Multiscale Testing for Equality of Nonparametric Trend Curves

Marina Khismatullina[1] University of Bonn

Michael Vogt[2] University of Bonn

We develop multiscale methods to test qualitative hypotheses about nonparametric time trends. In many applications, practitioners are interested in whether the observed time series has a time trend at all, that is, whether the trend function is non-constant. Moreover, they would like to get further information about the shape of the trend function. Among other things, they would like to know in which time regions there is an upward/downward movement in the trend. When multiple time series are observed, another important question is whether the observed time series all have the same time trend. We design multiscale tests to formally approach these questions. We derive asymptotic theory for the proposed tests and investigate their finite sample performance by means of simulations. In addition, we illustrate the methods by two applications to temperature data.

Key words: Multiscale statistics; nonparametric regression; time series errors; shape constraints; strong approximations; anti-concentration bounds.
AMS 2010 subject classifications: 62E20; 62G10; 62G20; 62M10.

1 State of the art and preliminary work

The comparison of nonparametric curves is a classic topic in econometrics and statistics. Depending on the specific application, the curves of interest are densities, distribution functions, time trends or regression curves. The problem of testing for equality of densities has been studied in Mammen (1992), Anderson et al. (1994) and Li et al. (2009) among others. Tests for equality of distribution functions can be found for example in Kiefer (1959), Anderson (1962) and Finner and Gontscharuk (2018). Tests for equality of trend or regression curves have been developed in Härdle and Marron (1990), Hall and Hart (1990), Delgado (1993), Degras et al. (2012), Zhang et al. (2012) and Hidalgo and Lee (2014) among many others. In the proposed project, we focus on the comparison of nonparametric trend curves.

The statistical problem of comparing trends has a wide range of applications in economics, finance and other fields such as climatology and biology. In economics, one may wish is to compare trends in real gross domestic product (GDP) across different countries (cp. Grier and Tullock, 1989). Another example concerns the dynamics of

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

long-term interest rates. To better understand these dynamics, researchers aim to compare the yields of US Treasury bills at different maturities over time (cp. Park et al., 2009). In finance, it is of interest to compare the volatility trends of different stocks (cp. Nyblom and Harvey, 2000). Finally, in climatology, researchers are interested in comparing the trending behaviour of temperature time series across different spatial locations (cp. Karoly and Wu, 2005).

Classically, time trends are modelled stochastically in econometrics; see e.g. Stock and Watson (1988). Recently, however, there has been a growing interest in econometric models with deterministic time trends; see Cai (2007), Atak et al. (2011), Robinson (2012) and Chen et al. (2012) among others. Non- and semiparametric trend modelling has attracted particular interest in a panel data context. Li et al. (2010), Atak et al. (2011), Robinson (2012) and Chen et al. (2012) considered panel models where the observed time series have a common time trend. In many applications, however, the assumption of a common time trend is quite harsh. In particular when the number of observed time series is large, it is quite natural to suppose that the time trend may differ across time series. More flexible panel settings with heterogeneous trends have been studied, for example, in Zhang et al. (2012) and Hidalgo and Lee (2014).

In what follows, we consider a general panel framework with heterogeneous trends which is useful for a number of economic and financial applications: Suppose we observe a panel of n time series $\mathcal{Z}_i = \{(Y_{it}, \mathbf{X}_{it}) : 1 \leq t \leq T\}$ for $1 \leq i \leq n$, where Y_{it} are real-valued random variables and $\mathbf{X}_{it} = (X_{it,1}, \ldots, X_{it,d})^\top$ are d-dimensional random vectors. Each time series $\mathcal{Z}_i$ is modelled by the equation

$$Y_{it} = m_i\Big(\frac{t}{T}\Big) + \boldsymbol{\beta}_i^\top \mathbf{X}_{it} + \alpha_i + \varepsilon_{it} \tag{1.1}$$

for $1 \leq t \leq T$, where $m_i : [0,1] \to \mathbb{R}$ is a nonparametric (deterministic) trend function, $\mathbf{X}_{it}$ is a vector of regressors or controls and $\boldsymbol{\beta}_i$ is the corresponding parameter vector. Moreover, α_i are so-called fixed effect error terms and ε_{it} are standard regression errors with $\mathbb{E}[\varepsilon_{it}|\mathbf{X}_{it}] = 0$ for all t. Model (1.1) nests a number of panel models which have recently been considered in the literature. Special cases of model (1.1) with a nonparametric trend specification are for example considered in Atak et al. (2011), Zhang et al. (2012) and Hidalgo and Lee (2014). Versions of model (1.1) with a parametric trend are studied in Vogelsang and Franses (2005), Sun (2011) and Xu (2012) among others. Within the general framework of model (1.1), we can formulate a number of interesting statistical questions concerning the set of trend functions $\{m_i : 1 \leq i \leq n\}$.

(a) Testing for equality of nonparametric trend curves

In many application contexts, an important question is whether the time trends m_i in model (1.1) are all the same. Put differently, the question is whether the observed time series have a common trend. This question can formally be addressed by a statistical

test of the null hypothesis

$$H_0 : \text{There exists a function } m : [0,1] \to \mathbb{R} \text{ such that } m_i = m \text{ for all } 1 \leq i \leq n.$$

A closely related question is whether all time trends have the same parametric form. To formulate the corresponding null hypothesis, let $m(\theta, \cdot) : [0,1] \to \mathbb{R}$ be a function which is known up to the finite-dimensional parameter $\theta \in \Theta$, where Θ denotes the parameter space. The null hypothesis of interest now reads as follows:

$$H_{0,\text{para}} : \text{ There exists } \theta \in \Theta \text{ such that } m_i(\cdot) = m(\theta, \cdot) \text{ for all } 1 \leq i \leq n.$$

If $m(\theta, w) = a + bw$ with $\theta = (a, b)$, for example, then H_0 is the hypothesis that all trends m_i are linear with the same intercept a and slope b. A somewhat simpler but yet important hypothesis is given by

$$H_{0,\text{const}} : m_i \equiv 0 \text{ for all } 1 \leq i \leq n.$$

Under this hypothesis, there is no time trend at all in the observed time series. Put differently, all the time trends m_i are constant. (Note that under the normalization constraint $\int_0^1 m_i(w)dw = 0$, m_i must be equal to zero if it is a constant function.) A major goal of our project is to develop new tests for the hypotheses H_0, $H_{0,\text{para}}$ and $H_{0,\text{const}}$ in model (1.1). In order to keep the exposition as clear as possible, we focus attention on the hypothesis H_0 in what follows. Tests of $H_{0,\text{para}}$, $H_{0,\text{const}}$ and related hypotheses have for example been studied in Lyubchich and Gel (2016) and Chen and Wu (2018).

In recent years, a number of different approaches have been developed to test the hypothesis H_0. Degras et al. (2012) consider the problem of testing H_0 within the model framework

$$Y_{it} = m_i\Big(\frac{t}{T}\Big) + \alpha_i + \varepsilon_{it} \qquad (1 \leq t \leq T,\, 1 \leq i \leq n), \tag{1.2}$$

where $\mathbb{E}[\varepsilon_{it}] = 0$ for all i and t and the terms α_i are assumed to be deterministic. Obviously, (1.2) is a special case of (1.1) which does not include additional regressors. Degras et al. (2012) construct an L_2-type statistic to test H_0. The statistic is based on the difference between estimators of the trend with and without imposing H_0. Let $\hat{m}_{i,h}$ be the estimator of m_i and $\hat{m}_h$ the estimator of the common trend m under H_0, where h denotes the bandwidth parameter. With these estimators, the authors define the statistic

$$\Delta_{n,T} = \sum_{i=1}^{n} \int_0^1 \big(\hat{m}_{i,h}(u) - \hat{m}_h(u)\big)^2 du, \tag{1.3}$$

which measures the L_2-distance between $\hat{m}_{i,h}$ and $\hat{m}_h$. In the theoretical part of their

paper, they derive the limit distribution of $\Delta_{n,T}$. Chen and Wu (2018) develop theory for test statistics closely related to those from Degras et al. (2012), but under more general conditions on the error terms.

Zhang et al. (2012) investigate the problem of testing the hypothesis H_0 in a slightly restricted version of model (1.1), where $\beta_i = \beta$ for all i. The regression coefficients β_i are thus assumed to be homogeneous in their setting. They construct a residual-based test statistic as follows: First, they obtain profile least squares estimators $\hat{\beta}$ and $\hat{m}_h(t/T)$ of the parameter vector β and the common trend m under H_0, where h denotes the bandwidth. With these estimators, they compute the residuals $\hat{u}_{it} = Y_{it} - \hat{\beta}^T X_{it} - \hat{m}_h(t/T)$. These residuals are shown to have the form $\hat{u}_{it} = \Delta_i(t/T) + \eta_{it}$, where Δ_i is a deterministic function with the property that $\Delta_i \equiv 0$ under H_0 and η_{it} denotes the error term. Testing H_0 is thus equivalent to testing the hypothesis $H_0' : \Delta_i \equiv 0$ for all $1 \leq i \leq n$. The authors construct a test statistic for the hypothesis H_0' on the basis of nonparametric kernel estimators of the functions Δ_i and derive its limit distribution.

