ARTICLE IN PRESS

Journal of Econometrics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Indirect inference estimation of dynamic panel data models

Yong Bao [a,*], Xuewen Yu [b]

[a] Department of Economics, Purdue University, 403 W State St, West Lafayette, IN 47907, USA
[b] Department of Applied Economics, School of Management, Fudan University, China

ARTICLE INFO

Article history:
Received 29 July 2021
Received in revised form 13 September 2022
Accepted 18 September 2022
Available online xxxx

JEL classification:
C23
C13
C26

Keywords:
Dynamic panel
Indirect inference
Within-group estimator
Convergence

ABSTRACT

This paper proposes an estimator for higher-order dynamic panel models based on the idea of indirect inference by matching the simple within-group estimator with its analytical approximate expectation. The resulting estimator is shown to be consistent and asymptotically normal. For the special case of first-order dynamic panel, the estimator yields numerically the same result from an existing procedure in the literature, but the inference to follow differs and this paper examines the differences and implications for hypothesis testing. Monte Carlo simulations show that the proposed estimator is virtually unbiased, achieves usually lower root mean squared error than competing estimators, and delivers very reliable empirical size across various parameter configurations and error distributions. This new estimator is used to estimate the convergence parameter in an inequality measure among 63 countries during 1985–2015. It shows strong evidence of convergence over long test horizons but much weaker evidence over a 5-year horizon for developing countries.

1. Introduction

Dynamic panel (DP) models with fixed effects have been used extensively in applied microeconomics, development, macroeconomics, and other disciplines of the social sciences. In the classical set-up of large N (the number of cross-sectional units) and fixed T (the number of time periods), the simple within-group (WG) or least squares dummy variables (LSDV) estimator for the first-order dynamic panel (DP(1) henceforth) is well known to be inconsistent, see Nickell (1981). The (Gaussian) maximum likelihood estimator (MLE) is found in Anderson and Hsiao (1981) to be consistent in most cases but is also sensitive to the assumptions on the initial conditions and the asymptotic plans.[1] In light of these findings, Anderson and Hsiao (1981) proposed using the so-called instrumental variables (IV) to estimate dynamic panels. Their seminal work inspired many influential publications in decades to follow, see, *inter alia*, Holtz-Eakin et al. (1988), Arellano and Bond (1991), Arellano and Bover (1995), Ahn and Schmidt (1995), Hahn (1997), Blundell and Bond (1998, 2000), Alvarez and Arellano (2003), Hahn et al. (2007), and Ashley and Sun (2016), and dynamic panels witnessed their widespread use in applied fields. Baltagi (2008) and Bun and Sarafidis (2015) provided comprehensive reviews on this literature of estimation strategies based on IV and the generalized method of moments (GMM).

For one to be able to use the popular IV/GMM approach, two questions need to be answered. The first is regarding the choice of IV or moment conditions to be used. One can use the lagged level, the first differenced variable, the lagged

* Corresponding author.
E-mail addresses: ybao@purdue.edu (Y. Bao), xuewenyu@fudan.edu.cn (X. Yu).

[1] The literature of fixed-T dynamic panels based on the likelihood approach also includes the random effects estimators as considered by Blundell and Smith (1991) and Alvarez and Arellano (2003) and the conditional likelihood estimator of Lancaster (2002). Very recently, Alvarez and Arellano (2022) proposed estimators that are robust to time-series heteroskedasticity.

https://doi.org/10.1016/j.jeconom.2022.09.003
0304-4076/

differenced variable, the long differenced variable, and so on, see Hsiao and Zhou (2017). There are also choices regarding which lags or differences to be used, as they are related to the quality of instruments. The second is regarding the number of IV or moment conditions, even if the first question is already settled. In general, for a moderate T, there are many instruments available, even for the baseline DP(1). Ideally, one would like to use as many as possible to improve asymptotic efficiency. But it has been documented that using more moments does not necessarily lead to better performance of the resulting IV/GMM estimator in finite samples and the strategy of exploiting all the moment conditions for estimation is actually not recommended for panel data applications, see Kiviet (1995), Wansbeek and Bekker (1996), Judson and Owen (1999), Ziliak (1997), and Bun and Kiviet (2006).

In view of these challenges to use IV/GMM in dynamic panels, a different strand of literature emerged that focuses explicitly on correcting the bias of the simple WG estimator in DP(1), see Kiviet (1995), Bun and Kiviet (2001), Hahn and Kuersteiner (2002), Bun and Carree (2005, BC henceforth), Everaert and Pozzi (2007), and Gouriéroux et al. (2010), among others.[2] Note that these works correct the bias by either using analytical approximate bias of the WG estimator, or via the approach of simulations, or through bootstrapping. The latter two approaches can be numerically intensive and this paper proposes an estimator by following the first approach of analytical bias approximation. The intuition behind the proposed estimator in this paper is straightforward: it tries to match the WG estimator, which is inconsistent and yet computationally simple, with its analytical approximate expectation. The latter is a function of model parameters and thus one can build a feasible sample binding function that links model parameters, the sample data, and the WG estimator. In turn, model parameters can be numerically solved from this sample binding function. The resulting estimator is named the indirect inference (II) estimator, largely in line with the spirit of Gouriéroux et al. (1993) and Smith (1993). Given that the appropriately recentered WG estimator is consistent, where the recentering term is based on the analytical approximate expectation, the resulting II estimator is necessarily consistent and its asymptotic distribution follows from the delta method.

There are three major contributions of this paper. First, it focuses on the traditional large-N and finite-T framework and considers higher-order dynamic panels, including DP(1) as a special case. The proposed estimator is based on the simple least squares procedure, simulation free, and does not rely on instruments. In comparison with the GMM estimator, whose performance depends crucially on the quality and quantity of instruments, the II estimator is straightforward to calculate and is shown in simulations to possess very good finite-sample performance. In particular, it is virtually unbiased and usually achieves higher estimation precision than competing estimators. Second, in contrast to the existing works, the results derived in this paper do not rely on normality assumption on the idiosyncratic error term, nor some restrictive assumption on the initial (latent) variable. The asymptotic variance of the II estimator is different from that existing in the literature for DP(1) and this paper explains the differences and provides insights to the implications for hypothesis testing. Third, the II estimator is asymptotically normal and the asymptotic variance can be consistently estimated once the II estimator is available. Thus inference procedures can be easily implemented. Simulations show that the II-based inference delivers very good size performance.

The plan of this paper is as follows. The next section contains the main results, starting with the classical set-up, where a feasible binding function in terms of model parameters and the sample data is constructed and the asymptotic distribution of the resulting estimator that solves for this function is derived. It also contains discussions over existing results on the simple DP(1) and analyzes a special type of DP in convergence studies. Further, a robust II estimator is proposed in the presence of time-series heteroskedasticity. Section 3 provides some simulations, where the new estimator is found to possess good finite-sample properties, in comparison with existing estimators. Using the II estimator developed in this paper, Section 4 contains an empirical study of convergence in inequality among 63 countries during the period 1985–2015. It finds strong evidence of convergence over the longer 10-year, 15-year, and 20-year horizons among all the countries but little evidence of convergence in developing counties in the shorter 5-year horizon. The last section concludes. All technical details and proofs are provided in the appendix and additional results are collected in supplementary appendices.

2. Main results

This section first focuses on the classical set-up of large panels with short time spans, where the idiosyncratic errors are i.i.d. across time and individual units. The general case of higher-order dynamic panel model is considered. Then the case of DP(1) is discussed and in comparison with existing results in the literature, some misleading points are clarified. The general result simplifies for a special type of higher-order DP in convergence studies. In the presence of time-series heteroskedasticity, a robust estimation strategy is proposed.

[2] For higher-order dynamic panels, to the best of the authors' knowledge, the literature is relatively scarce, see Bun (2003) that derived the asymptotic distribution of the bias-corrected LSDV in higher-order dynamic panels under large T and general covariance structure, Lee (2012) that studied bias-corrected LSDV when the lag order is possibly misspecified under large N and large T, and De Vos et al. (2015) that proposed the bootstrap approach to correcting the bias of LSDV. Chowdhury (1987) outlined a somewhat different approach, under which the bias of the least squares estimator based on the differenced model was derived. A very recent reference is Alvarez and Arellano (2022) that proposed bias-correcting the score function based on the likelihood approach.

2.1. Model specification and notation

The pth order dynamic panel model, DP(p) for short, with fixed effects and exogenous regressors considered in this paper is

$$y_{it} = \alpha_i + \phi_1 y_{i,t-1} + \cdots + \phi_p y_{i,t-p} + \boldsymbol{x}_{it}'\boldsymbol{\beta} + u_{it}, \ \ i = 1, \ldots, N, \ \ t = 1, \ldots, T, \tag{1}$$

where the dependent variable y_{it} is related to its lagged values, up to order p, the unobserved individual-specific effect α_i, the $k \times 1$ vector of exogenous variables $\boldsymbol{x}_{it}$, and an idiosyncratic disturbance term u_{it}.[3]

Throughout, $\boldsymbol{I}_r$ denotes the identity matrix of size r with the unit vector $\boldsymbol{e}_{r,j}$ as its jth column, $j = 1, \ldots, r$, $\boldsymbol{1}_T$ is a $T \times 1$ vector consisting of 1's, $\boldsymbol{M}_T = \boldsymbol{I}_T - T^{-1}\boldsymbol{1}_T\boldsymbol{1}_T'$, $\boldsymbol{A}_{NT} = \boldsymbol{I}_N \otimes \boldsymbol{M}_T$, $\boldsymbol{L}_T$ is a $T \times T$ strict lower triangular matrix with 1's on the first sub-diagonals and zero elsewhere, $\boldsymbol{O}_{r\times s}$ is an $r \times s$ matrix of zeros ($\boldsymbol{O}_r$ for the case when $r = s$ and $\boldsymbol{0}_r = \boldsymbol{O}_{r\times 1}$), and tr, $\odot$, and $\otimes$ denote matrix trace, Hadamard (element-wise) product, and Kronecker product operators, respectively. When there is no confusion from the context, matrix and vector dimension subscripts may be omitted. (So $\boldsymbol{I}$, $\boldsymbol{1}$, $\boldsymbol{e}_1$, $\boldsymbol{L}$, $\boldsymbol{M}$, and $\boldsymbol{A}$ are used to denote $\boldsymbol{I}_T$, $\boldsymbol{1}_T$, $\boldsymbol{e}_{T,1}$, $\boldsymbol{L}_T$, $\boldsymbol{M}_T$, and $\boldsymbol{A}_{NT}$, respectively, unless indicated otherwise. In particular, $\boldsymbol{A}$ is the well-known within transformation matrix that wipes out the fixed effects.) The subscript 0 is used to signify the true parameter value. When a term is denoted without its parameter argument, it means it is evaluated at the true parameter value.

For each cross-sectional unit $i = 1, \ldots, n$, let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})'$, $\boldsymbol{y}_{i,-s} = (y_{i,1-s}, \ldots, y_{i,T-s})'$, $s = 1, \ldots, p$, $\boldsymbol{Y}_i = (\boldsymbol{y}_{i,-1}, \ldots, \boldsymbol{y}_{i,-p})$, $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT})'$, $\boldsymbol{W}_i = (\boldsymbol{Y}_i, \boldsymbol{X}_i)$, and $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{iT})'$. Stacking over i, one can put $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_N')'$, $\boldsymbol{y}_{-s} = (\boldsymbol{y}_{1,-s}', \ldots, \boldsymbol{y}_{N,-s}')'$, $s = 1, \ldots, p$, $\boldsymbol{Y} = (\boldsymbol{Y}_1', \ldots, \boldsymbol{Y}_N')' = (\boldsymbol{y}_{-1}, \ldots, \boldsymbol{y}_{-p})$, $\boldsymbol{X} = (\boldsymbol{X}_1', \ldots, \boldsymbol{X}_N')'$, $\boldsymbol{W} = (\boldsymbol{Y}, \boldsymbol{X}) = (\boldsymbol{W}_1', \ldots, \boldsymbol{W}_N')'$, $\boldsymbol{u} = (\boldsymbol{u}_1', \ldots, \boldsymbol{u}_N')'$, and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)'$. Note that the $NT \times p$ matrix $\boldsymbol{Y}$ collects all the lagged observations on the outcome variable, vertically stacked across the $T \times p$ matrices $\boldsymbol{Y}_i$, $i = 1, \ldots, N$, and equivalently, horizontally stacked across the $NT \times 1$ vectors $\boldsymbol{y}_{-s}$, $s = 1, \ldots, p$. In matrix notation, (1) can be written as

$$\boldsymbol{y} = (\boldsymbol{I}_N \otimes \boldsymbol{1}_T)\boldsymbol{\alpha} + \boldsymbol{W}\boldsymbol{\theta} + \boldsymbol{u}, \tag{2}$$

where $\boldsymbol{\theta} = (\boldsymbol{\phi}', \boldsymbol{\beta}')'$ and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)'$. Let $m = p+k$ denote the dimension of the parameter vector $\boldsymbol{\theta}$ and $n = N(T-1)$ be the number of observations used in the ordinary least squares (OLS) regression for the WG estimator when the fixed effects are first wiped out.

2.2. Assumptions

The following conditions regarding the error terms, fixed effects, and initial values are assumed.

Assumption 1. The series of error terms u_{it}, $i = 1, \ldots, N$, $t = 1, \ldots, T$, is i.i.d. across time and individuals, $\mathrm{E}(u_{it}) = 0$, $\mathrm{Var}(u_{it}) = \sigma^2$, and has finite moments up to the fourth order.

Assumption 2. The series of fixed effects α_i, $i = 1, \ldots, N$, is i.i.d. across individuals with finite moments up to the fourth order.

Assumption 3. The error terms u_{it} and fixed effects α_i are independent for any $i = 1, \ldots, N$, $t = 1, \ldots, T$.

Assumption 4. The regressors $\boldsymbol{X}$, when present, are either fixed or random and $N^{-1}\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}$ converges (in probability) to a positive definite matrix as $N \to \infty$. When they are fixed, each one of them is $O(1)$. When they are random: (i) they are strictly exogenous with respect to error terms and i.i.d. across i; (ii) each one of them is $O_P(1)$ with finite moments up to the fourth order; (ii) $\mathrm{E}(\alpha_i^{r_1} x_{it,s}^{r_2}) = O(1)$ and $\mathrm{Cov}(\alpha_i^{r_1} x_{it,s_i}^{r_2}, \alpha_j^{r_1} x_{jt,s_j}^{r_2}) = 0$, $r_1 + r_2 \le 4$, $r_1 \ge 0$, $r_2 \ge 0$, $i \ne j$, $i, j = 1, \ldots, N$, $s, s_i, s_j = 1, \ldots, k$.

