CESifo WORKING PAPERS

10725
2023
October 2023

Trimmed Mean Group Estimation of Average Treatment Effects in Ultra Short T Panels under Correlated Heterogeneity

M. Hashem Pesaran, Liying Yang

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 10725*

Trimmed Mean Group Estimation of Average Treatment Effects in Ultra Short *T* Panels under Correlated Heterogeneity

Abstract

Under correlated heterogeneity, the commonly used two-way fixed effects estimator is biased and can lead to misleading inference. This paper proposes a new trimmed mean group (TMG) estimator which is consistent at the irregular rate of n $^{1/3}$ even if the time dimension of the panel is as small as the number of its regressors. Extensions to panels with time effects are provided, and a Hausman-type test of correlated heterogeneity is proposed. Small sample properties of the TMG estimator (with and without time effects) are investigated by Monte Carlo experiments and shown to be satisfactory and perform better than other trimmed estimators proposed in the literature. The proposed test of correlated heterogeneity is also shown to have the correct size and satisfactory power. The utility of the TMG approach is illustrated with an empirical application.

JEL-Codes: C210, C230.

Keywords: correlated heterogeneity, irregular estimators, two-way fixed effects, FE-TE, tests of correlated heterogeneity, calorie demand.

M. Hashem Pesaran
Department of Economics
University of Southern California
Los Angeles / CA / USA
pesaran@usc.edu

Liying Yang
Sauder School of Business
University of British Columbia
Vancouver / BC / Canada
liying.yang@sauder.ubc.ca

October 17, 2023
We are grateful to Cheng Hsiao, Ron Smith and Hayun Song for helpful comments and discussions. The research on this paper was completed when Liying Yang was a Ph.D. student at the Department of Economics, University of Southern California.

1 Introduction

Fixed effects estimation of average treatment effects has been predominantly utilized for program and policy evaluation. For static panel data models where slope heterogeneity is uncorrelated with treatment effects, standard fixed and time effects (FE-TE) estimators are consistent and if used in conjunction with robust standard errors lead to valid inference in short T (time dimension) panels when the number of cross sections (n) is sufficiently large. However, when the slope heterogeneity is correlated with the treatment and/or control variables the FE-TE estimators (also known as two-way fixed effects) become inconsistent even if both T and $n \to \infty$.[1] Such correlated heterogeneity arises endogenously in the case of dynamic panel data models considered by Pesaran and Smith (1995) even if the slope heterogeneity itself is purely random.

In the case of static panels, correlated heterogeneity could arise when treatment effects are correlated with the treatment itself and/or the control variables. For example, in estimation of returns to education, the choice of educational level is likely to be correlated with expected returns to education. In a review of active policies in labor markets, Crépon and Van Den Berg (2016) emphasize that when estimating the average impacts on workers' productivity and earnings, correlated heterogeneity should be accounted for, to better encourage enrollment in training programs. Banerjee et al. (2015) also consider identification and estimation of heterogeneous treatment effects in the case of micro-credit evaluation programs, and Bastagli et al. (2019) consider similar issues in studies of anti-poverty cash transfer programs.[2] In a recent survey de Chaisemartin and D'Haultfoeuille (2023) highlight the importance of allowing for correlated heterogeneity and draw attention to the misleading inferences that can result when FE-TE estimates are used in the case of heterogeneous policy effects.

Pesaran and Smith (1995) proposed mean group (MG) estimation for dynamic heterogeneous panel data models, where by construction represent examples of correlated heterogeneity. It was later shown that for panels with strictly exogenous regressors, the MG estimator is $\sqrt{n}$-consistent in the presence of correlated heterogeneity even if T is fixed as $n \to \infty$, so long as T is sufficiently large such that at least second order moment of the MG estimator exists. However, when T is ultra short such that T is close to the number of regressors, k, the MG estimator could fail. As shown by Chamberlain (1992) one needs T to be strictly larger than

[1]The concept of the correlated random coefficient model is due to Heckman and Vytlacil (1998). Wooldridge (2005) shows that FE-TE estimators continue to be consistent if slope heterogeneity is mean-independent of all the de-trended covariates. See also condition (2.19) given below.

[2]Reviews of recent advances in econometric methods for heterogeneous treatment effects of binary variables can be found in Athey and Imbens (2017) and Abadie and Cattaneo (2018).

k for regular identification of average effects under correlated heterogeneity.[3] Chamberlain (1992) calculated efficiency bounds for models defined by conditional moment restrictions with a nonparametric component, and proposed a $\sqrt{n}$-consistent Generalized Method of Moments (GMM) estimator for the mean of correlated random coefficients in panel data models provided that certain rank and moment conditions hold.[4] See also Bonhomme (2012) and Arellano and Bonhomme (2012). Assuming the errors follow autoregressive moving average processes, Arellano and Bonhomme (2012) provide rank conditions under which the GMM estimators they propose for variances and densities of correlated random coefficients can be regularly identified.

The above papers adopt the GMM approach to address identification and estimation of average treatment effects. Some researchers consider other regular estimators by imposing additional restrictions on the correlation between heterogeneous coefficients and regressors. Wooldridge (2005) proposes an alternative estimator for models with nonlinear individual-specific unobserved effects, where he imposes a condition that random coefficients are mean independent of the idiosyncratic deviations in regressors. To estimate the average effects of binary treatment variables for the sub-population with no time variations in treatment status, Verdier (2020) explicitly models selection into treatment. Assuming random coefficients are independent of regressors, Lee and Sul (2022) apply a double-sided trimming scheme to the MG estimators for static panels with common correlated effects developed by Chudik and Pesaran (2015), so as to eliminate effects of outlying individual estimates with too small or too large regressor sample variances.

This paper considers identification and estimation of average treatment effects in ultra short linear panel data models with continuous covariates, where T could be as small as k. Building on the pioneering work of Chamberlain (1992), Graham and Powell (2012) focus on panels with $T = k$, where identification issues of time effects and the mean coefficients arise especially when there are insufficient within-individual variations for some regressors. They derive an irregular estimator of the mean coefficients by excluding individual estimates from the estimation of the average treatment effects if the sample variance of regressor in question is smaller than a given threshold.[5] Exploiting the sub-population of "stayers"with no time variations in regressors, they then propose an estimator of time effects.[6] More recently Sasaki

[3]An unknown parameter, β_0, is said to be regularly identified if there exists an estimator that converges to β_0 in probability at the rate of $\sqrt{n}$. Any estimator that converges to its true value at a rate slower than $\sqrt{n}$ is said to be irregularly identified.

[4]The GMM estimator proposed by Chamberlain (1992) turns out to be the same as the MG estimator. See equation (4.8b) in Chamberlain (1992).

[5]The trimming idea of Graham and Powell has also been used recently by de Chaisemartin et al. (2023) for identification of the average slopes of switchers' potential outcomes with many "near stayers".

[6]Graham and Powell (2012) establish identification results based on moment equations conditional on the sub-population of "stayers", namely individuals with no time variations in their realized covariates. But in

and Ura (2021) propose an alternative procedure to deal with the possibility of many stayers and/or slow movers in the panel. They consider various distributions of within-variations and use local polynomial regressions to provide robust inference.

In this paper, we first derive asymptotic properties of MG and FE estimators in large n and short T heterogeneous static panel data models, and provide sufficient conditions under which MG and FE estimators are regular, in the sense that they are $\sqrt{n}$-consistent. In cases where these conditions are not met, we propose a new trimmed mean group (TMG) estimator which makes use of additional information on trimmed units not exploited by Graham and Powell (2012). In effect, information on all units (whether subject to trimming or not) are included in the computation of the average treatment effect. Following the literature, the decision on whether a unit i is subject to trimming is made with respect to the determinant of the sample covariance matrix of the regressors, denoted by d_i, and the individual estimates for unit i are trimmed uniformly if $d_i < a_n = \bar{d}_n n^{-\alpha}$, where $\bar{d}_n = n^{-1}\Sigma_{i=1}^{n} d_i$, and α measures the rate of trade-off between bias and variance of TMG. Our asymptotic derivations suggest setting α close to 1/3. Also noting that $d_i/\bar{d}_n$ is scale free, our choice of trimming threshold, a_n, does not involve any other tuning parameters.

We also consider heterogeneous panels with time effects and develop two new estimators of the average treatment effects in two-way fixed effects regressions, which we denote by TMG-TE and TMG-C, corresponding to cases where $T \geq k$ and $T > k$, respectively. The TMG-TE estimator is based on joint estimation of time and average effects, whilst the TMG-C estimator follows Chamberlain (1992) and eliminates the time effects before estimating the average treatments, which is possible only if $T > k$. We derive the asymptotic distributions of TMG-TE and TMG-C estimators under fairly general assumptions but require the identifying condition that the non-zero dependence between heterogeneous slope coefficients and the regressors is time-invariant. Note that this condition trivially holds in the case of FE-TE estimators whose validity requires zero dependence between the slope coefficients and the regressors.

As noted above the presence of heterogeneity by itself does not invalidate the use of the FE-TE estimator which continues to have the regular convergence rate of $\sqrt{n}$. The problem arises when slope heterogeneity is correlated with the covariates, such as the treatment variable. It is, therefore, important that before using the FE-TE estimator the assumption of uncorrelated heterogeneity is tested. To this end, we also propose Hausman-type tests of correlated heterogeneity by comparing the FE and FE-TE estimators with the associated TMG estimators, and derive their asymptotic distributions under fairly general conditions. The earlier Hausman tests of slope homogeneity developed by Pesaran et al. (1996) and Pe-

estimation, a sub-sample of “near stayers”is used instead.

saran and Yamagata (2008) are based on the difference between FE and MG estimators and do not apply when T is ultra short.

We also carry out an extensive set of Monte Carlo (MC) simulations to investigate the small sample properties of the TMG, TMG-TE and TMG-C estimators and how they compare with other estimators, including the trimmed estimators proposed by Graham and Powell (GP) and Sasaki and Ura (SU). The MC evidence on the size and empirical power of the Hausman-type tests of correlated heterogeneity in panel data models without and with time effects is provided, and the sensitivity of estimation results to the choice of the trimming threshold parameter, α, is also investigated. The MC and theoretical results of the paper are all in agreement. The TMG and TMG-TE estimators not only have the correct size but also achieve better finite sample properties compared with other trimmed estimators across a number of experiments with different data generating processes, allowing for heteroskedasticity (random and correlated), error serial correlations, and regressors with heterogeneous dynamics and interactive effects. The simulation results also confirm that the Hausman-type tests based on the difference between FE (FE-TE) and TMG (TMG-TE) estimators have the correct size and power against the alternative of correlated heterogeneity.

As an empirical illustration, we re-visit the example considered by GP who provide estimates of the average effect of household expenditures on calorie demand using a balanced panel of $n = 1,358$ households in poor rural communities in Nicaragua over the years 2001–2002 ($T = 2$) and 2000–2002 ($T = 3$). Comparing the FE and TMG estimates, for panels with and without time effects, we find that the Hausman tests reject the null of uncorrelated heterogeneity, thus shedding doubt on the use of FE or FE-TE estimates for this application. For the ultra short panel with $T = 2$, the FE and TMG estimates of the average treatment effects are 0.6568 (0.0287) and 0.5623 (0.0425). The figures in brackets are standard errors. Given the result of the Hausman test, most likely the FE estimate is biased upward, with a much lower standard error. These results do not change if we allow for time effects. Turning to the other trimmed estimators, the GP and SU estimates, 0.4549 (0.1003) and 0.6974 (0.1689), respectively, are wide apart, and both have larger standard errors as compared to the TMG estimate. Again these estimates are not much affected by allowing for time effects. But once we consider the panel with $T = 3$ the TMG and GP estimates (with or without time effects) become very similar, although the TMG estimates continue to be more precisely estimated. The gap between FE and TMG estimates also becomes closer but remains statistically highly significant.

The rest of the paper is organized as follows. Section 2 sets out the heterogeneous panel data model and investigates the asymptotic properties of the MG and FE estimators. Section 3 considers ultra short T panels and discusses the need for trimming as suggested by GP.

The proposed TMG estimator is introduced in Section 4, and its asymptotic properties are established in Section 5. Section 6 extends the TMG estimation to ultra short panel data models with time effects, distinguishing between cases where $T \geq k$ and $T > k$. Section 7 sets out the Hausman-type test of correlated heterogeneous slope coefficients. Section 8 describes the Monte Carlo experiments and reports the simulation results. Section 9 presents the empirical illustration. Section 10 concludes. The online supplement develops the test of correlated heterogeneity for panels with time effects, and provides supplementary information on Monte Carlo designs and additional Monte Carlo evidence.

2 Heterogeneous linear panel data models

Consider the panel data model where the outcome variable y_{it} for unit i at time t is explained linearly in terms of the $k \times 1$ vector of covariates $\boldsymbol{w}_{it}$

$$y_{it} = \boldsymbol{\theta}_i' \boldsymbol{w}_{it} + u_{it}, \text{ for } i = 1, 2, ..., n, \text{ and } t = 1, 2, ..., T, \tag{2.1}$$

where $\boldsymbol{\theta}_i$ is a $k \times 1$ vector of unknown unit-specific coefficients and u_{it} is the error terms. Stacking by time we have

$$\boldsymbol{y}_i = \boldsymbol{W}_i \boldsymbol{\theta}_i + \boldsymbol{u}_i, \text{ for } i = 1, 2, ..., n, \tag{2.2}$$

where $\boldsymbol{y}_i = (y_{i1}, y_{i2}, ..., y_{iT})'$, $\boldsymbol{W}_i = (\boldsymbol{w}_{i1}, \boldsymbol{w}_{i2}, ..., \boldsymbol{w}_{iT})'$, and $\boldsymbol{u}_i = (u_{i1}, u_{i2}, ..., u_{iT})'$. The parameter of interest is the $k \times 1$ vector of average treatment effects, $\boldsymbol{\theta}_0$, defined by

$$\boldsymbol{\theta}_0 = \operatorname*{plim}_{n \to \infty} \left(n^{-1} \sum_{i=1}^{n} \boldsymbol{\theta}_i \right). \tag{2.3}$$

When $T \geq k$, $\boldsymbol{\theta}_0$ can be estimated by the mean group estimator, $\hat{\boldsymbol{\theta}}_{MG}$, computed as a simple average of the least squares estimates of $\boldsymbol{\theta}_i$, namely (see Pesaran and Smith (1995))

$$\hat{\boldsymbol{\theta}}_{MG} = \frac{1}{n} \sum_{i=1}^{n} \hat{\boldsymbol{\theta}}_i, \tag{2.4}$$

where

$$\hat{\boldsymbol{\theta}}_i = \left(\boldsymbol{W}_i' \boldsymbol{W}_i \right)^{-1} \left(\boldsymbol{W}_i' \boldsymbol{y}_i \right). \tag{2.5}$$

To investigate the asymptotic properties of the MG estimator when T is short and $n \to \infty$, we make the following assumptions:

Assumption 1 (Errors) *Conditional on $\boldsymbol{W}_i$, (a) the errors, u_{it}, in (2.1) are cross-sectionally independent, (b) $E(\boldsymbol{u}_i |\boldsymbol{W}_i) = \boldsymbol{0}$, for $i = 1, 2, ..., n$, and (c) $E(\boldsymbol{u}_i\boldsymbol{u}_i' |\boldsymbol{W}_i) = \boldsymbol{H}_i(\boldsymbol{W}_i) = \boldsymbol{H}_i$, where $\boldsymbol{H}_i$ is a $T \times T$ bounded matrix with $0 < c < \inf_i \lambda_{min}(\boldsymbol{H}_i) < \sup_i \lambda_{max}(\boldsymbol{H}_i) < C$.*

Assumption 2 (Regression coefficients) *The $k \times 1$ vector of coefficients, $\boldsymbol{\theta}_i$, is allowed to depend on the distribution of $\boldsymbol{W}_i$ with $rank(\boldsymbol{W}_i) = k$. This dependence could be (a) deterministic with $\boldsymbol{\theta}_i$ fixed and bounded or (b) stochastic, with $\boldsymbol{\theta}_i$ jointly determined with $\boldsymbol{W}_i$.*

(a) $\boldsymbol{\theta}_i$ are deterministic with $\sup_i \|\boldsymbol{\theta}_i\| < C$ for $i = 1, 2, ..., n$, such that $\bar{\boldsymbol{\theta}}_n = n^{-1}\sum_{i=1}^n \boldsymbol{\theta}_i \to \boldsymbol{\theta}_0$, with $\|\boldsymbol{\theta}_0\| < C$.

(b) $\boldsymbol{\theta}_i$ are independent draws from a distribution with $E(\boldsymbol{\theta}_i) = \boldsymbol{\theta}_0$ and bounded variances for $i = 1, 2, ..., n$, where $\|\boldsymbol{\theta}_0\| < C$, and $\sup_i E\|\boldsymbol{\theta}_i\|^4 < C$.

Remark 1 *Under Assumption 1, the $k \times 1$ vector of covariates, $\boldsymbol{w}_{it}$, for $i = 1, 2, ..., n$ are strictly exogenous, but it allows the conditional variance of $\boldsymbol{u}_i$ to depend on $\boldsymbol{W}_i$, and for the errors, u_{it}, to be serially (over time t) correlated.*

Remark 2 *Part (c) of Assumption 1 rules out the possibility of unbounded random variations in $\boldsymbol{H}_i$, but can be relaxed if instead it is assumed that $0 < c < \inf_i \lambda^2_{min}(\boldsymbol{H}_i) < \sup_i \lambda^2_{max}(\boldsymbol{H}_i) < C$, with higher order moment conditions on $d_i = \det(\mathbf{W}_i'\mathbf{W}_i)$ and $\|(\boldsymbol{W}_i'\boldsymbol{W}_i)^*\|$.*

Remark 3 *Assumption 2 is an identification condition for $\boldsymbol{\theta}_0$. Under Assumption 2(b) where $\boldsymbol{\theta}_i$ follows a random coefficients model, $E(\bar{\boldsymbol{\theta}}_n) = \boldsymbol{\theta}_0$, and $n^{-1}\sum_{i=1}^n \boldsymbol{\theta}_i \to_p \boldsymbol{\theta}_0$.*

2.1 Properties of mean group estimator in short T panels

Substituting (2.2) in (2.5) we have

$$\hat{\boldsymbol{\theta}}_i = \boldsymbol{\theta}_i + \boldsymbol{\xi}_{iT}, \tag{2.6}$$

where

$$\boldsymbol{\xi}_{iT} = \boldsymbol{R}_i'\boldsymbol{u}_i, \tag{2.7}$$

and $\boldsymbol{R}_i = \boldsymbol{W}_i(\boldsymbol{W}_i'\boldsymbol{W}_i)^{-1}$. Averaging both sides of (2.6) over i, we have

$$\hat{\boldsymbol{\theta}}_{MG} = \bar{\boldsymbol{\theta}}_n + \bar{\boldsymbol{\xi}}_{nT}, \tag{2.8}$$

where

$$\bar{\boldsymbol{\theta}}_n = n^{-1}\sum_{i=1}^n \boldsymbol{\theta}_i, \text{ and } \bar{\boldsymbol{\xi}}_{nT} = n^{-1}\sum_{i=1}^n \boldsymbol{\xi}_{iT}. \tag{2.9}$$

Then under Assumption 1, $E(\boldsymbol{u}_i | \boldsymbol{W}_i) = \mathbf{0}$, and hence $E(\bar{\boldsymbol{\xi}}_{nT}) = E(n^{-1}\sum_{i=1}^{n}\boldsymbol{\xi}_{iT}) = n^{-1}\sum_{i=1}^{n} E[\boldsymbol{R}_i' E(\boldsymbol{u}_i | \boldsymbol{W}_i)] = \mathbf{0}$. Then using (2.8) $E(\hat{\boldsymbol{\theta}}_{MG}) = E(\bar{\boldsymbol{\theta}}_n) + E(\bar{\boldsymbol{\xi}}_{nT}) = \boldsymbol{\theta}_0$, namely $\hat{\boldsymbol{\theta}}_{MG}$ is an *unbiased* estimator of $\boldsymbol{\theta}_0$ irrespective of the possible dependence of $\boldsymbol{\theta}_i$ on $\boldsymbol{W}_i$. However, the MG estimator is likely to have a large variance when T is too small. This arises, for example, when the variance of $\bar{\boldsymbol{\xi}}_{nT}$ does not exist or is very large. The conditions under which $\hat{\boldsymbol{\theta}}_{MG}$ converges to $\boldsymbol{\theta}_0$ at the regular $n^{1/2}$ rate are given in the following proposition:

Proposition 1 (Sufficient conditions for $\sqrt{n}$-consistency of $\hat{\boldsymbol{\theta}}_{MG}$) *Suppose that y_{it} for $i = 1, 2, ..., n$ and $t = 1, 2, ..., T$ are generated by model (2.2), and Assumptions 1-2 hold. Then as $n \to \infty$, the MG estimator given by (2.4) is $\sqrt{n}$-consistent for fixed T panels if*

$$\sup_i E\left(d_i^{-2}\right) < C, \text{ and } \sup_i E\left[\|(\boldsymbol{W}_i'\boldsymbol{W}_i)^*\|_1^2\right] < C, \tag{2.10}$$

where $d_i = \det(\boldsymbol{W}_i'\boldsymbol{W}_i)$, and $(\boldsymbol{W}_i'\boldsymbol{W}_i)^*$ is the adjoint of $\boldsymbol{W}_i'\boldsymbol{W}_i$.

For a proof see A.2.1 in the Appendix.

Example 1 *In the simple case where $k = 2$, $\boldsymbol{w}_{it} = (1, x_{it})'$, and $\boldsymbol{\theta}_i = (\alpha_i, \beta_i)'$. Suppose $E(\boldsymbol{u}_i\boldsymbol{u}_i'|\boldsymbol{x}_i) = \sigma_i^2\boldsymbol{I}_T$ for $t = 1, 2, ..., T$ and $i = 1, 2, ..., n$ with $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{iT})'$, then the individual OLS estimator of the slope coefficient, $\hat{\beta}_i = (\boldsymbol{x}_i'\boldsymbol{M}_T\boldsymbol{x}_i)^{-1}\boldsymbol{x}_i'\boldsymbol{M}_T\boldsymbol{y}_i$, has first and second order moments if $E(u_{it}^2) < C$ and $E\left(d_{ix}^{-2}\right) < C$, where $d_{ix} = \det(\boldsymbol{x}_i'\boldsymbol{M}_T\boldsymbol{x}_i)$, $\boldsymbol{M}_T = \boldsymbol{I}_T - T^{-1}\boldsymbol{\tau}_T\boldsymbol{\tau}_T'$, and $\boldsymbol{\tau}_T = (1, 1, ..., 1)'$. In the case where x_{it} are Gaussian distributed with mean zeros and a finite variance, σ_x^2, it follows that $\frac{1}{d_{ix}} = \frac{\sigma_x^2}{\chi_{T-1}^2}$, where χ_{T-1}^2 is a Chi-squared variable with $T-1$ degrees of freedom. Hence, $E\left(d_{ix}^{-2}\right)$ exists if $T - 1 > 4$, or if $T > 5$. For panels with $T \leq 5$, the MG estimator would be irregular when first and/or second order moments of some individual estimates do not exist.*

2.2 A comparison of MG and FE estimators

Consider a panel data model with individual fixed effects, α_i, and heterogeneous slope coefficients, $\boldsymbol{\beta}_i$,

$$y_{it} = \alpha_i + \boldsymbol{\beta}_i'\boldsymbol{x}_{it} + u_{it}, \text{ for } i = 1, 2, ..., n, \text{ and } t = 1, 2, ..., T, \tag{2.11}$$

where $\boldsymbol{x}_{it}$ is a $k' \times 1$ vector of regressors ($k' = k - 1$). In matrix notations

$$\boldsymbol{y}_i = \alpha_i\boldsymbol{\tau}_T + \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{u}_i, \tag{2.12}$$

where $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{iT})'$. The FE and MG estimators of $\boldsymbol{\beta}_0$ are given by

$$\hat{\boldsymbol{\beta}}_{FE} = \left(n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{X}_i\right)^{-1}\left(n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{y}_i\right), \tag{2.13}$$

and

$$\hat{\boldsymbol{\beta}}_{MG} = n^{-1}\sum_{i=1}^{n}\hat{\boldsymbol{\beta}}_i, \tag{2.14}$$

where $\hat{\boldsymbol{\beta}}_i = (\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{y}_i$. In this setting the parameter of interest is given by $\boldsymbol{\beta}_0 = \text{plim}_{n\to\infty}\left(n^{-1}\sum_{i=1}^{n}\boldsymbol{\beta}_i\right)$. One of the main advantages of the FE estimator is its robustness to the dependence between α_i and the regressors. $\hat{\boldsymbol{\beta}}_{FE}$ is also well defined even if $T = k$ so long as the following standard assumption is met:

Assumption 3 (Data pooling assumption) *Let $\bar{\boldsymbol{\Psi}}_n = n^{-1}\sum_{i=1}^{n}\boldsymbol{\Psi}_{ix}$, where $\boldsymbol{\Psi}_{ix} = \boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{X}_i$. For $T \geq k$, there exists n_0 such that for all $n > n_0$, $\bar{\boldsymbol{\Psi}}_n$ is positive definite,*

$$\bar{\boldsymbol{\Psi}}_n \to_p \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}E\left(\boldsymbol{\Psi}_{ix}\right) = \bar{\boldsymbol{\Psi}} \succ \mathbf{0}, \tag{2.15}$$

and

$$\bar{\boldsymbol{\Psi}}_n^{-1} = \bar{\boldsymbol{\Psi}}^{-1} + o_p(1). \tag{2.16}$$

2.2.1 Conditions for $\sqrt{n}$−consistency of FE estimator

Under the heterogeneous specification (2.11) and noting that $\boldsymbol{M}_T\boldsymbol{\tau}_T = \mathbf{0}$, we have

$$\hat{\boldsymbol{\beta}}_{FE} - \boldsymbol{\beta}_0 = \bar{\boldsymbol{\Psi}}_n^{-1}\left[n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{X}_i(\boldsymbol{\beta}_i - \boldsymbol{\beta}_0)\right] + \bar{\boldsymbol{\Psi}}_n^{-1}\left(n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{u}_i\right). \tag{2.17}$$

Then by Assumption 1, $E\left(\boldsymbol{u}_i\,|\boldsymbol{X}_i\right) = \mathbf{0}$, and hence $E\left(\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{u}_i\right) = E\left[E\left(\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{u}_i|\,\boldsymbol{X}_i\right)\right] = E\left[\boldsymbol{X}_i'\boldsymbol{M}_T E\left(\boldsymbol{u}_i\,|\boldsymbol{X}_i\right)\right] = \mathbf{0}$. Under Assumptions 1, 2 and 3,

$$\hat{\boldsymbol{\beta}}_{FE} - \boldsymbol{\beta}_0 \to_p \bar{\boldsymbol{\Psi}}^{-1}\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}E\left(\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{X}_i\boldsymbol{\eta}_{i\beta}\right),$$

where $\boldsymbol{\eta}_{i\beta} = \boldsymbol{\beta}_i - \boldsymbol{\beta}_0$, and $\hat{\boldsymbol{\beta}}_{FE}$ is a consistent estimator of the average treatment effect, $\boldsymbol{\beta}_0$, if

$$\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}E\left(\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{X}_i\boldsymbol{\eta}_{i\beta}\right) = \mathbf{0}. \tag{2.18}$$

This condition is clearly met if

$$E\left[\left(\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{X}_i\right)\boldsymbol{\eta}_{i\beta}\right] = \boldsymbol{0}, \tag{2.19}$$

for all $i = 1, 2, ..., n$, and has been already derived by Wooldridge (2005).[7] But it is too restrictive, since it is possible for the average condition in (2.18) to hold even though condition (2.19) is violated for some units as $n \to \infty$. What is required is that a sufficiently large number of units satisfy the condition (2.19). Specifically, denote the number of units that *do not* satisfy (2.19) by $m_n = \ominus(n^{a_\eta})$ and note that $n^{-1}\sum_{i=1}^{n} E\left(\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{X}_i\boldsymbol{\eta}_{i\beta}\right) = \ominus(n^{a_\eta - 1})$, and condition (2.18) is met if $a_\eta < 1$. But for $\hat{\boldsymbol{\beta}}_{FE}$ to be a regular $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}_0$ a much more restrictive condition on a_η is required. Using (2.17) note that

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{FE} - \boldsymbol{\beta}_0\right) = \bar{\boldsymbol{\Psi}}_n^{-1}\left(n^{-1/2}\sum_{i=1}^{n}\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{X}_i\boldsymbol{\eta}_{i\beta}\right) + \bar{\boldsymbol{\Psi}}_n^{-1}\left(n^{-1/2}\sum_{i=1}^{n}\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{u}_i\right),$$

and $\sqrt{n}(\hat{\boldsymbol{\beta}}_{FE} - \boldsymbol{\beta}_0) \to_p \boldsymbol{0}$ if $n^{-1/2}\sum_{i=1}^{n}\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{X}_i\boldsymbol{\eta}_{i\beta} \to_p \boldsymbol{0}$. The bias term can be written as

$$\begin{aligned} n^{-1/2}\sum_{i=1}^{n}\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{X}_i\boldsymbol{\eta}_{i\beta} &= n^{-1/2}\sum_{i=1}^{n}\left[\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{X}_i\boldsymbol{\eta}_{i\beta} - E\left(\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{X}_i\boldsymbol{\eta}_{i\beta}\right)\right] \\ &\quad + n^{-1/2}\sum_{i=1}^{n}E\left(\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{X}_i\boldsymbol{\eta}_{i\beta}\right). \end{aligned}$$

The first term tends to zero in probability if $\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{X}_i\boldsymbol{\eta}_{i\beta}$ are weakly cross-correlated over i. For the second term to tend to zero we must have $m_n n^{-1/2} \to 0$, or if $a_\eta < 1/2$.

Proposition 2 (Condition for $\sqrt{n}$−consistency of the FE estimator) *Suppose that y_{it} for $i = 1, 2, ..., n$ and $t = 1, 2, ..., T$ are generated by the heterogeneous panel data model (2.12), and Assumptions 1, 2 and 3 hold. Then the FE estimator given by (2.13) is $\sqrt{n}$-consistent if*

$$n^{-1/2}\sum_{i=1}^{n}E\left[\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{X}_i(\boldsymbol{\beta}_i - \boldsymbol{\beta}_0)\right] \to \boldsymbol{0}, \tag{2.20}$$

and this condition is met if $a_\eta < 1/2$, with a_η defined by $m_n = \ominus(n^{a_\eta})$, where m_n denotes the number of units that are subject to correlated heterogeneity.

2.2.2 Relative efficiency of FE and MG estimators

Suppose now that conditions (2.20) and (2.10) hold and both FE and MG estimators are $\sqrt{n}$-consistent. The choice between the two estimators will then depend on their relative

[7] See equation (12) on page 387 of Wooldridge (2005).

efficiency, which we measure in terms of their asymptotic covariances, given by

$$Var\left(\sqrt{n}\hat{\boldsymbol{\beta}}_{MG}\,|\boldsymbol{X}\right)=\boldsymbol{\Omega}_{\beta}+n^{-1}\sum_{i=1}^{n}\boldsymbol{\Psi}_{ix}^{-1}\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{H}_i\boldsymbol{M}_T\boldsymbol{X}_i\boldsymbol{\Psi}_{ix}^{-1},$$

and

$$\begin{aligned}Var\left(\sqrt{n}\hat{\boldsymbol{\beta}}_{FE}\,|\boldsymbol{X}\right) &= \bar{\boldsymbol{\Psi}}_n^{-1}\left(n^{-1}\sum_{i=1}^{n}\boldsymbol{\Psi}_{ix}\boldsymbol{\Omega}_{\beta}\boldsymbol{\Psi}_{ix}\right)\bar{\boldsymbol{\Psi}}_n^{-1}\\ &\quad+\bar{\boldsymbol{\Psi}}_n^{-1}\left(n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{H}_i\boldsymbol{M}_T\boldsymbol{X}_i\right)\bar{\boldsymbol{\Psi}}_n^{-1},\end{aligned}$$

where $\boldsymbol{X}=(\boldsymbol{X}_1,\boldsymbol{X}_2,...,\boldsymbol{X}_n)$, $\boldsymbol{\Omega}_{\beta}=Var(\boldsymbol{\beta}_i|\boldsymbol{X}_i)\succeq\mathbf{0}$, $\boldsymbol{H}_i=E\left(\boldsymbol{u}_i\boldsymbol{u}_i'|\boldsymbol{X}_i\right)$, and as before $\boldsymbol{\Psi}_{ix}=\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{X}_i$, and $\bar{\boldsymbol{\Psi}}_n=n^{-1}\sum_{i=1}^{n}\boldsymbol{\Psi}_{ix}$. Hence

$$Var\left(\sqrt{n}\hat{\boldsymbol{\beta}}_{MG}\,|\boldsymbol{X}\right)-Var\left(\sqrt{n}\hat{\boldsymbol{\beta}}_{FE}\,|\boldsymbol{X}\right)=\boldsymbol{A}_n+\boldsymbol{B}_n, \tag{2.21}$$

where

$$\boldsymbol{A}_n=\boldsymbol{\Omega}_{\beta}-\bar{\boldsymbol{\Psi}}_n^{-1}\left(n^{-1}\sum_{i=1}^{n}\boldsymbol{\Psi}_{ix}\boldsymbol{\Omega}_{\beta}\boldsymbol{\Psi}_{ix}\right)\bar{\boldsymbol{\Psi}}_n^{-1}, \tag{2.22}$$

and

$$\boldsymbol{B}_n=\left(n^{-1}\sum_{i=1}^{n}\boldsymbol{\Psi}_{ix}^{-1}\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{H}_i\boldsymbol{M}_T\boldsymbol{X}_i\boldsymbol{\Psi}_{ix}^{-1}\right)-\bar{\boldsymbol{\Psi}}_n^{-1}\left(n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{H}_i\boldsymbol{M}_T\boldsymbol{X}_i\right)\bar{\boldsymbol{\Psi}}_n^{-1}. \tag{2.23}$$

$\boldsymbol{A}_n$ and $\boldsymbol{B}_n$ capture the effects of two different types of heterogeneity, namely slope heterogeneity and regressors/errors heterogeneity. The superiority of the FE over MG is readily established when the slope coefficients and error variances are homogeneous across i, and the errors are serially uncorrelated, namely if $\boldsymbol{\Omega}_{\beta}=\mathbf{0}$ and $\boldsymbol{H}_i=\sigma^2\boldsymbol{I}_T$ for all i. In this case $\boldsymbol{A}_n=\mathbf{0}$, and we have

$$\frac{Var\left(\sqrt{n}\hat{\boldsymbol{\beta}}_{MG}\,|\boldsymbol{X}\right)-Var\left(\sqrt{n}\hat{\boldsymbol{\beta}}_{FE}\,|\boldsymbol{X}\right)}{\sigma^2}=n^{-1}\sum_{i=1}^{n}\boldsymbol{\Psi}_{ix}^{-1}-\bar{\boldsymbol{\Psi}}_n^{-1},$$

which is the difference between the harmonic mean of $\boldsymbol{\Psi}_{ix}$ and the inverse of its arithmetic mean, which is a non-negative definite matrix.[8] However, this result may be reversed when we allow for heterogeneous $\boldsymbol{\Omega}_{\beta}\succeq\mathbf{0}$, and/or if $\boldsymbol{H}_i\neq\sigma^2\boldsymbol{I}_T$. The following proposition summarizes

[8] For a proof see the Appendix to Pesaran et al. (1996).

the results of the comparison between the FE and MG estimators.

Proposition 3 (Relative efficiency of MG and FE estimators) *Suppose that y_{it} for $i = 1, 2, ..., n$ and $t = 1, 2, ..., T$ are generated by the heterogeneous panel data model (2.12), and Assumptions 1, 2 and 3 hold, and the uncorrelated heterogeneity condition (2.20) is met. Then $Var\left(\sqrt{n}\hat{\boldsymbol{\beta}}_{MG}\left|\boldsymbol{X}\right.\right) - Var\left(\sqrt{n}\hat{\boldsymbol{\beta}}_{FE}\left|\boldsymbol{X}\right.\right) = \boldsymbol{A}_n + \boldsymbol{B}_n$, where $\boldsymbol{A}_n$ and $\boldsymbol{B}_n$ are given by (2.22) and (2.23), respectively. $\boldsymbol{A}_n$ is a non-positive definite matrix, and the sign of $\boldsymbol{B}_n$ is indeterminate. Under uncorrelated heterogeneity, the FE estimator, $\hat{\boldsymbol{\beta}}_{FE}$, is asymptotically more efficient than the MG estimator if the benefit from pooling (i.e. when $\boldsymbol{B}_n \succ \boldsymbol{0}$) outweighs the loss in efficiency due to slope heterogeneity (since $\boldsymbol{A}_n \preceq \boldsymbol{0}$).*

For a proof see Section A.2.2 of the Appendix.

Example 2 *Consider a simple case where $k' = 1$, $\boldsymbol{\Psi}_{ix} = \psi_{ix}$ and $\boldsymbol{\Omega}_\beta = \sigma_\beta^2$ are scalars, and suppose that $\boldsymbol{H}_i(\boldsymbol{X}_i) = E\left(\boldsymbol{u}_i\boldsymbol{u}_i'\left|\boldsymbol{W}_i\right.\right) = \sigma^2\psi_{ix}\boldsymbol{I}_T$. then*

$$Var\left(\sqrt{n}\hat{\boldsymbol{\beta}}_{MG}\left|\boldsymbol{X}\right.\right) - Var\left(\sqrt{n}\hat{\boldsymbol{\beta}}_{FE}\left|\boldsymbol{X}\right.\right) = -\left(\sigma_\beta^2 + \sigma^2\right)\left[n^{-1}\frac{\sum_{i=1}^n\left(\psi_{ix} - \bar{\psi}_n\right)^2}{\bar{\psi}_n^2}\right],$$

where $\bar{\psi}_n = n^{-1}\sum_{i=1}^n \psi_{ix}$. In this simple case the MG estimator is more efficient than the FE estimator even if $\sigma_\beta^2 = 0$.

In general, with uncorrelated heterogeneous coefficients, the relative efficiency of the MG and FE estimators depends on the relative magnitude of the two components in (2.21). Since $\boldsymbol{A}_n \preceq \boldsymbol{0}$, the outcome depends on the sign and the magnitude of $\boldsymbol{B}_n$, which in turn depends on the heterogeneity of error variances, $\boldsymbol{H}_i(\boldsymbol{X}_i)$ and $\boldsymbol{\Psi}_{ix}$ over i.

3 Irregular mean group estimators

So far we have argued that the MG estimator is robust to correlated heterogeneity, and its performance is comparable to the FE estimator even under uncorrelated heterogeneity. However, since the MG estimator is based on the individual estimates, $\hat{\boldsymbol{\theta}}_i$ for $i = 1, 2, ..., n$, its optimality and robustness critically depend on how well the individual coefficients can be estimated. This is particularly important when T is ultra short, which is the primary concern of this paper. In cases where T is small and/or the observations on $\boldsymbol{w}_{it}$ are highly correlated, or are slowly moving, $d_i = \det\left(\boldsymbol{W}_i'\boldsymbol{W}_i\right)$ is likely to be close to zero in finite samples for a large number of units $i = 1, 2, ..., n$. As a result, $\hat{\boldsymbol{\theta}}_i$ is likely to be a poor estimate of $\boldsymbol{\theta}_i$ for some i, and including it in $\hat{\boldsymbol{\theta}}_{MG}$ could be problematic, rendering the MG estimator inefficient and unreliable.

However, as discussed above, $\hat{\boldsymbol{\theta}}_{MG}$ continues to be an unbiased estimator of $\boldsymbol{\theta}_0$, even if $\boldsymbol{\theta}_i$ are correlated with $\boldsymbol{W}_i$ so long as the stochastic component of $\boldsymbol{w}_{it}$ is strictly exogenous with respect to u_{it}. By averaging over $\hat{\boldsymbol{\theta}}_i$ for $i = 1, 2, ..., n$, as $n \to \infty$, the MG estimator converges to $\boldsymbol{\theta}_0$ if T is sufficiently large such that $\hat{\boldsymbol{\theta}}_i$ have at least second order moments for all i. The existence of first order moments of $\hat{\boldsymbol{\theta}}_i$ is required for the MG estimator to be unbiased, and we need $\hat{\boldsymbol{\theta}}_i$ to have second order moments for $\sqrt{n}$-consistent estimation and valid inference about the average effects, $\boldsymbol{\theta}_0$. Accordingly, we need to distinguish between cases where $\hat{\boldsymbol{\theta}}_i$ have first and second order moments for all i, as compared to cases where some $\hat{\boldsymbol{\theta}}_i$ may not even have first order moments. We refer to the MG estimator based on individual estimates without first or second order moments as the "irregular MG estimator", which is the focus of our analysis. We consider the irregular MG estimator both for models with and without time effects and show how our proposed estimator relates to the literature.

3.1 Graham and Powell estimator

For panels with $T = k$, Graham and Powell (2012) propose a trimmed GMM estimator (denoted as "GP") whereby individual estimates with $|\det(\boldsymbol{W}_i)|$ smaller than a given threshold value, h_n, are omitted from the estimation of $\boldsymbol{\theta}_0$. For now, abstracting from time effects, the GP estimator can be viewed as a trimmed MG estimator given by

$$\hat{\boldsymbol{\theta}}_{GP} = \frac{\sum_{i=1}^{n} \mathbf{1}\{d_i > h_n^2\}\hat{\boldsymbol{\theta}}_i}{\sum_{i=1}^{n} \mathbf{1}\{d_i > h_n^2\}}. \tag{3.1}$$

In the special case where $T = k$, $d_i = |\det(\boldsymbol{W}_i)|^2$, and the trimming procedure based on $|\det(\boldsymbol{W}_i)| > h_n$ is algebraically the same as the one used in (3.1). GP show that to correctly center the limiting distribution of $\hat{\boldsymbol{\theta}}_{GP}$, h_n must be set such that $(nh_n)^{1/2}h_n \to 0$, as $n \to \infty$. For example, for the choice of $h_n = C_{GP}n^{-\alpha_{GP}}$, it is required that $\alpha_{GP} > 1/3$.[9] The GP approach can be viewed as trimming by exclusion and overlooks the information that might be contained in $(\boldsymbol{W}_i'\boldsymbol{W}_i)^*$ when $d_i \leq h_n^2$. In what follows we propose an alternative trimmed MG (TMG) estimator that makes use of this information.

[9]See Section 2 of GP and page 2125 where the use of $h_n = C_0 n^{-1/3}$ is recommended.

4 Trimmed mean group estimators

To motivate the TMG estimator we first introduce the following trimmed estimator of $\boldsymbol{\theta}_i$,

$$\tilde{\boldsymbol{\theta}}_i = \hat{\boldsymbol{\theta}}_i, \text{ if } d_i > a_n, \text{ and } \tilde{\boldsymbol{\theta}}_i = \hat{\boldsymbol{\theta}}_i^*, \text{ if } d_i \leq a_n,$$

where as before $\hat{\boldsymbol{\theta}}_i = (\boldsymbol{W}_i'\boldsymbol{W}_i)^{-1}\boldsymbol{W}_i'\boldsymbol{y}_i$, $d_i = \det(\boldsymbol{W}_i'\boldsymbol{W}_i)$, $\hat{\boldsymbol{\theta}}_i^* = a_n^{-1}(\boldsymbol{W}_i'\boldsymbol{W}_i)^*\boldsymbol{W}_i'\boldsymbol{y}_i$, and

$$a_n = C_n n^{-\alpha}, \tag{4.1}$$

with $\alpha > 0$, and $C_n > 0$ bounded in n. The choice of α and C_n will be discussed below. Written more compactly, we have

$$\tilde{\boldsymbol{\theta}}_i = \mathbf{1}\{d_i > a_n\}\hat{\boldsymbol{\theta}}_i + \mathbf{1}\{d_i \leq a_n\}\hat{\boldsymbol{\theta}}_i^* = (1+\delta_i)\hat{\boldsymbol{\theta}}_i, \tag{4.2}$$

where δ_i is given by

$$\delta_i = \left(\frac{d_i - a_n}{a_n}\right)\mathbf{1}\{d_i \leq a_n\} \leq 0. \tag{4.3}$$

We considered two versions of TMG estimators depending on how individual trimmed estimators, $\tilde{\boldsymbol{\theta}}_i$, are combined. An obvious choice was to use a simple average of $\tilde{\boldsymbol{\theta}}_i$, namely

$$\bar{\tilde{\boldsymbol{\theta}}}_n = n^{-1}\sum_{i=1}^{n}\tilde{\boldsymbol{\theta}}_i = n^{-1}\sum_{i=1}^{n}(1+\delta_i)\hat{\boldsymbol{\theta}}_i, \tag{4.4}$$

which can also be viewed as a weighted average estimator with the weights $w_i = (1+\delta_i)/n < 1/n$. But it is easily seen that these weights do not add up to unity, and it might be desirable to use the scaled weights $w_i/(1+\bar{\delta}_n) = n^{-1}(1+\delta_i)/(1+\bar{\delta}_n)$, where $\bar{\delta}_n = n^{-1}\sum_{i=1}^{n}\delta_i$. Using these modified weights we consider

$$\hat{\boldsymbol{\theta}}_{TMG} = n^{-1}\sum_{i=1}^{n}\left(\frac{1+\delta_i}{1+\bar{\delta}_n}\right)\hat{\boldsymbol{\theta}}_i = \frac{\bar{\tilde{\boldsymbol{\theta}}}_n}{1+\bar{\delta}_n}. \tag{4.5}$$

Although the difference between the two TMG estimators is small for sufficiently large n, it turns out that $\hat{\boldsymbol{\theta}}_{TMG}$ behaves much better in small samples and will be the focus of this paper.