The tests of Zhang et al. (2012), Degras et al. (2012) and Chen and Wu (2018) are based on nonparametric estimators of the trend functions m_i that depend on one or several bandwidth parameters. Unfortunately, it is far from clear how to choose these bandwidths in an appropriate way. This is a general problem concerning essentially all tests based on nonparametric curve estimators. There are of course many theoretical results on optimal bandwidth choice for estimation purposes. However, the optimal bandwidth for curve estimation is usually not optimal for testing. Optimal bandwidth choice for tests is indeed an open problem, and only little theory for simple cases is available (cp. Gao and Gijbels, 2008). Since tests based on nonparametric curve estimators are commonly quite sensitive to the choice of bandwidth and theory for optimal bandwidth selection is not available, it appears preferable to work with bandwidth-free tests.

A classical way to obtain a bandwidth-free test of the hypothesis H_0 is to use CUSUM-type statistics which are based on partial sum processes. This approach is taken in Hidalgo and Lee (2014). A more modern approach to obtain a bandwidth-free test is to employ multiscale methods. These methods avoid the need to choose a bandwidth by considering a large collection of bandwidths simultaneously. More specifically, the basic idea is as follows: Let S_h be a test statistic for the null hypothesis of interest, which depends on the bandwidth h. Rather than considering only a single statistic S_h for a specific bandwidth h, a multiscale approach simultaneously considers a whole family of statistics $\{S_h : h \in \mathcal{H}\}$, where $\mathcal{H}$ is a set of bandwidth values. The multiscale test then proceeds as follows: For each bandwidth or scale h, one checks whether $S_h > q_h(\alpha)$, where $q_h(\alpha)$ is a bandwidth-dependent critical value (for given significance level α). The multiscale test rejects if $S_h > q_h(\alpha)$ for at least one scale h. The main theoretical difficulty in this approach is of course to derive appropri-

ate critical values $q_h(\alpha)$. Specifically, the critical values $q_h(\alpha)$ need to be determined such that the multiscale test has the correct (asymptotic) level, that is, such that $\mathbb{P}(S_h > q_h(\alpha) \text{ for some } h \in \mathcal{H}) = (1-\alpha) + o(1)$.

Multiscale methods have been developed for a variety of different test problems in recent years. Chaudhuri and Marron (1999, 2000) introduced the so-called SiZer method which has been extended in various directions; see for example Hannig and Marron (2006) and Rondonotti et al. (2007). Horowitz and Spokoiny (2001) proposed a multiscale test for the parametric form of a regression function. Dümbgen and Spokoiny (2001) constructed a multiscale approach which works with additively corrected supremum statistics. This general approach has been very influential in recent years and has been further developed in numerous ways; see for example Dümbgen (2002), Rohde (2008) and Proksch et al. (2018) for multiscale methods in the regression context and Dümbgen and Walther (2008), Rufibach and Walther (2010), Schmidt-Hieber et al. (2013) and Eckle et al. (2017) for methods in the context of density estimation. Importantly, all of these studies are restricted to the case of independent data. It turns out that it is highly non-trivial to extend the multiscale approach of Dümbgen and Spokoiny (2001) to the case of dependent data. A first step into this direction has recently been made in Khismatullina and Vogt (2018). They developed multiscale methods to test for local increases/decreases of the nonparametric trend function m in the univariate time series model $Y_t = m(t/T) + \varepsilon_t$.

To the best of our knowledge, multiscale tests of the hypotheses H_0, $H_{0,\text{para}}$ and $H_{0,\text{const}}$ in model (1.1) are not available in the literature. The only exception is Park et al. (2009) who developed SiZer methods for the comparison of nonparametric trend curves in a strongly simplified version of model (1.1). Their analysis, however, is mainly methodological and not fully backed up by theory. Indeed, theory has only been derived for the special case $n = 2$, that is, for the case that only two time series are observed.

(b) Clustering of nonparametric trend curves

Consider the situation that the null hypothesis $H_0 : m_1 = \ldots = m_n$ is violated in the general panel data model (1.1). Even though some of the trend functions m_i are different in this case, there may still be groups of time series with the same time trend. Formally, a group stucture can be defined as follows within the framework of model (1.1): There exist sets or groups of time series $G_1, \ldots, G_{K_0}$ with $\{1, \ldots, n\} = \dot{\bigcup}_{k=1}^{K_0} G_k$ such that for each $1 \leq k \leq K_0$,

$$m_i = m_j \quad \text{for all } i, j \in G_k. \tag{1.4}$$

According to (1.4), the time series of a given group G_k all have the same time trend. In many applications, it is very natural to suppose that there is such a group structure in the data. An interesting statistical problem which we aim to investigate in our project

is how to estimate the unknown groups $G_1, \ldots, G_{K_0}$ and their unknown number K_0 from the data.

Several approaches to this problem have been proposed in the context of models closely related to (1.1). Degras et al. (2012) used a repeated testing procedure based on L_2-type test statistics of the form (1.3) in order to estimate the unknown group structure in model (1.2). Zhang (2013) developed a clustering method within the same model framework which makes use of an extended Bayesian information criterion. Vogt and Linton (2017) constructed a thresholding method to estimate the unknown group structure in the panel model $Y_{it} = m_i(X_{it}) + u_{it}$, where X_{it} are random regressors and u_{it} are general error terms that may include fixed effects. Their approach can also be adapted to the case of fixed regressors $X_{it} = t/T$. As an alternative to a group structure, factor-type structures have been imposed on the trend and regression functions in panel models. Such factor-type structures are studied in Kneip et al. (2012), Boneva et al. (2015) and Boneva et al. (2016) among others.

The problem of estimating the unknown groups $G_1, \ldots, G_{K_0}$ and their unknown number K_0 in model (1.1) has close connections to functional data clustering. There, the aim is to cluster smooth random curves that are functions of (rescaled) time and that are observed with or without noise. A number of different clustering approaches have been proposed in the context of functional data models; see for example Abraham et al. (2003), Tarpey and Kinateder (2003) and Tarpey (2007) for procedures based on k-means clustering, James and Sugar (2003) and Chiou and Li (2007) for model-based clustering approaches and Jacques and Preda (2014) for a recent survey.

The problem of finding the unknown group structure in model (1.1) is also closely related to a developing literature in econometrics which aims to identify unknown group structures in parametric panel regression models. In its simplest form, the panel regression model under consideration is given by the equation $Y_{it} = \beta_i^\top X_{it} + u_{it}$ for $1 \leq t \leq T$ and $1 \leq i \leq n$, where the coefficient vectors β_i are allowed to vary across individuals i and the error terms u_{it} may include fixed effects. Similar to the trend functions in model (1.1), the coefficients β_i are assumed to belong to a number of groups: there are K_0 groups $G_1, \ldots, G_{K_0}$ such that $\beta_i = \beta_j$ for all $i, j \in G_k$ and all $1 \leq k \leq K_0$. The problem of estimating the unknown groups and their unknown number has been studied in different versions of this modelling framework; cp. Su et al. (2016), Su and Ju (2018) and Wang et al. (2018) among others. Bonhomme and Manresa (2015) considered a related model where the group structure is not imposed on the regression coefficients but rather on some unobserved time-varying fixed effect components of the panel model.

Virtually all the proposed procedures to cluster nonparametric curves in panel and functional data models related to (1.1) depend on a number of bandwidth or smoothing parameters required to estimate the nonparametric functions m_i. In general, nonparametric curve estimators are strongly affected by the chosen bandwidth parameters. A clustering procedure which is based on such estimators can be expected to be strongly

influenced by the choice of bandwidths as well. Moreover, as in the context of statistical testing, there is no theory available on how to pick the bandwidths optimally for the clustering problem. Hence, as in the context of testing, it is desirable to construct a clustering procedure which is free of bandwidth or smoothing parameters that need to be selected.

