# 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 the presence of covariates. In many applications, practitioners are interested whether the observed time series all have the same time trend. Moreover, when some of the trends are different, there may still be groups of time series with the same trend. In this case, it is often of interest to estimate the unknown groups from the data. In addition, when two trends are not the same, it may also be relevant to know in which time regions they differ from each other. 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.

**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 Introduction

Comparison of several regression curves is a classical topic in econometrics and statistics. In many cases of practical interest, the objective regression curves are of unknown functional form and the parametric approach is not applicable. In this paper, we are interested in performing the comparison of several regression curves in a nonparametric context. Specifically, we present a new testing procedure for detecting differences in the nonparametric trends curves.

In what follows, we consider a general panel framework with heterogeneous trends. 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 $\boldsymbol{\beta}_i$ is a $d \times 1$ vector of unknown parameters, $\mathbf{X}_{it}$ is a $d \times 1$ vector of individual covariates or controls, $m_i$ is an unknown nonparametric (deterministic) trend function defined on $[0,1]$, $\alpha_i$ are so-called fixed effect error terms and $\mathcal{E}_i = \{\varepsilon_{it} : 1 \leq t \leq T\}$ is a zero-mean stationary error process.

[1]Corresponding author. Address: Bonn Graduate School of Economics, University of Bonn, 53113 Bonn, Germany. Email: marina.k@uni-bonn.de.

[2]Address: Institute of Statistics, Department of Mathematics and Economics, Ulm University, 89081 Ulm, Germany. Email: m.vogt@uni-ulm.de.

An important question in many applications is whether the observed time series have the common trend. In other words, the researchers would like to know if $m_i$ are the same for all $i$. Moreover, when some of the trends are different, there may still be groups of time series with the same trend. In this case, it is often of interest to estimate the unknown groups from the data. In addition, when two trends $m_i$ and $m_j$ are not the same, it may also be relevant to know in which time regions they differ from each other. In this paper, we introduce new statistical methods to approach these questions. In particular, we develop a test of the hypothesis that all time trends in model (1.1) are the same. In this setting, the null hypothesis is formulated as

$$H_0 : m_1 = m_2 = \ldots = m_n, \tag{1.2}$$

whereas the alternative hypothesis is

$$H_1 : \text{ there exists } x \in [0,1] \text{ such that } m_i(x) \neq m_j(x) \text{ for some } 1 \leq i < j \leq n.$$

The method that we propose does not only allow to test whether the null hypothesis is violated. It also allows to detect, with a given statistical confidence, which time trends are different and in which time regions they differ. More specifically, for any given interval $[u-h, u+h] \subseteq [0,1]$, consider the hypothesis

$$H_0^{[i,j]}(u,h) : m_i(w) = m_j(w) \text{ for all } w \in [u-h, u+h].$$

Here, we can regard $h$ as a bandwidth parameter, a common tuning parameter in nonparametric estimation. The given interval $\mathcal{I}_{(u,h)} = [u-h, u+h] \subseteq [0,1]$ is then fully characterized by $u$, its center (the location parameter), and $h$, the bandwidth. In order to determine the regions where the time trends are different, we consider a broad range of pairs $(u,h)$ with the property that they cover the unit interval $[0,1]$. Formally, let $\mathcal{G} := \{(u,h) : \mathcal{I}_{(u,h)} = [u-h, u+h] \subseteq [0,1]\}$ be a grid of location-bandwidth points such that

$$\bigcup_{(u,h)\in\mathcal{G}} \mathcal{I}_{(u,h)} = [0,1].$$

We then reformulate our null hypothesis (1.2) as

$$H_0 : \text{The hypothesis } H_0^{[i,j]}(u,h) \text{ holds true for all intervals } [u-h, u+h], (u,h) \in \mathcal{G}, \text{ and for all } 1 \leq i < j \leq n.$$

In this paper, we introduce a method that allows to test the hypothesis $H_0^{[i,j]}(u,h)$ simultaneously for all pairs $(i,j)$ and for all intervals $\mathcal{I}_{(u,h)}$ under consideration. The method that we propose is a multiscale test. The underlying idea of a multiscale test is to consider a number of test statistics (each of which corresponds to different values of some tuning parameters) simultaneously rather to consider a single test statistic. In our

case, this results in testing many separate null hypotheses $H_0^{[i,j]}(u,h)$. In the paper, we also show how to derive appropriate critical values and we prove the main theoretical result that our multiscale test has the correct (asymptotic) level.

controls the familywise error rate, that is, the probability of wrongly rejecting at least one null hypothesis H(ijk) 0 . As we will see, this allows us to make simultaneous confidence statements of the following form for a given signifcance level We now set up a multiscale method which simultaneously tests the hypothesis $H_0^{[i,j]}(u,h)$ for all possible points $(u,h)$ and all pairs $(i,j)$ with $i<j$.[1] Our strategy to derive such a method can be outlined as follows:

The problem of testing whether the observed time series all have the same trend has been widely studied and tests for equality of trend or regression curves have been developed in Härdle and Marron (1990), Hall and Hart (1990), Delgado (1993) and Degras et al. (2012) among many others. Versions of model (1.1) with a parametric trend are studied in Vogelsang and Franses (2005), Sun (2011) and Xu (2012) among others. In the nonparametric 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 Degras et al. (2012), Zhang et al. (2012) and Hidalgo and Lee (2014). Degras et al. (2012) consider the problem of testing $H_0$ within the model framework that is a special case of (1.1) which does not include additional regressors. 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 $i$.

Recently, Khismatullina and Vogt (2021) proposed a new inference method that allows to detect differences between epidemic time trends in the context of the COVID-19 pandemic. They presented a statistically rigorous procedure that not only allows to compare trends across different countries, but also to pinpoint the time intervals where the differences occur. However, they examined a special case of the model (1.1) which does not include neither the covariates $\mathbf{X}_{it}$, nor the fixed effects $\alpha_i$, and restricts the error terms $\varepsilon_{it}$ to be independent across $t$. Our model (1.1) can be regarded as a generalization of their model, which allows for a wider range of economic and financial

[1] Obviously, in practice, we cannot consider all points $u \in (0,1)$ and all $h>0$ but have to restrict attention to a finite subset of points. We ignore this in our presentation for simplicity.

applications.
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 (Grier and Tullock, 1989). Another example concerns the dynamics of 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 (Nyblom and Harvey, 2000). Finally, in climatology, researchers are interested in comparing the trending behaviour of temperature time series across different spatial locations (Karoly and Wu, 2005).

In this article we discuss testing the equality of two regres- sion functions that are specified in terms of only some smoothness conditions and, consequently, allow the proposed tests to be used for a wide class of functions.
Within the general framework of model (1.1), we can formulate a number of interesting statistical questions concerning the set of trend functions $\{m_i : 1 \leq i \leq n\}$.
Tests of $H_{0,\text{para}}$, $H_{0,\text{const}}$ and related hypotheses have for example been studied in Lyubchich and Gel (2016).
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. In most cases, it is far from clear how to choose these bandwidths in an appropriate way. On the contrary, our method allows us consider a large collection of bandwidths simultaneously and thus, we avoid the need to choose only one bandwidth.
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 $\alpha$). 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 appropriate 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 (2020). They developed multiscale methods to test for local increases/decreases of the nonparametric trend function $m$ in the univariate time series model $Y_t = m(t/T) + \varepsilon_t$.

The main theoretical contribution of the current paper is the multiscale method that allows to make simultaneous confidence statements about the regions where the time trends differ. To the best of our knowledge, currently there are no equivalent statistical methods. Even though tests for equality of the trends have been developed already for a while, most existing procedures allow only to test whether the trend curves are all the same or not, but they almost never allow to infer which curves are different and where. The only two exceptions are Khismatullina and Vogt (2021) whose contribution is briefly discussed above and Park et al. (2009) who developed SiZer methods for the comparison of nonparametric trend curves in a strongly simplified version of model (1.1). The analysis of Park et al. (2009), however, is mainly methodological with the theory derived only for the special case $n = 2$, that is, when only two time series are observed.

The paper is structured as follows. Section 2 introduces the model setting and the necessary technical assumptions that are required for the theory. The multiscale test is developed step by step in Section 3. The main theoretical results are presented in Section 4. To keep the discussion as clear as possible, we include in the main text of the paper only the essential parts of the theoretical arguments, whereas the technical details and extended proofs are deferred to the Appendix. Section 5 concludes.

# 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 $|\mathbf{v}| = \big(\sum_{i=1}^m v_i^2\big)^{1/2}$ and $|\mathbf{v}|_q = \big(\sum_{i=1}^m v_i^q\big)^{1/q}$ respectively. 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 the particular 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 $\mathcal{F}_t = (\ldots, \epsilon_{-1}, \epsilon_0, \epsilon_1, \ldots, \epsilon_{t-1}, \epsilon_t)$ and $\mathcal{F}_t' = (\ldots, \epsilon_{-1}, \epsilon_0', \epsilon_1, \ldots, \epsilon_{t-1}, \epsilon_t)$ is a coupled process of $\mathcal{F}_t$ with $\epsilon_0'$ being an i.i.d. copy of $\epsilon_0$.

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 $[0, 1]$, $\alpha_i$ 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 $\alpha_i$. 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 $\alpha_i$ 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 $\alpha_i$ 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 $\alpha_i$ may be correlated, the error processes $\mathcal{E}_i$ are assumed to be independent across $i$. Technical conditions regarding the model are discussed further in this section.

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.

## 2.1 Assumptions

Each process $\mathcal{E}_i$ is supposed to satisfy the following conditions:

(C1) For each $i$ the variables $\varepsilon_{it}$ allow for the representation $\varepsilon_{it} = G_i(\ldots, \eta_{it-1}, \eta_{it})$, where $\eta_{it}$ 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 (2020).
Regarding the independent variables $\mathbf{X}_{it}$, we need the following additional assumptions for each $i$:

(C4) The covariates $\mathbf{X}_{it}$ allow for the representation $\mathbf{X}_{it} = \mathbf{H}_i(\ldots, u_{it-1}, u_{it})$ 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. Denote $\mathcal{U}_{it} = (\ldots, u_{it-1}, u_{it})$.

(C5) 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.

(C6) Let $\mathbb{E}[\mathbf{H}_i(\mathcal{U}_{i0})] = \mathbf{0}$ and $||\mathbf{H}_i(\mathcal{U}_{it})||_{q'} < \infty$ for some $q' > \max\{2\theta, 4\}$, where $\theta$ will be introduced further in Assumption (C12).

(C7) $\sum_{s=0}^{\infty} \delta_{q'}(\mathbf{H}_i, s) < \infty$ for $q'$ from Assumption (C6).

(C8) For each $i$ it holds that $\sum_{s=t}^{\infty} \delta_{q'}(\mathbf{H}_i, s) = O(t^{-\alpha})$ for $q'$ from Assumption (C6) and for some $\alpha > 1/2 - 1/q'$.

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

(C9) $\mathbf{X}_{it}$ (elementwise) and $\varepsilon_{is}$ are uncorrelated for each $t, s \in \{1, \ldots, T\}$.

(C10) Let $\zeta_{i,t} = (u_{it}, \eta_{it})^\top$. Define $\mathcal{I}_{i,t} = (\ldots, \zeta_{i,t-1}, \zeta_{i,t})$ and $\mathbf{U}_i(\mathcal{I}_{i,t}) = \mathbf{H}_i(\mathcal{U}_{it})G_i(\mathcal{J}_{it})$. With this notation at hand, we assume that $\sum_{s=0}^{\infty} \delta_2(\mathbf{U}_i, s) < \infty$.

# 3 Testing for equality of time trends

In this section, we adapt the multiscale method developed in **?** 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 continuously differentiable on $[0, 1]$.

It is obvious that if $\alpha_i$ 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 $Y_{it}^\circ$ are not observed since the intercept $\alpha_i$ and the coefficients $\boldsymbol{\beta}_i$ are not known. Given appropriate estimators $\widehat{\boldsymbol{\beta}}_i$ and $\widehat{\alpha}_i$, we can then consider

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

Then our unobserved variables $Y_{it}^\circ$ can be approximated by $\widehat{Y}_{it}$ and we compute our test statistic based on $\widehat{Y}_{it}$. In what follows, we assume that an estimator with the property that $\boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_i = O_P(T^{-1/2})$ is given. Details on one of the possible ways to construct $\widehat{\boldsymbol{\beta}}_i$ are deferred to Section 4.1.

Given $\widehat{\boldsymbol{\beta}}_i$, consider an appropriate estimator $\widehat{\alpha}_i$ for the intercept $\alpha_i$ calculated by

$$\begin{aligned} \widehat{\alpha}_i &= \frac{1}{T}\sum_{t=1}^{T}\big(Y_{it} - \widehat{\boldsymbol{\beta}}_i^\top \mathbf{X}_{it}\big) = \frac{1}{T}\sum_{t=1}^{T}\big(\boldsymbol{\beta}_i^\top \mathbf{X}_{it} - \widehat{\boldsymbol{\beta}}_i^\top \mathbf{X}_{it} + \alpha_i + m_i(t/T) + \varepsilon_{it}\big) = \qquad (3.1) \\ &= \big(\boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_i\big)^\top \frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_{it} + \alpha_i + \frac{1}{T}\sum_{i=1}^{T} m_i(t/T) + \frac{1}{T}\sum_{i=1}^{T}\varepsilon_{it}. \end{aligned}$$

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 $\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i = O_P(T^{-1/2})$. Plugging all these results together

in (3.1), we get that $\widehat{\alpha}_i - \alpha_i = O_P(T^{-1/2})$.Thus, 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 $\widehat{Y}_{it} = Y_{it} - \widehat{\alpha}_i - \widehat{\boldsymbol{\beta}}_i^{\top}\mathbf{X}_{it} = Y_{it}^{\circ} + O_P(T^{-1/2})$.