To relate $\bar{\tilde{\boldsymbol{\theta}}}_n$ to the GP estimator given by (3.1), using the above results we note that

$$\bar{\tilde{\boldsymbol{\theta}}}_n = (1-\pi_n)\left(\frac{\sum_{i=1}^{n}\mathbf{1}\{d_i > a_n\}\hat{\boldsymbol{\theta}}_i}{\sum_{i=1}^{n}\mathbf{1}\{d_i > a_n\}}\right) + \pi_n\left(\frac{\sum_{i=1}^{n}\mathbf{1}\{d_i \leq a_n\}\hat{\boldsymbol{\theta}}_i^*}{\sum_{i=1}^{n}\mathbf{1}\{d_i \leq a_n\}}\right), \tag{4.6}$$

where π_n is the fraction of the estimates being trimmed

$$\pi_n = \frac{\sum_{i=1}^{n} \mathbf{1}\{d_i \leq a_n\}}{n}. \tag{4.7}$$

Compared to $\bar{\tilde{\boldsymbol{\theta}}}_n$, the GP estimator ignores the second term in (4.6), and hence places zero weights on the estimates with $d_i \leq a_n$. In contrast, both $\bar{\tilde{\boldsymbol{\theta}}}_n$ and hence $\hat{\boldsymbol{\theta}}_{TMG}$ place non-zero weights on all the individual estimates, $\tilde{\boldsymbol{\theta}}_i$.

5 Asymptotic properties of the TMG estimator

To investigate the asymptotic properties of the TMG estimator, $\hat{\boldsymbol{\theta}}_{TMG}$, we introduce the following additional assumptions:

Assumption 4 *For $i = 1, 2, ..., n$, denote by $d_i = \det(\boldsymbol{W}_i'\boldsymbol{W}_i)$ where $\boldsymbol{W}_i = (\boldsymbol{w}_{i1}, \boldsymbol{w}_{i2}, ..., \boldsymbol{w}_{iT})'$ is the $T \times k$ matrix of observations on $\boldsymbol{w}_{it}$ in the heterogeneous panel data model (2.2). $\inf_i(d_i) > 0$, $\inf_i \lambda_{\min}(\boldsymbol{W}_i'\boldsymbol{W}_i)^* > c > 0$, and $\sup_i E\left[\left\|(\boldsymbol{W}_i'\boldsymbol{W}_i)^*\right\|^2\right] < C$, where $(\boldsymbol{W}_i'\boldsymbol{W}_i)^* = d_i(\boldsymbol{W}_i'\boldsymbol{W}_i)^{-1}$ is the adjoint of $\boldsymbol{W}_i'\boldsymbol{W}_i$.*

Assumption 5 (Distribution of d_i) *For $i = 1, 2, ..., n$, d_i are random draws from the probability distribution function, $F_d(u)$, with the continuously differentiable density function, $f_d(u)$, over $u \in (0, \infty)$, such that $F_d(0) = 0$, $f_d(\bar{a}_n) < C$, and $\left|f_d'(\bar{a}_n)\right| < C$, where $f_d'(\bar{a}_n)$ is the first derivative of $f_d(u)$ evaluated at $\bar{a}_n \in (0, a_n)$.*

Assumption 6 (Characterization of correlation between $\boldsymbol{\theta}_i$ and d_i) *For $i = 1, 2, ..., n$, the dependence of $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, ..., \theta_{ik})'$ on d_i is characterized by (a):*

$$\boldsymbol{\theta}_i = E(\boldsymbol{\theta}_i|d_i) + \boldsymbol{\epsilon}_i, \tag{5.1}$$

where $E(\boldsymbol{\epsilon}_i|d_i) = \mathbf{0}$, and $\sup_i E\left\|\boldsymbol{\epsilon}_i\right\|^4 < C$. (b): Denote

$$\boldsymbol{\eta}_i = \boldsymbol{\theta}_i - \boldsymbol{\theta}_0, \tag{5.2}$$

and

$$\boldsymbol{\psi}_i = E(\boldsymbol{\eta}_i|d_i) = \boldsymbol{B}_i\left\{\boldsymbol{g}(d_i) - E\left[\boldsymbol{g}(d_i)\right]\right\}, \tag{5.3}$$

where $\boldsymbol{g}(u) = (g_1(u), g_2(u), ..., g_k(u))'$ and $g_j(u)$ for $j = 1, 2, ..., k$ are bounded and continuously differentiable functions of u on $(0, \infty)$, and $\boldsymbol{B}_i$ are bounded $k \times k$ matrices of fixed constants with $\sup_i \left\|\boldsymbol{B}_i\right\| < C$. (c) $\boldsymbol{\eta}_i$ are distributed independently over i.

Remark 4 *Under Assumption 4, by imposing* $\inf_i (d_i) > 0$ *and* $F_d(0) = 0$*, we do not consider the case where there is a positive mass of "stayers"in the population, which is the focus of Sasaki and Ura (2021).*

Remark 5 *Under Assumption 5,* d_i *are distributed independently over* i*, which also implies that* δ_i*, defined by (4.3), are also distributed independently over* i*.*

Remark 6 *Under Assumption 6,* $\boldsymbol{\eta}_i$ *can be written as*

$$\boldsymbol{\eta}_i = \boldsymbol{\psi}_i + \boldsymbol{\epsilon}_i, \tag{5.4}$$

where $\boldsymbol{\psi}_i$ *represents the part of the heterogeneity of* $\boldsymbol{\theta}_i$ *that is correlated with* d_i*, and* $\boldsymbol{\epsilon}_i$ *represents random or idiosyncratic heterogeneity which is distributed independently of* d_i*, with* $E(\boldsymbol{\epsilon}_i) = \mathbf{0}$*, for all* i*.*

Remark 7 *Assumptions 5 and 6 can be relaxed by requiring that* $\boldsymbol{\eta}_i$ *and* δ_i *to be weakly cross-sectionally correlated. The cross-sectional independence assumption is maintained to simplify the mathematical exposition.*

Remark 8 *Under Assumption 6, it also follows that* $(1+\delta_i)\boldsymbol{\eta}_i$ *are distributed independently over* i*, although in general* $E(\delta_i \boldsymbol{\eta}_i) \neq \mathbf{0}$*.*

Using (2.6) and (5.2) in (4.2) we have $\tilde{\boldsymbol{\theta}}_i = (1+\delta_i)\boldsymbol{\theta}_0 + \boldsymbol{\zeta}_{iT}$, where $\boldsymbol{\zeta}_{iT} = (1+\delta_i)\left(\boldsymbol{\eta}_i + \boldsymbol{\xi}_{iT}\right)$, and $\hat{\boldsymbol{\theta}}_{TMG}$ defined by (4.5) can be written as

$$\hat{\boldsymbol{\theta}}_{TMG} - \boldsymbol{\theta}_0 = \left(\frac{1}{1+\bar{\delta}_n}\right)\bar{\boldsymbol{\zeta}}_{nT}, \tag{5.5}$$

where $\bar{\boldsymbol{\zeta}}_{nT} = n^{-1}\sum_{i=1}^{n}\boldsymbol{\zeta}_{iT}$. (5.5) can be written equivalently as

$$\hat{\boldsymbol{\theta}}_{TMG} - \boldsymbol{\theta}_0 = \left(\frac{1+E\left(\bar{\delta}_n\right)}{1+\bar{\delta}_n}\right)\left(\boldsymbol{b}_n + n^{-1}\sum_{i=1}^{n}\left[\boldsymbol{p}_i - E\left(\boldsymbol{p}_i\right)\right] + \bar{\boldsymbol{q}}_{nT}\right), \tag{5.6}$$

where

$$\boldsymbol{b}_n = n^{-1}\sum_{i=1}^{n}E\left(\boldsymbol{p}_i\right) = n^{-1}\sum_{i=1}^{n}\frac{E(\delta_i\boldsymbol{\eta}_i)}{1+E\left(\bar{\delta}_n\right)}, \text{ and } \bar{\boldsymbol{q}}_{nT} = n^{-1}\sum_{i=1}^{n}\boldsymbol{q}_{iT}, \tag{5.7}$$

with

$$\boldsymbol{p}_i = \frac{(1+\delta_i)\,\boldsymbol{\eta}_i}{1+E\left(\bar{\delta}_n\right)}, \text{ and } \boldsymbol{q}_{iT} = \frac{(1+\delta_i)\,\boldsymbol{\xi}_{iT}}{1+E\left(\bar{\delta}_n\right)}. \tag{5.8}$$

Under Assumptions 4, 5 and 6 $\delta_i - E(\delta_i)$ and $\boldsymbol{p}_i - E(\boldsymbol{p}_i)$ are distributed independently over i with zero means and bounded variances, and we have

$$\bar{\delta}_n - E\left(\bar{\delta}_n\right) = n^{-1}\sum_{i=1}^{n}\left[\delta_i - E(\delta_i)\right] = O_p(n^{-1/2}), \text{ and } n^{-1}\sum_{i=1}^{n}\left[\boldsymbol{p}_i - E\left(\boldsymbol{p}_i\right)\right] = O_p(n^{-1/2}).$$

Furthermore by Lemma A.1, $E(\delta_i) = O(a_n)$, $E\left(\delta_i\boldsymbol{\eta}_i\right) = O(a_n)$, and it follows that

$$\boldsymbol{b}_n = \frac{1}{1+E\left(\bar{\delta}_n\right)}\left[n^{-1}\sum_{i=1}^{n}E(\delta_i\boldsymbol{\eta}_i)\right] = \frac{O(a_n)}{1+O(a_n)} = O(a_n), \tag{5.9}$$

and

$$\frac{1+E\left(\bar{\delta}_n\right)}{1+\bar{\delta}_n} = 1 - \frac{\bar{\delta}_n - E\left(\bar{\delta}_n\right)}{1+E\left(\bar{\delta}_n\right)+\left(\bar{\delta}_n - E\left(\bar{\delta}_n\right)\right)} = 1 + O_p(n^{-1/2}). \tag{5.10}$$

Also conditional on $\boldsymbol{W}_i$, $\boldsymbol{q}_{iT}$ are distributed independently with mean zeros, and since $\bar{\boldsymbol{q}}_{nt} = n^{-1}\sum_{i=1}^{n}\boldsymbol{q}_{iT} = \left(\frac{1}{1+E\left(\bar{\delta}_n\right)}\right)\bar{\boldsymbol{\xi}}_{\delta,nT}$, where $\bar{\boldsymbol{\xi}}_{\delta,nT} = n^{-1}\sum_{i=1}^{n}\left(1+\delta_i\right)\boldsymbol{\xi}_{iT}$, using results in Lemma A.2 we have $E\left(\bar{\boldsymbol{q}}_{nt}\right) = \boldsymbol{0}$ and

$$Var\left(\bar{\boldsymbol{q}}_{nt}\right) = \left(\frac{1}{1+E\left(\bar{\delta}_n\right)}\right)^2 Var\left(\bar{\boldsymbol{\xi}}_{\delta,nT}\right) = O(n^{-1+\alpha}).$$

Hence, $\bar{\boldsymbol{q}}_{nt} = O_p\left(n^{-1/2+\alpha/2}\right)$. Using these results in (5.6) we have

$$\hat{\boldsymbol{\theta}}_{TMG} - \boldsymbol{\theta}_0 = O(n^{-\alpha}) + O_p\left(n^{-\frac{(1-\alpha)}{2}}\right). \tag{5.11}$$

Hence $\hat{\boldsymbol{\theta}}_{TMG}$ asymptotically converges to $\boldsymbol{\theta}_0$, so long as $0 < \alpha < 1$ as $n \to \infty$. The convergence rate of $\hat{\boldsymbol{\theta}}_{TMG}$ to $\boldsymbol{\theta}_0$ will depend on the trade-off between the asymptotic bias and variance of $\hat{\boldsymbol{\theta}}_{TMG}$. Though it is possible to reduce the bias of $\hat{\boldsymbol{\theta}}_{TMG}$ by choosing a value of α close to unity, it will be at the expense of a large variance. In what follows we shed light on the choice of α by considering the conditions under which the asymptotic distribution of $\hat{\boldsymbol{\theta}}_{TMG}$ is centered around $\boldsymbol{\theta}_0$ so that $Var(\hat{\boldsymbol{\theta}}_{TMG})$ also tends to zero at a reasonably fast rate.

5.1 The choice of the trimming threshold

We begin by assuming that the rate at which $\hat{\boldsymbol{\theta}}_{TMG}$ converges to $\boldsymbol{\theta}_0$ is given by n^{γ}, where γ is set in relation to α. Given the irregular nature of the individual estimators of $\boldsymbol{\theta}_i$ when T is ultra short (for example $T = k$), we expect the rate, n^{γ}, to be below the standard rate of $n^{1/2}$.[10] Using (5.6) and (5.10) and noting that $\gamma \leq 1/2$ (with equality holding only under regular convergence), we have

$$n^{\gamma}\left(\hat{\boldsymbol{\theta}}_{TMG}-\boldsymbol{\theta}_0\right)=n^{\gamma}\boldsymbol{b}_n+n^{\gamma-(1-\alpha)/2}\left[n^{-(1+\alpha)/2}\sum_{i=1}^{n}\left[\boldsymbol{p}_i-E\left(\boldsymbol{p}_i\right)\right]+n^{-(1+\alpha)/2}\sum_{i=1}^{n}\boldsymbol{q}_{iT}\right]+o_p(1). \tag{5.12}$$

To ensure that the asymptotic distribution of $\hat{\boldsymbol{\theta}}_{TMG}$ is correctly centered, we must have $n^{\gamma}\boldsymbol{b}_n \to \mathbf{0}$ as $n \to \infty$. Since $n^{\gamma}\boldsymbol{b}_n = O(n^{\gamma}a_n) = O(n^{\gamma-\alpha})$, this condition is ensured if $\gamma < \alpha$. Turning to the second term of the above, we also note that to obtain a non-degenerate distribution we also need to set $\gamma = (1-\alpha)/2$. Combining these two requirements yields

$$\left(\frac{1-\alpha}{2}\right) < \alpha, \text{ or } \alpha > 1/3, \tag{5.13}$$

which implies that at most the convergence rate of $\hat{\boldsymbol{\theta}}_{TMG}$ can be $n^{1/3}$, well below the standard convergence rate, $n^{1/2}$, which is achieved only if individual estimators of $\boldsymbol{\theta}_i$ have at least second order moments for all i. In practice, we suggest setting α at the boundary value of 1/3 or just above 1/3, which yields the familiar non-parametric convergent rate of 1/3.

5.2 Trimming condition

The condition $\alpha > 1/3$ whilst necessary, it is not sufficient. It is also required that the asymptotic variance of $n^{\gamma}\left(\hat{\boldsymbol{\theta}}_{TMG}-\boldsymbol{\theta}_0\right)$ tends to a positive definite matrix. To this end, setting $\gamma = (1-\alpha)/2$ we first write (5.12) as

$$n^{(1-\alpha)/2}\left(\hat{\boldsymbol{\theta}}_{TMG}-\boldsymbol{\theta}_0\right)=n^{(1-\alpha)/2}\boldsymbol{b}_n+\boldsymbol{z}_{p,n}+\boldsymbol{z}_{q,nT}+o_p(1),$$

where $\boldsymbol{z}_{p,n} = n^{-(1+\alpha)/2}\sum_{i=1}^{n}\left[\boldsymbol{p}_i - E\left(\boldsymbol{p}_i\right)\right]$, and $\boldsymbol{z}_{q,nT} = n^{-(1+\alpha)/2}\sum_{i=1}^{n}\boldsymbol{q}_{iT}$. Recall also that $n^{(1-\alpha)/2}\boldsymbol{b}_n = O\left(n^{(1-3\alpha)/2}\right)$ which becomes negligible since $\alpha > 1/3$, and under Assumption 6, $\boldsymbol{p}_i$ are cross-sectionally independent and we have $Var\left(\boldsymbol{z}_{p,n}\right) = n^{-\alpha}\left[n^{-1}\sum_{i=1}^{n}Var\left(\boldsymbol{p}_i\right)\right] = O(n^{-\alpha})$. Since $E\left(\boldsymbol{z}_{p,n}\right)=\mathbf{0}$, it follows that $\boldsymbol{z}_{p,n} \to_p \mathbf{0}$ at the rate of $a_n^{1/2}$as $n \to \infty$, and hence

$$n^{(1-\alpha)/2}\left(\hat{\boldsymbol{\theta}}_{TMG}-\boldsymbol{\theta}_0\right)=\boldsymbol{z}_{q,nT}+O_p(n^{-\alpha/2})+o_p(1).$$

The first term can be written as $\boldsymbol{z}_{q,nT} = n^{(1-\alpha)/2}\left(\frac{1}{1+E\left(\bar{\delta}_n\right)}\right)\bar{\boldsymbol{\xi}}_{\delta,nT}$. By (A.1.14) of Lemma A.2 and recalling that $E\left(\bar{\delta}_n\right) = O(a_n)$, we have $Var\left(\boldsymbol{z}_{q,nT}\right) = n^{(1-\alpha)}\left(\frac{1}{1+O(a_n)}\right)^2 O(n^{-1+\alpha})$

[10] This issue has also been addressed by Graham and Powell (2012) and Sasaki and Ura (2021).

$= O(1)$, and the asymptotic distribution of $\hat{\boldsymbol{\theta}}_{TMG}$ is determined by that of $\boldsymbol{z}_{q,nT}$. Under Assumption 1, conditional on $\boldsymbol{W}_i$, $\boldsymbol{q}_{iT}$ are independently distributed over i with zero means and $\boldsymbol{z}_{q,nT}$ tends to a normal distribution if $\lim_{n\to\infty} Var\left(\boldsymbol{z}_{q,nT}\right)$ is a positive definite matrix. Using (A.1.13) of Lemma A.2 we note that

$$\begin{aligned} Var\left(\boldsymbol{z}_{q,nT}\right) &= \left(\frac{1}{1+E\left(\bar{\delta}_n\right)}\right)^2 \left\{n^{-1-\alpha}\sum_{i=1}^{n} E\left[\mathbf{1}\{d_i > a_n\}\boldsymbol{R}_i'\boldsymbol{H}_i\boldsymbol{R}_i\right]\right\} \\ &\quad + \left(\frac{1}{1+E\left(\bar{\delta}_n\right)}\right)^2 \left\{n^{-1-\alpha}\sum_{i=1}^{n} a_n^{-2} E\left[d_i^2\mathbf{1}\{d_i \le a_n\}\boldsymbol{R}_i'\boldsymbol{H}_i\boldsymbol{R}_i\right]\right\}, \quad (5.14) \end{aligned}$$

which can be written equivalently as

$$\begin{aligned} Var\left(\boldsymbol{z}_{q,nT}\right) &= C_n^{-1}\left[1+E\left(\bar{\delta}_n\right)\right]^{-2}\left[n^{-1}\sum_{i=1}^{n} E\left[a_n\mathbf{1}\{d_i > a_n\}\boldsymbol{R}_i'\boldsymbol{H}_i\boldsymbol{R}_i\right]\right] \\ &\quad + C_n^{-1}\left[1+E\left(\bar{\delta}_n\right)\right]^{-2}\left[n^{-1}\sum_{i=1}^{n} a_n^{-1} E\left[d_i^2\mathbf{1}\{d_i \le a_n\}\boldsymbol{R}_i'\boldsymbol{H}_i\boldsymbol{R}_i\right]\right]. \quad (5.15) \end{aligned}$$

By (A.1.15) in Lemma A.2 $E\left[n^{-1}\sum_{i=1}^{n} a_n^{-1} d_i^2\mathbf{1}\{d_i \le a_n\}\boldsymbol{R}_i'\boldsymbol{H}_i\boldsymbol{R}_i\right] = O\left(a_n^{1/2}\right)$, and since $E\left(\bar{\delta}_n\right) = O(a_n)$ it then follows that (recall that $0 < C_n < C$)

$$\lim_{n\to\infty} Var\left(\boldsymbol{z}_{q,nT}\right) = C^{-1}\lim_{n\to\infty}\left[n^{-1}\sum_{i=1}^{n} E\left(a_n\mathbf{1}\{d_i > a_n\}\boldsymbol{R}_i'\boldsymbol{H}_i\boldsymbol{R}_i\right)\right].$$

To establish conditions under which $\lim_{n\to\infty} Var\left(\boldsymbol{z}_{q,nT}\right) \succ \mathbf{0}$, note that a $k\times k$ symmetric matrix $\boldsymbol{A}$ is positive definite if $\boldsymbol{p}'\boldsymbol{A}\boldsymbol{p} > 0$, for *all* non-zero vectors $\boldsymbol{p} \in \mathbb{R}^k$. Accordingly, consider

$$\boldsymbol{p}'\left[n^{-1}\sum_{i=1}^{n} a_n\mathbf{1}\{d_i > a_n\}\boldsymbol{R}_i'\boldsymbol{H}_i\boldsymbol{R}_i\right]\boldsymbol{p} = n^{-1}\sum_{i=1}^{n} a_n\mathbf{1}\{d_i > a_n\}\boldsymbol{p}'\boldsymbol{R}_i'\boldsymbol{H}_i\boldsymbol{R}_i\boldsymbol{p},$$

for some $\boldsymbol{p}$ such that $\boldsymbol{p}'\boldsymbol{p} > 0$. Note that

$$n^{-1}\sum_{i=1}^{n} a_n\mathbf{1}\{d_i > a_n\}\boldsymbol{p}'\boldsymbol{R}_i'\boldsymbol{H}_i\boldsymbol{R}_i\boldsymbol{p} \ge n^{-1}\sum_{i=1}^{n} a_n\mathbf{1}\{d_i > a_n\}\left(\boldsymbol{p}'\boldsymbol{R}_i'\boldsymbol{R}_i\boldsymbol{p}\right)\lambda_{min}\left(\boldsymbol{H}_i\right),$$

and since $\boldsymbol{R}_i'\boldsymbol{R}_i = (\boldsymbol{W}_i'\boldsymbol{W}_i)^{-1} = d_i^{-1}(\boldsymbol{W}_i'\boldsymbol{W}_i)^*$, then

$$\boldsymbol{p}'\left[\frac{1}{n}\sum_{i=1}^{n} a_n \mathbf{1}\{d_i > a_n\}\boldsymbol{R}_i'\boldsymbol{H}_i\boldsymbol{R}_i\right]\boldsymbol{p} \geq (\boldsymbol{p}'\boldsymbol{p})\,\frac{1}{n}\sum_{i=1}^{n}\left(\frac{a_n}{d_i}\right)\mathbf{1}\{d_i > a_n\}\lambda_{min}\left[(\boldsymbol{W}_i'\boldsymbol{W}_i)^*\right]\lambda_{min}\left(\boldsymbol{H}_i\right).$$

But by assumption $\inf_i \lambda_{min}(\boldsymbol{H}_i) > c > 0$, and $\inf_i \lambda_{min}\left[(\boldsymbol{W}_i'\boldsymbol{W}_i)^*\right] > c > 0$, (see Assumptions 1 and 4). Hence, a necessary and sufficient condition for $Var\left(\boldsymbol{z}_{q,nT}\right)$ to tend to a positive definite matrix is given by

$$\lim_{n\to\infty}\left[n^{-1}\sum_{i=1}^{n}\left(\frac{a_n}{d_i}\right)\mathbf{1}\{d_i > a_n\}\right] > 0. \tag{5.16}$$

Assumption 7 (Trimming condition) *d_i and $(\boldsymbol{W}_i'\boldsymbol{W}_i)^*$ are jointly distributed such that*

$$\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} E\left[\left(\frac{a_n}{d_i}\right)\mathbf{1}\{d_i > a_n\}\right] > 0 \tag{5.17}$$

where $a_n = C_n n^{-\alpha}$, for $\alpha > 1/3$ and $0 < C_n < C$.

Theorem 1 (Asymptotic distribution of TMG estimator) *Suppose for $i = 1, 2, ..., n$ and $t = 1, 2, ..., T$, y_{it} are generated by the heterogeneous panel data model (2.2), and Assumptions 1-7 hold. Then as $n \to \infty$, for $\alpha > 1/3$, we have*

$$n^{(1-\alpha)/2}\left(\hat{\boldsymbol{\theta}}_{TMG} - \boldsymbol{\theta}_0\right) \to_d N\left(\mathbf{0}_k, \boldsymbol{V}_\theta\right), \tag{5.18}$$

where $\hat{\boldsymbol{\theta}}_{TMG}$ is given by (4.5), and

$$\boldsymbol{V}_\theta = \lim_{n\to\infty}\left(\frac{1}{1 + E\left(\bar{\delta}_n\right)}\right)^2 n^{-(1+\alpha)}\sum_{i=1}^{n} E\left[(1+\delta_i)^2\boldsymbol{R}_i'\boldsymbol{H}_i\boldsymbol{R}_i\right], \tag{5.19}$$

where $\boldsymbol{H}_i = \boldsymbol{H}_i(\boldsymbol{W}_i) = E\left(\boldsymbol{u}_i\boldsymbol{u}_i' \,|\boldsymbol{W}_i\right)$, $\boldsymbol{R}_i = \boldsymbol{W}_i\left(\boldsymbol{W}_i'\boldsymbol{W}_i\right)^{-1}$, $E\left(\bar{\delta}_n\right) = n^{-1}\sum_{i=1}^{n} E(\delta_i)$, $(1+\delta_i)^2 = \mathbf{1}\{d_i > a_n\} + a_n^{-2}d_i^2\mathbf{1}\{d_i \leq a_n\}$, and $d_i = det\left(\boldsymbol{W}_i'\boldsymbol{W}_i\right)$.

5.3 Robust estimation of the covariance matrix of the trimmed MG estimator

As with standard MG estimation, consistent estimation of $\boldsymbol{V}_\theta$ using (5.19) requires knowledge of $\boldsymbol{H}_i$ which cannot be estimated consistently when T is short. Here we follow the literature and propose a robust covariance estimator of $\boldsymbol{V}_\theta$ which is asymptotically unbiased for a wide class of error variances, $E\left(\boldsymbol{u}_i\boldsymbol{u}_i' \,|\boldsymbol{W}_i\right) = \boldsymbol{H}_i(\boldsymbol{W}_i)$, thus allowing for serially correlated and conditionally heteroskedastic errors. The main result is summarized in the following theorem.

Theorem 2 (Robust covariance matrix of TMG estimator) *Suppose Assumptions 4-7 hold, and $\boldsymbol{\theta}_0$ is estimated by $\hat{\boldsymbol{\theta}}_{TMG}$ given by (4.5). Then as $n \to \infty$, for $\alpha > 1/3$,*

$$\lim_{n\to\infty}\left[nVar\left(\hat{\boldsymbol{\theta}}_{TMG}\right)\right] = \operatorname*{plim}_{n\to\infty}\left[n^{-1}\sum_{i=1}^{n}\left(\tilde{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_{TMG}\right)\left(\tilde{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_{TMG}\right)'\right], \tag{5.20}$$

and $Var\left(\hat{\boldsymbol{\theta}}_{TMG}\right)$ can be consistently estimated by $n^{-2}\sum_{i=1}^{n}\left(\tilde{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_{TMG}\right)\left(\tilde{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_{TMG}\right)'$.

See Section A.2.3 of the Appendix for a proof.

Remark 9 *Following the literature on MG estimation here we also consider the following bias-adjusted and scaled version*

$$\widehat{Var(\hat{\boldsymbol{\theta}}_{TMG})} = \frac{1}{n(n-1)(1+\bar{\delta}_n)^2}\sum_{i=1}^{n}\left(\tilde{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_{TMG}\right)\left(\tilde{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_{TMG}\right)'. \tag{5.21}$$

The above results can be readily extended to panel data models with time effects.

6 Ultra short panels with time effects

Allowing for time effects the panel data model (2.11) can be written as

$$y_{it} = \alpha_i + \phi_t + \boldsymbol{x}_{it}'\boldsymbol{\beta}_i + u_{it}, \tag{6.1}$$

where ϕ_t for $t = 1, 2, ..., T$ are the time effects. Without loss of generality we adopt the normalization $\boldsymbol{\tau}_T'\boldsymbol{\phi} = 0$,[11] where $\boldsymbol{\phi} = (\phi_1, \phi_2, ..., \phi_T)'$, and make the following additional assumption:

Assumption 8

$$E\left(\boldsymbol{x}_{it}'\boldsymbol{\eta}_{i\beta}\right) = E\left(\boldsymbol{x}_{is}'\boldsymbol{\eta}_{i\beta}\right), \textit{ for all } t, s = 1, 2, ..., T, \tag{6.2}$$

where $\boldsymbol{\eta}_{i\beta} = \boldsymbol{\beta}_i - \boldsymbol{\beta}_0$, and $\|\boldsymbol{\beta}_0\| < C$.

Remark 10 *Assumption 8 allows for dependence between $\boldsymbol{x}_{it}$ and $\boldsymbol{\eta}_{i\beta}$, but requires this dependence to be time-invariant.*

[11] Graham and Powell (2012) use the normalization $\phi_1 = 0$. The choice of normalization is innocuous for the estimation of the average treatment effects, $\boldsymbol{\beta}_0$.

Remark 11 *The irregular identification of* $\boldsymbol{\phi}$ *when* $T = k$ *in Graham and Powell (2012) is based on moments conditional on the sub-population of "stayers". Under Assumption 4* $d_i > 0$ *for all* i*, i.e., there are no "stayers"in the population, this identification strategy cannot be used. Moreover, GP assume that the joint distribution of* $(u_{it}, \boldsymbol{\theta}_i')'$ *given* $\boldsymbol{W}_i$ *does not depend on* t*, which is similar to Assumption 8. See interpretations of Assumption 1.1 part (ii) on page 2111 in Graham and Powell (2012).*

To estimate $\boldsymbol{\theta}_0 = (\alpha_0, \boldsymbol{\beta}_0')'$, initially we suppose $\boldsymbol{\phi}$ is known. Let

$$\boldsymbol{Q}_i = (1+\delta_i)\,\boldsymbol{W}_i\left(\boldsymbol{W}_i'\boldsymbol{W}_i\right)^{-1}. \tag{6.3}$$

Then the trimmed estimator of $\boldsymbol{\theta}_i = (\alpha_i, \boldsymbol{\beta}_i')'$ is given by $\tilde{\boldsymbol{\theta}}_i(\boldsymbol{\phi}) = \boldsymbol{Q}_i'(\boldsymbol{y}_i - \boldsymbol{\phi}) = \tilde{\boldsymbol{\theta}}_i - \boldsymbol{Q}_i'\boldsymbol{\phi}$, and the associated TMG-TE estimator of $\boldsymbol{\theta}_0$ follows as

$$\hat{\boldsymbol{\theta}}_{TMG-TE}(\boldsymbol{\phi}) = n^{-1}\sum_{i=1}^{n}\left(1+\bar{\delta}_n\right)^{-1}\tilde{\boldsymbol{\theta}}_i(\boldsymbol{\phi}) = \hat{\boldsymbol{\theta}}_{TMG} - \bar{\boldsymbol{Q}}_n'\boldsymbol{\phi},$$

where $\hat{\boldsymbol{\theta}}_{TMG}$ is given by (4.5), and

$$\bar{\boldsymbol{Q}}_n = \frac{1}{1+\bar{\delta}_n}\left(n^{-1}\sum_{i=1}^{n}\boldsymbol{Q}_i\right). \tag{6.4}$$

From our earlier analysis, it is clear that for a known $\boldsymbol{\phi}$, $\hat{\boldsymbol{\theta}}_{TMG-TE}(\boldsymbol{\phi})$ has the same asymptotic distribution as $\hat{\boldsymbol{\theta}}_{TMG}$ with $\boldsymbol{y}_i$ replaced by $\boldsymbol{y}_i - \boldsymbol{\phi}$. We first propose an estimator of $\boldsymbol{\phi}$ for the case where $T \geq k$, and then following Chamberlain (1992) we consider an alternative estimator of $\boldsymbol{\phi}$ with better small sample properties when $T > k$.

6.1 TMG-TE estimator with $T \geq k$

Averaging (6.1) over i,

$$\bar{y}_{\circ t} = \bar{\alpha}_n + \phi_t + \bar{\boldsymbol{x}}_{\circ t}'\boldsymbol{\beta}_0 + \bar{\nu}_{\circ t}, \tag{6.5}$$

where $\bar{\nu}_{\circ t} = n^{-1}\sum_{i=1}^{n}\nu_{it}$, $\nu_{it} = \boldsymbol{x}_{it}'\boldsymbol{\eta}_{i\beta} + u_{it}$, $\bar{y}_{\circ t} = n^{-1}\sum_{i=1}^{n} y_{it}$, $\bar{\boldsymbol{x}}_{\circ t} = n^{-1}\sum_{i=1}^{n}\boldsymbol{x}_{it}$, $\bar{u}_{\circ t} = n^{-1}\sum_{i=1}^{n} u_{it}$, and $\bar{\alpha}_n = n^{-1}\sum_{i=1}^{n}\alpha_i$. Averaging over t, under the normalization $\sum_{t=1}^{T}\phi_t = 0$,

$$\bar{y}_{\circ\circ} = \bar{\alpha} + \bar{\boldsymbol{x}}_{\circ\circ}'\boldsymbol{\beta}_0 + n^{-1}\sum_{i=1}^{n}\bar{\boldsymbol{x}}_{i\circ}'\boldsymbol{\eta}_{i\beta} + \bar{u}_{\circ\circ}, \tag{6.6}$$

where $\bar{y}_{\circ\circ} = T^{-1}\sum_{t=1}^{T}\bar{y}_{\circ t}$, $\bar{\boldsymbol{x}}_{\circ\circ} = T^{-1}\sum_{t=1}^{T}\bar{\boldsymbol{x}}_{\circ t}$ and $\bar{u}_{\circ\circ} = T^{-1}\sum_{t=1}^{T}\bar{u}_{\circ t}$. Subtracting (6.6) from (6.5), yields (noting that $(\bar{\boldsymbol{x}}_{\circ t} - \bar{\boldsymbol{x}}_{\circ\circ})'\boldsymbol{\beta}_0 = (\bar{\boldsymbol{w}}_{\circ t} - \bar{\boldsymbol{w}}_{\circ\circ})'\boldsymbol{\theta}_0$)

$$\phi_t = (\bar{y}_{\circ t} - \bar{y}_{\circ\circ}) - (\bar{\boldsymbol{w}}_{\circ t} - \bar{\boldsymbol{w}}_{\circ\circ})'\boldsymbol{\theta}_0 - (\bar{\nu}_{\circ t} - \bar{\nu}_{\circ\circ}), \text{ for } t = 1, 2, ..., T, \tag{6.7}$$

where $\bar{\nu}_{\circ t} - \bar{\nu}_{\circ\circ} = (\bar{u}_{\circ t} - \bar{u}_{\circ\circ}) + n^{-1}\sum_{i=1}^{n}(\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_{i\circ})'\boldsymbol{\eta}_{i\beta}$. Under Assumptions 1, 6 and 8, $\bar{\nu}_{\circ t} - \bar{\nu}_{\circ\circ} = O_p(n^{-1/2})$,[12] which suggests the following estimator of ϕ_t

$$\hat{\phi}_t = (\bar{y}_{\circ t} - \bar{y}_{\circ\circ}) - (\bar{\boldsymbol{w}}_{\circ t} - \bar{\boldsymbol{w}}_{\circ\circ})'\hat{\boldsymbol{\theta}}_{TMG-TE}, \text{ for } t = 1, 2, ..., T, \tag{6.8}$$

where

$$\hat{\boldsymbol{\theta}}_{TMG-TE} = \hat{\boldsymbol{\theta}}_{TMG} - \bar{\boldsymbol{Q}}_n'\hat{\boldsymbol{\phi}}. \tag{6.9}$$

Stacking the equations in (6.8) over $t = 1, 2, ..., T$ we have

$$\hat{\boldsymbol{\phi}} = \boldsymbol{M}_T\left(\bar{\boldsymbol{y}} - \bar{\boldsymbol{W}}\hat{\boldsymbol{\theta}}_{TMG-TE}\right), \tag{6.10}$$

where $\boldsymbol{M}_T = \boldsymbol{I}_T - T^{-1}\boldsymbol{\tau}_T\boldsymbol{\tau}_T'$, $\bar{\boldsymbol{y}} = n^{-1}\sum_{i=1}^{n}\boldsymbol{y}_i$, and $\bar{\boldsymbol{W}} = n^{-1}\sum_{i=1}^{n}\boldsymbol{W}_i$. The above system of equations can now be solved in terms of $\hat{\boldsymbol{\theta}}_{TMG}$ if $(\boldsymbol{I}_T - \boldsymbol{M}_T\bar{\boldsymbol{W}}\bar{\boldsymbol{Q}}_n')$ is non-singular. Under this condition we have

$$\hat{\boldsymbol{\phi}} = \left(\boldsymbol{I}_T - \boldsymbol{M}_T\bar{\boldsymbol{W}}\bar{\boldsymbol{Q}}_n'\right)^{-1}\boldsymbol{M}_T\left(\bar{\boldsymbol{y}} - \bar{\boldsymbol{W}}\hat{\boldsymbol{\theta}}_{TMG}\right) \tag{6.11}$$

and substituting $\hat{\boldsymbol{\phi}}$ from (6.10) in (6.9) we have

$$\hat{\boldsymbol{\theta}}_{TMG-TE} = \left(\boldsymbol{I}_k - \bar{\boldsymbol{Q}}_n'\boldsymbol{M}_T\bar{\boldsymbol{W}}\right)^{-1}\left(\hat{\boldsymbol{\theta}}_{TMG} - \bar{\boldsymbol{Q}}_n'\boldsymbol{M}_T\bar{\boldsymbol{y}}\right). \tag{6.12}$$

Remark 12 *Note that* $\left(\boldsymbol{M}_T\bar{\boldsymbol{W}}\right)\bar{\boldsymbol{Q}}_n'$ *and* $\bar{\boldsymbol{Q}}_n'\left(\boldsymbol{M}_T\bar{\boldsymbol{W}}\right)$ *have the same* k *(*$k \leq T$*) non-zero eigenvalues,* $\det\left(\boldsymbol{I}_T - \boldsymbol{M}_T\bar{\boldsymbol{W}}\bar{\boldsymbol{Q}}_n'\right) = \det\left(\boldsymbol{I}_k - \bar{\boldsymbol{Q}}_n'\boldsymbol{M}_T\bar{\boldsymbol{W}}\right)$*, and if* $\left(\boldsymbol{I}_T - \boldsymbol{M}_T\bar{\boldsymbol{W}}\bar{\boldsymbol{Q}}_n'\right)$ *is invertible so will* $\left(\boldsymbol{I}_k - \bar{\boldsymbol{Q}}_n'\boldsymbol{M}_T\bar{\boldsymbol{W}}\right)$.

The following theorem provides a summary of the results for estimation of $\boldsymbol{\phi}_0$ and $\boldsymbol{\theta}_0$, and their asymptotic distributions.

Theorem 3 (Asymptotic distribution of $\hat{\boldsymbol{\theta}}_{TMG-TE}$ and the time effects $\hat{\boldsymbol{\phi}}$ when $T \geq k$) *Suppose that for* $i = 1, 2, ..., n$ *and* $t = 1, 2, ..., T$, y_{it} *are generated by (6.1),* $T \geq k$*, Assumptions 1-8 hold, and* $\boldsymbol{I}_k - \bar{\boldsymbol{Q}}_n'\boldsymbol{M}_T\bar{\boldsymbol{W}}$ *is invertible where* $\bar{\boldsymbol{Q}}_n$ *is given by (6.4), and*

[12] For a proof see Lemma A.3.

$\bar{\boldsymbol{W}} = n^{-1}\sum_{i=1}^{n}\boldsymbol{W}_i$. *Then as* $n \to \infty$, *for* $\alpha > 1/3$,

$$n^{(1-\alpha)/2}\left(\hat{\boldsymbol{\theta}}_{TMG-TE} - \boldsymbol{\theta}_0\right) \to_d N\left(\mathbf{0}_k, \boldsymbol{V}_{\theta,TMG-TE}\right), \tag{6.13}$$

where $\hat{\boldsymbol{\theta}}_{TMG-TE}$ *is given by (6.12),*

$$\boldsymbol{V}_{\theta,TMG-TE} = \left(\boldsymbol{I}_k - \boldsymbol{G}_w\right)^{-1}\boldsymbol{V}_{\theta}(\boldsymbol{\phi})\left(\boldsymbol{I}_k - \boldsymbol{G}_w'\right)^{-1},$$

$\boldsymbol{G}_w = \lim_{n\to\infty}\left(\bar{\boldsymbol{Q}}_n'\boldsymbol{M}_T\bar{\boldsymbol{W}}\right)$, *and* $\boldsymbol{V}_{\theta}(\boldsymbol{\phi}) = \lim_{n\to\infty} Var\left[n^{(1-\alpha)/2}\hat{\boldsymbol{\theta}}_{TMG-TE}(\boldsymbol{\phi})\right]$. *Also*

$$\hat{\boldsymbol{\phi}} = \boldsymbol{M}_T\left(\bar{\boldsymbol{y}} - \bar{\boldsymbol{W}}\hat{\boldsymbol{\theta}}_{TMG-TE}\right) = \boldsymbol{M}_T\left(\bar{\boldsymbol{y}} - \bar{\boldsymbol{X}}\hat{\boldsymbol{\beta}}_{TMG-TE}\right).$$

(a) If $\operatorname{plim}_{n\to\infty}\boldsymbol{M}_T\bar{\boldsymbol{X}} = \mathbf{0}$, *we have*

$$\sqrt{n}\left(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0\right) \to_d N(\mathbf{0}_T, \boldsymbol{M}_T\boldsymbol{\Omega}_{\nu}\boldsymbol{M}_T), \tag{6.14}$$

where $\boldsymbol{\Omega}_{\nu} = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}E\left(\boldsymbol{\nu}_i\boldsymbol{\nu}_i'\right)$, $\boldsymbol{\nu}_i = (\nu_{i1}, \nu_{i2}, ..., \nu_{iT})'$, *and* $\nu_{it} = u_{it} + \boldsymbol{x}_{it}'\boldsymbol{\eta}_{i\beta}$.

(b) If $\operatorname{plim}_{n\to\infty}\boldsymbol{M}_T\bar{\boldsymbol{X}} \neq \mathbf{0}$, *for* $\alpha > 1/3$, *we have*

$$n^{(1-\alpha)/2}\left(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0\right) \to_d N\left(\mathbf{0}_T, \boldsymbol{V}_{\phi}\right), \tag{6.15}$$

where $\boldsymbol{V}_{\phi} = \operatorname{plim}_{n\to\infty}\boldsymbol{M}_T\bar{\boldsymbol{X}}Var\left(n^{(1-\alpha)/2}\hat{\boldsymbol{\beta}}_{TMG-TE}\right)\bar{\boldsymbol{X}}'\boldsymbol{M}_T$.

A proof is given in Section A.3 of the Appendix. Using results similar to the ones employed to establish Theorem 2, robust covariance matrices for $\hat{\boldsymbol{\theta}}_{TMG-TE}$ and $\hat{\boldsymbol{\phi}}$ are given by (A.3.4) and (A.3.7), respectively, in Section A.3 of the Appendix. In particular, the asymptotic covariance of $\hat{\boldsymbol{\phi}}$ is applicable to both cases (a) and (b) of Theorem 3, and does not require knowing if $\operatorname{plim}_{n\to\infty}\boldsymbol{M}_T\bar{\boldsymbol{X}} = \mathbf{0}$, or not.