One way to obtain a clustering method which does not require to select any bandwidth parameter is to use multiscale methods. This approach has recently been taken in Vogt and Linton (2018). They develop a clustering approach in the context of the panel model $Y_{it} = m_i(X_{it}) + u_{it}$, where X_{it} are random regressors and u_{it} are general error terms that may include fixed effects. Imposing the same group structure as in (1.4) on their model, they construct estimators of the unknown groups and their unknown number as follows: In a first step, they develop bandwidth-free multiscale statistics $\hat{d}_{ij}$ which measure the distance between pairs of functions m_i and m_j. To construct them, they make use of the multiscale testing methods described in part (a) of this section. In a second step, the statistics $\hat{d}_{ij}$ are employed as dissimilarity measures in a hierarchical clustering algorithm.

2 The model

Before we proceed any further, we need to introduce some notation used throughout the paper. For a vector $\mathbf{v} = (v_1, \dots, v_m) \in \mathbb{R}^m$, we write $|\mathbf{v}| = \big(\sum_{i=1}^m v_i^2\big)^{1/2}$. For a random vector $\mathbf{V}$, we define it's $\mathcal{L}^q, q > 1$ norm as $||\mathbf{V}||_q = \big(\mathbb{E}|\mathbf{V}|^q\big)^{1/q}$. For a patricular case $q = 2$, we write $||\mathbf{V}|| := ||\mathbf{V}||_2$.

Following Wu (2005), we define the *physical dependence measure* for the process $\mathbf{L}(\mathcal{F}_t)$ as the following:

$$\delta_q(\mathbf{L}, t) = ||\mathbf{L}(\mathcal{F}_t) - \mathbf{L}(\mathcal{F}_t')||_q,$$

where $\mathcal{F}_t = (\dots, \epsilon_{-1}, \epsilon_0, \epsilon_1, \dots, \epsilon_{t-1}, \epsilon_t)$ and $\mathcal{F}_t' = (\dots, \epsilon_{-1}, \epsilon_0', \epsilon_1, \dots, \epsilon_{t-1}, \epsilon_t)$ is a coupled process of $\mathcal{F}_t$ with ϵ_0' being an i.i.d. copy of ϵ_0.

The model setting is as follows. We observe a panel of n time series $\mathcal{Z}_i = \{(Y_{it}, \mathbf{X}_{it}) : 1 \leq t \leq T\}$ of length T for $1 \leq i \leq n$. Each time series $\mathcal{Z}_i$ satisfies the model equation

$$Y_{it} = \boldsymbol{\beta}_i^\top \mathbf{X}_{it} + m_i\Big(\frac{t}{T}\Big) + \alpha_i + \varepsilon_{it} \tag{2.1}$$

for $1 \leq t \leq T$, where $\boldsymbol{\beta}_i$ is a $d \times 1$ vector of unknown parameters, $\mathbf{X}_{it}$ is a $d \times 1$ vector of individual covariates, m_i is an unknown nonparametric trend function defined on $[0, 1]$, α_i is a (deterministic or random) intercept term and $\mathcal{E}_i = \{\varepsilon_{it} : 1 \leq t \leq T\}$ is a zero-mean stationary error process. As usual in nonparametric regression, the trend functions m_i in model (2.1) depend on rescaled time t/T rather than on real time t; cp. Robinson (1989), Dahlhaus (1997) and Vogt and Linton (2014) for the

use and some discussion of the rescaled time argument. The functions m_i are only identified up to an additive constant in model (2.1): One can reformulate the model as $Y_{it} = [m_i(t/T) + c_i] + \boldsymbol{\beta}_i^\top \mathbf{X}_{it} + [\alpha_i - c_i] + \varepsilon_{it}$, that is, one can freely shift additive constants c_i between the trend $m_i(t/T)$ and the error component α_i. In order to obtain identification, one may impose different normalization constraints on the trends m_i. One possibility is to normalize them such that $\int_0^1 m_i(u)du = 0$ for all i. In what follows, we take for granted that the trends m_i satisfy this constraint. The term α_i can also be regarded as an additional error component. In the econometrics literature, it is commonly called a fixed effect error term. It can be interpreted as capturing unobserved characteristics of the time series $\mathcal{Z}_i$ which remain constant over time. We allow the error terms α_i to be dependent across i in an arbitrary way. Hence, by including them in model equation (2.1), we allow the n time series $\mathcal{Z}_i$ in our panel to be correlated with each other. Whereas the terms α_i may be correlated, the error processes $\mathcal{E}_i$ are assumed to be independent across i. In addition, each process $\mathcal{E}_i$ is supposed to satisfy the following conditions:

(C1) For each i the variables ε_{it} allow for the representation $\varepsilon_{it} = G_i(\ldots, \eta_{it-1}, \eta_{it})$, where η_{it} are i.i.d. random variables across t and $G_i : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}$ is a measurable function. Denote $\mathcal{J}_{it} = (\ldots, \eta_{it-2}, \eta_{it-1}, \eta_{it})$.

(C2) For all i it holds that $\mathbb{E}[\varepsilon_{it}] = 0$ and $\|\varepsilon_{it}\|_q < \infty$ for some $q > 4$.

Following Wu (2005), we impose conditions on the dependence structure of the error processes $\mathcal{E}_i$ in terms of the physical dependence measure $\delta_q(G_i, t)$. In particular, we assume the following:

(C3) Define $\Theta_{i,t,q} = \sum_{s \geq t} \delta_q(G_i, s)$ for $t \geq 0$. For each i it holds that $\Theta_{i,t,q} = O(t^{-\tau_q}(\log t)^{-A})$, where $A > \frac{2}{3}(1/q+1+\tau_q)$ and $\tau_q = \{q^2-4+(q-2)\sqrt{q^2+20q+4}\}/8q$.

The conditions (C1)–(C3) are fulfilled by a wide range of stationary processes $\mathcal{E}_i$. For a detailed discussion of these properties, see Khismatullina and Vogt (2018).
Finally note that throughout the paper, we restrict attention to the case where the number of time series n in model (2.1) is fixed. Extending our theoretical results to the case where n slowly grows with the sample size T is a possible topic for further research.

3 Testing for equality of time trends

In this section, we adapt the multiscale method developed in Khismatullina and Vogt (2018) to the problem of comparison of the trend curves m_i in model (2.1). As we will see, the proposed multiscale method does not only allow to test whether the null hypothesis is violated. It also provides information on where violations occur. More

specifically, it allows to identify, with a pre-specified confidence, (i) trend functions which are different from each other and (ii) time intervals where these trend functions differ.

3.1 Construction of the test statistic

In what follows, we describe the construction of the test statistic that adresses the question of comparing different trend curves. More specifically, we test the null hypothesis $H_0 : m_1 = m_2 = \ldots = m_n$ in model (2.1). We assume that all the trend functions $m_i(\cdot)$ are continuosly differentiable on $[0, 1]$.

It is obvious that if α_i and $\boldsymbol{\beta}_i$ are known, the problem of testing for the common time trend would be greatly simplified. That is, we would test $H_0 : m_1 = m_2 = \ldots = m_n$ in the model

$$\begin{aligned} Y_{it} - \alpha_i - \boldsymbol{\beta}_i^\top \mathbf{X}_{it} =: Y_{it}^\circ \\ = m_i\Big(\frac{t}{T}\Big) + \varepsilon_{it}, \end{aligned}$$

which is a standard nonparametric regression equation. The variables Y_{it}° are not observed since the intercept α_i and the coefficients $\boldsymbol{\beta}_i$ are not known. Given appropriate estimators $\widehat{\boldsymbol{\beta}}_i$ and $\widehat{\alpha}_i$, we can then consider

$$\widehat{Y}_{it} := Y_{it} - \widehat{\alpha}_i - \widehat{\boldsymbol{\beta}}_i^\top \mathbf{X}_{it} = (\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)^\top \mathbf{X}_{it} + m_i\Big(\frac{t}{T}\Big) + \big(\alpha_i - \widehat{\alpha}_i\big) + \varepsilon_{it}.$$

Then our unobserved variables Y_{it}° can be approximated by $\widehat{Y}_{it}$ and we compute our test statistic based on $\widehat{Y}_{it}$. In what follows, we assume that the such estimators with the property that $\widehat{\alpha}_i - \alpha_i = O_P(T^{-1/2})$ and $\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i = O_P(T^{-1/2})$ are given. Since $\frac{1}{T}\sum_{t=1}^T \mathbf{X}_{it} = O_P(1)$ by Chebyshev's inequality, the unobserved variable $Y_{it}^\circ := Y_{it} - \boldsymbol{\beta}_i^\top \mathbf{X}_{it} - \alpha_i = m_i(t/T) + \varepsilon_{it}$ can be well approximated by $\widehat{Y}_{it} = Y_{it} - \widehat{\alpha}_i - \widehat{\boldsymbol{\beta}}_i^\top \mathbf{X}_{it} = Y_{it}^\circ + O_P(T^{-1/2})$. Details on one of the possible ways to construct $\widehat{\alpha}_i$ and $\widehat{\boldsymbol{\beta}}_i^\top$ are deferred to Section 3.4.