We now turn to the estimator of the long-run error variance $\sigma_i^2 = \sum_{\ell=-\infty}^{\infty} \mathrm{Cov}(\varepsilon_{i0}, \varepsilon_{i\ell})$. For the moment, we assume that the long-run variance does not depend on $i$, that is $\sigma_i^2 = \sigma^2$ for all $i$. We will need this further for conducting the testing procedure properly. Nevertheless, we keep the indices throughout the paper in order to be congruous in notation. We further let $\widehat{\sigma}_i^2$ be an estimator of $\sigma_i^2$ 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 $1 \le t \le T$. Hence, whereas the true long-run variance is the sme for all time series, the estimators are different. Throughout the section, we assume that $\widehat{\sigma}_i^2 = \sigma_i^2 + o_p(\rho_T)$ with $\rho_T = o(\sqrt{h_{\min}}/\log T)$. Details on how to construct $\widehat{\sigma}_i^2$ are deferred to Section **??**.

Moreover, in the proof of our main theorem 4.1 we will need additional auxiliary statistics that do not include the covariates $\mathbf{X}_{it}$. Hence, we imagine that we know the parameters $\boldsymbol{\beta}_i$ and consider the unobserved variables

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

For this auxiliary statistics we will use the auxiliary estimator $\widehat{\widehat{\sigma}}_i^2$ of the long-run error variance $\sigma_i^2$ which is computed from the augmented sample $\{\widehat{\widehat{Y}}_{it} : 1 \le t \le T\}$. We thus regard $\widehat{\widehat{\sigma}}_i^2 = \widehat{\widehat{\sigma}}_i^2(\widehat{\widehat{Y}}_{i1}, \ldots, \widehat{\widehat{Y}}_{iT})$ as a function of the variables $\widehat{\widehat{Y}}_{it}$ for $1 \le t \le T$. As with $\widehat{\sigma}_i^2$, we assume that $\widehat{\widehat{\sigma}}_i^2 = \sigma_i^2 + o_p(\rho_T)$ with $\rho_T = o(\sqrt{h_{\min}}/\log T)$.

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.

$$w_{t,T}(u,h) = \frac{\Lambda_{t,T}(u,h)}{\{\sum_{t=1}^{T}\Lambda_{t,T}(u,h)^2\}^{1/2}}, \tag{3.2}$$

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:

(C11) 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 now combine the test statistics $\widehat{\psi}_{ij,T}(u,h)$ for a wide range of different locations $u$ and bandwidths or scales $h$ in a following way:

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

where $\lambda(h) = \sqrt{2\log\{1/(2h)\}}$ and the set $\mathcal{G}_T$ is the set of points $(u,h)$ that are taken into consideration. The statistic $\widehat{\Psi}_{ij,T}$ can be interpreted as a global distance measure between the two curves $m_i$ and $m_j$. Thus, the multiscale statistic $\widehat{\Psi}_{ij,T}$ simultaneously takes into account all locations $u$ and bandwidths $h$ with $(u,h) \in \mathcal{G}_T$. Throughout the paper, we suppose that $\mathcal{G}_T$ is some subset of $\mathcal{G}_T^{\text{full}} = \{(u,h) : u = t/T \text{ for some } 1 \leq t \leq T \text{ and } h \in [h_{\min}, h_{\max}]\}$, where $h_{\min}$ and $h_{\max}$ denote some minimal and maximal bandwidth value, respectively. For our theory to work, we require the following conditions to hold:

(C12) $|\mathcal{G}_T| = O(T^\theta)$ for some arbitrarily large but fixed constant $\theta > 0$, where $|\mathcal{G}_T|$ denotes the cardinality of $\mathcal{G}_T$.

(C13) $h_{\min} \gg T^{-(1-\frac{2}{q})} \log T$, that is, $h_{\min}/\{T^{-(1-\frac{2}{q})} \log T\} \to \infty$ with $q > 4$ defined in (C2) and $h_{\max} < 1/2$.

We finally define the multiscale statistic for testing the null hypothesis $H_0 : m_1 = m_2 = \ldots = m_n$ as

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

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 which are independent of the error terms $\varepsilon_{it}$ 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

$$\Phi_{ij,T} = \max_{(u,h)\in\mathcal{G}_T} \Big\{ \Big| \frac{\phi_{ij,T}(u,h)}{(\sigma_i^2 + \sigma_j^2)^{1/2}} \Big| - \lambda(h) \Big\}, \tag{3.4}$$

where $\phi_{ij,T}(u,h) = \sum_{t=1}^{T} w_{t,T}(u,h)\left\{\sigma_i(Z_{it}-\bar{Z}_i)-\sigma_j(Z_{jt}-\bar{Z}_j)\right\}$. Since by our assumption $\sigma_i^2 = \sigma_j^2 = \sigma^2$, we can rewrite the Gaussian statistic as follows:

$$\Phi_{ij,T} = \max_{(u,h)\in\mathcal{G}_T}\left\{\frac{1}{\sqrt{2}}\Big|\sum\nolimits_{t=1}^{T} w_{t,T}(u,h)\left\{(Z_{it}-\bar{Z}_i)-(Z_{jt}-\bar{Z}_j)\right\}\Big| - \lambda(h)\right\},$$

which means that $\Phi_{ij,T}$ does not depend on any unknown quantities such as $\sigma_i^2$ or $\sigma_j^2$ and the distribution of this random valuable is fully known. However, we will stick to the notation in (3.4) for the sake of similarity to $\widehat{\Psi}_{ij,T}$.
Moreover, we define the statistic

$$\Phi_{n,T} = \max_{1\le i<j\le n} \Phi_{ij,T} \tag{3.5}$$

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 $\widehat{\Psi}_{n,T} > q_{n,T}(\alpha)$.

# 4 Theoretical properties of the test

To start with, we introduce the auxiliary statistic

$$\widehat{\Phi}_{n,T} = \max_{1\le i<j\le n} \widehat{\Phi}_{ij,T}, \tag{4.1}$$

where

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

and $\widehat{\phi}_{ij,T}(u,h) = \sum_{t=1}^{T} w_{t,T}(u,h)\big\{(\varepsilon_{it}-\bar{\varepsilon}_i) + (\boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_i)^\top(\mathbf{X}_{it}-\bar{\mathbf{X}}_i) - (\varepsilon_{jt}-\bar{\varepsilon}_j) - -(\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j)^\top(\mathbf{X}_{jt}-\bar{\mathbf{X}}_j)\big\}$ 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 4.1.** *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)–(C13) be fulfilled and assume that for all $i \in \{1,\ldots,n\}$ we have $\sigma_i^2 = \sigma^2$, $\widehat{\sigma}_i^2 = \sigma_i^2 + o_p(\rho_T)$ and $\widehat{\widehat{\sigma}}_i^2 = \sigma_i^2 + o_p(\rho_T)$ with $\rho_T = o(\sqrt{h_{\min}}/\log T)$. Then*

$$\mathbb{P}\big(\widehat{\Phi}_{n,T} \le q_{n,T}(\alpha)\big) = (1-\alpha) + o(1)$$

Theorem 4.1 is the main stepping stone to derive the theoretical properties of our multiscale test. The proof of the theorem is provided in the Section 6.2.

### 4.1 Estimation of the parameters $\boldsymbol{\beta}_i$

We now focus on finding an appropriate estimator $\widehat{\boldsymbol{\beta}}_i$ of $\boldsymbol{\beta}_i$. For that purpose, for each $i$ we consider the time series $\{\Delta Y_{it} : 2 \leq t \leq T\}$ of the differences $\Delta Y_{it} = Y_{it} - Y_{it-1}$. We then have

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

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 $\big|m_i\big(\frac{t}{T}\big) - m_i\big(\frac{t-1}{T}\big)\big| = O\big(\frac{1}{T}\big)$ and rewrite

$$\Delta Y_{it} = \boldsymbol{\beta}_i^\top \Delta \mathbf{X}_{it} + \Delta\varepsilon_{it} + O\Big(\frac{1}{T}\Big). \tag{4.2}$$

In particular, for each $i$ we employ the least squares estimation method to estimate $\boldsymbol{\beta}_i$ in (4.2), treating $\Delta \mathbf{X}_{it}$ as the regressors and $\Delta Y_{it}$ as the response variable. That is, we propose the following differencing estimator:

$$\widehat{\boldsymbol{\beta}}_i = \Big( \sum_{t=2}^{T} \Delta \mathbf{X}_{it} \Delta \mathbf{X}_{it}^\top \Big)^{-1} \sum_{t=2}^{T} \Delta \mathbf{X}_{it} \Delta Y_{it} \tag{4.3}$$

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

**Theorem 4.2.** *Under Assumptions (C1) - (C10), we have*

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

*where $\widehat{\boldsymbol{\beta}}_i$ is the differencing estimator given by* (4.3).

# 5 Conclusion

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 $1 \leq k \leq K_0$,

$$m_i = m_j \quad \text{for all } i, j \in G_k. \tag{5.1}$$

According to (5.1), 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.

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 $\beta_i$ 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 $\beta_i$ 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 (5.1) 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.

# 6 Appendix

In this section, we prove the theoretical results from Section **??**. We use the following notation: The symbol $C$ denotes a universal real constant which may take a different value on each occurrence. For $a, b \in \mathbb{R}$, we write $a_+ = \max\{0, a\}$ and $a \vee b = \max\{a, b\}$. For any set $A$, the symbol $|A|$ denotes the cardinality of $A$. The notation $X \overset{\mathcal{D}}{=} Y$ means that the two random variables $X$ and $Y$ have the same distribution. Finally, $f_0(\cdot)$ and $F_0(\cdot)$ denote the density and the distribution function of the standard normal distribution, respectively.

## 6.1 Statistics used in the Appendix

Table 1: Relationship between statistics used in the proofs

| | $\widehat{\Phi}_{n,T}$ | $\widehat{\widehat{\Phi}}_{n,T}$ | $\widetilde{\Phi}_{n,T}$ | $\Phi_{n,T}$ |
|---|---|---|---|---|
| $\widehat{\Psi}_{n,T}$ | Equal under $H_0$ | | | |
| $\widehat{\Phi}_{n,T}$ | | Close due to Prop. A.2 | | |
| $\widehat{\widehat{\Phi}}_{n,T}$ | | | Same distribution (Prop. A.3) | |
| $\widetilde{\Phi}_{n,T}$ | | | | Lemma A.6 with the help of Prop. A.3 and Prop. A.5 |

In the proof of Theorem 4.1, we use a number of different test statistics, either defined in Section **??** or the auxuliary statistics defined below. Each of these statistics plays an important role in one or more steps of the proof. In the following list, we present these test statistics, describe how they are constructed and explain in which parts of the proof they are used. Table 1 is a useful guide for connecting these statistics with their places in the proof strategy presented below.

- Multiscale statistic that is calculated based on data (defined in (3.3)):

$$\begin{aligned} \widehat{\Psi}_{n,T} &= \max_{1 \le i < j \le n} \widehat{\Psi}_{ij,T} \\ \widehat{\Psi}_{ij,T} &= \max_{(u,h) \in \mathcal{G}_T} \Big\{ \Big| \frac{\widehat{\psi}_{ij,T}(u,h)}{(\widehat{\sigma}_i^2 + \widehat{\sigma}_j^2)^{1/2}} \Big| - \lambda(h) \Big\}, \\ \widehat{\psi}_{ij,T}(u,h) &= \sum_{t=1}^{T} w_{t,T}(u,h)(\widehat{Y}_{it} - \widehat{Y}_{jt}). \end{aligned} \tag{6.1}$$

- Auxiliary statistic that can be regarded as the version of our multiscale statistic $\widehat{\Psi}_{n,T}$ under $H_0$ (defined in (4.1)):

$$\begin{aligned}
\widehat{\Phi}_{n,T} &= \max_{1\le i<j\le n} \widehat{\Phi}_{ij,T}, \\
\widehat{\Phi}_{ij,T} &= \max_{(u,h)\in\mathcal{G}_T} \Big\{ \Big| \frac{\widehat{\phi}_{ij,T}(u,h)}{\{\widehat{\sigma}_i^2+\widehat{\sigma}_j^2\}^{1/2}} \Big| - \lambda(h) \Big\}, \qquad (6.2) \\
\widehat{\phi}_{ij,T}(u,h) &= \sum_{t=1}^{T} w_{t,T}(u,h) \big\{ (\varepsilon_{it}-\bar{\varepsilon}_i) + (\boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_i)^\top (\mathbf{X}_{it}-\bar{\mathbf{X}}_i) \\
&\qquad - (\varepsilon_{jt}-\bar{\varepsilon}_j) - (\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j)^\top (\mathbf{X}_{jt}-\bar{\mathbf{X}}_j) \big\}.
\end{aligned}$$