Example 3 *As an example of case (a) in Theorem 3, suppose* $\boldsymbol{x}_{it} = \boldsymbol{\alpha}_{ix} + \boldsymbol{u}_{x,it}$, *where* $\boldsymbol{u}_{x,it}$ *are distributed independently over i with zero means. Then* $\bar{\boldsymbol{x}}_{\circ t} - \bar{\boldsymbol{x}}_{\circ\circ} = \bar{\boldsymbol{u}}_{x,\circ t} - \bar{\boldsymbol{u}}_{x,\circ\circ} \to_p \mathbf{0}$, *and we have* $\operatorname{plim}_{n\to\infty}\boldsymbol{M}_T\bar{\boldsymbol{X}} = \mathbf{0}$. *An example of case (b) arises when* $\boldsymbol{x}_{it}$ *contains an interactive effect, namely* $\boldsymbol{x}_{it} = \boldsymbol{\alpha}_{ix} + \boldsymbol{\Gamma}_i\boldsymbol{f}_t + \boldsymbol{u}_{x,it}$. *In this case* $\bar{\boldsymbol{x}}_{\circ t} - \bar{\boldsymbol{x}}_{\circ\circ} = \bar{\boldsymbol{\Gamma}}\left(\boldsymbol{f}_t - \bar{\boldsymbol{f}}\right) + \bar{\boldsymbol{u}}_{x,\circ t} - \bar{\boldsymbol{u}}_{x,\circ\circ}$, *where* $\bar{\boldsymbol{\Gamma}} = n^{-1}\sum_{i=1}^{n}\boldsymbol{\Gamma}_i \to_p \boldsymbol{\Gamma}$, *and it follows that* $\bar{\boldsymbol{x}}_{\circ t} - \bar{\boldsymbol{x}}_{\circ\circ} \to_p \boldsymbol{\Gamma}\left(\boldsymbol{f}_t - \bar{\boldsymbol{f}}\right)$ *which is non-zero if* $\boldsymbol{f}_t$ *varies over time and* $\boldsymbol{\Gamma} \neq \mathbf{0}$, *namely at least one of the factors has loadings with non-zero means.*

6.2 TMG-C estimator when $T > k$

When $T > k$, we can follow Chamberlain (1992) and eliminate the time effects by the demeaning transformation $\boldsymbol{M}_i = \boldsymbol{I}_T - \boldsymbol{M}_T\boldsymbol{X}_i(\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\boldsymbol{M}_T$. Under the normalization $\boldsymbol{\tau}_T'\boldsymbol{\phi} = 0$, $\boldsymbol{M}_T\boldsymbol{\phi} = \boldsymbol{\phi}$, and we have $\boldsymbol{M}_T\boldsymbol{y}_i = \boldsymbol{M}_T\boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{\phi} + \boldsymbol{M}_T\boldsymbol{u}_i$. Then $\boldsymbol{M}_i\boldsymbol{M}_T\boldsymbol{y}_i = \boldsymbol{M}_i\boldsymbol{\phi} + \boldsymbol{M}_i\boldsymbol{M}_T\boldsymbol{u}_i$, and averaging over i we obtain

$$n^{-1}\sum_{i=1}^{n}\boldsymbol{M}_i\boldsymbol{M}_T\boldsymbol{y}_i = \left(n^{-1}\sum_{i=1}^{n}\boldsymbol{M}_i\right)\boldsymbol{\phi} + n^{-1}\sum_{i=1}^{n}\boldsymbol{M}_i\boldsymbol{M}_T\boldsymbol{u}_i. \tag{6.16}$$

Hence, $\boldsymbol{\phi}$ can be estimated if $\bar{\boldsymbol{M}}_n = n^{-1}\sum_{i=1}^{n}\boldsymbol{M}_i$ is a positive definite matrix, without knowing $\boldsymbol{\theta}_0$. This requires $T > k$, since $\bar{\boldsymbol{M}}_n$ is singular if $T = k$. Therefore, to implement the Chamberlain estimation approach we require the following additional assumption:

Assumption 9 *For* $T > k$, $\bar{\boldsymbol{M}}_n = n^{-1}\sum_{i=1}^{n}\boldsymbol{M}_i \to_p \boldsymbol{M} \succ \boldsymbol{0}$, *where*

$$\boldsymbol{M}_i = \boldsymbol{I}_T - \boldsymbol{M}_T\boldsymbol{X}_i(\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\boldsymbol{M}_T.$$

Under this Assumption $\boldsymbol{\phi}$ can be estimated by

$$\hat{\boldsymbol{\phi}}_C = \left(n^{-1}\sum_{i=1}^{n}\boldsymbol{M}_i\right)^{-1}\left(n^{-1}\sum_{i=1}^{n}\boldsymbol{M}_i\boldsymbol{M}_T\boldsymbol{y}_i\right), \tag{6.17}$$

and its asymptotic distribution follows straightforwardly. Specifically, using (6.16) we have

$$\sqrt{n}\left(\hat{\boldsymbol{\phi}}_C - \boldsymbol{\phi}\right) = \bar{\boldsymbol{M}}_n^{-1}\left(n^{-1/2}\sum_{i=1}^{n}\boldsymbol{M}_i\boldsymbol{M}_T\boldsymbol{u}_i\right), \tag{6.18}$$

and $\sqrt{n}\left(\hat{\boldsymbol{\phi}}_C - \boldsymbol{\phi}_0\right) \to_d N(\boldsymbol{0}, \boldsymbol{V}_{\phi,C})$, where

$$\boldsymbol{V}_{\phi,C} = \boldsymbol{M}^{-1}\lim_{n\to\infty} E\left(n^{-1}\sum_{i=1}^{n}\boldsymbol{M}_i\boldsymbol{M}_T\boldsymbol{u}_i\boldsymbol{u}_i'\boldsymbol{M}_T\boldsymbol{M}_i\right)\boldsymbol{M}^{-1}.$$

Since $\boldsymbol{M}_i\boldsymbol{M}_T\boldsymbol{u}_i = \boldsymbol{M}_i\boldsymbol{M}_T(\boldsymbol{y}_i - \boldsymbol{\phi})$, $Var\left(\hat{\boldsymbol{\phi}}_C\right)$ can be consistently estimated by

$$\widehat{Var\left(\hat{\boldsymbol{\phi}}_C\right)} = n^{-1}\bar{\boldsymbol{M}}_n^{-1}\left[n^{-1}\sum_{i=1}^{n}\boldsymbol{M}_i\boldsymbol{M}_T(\boldsymbol{y}_i - \hat{\boldsymbol{\phi}}_C)(\boldsymbol{y}_i - \hat{\boldsymbol{\phi}}_C)'\boldsymbol{M}_T\boldsymbol{M}_i\right]\bar{\boldsymbol{M}}_n^{-1}. \tag{6.19}$$

Using $\hat{\boldsymbol{\phi}}_C$, the TMG-C estimator of $\boldsymbol{\theta}_0$ is now given by

$$\hat{\boldsymbol{\theta}}_{TMG-C} = \frac{1}{1+\bar{\delta}_n}\left[n^{-1}\sum_{i=1}^{n}\boldsymbol{Q}_i'(\boldsymbol{y}_i - \hat{\boldsymbol{\phi}}_C)\right]. \tag{6.20}$$

Also since $\hat{\boldsymbol{\theta}}_{TMG-C} = \hat{\boldsymbol{\theta}}_{TMG-C}(\boldsymbol{\phi}) - \bar{\boldsymbol{Q}}_n'(\hat{\boldsymbol{\phi}}_C - \boldsymbol{\phi})$, the asymptotic variance of $\hat{\boldsymbol{\theta}}_{TMG-C}$ can be consistently estimated by

$$\widehat{Var\left(\hat{\boldsymbol{\theta}}_{TMG-C}\right)} = \widehat{Var\left(\hat{\boldsymbol{\theta}}_{TMG-C}(\boldsymbol{\phi})\right)} + \bar{\boldsymbol{Q}}_n'\widehat{Var\left(\hat{\boldsymbol{\phi}}_C\right)}\bar{\boldsymbol{Q}}_n, \tag{6.21}$$

where $\widehat{Var\left(\hat{\boldsymbol{\theta}}_{TMG-C}(\boldsymbol{\phi})\right)} = n^{-1}(n-1)^{-1}(1+\bar{\delta}_n)^{-2}\sum_{i=1}^{n}\left(\tilde{\boldsymbol{\theta}}_{i,C} - \hat{\boldsymbol{\theta}}_{TMG-C}\right)\left(\tilde{\boldsymbol{\theta}}_{i,C} - \hat{\boldsymbol{\theta}}_{TMG-C}\right)'$, and $\tilde{\boldsymbol{\theta}}_{i,C} = \boldsymbol{Q}_i'(\boldsymbol{y}_i - \hat{\boldsymbol{\phi}}_C)$.

7 A Hausman-type test of the validity of the FE estimator

As summarized by Proposition 2, the validity of the FE estimator depends on the independence of slope heterogeneity, $\boldsymbol{\eta}_{i\beta} = \boldsymbol{\beta}_i - \boldsymbol{\beta}_0$, from the covariates, $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{iT})'$. Here we propose a Hausman-type test of this condition when T is ultra short, under the null hypothesis

$$H_0 : E\left(\boldsymbol{\eta}_{i\beta}\,|\boldsymbol{x}_{it}\right) = \mathbf{0}, \text{ for all } i \text{ and } t. \tag{7.1}$$

It is clear that the homogeneous alternative, $\boldsymbol{\eta}_{i\beta} = \mathbf{0}$, for all i, and the uncorrelated alternative, $E\left[(\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{X}_i)\boldsymbol{\eta}_{i\beta}\right] = \mathbf{0}$, for all i, are both implied by H_0. But a less restrictive null can also be entertained by allowing $E\left(\boldsymbol{\eta}_{i\beta}\,|\boldsymbol{X}_i\right) \neq \mathbf{0}$, for $i = 1, 2, ..., n^{a_\eta}$, so long as $a_\eta < 1/2$, namely the number of violations of the null over the units $i = 1, 2, ..., n$ is relatively few. This is in line with condition (2.20) that requires $n^{-1/2}\sum_{i=1}^{n}E\left(\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{X}_i\boldsymbol{\eta}_{i\beta}\right) \to \mathbf{0}$, which is the implicit null of the Hausman-type test. But to simplify the derivations we derive the tests under H_0.

Consider the FE and TMG estimators defined by (2.13) and (4.5) respectively. Then a Hausman-type test of H_0 can be constructed based on the difference $\hat{\boldsymbol{\Delta}}_\beta = \hat{\boldsymbol{\beta}}_{FE} - \hat{\boldsymbol{\beta}}_{TMG}$. Such a test has been considered by Pesaran et al. (1996) and Pesaran and Yamagata (2008), assuming the MG estimator has at least the second order moment.[13] Here we extend this test to cover cases when T is ultra short. Also, the earlier tests were derived under the null of homogeneity (namely $\boldsymbol{\eta}_{i\beta} = \mathbf{0}$, for all i), whilst the null that we are considering is more

[13] See pages 160–162 of Pesaran et al. (1996), and page 53 of Pesaran and Yamagata (2008).

general and covers the null of homogeneity as a special case.

First recall from (2.17) and (4.5) that $\hat{\boldsymbol{\beta}}_{FE} - \boldsymbol{\beta}_0 = \bar{\boldsymbol{\Psi}}_n^{-1}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{\nu}_i\right)$ and $\hat{\boldsymbol{\beta}}_{TMG} - \boldsymbol{\beta}_0 = \frac{1}{n}\sum_{i=1}^n \left(\frac{1+\delta_i}{1+\bar{\delta}_n}\right)\boldsymbol{\Psi}_{ix}^{-1}\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{\nu}_i$, where $\bar{\boldsymbol{\Psi}}_n = n^{-1}\sum_{i=1}^n \boldsymbol{\Psi}_{ix}$, $\boldsymbol{\Psi}_{ix} = \boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{X}_i$, δ_i is given by (4.3), and $\boldsymbol{\nu}_i = \boldsymbol{u}_i + \boldsymbol{X}_i\boldsymbol{\eta}_{i\beta}$. Also by Assumption 3, $\bar{\boldsymbol{\Psi}}_n \to_p \lim_{n\to\infty} n^{-1}\sum_{i=1}^n E\left(\boldsymbol{\Psi}_{ix}\right) = \bar{\boldsymbol{\Psi}} \succ \mathbf{0}$, and $\bar{\boldsymbol{\Psi}}_n^{-1} - \bar{\boldsymbol{\Psi}}^{-1} = o_p(1)$. Using these results it follows that

$$\sqrt{n}\hat{\boldsymbol{\Delta}}_\beta = \frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{G}_i'\boldsymbol{M}_T\boldsymbol{\nu}_i + o_p(1),$$

where $\boldsymbol{G}_i' = \left[\bar{\boldsymbol{\Psi}}^{-1} - \left(\frac{1+\delta_i}{1+\bar{\delta}_n}\right)\boldsymbol{\Psi}_{ix}^{-1}\right]\boldsymbol{X}_i'$. Under H_0 and Assumption 1, $E(\nu_{it}|\boldsymbol{G}_i) = 0$ for all i and t, and since by Assumptions 1 and part (c) of Assumption 6 u_{it} and $\boldsymbol{\eta}_{i\beta}$ are cross-sectionally independent, then conditional on $\boldsymbol{X}_i$, $\boldsymbol{\nu}_i$ are also cross-sectionally independent and we have $\sqrt{n}\hat{\boldsymbol{\Delta}}_\beta \to_d N(\mathbf{0}, \boldsymbol{V}_\Delta)$ as $n \to \infty$, so long as

$$\boldsymbol{V}_\Delta = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n E\left(\boldsymbol{G}_i'\boldsymbol{M}_T E\left(\boldsymbol{\nu}_i\boldsymbol{\nu}_i'\,|\boldsymbol{X}_i\right)\boldsymbol{M}_T\boldsymbol{G}_i\right) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n E\left(\boldsymbol{G}_i'\boldsymbol{M}_T\boldsymbol{V}_{i\nu}\boldsymbol{M}_T\boldsymbol{G}_i\right) \succ \mathbf{0},$$

where $\boldsymbol{V}_{i\nu} = \boldsymbol{H}_i + \boldsymbol{X}_i\boldsymbol{\Omega}_\beta\boldsymbol{X}_i'$, and $\boldsymbol{\Omega}_\beta = E\left(\boldsymbol{\eta}_{i\beta}\boldsymbol{\eta}_{i\beta}'\right)$. Hence $H_\beta = n\hat{\boldsymbol{\Delta}}_\beta'\boldsymbol{V}_\Delta^{-1}\hat{\boldsymbol{\Delta}}_\beta \to_d \chi^2_{k'}$, where $\chi^2_{k'}$ is a chi-squared distribution with $k' = dim(\boldsymbol{\beta})$ degree of freedom. Note that $\boldsymbol{V}_\Delta$ can be written equivalently as $\boldsymbol{V}_\Delta = n^{-1}\sum_{i=1}^n\sum_{t=1}^T\sum_{t'=1}^T E\left(\boldsymbol{g}_{it}\boldsymbol{g}_{it'}'\tilde{\nu}_{it}\tilde{\nu}_{it'}\right)$, where $\tilde{\nu}_{it} = \nu_{it} - \bar{\nu}_{i\circ}$, and $\boldsymbol{g}_{it}$ is the t^{th} column of $\boldsymbol{G}_i'$. For fixed T, a consistent estimator of $\boldsymbol{V}_\Delta$, which is robust to the choices of $\boldsymbol{H}_i$ and $\boldsymbol{\Omega}_\beta$, can be obtained given by $\widehat{\boldsymbol{V}}_\Delta = \frac{1}{n}\sum_{i=1}^n\sum_{t=1}^T\sum_{t'=1}^T \hat{\boldsymbol{g}}_{it}\hat{\boldsymbol{g}}_{it'}'\hat{\tilde{\nu}}_{it}\hat{\tilde{\nu}}_{it'}$, where $\hat{\tilde{\nu}}_{it} = (y_{it} - \bar{y}_{i\circ}) - \hat{\boldsymbol{\beta}}_{FE}'(\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_{i\circ})$, $\bar{y}_{i\circ} = T^{-1}\sum_{t=1}^T y_{it}$, $\bar{\boldsymbol{x}}_{i\circ} = T^{-1}\sum_{t=1}^T \boldsymbol{x}_{it}$, with $\hat{\boldsymbol{g}}_{it}$ being the t^{th} column of $\hat{\boldsymbol{G}}_i'$ given by

$$\hat{\boldsymbol{G}}_i' = \left[\left(n^{-1}\sum_{i=1}^n \boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{X}_i\right)^{-1} - \left(\frac{1+\delta_i}{1+\bar{\delta}_n}\right)\left(\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{X}_i\right)^{-1}\right]\boldsymbol{X}_i'. \tag{7.2}$$

Using the above estimator of $\boldsymbol{V}_\Delta$, the Hausman-type test statistic for correlated slope heterogeneity is given by

$$\hat{H}_\beta = n\left(\hat{\boldsymbol{\beta}}_{FE} - \hat{\boldsymbol{\beta}}_{TMG}\right)'\widehat{\boldsymbol{V}}_\Delta^{-1}\left(\hat{\boldsymbol{\beta}}_{FE} - \hat{\boldsymbol{\beta}}_{TMG}\right). \tag{7.3}$$

Under the alternative hypothesis that $H_1 : \lim_{n\to\infty} n^{-1}\sum_{i=1}^n E\left(\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{X}_i\boldsymbol{\eta}_{i\beta}\right) = \boldsymbol{Q}_\eta \succ \mathbf{0}$, $\hat{H}_\beta \to_p \infty$, as $n \to \infty$, and the test is consistent. The extension of the $\hat{H}_\beta$ test to panel data models with time effects are provided in Section S.2 of the online supplement.

8 Monte Carlo evidence on small sample properties

Using Monte Carlo (MC) techniques, we now consider the small-sample properties of the TMG estimator and compare its performance with the FE, MG and GP estimators, as well as the recent estimator proposed by Sasaki and Ura (SU).[14] We also provide MC evidence on the estimation of panels with time effects. The finite-sample performance of the Hausman-type test of correlated slope heterogeneity is also examined.

8.1 Monte Carlo designs

8.1.1 Data generating processes (DGP)

The outcome variable, y_{it}, is generated as

$$y_{it} = \alpha_i + \phi_t + \beta_i x_{it} + \kappa \sigma_{it} e_{it}, \text{ for } i = 1, 2, ..., n, \text{ and } t = 1, 2, ..., T, \tag{8.1}$$

where we allow for heteroskedastic and serially correlated errors. We generate e_{it} as first order autoregressive (AR(1)) processes

$$e_{it} = \rho_{ie} e_{i,t-1} + \left(1 - \rho_{ie}^2\right)^{1/2} \varsigma_{it}, \tag{8.2}$$

and consider two scenarios for ς_{it}, namely Gaussian $\varsigma_{it} \sim IIDN(0,1)$, and chi-squared, $\varsigma_{it} \sim IID\frac{1}{2}\left(\chi_2^2 - 2\right)$. We also allow the shocks in the outcome equation, denoted by $u_{it} = \sigma_{it} e_{it}$, to be cross-sectionally heteroskedastic. In the baseline model we generate $\sigma_{it} = \sigma_{iu}$ for all t where $\sigma_{iu}^2 \sim IID\frac{1}{2}\left(1 + z_{iu}^2\right)$, with $z_{iu} \sim IIDN(0,1)$. We also consider the robustness of the MC results to cases where σ_{it}^2 also varies with x_{it}, as detailed below in Section 8.1.3.

The regressors, x_{it}, are generated as factor-augmented AR processes

$$x_{it} = \alpha_{ix}(1 - \rho_{ix}) + \gamma_{ix} f_t + \rho_{ix} x_{i,t-1} + \left(1 - \rho_{ix}^2\right)^{1/2} u_{x,it}, \tag{8.3}$$

where $u_{x,it} = \sigma_{ix} e_{x,it}$, for $i = 1, 2, ..., n$, and $t = 1, 2, ..., T$. We generate the individual effects in x_{it}, α_{ix}, as $\alpha_{ix} \sim IIDN(1,1)$, with $e_{x,it} \sim IID(0,1)$, $\sigma_{ix}^2 = \frac{1}{2}\left(1 + z_{ix}^2\right)$, and $z_{ix} \sim IIDN(0,1)$. When time effects are included in the model, we set $\phi_t = t$, for $t = 1, 2, ..., T-1$, and $\phi_T = -T(T-1)/2$, so that $\boldsymbol{\tau}_T' \boldsymbol{\phi} = 0$.

[14] Perhaps it should be noted that the SU estimator is intended for a more general setup that allows for many stayers (units with $\boldsymbol{x}_{it} = \boldsymbol{x}_{it'}$ for some $t \neq t'$) which we do not allow in our analysis. We are grateful to Sasaki and Ura for providing us with their codes written specifically for the case when $T = k = 2$.

8.1.2 Generation of the heterogeneous coefficients

We consider both correlated and uncorrelated effects specifications and generate $\boldsymbol{\theta}_i = (\alpha_i, \beta_i)'$ as

$$\boldsymbol{\theta}_i = \begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix} + \begin{pmatrix} \eta_{i\alpha} \\ \eta_{i\beta} \end{pmatrix} = \boldsymbol{\theta}_0 + \boldsymbol{\eta}_i, \tag{8.4}$$

where $\boldsymbol{\eta}_i = \boldsymbol{\psi}\lambda_i + \boldsymbol{\epsilon}_i$, with $\boldsymbol{\psi} = (\psi_\alpha, \psi_\beta)'$ and $\boldsymbol{\epsilon}_i = (\epsilon_{i\alpha}, \epsilon_{i\beta})'$. To generate correlated effects we set λ_i to be a function of the innovations to the x_{it} process:

$$\lambda_i = \frac{\boldsymbol{e}_{ix}'\boldsymbol{M}_T\boldsymbol{e}_{ix} - E\left(\boldsymbol{e}_{ix}'\boldsymbol{M}_T\boldsymbol{e}_{ix}\right)}{\sqrt{Var\left(\boldsymbol{e}_{ix}'\boldsymbol{M}_T\boldsymbol{e}_{ix}\right)}}. \tag{8.5}$$

Since $\boldsymbol{e}_{ix} = (e_{x,i1}, e_{x,i2}, ..., e_{x,iT})' \sim IID(\mathbf{0}, \boldsymbol{I}_T)$, it follows that λ_i is $IID(0,1)$.[15] The random components of $\boldsymbol{\eta}_i$, namely $\boldsymbol{\epsilon}_i$, are generated independently of $\boldsymbol{W}_i = (\boldsymbol{\tau}_T, \boldsymbol{x}_i)$, as $\boldsymbol{\epsilon}_i \sim IIDN\left(\mathbf{0}, \boldsymbol{V}_\epsilon\right)$, where $\boldsymbol{V}_\epsilon = Diag(\sigma^2_{\epsilon\alpha}, \sigma^2_{\epsilon\beta})$. Namely,

$$E(\boldsymbol{\eta}_i) = \mathbf{0}, \text{ and } \boldsymbol{V}_{\boldsymbol{\eta}} = E(\boldsymbol{\eta}_i\boldsymbol{\eta}_i') = \begin{pmatrix} \sigma^2_\alpha & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & \sigma^2_\beta \end{pmatrix} = \boldsymbol{\psi}\boldsymbol{\psi}' + \boldsymbol{V}_\epsilon.$$

The degree of correlated heterogeneity is determined by $\boldsymbol{\psi}\boldsymbol{\psi}'$, and it is zero if $\boldsymbol{\psi} = \mathbf{0}$. Also $Cov(\alpha_i, \beta_i) = \sigma_{\alpha\beta}$ will be non-zero when both ψ_α and ψ_β are non-zero. Specifically $\sigma^2_\alpha = \psi^2_\alpha + \sigma^2_{\epsilon\alpha}$, $\sigma_{\alpha\beta} = \psi_\alpha\psi_\beta$, and $\sigma^2_\beta = \psi^2_\beta + \sigma^2_{\epsilon\beta}$. Therefore, the correlation coefficients of $\boldsymbol{\theta}_i$ and λ_i are given by $\rho_{\alpha\lambda} = Corr(\alpha_i, \lambda_i) = \psi_\alpha / \sqrt{\psi^2_\alpha + \sigma^2_{\epsilon\alpha}}$, and

$$\rho_\beta = \rho_{\beta\lambda} = Corr(\beta_i, \lambda_i) = \frac{\psi_\beta}{\sqrt{\psi^2_\beta + \sigma^2_{\epsilon\beta}}}. \tag{8.6}$$

Solving the above equations for ψ_α and ψ_β, we have

$$\psi_\alpha = \left(\frac{\rho^2_{\alpha\lambda}}{1 - \rho^2_{\alpha\lambda}}\right)^{1/2} \sigma_{\epsilon\alpha}, \text{ and } \psi_\beta = \left(\frac{\rho^2_{\beta\lambda}}{1 - \rho^2_{\beta\lambda}}\right)^{1/2} \sigma_{\epsilon\beta}. \tag{8.7}$$

Also recall that $\sigma^2_\alpha = \psi^2_\alpha + \sigma^2_{\epsilon\alpha}$, and $\sigma^2_\beta = \psi^2_\beta + \sigma^2_{\epsilon\beta}$, then $\sigma^2_{\epsilon\alpha} = (1 - \rho^2_{\alpha\lambda})\sigma^2_\alpha$, and $\sigma^2_{\epsilon\beta} = (1 - \rho^2_{\beta\lambda})\sigma^2_\beta$. Hence, the key drivers of heterogeneity are σ^2_α, σ^2_β, $\rho^2_{\alpha\lambda}$, and $\rho^2_{\beta\lambda}$. The scaling parameter κ in (8.1) is set to achieve a given level of overall fit, PR^2, given by (S.3.3) in Section S.3.2 of the online supplement.

Section 8.1.3 summarizes parameters and details of the baseline model and the other ex-

[15] In Section S.3.1 of the online supplement, we show that when x_{it} is serially independent with no interactive effects ($\rho_{ix} = 0$ and $\gamma_{ix} = 0$), then λ_i can be written as a standardized version of $d_i = \det(\boldsymbol{W}_i'\boldsymbol{W}_i)$.

periments where we allow for heterogeneity in the autoregressive processes assumed for $\{u_{it}\}$ and $\{x_{it}\}$. We also consider cases where $\{x_{it}\}$ is generated with and without an interactive factor.[16] All estimations are based on R simulated observations $\left(y_{it}^{(r)}, x_{it}^{(r)}\right)$, for $r = 1, 2, ..., R$; $i = 1, 2, ..., n$; and $t = 1, 2, ..., T$.

8.1.3 Baseline and other experiments

For all experiments we set $\alpha_0 = \beta_0 = 1$, $\sigma_\alpha^2 = 0.2$, and $\sigma_\beta^2 = 0.5$, with $Corr(\alpha_i, \beta_i) = 0.25$, and experiment with two levels of fit: $PR^2 = 0.2$ and 0.4. We also consider two options when generating $e_{x,it}$, the shocks to the x_{it} process, namely Gaussian, $e_{x,it} \sim IIDN(0,1)$, and uniformly distributed errors, $e_{x,it} = \sqrt{12}(\mathfrak{z}_{it} - 1/2)$, with $\mathfrak{z}_{it} \sim IIDU(0,1)$.

For the baseline experiments, we set $PR^2 = 0.2$, generate the errors in the outcome equation as chi-squared without serial correlation ($\rho_{ie} = 0$ in (8.2)). For x_{it}, we allow for heterogeneous serial correlation, with $\rho_{ix} \sim IIDU(0, 0.95)$, but did not include the interactive effects in x_{it} (setting $\gamma_{ix} = 0$ in (8.3)). We consider both uncorrelated and correlated heterogeneity and set ρ_β, defined by (8.6), to (a) zero correlation, $\rho_\beta = 0$, (b) a medium level of correlation, $\rho_\beta = 0.25$, and (c) a high level of correlation, $\rho_\beta = 0.5$. For each choice of ρ_β, the scalar variable, κ, in the outcome equation, (8.1), is set such that $PR_T^2 = 0.2$, on average. This is achieved by stochastic simulation for each T, as described in Section S.3.2 of the online supplement.

To check the robustness of the TMG estimator, the following variations in the DGP of the errors and regressors are considered. When the errors in y_{it} are serially correlated, we generate $\rho_{ie} \sim IIDU(0, 0.95)$, and $e_{i0} \sim IIDN(0,1)$ for all i. When there is an interactive effect in $\{x_{it}\}$, $\gamma_{ix} \sim IIDU(0,2)$ and $f_t = 0.9 f_{t-1} + (1 - 0.9^2)^{1/2} v_t$, for $t = -49, -48, ..., -1, 0, 1, ..., T$, where $v_t \sim IIDN(0,1)$, with $f_{-50} = 0$.

To examine the relative efficiency of TMG and FE estimators, we set $\rho_\beta = 0$ (uncorrelated heterogeneity) but allow for error heteroskedasticity to be correlated with the processes generating x_{it}. We consider the following two scenarios: (a) cross-sectional heteroskedasticity, $\sigma_{it}^2 = \lambda_i^2$, for all i and t, where λ_i is given by (8.5); and (b) cross-sectional and time series heteroskedasticity, $\sigma_{it}^2 = e_{x,it}^2$, for all i and t, where $e_{x,it}$ is the innovation to the x_{it} process. In both cases we have $E(\sigma_{it}^2) = 1$, which match the case of randomly generated heteroskedasticity.

To investigate the small sample properties of the Hausman-type test, we allow individual effects, α_i, to be correlated with x_{it}, irrespective of whether x_{it} and β_i are correlated. Recall that FE and MG estimators are both robust to the correlation of x_{it} and α_i. The focus of

[16] When there are feedbacks, we generate x_{it} or e_{it} for $t = -49, -48, ..., -1, 0, 1, ..., T$, then drop the first 50 observations.

the Hausman-type test is on the degree of heterogeneity of β_i and the nature of correlation between β_i and x_{it}. We carry out $R = 2,000$ replications for all experiments.

8.2 Monte Carlo findings

8.2.1 Comparison of TMG, FE, and MG estimators

We first compare the performance of the TMG estimator with FE and MG estimators under both uncorrelated and correlated heterogeneity for the sample size combinations $n = 1,000, 2,000, 5,000, 10,000$ and $T = 2, 3, 4, 5, 6, 8$. The TMG estimator depends on the indicator, $\mathbf{1}\{d_i > a_n\}$, where $a_n = C_n n^{-\alpha}$. In view of the discussion in Section 5.1 on the choice of α, we consider the values of $\alpha = 1/3, 0.35$ and $1/2$, and set $C_n = \bar{d}_n = n^{-1}\sum_{i=1}^{n} d_i > 0$, where $d_i = \det(\boldsymbol{W}_i'\boldsymbol{W}_i)$. This choice of C_n ensures that the value of $\mathbf{1}\{d_i > a_n\} = \mathbf{1}\{d_i/\bar{d}_n > n^{-\alpha}\}$ is unaffected by the scale of x_{it}. In what follows we report the results for the TMG estimator with $\alpha = 1/3$, but discuss the sensitivity of the TMG estimator to the choice of α in sub-section 8.2.4.

Table 1 reports bias, root mean squared errors (RMSE) and size for estimation of β_0. The first column of the table gives estimates of the fraction of individual estimates being trimmed as defined by (4.7). The left panel gives the estimates under uncorrelated heterogeneity, with $\rho_\beta = 0$, and the right panel reports the results for the case of correlated heterogeneity, with $\rho_\beta = 0.5$. The estimates of π_n tend to be quite large for the case of ultra short T but fall quite rapidly as T is increased. For example, for $T = 2$ and $n = 1,000$ as many as 31.2 per cent of the individual estimates are trimmed when computing the TMG estimates, but it falls to 3.2 per cent when T is increased to $T = 8$. However, recall that the TMG estimator continues to make use of the trimmed estimates, as can be seen from (4.6), and the TMG estimator shows little bias compared to the untrimmed MG estimator. The TMG and MG estimators converge as T is increased and they are almost identical for the panels with $T = 8$. The results in Table 1 clearly show the effectiveness of trimming in dealing with outlying individual estimates.

Comparing TMG and FE estimators, we first note that in line with the theory, the FE estimator performs very well under uncorrelated heterogeneity but is badly biased when heterogeneity is correlated, and this bias does not diminish if n and T are increased. When heterogeneity is correlated, the FE estimator also exhibits substantial size distortions which tend to get accentuated as n is increased for a given T. In contrast, the TMG estimator is robust to the correlation between β_i and d_i, and delivers size around the 5 per cent nominal level in all cases.[17]

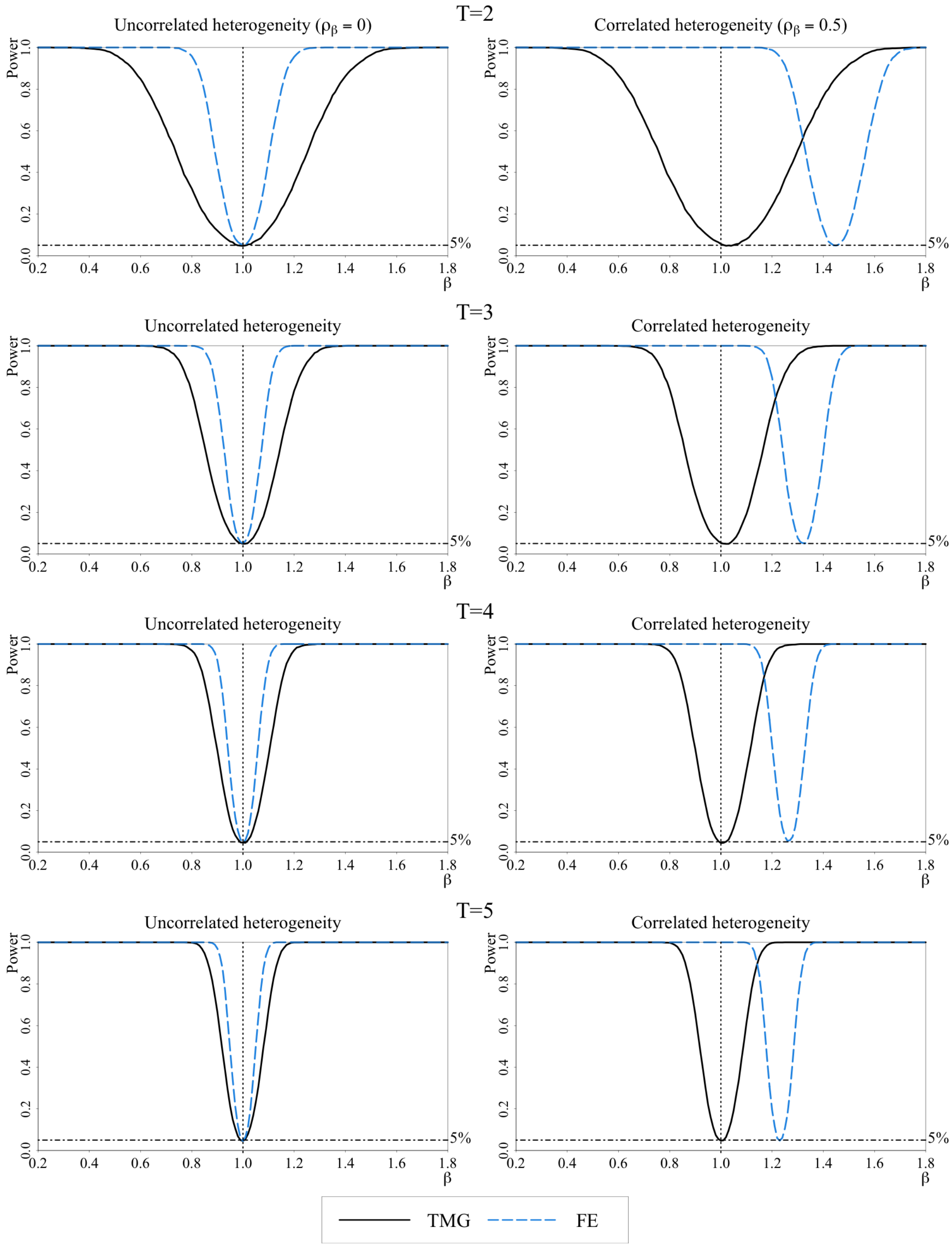
[17]Increasing PR^2 from 0.2 to 0.4 does not affect the bias and RMSE of the FE estimator, but results in a

Table 1: Bias, RMSE and size of FE, MG and TMG estimators of β_0 ($E(\beta_i) = \beta_0 = 1$) in the baseline model without time effects

	Uncorrelated heterogeneity: $\rho_\beta = 0$, $PR^2 = 0.2$										Correlated heterogeneity: $\rho_\beta = 0.5$, $PR^2 = 0.2$									
	$\hat{\pi}$ (×100)	Bias			RMSE			Size (×100)			$\hat{\pi}$ (×100)	Bias			RMSE			Size (×100)		
T	TMG	FE	MG	TMG	FE	MG	TMG	FE	MG	TMG	TMG	FE	MG	TMG	FE	MG	TMG	FE	MG	TMG
$n = 1,000$																				
2	31.2	-0.004	-7.325	-0.002	0.17	353.13	0.33	5.0	2.1	4.7	31.2	0.444	-7.714	0.048	0.48	371.86	0.35	66.6	2.1	4.9
3	16.5	0.001	-0.011	-0.002	0.11	0.57	0.19	4.9	4.0	4.7	16.5	0.322	-0.011	0.023	0.34	0.60	0.20	74.1	4.2	5.2
4	10.4	-0.002	-0.001	-0.002	0.09	0.21	0.14	4.8	5.1	5.5	10.4	0.265	0.000	0.013	0.28	0.22	0.15	76.7	5.3	5.2
5	7.1	-0.002	-0.001	0.000	0.08	0.13	0.11	5.6	4.0	3.6	7.1	0.230	0.000	0.009	0.25	0.14	0.12	77.5	4.2	4.0
6	5.2	0.002	0.000	0.001	0.07	0.11	0.10	4.5	4.5	4.7	5.2	0.211	0.000	0.008	0.22	0.11	0.10	80.0	4.8	5.1
8	3.2	0.002	0.000	0.000	0.06	0.08	0.08	4.9	4.6	4.6	3.2	0.179	0.000	0.003	0.19	0.08	0.08	80.0	4.6	4.5
$n = 2,000$																				
2	28.5	-0.005	-2.398	-0.001	0.12	149.72	0.26	4.6	2.0	5.5	28.5	0.445	-2.525	0.044	0.46	157.67	0.27	90.8	2.0	5.3
3	14.1	0.002	0.007	-0.002	0.08	0.44	0.15	4.9	4.3	5.6	14.1	0.323	0.007	0.018	0.34	0.46	0.16	95.2	4.5	5.4
4	8.4	0.000	-0.003	-0.003	0.07	0.15	0.11	4.9	4.3	4.4	8.4	0.266	-0.003	0.008	0.27	0.15	0.11	95.7	4.8	4.7
5	5.6	0.001	-0.004	-0.001	0.06	0.10	0.09	5.6	5.1	5.1	5.6	0.233	-0.004	0.006	0.24	0.11	0.09	96.3	5.2	5.1
6	4.0	0.000	-0.001	-0.001	0.05	0.08	0.07	3.8	4.2	4.3	4.0	0.208	-0.001	0.004	0.21	0.08	0.07	97.4	4.0	4.4
8	2.4	0.000	0.000	0.000	0.04	0.06	0.06	4.2	5.0	4.6	2.4	0.178	0.000	0.003	0.18	0.06	0.06	97.4	4.7	4.8
$n = 5,000$																				
2	24.7	0.002	-4.275	0.000	0.08	426.12	0.17	5.1	1.8	4.3	24.7	0.452	-4.502	0.037	0.46	448.73	0.18	100.0	1.8	4.7
3	10.8	0.001	-0.006	0.000	0.05	0.28	0.10	4.3	4.7	4.8	10.8	0.323	-0.006	0.016	0.33	0.29	0.11	100.0	4.7	5.3
4	5.8	0.001	0.000	0.000	0.04	0.10	0.07	5.1	4.9	5.7	5.8	0.265	0.000	0.008	0.27	0.10	0.08	100.0	5.3	5.6
5	3.5	0.000	-0.001	0.000	0.04	0.07	0.06	5.1	4.6	4.0	3.5	0.231	-0.001	0.004	0.23	0.07	0.06	100.0	4.3	4.0
6	2.3	-0.001	0.000	0.000	0.03	0.05	0.05	4.7	5.1	5.2	2.3	0.207	0.000	0.003	0.21	0.06	0.05	100.0	5.1	5.4
8	1.2	0.000	0.000	0.000	0.03	0.04	0.04	5.2	5.0	5.1	1.2	0.178	0.000	0.001	0.18	0.04	0.04	100.0	5.1	5.1
$n = 10,000$																				
2	22.1	-0.001	-1.536	-0.004	0.05	175.33	0.13	5.3	2.2	4.7	22.1	0.449	-1.617	0.029	0.45	184.63	0.14	100.0	2.2	5.6
3	8.8	-0.002	0.004	0.000	0.04	0.19	0.07	5.1	4.2	4.9	8.8	0.321	0.004	0.013	0.32	0.20	0.08	100.0	4.3	5.3
4	4.4	0.000	0.000	0.001	0.03	0.07	0.05	4.6	4.6	4.6	4.4	0.265	0.000	0.007	0.27	0.07	0.05	100.0	4.7	4.4
5	2.5	0.000	-0.001	-0.001	0.03	0.05	0.04	4.5	4.4	4.6	2.5	0.231	-0.001	0.002	0.23	0.05	0.04	100.0	4.7	4.7
6	1.6	0.000	0.000	0.000	0.02	0.04	0.03	5.6	4.2	4.0	1.6	0.208	0.000	0.002	0.21	0.04	0.04	100.0	4.4	4.1
8	0.8	0.000	0.000	0.001	0.02	0.03	0.03	5.1	4.9	4.8	0.8	0.177	0.000	0.001	0.18	0.03	0.03	100.0	4.8	4.7

Notes: (i) The baseline model is generated as $y_{it} = \alpha_i + \beta_i x_{it} + u_{it}$, where the errors processes for y_{it} and x_{it} equations are chi-squared and Gaussian, respectively, x_{it} are generated as heterogeneous AR(1) processes, and ρ_β (the degree of correlated heterogeneity) is defined by (8.6). For further details see Section 8.1.3. (ii) FE and MG estimators are given by (2.13) and (2.4). The trimmed mean group (TMG) estimator and its asymptotic variance are given by (4.5) and (5.21), respectively. (iii) The trimming threshold for the TMG estimator is given by $a_n = \bar{d}_n n^{-\alpha}$, where $\bar{d}_n = \frac{1}{n}\sum_i^n d_i$, $d_i = \det(\boldsymbol{W}_i'\boldsymbol{W}_i)$, $\boldsymbol{W}_i = (\boldsymbol{w}_{i1}, \boldsymbol{w}_{i2}, ..., \boldsymbol{w}_{iT})'$ and $\boldsymbol{w}_{it} = (1, x_{it})'$. α is set to 1/3. $\hat{\pi}$ is the simulated fraction of individual estimates being trimmed, defined by (4.7).

Figure 1: Empirical power functions for FE and TMG estimators of β_0 ($E(\beta_i) = \beta_0 = 1$) in the baseline model without time effects for $n = 10,000$ and $T = 2, 3, 4, 5$



Notes: For details of the baseline model without time effects, see footnote (i) to Table 1. For the FE estimator, see footnote (ii) to Table 1. For the TMG estimator, see footnotes (ii) and (iii) to Table 1.

higher degree of size distortion under correlated heterogeneity. Compare the results summarized in the right panel of Table 1 and Table S.4, respectively, in the online supplement.

Figure 1 shows the plots of the empirical power functions for TMG and FE estimators for $n = 10,000$ and $T = 2, 3, 4$ and 5. The left panel gives the power functions for the case of uncorrelated heterogeneity ($\rho_\beta = 0$), and as can be seen, both estimators are centered correctly around $\beta_0 = 1$, with the FE estimator having better power properties. But the differences between the power of FE and TMG estimators shrink rapidly and become negligible as T is increased from $T = 2$ to $T = 5$.[18] The right panel of the figure provides the same results but under correlated heterogeneity with $\rho_\beta = 0.5$. In this case, the empirical power functions of the FE estimator now shift dramatically to the right, away from the true value, an outcome that becomes more concentrated as T is increased. In contrast, the empirical power functions for the TMG estimator tend to be reasonably robust to the choice of ρ_β.

To summarize, in the case of uncorrelated heterogeneity, the FE estimator performs well despite of the heterogeneity and is more efficient than the TMG estimator in the case of baseline model used in our MCs, but in general the relative efficiency of TMG and FE estimators depends on the underlying DGP. The situation is markedly different when heterogeneity is correlated, and the FE estimator can be badly biased, leading to incorrect inference, whilst the TMG estimator provides valid inference with size around the nominal five per cent level and reasonable power, irrespective of whether β_i is correlated with x_{it} or not.

8.2.2 Comparison of TMG, GP, and SU estimators

Focusing on the case of correlated heterogeneity, we now compare the performance of the TMG estimator with GP and SU estimators. To implement the GP estimator, defined by (3.1), for $T = 2$ we follow GP and set $h_n = C_{GP} n^{-\alpha_{GP}}$, with $\alpha_{GP} = 1/3$, and $C_{GP} = \frac{1}{2} \min(\hat{\sigma}_D, \hat{r}_D/1.34)$, where $\hat{\sigma}_D$ and $\hat{r}_D$ are the respective sample standard deviation and interquartile range of $\det(\boldsymbol{W}_i)$. See page 2138 in Graham and Powell (2012).[19] There is no clear guidance in GP as to the choice of h_n when $T = 3$.[20] For consistency, for GP estimates we continue to use their bandwidth, $h_n = C_{GP} n^{-\alpha_{GP}}$ with $\alpha_{GP} = 1/3$, but set $C_{GP} = (\bar{d}_n)^{1/2}$ and trim if $d_i = \det(\mathbf{W}_i'\mathbf{W}_i) < h_n^2$. The sensitivity of the results to the other choices of α_{GP} is considered below. For SU we use the code made available to us by the authors, which is applicable only when $T = 2$.

The bias, RMSE and size for all three estimators are summarized in Table 2 for $T = 2$,

[18] But see the left panel of Table S.3 and Figure S.1 in the online supplement where it is shown that it does not necessarily follow that the FE estimator will dominate the TMG estimator in terms of efficiency even when T is ultra short.

[19] However, GP seem to be using the larger scaling value of $C_{GP} = \min(\hat{\sigma}_D, \hat{r}_D/1.34)$, when they generate the histograms in Figure 1, on page 2137.

[20] For $T = 3$, GP do not use the bandwidth parameter, h_n, but directly select the "percent trimmed", π_n. In their empirical application for $T = 3$ they report estimates with 4 per cent being trimmed. See the last column of Table 3 on page 2136 of GP.

and $n = 1,000, 2,000, 5,000, 10,000$. For $T = 3$ the results are provided for TMG and GP estimators only. The associated empirical power functions are provided in Figure 2.

Table 2: Bias, RMSE and size of TMG, GP and SU estimators of β_0 ($E(\beta_i) = \beta_0 = 1$) in the baseline model without time effects and with correlated heterogeneity, $\rho_\beta = 0.5$

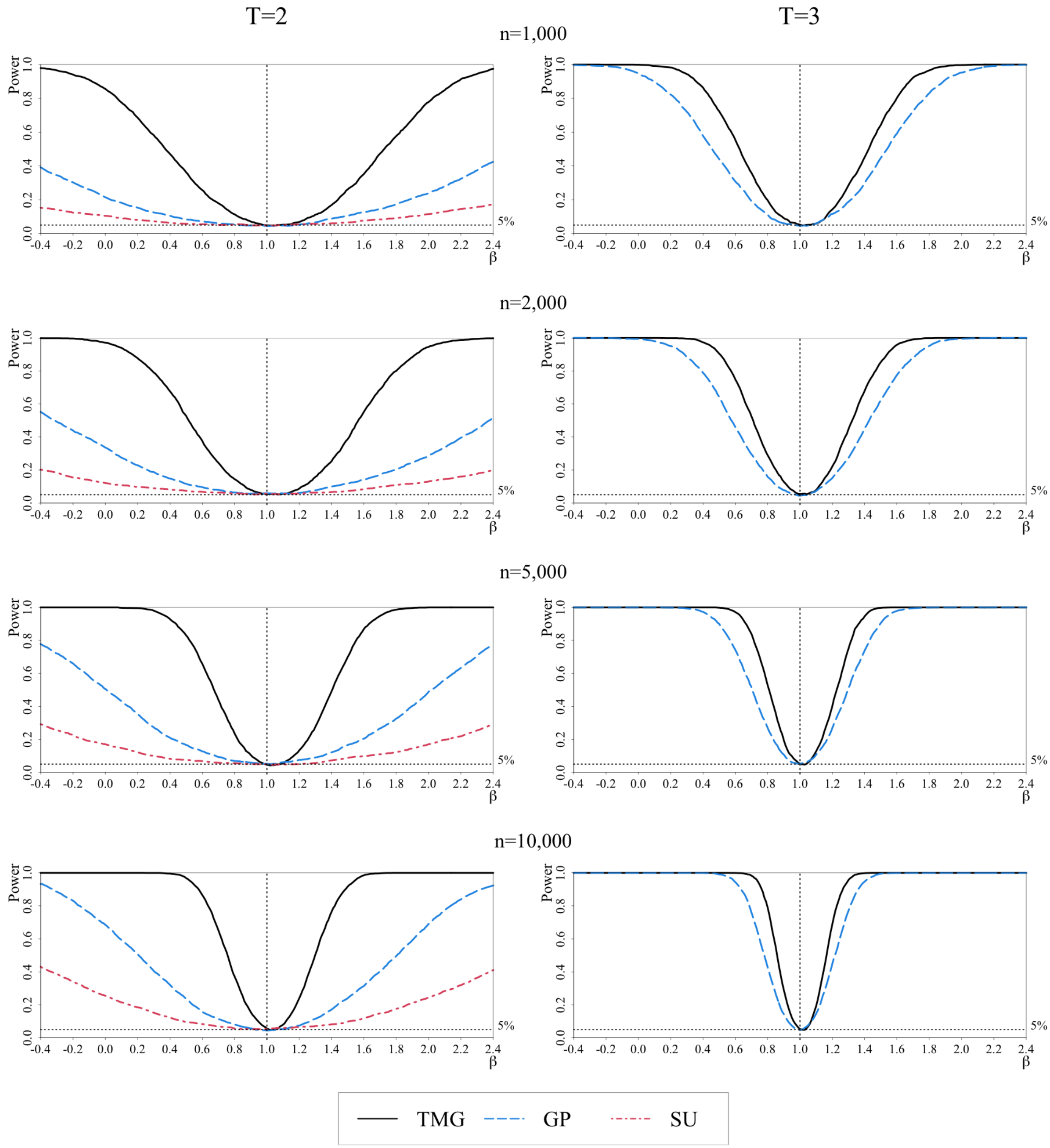
	$T=2$				$T=3$			
Estimator	$\hat{\pi}$ ($\times 100$)	Bias	RMSE	Size ($\times 100$)	$\hat{\pi}$ ($\times 100$)	Bias	RMSE	Size ($\times 100$)
				$n = 1,000$				
TMG	31.2	0.048	0.35	4.9	16.5	0.023	0.20	5.2
GP	4.2	-0.029	0.83	4.5	2.0	-0.002	0.27	4.6
SU	4.2	-0.045	1.62	4.9	...	...	...	...
				$n = 2,000$				
TMG	28.5	0.044	0.27	5.3	14.1	0.018	0.16	5.4
GP	3.4	0.031	0.70	5.8	1.3	0.003	0.22	4.4
SU	3.4	0.008	1.39	5.5	...	...	...	...
				$n = 5,000$				
TMG	24.7	0.037	0.18	4.7	10.8	0.016	0.11	5.3
GP	2.5	0.009	0.52	5.2	0.7	-0.001	0.15	5.2
SU	2.5	0.003	1.01	4.9	...	...	...	...
				$n = 10,000$				
TMG	22.1	0.029	0.14	5.6	8.8	0.013	0.08	5.3
GP	2.0	0.002	0.41	4.3	0.5	-0.002	0.11	4.9
SU	2.0	0.007	0.82	5.2	...	...	...	...