Assumption 5. The initial values $y_{i,-s}$, $i = 1, \ldots, N$, $s = 0, 1, \ldots, p$, are either fixed or random. When they are fixed, each one of them is $O(1)$. When they are random: (i) each one of them is $O_P(1)$ with finite moments up to the fourth order; (ii) $\mathrm{E}(\alpha_i^{r_1} y_{i,-s}^{r_2}) = O(1)$ and $\mathrm{Cov}(\alpha_i^{r_1} y_{i,-s}^{r_2}, \alpha_j^{r_1} y_{j,-s}^{r_2}) = 0$, $r_1 + r_2 \le 4$, $r_1 \ge 0$, $r_2 \ge 0$, $i \ne j$, $i, j = 1, \ldots, N$, $s = 0, \ldots, p$; (iii) $\mathrm{Cov}(y_{i,-s}, u_{it}) = 0$, $s = 0, \ldots, p$, $t = 1, \ldots, T$, $i = 1, \ldots, N$.

Assumptions 1–3 are classical ones regarding the idiosyncratic errors and fixed effects. Their distributions are unspecified, so long as their moments up to the fourth order exist. Assumption 4 does not rule out possible correlation between the exogenous x variables and the fixed effects. Assumption 5 does not specify how the initial observations are generated, except conditions regarding how they may be correlated with the fixed effects and idiosyncratic errors. In Assumption 5, (iii) is most natural and (ii) does allow for possible correlation between the initial condition and fixed

[3] In this paper, unless stated otherwise, it is assumed that the pre-time/initial observations $y_{i0}, \ldots, y_{i,-p}$ are available so that the effective sample size (over time) is $T + p$. Otherwise, all the results should be modified with T replaced by $T - p$ in various places. Sections 2.5, 3.5 and 4 discuss the scenario when the pre-time observations are not available in convergence studies.

effects for each unit i, but it rules out possible cross-sectional correlation among (products of) them. No assumption has been made regarding the parameter vector $\boldsymbol{\phi}$. In other words, it is not necessary to assume that the DP(p) in this paper is dynamically stable. This is innocuous, given that the focus in this paper is on the traditional panels of short time spans.

2.3. The general case

The WG estimator of $\boldsymbol{\theta}_0$ is given by

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{y} = \boldsymbol{\theta}_0 + (\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u}. \tag{3}$$

If one allows $T \to \infty$, from Supplementary Appendices B and C, $T^{-\eta}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W} = O_P(N)$ and $T^{-(\eta-1)}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} = O_P(N)$, where $\eta = 1$ when the panel is dynamically stable, it follows that

$$\operatorname*{plim}_{T\to\infty}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \left(\operatorname*{plim}_{T\to\infty}\frac{1}{N}T^{-\eta}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W}\right)^{-1}\left(\operatorname*{plim}_{T\to\infty}\frac{1}{T}\frac{1}{N}T^{-(\eta-1)}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u}\right) = \boldsymbol{0}, \tag{4}$$

indicating that $\hat{\boldsymbol{\theta}}$ is in fact consistent when $T \to \infty$.

In this paper, the focus is on the case when T is finite and consider the asymptotic distribution of an estimator that is based on $\hat{\boldsymbol{\theta}}$ when $N \to \infty$.[4] To handle the bias, a popular approach is to recenter $\hat{\boldsymbol{\theta}}$ by $\boldsymbol{\delta} = T^{-1}(\operatorname{plim}_{N\to\infty} N^{-1}T^{-\eta}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}(\operatorname{plim}_{N\to\infty} N^{-1}T^{-(\eta-1)}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})$ (or $T^{-1}[N^{-1}T^{-\eta}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}[N^{-1}T^{-(\eta-1)}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})]$), namely, its asymptotic bias. The appendix shows that $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \boldsymbol{\delta}) \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Delta})$, where $\boldsymbol{\Delta} = \lim_{N\to\infty} N[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}\mathrm{Var}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} - \boldsymbol{W}'\boldsymbol{A}\boldsymbol{W}\boldsymbol{\delta})[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}$. While $\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})$ is straightforward to derive (see Supplementary Appendix C), given by

$$\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u}) = \begin{pmatrix} N\sigma^2\mathrm{tr}(\boldsymbol{M}\boldsymbol{\Phi}_p^{-1}\boldsymbol{L}) \\ \vdots \\ N\sigma^2\mathrm{tr}(\boldsymbol{M}\boldsymbol{\Phi}_p^{-1}\boldsymbol{L}^p) \\ \boldsymbol{0}_k \end{pmatrix} = -\frac{N\sigma^2}{T}\begin{pmatrix} \boldsymbol{1}'\boldsymbol{\Phi}_p^{-1}\boldsymbol{L}\boldsymbol{1} \\ \vdots \\ \boldsymbol{1}'\boldsymbol{\Phi}_p^{-1}\boldsymbol{L}^p\boldsymbol{1} \\ \boldsymbol{0}_k \end{pmatrix} \equiv -\frac{N\sigma^2}{T}\boldsymbol{r}, \tag{5}$$

where $\boldsymbol{r} = (\boldsymbol{r}_p', \boldsymbol{0}_k')'$, $\boldsymbol{r}_p = (\boldsymbol{1}'\boldsymbol{\Phi}_p^{-1}\boldsymbol{L}\boldsymbol{1}, \ldots, \boldsymbol{1}'\boldsymbol{\Phi}_p^{-1}\boldsymbol{L}^p\boldsymbol{1})'$, $\boldsymbol{\Phi}_p = \boldsymbol{\Phi}_p(\boldsymbol{\phi}_0)$, and $\boldsymbol{\Phi}_p(\boldsymbol{\phi}) = \boldsymbol{I} - \phi_1\boldsymbol{L} - \cdots - \phi_p\boldsymbol{L}^p$ is a $T \times T$ matrix, a daunting task in practice is to derive $\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})$. Given that $N^{-1}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W}$ is consistent for $N^{-1}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})$, a second choice is to use $\boldsymbol{d}_N = \boldsymbol{d}_N(\boldsymbol{\phi}_0, \sigma^2, \boldsymbol{W}) = (\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})$ as the recentering term. This recentering term $\boldsymbol{d}_N$ is in terms of model parameters as well as the observable data $\boldsymbol{W}$, standing in contrast to the asymptotic bias $\boldsymbol{\delta}$. (And thus the subscript N is added explicitly to emphasize its dependence on the sample data. For notational convenience, in what follows, if a term has the subscript N, it means it is a function of the sample data $\boldsymbol{W}$ and $\boldsymbol{W}$ is suppressed.) The appendix shows that $\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0 - \boldsymbol{d}_N) \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega} = \lim_{N\to\infty} N[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}\mathrm{Var}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}$. In spite of $\boldsymbol{d}_N$ being random, there are three advantages of using $\boldsymbol{d}_N$ as the recentering term. (i) The resulting recentered estimator has a simpler asymptotic variance matrix $\boldsymbol{\Omega}$ compared with $\boldsymbol{\Delta}$, whereas the latter involves additionally the (limits of properly scaled) variance of $\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W}$ and covariance of $\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u}$ and $\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W}$. (ii) The asymptotic bias $\boldsymbol{\delta}$ involves either $\operatorname{plim}_{N\to\infty}N^{-1}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W}$, which is unknown, or $N^{-1}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})$, which depends on $\boldsymbol{\theta}_0$, the initial conditions $(y_{1,-s}, \ldots, y_{N,-s})$, $s = 0, \ldots, p-1$, and the fixed effects $\boldsymbol{\alpha}$ (see Supplementary Appendix B). In contrast, $\boldsymbol{d}_N$ is directly a function of the observable sample data ($(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}$) and parameter vector $\boldsymbol{\phi}_0$ (and σ^2, both appearing in $\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})$). (iii) Since $\boldsymbol{d}_N$ involves only the observable data and model parameters, it facilitates the construction of a new estimator that aims to correct the inconsistency of the WG estimator. If one is to design a bias-correction procedure based on the asymptotic bias, one may be tempted to replace $\operatorname{plim}_{N\to\infty}N^{-1}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W}$ or $N^{-1}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})$ in $\boldsymbol{\delta}$ with the sample average $N^{-1}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W}$. This is exactly the approach taken by BC and is of the same spirit of the procedure in this paper.

The estimator to be introduced in this paper is to invert a function that is related to the recentering term $\boldsymbol{d}_N$, given the sample data and the WG estimator.[5] Note that $\boldsymbol{\beta}_0$ does not appear directly in this recentering term. There is still the nuisance parameter σ^2 appearing in $\boldsymbol{d}_N$.[6] A natural choice is to replace it with $\boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}/n = (\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\theta}_0)'\boldsymbol{A}(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\theta}_0)/n$ and define

$$\boldsymbol{d}_N^{\dagger} = \boldsymbol{d}_N^{\dagger}(\boldsymbol{\theta}_0, \boldsymbol{W}) = -(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\theta}_0)'\boldsymbol{A}(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\theta}_0)(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}\boldsymbol{h}, \tag{6}$$

where $\boldsymbol{h} = \boldsymbol{r}/[T(T-1)]$.

---

[4] Since T is finite, the orders of various terms to follow are expressed in terms of N. When a normalization constant $T^{-\eta}$ is used, it intends to emphasize that the relevant terms are proportional to T^{η}. When there is a unit root, different blocks of $\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W}$ and $\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u}$ will have different normalization constants. To facilitate presentation, the normalization constants under the unit root case are not discussed explicitly and in what follows it is assumed that the panel is dynamically stable (unless stated otherwise), but this does not affect the analysis in this paper when T is finite.

[5] Using the partitioned inverse formula, one may follow the notation in BC to rewrite the recentering term as $\boldsymbol{d}_N = -\sigma^2(\boldsymbol{I}_p, -\hat{\boldsymbol{\Xi}}')'\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}|\boldsymbol{X}}^{-1}\boldsymbol{r}_p/[T(T-1)]$, where $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}|\boldsymbol{X}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}}(\boldsymbol{I}_p - \hat{\boldsymbol{\Psi}})$, $\hat{\boldsymbol{\Xi}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{XX}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{XY}}$, $\hat{\boldsymbol{\Psi}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{XY}}'\hat{\boldsymbol{\Sigma}}_{\boldsymbol{XX}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{XY}}$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{XX}} = \boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}/n$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}} = \boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}/n$, and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{XY}} = \boldsymbol{X}'\boldsymbol{A}\boldsymbol{Y}/n$.

[6] One may also work out explicitly the asymptotic bias of $\hat{\sigma}^2$ that is based on the LSDV residuals and augment this to $\boldsymbol{d}_N$ to recenter $(\hat{\boldsymbol{\theta}}', \hat{\sigma}^2)'$.

With the recentering term $\boldsymbol{d}_N^{\dagger}$ in (6) as a function of $\boldsymbol{\theta}_0$ (as well as the sample data $\boldsymbol{W}$), one can define the *sample* binding function, given the sample data $\boldsymbol{W}$, for any $\boldsymbol{\theta}$, as

$$\boldsymbol{b}_N(\boldsymbol{\theta}) = \boldsymbol{\theta} + \boldsymbol{d}_N^{\dagger}(\boldsymbol{\theta}) \tag{7}$$

and consider the II estimator that inverts this function, namely,

$$\check{\boldsymbol{\theta}} = \boldsymbol{b}_N^{-1}(\hat{\boldsymbol{\theta}}). \tag{8}$$

Intuitively, the II estimator defined as such tries to match the biased WG estimator $\hat{\boldsymbol{\theta}}$ from the observed sample data to its expected value, at least approximately. Typically, the expectation may be approximated via the method of simulations, in line with the original spirit of Gouriéroux et al. (1993) and Smith (1993).[7] The probability limit of $\boldsymbol{b}_N(\boldsymbol{\theta})$, namely, the binding function, is $\operatorname{plim}_{N\to\infty} \boldsymbol{b}_N(\boldsymbol{\theta}) \equiv \boldsymbol{b}(\boldsymbol{\theta}) = \boldsymbol{\theta} + \boldsymbol{\delta}(\boldsymbol{\theta})$. Let $\boldsymbol{G}(\boldsymbol{\theta}) = \partial \boldsymbol{b}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}' = \boldsymbol{I}_m + \partial \boldsymbol{\delta}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}'$ be the Jacobian. For (local) identification, one needs to impose some condition on this Jacobian.

Theorem 1. *Under Assumptions 1 to 5 and that $\boldsymbol{G}(\boldsymbol{\theta})$ is nonsingular in a neighborhood of $\boldsymbol{\theta}_0$, $\check{\boldsymbol{\theta}}$ based on the sample binding function (7) has the following asymptotic distribution:*

$$\sqrt{N}(\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \boldsymbol{V}_{II}), \tag{9}$$

where $\boldsymbol{V}_{II} = \boldsymbol{G}^{-1}\boldsymbol{V}\boldsymbol{G}^{-1\prime}$, $\boldsymbol{G} = \boldsymbol{G}(\boldsymbol{\theta}_0)$, and

$$\boldsymbol{V} = \lim_{N\to\infty} N[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}\mathrm{Var}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} + \boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}\boldsymbol{h})[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}.$$

Global identification, namely, nonsingularity of $\boldsymbol{G}(\boldsymbol{\theta})$ (or invertibility of the binding function) is extremely difficult, if not impossible, to verify, given that neither $\boldsymbol{\delta}(\boldsymbol{\theta})$ nor $\boldsymbol{G}(\boldsymbol{\theta})$ has a tractable analytical form.[8] Note that given $\boldsymbol{\delta}(\boldsymbol{\theta}_0) = \operatorname{plim}_{N\to\infty} \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ and $\operatorname{plim}_{N\to\infty} \boldsymbol{b}_N(\boldsymbol{\theta}_0) = \operatorname{plim}_{N\to\infty} \hat{\boldsymbol{\theta}}$, one can interpret $\boldsymbol{G}(\boldsymbol{\theta}_0)$ as $\partial \operatorname{plim}_{N\to\infty} \hat{\boldsymbol{\theta}}/\partial \boldsymbol{\theta}$ evaluated at $\boldsymbol{\theta}_0$. So the nonsingularity of $\boldsymbol{G}(\boldsymbol{\theta})$ in a neighborhood of $\boldsymbol{\theta}_0$ is equivalent to assuming that there is a one-to-one mapping from $\boldsymbol{\theta}$ to $\operatorname{plim}_{N\to\infty} \hat{\boldsymbol{\theta}}(\boldsymbol{\theta})$ in this neighborhood. For a typical consistent estimator $\tilde{\boldsymbol{\theta}}$, namely, $\operatorname{plim}_{N\to\infty} \tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$, this mapping is obviously one-to-one. For the inconsistent WG estimator $\hat{\boldsymbol{\theta}}$ with $\operatorname{plim}_{N\to\infty} \hat{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \boldsymbol{\theta} + \boldsymbol{\delta}(\boldsymbol{\theta})$, where $T\boldsymbol{\delta}(\boldsymbol{\theta}) = O(1)$ for $\boldsymbol{\theta}$ in this neighborhood, it is likely that $\boldsymbol{\theta}$ dominates $\boldsymbol{\delta}(\boldsymbol{\theta})$ in magnitude for a panel with moderate T and thus the mapping is likely to be one-to-one.

