delta)}$ with some small $\delta > 0$. We thus have

$$\Psi(x) - \mathbb{E}[\Psi(x)] = (\Psi^{\leq}(x) - \mathbb{E}[\Psi^{\leq}(x)]) + (\Psi^{>}(x) - \mathbb{E}[\Psi^{>}(x)]).$$

Straightforward arguments show that $\sup_{x\in[0,1]} |\Psi^{>}(x) - \mathbb{E}[\Psi^{>}(x)]| = O_p(a_{n,T})$. To analyze the term $\sup_{x\in[0,1]} |\Psi^{\leq}(x) - \mathbb{E}[\Psi^{\leq}(x)]|$, we cover the unit interval by a grid of points $G_{n,T}$ that gets finer and finer as the sample size increases. We then replace the supremum over x by the maximum over the grid points $x \in G_{n,T}$ and show that the resulting error is negligible. To complete the proof, we write

$$\mathbb{P}\Big(\max_{x\in G_{n,T}} \big|\Psi^{\leq}(x) - \mathbb{E}[\Psi^{\leq}(x)]\big| > Ca_{n,T}\Big) \leq \sum_{x\in G_{n,T}} \mathbb{P}\big(\big|\Psi^{\leq}(x) - \mathbb{E}[\Psi^{\leq}(x)]\big| > Ca_{n,T}\big)$$

and bound the probabilities $\mathbb{P}(|\Psi^{\leq}(x) - \mathbb{E}[\Psi^{\leq}(x)]| > Ca_{n,T})$ for each grid point with the help of an exponential inequality. To do so, let

$$\Psi^{\leq}(x) - \mathbb{E}[\Psi^{\leq}(x)] = \sum_{i=1}^{n}\sum_{t=1}^{T} W_{it}(x)$$

with $W_{it}(x) = \frac{1}{nT}\{K_h(X_{it}-x)Z_{it}I(|Z_{it}| \leq \tau_{n,T}) - \mathbb{E}[K_h(X_{it}-x)Z_{it}I(|Z_{it}| \leq \tau_{n,T})]\}$ and split up the expression $\sum_{t=1}^{T} W_{it}(x)$ into a growing number of blocks of increasing size. Using Bradley's lemma (see Lemma 1.2 in Bosq (1998)), we can replace these blocks by independent versions and apply an exponential inequality. □

Proof of Lemma B3. Throughout the proof, we use the following notation. Let

$$\begin{array}{ll} \mathcal{C}_T: & \text{the event that } \max_i \sup_x |V_i(x)^{1/\nu}| \le C\sqrt{\log T/Th} \text{ and} \\ & \max_i \sup_x T^{-1}\sum_{t=1}^T K_h(X_{it}-x) \le C \\ \mathcal{C}_{iT}: & \text{the event that } \sup_x |V_i(x)^{1/\nu}| \le C\sqrt{\log T/Th} \text{ and} \\ & \sup_x T^{-1}\sum_{t=1}^T K_h(X_{it}-x) \le C \end{array}$$

for a fixed large constant C. Moreover, write $\mathcal{C}_T^c$ and $\mathcal{C}_{iT}^c$ to denote the complements of $\mathcal{C}_T$ and $\mathcal{C}_{iT}$, respectively. Inspecting the proof of Lemma B1, it is easily seen that $P(\mathcal{C}_T^c) = o(1)$ and $P(\mathcal{C}_{iT}^c) = o(1)$, given that the constant C in the definition of the events $\mathcal{C}_T$ and $\mathcal{C}_{iT}$ is chosen sufficiently large. With this notation at hand, we obtain that

$$\begin{aligned} \mathbb{P}\Big(\sup_{x\in[0,1]}\big|\Psi(x)\big| > Ma_{n,T}\Big) &\le \mathbb{P}\Big(\sup_{x\in[0,1]}\big|\Psi(x)\big| > Ma_{n,T}, \mathcal{C}_T\Big) \\ &\quad + \mathbb{P}\Big(\sup_{x\in[0,1]}\big|\Psi(x)\big| > Ma_{n,T}, \mathcal{C}_T^c\Big) \\ &= \mathbb{P}\Big(\sup_{x\in[0,1]}\big|\Psi(x)\big| > Ma_{n,T}, \mathcal{C}_T\Big) + o(1), \end{aligned}$$

where $a_{n,T} = (\log nT\sqrt{nTh})^{-1}$ and M is a large positive constant. Moreover,

$$\begin{aligned} \mathbb{P}\Big(\sup_{x\in[0,1]}\big|\Psi(x)\big| > Ma_{n,T}, \mathcal{C}_T\Big) &= \mathbb{P}\Big(\sup_{x\in[0,1]}\Big|\frac{1}{n}\sum_{i=1}^n V_i(x)W_i(x)\Big| > Ma_{n,T}, \mathcal{C}_T\Big) \\ &= \mathbb{P}\Big(\sup_{x\in[0,1]}\Big|\frac{1}{n}\sum_{i=1}^n I(\mathcal{C}_T)V_i(x)W_i(x)\Big| > Ma_{n,T}\Big) \\ &\le \mathbb{P}\Big(\sup_{x\in[0,1]}\Big|\frac{1}{n}\sum_{i=1}^n I(\mathcal{C}_{iT})V_i(x)W_i(x)\Big| > Ma_{n,T}\Big). \end{aligned}$$

Now write

$$\frac{1}{n}\sum_{i=1}^n I(\mathcal{C}_{iT})V_i(x)W_i(x) = Q^{\le}(x) + Q^{>}(x)$$

with the two terms on the right-hand side being defined as

$$\begin{aligned} Q^{\le}(x) &= \frac{1}{n}\sum_{i=1}^n I(\mathcal{C}_{iT})V_i(x)W_i^{\le}(x) \\ Q^{>}(x) &= \frac{1}{n}\sum_{i=1}^n I(\mathcal{C}_{iT})V_i(x)W_i^{>}(x). \end{aligned}$$