We further let $\widehat{\sigma}_i^2$ be an estimator of the long-run error variance $\sigma_i^2 = \sum_{\ell=-\infty}^{\infty} \mathrm{Cov}(\varepsilon_{i0}, \varepsilon_{i\ell})$ which is computed from the constructed sample $\{\widehat{Y}_{it} : 1 \le t \le T\}$. We thus regard $\widehat{\sigma}_i^2 = \widehat{\sigma}_i^2(\widehat{Y}_{i1}, \ldots, \widehat{Y}_{iT})$ as a function of the variables $\widehat{Y}_{it}$ for $1 \le t \le T$. Throughout the section, we assume that $\widehat{\sigma}_i^2 = \sigma_i^2 + o_p(\rho_T)$ with $\rho_T = o(1/\log T)$. Details on how to construct estimators of σ_i^2 are deferred to Section **??**.

We are now ready to introduce the multiscale statistic for testing the hypothesis $H_0 : m_1 = m_2 = \ldots = m_n$. For any pair of time series i and j, we define the kernel averages

$$\widehat{\psi}_{ij,T}(u, h) = \sum_{t=1}^{T} w_{t,T}(u, h)(\widehat{Y}_{it} - \widehat{Y}_{jt}),$$

where $w_{t,T}(u,h)$ are the local linear kernel weights calculated by the following formula.

$$w_{t,T}(u,h) = \frac{\Lambda_{t,T}(u,h)}{\{\sum_{t=1}^{T} \Lambda_{t,T}(u,h)^2\}^{1/2}}, \tag{3.1}$$

where

$$\Lambda_{t,T}(u,h) = K\Big(\frac{\frac{t}{T}-u}{h}\Big)\Big[S_{T,2}(u,h) - \Big(\frac{\frac{t}{T}-u}{h}\Big)S_{T,1}(u,h)\Big],$$

$S_{T,\ell}(u,h) = (Th)^{-1}\sum_{t=1}^{T} K(\frac{\frac{t}{T}-u}{h})(\frac{\frac{t}{T}-u}{h})^{\ell}$ for $\ell = 0,1,2$ and K is a kernel function with the following properties:

(C4) The kernel K is non-negative, symmetric about zero and integrates to one. Moreover, it has compact support $[-1,1]$ and is Lipschitz continuous, that is, $|K(v) - K(w)| \leq C|v-w|$ for any $v, w \in \mathbb{R}$ and some constant $C > 0$.

The kernel average $\widehat{\psi}_{ij,T}(u,h)$ can be regarded as measuring the distance between the two trend curves m_i and m_j on the interval $[u-h, u+h]$. We aggregate the kernel averages $\widehat{\psi}_{ij,T}(u,h)$ for all $(u,h) \in \mathcal{G}_T$ by the multiscale statistic

$$\widehat{\Psi}_{ij,T} = \max_{(u,h)\in\mathcal{G}_T} \Big\{\Big|\frac{\widehat{\psi}_{ij,T}(u,h)}{(\widehat{\sigma}_i^2 + \widehat{\sigma}_j^2)^{1/2}}\Big| - \lambda(h)\Big\},$$

where $\lambda(h) = \sqrt{2\log\{1/(2h)\}}$ and the set $\mathcal{G}_T$ has been introduced in Section **??**. The statistic $\widehat{\Psi}_{ij,T}$ can be interpreted as a distance measure between the two curves m_i and m_j. We finally define the multiscale statistic for testing the null hypothesis $H_0 : m_1 = m_2 = \ldots = m_n$ as

$$\widehat{\Psi}_{n,T} = \max_{1\leq i<j\leq n} \widehat{\Psi}_{ij,T},$$

that is, we define it as the maximal distance $\widehat{\Psi}_{ij,T}$ between any pair of curves m_i and m_j with $i \neq j$.

3.2 The test procedure

Let Z_{it} for $1 \leq t \leq T$ and $1 \leq i \leq n$ be independent standard normal random variables and independent standard normal random vectors respectively which are independent of the error terms ε_{it} and the covariates $\mathbf{X}_{it}$. Denote the empirical average of the variables $Z_{i1}, \ldots, Z_{iT}$ by $\bar{Z}_{i,T} = T^{-1}\sum_{t=1}^{T} Z_{it}$. To simplify notation, we write $\bar{Z}_i = \bar{Z}_{i,T}$ in what follows. For each i and j, we introduce the Gaussian statistic

$$\Phi_{ij,T} = \max_{(u,h)\in\mathcal{G}_T} \Big\{\Big|\frac{\phi_{ij,T}(u,h)}{(\widehat{\sigma}_i^2 + \widehat{\sigma}_j^2)^{1/2}}\Big| - \lambda(h)\Big\},$$

where $\phi_{ij,T}(u,h) = \sum_{t=1}^{T} w_{t,T}(u,h)\left\{\widehat{\sigma}_i(Z_{it}-\bar{Z}_i) - \widehat{\sigma}_j(Z_{jt}-\bar{Z}_j)\right\}$. Moreover, we define the statistic

$$\Phi_{n,T} = \max_{1\le i<j\le n} \Phi_{ij,T} \tag{3.2}$$

and denote its $(1-\alpha)$-quantile by $q_{n,T}(\alpha)$. Our multiscale test of the hypothesis $H_0 : m_1 = m_2 = \ldots = m_n$ is defined as follows: For a given significance level $\alpha \in (0,1)$, we reject H_0 if $\widehat{\Psi}_{n,T} > q_{n,T}(\alpha)$.

3.3 Theoretical properties of the test

To start with, we introduce the auxiliary statistic

$$\widehat{\Phi}_{n,T} = \max_{1\le i<j\le n} \widehat{\Phi}_{ij,T}, \tag{3.3}$$

where

$$\widehat{\Phi}_{ij,T} = \max_{(u,h)\in\mathcal{G}_T} \left\{ \left| \frac{\widehat{\phi}_{ij,T}(u,h)}{\{\widehat{\sigma}_i^2 + \widehat{\sigma}_j^2\}^{1/2}} \right| - \lambda(h) \right\}$$

and $\widehat{\phi}_{ij,T}(u,h) = \sum_{t=1}^{T} w_{t,T}(u,h)\big\{(\varepsilon_{it}-\bar{\varepsilon}_i) + \boldsymbol{\beta}_i^\top(\mathbf{X}_{it}-\bar{\mathbf{X}}_i) - (\varepsilon_{jt}-\bar{\varepsilon}_j) - \boldsymbol{\beta}_j^\top(\mathbf{X}_{jt}-\bar{\mathbf{X}}_j)\big\}$ with $\bar{\varepsilon}_i = \bar{\varepsilon}_{i,T} = T^{-1}\sum_{t=1}^{T}\varepsilon_{it}$ and $\bar{\mathbf{X}}_i = \bar{\mathbf{X}}_{i,T} = T^{-1}\sum_{t=1}^{T}\mathbf{X}_{it}$ respectively. Our first theoretical result characterizes the asymptotic behaviour of the statistic $\widehat{\Phi}_{n,T}$.

Theorem 3.1. *Suppose that the error processes $\mathcal{E}_i = \{\varepsilon_{it} : 1 \le t \le T\}$ are independent across i and satisfy (C1)–(C3) for each i. Moreover, let (C4)–**??** be fulfilled and assume that $\widehat{\sigma}_i^2 = \sigma_i^2 + o_p(\rho_T)$ with $\rho_T = o(1/\log T)$ for each i. Then*

$$\mathbb{P}\big(\widehat{\Phi}_{n,T} \le q_{n,T}(\alpha) | \{\mathbf{X}_{it} : 1 \le t \le T, 1 \le i \le n\}\big) = (1-\alpha) + o(1) \text{ a.s.}$$

Theorem 3.1 is the main stepping stone to derive the theoretical properties of our multiscale test. The details are provided in the Appendix. The following proposition characterizes the behaviour of our multiscale test under the null hypothesis and under local alternatives.