Gospodinov et al. (2017) considered dynamic models with measurement errors. In their framework, the binding function is built on the auxiliary statistics consisting of the OLS estimator, sample moments of OLS residuals, and cross sample moments of regressors and residuals, whereas the parameter vector includes additionally variance of the structural error and variance and auto-covariances of the measurement error. For the first-order distributed lag model, invertibility of the mapping from the parameter vector to the binding function is verifiable (see their Lemma 1). For higher-order models, invertibility of the mapping is not verifiable and Gospodinov et al. (2017) suggested using simulations to approximate the binding function and proposed a novel simulation algorithm.[9] They recommended checking numerically invertibility of the approximated binding function from simulations. The sample binding function $\boldsymbol{b}_N(\boldsymbol{\theta})$ in this paper is neither a pure auxiliary statistic nor the binding function in the strict sense of a probability limit. If one uses simulations and defines, say, $\hat{\boldsymbol{\theta}}^S(\boldsymbol{\theta}) = S^{-1}\sum_{s=1}^{S} \hat{\boldsymbol{\theta}}^s(\boldsymbol{\theta})$, where $\hat{\boldsymbol{\theta}}^s(\boldsymbol{\theta})$ is the WG estimator from the sth simulated sample with parameter $\boldsymbol{\theta}$, $s = 1, \ldots, S$, then for a given N, $\hat{\boldsymbol{\theta}}^S(\boldsymbol{\theta}) = \boldsymbol{\theta} + \boldsymbol{\delta}(\boldsymbol{\theta}) + O_P(N^{-1/2})$, no matter how large S is. The requirement is that $\operatorname{plim}_{N\to\infty} \hat{\boldsymbol{\theta}}^S(\boldsymbol{\theta}) = \boldsymbol{\theta} + \boldsymbol{\delta}(\boldsymbol{\theta})$ so that the auxiliary statistic from the simulated data provides a consistent functional estimator of the binding function. Instead of using $\hat{\boldsymbol{\theta}}^S(\boldsymbol{\theta})$, this paper uses the sample binding function $\boldsymbol{b}_N(\boldsymbol{\theta})$, which is also a consistent functional estimator of the binding function.

At a given $\boldsymbol{\theta}$, the sample Jacobian is

$$\begin{aligned}\frac{\partial \boldsymbol{b}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} &= \boldsymbol{I}_m + 2(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}\boldsymbol{h}(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\theta})'\boldsymbol{A}\boldsymbol{W} \\ &\quad - (\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\theta})'\boldsymbol{A}(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\theta})(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}\boldsymbol{H}(\boldsymbol{\theta}) \\ &\equiv \boldsymbol{G}_N(\boldsymbol{\theta}),\end{aligned} \tag{10}$$

where

$$\boldsymbol{H}(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{h}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = \frac{1}{T(T-1)}\begin{bmatrix} \boldsymbol{R}(\boldsymbol{\phi}) & \mathbf{O}_{p\times k} \\ \mathbf{O}_{k\times p} & \mathbf{O}_k \end{bmatrix},$$

7 There are also nonparametric sieves simulation estimators that do not require distributional assumptions on the pseudo errors in simulations (e.g., Forneron, 2020, and references therein).

8 The authors thank an associate editor and the referees for raising the issue of invertibility of the binding function.

9 Forneron and Ng (2018) also discussed other simulation-based methods and provided an example of applying them to the first-order dynamic panel model.

and $\boldsymbol{R}(\boldsymbol{\phi}) = \partial \boldsymbol{r}_p(\boldsymbol{\phi})/\partial \boldsymbol{\phi}'$ is a $p \times p$ matrix with $\mathbf{1}'\boldsymbol{\Phi}_p^{-1}(\boldsymbol{\phi})\boldsymbol{L}^{j_2}\boldsymbol{\Phi}_p^{-1}(\boldsymbol{\phi})\boldsymbol{L}^{j_1}\mathbf{1}$ in its (j_1, j_2) position, $j_1, j_2 = 1, \ldots, p$. So even though in general one cannot check nonsingularity of $\boldsymbol{G}$, one can always check numerically the identification condition for a given sample $\boldsymbol{W}$ by examining the determinant of $\boldsymbol{G}_N(\boldsymbol{\theta})$ over a grid of values of $\boldsymbol{\theta}$.

To make inference feasible in practice, one needs to estimate $\boldsymbol{V}$ and $\boldsymbol{G}$ in $\boldsymbol{V}_{II} = \boldsymbol{G}^{-1}\boldsymbol{V}\boldsymbol{G}^{-1\prime}$. The estimation of $\boldsymbol{V}$ is more complicated, even though $\boldsymbol{G}$ can be naturally replaced by $\hat{\boldsymbol{G}} = \boldsymbol{G}_N(\check{\boldsymbol{\theta}})$.[10] One may replace $N^{-1}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})$ in $\boldsymbol{V}$ by $N^{-1}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W}$. Supplementary Appendix D shows that $\mathrm{Var}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} + \boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}\boldsymbol{h}) = \mathrm{Var}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u}) - \sigma^4 N[2(T-1) + \gamma_2 T^{-1}(T-1)^2]\boldsymbol{h}\boldsymbol{h}'$ (where γ_2 is the excess kurtosis of u_{it}) and Supplementary Appendix C discusses the various terms appearing in $\mathrm{Var}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})$. Specifically, $N^{-1}\mathrm{Var}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} + \boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}\boldsymbol{h})$ involves, in addition to $\boldsymbol{\theta}_0$, the fixed effects and the possible interactions between the fixed effects and initial conditions, as well as the skewness (γ_1) and kurtosis (γ_2) of u_{it}. (The variance and covariance matrices pertaining to the initial conditions, when they are not assumed to be fixed, drop out in the asymptotic variance when scaled by N^{-1}.) Thus simply replacing $\boldsymbol{\theta}_0$ by $\check{\boldsymbol{\theta}}$ does not yield a meaningful estimator. Instead, one may use

$$\hat{\boldsymbol{V}} = N(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}\left(\sum_{i=1}^{N}\hat{\boldsymbol{v}}_i\hat{\boldsymbol{v}}_i'\right)(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}, \tag{11}$$

where

$$\hat{\boldsymbol{v}}_i = \boldsymbol{W}_i'\boldsymbol{M}(\boldsymbol{y}_i - \boldsymbol{W}_i\check{\boldsymbol{\theta}}) + (\boldsymbol{y}_i - \boldsymbol{W}_i\check{\boldsymbol{\theta}})'\boldsymbol{M}(\boldsymbol{y}_i - \boldsymbol{W}_i\check{\boldsymbol{\theta}})\hat{\boldsymbol{h}}, \tag{12}$$

in which $\hat{\boldsymbol{h}} = \hat{\boldsymbol{r}}/[T(T-1)]$, $\hat{\boldsymbol{r}} = (\boldsymbol{r}_p(\check{\boldsymbol{\phi}})', \mathbf{0}_k')'$. (Recall that $\boldsymbol{M} = \boldsymbol{I}_T - T^{-1}\mathbf{1}_T\mathbf{1}_T'$.) In the end, a consistent estimator of $\boldsymbol{V}_{II}$ is

$$\hat{\boldsymbol{V}}_{II} = \hat{\boldsymbol{G}}^{-1}\hat{\boldsymbol{V}}\hat{\boldsymbol{G}}^{-1\prime}. \tag{13}$$

One may wonder what will happen to the II estimator when in fact T is large. From $\boldsymbol{d}_N^{\dagger} = -(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\theta}_0)'\boldsymbol{A}(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\theta}_0)(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}\boldsymbol{h}$, where $(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\theta}_0)'\boldsymbol{A}(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\theta}_0) = O_P(NT)$, $\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W} = O_P(NT)$ when the process is dynamically stable or has its leading block being $O_P(NT^3)$ under unit root, $\boldsymbol{h} = \boldsymbol{r}/[T(T-1)] = O(T^{-1})$ when the process is dynamically stable or $O(1)$ when it contains a unit root, one has $\mathrm{plim}_{T\to\infty}\boldsymbol{d}_N^{\dagger} = \mathbf{0}$, namely, the II estimator tries to correct the consistent WG estimator. As Hahn and Kuersteiner (2002) pointed out, the consistent WG estimator may still possess an asymptotic bias, depending on the relative rates of increase of T and N. In practice, regardless of T, one can always apply the II procedure and it yields a consistent estimator. When there is no $\boldsymbol{X}$ and if one assumes stationarity, the correction term from the II procedure for DP(1) is $-(1+\phi_0)/(T-1)$ (namely, the leading term of the Nickel bias, see (18) in Section 2.4.2 to follow). Thus, the resulting II estimator becomes $[(T-1)\hat{\phi} + 1]/(T-2)$. In contrast, Hahn and Kuersteiner (2002) corrected the bias by directly subtracting from $\hat{\phi}$ the estimated bias to arrive at $[(T+1)\hat{\phi} + 1]/T$.[11] When there is a unit root, a direct bias-correction procedure as in Hahn and Kuersteiner (2002) is not available, since the asymptotic bias involves the fixed effects (see their Theorem 5), but the II procedure in this paper can still be used. Of course, one would expect that the asymptotic distribution as given by (9) is not valid. If there is no unit root, one can show that $\boldsymbol{H} = \partial\boldsymbol{h}(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}' = O(T^{-1})$. This, together with $\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} = O(N) + O_P(\sqrt{NT})$ and $\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W} = O_P(NT)$, implies that $\mathrm{plim}_{T\to\infty}\boldsymbol{G}_N = \boldsymbol{I}_m$. Moreover, from $\mathrm{Var}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} + \boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}\boldsymbol{h}) = \mathrm{Var}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u}) - \sigma^4 N[2(T-1) + \gamma_2 T^{-1}(T-1)^2]\boldsymbol{h}\boldsymbol{h}'$ (see Supplementary Appendix D), one has $\lim_{T\to\infty}(NT)^{-1}\mathrm{Var}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} + \boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}\boldsymbol{h}) = \lim_{T\to\infty}(NT)^{-1}\mathrm{Var}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})$. After some algebra, one can show that $\lim_{T\to\infty}(NT)^{-1}\mathrm{Var}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u}) = \sigma^2\mathrm{plim}_{T\to\infty}(NT)^{-1}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W}$. Therefore, the asymptotic distribution result becomes $\sqrt{NT}(\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \sigma^2(\mathrm{plim}_{T\to\infty}(NT)^{-1}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1})$ if there is no unit root. As shown in Hahn and Kuersteiner (2002) and Bai (2013), if the errors are i.i.d. Gaussian, this asymptotic variance equals the lower variance bound. In other words, the II estimator is most efficient under the double asymptotic regime. Importantly, the II estimator does not involve an asymptotic bias and one does not need to apply bias-correction as in Hahn and Kuersteiner (2002), nor does it involve simulations as in Gouriéroux et al. (2010). For the special case of DP(1) with no $\boldsymbol{X}$, one can show that $\sigma^2(\mathrm{plim}_{T\to\infty}(NT)^{-1}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1} = 1 - \phi_0^2$, which is also the variance result from Hahn and Kuersteiner (2002) for their bias-corrected estimator and Gouriéroux et al. (2010) for their simulation-based II estimator.[12]

2.4. The special case of DP(1)

When $p = 1$ in (1), with $\boldsymbol{W} = (\boldsymbol{y}_{-1}, \boldsymbol{X})$ and $\boldsymbol{\theta} = (\phi, \boldsymbol{\beta}')'$, one has

$$\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u}) = -\frac{N\sigma^2}{T}\begin{pmatrix}\mathbf{1}'(\boldsymbol{I} - \phi_0\boldsymbol{L})^{-1}\boldsymbol{L}\mathbf{1} \\ \mathbf{0}_k\end{pmatrix} = -\frac{N\sigma^2}{T}\mathbf{1}'(\boldsymbol{I} - \phi_0\boldsymbol{L})^{-1}\boldsymbol{L}\mathbf{1}\boldsymbol{e}_{k+1,1}, \tag{14}$$

[10] Similar to DP(1) considered in BC, the term $(\boldsymbol{y} - \boldsymbol{W}\check{\boldsymbol{\theta}})'\boldsymbol{A}\boldsymbol{W}$ that appears in $\hat{\boldsymbol{G}}$ can be simplified as $((\boldsymbol{y} - \boldsymbol{W}\check{\boldsymbol{\theta}})'\boldsymbol{A}\boldsymbol{Y}, \mathbf{0}_k')$.

[11] Obviously, under large T, they are asymptotically equivalent. In finite samples, the bias-corrected estimator from Hahn and Kuersteiner (2002) should display a slightly smaller root mean squared error than the II estimator, see Table 3 in Section 3.2 for Monte Carlo evidence.

[12] The authors thank a referee for suggesting this line of discussion regarding large T and the issue of efficiency. When there is a unit root, the derivation is more complicated and is left for future research. Under large N and finite T, the issue of efficiency of the II estimator is more subtle. From $\boldsymbol{V}_{II} = \boldsymbol{G}^{-1}\boldsymbol{V}\boldsymbol{G}^{-1\prime}$, where $\boldsymbol{V}$ is the asymptotic variance of $\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0 - \boldsymbol{d}_N^{\dagger})$ and $\boldsymbol{G}$ can be interpreted as $\partial\,\mathrm{plim}_{N\to\infty}\hat{\boldsymbol{\theta}}/\partial\boldsymbol{\theta}$ evaluated at $\boldsymbol{\theta}_0$, one can follow the interpretation in Gouriéroux et al. (2010) (see their Eq. (16)) that the II estimator should inherit some of the "efficiency" properties of the bias-corrected estimator $\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0 - \boldsymbol{d}_N^{\dagger}$.

where $\mathbf{1}'(\boldsymbol{I}-\phi_0\boldsymbol{L})^{-1}\boldsymbol{L}\mathbf{1} = T/(1-\phi_0)-(1-\phi_0^T)/(1-\phi_0)^2$ when the panel is dynamically stable and $T(T-1)/2$ if it has a unit root. Correspondingly, the recentering term, with σ^2 replaced by $\boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}/n = (\boldsymbol{y}-\boldsymbol{W}\boldsymbol{\theta}_0)'\boldsymbol{A}(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{\theta}_0)/n$, is defined as

$$\boldsymbol{d}_N^{\dagger} = -h(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{\theta}_0)'\boldsymbol{A}(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{\theta}_0)(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}\boldsymbol{e}_{k+1,1}, \tag{15}$$

where $h = h(\phi_0)$, $h(\phi) = \mathbf{1}'(\boldsymbol{I}-\phi\boldsymbol{L})^{-1}\boldsymbol{L}\mathbf{1}/[T(T-1)]$. The sample Jacobian from the sample binding function $\boldsymbol{b}_N(\boldsymbol{\theta}) = \boldsymbol{\theta}+\boldsymbol{d}_N^{\dagger}(\boldsymbol{\theta})$ is then

$$\begin{aligned}\boldsymbol{G}_N(\boldsymbol{\theta}) &= \boldsymbol{I}_{k+1} + 2h(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}\boldsymbol{e}_{k+1,1}(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{\theta})'\boldsymbol{A}\boldsymbol{W}\\ &\quad -(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{\theta})'\boldsymbol{A}(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{\theta})(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}\begin{bmatrix} H(\phi) & \mathbf{0}_k' \\ \mathbf{0}_k & \mathbf{O}_k\end{bmatrix},\end{aligned}$$

where $H(\phi) = \partial h(\phi)/\partial\phi = \mathbf{1}'(\boldsymbol{I}-\phi\boldsymbol{L})^{-1}\boldsymbol{L}(\boldsymbol{I}-\phi\boldsymbol{L})^{-1}\boldsymbol{L}\mathbf{1}/[T(T-1)]$. Theorem 1 carries through with $\boldsymbol{V} = \lim_{N\to\infty} N[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}\mathrm{Var}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u}+h\boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}\boldsymbol{e}_{k+1,1})[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}$.