- Intermediate statistic that is close to $\widehat{\Phi}_{n,T}$ but is based on the kernel averages that do not include the covariates:

$$\begin{aligned}
\widehat{\widehat{\Phi}}_{n,T} &= \max_{1\le i<j\le n} \widehat{\widehat{\Phi}}_{ij,T}, \\
\widehat{\widehat{\Phi}}_{ij,T} &= \max_{(u,h)\in\mathcal{G}_T} \Big\{ \Big| \frac{\widehat{\widehat{\phi}}_{ij,T}(u,h)}{\{\widehat{\widehat{\sigma}}_i^2+\widehat{\widehat{\sigma}}_j^2\}^{1/2}} \Big| - \lambda(h) \Big\}, \qquad (6.3) \\
\widehat{\widehat{\phi}}_{ij,T}(u,h) &= \sum_{t=1}^{T} w_{t,T}(u,h) \big\{ (\varepsilon_{it}-\bar{\varepsilon}_i) - (\varepsilon_{jt}-\bar{\varepsilon}_j) \big\}.
\end{aligned}$$

- Auxiliary statistic that has the same distribution as $\widehat{\widehat{\Phi}}_{n,T}$ for each $T = 1, 2, \ldots$

$$\begin{aligned}
\widetilde{\Phi}_{n,T} &= \max_{1\le i<j\le n} \widetilde{\Phi}_{ij,T}, \\
\widetilde{\Phi}_{ij,T} &= \max_{(u,h)\in\mathcal{G}_T} \Big\{ \Big| \frac{\widetilde{\phi}_{ij,T}(u,h)}{\{\widetilde{\sigma}_i^2+\widetilde{\sigma}_j^2\}^{1/2}} \Big| - \lambda(h) \Big\}, \qquad (6.4) \\
\widetilde{\phi}_{ij,T}(u,h) &= \sum\nolimits_{t=1}^{T} w_{t,T}(u,h) \big\{ (\widetilde{\varepsilon}_{it}-\bar{\widetilde{\varepsilon}}_i) - (\widetilde{\varepsilon}_{jt}-\bar{\widetilde{\varepsilon}}_j) \big\}
\end{aligned}$$

  with $[\widetilde{\varepsilon}_{i1}, \ldots, \widetilde{\varepsilon}_{iT}] \stackrel{\mathcal{D}}{=} [\varepsilon_{i1}, \ldots, \varepsilon_{iT}]$ for each $i$ and $T$.

- The Gaussian statistic that is used to calculate the critical values (defined in (3.5)):

$$\begin{aligned}
\Phi_{n,T} &= \max_{1\le i<j\le n} \Phi_{ij,T}, \\
\Phi_{ij,T} &= \max_{(u,h)\in\mathcal{G}_T} \Big\{ \Big| \frac{\phi_{ij,T}(u,h)}{(\sigma_i^2+\sigma_j^2)^{1/2}} \Big| - \lambda(h) \Big\}, \qquad (6.5) \\
\phi_{ij,T}(u,h) &= \sum\nolimits_{t=1}^{T} w_{t,T}(u,h) \big\{ \sigma_i(Z_{it}-\bar{Z}_i) - \sigma_j(Z_{jt}-\bar{Z}_j) \big\}.
\end{aligned}$$

## 6.2 Proof of Theorem 4.1

We will build the proof of Theorem 4.1 on the auxiliary results derived below. The steps of the proof are as follows.

1. First, we introduce the intermediate statistic $\widehat{\widehat{\Phi}}_{n,T}$ that can be regarded as the version of the statistics $\widehat{\Phi}_{n,T}$ but without the regressors and in Propositions A.1 and A.2 we show that this intermediate statistic is close to $\widehat{\Phi}_{n,T}$, i.e. there exists a sequence of positive numbers $\gamma_{n,T}$ that converges to 0 as $T \to \infty$ such that for all $x \in \mathbb{R}$

$$\begin{aligned}\mathbb{P}\Big(\widehat{\widehat{\Phi}}_{n,T} \leq x - \gamma_{n,T}\Big) - \mathbb{P}\Big(\big|\widehat{\widehat{\Phi}}_{n,T} - \widehat{\Phi}_{n,T}\big| > \gamma_{n,T}\Big) &\leq \mathbb{P}(\widehat{\Phi}_{n,T} \leq x) \\ &\leq \mathbb{P}\Big(\widehat{\widehat{\Phi}}_{n,T} \leq x + \gamma_{n,T}\Big) + \mathbb{P}\Big(\big|\widehat{\widehat{\Phi}}_{n,T} - \widehat{\Phi}_{n,T}\big| > \gamma_{n,T}\Big),\end{aligned}$$

   where

$$\mathbb{P}\Big(\big|\widehat{\widehat{\Phi}}_{n,T} - \widehat{\Phi}_{n,T}\big| > \gamma_{n,T}\Big) = o(1).$$

2. Second, by Proposition A.3, there exist statistics $\widetilde{\Phi}_{n,T}$ for $T = 1, 2, \ldots$ which are distributed as $\widehat{\widehat{\Phi}}_{n,T}$ for any $T \geq 1$ and which have the property that

$$\big|\widetilde{\Phi}_{n,T} - \Phi_{n,T}\big| = o_p\Big(\frac{T^{1/q}}{\sqrt{T h_{\min}}} + \rho_T \sqrt{\log T}\Big),$$

   where $\Phi_{n,T}$ is the Gaussian statistic defined in (3.5). This approximation result allows us to replace the multiscale statistic $\widehat{\widehat{\Phi}}_{n,T}$ by an identically distributed version $\widetilde{\Phi}_{n,T}$ which is close to $\Phi_{n,T}$.

3. Then, by Proposition A.5 we show that $\Phi_{n,T}$ does not concentrate too strongly in small regions of the form $[x - \delta_T, x + \delta_T]$ with $\delta_T$ converging to zero. Or, in other words, it holds that

$$\sup_{x \in \mathbb{R}} \mathbb{P}\big(|\Phi_{n,T} - x| \leq \delta_T\big) = o(1),$$

   where $\delta_T = T^{1/q}/\sqrt{T h_{\min}} + \rho_T \sqrt{\log T}$.

4. In the fourth step we make use of Lemma A.6 to show that

$$\sup_{x \in \mathbb{R}} \big|\mathbb{P}(\widehat{\widehat{\Phi}}_{n,T} \leq x) - \mathbb{P}(\Phi_{n,T} \leq x)\big| = o(1).$$

   This statement directly follows from the previous two steps and the fact that $\widetilde{\Phi}_{n,T}$ is distributed as $\widehat{\widehat{\Phi}}_{n,T}$ for any $n \geq 2$, $T \geq 1$.

5. And finally, by the means of Proposition A.7 we prove that

$$\sup_{x \in \mathbb{R}} \big|\mathbb{P}(\widehat{\Phi}_{n,T} \leq x) - \mathbb{P}(\Phi_{n,T} \leq x)\big| = o(1),$$

   which immediately implies the statement of Theorem 4.1.

## Step 1

The auxiliary statistic $\widehat{\Phi}_{n,T}$ defined in Section **??** (which is equal to our multiscale statistics $\widehat{\Psi}_{n,T}$ under the null hypothesis) heavily depends on the known covariates $\mathbf{X}_{it}$, whereas the Gaussian version $\Phi_{n,T}$ does not. This is the reason why we need to introduce additional intermediate test statistic without the covariates that connects $\widehat{\Phi}_{n,T}$ and $\Phi_{n,T}$.

We do it in the following way. For each $i$ and $j$, consider the kernel averages

$$\widehat{\widehat{\phi}}_{ij,T}(u,h) = \sum_{t=1}^{T} w_{t,T}(u,h)\big\{(\varepsilon_{it} - \bar{\varepsilon}_i) - (\varepsilon_{jt} - \bar{\varepsilon}_j)\big\}.$$

The intermediate statistic is then defined as

$$\widehat{\widehat{\Phi}}_{n,T} = \max_{1\le i<j\le n} \widehat{\widehat{\Phi}}_{ij,T} \quad \text{with}$$

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

with $\widehat{\widehat{\sigma}}_i^2$ being an estmator 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 unobserved sample $\{\widehat{\widehat{Y}}_{it} : 1 \le t \le T\}$. We thus regard $\widehat{\widehat{\sigma}}_i^2 = \widehat{\widehat{\sigma}}_i^2(\widehat{\widehat{Y}}_{i1}, \ldots, \widehat{\widehat{Y}}_{iT})$ as a function of the variables $\widehat{\widehat{Y}}_{it}$ for $1 \le t \le T$. As with the estimator $\widehat{\sigma}_i^2$, we assume that $\widehat{\widehat{\sigma}}_i^2 = \sigma_i^2 + o_p(\rho_T)$ with $\rho_T = o(\sqrt{h_{\min}}/\log T)$.

This statistic can thus be regarded as an approximation of the statistic $\widehat{\Phi}_{n,T}$. Heuristically, the kernel averages $\widehat{\widehat{\phi}}_{ij,T}(u,h)$ are close to the kernel averages $\widehat{\phi}_{ij,T}(u,h)$ because of the properties of our estimators $\widehat{\boldsymbol{\beta}}_i$, $\widehat{\widehat{\sigma}}_i^2$ and assumptions on $\mathbf{X}_{it}$. In the following two propositions we prove it formally.

**Proposition A.1.** *For any $x \in \mathbb{R}$ and any $\gamma > 0$, we have*

$$\begin{aligned} \mathbb{P}\Big(\widehat{\widehat{\Phi}}_{n,T} \le x - \gamma\Big) - \mathbb{P}\Big(\big|\widehat{\widehat{\Phi}}_{n,T} - \widehat{\Phi}_{n,T}\big| > \gamma\Big) &\le \mathbb{P}(\widehat{\Phi}_{n,T} \le x) \\ &\le \mathbb{P}\Big(\widehat{\widehat{\Phi}}_{n,T} \le x + \gamma\Big) + \mathbb{P}\Big(\big|\widehat{\widehat{\Phi}}_{n,T} - \widehat{\Phi}_{n,T}\big| > \gamma\Big). \end{aligned} \tag{6.6}$$

**Proof of Proposition A.1**. From the law of total probability and the monotonic property of the probability function, we have

$$\begin{aligned} \mathbb{P}(\widehat{\Phi}_{n,T} \le x) &= \mathbb{P}\Big(\widehat{\Phi}_{n,T} \le x, \big|\widehat{\widehat{\Phi}}_{n,T} - \widehat{\Phi}_{n,T}\big| \le \gamma\Big) + \mathbb{P}\Big(\widehat{\Phi}_{n,T} \le x, \big|\widehat{\widehat{\Phi}}_{n,T} - \widehat{\Phi}_{n,T}\big| > \gamma\Big) \\ &\le \mathbb{P}\Big(\widehat{\Phi}_{n,T} \le x, \widehat{\Phi}_{n,T} - \gamma \le \widehat{\widehat{\Phi}}_{n,T} \le \widehat{\Phi}_{n,T} + \gamma\Big) + \mathbb{P}\Big(\big|\widehat{\widehat{\Phi}}_{n,T} - \widehat{\Phi}_{n,T}\big| > \gamma\Big) \\ &\le \mathbb{P}\Big(\widehat{\widehat{\Phi}}_{n,T} \le x + \gamma\Big) + \mathbb{P}\Big(\big|\widehat{\widehat{\Phi}}_{n,T} - \widehat{\Phi}_{n,T}\big| > \gamma\Big). \end{aligned}$$

Analogously,

$$\mathbb{P}(\widehat{\widehat{\Phi}}_{n,T} \le x - \gamma) \le \mathbb{P}\Big(\widehat{\Phi}_{n,T} \le x\Big) + \mathbb{P}\Big(\big|\widehat{\widehat{\Phi}}_{n,T} - \widehat{\Phi}_{n,T}\big| > \gamma\Big).$$

Combining these two inequalities together, we arrive at the desired result. □

The aim of the next proposition is to determine the value of $\gamma_{n,T}$, that may depend on $n$ and $T$, such that the difference between the distributions of $\widehat{\Phi}_{n,T}$ and $\widehat{\widehat{\Phi}}_{n,T}$ is not too big. In other words,

**Proposition A.2.** *There exists a sequence of positive random numbers $\{\gamma_{n,T}\}_T$, that converges to 0 as $T \to \infty$, such that*

$$\mathbb{P}\Big(\big|\widehat{\widehat{\Phi}}_{n,T} - \widehat{\Phi}_{n,T}\big| > \gamma_{n,T}\Big) = o(1). \tag{6.7}$$

**Proof of Proposition A.2**. Straightforward calculations yield that

$$\begin{aligned}\big|\widehat{\widehat{\Phi}}_{n,T} - \widehat{\Phi}_{n,T}\big| \leq & \max_{1\leq i<j\leq n} \max_{(u,h)\in\mathcal{G}_T} \left| \frac{\widehat{\widehat{\phi}}_{ij,T}(u,h)}{\big(\widehat{\widehat{\sigma}}_i^2 + \widehat{\widehat{\sigma}}_j^2\big)^{1/2}} - \frac{\widehat{\widehat{\phi}}_{ij,T}(u,h)}{(\widehat{\sigma}_i^2 + \widehat{\sigma}_j^2)^{1/2}} \right| \\ & + \max_{1\leq i<j\leq n} \max_{(u,h)\in\mathcal{G}_T} \left| \frac{\widehat{\widehat{\phi}}_{ij,T}(u,h)}{(\widehat{\sigma}_i^2 + \widehat{\sigma}_j^2)^{1/2}} - \frac{\widehat{\phi}_{ij,T}(u,h)}{\{\widehat{\sigma}_i^2 + \widehat{\sigma}_j^2\}^{1/2}} \right|.\end{aligned}$$