Here, $W_i(x) = W_i^{\le}(x) + W_i^{>}(x)$ with

$$W_i^{\le}(x) = \frac{1}{T}\sum_{t=1}^T K_h(X_{it}-x)Z_{it}^{\le}$$

$$W_i^{>}(x) = \frac{1}{T} \sum_{t=1}^{T} K_h(X_{it} - x) Z_{it}^{>}$$

and $Z_{it} = Z_{it}^{\leq} + Z_{it}^{>}$ with

$$\begin{aligned} Z_{it}^{\leq} &= Z_{it} I(|Z_{it}| \leq \tau_{n,T}) - \mathbb{E}[Z_{it} I(|Z_{it}| \leq \tau_{n,T}) | X_{it}] \\ Z_{it}^{>} &= Z_{it} I(|Z_{it}| > \tau_{n,T}) - \mathbb{E}[Z_{it} I(|Z_{it}| > \tau_{n,T}) | X_{it}], \end{aligned}$$

where the truncation sequence $\tau_{n,T}$ is chosen to equal $\tau_{n,T} = (nT)^{1/(\theta - \delta)}$ for some small $\delta > 0$. We now arrive at

$$\begin{aligned} &\mathbb{P}\Big(\sup_{x \in [0,1]} \Big| \frac{1}{n} \sum_{i=1}^{n} I(\mathcal{C}_{iT}) V_i(x) W_i(x) \Big| > M a_{n,T} \Big) \\ &\qquad \leq \mathbb{P}\Big(\sup_{x \in [0,1]} |Q^{\leq}(x)| > \frac{M}{2} a_{n,T} \Big) + \mathbb{P}\Big(\sup_{x \in [0,1]} |Q^{>}(x)| > \frac{M}{2} a_{n,T} \Big). \end{aligned}$$

In the remainder of the proof, we show that the two terms on the right-hand side converge to zero as the sample size goes to infinity. To do so, we proceed in several steps.

Step 1. We start by considering the term $Q^{>}(x)$. It holds that

$$\begin{aligned} &\mathbb{P}\Big(\sup_{x \in [0,1]} \Big| \frac{1}{n} \sum_{i=1}^{n} I(\mathcal{C}_{iT}) V_i(x) \Big(\frac{1}{T} \sum_{t=1}^{T} K_h(X_{it} - x) Z_{it} I(|Z_{it}| > \tau_{n,T}) \Big) \Big| > C a_{n,T} \Big) \\ &\leq \mathbb{P}\Big(|Z_{it}| > \tau_{n,T} \text{ for some } 1 \leq i \leq n \text{ and } 1 \leq t \leq T \Big) \\ &\leq \sum_{i=1}^{n} \sum_{t=1}^{T} \mathbb{P}(|Z_{it}| > \tau_{n,T}) \leq \sum_{i=1}^{n} \sum_{t=1}^{T} \mathbb{E}\Big[\frac{|Z_{it}|^{\theta}}{\tau_{n,T}^{\theta}} \Big] \leq C \frac{nT}{\tau_{n,T}^{\theta}} \to 0. \end{aligned}$$

In addition,

$$\begin{aligned} &\sup_{x \in [0,1]} \Big| \frac{1}{n} \sum_{i=1}^{n} I(\mathcal{C}_{iT}) V_i(x) \Big(\frac{1}{T} \sum_{t=1}^{T} K_h(X_{it} - x) \mathbb{E}[Z_{it} I(|Z_{it}| > \tau_{n,T}) | X_{it}] \Big) \Big| \\ &\leq C \sqrt{\frac{\log T}{Th}} \max_{1 \leq i \leq n} \max_{1 \leq t \leq T} \mathbb{E}\big[|Z_{it}| I(|Z_{it}| > \tau_{n,T}) | X_{it} \big] \\ &\leq C \sqrt{\frac{\log T}{Th}} \frac{1}{\tau_{n,T}^{\theta - 1}} \leq C a_{n,T}, \end{aligned}$$

where the third line follows by (A4'). As a result,

$$\mathbb{P}\Big(\sup_{x \in [0,1]} |Q^{>}(x)| > \frac{M}{2} a_{n,T} \Big) = o(1)$$

for M sufficiently large.

Step 2. We now turn to the analysis of the term $Q^{\leq}(x)$. Cover the region $[0,1]$ with open intervals J_l ($l = 1, \dots, L_{n,T}$) of length $C/L_{n,T}$ and let x_l be the midpoint of the interval J_l. Then

$$\sup_{x \in [0,1]} |Q^{\leq}(x)| \leq \max_{1 \leq l \leq L_{n,T}} |Q^{\leq}(x_l)| + \max_{1 \leq l \leq L_{n,T}} \sup_{x \in J_l} |Q^{\leq}(x) - Q^{\leq}(x_l)|.$$

For any point $x \in J_l$, we have

$$I(\mathcal{C}_{iT}) \big| V_i(x) W_i^{\leq}(x) - V_i(x_l) W_i^{\leq}(x_l) \big| \leq \frac{C \tau_{n,T}}{h^2} |x - x_l| \leq \frac{C \tau_{n,T}}{h^2 L_{n,T}}.$$

Therefore,

$$\max_{1 \leq l \leq L_{n,T}} \sup_{x \in J_l} |Q^{\leq}(x) - Q^{\leq}(x_l)| \leq \frac{C \tau_{n,T}}{h^2 L_{n,T}}.$$

Choosing $L_{n,T} \to \infty$ with $L_{n,T} = C \tau_{n,T} / a_{n,T} h^2$, we obtain that

$$\max_{1 \leq l \leq L_{n,T}} \sup_{x \in J_l} |Q^{\leq}(x) - Q^{\leq}(x_l)| \leq C a_{n,T}.$$

If we pick the constant M large enough, we thus arrive at

$$\mathbb{P}\Big(\sup_{x \in [0,1]} |Q^{\leq}(x)| > \frac{M}{2} a_{n,T} \Big) \leq \mathbb{P}\Big(\max_{1 \leq l \leq L_{n,T}} |Q^{\leq}(x_l)| > \frac{M}{4} a_{n,T} \Big) + o(1).$$

Step 3. It remains to show that

$$\mathbb{P}\Big(\max_{1 \leq l \leq L_{n,T}} |Q^{\leq}(x_l)| > \frac{M}{4} a_{n,T} \Big) = o(1)$$

for some large fixed constant M. To do so, we write

$$\mathbb{P}\Big(\max_{1 \leq l \leq L_{n,T}} |Q^{\leq}(x_l)| > \frac{M}{4} a_{n,T} \Big) \leq P_1 + P_2$$

with

$$P_1 = \mathbb{P}\Big(\max_{1 \leq l \leq L_{n,T}} |Q^{\leq}(x_l) - \mathbb{E} Q^{\leq}(x_l)| > \frac{M}{8} a_{n,T} \Big)$$
$$P_2 = \mathbb{P}\Big(\max_{1 \leq l \leq L_{n,T}} |\mathbb{E} Q^{\leq}(x_l)| > \frac{M}{8} a_{n,T} \Big).$$