2.4.1. In relation to Bun and Carree (2005)

BC used $\boldsymbol{\delta} = (\mathrm{plim}_{N\to\infty}N^{-1}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}(\mathrm{plim}_{N\to\infty}N^{-1}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})$ as the recentering vector, not $\boldsymbol{d}_N = (\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})$ as in this paper. The asymptotic variance of the recentered WG estimator, as well as that of the resulting II estimator, depends on the recentering vector used. In practice, since the recentering vector $\boldsymbol{\delta}$ involves, in addition to $\boldsymbol{\theta}_0$ and σ^2, the unknown $\mathrm{plim}_{N\to\infty}N^{-1}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W}$, BC suggested using its sample analogue. When σ^2 in $\boldsymbol{\delta}$ is also replaced by $\boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}/n$, the recentering vector effectively becomes $\boldsymbol{d}_N^{\dagger}$ in this paper and the bias-correction procedure proposed in BC is in fact the same as the II procedure. However, their asymptotic variance expression is not stated correctly. If one uses the (infeasible) recentering quantity $\boldsymbol{\delta}$, then the asymptotic variance of $\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0-\boldsymbol{\delta})$ should be $\boldsymbol{\Delta}$, not their $\boldsymbol{V}_{\boldsymbol{X}}$; if one uses the feasible recentering quantity $\boldsymbol{d}_N^{\dagger}$, then the asymptotic variance of $\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0-\boldsymbol{d}_N^{\dagger})$ is $\boldsymbol{V}$. The resulting bias-corrected estimator $\check{\boldsymbol{\theta}}$ has its asymptotic variance $\boldsymbol{V}_{II} = \boldsymbol{G}^{-1}\boldsymbol{V}\boldsymbol{G}^{-1\prime}$, not $\boldsymbol{V}_{BC} = \boldsymbol{G}^{-1}\boldsymbol{V}_{\boldsymbol{X}}\boldsymbol{G}^{-1\prime}$ as in BC. Thus, the standard errors may be different from those from BC, because of (i) the initial latent variable, (ii) the (non-normal) distribution of u_{it}, and (iii) the estimated error variance in the recentering term. Note that (i) and (ii) give rise to different expressions of $\mathrm{Var}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})$ (one under the general set-up in this paper and one under normality and the assumption on the initial latent variable in BC) and (iii) is related to the usage of $\mathrm{Var}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u}+h\boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}\boldsymbol{e}_{k+1,1}) = \mathrm{Var}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u}) - h^2\sigma^4N\left[2(T-1)+\gamma_2T^{-1}(T-1)^2\right]\boldsymbol{e}_{k+1,1}\boldsymbol{e}_{k+1,1}'$ (see Supplementary Appendix E.1) versus $\mathrm{Var}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})$ in the sandwich form of the variance formula. In practice, the three factors may interact with each other and one may not have a clear-cut idea of their composite effect on the standard errors.

To better understand how the differences would affect the inference procedures, one can compare carefully the meat parts of $\boldsymbol{G}^{-1}\boldsymbol{V}\boldsymbol{G}^{-1\prime}$ and $\boldsymbol{G}^{-1}\boldsymbol{V}_{\boldsymbol{X}}\boldsymbol{G}^{-1\prime}$, namely, $\boldsymbol{V}$ in this paper and $\boldsymbol{V}_{\boldsymbol{X}}$ in BC. Note that $\boldsymbol{V}$ and $\boldsymbol{V}_{\boldsymbol{X}}$ are also constructed in sandwich forms with the corresponding meat parts $\mathrm{Var}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u}+h\boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}\boldsymbol{e}_{k+1,1}) \equiv \boldsymbol{V}_{0,II}$ and $\mathrm{Var}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u}) \equiv \boldsymbol{V}_{0,BC}$ (which is $\sigma^2\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})+N\sigma^4\mathrm{tr}(\boldsymbol{MCMC})\boldsymbol{e}_{k+1,1}\boldsymbol{e}_{k+1,1}'$ under normality and the initial latent variable condition), respectively, while sharing the same bread part $[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}$. From (E.4), (E.6), and (E.7) in Supplementary Appendix E.1, one has

$$\begin{aligned}\boldsymbol{V}_{0,II} &= \boldsymbol{V}_{0,BC}\\ &\quad -2h^2\sigma^4N(T-1)\boldsymbol{e}_{k+1,1}\boldsymbol{e}_{k+1,1}'\\ &\quad +\sigma^2\mathrm{E}[(\boldsymbol{Mf}\tilde{\boldsymbol{y}}_0+\boldsymbol{MC}\mathbf{1})'(\boldsymbol{Mf}\tilde{\boldsymbol{y}}_0+\boldsymbol{MC}\mathbf{1})]\boldsymbol{e}_{k+1,1}\boldsymbol{e}_{k+1,1}'\\ &\quad +\sigma^3\gamma_1\left\{\overline{\boldsymbol{W}}'[\mathbf{1}_N\otimes\boldsymbol{MC}\mathrm{dg}(\boldsymbol{MC})]\boldsymbol{e}_{k+1,1}'+\boldsymbol{e}_{k+1,1}[\mathbf{1}_N\otimes\boldsymbol{MC}\mathrm{dg}(\boldsymbol{MC})]\overline{\boldsymbol{W}}\right\}\\ &\quad +\sigma^4N\gamma_2\left[\mathrm{tr}(\boldsymbol{MC}\odot\boldsymbol{MC})-\frac{h^2(T-1)^2}{T}\right]\boldsymbol{e}_{k+1,1}\boldsymbol{e}_{k+1,1}',\end{aligned} \tag{16}$$

where $\boldsymbol{C} = \boldsymbol{\Phi}_1^{-1}\boldsymbol{L}$, $\boldsymbol{f} = \boldsymbol{\Phi}_1^{-1}\boldsymbol{e}_1$, $\overline{\boldsymbol{W}} = \mathrm{E}(\boldsymbol{W})$, $\tilde{\boldsymbol{y}}_0 = \boldsymbol{y}_0-\mathrm{E}(\boldsymbol{y}_0)$, $\boldsymbol{y}_0 = (y_{1,0},\ldots,y_{N,0})'$ collects the initial observations, and $\mathrm{dg}(\cdot)$ collects in order the diagonal elements of its argument as a column vector.

First, suppose one assumes normality and also the initial observations are fixed ($y_{i0} = \alpha_i/(1-\phi_0)$), then $\boldsymbol{V}_{0,II} = \boldsymbol{V}_{0,BC}-2h^2\sigma^4N(T-1)\boldsymbol{e}_{k+1,1}\boldsymbol{e}_{k+1,1}'$ and the difference in $\boldsymbol{V}_{0,II}$ and $\boldsymbol{V}_{0,BC}$ is solely related to (iii), namely, whether the effect of the estimated error variance in the recentering term is taken into account. Since $2h^2\sigma^4N(T-1) > 0$ holds for all $T \geq 2$, one would expect that the variance of $\check{\phi}$ is to be over-estimated by BC and the resulting t-test may be under-sized. Further, the degree of variance over-estimation increases in h. Section 3 provides simulation evidence to support this aspect of the difference between $\boldsymbol{V}_{0,II}$ and $\boldsymbol{V}_{0,BC}$.

Second, consider the effects of the error distribution. Fixing the initial observations, suppose $\gamma_1 = 0$ and $\gamma_2 \neq 0$, namely, the distribution is symmetric and may have thinner or fatter tails than a normal distribution. Supplementary Appendix E.3 shows that $\mathrm{tr}(\boldsymbol{MC}\odot\boldsymbol{MC})-h^2(T-1)^2/T$ is indeed positive for all $\phi_0 \in (-1, 1]$. Therefore, a negative γ_2 will aggravate the over-estimation of variance by BC while a positive γ_2 will mitigate this problem. Furthermore, if the error kurtosis is high enough to offset the negative part $-2h^2\sigma^4N(T-1)$, then BC may start to under estimate the variance of $\check{\phi}$. This is confirmed through simulations in Section 3. One may also wonder about the case of $\gamma_2 = 0$ and $\gamma_1 \neq 0$. Now the magnitude of the term associated with γ_1 depends on the quadratic form in $\overline{\boldsymbol{W}}$, whose sign is indeterminate, so its overall effect is infeasible to determine.

Third, suppose the error term is normally distributed and consider the possible effects of the initial latent variable. Obviously, $\sigma^2\mathrm{E}[(\boldsymbol{Mf}\widetilde{\boldsymbol{y}}_0+\boldsymbol{MC1})'(\boldsymbol{Mf}\widetilde{\boldsymbol{y}}_0+\boldsymbol{MC1})]$ is non-negative. BC fixed this initial condition by assuming $\widetilde{\boldsymbol{y}}_0=(1-\phi_0)^{-1}\widetilde{\boldsymbol{\alpha}}$ (where $\widetilde{\boldsymbol{\alpha}}=\boldsymbol{\alpha}-\mathrm{E}(\boldsymbol{\alpha})$), and since $(1-\phi_0)^{-1}\boldsymbol{Mf}+\boldsymbol{MC1}=\boldsymbol{0}$, this part reduces to $\boldsymbol{0}$ (see Supplementary Appendix E.1 for more details). Consequently, one would expect that any other initial conditions would result in the $\boldsymbol{V}_0$ part of the variance being under-estimated by BC. Its overall effect is not clear though given that it also has impact on the bread part $[\mathrm{E}(\boldsymbol{W}'\boldsymbol{AW})]^{-1}$.

Finally, one needs to keep in mind that while $\boldsymbol{V}_{0,BC}$ can be estimated by a plug-in method, it is advised that $\boldsymbol{V}_{0,II}$ not be estimated by such a method, due to the additional complications in estimating the skewness, kurtosis, and the relevant terms pertaining to the initial conditions (when they are not assumed to be fixed) and the unobservable fixed effects $\boldsymbol{\alpha}$. Instead, one can use the White-type estimator, see (11). So in the end, when one attempts to compare inferences using the estimated $\boldsymbol{V}_{II}$ and $\boldsymbol{V}_{BC}$, a direct conclusion is in general not feasible given that one cannot disentangle the aforementioned three factors and also that they may be based on different estimation methods. Nevertheless, the analysis here may help one understand the possible causes of the difference if one is able to fix some aspects of the model specification.

2.4.2. In relation to Nickell (1981)

When there are no exogenous regressors present, namely,

$$y_{it}=\phi y_{it-1}+\alpha_i+u_{it},\quad i=1,\ldots,N,\quad t=1,\ldots,T,$$

Nickell (1981) derived the asymptotic bias analytically under the stationarity assumption such that

$$y_{i0}=\frac{\alpha_i}{1-\phi_0}+\frac{u_{i0}}{\sqrt{1-\phi_0^2}},\quad |\phi_0|<1. \tag{17}$$

The asymptotic bias, following the notation in the previous subsection, is given by

$$\begin{aligned}\delta&=\frac{\mathrm{E}(\boldsymbol{y}_{-1}'\boldsymbol{Au})}{\mathrm{E}(\boldsymbol{y}_{-1}'\boldsymbol{Ay}_{-1})}\\&=-\frac{(1+\phi_0)}{T-1}\left[1-\frac{1-\phi_0^T}{T(1-\phi_0)}\right]\left\{1-\frac{2\phi_0}{(T-1)(1-\phi_0)}\left[1-\frac{1-\phi_0^T}{T(1-\phi_0)}\right]\right\}^{-1},\end{aligned} \tag{18}$$

where $\mathrm{E}(\boldsymbol{y}_{-1}'\boldsymbol{Au})$ and $\mathrm{E}(\boldsymbol{y}_{-1}'\boldsymbol{Ay}_{-1})$ (under the stationarity assumption) are given by (E.8) and (E.14), respectively, in Supplementary Appendix E.2.[13] A striking feature of this asymptotic bias is that it is solely a function of the true parameter ϕ_0 (and T), independent of the variance and distribution of the error term. The recentered WG estimator has the following asymptotic distribution:

$$\begin{aligned}&\sqrt{N}(\hat{\phi}-\phi_0-\delta)\xrightarrow{d}\\&\mathrm{N}\left(0,\lim_{N\to\infty}N[\mathrm{E}(\boldsymbol{y}_{-1}'\boldsymbol{Ay}_{-1})]^{-2}[\mathrm{Var}(\boldsymbol{y}_{-1}'\boldsymbol{Au})+\delta^2\mathrm{Var}(\boldsymbol{y}_{-1}'\boldsymbol{Ay}_{-1})-2\delta\mathrm{Cov}(\boldsymbol{y}_{-1}'\boldsymbol{Au},\boldsymbol{y}_{-1}'\boldsymbol{Ay}_{-1})]\right),\end{aligned}$$

where $\mathrm{Var}(\boldsymbol{y}_{-1}'\boldsymbol{Au})$, $\mathrm{Var}(\boldsymbol{y}_{-1}'\boldsymbol{Ay}_{-1})$, and $\mathrm{Cov}(\boldsymbol{y}_{-1}'\boldsymbol{Au},\boldsymbol{y}_{-1}'\boldsymbol{Ay}_{-1})$ are given by (E.15), (E.16), and (E.17), respectively, in Supplementary Appendix E.2.

Without the stationarity condition (17), $\mathrm{E}(\boldsymbol{y}_{-1}'\boldsymbol{Ay}_{-1})$ depends on the initial condition, the unobservable individual effects, and the error variance. Following the derivations previously, one may define

$$d_N=\frac{\mathrm{E}(\boldsymbol{y}_{-1}'\boldsymbol{Au})}{\boldsymbol{y}_{-1}'\boldsymbol{Ay}_{-1}}=-\frac{N(T-1)h\sigma^2}{\boldsymbol{y}_{-1}'\boldsymbol{Ay}_{-1}},\quad d_N^{\dagger}=-\frac{h(\boldsymbol{y}-\phi_0\boldsymbol{y}_{-1})'\boldsymbol{A}(\boldsymbol{y}-\phi_0\boldsymbol{y}_{-1})}{\boldsymbol{y}_{-1}'\boldsymbol{Ay}_{-1}}, \tag{19}$$

with the resulting sample binding function $b_N(\phi)=\phi+d_N^{\dagger}(\phi)$ and the associated II estimator $\check{\phi}=b_N^{-1}(\hat{\phi})$.