Obviously,

$$\begin{aligned}&\max_{1\leq i<j\leq n} \max_{(u,h)\in\mathcal{G}_T} \left| \frac{\widehat{\widehat{\phi}}_{ij,T}(u,h)}{\big(\widehat{\widehat{\sigma}}_i^2 + \widehat{\widehat{\sigma}}_j^2\big)^{1/2}} - \frac{\widehat{\widehat{\phi}}_{ij,T}(u,h)}{(\widehat{\sigma}_i^2 + \widehat{\sigma}_j^2)^{1/2}} \right| \\ &\quad \leq \max_{1\leq i<j\leq n} \left( \big|\{\widehat{\widehat{\sigma}}_i^2 + \widehat{\widehat{\sigma}}_j^2\}^{-1/2} - \{\widehat{\sigma}_i^2 + \widehat{\sigma}_j^2\}^{-1/2}\big| \max_{(u,h)\in\mathcal{G}_T} \left|\widehat{\widehat{\phi}}_{ij,T}(u,h)\right| \right)\end{aligned}$$

and

$$\begin{aligned}&\max_{1\leq i<j\leq n} \max_{(u,h)\in\mathcal{G}_T} \left| \frac{\widehat{\widehat{\phi}}_{ij,T}(u,h)}{(\widehat{\sigma}_i^2 + \widehat{\sigma}_j^2)^{1/2}} - \frac{\widehat{\phi}_{ij,T}(u,h)}{\{\widehat{\sigma}_i^2 + \widehat{\sigma}_j^2\}^{1/2}} \right| \\ &\quad \leq \max_{1\leq i<j\leq n} \left( \{\widehat{\sigma}_i^2 + \widehat{\sigma}_j^2\}^{-1/2} \max_{(u,h)\in\mathcal{G}_T} \left|\widehat{\widehat{\phi}}_{ij,T}(u,h) - \widehat{\phi}_{ij,T}(u,h)\right| \right).\end{aligned}$$

Then, the difference of the kernel averages is

$$\begin{aligned}\left|\widehat{\widehat{\phi}}_{ij,T}(u,h) - \widehat{\phi}_{ij,T}(u,h)\right| &= \Big| \sum_{t=1}^{T} w_{t,T}(u,h)\big\{(\boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_i)^\top(\mathbf{X}_{it} - \bar{\mathbf{X}}_i) - (\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j)^\top(\mathbf{X}_{jt} - \bar{\mathbf{X}}_j)\big\}\Big| \\ &\leq \Big|(\boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_i)^\top \sum_{t=1}^{T} w_{t,T}(u,h)\mathbf{X}_{it}\Big| + \big|(\boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_i)^\top \bar{\mathbf{X}}_i\big| \Big|\sum_{t=1}^{T} w_{t,T}(u,h)\Big| \\ &\quad + \Big|(\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j)^\top \sum_{t=1}^{T} w_{t,T}(u,h)\mathbf{X}_{jt}\Big| + \big|(\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j)^\top \bar{\mathbf{X}}_j\big| \Big|\sum_{t=1}^{T} w_{t,T}(u,h)\Big|\end{aligned}$$

Hence,

$$\begin{aligned}\left|\widehat{\widehat{\Phi}}_{n,T}-\widehat{\Phi}_{n,T}\right| \leq & \max_{1\leq i<j\leq n}\left|\{\widehat{\widehat{\sigma}}_i^2+\widehat{\widehat{\sigma}}_j^2\}^{-1/2}-\{\widehat{\sigma}_i^2+\widehat{\sigma}_j^2\}^{-1/2}\right| \max_{1\leq i<j\leq n}\max_{(u,h)\in\mathcal{G}_T}\left|\widehat{\widehat{\phi}}_{ij,T}(u,h)\right| \\ & + 2\max_{1\leq i<j\leq n}\{\widehat{\sigma}_i^2+\widehat{\sigma}_j^2\}^{-1/2}\max_{1\leq i\leq n}\max_{(u,h)\in\mathcal{G}_T}\left|(\boldsymbol{\beta}_i-\widehat{\boldsymbol{\beta}}_i)^\top\sum_{t=1}^{T}w_{t,T}(u,h)\mathbf{X}_{it}\right| \\ & + 2\max_{1\leq i<j\leq n}\{\widehat{\sigma}_i^2+\widehat{\sigma}_j^2\}^{-1/2}\max_{1\leq i\leq n}\left|(\boldsymbol{\beta}_i-\widehat{\boldsymbol{\beta}}_i)^\top\bar{\mathbf{X}}_i\right|\max_{(u,h)\in\mathcal{G}_T}\left|\sum_{t=1}^{T}w_{t,T}(u,h)\right|. \end{aligned} \tag{6.8}$$

We consider each of the three summands separately.
We start by looking at the first summand in (6.8). Since $\widehat{\widehat{\sigma}}_i^2=\sigma_i^2+o_P(\rho_T)$ and $\widehat{\sigma}_i^2=\sigma_i^2+o_P(\rho_T)$ by our assumptions, we have that

$$\max_{1\leq i<j\leq n}\left|\{\widehat{\widehat{\sigma}}_i^2+\widehat{\widehat{\sigma}}_j^2\}^{-1/2}-\{\widehat{\sigma}_i^2+\widehat{\sigma}_j^2\}^{-1/2}\right|=o_P(\rho_T). \tag{6.9}$$

Then, we investigate $\max_{(u,h)\in\mathcal{G}_T}\left|\widehat{\widehat{\phi}}_{ij,T}(u,h)\right|$. Since $\widehat{\widehat{\phi}}_{ij,T}(u,h)$ has the same distribution as $\widetilde{\phi}_{ij,T}(u,h)$ for all $1\leq i<j\leq n$ and all $(u,h)\in\mathcal{G}_T$, we now look at the distribution of $\max_{(u,h)\in\mathcal{G}_T}\left|\widetilde{\phi}_{ij,T}(u,h)\right|$ instead.

$$\mathbb{P}\left(\max_{(u,h)\in\mathcal{G}_T}\left|\widehat{\widehat{\phi}}_{ij,T}(u,h)\right|\leq c_T\right)=\mathbb{P}\left(\max_{(u,h)\in\mathcal{G}_T}\left|\widetilde{\phi}_{ij,T}(u,h)\right|\leq c_T\right).$$

In bounding this probability, we can use the strategy from the second part of the proof of Propostion A.1. Similarly, we have

$$\begin{aligned}&\mathbb{P}\left(\max_{(u,h)\in\mathcal{G}_T}|\phi_{ij,T}(u,h)|\leq c_T/2\right) \\ &\leq \mathbb{P}\left(\max_{(u,h)\in\mathcal{G}_T}\left|\widetilde{\phi}_{ij,T}(u,h)\right|\leq c_T\right)+\mathbb{P}\left(\left|\max_{(u,h)\in\mathcal{G}_T}\left|\widetilde{\phi}_{ij,T}(u,h)\right|-\max_{(u,h)\in\mathcal{G}_T}|\phi_{ij,T}(u,h)|\right|>\frac{c_T}{2}\right) \\ &\leq \mathbb{P}\left(\max_{(u,h)\in\mathcal{G}_T}\left|\widetilde{\phi}_{ij,T}(u,h)\right|\leq c_T\right)+\mathbb{P}\left(\max_{(u,h)\in\mathcal{G}_T}\left|\widetilde{\phi}_{ij,T}(u,h)-\phi_{ij,T}(u,h)\right|>\frac{c_T}{2}\right).\end{aligned}$$

Hence,

$$\begin{aligned}&\mathbb{P}\left(\max_{(u,h)\in\mathcal{G}_T}\left|\widehat{\widehat{\phi}}_{ij,T}(u,h)\right|\leq c_T\right)=\mathbb{P}\left(\max_{(u,h)\in\mathcal{G}_T}\left|\widetilde{\phi}_{ij,T}(u,h)\right|\leq c_T\right) \\ &\quad\geq \mathbb{P}\left(\max_{(u,h)\in\mathcal{G}_T}|\phi_{ij,T}(u,h)|\leq c_T/2\right)-\mathbb{P}\left(\max_{(u,h)\in\mathcal{G}_T}\left|\widetilde{\phi}_{ij,T}(u,h)-\phi_{ij,T}(u,h)\right|>\frac{c_T}{2}\right).\end{aligned} \tag{6.10}$$

Now we will need one result that we will prove further: by (6.31) we have

$$\max_{(u,h)\in\mathcal{G}_T}\left|\widetilde{\phi}_{ij,T}(u,h)-\phi_{ij,T}(u,h)\right|=o_P\left(\frac{T^{1/q}}{\sqrt{Th_{\min}}}\right).$$

Furthermore, $\phi_{ij,T}(u,h)$ is distributed as $N(0,\sigma_i^2+\sigma_j^2)$ for all $(u,h)\in\mathcal{G}_T$ and all $1\le i<j\le n$ and $|\mathcal{G}_T|=O(T^\theta)$ for some large but fixed constant $\theta$ by Assumption (C12). By the standard results from the probability theory, we know that

$$\max_{(u,h)\in\mathcal{G}_T}|\phi_{ij,T}(u,h)|=O_P(\sqrt{\log T}).$$

Since $T^{1/q}/\sqrt{Th_{\min}}\ll\sqrt{\log T}$, we can take $c_T=o(\sqrt{\log T})$ in (6.10) to get the following:

$$\begin{aligned}
&\mathbb{P}\left(\max_{(u,h)\in\mathcal{G}_T}\left|\widehat{\widetilde{\phi}}_{ij,T}(u,h)\right|\le c_T\right)\\
&\ge\mathbb{P}\left(\max_{(u,h)\in\mathcal{G}_T}|\phi_{ij,T}(u,h)|\le\frac{c_T}{2}\right)-\mathbb{P}\left(\max_{(u,h)\in\mathcal{G}_T}\left|\widetilde{\phi}_{ij,T}(u,h)-\phi_{ij,T}(u,h)\right|>\frac{c_T}{2}\right)\\
&=1-o(1)-o(1)\\
&=1-o(1),
\end{aligned}$$

which means that

$$\max_{(u,h)\in\mathcal{G}_T}\left|\widehat{\widetilde{\phi}}_{ij,T}(u,h)\right|=o_P(\sqrt{\log T}). \tag{6.11}$$

Combining (6.9) and (6.11), we get the following:

$$\begin{aligned}
\max_{1\le i<j\le n}\left|\{\widehat{\widetilde{\sigma}}_i^2+\widehat{\widetilde{\sigma}}_j^2\}^{-1/2}-\{\widehat{\sigma}_i^2+\widehat{\sigma}_j^2\}^{-1/2}\right|\max_{1\le i<j\le n}\max_{(u,h)\in\mathcal{G}_T}\left|\widehat{\widetilde{\phi}}_{ij,T}(u,h)\right|&\\
=o_P(\rho_T)\cdot o_P(\sqrt{\log T})&\\
=o_P(1)&
\end{aligned} \tag{6.12}$$

since by our assumption $\rho_T=O(\sqrt{h_{\min}}/\log T)$.
Now we evaluate the second summand in (6.8).
First, by our assumptions $\widehat{\sigma}_i^2=\sigma_i^2+o_P(\rho_T)$. Moreover, for all $i\in\{1,\dots,n\}$ we know $\sigma_i^2\neq 0$. Hence,

$$\max_{1\le i<j\le n}\{\widehat{\sigma}_i^2+\widehat{\sigma}_j^2\}^{-1/2}=O_P(1). \tag{6.13}$$

Then, by Theorem 4.2, we know that

$$\boldsymbol{\beta}_i-\widehat{\boldsymbol{\beta}}_i=O_P(1/\sqrt{T}). \tag{6.14}$$

Now consider $\sum_{t=1}^T w_{t,T}(u,h)\mathbf{X}_{it}$. Without loss of generality, we can regard the covariates $\mathbf{X}_{it}$ to be scalars $X_{it}$, not vectors. The proof in case of vectors proceeds analogously. By construction the weights $w_{t,T}(u,h)$ are not equal to 0 if and only if $T(u-h)\le t\le T(u+h)$. We can use this fact to rewrite

$$\left|\sum_{t=1}^T w_{t,T}(u,h)X_{it}\right|=\left|\sum_{t=\lfloor T(u-h)\rfloor}^{\lceil T(u+h)\rceil} w_{t,T}(u,h)X_{it}\right|.$$

Note that

$$\sum_{t=\lfloor T(u-h)\rfloor}^{\lceil T(u+h)\rceil} w_{t,T}^2(u,h) = \sum_{t=1}^{T} w_{t,T}^2(u,h) = \sum_{t=1}^{T} \frac{\Lambda_{t,T}^2(u,h)}{\sum_{s=1}^{T}\Lambda_{s,T}^2(u,h)} = 1. \tag{6.15}$$

Denoting by $D_{T,u,h}$ the number of integers between $\lfloor T(u-h)\rfloor$ and $\lceil T(u+h)\rceil$ incl. (with obvious bounds $2Th \le D_{T,u,h} \le 2Th+2$), we can normalize the weights as follows:

$$\sum_{t=\lfloor T(u-h)\rfloor}^{\lceil T(u+h)\rceil} \left(\sqrt{D_{T,u,h}}\cdot w_{t,T}(u,h)\right)^2 = D_{T,u,h}.$$

According to Theorem A.1 (Theorem 2(ii) in Wu et al. (2016)), if we denote the weights from the theorem as $a_t = \sqrt{D_{T,u,h}}\cdot w_{t,T}(u,h)$ we can bound the probability as follows:

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{t=\lfloor T(u-h)\rfloor}^{\lceil T(u+h)\rceil} \sqrt{D_{T,u,h}}\cdot w_{t,T}(u,h)X_{it}\right| \ge x\right) \\
&\quad\le C_1 \frac{\left(\sum_{t=\lfloor T(u-h)\rfloor}^{\lceil T(u+h)\rceil} |\sqrt{D_{T,u,h}}\cdot w_{t,T}(u,h)|^{q'}\right)||X_{i\cdot}||_{q',\alpha}^{q'}}{x^{q'}} + C_2\exp\left(-\frac{C_3x^2}{D_{T,u,h}||X_{i\cdot}||_{2,\alpha}^2}\right),
\end{aligned} \tag{6.16}$$

where $||X_{i\cdot}||_{q,\alpha}^q$ is the dependence adjusted norm as defined by Definition A.1. Taking any $\delta > 0$ and applying Boole's inequality and (6.16) subsequently, we get

$$\begin{aligned}
&\mathbb{P}\left(\max_{(u,h)\in\mathcal{G}_T}\left|\sum_{t=\lfloor T(u-h)\rfloor}^{\lceil T(u+h)\rceil} w_{t,T}(u,h)X_{it}\right| \ge \delta\sqrt{T}\right) \\
&\le \sum_{(u,h)\in\mathcal{G}_T}\mathbb{P}\left(\left|\sum_{t=\lfloor T(u-h)\rfloor}^{\lceil T(u+h)\rceil} w_{t,T}(u,h)X_{it}\right| \ge \delta\sqrt{T}\right) \\
&= \sum_{(u,h)\in\mathcal{G}_T}\mathbb{P}\left(\left|\sum_{t=\lfloor T(u-h)\rfloor}^{\lceil T(u+h)\rceil} \sqrt{D_{T,u,h}}\cdot w_{t,T}(u,h)X_{it}\right| \ge \delta\sqrt{D_{T,u,h}T}\right) \\
&\le \sum_{(u,h)\in\mathcal{G}_T}\left[C_1\frac{(\sqrt{D_{T,u,h}})^{q'}\left(\sum|w_{t,T}(u,h)|^{q'}\right)||X_{i\cdot}||_{q',\alpha}^{q'}}{\left(\delta\sqrt{D_{T,u,h}T}\right)^{q'}} + C_2\exp\left(-\frac{C_3\left(\delta\sqrt{D_{T,u,h}T}\right)^2}{D_{T,u,h}||X_{i\cdot}||_{2,\alpha}^2}\right)\right] \\
&= \sum_{(u,h)\in\mathcal{G}_T}\left[C_1\frac{\left(\sum|w_{t,T}(u,h)|^{q'}\right)||X_{i\cdot}||_{q',\alpha}^{q'}}{\left(\delta\sqrt{T}\right)^{q'}} + C_2\exp\left(-\frac{C_3\delta^2T}{||X_{i\cdot}||_{2,\alpha}^2}\right)\right] \\
&\le C_1\frac{T^\theta||X_{i\cdot}||_{q',\alpha}^{q'}}{T^{q'/2}\cdot\delta^{q'}}\max_{(u,h)\in\mathcal{G}_T}\left(\sum\nolimits_{t=\lfloor T(u-h)\rfloor}^{\lceil T(u+h)\rceil}|w_{t,T}(u,h)|^{q'}\right) + C_2T^\theta\exp\left(-\frac{C_3\delta^2T}{||X_{i\cdot}||_{2,\alpha}^2}\right) \\
&= C\frac{T^{\theta-q'/2}}{\delta^{q'}} + CT^\theta\exp\left(-CT\delta^2\right).
\end{aligned}$$

where the symbol $C$ denotes a universal real constant that does not depend neither on $T$ nor on $\delta$ and that takes a different value on each occurrence. Here in the last equality we used the following facts:

1. $||X_{i\cdot}||_{q',\alpha}^{q'} = \sup_{t\geq 0}(t+1)^{\alpha}\sum_{s=t}^{\infty}\delta_{q'}(H_i,s) < \infty$ holds true since $\sum_{s=t}^{\infty}\delta_{q'}(H_i,s) = O(t^{-\alpha})$ by Assumption (C8);

2. $\max_{(u,h)\in\mathcal{G}_T}\left(\sum_{t=\lfloor T(u-h)\rfloor}^{\lceil T(u+h)\rceil}|w_{t,T}(u,h)|^{q'}\right) \leq 1$ because for every $x\in[0,1]$ we have $0 \leq x^{q'/2} \leq x \leq 1$. Thus, since $\sum_{t=1}^{T} w_{t,T}^2(u,h) = 1$ by (6.15) we have $0 \leq w_{t,T}^2(u,h) \leq 1$ for all $t\in\{1,\ldots,T\}$ and all $(u,h)\in\mathcal{G}_T$, we get

$$0 \leq |w_{t,T}(u,h)|^{q'} = (w_{t,T}^2(u,h))^{q'/2} \leq w_{t,T}^2(u,h) \leq 1.$$

   This leads to a bound

$$\max_{(u,h)\in\mathcal{G}_T}\left(\sum_{t=\lfloor T(u-h)\rfloor}^{\lceil T(u+h)\rceil}|w_{t,T}(u,h)|^{q'}\right) \leq \max_{(u,h)\in\mathcal{G}_T}\left(\sum_{t=\lfloor T(u-h)\rfloor}^{\lceil T(u+h)\rceil} w_{t,T}^2(u,h)\right) = 1.$$

3. $||X_{i\cdot}||_{2,\alpha}^2 < \infty$ (follows from 1).

By Assumption (C6), $\theta - q'/2 < 0$ and the term on the RHS of the above inequality is converging to 0 as $T\to\infty$ for any fixed $\delta > 0$. Hence,

$$\max_{(u,h)\in\mathcal{G}_T}\left|\sum_{t=\lfloor T(u-h)\rfloor}^{\lceil T(u+h)\rceil} w_{t,T}(u,h)X_{it}\right| = o_P(\sqrt{T}),$$

and similarly,

$$\max_{(u,h)\in\mathcal{G}_T}\left|\sum_{t=\lfloor T(u-h)\rfloor}^{\lceil T(u+h)\rceil} w_{t,T}(u,h)\mathbf{X}_{it}\right| = o_P(\sqrt{T}). \tag{6.17}$$

Combining (6.13), (6.14) and (6.17), we get the following:

$$\begin{aligned}\max_{1\leq i<j\leq n}\{\widehat{\sigma}_i^2+\widehat{\sigma}_j^2\}^{-1/2}\max_{1\leq i\leq n}\max_{(u,h)\in\mathcal{G}_T}\left|(\boldsymbol{\beta}_i-\widehat{\boldsymbol{\beta}}_i)^{\top}\sum_{t=1}^{T} w_{t,T}(u,h)\mathbf{X}_{it}\right| \\ = O_P(1)\cdot O_P(1/\sqrt{T})\cdot o_P(\sqrt{T}) \\ = o_P(1).\end{aligned} \tag{6.18}$$

Now consider the third summand in (6.8). Similarly as before,

$$\max_{1\leq i<j\leq n}\{\widehat{\sigma}_i^2+\widehat{\sigma}_j^2\}^{-1/2} = O_P(1) \tag{6.19}$$

and

$$\boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_i = O_P(1/\sqrt{T}). \tag{6.20}$$

Then, by Proposition A.9 $\bar{\mathbf{X}}_i = o_P(1)$.

Finally, consider the local linear kernel weights $w_{t,T}(u,h)$ defined in (3.2). Again, by construction the weights $w_{t,T}(u,h)$ are not equal to 0 if and only if

$T(u-h) \leq t \leq T(u+h)$. We can use this fact to bound $\left|\sum_{t=1}^{T} w_{t,T}(u,h)\right|$ for all $(u,h) \in \mathcal{G}_T$ using the Cauchy-Schwarz inequality:

$$\begin{aligned}
\Big|\sum_{t=1}^{T} w_{t,T}(u,h)\Big| &= \left|\sum_{t=\lfloor T(u-h)\rfloor}^{\lceil T(u+h)\rceil} w_{t,T}(u,h)\cdot 1\right| \\
&\leq \sqrt{\sum_{t=\lfloor T(u-h)\rfloor}^{\lceil T(u+h)\rceil} w_{t,T}^2(u,h)}\sqrt{\sum_{t=\lfloor T(u-h)\rfloor}^{\lceil T(u+h)\rceil} 1^2} \\
&= \sqrt{1}\cdot\sqrt{D_{T,u,h}} \\
&\leq \sqrt{2Th+2} \\
&\leq \sqrt{2Th_{\max}+2} \\
&\leq \sqrt{T+2}.
\end{aligned} \tag{6.21}$$

Hence,

$$\max_{(u,h)\in\mathcal{G}_T}\Big|\sum_{t=1}^{T} w_{t,T}(u,h)\Big| = O(\sqrt{T}). \tag{6.22}$$

Combining (6.19), (6.20), Proposition A.9 and (6.22), we get the following:

$$\begin{aligned}
&\max_{1\leq i<j\leq n}\{\widehat{\sigma}_i^2+\widehat{\sigma}_j^2\}^{-1/2}\max_{1\leq i\leq n}\big|(\boldsymbol{\beta}_i-\widehat{\boldsymbol{\beta}}_i)^\top\bar{\mathbf{X}}_i\big|\max_{(u,h)\in\mathcal{G}_T}\Big|\sum_{t=1}^{T} w_{t,T}(u,h)\Big| \\
&\qquad = O_P(1)\cdot O_P(1/\sqrt{T})\cdot o_P(1)\cdot O(\sqrt{T}) \\
&\qquad = o_P(1).
\end{aligned} \tag{6.23}$$

Plugging (6.12), (6.18) and (6.23) in (6.8), we get that $\big|\widehat{\widehat{\Phi}}_{n,T}-\widehat{\Phi}_{n,T}\big| = o_P(1)$ and the statement of the theorem follows. □

## Step 2

The main purpose of this section is to prove that there is a version of the multiscale statistic $\widehat{\widehat{\Phi}}_{n,T}$ which is close to the Gaussian statistic $\Phi_{n,T}$ (defined in (6.5)) whose distribution is known. More specifically, we prove the following result.

**Proposition A.3.** *Under the conditions of Theorem 4.1, there exist statistics $\widetilde{\Phi}_{n,T}$ for $T = 1, 2, \ldots$ with the following two properties: (i) $\widetilde{\Phi}_{n,T}$ has the same distribution as $\widehat{\widehat{\Phi}}_{n,T}$ for any $T$, and (ii)*

$$\big|\widetilde{\Phi}_{n,T}-\Phi_{n,T}\big| = o_p\Big(\frac{T^{1/q}}{\sqrt{Th_{\min}}}+\rho_T\sqrt{\log T}\Big), \tag{6.24}$$

*where $\Phi_{n,T}$ is a Gaussian statistic as defined in* (6.5).

**Proof of Proposition A.3**. For the proof, we draw on strong approximation theory for each stationary process $\mathcal{E}_i = \{\varepsilon_{it} : 1 \leq t \leq T\}$ that fulfill the conditions (C1)–(C3). By Theorem 2.1 and Corollary 2.1 in Berkes et al. (2014), the following strong approximation result holds true: On a richer probability space, there exists a standard Brownian motion $\mathbb{B}_i$ and a sequence $\{\widetilde{\varepsilon}_{it} : t \in \mathbb{N}\}$ such that $[\widetilde{\varepsilon}_{i1}, \ldots, \widetilde{\varepsilon}_{iT}] \stackrel{\mathcal{D}}{=} [\varepsilon_{i1}, \ldots, \varepsilon_{iT}]$ for each $T$ and

$$\max_{1 \leq t \leq T} \Big| \sum_{s=1}^{t} \widetilde{\varepsilon}_{is} - \sigma_i \mathbb{B}_i(t) \Big| = o\big(T^{1/q}\big) \quad \text{a.s.}, \tag{6.25}$$

where $\sigma_i^2 = \sum_{k \in \mathbb{Z}} \mathrm{Cov}(\varepsilon_{i0}, \varepsilon_{ik})$ denotes the long-run error variance.
We apply this result for each stationary process $\mathcal{E}_i = \{\varepsilon_{it} : 1 \leq t \leq T\}$ so that each process $\widetilde{\mathcal{E}}_i = \{\widetilde{\varepsilon}_{it} : t \in \mathbb{N}\}$ is independent of $\widetilde{\mathcal{E}}_j = \{\widetilde{\varepsilon}_{jt} : t \in \mathbb{N}\}$ for $i \neq j$.
Furthermore, we define

$$\begin{aligned}
\widetilde{\Phi}_{n,T} &= \max_{1 \leq i < j \leq n} \widetilde{\Phi}_{ij,T}, \\
\widetilde{\Phi}_{ij,T} &= \max_{(u,h) \in \mathcal{G}_T} \Big\{ \Big| \frac{\widetilde{\phi}_{ij,T}(u,h)}{\{\widetilde{\sigma}_i^2 + \widetilde{\sigma}_j^2\}^{1/2}} \Big| - \lambda(h) \Big\},
\end{aligned}$$

where $\widetilde{\phi}_{ij,T}(u,h) = \sum_{t=1}^{T} w_{t,T}(u,h)\big\{(\widetilde{\varepsilon}_{it} - \bar{\widetilde{\varepsilon}}_i) - (\widetilde{\varepsilon}_{jt} - \bar{\widetilde{\varepsilon}}_j)\big\}$ and $\widetilde{\sigma}_i^2$ are the same estimators as $\widehat{\sigma}_i^2$ with $\widehat{Y}_{it} = (\boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_i)^\top \mathbf{X}_{it} + m_i(t/T) + \big(\alpha_i - \widehat{\alpha}_i\big) + \varepsilon_{it}$ replaced by $\widetilde{Y}_{it} = = (\boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_i)^\top \mathbf{X}_{it} + m_i(t/T) + \big(\alpha_i - \widehat{\alpha}_i\big) + \widetilde{\varepsilon}_{it}$ for $1 \leq t \leq T$. Since $[\widetilde{\varepsilon}_{i1}, \ldots, \widetilde{\varepsilon}_{iT}] \stackrel{\mathcal{D}}{=} [\varepsilon_{i1}, \ldots, \varepsilon_{iT}]$, we have $\sum_{\ell=-\infty}^{\infty} \mathrm{Cov}(\widetilde{\varepsilon}_{i0}, \widetilde{\varepsilon}_{i\ell}) = \sum_{\ell=-\infty}^{\infty} \mathrm{Cov}(\varepsilon_{i0}, \varepsilon_{i\ell}) = \sigma_i^2$. Hence, by construction $\widetilde{\sigma}_i^2 = \sigma_i^2 + o_P(\rho_T)$.
In addition, we let