First consider the term P_2. If $\nu \geq 3$, then

$$\begin{aligned} |\mathbb{E} Q^{\leq}(x_l)| &= \Big| \frac{1}{n} \sum_{i=1}^n \mathbb{E}[I(\mathcal{C}_{iT}) V_i(x_l) W_i^{\leq}(x_l)] \Big| \\ &\leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}\big[I(\mathcal{C}_{iT}) V_i(x_l)^2 \big]^{1/2} \mathbb{E}\big[W_i^{\leq}(x_l)^2 \big]^{1/2} \end{aligned}$$

$$\leq \frac{C}{\sqrt{Th}}\Big(\frac{\log T}{Th}\Big)^{\nu/2} = o(a_{n,T}).$$

For $\nu \leq 2$, we write

$$\begin{aligned}
|\mathbb{E}Q^{\leq}(x_l)| &= \Big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[I(\mathcal{C}_{iT})V_i(x_l)W_i^{\leq}(x_l)]\Big| \\
&\leq \Big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[V_i(x_l)W_i^{\leq}(x_l)]\Big| + \Big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[I(\mathcal{C}_{iT}^c)V_i(x_l)W_i^{\leq}(x_l)]\Big|.
\end{aligned}$$

If $\nu = 1$, we have

$$\begin{aligned}
\big|\mathbb{E}[V_i(x_l)W_i^{\leq}(x_l)]\big| &= \Big|\frac{1}{T^2}\sum_{s,t=1}^{T}\mathbb{E}\big[(K_h(X_{is}-x_l) - \mathbb{E}[K_h(X_{is}-x_l)])K_h(X_{it}-x_l)Z_{it}^{\leq}\big]\Big| \\
&= \Big|\frac{1}{T^2}\sum_{\substack{s,t=1 \\ s\neq t}}^{T}\mathbb{E}\big[(K_h(X_{is}-x_l) - \mathbb{E}[K_h(X_{is}-x_l)])K_h(X_{it}-x_l)Z_{it}^{\leq}\big]\Big| \\
&\leq \frac{C\log T}{T} = o(a_{n,T}),
\end{aligned}$$

the last line following with the help of Davydov's inequality and (A4'). For $\nu = 2$, it holds that

$$\begin{aligned}
\big|\mathbb{E}[V_i(x_l)W_i^{\leq}(x_l)]\big| &= \Big|\frac{1}{T^3}\sum_{s,s',t=1}^{T}\mathbb{E}\big[(K_h(X_{is}-x_l) - \mathbb{E}[K_h(X_{is}-x_l)]) \\
&\qquad \times (K_h(X_{is'}-x_l) - \mathbb{E}[K_h(X_{is'}-x_l)])K_h(X_{it}-x_l)Z_{it}^{\leq}\big]\Big| \\
&\leq \frac{CT(\log T)^2}{T^3h^2} = C\Big(\frac{\log T}{Th}\Big)^2 = o(a_{n,T}),
\end{aligned}$$

the last line again following by Davydov's inequality and (A4'). In addition,

$$\mathbb{E}[I(\mathcal{C}_{iT}^c)V_i(x_l)W_i^{\leq}(x_l)] \leq \mathbb{E}[I(\mathcal{C}_{iT}^c)]^{1/2}\mathbb{E}[V_i(x_l)^2W_i^{\leq}(x_l)^2]^{1/2}.$$

Repeating the usual strategy to prove uniform convergence for kernel estimates, it can be shown that under our assumptions, $\mathbb{E}[I(\mathcal{C}_{iT}^c)] = \mathbb{P}(\mathcal{C}_{iT}^c) \leq T^{-C}$ for an arbitrarily large constant C. This yields that $\mathbb{E}[I(\mathcal{C}_{iT}^c)V_i(x_l)W_i^{\leq}(x_l)] = o(a_{n,T})$, which in turn implies that $|\mathbb{E}Q^{\leq}(x_l)| = o(a_{n,T})$ for $\nu = 1, 2$. As a result, $P_2 = o(1)$ for any $\nu \geq 1$.

To cope with the term P_1, we apply the bound

$$P_1 \leq \sum_{l=1}^{L_{n,T}}\mathbb{P}\Big(|Q^{\leq}(x_l) - \mathbb{E}Q^{\leq}(x_l)| > \frac{M}{8}a_{n,T}\Big)$$

and consider the probability $\mathbb{P}(|Q^{\leq}(x_l) - \mathbb{E}Q^{\leq}(x_l)| > Ma_{n,T}/8)$ for an arbitrary fixed grid point x_l. Write

$$Q^{\leq}(x_l) - \mathbb{E}Q^{\leq}(x_l) = \sum_{i=1}^{n} \xi_i(x_l)$$

with $\xi_i(x_l) = n^{-1}\{I(\mathcal{C}_{iT})V_i(x_l)W_i^{\leq}(x_l) - \mathbb{E}[I(\mathcal{C}_{iT})V_i(x_l)W_i^{\leq}(x_l)]\}$. Recalling the definition of the events $\mathcal{C}_{iT}$, the variables $\xi_i(x_l)$ can be bounded as follows:

$$|\xi_i(x_l)| \leq C\sqrt{\frac{\log T}{Th}}\frac{\tau_{n,T}}{n} \leq \frac{C}{(nTh)^{1/2+\delta}} := \overline{C}_{n,T}$$

with some sufficiently large constant C and a small $\delta > 0$, given that $n \gg T^{2/3}$ and $\theta > 5$. With $\lambda_{n,T} = \overline{C}_{n,T}^{-1}/2$, we obtain that $\lambda_{n,T}|\xi_i(x_l)| \leq 1/2$. As $\exp(x) \leq 1 + x + x^2$ for $|x| \leq 1/2$,

$$\mathbb{E}\Big[\exp\big(\pm\lambda_{n,T}\xi_i(x_l)\big)\Big] \leq 1 + \lambda_{n,T}^2\mathbb{E}[\xi_i(x_l)^2] \leq \exp\Big(\lambda_{n,T}^2\mathbb{E}[\xi_i(x_l)^2]\Big).$$