Notes: (i) GP and SU estimators are proposed by Graham and Powell (2012) and Sasaki and Ura (2021), respectively. The GP estimator is given by (3.1). For $T = 2$, GP compare $d_i^{1/2}$ with the bandwidth $h_n = C_{GP} n^{-\alpha_{GP}}$. α_{GP} is set to 1/3. $C_{GP} = \frac{1}{2} \min(\hat{\sigma}_D, \hat{r}_D/1.34)$, where $\hat{\sigma}_D$ and $\hat{r}_D$ are the respective sample standard deviation and interquartile range of $\det(\boldsymbol{W}_i)$. For $T = 3$, we continue using the bandwidth h_n with $C_{GP} = (\bar{d}_n)^{1/2}$. See Section 8.2.2 for details. When $T = 2$, the SU estimator uses the same bandwidth as GP. (ii) For details of the baseline model without time effects, see footnote (i) to Table 1. For the TMG estimator and its trimming threshold, see footnotes (ii) and (iii) to Table 1. $\hat{\pi}$ is the simulated fraction of individual estimates being trimmed, defined by (4.7). The estimation algorithm for the SU estimator is not available for $T = 3$, denoted by "...".

The fraction of the trimmed estimates, π_n, defined by (4.7), for the TMG estimator is quite high when $T = 2$, but declines markedly when T is raised to 3, and to a lesser extent as n is increased. This is not the case for the other two estimators. For example, when $T = 2$ and $n = 1,000$, the fraction of trimmed estimates for the TMG estimator is around 31.2 per cent as compared to 4.2 per cent for GP and SU estimators, and falls to 22.1 per cent as n is increased to $10,000$. Increasing T from 2 to 3 with $n = 1,000$ reduces this fraction to 16.5 per cent as compared to 2 per cent for the GP estimator.[21] The heavy trimming causes the TMG estimator to have a much larger bias than GP and SU estimators, particularly when $T = 2$ and n is sufficiently large. However, the TMG estimator continues to have better

[21]Recall that the codes released by SU do not allow us to compute their estimator when $T = 3$.

Figure 2: Empirical power functions for TMG, GP and SU estimators of β_0 ($E(\beta_i) = \beta_0 = 1$) in the baseline model without time effects and with correlated heterogeneity, $\rho_\beta = 0.5$, for $n = 1,000$, $2,000$, $5,000$, $10,000$ and $T = 2, 3$



Notes: For details of the baseline model without time effects, see footnote (i) to Table 1. For the TMG estimator, see footnote (ii) to Table 1. For GP and SU estimators, see footnote (i) to Table 2.

overall small sample performance due to its higher efficiency. Recall that the TMG estimator makes use of the trimmed estimates, as set out in the second term of (4.6), but the trimmed estimates are not used in the GP estimator. This difference in the way trimmed estimates are treated is reflected in the lower RMSE of the TMG estimator as compared to the other two estimators for all T and n combinations. For example, when $T = 2$ and $n = 1,000$, the RMSE of the TMG is 0.35 as compared to 0.83 and 1.62 for GP and SU estimators. The relative advantage of the TMG estimator continues when T is increased from 2 to 3, but its relative advantage declines. For $T = 3$, the RMSE of the TMG estimator stands at 0.20 compared to 0.27 for the GP estimator. The larger the value of T, the less important the trimming becomes.

The empirical power functions for all three estimators are shown in Figure 2. As can be seen, the TMG estimator is uniformly more powerful than the GP estimator and the GP estimator is more powerful than the SU estimator.

8.2.3 Models with time effects

Adding time effects to the panel regressions does not alter the above conclusions. The MC results for estimation of β_0 and the time effects $\boldsymbol{\phi} = (\phi_1, \phi_2)'$ are summarized in Tables 3 and 4, respectively. The small sample properties of the TMG-TE estimator of β_0 are very close to those reported for the TMG estimator in Table 2. Interestingly, there are also little differences between TMG-TE and TMG-C estimators of β_0 when $T = 3$, as can be seen from the right panel of Table 3. Also, the time effects are precisely estimated. Bias, RMSE and size for TMG-TE and GP estimators of $\boldsymbol{\phi} = (\phi_1, \phi_2)'$ are summarized in Table 4. The bias of TMG-TE and GP estimators of ϕ_1 are similar, but the TMG-TE estimator has much lower RMSEs and higher power when $T = 2$. A comparison of the empirical powers of these two estimators is given in Figures S.9–S.11 in the online supplement.

8.2.4 Sensitivity of TMG and GP estimators to the choice of the threshold values

Finally, we consider the sensitivity of TMG and GP estimators to the choice of threshold values. The baseline value of the threshold value for the GP estimator, $\alpha_{GP} = 1/3$ as recommended by GP.[22] But for the purpose of comparison with the TMG estimator computed for $\alpha = 1/3, 0.35$ and $1/2$, we also consider $\alpha_{GP} = 0.35/2$ and $1/4$. Recall that the bandwidth, h_n^2, used by GP corresponds to a_n used in the specification of TMG. Hence, $2\alpha_{GP}$ corresponds to α. For comparability, we decided to consider values of α_{GP} below $1/3$ required by GP's

[22] See equation (3.1) and the related discussion for the implementation of the GP estimator in sub-section 8.2.2.

Table 3: Bias, RMSE and size of TMG-TE, TMG-C, GP and SU estimators of β_0 ($E(\beta_i) = \beta_0 = 1$) in the baseline model with time effects and correlated heterogeneity, $\rho_\beta = 0.5$

	$T=2$				$T=3$			
Estimator	$\hat{\pi}$ (×100)	Bias	RMSE	Size (×100)	$\hat{\pi}$ (×100)	Bias	RMSE	Size (×100)
				$n=1,000$				
TMG-TE	31.2	0.048	0.35	5.0	16.5	0.023	0.20	5.4
TMG-C	...	...	...	...	16.5	0.023	0.20	5.2
GP	4.2	-0.034	0.84	3.9	2.0	-0.002	0.27	4.6
SU	4.2	-0.052	1.67	5.3	...	...	...	...
				$n=2,000$				
TMG-TE	28.5	0.044	0.27	5.6	14.1	0.018	0.16	5.5
TMG-C	...	...	...	...	14.1	0.018	0.16	5.6
GP	3.4	0.032	0.71	5.2	1.3	0.003	0.22	4.6
SU	3.4	0.012	1.40	5.8	...	...	...	...
				$n=5,000$				
TMG-TE	24.7	0.037	0.18	4.7	10.8	0.016	0.11	5.3
TMG-C	...	...	...	...	10.8	0.016	0.11	5.3
GP	2.5	0.008	0.53	5.0	0.7	-0.001	0.15	5.1
SU	2.5	0.006	1.02	4.7	...	...	...	...
				$n=10,000$				
TMG-TE	22.1	0.028	0.14	5.7	8.8	0.013	0.08	5.3
TMG-C	...	...	...	...	8.8	0.013	0.08	5.3
GP	2.0	0.003	0.41	4.4	0.5	-0.002	0.11	5.0
SU	2.0	0.011	0.82	5.5	...	...	...	...

Notes: (i) The baseline model is generated as $y_{it} = \alpha_i + \phi_t + \beta_i x_{it} + u_{it}$, with time effects given by $\phi_t = t$ for $t = 1, 2, ..., T-1$, and $\phi_T = -T(T-1)/2$. The errors processes for y_{it} and x_{it} equations are chi-squared and Gaussian, respectively, x_{it} are generated as heterogeneous AR(1) processes, and ρ_β (the degree of correlated heterogeneity) is defined by (8.6). For further details see Section 8.1.3. (ii) The TMG-TE estimators of $\boldsymbol{\theta}_0$ and $\boldsymbol{\phi}$ are given by (6.9) and (6.11), respectively, and their asymptotic variances are estimated by (A.3.4) and (A.3.7), respectively, in the Appendix. The TMG-C estimators of $\boldsymbol{\theta}_0$ and $\boldsymbol{\phi}$ are given by (6.20) and (6.17), respectively, and their asymptotic variances are estimated by (6.21) and (6.19), respectively. For the trimming threshold, see footnote (iii) to Table 1. (iii) For GP and SU estimators, see footnote (i) to Table 2. $\hat{\pi}$ is the simulated fraction of individual estimates being trimmed, defined by (4.7). "..." denotes the estimation algorithms are not available or not applicable.

Table 4: Bias, RMSE and size of TMG-TE and GP estimators of the time effects, ϕ_1 and ϕ_2, in the baseline model with correlated heterogeneity, $\rho_\beta = 0.5$

		$n=1,000$			$n=5,000$		
	Estimator	Bias	RMSE	Size (×100)	Bias	RMSE	Size (×100)
			$T=2$			$T=2$	
$\phi_1 = 1$	TMG-TE	0.002	0.09	6.1	-0.001	0.04	4.8
	GP	0.001	0.54	7.1	-0.008	0.35	6.9
			$T=3$			$T=3$	
$\phi_1 = 1$	TMG-TE	-0.002	0.10	5.6	0.001	0.05	4.9
	GP	0.004	0.15	5.7	0.001	0.07	5.0
$\phi_2 = 2$	TMG-TE	-0.006	0.10	5.7	0.000	0.05	4.9
	GP	-0.010	0.13	5.3	0.000	0.06	4.2

Notes: For the baseline model with time effects, see footnote (i) to Table 3. For the TMG-TE estimator, see footnote (ii) to Table 3. For the GP estimator, see footnote (i) to Table 2.

theory. This allows us to compare GP and TMG focusing on the utility of including both trimmed and untrimmed estimates of $\boldsymbol{\theta}_i$ in estimation of average treatment effects.

The results are summarized in Section S.4.2 of the online supplement. As can be seen from Table S.5 there is a clear trade-off between bias and variance as α and α_{GP} are increased. For $T = 2$, the TMG estimator is biased when $\alpha = 1/3$ (as predicted by the theory), but has a lower variance with its RMSE declining as α is increased from $1/3$ to $1/2$. This trade-off is less consequential when T is increased to $T = 3$. The same is also true for the GP estimator. But for all choices of α and α_{GP} the TMG performs better in terms of RMSE when $T = 2$. For $T = 3$, TMG and GP estimators share the same trimming threshold when $\alpha = 2\alpha_{GP}$, resulting in identical trimmed fractions for $\alpha = 2\alpha_{GP} \in \{0.35, 1/2\}$. While RMSEs are similar, the GP estimator exhibits significantly higher bias than the TMG estimator as observations of the trimmed units are not exploited by the GP estimator.

Figure S.2 compares power functions of TMG and GP estimators with $\alpha = 2\alpha_{GP} = 0.35$. For $T = 2$, a higher trimmed fraction results in a steeper power function for the TMG estimator as compared to that of the GP estimator. When $T = 3$, with the same trimmed fraction, the power function of the GP estimator shifts to the right, away from the true value. The substantial differences in power performance of the TMG estimator with $\alpha = 1/3$ and the GP estimator with $\alpha_{GP} = 1/3$ are also illustrated in Figure S.3.

The power comparisons of TMG and GP estimators for different values of α and α_{GP} are given in Figures S.4 and S.5, respectively, and convey the same message, suggesting that for the TMG estimator the boundary choice of $\alpha = 1/3$ tends to produce the best bias-variance trade-off. Increasing α reduces the bias but increases the variance, and the boundary value derived theoretically seems to strike a sensible balance and is recommended.

8.2.5 MC evidence on the Hausman-type test of correlated heterogeneity

Table 5 reports empirical size and power of the Hausman-type test of correlated heterogeneity given by (7.3). The left, middle and right panels report the results under homogeneity, uncorrelated heterogeneity and correlated heterogeneity in slope coefficients. In the left and middle panels, the size of the test is around the nominal level of 5 per cent. As shown in the paper, when $\boldsymbol{x}_{it}$ is strictly exogenous and $\boldsymbol{\beta}_i$ is mean independent of $\boldsymbol{X}_i$, FE, MG and TMG estimators are all consistent under homogeneity and uncorrelated heterogeneity, and in this case the Hausman-type test does not have power against uncorrelated heterogeneity. However, in the case where slope coefficients are heterogeneous *and* correlated with the regressors, the MG and TMG estimators are consistent when they have at least finite second moments, while the FE estimator is biased for all T. In this case, we would expect the proposed test to have power, and this is indeed evident in the right panel of Table 5. Also

Table 5: Empirical size and power of the Hausman-type test of correlated heterogeneity in the baseline model without time effects

	Under H_0								Under H_1			
	Homogeneity: $\sigma_\beta^2 = 0$				Uncorrelated hetro.: $\rho_\beta = 0$, $\sigma_\beta^2 = 0.5$				Correlated hetro.: $\rho_\beta = 0.5$, $\sigma_\beta^2 = 0.5$			
T/n	1,000	2,000	5,000	10,000	1,000	2,000	5,000	10,000	1,000	2,000	5,000	10,000
2	4.1	4.4	5.3	5.5	5.6	5.5	4.8	5.9	26.0	39.0	69.8	90.8
3	4.5	5.7	5.8	4.8	5.3	4.5	5.7	4.7	41.0	61.5	91.3	99.6
4	4.9	5.7	5.6	4.9	4.8	4.5	5.5	4.7	53.5	77.2	97.8	100.0
5	5.3	4.6	4.9	5.2	4.3	4.6	4.4	4.3	63.5	86.1	99.5	100.0
6	4.8	5.2	5.5	5.3	5.2	4.4	4.6	4.4	72.0	91.7	100.0	100.0
8	5.0	4.8	5.0	5.3	4.6	4.9	6.0	5.7	81.1	96.2	100.0	100.0

Notes: (i) The baseline model for the Hausman-type test is generated as $y_{it} = \alpha_i + \beta_i x_{it} + u_{it}$, with α_i correlated with x_{it} under both the null and alternative hypotheses. The errors processes for y_{it} and x_{it} equations are chi-squared and Gaussian, respectively, and x_{it} are generated as heterogeneous AR(1) processes. For further details see Section 8.1.3. (ii) The null hypothesis is given by (7.1), including the case of homoegenity with $\sigma_\beta^2 = 0$ and the case of uncorrelated heterogeneity with $\rho_\beta = 0$ (the degree of correlated heterogeneity defined by (8.6) and $\sigma_\beta^2 = 0.5$. The alternative of correlated heterogeneity is generated with $\rho_\beta = 0.5$ and $\sigma_\beta^2 = 0.5$. (iii) The test statistic is calculated based on the difference between FE and TMG estimators, given by (7.3). Under H_0, the test statistic is asymptotically distributed as χ^2_{k-1} as $n \to \infty$. For the FE estimator, see footnote (ii) to Table 1. For the TMG estimator, see footnotes (ii) and (iii) to Table 1. Size and power are in per cent.

the power of the test rises with increases in n even when $T = 2$, illustrating the (ultra) small T consistency of the proposed test.

The MC evidence on the performance of our proposed test of correlated heterogeneity in the case of panels with time effects is given in Table S.15 of the online supplement. We consider two versions of the test, depending on how time effects are filtered out, namely TMG-TE and TMG-C estimators (see equations (S.2.15) and (S.2.21)).[23] The empirical size and power of these two test statistics are comparable for $T > 2$. More importantly, allowing for time effects has negligible effects on the small sample performance of the test, while the power of the test is slightly lower than the power reported in Table 5 for models without time effects. Increases in n and/or T result in a rapid rise in power, illustrating the consistent property of the proposed test.

9 Empirical illustration

In this section, we re-visit the empirical application in Graham and Powell (2012) who provide estimates of the average effect of household expenditures on calorie demand, based on a sample of households from poor rural communities in Nicaragua that participated in a conditional cash transfer program *Red de Proteccion Social* (RPS). The data set is a balanced panel with $n = 1,358$ households observed from 2000 to 2002. We present estimates of the average treatment effects using the following panel data model with time effects:

$$\ln(Cal_{it}) = \alpha_i + \phi_t + \beta_i \ln(Exp_{it}) + u_{it}, \tag{9.1}$$

where $\ln(Cal_{it})$ denotes the logarithm of household calorie availability per capita in year t of household i, and $\ln(Exp_{it})$ denotes the logarithm of real household expenditures per capita (in thousands of 2001 cordobas) of household i in year t. The parameter of interest is the average treatment effect defined by $\beta_0 = E(\beta_i)$, allowing for possible dependence between β_i and $\ln(Exp_{it})$. Correlated heterogeneity could arise for a number of reasons, such as model misspecification, individuals responding strategically to treatments, and common factors that simultaneously affect β_i and the treatment, $\ln(Exp_{it})$. It is, therefore, prudent to first test for correlated heterogeneity before estimating β_0 by fixed effects, which is the standard approach when T is ultra short. We provide test statistics and estimates of β_0 for the panels of 2001–2002 ($T = 2$) and 2000–2002 ($T = 3$) with and without time effects.

Table 6 reports results of the Hausman-type test of correlated heterogeneity in the effects of household expenditures on calorie demand. The null hypothesis is rejected for both panels

[23] For further details, see Section S.2 in the online supplement.

covering the periods 2001–2002 and 2000–2002, and irrespective of whether time effects are allowed. As shown in the Monte Carlo experiments, the test only has power against the alternative of correlated heterogeneity. Therefore, these results provide strong evidence of heterogeneity in the treatment effects that are correlated with the level of household expenditures, which in turn sheds doubt on the validity of the FE estimation of the average treatment effect for this application.

Table 6: Hausman-type statistics for testing correlated heterogeneity in the effects of household expenditures on calorie demand in Nicaragua

	Without time effects		With time effects		
	2001–2002	2000–2002	2001–2002	2000–2002	
			TMG-TE	TMG-TE	TMG-C
Statistics	5.918	7.626	5.959	6.772	7.653
p-value	0.015	0.006	0.015	0.009	0.006
T	2	3	2	3	3

Notes: The test is applied to the average effect $\beta_0 = E(\beta_i)$ in the model (9.1) based on the RPS panel of $1,358$ households. The test statistic for panels without time effects is described in footnote (iii) to Table 5. The test statistics for panels with time effects are based on the difference between the FE-TE and TMG-TE estimators given by (S.2.15) with $T \geq 2$, and the difference between the FE-TE and TMG-C estimators given by (S.2.21) with $T > 2$. For further details see Section S.2 in the online supplement.

Table 7: Alternative estimates of the average effect of household expenditures on calorie demand in Nicaragua over the period 2001–2002 ($T = 2$)

	Without time effects				With time effects			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	FE	GP	SU	TMG	FE-TE	GP	SU	TMG-TE
$\hat{\beta}_0$	0.6568	0.4549	0.6974	0.5623	0.6554	0.4629	0.6952	0.5612
	(0.0287)	(0.1003)	(0.1689)	(0.0425)	(0.0284)	(0.1025)	(0.1650)	(0.0424)
$\hat{\phi}_{2002}$	...	...	...	...	0.0172	-0.0181	...	0.0178
	...	...	...	...	(0.0063)	(0.0296)	...	(0.0064)
$\hat{\pi}$ ($\times 100$)	...	3.8	3.8	27.1	...	3.8	3.8	27.1

Notes: The estimates of $\beta_0 = E(\beta_i)$ and ϕ_{2002} in the model (9.1) are based on the RPS panel of $1,358$ households. The FE estimator is described in the footnote (ii) to Table 1. GP and SU estimators are described in footnote (i) to Table 2. The TMG estimator is described in footnotes (ii) and (iii) to Table 1. The FE-TE estimator is the two-way fixed effects estimator given by (S.2.1) in the online supplement. TMG-TE and TMG-C estimators are described in footnote (ii) to Table 3. $\hat{\pi}$ is the estimated fraction of individual estimates being trimmed, defined by (4.7). The numbers in brackets are standard errors. "..." denotes the estimation algorithms are not available or not applicable.

Table 7 presents the estimates of β_0 based on the panel of 2001–2002 (with $T = 2$) without time effects (left panel), and with time effects (right panel). The estimates are not affected by the inclusion of time effects but differ considerably across different methods.[24] Based on the test results reported in Table 6, the FE estimates are most likely biased. Turning to the trimmed estimators, we find that only the TMG estimator is heavily trimmed with 27.1 per cent of the estimates being trimmed, whilst the rate of trimming is only around 3.8 per cent for GP and SU estimators.[25] Focussing on the estimates without time effects, we find the FE estimate, 0.6568 (0.0287), is much larger and more precisely estimated than either the GP or TMG estimates, given by 0.4549 (0.1003) and 0.5623 (0.0425), respectively.[26] Judging by the standard errors, it is also noticeable that the TMG is more precisely estimated than the GP estimate and lies somewhere between the FE and GP estimates. In contrast, the SU estimate of 0.6974 (0.1689) is close to the FE estimate but with a much larger degree of uncertainty. These estimates are in line with the MC results reported in the previous section, where we found that in the presence of correlated heterogeneity FE estimates are biased with smaller standard errors (thus leading to incorrect inference), whilst GP and TMG estimators are correctly centered with the TMG estimator being more efficient.

Table 8: Alternative estimates of the average effect of household expenditures on calorie demand in Nicaragua over the period 2000–2002 ($T = 3$)

	Without time effects			With time effects			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	FE	GP	TMG	FE-TE	GP	TMG-TE	TMG-C
$\hat{\beta}_0$	0.6588	0.6034	0.5900	0.6968	0.6448	0.6370	0.6338
	(0.0233)	(0.0350)	(0.0284)	(0.0211)	(0.0330)	(0.0263)	(0.0261)
$\hat{\phi}_{2001}$	...	...	...	0.0727	0.0682	0.0708	0.0682
	...	...	...	(0.0087)	(0.0123)	(0.0088)	(0.0123)
$\hat{\phi}_{2002}$	...	...	...	0.1066	0.0954	0.1054	0.0682
	...	...	...	(0.0080)	(0.0108)	(0.0080)	(0.0123)
$\hat{\pi}$ ($\times 100$)	...	1.2	10.9	...	1.2	10.9	10.9

Notes: The estimates of $\beta_0 = E(\beta_i)$ and $(\phi_{2001}, \phi_{2002})'$ in the model (9.1) are based on the RPS panel of 1,358 households. $\hat{\pi}$ is the estimated fraction of individual estimates being trimmed, defined by (4.7). See also footnotes to Table 7.

Table 8 gives the estimates of β_0 for the extended panel, 2000–2002 (with $T = 3$), both with and without time effects. When time effects are included we provide two versions of the TMG estimates (TMG-TE and TMG-C), depending on how time effects are estimated.

[24] When $T = 2$, $\hat{\phi}_{2002}$ is not significant, and adding time effects does not change the estimated average effect.

[25] For the 2001–2002 panel, $\hat{\pi}$ of the GP estimator is identical to the one reported in Table 3 of Graham and Powell (2012). Graham and Powell (2012) estimated a model with time-varying coefficients, $y_{it} = \alpha_i + \phi_t + (\beta_i + \phi_{t,\beta})\, x_{it} + u_{it}$, where $(\boldsymbol{\phi}, \boldsymbol{\phi}_\beta)$ are identified by stayers but estimated by near stayers. While $\boldsymbol{\phi}_\beta$ is not included in (9.1), the GP estimates we compute are close to the trimmed estimates in Table 3 of Graham and Powell (2012).

[26] The bracketed figures are standard errors.

As in the case of the 2001–2002 panel, the FE estimates are larger than the GP and TMG estimates, but these differences are reduced somewhat, particularly when time effects are included in the panel regressions. Further, as expected, increasing T reduces the rate of trimming and brings the GP and TMG estimators closer to one another. The trimming rate for the GP estimator is very small indeed (only 1.2 per cent), as compared to around 11 per cent for the TMG estimator in the case of the 2000–2002 panel. The TMG-TE and TMG-C estimates of the time effects (ϕ_{2001} and ϕ_{2002}) are quite close and are both highly statistically significant and positive, capturing the upward trend in the calorie intake.

10 Conclusions

This paper studies estimation of average treatment effects in panel data models with possibly correlated heterogeneous coefficients, when the number of cross-sectional units is large, but the number of time periods can be as small as the number of regressors. We recall that the FE estimator is inconsistent under correlated heterogeneity, and the MG estimator can have unbounded first or second moments when applied to ultra short panels. Thus, the TMG estimator is proposed, where the trimming process is derived by a careful examination of the bias/efficiency trade-off in the asymptotic distribution. Conditions under which the TMG estimator is consistent and asymptotically normally distributed are provided. We also propose new estimators for ultra short panel data models with time effects, distinguishing between cases where $T > k$ and $T \geq k$, and derive their asymptotic distributions under the identifying condition that the dependence between heterogeneous slope coefficients and the regressors is time-invariant. Moreover, based on differences between the TMG and FE estimators (without and with time effects), we propose Hausman-type tests of correlated slope heterogeneity which must be applied before using FE or FE-TE estimators in practice.

Using Monte Carlo experiments, we highlight the bias and size distortion properties of the FE and FE-TE estimators under correlated heterogeneity. In contrast, the TMG and TMG-TE estimators are shown to have desirable finite sample performance under a number of different MC designs, allowing for Gaussian and non-Gaussian heteroskedastic error processes, dynamic heterogeneity and interactive time effects in the covariates, and different choices of the trimming threshold parameter, α. In particular, since the TMG and TMG-TE estimators exploit information on untrimmed and trimmed estimates alike, they have the smallest RMSE, and tests based on them have the correct size and are more powerful compared with the other trimmed estimators currently proposed in the literature.

The Hausman-type tests based on TMG and TMG-TE estimators are also shown to have very good small sample properties, with their size controlled and their power rising strongly with n even when $T = k = 2$. It is hoped that the new Hausman test provides empirical investigators with a diagnostic test that can be used before the application of the FE-TE estimators that are commonly used in the empirical literature. It is hoped that the use of this diagnostic test can help researchers in avoiding biased estimates and possibly misleading inferences. When the TMG and TMG-TE estimators and the Hausman-type tests of correlated heterogeneity are applied to a panel of households in poor rural communities in Nicaragua, the results provide clear evidence of correlated heterogeneity in the average effect of household expenditures on calorie demand, which sheds doubt on the application of FE-TE estimators to this data set.

Finally, we would like to end by acknowledging that, similarly to the FE-TE estimators, the validity of the TMG-TE and TMG-C estimators requires the so-called parallel trends assumption where time effects are assumed to have homogeneous effects across all individual units in the panel. Relaxing the parallel trends assumption in short T panels has been an important area of active research, but most of these contributions either assume homogeneous slopes or restricted forms of slope heterogeneity, or place restrictions on the time effects. The development of techniques for estimation and inference in ultra short panels with correlated heterogeneous slopes and interactive effects is a topic for future research.

References

Abadie, A. and M. D. Cattaneo (2018). Econometric methods for program evaluation. *Annual Review of Economics 10*, 465–503.

Arellano, M. and S. Bonhomme (2012). Identifying distributional characteristics in random coefficients panel data models. *The Review of Economic Studies 79*, 987–1020.

Athey, S. and G. W. Imbens (2017). The state of applied econometrics: causality and policy evaluation. *Journal of Economic Perspectives 31*, 3–32.

Banerjee, A., D. Karlan, and J. Zinman (2015). Six randomized evaluations of microcredit: introduction and further steps. *American Economic Journal: Applied Economics 7*, 1–21.

Bastagli, F., J. Hagen-Zanker, L. Harman, V. Barca, G. Sturge, and T. Schmidt (2019). The impact of cash transfers: a review of the evidence from low-and middle-income countries. *Journal of Social Policy 48*, 569–594.

Bonhomme, S. (2012). Functional differencing. *Econometrica 80*, 1337–1385.

Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica 60*, 567–596.

Chudik, A. and M. H. Pesaran (2015). Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors. *Journal of Econometrics 188*, 393–420.

Crépon, B. and G. J. Van Den Berg (2016). Active labor market policies. *Annual Review of Economics 8*, 521–546.

de Chaisemartin, C. and X. D'Haultfoeuille (2023). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: a survey. *The Econometrics Journal 26*, C1–C30.

de Chaisemartin, C., X. D'Haultfoeuille, F. Pasquier, and G. Vazquez-Bare (2023). Difference-in-differences estimators for treatments continuously distributed at every period. arXiv preprint arXiv:2201.06898.

Graham, B. S. and J. L. Powell (2012). Identification and estimation of average partial effects in "irregular" correlated random coefficient panel data models. *Econometrica 80*, 2105–2152.

Heckman, J. and E. Vytlacil (1998). Instrumental variables methods for the correlated random coefficient model: estimating the average rate of return to schooling when the return is correlated with schooling. *Journal of Human Resources 33*, 974–987.

Lee, Y. and D. Sul (2022). Trimmed mean group estimation. In *Essays in Honor of M. Hashem Pesaran: Panel Modeling, Micro Applications, and Econometric Methodology*. Emerald Publishing Limited.

Pesaran, M. H. and R. Smith (1995). Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics 68*, 79–113.

Pesaran, M. H., R. Smith, and K. S. Im (1996). Dynamic linear models for heterogenous panels. In P. Mátyás, Lászlóand Sevestre (Ed.), *The Econometrics of Panel Data: A Handbook of the Theory with Applications*. Springer Netherlands.

Pesaran, M. H. and T. Yamagata (2008). Testing slope homogeneity in large panels. *Journal of Econometrics 142*, 50–93.

Pesaran, M. H. and T. Yamagata (2023). Testing for Alpha in linear factor pricing models with a large number of securities. Published online in *Journal of Financial Econometrics*.

Sasaki, Y. and T. Ura (2021). Slow Movers in Panel Data. arXiv preprint arXiv:2110.12041.

Verdier, V. (2020). Average treatment effects for stayers with correlated random coefficient models of panel data. *Journal of Applied Econometrics 35*, 917–939.

Wooldridge, J. M. (2005). Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models. *Review of Economics and Statistics 87*, 385–390.

Appendix

Notations: Generic positive finite constants are denoted by C when large, and c when small. They can take different values at different instances. $\lambda_{\max}(\boldsymbol{A})$ and $\lambda_{\min}(\boldsymbol{A})$ denote the maximum and minimum eigenvalues of matrix $\boldsymbol{A}$. $\boldsymbol{A} \succ \boldsymbol{0}$ and $\boldsymbol{A} \succeq \boldsymbol{0}$ denote that $\boldsymbol{A}$ is a positive definite and a non-negative definite matrix, respectively. $\|\boldsymbol{A}\| = \lambda_{\max}^{1/2}(\boldsymbol{A}'\boldsymbol{A})$ and $\|\boldsymbol{A}\|_1$ denote the spectral and column norms of matrix $\boldsymbol{A}$, respectively. $\boldsymbol{A}^*$ denotes the adjoint of $\boldsymbol{A}$, such that $\boldsymbol{A}^{-1} = d^{-1}\boldsymbol{A}^*$, and $d = det(\boldsymbol{A})$. $\|\boldsymbol{x}\|_p = [E(\|\boldsymbol{x}\|^p)]^{1/p}$. If $\{f_n\}_{n=1}^{\infty}$ is any real sequence and $\{g_n\}_{n=1}^{\infty}$ is a sequences of positive real numbers, then $f_n = O(g_n)$, if there exists C such that $|f_n|/g_n \le C$ for all n. $f_n = o(g_n)$ if $f_n/g_n \to 0$ as $n \to \infty$. Similarly, $f_n = O_p(g_n)$ if f_n/g_n is stochastically bounded, and $f_n = o_p(g_n)$, if $f_n/g_n \to_p 0$. The operator $\to_p$ denotes convergence in probability, and $\to_d$ denotes convergence in distribution.

A.1 Lemmas

Lemma A.1 *Suppose that Assumptions 2, 5 and 6 hold. Then for each i, we have*

$$E\left[d_i^s \mathbf{1}\{d_i \le a_n\}\right] = O(a_n^{s+1}), \text{ for } s = 1, 2, ..., \tag{A.1.1}$$

$$E(\delta_i) = O(a_n), \text{ and } E(\delta_i^2) = O(a_n), \tag{A.1.2}$$

$$E(\delta_i \boldsymbol{\eta}_i) = O(a_n), \text{ and } E\left(\delta_i^2 \boldsymbol{\eta}_i\right) = O(a_n), \tag{A.1.3}$$

$$n^{-1}\sum_{i=1}^{n}\left\{E\left[d_i^2 \mathbf{1}\{d_i \le a_n\}\right]\right\}^{1/2} = O(a_n^{3/2}), \tag{A.1.4}$$

$$n^{-1}\sum_{i=1}^{n}\left\{E\left[d_i^{-2} \mathbf{1}\{d_i > a_n\}\right]\right\}^{1/2} = O(a_n^{-1}). \tag{A.1.5}$$

Proof. By mean value theorem (MVT), under Assumption 5, we have

$$F_d(a_n) = F_d(0) + f_d(\bar{a}_n)a_n = f_d(\bar{a}_n)a_n = O(a_n), \tag{A.1.6}$$

where $\bar{a}_n$ lies on the line segment between 0 and a_n. Similarly, let $\psi(a_n) = \int_0^{a_n} u^s f_d(u)du$ and note that $\psi'(a_n) = a_n^s f_d(a_n)$. Then by MVT $\psi(a_n) = \psi(0) + [\bar{a}_n^s f_d(\bar{a}_n)]\, a_n$, and we have

$$E\left[d_i^s \mathbf{1}\{d_i \le a_n\}\right] = \int_0^{a_n} u^s f_d(u)du = \bar{a}_n^s f_d(\bar{a}_n)a_n = O(a_n^{s+1}), \text{ for } s = 1, 2, ... \tag{A.1.7}$$

Using the above results

$$E(\delta_i) = E\left[\left(\frac{d_i - a_n}{a_n}\right)\mathbf{1}\{d_i \leq a_n\}\right] = a_n^{-1}E\left[d_i\mathbf{1}\{d_i \leq a_n\}\right] - E\left[\mathbf{1}\{d_i \leq a_n\}\right]$$
$$= a_n^{-1}O(a_n^2) - F_d(a_n) = O(a_n), \tag{A.1.8}$$

and

$$E(\delta_i^2) = E\left[\left(\frac{d_i - a_n}{a_n}\right)^2\mathbf{1}\{d_i \leq a_n\}\right]$$
$$= a_n^{-2}E\left[d_i^2\mathbf{1}\{d_i \leq a_n\}\right] + E\left[\mathbf{1}\{d_i \leq a_n\}\right] - 2a_n^{-1}E\left[d_i\mathbf{1}\{d_i \leq a_n\}\right]$$
$$= a_n^{-2}O(a_n^3) + F_d(a_n) - 2a_n^{-1}O(a_n^2) = O(a_n). \tag{A.1.9}$$

Consider now the terms involving the products of δ_i and $\boldsymbol{\eta}_i$

$$E\left(\delta_i\boldsymbol{\eta}_i\right) = \boldsymbol{B}_i E\left[\left(\frac{d_i - a_n}{a_n}\right)\mathbf{1}\{d_i \leq a_n\}\left[\boldsymbol{g}(d_i) - E\left[\boldsymbol{g}(d_i)\right]\right]\right]. \tag{A.1.10}$$

Since $\boldsymbol{B}_i$ is bounded and does not depend on d_i, without loss of generality we set $\boldsymbol{B}_i = \boldsymbol{I}_k$ and consider the the j^{th} term of (A.1.10), namely

$$s_j(a_n) = E\left\{\left(\frac{d_i - a_n}{a_n}\right)\mathbf{1}\{d_i \leq a_n\}\left[g_j(d_i) - E\left[g_j(d_i)\right]\right]\right\}$$
$$= \frac{1}{a_n}\int_0^{a_n} u g_j(u) f_d(u)du - \int_0^{a_n} g_j(u) f_d(u)du$$
$$- E\left[g_j(d_i)\right]\left[\frac{1}{a_n}\int_0^{a_n} u f_d(u)du\right] + E\left[g_j(d_i)\right]\left[\int_0^{a_n} f_d(u)du\right].$$

By Assumption 6 $E\left[g_j(d_i)\right] < C$, and using (A.1.6) and (A.1.1) we have

$$\int_0^{a_n} f_d(u)du = O(a_n), \text{ and } a_n^{-1}\int_0^{a_n} u f_d(u)du = O(a_n).$$

Also by the mean value theorem

$$\int_{u=0}^{a_n} g_j(u) f_d(u)du = g_j(\bar{a}_n) f_d(\bar{a}_n) a_n = O(a_n),$$
$$\frac{1}{a_n}\int_{u=0}^{a_n} u g_j(u) f_d(u)du = \frac{1}{a_n}\left[\bar{a}_n g_j(\bar{a}_n) f_d(\bar{a}_n) a_n\right] = O(a_n).$$

Hence, $E\left(\delta_i \boldsymbol{\eta}_i\right) = O(a_n)$. Similarly the j^{th} term of $E\left(\delta_i^2 \boldsymbol{\eta}_i\right)$ (setting $\boldsymbol{B}_i = \boldsymbol{I}_k$) is given by

$$\begin{aligned} s_{j2}(a_n) &= E\left\{\left(\frac{d_i - a_n}{a_n}\right)^2 \mathbf{1}\{d_i \leq a_n\}\left[g_j(d_i) - E\left[g_j(d_i)\right]\right]\right\} \\ &= E\left\{\left(\frac{d_i^2}{a_n^2} - 1 - 2\frac{d_i - a_n}{a_n}\right) \mathbf{1}\{d_i \leq a_n\}\left[g_j(d_i) - E\left[g_j(d_i)\right]\right]\right\}. \quad \text{(A.1.11)} \end{aligned}$$

Consider the first term

$$\begin{aligned} & E\left\{\frac{d_i^2}{a_n^2} \mathbf{1}\{d_i \leq a_n\}\left[g_j(d_i) - E\left[g_j(d_i)\right]\right]\right\} \\ = & \frac{1}{a_n^2} \int_0^{a_n} u^2 g_j(u) f_d(u) du - \frac{1}{a_n^2} E\left[g_j(d_i)\right] E\left[d_i^2 \mathbf{1}\{d_i \leq a_n\}\right], \end{aligned}$$

and again by mean value theorem $a_n^{-2} \int_0^{a_n} u^2 g_j(u) f_d(u) du = O(a_n)$, $E\left[g_j(d_i)\right] < C$, and using (A.1.7) $E\left[d_i^2 \mathbf{1}\{d_i \leq a_n\}\right] = O(a_n^3)$. Hence, the first term of (A.1.11) is $O(a_n)$. For its second term, we have

$$E\left\{\mathbf{1}\{d_i \leq a_n\}\left[g_j(d_i) - E\left[g_j(d_i)\right]\right]\right\} = \int_0^{a_n} g_j(u) f_d(u) du - E\left[g_j(d_i)\right] \int_0^{a_n} f_d(u) du = O(a_n),$$

and the order of the third term is already established to be $O(a_n)$. Hence, it follows that $E\left(\delta_i^2 \boldsymbol{\eta}_i\right) = O(a_n)$. Finally, result (A.1.4) follows from (A.1.1) and (A.1.5) follows noting that $d_i^{-2} \mathbf{1}\{d_i > a_n\} \leq a_n^{-2}$. ■

Lemma A.2 *Suppose that Assumptions 1, 2, 4, 5 and 6, hold. Let*

$$\bar{\boldsymbol{\xi}}_{\delta,nT} = n^{-1} \sum_{i=1}^n \left(1 + \delta_i\right) \boldsymbol{\xi}_{iT},$$

where $\delta_i = \left(\frac{d_i - a_n}{a_n}\right) \mathbf{1}\{d_i \leq a_n\}$, $a_n = C_n n^{-\alpha}$, $C_n < C$, $d_i = \det(\boldsymbol{W}_i' \boldsymbol{W}_i)$, *and* $\boldsymbol{\xi}_{iT} = \left(\boldsymbol{W}_i' \boldsymbol{W}_i\right)^{-1} \boldsymbol{W}_i' \boldsymbol{u}_i = \boldsymbol{R}_i' \boldsymbol{u}_i$. *Then*

$$E\left(\bar{\boldsymbol{\xi}}_{\delta,nT}\right) = \mathbf{0}, \quad \text{(A.1.12)}$$

$$Var\left(\bar{\boldsymbol{\xi}}_{\delta,nT}\right) = n^{-2} \sum_{i=1}^n E\left[\mathbf{1}\{d_i > a_n\} \boldsymbol{R}_i' \boldsymbol{H}_i \boldsymbol{R}_i\right] + n^{-2} \sum_{i=1}^n a_n^{-2} E\left[d_i^2 \mathbf{1}\{d_i \leq a_n\} \boldsymbol{R}_i' \boldsymbol{H}_i \boldsymbol{R}_i\right], \quad \text{(A.1.13)}$$

$$Var\left(\bar{\boldsymbol{\xi}}_{\delta,nT}\right) = O\left(n^{-1+\alpha}\right), \quad \text{(A.1.14)}$$

and

$$E\left[n^{-1} \sum_{i=1}^n a_n^{-1} d_i^2 \mathbf{1}\{d_i \leq a_n\} \boldsymbol{R}_i' \boldsymbol{H}_i(\boldsymbol{W}_i) \boldsymbol{R}_i\right] = O\left(a_n^{1/2}\right). \quad \text{(A.1.15)}$$

Proof. Under Assumptions 1 and conditional on $\boldsymbol{W}_i$ (and hence on d_i), $(1+\delta_i)\boldsymbol{\xi}_{iT}$ are distributed independently over i and

$$\begin{aligned}
E\left(\bar{\boldsymbol{\xi}}_{\delta,nT}\left|\boldsymbol{W}_i\right.\right) &= n^{-1}\sum_{i=1}^{n}(1+\delta_i)\boldsymbol{R}_i' E\left(\boldsymbol{u}_i\left|\boldsymbol{W}_i\right.\right)=\boldsymbol{0},\\
Var\left(\bar{\boldsymbol{\xi}}_{\delta,nT}\left|\boldsymbol{W}_i\right.\right) &= n^{-2}\sum_{i=1}^{n}(1+\delta_i)^2 E\left(\boldsymbol{\xi}_{iT}\boldsymbol{\xi}_{iT}'\left|\boldsymbol{W}_i\right.\right)\\
&= n^{-2}\sum_{i=1}^{n}(1+\delta_i)^2\boldsymbol{R}_i' E\left(\boldsymbol{u}_i\boldsymbol{u}_i'\left|\boldsymbol{W}_i\right.\right)\boldsymbol{R}_i = n^{-2}\sum_{i=1}^{n}(1+\delta_i)^2\boldsymbol{R}_i'\boldsymbol{H}_i\boldsymbol{R}_i,
\end{aligned}$$

where $\boldsymbol{H}_i = E\left(\boldsymbol{u}_i\boldsymbol{u}_i'\left|\boldsymbol{W}_i\right.\right)$. We have suppressed the dependence of $\boldsymbol{H}_i$ on $\boldsymbol{W}_i$ to simplify the exposition. Hence, $E\left(\bar{\boldsymbol{\xi}}_{\delta,nT}\right)=\boldsymbol{0}$, and $Var\left(\bar{\boldsymbol{\xi}}_{\delta,nT}\right)=E\left[Var\left(\bar{\boldsymbol{\xi}}_{\delta,nT}\left|\boldsymbol{W}_i\right.\right)\right]$. To establish (A.1.13) note that

$$(1+\delta_i)^2 = \mathbf{1}\{d_i>a_n\}+a_n^{-2}d_i^2\mathbf{1}\{d_i\le a_n\}, \tag{A.1.16}$$

and

$$Var\left(\bar{\boldsymbol{\xi}}_{\delta,nT}\right)=n^{-2}\sum_{i=1}^{n}E\left[\mathbf{1}\{d_i>a_n\}\boldsymbol{R}_i'\boldsymbol{H}_i\boldsymbol{R}_i\right]+n^{-2}E\left[\sum_{i=1}^{n}a_n^{-2}d_i^2\mathbf{1}\{d_i\le a_n\}\boldsymbol{R}_i'\boldsymbol{H}_i\boldsymbol{R}_i\right].$$