3.4 Estimation of the parameters $\boldsymbol{\beta}_i$ and α_i

We now focus on finding an appropriate estimator $\widehat{\boldsymbol{\beta}}_i$ of $\boldsymbol{\beta}_i$. For that purpose, we consider the time series $\{\Delta Y_{it}\}$ of the differences $\Delta Y_{it} = Y_{it} - Y_{it-1}$ for each i. We then have

$$\Delta Y_{it} = Y_{it} - Y_{it-1} = \boldsymbol{\beta}_i^\top \Delta \mathbf{X}_{it} + \left(m_i\Big(\frac{t}{T}\Big) - m_i\Big(\frac{t-1}{T}\Big)\right) + \Delta\varepsilon_{it},$$

where $\Delta \mathbf{X}_{it} = \mathbf{X}_{it} - \mathbf{X}_{it-1}$ and $\Delta \varepsilon_{it} = \varepsilon_{it} - \varepsilon_{it-1}$. Since $m_i(\cdot)$ is Lipschitz, we can use the fact that $\left|m_i\left(\frac{t}{T}\right) - m_i\left(\frac{t-1}{T}\right)\right| = O\left(\frac{1}{T}\right)$ and rewrite

$$\Delta Y_{it} = \boldsymbol{\beta}_i^\top \Delta \mathbf{X}_{it} + \Delta \varepsilon_{it} + O\Big(\frac{1}{T}\Big). \tag{3.4}$$

In particular, for each i we employ the least squares estimation method to estimate $\boldsymbol{\beta}_i$ in (3.4), treating $\Delta \mathbf{X}_{it}$ as the regressors and ΔY_{it} as the response variable. That is, we propose the following differencing estimator:

$$\widehat{\boldsymbol{\beta}}_i = \Big(\sum_{t=1}^{T} \Delta \mathbf{X}_{it} \Delta \mathbf{X}_{it}^\top\Big)^{-1} \sum_{t=1}^{T} \Delta \mathbf{X}_{it} \Delta Y_{it} \tag{3.5}$$

We need the following assumptions on the independent variables $\mathbf{X}_{it}$ for each i:

(C5) The covariates $\mathbf{X}_{it}$ allow for the representation $\mathbf{X}_{it} = \mathbf{H}_i(\mathcal{U}_{it})$, where $\mathcal{U}_{it} = (\ldots, u_{it-1}, u_{it})$ with u_{it} being i.i.d. random variables and $\mathbf{H}_i := (H_{i1}, H_{i2}, \ldots, H_{id})^\top : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}^d$ is a measurable function such that $\mathbf{H}_i(\mathcal{U}_{it})$ is well defined.

(C6) Let N_i be the $d \times d$ matrix with kl-th entry $n_{i,kl} = \mathbb{E}[H_{ik}(\mathcal{U}_{i0}) H_{il}(\mathcal{U}_{i0})]$. We assume that the smallest eigenvalue of N_i is strictly bigger than 0.

(C7) Let $\mathbb{E}[\mathbf{H}_i(\mathcal{U}_{i0})] = \mathbf{0}$ and $||\mathbf{H}_i(\mathcal{U}_{it})||_4 < \infty$.

To be able to prove the main theorems in this section, we need additional assumptions on the relationship between the covariates and the error process.

(C8) $\mathbf{X}_{it}$ (elementwise) and ε_{is} are uncorrelated for each $t, s \in \{1, \ldots, T\}$.

(C9) Let $\zeta_{i,t} = (u_{it}, \eta_{it})^\top$. Define $\mathcal{I}_{i,t} = (\ldots, \zeta_{i,t-1}, \zeta_{i,t})$ and $\mathbf{U}_i(\mathcal{I}_{i,t}) = \mathbf{H}_i(\mathcal{U}_{it}) G_i(\mathcal{J}_{it})$. Then, $\sum_{s=0}^{\infty} \delta_2(\mathbf{U}_i, s) < \infty$.

(C10) $\sum_{s=0}^{\infty} \delta_4(\mathbf{H}_i, s) < \infty$.

Then the asymptotic consistency for this differencing estimator is given by the following theorem:

Theorem 3.2. *Under Assumptions (C1) - (C10), we have*

$$\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i = O_P\Big(\frac{1}{\sqrt{T}}\Big),$$

where $\widehat{\boldsymbol{\beta}}_i$ is the differencing estimator given by (3.5).

Now consider an appropriate estimator $\widehat{\alpha}_i$ for the intercept α_i calculated by

$$\widehat{\alpha}_i = \frac{1}{T}\sum_{t=1}^{T}\left(Y_{it} - \widehat{\boldsymbol{\beta}}_i^\top \mathbf{X}_{it}\right) = \frac{1}{T}\sum_{t=1}^{T}\left(\boldsymbol{\beta}_i^\top \mathbf{X}_{it} - \widehat{\boldsymbol{\beta}}_i^\top \mathbf{X}_{it} + \alpha_i + m_i(t/T) + \varepsilon_{it}\right) = \tag{3.6}$$
$$= \left(\boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_i\right)^\top \frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_{it} + \alpha_i + \frac{1}{T}\sum_{i=1}^{T} m_i(t/T) + \frac{1}{T}\sum_{i=1}^{T}\varepsilon_{it}.$$

Note that $\frac{1}{T}\sum_{i=1}^{T}\varepsilon_{it} = O_P(T^{-1/2})$ and $\frac{1}{T}\sum_{i=1}^{T} m_i(t/T) = O(T^{-1})$ due to Lipschitz continuity of m_i and normalization $\int_0^1 m_i(u)du = 0$. Furthermore, $\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_{it} = O_P(1)$ by Chebyshev's inequality and $\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i = O_P(T^{-1/2})$ by Theorem 3.2. Plugging all these results together in (3.6), we get that $\widehat{\alpha}_i - \alpha_i = O_P(T^{-1/2})$.

3.5 Proof of Theorem 3.1

References

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

Anderson, N. H., Hall, P. and Titterington, D. M. (1994). Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, **50** 41–54.

Anderson, T. W. (1962). On the distribution of the two-sample Cramér-von Mises criterion. *Annals of Mathematical Statistics*, **33** 1148–1159.

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

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

Boneva, L., Linton, O. and Vogt, M. (2016). The effect of fragmentation in trading on market quality in the UK equity market. *Journal of Applied Econometrics*, **31** 192–213.

Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, **83** 1147–1184.

Cai, Z. (2007). Trending time-varying coefficients time series models with serially correlated errors. *Journal of Econometrics*, **136** 163–188.

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

Chaudhuri, P. and Marron, J. S. (2000). Scale space view of curve estimation. *Annals of Statistics*, **28** 408–428.

Chen, J., Gao, J. and Li, D. (2012). Semiparametric trending panel data models with cross-sectional dependence. *Journal of Econometrics*, **171** 71–85.

Chen, L. and Wu, W. B. (2018). Testing for trends in high-dimensional time series. *Forthcoming in Journal of the American Statistical Association.*

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

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *The Annals of Statistics*, **25** 1–37.

Degras, D., Xu, Z., Zhang, T. and Wu, W. B. (2012). Testing for parallelism among trends in multiple time series. *IEEE Transactions on Signal Processing*, **60** 1087–1097.

Delgado, M. A. (1993). Testing the equality of nonparametric regression curves. *Statistics & Probability Letters*, **17** 199–204.

Dümbgen, L. (2002). Application of local rank tests to nonparametric regression. *Journal*

of Nonparametric Statistics, **14** 511–537.

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

Dümbgen, L. and Walther, G. (2008). Multiscale inference about a density. *Annals of Statistics*, **36** 1758–1785.

Eckle, K., Bissantz, N. and Dette, H. (2017). Multiscale inference for multivariate deconvolution. *Electronic Journal of Statistics*, **11** 4179–4219.

Finner, H. and Gontscharuk, V. (2018). Two-sample Kolmogorov-Smirnov-type tests revisited: old and new tests in terms of local levels. *Annals of Statistics*, **46** 3014–3037.

Gao, J. and Gijbels, I. (2008). Bandwidth selection in nonparametric kernel testing. *Journal of the American Statistical Association*, **103** 1584–1594.

Grier, K. B. and Tullock, G. (1989). An empirical analysis of cross-national economic growth, 1951–1980. *Journal of Monetary Economics*, **24** 259–276.

Hall, P. and Hart, J. D. (1990). Bootstrap test for difference between means in nonparametric regression. *Journal of the American Statistical Association*, **85** 1039–1049.

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

Härdle, W. and Marron, J. S. (1990). Semiparametric comparison of regression curves. *Annals of Statistics*, **18** 63–89.

Hidalgo, J. and Lee, J. (2014). A CUSUM test for common trends in large heterogeneous panels. In *Essays in Honor of Peter C. B. Phillips*. Emerald Group Publishing Limited, 303–345.

Horowitz, J. L. and Spokoiny, V. G. (2001). An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. *Econometrica*, **69** 599–631.

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

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

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

Khismatullina, M. and Vogt, M. (2018). Multiscale inference and long-run variance estimation in nonparametric regression with time series errors. *Preprint*.

Kiefer, J. (1959). K-sample analogues of the Kolmogorov-Smirnov and Cramér-v. Mises tests. *Annals of Mathematical Statistics*, **30** 420–447.