Using this together with Markov's inequality, we arrive at

$$\begin{aligned}
&\mathbb{P}\Big(\Big|\sum_{i=1}^{n}\xi_i(x_l)\Big| > \frac{M}{8}a_{n,T}\Big) \\
&\leq \exp\Big(-\frac{M}{8}\lambda_{n,T}a_{n,T}\Big)\Big\{\mathbb{E}\Big[\exp\big(\lambda_{n,T}\sum_{i=1}^{n}\xi_i(x_l)\big)\Big] + \mathbb{E}\Big[\exp\big(-\lambda_{n,T}\sum_{i=1}^{n}\xi_i(x_l)\big)\Big]\Big\} \\
&\leq 2\exp\Big(-\frac{M}{8}\lambda_{n,T}a_{n,T}\Big)\prod_{i=1}^{n}\exp\big(\lambda_{n,T}^2\mathbb{E}[\xi_i(x_l)^2]\big) \\
&= 2\exp\Big(-\frac{M}{8}\lambda_{n,T}a_{n,T}\Big)\exp\Big(\lambda_{n,T}^2\sum_{i=1}^{n}\mathbb{E}[\xi_i(x_l)^2]\Big).
\end{aligned}$$

Now note that

$$\mathbb{E}[\xi_i(x_l)^2] \leq \frac{1}{n^2}\mathbb{E}[I(\mathcal{C}_{iT})V_i(x_l)^2W_i^{\leq}(x_l)^2] \leq \frac{C\log T}{n^2Th}\mathbb{E}[W_i^{\leq}(x_l)^2]$$

and

$$\begin{aligned}
\mathbb{E}[W_i^{\leq}(x_l)^2] &= \frac{1}{T^2}\sum_{s,t=1}^{T}\mathbb{E}\big[K_h(X_{is}-x_l)K_h(X_{it}-x_l)Z_{is}^{\leq}Z_{it}^{\leq}\big] \\
&= \frac{1}{T^2}\sum_{s,t=1}^{T}\mathrm{Cov}\big(K_h(X_{is}-x_l)Z_{is}^{\leq}, K_h(X_{it}-x_l)Z_{it}^{\leq}\big) \leq \frac{C}{Th}.
\end{aligned}$$

Hence, $\mathbb{E}[\xi_i(x_l)^2] \leq C\log T/(nTh)^2$ and

$$\lambda_{n,T}^2\sum_{i=1}^{n}\mathbb{E}[\xi_i(x_l)^2] \leq C(nTh)^{1+2\delta}\frac{\log T}{n(Th)^2} \leq C\frac{(nT)^{2\delta}}{Th} = o(1).$$

Moreover,

$$\lambda_{n,T} a_{n,T} = \frac{(nTh)^{1/2+\delta}}{\log nT(nTh)^{1/2}} \to \infty$$

at polynomial rate. As a result,

$$\mathbb{P}\Big(\Big| \sum_{i=1}^{n} \xi_i(x_l) \Big| > \frac{M}{8} a_{n,T} \Big) \le C T^{-p},$$

where the constant $p > 0$ can be chosen arbitrarily large. This completes the proof. $\square$

Proof of Lemma B4. The proof is similar to that of Lemma B3 with the roles of i and t being reversed. Let $a_{n,T} = (\log nT\sqrt{nTh})^{-1}$ and $\tau_{n,T} = (nT)^{1/(\theta-\delta)}$ for some small $\delta > 0$. Arguments analogous to those for Step 1 in the proof of Lemma B3 yield that $\Psi(x)$ can be replaced by the term

$$Q^{\le}(x) = \frac{1}{T} \sum_{t=1}^{T} I(\mathcal{C}_{tn}) V_t(x) W_t^{\le},$$

where $W_t^{\le} = \frac{1}{n}\sum_{i=1}^n Z_{it}^{\le}$ with $Z_{it}^{\le} = Z_{it} I(|Z_{it}| \le \tau_{n,T}) - \mathbb{E}[Z_{it} I(|Z_{it}| \le \tau_{n,T})]$ and $\mathcal{C}_{tn}$ is the event that $\sup_x |V_t(x)| \le C\sqrt{\log n/nh}$ for some sufficiently large constant C. Next cover the unit interval by a grid of $L_{n,T} = C\tau_{n,T}/a_{n,T}h^2$ points. As in the proof of Lemma B3, we can show that

$$\sup_{x\in[0,1]} |Q^{\le}(x)| = \max_{1\le l\le L_{n,T}} |Q^{\le}(x_l)| + O(a_{n,T}).$$

Moreover, again repeating the arguments from Lemma B3, we obtain that for some sufficiently large constant M,

$$\begin{aligned} \mathbb{P}\Big(\max_{1\le l\le L_{n,T}} |Q^{\le}(x_l)| > M a_{n,T} \Big) &\le \mathbb{P}\Big(\max_{1\le l\le L_{n,T}} |Q^{\le}(x_l) - \mathbb{E}Q^{\le}(x_l)| > \frac{M}{2} a_{n,T} \Big) + o(1) \\ &\le \sum_{l=1}^{L_{n,T}} \mathbb{P}\Big(|Q^{\le}(x_l) - \mathbb{E}Q^{\le}(x_l)| > \frac{M}{2} a_{n,T} \Big) + o(1). \end{aligned}$$

To complete the proof, we bound the probability $\mathbb{P}(|Q^{\le}(x) - \mathbb{E}Q^{\le}(x)| > \frac{M}{2}a_{n,T})$ for an arbitrary point x by an exponential inequality. To do so, we must slightly vary the arguments for Lemma B3, taking into account the fact that $Q^{\le}(x)$ is not a sum of independent terms any more. In particular, we write

$$Q^{\le}(x) - \mathbb{E}Q^{\le}(x) = \sum_{t=1}^{T} \xi_t(x)$$

with $\xi_t(x) = T^{-1}\{I(\mathcal{C}_{tn})V_t(x)W_t^{\le} - \mathbb{E}[I(\mathcal{C}_{tn})V_t(x)W_t^{\le}]\}$ and split up the expression

$\sum_{t=1}^T \xi_t(x)$ into blocks as follows:

$$\sum_{t=1}^T \xi_t(x) = \sum_{s=1}^{q_{n,T}} B_{2s-1}(x) + \sum_{s=1}^{q_{n,T}} B_{2s}(x)$$

with $B_s(x) = \sum_{t=(s-1)r_{n,T}+1}^{sr_{n,T}} \xi_t(x)$, where $2q_{n,T}$ is the number of blocks and $r_{n,T} = T/2q_{n,T}$ is the block length. We now get

$$\begin{aligned} \mathbb{P}\Big(\Big|\sum_{t=1}^T \xi_t(x)\Big| > \frac{M}{2} a_{n,T}\Big) \le \mathbb{P}\Big(\Big|\sum_{s=1}^{q_{n,T}} B_{2s-1}(x)\Big| > \frac{M}{4} a_{n,T}\Big) \\ + \mathbb{P}\Big(\Big|\sum_{s=1}^{q_{n,T}} B_{2s}(x)\Big| > \frac{M}{4} a_{n,T}\Big). \end{aligned}$$