Proceeding similarly as before, one has

$$\sqrt{N}(\hat{\phi}-\phi_0-d_N)\xrightarrow{d}\mathrm{N}\left(0,\lim_{N\to\infty}\frac{N\mathrm{Var}(\boldsymbol{y}_{-1}'\boldsymbol{Au})}{[\mathrm{E}(\boldsymbol{y}_{-1}'\boldsymbol{Ay}_{-1})]^2}\right), \tag{20}$$

$$\sqrt{N}(\hat{\phi}-\phi_0-d_N^{\dagger})\xrightarrow{d}\mathrm{N}\left(0,\lim_{N\to\infty}\frac{N[\mathrm{Var}(\boldsymbol{y}_{-1}'\boldsymbol{Au})-h^2\mathrm{Var}(\boldsymbol{u}'\boldsymbol{Au})]}{[\mathrm{E}(\boldsymbol{y}_{-1}'\boldsymbol{Ay}_{-1})]^2}\right), \tag{21}$$

$$\sqrt{N}(\check{\phi}-\phi_0)\xrightarrow{d}\mathrm{N}\left(0,\lim_{N\to\infty}\frac{N[\mathrm{Var}(\boldsymbol{y}_{-1}'\boldsymbol{Au})-h^2\mathrm{Var}(\boldsymbol{u}'\boldsymbol{Au})]}{g^2[\mathrm{E}(\boldsymbol{y}_{-1}'\boldsymbol{Ay}_{-1})]^2}\right), \tag{22}$$

[13] Nickell (1981) further assumed that the error terms are random drawings from a normal distribution in his brief derivations, though in his footnote 7 he mentioned that his analysis of the bias does not depend on the normality assumption. It is not clear exactly how he derived his result independent of the normality assumption.

where $\mathrm{E}(\boldsymbol{y}'_{-1}\boldsymbol{A}\boldsymbol{y}_{-1})$ is given by (E.9), $\mathrm{Var}(\boldsymbol{y}'_{-1}\boldsymbol{A}\boldsymbol{u})$ by (E.10) in Supplementary Appendix E.2, $\mathrm{Var}(\boldsymbol{u}'\boldsymbol{A}\boldsymbol{u})$ by (D.1) in Supplementary Appendix D, and

$$g = \operatorname*{plim}_{N\to\infty}\left[1 - \frac{H(\boldsymbol{y}-\phi_0\boldsymbol{y}_{-1})'\boldsymbol{A}(\boldsymbol{y}-\phi_0\boldsymbol{y}_{-1})}{\boldsymbol{y}'_{-1}\boldsymbol{A}\boldsymbol{y}_{-1}} + \frac{2h\boldsymbol{y}'_{-1}\boldsymbol{A}(\boldsymbol{y}-\phi_0\boldsymbol{y}_{-1})}{\boldsymbol{y}'_{-1}\boldsymbol{A}\boldsymbol{y}_{-1}}\right]. \tag{23}$$

2.5. Dynamic panel in convergence studies

Caselli et al. (1996) suggested the following dynamic panel model in convergence studies,

$$y_{it} = \alpha_i + \phi y_{i,t-\tau} + u_{it}, \tag{24}$$

where y_{it} is some measure of income distribution, say, the Gini coefficient (in terms of deviation from period mean), and $\tau \geq 2$ is the time horizon over which one wants to test convergence. (For convenience, (24) is called DP-τ henceforth.) When $\phi - 1 < 0$, it signals (beta-) convergence in income distribution. Caselli et al. (1996) took a τth-order difference approach and proposed using the GMM estimator based on the differenced model. Bao and Dhongde (2009) instead considered the simple OLS estimator based on a first-order difference (OLS1 for short). If one views (24) as a special case of (1), then the previous results follow directly.

Let $\boldsymbol{\Phi} = \boldsymbol{I} - \phi_0\boldsymbol{L}^\tau$ and $\boldsymbol{C}_\tau = \boldsymbol{\Phi}^{-1}\boldsymbol{L}^\tau$. One can check that $\boldsymbol{C}_\tau$ is a strict lower triangular matrix with ϕ_0^{s-1} in its (s,τ)th sub-diagonal position, $s = 1,\ldots,j$, where $j = \lfloor (T-1)/\tau \rfloor$. Then one can write (see (A.1) from Supplementary Appendix A)

$$\boldsymbol{y}_{-\tau} = (\boldsymbol{I}_N \otimes \boldsymbol{C}_\tau)\boldsymbol{u} + (\boldsymbol{I}_N \otimes \boldsymbol{C}_\tau\boldsymbol{1})\boldsymbol{\alpha} + \sum_{s=0}^{\tau-1}(\boldsymbol{I}_N \otimes \boldsymbol{C}_{\tau-1-s}\boldsymbol{e}_1)\boldsymbol{y}_{-s},$$

where $\boldsymbol{y}_{-s} = (y_{1,-s},\ldots,y_{N,-s})'$. It follows that

$$\mathrm{E}(\boldsymbol{y}'_{-\tau}\boldsymbol{A}\boldsymbol{u}) = N\sigma^2\mathrm{tr}(\boldsymbol{M}\boldsymbol{\Phi}^{-1}\boldsymbol{L}^\tau), \tag{25}$$

where, when $|\phi_0| \neq 1$,

$$\mathrm{tr}(\boldsymbol{M}\boldsymbol{\Phi}^{-1}\boldsymbol{L}^\tau) = -\frac{1}{T}\left[\frac{T(1-\phi_0^j)}{1-\phi_0} + \frac{j\tau\phi_0^j}{1-\phi_0} - \frac{\tau(1-\phi_0^j)}{(1-\phi_0)^2}\right].$$

Correspondingly, the recentering quantity and its feasible version are

$$d_N = \frac{\mathrm{E}(\boldsymbol{y}'_{-\tau}\boldsymbol{A}\boldsymbol{u})}{\boldsymbol{y}'_{-\tau}\boldsymbol{A}\boldsymbol{y}_{-\tau}} = -\frac{N(T-1)h\sigma^2}{\boldsymbol{y}'_{-\tau}\boldsymbol{A}\boldsymbol{y}_{-\tau}}, \quad d_N^\dagger = -\frac{h(\boldsymbol{y}-\phi_0\boldsymbol{y}_{-\tau})'\boldsymbol{A}(\boldsymbol{y}-\phi_0\boldsymbol{y}_{-\tau})}{\boldsymbol{y}'_{-\tau}\boldsymbol{A}\boldsymbol{y}_{-\tau}},$$

where h is now defined as $h = \mathrm{tr}(\boldsymbol{M}\boldsymbol{\Phi}^{-1}\boldsymbol{L}^\tau)/[T(T-1)]$. The associated II estimator $\check{\phi} = b_N^{-1}(\hat{\phi})$ is defined from the sample binding function $b_N(\phi) = \phi + d_N^\dagger(\phi)$. Proceeding similarly as before, one has the asymptotic distribution results as in (20)–(22) with $\boldsymbol{y}_{-\tau}$ replacing $\boldsymbol{y}_{-1}$ (including that in g), where $\mathrm{E}(\boldsymbol{y}'_{-\tau}\boldsymbol{A}\boldsymbol{y}_{-\tau})$, $\mathrm{Var}(\boldsymbol{y}'_{-\tau}\boldsymbol{A}\boldsymbol{u})$, and $\mathrm{Cov}(\boldsymbol{y}'_{-\tau}\boldsymbol{A}\boldsymbol{u}, \boldsymbol{u}'\boldsymbol{A}\boldsymbol{u})$ are of the same expressions as (E.9), (E.10), and (E.13) in Supplementary Appendix E.2, respectively, with $\boldsymbol{C}_\tau$ replacing $\boldsymbol{C}$ everywhere, and $H = \mathrm{tr}(\boldsymbol{M}\boldsymbol{\Phi}^{-1}\boldsymbol{L}^\tau\boldsymbol{\Phi}^{-1}\boldsymbol{L}^\tau)/[T(T-1)]$.

On many occasions, one may be interested in the question of convergence over a long time horizon, under which T and τ are of comparable size, without assuming availability of the pre-time observations. (Now the effective sample size is T instead of $T+\tau$, cf. Footnote 3.) For example, given a panel of cross-country observations over 10 years, one may ask whether the inequality measures in these countries under study converge over a horizon of 8 years. In general, the τth-order difference GMM approach of Caselli et al. (1996) does not work given that no data is available for one to take the τth order difference when τ and T are of comparable size. First consider $T = \tau+1$, namely, $y_{i,\tau+1} = \alpha_i + \phi y_{i,1} + u_{i,\tau+1}$. Without any pre-time observations, none of the estimation strategies will work, given that no IV is available and one cannot wipe out the individual effects α_i in WG or OLS1. When $T = \tau+2$, surprisingly, the WG, OLS1, and II procedures yield the same *consistent* estimator. The intuition is as follows. For the WG procedure, one is actually regressing $y_{i,\tau+2} - (y_{i,\tau+2} + y_{i,\tau+1})/2$ on $y_{i,2} - (y_{i,2} + y_{i,1})/2$ and for the OLS1 estimator, one is regressing $y_{i,\tau+2} - y_{i,\tau+1}$ on $y_{i,2} - y_{i,1}$, which is essentially the same. According to (25), with T replaced by $T - \tau = 2$, $j = \lfloor (2-1)/\tau \rfloor = 0$ and thus $\mathrm{E}(\boldsymbol{y}'_{-\tau}\boldsymbol{A}\boldsymbol{u}) = 0$. (It should be noted that their standard errors may still be different since their variances are constructed differently.) One can imagine that when II tries to bias-correct the consistent WG estimator, it may perform worse than the WG estimator in finite samples. For the differenced equation, namely, $y_{i,\tau+2} - y_{i,\tau+1} = \phi(y_{i,2} - y_{i,1}) + u_{i,\tau+2} - u_{i,\tau+1}$, y_{i1} and y_{i2} are legitimate instruments if one wants to use the GMM. When $T = \tau + 3$ and if $\tau > 2$, then $j = \lfloor (3-1)/\tau \rfloor = 0$ and both WG and II yield the same consistent estimator. Again, the II procedure may be redundant. For the differenced equation, namely, $y_{i,t} - y_{i,t-1} = \phi(y_{i,t-\tau} - y_{i,t-\tau-1}) + u_{i,t} - u_{i,t-1}$, $t = \tau+3, \tau+2$, one can use y_{i1} and y_{i2} as instruments for $t = \tau+2$ and y_{i1}, y_{i2}, and y_{i3} as instruments for $t = \tau+3$ in the GMM framework.[14] One can continue this analysis and depending on

[14] If one uses the variable in level in regression, then the differenced lagged variables as IV will also work, see Arellano and Bover (1995).

the magnitudes of $T-\tau$ and τ, the WG estimator may be consistent and there are many different choices of instruments for the GMM estimator, depending on whether one uses the variable in difference or level in the regression. In the next section, simulations are conducted to verify these predictions.

2.6. Robust II estimator

One may wonder whether the II approach, when one matches the inconsistent WG estimator with its analytical approximate expectation, can be extended to cases when some of the classical assumptions are relaxed. Consider the scenario of cross-sectional heteroskedasticity. If heteroskedasticity exists only among α_i, then the analysis in this paper is not affected. If heteroskedasticity exists instead among u_{it}, it will make the analysis more complicated. Suppose $\mathrm{E}(u_{it}u_{jt})=\sigma_{ij,t}$ (when $i=j$, σ_{it}^2). Now u_{it} has not only heteroskedasticity (over t or across i or both) but also possible cross-sectional correlation (though no intertemporal correlation). Let $\boldsymbol{\Sigma}_i=\boldsymbol{\Sigma}_{ii}=\mathrm{Dg}(\sigma_{i1}^2,\ldots,\sigma_{iT}^2)$ and $\boldsymbol{\Sigma}_{ij}=\mathrm{Dg}(\sigma_{ij,1},\ldots,\sigma_{ij,T})$, where $\mathrm{Dg}(\cdot)$ generates a diagonal matrix with its arguments spanning the main diagonal. (When Dg has a matrix argument, it means that it applies to the diagonal elements of the matrix argument.) The covariance matrix of $\boldsymbol{u}$, denoted by $\boldsymbol{\Sigma}$, consists of $N\times N$ blocks of $T\times T$ matrices $\boldsymbol{\Sigma}_{ij}$, $i,j=1,\ldots,N$. Note that all the $T\times T$ blocks are diagonal, but $\boldsymbol{\Sigma}$ itself is not diagonal. Then, for $t=1,\ldots,p$, by using $\mathrm{tr}(\boldsymbol{D}_t\boldsymbol{M}\boldsymbol{\Sigma}_i\boldsymbol{M})=\mathrm{tr}(\boldsymbol{D}_t\boldsymbol{\Sigma}_i)+T^{-2}\mathrm{tr}(\boldsymbol{D}_t)\mathrm{tr}(\boldsymbol{\Sigma}_i)-2T^{-1}\mathrm{tr}(\boldsymbol{D}_t\boldsymbol{\Sigma}_i)$ and $\mathrm{tr}(\boldsymbol{M}\boldsymbol{\Sigma}_i)=(1-T^{-1})\mathrm{tr}(\boldsymbol{\Sigma}_i)$, one has

$$\begin{aligned}\mathrm{E}(\boldsymbol{u}_i'\boldsymbol{M}\boldsymbol{\Phi}_p^{-1}\boldsymbol{L}^t\boldsymbol{u}_i)&=\mathrm{tr}(\boldsymbol{M}\boldsymbol{\Phi}_p^{-1}\boldsymbol{L}^t\boldsymbol{\Sigma}_i)=\mathrm{tr}(\boldsymbol{D}_t\boldsymbol{\Sigma}_i)\\&=\frac{T}{T-2}\mathrm{tr}(\boldsymbol{D}_t\boldsymbol{M}\boldsymbol{\Sigma}_i\boldsymbol{M})-\frac{1}{(T-1)(T-2)}\mathrm{tr}(\boldsymbol{D}_t)\mathrm{tr}(\boldsymbol{M}\boldsymbol{\Sigma}_i)\\&=\frac{T}{T-2}\mathrm{E}(\boldsymbol{u}_i'\boldsymbol{M}\boldsymbol{D}_t\boldsymbol{M}\boldsymbol{u}_i)-\frac{1}{(T-1)(T-2)}\mathrm{tr}(\boldsymbol{D}_t)\mathrm{E}(\boldsymbol{u}_i'\boldsymbol{M}\boldsymbol{u}_i)\\&=\mathrm{E}(\boldsymbol{u}_i'\boldsymbol{M}\boldsymbol{E}_t\boldsymbol{M}\boldsymbol{u}_i),\end{aligned}\tag{26}$$

where

$$\boldsymbol{D}_t=\mathrm{Dg}(\boldsymbol{M}\boldsymbol{\Phi}_p^{-1}\boldsymbol{L}^t),\quad \boldsymbol{E}_t=\frac{T}{T-2}\boldsymbol{D}_t-\frac{\mathrm{tr}(\boldsymbol{D}_t)}{(T-1)(T-2)}\boldsymbol{I}.\tag{27}$$