Kneip, A., Sickles, R. C. and Song, W. (2012). A new panel data treatment for heterogeneity in time trends. *Econometric Theory*, **28** 590–628.

Li, D., Chen, J. and Gao, J. (2010). Nonparametric time-varying coefficient panel data models with fixed effects. *The Econometrics Journal*, **14** 387–408.

Li, Q., Maasoumi, E. and Racine, J. S. (2009). A nonparametric test for equality of distributions with mixed categorical and continuous data. *Journal of Econometrics*, **148** 186–200.

Lyubchich, V. and Gel, Y. R. (2016). A local factor nonparametric test for trend synchronism in multiple time series. *Journal of Multivariate Analysis*, **150** 91–104.

Mammen, E. (1992). *When does bootstrap work? Asymptotic results and simulations*. New York, Springer.

Nyblom, J. and Harvey, A. (2000). Tests of common stochastic trends. *Econometric Theory*, **16** 176–199.

Park, C., Vaughan, A., Hannig, J. and Kang, K.-H. (2009). SiZer analysis for the comparison of time series. *Journal of Statistical Planning and Inference*, **139** 3974–3988.

Proksch, K., Werner, F. and Munk, A. (2018). Multiscale scanning in inverse problems. *Forthcoming in Annals of Statistics*.

Robinson, P. M. (1989). Nonparametric estimation of time-varying parameters. In *Statistical Analysis and Forecasting of Economic Structural Change*. Springer, 253–264.

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

Rohde, A. (2008). Adaptive goodness-of-fit tests based on signed ranks. *Annals of Statistics*, **36** 1346–1374.

Rondonotti, V., Marron, J. S. and Park, C. (2007). SiZer for time series: a new approach to the analysis of trends. *Electronic Journal of Statistics*, **1** 268–289.

Rufibach, K. and Walther, G. (2010). The block criterion for multiscale inference about a density, with applications to other multiscale problems. *Journal of Computational and Graphical Statistics*, **19** 175–190.

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

Stock, J. H. and Watson, M. W. (1988). Testing for common trends. *Journal of the American Statistical Association*, **83** 1097–1107.

Su, L. and Ju, G. (2018). Identifying latent grouped patterns in panel data models with interactive fixed effects. *Journal of Econometrics*, **206** 554–573.

Su, L., Shi, Z. and Phillips, P. C. (2016). Identifying latent structures in panel data. *Econometrica*, **84** 2215–2264.

Sun, Y. (2011). Robust trend inference with series variance estimator and testing-optimal

smoothing parameter. *Journal of Econometrics*, **164** 345–366.

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

Tarpey, T. and Kinateder, K. K. (2003). Clustering functional data. *Journal of Classification*, **20** 093–114.

Vogelsang, T. J. and Franses, P. H. (2005). Testing for common deterministic trend slopes. *Journal of Econometrics*, **126** 1–24.

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

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

Vogt, M. and Linton, O. (2018). Multiscale clustering of nonparametric regression curves. *Preprint*.

Wang, W., Phillips, P. C. and Su, L. (2018). Homogeneity pursuit in panel data models: theory and application. *Journal of Applied Econometrics*, **33** 797–815.

Wu, W. B. (2005). Nonlinear system theory: another look at dependence. *Proc. Natn. Acad. Sci. USA*, **102** 14150–14154.

Xu, K.-L. (2012). Robustifying multivariate trend tests to nonstationary volatility. *Journal of Econometrics*, **169** 147–154.

Zhang, T. (2013). Clustering high-dimensional time series based on parallelism. *Journal of the American Statistical Association*, **108** 577–588.

Zhang, Y., Su, L. and Phillips, P. C. (2012). Testing for common trends in semiparametric panel data models with fixed effects. *The Econometrics Journal*, **15** 56–100.

4 Appendix

4.1 Proof of Theorem 3.2

We define the first-differenced regressors as follows.

$$\Delta \mathbf{X}_{it} = \mathbf{H}_i(\mathcal{U}_{it}) - \mathbf{H}_i(\mathcal{U}_{it-1}) := \Delta \mathbf{H}_i(\mathcal{U}_{it}).$$

Similarly,

$$\Delta \varepsilon_{it} = \varepsilon_{it} - \varepsilon_{it-1} = G_i(\mathcal{J}_{it}) - G_i(\mathcal{J}_{it-1}) = \Delta G_i(\mathcal{J}_{it}).$$

With these assumptions we can prove the following propositions.

Proposition 4.1. *Under Assumptions (C5) and (C7),* $||\Delta \mathbf{H}_i(\mathcal{U}_{it})||_4 < \infty$.

Proof of Proposition 4.1. By Assumption (C7),

$$||\Delta \mathbf{H}_i(\mathcal{U}_{it})||_4 \leq ||\mathbf{H}_i(\mathcal{U}_{it})||_4 + ||\mathbf{H}_i(\mathcal{U}_{it-1})||_4 < \infty.$$

□

Proposition 4.2. *Under Assumption (C8), $\Delta \mathbf{X}_{it}$ (elementwise) and $\Delta \varepsilon_{it}$ are uncorrelated for each $t \in \{1, \ldots, T\}$.*

Proof of Proposition 4.2. By Assumption (C8),

$$\begin{aligned}
\mathbb{E}[\Delta \mathbf{X}_{it} \Delta \varepsilon_{it}] = \mathbb{E}\big[(\mathbf{X}_{it} - \mathbf{X}_{it-1})(\varepsilon_{it} - \varepsilon_{it-1})\big] = \\
= \mathbb{E}[\mathbf{X}_{it}\varepsilon_{it}] - \mathbb{E}[\mathbf{X}_{it-1}\varepsilon_{it}] - \mathbb{E}[\mathbf{X}_{it}\varepsilon_{it-1}] + \mathbb{E}[\mathbf{X}_{it-1}\varepsilon_{it-1}] = \\
= \mathbb{E}[\mathbf{X}_{it}]\mathbb{E}[\varepsilon_{it}] - \mathbb{E}[\mathbf{X}_{it-1}]\mathbb{E}[\varepsilon_{it}] - \mathbb{E}[\mathbf{X}_{it}]\mathbb{E}[\varepsilon_{it-1}] + \mathbb{E}[\mathbf{X}_{it-1}]\mathbb{E}[\varepsilon_{it-1}] = \\
= \big(\mathbb{E}[\mathbf{X}_{it}] - \mathbb{E}[\mathbf{X}_{it-1}]\big)\big(\mathbb{E}[\varepsilon_{it}] - \mathbb{E}[\varepsilon_{it-1}]\big) = \mathbb{E}[\Delta \mathbf{X}_{it}]\mathbb{E}[\Delta \varepsilon_{it}]
\end{aligned}$$

□

Proposition 4.3. *Define*

$$\Delta \mathbf{U}_i(\mathcal{I}_{i,t}) := \Delta \mathbf{H}_i(\mathcal{U}_{it}) \Delta G_i(\mathcal{J}_{it}).$$

Under Assumptions (C2) - (C3), (C9) - (C10), we have that $\sum_{s=0}^{\infty} \delta_2(\Delta \mathbf{U}_i, s) < \infty$.