In what follows, we restrict attention to the first term on the right-hand side of the above display. The second one can be analyzed by analogous arguments. We make use of the following two facts:

(1) Let $\mathcal{V}^{(i)} = \{\mathcal{V}_t^{(i)} : t = 1, \dots, T\} = \{(X_{it}, Z_{it}) : t = 1, \dots, T\}$ be the time series of the i-th individual and consider the time series $\mathcal{W} = \{\mathcal{W}_t : t = 1, \dots, T\}$ with $\mathcal{W}_t = h_t(\mathcal{V}_t^{(1)}, \dots, \mathcal{V}_t^{(n)}) = h_t(X_{1t}, Z_{1t}, \dots, X_{nt}, Z_{nt})$ for some Borel functions h_t. Then by Theorem 5.2 in Bradley (2005) and the comments thereafter, the mixing coefficients $\alpha^{\mathcal{W}}(k)$ of the time series $\mathcal{W}$ are such that $\alpha^{\mathcal{W}}(k) \le \sum_{i=1}^n \alpha_i(k) \le n\alpha(k)$ for each $k \in \mathbb{N}$. In particular, letting $\alpha^\xi(k)$ be the mixing coefficients of the time series $\{\xi_t(x)\}$, it holds that $\alpha^\xi(k) \le n\alpha(k)$.

(2) By Bradley's lemma (see Lemma 1.2 in Bosq (1998)), we can construct a sequence of random variables $B_1^*(x), B_3^*(x), \dots$ such that (i) $B_1^*(x)$, $B_3^*(x)$, ... are independent, (ii) $B_{2s-1}^*(x)$ has the same distribution as $B_{2s-1}(x)$, and (iii) for $0 < \mu \le \|B_{2s-1}(x)\|_\infty$, it holds that

$$\mathbb{P}\big(|B_{2s-1}^*(x) - B_{2s-1}(x)| > \mu\big) \le 18\Big(\frac{\|B_{2s-1}(x)\|_\infty}{\mu}\Big)^{1/2} \alpha^\xi(r_{n,T}). \tag{S1}$$

Using fact (2), we can write

$$\mathbb{P}\Big(\Big|\sum_{s=1}^{q_{n,T}} B_{2s-1}(x)\Big| > \frac{M}{4} a_{n,T}\Big) \le P_1 + P_2$$

with

$$\begin{aligned} P_1 &= \mathbb{P}\Big(\Big|\sum_{s=1}^{q_{n,T}} B_{2s-1}^*(x)\Big| > \frac{M}{8} a_{n,T}\Big) \\ P_2 &= \mathbb{P}\Big(\Big|\sum_{s=1}^{q_{n,T}} \big(B_{2s-1}(x) - B_{2s-1}^*(x)\big)\Big| > \frac{M}{8} a_{n,T}\Big). \end{aligned}$$

We first consider P_1. Picking the block length to equal $r_{n,T} = (nT)^\eta$ for some small $\eta > 0$, it holds that $|B_{2s-1}(x)| \le C\sqrt{\frac{\log n}{nh}} \frac{\tau_{n,T} r_{n,T}}{T} \le \frac{C}{(nTh)^{1/2+\delta}} =: \overline{C}_{n,T}$ with some sufficiently large constant C and a small $\delta > 0$. Choosing $\lambda_{n,T} = \overline{C}_{n,T}^{-1}/2$ and applying Markov's inequality, the same arguments as in Lemma B3 yield that

$$P_1 \le 2\exp\Big(-\frac{M}{8}\lambda_{n,T} a_{n,T} + \lambda_{n,T}^2 \sum_{s=1}^{q_{n,T}} \mathbb{E}[B_{2s-1}^*(x)^2]\Big).$$

Since $\sum_{s=1}^{q_{n,T}} \mathbb{E}[B_{2s-1}^*(x)^2] \le C\log n \log T / n^2 Th$, we finally arrive at

$$P_1 \le 2\exp\Big(-\frac{M}{8}\lambda_{n,T} a_{n,T} + C\lambda_{n,T}^2 \frac{\log n \log T}{n^2 Th}\Big).$$

Direct calculations show that $\lambda_{n,T} a_{n,T} \to \infty$, whereas $\lambda_{n,T}^2 \frac{\log n \log T}{n^2 Th} = o(1)$. This implies that P_1 converges to zero at an arbitarily fast polynomial rate. Moreover, using (S1) together with the fact that $\alpha^\xi(k) \le n\alpha(k)$ and recalling that the coefficients $\alpha(k)$ decay exponentially fast to zero, it immediately follows that P_2 converges to zero at an arbitrarily fast polynomial rate as well. From this, the result easily follows. □

Proof of Lemma B5. Let $\mathcal{C}_T$ be the event that $\max_{1\le i\le n}\sup_{x\in[0,1]}|\widehat{\phi}_i(x)| \le Cb_{n,T}$ and $\mathcal{C}_{iT}$ the event that $\sup_{x\in[0,1]}|\widehat{\phi}_i(x)| \le Cb_{n,T}$. Moreover, write $\mathcal{C}_T^c$ and $\mathcal{C}_{iT}^c$ to denote the complements of $\mathcal{C}_T$ and $\mathcal{C}_{iT}$, respectively. By assumption, $P(\mathcal{C}_T^c) = o(1)$ and $P(\mathcal{C}_{iT}^c) = o(1)$. With this notation at hand, we have

$$\begin{aligned}
\mathbb{P}\Big(\sup_{x\in[0,1]}\big|\Psi(x)\big| > Ma_{n,T}\Big) &\le \mathbb{P}\Big(\sup_{x\in[0,1]}\big|\Psi(x)\big| > Ma_{n,T}, \mathcal{C}_T\big) \\
&\quad + \mathbb{P}\big(\sup_{x\in[0,1]}\big|\Psi(x)\big| > Ma_{n,T}, \mathcal{C}_T^c\Big) \\
&\le \mathbb{P}\Big(\sup_{x\in[0,1]}\big|\Psi(x)\big| > Ma_{n,T}, \mathcal{C}_T\Big) + o(1),
\end{aligned}$$

where $a_{n,T} = \sqrt{\frac{\log nT}{nTh(nT)^\eta}}$ and M is a positive constant. Moreover,