Correspondingly, in view of (A.4) in Supplementary Appendix A,

$$\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})=\sum_{i=1}^{N}\begin{pmatrix}\mathrm{E}(\boldsymbol{u}_i'\boldsymbol{M}\boldsymbol{E}_1\boldsymbol{M}\boldsymbol{u}_i)\\ \vdots\\ \mathrm{E}(\boldsymbol{u}_i'\boldsymbol{M}\boldsymbol{E}_p\boldsymbol{M}\boldsymbol{u}_i)\\ \boldsymbol{0}_k\end{pmatrix}.\tag{28}$$

Since $\boldsymbol{M}\boldsymbol{u}_i=\boldsymbol{M}\boldsymbol{y}_i-\boldsymbol{M}\boldsymbol{Y}_i\boldsymbol{\phi}_0-\boldsymbol{M}\boldsymbol{X}_i\boldsymbol{\beta}_0$, now consider

$$\begin{aligned}\hat{\mathrm{E}}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})&=\sum_{i=1}^{N}\begin{pmatrix}(\boldsymbol{M}\boldsymbol{y}_i-\boldsymbol{M}\boldsymbol{Y}_i\boldsymbol{\phi}_0-\boldsymbol{M}\boldsymbol{X}_i\boldsymbol{\beta}_0)'\boldsymbol{E}_1(\boldsymbol{M}\boldsymbol{y}_i-\boldsymbol{M}\boldsymbol{Y}_i\boldsymbol{\phi}_0-\boldsymbol{M}\boldsymbol{X}_i\boldsymbol{\beta}_0)\\ \vdots\\ (\boldsymbol{M}\boldsymbol{y}_i-\boldsymbol{M}\boldsymbol{Y}_i\boldsymbol{\phi}_0-\boldsymbol{M}\boldsymbol{X}_i\boldsymbol{\beta}_0)'\boldsymbol{E}_p(\boldsymbol{M}\boldsymbol{y}_i-\boldsymbol{M}\boldsymbol{Y}_i\boldsymbol{\phi}_0-\boldsymbol{M}\boldsymbol{X}_i\boldsymbol{\beta}_0)\\ \boldsymbol{0}_k\end{pmatrix},\\&=\begin{pmatrix}(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{\theta}_0)'\boldsymbol{A}(\boldsymbol{I}_N\otimes\boldsymbol{E}_1)\boldsymbol{A}(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{\theta}_0)\\ \vdots\\ (\boldsymbol{y}-\boldsymbol{W}\boldsymbol{\theta}_0)'\boldsymbol{A}(\boldsymbol{I}_N\otimes\boldsymbol{E}_p)\boldsymbol{A}(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{\theta}_0)\\ \boldsymbol{0}_k\end{pmatrix}\equiv\boldsymbol{f}(\boldsymbol{\theta}_0,\boldsymbol{W})=\boldsymbol{f}_N(\boldsymbol{\theta}_0),\end{aligned}\tag{29}$$

which is a function of $\boldsymbol{\theta}_0$ and the observable sample data. It is obvious that $\mathrm{E}[\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u}-\hat{\mathrm{E}}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})]=\boldsymbol{0}$ and one may define the recentering term $\boldsymbol{d}_N^{\dagger}=(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}\hat{\mathrm{E}}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})$. Numerically, one can still follow the II procedure to solve for the unknown $\boldsymbol{\theta}_0$ based on the sample binding function. The challenge though is regarding inference, since the asymptotic variance matrix of the resulting estimator involves $\boldsymbol{\Sigma}$, which is not diagonal and thus a typical White-type asymptotic variance estimator is not available.

When $\mathrm{E}(u_{it}^2)=\sigma_t^2$ and $\mathrm{Var}(\boldsymbol{u}_i)=\mathrm{Dg}(\sigma_1^2,\ldots,\sigma_T^2)\equiv\boldsymbol{\Sigma}_T$, namely, when there is time-series heteroskedasticity only, as considered in Alvarez and Arellano (2022), the II estimation procedure based on $\boldsymbol{d}_N^{\dagger}(\boldsymbol{\theta})=(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}\boldsymbol{f}_N(\boldsymbol{\theta})$ carries through. Formally, Assumption 1 is modified as follows:

Assumption 1′. The series of error terms u_{it}, $i=1,\ldots,N$, $t=1,\ldots,T$, is independent across time and individuals, $\mathrm{E}(u_{it})=0$, $\mathrm{Var}(u_{it})=\sigma_t^2$, and has finite moments up to the fourth order.

Now $\boldsymbol{\Sigma} = \boldsymbol{I}_N \otimes \boldsymbol{\Sigma}_T$,

$$\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u}) = \begin{pmatrix} N\mathrm{tr}(\boldsymbol{\Sigma}_T\boldsymbol{M}\boldsymbol{\Phi}_p^{-1}\boldsymbol{L}) \\ \vdots \\ N\mathrm{tr}(\boldsymbol{\Sigma}_T\boldsymbol{M}\boldsymbol{\Phi}_p^{-1}\boldsymbol{L}^p) \\ \boldsymbol{0}_k \end{pmatrix} = \mathrm{E}[\boldsymbol{f}_N(\boldsymbol{\theta}_0)],$$

and

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \boldsymbol{d}_N^{\dagger}) = \sqrt{N}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}\left[\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} - \boldsymbol{f}_N(\boldsymbol{\theta}_0)\right], \\ \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \boldsymbol{V}), \tag{30}$$

where

$$\boldsymbol{V} = \lim_{N\to\infty}[N^{-1}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}N^{-1}\mathrm{Var}[\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} - \boldsymbol{f}_N(\boldsymbol{\theta}_0)][N^{-1}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}. \tag{31}$$

The sample Jacobian at $\boldsymbol{\theta}$ is, in view of (29),

$$\begin{aligned}\boldsymbol{G}_N(\boldsymbol{\theta}) &= \boldsymbol{I}_m + (\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}\frac{\partial \boldsymbol{f}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \\ &= \boldsymbol{I}_m - 2(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}\begin{pmatrix} (\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\theta})'\boldsymbol{A}(\boldsymbol{I}_N \otimes \boldsymbol{E}_1)\boldsymbol{A}\boldsymbol{W} \\ \vdots \\ (\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\theta})'\boldsymbol{A}(\boldsymbol{I}_N \otimes \boldsymbol{E}_p)\boldsymbol{A}\boldsymbol{W} \\ \boldsymbol{0}_{k\times m} \end{pmatrix} \\ &\quad + (\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}\sum_{i=1}^{N}\begin{pmatrix} \boldsymbol{Z}_{ip} & \boldsymbol{0}_{p\times k} \\ \boldsymbol{0}_{k\times p} & \boldsymbol{0}_k \end{pmatrix},\end{aligned} \tag{32}$$

where $\boldsymbol{Z}_{ip}$ is a $p \times p$ matrix consisting of $\mathrm{tr}[\boldsymbol{z}_i(\boldsymbol{\theta})\boldsymbol{z}_i(\boldsymbol{\theta})'\partial \boldsymbol{E}_t(\boldsymbol{\theta})/\partial \phi_s]$ in its (t, s) position, $t, s = 1, \ldots, p$ with $\boldsymbol{z}_i(\boldsymbol{\theta}) = \boldsymbol{M}\boldsymbol{y}_i - \boldsymbol{M}\boldsymbol{Y}_i\boldsymbol{\phi} - \boldsymbol{M}\boldsymbol{X}_i\boldsymbol{\beta}$. From (27), one has

$$\frac{\partial \boldsymbol{E}_t(\boldsymbol{\theta})}{\partial \phi_s} = \frac{T}{T-2}\mathrm{Dg}(\boldsymbol{M}\boldsymbol{\Phi}_p^{-1}(\boldsymbol{\phi})\boldsymbol{L}^s\boldsymbol{\Phi}_p^{-1}(\boldsymbol{\phi})\boldsymbol{L}^t) - \frac{\mathrm{tr}(\boldsymbol{M}\boldsymbol{\Phi}_p^{-1}(\boldsymbol{\phi})\boldsymbol{L}^s\boldsymbol{\Phi}_p^{-1}(\boldsymbol{\phi})\boldsymbol{L}^t)}{(T-1)(T-2)}\boldsymbol{I}.$$

Note that this Jacobian function does not depend on $\boldsymbol{\Sigma}_T$.

Theorem 2. *Under Assumptions 2 to 5 and 1′ and that $\boldsymbol{G}(\boldsymbol{\theta}) = \mathrm{plim}_{N\to\infty}\boldsymbol{G}_N(\boldsymbol{\theta})$ is nonsingular in a neighborhood of $\boldsymbol{\theta}_0$, $\check{\boldsymbol{\theta}}$ based on the sample binding function $\mathbf{b}_N(\boldsymbol{\theta}) = \boldsymbol{\theta} + (\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}\boldsymbol{f}_N(\boldsymbol{\theta})$ has the following asymptotic distribution:*

$$\sqrt{N}(\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \boldsymbol{V}_{II}), \tag{33}$$

where $\boldsymbol{V}_{II} = \boldsymbol{G}^{-1}\boldsymbol{V}\boldsymbol{G}^{-1\prime}$, $\boldsymbol{G} = \boldsymbol{G}(\boldsymbol{\theta}_0)$, and $\boldsymbol{V}$ is given by (31).

A consistent estimator of $\boldsymbol{V}_{II}$ can be constructed by using $\hat{\boldsymbol{G}} = \boldsymbol{G}_N(\check{\boldsymbol{\theta}})$ and the White-type estimator of $\boldsymbol{V}$ in (31). That is,

$$\hat{\boldsymbol{V}}_{II} = \hat{\boldsymbol{G}}^{-1}N(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}\left(\sum_{i=1}^{N}\hat{\boldsymbol{v}}_i\hat{\boldsymbol{v}}_i'\right)(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}\hat{\boldsymbol{G}}^{-1\prime}, \tag{34}$$

where

$$\hat{\boldsymbol{v}}_i = \boldsymbol{W}_i'\boldsymbol{M}(\boldsymbol{y}_i - \boldsymbol{W}_i\check{\boldsymbol{\theta}}) - \begin{pmatrix} (\boldsymbol{y}_i - \boldsymbol{W}_i\check{\boldsymbol{\theta}})'\boldsymbol{M}\boldsymbol{E}_1\boldsymbol{M}(\boldsymbol{y}_i - \boldsymbol{W}_i\check{\boldsymbol{\theta}}) \\ \vdots \\ (\boldsymbol{y}_i - \boldsymbol{W}_i\check{\boldsymbol{\theta}})'\boldsymbol{M}\boldsymbol{E}_p\boldsymbol{M}(\boldsymbol{y}_i - \boldsymbol{W}_i\check{\boldsymbol{\theta}}) \\ \boldsymbol{0}_k \end{pmatrix}. \tag{35}$$

It should be pointed out that in this paper the interest centers on estimating the mean parameters $\boldsymbol{\theta}_0$ and treat the variance parameter(s) as nuisance. This stands in contrast to Alvarez and Arellano's (2022) approach that also estimates the variance parameter(s). One can imagine that for cases of moderate T, the finite-sample performance of such an estimator may not be good. Alvarez and Arellano (2022) did not explicitly provide the asymptotic variance result under a general non-normal distribution and they recommended using numerical score and Hessian functions.

Note that the definition of $\boldsymbol{E}_t$ as in (27) assumes that $T > 2$. When $T = 2$, $\boldsymbol{\Sigma}_2 = \mathrm{Dg}(\sigma_1^2, \sigma_2^2)$. From Supplementary Appendix A, $\boldsymbol{\Phi}_1^{-1}\boldsymbol{L}\boldsymbol{\Sigma}_2$ is a 2×2 matrix with σ_1^2 in the lower left position and zero elsewhere, and $\boldsymbol{\Phi}_s^{-1}\boldsymbol{L}\boldsymbol{\Sigma}_2$ is a 2×2

matrix of zeros for any $s = 2, \ldots, p$. Then the only non-zero component of $\mathrm{E}(\boldsymbol{W}_i'\boldsymbol{M}\boldsymbol{u}_i)$ is $\mathrm{tr}(\boldsymbol{M}\boldsymbol{\Phi}_1^{-1}\boldsymbol{L}\boldsymbol{\Sigma}_2) = -\sigma_1^2/2$. In this case, a consistent estimator of σ_1^2 is not straightforward.[15] Similarly, one can show that taking the first-order difference cannot distinguish $\sigma_1^2 \neq \sigma_2^2$ from $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and the II procedure will not work. Instead, the IV estimator of Anderson and Hsiao (1981) can be used, which is shown in Alvarez and Arellano (2022) to be the same as their random effects maximum likelihood (RML) estimator.

3. Monte Carlo simulations

This section presents Monte Carlo experiments designed to assess the finite-sample performance of the proposed estimator and the resulting inference procedure in comparison with existing approaches. In particular, the first subsection compare the size performances from II and BC for DP(1). Recall that numerically the II estimator and that from BC are identical, but the resulting inferences may be different because the estimated variances are calculated differently. The second subsection compared the simulation-free II estimator in this paper and the simulation-based II estimator of Gouriéroux et al. (2010). The third subsection considers a correlated design when the covariate is correlated with the fixed effects. The next two subsections include results for higher-order and DP-τ models. The number of Monte Carlo replications is 10,000 for DP(1) and higher-order dynamic panels and it is 1000 for DP-τ.[16] "Size(%)" in the relevant tables refers to the empirical rejection rate (%) of the two-sided 5% t test of the relevant parameter equal to its true value under the asymptotic distribution pertaining to the corresponding estimation strategy. All the reported bias and root mean squared error (RMSE) results in the relevant tables are multiplied by 100.

3.1. DP(1): II versus BC

To facilitate a direct comparison, the same design in BC is adopted. Specifically, the data generating process (DGP) used is as follows:

$$\begin{aligned} y_{it} &= \alpha_i + \phi y_{i,t-1} + \beta x_{it} + u_{it}, \ \ \alpha_i \sim \text{i.i.d.N}(0,1), \ \ u_{it} \sim \text{i.i.d.}(0,1), \\ x_{it} &= 0.8 x_{i,t-1} + \xi_{it}, \ \ \xi_{it} \sim \text{i.i.d.N}(0,1), \end{aligned} \tag{36}$$

where the true parameter values are $\phi_0 = 0.8$ and $\beta_0 = 1$ and the combinations of N and T satisfy $NT = 600$. In addition to the baseline normal error distribution, four non-normal distributions are considered: uniform on [0, 1], student-t distribution with 5 degrees of freedom (t_5), log-normal distribution lnN(0, 1), and mixture of N(−3, 1) and N(3, 1) with half probability each. They are recentered and standardized so that each has mean 0 and variance 1. Among the four non-normal distributions, lnN(0, 1) is skewed (with $\gamma_1 = 6.1849$), the uniform and normal mixture distributions have negative excess kurtosis coefficients ($\gamma_2 = -1.2$ and -1.62, respectively) and the t_5 and log-normal distribution have positive excess kurtosis values of 6 and 110.9364, respectively.