Proof of Proposition 4.3. Note the following

$$\begin{aligned}
&\delta_2(\Delta \mathbf{U}_i, t) = ||\Delta \mathbf{U}_i(\mathcal{I}_{i,t}) - \Delta \mathbf{U}_i(\mathcal{I}'_{i,t})||_2 = \\
&= ||\Delta \mathbf{H}_i(\mathcal{U}_{it}) \Delta G_i(\mathcal{J}_{it}) - \Delta \mathbf{H}_i(\mathcal{U}'_{it}) \Delta G_i(\mathcal{J}'_{it})||_2 = \\
&= ||\big(\mathbf{H}_i(\mathcal{U}_{it}) - \mathbf{H}_i(\mathcal{U}_{it-1})\big)\big(G_i(\mathcal{J}_{it}) - G_i(\mathcal{J}_{it-1})\big) - \big(\mathbf{H}_i(\mathcal{U}'_{it}) - \mathbf{H}_i(\mathcal{U}'_{it-1})\big)\big(G_i(\mathcal{J}'_{it}) - G_i(\mathcal{J}'_{it-1})\big)||_2 = \\
&= ||\mathbf{H}_i(\mathcal{U}_{it}) G_i(\mathcal{J}_{it}) - \mathbf{H}_i(\mathcal{U}_{it-1}) G_i(\mathcal{J}_{it}) - \mathbf{H}_i(\mathcal{U}_{it}) G_i(\mathcal{J}_{it-1}) + \mathbf{H}_i(\mathcal{U}_{it-1}) G_i(\mathcal{J}_{it-1}) - \\
&\quad - \mathbf{H}_i(\mathcal{U}'_{it}) G_i(\mathcal{J}'_{it}) + \mathbf{H}_i(\mathcal{U}'_{it-1}) G_i(\mathcal{J}'_{it}) + \mathbf{H}_i(\mathcal{U}'_{it}) G_i(\mathcal{J}'_{it-1}) - \mathbf{H}_i(\mathcal{U}'_{it-1}) G_i(\mathcal{J}'_{it-1})||_2 \leq \\
&\leq ||\mathbf{H}_i(\mathcal{U}_{it}) G_i(\mathcal{J}_{it}) - \mathbf{H}_i(\mathcal{U}'_{it}) G_i(\mathcal{J}'_{it})||_2 + ||\mathbf{H}_i(\mathcal{U}_{it-1}) G_i(\mathcal{J}_{it-1}) - \mathbf{H}_i(\mathcal{U}'_{it-1}) G_i(\mathcal{J}'_{it-1})||_2 + \\
&\quad + ||\mathbf{H}_i(\mathcal{U}_{it-1}) G_i(\mathcal{J}_{it}) - \mathbf{H}_i(\mathcal{U}'_{it-1}) G_i(\mathcal{J}'_{it})||_2 + ||\mathbf{H}_i(\mathcal{U}_{it}) G_i(\mathcal{J}_{it-1}) - \mathbf{H}_i(\mathcal{U}'_{it}) G_i(\mathcal{J}'_{it-1})||_2 = \\
&= \delta_2(\mathbf{U}_i, t) + \delta_2(\mathbf{U}_i, t-1) + \\
&\quad + ||\mathbf{H}_i(\mathcal{U}_{it-1}) G_i(\mathcal{J}_{it}) - \mathbf{H}_i(\mathcal{U}'_{it-1}) G_i(\mathcal{J}_{it}) + \mathbf{H}_i(\mathcal{U}'_{it-1}) G_i(\mathcal{J}_{it}) - \mathbf{H}_i(\mathcal{U}'_{it-1}) G_i(\mathcal{J}'_{it})||_2 + \\
&\quad + ||\mathbf{H}_i(\mathcal{U}_{it}) G_i(\mathcal{J}_{it-1}) - \mathbf{H}_i(\mathcal{U}'_{it}) G_i(\mathcal{J}_{it-1}) + \mathbf{H}_i(\mathcal{U}'_{it}) G_i(\mathcal{J}_{it-1}) - \mathbf{H}_i(\mathcal{U}'_{it}) G_i(\mathcal{J}'_{it-1})||_2 \leq \\
&\leq \delta_2(\mathbf{U}_i, t) + \delta_2(\mathbf{U}_i, t-1) + \\
&\quad + ||\big(\mathbf{H}_i(\mathcal{U}_{it-1}) - \mathbf{H}_i(\mathcal{U}'_{it-1})\big) G_i(\mathcal{J}_{it})||_2 + ||\mathbf{H}_i(\mathcal{U}'_{it-1})\big(G_i(\mathcal{J}_{it}) - G_i(\mathcal{J}'_{it})\big)||_2 + \\
&\quad + ||\big(\mathbf{H}_i(\mathcal{U}_{it}) - \mathbf{H}_i(\mathcal{U}'_{it})\big) G_i(\mathcal{J}_{it-1})||_2 + ||\mathbf{H}_i(\mathcal{U}'_{it})\big(G_i(\mathcal{J}_{it-1}) - G_i(\mathcal{J}'_{it-1})\big)||_2 \leq \\
&\leq \delta_2(\mathbf{U}_i, t) + \delta_2(\mathbf{U}_i, t-1) + \big(\delta_2(\mathbf{H}_i, t-1) + \delta_2(\mathbf{H}_i, t)\big)||G_i||_2 + \big(\delta_2(G_i, t-1) + \delta_2(G_i, t)\big)||\mathbf{H}_i||_2
\end{aligned}$$

Here $\mathcal{U}'_{it} = (\ldots, u_{i(-1)}, u'_{i0}, u_{i1}, \ldots, u_{it-1}, u_{it})$, $\mathcal{U}'_{it-1} = (\ldots, u_{i(-1)}, u'_{i0}, u_{i1}, \ldots, u_{it-1})$, $\mathcal{J}'_{it} = (\ldots, \eta_{i(-1)}, \eta'_{i0}, \eta_{i1}, \ldots, \eta_{it-1}, \eta_{it})$, $\mathcal{J}'_{it-1} = (\ldots, \eta_{i(-1)}, \eta'_{i0}, \eta_{i1}, \ldots, \eta_{it-1})$ are coupled processes with u'_{i0} being an i.i.d. copy of u_{i0} and η'_{i0} being an i.i.d. copy of η_{i0}. This leads us to

$$\sum_{s=0}^{\infty} \delta_2(\Delta \mathbf{U}_i, s) \leq \sum_{s=0}^{\infty} \delta_2(\mathbf{U}_i, s) + \sum_{s=1}^{\infty} \delta_2(\mathbf{U}_i, s-1) + \\ + \sum_{s=1}^{\infty} \big(\delta_2(\mathbf{H}_i, s-1) + \delta_2(\mathbf{H}_i, s)\big)||G_i||_2 + \sum_{s=1}^{\infty} \big(\delta_2(G_i, s-1) + \delta_2(G_i, s)\big)||\mathbf{H}_i||_2 < \infty$$

□

Proposition 4.4. *Under Assumptions (C1) - (C10),*

$$\Big|\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \Delta \mathbf{X}_{it} \Delta \varepsilon_{it}\Big| = O_P(1).$$

Proof of Proposition 4.4. We need the following notation:

$$\begin{aligned} \mathcal{P}_{i,t}(\cdot) &:= \mathbb{E}[\cdot|\mathcal{I}_{i,t}] - \mathbb{E}[\cdot|\mathcal{I}_{i,t-1}], \\ \kappa_i &:= \frac{1}{T} \sum_{t=1}^{T} \Delta \mathbf{X}_{it} \Delta \varepsilon_{it}, \\ \kappa^{\mathcal{P}}_{i,s} &:= \frac{1}{T} \sum_{t=1}^{T} \mathcal{P}_{i,t-s}\big(\Delta \mathbf{X}_{it} \Delta \varepsilon_{it}\big). \end{aligned}$$

Then,

$$\begin{aligned} ||\kappa^{\mathcal{P}}_{i,s}||^2 &= \Big|\Big|\frac{1}{T} \sum_{t=1}^{T} \mathcal{P}_{i,t-s}\big(\Delta \mathbf{X}_{it} \Delta \varepsilon_{it}\big)\Big|\Big|^2 \leq \\ &\leq \frac{1}{T^2} \sum_{t=1}^{T} \Big|\Big|\mathbb{E}(\Delta \mathbf{X}_{it} \Delta \varepsilon_{it}|\mathcal{I}_{i,t-s}) - \mathbb{E}(\Delta \mathbf{X}_{it} \Delta \varepsilon_{it}|\mathcal{I}_{i,t-s-1})\Big|\Big|^2 = \\ &= \frac{1}{T^2} \sum_{t=1}^{T} \Big|\Big|\mathbb{E}(\Delta \mathbf{X}_{it} \Delta \varepsilon_{it}|\mathcal{I}_{i,t-s}) - \mathbb{E}(\Delta \mathbf{X}'_{it,s} \Delta \varepsilon'_{it,s}|\mathcal{I}_{i,t-s})\Big|\Big|^2, \end{aligned}$$

where $\Delta \mathbf{X}'_{it,s} \Delta \varepsilon'_{it,s}$ denotes $\Delta \mathbf{X}_{it} \Delta \varepsilon_{it}$ with $\{\zeta_{i,t-s}\}$ replaced by its i.i.d. copy $\{\zeta'_{i,t-s}\}$. In this case $\mathbb{E}(\Delta \mathbf{X}'_{it,s} \Delta \varepsilon'_{it,s}|\mathcal{I}_{i,t-s-1}) = \mathbb{E}(\Delta \mathbf{X}'_{it,s} \Delta \varepsilon'_{it,s}|\mathcal{I}_{i,t-s})$. Furthermore, by linearity of the expectation and Jensen's inequality, we have

$$||\kappa^{\mathcal{P}}_{i,s}||^2 \leq \frac{1}{T^2} \sum_{t=1}^{T} \Big|\Big|\mathbb{E}(\Delta \mathbf{X}_{it} \Delta \varepsilon_{it}|\mathcal{I}_{i,t-s}) - \mathbb{E}(\Delta \mathbf{X}'_{it,s} \Delta \varepsilon'_{it,s}|\mathcal{I}_{i,t-s})\Big|\Big|^2 \leq$$