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{x\in[0,1]}\big|\Psi(x)\big| > Ma_{n,T}, \mathcal{C}_T\Big) \\
&= \mathbb{P}\Big(\sup_{x\in[0,1]}\Big|\frac{1}{n}\sum_{i=1}^n\Big(\frac{1}{nT}\sum_{j\ne i}\sum_{t=1}^T I(\mathcal{C}_T)\varphi_{it}(x)Z_{jt}\Big)\Big| > Ma_{n,T}\Big) \\
&\le \mathbb{P}\Big(\sup_{x\in[0,1]}\Big|\frac{1}{n}\sum_{i=1}^n\Big(\frac{1}{nT}\sum_{j\ne i}\sum_{t=1}^T I(\mathcal{C}_{iT})\varphi_{it}(x)Z_{jt}\Big)\Big| > Ma_{n,T}\Big).
\end{aligned}$$

Defining

$$Z_{jt}^{\le} = Z_{jt}I(|Z_{jt}| \le \tau_{n,T}) - \mathbb{E}\big[Z_{jt}I(|Z_{jt}| \le \tau_{n,T})\big]$$

$$Z_{jt}^{>} = Z_{jt} I(|Z_{jt}| > \tau_{n,T}) - \mathbb{E}\big[Z_{jt} I(|Z_{jt}| > \tau_{n,T})\big]$$

with $\tau_{n,T} = (nT)^{1/(\theta-\delta)}$ for some small $\delta > 0$, we further get that

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{1}{nT}\sum_{j\neq i}\sum_{t=1}^{T} I(\mathcal{C}_{iT})\varphi_{it}(x)Z_{jt}\Big) = Q^{\leq}(x) + Q^{>}(x)$$

with

$$\begin{aligned} Q^{\leq}(x) &= \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{1}{nT}\sum_{j\neq i}\sum_{t=1}^{T} I(\mathcal{C}_{iT})\varphi_{it}(x)Z_{jt}^{\leq}\Big) \\ Q^{>}(x) &= \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{1}{nT}\sum_{j\neq i}\sum_{t=1}^{T} I(\mathcal{C}_{iT})\varphi_{it}(x)Z_{jt}^{>}\Big). \end{aligned}$$

Hence,

$$\begin{aligned} &\mathbb{P}\Big(\sup_{x\in[0,1]}\Big|\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{1}{nT}\sum_{j\neq i}\sum_{t=1}^{T} I(\mathcal{C}_{iT})\varphi_{it}(x)Z_{jt}\Big)\Big| > Ma_{n,T}\Big) \\ &\leq \mathbb{P}\Big(\sup_{x\in[0,1]}\big|Q^{\leq}(x)\big| > \frac{M}{2}a_{n,T}\Big) + \mathbb{P}\Big(\sup_{x\in[0,1]}\big|Q^{>}(x)\big| > \frac{M}{2}a_{n,T}\Big). \end{aligned}$$

In what follows, we show that the two terms on the right-hand side converge to zero as the sample size increases. The proof splits up into several steps.

Step 1. We first consider $Q^{>}(x)$. Similarly to Lemma B3, it holds that

$$\begin{aligned} &\mathbb{P}\Big(\sup_{x\in[0,1]}\Big|\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{1}{nT}\sum_{j\neq i}\sum_{t=1}^{T} I(\mathcal{C}_{iT})\varphi_{it}(x)Z_{jt}I(|Z_{jt}| > \tau_{n,T})\Big)\Big| > Ca_{n,T}\Big) \\ &\leq \mathbb{P}\Big(|Z_{jt}| > \tau_{n,T} \text{ for some } 1 \leq j \leq n \text{ and } 1 \leq t \leq T\Big) \to 0 \end{aligned}$$

and

$$\begin{aligned} &\sup_{x\in[0,1]}\Big|\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{1}{nT}\sum_{j\neq i}\sum_{t=1}^{T} I(\mathcal{C}_{iT})\varphi_{it}(x)\mathbb{E}\big[Z_{jt}I(|Z_{jt}| > \tau_{n,T})\big]\Big)\Big| \\ &\leq \frac{Cb_{n,T}}{n^2Th}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{t=1}^{T}\mathbb{E}\big[|Z_{jt}|I(|Z_{jt}| > \tau_{n,T})\big] \leq \frac{Cb_{n,T}}{\tau_{n,T}^{\theta-1}h} \leq Ca_{n,T}. \end{aligned}$$

From this, it immediately follows that $\mathbb{P}(\sup_{x\in[0,1]}|Q^{>}(x)| > Ma_{n,T}/2) = o(1)$ for M sufficiently large.

Step 2. We now turn to the analysis of $Q^{\leq}(x)$. Let $L_{n,T} \to \infty$ with $L_{n,T} = \max\{\frac{\tau_{n,T}c_{n,T}}{ha_{n,T}}, \frac{b_{n,T}\tau_{n,T}}{h^2a_{n,T}}, (nT)^{\delta}\}$ for some small $\delta > 0$. Cover the region $[0,1]$ with open intervals J_l

$(l = 1, \dots, L_{n,T})$ of length $C/L_{n,T}$ and let x_l be the midpoint of the interval J_l. Then for $x \in J_l$,

$$
\begin{aligned}
\big|Q^{\le}(x) - Q^{\le}(x_l)\big| &\le \frac{C\tau_{n,T}}{n} \sum_{i=1}^{n} \Big(\frac{1}{nT} \sum_{j \ne i} \sum_{t=1}^{T} I(\mathcal{C}_{iT}) |\varphi_{it}(x) - \varphi_{it}(x_l)| \Big) \\
&\le \frac{C\tau_{n,T}}{n} \sum_{i=1}^{n} \Big(\frac{1}{nT} \sum_{j \ne i} \sum_{t=1}^{T} I(\mathcal{C}_{iT}) \big\{ K_h(X_{it} - x) |\widehat{\phi}_i(x) - \widehat{\phi}_i(x_l)| \\
&\qquad\qquad + |\widehat{\phi}_i(x_l)| |K_h(X_{it} - x) - K_h(X_{it} - x_l)| \big\} \Big) \\
&\le C\tau_{n,T} \Big(\frac{c_{n,T}}{h} + \frac{b_{n,T}}{h^2} \Big) |x - x_l| \le C \frac{\tau_{n,T}}{L_{n,T}} \Big(\frac{c_{n,T}}{h} + \frac{b_{n,T}}{h^2} \Big) \le C a_{n,T}
\end{aligned}
$$

with probability tending to one. From this, it immediately follows that for sufficiently large M,

$$
\mathbb{P}\Big(\sup_{x \in [0,1]} \big|Q^{\le}(x)\big| > \frac{M}{2} a_{n,T} \Big) \le \mathbb{P}\Big(\max_{1 \le l \le L_{n,T}} \big|Q^{\le}(x_l)\big| > \frac{M}{4} a_{n,T} \Big) + o(1).
$$