Table 1 presents the size performances of inferences based on the estimated variances calculated from $\boldsymbol{V}_{BC}$ and $\boldsymbol{V}_{II}$ in this paper.[17] To better understand the effects of different error distributions, the initial observations in Table 1 are fixed ($y_{i0} = \alpha_i/(1-\phi_0)$) so that a direct comparison with BC is possible.

The overall picture is that when one uses $\boldsymbol{V}_{BC}$ to conduct inference on ϕ, it can result in nonnegligible size distortions, depending on the specific error distributions, but $\boldsymbol{V}_{II}$ delivers very reliable size performance, regardless of the error distribution. In more detail, one observes that under normally distributed errors, the empirical size of the t test from BC is somewhat low when T is small. Given the fixed initial conditions, this phenomenon is consistent with earlier discussion of the difference between $\boldsymbol{V}_{BC}$ and $\boldsymbol{V}_{II}$ in Section 2.4.1, where it is claimed that $\boldsymbol{V}_{BC}$ over-estimates the variance of $\hat{\phi}_{II}$ inherently. Furthermore, when T is small ($T = 2, 3$), under the uniform distribution, where the errors exhibit zero skewness but negative excess kurtosis, the downward size distortion from BC is more severe, in comparison with the normal case. This is again consistent with the previous argument that negative kurtosis exacerbates the over-estimation problem of $\boldsymbol{V}_{BC}$ for $\hat{\phi}_{II}$. When the errors possess positive excess kurtosis such as the t_5 distribution, the under-size problem may disappear, but when the excess kurtosis is of a very large magnitude such as the log-normal errors, the t test from BC becomes more liberal, especially when T is small. This is in line with earlier analysis that positive excess kurtosis may be large enough to offset, and possibly overtake, the inherent negative term missing in $\boldsymbol{V}_{0,BC}$. For β, both $\boldsymbol{V}_{BC}$ and $\boldsymbol{V}_{II}$ deliver good sizes across T and different error distributions, with an exception that under the log-normal errors, $\boldsymbol{V}_{BC}$ results in slight size distortion (around 9%) when $T = 2$.

---

15 If $\boldsymbol{\theta}_0$ is known, let $\boldsymbol{K}$ be a $N \times 2N$ selection matrix such that $\boldsymbol{K}\boldsymbol{A}(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\theta}_0)$ picks up elements of $\boldsymbol{A}(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\theta}_0)$ corresponding to $t = 1$. Then one may be tempted to use $(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\theta}_0)'\boldsymbol{A}\boldsymbol{K}'\boldsymbol{K}\boldsymbol{A}(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\theta}_0)/N$ to estimate σ_1^2, but it in fact estimates $\mathrm{Var}(u_{i1} - (u_{i1} + u_{i2})/2) = (\sigma_1^2 + \sigma_2^2)/4$. The challenge here is that $\mathrm{tr}(\boldsymbol{M}\boldsymbol{\Phi}_1^{-1}\boldsymbol{L}\boldsymbol{\Sigma}_2) = -\sigma_1^2/2$ cannot distinguish the case $\sigma_1^2 \neq \sigma_2^2$ from $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

16 Matlab's `fminsearch` is used to search for the solution that numerically minimizes the Euclidean distance between $\boldsymbol{b}_N(\boldsymbol{\theta})$ and $\hat{\boldsymbol{\theta}}$, which is equivalent to numerically inverting the sample binding function. It results in no numerical failure, up to the default termination tolerance, in all the simulations in this section. `fsolve`, however, occasionally yields numerical failures and thus it is not used.

17 The bias and RMSE results match almost exactly those from BC and they are omitted. Results related to GMM and WG, which were considered in BC, are also omitted due to the same reason.

Table 1
Rejection rates (%) of 5% t tests from BC and II in DP(1).

Error distribution		(N, T)	(300,2)	(200,3)	(150,4)	(100,6)	(60,10)	(40,15)
Normal	ϕ	BC	2.68	3.55	3.75	4.19	4.86	4.71
		II	5.34	5.41	5.28	5.67	6.02	5.80
	β	BC	4.94	5.22	5.09	5.17	5.39	4.78
		II	5.41	5.61	5.63	5.53	5.48	5.29
Uniform	ϕ	BC	1.82	2.79	3.34	4.06	4.60	5.46
		II	5.16	5.23	5.49	5.81	5.68	6.48
	β	BC	4.71	5.02	5.13	5.09	4.84	5.11
		II	5.71	5.45	5.39	5.17	5.27	6.09
t_5	ϕ	BC	5.12	4.99	4.64	4.95	5.04	4.82
		II	4.95	5.42	5.53	5.62	5.85	5.70
	β	BC	5.35	5.26	5.03	4.81	5.01	5.15
		II	5.38	5.27	5.25	5.67	5.57	5.53
lnN(0, 1)	ϕ	BC	22.44	16.26	12.79	8.82	6.33	5.70
		II	4.25	3.88	4.19	4.85	5.01	5.84
	β	BC	8.68	6.80	5.75	5.30	5.32	5.02
		II	5.11	5.06	4.75	5.13	5.26	5.05
Mixture	ϕ	BC	1.58	2.96	3.51	3.95	4.77	5.59
		II	5.33	5.58	5.74	5.62	5.72	6.53
	β	BC	4.41	4.76	4.68	5.42	4.75	4.72
		II	5.23	5.29	5.09	5.94	5.28	5.64

Notes: $\phi_0 = 0.8$, $\beta_0 = 1$, $\alpha_i \sim$ i.i.d.N(0, 1), $y_{i0} = \alpha_i/(1-\phi_0)$, $x_{it} = 0.8x_{i,t-1} + \xi_{it}$, $\xi_{it} \sim$ i.i.d.N(0, 1).

In summary, BC's size performance is not stable across different error specifications, especially when T is small. To further demonstrate the effects on hypothesis testing due to the difference between $\boldsymbol{V}_{BC}$ and $\boldsymbol{V}_{II}$, a second set of simulations are conducted. Given that there are multiple factors that contribute to the difference, each time only one factor is adjusted while the other factors are fixed. The sample size is set at $(N, T) = (300, 2)$. In Fig. 1, the top two sub-figures (first row) plot the empirical rejection probabilities of the t tests for ϕ using $\boldsymbol{V}_{BC}$ and $\boldsymbol{V}_{II}$ under normal errors and mixture normal with negative excess kurtosis across different values of ϕ_0 (with initial observations $y_{i0} = \alpha_i/(1-\phi_0)$). These two sub-figures clearly indicate the persistence of BC's under-size issue. In the second row, the left sub-figure conducts an experiment where the errors display zero excess kurtosis but its skewness changes.[18] With other factors fixed, the size of BC-based t test remains stable across different degrees of skewness and it is under-sized. The right sub-figure uses the same error distribution at a given skewness level and varies the value of ϕ_0. Again, it documents the persistent under-size issue of the t test from BC. The third row investigates the effects of the excess kurtosis parameter γ_2. The left sub-figure uses the mixture of normal errors (see Supplementary Appendix F) with negative excess kurtosis, and it is clear that as the magnitude of negative excess kurtosis increases, the under-size issue of BC-based t test is more pronounced, confirming that negative excess kurtosis leads to over-estimation of the variance of $\hat{\phi}_{II}$ from $\boldsymbol{V}_{BC}$. The right sub-figure uses a certain t-distribution to fix the condition $\gamma_1 = 0$ and varies γ_2 in the positive domain.[19] Interestingly, within the positive region of γ_2, the empirical size of BC-based t test increases as γ_2 grows and after a certain point (around $\gamma_2 = 9$), it starts to exceed the nominal 5% rate. Again, this confirms the previous analysis of the effects of γ_2 in Section 2.4.1.

The last row in Fig. 1 studies the role of the initial conditions. The errors are simulated as standard normal but the initial observations are generated by $y_{i0} = \psi\alpha_i + \omega_i$, where $\omega_i \sim$ i.i.d.N$(0, \sigma_\omega^2)$. By changing ψ and σ_ω^2, one sees that the size performance of BC-based t test is pretty stable (between 2% and 3% throughout). Keep in mind that changing values of y_{i0} affects also the bread part of the sandwich variance matrices and its overall effect may not be obvious. Across all configurations, the II procedure achieves outstanding size performance, with its empirical rejection rate very close to the nominal size.

3.2. DP(1): II versus simulation-based II

This subsection compares the simulation-free II estimator proposed in this paper and its robust version (denoted by "IIr") with the simulation-based II estimator of Gouriéroux et al. (2010) (denoted by "II*") under both homoskedasticity

[18] Supplementary Appendix F provides details of constructing such errors by employing a location mixture of two normal distributions.

[19] It is infeasible to construct an error distribution that satisfies zero mean, unit variance, zero skewness but an arbitrary magnitude of positive excess kurtosis at the same time using the location mixture (see Supplementary Appendix F for details). Alternatively, one can easily construct such a distribution via a t distribution. For a given γ_2 level, say $\gamma_2 = a$, $a \in (0, \infty)$, the error term can be simulated as $u_{it} = \sqrt{(a+3)/(2a+3)}e_{it}$, where e_{it} follows a t distribution with degrees of freedom $6/a + 4$.

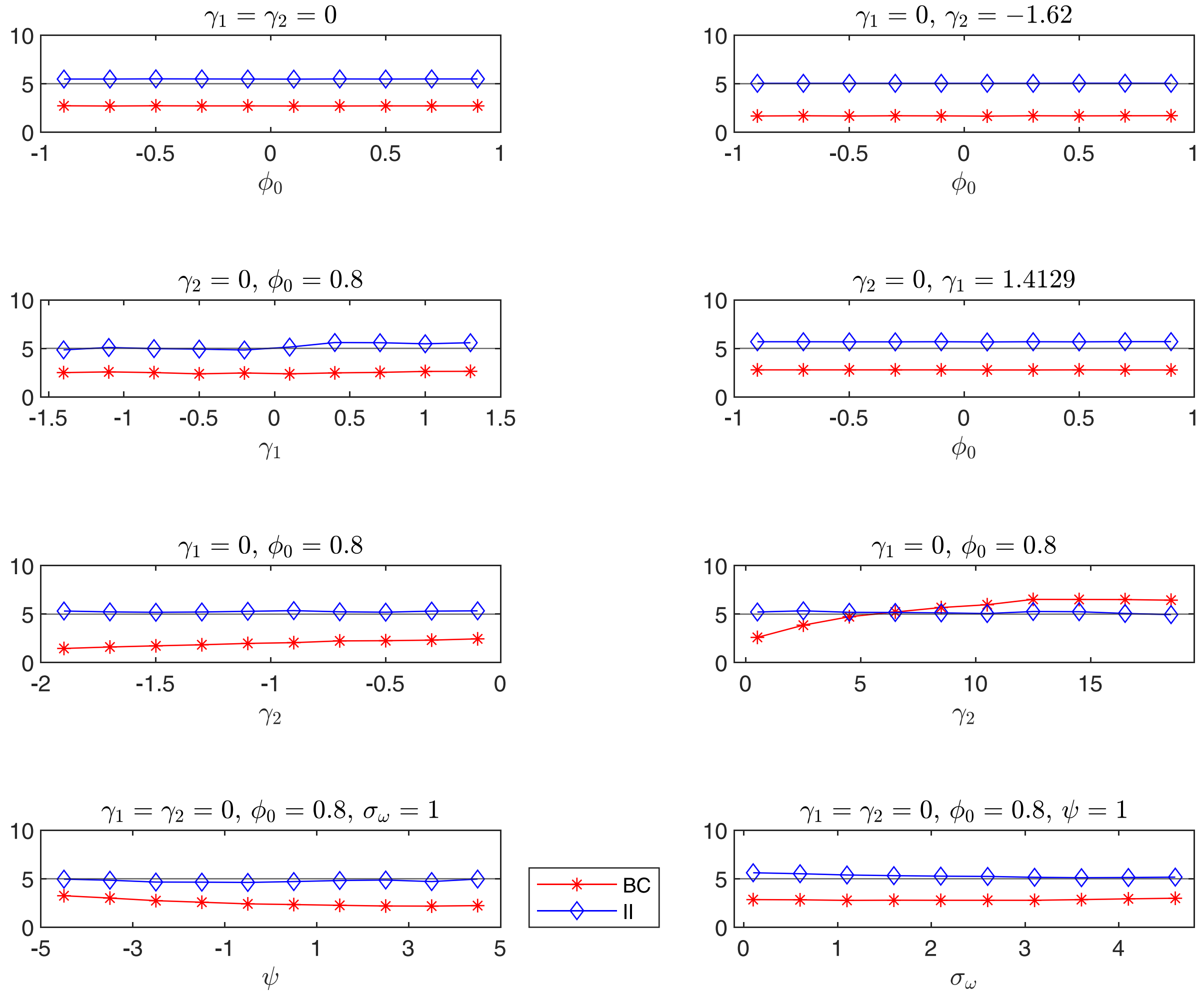

Fig. 1. Rejection rates (%) of 5% t tests from BC and II ($N = 300$, $T = 2$).

and heteroskedasticity ($\text{Var}(u_{it}) = t$). The DGP is the same as (36) and Table 2 reports results under normal errors.[20] Note that Gouriéroux et al. (2010) did not estimate β and the asymptotic distribution result is under the assumptions of large N and large T. The resulting variance formula gives rise to empirical rejection rates of almost always around 100% in the experiment configurations considered (when T is relatively small and there is $\boldsymbol{X}$) for the 5% t test and thus only the bias and RMSE results (pertaining to the estimated ϕ) are reported in Table 2. (This is also the case for II* in the next subsection.) Further, Gouriéroux et al. (2010) assumed $|\phi_0| < 1$ and thus Table 2 does not contain the case of $\phi_0 = 1$.

Notably, the simulation-based II estimator performs worse in terms of bias than the simulation-free II estimator (and its robust version) under homoskedasticity. Its RMSE also performs worse than the simulation-free one, especially when ϕ_0 is small (0.4) and occasionally better when ϕ_0 is large. Under heteroskedasticity, the simulation-based II does not necessarily perform worse, even though homoskedastic pseudo errors are simulated. The simulation-free robust II is dominantly better in terms of bias. It also fares much better in terms of RMSE on most occasions.