$$\leq \frac{1}{T^2}\sum_{t=1}^{T}\Big|\Big|\Delta\mathbf{X}_{it}\Delta\varepsilon_{it}-\Delta\mathbf{X}'_{it,s}\Delta\varepsilon'_{it,s}\Big|\Big|^2 =$$

$$= \frac{1}{T^2}\sum_{t=1}^{T}\Big|\Big|\Delta\mathbf{H}_i(\mathcal{U}_{it})\Delta G_i(\mathcal{J}_{it})-\Delta\mathbf{H}_i(\mathcal{U}'_{it,s})\Delta G_i(\mathcal{J}'_{it,s})\Big|\Big|^2 =$$

$$= \frac{1}{T^2}\sum_{t=1}^{T}\Big|\Big|\Delta\mathbf{U}_i(\mathcal{I}_{i,t})-\Delta\mathbf{U}_i(\mathcal{I}'_{i,t,s})\Big|\Big|^2 \leq \frac{1}{T^2}\sum_{t=1}^{T}\delta_2^2(\Delta\mathbf{U}_i,s) = \frac{1}{T}\delta_2^2(\Delta\mathbf{U}_i,s)$$

with $\mathcal{U}'_{it,s}=(\ldots,u_{it-s-1},u'_{it-s},u_{it-s+1},\ldots,u_{it})$, $\mathcal{J}'_{it,s}=(\ldots,\eta_{it-s-1},\eta'_{it-s},\eta_{it-s+1},\ldots,\eta_{it})$, $\zeta'_{it}=(u'_{it},\eta'_{it})^\top$ and $\mathcal{I}'_{i,t,s}=(\ldots,\zeta_{it-s-1},\zeta'_{it-s},\zeta_{it-s+1},\ldots,\zeta_{it})$.
Moreover,

$$\begin{aligned}\kappa_i-\mathbb{E}\kappa_i &= \frac{1}{T}\sum_{t=1}^{T}\Delta\mathbf{X}_{it}\Delta\varepsilon_{it}-\mathbb{E}\kappa_i = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(\Delta\mathbf{X}_{it}\Delta\varepsilon_{it}|\mathcal{I}_{i,t})-\mathbb{E}\kappa_i = \\ &= \frac{1}{T}\sum_{t=1}^{T}\big(\mathbb{E}(\Delta\mathbf{X}_{it}\Delta\varepsilon_{it}|\mathcal{I}_{i,t})-\mathbb{E}(\mathbf{X}_{it}\Delta\varepsilon_{it})\big) = \\ &= \frac{1}{T}\sum_{t=1}^{T}\sum_{s=0}^{\infty}\big(\mathbb{E}(\Delta\mathbf{X}_{it}\Delta\varepsilon_{it}|\mathcal{I}_{i,t-s})-\mathbb{E}(\Delta\mathbf{X}_{it}\Delta\varepsilon_{it}|\mathcal{I}_{i,t-s-1})\big) = \\ &= \frac{1}{T}\sum_{t=1}^{T}\sum_{s=0}^{\infty}\mathcal{P}_{i,t-s}(\Delta\mathbf{X}_{it}\Delta\varepsilon_{it}) = \sum_{s=0}^{\infty}\kappa_{i,s}^{\mathcal{P}}.\end{aligned}$$

Thus, by Proposition 4.3,

$$||\kappa_i-\mathbb{E}\kappa_i|| \leq \sum_{s=0}^{\infty}||\kappa_{i,s}^{\mathcal{P}}|| \leq \frac{1}{\sqrt{T}}\sum_{s=0}^{\infty}\delta_2(\Delta\mathbf{U}_i,s) = O\Big(\frac{1}{\sqrt{T}}\Big)$$

Since $\mathbb{E}\kappa_i=0$ by Proposition 4.2, we conclude that

$$\Big|\Big|\frac{1}{T}\sum_{t=1}^{T}\Delta\mathbf{X}_{it}\Delta\varepsilon_{it}\Big|\Big| = O\Big(\frac{1}{\sqrt{T}}\Big).$$

Therefore, the proposition follows. □

Proof of Theorem 3.2. Recall the differencing estimator $\widehat{\boldsymbol{\beta}}_i$:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_i &= \Big(\sum_{t=1}^{T}\Delta\mathbf{X}_{it}\Delta\mathbf{X}_{it}^\top\Big)^{-1}\sum_{t=1}^{T}\Delta\mathbf{X}_{it}\Delta Y_{it} = \\ &= \Big(\sum_{t=1}^{T}\Delta\mathbf{X}_{it}\Delta\mathbf{X}_{it}^\top\Big)^{-1}\sum_{t=1}^{T}\Delta\mathbf{X}_{it}\Big(\Delta\mathbf{X}_{it}^\top\boldsymbol{\beta}_i+\Delta\varepsilon_{it}+O\Big(\frac{1}{T}\Big)\Big) = \\ &= \boldsymbol{\beta}_i+\Big(\sum_{t=1}^{T}\Delta\mathbf{X}_{it}\Delta\mathbf{X}_{it}^\top\Big)^{-1}\sum_{t=1}^{T}\Delta\mathbf{X}_{it}\Delta\varepsilon_{it}+O\Big(\frac{1}{T}\Big)\Big(\sum_{t=1}^{T}\Delta\mathbf{X}_{it}\Delta\mathbf{X}_{it}^\top\Big)^{-1}\sum_{t=1}^{T}\Delta\mathbf{X}_{it}.\end{aligned}$$

This leads to

$$\sqrt{T}(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i) = \Big(\frac{1}{T}\sum_{t=1}^{T}\Delta\mathbf{X}_{it}\Delta\mathbf{X}_{it}^{\top}\Big)^{-1}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\Delta\mathbf{X}_{it}\Delta\varepsilon_{it} + \\ + O\Big(\frac{1}{\sqrt{T}}\Big)\Big(\frac{1}{T}\sum_{t=1}^{T}\Delta\mathbf{X}_{it}\Delta\mathbf{X}_{it}^{\top}\Big)^{-1}\frac{1}{T}\sum_{t=1}^{T}\Delta\mathbf{X}_{it}.$$

Since

$$\mathbb{E}\Big[\frac{1}{T}\sum_{t=1}^{T}\Delta H_{ij}(\mathcal{U}_{it})\Big] = 0$$

and

$$\mathrm{Var}\Big[\frac{1}{T}\sum_{t=1}^{T}\Delta H_{ij}(\mathcal{U}_{it})\Big] \leq \frac{4}{T^2}\mathbb{E}\big[H_{ij}^2(\mathcal{U}_{it})\big],$$

by Chebyshev's inequality we have that $\Big|\frac{1}{T}\sum_{t=1}^{T}\Delta H_{ij}(\mathcal{U}_{it})\Big| = O_P(1)$ for each $j \in \{1,\dots,d\}$. And this in turn implies that

$$\Big|\frac{1}{T}\sum_{t=1}^{T}\Delta\mathbf{H}_i(\mathcal{U}_{it})\Big| = \Big|\frac{1}{T}\sum_{t=1}^{T}\Delta\mathbf{X}_{it}\Big| = O_P(1). \tag{4.1}$$

Similarly, by Proposition 4.1 and Chebyshev's inequality, we have that for each $j,k \in \{1,\dots,d\}$

$$\Big|\frac{1}{T}\sum_{t=1}^{T}\Delta H_{ij}(\mathcal{U}_{it})\Delta H_{ik}(\mathcal{U}_{it})\Big| = O_P(1),$$

which leads to

$$\Big|\Big|\frac{1}{T}\sum_{t=1}^{T}\Delta\mathbf{H}_i(\mathcal{U}_{it})\Delta\mathbf{H}_i(\mathcal{U}_{it})^{\top}\Big|\Big| = \Big|\Big|\frac{1}{T}\sum_{t=1}^{T}\Delta\mathbf{X}_{it}\Delta\mathbf{X}_{it}^{\top}\Big|\Big| = O_P(1), \tag{4.2}$$

where $||A||$ with A being a matrix is any matrix norm.
By Assumption (C6), we know that $\mathbb{E}[\Delta\mathbf{X}_{it}\Delta\mathbf{X}_{it}^{\top}] = \mathbb{E}[\Delta\mathbf{X}_{i0}\Delta\mathbf{X}_{i0}^{\top}]$ is invertible, thus,

$$\Big|\Big|\Big(\frac{1}{T}\sum_{t=1}^{T}\Delta\mathbf{X}_{it}\Delta\mathbf{X}_{it}^{\top}\Big)^{-1}\Big|\Big| = O_P(1).$$

By applying Proposition 4.4, (4.1) and (4.2), the statement of the theorem follows. □

4.2 Proof of Theorem ??