Step 3. It remains to show that

$$
\mathbb{P}\Big(\max_{1 \le l \le L_{n,T}} \big|Q^{\le}(x_l)\big| > \frac{M}{4} a_{n,T} \Big) = o(1)
$$

for some sufficiently large constant M. Writing

$$
\max_{1 \le l \le L_{n,T}} \big|Q^{\le}(x_l)\big| \le \max_{\substack{1 \le i \le n \\ 1 \le l \le L_{n,T}}} \Big| \sum_{j \ne i} \sum_{t=1}^{T} I(\mathcal{C}_{iT}) \varphi_{it}(x_l) W_{jt} \Big|
$$

with $W_{jt} = \frac{1}{nT} \{ Z_{jt} I(|Z_{jt}| \le \tau_{n,T}) - \mathbb{E}[Z_{jt} I(|Z_{jt}| \le \tau_{n,T})] \}$, we obtain

$$
\begin{aligned}
&\mathbb{P}\Big(\max_{1 \le l \le L_{n,T}} \big|Q^{\le}(x_l)\big| > \frac{M}{4} a_{n,T} \Big) \\
&\le \mathbb{P}\Big(\max_{\substack{1 \le i \le n \\ 1 \le l \le L_{n,T}}} \Big| \sum_{j \ne i} \sum_{t=1}^{T} I(\mathcal{C}_{iT}) \varphi_{it}(x_l) W_{jt} \Big| > \frac{M}{4} a_{n,T} \Big) \\
&\le \sum_{i=1}^{n} \sum_{l=1}^{L_{n,T}} \mathbb{P}\Big(\Big| \sum_{j \ne i} \sum_{t=1}^{T} I(\mathcal{C}_{iT}) \varphi_{it}(x_l) W_{jt} \Big| > \frac{M}{4} a_{n,T} \Big).
\end{aligned}
$$

We now bound the probability $\mathbb{P}(|\sum_{j \ne i} \sum_{t=1}^{T} I(\mathcal{C}_{iT}) \varphi_{it}(x) W_{jt}| > M a_{n,T}/4)$ for an arbitrary point x with the help of an exponential inequality. To do so, we rewrite the expression $\sum_{j \ne i} \sum_{t=1}^{T} I(\mathcal{C}_{iT}) \varphi_{it}(x) W_{jt}$. In particular, we split up the inner sum over t

into blocks as follows:

$$\sum_{t=1}^{T} I(\mathcal{C}_{iT})\varphi_{it}(x)W_{jt} = \sum_{s=1}^{q_{n,T}} B_{j,2s-1}(x) + \sum_{s=1}^{q_{n,T}} B_{j,2s}(x)$$

with

$$B_{j,s}(x) = \sum_{t=(s-1)r_{n,T}+1}^{sr_{n,T}} I(\mathcal{C}_{iT})\varphi_{it}(x)W_{jt},$$

where as in Lemma B4, $2q_{n,T}$ is the number of blocks and $r_{n,T} = T/2q_{n,T}$ is the block length. We thus get

$$\begin{aligned} \mathbb{P}\Big(\Big|\sum_{j\neq i}\sum_{t=1}^{T} I(\mathcal{C}_{iT})\varphi_{it}(x)W_{jt}\Big| > \frac{M}{4}a_{n,T}\Big) &\leq \mathbb{P}\Big(\Big|\sum_{j\neq i}\sum_{s=1}^{q_{n,T}} B_{j,2s-1}(x)\Big| > \frac{M}{8}a_{n,T}\Big) \\ &\quad + \mathbb{P}\Big(\Big|\sum_{j\neq i}\sum_{s=1}^{q_{n,T}} B_{j,2s}(x)\Big| > \frac{M}{8}a_{n,T}\Big). \end{aligned}$$

In what follows, we restrict attention to the first term on the right-hand side. The second one can be analyzed by similar arguments.

To indicate the dependence of the block $B_{j,s}(x)$ on the i-th time series $\{X_{it}\}_{t=1}^{T}$, we use the notation $B_{j,s}(x) = B_{j,s}(x, \{X_{it}\}_{t=1}^{T})$. Moreover, we employ the shorthand $\overline{B}_{j,s}(x) = B_{j,s}(x, \{x_{it}\}_{t=1}^{T})$ to denote the s-th block for a fixed realization $\{x_{it}\}_{t=1}^{T}$ of $\{X_{it}\}_{t=1}^{T}$. With this notation at hand, we write

$$\begin{aligned} &\mathbb{P}\Big(\Big|\sum_{j\neq i}\sum_{s=1}^{q_{n,T}} B_{j,2s-1}(x)\Big| > \frac{M}{8}a_{n,T}\Big) \\ &\qquad = \mathbb{E}\Big[\mathbb{P}\Big(\Big|\sum_{j\neq i}\sum_{s=1}^{q_{n,T}} B_{j,2s-1}(x)\Big| > \frac{M}{8}a_{n,T}\Big|\{X_{it}\}_{t=1}^{T}\Big)\Big] \end{aligned}$$

and bound the term

$$\begin{aligned} &\mathbb{P}\Big(\Big|\sum_{j\neq i}\sum_{s=1}^{q_{n,T}} B_{j,2s-1}(x)\Big| > \frac{M}{8}a_{n,T}\Big|\{X_{it}\}_{t=1}^{T} = \{x_{it}\}_{t=1}^{T}\Big) \\ &\qquad\qquad = \mathbb{P}\Big(\Big|\sum_{j\neq i}\sum_{s=1}^{q_{n,T}} \overline{B}_{j,2s-1}(x)\Big| > \frac{M}{8}a_{n,T}\Big) \end{aligned}$$

for an arbitrary but fixed realization $\{x_{it}\}_{t=1}^{T}$. By Bradley's lemma, we can construct a sequence of random variables $\overline{B}^{*}_{j,1}(x), \overline{B}^{*}_{j,3}(x), \ldots$ such that (i) $\overline{B}^{*}_{j,1}(x)$, $\overline{B}^{*}_{j,3}(x)$, $\ldots$ are independent, (ii) $\overline{B}^{*}_{j,2s-1}(x)$ has the same distribution as $\overline{B}_{j,2s-1}(x)$, and (iii) for $0 < \mu \leq$