Since $\boldsymbol{H}_i$ is positive definite and by Assumption 1 $\sup_i\lambda_{max}\left(\boldsymbol{H}_i\right)<C$,

$$\left\|\boldsymbol{R}_i'\boldsymbol{H}_i\boldsymbol{R}_i\right\|\le\lambda_{max}\left(\boldsymbol{H}_i\right)\left\|\boldsymbol{R}_i'\boldsymbol{R}_i\right\|=\lambda_{max}\left(\boldsymbol{H}_i\right)\left\|\left(\boldsymbol{W}_i'\boldsymbol{W}_i\right)^{-1}\right\|<Cd_i^{-1}\left\|\left(\boldsymbol{W}_i'\boldsymbol{W}_i\right)^{*}\right\| \tag{A.1.17}$$

and

$$\begin{aligned}
\left\|Var\left(\bar{\boldsymbol{\xi}}_{\delta,nT}\right)\right\| &\le Cn^{-2}\sum_{i=1}^{n}E\left[\mathbf{1}\{d_i>a_n\}d_i^{-1}\left\|\left(\boldsymbol{W}_i'\boldsymbol{W}_i\right)^{*}\right\|\right]\\
&\quad +Cn^{-2}E\left[\sum_{i=1}^{n}a_n^{-2}d_i\mathbf{1}\{d_i\le a_n\}\left\|\left(\boldsymbol{W}_i'\boldsymbol{W}_i\right)^{*}\right\|\right].
\end{aligned}$$

By Cauchy-Schwarz inequality

$$\begin{aligned}
E\left[\mathbf{1}\{d_i>a_n\}d_i^{-1}\left\|\left(\boldsymbol{W}_i'\boldsymbol{W}_i\right)^{*}\right\|\right] &\le \left\{E\left[d_i^{-2}\mathbf{1}\{d_i>a_n\}\right]\right\}^{1/2}\left\{E\left[\left\|\left(\boldsymbol{W}_i'\boldsymbol{W}_i\right)^{*}\right\|^2\right]\right\}^{1/2},\\
E\left[d_i\mathbf{1}\{d_i\le a_n\}\lambda_{max}\left(\boldsymbol{H}_i\right)\left\|\left(\boldsymbol{W}_i'\boldsymbol{W}_i\right)^{*}\right\|\right] &\le \left\{E\left[d_i^{2}\mathbf{1}\{d_i\le a_n\}\right]\right\}^{1/2}\left\{E\left[\left\|\left(\boldsymbol{W}_i'\boldsymbol{W}_i\right)^{*}\right\|^2\right]\right\}^{1/2},
\end{aligned}$$

and since by Assumption 4 $\sup_i E\left[\left\|(\boldsymbol{W}_i'\boldsymbol{W}_i)^*\right\|^2\right] < C$, then

$$\left\|Var\left(\bar{\boldsymbol{\xi}}_{\delta,nT}\ \right)\right\| \leq C\left[n^{-2}\sum_{i=1}^{n}\left\{E\left[d_i^{-2}\mathbf{1}\{d_i > a_n\}\right]\right\}^{1/2} + a_n^{-2}n^{-2}\sum_{i=1}^{n}\left\{E\left[d_i^{2}\mathbf{1}\{d_i \leq a_n\}\right]\right\}^{1/2}\right].$$

Now using results (A.1.4) and(A.1.5) of Lemma A.1, we have

$$\begin{aligned} n^{-2}\sum_{i=1}^{n}\left\{E\left[d_i^{-2}\mathbf{1}\{d_i > a_n\}\right]\right\}^{1/2} &= O(n^{-1}a_n^{-1}), \\ \text{and } n^{-2}\sum_{i=1}^{n}\left\{E\left[d_i^{2}\mathbf{1}\{d_i \leq a_n\}\right]\right\}^{1/2} &= O(n^{-1}a_n^{3/2}), \end{aligned}$$

then $\left\|Var\left(\bar{\boldsymbol{\xi}}_{\delta,nT}\right)\right\| = O(n^{-1}a_n^{-1}) + O(n^{-1}a_n^{-1/2})$, and since $a_n = C_n n^{-\alpha}$ result (A.1.14) follows. To establish (A.1.15) using (A.1.17) we have

$$\left\|n^{-1}\sum_{i=1}^{n}a_n^{-1}d_i^2\mathbf{1}\{d_i \leq a_n\}\boldsymbol{R}_i'\boldsymbol{H}_i\boldsymbol{R}_i\right\| \leq Cn^{-1}\sum_{i=1}^{n}a_n^{-1}d_i\mathbf{1}\{d_i \leq a_n\}\left\|(\boldsymbol{W}_i'\boldsymbol{W}_i)^*\right\|, \quad \text{(A.1.18)}$$

and by Cauchy-Schwarz inequality,

$$\begin{aligned} & E\left\|n^{-1}\sum_{i=1}^{n}a_n^{-1}d_i^2\mathbf{1}\{d_i \leq a_n\}\boldsymbol{R}_i'\boldsymbol{H}_i\boldsymbol{R}_i\right\| \\ \leq\ & Cn^{-1}\sum_{i=1}^{n}a_n^{-1}\left\{E\left[d_i^2\mathbf{1}\{d_i \leq a_n\}\right]\right\}^{1/2}\left[E\left\|(\boldsymbol{W}_i'\boldsymbol{W}_i)^*\right\|^2\right]^{1/2}. \end{aligned}$$

Under Assumption 4, $\sup_i E\left\|(\boldsymbol{W}_i'\boldsymbol{W}_i)^*\right\|^2 < C$, and we have

$$E\left\|n^{-1}\sum_{i=1}^{n}a_n^{-1}d_i^2\mathbf{1}\{d_i \leq a_n\}\boldsymbol{R}_i'\boldsymbol{H}_i(\boldsymbol{W}_i)\boldsymbol{R}_i\right\| \leq C\left[n^{-1}\sum_{i=1}^{n}\left\{E\left[a_n^{-2}d_i^2\mathbf{1}\{d_i \leq a_n\}\right]\right\}^{1/2}\right].$$

Now using (A.1.1) $E\left[a_n^{-2}d_i^2\mathbf{1}\{d_i \leq a_n\}\right] = a_n^{-2}O(a_n^3) = O\left(a_n\right)$, and result (A.1.15) follows. ■

Lemma A.3 *Let*

$$\begin{aligned} v_{it} - \bar{v}_{i\circ} &= u_{it} - u_{i\circ} + \left(\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_{i\circ}\right)'\boldsymbol{\eta}_{i\beta},\ \textit{for } i = 1,2,...,n; t = 1,2,...,T, \\ \bar{v}_{\circ t} - \bar{v}_{\circ\circ} &= n^{-1}\sum_{i=1}^{n}\left(v_{it} - \bar{v}_{i\circ}\right) = \left(\bar{u}_{\circ t} - \bar{u}_{\circ\circ}\right) + n^{-1}\sum_{i=1}^{n}\left(\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_{i\circ}\right)'\boldsymbol{\eta}_{i\beta}, \end{aligned}$$

where $\boldsymbol{\eta}_{i\beta} = \boldsymbol{\beta}_i - \boldsymbol{\beta}_0$*, and suppose that Assumptions 1, 6 and 8 hold. Then*

$$E\left(v_{it} - \bar{v}_{i\circ}\right) = 0, \text{ for } i = 1, 2, ..., n; t = 1, 2, ..., T, \tag{A.1.19}$$

$$\bar{v}_{\circ t} - \bar{v}_{\circ\circ} = O_p(n^{-1/2}), \text{ for } t = 1, 2, ..., T, \tag{A.1.20}$$

and (noting that T *is fixed as* $n \to \infty$*)*

$$\sqrt{n}\left(\bar{\boldsymbol{v}}_T - \bar{v}_{\circ\circ}\boldsymbol{\tau}_T\right) \to_d N(\mathbf{0}, \boldsymbol{\Omega}_\nu), \tag{A.1.21}$$

where $\bar{\boldsymbol{v}}_T = (\bar{v}_{\circ 1}, \bar{v}_{\circ 2}, ..., \bar{v}_{\circ T})' = n^{-1}\sum_{i=1}^{n} \boldsymbol{\nu}_{i\circ}$*,* $\boldsymbol{\nu}_{i\circ} = (\nu_{i1}, \nu_{i2}, ..., \nu_{iT})'$*,*

$$\boldsymbol{\Omega}_\nu = \boldsymbol{M}_T\left[\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} E\left(\boldsymbol{\nu}_{i\circ}\boldsymbol{\nu}_{i\circ}'\right)\right]\boldsymbol{M}_T, \tag{A.1.22}$$

and $\boldsymbol{M}_T = \boldsymbol{I}_T - T^{-1}\boldsymbol{\tau}_T\boldsymbol{\tau}_T'$*.*

Proof. Under Assumptions 1 and 8, $E(u_{it}) = 0$ and $E\left(\boldsymbol{x}_{it}'\boldsymbol{\eta}_{i\beta}\right) = E\left(\boldsymbol{x}_{is}'\boldsymbol{\eta}_{i\beta}\right)$ for all t and s. Hence

$$E\left(u_{it} - u_{i\circ}\right) = 0, \text{ and } E\left[\left(\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_{i\circ}\right)'\boldsymbol{\eta}_{i\beta}\right] = E\left(\boldsymbol{x}_{it}'\boldsymbol{\eta}_{i\beta}\right) - T^{-1}\sum_{t'=1}^{T} E\left(\boldsymbol{x}_{it'}'\boldsymbol{\eta}_{i\beta}\right) = 0,$$

then result (A.1.19) follows. Result (A.1.20) also follows noting that under Assumptions 1 and 6, $\{v_{it} - \bar{v}_{i\circ}$, for $i = 1, 2, ..., n\}$, are cross-sectionally independent with mean zero and finite variances. To establish A.1.21 we first note that $\bar{v}_{\circ\circ} = T^{-1}\left(\boldsymbol{\tau}_T'\bar{\boldsymbol{v}}_T\right)$, and hence

$$\sqrt{n}\left(\bar{\boldsymbol{v}}_T - \bar{v}_{\circ\circ}\boldsymbol{\tau}_T\right) = \boldsymbol{M}_T\sqrt{n}\bar{\boldsymbol{v}}_T = n^{-1/2}\sum_{i=1}^{n}\boldsymbol{M}_T\boldsymbol{\nu}_{i\circ},$$

where $\boldsymbol{M}_T\boldsymbol{\nu}_{i\circ}$ is a $T \times 1$ vector (T is fixed) with zero means and finite variances, and by Assumption 6 are cross-sectionally independent. Therefore, result A.1.21 follows by standard central limit theorems for independent but not identically distributed random variables. ■

A.2 Proof of Propositions and Theorems

A.2.1 Proof of Proposition 1

Proof. Under Assumption 2, $\hat{\boldsymbol{\theta}}_{MG} \to_p \boldsymbol{\theta}_0$ if $\bar{\boldsymbol{\xi}}_{nT} \to_p \mathbf{0}$. A sufficient (but not necessary) condition for the latter to hold can be obtained by applying Markov inequality to $\bar{\boldsymbol{\xi}}_{nT}$, i.e., for any fixed $\epsilon > 0$, $Pr\left(\left\|\bar{\boldsymbol{\xi}}_{nT}\right\| \geq \epsilon\right) \leq \frac{E\left\|\bar{\boldsymbol{\xi}}_{nT}\right\|^2}{\epsilon^2}$. Thus for $\bar{\boldsymbol{\xi}}_{nT} \rightarrow_p \mathbf{0}$, it is sufficient to show that $E\left\|\bar{\boldsymbol{\xi}}_{nT}\right\|^2 \rightarrow 0$. In what follows we find conditions under which $E\left\|\bar{\boldsymbol{\xi}}_{nT}\right\|^2 = O(n^{-1})$, and hence establish that $\bar{\boldsymbol{\xi}}_{nT} \rightarrow_p \mathbf{0}$ at the regular rate of $n^{-1/2}$. Note that

$$\left\|\bar{\boldsymbol{\xi}}_{nT}\right\|^2 = n^{-2}\left\|\sum_{i=1}^{n}\boldsymbol{\xi}_{iT}\right\|^2 = n^{-2}\left(\sum_{i=1}^{n}\boldsymbol{\xi}_{iT}\right)'\left(\sum_{i=1}^{n}\boldsymbol{\xi}_{iT}\right) = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{\xi}_{iT}'\boldsymbol{\xi}_{jT}.$$

Hence $E\left\|\bar{\boldsymbol{\xi}}_{nT}\right\|^2 = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}E\left(\boldsymbol{\xi}_{iT}'\boldsymbol{\xi}_{jT}\right)$. Since under Assumption 1 $u_{it}'s$ are cross-sectionally independent and we have

$$E\left\|\bar{\boldsymbol{\xi}}_{nT}\right\|^2 = n^{-2}\sum_{i=1}^{n}E\left(\boldsymbol{\xi}_{iT}'\boldsymbol{\xi}_{iT}\right). \tag{A.2.1}$$

Then using (2.7),

$$E\left(\boldsymbol{\xi}_{iT}'\boldsymbol{\xi}_{iT}|\boldsymbol{W}_i\right) = E\left(\boldsymbol{u}_i'\boldsymbol{R}_i\boldsymbol{R}_i'\boldsymbol{u}_i|\boldsymbol{W}_i\right) = E\left[Tr\left(\boldsymbol{R}_i'\boldsymbol{u}_i\boldsymbol{u}_i'\boldsymbol{R}_i\right)|\boldsymbol{W}_i\right] = Tr\left(\boldsymbol{R}_i'\boldsymbol{H}_i(\boldsymbol{W}_i)\boldsymbol{R}_i\right),$$

where by Assumption 1, $\boldsymbol{H}_i(\boldsymbol{W}_i) = E(\boldsymbol{u}_i\boldsymbol{u}_i'|\boldsymbol{W}_i)$. Also

$$\begin{aligned}Tr\left(\boldsymbol{R}_i'\boldsymbol{H}_i(\boldsymbol{W}_i)\boldsymbol{R}_i\right) &\leq T\lambda_{max}\left[\boldsymbol{H}_i(\boldsymbol{W}_i)\right]Tr\left(\boldsymbol{R}_i'\boldsymbol{R}_i\right)\\ &= T\lambda_{max}\left[\boldsymbol{H}_i(\boldsymbol{W}_i)\right]Tr\left[\left(\boldsymbol{W}_i'\boldsymbol{W}_i\right)^{-1}\right] \leq T\lambda_{max}\left[\boldsymbol{H}_i(\boldsymbol{W}_i)\right]\left\{k\lambda_{max}\left[\left(\boldsymbol{W}_i'\boldsymbol{W}_i\right)^{-1}\right]\right\}.\end{aligned}$$

Since T and k are finite, and under Assumption 1, $\sup_i \lambda_{max}\left[\boldsymbol{H}_i(\boldsymbol{W}_i)\right] < C$,

$$Tr\left(\boldsymbol{R}_i'\boldsymbol{H}_i(\boldsymbol{W}_i)\boldsymbol{R}_i\right) \leq C\lambda_{max}\left(\boldsymbol{W}_i'\boldsymbol{W}_i\right)^{-1}.$$

Then given (A.2.1) we have

$$E\left\|\bar{\boldsymbol{\xi}}_{nT}\right\|^2 = n^{-2}\sum_{i=1}^{n}E\left[Tr\left(\boldsymbol{R}_i'\boldsymbol{H}_i(\boldsymbol{W}_i)\boldsymbol{R}_i\right)\right] \leq Cn^{-2}\sum_{i=1}^{n}E\left\{\lambda_{max}\left[\left(\boldsymbol{W}_i'\boldsymbol{W}_i\right)^{-1}\right]\right\}.$$

Hence, $E\left\|\bar{\boldsymbol{\xi}}_{nT}\right\|^2 = O\left(n^{-1}\right)$, if

$$\sup_i E\left\{\lambda_{max}\left[\left(\boldsymbol{W}_i'\boldsymbol{W}_i\right)^{-1}\right]\right\} < C < \infty. \tag{A.2.2}$$

It is also worth noting that condition (A.2.2) can be written in terms of column or row norms of $(\boldsymbol{W}_i'\boldsymbol{W}_i)^{-1}$ which is easier to use in practice. Since $\boldsymbol{W}_i'\boldsymbol{W}_i$ is a symmetric matrix then it follows that $\lambda_{\max}\left[(\boldsymbol{W}_i'\boldsymbol{W}_i)^{-1}\right] \leq \left\|(\boldsymbol{W}_i'\boldsymbol{W}_i)^{-1}\right\|_1$, where $\left\|\boldsymbol{A}\right\|_1$ denotes the column norm of

$\boldsymbol{A}$. Also $(\boldsymbol{W}_i'\boldsymbol{W}_i)^{-1} = d_i^{-1}(\boldsymbol{W}_i'\boldsymbol{W}_i)^*$, where $d_i = \det(\boldsymbol{W}_i'\boldsymbol{W}_i)$, and $(\boldsymbol{W}_i'\boldsymbol{W}_i)^*$ is the adjoint of $\boldsymbol{W}_i'\boldsymbol{W}_i$. Then $\lambda_{\max}\left[(\boldsymbol{W}_i'\boldsymbol{W}_i)^{-1}\right] \leq d_i^{-1}\left\|(\boldsymbol{W}_i'\boldsymbol{W}_i)^*\right\|_1$, and by Cauchy–Schwarz inequality

$$E\left\{\lambda_{\max}\left[(\boldsymbol{W}_i'\boldsymbol{W}_i)^{-1}\right]\right\} \leq \left[E\left(d_i^{-2}\right)\right]^{1/2}\left\{E\left[\|(\boldsymbol{W}_i'\boldsymbol{W}_i)^*\|_1^2\right]\right\}^{1/2},$$

hence equation (A.2.2) will hold under the following conditions

$$\sup_i E\left(d_i^{-2}\right) < C, \text{ and } \sup_i E\left[\|(\boldsymbol{W}_i'\boldsymbol{W}_i)^*\|_1^2\right] < C, \text{ for } i = 1, 2, ..., n.$$

Under the above conditions $\hat{\boldsymbol{\theta}}_{MG}$ converges in probability to $\boldsymbol{\theta}_0$ at the regular rate of $n^{-1/2}$, irrespective of whether $\boldsymbol{\theta}_i$ are correlated with the regressors or not, and it is robust to error serial correlation and conditional heteroskedasticity. ■

A.2.2 Proof of Proposition 3

Proof. Consider

$$\bar{\boldsymbol{\Psi}}_n\boldsymbol{A}_n\bar{\boldsymbol{\Psi}}_n = \bar{\boldsymbol{\Psi}}_n\boldsymbol{\Omega}_\beta\bar{\boldsymbol{\Psi}}_n - \left(n^{-1}\sum_{i=1}^n \boldsymbol{\Psi}_{ix}\boldsymbol{\Omega}_\beta\boldsymbol{\Psi}_{ix}\right),$$

and without loss of generality suppose that $\boldsymbol{\Omega}_\beta$ is positive definite. Then

$$\bar{\boldsymbol{\Psi}}_n\boldsymbol{A}_n\bar{\boldsymbol{\Psi}}_n = -\left[n^{-1}\sum_{i=1}^n \boldsymbol{P}_i\boldsymbol{P}_i' - \bar{\boldsymbol{P}}_n\bar{\boldsymbol{P}}_n'\right] = -n^{-1}\sum_{i=1}^n\left(\boldsymbol{P}_i - \bar{\boldsymbol{P}}_n\right)\left(\boldsymbol{P}_i - \bar{\boldsymbol{P}}_n\right)',$$

where $\boldsymbol{P}_i = \boldsymbol{\Psi}_{ix}\boldsymbol{\Omega}_\beta^{1/2}$ and $\bar{\boldsymbol{P}}_n = n^{-1}\sum_{i=1}^n \boldsymbol{P}_i$. Hence $\boldsymbol{A}_n = -\bar{\boldsymbol{\Psi}}_n^{-1}\boldsymbol{V}_n^P\bar{\boldsymbol{\Psi}}_n^{-1}$, where

$$\boldsymbol{V}_n^P = \left[n^{-1}\sum_{i=1}^n\left(\boldsymbol{P}_i - \bar{\boldsymbol{P}}_n\right)\left(\boldsymbol{P}_i - \bar{\boldsymbol{P}}_n\right)'\right].$$

It is clear that $\boldsymbol{V}_n^P$ is semi-positive definite and by Assumption 3 $\bar{\boldsymbol{\Psi}}_n$ is positive definite. Then it follows that $\bar{\boldsymbol{\Psi}}_n^{-1}\boldsymbol{V}_n^P\bar{\boldsymbol{\Psi}}_n^{-1}$ is also semi-positive definite and hence $\boldsymbol{A}_n$ is non-positive definite, $\boldsymbol{A}_n \preceq \boldsymbol{0}$. For $\boldsymbol{B}_n$ we have

$$\begin{aligned}\bar{\boldsymbol{\Psi}}_n\boldsymbol{B}_n\bar{\boldsymbol{\Psi}}_n &= \bar{\boldsymbol{\Psi}}_n\left[n^{-1}\sum_{i=1}^n \boldsymbol{\Psi}_{ix}^{-1}\boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{H}_i(\boldsymbol{X}_i)\boldsymbol{M}_T\boldsymbol{X}_i\boldsymbol{\Psi}_{ix}^{-1}\right]\bar{\boldsymbol{\Psi}}_n \\ &\quad - \left[n^{-1}\sum_{i=1}^n \boldsymbol{X}_i'\boldsymbol{M}_T\boldsymbol{H}_i(\boldsymbol{X}_i)\boldsymbol{M}_T\boldsymbol{X}_i\right],\end{aligned}$$

and in general it is not possible to sign $\bar{\boldsymbol{\Psi}}_n\boldsymbol{B}_n\bar{\boldsymbol{\Psi}}_n$. The outcome depends on the heterogeneity of error variances and their interactions with the heterogeneity of regressors. We have already seen that $\boldsymbol{B}_n \succeq \mathbf{0}$ when $\boldsymbol{H}_i(\boldsymbol{X}_i) = \sigma^2 \boldsymbol{I}_T$, but this result need not hold in a more general setting where $\boldsymbol{H}_i(\boldsymbol{X}_i)$ varies across i. ■

A.2.3 Proof of Theorem 2

Proof. Using (2.5), (4.2) and (5.2), we have

$$\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_0 = \delta_i \boldsymbol{\theta}_0 + \boldsymbol{\zeta}_{iT}, \tag{A.2.3}$$

where $\boldsymbol{\zeta}_{iT} = (1+\delta_i)\boldsymbol{\eta}_i + (1+\delta_i)\boldsymbol{\xi}_{iT}$, and using (5.5)

$$\hat{\boldsymbol{\theta}}_{TMG} - \boldsymbol{\theta}_0 = \left(\frac{1}{1+\bar{\delta}_n}\right) \bar{\boldsymbol{\zeta}}_{nT}, \tag{A.2.4}$$

where $\bar{\boldsymbol{\zeta}}_{nT} = n^{-1}\sum_{i=1}^n \boldsymbol{\eta}_i + n^{-1}\sum_{i=1}^n \delta_i \boldsymbol{\eta}_i + n^{-1}\sum_{i=1}^n (1+\delta_i)\boldsymbol{\xi}_{iT}$. Subtracting (A.2.4) from (A.2.3) now yields

$$\tilde{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_{TMG} = \boldsymbol{\zeta}_{iT} + \delta_i \boldsymbol{\theta}_0 - \left(\frac{1}{1+\bar{\delta}_n}\right) \bar{\boldsymbol{\zeta}}_{nT},$$

and we have

$$\begin{aligned}
& n^{-1}\sum_{i=1}^n \left(\tilde{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_{TMG}\right)\left(\tilde{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_{TMG}\right)' \\
&= n^{-1}\sum_{i=1}^n \boldsymbol{\zeta}_{iT}\boldsymbol{\zeta}_{iT}' + \left(n^{-1}\sum_{i=1}^n \delta_i^2\right)\boldsymbol{\theta}_0\boldsymbol{\theta}_0' + \left(n^{-1}\sum_{i=1}^n \delta_i \boldsymbol{\zeta}_{iT}\right)\boldsymbol{\theta}_0' + \boldsymbol{\theta}_0\left(n^{-1}\sum_{i=1}^n \delta_i \boldsymbol{\zeta}_{iT}'\right) \\
&\quad + \left[\left(\frac{1}{1+\bar{\delta}_n}\right)^2 - 2\left(\frac{1}{1+\bar{\delta}_n}\right)\right]\bar{\boldsymbol{\zeta}}_{nT}\bar{\boldsymbol{\zeta}}_{nT}' - \left(\frac{\bar{\delta}_n}{1+\bar{\delta}_n}\right)\boldsymbol{\theta}_0\bar{\boldsymbol{\zeta}}_{nT}' - \left(\frac{\bar{\delta}_n}{1+\bar{\delta}_n}\right)\bar{\boldsymbol{\zeta}}_{nT}\boldsymbol{\theta}_0'.
\end{aligned} \tag{A.2.5}$$

By the results in Lemma A.1, $E\left(\bar{\delta}_n\right) = O(a_n)$, and $E\left(\bar{\boldsymbol{\zeta}}_{nT}\right) = E\left(\delta_i \boldsymbol{\eta}_i\right) = O(a_n)$, $\bar{\delta}_n = O(a_n) + o_p(1)$ and $n^{-1}\sum_{i=1}^n \delta_i^2 = O(a_n) + o_p(1)$. Also using (5.11) we have

$$\bar{\boldsymbol{\zeta}}_{nT} = O(n^{-\alpha}) + O_p\left(n^{-\frac{(1-\alpha)}{2}}\right),$$

and since $\alpha > 1/3$ using (5.10),

$$\left[\left(\frac{1}{1+\bar{\delta}_n}\right)^2 - 2\left(\frac{1}{1+\bar{\delta}_n}\right)\right]\bar{\boldsymbol{\zeta}}_{nT}\bar{\boldsymbol{\zeta}}_{nT}' = O_p\left(n^{-2\alpha}\right) + O_p\left(n^{-(1-\alpha)}\right) = O_p\left(n^{-(1-\alpha)}\right), \tag{A.2.6}$$

$$\left(\frac{\bar{\delta}_n}{1+\bar{\delta}_n}\right)\bar{\boldsymbol{\zeta}}_{nT}\boldsymbol{\theta}_0' = O(a_n n^{-\alpha}) + O_p\left(a_n n^{-\frac{(1-\alpha)}{2}}\right) = O_p\left(n^{-\frac{(1+\alpha)}{2}}\right). \tag{A.2.7}$$

Consider now

$$\bar{\boldsymbol{\zeta}}_{\delta,nT} = n^{-1}\sum_{i=1}^{n}\delta_i\boldsymbol{\zeta}_{iT} = n^{-1}\sum_{i=1}^{n}\delta_i\left(1+\delta_i\right)\boldsymbol{\eta}_i + n^{-1}\sum_{i=1}^{n}\delta_i\left(1+\delta_i\right)\boldsymbol{\xi}_{iT}. \tag{A.2.8}$$

By (A.1.3) in Lemma A.1, $E\left[\delta_i\left(1+\delta_i\right)\boldsymbol{\eta}_i\right] = O(a_n)$, and since $\delta_i\left(1+\delta_i\right)\boldsymbol{\eta}_i$ are distributed independently over i we have

$$n^{-1}\sum_{i=1}^{n}\delta_i\left(1+\delta_i\right)\boldsymbol{\eta}_i = O_p(a_n). \tag{A.2.9}$$

Since conditional on $\boldsymbol{W}_i$, $\delta_i\left(1+\delta_i\right)\boldsymbol{\xi}_{iT}$ are distributed over i with zero means, then following the same line of argument as in the proof of Lemma A.2, we have $E\left[\delta_i\left(1+\delta_i\right)\boldsymbol{\xi}_{iT}\right] = 0$ and

$$\begin{aligned} Var\left[n^{-1}\sum_{i=1}^{n}\delta_i\left(1+\delta_i\right)\boldsymbol{\xi}_{iT}\right] &= n^{-2}\sum_{i=1}^{n}E\left[\delta_i^2\left(1+\delta_i\right)^2\boldsymbol{R}_i'\boldsymbol{H}_i\boldsymbol{R}_i\right] \\ &\leq Cn^{-2}\sum_{i=1}^{n}E\left[\delta_i^2\left(1+\delta_i\right)^2 d_i^{-1}\left\|(\boldsymbol{W}_i'\boldsymbol{W}_i)^*\right\|\right]. \end{aligned}$$

Further using (A.1.16)

$$\begin{aligned} \left(1+\delta_i\right)^2\delta_i^2 &= \left[\mathbf{1}\{d_i > a_n\} + a_n^{-2}d_i^2\mathbf{1}\{d_i \leq a_n\}\right]\left(\frac{d_i - a_n}{a_n}\right)^2\mathbf{1}\{d_i \leq a_n\} \\ &= a_n^{-2}d_i^2\left(\frac{d_i - a_n}{a_n}\right)^2\mathbf{1}\{d_i \leq a_n\}, \end{aligned}$$

and

$$Var\left[n^{-1}\sum_{i=1}^{n}\delta_i\left(1+\delta_i\right)\boldsymbol{\xi}_{iT}\right] \leq Cn^{-2}\sum_{i=1}^{n}E\left[a_n^{-2}d_i\left(\frac{d_i - a_n}{a_n}\right)^2\mathbf{1}\{d_i \leq a_n\}\left\|(\boldsymbol{W}_i'\boldsymbol{W}_i)^*\right\|\right].$$

By Cauchy-Schwarz inequality

$$\begin{aligned} &E\left[a_n^{-2}d_i\left(\frac{d_i - a_n}{a_n}\right)^2\mathbf{1}\{d_i \leq a_n\}\left\|(\boldsymbol{W}_i'\boldsymbol{W}_i)^*\right\|\right] \\ \leq\ & a_n^{-2}\left\{E\left[d_i^2\left(\frac{d_i - a_n}{a_n}\right)^4\mathbf{1}\{d_i \leq a_n\}\right]\right\}^{1/2}\left[E\left\|(\boldsymbol{W}_i'\boldsymbol{W}_i)^*\right\|^2\right]^{1/2}, \end{aligned}$$

and since under Assumption 4, $\sup_i E\left\|(\boldsymbol{W}_i'\boldsymbol{W}_i)^*\right\|^2 < C$, we have

$$Var\left[n^{-1}\sum_{i=1}^{n}\delta_i\left(1+\delta_i\right)\boldsymbol{\xi}_{iT}\right] \leq Cn^{-2}a_n^{-2}\sum_{i=1}^{n}\left\{E\left[d_i^2\left(\frac{d_i-a_n}{a_n}\right)^4\mathbf{1}\{d_i\leq a_n\}\right]\right\}^{1/2}.$$

Also using (A.1.1) of Lemma A.1

$$E\left[d_i^2\left(\frac{d_i-a_n}{a_n}\right)^4\mathbf{1}\{d_i\leq a_n\}\right] = a_n^{-4}E\left[\left(d_i^6-3d_i^5a_n+3a_n^3d_i^3-a_n^4d_i^2\right)\mathbf{1}\{d_i\leq a_n\}\right] = O\left(a_n^3\right),$$

which yields

$$Var\left[n^{-1}\sum_{i=1}^{n}\delta_i\left(1+\delta_i\right)\boldsymbol{\xi}_{iT}\right] = O\left(n^{-1}a_n^{-2}a_n^{3/2}\right) = O\left(n^{-1}a_n^{-1/2}\right),$$

and by Markov inequality

$$n^{-1}\sum_{i=1}^{n}\delta_i\left(1+\delta_i\right)\boldsymbol{\xi}_{iT} = O\left(n^{-1/2}a_n^{-1/4}\right) = O\left(n^{-1/2+\alpha/4}\right). \tag{A.2.10}$$

Using (A.2.9) and (A.2.10) in (A.2.8), we have $\bar{\boldsymbol{\zeta}}_{\delta,nT} = O_p(n^{-\alpha}) + O\left(n^{-1/2+\alpha/4}\right)$, which if used with (A.2.6) and (A.2.7) in (A.2.5) now yields (for $\alpha > 1/3$)

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(\tilde{\boldsymbol{\theta}}_i-\hat{\boldsymbol{\theta}}_{TMG})(\tilde{\boldsymbol{\theta}}_i-\hat{\boldsymbol{\theta}}_{TMG})' =& \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\zeta}_{iT}\boldsymbol{\zeta}_{iT}' \\ &+ O_p\left(n^{-(1-\alpha)}\right)+O_p\left(n^{-\frac{(1+\alpha)}{2}}\right)+O(n^{-\alpha})+O\left(n^{-1/2+\alpha/4}\right),\end{aligned}$$

and since $\boldsymbol{\zeta}_{iT}$ are independently distributed over i, we have

$$\operatorname*{plim}_{n\to\infty} n^{-1}\sum_{i=1}^{n}\left(\tilde{\boldsymbol{\theta}}_i-\hat{\boldsymbol{\theta}}_{TMG}\right)\left(\tilde{\boldsymbol{\theta}}_i-\hat{\boldsymbol{\theta}}_{TMG}\right)' = \lim_{n\to\infty}\left[n^{-1}\sum_{i=1}^{n}E\left(\boldsymbol{\zeta}_{iT}\boldsymbol{\zeta}_{iT}'\right)\right].$$

But using (A.2.4) and recalling that $\bar{\delta}_n = O(a_n)$ then

$$\lim_{n\to\infty} nVar\left(\hat{\boldsymbol{\theta}}_{TMG}\right) = \lim_{n\to\infty} nVar\left(\bar{\boldsymbol{\zeta}}_{nT}\right).$$

Also (recall that $E\left(\bar{\boldsymbol{\zeta}}_{nT}\right) = O(a_n)$)

$$nVar\left(\bar{\boldsymbol{\zeta}}_{nT}\right) = E\left\{n^{-1}\sum_{i=1}^{n}\left[\boldsymbol{\zeta}_{iT} - E\left(\boldsymbol{\zeta}_{iT}\right)\right]\left[\boldsymbol{\zeta}_{iT} - E\left(\boldsymbol{\zeta}_{iT}\right)\right]'\right\} = n^{-1}\sum_{i=1}^{n} E\left(\boldsymbol{\zeta}_{iT}\boldsymbol{\zeta}_{iT}'\right) + O(a_n^2).$$

Hence

$$\lim_{n\to\infty} nVar\left(\hat{\boldsymbol{\theta}}_{TMG}\right) = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} E\left(\boldsymbol{\zeta}_{iT}\boldsymbol{\zeta}_{iT}'\right) = \operatorname*{plim}_{n\to\infty} n^{-1}\sum_{i=1}^{n}\left(\tilde{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_{TMG}\right)\left(\tilde{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_{TMG}\right)',$$

and $n^{-1}\sum_{i=1}^{n}\left(\tilde{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_{TMG}\right)\left(\tilde{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_{TMG}\right)'$ is a consistent estimator of $nVar\left(\hat{\boldsymbol{\theta}}_{TMG}\right)$. ■

A.3 Proof of Theorem 3 (Asymptotic distribution of the TMG-TE estimator)

Proof. Initially, we consider the case where $T \geq k$. To derive the asymptotic distribution of $\hat{\boldsymbol{\theta}}_{TMG-TE}$ we first note that $\hat{\boldsymbol{\theta}}_{TMG-TE}(\boldsymbol{\phi}) = \hat{\boldsymbol{\theta}}_{TMG} - \bar{\boldsymbol{Q}}_n'\boldsymbol{\phi}$, and $\hat{\boldsymbol{\theta}}_{TMG-TE} = \hat{\boldsymbol{\theta}}_{TMG} - \bar{\boldsymbol{Q}}_n'\hat{\boldsymbol{\phi}}$. Hence

$$\left(\hat{\boldsymbol{\theta}}_{TMG-TE} - \boldsymbol{\theta}_0\right) - \left(\hat{\boldsymbol{\theta}}_{TMG-TE}(\boldsymbol{\phi}) - \boldsymbol{\theta}_0\right) = -\bar{\boldsymbol{Q}}_n'\left(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}\right). \tag{A.3.1}$$

Also stacking (6.7) over t and subtracting the results from (6.10) yields

$$\hat{\boldsymbol{\phi}} - \boldsymbol{\phi} = -\boldsymbol{M}_T\bar{\boldsymbol{W}}\left(\hat{\boldsymbol{\theta}}_{TMG-TE} - \boldsymbol{\theta}_0\right) + \boldsymbol{M}_T\bar{\boldsymbol{\nu}}, \tag{A.3.2}$$

where $\bar{\boldsymbol{\nu}} = n^{-1}\sum_{i=1}^{n}\boldsymbol{\nu}_i$, and $\boldsymbol{\nu}_i = (\nu_{i1}, \nu_{i2}, ..., \nu_{iT})'$ with $\nu_{it} = u_{it} + \boldsymbol{x}_{it}'\boldsymbol{\eta}_{\beta_i}$. Using this result in (A.3.1) we have

$$\left(\boldsymbol{I}_k - \bar{\boldsymbol{Q}}_n'\boldsymbol{M}_T\bar{\boldsymbol{W}}\right)\left(\hat{\boldsymbol{\theta}}_{TMG-TE} - \boldsymbol{\theta}_0\right) = \left(\hat{\boldsymbol{\theta}}_{TMG-TE}(\boldsymbol{\phi}) - \boldsymbol{\theta}_0\right) - \bar{\boldsymbol{Q}}_n'\boldsymbol{M}_T\bar{\boldsymbol{\nu}}.$$

For a known value of $\boldsymbol{\phi}$, the asymptotic distribution of $\left(\hat{\boldsymbol{\theta}}_{TMG-TE}(\boldsymbol{\phi}) - \boldsymbol{\theta}_0\right)$ is the same as $\hat{\boldsymbol{\theta}}_{TMG}$ with $\boldsymbol{y}_i$ replaced by $\boldsymbol{y}_i - \boldsymbol{\phi}$. Under the assumption that $\boldsymbol{I}_k - \bar{\boldsymbol{Q}}_n'\boldsymbol{M}_T\bar{\boldsymbol{W}}$ is invertible, we have

$$\hat{\boldsymbol{\theta}}_{TMG-TE} - \boldsymbol{\theta}_0 = \left(\boldsymbol{I}_k - \bar{\boldsymbol{Q}}_n'\boldsymbol{M}_T\bar{\boldsymbol{W}}\right)^{-1}\left(\hat{\boldsymbol{\theta}}_{TMG-TE}(\boldsymbol{\phi}) - \boldsymbol{\theta}_0\right) - \left(\boldsymbol{I}_k - \bar{\boldsymbol{Q}}_n'\boldsymbol{M}_T\bar{\boldsymbol{W}}\right)^{-1}\bar{\boldsymbol{Q}}_n'\boldsymbol{M}_T\bar{\boldsymbol{\nu}}.$$

Hence using Lemma A.3, $\bar{\boldsymbol{\nu}} = O_p\left(n^{-1/2}\right)$, and we have

$$n^{(1-\alpha)/2}\left(\hat{\boldsymbol{\theta}}_{TMG-TE} - \boldsymbol{\theta}_0\right) = \left(\boldsymbol{I}_k - \bar{\boldsymbol{Q}}_n'\boldsymbol{M}_T\bar{\boldsymbol{W}}\right)^{-1}\left[n^{(1-\alpha)/2}\left(\hat{\boldsymbol{\theta}}_{TMG-TE}(\boldsymbol{\phi}) - \boldsymbol{\theta}_0\right)\right] + O_p(n^{-\alpha/2}),$$

where for a known $\boldsymbol{\phi}$ we have already established in Theorem 1 that

$$n^{(1-\alpha)/2}\left(\hat{\boldsymbol{\theta}}_{TMG-TE}(\boldsymbol{\phi})-\boldsymbol{\theta}_0\right)\rightarrow_d N\left(\mathbf{0},\boldsymbol{V}_{\boldsymbol{\theta}}(\boldsymbol{\phi})\right),$$

with $\boldsymbol{V}_{\theta}(\boldsymbol{\phi})=\lim_{n\rightarrow\infty}Var\left[n^{(1-\alpha)/2}\hat{\boldsymbol{\theta}}_{TMG-TE}(\boldsymbol{\phi})\right]$. Suppose further that $\text{plim}_{n\rightarrow\infty}\left(\bar{\boldsymbol{Q}}_n'\boldsymbol{M}_T\bar{\boldsymbol{W}}\right)$ $=\boldsymbol{G}_w$, where $\boldsymbol{I}_k-\boldsymbol{G}_w$ is non-singular. For $\alpha>1/3$, we have $n^{(1-\alpha)/2}\left(\hat{\boldsymbol{\theta}}_{TMG-TE}-\boldsymbol{\theta}_0\right)\rightarrow_d$ $N\left(\mathbf{0},\boldsymbol{V}_{\theta,TMG-TE}\right)$, where

$$\boldsymbol{V}_{\theta,TMG-TE}=\left(\boldsymbol{I}_k-\boldsymbol{G}_w\right)^{-1}\boldsymbol{V}_{\theta}(\boldsymbol{\phi})\left[\left(\boldsymbol{I}_k-\boldsymbol{G}_w\right)^{-1}\right]'. \tag{A.3.3}$$

A consistent estimator of the asymptotic variance of $\hat{\boldsymbol{\theta}}_{TMG-TE}$ is given by

$$\widehat{Var(\hat{\boldsymbol{\theta}}_{TMG-TE})}=\frac{1}{n-1}\left(\boldsymbol{I}_k-\bar{\boldsymbol{Q}}_n'\boldsymbol{M}_T\bar{\boldsymbol{W}}\right)^{-1}\widehat{\boldsymbol{V}}_{\theta}\left[\left(\boldsymbol{I}_k-\bar{\boldsymbol{Q}}_n'\boldsymbol{M}_T\bar{\boldsymbol{W}}\right)^{-1}\right]', \tag{A.3.4}$$

where

$$\widehat{\boldsymbol{V}}_{\theta}=\frac{1}{(n-1)(1+\bar{\delta}_n)^2}\sum_{i=1}^{n}(\tilde{\boldsymbol{\theta}}_i-\boldsymbol{Q}_i'\hat{\boldsymbol{\phi}}-\hat{\boldsymbol{\theta}}_{TMG-TE})(\tilde{\boldsymbol{\theta}}_i-\boldsymbol{Q}_i'\hat{\boldsymbol{\phi}}-\hat{\boldsymbol{\theta}}_{TMG-TE})', \tag{A.3.5}$$

Consider now the asymptotic distribution of $\hat{\boldsymbol{\phi}}$. Using (A.3.2) and noting that $\boldsymbol{M}_T\bar{\boldsymbol{W}}\left(\hat{\boldsymbol{\theta}}_{TMG-TE}-\boldsymbol{\theta}_0\right)=\boldsymbol{M}_T\bar{\boldsymbol{X}}\left(\hat{\boldsymbol{\beta}}_{TMG-TE}-\boldsymbol{\beta}_0\right)$, we have

$$\hat{\boldsymbol{\phi}}-\boldsymbol{\phi}_0=-\boldsymbol{M}_T\bar{\boldsymbol{X}}\left(\hat{\boldsymbol{\beta}}_{TMG-TE}-\boldsymbol{\beta}_0\right)+\boldsymbol{M}_T\bar{\boldsymbol{\nu}}.$$

Two cases can arise depending on whether the probability limit of $\boldsymbol{M}_T\bar{\boldsymbol{X}}$ tends to zero as $n\rightarrow\infty$, or not. Under (a) $\text{plim}_{n\rightarrow\infty}\boldsymbol{M}_T\bar{\boldsymbol{X}}=\mathbf{0}$, we have $n^{1/2}(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi}_0)\rightarrow_d N(\mathbf{0},\boldsymbol{M}_T\boldsymbol{\Omega}_{\nu}\boldsymbol{M}_T)$, where $\boldsymbol{\Omega}_{\nu}$ is given by (A.1.22), namely $\hat{\boldsymbol{\phi}}\rightarrow_p\boldsymbol{\phi}_0$ at the regular rate of $n^{-1/2}$. Also since

$$\nu_{it}-\bar{\nu}_{i\circ}=(u_{it}-\bar{u}_{i\circ})+(\boldsymbol{x}_{it}-\bar{\boldsymbol{x}}_{i\circ})'\boldsymbol{\eta}_{i\beta}=y_{it}-\bar{y}_{i\circ}-(\boldsymbol{x}_{it}-\boldsymbol{x}_{i\circ})'\boldsymbol{\beta}-\phi_t,$$

$\boldsymbol{\Omega}_{\nu}$ can be consistently estimated by

$$\widehat{\boldsymbol{\Omega}}_{\nu}=\frac{1}{n-1}\sum_{i=1}^{n}\left(\boldsymbol{y}_i-\boldsymbol{X}_i\hat{\boldsymbol{\beta}}_{TMG-TE}-\hat{\boldsymbol{\phi}}\right)\left(\boldsymbol{y}_i-\boldsymbol{X}_i\hat{\boldsymbol{\beta}}_{TMG-TE}-\hat{\boldsymbol{\phi}}\right)'. \tag{A.3.6}$$

Under case (b), $\text{plim}_{n\rightarrow\infty}\boldsymbol{M}_T\bar{\boldsymbol{X}}\neq\mathbf{0}$, and convergence of $\hat{\boldsymbol{\phi}}$ to $\boldsymbol{\phi}_0$ cannot achieve the regular rate. To see this note that

$$n^{(1-\alpha)/2}\left(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi}_0\right)=-\boldsymbol{M}_T\bar{\boldsymbol{X}}\left[n^{(1-\alpha)/2}\left(\hat{\boldsymbol{\beta}}_{TMG-TE}-\boldsymbol{\beta}_0\right)\right]+n^{-\alpha/2}\boldsymbol{M}_T\left(n^{1/2}\bar{\boldsymbol{\nu}}\right),$$

where $\boldsymbol{M}_T\left(n^{1/2}\bar{\boldsymbol{\nu}}\right)=O_p(1)$ and and since $\alpha>0$ the second term tends to zero, but rather slowly. In practice, where it is not known whether $\boldsymbol{M}_T\bar{\boldsymbol{X}}\rightarrow\mathbf{0}$ or not, one can consistently estimate the asymptotic variance of $\hat{\boldsymbol{\phi}}$ by

$$\widehat{Var\left(\hat{\boldsymbol{\phi}}\right)}=\boldsymbol{M}_T\left[\bar{\boldsymbol{X}}\widehat{Var\left(\hat{\boldsymbol{\beta}}_{TMG-TE}\right)}\bar{\boldsymbol{X}}'+n^{-1}\widehat{\boldsymbol{\Omega}}_\nu\right]\boldsymbol{M}_T, \tag{A.3.7}$$

where $\widehat{Var\left(\hat{\boldsymbol{\beta}}_{TMG-TE}\right)}$ and $\widehat{\boldsymbol{\Omega}}_\nu$ are given by (A.3.4) and (A.3.6), respectively. Note that $\widehat{Var\left(\hat{\boldsymbol{\phi}}\right)}$ is singular as $\widehat{Var\left(\hat{\boldsymbol{\phi}}\right)}\boldsymbol{\tau}_T=\mathbf{0}$, but its diagonal elements can be used to test if $\hat{\phi}_t$ for $t=1,2,..,T$ are individually or jointly statistically significant subject to $\boldsymbol{\phi}'\boldsymbol{\tau}_T=0$. ■

Online Supplement to "Trimmed Mean Group Estimation of Average Treatment Effects in Ultra Short T Panels with Correlated Heterogeneous Coefficients"

M. Hashem Pesaran
University of Southern California, and Trinity College, Cambridge

Liying Yang
Postdoctoral research fellow, Sauder School of Business, University of British Columbia

October 17, 2023

S.1 Introduction

This online supplement is structured as follows. Section S.2 derives the Hausman-type test of correlated heterogeneity in panel data models with time effects, with corresponding Monte Carlo (MC) evidence provided in Section S.9. Section S.3 provides details of the MC design. Sections S.4 and S.6 summarize MC results for the TMG estimator with Gaussian and uniformly distributed errors in the regressor (x_{it}) process, respectively. Section S.4.2 provides MC evidence for the TMG and GP estimators using different α and α_{GP}, exponents in the threshold values. Section S.5 investigates the robustness of the TMG estimator to a number of variations on the baseline DGP. Section S.7 shows and discusses the MC results when there is an interactive effect in the regressor process. Section S.8 presents the empirical power functions for the baseline model with correlated heterogeneity and time effects.