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

with $\phi_{ij,T}(u,h)$ defined in (6.5) and $Z_{it} = \mathbb{B}_i(t) - \mathbb{B}_i(t-1)$. With this notation, we can write

$$\big|\widetilde{\Phi}_{n,T} - \Phi'_{n,T}\big| \leq \big|\widetilde{\Phi}_{n,T} - \Phi_{n,T}^{\diamond}\big| + \big|\Phi_{n,T}^{\diamond} - \Phi_{n,T}\big|. \tag{6.26}$$

First consider $|\widetilde{\Phi}_{n,T} - \Phi_{n,T}^{\diamond}|$. Straightforward calculations yield that

$$\big|\widetilde{\Phi}_{n,T} - \Phi_{n,T}^{\diamond}\big| \leq \max_{1 \leq i < j \leq n} \Big( \{\widetilde{\sigma}_i^2 + \widetilde{\sigma}_j^2\}^{-1/2} \max_{(u,h) \in \mathcal{G}_T} \big|\widetilde{\phi}_{ij,T}(u,h) - \phi_{ij,T}(u,h)\big| \Big). \tag{6.27}$$

Using summation by parts, ($\sum_{i=1}^{n} a_i b_i = \sum_{i=1}^{n-1} A_i(b_i - b_{i+1}) + A_n b_n$ with $A_j = \sum_{j=1}^{i} a_j$) we further obtain that

$$\begin{aligned}
&\big|\widetilde{\phi}_{ij,T}(u,h) - \phi_{ij,T}(u,h)\big| \\
&\quad = \Big| \sum_{t=1}^{T} w_{t,T}(u,h)\big\{(\widetilde{\varepsilon}_{it} - \bar{\widetilde{\varepsilon}}_i) - (\widetilde{\varepsilon}_{jt} - \bar{\widetilde{\varepsilon}}_j) - \sigma_i(Z_{it} - \bar{Z}_i) + \sigma_j(Z_{jt} - \bar{Z}_j)\big\} \Big| \\
&\quad = \Big| \sum_{t=1}^{T-1} A_{ij,t}\big(w_{t,T}(u,h) - w_{t+1,T}(u,h)\big) + A_{ij,T} w_{T,T}(u,h) \Big|,
\end{aligned}$$

where

$$A_{ij,t} = \sum_{s=1}^{t} \big\{ (\widetilde{\varepsilon}_{is} - \bar{\widetilde{\varepsilon}}_i) - (\widetilde{\varepsilon}_{js} - \bar{\widetilde{\varepsilon}}_j) - \sigma_i(Z_{it} - \bar{Z}_i) + \sigma_j(Z_{jt} - \bar{Z}_j) \big\}.$$

Note that by construction $A_{ij,T} = 0$ for all pairs $(i, j)$. Denoting

$$W_T(u,h) = \sum_{t=1}^{T-1} |w_{t+1,T}(u,h) - w_{t,T}(u,h)|,$$

we have

$$\big|\widetilde{\phi}_{ij,T}(u,h) - \phi_{ij,T}(u,h)\big| = \Big| \sum_{t=1}^{T-1} A_{ij,t}\big(w_{t,T}(u,h) - w_{t+1,T}(u,h)\big) \Big| \le W_T(u,h) \max_{1\le t\le T} |A_{ij,t}|. \tag{6.28}$$

Now consider $\max_{1\le t\le T} |A_{ij,t}|$. Straightforward calculations yield the following bound:

$$\begin{aligned}
\max_{1\le t\le T} |A_{ij,t}| \le & \max_{1\le t\le T} \Big| \sum_{s=1}^{t} \widetilde{\varepsilon}_{is} - \sigma_i \sum_{s=1}^{t} Z_{is} \Big| + \max_{1\le t\le T} \Big| t(\bar{\widetilde{\varepsilon}}_i - \sigma_i \bar{Z}_i) \Big| \\
& + \max_{1\le t\le T} \Big| \sum_{s=1}^{t} \widetilde{\varepsilon}_{js} - \sigma_j \sum_{s=1}^{t} Z_{js} \Big| + \max_{1\le t\le T} \Big| t(\bar{\widetilde{\varepsilon}}_j - \sigma_j \bar{Z}_j) \Big| \\
\le & 2 \max_{1\le t\le T} \Big| \sum_{s=1}^{t} \widetilde{\varepsilon}_{is} - \sigma_i \sum_{s=1}^{t} Z_{is} \Big| + 2 \max_{1\le t\le T} \Big| \sum_{s=1}^{t} \widetilde{\varepsilon}_{js} - \sigma_j \sum_{s=1}^{t} Z_{js} \Big| \\
= & 2 \max_{1\le t\le T} \Big| \sum_{s=1}^{t} \widetilde{\varepsilon}_{is} - \sigma_i \sum_{s=1}^{t} \big(\mathbb{B}_i(s) - \mathbb{B}_i(s-1)\big) \Big| \\
& + 2 \max_{1\le t\le T} \Big| \sum_{s=1}^{t} \widetilde{\varepsilon}_{js} - \sigma_j \sum_{s=1}^{t} \big(\mathbb{B}_j(s) - \mathbb{B}_j(s-1)\big) \Big| \\
= & 2 \max_{1\le t\le T} \Big| \sum_{s=1}^{t} \widetilde{\varepsilon}_{is} - \sigma_i \mathbb{B}_i(t) \Big| + 2 \max_{1\le t\le T} \Big| \sum_{s=1}^{t} \widetilde{\varepsilon}_{js} - \sigma_j \mathbb{B}_j(t) \Big|.
\end{aligned}$$

Applying the strong approximation result (6.25), we can infer that

$$\max_{1\le t\le T} |A_{ij,t}| = o_P\big(T^{1/q}\big). \tag{6.29}$$

Standard arguments show that $\max_{(u,h)\in\mathcal{G}_T} W_T(u,h) = O(1/\sqrt{Th_{\min}})$. Plugging (6.29) in (6.28) and then in (6.27), we can thus infer that

$$\begin{aligned}
\big|\widetilde{\Phi}_{n,T} - \Phi^{\diamond}_{n,T}\big| &\le \{\widetilde{\sigma}_i^2 + \widetilde{\sigma}_j^2\}^{-1/2} \max_{(u,h)\in\mathcal{G}_T} W_T(u,h) \max_{1\le i<j\le n} \max_{1\le t\le T} |A_{ij,t}| \\
&= o_P\Big(\frac{T^{1/q}}{\sqrt{Th_{\min}}}\Big).
\end{aligned} \tag{6.30}$$

Now consider $|\Phi^{\diamond}_{n,T} - \Phi_{n,T}|$. Since $\phi_{ij,T}(u,h)$ is distributed as $N(0, \sigma_i^2 + \sigma_j^2)$ for all $(u,h) \in \mathcal{G}_T$ and all $1 \leq i < j \leq n$, $|\mathcal{G}_T| = O(T^\theta)$ for some large but fixed constant $\theta$ by Assumption (C12), $n$ is fixed and $\widetilde{\sigma}_i^2 = \sigma_i^2 + o_p(\rho_T)$, we can establish that

$$\big|\Phi^{\diamond}_{n,T} - \Phi_{n,T}\big| \leq \max_{1\leq i<j\leq n} \max_{(u,h)\in\mathcal{G}_T} \Big| \frac{\phi_{ij,T}(u,h)}{\{\widetilde{\sigma}_i^2 + \widetilde{\sigma}_j^2\}^{1/2}} - \frac{\phi_{ij,T}(u,h)}{\{\sigma_i^2 + \sigma_j^2\}^{1/2}} \Big| = o_P(\rho_T \sqrt{\log T}). \tag{6.31}$$

Plugging (6.30) and (6.31) in (6.26) completes the proof. □

## Step 3

In this section, we establish some properties of the Gaussian statistic $\Phi_{n,T}$ defined in (6.5). We in particular show that $\Phi_{n,T}$ does not concentrate too strongly in small regions of the form $[x - \delta_T, x + \delta_T]$ with $\delta_T$ converging to zero.
The main technical tool for proving Proposition A.5 are anti-concentration bounds for Gaussian random vectors. The following proposition slightly generalizes anti-concentration results derived in Chernozhukov et al. (2015), in particular Theorem 3 therein.

**Proposition A.4.** *Let $(X_1, \ldots, X_p)^\top$ be a Gaussian random vector in $\mathbb{R}^p$ with $\mathbb{E}[X_j] = \mu_j$ and $\mathrm{Var}(X_j) = \sigma_j^2 > 0$ for $1 \leq j \leq p$. Define $\overline{\mu} = \max_{1\leq j\leq p} |\mu_j|$ together with $\underline{\sigma} = \min_{1\leq j\leq p} \sigma_j$ and $\overline{\sigma} = \max_{1\leq j\leq p} \sigma_j$. Moreover, set $a_p = \mathbb{E}[\max_{1\leq j\leq p}(X_j - \mu_j)/\sigma_j]$ and $b_p = \mathbb{E}[\max_{1\leq j\leq p}(X_j - \mu_j)]$. For every $\delta > 0$, it holds that*

$$\sup_{x\in\mathbb{R}} \mathbb{P}\Big( \big| \max_{1\leq j\leq p} X_j - x \big| \leq \delta \Big) \leq C\delta \big\{ \overline{\mu} + a_p + b_p + \sqrt{1 \vee \log(\underline{\sigma}/\delta)} \big\},$$

*where $C > 0$ depends only on $\underline{\sigma}$ and $\overline{\sigma}$.*

The proof of Proposition A.4 is provided in **?**.

**Proposition A.5.** *Under the conditions of Theorem 4.1, it holds that*

$$\sup_{x\in\mathbb{R}} \mathbb{P}\big( |\Phi_{n,T} - x| \leq \delta_T \big) = o(1), \tag{6.32}$$

*where $\delta_T = T^{1/q}/\sqrt{T h_{\min}} + \rho_T \sqrt{\log T}$.*

**Proof of Proposition A.5**. We write $x = (u,h)$ along with $\mathcal{G}_T = \{x : x \in \mathcal{G}_T\} = \{x_1, \ldots, x_p\}$, where $p := |\mathcal{G}_T| \leq O(T^\theta)$ for some large but fixed $\theta > 0$ by our assumptions. Moreover, for $k = 1, \ldots, p$, we set

$$U_{ij,2k-1} = \frac{\phi_{ij,T}(x_{k1}, x_{k2})}{\{\sigma_i^2 + \sigma_j^2\}^{1/2}} - \lambda(x_{k2})$$

$$U_{ij,2k} = -\frac{\phi_{ij,T}(x_{k1}, x_{k2})}{\{\sigma_i^2 + \sigma_j^2\}^{1/2}} - \lambda(x_{k2})$$

with $x_k = (x_{k1}, x_{k2})$. This notation allows us to write

$$\Phi_{n,T} = \max_{1\le i<j\le n} \max_{1\le k\le 2p} U_{ij,k} = \max_{1\le l\le (n-1)np} U_l'$$

where $(U_1', \dots, U_{(n-1)np}')^\top \in \mathbb{R}^{n(n-1)p}$ is a Gaussian random vector with the following properties: (i) $\mu_l := \mathbb{E}[U_l'] = \{\mathbb{E}[U_{ij,2k}] \text{ or } \mathbb{E}[U_{ij,2k-1}]\} = -\lambda(x_{k2})$ and thus

$$\overline{\mu} = \max_{1\le l\le (n-1)np} |\mu_l| \le C\sqrt{\log T},$$

and (ii) $\sigma_l^2 := \mathrm{Var}(U_l') = 1$ for all $1 \le l \le (n-1)np$. Hence, $a_{(n-1)np} = b_{(n-1)np}$. Moreover, as the variables $(U_l' - \mu_l)/\sigma_l$ are standard normal, we have that $a_{(n-1)np} = b_{(n-1)np} \le C\sqrt{\log((n-1)np)} \le C\sqrt{\log T}$. With this notation at hand, we can apply Proposition A.4 to obtain that

$$\sup_{x\in\mathbb{R}} \mathbb{P}\Big(\big|\Phi_{n,T} - x\big| \le \delta_T\Big) \le C\delta_T\Big[\sqrt{\log T} + \sqrt{\log(1/\delta_T)}\Big] = o(1)$$

with $\delta_T = T^{1/q}/\sqrt{Th_{\min}} + \rho_T\sqrt{\log T}$, which is the statement of Proposition A.5. □