$\|\overline{B}_{j,2s-1}(x)\|_\infty$,

$$\mathbb{P}\big(|\overline{B}^*_{j,2s-1}(x) - \overline{B}_{j,2s-1}(x)| > \mu\big) \le 18\Big(\frac{\|\overline{B}_{j,2s-1}(x)\|_\infty}{\mu}\Big)^{1/2} \alpha(r_{n,T}). \tag{S2}$$

This allows us to write

$$\mathbb{P}\Big(\Big|\sum_{j\neq i}\sum_{s=1}^{q_{n,T}} \overline{B}_{j,2s-1}(x)\Big| > \frac{M}{8} a_{n,T}\Big) \le P_1 + P_2$$

with

$$\begin{aligned} P_1 &= \mathbb{P}\Big(\Big|\sum_{j\neq i}\sum_{s=1}^{q_{n,T}} \overline{B}^*_{j,2s-1}(x)\Big| > \frac{M}{16} a_{n,T}\Big) \\ P_2 &= \mathbb{P}\Big(\Big|\sum_{j\neq i}\sum_{s=1}^{q_{n,T}} \big(\overline{B}_{j,2s-1}(x) - \overline{B}^*_{j,2s-1}(x)\big)\Big| > \frac{M}{16} a_{n,T}\Big). \end{aligned}$$

First consider P_1. It holds that

$$|\overline{B}_{j,2s-1}(x)| \le \frac{C\tau_{n,T} r_{n,T} b_{n,T}}{nTh} \le \frac{C\tau_{n,T} r_{n,T} (b_{n,T}/\sqrt{h})}{nTh} \le \frac{C\tau_{n,T} r_{n,T}}{nTh(nT)^{\eta/2}} =: \overline{C}_{n,T}.$$

Choosing $\lambda_{n,T} = \overline{C}_{n,T}^{-1}/2$ and applying Markov's inequality, the same arguments as in Lemma B3 yield that

$$P_1 \le 2\exp\Big(-\frac{M}{16}\lambda_{n,T} a_{n,T} + \lambda_{n,T}^2 \sum_{j\neq i}\sum_{s=1}^{q_{n,T}} \mathbb{E}[\overline{B}^*_{j,2s-1}(x)^2]\Big).$$

Noting that

$$\begin{aligned} \sum_{j\neq i}\sum_{s=1}^{q_{n,T}} \mathbb{E}[\overline{B}^*_{j,2s-1}(x)^2] &= \sum_{j\neq i}\sum_{s=1}^{q_{n,T}} \mathbb{E}[\overline{B}_{j,2s-1}(x)^2] \\ &= \sum_{j\neq i}\sum_{s=1}^{q_{n,T}} \mathbb{E}[B_{j,2s-1}(x)^2 | \{X_{it}\}_{t=1}^T = \{x_{it}\}_{t=1}^T] \\ &\le \sum_{j\neq i}\sum_{s,t=1}^{T} I(\mathcal{C}_{iT})|\varphi_{is}(x)\varphi_{it}(x)|\big|\mathbb{E}[W_{js}W_{jt}]\big| \\ &\le C b_{n,T}^2 \sum_{j\neq i}\sum_{s,t=1}^{T} K_h(x_{is}-x)K_h(x_{it}-x)\big|\mathbb{E}[W_{js}W_{jt}]\big| \\ &\le \frac{C b_{n,T}^2}{h^2}\sum_{j\neq i}\Big(\sum_{t=1}^{T} |\mathbb{E}[W_{jt}^2]| + 2\sum_{l=1}^{T-1}\sum_{t=1}^{T-l} |\mathbb{E}[W_{jt}W_{jt+l}]|\Big) \end{aligned}$$

$$\leq \frac{C}{nTh(nT)^{\eta}},$$

we arrive at

$$P_1 \leq C \exp\Big(-\frac{M}{16}\lambda_{n,T} a_{n,T} + C\frac{\lambda_{n,T}^2}{nTh(nT)^{\eta}}\Big).$$

Moreover, choosing

$$r_{n,T} = \sqrt{\frac{nTh}{\tau_{n,T}^2 \log nT}},$$

we obtain that $\frac{\lambda_{n,T}^2}{nTh(nT)^{\eta}} = \log(nT)$ and $\lambda_{n,T} a_{n,T} = \log(nT)$. As a result,

$$P_1 \leq C \exp\Big(\Big[C - \frac{M}{16}\Big] \log nT\Big) \leq C(nT)^{-p},$$

where p can be made arbitrarily large by choosing M large enough. We next turn to P_2. Using (S2), we obtain that

$$\begin{aligned} P_2 &\leq \sum_{j \neq i} \sum_{s=1}^{q_{n,T}} \mathbb{P}\Big(\big|\overline{B}_{j,2s-1}(x) - \overline{B}_{j,2s-1}^{*}(x)\big| > \frac{M a_{n,T}}{16 n q_{n,T}}\Big) \\ &\leq C \sum_{j \neq i} \sum_{s=1}^{q_{n,T}} \Big(\frac{\overline{C}_{n,T}}{a_{n,T}/n q_{n,T}}\Big)^{1/2} \alpha(r_{n,T}) \leq C(nT)^{-q}, \end{aligned}$$

where q can be chosen arbitrarily large as the α-coefficients decay exponentially fast.

Putting everything together, we arrive at

$$\begin{aligned} \mathbb{P}\Big(\max_{1 \leq l \leq L_{n,T}} \big|Q^{\leq}(x_l)\big| > \frac{M}{4} a_{n,T}\Big) &\leq \sum_{i=1}^{n} \sum_{l=1}^{L_{n,T}} \mathbb{P}\Big(\Big| \sum_{j \neq i} \sum_{t=1}^{T} I(\mathcal{C}_{iT}) \varphi_{it}(x_l) W_{jt}\Big| > \frac{M}{4} a_{n,T}\Big) \\ &\leq C n L_{n,T} \big[(nT)^{-p} + (nT)^{-q}\big]. \end{aligned}$$

If we choose the exponents p and q sufficiently large, then the right-hand side converges to zero at an arbitrarily fast polynomial rate. This completes the proof. □