S.2 The Hausman-type test of correlated heterogeneity with time effects

Given the panel data model with time effects in (6.1), a Hausman-type test can be constructed based on the difference between the TMG and FE-TE estimators when $T \geq k$. The FE-TE estimator is given by

$$\hat{\boldsymbol{\beta}}_{FE-TE} = \bar{\boldsymbol{\Psi}}_{n,TE}^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)'\boldsymbol{M}_T\left(\boldsymbol{y}_i - \bar{\boldsymbol{y}}\right)\right], \tag{S.2.1}$$

and

$$\hat{\boldsymbol{\phi}}_{FE-TE} = \boldsymbol{M}_T(\bar{\boldsymbol{y}} - \bar{\boldsymbol{X}}\hat{\boldsymbol{\beta}}_{FE-TE}), \tag{S.2.2}$$

where

$$\bar{\boldsymbol{\Psi}}_{n,TE} = \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)'\boldsymbol{M}_T\left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right).$$

Then,

$$\hat{\boldsymbol{\beta}}_{FE-TE} - \boldsymbol{\beta}_0 = \bar{\boldsymbol{\Psi}}_{n,TE}^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)'\boldsymbol{M}_T\tilde{\boldsymbol{\nu}}_i\right], \tag{S.2.3}$$

where $\tilde{\boldsymbol{\nu}}_i = \boldsymbol{\nu}_i - \bar{\boldsymbol{\nu}}$, $\bar{\boldsymbol{\nu}} = n^{-1}\sum_{i=1}^{n}\boldsymbol{\nu}_i$, and $\boldsymbol{\nu}_i = \boldsymbol{u}_i + \boldsymbol{X}_i\boldsymbol{\eta}_{i\beta}$. We derive the test statistics under the null hypothesis in (7.1), for two cases: (a) when $T \geq k$ the TMG-TE estimator in (6.12) is used, and (b) the TMG-C estimator in (6.20) is used when $T > k$. The implicit null is given by $n^{-1/2}\sum_{i=1}^{n} E\left[\left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)'\boldsymbol{M}_T\left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)\boldsymbol{\eta}_{i\beta}\right] \to \boldsymbol{0}$, which is implied by (7.1) but not *vice versa*. We make the following assumption that corresponds to the pooling Assumption 3:

Assumption S.1 (FE-TE pooling assumption) *Let* $\bar{\boldsymbol{\Psi}}_{n,TE} = n^{-1}\sum_{i=1}^{n}\boldsymbol{\Psi}_{i,TE}$, *where* $\boldsymbol{\Psi}_{i,TE} = \left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)'\boldsymbol{M}_T\left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)$. *For* $T \geq k$, *there exists* n_0 *such that for all* $n > n_0$, $\bar{\boldsymbol{\Psi}}_{n,TE}$ *is positive definite,*

$$\bar{\boldsymbol{\Psi}}_{n,TE} \to_p \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} E\left(\boldsymbol{\Psi}_{i,TE}\right) = \bar{\boldsymbol{\Psi}}_{TE} \succ \mathbf{0}, \tag{S.2.4}$$

and

$$\bar{\boldsymbol{\Psi}}_{n,TE}^{-1} = \bar{\boldsymbol{\Psi}}_{TE}^{-1} + o_p(1). \tag{S.2.5}$$

S.2.1 $T \geq k$

When $T \geq k$, consider

$$\hat{\boldsymbol{\Delta}}_{\beta,TE} = \hat{\boldsymbol{\beta}}_{FE-TE} - \hat{\boldsymbol{\beta}}_{TMG-TE},$$

where $\hat{\boldsymbol{\beta}}_{FE-TE}$ and $\hat{\boldsymbol{\beta}}_{TMG-TE}$ are given by (S.2.1) and (6.12), respectively. Given (A.3.1), (A.3.2) and $\hat{\boldsymbol{\phi}} = \boldsymbol{M}_T(\bar{\boldsymbol{y}} - \bar{\boldsymbol{X}}\hat{\boldsymbol{\beta}}_{TMG-TE})$, we have

$$\hat{\boldsymbol{\beta}}_{TMG-TE} - \boldsymbol{\beta}_0 = (\boldsymbol{I}_{k'} - \bar{\boldsymbol{Q}}'_{nx}\boldsymbol{M}_T\bar{\boldsymbol{X}})^{-1}\left[\frac{1}{n(1+\bar{\delta}_n)}\sum_{i=1}^{n}\boldsymbol{Q}'_{ix}\boldsymbol{M}_T\tilde{\boldsymbol{\nu}}_i\right], \tag{S.2.6}$$

where δ_i is given by (4.3), and partitioning $\boldsymbol{Q}_i$ conformably with $\boldsymbol{W}_i = (\boldsymbol{\tau}_T, \boldsymbol{X}_i)$ we have

$$\boldsymbol{Q}_{ix} = (1+\delta_i)\boldsymbol{M}_T\boldsymbol{X}_i(\boldsymbol{X}'_i\boldsymbol{M}_T\boldsymbol{X}_i)^{-1}, \tag{S.2.7}$$

and

$$\bar{\boldsymbol{Q}}_{nx} = \frac{1}{n(1+\bar{\delta}_n)}\sum_{i=1}^{n}\boldsymbol{Q}_{ix}. \tag{S.2.8}$$

Using (S.2.3) and under Assumption S.1

$$\hat{\boldsymbol{\beta}}_{FE-TE} - \boldsymbol{\beta}_0 = \bar{\boldsymbol{\Psi}}_{TE}^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)'\boldsymbol{M}_T\tilde{\boldsymbol{\nu}}_i\right] + o_p(1), \tag{S.2.9}$$

which in conjunction with (S.2.6) yields

$$\hat{\boldsymbol{\Delta}}_{\beta,TE} = \hat{\boldsymbol{\beta}}_{FE-TE} - \hat{\boldsymbol{\beta}}_{TMG-TE} = n^{-1}\sum_{i=1}^{n}\boldsymbol{G}'_{i,TE}\boldsymbol{M}_T\tilde{\boldsymbol{\nu}}_i,$$

where $\boldsymbol{G}_{i,TE}$ is a $T \times k'$ matrix given by

$$\boldsymbol{G}_{i,TE} = \left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right) \bar{\boldsymbol{\Psi}}_{TE}^{-1} - (1+\bar{\delta}_n)^{-1}\boldsymbol{Q}_{ix} \left[\left(\boldsymbol{I}_{k'} - \bar{\boldsymbol{Q}}_{nx}'\boldsymbol{M}_T\bar{\boldsymbol{X}}\right)^{-1}\right]'.$$

Under Assumption 1 and the null hypothesis given by (7.1), $E(\tilde{\nu}_{it}|\boldsymbol{G}_{i,TE}) = 0$ for all i and t. Also by Assumptions 1 and 6, u_{it} and $\boldsymbol{\eta}_{i\beta}$ are cross-sectionally independent so that $\tilde{\nu}_{it}$ conditional on $\boldsymbol{X}_i$ are also cross-sectionally independent. Then as $n \to \infty$,

$$\sqrt{n}\hat{\boldsymbol{\Delta}}_{\beta,TE} \to_d N(\boldsymbol{0}, \boldsymbol{V}_{\Delta,TE}),$$

as long as

$$\boldsymbol{V}_{\Delta,TE} = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} E\left(\boldsymbol{G}_{i,TE}'\boldsymbol{M}_T\tilde{\boldsymbol{\nu}}_i\tilde{\boldsymbol{\nu}}_i'\boldsymbol{M}_T\boldsymbol{G}_{i,TE}\right) \succ \boldsymbol{0}. \tag{S.2.10}$$

Hence when $T \geq k$, the Hausman-type test for panels with time effects is given by

$$H_{\beta,TE} = n\hat{\boldsymbol{\Delta}}_{\beta,TE}'\boldsymbol{V}_{\Delta,TE}^{-1}\hat{\boldsymbol{\Delta}}_{\beta,TE}, \tag{S.2.11}$$

and as $n \to \infty$, $H_{\beta,TE} \to_d \chi^2_{k'}$. For fixed T, $\boldsymbol{V}_{\Delta,TE}$ can be consistently estimated by

$$\widehat{\boldsymbol{V}}_{\Delta,TE} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\boldsymbol{G}}_{i,TE}'\boldsymbol{M}_T\hat{\tilde{\boldsymbol{\nu}}}_{i,FE}\hat{\tilde{\boldsymbol{\nu}}}_{i,FE}'\boldsymbol{M}_T\hat{\boldsymbol{G}}_{i,TE}\right), \tag{S.2.12}$$

where

$$\hat{\tilde{\boldsymbol{\nu}}}_{i,FE} = \widehat{\boldsymbol{\nu}_i - \bar{\boldsymbol{\nu}}} = (\boldsymbol{y}_i - \bar{\boldsymbol{y}}) - (\boldsymbol{X}_i - \bar{\boldsymbol{X}})\hat{\boldsymbol{\beta}}_{FE-TE}, \tag{S.2.13}$$

and

$$\hat{\boldsymbol{G}}_{i,TE} = \left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right) \bar{\boldsymbol{\Psi}}_{n,TE}^{-1} - (1+\bar{\delta}_n)^{-1}\boldsymbol{Q}_{ix} \left[\left(\boldsymbol{I}_{k'} - \bar{\boldsymbol{Q}}_{nx}'\boldsymbol{M}_T\bar{\boldsymbol{X}}\right)^{-1}\right]'. \tag{S.2.14}$$

Using the above estimate of $\widehat{\boldsymbol{V}}_{\Delta,TE}$, a feasible statistic for testing the null hypothesis, H_0 given by (7.1), (in the case of panel regression models with time effects and $T \geq k$) is given by

$$\hat{H}_{\beta,TE} = n\left(\hat{\boldsymbol{\beta}}_{FE-TE} - \hat{\boldsymbol{\beta}}_{TMG-TE}\right)'\widehat{\boldsymbol{V}}_{\Delta,TE}^{-1}\left(\hat{\boldsymbol{\beta}}_{FE-TE} - \hat{\boldsymbol{\beta}}_{TMG-TE}\right). \tag{S.2.15}$$

S.2.2 $T > k$

For panels with $T > k$, we consider

$$\hat{\boldsymbol{\Delta}}_{\beta,C} = \hat{\boldsymbol{\beta}}_{FE-TE} - \hat{\boldsymbol{\beta}}_{TMG-C},$$

where $\hat{\boldsymbol{\beta}}_{FE-TE}$ is given by (S.2.1), and $\hat{\boldsymbol{\beta}}_{TMG-C}$ is the TMG-C estimator based on $\hat{\boldsymbol{\phi}}_C$ as the estimator of the time effects given by (6.20) and (6.17), respectively. Using (6.18) and noting that $\boldsymbol{M}_i \boldsymbol{M}_T \boldsymbol{X}_i = \boldsymbol{0}$ we have

$$\hat{\boldsymbol{\phi}}_C - \boldsymbol{\phi} = \bar{\boldsymbol{M}}_n^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{M}_i \boldsymbol{M}_T \boldsymbol{\nu}_i \right) \tag{S.2.16}$$

where $\boldsymbol{\nu}_i = \boldsymbol{X}_i \boldsymbol{\eta}_{i\beta} + \boldsymbol{u}_i$, and $\bar{\boldsymbol{M}}_n = n^{-1} \sum_{i=1}^{n} \boldsymbol{M}_i$. Using (6.20) and partitioning $\hat{\boldsymbol{\theta}}_{TMG-C} = (\hat{\alpha}_{TMG-C}, \hat{\boldsymbol{\beta}}'_{TMG-C})'$, we have

$$\hat{\boldsymbol{\beta}}_{TMG-C} = \frac{1}{1+\bar{\delta}_n} \left[n^{-1} \sum_{i=1}^{n} \boldsymbol{Q}'_{ix} \boldsymbol{M}_T (\boldsymbol{y}_i - \hat{\boldsymbol{\phi}}_C) \right],$$

where $\boldsymbol{Q}_{ix}$ is defined by (S.2.7). Also, since $\boldsymbol{y}_i = \alpha_i \boldsymbol{\tau}_T + \boldsymbol{\phi} + \boldsymbol{X}_i \boldsymbol{\beta}_0 + \boldsymbol{\nu}_i$, then noting that $n^{-1} \sum_{i=1}^{n} \left(1+\bar{\delta}_n\right)^{-1} \boldsymbol{Q}'_{ix} \boldsymbol{M}_T \boldsymbol{X}_i = \boldsymbol{I}_{k'}$, we have

$$\hat{\boldsymbol{\beta}}_{TMG-C} - \boldsymbol{\beta}_0 = \frac{1}{1+\bar{\delta}_n} \left[n^{-1} \sum_{i=1}^{n} \boldsymbol{Q}'_{ix} \boldsymbol{M}_T (\boldsymbol{\nu}_i + \boldsymbol{\phi} - \hat{\boldsymbol{\phi}}_C) \right].$$

Also using (S.2.16),

$$\frac{1}{n(1+\bar{\delta}_n)} \sum_{i=1}^{n} \boldsymbol{Q}'_{ix} \boldsymbol{M}_T \left(\hat{\boldsymbol{\phi}}_C - \boldsymbol{\phi} \right) = \bar{\boldsymbol{Q}}'_{nx} \boldsymbol{M}_T \bar{\boldsymbol{M}}_n^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{M}_i \boldsymbol{M}_T \boldsymbol{\nu}_i \right)$$

where $\bar{\boldsymbol{Q}}_{nx}$ is given by (S.2.8). Hence

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{TMG-C} - \boldsymbol{\beta}_0 &= \frac{1}{1+\bar{\delta}_n} n^{-1} \sum_{i=1}^{n} \boldsymbol{Q}'_{ix} \boldsymbol{M}_T \boldsymbol{\nu}_i - \bar{\boldsymbol{Q}}'_{nx} \boldsymbol{M}_T \bar{\boldsymbol{M}}_n^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{M}_i \boldsymbol{M}_T \boldsymbol{\nu}_i \right) \\ &= n^{-1} \sum_{i=1}^{n} \left[(1+\bar{\delta}_n)^{-1} \boldsymbol{Q}'_{ix} - \bar{\boldsymbol{Q}}'_{nx} \boldsymbol{M}_T \bar{\boldsymbol{M}}_n^{-1} \boldsymbol{M}_i \right] \boldsymbol{M}_T \boldsymbol{\nu}_i, \end{aligned}$$

or equivalently in terms of $\tilde{\boldsymbol{\nu}}_i = \boldsymbol{\nu}_i - \bar{\boldsymbol{\nu}}$,

$$\hat{\boldsymbol{\beta}}_{TMG-C} - \boldsymbol{\beta}_0 = n^{-1} \sum_{i=1}^{n} \left[(1+\bar{\delta}_n)^{-1} \boldsymbol{Q}_{ix}' - \bar{\boldsymbol{Q}}'_{nx} \boldsymbol{M}_T \bar{\boldsymbol{M}}_n^{-1} \boldsymbol{M}_i \right] \boldsymbol{M}_T \tilde{\boldsymbol{\nu}}_i,$$

since $\frac{1}{n}\sum_{i=1}^{n} \left[(1+\bar{\delta}_n)^{-1} \boldsymbol{Q}_{ix}' - \bar{\boldsymbol{Q}}_{nx}' \boldsymbol{M}_T \bar{\boldsymbol{M}}_n^{-1} \boldsymbol{M}_i \right] \boldsymbol{M}_T \tilde{\boldsymbol{\nu}} = \mathbf{0}$ given $\frac{1}{n}\sum_{i=1}^{n} \bar{\boldsymbol{M}}_n^{-1} \boldsymbol{M}_i = \boldsymbol{I}_T$. Using this result together with (S.2.9), we now have

$$\hat{\boldsymbol{\Delta}}_{\beta,C} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{G'}_{i,C} \boldsymbol{M}_T \tilde{\boldsymbol{\nu}}_i,$$

where

$$\boldsymbol{G}_{i,C} = \left(\boldsymbol{X}_i - \bar{\boldsymbol{X}} \right) \bar{\boldsymbol{\Psi}}_{TE}^{-1} - \left[(1+\bar{\delta}_n)^{-1} \boldsymbol{Q}_{ix} - \boldsymbol{M}_i \bar{\boldsymbol{M}}_n^{-1} \boldsymbol{M}_T \bar{\boldsymbol{Q}}_{nx} \right],$$

and by Assumption S.1 $\bar{\boldsymbol{\Psi}}_{TE} \succ \mathbf{0}$. Also by Assumptions 1 and 9, $\text{plim}_{n\to\infty} \bar{\boldsymbol{M}}_n = \boldsymbol{M} \succ \mathbf{0}$, and under the null hypothesis given by (7.1), we have $E(\tilde{\nu}_{it}|\boldsymbol{G}_{i,C}) = 0$, for all i and t. Also by Assumptions 1 and 6, u_{it} and $\boldsymbol{\eta}_{i\beta}$ are cross-sectionally independent so that $\tilde{\nu}_{it}$ conditional on $\boldsymbol{X}_i$ are also cross-sectionally independent. Then when $T > k$, as $n \to \infty$, $\sqrt{n}\hat{\boldsymbol{\Delta}}_{\beta,C} \to_d N(\mathbf{0}, \boldsymbol{V}_{\Delta,C})$, so long as

$$\boldsymbol{V}_{\Delta,C} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} E\left(\boldsymbol{G}_{i,C}' \boldsymbol{M}_T \tilde{\boldsymbol{\nu}}_i \tilde{\boldsymbol{\nu}}_i' \boldsymbol{M}_T \boldsymbol{G}_{i,C} \right) \succ \mathbf{0}. \tag{S.2.17}$$

Thus, when $T > k$, the Hausman-type test for panels with time effects is given by

$$H_{\beta,C} = n \hat{\boldsymbol{\Delta}}_{\beta,TE}' \boldsymbol{V}_{\Delta,C}^{-1} \hat{\boldsymbol{\Delta}}_{\beta,TE}, \tag{S.2.18}$$

and as $n \to \infty$, $H_{\beta,C} \to_d \chi^2_{k'}$. A consistent estimator of $\boldsymbol{V}_{\Delta,C}$ for a fixed T is given by

$$\widehat{\boldsymbol{V}}_{\Delta,C} = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\boldsymbol{G}}_{i,C}' \boldsymbol{M}_T \hat{\tilde{\boldsymbol{\nu}}}_{i,FE} \hat{\tilde{\boldsymbol{\nu}}}_{i,FE}' \boldsymbol{M}_T \hat{\boldsymbol{G}}_{i,C} \right) \tag{S.2.19}$$

where

$$\hat{\boldsymbol{G}}_{i,C} = \left(\boldsymbol{X}_i - \bar{\boldsymbol{X}} \right) \bar{\boldsymbol{\Psi}}_{n,TE}^{-1} - \left[(1+\bar{\delta}_n)^{-1} \boldsymbol{Q}_{ix} - \boldsymbol{M}_i \bar{\boldsymbol{M}}_n^{-1} \boldsymbol{M}_T \bar{\boldsymbol{Q}}_{nx} \right], \tag{S.2.20}$$

with $\hat{\tilde{\boldsymbol{\nu}}}_{i,FE}$ given by (S.2.13). Then the test statistics given by (S.2.18) for panel regressions with time effects and $T > k$ can be consistently estimated by

$$\hat{H}_{\beta,C} = n \left(\hat{\boldsymbol{\beta}}_{FE-TE} - \hat{\boldsymbol{\beta}}_{TMG-C} \right)' \widehat{\boldsymbol{V}}_{\Delta,C}^{-1} \left(\hat{\boldsymbol{\beta}}_{FE-TE} - \hat{\boldsymbol{\beta}}_{TMG-C} \right). \tag{S.2.21}$$

S.3 Parameters of Monte Carlo experiments

The DGP for y_{it} and x_{it} is described in Section 8.1 of the main paper. Section S.3.1 describes different DGPs considered in the MC experiments, with their key parameters summarized in Table S.1. Section S.3.2 describes how the value of κ_T in the y_{it} process has been calibrated by stochastic simulations to achieve a given level of overall fit, PR^2 for a given value of T. The simulated values of κ_T^2 for different DGPs are reported in Table S.2.

S.3.1 Calibration of the parameters

1. Generation of y_{it} and x_{it} , for $i = 1, 2, ..., n$, and $t = 1, 2, ..., T$.

   (a) x_{it} are generated as heterogeneous AR(1) processes with $\rho_{ix} = 0$ for all i (in the static case), and $\rho_{ix} \sim IIDU(0, 0.95)$ for the dynamic case. See (8.3) in the main paper. The errors $e_{x,it}$ of the x_{it} equation are generated according to the following two distributions:

      i. Gaussian with $e_{x,it} \sim IIDN(0, 1)$, where $E(e_{x,it}) = 0$, $E(e^2_{x,it}) = 1$, and $\gamma_2 = E(e^4_{x,it}) - 3 = 0$.

      ii. Uniform distribution with $e_{x,it} = \sqrt{12}(\mathfrak{z}_{it} - 1/2)$, where $\mathfrak{z}_{it} \sim IIDU(0, 1)$, with $E(\mathfrak{z}_{it}) = 1/2$ and $Var(\mathfrak{z}_{it}) = \frac{1}{12}$. Hence, $E(e_{x,it}) = 0$, $E(e^2_{x,it}) = 1$ and $\gamma_2 = E(e^4_{x,it}) - 3 = \frac{\left(\sqrt{3}\right)^5 - \left(\sqrt{-3}\right)^5}{(4+1)(\sqrt{3}+\sqrt{3})} - 3 = -\frac{6}{5}$.

   (b) $\alpha_{ix} \sim IIDN(1, 1)$, and $\sigma^2_{ix} \sim IID\frac{1}{2}(z^2_{ix} + 1)$, with $z_{ix} \sim IIDN(0, 1)$.

   (c) The errors in the y_{it} equation are composed of three components, $\kappa\sigma_{it}e_{it}$. See (8.1) in the main paper. e_{it} are generated as heterogeneous AR(1) processes given by (8.2) in the main paper, with $\rho_{ie} = 0$ for all i (serially uncorrelated case) and $\rho_{ie} \sim IIDU(0, 0.95)$ (serially correlated case). The innovations to the y_{it}, ς_{it}, are generated as $\varsigma_{it} \sim IIDN(0, 1)$, or $IID\frac{1}{2}(\chi^2_2 - 2)$. σ^2_{it} are generated based on different cases as described in Section 8.1.3 with $E(\sigma^2_{it}) = 1$. The scalar, κ, is calibrated for each T to achieve a given level of fit, $PR^2 \in \{0.2, 0.4\}$, See sub-section S.3.2 below.

2. Generation of heterogeneous coefficients, $\boldsymbol{\theta}_i = (\alpha_i, \beta_i)'$ for $i = 1, 2, ..., n$.

   (a) $\boldsymbol{\theta}_i = (\alpha_i, \beta_i)'$ are generated using (8.4) in sub-section 8.1.2 of the main paper, with $\alpha_0 = E(\alpha_i) = 1$ and $\beta_0 = E(\beta_i) = 1$.

(b) $(\rho_{\alpha\phi}, \rho_{\beta\phi})' \in \{0, 0.5\}$, $\sigma_{\alpha}^{2} = 0.2$ and $\sigma_{\beta}^{2} \in \{0.2, 0.5, 0.75\}$.

(c) $\epsilon_{i\alpha} \sim IID(0, \sigma_{\epsilon\alpha}^{2})$ and $\epsilon_{i\beta} \sim IID(0, \sigma_{\epsilon\beta}^{2})$, where $\sigma_{\epsilon\alpha}^{2} = (1 - \rho_{\alpha\lambda}^{2})\sigma_{\alpha}^{2}$ and $\sigma_{\epsilon\beta}^{2} = (1 - \rho_{\beta\lambda}^{2})\sigma_{\beta}^{2}$.

(d) For correlated heterogeneous coefficients we set

$$\psi_{\alpha} = \left(\frac{\rho_{\alpha\lambda}^{2}}{1 - \rho_{\alpha\lambda}^{2}}\right)^{1/2} \sigma_{\epsilon\alpha} = \rho_{\alpha\lambda}\sigma_{\alpha}, \text{ and } \psi_{\beta} = \left(\frac{\rho_{\beta\lambda}^{2}}{1 - \rho_{\beta\lambda}^{2}}\right)^{1/2} \sigma_{\epsilon\beta} = \rho_{\beta\lambda}\sigma_{\beta}.$$

Example 4 *In the simple case where $\rho_{ix} = 0$ and $\gamma_{ix} = 0$, $x_{it} = \alpha_{ix} + \sigma_{ix} e_{x,it}$, then*

$$d_i = \det(\boldsymbol{W}_i'\boldsymbol{W}_i) = T\boldsymbol{x}_i'\boldsymbol{M}_T\boldsymbol{x}_i = T\sigma_{ix}^{2}\left(\boldsymbol{e}_{ix}'\boldsymbol{M}_T\boldsymbol{e}_{ix}\right). \tag{S.3.1}$$

Using Lemma 6 in the online supplement of Pesaran and Yamagata (2023), we have

$$E(\boldsymbol{e}_{ix}'\boldsymbol{M}_T\boldsymbol{e}_{ix}) = Tr(\boldsymbol{M}_T) = T - 1,$$
$$E\left[(\boldsymbol{e}_{ix}'\boldsymbol{M}_T\boldsymbol{e}_{ix})^2\right] = \gamma_2 Tr[\boldsymbol{M}_T \odot \boldsymbol{M}_T] + Tr(\boldsymbol{M}_T\boldsymbol{M}_T) + 2Tr(\boldsymbol{M}_T)Tr(\boldsymbol{M}_T),$$

where $\odot$ denotes the element-wise product, and γ_2 measures excess kurtosis of $e_{x,it} \sim IID(0,1)$ given by $\gamma_2 = E(e_{x,it}^4) - 3$, which depends on the specific distribution of $e_{x,it}$. Denote the diagonal elements of $\boldsymbol{M}_T$ as m_{tt} for $t = 1, 2, ..., T$, then $m_{tt} = 1 - \frac{1}{T}$, and we have $Tr[\boldsymbol{M}_T \odot \boldsymbol{M}_T] = \sum_{t=1}^{T} m_{tt}^2 = \frac{(T-1)^2}{T}$. It is now easily seen that

$$E\left[(\boldsymbol{e}_{ix}'\boldsymbol{M}_T\boldsymbol{e}_{ix})^2\right] = (T-1)(T+1) + \frac{\gamma_2(T-1)^2}{T},$$

and hence

$$Var(\boldsymbol{e}_{ix}'\boldsymbol{M}_T\boldsymbol{e}_{ix}) = 2(T-1) + \frac{\gamma_2(T-1)^2}{T}.$$

Using the above results we now have

$$\lambda_i = \frac{\boldsymbol{e}_{ix}'\boldsymbol{M}_T\boldsymbol{e}_{ix} - (T-1)}{\left[2(T-1) + \frac{\gamma_2(T-1)^2}{T}\right]^{1/2}}, \tag{S.3.2}$$

which can be viewed as a standardized version of $d_i = T\sigma_{ix}^2\left(\boldsymbol{e}_{ix}'\boldsymbol{M}_T\boldsymbol{e}_{ix}\right)$, conditional on σ_{ix}^2, namely

$$\lambda_i = \frac{d_i/T\sigma_{ix}^2 - E\left(d_i/T\sigma_{ix}^2\right)}{\sqrt{Var\left(d_i/T\sigma_{ix}^2\right)}} = \frac{d_i - E(d_i)}{\sqrt{Var\left(d_i\right)}}.$$

Table S.1: Summary of key parameters in the Monte Carlo experiments with Gaussian and uniformly distributed errors

	Gaussian				Uniform			
Case	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
$E(\alpha_i)$	1	1	1	1	1	1	1	1
$E(\beta_i)$	1	1	1	1	1	1	1	1
$E(\sigma_{it}^2)$	1	1	1	1	1	1	1	1
$E(\sigma_{ix}^2)$	1	1	1	1	1	1	1	1
γ_2	0	0	0	0	-1.2	-1.2	-1.2	-1.2
$E(d_i)$	2	2	2	2	2	2	2	2
$Var(d_i)$	14	14	14	14	10.4	10.4	10.4	10.4
PR^2	0.2	0.2	0.2	0.4	0.2	0.2	0.2	0.4
σ_α^2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
σ_β^2	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
$\rho_{\lambda\alpha}$	0	0.25	0.5	0.5	0	0.25	0.5	0.5
$\rho_{\lambda\beta}$	0	0.25	0.5	0.5	0	0.25	0.5	0.5
$\sigma_{\epsilon\alpha}^2$	0.2	0.19	0.15	0.15	0.2	0.19	0.15	0.15
$\sigma_{\epsilon\beta}^2$	0.5	0.469	0.375	0.375	0.5	0.469	0.375	0.375
ψ_α	0	0.11	0.22	0.22	0	0.11	0.22	0.22
ψ_β	0	0.18	0.35	0.35	0	0.18	0.35	0.35
$Corr(\alpha_i, \beta_i)$	0	0.0625	0.25	0.25	0	0.0625	0.25	0.25

Notes: The values of key parameters under columns (1), (2) and (3) are set according to the baseline model described in Section 8.1.3 with zero, medium and large degrees of correlated heterogeneity, where $\rho_{\lambda\beta}$ is defined by (8.6) in the main paper. For further details, see Section 8.1.3 of the main paper. The calibrated parameter values are reported for $T = 2$, to save space.

S.3.2 Calibration of κ^2 by stochastic simulation

The scaling parameter κ in (8.1) is set to achieve a given level of fit as measured by the pooled PR^2

$$\begin{aligned} PR^2 &= \lim_{n\to\infty} PR_n^2 = 1 - \frac{\lim_{n\to\infty} n^{-1}T^{-1}\sum_{i=1}^n \sum_{t=1}^T Var(u_{it})}{\lim_{n\to\infty} n^{-1}T^{-1}\sum_{i=1}^n \sum_{t=1}^T Var(y_{it} - \alpha_i - \phi_t)} \\ &= 1 - \frac{\kappa^2}{\lim_{n\to\infty} n^{-1}T^{-1}\sum_{i=1}^n \sum_{t=1}^T Var(\beta_i x_{it}) + \kappa^2}. \end{aligned} \tag{S.3.3}$$

Since T is fixed the value of κ in general depends on T and we have (noting that $Var(\beta_i x_{it}) = E(\beta_i^2 x_{it}^2) - [E(\beta_i x_{it})]^2$)

$$\kappa_T^2 = \left(\frac{1 - PR^2}{PR^2}\right) \lim_{n\to\infty} \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T \left\{ E(\beta_i^2 x_{it}^2) - [E(\beta_i x_{it})]^2 \right\}. \tag{S.3.4}$$

Due to the non-linear dependence of β_i on x_{it} (through d_i) we use stochastic simulations to compute $E(\beta_i^2 x_{it}^2)$ and $E(\beta_i x_{it})$, which can be carried out in a straightforward manner since the values of x_{it} and β_i do not depend on κ and can be jointly simulated using the relations (8.3) and (8.4).

The total number of simulations is $R_\kappa = 1,000$ with $n = 5,000$ and $T = 2,3,4,5,6,8$. For each replication $r = 1,2,...,R_\kappa$, we generate a new sample of $\beta_i^{(r)}$ and $\{x_{it}^{(r)}\}$ given the DGP set up in our paper. The random variables which are drawn independently across replications are denoted with a superscript (r). The random variables that are drawn once and used for all replications are denoted without a superscript (r).

1. Generate $x_{it}^{(r)}$:

   (a) First generate $e_{x,it}^{(r)}$ as $IID(0,1)$ according to the two distributions specified in Section S.3.1, namely Gaussian or uniform distributions, and generate $(\sigma_{ix}^2)^{(r)}$ as $IID\frac{1}{2}\left[\left(z_{ix}^{(r)}\right)^2 + 1\right]$, where $z_{ix}^{(r)}$ are generated as $IIDN(0,1)$.

   (b) Generate
   $$\lambda_i^{(r)} = \frac{\boldsymbol{e}_{ix}^{(r)\prime}\boldsymbol{M}_T\boldsymbol{e}_{ix}^{(r)} - (T-1)}{\sqrt{2(T-1) + \frac{\gamma_2}{T}(T-1)^2}},$$
   where γ_2 depends on the chosen distribution: (i) with Gaussian distribution $\gamma_2 = 0$, and (ii) with uniform distribution $\gamma_2 = -\frac{6}{5}$.

   (c) Generate $\rho_{ix} = 0\,\forall i$ for static x_{it}, or $\rho_{ix}^{(r)} \sim IIDU(0, 0.95)$ for dynamic x_{it}. Then generate $\epsilon_{ix}^{(r)}$ as $IIDN(0,1)$, and $x_{it}^{(r)}$ iteratively for $t = -49, -48, ..., -1, 0, 1, ..., T$ for the dynamic case
   $$x_{it}^{(r)} = \alpha_{ix}^{(r)}\left(1 - \rho_{ix}^{(r)}\right) + \gamma_{ix}^{(r)} f_t + \rho_i^{(r)} x_{i,t-1}^{(r)} + \left[1 - \left(\rho_{ix}^{(r)}\right)^2\right]^{1/2} \sigma_{ix}^{(r)} e_{x,it}^{(r)},$$
   without or with interactive effects, $\gamma_{ix}^{(r)} \sim IIDU(0,2)$, $f_t = 0.9 f_{t-1} + (1-0.9^2)^{1/2} v_t$, and $v_t \sim IIDN(0,1)$, where $x_{i,-50} = 0$ and $f_{-50} = 0$. The first 50 observations are dropped, and $\{x_{i1}, x_{i2}, ..., x_{iT}\}$ are used in the simulations.

2. Generate $\beta_i^{(r)}$

   (a) Generate $\epsilon_{i\beta}^{(r)}$ as $IIDN(0, \sigma_{\epsilon\beta}^2)$ where $\sigma_{\epsilon\beta}^2 = (1 - \rho_{\lambda\beta}^2)\sigma_\beta^2$.

   (b) Given ψ_β, $\lambda_i^{(r)}$ and $\epsilon_{i\beta}^{(r)}$, set $\beta_i^{(r)} = \beta_0 + \eta_{i\beta}^{(r)}$. Then $\beta_i^{(r)} = \beta_0 + \eta_{i\beta}^{(r)}$.

3. Given $\beta_i^{(r)}$ and $x_{it}^{(r)}$, we then simulate

$$A_{RT} = R^{-1}T^{-1}n^{-1}\sum_{r=1}^{R}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(\beta_i^{(r)}\right)^2\left(x_{it}^{(r)}\right)^2$$

$$B_{RT} = R^{-1}T^{-1}n^{-1}\sum_{r=1}^{R}\sum_{i=1}^{n}\sum_{t=1}^{T}\beta_i^{(r)}x_{it}^{(r)}$$

$$\widehat{Var(\beta_i x_{it})}_{RT} = A_{RT} - B_{RT}^2.$$

Then for given values of $PR^2 = 0.2$ or 0.4, compute κ_T^2 as

$$\kappa_T^2 = \left(\frac{1 - PR^2}{PR^2}\right)\widehat{Var(\beta_i x_{it})}_{RT}.$$

Table S.2: Simulated values of κ_T^2 for $T = 2, 3, 4, 5, 6, 8$

	Gaussian				Uniform			
Case	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
There is no autoregressions or factors in x_{it}.								
$T = 2$	14.02	14.62	17.02	6.38	14.01	14.08	15.30	5.74
$T = 3$	14.00	14.61	16.56	6.21	14.00	14.10	15.16	5.69
$T = 4$	14.00	14.58	16.24	6.09	14.00	14.13	15.06	5.65
$T = 5$	14.00	14.56	16.00	6.00	14.00	14.13	14.98	5.62
$T = 6$	14.00	14.56	15.83	5.93	14.00	14.15	14.93	5.60
$T = 8$	14.00	14.50	15.57	5.84	13.99	14.14	14.82	5.56
x_{it} are generated as heterogeneous AR(1) processes.								
$T = 2$	13.98	14.62	15.50	5.81	13.98	14.62	15.50	5.81
$T = 3$	13.99	14.61	15.43	5.79	13.99	14.61	15.43	5.79
$T = 4$	13.97	14.58	15.33	5.75	13.97	14.58	15.33	5.75
$T = 5$	13.99	14.56	15.26	5.72	13.99	14.56	15.26	5.72
$T = 6$	14.01	14.56	15.22	5.71	14.01	14.56	15.22	5.71
$T = 8$	13.98	14.50	15.10	5.66	13.98	14.50	15.10	5.66
x_{it} are generated as heterogeneous AR(1) processes with interactive effects.								
$T = 2$	23.95	24.59	25.47	9.55	23.94	24.02	24.16	9.06
$T = 3$	25.05	25.67	26.48	9.93	25.07	25.20	25.37	9.51
$T = 4$	24.53	25.14	25.91	9.71	24.55	24.70	24.88	9.33
$T = 5$	24.89	25.47	26.17	9.81	24.91	25.05	25.23	9.46
$T = 6$	24.36	24.91	25.58	9.59	24.34	24.51	24.71	9.27
$T = 8$	23.20	23.71	24.30	9.11	23.20	23.37	23.57	8.84

Notes: The values of κ_T^2 are computed according to stochastic simulations described in Section S.3.2, with $1,000$ replications. The values of key parameters for different cases are summarized in Table S.1.

S.4 Monte Carlo evidence with Gaussian distributed errors in the regressor

S.4.1 Comparison of FE, MG and TMG estimators

Table S.3 summarizes the MC results for the FE, MG and TMG estimators in panel data models under uncorrelated heterogeneity ($\rho_\beta = 0$), but with correlated heteroskedastic errors (in the y_{it} equation), for $T = 2, 3, 4, 5, 6, 8$ and $n = 1,000, 2,000, 5,000$ and $10,000$. It gives bias, RMSE and size for case (a) $\sigma_{it}^2 = \lambda_i^2$, on the left panel and for case (b) $\sigma_{it}^2 = e_{x,it}^2$, on the right panel of the table.[S1] As to be expected, the MG estimator performs very poorly when T is ultra short, and suffers from substantial bias. In contrast, the bias of the TMG estimator remains small even when $T = 2$. Turning to the comparison of FE and TMG estimators, we note that under both specifications of error heteroskedasticity, the bias of FE and TMG estimators are close to zero, and both estimators have the correct size for all T and n combinations. The main difference between FE and TMG estimators lies in their relative efficiency (in the RMSE sense), when T is ultra short. For example, when $T = 2$ and $n = 1,000$ the FE estimator is more efficient than the TMG estimator under case (b), whilst the reverse is true under case (a). This ranking of the two estimators is also reflected in their empirical power functions shown on the left and right panels of Figure S.1, for $T = 2, 3, 4, 5$ and $n = 10,000$. The empirical power functions for both estimators are correctly centered around $\beta_0 = 1$. But under heteroskedasticity of type (a), the empirical power function of the TMG estimator is steeper and for $T = 2$ lies within that of the FE estimator, with the reverse being true when error heteroskedasticity is generated under case (b).[S2] However, differences between FE, MG and TMG estimators vanish very rapidly as n and T are increased.

As a general rule, the FE estimator performs well when heterogeneity is uncorrelated. But in line with our theoretical results, the FE estimator suffers from substantial bias and size distortions under correlated heterogeneity, irrespective of whether the errors are heteroskedastic. The degree of bias and size distortion of the FE estimator rises with the degree of heterogeneity, ρ_β. Table S.4 provides additional MC results for $\rho_\beta = 0.25$ and $PR^2 = 0.2$ on the left panel and for $\rho_\beta = 0.5$ and $PR^2 = 0.4$ on the right panel for sample size combinations $T = 2, 3, 4, 5, 6, 8$, and $n = 1,000$, $2,000$, $5,000$, $10,000$. Comparing these results with those already reported in Table 1 we also note that the FE estimator shows a higher degree of distortion when PR^2 is increased from 0.20 to 0.40, with ρ_β fixed at 0.50.

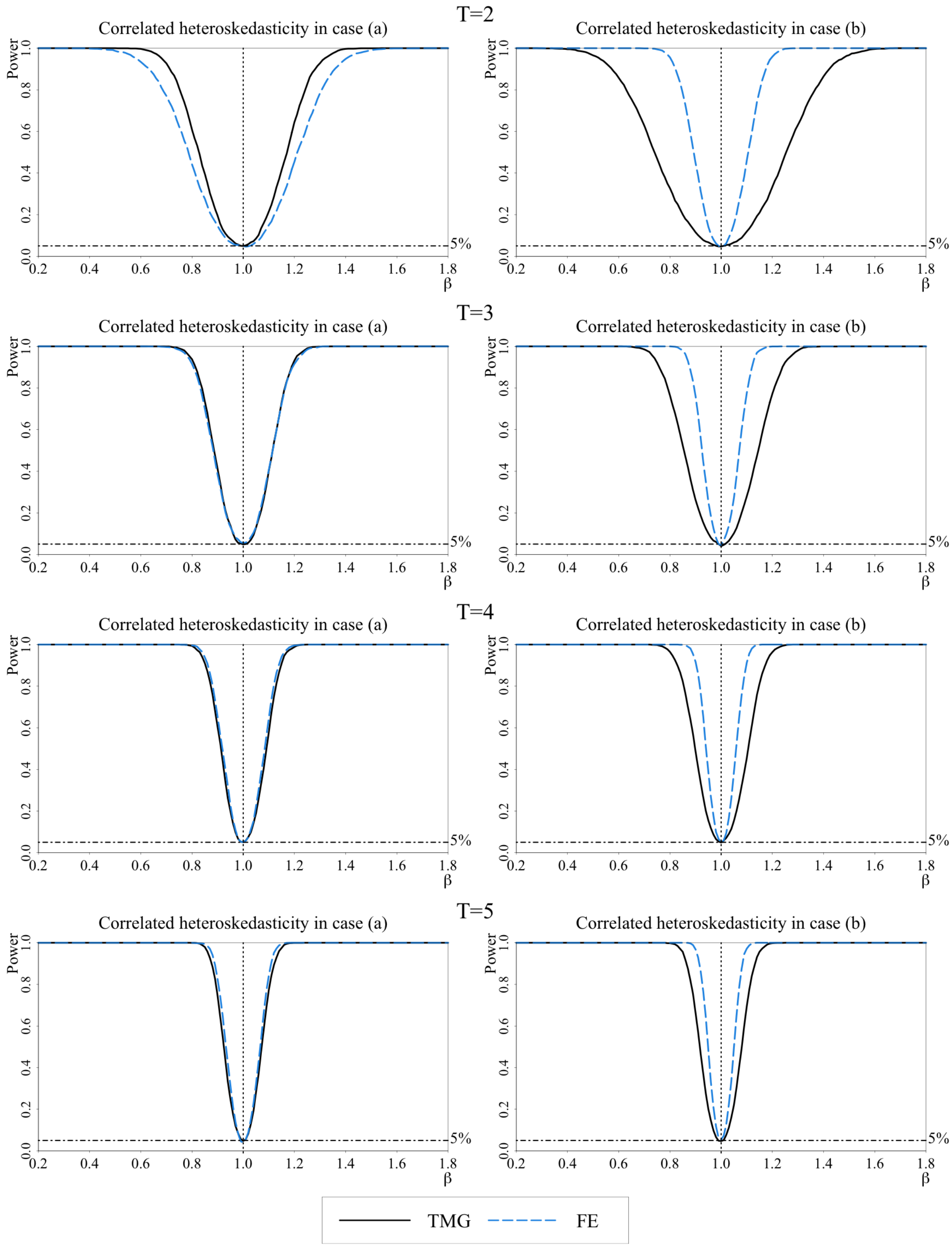
[S1]For details on the DGP and the rationale behind these specifications see Section 8.1.3 in the main paper.
[S2]These results are in line with Proposition 3 and Example 2 in the main paper.