## Step 4

**Lemma A.6.** *Let $V_T$ and $W_T$ be real-valued random variables for $T = 1, 2, \dots$ such that $V_T - W_T = o_p(\delta_T)$ with some $\delta_T = o(1)$. If*

$$\sup_{x\in\mathbb{R}} \mathbb{P}(|V_T - x| \le \delta_T) = o(1), \tag{6.33}$$

*then*

$$\sup_{x\in\mathbb{R}} \big|\mathbb{P}(V_T \le x) - \mathbb{P}(W_T \le x)\big| = o(1). \tag{6.34}$$

Proof of this lemma is provided in **?**.
Applying Lemma A.6 to $\widetilde{\Phi}_{n,T}$ and $\Phi_{n,T}$ together with the results (6.24) and (6.32) and noting the fact that $\widetilde{\Phi}_{n,T}$ is distributed as $\widehat{\widehat{\Phi}}_{n,T}$ for any $n \ge 2$, $T \ge 1$ leads us to

$$\sup_{x\in\mathbb{R}} \big|\mathbb{P}(\widehat{\widehat{\Phi}}_{n,T} \le x) - \mathbb{P}(\Phi_{n,T} \le x)\big| = o(1).$$

## Step 5

**Proposition A.7.** *Under the conditions of Theorem 4.1, it holds that*

$$\sup_{x\in\mathbb{R}} \big|\mathbb{P}(\widehat{\Phi}_{n,T} \le x) - \mathbb{P}(\Phi_{n,T} \le x)\big| = o(1). \tag{6.35}$$

**Proof of Proposition A.7**. First, we consider those $x \in \mathbb{R}$ such that $\mathbb{P}(\widehat{\Phi}_{n,T} \le x) \ge \mathbb{P}(\Phi_{n,T} \le x)$. Then by Proposition A.1 for $\gamma_{n,T} > 0$ from the Proposition

A.2 we have

$$\begin{aligned}\left|\mathbb{P}(\widehat{\Phi}_{n,T} \le x) - \mathbb{P}(\Phi_{n,T} \le x)\right| &= \mathbb{P}(\widehat{\Phi}_{n,T} \le x) - \mathbb{P}(\Phi_{n,T} \le x) \\ &\le \mathbb{P}\Big(\widehat{\widehat{\Phi}}_{n,T} \le x + \gamma_{n,T}\Big) + \mathbb{P}\Big(\big|\widehat{\widehat{\Phi}}_{n,T} - \widehat{\Phi}_{n,T}\big| > \gamma_{n,T}\Big) - \mathbb{P}\Big(\Phi_{n,T} \le x\Big) \\ &= \mathbb{P}\Big(\widehat{\widehat{\Phi}}_{n,T} \le x + \gamma_{n,T}\Big) - \mathbb{P}\Big(\Phi_{n,T} \le x + \gamma_{n,T}\Big) \\ &\quad + \mathbb{P}\Big(\Phi_{n,T} \le x + \gamma_{n,T}\Big) - \mathbb{P}\Big(\Phi_{n,T} \le x\Big) + \mathbb{P}\Big(\big|\widehat{\widehat{\Phi}}_{n,T} - \widehat{\Phi}_{n,T}\big| > \gamma_{n,T}\Big) \\ &\le \mathbb{P}\Big(\widehat{\widehat{\Phi}}_{n,T} \le x + \gamma_{n,T}\Big) - \mathbb{P}\Big(\Phi_{n,T} \le x + \gamma_{n,T}\Big) \\ &\quad + \mathbb{P}\Big(\big|\Phi_{n,T} - x\big| \le \gamma_{n,T}\Big) + \mathbb{P}\Big(\big|\widehat{\widehat{\Phi}}_{n,T} - \widehat{\Phi}_{n,T}\big| > \gamma_{n,T}\Big).\end{aligned}$$

Now consider such $x \in \mathbb{R}$ that $\mathbb{P}(\widehat{\Phi}_{n,T} \le x) < \mathbb{P}(\Phi_{n,T} \le x)$. Analogously,

$$\begin{aligned}\left|\mathbb{P}(\widehat{\Phi}_{n,T} \le x) - \mathbb{P}(\Phi_{n,T} \le x)\right| \le &\mathbb{P}\Big(\big|\Phi_{n,T} - x\big| \le \gamma_{n,T}\Big) + \mathbb{P}\Big(\Phi_{n,T} \le x - \gamma_{n,T}\Big) \\ &- \mathbb{P}\Big(\widehat{\widehat{\Phi}}_{n,T} \le x - \gamma_{n,T}\Big) + \mathbb{P}\Big(\big|\widehat{\widehat{\Phi}}_{n,T} - \widehat{\Phi}_{n,T}\big| > \gamma_{n,T}\Big).\end{aligned}$$

Note that since $\gamma_{n,T} \to 0$, we can use the anticoncentration results (6.32) for the Gaussian statistic $\Phi_{n,T}$ to get $\sup_{x\in\mathbb{R}} \mathbb{P}\Big(\big|\Phi_{n,T} - x\big| \le \gamma_{n,T}\Big) = o(1)$. Moreover,

$$\mathbb{P}\Big(\big|\widehat{\widehat{\Phi}}_{n,T} - \widehat{\Phi}_{n,T}\big| > \gamma_{n,T}\Big) = o(1)$$

by Proposition A.2 and this probability does not depend on $x$.
Thus,

$$\begin{aligned}&\sup_{x\in\mathbb{R}} \left|\mathbb{P}(\widehat{\Phi}_{n,T} \le x) - \mathbb{P}(\Phi_{n,T} \le x)\right| \le \\ &\quad \le \max\Bigg\{ \sup_{x\in\mathbb{R}} \left|\mathbb{P}\Big(\Phi_{n,T} \le x - \gamma_{n,T}\Big) - \mathbb{P}\Big(\widehat{\widehat{\Phi}}_{n,T} \le x - \gamma_{n,T}\Big)\right|, \\ &\qquad \sup_{x\in\mathbb{R}} \left|\mathbb{P}\Big(\Phi_{n,T} \le x + \gamma_{n,T}\Big) - \mathbb{P}\Big(\widehat{\widehat{\Phi}}_{n,T} \le x + \gamma_{n,T}\Big)\right| \Bigg\} + \\ &\qquad + \sup_{x\in\mathbb{R}} \mathbb{P}\Big(\big|\Phi_{n,T} - x\big| \le \gamma_{n,T}\Big) + \sup_{x\in\mathbb{R}} \mathbb{P}\Big(\big|\widehat{\widehat{\Phi}}_{n,T} - \widehat{\Phi}_{n,T}\big| > \gamma_{n,T}\Big) = \\ &\quad = \sup_{y\in\mathbb{R}} \left|\mathbb{P}\Big(\Phi_{n,T} \le y\Big) - \mathbb{P}\Big(\widehat{\widehat{\Phi}}_{n,T} \le y\Big)\right| + o(1) + o(1) = o(1).\end{aligned}$$

□

## Auxiliary results

**Definition A.1.** *For a given $q > 0$ and $\alpha > 0$, we define dependence adjusted norm as $||X_\cdot||_{q,\alpha}^q = \sup_{m\ge 0}(m+1)^\alpha \sum_{t=m}^\infty \delta_q(X,t)$.*

**Theorem A.1.** *Wu et al. (2016) Assume that* $||X_\cdot||_{q,\alpha}^q < \infty$, *where* $q > 2$ *and* $\alpha > 0$, *and* $\sum_{t=1}^T a_t^2 = T$. *Moreover, assume that* $\alpha > 1/2 - 1/q$. *Denote* $S_T = a_1 X_1 + \ldots + a_T X_T$. *Then for all* $x > 0$,

$$\mathbb{P}(|S_T| \geq x) \leq C_1 \frac{|a|_q^q ||X_\cdot||_{q,\alpha}^q}{x^q} + C_2 \exp\left(-\frac{C_3 x^2}{T||X_\cdot||_{2,\alpha}^2}\right),$$

*where* $C_1, C_2, C_3$ *are constants that only depend on* $q$ *and* $\alpha$.

**Theorem A.2.** *Wu (2007) Let* $(\xi_i)_{i\in\mathbb{Z}}$ *be a stationary and ergodic Markov chain and* $g(\cdot)$ *be a measurable function. Let* $g(\xi_1) \in \mathcal{L}^q, q > 2, \mathbb{E}[g(\xi_0)] = 0$ *and* $l$ *be a positive, nondecreasing slowly varying function. Assume that*

$$\sum_{i=n}^{\infty} \big|\big|\mathbb{E}[g(\xi_i)|\xi_0] - \mathbb{E}[g(\xi_i)|\xi_{-1}]\big|\big|_q = O\big([\log n]^{-\beta}\big),$$

*where* $0 \leq \beta < 1/q$ *and*

$$\sum_{k=1}^{\infty} \frac{k^{-\beta q}}{[l(2^k)]^q} < \infty.$$

*Then* $S_n = g(\xi_1) + \ldots + g(\xi_n) = o_{a.s.}[\sqrt{n} l(n)]$.

**Proposition A.8.** *Wu (2007) Let* $(\epsilon_n)_{n\in\mathbb{Z}}$ *be i.i.d. random variables,* $\xi_n = (\ldots, \epsilon_{n-1}, \epsilon_n)$ *and* $g(\cdot)$ *be a measurable function such that* $g(\xi_n)$ *is a proper random variable for each* $n \geq 0$. *For* $k \geq 0$ *let* $\tilde{\xi}_k = (\ldots, \epsilon_{-1}, \epsilon_0', \epsilon_1, \ldots, \epsilon_{k-1}, \epsilon_k)$, *where* $\epsilon_0'$ *is an i.i.d. copy of* $\epsilon_0$. *Let* $g(\xi_0) \in \mathcal{L}^q, q > 1$ *and* $\mathbb{E}[g(\xi_0)] = 0$. *For* $n \geq 1$ *we have*

$$\big|\big|\mathbb{E}[g(\xi_n)|\xi_0] - \mathbb{E}[g(\xi_n)|\xi_{-1}]\big|\big|_q \leq 2\big|\big|g(\xi_n) - g(\tilde{\xi}_n)\big|\big|_q.$$

**Proposition A.9.** *Under the conditions of Theorem 4.1, it holds that*

$$\bar{\mathbf{X}}_i = \frac{1}{T}\sum_{t=1}^{T} \mathbf{H}_i(\mathcal{U}_{it}) = o_P(1).$$

**Proof of Proposition A.9**. To prove this fact, we will use two results from Wu (2007) stated above. First, fix $j \in \{1, \ldots, d\}$. Denote $\xi_t = \mathcal{U}_{it}, \tilde{\xi}_t = \mathcal{U}_{it}'$ and $g(\cdot) = H_{i,j}(\cdot)$. Then by Assumption (C6), $g(\xi_0) = H_{i,j}(\mathcal{U}_{i0}) \in \mathcal{L}^{q'}$ for $q' > 4$ and $\mathbb{E}[g(\xi_0)] = \mathbb{E}[H_{i,j}(\mathcal{U}_{i0})] = 0$ and we can apply Proposition A.8 (Proposition 3(ii) in Wu (2007)) that says that for all $s \geq 1$ we have:

$$\big|\big|\mathbb{E}[g(\xi_s)|\xi_0] - \mathbb{E}[g(\xi_s)|\xi_{-1}]\big|\big|_{q'} \leq 2\big|\big|g(\xi_s) - g(\tilde{\xi}_s)\big|\big|_{q'},$$

or, equivalently,

$$\big|\big|\mathbb{E}[H_{i,j}(\mathcal{U}_{is})|\mathcal{U}_{i0}] - \mathbb{E}[H_{i,j}(\mathcal{U}_{is})|\mathcal{U}_{i(-1)}]\big|\big|_{q'} \leq 2\big|\big|H_{i,j}(\mathcal{U}_{is}) - H_{i,j}(\mathcal{U}_{is}')\big|\big|_{q'}.$$

Since this holds simultaneously for all $j \in \{1, \ldots, d\}$, we can use the obvious bound $\big|\big|H_{i,j}(\mathcal{U}_{is}) - H_{i,j}(\mathcal{U}'_{is})\big|\big|_{q'} \leq \big|\big|\mathbf{H}_i(\mathcal{U}_{is}) - \mathbf{H}_i(\mathcal{U}'_{is})\big|\big|_{q'} = \delta_{q'}(\mathbf{H}_i, s)$ and Assumption (C8) to write

$$0 \leq \sum_{s=t}^{\infty} \big|\big|\mathbb{E}[g(\xi_s)|\xi_0] - \mathbb{E}[g(\xi_s)|\xi_{-1}]\big|\big|_{q'} \leq \sum_{s=t}^{\infty} \delta_{q'}(\mathbf{H}_i, s) = O(t^{-\alpha}),$$

where $\alpha > 1/2 - 1/q'$.

Now we want to apply Theorem A.2 (Corollary 2(i) in Wu (2007)). As a parameter $\beta$ in the theorem we can take any value satisfying assumption $0 \leq \beta < 1/q'$ because for every $\beta \geq 0$ we have

$$\sum_{s=t}^{\infty} \big|\big|\mathbb{E}[g(\xi_s)|\xi_0] - \mathbb{E}[g(\xi_s)|\xi_{-1}]\big|\big|_{q'} \leq \sum_{s=t}^{\infty} \delta_{q'}(\mathbf{H}_i, s) = O(t^{-\alpha}) = O\big([\log t]^{-\beta}\big).$$

Furthermore, as a positive, nondecreasing slowly varying function $l$ we can take $l(x) = \log^{2/q'-\beta}(x)$. Then,

$$\begin{aligned}
\sum_{k=1}^{\infty} \frac{k^{-\beta q'}}{[l(2^k)]^{q'}} &= \sum_{k=1}^{\infty} \frac{k^{-\beta q'}}{\big[\log^{2/q'-\beta}(2^k)\big]^{q'}} \\
&= \sum_{k=1}^{\infty} \frac{k^{-\beta q'}}{k^{2-\beta q'}(\log 2)^{2-\beta q'}} \\
&= \frac{1}{(\log 2)^{2-\beta q'}} \sum_{k=1}^{\infty} \frac{1}{k^2} \\
&< \infty.
\end{aligned}$$

Hence, $S_T = g(\xi_1) + \ldots + g(\xi_T) = o_{a.s.}[\sqrt{T}\log^{2/q'-\beta}(T)]$, or, equivalently, $\bar{X}_{i,j} = S_T/T = o_{a.s.}[\log^{2/q'-\beta}(T)/\sqrt{T}] = o_P(1)$ for each $j \in \{1, \ldots, d\}$. Obviously, this means that $\bar{\mathbf{X}}_i = o_P(1)$. □

## 6.3 Proof of Theorem 4.2

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 A.10.** *Under Assumptions (C4) and (C6),* $||\Delta\mathbf{H}_i(\mathcal{U}_{it})||_4 < \infty$.

**Proof of Proposition A.10**. By Assumption (C6),

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

□

**Proposition A.11.** *Under Assumption (C9), $\Delta\mathbf{X}_{it}$ (elementwise) and $\Delta\varepsilon_{it}$ are uncorrelated for each $t \in \{1, \ldots, T\}$.*

**Proof of Proposition A.11**. By Assumption (C9),

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

□

**Proposition A.12.** *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), (C6), (C7) and (C10), we have that $\sum_{s=0}^{\infty}\delta_2(\Delta\mathbf{U}_i, s) < \infty$.*

**Proof of Proposition A.12**. It is easy to check that

$$\begin{aligned}
\delta_2(\Delta\mathbf{U}_i, s) \leq\ & \delta_2(\mathbf{U}_i, s) + \delta_2(\mathbf{U}_i, s-1) \\
& + \big(\delta_2(\mathbf{H}_i, s-1) + \delta_2(\mathbf{H}_i, s)\big)||G_i||_2 + \big(\delta_2(G_i, s-1) + \delta_2(G_i, s)\big)||\mathbf{H}_i||_2.
\end{aligned}$$

Hence,

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

By Assumptions (C2), (C3), (C6), (C7) and (C10), the RHS is finite. This proves the theorem. □

**Proposition A.13.** *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 A.13**. We need the following notation:

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

Then,

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

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

$$
\begin{aligned}
||\kappa_{i,s}^{\mathcal{P}}||^2 &\leq \frac{1}{T^2}\sum_{t=1}^{T}\Big|\Big|\mathbb{E}(\Delta\mathbf{X}_{it}\Delta\varepsilon_{it}|\mathcal{I}_{i,t-s}) - \mathbb{E}(\Delta\mathbf{X}'_{it,s}\Delta\varepsilon'_{it,s}|\mathcal{I}_{i,t-s})\Big|\Big|^2 \\
&\leq \frac{1}{T^2}\sum_{t=1}^{T}\Big|\Big|\Delta\mathbf{X}_{it}\Delta\varepsilon_{it} - \Delta\mathbf{X}'_{it,s}\Delta\varepsilon'_{it,s}\Big|\Big|^2 \\
&= \frac{1}{T^2}\sum_{t=1}^{T}\Big|\Big|\Delta\mathbf{H}_i(\mathcal{U}_{it})\Delta G_i(\mathcal{J}_{it}) - \Delta\mathbf{H}_i(\mathcal{U}'_{it,s})\Delta G_i(\mathcal{J}'_{it,s})\Big|\Big|^2 \\
&= \frac{1}{T^2}\sum_{t=1}^{T}\Big|\Big|\Delta\mathbf{U}_i(\mathcal{I}_{i,t}) - \Delta\mathbf{U}_i(\mathcal{I}'_{i,t,s})\Big|\Big|^2 \\
&\leq \frac{1}{T^2}\sum_{t=1}^{T}\delta_2^2(\Delta\mathbf{U}_i, s) \\
&= \frac{1}{T}\delta_2^2(\Delta\mathbf{U}_i, s)
\end{aligned}
$$

with $\mathcal{U}'_{it,s} = (\ldots, u_{i(t-s-1)}, u'_{i(t-s)}, u_{i(t-s+1)}, \ldots, u_{it})$, $u'_{i(t-s)}$ being an i.i.d. copy of $u_{i(t-s)}$, $\mathcal{J}'_{it,s} = (\ldots, \eta_{i(t-s-1)}, \eta'_{i(t-s)}, \eta_{i(t-s+1)}, \ldots, \eta_{it})$, $\eta'_{i(t-s)}$ being an i.i.d. copy of $\eta_{i(t-s)}$, and $\zeta'_{it} = (u'_{it}, \eta'_{it})^\top$ and $\mathcal{I}'_{i,t,s} = (\ldots, \zeta_{i(t-s-1)}, \zeta'_{i(t-s)}, \zeta_{i(t-s+1)}, \ldots, \zeta_{it})$. 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 A.12,

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

Since $\mathbb{E}\kappa_i = 0$ by Proposition A.11, we conclude that

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

Therefore, the proposition follows. □

**Proof of Theorem 4.2**. Define $\Delta m_{it} = m_i\left(\frac{t}{T}\right) - m_i\left(\frac{t-1}{T}\right)$.
Recall the differencing estimator $\widehat{\boldsymbol{\beta}}_i$:

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

This leads to

$$\begin{aligned}
\sqrt{T}(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i) = &\Big(\frac{1}{T}\sum_{t=1}^{T} \Delta \mathbf{X}_{it} \Delta \mathbf{X}_{it}^{\top}\Big)^{-1} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \Delta \mathbf{X}_{it} \Delta m_{it} \\
&+ \Big(\frac{1}{T}\sum_{t=1}^{T} \Delta \mathbf{X}_{it} \Delta \mathbf{X}_{it}^{\top}\Big)^{-1} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \Delta \mathbf{X}_{it} \Delta \varepsilon_{it}.
\end{aligned} \tag{6.36}$$

To begin with, we take a closer look at the first summand in (6.36). Dealing with scalars is much more understandable for many readers, therefore, we prove everything for each of the elements of the vector separately.
Fix $j \in 1, \ldots, d$. By Chebyshev's inequality we have

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T} |\Delta H_{ij}(\mathcal{U}_{it})| > a\right) \leq \frac{\mathbb{E}\Big[\big(\sum_{t=1}^{T} |\Delta H_{ij}(\mathcal{U}_{it})|\big)^2\Big]}{T^2 a^2} \tag{6.37}$$

and

$$\mathbb{E}\Big[\big(\sum_{t=1}^{T} |\Delta H_{ij}(\mathcal{U}_{it})|\big)^2\Big] = \sum_{t=1}^{T} \mathbb{E}\left[\Delta H_{ij}^2(\mathcal{U}_{it})\right] + \sum_{\substack{t=1,s=1,\\ t\neq s}}^{T} \mathbb{E}\big[\,|\Delta H_{ij}(\mathcal{U}_{it}) \Delta H_{ij}(\mathcal{U}_{is})|\,\big]. \tag{6.38}$$

Note that by the Cauchy-Schwarz inequality for all $t$ and $s$ we have

$$\mathbb{E}\big[\,|H_{ij}(\mathcal{U}_{it})H_{ij}(\mathcal{U}_{is})|\,\big] \le \sqrt{\mathbb{E}\big[H_{ij}^2(\mathcal{U}_{it})\big]}\sqrt{\mathbb{E}\big[H_{ij}^2(\mathcal{U}_{is})\big]} = \mathbb{E}\left[H_{ij}^2(\mathcal{U}_{i0})\right]$$

and

$$\big|\mathbb{E}\big[H_{ij}(\mathcal{U}_{it})H_{ij}(\mathcal{U}_{is})\big]\big| \le \mathbb{E}\big[\,|H_{ij}(\mathcal{U}_{it})H_{ij}(\mathcal{U}_{is})|\,\big] \le \mathbb{E}\left[H_{ij}^2(\mathcal{U}_{i0})\right]. \tag{6.39}$$

Hence,

$$\begin{aligned}\mathbb{E}\left[\Delta H_{ij}^2(\mathcal{U}_{it})\right] &= \mathbb{E}\left[H_{ij}^2(\mathcal{U}_{it})\right] - 2\mathbb{E}\left[H_{ij}(\mathcal{U}_{it})H_{ij}(\mathcal{U}_{it-1})\right] + \mathbb{E}\left[H_{ij}^2(\mathcal{U}_{it-1})\right]\\ &\le \mathbb{E}\left[H_{ij}^2(\mathcal{U}_{i0})\right] + 2\mathbb{E}\left[H_{ij}^2(\mathcal{U}_{i0})\right] + \mathbb{E}\left[H_{ij}^2(\mathcal{U}_{i0})\right]\\ &= 4\mathbb{E}\left[H_{ij}^2(\mathcal{U}_{i0})\right]\end{aligned}$$

and the first summand in (6.38) can be bounded by $4T\mathbb{E}\left[H_{ij}^2(\mathcal{U}_{i0})\right]$.
Now to the second summand in (6.38):

$$\begin{aligned}\mathbb{E}\big[\,|\Delta H_{ij}(\mathcal{U}_{it})\Delta H_{ij}(\mathcal{U}_{is})|\,\big] &\le \mathbb{E}\big[\,|H_{ij}(\mathcal{U}_{it})H_{ij}(\mathcal{U}_{is})|\,\big] + \mathbb{E}\big[\,|H_{ij}(\mathcal{U}_{it-1})H_{ij}(\mathcal{U}_{is})|\,\big]\\ &\quad + \mathbb{E}\big[\,|H_{ij}(\mathcal{U}_{it})H_{ij}(\mathcal{U}_{is-1})|\,\big] + \mathbb{E}\big[\,|H_{ij}(\mathcal{U}_{it-1})H_{ij}(\mathcal{U}_{is-1})|\,\big]\\ &\le 4\mathbb{E}\left[H_{ij}^2(\mathcal{U}_{i0})\right],\end{aligned}$$

where in the last inequality we used (6.39). This means that the second summand in (6.38) can be bounded by $4T(T-1)\mathbb{E}\left[H_{ij}^2(\mathcal{U}_{i0})\right]$.
Plugging these bounds in (6.38), we get

$$\mathbb{E}\left[\left(\sum_{t=1}^{T}|\Delta H_{ij}(\mathcal{U}_{it})|\right)^2\right] \le 4T\mathbb{E}\left[H_{ij}^2(\mathcal{U}_{i0})\right] + 4T(T-1)\mathbb{E}\left[H_{ij}^2(\mathcal{U}_{i0})\right] = 4T^2\mathbb{E}\left[H_{ij}^2(\mathcal{U}_{i0})\right],$$

which together with (6.37) leads to $\frac{1}{T}\sum_{t=1}^{T}\big|\Delta H_{ij}(\mathcal{U}_{it})\big| = O_P(1)$. By the assumption in Theorem 4.2, $m_i(\cdot)$ is Lipschitz continuous, that is, $|\Delta m_{it}| = \big|m_i\left(\frac{t}{T}\right) - m_i\left(\frac{t-1}{T}\right)\big| \le C\frac{1}{T}$ for all $t \in \{1,\dots,T\}$ and some constant $C > 0$. Hence,

$$\begin{aligned}\Big|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\Delta H_{ij}(\mathcal{U}_{it})\Delta m_{it}\Big| &\le \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\big|\Delta H_{ij}(\mathcal{U}_{it})\big|\cdot\big|\Delta m_{it}\big|\\ &\le \frac{C}{\sqrt{T}}\cdot\frac{1}{T}\sum_{t=1}^{T}|\Delta H_{ij}(\mathcal{U}_{it})|\\ &= O_P\Big(\frac{1}{\sqrt{T}}\Big).\end{aligned} \tag{6.40}$$

Since it holds for each $j \in \{1,\dots,d\}$, it is obvious that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\Delta\mathbf{X}_{it}\Delta m_{it} = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\Delta\mathbf{H}_i(\mathcal{U}_{it})\Delta m_{it} = O_P\Big(\frac{1}{\sqrt{T}}\Big). \tag{6.41}$$

Similarly, by Proposition A.10 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

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

where $|A|$ with $A$ being a matrix is any matrix norm.
By Assumption (C5), 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,

$$\left|\Big(\frac{1}{T}\sum_{t=1}^{T} \Delta \mathbf{X}_{it}\Delta \mathbf{X}_{it}^{\top}\Big)^{-1}\right| = O_P(1). \tag{6.42}$$

Plugging (6.41) into (6.40) and combining it with (6.42), we get that the first summand in (6.36) is $O_P(1/\sqrt{T})$.
To estimate the other term, we can apply the Proposition A.13 together with (6.42) to get that the second summand in (6.36) is $O_P(1)$.
The statement of the theorem follows. □