Table S.3: Bias, RMSE and size of FE, MG and TMG estimators of β_0 ($E(\beta_i) = \beta_0 = 1$) in panel data models without time effects and with uncorrelated heterogeneity, $\rho_\beta = 0$, and correlated heteroskedasticity (cases (a) and (b))

	Case (a) of correlated heteroskedasticity											Case (b) of correlated heteroskedasticity										
	$\hat{\pi}$ (×100)	Bias			RMSE			Size (×100)			$\hat{\pi}$ (×100)	Bias			RMSE			Size (×100)		
T	TMG	FE	MG	TMG	FE	MG	TMG	FE	MG	TMG	TMG	FE	MG	TMG	FE	MG	TMG	FE	MG	TMG
										$n = 1,000$										
2	31.2	-0.001	-6.872	-0.001	0.36	301.20	0.23	5.8	2.1	4.9	31.2	-0.006	-6.617	0.000	0.17	287.87	0.35	5.1	2.1	5.4
3	16.5	-0.002	-0.009	-0.001	0.18	0.47	0.15	5.7	4.6	4.4	16.5	0.000	0.005	-0.001	0.11	0.65	0.20	5.1	4.0	4.9
4	10.4	0.000	0.003	0.001	0.13	0.19	0.12	4.6	4.2	4.3	10.4	0.000	0.000	0.000	0.09	0.21	0.15	4.8	5.0	5.1
5	7.1	-0.005	-0.002	-0.002	0.11	0.13	0.10	5.1	4.0	4.2	7.1	-0.003	-0.005	-0.003	0.08	0.14	0.12	5.0	4.4	4.3
6	5.2	0.001	-0.001	0.000	0.09	0.11	0.09	5.4	4.4	5.6	5.2	0.002	0.001	0.001	0.07	0.11	0.10	5.2	5.6	5.2
8	3.2	0.003	0.001	0.001	0.07	0.08	0.08	5.2	5.3	5.2	3.2	0.003	0.001	0.001	0.06	0.09	0.08	5.2	5.6	5.5
										$n = 2,000$										
2	28.5	-0.006	-2.813	-0.001	0.25	86.06	0.18	5.2	2.6	5.1	28.5	-0.005	-1.826	-0.001	0.12	156.16	0.26	5.3	2.2	5.2
3	14.1	0.002	0.004	-0.003	0.14	0.38	0.12	5.1	4.3	5.6	14.1	0.002	0.005	0.000	0.09	0.45	0.15	5.3	4.0	5.1
4	8.4	0.001	-0.004	-0.004	0.10	0.14	0.09	5.4	4.5	4.8	8.4	0.000	-0.002	-0.002	0.07	0.15	0.11	5.1	5.1	4.5
5	5.6	0.001	-0.004	-0.001	0.08	0.10	0.08	5.0	4.8	5.1	5.6	0.000	-0.001	0.001	0.06	0.11	0.09	4.9	4.6	4.6
6	4.0	0.000	-0.001	-0.001	0.06	0.08	0.07	4.7	4.9	4.2	4.0	0.000	-0.001	-0.001	0.05	0.08	0.07	3.8	4.2	5.3
8	2.4	0.000	0.000	0.000	0.05	0.06	0.06	4.8	4.8	5.0	2.4	-0.001	-0.001	-0.001	0.04	0.06	0.06	4.6	5.5	5.3
										$n = 5,000$										
2	24.7	-0.001	-7.430	-0.002	0.16	363.42	0.12	4.7	1.7	4.2	24.7	0.000	-6.647	-0.004	0.08	485.77	0.17	4.9	1.9	4.6
3	10.8	0.001	-0.004	0.000	0.09	0.22	0.08	5.1	5.5	5.1	10.8	0.000	-0.007	-0.002	0.05	0.25	0.10	4.2	4.7	5.1
4	5.8	0.002	0.001	0.001	0.06	0.09	0.06	5.4	4.8	5.4	5.8	0.000	-0.001	-0.002	0.04	0.10	0.07	4.6	5.0	5.3
5	3.5	0.000	-0.001	0.000	0.05	0.06	0.05	4.8	4.5	5.1	3.5	-0.001	-0.002	-0.001	0.04	0.07	0.06	5.2	4.5	5.0
6	2.3	-0.001	0.001	0.000	0.04	0.05	0.05	4.6	5.3	5.9	2.3	-0.001	0.000	0.000	0.03	0.05	0.05	4.6	5.3	4.7
8	1.2	0.000	-0.001	-0.001	0.03	0.04	0.04	4.9	5.3	5.0	1.2	0.001	0.000	0.000	0.03	0.04	0.04	5.4	4.7	4.6
										$n = 10,000$										
2	22.1	-0.002	-0.902	-0.002	0.11	102.44	0.09	5.0	2.2	5.0	22.1	-0.001	-1.820	-0.001	0.05	183.52	0.13	4.6	1.3	4.8
3	8.8	-0.001	0.007	0.001	0.06	0.17	0.06	5.6	4.6	4.9	8.8	-0.001	0.003	0.000	0.04	0.20	0.07	4.7	4.8	4.0
4	4.4	0.001	0.001	0.001	0.04	0.06	0.05	5.0	5.4	5.1	4.4	0.000	0.002	0.001	0.03	0.07	0.05	4.9	5.4	4.9
5	2.5	-0.001	-0.001	-0.001	0.03	0.04	0.04	4.0	4.9	4.8	2.5	0.000	-0.002	-0.001	0.03	0.05	0.04	5.4	5.4	4.7
6	1.6	0.000	0.000	0.000	0.03	0.04	0.03	4.2	4.2	4.4	1.6	0.001	0.000	0.000	0.02	0.04	0.03	5.1	4.9	4.7
8	0.8	0.001	0.001	0.001	0.02	0.03	0.03	4.3	5.3	5.1	0.8	0.000	0.001	0.001	0.02	0.03	0.03	5.2	4.9	5.1

Notes: (i) The data generating process is given by $y_{it} = \alpha_i + \beta_i x_{it} + \sigma_{it} e_{it}$, where σ_{it}^2 are generated as case (a): $\sigma_{it}^2 = \lambda_i^2$, and case (b): $\sigma_{it}^2 = e_{x,it}^2$, for all i and t. The errors processes for y_{it} and x_{it} equations are chi-squared and Gaussian, respectively, and x_{it} are generated as heterogeneous AR(1) processes. For further details see Section 8.1.3 in the main paper and Section S.3. (ii) FE and MG estimators are given by (2.13) and (2.4), respectively, in the main paper. The trimmed mean group (TMG) estimator and its asymptotic variance are given by (4.5) and (5.21) in the main paper. (iii) The trimming threshold for the TMG estimator is given by $a_n = \bar{d}_n n^{-\alpha}$, where $\bar{d}_n = \frac{1}{n}\sum_i^n d_i$, $d_i = \det(\boldsymbol{W}_i'\boldsymbol{W}_i)$, $\boldsymbol{W}_i = (\boldsymbol{w}_{i1}, \boldsymbol{w}_{i2}, ..., \boldsymbol{w}_{iT})'$ and $\boldsymbol{w}_{it} = (1, x_{it})'$. α is set to 1/3. $\hat{\pi}$ is the simulated fraction of individual estimates being trimmed, defined by (4.7) in the main paper.

Figure S.1: Empirical power functions for FE and TMG estimators of β_0 ($E(\beta_i) = \beta_0 = 1$) in panel data models without time effects and with uncorrelated heterogeneity, $\rho_\beta = 0$, and correlated heteroskedasticity (cases (a) and (b))



Notes: For details of the DGP for the left and right panels, see footnote (i) to Table S.3. For the FE estimator, see footnote (ii) to Table S.3. For the TMG estimator, see footnotes (ii) and (iii) to Table S.3.

Table S.4: Bias, RMSE and size of FE, MG and TMG estimators of β_0 ($E(\beta_i) = \beta_0 = 1$) in panel data models without time effects and with different degrees of correlated heterogeneity, ρ_β, and overall fit, PR^2

	(2) Medium correlated heterogeneity $\rho_\beta = 0.25$ with $PR^2 = 0.2$										(4) Large correlated heterogeneity $\rho_\beta = 0.5$ with $PR^2 = 0.4$									
	$\hat{\pi}$ (×100)	Bias			RMSE			Size (×100)			$\hat{\pi}$ (×100)	Bias			RMSE			Size (×100)		
T	TMG	FE	MG	TMG	FE	MG	TMG	FE	MG	TMG	TMG	FE	MG	TMG	FE	MG	TMG	FE	MG	TMG
										$n = 1,000$										
2	31.2	0.220	-7.516	0.023	0.28	362.16	0.34	26.2	2.1	4.5	31.2	0.445	-4.724	0.049	0.47	227.71	0.22	94.1	2.1	5.4
3	16.5	0.161	-0.011	0.010	0.20	0.58	0.20	26.8	4.0	4.8	16.5	0.322	-0.007	0.023	0.33	0.37	0.13	96.8	4.0	5.6
4	10.4	0.132	-0.001	0.006	0.16	0.21	0.15	29.4	5.2	5.2	10.4	0.266	0.000	0.014	0.28	0.13	0.09	96.7	5.1	5.4
5	7.1	0.114	0.000	0.005	0.14	0.14	0.12	28.8	4.2	3.9	7.1	0.230	0.000	0.009	0.24	0.09	0.07	96.9	4.0	4.3
6	5.2	0.106	0.000	0.004	0.13	0.11	0.10	30.4	4.6	4.9	5.2	0.210	0.000	0.007	0.22	0.07	0.06	96.7	4.9	5.1
8	3.2	0.091	0.000	0.002	0.11	0.08	0.08	29.8	4.5	4.6	3.2	0.179	0.000	0.004	0.19	0.05	0.05	96.6	4.8	4.2
										$n = 2,000$										
2	28.5	0.220	-2.449	0.021	0.25	153.07	0.26	42.2	2.0	5.4	28.5	0.446	-1.546	0.044	0.46	96.55	0.17	99.8	2.0	5.7
3	14.1	0.162	0.007	0.008	0.18	0.45	0.16	47.3	4.3	5.5	14.1	0.323	0.005	0.020	0.33	0.28	0.10	100.0	4.4	5.9
4	8.4	0.133	-0.003	0.002	0.15	0.15	0.11	50.6	4.5	4.5	8.4	0.265	-0.002	0.009	0.27	0.10	0.07	99.9	4.4	4.9
5	5.6	0.117	-0.004	0.003	0.13	0.11	0.09	51.0	5.1	5.2	5.6	0.233	-0.003	0.007	0.24	0.07	0.06	99.9	5.4	5.0
6	4.0	0.104	-0.001	0.002	0.12	0.08	0.07	50.3	4.0	4.2	4.0	0.208	0.000	0.005	0.21	0.05	0.05	100.0	4.1	4.0
8	2.4	0.089	0.000	0.001	0.10	0.06	0.06	50.9	5.0	4.6	2.4	0.178	0.000	0.003	0.18	0.04	0.04	100.0	4.8	4.2
										$n = 5,000$										
2	24.7	0.227	-4.227	0.019	0.24	419.33	0.18	81.6	1.8	4.2	24.7	0.451	-2.757	0.037	0.45	274.79	0.12	100.0	1.8	5.6
3	10.8	0.162	-0.006	0.008	0.17	0.29	0.10	84.7	4.7	5.0	10.8	0.323	-0.004	0.016	0.33	0.18	0.07	100.0	4.7	6.0
4	5.8	0.133	0.000	0.004	0.14	0.10	0.08	86.9	5.0	5.6	5.8	0.265	0.000	0.008	0.27	0.06	0.05	100.0	5.1	6.0
5	3.5	0.115	-0.001	0.002	0.12	0.07	0.06	85.9	4.4	4.2	3.5	0.231	-0.001	0.004	0.23	0.04	0.04	100.0	4.3	4.4
6	2.3	0.103	0.000	0.002	0.11	0.06	0.05	86.4	5.3	5.5	2.3	0.207	0.000	0.003	0.21	0.04	0.03	100.0	5.3	5.9
8	1.2	0.089	0.000	0.000	0.09	0.04	0.04	87.4	5.0	4.9	1.2	0.177	0.000	0.001	0.18	0.03	0.03	100.0	4.7	4.7
										$n = 10,000$										
2	22.1	0.224	-1.595	0.012	0.23	179.03	0.14	97.9	2.2	4.8	22.1	0.450	-0.990	0.030	0.45	113.06	0.09	100.0	2.2	6.9
3	8.8	0.160	0.004	0.006	0.16	0.20	0.07	98.4	4.3	5.2	8.8	0.321	0.003	0.013	0.32	0.13	0.05	100.0	4.3	6.1
4	4.4	0.132	0.000	0.004	0.14	0.07	0.05	99.1	4.7	4.7	4.4	0.265	0.000	0.007	0.27	0.04	0.03	100.0	4.9	4.9
5	2.5	0.115	-0.001	0.001	0.12	0.05	0.04	99.2	4.3	4.6	2.5	0.231	0.000	0.003	0.23	0.03	0.03	100.0	4.6	4.7
6	1.6	0.104	0.000	0.001	0.11	0.04	0.04	99.4	4.1	4.0	1.6	0.208	0.000	0.002	0.21	0.02	0.02	100.0	4.2	3.9
8	0.8	0.089	0.000	0.001	0.09	0.03	0.03	99.2	4.9	4.9	0.8	0.177	0.000	0.001	0.18	0.02	0.02	100.0	5.0	5.1

Notes: The data generating process is given by $y_{it} = \alpha_i + \beta_i x_{it} + \sigma_{it} e_{it}$ with random heteroskedasticity, different degrees of correlated heterogeneity, ρ_β (defined by (8.6) in the main paper), and levels of overall fit, PR^2 (defined by (S.3.3)). The errors processes for y_{it} and x_{it} equations are chi-squared and Gaussian, respectively, and x_{it} are generated as heterogeneous AR(1) processes. For further details see Section 8.1.3 in the main paper and Section S.3. For FE and MG estimators, see footnote (ii) to Table S.3. For the TMG estimator, see footnotes (ii) and (iii) to Table S.3. $\hat{\pi}$ is the simulated fraction of individual estimates being trimmed, defined by (4.7) in the main paper.

S.4.2 Comparison of TMG and GP estimators for different exponents, α and α_{GP}, used in the threshold values

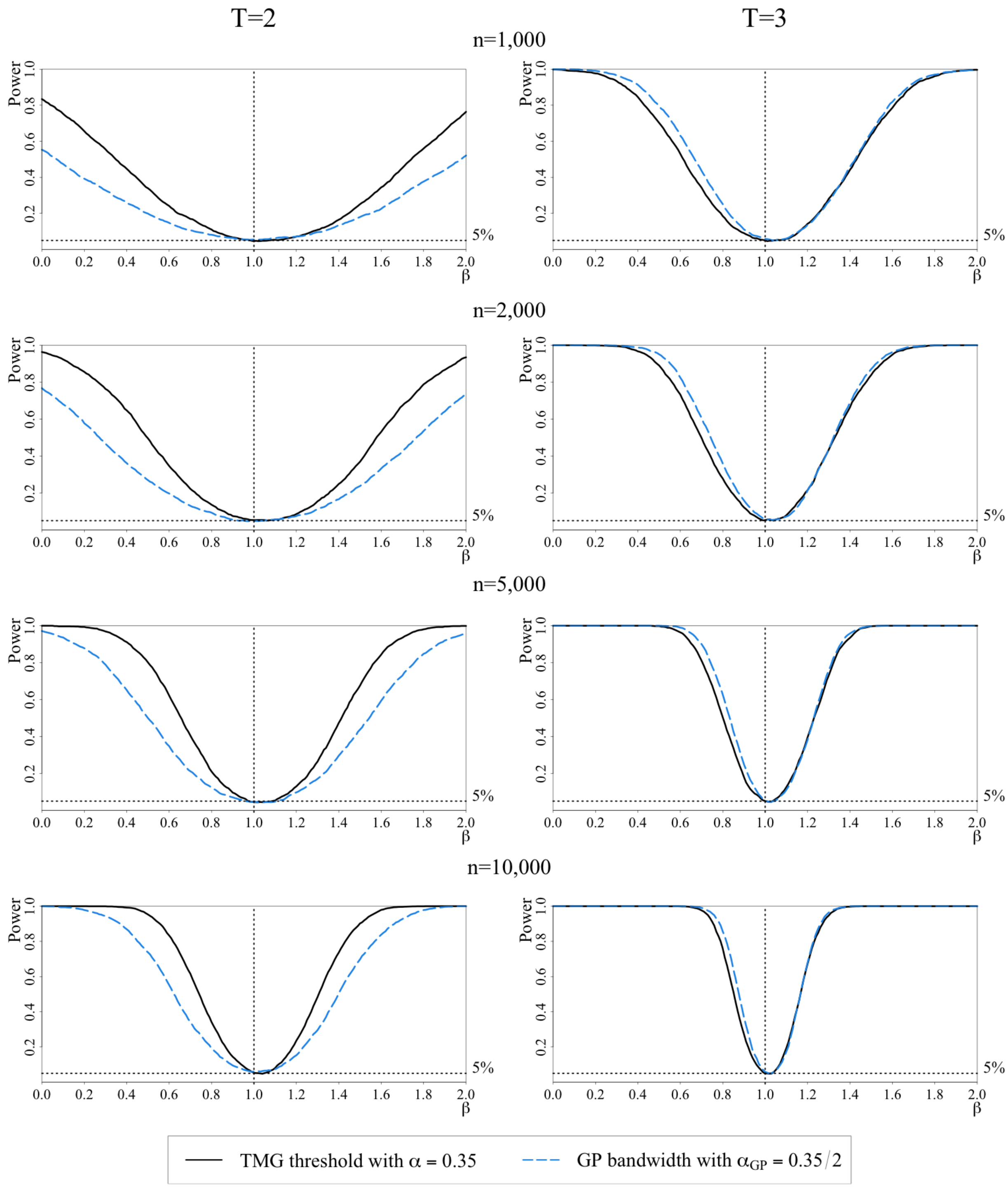
Table S.5 compares small sample properties of TMG and GP estimators of β_0 for different choices of α and α_{GP} used for their computations, respectively. The results are for the baseline model with correlated heterogeneity, but without time effects, for ultra short values of $T = 2, 3$, and $n = 1,000$, $2,000$, $5,000$ and $10,000$. The empirical power functions for TMG and GP estimators are shown in Figures S.4 and S.5. The figures suggest that both estimators benefit from setting α and $2\alpha_{GP}$ close to 1/3 value obtained from our theoretical derivations. The TMG estimator is less sensitive to the choice of α as compared to the GP estimator to α_{GP}. For further discussions see Section 8.2.4 of the main paper.

Table S.5: Bias, RMSE and size of TMG and GP estimators of β_0 ($E(\beta_i) = \beta_0 = 1$) for different values of the parameters, α and α_{GP}, in the baseline model without time effects and with correlated heterogeneity, $\rho_\beta = 0.5$

		$T=2$				$T=3$			
		$\hat{\pi}$			Size	$\hat{\pi}$			Size
Estimator	α/α_{GP}	($\times$100)	Bias	RMSE	($\times$100)	($\times$100)	Bias	RMSE	($\times$100)
					$n = 1,000$				
TMG	1/3	31.2	0.048	0.35	4.9	16.5	0.023	0.20	5.2
TMG	0.35	29.6	0.045	0.36	4.6	15.0	0.020	0.21	5.0
TMG	1/2	18.1	0.017	0.47	5.7	6.0	0.004	0.25	4.7
GP	0.35/2	12.6	0.017	0.51	5.5	15.0	0.045	0.20	5.9
GP	1/4	7.5	0.001	0.64	5.4	6.0	0.014	0.23	5.6
GP	1/3	4.2	-0.029	0.83	4.5	1.4	0.011	0.31	4.5
					$n = 2,000$				
TMG	1/3	28.5	0.044	0.27	5.3	14.1	0.018	0.16	5.4
TMG	0.35	26.8	0.041	0.28	5.3	12.7	0.016	0.16	5.3
TMG	1/2	15.6	0.028	0.37	5.1	4.5	0.004	0.20	5.0
GP	0.35/2	11.2	0.023	0.38	4.9	12.7	0.035	0.16	6.1
GP	1/4	6.4	0.035	0.50	4.7	4.5	0.009	0.18	5.0
GP	1/3	3.4	0.031	0.70	5.8	0.9	-0.001	0.23	5.0
					$n = 5,000$				
TMG	1/3	24.7	0.037	0.18	4.7	10.8	0.016	0.11	5.3
TMG	0.35	23.1	0.034	0.19	4.4	9.5	0.014	0.11	5.1
TMG	1/2	12.4	0.016	0.26	4.6	2.9	0.002	0.14	5.2
GP	0.35/2	9.6	0.018	0.26	4.6	9.5	0.029	0.10	5.1
GP	1/4	5.1	0.012	0.36	4.7	2.9	0.008	0.13	4.7
GP	1/3	2.5	0.009	0.52	5.2	0.5	0.002	0.16	3.9
					$n = 10,000$				
TMG	1/3	22.1	0.029	0.14	5.6	8.8	0.013	0.08	5.3
TMG	0.35	20.6	0.025	0.15	5.3	7.7	0.011	0.08	4.9
TMG	1/2	10.5	0.009	0.21	4.7	2.1	0.002	0.10	4.8
GP	0.35/2	8.5	0.012	0.20	6.0	7.7	0.022	0.08	5.8
GP	1/4	4.3	-0.002	0.28	5.1	2.1	0.005	0.09	4.8
GP	1/3	2.0	0.002	0.41	4.3	0.3	0.001	0.12	5.2

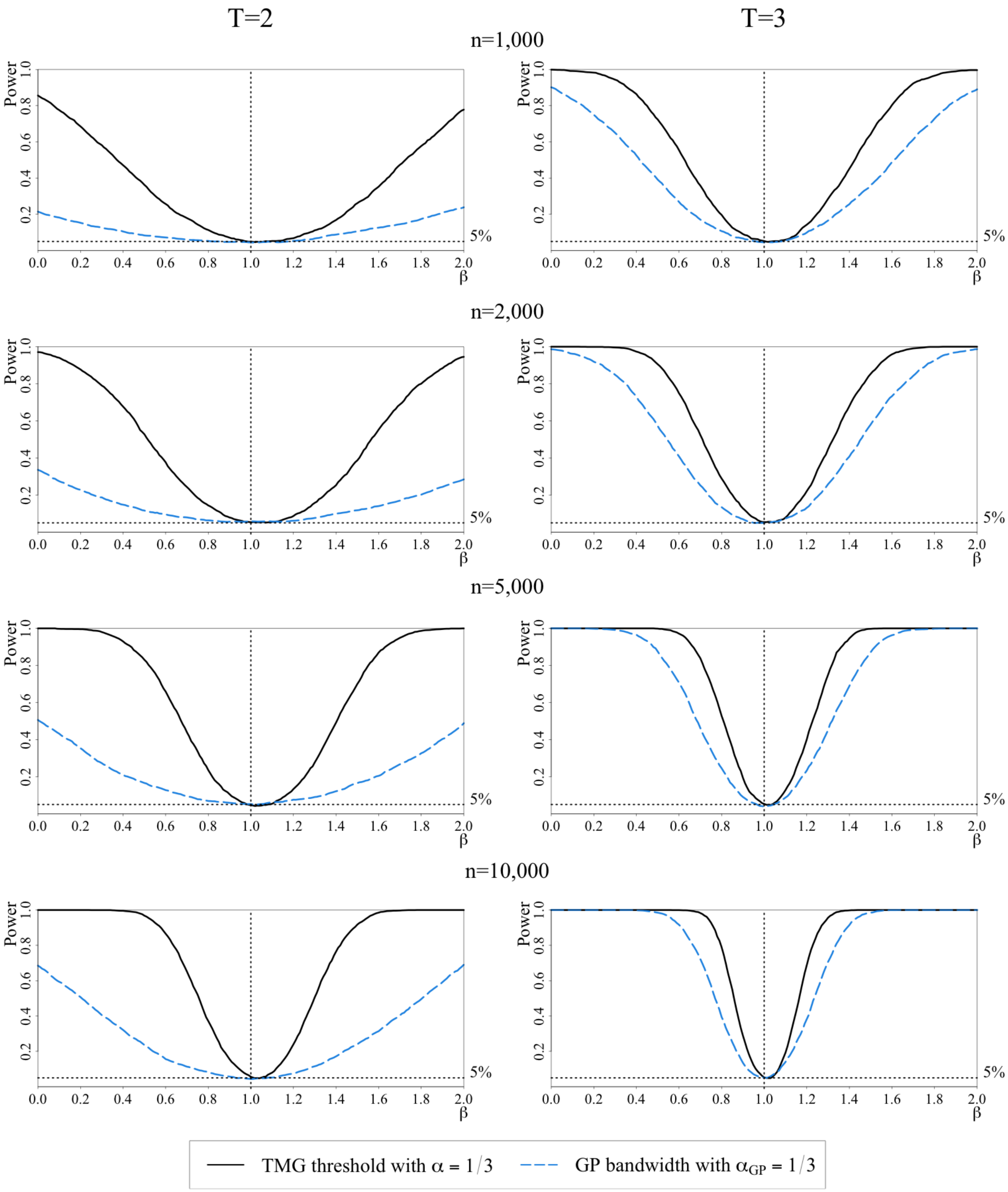
Notes: (i) The GP estimator is given by (3.1) in the main paper. For $T = 2$, GP compare $d_i^{1/2}$ with the bandwidth $h_n = C_{GP} n^{-\alpha_{GP}}$. α_{GP} is set to 1/3. $C_{GP} = \frac{1}{2}\min(\hat{\sigma}_D, \hat{r}_D/1.34)$, where $\hat{\sigma}_D$ and $\hat{r}_D$ are the respective sample standard deviation and interquartile range of $\det(\boldsymbol{W}_i)$. For $T = 3$, we continue using the bandwidth h_n with $C_{GP} = (\bar{d}_n)^{1/2}$. See Section 8.2.2 in the main paper for details. (ii) For details of the baseline model, see footnote (i) to Table 1 in the main paper. For the TMG estimator and its threshold, see footnotes (ii) and (iii) to Table S.3. $\hat{\pi}$ is the simulated fraction of individual estimates being trimmed, defined by (4.7) in the main paper.

Figure S.2: Empirical power functions for TMG and GP estimator of β_0 ($E(\beta_i) = \beta_0 = 1$) for parameters $\alpha = 0.35$ and $\alpha_{GP} = 0.35/2$ in the baseline model without time effects and with correlated heterogeneity, $\rho_\beta = 0.5$



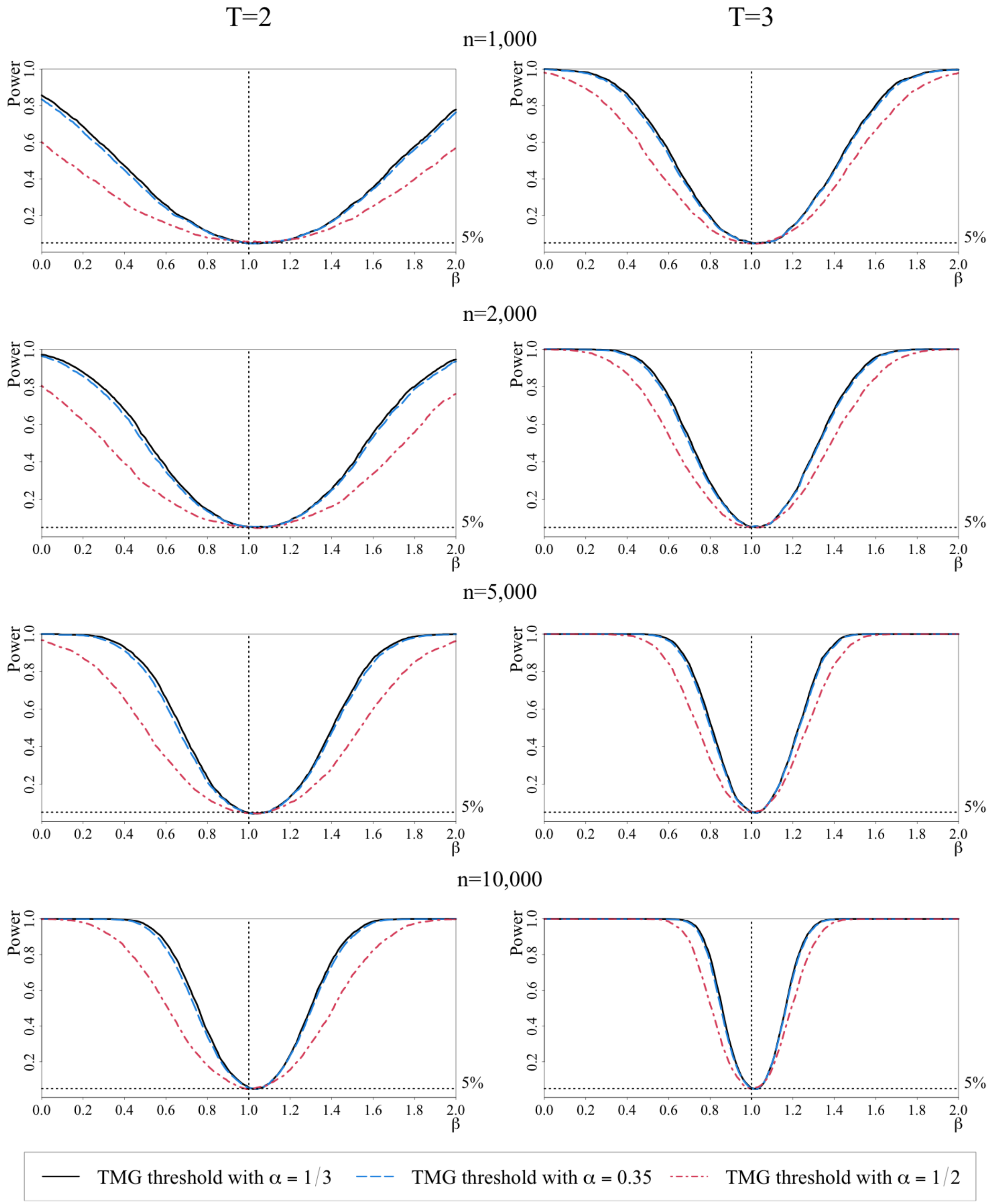
Notes: For details of the baseline model without time effects, see footnote (i) to Table 1 in the main paper. See also footnotes (ii) and (iii) to Table S.3 and footnote (i) to Table S.5.

Figure S.3: Empirical power functions for TMG and GP estimator of β_0 ($E(\beta_i) = \beta_0 = 1$) for parameters $\alpha = 1/3$ and $\alpha_{GP} = 1/3$ in the baseline model without time effects and with correlated heterogeneity, $\rho_\beta = 0.5$



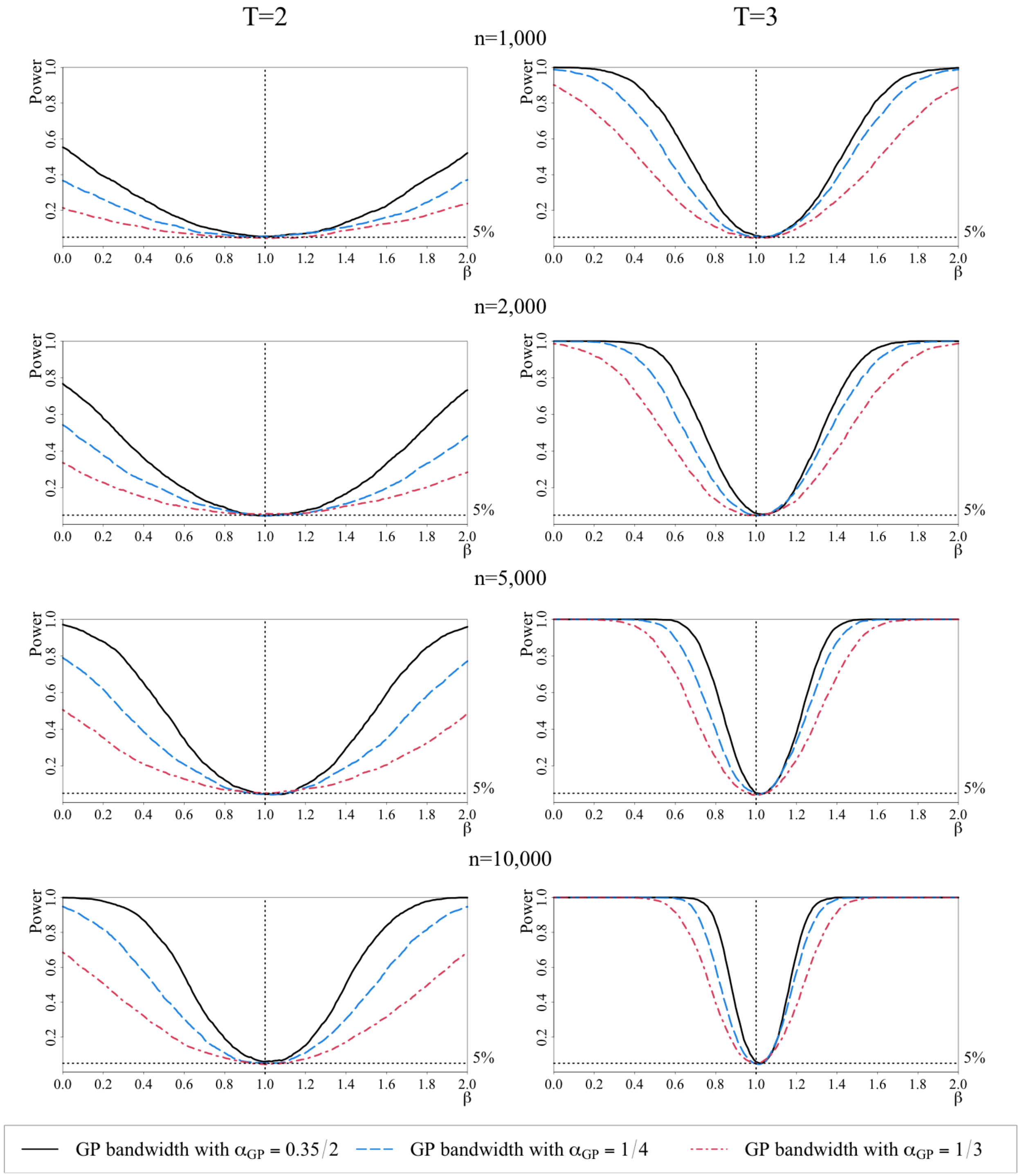
Notes: For details of the baseline model without time effects, see footnote (i) to Table 1 in the main paper. See also footnotes (ii) and (iii) to Table S.3 and footnote (i) to Table S.5.

Figure S.4: Empirical power functions for the TMG estimator of β_0 ($E(\beta_i) = \beta_0 = 1$) for different values of the threshold parameter, $\alpha \in \{1/3, 0.35, 1/2\}$, in the baseline model without time effects and with correlated heterogeneity, $\rho_\beta = 0.5$



Notes: For details of the baseline model without time effects, see footnote (i) to Table 1 in the main paper. See also footnotes (ii) and (iii) to Table S.3.

Figure S.5: Empirical power functions for the GP estimator of β_0 ($E(\beta_i) = \beta_0 = 1$) for different values of the bandwidth parameter, $\alpha_{GP} \in \{0.35/2, 1/4, 1/3\}$ in the baseline model without time effects and with correlated heterogeneity, $\rho_\beta = 0.5$



Notes: For details of the baseline model without time effects, see footnote (i) to Table 1 in the main paper. See also footnote (i) to Table S.5.

S.4.3 Comparison of TMG and GP estimators with correlated heteroskedasticity

Table S.6 provides additional MC results on small sample properties of TMG and GP estimators of β_0 in panel data models with correlated heterogeneity, $\rho_\beta = 0.5$, as well as correlated error heteroskedasticity which are generated as case (a) $\sigma_{it}^2 = \lambda_i^2$, and case (b) $\sigma_{it}^2 = e_{x,it}^2$, for all i and t. These results are to be compared to the ones in Table 2 in Section 8.2.2 of the main paper which are for random heteroskedasticity. The TMG estimator continues to perform better than the GP estimator when $T = 2$ or 3, and allowing for correlated heteroskedasticity does not alter this conclusion.

Table S.6: Bias, RMSE and size of TMG and GP estimators of β_0 ($E(\beta_i) = \beta_0 = 1$) in panel data models without time effects and with correlated heterogeneity, $\rho_\beta = 0.5$, and correlated heteroskedasticity (cases (a) and (b))

	$T=2$				$T=3$			
Estimator	$\hat{\pi}$ (×100)	Bias	RMSE	Size (×100)	$\hat{\pi}$ (×100)	Bias	RMSE	Size (×100)
Case (a): $\sigma_{it}^2 = \lambda_i^2$								
				$n = 1,000$				
TMG	31.2	0.049	0.25	5.4	16.5	0.023	0.16	4.9
GP	4.2	-0.012	0.53	4.7	2.0	0.000	0.22	4.3
				$n = 2,000$				
TMG	28.5	0.044	0.19	5.6	14.1	0.018	0.13	5.7
GP	3.4	0.022	0.46	6.0	1.3	0.004	0.18	5.2
				$n = 5,000$				
TMG	24.7	0.036	0.13	4.5	10.8	0.016	0.09	5.5
GP	2.5	0.000	0.33	5.4	0.7	0.000	0.13	5.4
				$n = 10,000$				
TMG	22.1	0.031	0.10	6.3	8.8	0.014	0.06	5.1
GP	2.0	0.004	0.25	4.3	0.5	0.000	0.09	4.9
Case (b): $\sigma_{it}^2 = e_{x,it}^2$								
				$n = 1,000$				
TMG	31.2	0.050	0.37	5.5	16.5	0.024	0.21	5.1
GP	4.2	-0.018	0.85	5.1	2.0	-0.003	0.28	4.6
				$n = 2,000$				
TMG	28.5	0.043	0.28	5.7	14.1	0.020	0.16	5.1
GP	3.4	0.038	0.72	5.4	1.3	0.003	0.22	4.7
				$n = 5,000$				
TMG	24.7	0.033	0.19	4.8	10.8	0.014	0.11	5.3
GP	2.5	-0.007	0.53	5.9	0.7	-0.002	0.15	4.3
				$n = 10,000$				
TMG	22.1	0.032	0.14	5.7	8.8	0.013	0.08	5.5
GP	2.0	0.006	0.41	5.3	0.5	-0.001	0.11	4.5

Notes: (i) The data generating process is given by $y_{it} = \alpha_i + \beta_i x_{it} + \sigma_{it} e_{it}$, where σ_{it}^2 are generated as case (a): $\sigma_{it}^2 = \lambda_i^2$, and case (b): $\sigma_{it}^2 = e_{x,it}^2$, for all i and t. The errors processes for y_{it} and x_{it} equations are chi-squared and Gaussian, respectively. x_{it} are generated as heterogeneous AR(1) processes. ρ_β is defined by (8.6) in the main paper. For further details see Section 8.1.3 in the main paper and Section S.3. (ii) For the TMG estimator, see footnotes (ii) and (iii) to Table S.3. For the GP estimator, see footnote (i) to Table S.5. $\hat{\pi}$ is the simulated fraction of individual estimates being trimmed, given by (4.7) in the main paper.

S.5 The effects of increasing PR^2 on the small sample properties of TMG, TMG-TE and TMG-C, GP and SU estimators

Table S.7 provides summary MC results with the higher level of fit, $PR^2 = 0.4$, for $T = 2$ and 3 and $n = 1,000$, $2,000$, $5,000$, $10,000$. These results are comparable with the ones reported in Table 2 for the baseline model where $PR^2 = 0.2$. Table S.8 gives the same results for β_0 but under DGPs with time effects, and Table S.9 provides the results for the time effect ϕ_1 when $T = 2$, and the time effects, ϕ_1 and ϕ_2, when $T = 3$.

Table S.10 reports bias, RMSE and size for the TMG and GP estimators for models with and without time effects, and for different x_{it} processes, for $T = 2$ and $n = 1,000$, $2,000$, $5,000$, $10,000$.

Table S.7: Bias, RMSE and size of TMG, GP and SU estimators of β_0 ($E(\beta_i) = \beta_0 = 1$) in panel data models without time effects and with correlated heterogeneity, $\rho_\beta = 0.5$, and the level of overall fit, $PR^2 = 0.4$

	$T = 2$				$T = 3$			
Estimator	$\hat{\pi}$ (×100)	Bias	RMSE	Size (×100)	$\hat{\pi}$ (×100)	Bias	RMSE	Size (×100)
				$n = 1,000$				
TMG	31.2	0.049	0.22	5.4	16.5	0.023	0.13	5.6
GP	4.2	-0.014	0.51	4.6	2.0	0.001	0.17	4.7
SU	4.2	-0.028	0.99	4.9	...	...	...	...
				$n = 2,000$				
TMG	28.5	0.044	0.17	5.7	14.1	0.020	0.10	5.9
GP	3.4	0.022	0.43	5.8	1.3	0.004	0.13	4.7
SU	3.4	0.005	0.85	5.5	...	...	...	...
				$n = 5,000$				
TMG	24.7	0.037	0.12	5.6	10.8	0.016	0.07	6.0
GP	2.5	0.007	0.32	5.3	0.7	0.000	0.09	5.1
SU	2.5	0.002	0.62	4.9	...	...	...	...
				$n = 10,000$				
TMG	22.1	0.030	0.09	6.9	8.8	0.013	0.05	6.1
GP	2.0	0.003	0.25	4.4	0.5	0.000	0.07	4.9
SU	2.0	0.004	0.50	5.1	...	...	...	...

Notes: (i) The data generating process is given by $y_{it} = \alpha_i + \beta_i x_{it} + \sigma_{it} e_{it}$ with random heteroskedasticity, ρ_β is defined by (8.6) in the main paper, and PR^2 is defined by (S.3.3). The errors processes for y_{it} and x_{it} equations are chi-squared and Gaussian, respectively, and x_{it} are generated as heterogeneous AR(1) processes. For further details see Section 8.1.3 in the main paper and Section S.3. (ii) For the TMG estimator, see footnotes (ii) and (iii) to Table S.3. For the GP estimator, see footnote (i) to Table S.5. The SU estimator is proposed by Sasaki and Ura (2021), with the same threshold as the GP estimator. $\hat{\pi}$ is the simulated fraction of individual estimates being trimmed, defined by (4.7) in the main paper. The estimation algorithm for the SU estimator is not available for $T = 3$, denoted by "...".

Table S.8: Bias, RMSE and size of TMG-TE, TMG-C, GP and SU estimators of β_0 ($E(\beta_i) = \beta_0 = 1$) in panel data models with time effects, correlated heterogeneity, $\rho_\beta = 0.5$, and the level of overall fit, $PR^2 = 0.4$

	$T = 2$				$T = 3$			
Estimator	$\hat{\pi}$ (×100)	Bias	RMSE	Size (×100)	$\hat{\pi}$ (×100)	Bias	RMSE	Size (×100)
				$n = 1,000$				
TMG-TE	31.2	0.049	0.22	5.5	16.5	0.023	0.13	5.6
TMG-C	...	...	...	...	16.5	0.023	0.13	5.5
GP	4.2	-0.018	0.52	4.0	2.0	0.001	0.17	4.6
SU	4.2	-0.032	1.02	5.4	...	...	...	...
				$n = 2,000$				
TMG-TE	28.5	0.044	0.17	5.6	14.1	0.020	0.10	5.7
TMG-C	...	...	...	...	14.1	0.020	0.10	5.7
GP	3.4	0.022	0.44	5.2	1.3	0.004	0.13	4.7
SU	3.4	0.007	0.86	5.8	...	...	...	...
				$n = 5,000$				
TMG-TE	24.7	0.037	0.12	5.6	10.8	0.016	0.07	6.0
TMG-C	...	...	...	...	10.8	0.016	0.07	6.0
GP	2.5	0.007	0.32	5.0	0.7	0.000	0.09	5.1
SU	2.5	0.004	0.63	4.7	...	...	...	...
				$n = 10,000$				
TMG-TE	22.1	0.030	0.09	6.9	8.8	0.013	0.05	6.1
TMG-C	...	...	...	...	8.8	0.013	0.05	6.1
GP	2.0	0.003	0.25	4.3	0.5	0.000	0.07	4.9
SU	2.0	0.007	0.50	5.4	...	...	...	...

Notes: (i) The data generating process is given by $y_{it} = \alpha_i + \beta_i x_{it} + \sigma_{it} e_{it}$ with random heteroskedasticity, ρ_β is defined by (8.6) in the main paper, and PR^2 is defined by (S.3.3). Time effects are set as $\phi_t = t$ for $t = 1, 2, ..., T-1$, and $\phi_T = -T(T-1)/2$. The errors processes for y_{it} and x_{it} equations are chi-squared and Gaussian, respectively, and x_{it} are generated as heterogeneous AR(1) processes. For further details see Section 8.1.3 in the main paper and Section S.3. (ii) The TMG-TE estimators of $\boldsymbol{\theta}_0$ and $\boldsymbol{\phi}$ are given by (6.9) and (6.11) in the main paper, respectively, and their asymptotic variances are given by (A.3.4) and (A.3.7), respectively. The TMG-C estimators of $\boldsymbol{\theta}_0$ and $\boldsymbol{\phi}$ are given by (6.20) and (6.17) in the main paper, respectively, and their asymptotic variances are given by (6.21) and (6.19) in the main paper, respectively. For the trimming threshold, see footnote (iii) to Table S.3. (iii) GP and SU estimators are proposed by Graham and Powell (2012) and Sasaki and Ura (2021). For their trimming threshold, see footnote (i) to Table S.5. $\hat{\pi}$ is the simulated fraction of individual estimates being trimmed, defined by (4.7) in the main paper. "..." denotes the estimation algorithms are not available or not applicable.

Table S.9: Bias, RMSE and size of TMG-TE and GP estimators of the time effects, ϕ_1 and ϕ_2, in panel data model with correlated heterogeneity, $\rho_\beta = 0.5$, and the overall fit, $PR^2 = 0.4$

		$n = 1,000$			$n = 5,000$		
	Estimator	Bias	RMSE	Size (×100)	Bias	RMSE	Size (×100)
			$T = 2$			$T = 2$	
$\phi_1 = 1$	TMG-TE	0.002	0.06	6.1	-0.001	0.02	4.7
	GP	0.001	0.33	7.1	-0.005	0.21	6.9
			$T = 3$			$T = 3$	
$\phi_1 = 1$	TMG-TE	0.165	0.00	6.5	0.108	0.00	2.9
	GP	0.020	0.00	9.1	0.007	0.00	4.1
$\phi_2 = 2$	TMG-TE	0.165	0.00	6.5	0.108	0.00	2.9
	GP	0.020	-0.01	8.1	0.007	0.00	3.7

Notes: See the notes to Table S.8.

Table S.10: Bias, RMSE and size of TMG, GP and TMG-TE estimators of β_0 ($E(\beta_i) = \beta_0 = 1$) under different model specifications with correlated heterogeneity, $\rho_\beta = 0.5$, and the level of overall fit, $PR^2 = 0.2$, for $T = 2$

	$n = 1,000$				$n = 2,000$				$n = 5,000$				$n = 10,000$			
Estimator	$\hat{\pi}$ (×100)	Bias	RMSE	Size (×100)	$\hat{\pi}$ (×100)	Bias	RMSE	Size (×100)	$\hat{\pi}$ (×100)	Bias	RMSE	Size (×100)	$\hat{\pi}$ (×100)	Bias	RMSE	Size (×100)
I. Without time effects in y_{it}																
(a) Gaussian errors in y_{it}																
TMG	27.3	0.051	0.263	5.3	24.5	0.057	0.200	5.7	21.1	0.039	0.139	6.2	18.9	0.034	0.108	7.3
GP	4.0	0.015	0.558	4.6	3.2	0.024	0.448	4.2	2.4	0.001	0.343	5.6	1.9	0.002	0.274	5.6
(b) Chi-squared errors in y_{it}																
TMG	27.3	0.051	0.260	5.8	24.5	0.049	0.199	5.1	21.1	0.039	0.138	5.8	18.9	0.033	0.105	6.2
GP	4.0	0.008	0.569	4.9	3.2	0.012	0.451	4.8	2.4	0.003	0.338	4.7	1.9	-0.003	0.267	4.7
(c) Heterogeneous AR(1) x_{it} processes																
TMG	31.2	0.048	0.353	4.9	28.5	0.044	0.272	5.3	24.7	0.037	0.184	4.7	22.1	0.029	0.143	5.6
GP	4.2	-0.029	0.830	4.5	3.4	0.031	0.701	5.8	2.5	0.009	0.523	5.2	2.0	0.002	0.408	4.3
(d) Heterogeneous AR(1) x_{it} processes with interactive effects																
TMG	29.9	0.037	0.413	5.4	26.9	0.028	0.309	4.7	23.4	0.025	0.213	5.1	20.8	0.019	0.162	5.7
GP	4.2	-0.017	0.938	4.9	3.3	0.002	0.765	4.8	2.5	0.017	0.584	5.4	1.9	-0.010	0.459	4.8
II. With time effects in y_{it}																
(a) Chi-squared errors in y_{it}																
TMG-TE	27.3	0.055	0.256	4.9	24.5	0.060	0.209	7.1	21.2	0.038	0.141	6.8	18.9	0.038	0.109	7.2
GP	4.1	-0.005	0.588	4.8	3.2	0.026	0.456	4.2	2.4	-0.001	0.356	5.4	1.9	0.009	0.271	5.4
(b) Heterogeneous AR(1) x_{it} processes																
TMG-TE	31.2	0.048	0.353	5.0	28.5	0.044	0.272	5.6	24.7	0.037	0.184	4.7	22.1	0.028	0.143	5.7
GP	4.2	-0.034	0.844	3.9	3.4	0.032	0.712	5.2	2.5	0.008	0.527	5.0	2.0	0.003	0.410	4.4
(c) Heterogeneous AR(1) x_{it} processes with interactive effects																
TMG-TE	29.9	0.037	0.433	5.2	26.9	0.029	0.319	4.2	23.4	0.026	0.220	4.9	20.8	0.020	0.167	5.5
GP	4.2	-0.017	0.968	4.4	3.3	-0.004	0.791	4.4	2.5	0.014	0.592	4.7	1.9	-0.008	0.470	4.7

Notes: (i) The data generating process is given by $y_{it} = \alpha_i + \phi_t + \beta_i x_{it} + \sigma_{it} e_{it}$ with random heteroskedasticity. Time effects are set as $\phi_t = t$ for $t = 1, 2, ..., T-1$, and $\phi_T = -T(T-1)/2$. For further details see Section 8.1.3 in the main paper and Section S.3. (ii) For the TMG estimator, see footnote (ii) to Table S.3. For the GP estimator, see footnote (i) to Table S.5. For the TMG-TE estimator, see footnote (ii) to Table S.8. $\hat{\pi}$ is the simulated fraction of individual estimates being trimmed, defined by (4.7) in the main paper.

S.6 Monte Carlo evidence when the errors of x_{it} are uniformly distributed

In this section, the DGP is generated using the baseline model without time effects, but with the errors in the x_{it} equation drawn from a uniform distribution, as compared to the Gaussian errors used in the baseline model. Table S.11 compares small sample performance of the TMG estimator of β_0 using different $\alpha \in \{1/3, 0.35, 1/2\}$ for threshold values with correlated heterogeneity, $\rho_\beta = 0.5$ for $T = 2, 3$, and $n = 1,000$, $2,000$, $5,000$ and $10,000$. Table S.12 reports bias, RMSE and size of FE, MG and TMG estimators of β_0 for $T = 2, 3, 4, 5, 6, 8$, and $n = 1,000$, $2,000$, $5,000$ and $10,000$.

The choice of the error distribution does not seem to be consequential.

Table S.11: Bias, RMSE and size of the TMG estimator of β_0 ($E(\beta_i) = \beta_0 = 1$) for different values of the threshold parameter, α, in the baseline model without time effects and with correlated heterogeneity, $\rho_\beta = 0.5$, using uniformly distributed errors in the x_{it} equation

		$T=2$				$T=3$			
Estimator	α	$\hat{\pi}$ ($\times$100)	Bias	RMSE	Size ($\times$100)	$\hat{\pi}$ ($\times$100)	Bias	RMSE	Size ($\times$100)
					$n = 1,000$				
TMG	1/3	27.0	0.051	0.33	5.0	12.6	0.027	0.20	5.2
TMG	0.35	25.5	0.048	0.34	5.0	11.3	0.025	0.20	5.3
TMG	0.50	15.2	0.025	0.43	4.4	4.0	0.012	0.23	4.9
					$n = 2,000$				
TMG	1/3	24.4	0.045	0.26	5.2	10.5	0.024	0.15	5.6
TMG	0.35	22.9	0.042	0.27	5.2	9.3	0.022	0.15	5.1
TMG	0.50	13.0	0.024	0.34	5.0	3.0	0.012	0.18	5.1
					$n = 5,000$				
TMG	1/3	20.9	0.041	0.18	6.0	7.7	0.015	0.10	5.0
TMG	0.35	19.5	0.038	0.19	5.9	6.7	0.013	0.10	5.3
TMG	0.50	10.3	0.020	0.25	5.0	1.9	0.006	0.12	4.9
					$n = 10,000$				
TMG	1/3	18.7	0.036	0.13	5.8	6.2	0.011	0.07	5.0
TMG	0.35	17.4	0.034	0.14	5.8	5.3	0.009	0.07	5.2
TMG	0.50	8.7	0.024	0.19	4.2	1.3	0.002	0.09	5.0

Notes: (i) The baseline model is generated as $y_{it} = \alpha_i + \beta_i x_{it} + u_{it}$, where the errors processes for y_{it} and x_{it} equations are chi-squared and uniformly distributed, respectively, x_{it} are generated as heterogeneous AR(1) processes, and ρ_β is defined by (8.6) in the main paper. For further details see Section 8.1.3 in the main paper and Section S.3. (ii) For the TMG estimator and its threshold, see footnotes (ii) and (iii) to Table S.3. $\hat{\pi}$ is the simulated fraction of individual estimates being trimmed, defined by (4.7) in the main paper.

Table S.12: Bias, RMSE and size of FE, MG and TMG estimators of β_0 ($E(\beta_i) = \beta_0 = 1$) in the baseline model without time effects, using uniformly distributed errors in the x_{it} equation

	Uncorrelated heterogeneity: $\rho_\beta = 0$										Correlated heterogeneity: $\rho_\beta = 0.5$									
	$\hat{\pi}$ (×100)	Bias			RMSE			Size (×100)			$\hat{\pi}$ (×100)	Bias			RMSE			Size (×100)		
T	TMG	FE	MG	TMG	FE	MG	TMG	FE	MG	TMG	TMG	FE	MG	TMG	FE	MG	TMG	FE	MG	TMG
										$n = 1,000$										
2	27.0	0.003	-4.499	0.001	0.17	188.74	0.31	5.5	2.0	4.8	27.0	0.367	-4.535	0.051	0.40	190.25	0.32	60.1	2.0	5.2
3	12.6	0.002	-0.015	0.005	0.11	0.61	0.19	4.7	4.3	5.3	12.6	0.243	-0.015	0.027	0.27	0.62	0.19	57.6	4.6	5.1
4	7.3	0.000	0.005	0.003	0.09	0.17	0.13	5.2	4.8	4.7	7.3	0.189	0.004	0.014	0.21	0.17	0.14	54.1	4.8	5.1
5	4.8	-0.003	0.000	-0.001	0.08	0.12	0.11	6.2	5.3	5.4	4.8	0.158	0.000	0.005	0.18	0.12	0.11	50.6	5.3	5.7
6	3.4	-0.001	-0.002	-0.001	0.07	0.10	0.09	4.3	5.5	5.2	3.4	0.142	-0.001	0.003	0.16	0.10	0.09	51.7	5.5	5.6
8	2.1	-0.002	-0.001	0.000	0.06	0.08	0.08	5.4	5.6	5.4	2.1	0.117	0.000	0.002	0.13	0.08	0.08	48.9	5.6	5.1
										$n = 2,000$										
2	24.4	-0.003	1.103	0.002	0.12	69.57	0.24	4.9	2.0	5.0	24.4	0.361	1.112	0.045	0.38	70.12	0.25	84.0	2.0	5.2
3	10.5	0.000	0.003	0.006	0.08	0.45	0.14	5.6	5.1	5.1	10.5	0.240	0.003	0.024	0.25	0.46	0.14	81.9	5.1	5.6
4	5.7	0.000	0.003	0.002	0.06	0.13	0.10	4.2	4.8	4.7	5.7	0.188	0.003	0.010	0.20	0.13	0.10	81.3	4.6	4.6
5	3.6	-0.002	-0.006	-0.005	0.06	0.09	0.08	5.1	4.6	4.6	3.6	0.159	-0.006	0.000	0.17	0.09	0.08	77.8	4.4	4.6
6	2.6	-0.001	-0.002	-0.002	0.05	0.08	0.07	4.9	5.9	6.2	2.6	0.140	-0.002	0.001	0.15	0.08	0.07	78.1	5.9	6.0
8	1.5	0.000	-0.001	-0.001	0.04	0.06	0.06	4.7	5.1	5.1	1.5	0.118	-0.001	0.001	0.13	0.06	0.06	77.9	5.2	5.3
										$n = 5,000$										
2	20.9	0.001	-0.277	0.004	0.07	171.87	0.17	4.5	2.2	5.1	20.9	0.365	-0.279	0.041	0.37	173.24	0.17	99.7	2.2	6.1
3	7.7	0.001	0.002	0.001	0.05	0.25	0.10	5.4	3.9	4.6	7.7	0.240	0.003	0.015	0.25	0.25	0.10	99.6	3.8	5.0
4	3.6	0.002	0.000	0.001	0.04	0.08	0.07	5.2	5.1	4.5	3.6	0.190	0.000	0.007	0.19	0.08	0.07	99.4	5.0	4.7
5	2.1	0.000	0.001	0.001	0.04	0.06	0.05	3.9	5.4	4.7	2.1	0.160	0.001	0.004	0.16	0.06	0.05	99.2	5.1	4.8
6	1.3	0.000	0.000	0.000	0.03	0.05	0.05	5.1	5.1	4.9	1.3	0.141	0.000	0.001	0.14	0.05	0.05	99.3	4.9	4.4
8	0.7	0.000	-0.001	-0.001	0.03	0.04	0.04	5.3	4.9	4.5	0.7	0.118	-0.001	0.000	0.12	0.04	0.04	98.9	5.0	4.6
										$n = 10,000$										
2	18.7	0.000	5.925	0.003	0.06	223.71	0.12	5.8	2.5	4.8	18.7	0.366	5.972	0.036	0.37	225.49	0.13	100.0	2.5	5.9
3	6.2	0.000	-0.006	0.000	0.04	0.21	0.07	4.1	4.9	4.7	6.2	0.239	-0.006	0.011	0.24	0.21	0.07	100.0	4.8	5.0
4	2.6	0.000	0.002	0.000	0.03	0.06	0.05	5.4	5.6	5.6	2.6	0.189	0.002	0.004	0.19	0.06	0.05	100.0	5.2	5.6
5	1.4	0.000	0.001	0.001	0.03	0.04	0.04	4.2	4.9	5.6	1.4	0.160	0.001	0.003	0.16	0.04	0.04	100.0	5.3	5.2
6	0.8	0.000	0.000	0.000	0.02	0.03	0.03	5.0	5.0	5.2	0.8	0.142	0.000	0.001	0.14	0.03	0.03	100.0	5.0	5.3
8	0.4	0.000	0.001	0.001	0.02	0.03	0.03	4.9	4.4	4.6	0.4	0.119	0.001	0.001	0.12	0.03	0.03	100.0	4.3	4.3

Notes: (i) For details of the baseline model, see footnote (i) to Table S.11. (ii) For FE and MG estimators, see footnote (ii) to Table S.3. For the TMG estimator and its threshold, see footnotes (ii) and (iii) to Table S.3. $\hat{\pi}$ is the simulated fraction of individual estimates being trimmed, defined by (4.7) in the main paper.

S.7 Monte Carlo evidence of estimation with interactive effects in the regressor

To further examine the robustness of TMG-TE and TMG-C estimators to the choice of DGP, we generate the regressor, x_{it}, with interactive effects, namely

$$x_{it} = \alpha_{ix}(1-\rho_{ix}) + \gamma_{ix} f_t + \rho_{ix} x_{i,t-1} + \left(1-\rho_{ix}^2\right)^{1/2} u_{x,it},$$

where $\gamma_{ix} \sim IIDU(0,2)$, $f_t = 0.9 f_{t-1} + (1-0.9^2)^{1/2} v_t$, and $v_t \sim IIDN(0,1)$ for $t = -49, -48, ..., -1, 0, 1, ..., T$, with $f_{-50} = 0$. We also calibrate κ in the outcome equation given by (8.1) in the main paper, to achieve $PR^2 = 0.2$ by stochastic simulation (see Table S.2). The rest of the parameters are set as in the baseline model. See Section 8.1.3 in the main paper for details.

Tables S.13–S.14 summarize the results for TMG-TE, TMG-C, GP and SU estimators of β_0 and the time effect ϕ_1 when $T = 2$, and the time effects ϕ_1 and ϕ_2 when $T = 3$, for $T = 2, 3$ and $n = 1,000$, $2,000$, $5,000$ and $10,000$.

The comparative performance of TMG-TE and GP estimators is unaffected by the addition of interactive effects to the x_{it} process. The inclusion of interactive effects, by increasing the variance of x_{it}, results in improved estimates with a higher degree of precision, and a smaller number of estimates being trimmed. This can be seen in the estimates of $\hat{\pi}$ (the fraction of estimates trimmed) which are slightly lower than those reported in Table 3 in the main paper for the baseline model. The bias is also slightly smaller but the RMSE is larger. The RMSE of TMG-TE estimator of ϕ_1 and ϕ_2 are also higher as compared with those reported in Table S.14 in the main paper.

The comparative empirical power functions for β_0 and ϕ_1 and ϕ_2 are shown in Figures S.6–S.8. As can be seen from Figure S.6, the TMG-C estimator of β_0 has *marginally* higher power than the TMG-TE estimator when x_{it} includes interactive effects, compared to the baseline model without interactive effects where TMG-TE and TMG-C estimators have very similar power functions. See Figure S.9.

Table S.13: Bias, RMSE and size of TMG-TE, TMG-C, GP and SU estimators of β_0 ($E(\beta_i) = \beta_0 = 1$) in panel data models with time effects, correlated heterogeneity, $\rho_\beta = 0.5$, and interactive effects in the x_{it} equation

	$T=2$				$T=3$			
Estimator	$\hat{\pi}$ (×100)	Bias	RMSE	Size (×100)	$\hat{\pi}$ (×100)	Bias	RMSE	Size (×100)
				$n = 1,000$				
TMG-TE	29.9	0.037	0.43	5.2	13.1	0.022	0.26	4.8
TMG-C	...	...	...	...	13.1	0.020	0.25	5.0
GP	4.2	-0.017	0.97	4.4	0.5	0.016	0.36	5.1
SU	4.2	-0.060	1.93	5.9	...	...	...	...
				$n = 2,000$				
TMG-TE	26.9	0.029	0.32	4.2	10.7	0.012	0.19	3.8
TMG-C	...	...	...	...	10.7	0.011	0.18	4.2
GP	3.3	-0.004	0.79	4.4	0.4	0.000	0.27	4.9
SU	3.3	-0.012	1.55	5.2	...	...	...	...
				$n = 5,000$				
TMG-TE	23.4	0.026	0.22	4.9	8.1	0.009	0.13	3.8
TMG-C	...	...	...	...	8.1	0.008	0.12	4.3
GP	2.5	0.014	0.59	4.7	0.2	0.002	0.18	4.2
SU	2.5	0.059	1.15	4.9	...	...	...	...
				$n = 10,000$				
TMG-TE	20.8	0.020	0.17	5.5	6.5	0.008	0.10	4.7
TMG-C	...	...	...	...	6.5	0.007	0.09	5.6
GP	1.9	-0.008	0.47	4.7	0.1	0.001	0.13	5.5
SU	1.9	-0.030	0.91	4.9	...	...	...	...

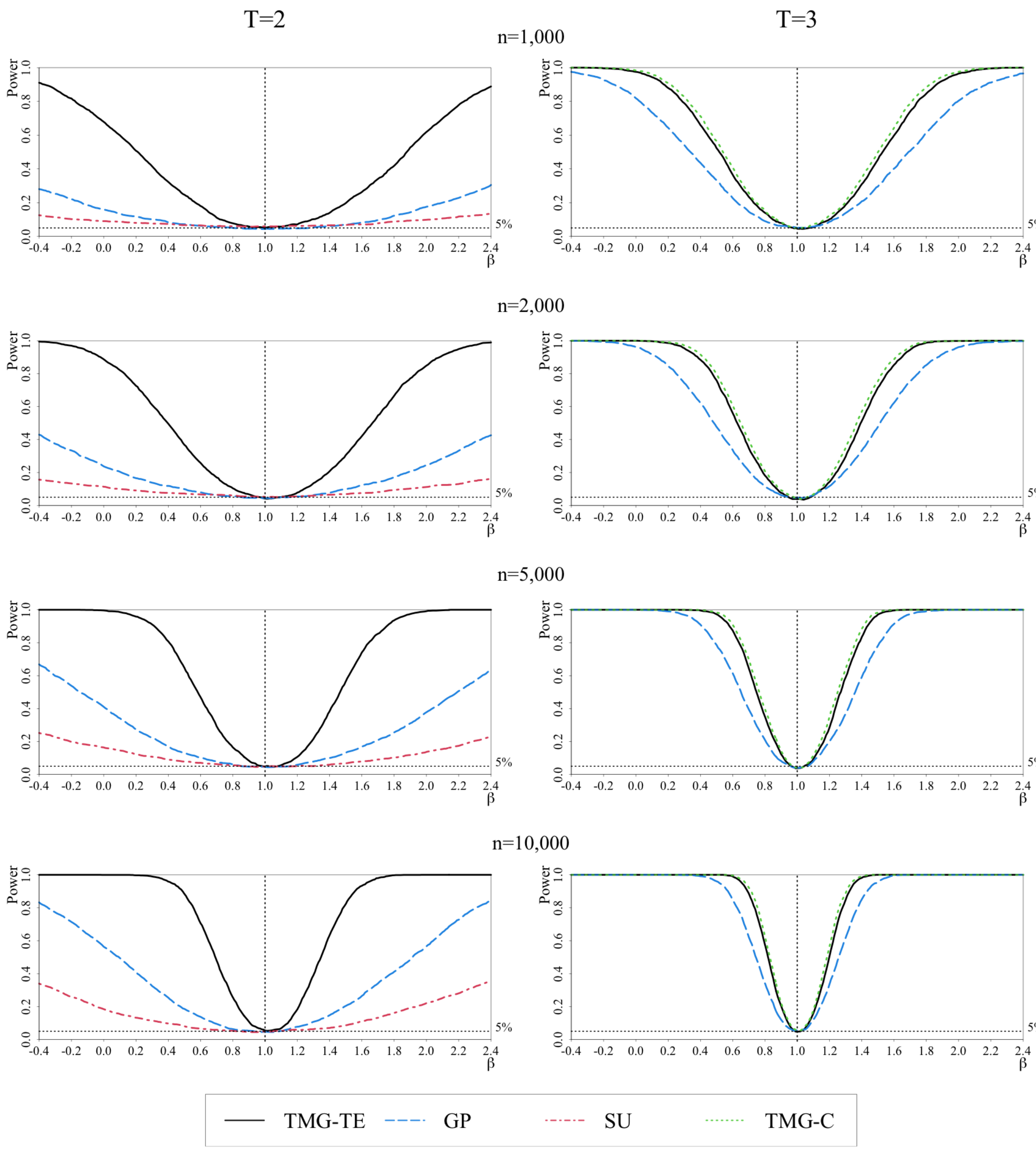
Notes: (i) The data generating process is given by $y_{it} = \alpha_i + \phi_t + \beta_i x_{it} + \sigma_{it} e_{it}$ with random heteroskedasticity, x_{it} are generated as heterogeneous AR(1) processes with interactive effects, and ρ_β is defined by (8.6) in the main paper. Time effects are set as $\phi_t = t$ for $t = 1, 2, ..., T-1$, and $\phi_T = -T(T-1)/2$. The errors processes for y_{it} and x_{it} equations are chi-squared and Gaussian, respectively. For further details see Section 8.1.3 in the main paper. (ii) For TMG-TE and TMG-C estimators, see footnote (ii) to Table S.8. For GP and SU estimators, see footnote (iii) to Table 3. $\hat{\pi}$ is the simulated fraction of individual estimates being trimmed, given by (4.7) in the main paper. "..." denotes the estimation algorithms are not applicable or available.

Table S.14: Bias, RMSE and size of TMG-TE and GP estimators of the time effects, ϕ_1 and ϕ_2, in panel data models with correlated heterogeneity, $\rho_\beta = 0.5$, and interactive effects in the x_{it} equation

		$n = 1,000$			$n = 5,000$		
	Estimator	Bias	RMSE	Size (×100)	Bias	RMSE	Size (×100)
			$T=2$			$T=2$	
$\phi_1 = 1$	TMG-TE	0.000	0.13	4.6	-0.005	0.06	4.5
	GP	0.017	0.72	7.6	0.007	0.44	6.2
			$T=3$			$T=3$	
$\phi_1 = 1$	TMG-TE	0.131	0.00	13.2	0.081	0.00	6.0
	GP	0.005	0.01	18.1	0.002	0.00	8.4
$\phi_2 = 2$	TMG-TE	0.131	0.01	14.9	0.081	0.00	6.7
	GP	0.005	0.00	17.5	0.002	0.00	7.9

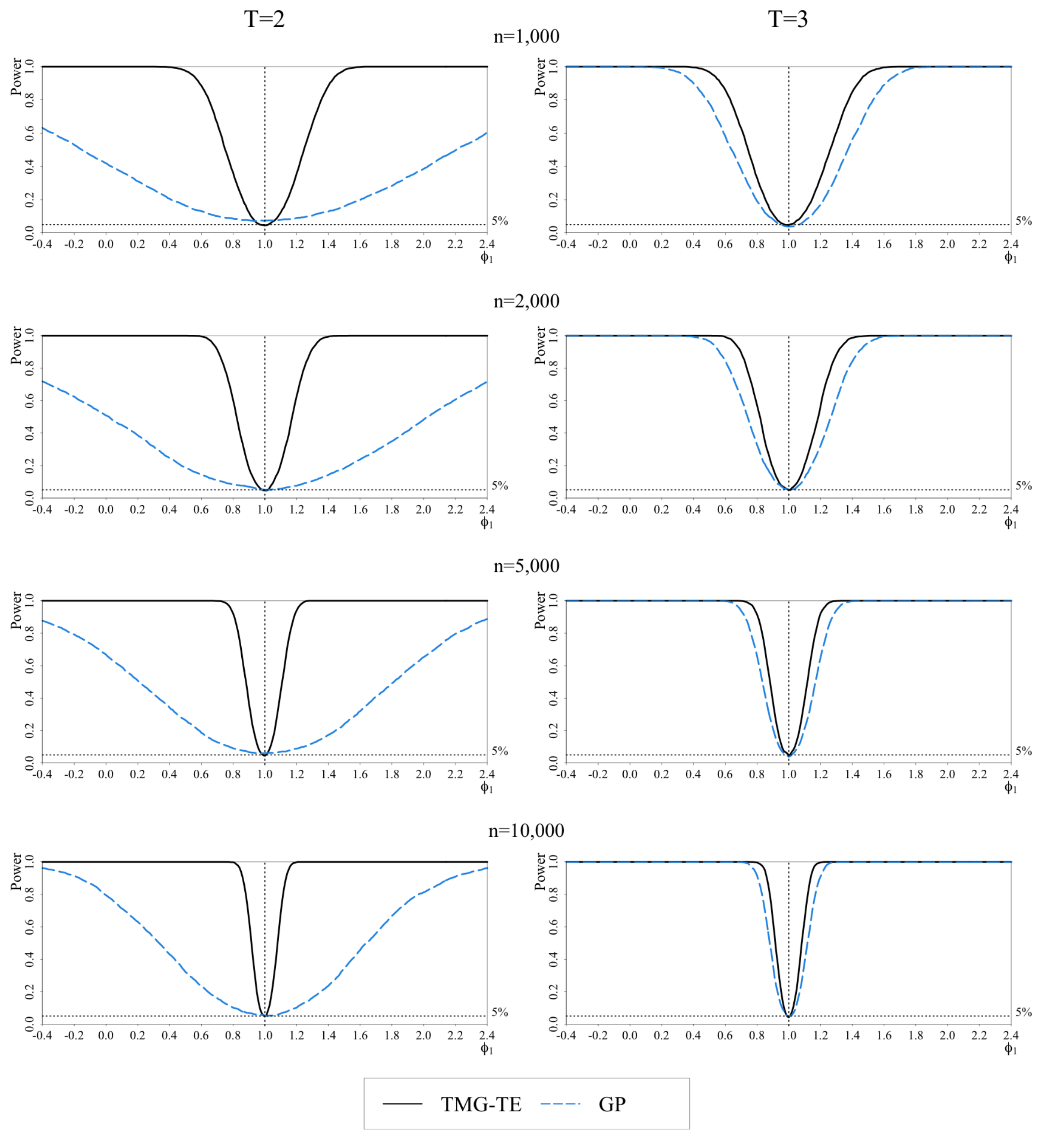
Notes: For details of models with time effects, where x_{it} are generated as heterogeneous AR(1) processes with interactive effects, see footnote (i) to Table S.13. For the TMG-TE estimator, see footnote (ii) to Table S.8. For the GP estimator, see footnote (i) to Table S.5.

Figure S.6: Empirical power functions for TMG-TE, GP, SU and TMG-C estimators of β_0 ($E(\beta_i) = \beta_0 = 1$) in panel data models with time effects, correlated heterogeneity, $\rho_\beta = 0.5$, and interactive effects in the x_{it} equation



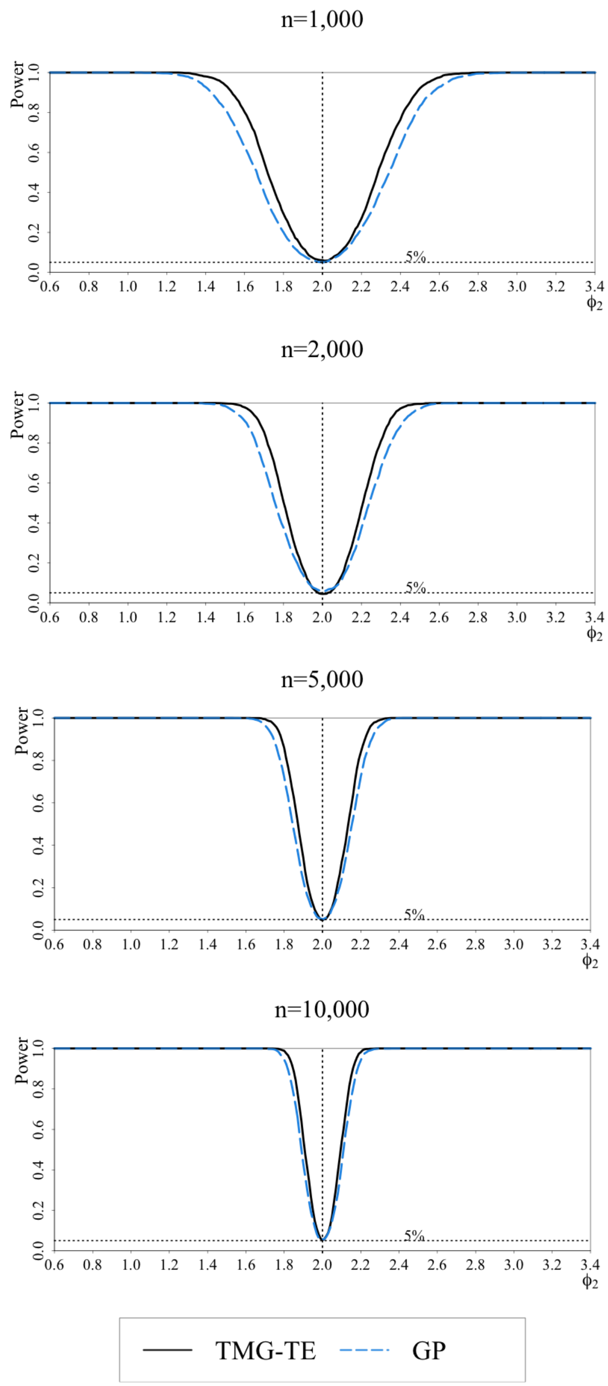
Notes: For details of models with time effects, where x_{it} are generated as heterogeneous AR(1) processes with interactive effects, see footnote (i) to Table S.13. For the TMG-TE estimator, see footnote (ii) to Table S.8. For GP and SU estimators, see footnote (iii) to Table S.8.

Figure S.7: Empirical power functions for TMG-TE and GP estimators of the time effect $\phi_1 = 1$ in panel data models with correlated heterogeneity, $\rho_\beta = 0.5$, and interactive effects in the x_{it} equation



Notes: For details of models with time effects, where x_{it} are generated as heterogeneous AR(1) processes with interactive effects, see footnote (i) to Table S.13. For the TMG-TE estimator, see footnote (ii) to Table S.8. For the GP estimator, see footnote (iii) to Table S.8.

Figure S.8: Empirical power functions for TMG-TE and GP estimators of the time effect $\phi_2 = 2$ in panel data models with $T = 3$, correlated heterogeneity, $\rho_\beta = 0.5$, and interactive effects in the x_{it} equation

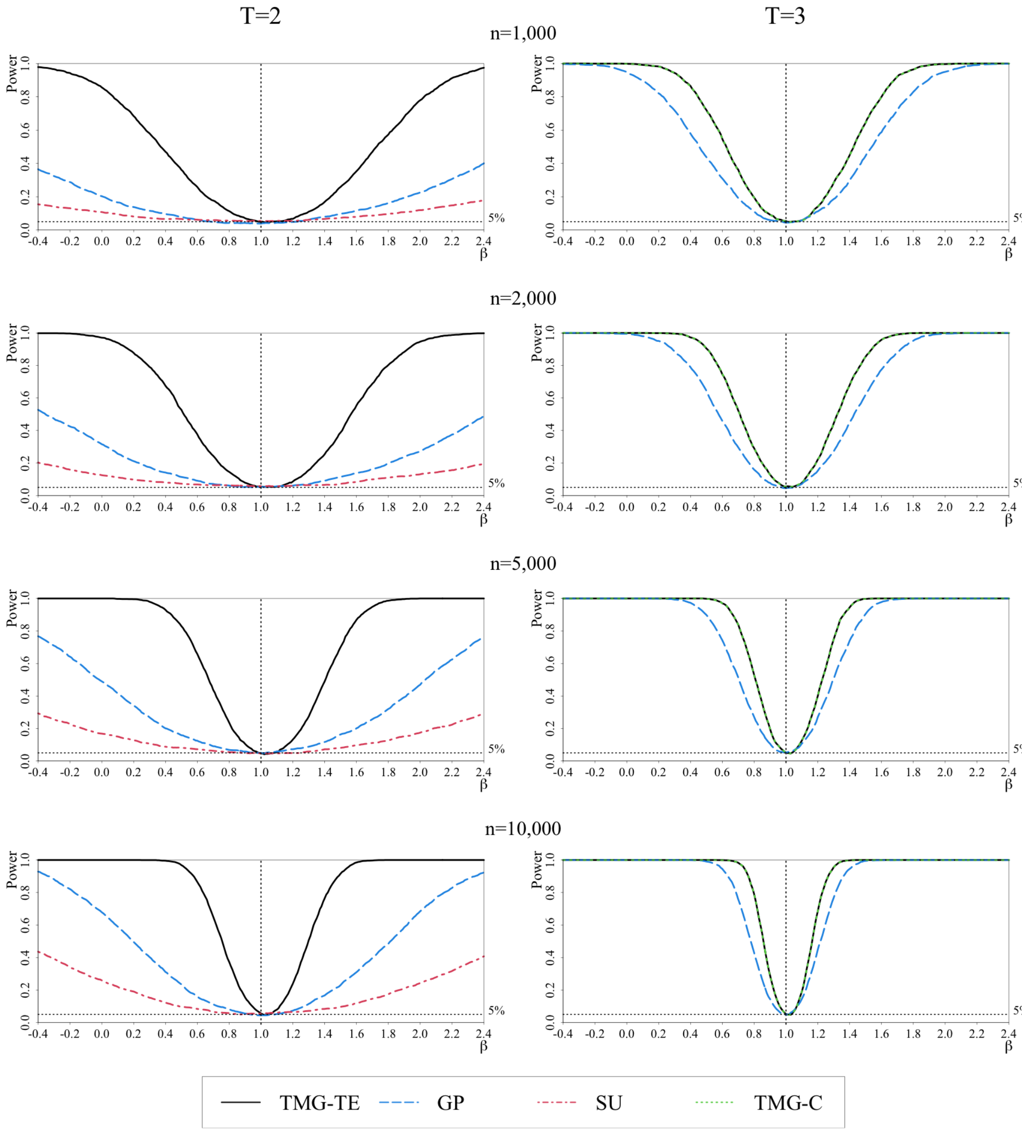


Notes: For details of models with time effects, where x_{it} are generated as heterogeneous AR(1) processes with interactive effects, see footnote (i) to Table S.13. For the TMG-TE estimator, see footnote (ii) to Table S.8. For the GP estimator, see footnote (iii) to Table S.8.

S.8 Empirical power functions for TMG-TE, TMG-C, GP and SU estimators in the baseline model with time effects and correlated heterogeneity

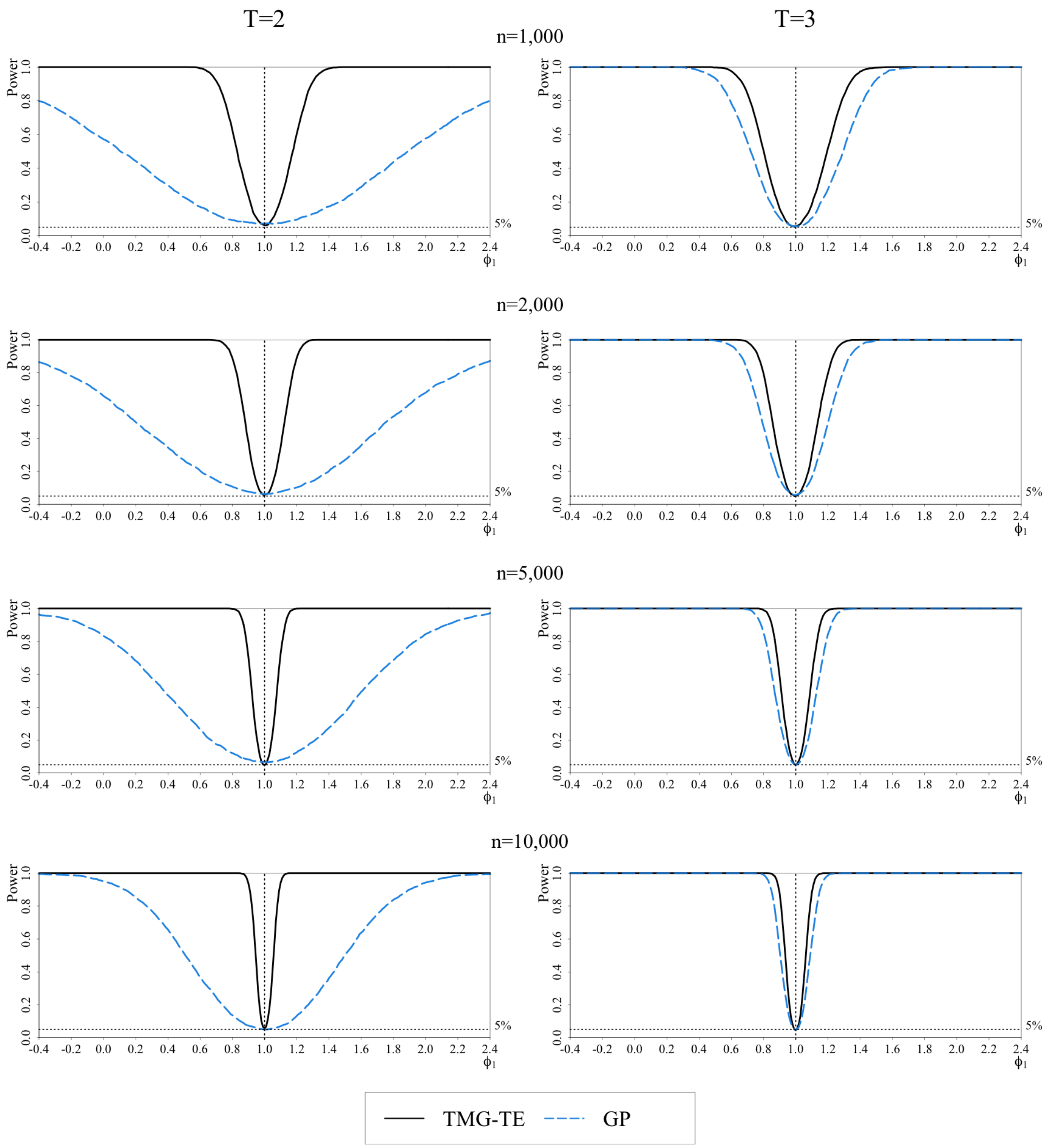
Figures S.9–S.11 show empirical power functions for TMG-TE, GP, SU (for $T = 2$) and TMG-C (for $T = 3$) estimators of β_0, and the time effects, ϕ_1 and ϕ_2, for the baseline model with correlated heterogeneity, $\rho_\beta = 0.5$, as discussed in Section 8.2.2 of the main paper.

Figure S.9: Empirical power functions for TMG-TE, GP, SU and TMG-C estimators of β_0 ($E(\beta_i) = \beta_0 = 1$) in the baseline model with time effects and correlated heterogeneity, $\rho_\beta = 0.5$



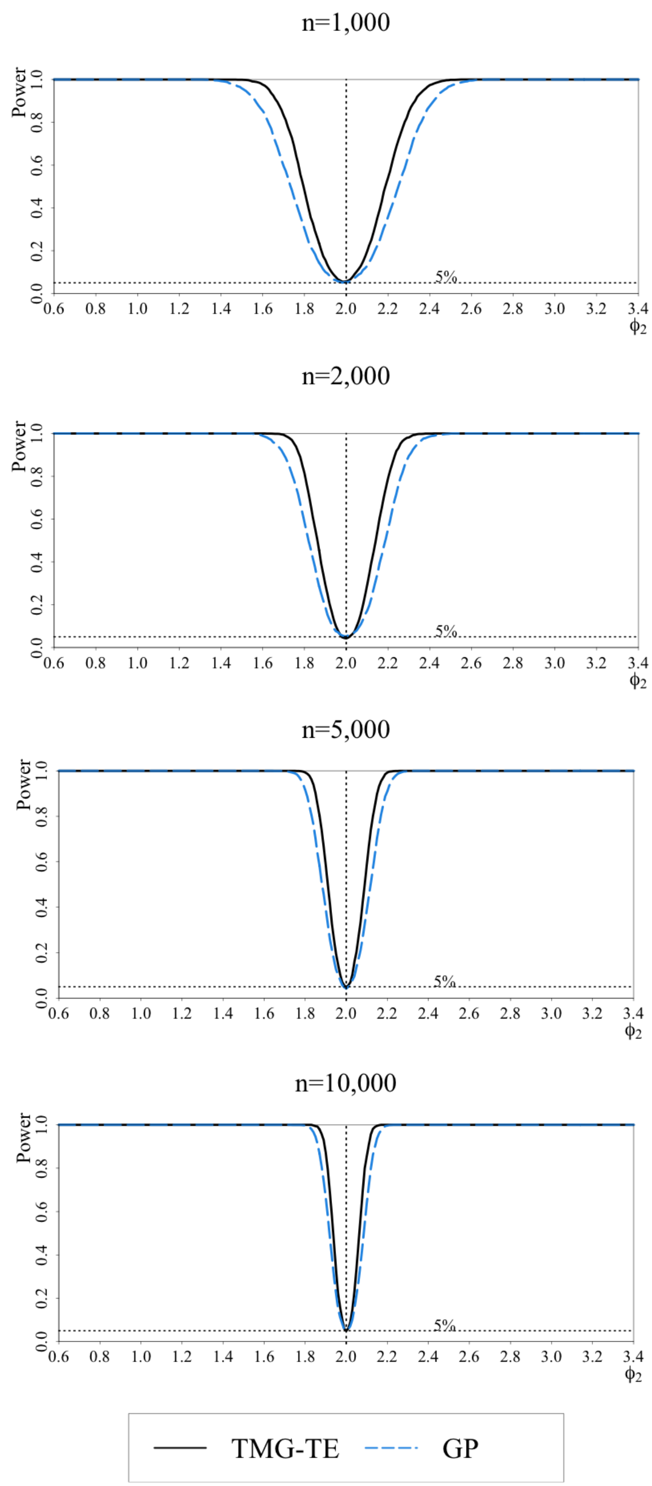
Notes: For details of the baseline model with time effects, see footnote (i) to Table 3 in the main paper.

Figure S.10: Empirical power functions for TMG-TE and GP estimators of the time effect $\phi_1 = 1$ in the baseline model with correlated heterogeneity, $\rho_\beta = 0.5$



Notes: For details of the baseline model with time effects, see footnote (i) to Table 3 in the main paper.

Figure S.11: Empirical power functions for TMG-TE and GP estimators of the time effect $\phi_2 = 2$ in the baseline model with $T = 3$ and correlated heterogeneity, $\rho_\beta = 0.5$



Notes: For details of the baseline model with time effects, see footnote (i) to Table 3 in the main paper.

S.9 Monte Carlo evidence of the Hausman-type test of correlated heterogeneity in panel data models with time effects

Table S.15: Empirical size and power of the Hausman-type test of correlated heterogeneity in the baseline model with time effects

	Under H_0								Under H_1			
	Homogeneity: $\sigma^2_\beta = 0$				Uncorrelated hetro.: $\rho_\beta = 0$, $\sigma^2_\beta = 0.5$				Correlated hetro.: $\rho_\beta = 0.5$, $\sigma^2_\beta = 0.5$			
T/n	1,000	2,000	5,000	10,000	1,000	2,000	5,000	10,000	1,000	2,000	5,000	10,000
TMG-TE												
2	4.3	4.4	5.3	5.5	5.7	5.4	4.9	5.9	25.0	37.1	67.0	89.5
3	4.5	5.6	5.8	4.8	5.5	4.5	5.8	4.7	39.6	59.9	90.0	99.3
4	4.9	5.7	5.6	4.9	5.0	4.2	5.5	4.7	52.1	75.5	97.3	100.0
5	5.2	4.6	4.7	5.2	4.4	4.5	4.5	4.4	61.9	84.5	99.4	100.0
6	4.7	5.2	5.5	5.3	5.2	4.3	4.6	4.4	70.4	90.9	100.0	100.0
8	5.0	4.7	5.0	5.3	4.6	4.9	5.9	5.8	79.7	95.6	100.0	100.0
TMG-C												
3	4.6	5.7	5.7	4.8	5.3	4.4	5.7	4.8	39.7	59.8	89.9	99.2
4	5.0	5.6	5.7	4.9	4.9	4.5	5.5	4.7	51.9	75.6	97.3	100.0
5	5.2	4.6	4.7	5.3	4.3	4.4	4.5	4.5	62.0	84.5	99.4	100.0
6	4.8	5.2	5.5	5.3	5.2	4.3	4.7	4.5	70.8	91.0	100.0	100.0
8	4.9	4.8	5.0	5.3	4.7	5.0	5.8	5.8	80.0	95.6	100.0	100.0

Notes: (i) The baseline model for the test is generated as $y_{it} = \alpha_i + \phi_t + \beta_i x_{it} + u_{it}$, with α_i correlated with x_{it} under both the null and alternative hypotheses. Time effects are set as $\phi_t = t$ for $t = 1, 2, ..., T-1$, and $\phi_T = -T(T-1)/2$. The errors processes for y_{it} and x_{it} equations are chi-squared and Gaussian, respectively, and x_{it} are generated as heterogeneous AR(1) processes. For further details see Section 8.1.3 in the main paper. (ii) The null hypothesis is given by (7.1) in the main paper, including the case of homogeneity with $\sigma^2_\beta = 0$ and the case of uncorrelated heterogeneity with $\rho_\beta = 0$ (the degree of correlated heterogeneity defined by (8.6) in the main paper) and $\sigma^2_\beta = 0.5$. The alternative of correlated heterogeneity is generated with $\rho_\beta = 0.5$ and $\sigma^2_\beta = 0.5$. (iii) The test statistics for panels with time effects are based on the difference between the FE-TE and TMG-TE estimator given by (S.2.15) with $T \geq k$, and the difference between the FE-TE and TMG-C estimators given by (S.2.21) with $T > k$. For further details see Section S.2. Size and power are in per cent.

References

Pesaran, M. H. and T. Yamagata (2023). Testing for Alpha in linear factor pricing models with a large number of securities. Published online in *Journal of Financial Econometrics.*

Graham, B. S. and J. L. Powell (2012). Identification and estimation of average partial effects in “irregular”correlated random coefficient panel data models. *Econometrica* 80, 2105–2152.

Sasaki, Y. and T. Ura (2021). Slow Movers in Panel Data. arXiv preprint arXiv:2110.12041.
.