Arguably, the somewhat disappointing performance of II* relative to the simulation-free II may be due to two factors. First, Gouriéroux et al. (2010) did not provide Monte Carlo results when there are exogenous regressors. Second, the inference procedure in Gouriéroux et al. (2010) requires large T. Table 3 presents results when T is in fact relatively large and there is no $\boldsymbol{X}$. Now the simulated-based II performs relatively better on some occasions, though not universally, than the simulation-free II in terms of bias and RMSE. Included also in Table 3 are the bias-corrected estimator of Hahn and

[20] See equation (19) of Gouriéroux et al. (2010) with exogenous regressors. In Table 2, $S = 250$ simulations are used in building the binding function for II*. As documented in Gouriéroux et al. (2010), increasing the number of simulations S gives rise to little improvement of their estimator. When u_{it} is independently and normally distributed with mean 0 and variance t, it is denoted by $u_{it} \sim \text{i.d.N}(0, t)$. Under heteroskedasticity, the pseudo error term in the simulation-based II estimator of Gouriéroux et al. (2010) is simulated as i.i.d.N(0, 1) given that one typically has no knowledge of the variance structure.

Table 2
Bias and RMSE results from II and simulation-based II (II*) in DP(1).

u_{it}	ϕ_0		(N,T)	(300,2)	(200,3)	(150,4)	(100,6)	(60,10)	(40,15)
i.i.d.N(0, 1)	0.4	Bias	II	0.22	0.06	0.00	−0.06	−0.11	−0.08
			IIr	0.00	0.13	0.01	−0.06	−0.11	−0.08
			II*	28.35	22.09	23.46	23.09	18.04	12.86
		RMSE	II	5.79	4.31	3.67	3.12	2.72	2.44
			IIr	0.00	4.81	3.81	3.14	2.72	2.45
			II*	28.68	22.20	23.56	23.22	18.28	13.14
	0.8	Bias	II	0.22	0.09	0.05	0.00	−0.06	−0.05
			IIr	0.00	0.16	0.06	0.00	−0.06	−0.05
			II*	6.74	3.16	5.50	7.63	9.14	9.60
		RMSE	II	5.79	3.80	2.99	2.27	1.74	1.43
			IIr	0.00	4.27	3.13	2.31	1.74	1.43
			II*	7.44	3.50	5.63	7.70	9.18	9.64
	0.95	Bias	II	0.22	0.09	0.07	0.03	−0.02	−0.01
			IIr	0.00	0.16	0.08	0.03	−0.02	−0.01
			II*	−3.17	−4.64	−2.22	−0.10	1.32	1.93
		RMSE	II	5.79	3.57	2.66	1.85	1.25	0.96
			IIr	0.00	4.02	2.79	1.88	1.25	0.96
			II*	4.13	4.83	2.45	0.77	1.43	1.98
i.d.N(0, t)	0.4	Bias	II	29.40	19.04	12.02	5.49	1.82	0.67
			IIr	0.00	0.64	0.15	−0.07	−0.24	−0.26
			II*	17.31	13.66	15.61	15.13	10.96	7.46
		RMSE	II	34.55	22.98	15.21	8.29	5.35	4.65
			IIr	0.00	8.42	6.32	5.23	4.75	4.50
			II*	19.78	15.07	16.04	15.71	11.95	8.78
	0.8	Bias	II	29.38	20.49	16.07	11.33	6.34	3.30
			IIr	0.00	0.78	0.27	0.07	−0.17	−0.22
			II*	0.36	−2.89	0.20	3.11	5.16	5.77
		RMSE	II	34.51	24.02	18.99	13.75	8.68	5.72
			IIr	0.00	8.19	5.82	4.47	3.76	3.41
			II*	3.86	3.69	2.01	3.61	5.49	6.15
	0.95	Bias	II	29.45	19.56	15.03	10.86	7.46	5.56
			IIr	0.00	0.78	0.29	0.12	−0.07	−0.09
			II*	−8.59	−10.23	−6.92	−3.81	−1.52	−0.38
		RMSE	II	34.62	22.97	17.73	12.93	9.03	6.88
			IIr	0.00	7.90	5.35	3.84	2.98	2.60
			II*	9.27	10.43	7.13	4.07	1.92	1.12

Notes: $\beta_0 = 1$, $\alpha_i \sim$ i.i.d.N(0, 1), $y_{i0} = \alpha_i/(1-\phi_0)$, $x_{it} = 0.8x_{i,t-1} + \xi_{it}$, $\xi_{it} \sim$ i.i.d.N(0, 1). Bias and RMSE are multiplied by 100.

Kuersteiner (2002, HK for short) and the first difference least squares estimator of Han and Phillips (2010, HP for short), both of which require large T.[21] The WG estimator, which is consistent under large T, is included as well. One can find that the II estimator and its robust version typically correct the bias well and have very good RSME performance, but with upward size distortions when one uses the asymptotic distributions in (9) and (33). Recall that the asymptotic distribution of the II estimator derived in this paper is under the assumption of finite T. The simulation-based II estimator usually achieves the lowest RMSE, though also with upward size distortions. The estimator of Han and Phillips (2010) achieves the best size performance, but when $\phi_0 = 1$, it delivers much higher RMSE than the II estimator. The estimator of Hahn and Kuersteiner (2002) is slightly better than II in terms of RMSE, but may perform worse in terms of bias. It also has upward size distortions, more severely as ϕ_0 goes up. The WG estimator performs reasonably well, as expected, in terms of bias and RMSE, but the associated inference becomes more and more unreliable as ϕ_0 goes up.

Supplementary Appendix H contains results under heteroskedasticity and additional results under other error distributions are available upon request. They lead to similar observations.

3.3. DP(1) under correlated design

In the previous two subsections, the scalar exogenous variable, when present, is not correlated with the fixed effects. For panel models with fixed effects, it may be of more interest to allow for correlation. Consider the following design: $y_{it} = \alpha_i + \phi y_{i,t-1} + \beta x_{it} + u_{it}$, $x_{it} = \rho_i \alpha_i + \xi_{it}$, $\rho_i \sim$ i.i.d. uniform on [0, 1] (i.i.d.U[0, 1] for short), $\xi_{it} \sim$ i.i.d.N(0, 1), $\alpha_i \sim$ i.i.d.N(0, 1). Table 4 reports the bias, RMSE, and empirical rejection rates when $\phi_0 = 0.8$, $\beta_0 = 1$, and u_{it} is

21 Hahn and Kuersteiner (2002) require $|\phi_0| < 1$ (see their Eq. (6)) and Han and Phillips (2010) require $\phi_0 \in (-1, 1]$ (see their Theorem 1) under large T. Further, Han and Phillips (2010) specify the fixed effects as $\alpha_i(1-\phi_0)$ instead of α_i to ensure continuity of their estimator and the resulting asymptotics at $\phi_0 = 1$. When $|\phi_0| < 1$, the two specifications of fixed effects are indistinguishable. Simulations results in Table 3 are based on the specification of Han and Phillips (2010) and for the initial conditions, $y_{i0} = \alpha_i(1-\phi_0) + u_{i0}/\sqrt{1-\phi_0^2}$ when $|\phi_0| < 1$ and $y_{i0} = 0$ when $\phi_0 = 1$.

Table 3
Simulation results for DP(1) (no $\boldsymbol{X}$ and large T) with $u_{it} \sim$ i.i.d.N(0, 1).

N		(T, ϕ_0)	(50,0.4)	(50,0.8)	(50,0.95)	(50,1)	(100,0.4)	(100,0.8)	(100,0.95)	(100,1)
10	Bias	WG	−2.95	−4.17	−5.28	−6.32	−1.53	−2.05	−2.47	−3.21
		HK	−0.21	−0.65	−1.48		−0.14	−0.27	−0.54	
		HP	0.34	0.34	0.22	−0.13	0.08	0.10	0.13	0.06
		II	−0.13	−0.34	−0.01	−0.94	−0.12	−0.19	0.71	−0.44
		IIr	−0.13	−0.34	−0.01	−0.96	−0.12	−0.19	0.72	−0.44
		II*	0.04	0.00	−1.01		−0.04	−0.02	−0.13	
	RMSE	WG	5.20	5.19	5.77	6.65	3.34	2.92	2.82	3.37
		HK	4.36	3.23	2.80		3.00	2.11	1.49	
		HP	7.54	8.59	8.89	8.93	5.34	5.99	6.28	6.43
		II	4.37	3.28	3.24	2.93	3.00	2.12	2.33	1.47
		IIr	4.37	3.28	3.24	2.96	3.00	2.12	2.34	1.48
		II*	4.40	3.22	2.29		3.01	2.11	1.38	
	Size (%)	WG	11.10	27.97	75.16	98.50	8.43	17.09	51.26	98.85
		HK	6.66	9.10	23.59		5.54	7.43	15.27	
		HP	4.86	4.96	4.90	4.46	4.86	4.65	5.22	4.97
		II	9.09	9.47	14.84	21.17	8.78	8.72	20.70	20.07
		IIr	9.11	9.41	14.30	21.36	8.76	8.79	20.80	20.10
		II*	6.93	10.77	13.95		5.72	7.99	15.74	
20	Bias	WG	−2.91	−4.00	−5.07	−6.09	−1.47	−1.96	−2.37	−3.09
		HK	−0.16	−0.48	−1.27		−0.08	−0.18	−0.44	
		HP	0.21	0.35	0.17	0.02	0.06	0.09	0.10	−0.00
		II	−0.08	−0.16	0.38	−0.57	−0.06	−0.1	0.57	−0.27
		IIr	−0.08	−0.16	0.36	−0.57	−0.06	−0.1	0.57	−0.27
		II*	0.01	0.01	−0.92		−0.02	−0.01	−0.12	
	RMSE	WG	4.17	4.56	5.33	6.26	2.55	2.43	2.56	3.18
		HK	3.05	2.30	2.11		2.11	1.47	1.07	
		HP	5.35	6.11	6.29	6.34	3.77	4.23	4.45	4.55
		II	3.06	2.32	2.74	2.14	2.11	1.47	1.86	1.08
		IIr	3.06	2.32	2.72	2.15	2.11	1.47	1.86	1.08
		II*	3.07	2.29	1.71		2.11	1.47	0.98	
	Size (%)	WG	16.18	47.52	95.50	99.99	11.04	28.02	78.87	100.00
		HK	6.19	9.92	29.19		5.63	6.95	16.47	
		HP	4.98	4.88	4.83	4.95	4.93	4.65	5.53	5.01
		II	6.93	7.59	16.86	17.22	6.60	6.74	14.44	17.00
		IIr	6.92	7.56	16.26	17.17	6.60	6.77	14.60	16.73
		II*	6.36	10.96	17.65		5.69	7.63	15.61	
50	Bias	WG	−2.88	−3.90	−4.94	−5.96	−1.41	−1.90	−2.29	−3.02
		HK	−0.14	−0.38	−1.14		−0.02	−0.12	−0.37	
		HP	0.13	0.30	0.07	0.01	0.08	0.09	0.06	0.01
		II	−0.06	−0.07	0.79	−0.49	0.00	−0.04	0.15	−0.24
		IIr	−0.06	−0.07	0.79	−0.49	0.00	−0.04	0.14	−0.25
		II*	−0.02	0.01	−0.85		0.02	0.00	−0.10	
	RMSE	WG	3.44	4.13	5.04	6.03	1.93	2.10	2.37	3.05
		HK	1.92	1.45	1.54		1.34	0.92	0.70	
		HP	3.36	3.83	3.95	4.00	2.37	2.69	2.82	2.86
		II	1.92	1.44	2.52	1.49	1.34	0.92	1.03	0.76
		IIr	1.92	1.44	2.52	1.51	1.34	0.92	1.03	0.77
		II*	1.93	1.42	1.24		1.34	0.92	0.61	
	Size (%)	WG	33.46	84.58	99.99	100.00	18.68	57.00	99.21	100.00
		HK	5.93	10.06	39.98		5.95	6.88	18.66	
		HP	4.55	4.75	4.68	4.55	4.73	5.01	4.92	5.06
		II	5.38	5.76	24.63	16.01	5.96	5.64	6.11	17.00
		IIr	5.39	5.84	24.73	15.43	5.95	5.58	6.04	16.78
		II*	6.00	10.09	26.32		5.98	7.07	14.45	

Notes: $\alpha_i \sim$ i.i.d.N(0, 1), $y_{i0} = \alpha_i(1-\phi_0) + u_{i0}/\sqrt{1-\phi_0^2}$ (= 0 when $\phi_0 = 1$). Bias and RMSE are multiplied by 100.

independently and normally distributed under both homoskedasticity and heteroskedasticity (Var$(u_{it}) = t$). Results related to β are in Supplementary Appendix H. The bias-corrected score, random effects maximum likelihood, concentrated random effects maximum likelihood of Alvarez and Arellano (2022), and their robust versions, denoted by "BCS", "RML", "RMLc", "BCSr", "RMLr", and "RMLcr", respectively, are also included for comparison. "GMM" denotes the (one-step) GMM estimator of Arellano and Bond (1991).

One can observe that the GMM-based inference displays substantial size distortions as T goes up. This is not surprising, given that the GMM estimator is consistent under small T but suffers from high estimation uncertainty when T is not small, as the number of moment conditions is proportional to T^2 when T grows while the number of endogenous variables is fixed in the current context. Under homoskedasticity, the performance of BC is similar to that in Section 3.1, but under heteroskedasticity, it performs much worse, especially in terms of its size. Recall the BC is not designed under heteroskedasticity. The robust II performs very well under both homoskedasticity and heteroskedasticity. The various estimators proposed in Alvarez and Arellano (2022) are beaten on many occasions by the robust II, especially in terms of size performance. As in the previous subsection, the simulation-based II performs reasonably well under homoskedasticity,

Table 4
Bias, RMSE, and rejection rates (%) of 5% t tests from various estimators of ϕ_0 in DP(1).

u_{it}		(N,T)	(300,2)	(200,3)	(150,4)	(100,6)	(60,10)	(40,15)
i.i.d.N(0, 1)	Bias	GMM	−0.10	−1.11	−1.58	−2.52	−3.70	−4.86
		BC	1.75	0.75	0.45	0.05	−0.03	−0.10
		II	1.75	0.75	0.45	0.05	−0.03	−0.10
		BCS	1.77	0.72	0.43	0.02	−0.05	−0.11
		RML	0.23	0.03	−0.05	−0.12	−0.31	−1.19
		RMLc	−0.01	−0.04	−0.06	−0.10	−0.08	−0.09
		IIr		1.74	0.74	0.13	0.00	−0.09
		BCSr		2.02	0.91	0.24	0.15	−0.01
		RMLr	34.93	−8.64	−12.50	−12.65	−10.04	−6.86
		RMLcr	134.43	−0.82	−0.58	−0.49	−0.43	−0.48
		II*	−4.42	−7.82	−3.16	1.18	4.24	5.00
	RMSE	GMM	9.10	6.26	5.11	4.55	4.70	5.46
		BC	13.42	8.02	5.91	4.28	3.14	2.63
		II	13.42	8.02	5.91	4.28	3.14	2.63
		BCS	13.60	8.02	5.91	4.27	3.14	2.63
		RML	6.32	4.33	3.40	2.74	2.73	3.53
		RMLc	5.35	3.89	3.27	2.69	2.22	1.99
		IIr		10.97	7.00	4.54	3.18	2.64
		BCSr		11.72	7.55	4.78	3.36	2.80
		RMLr	110.18	23.16	68.99	16.03	10.68	7.32
		RMLcr	282.99	4.11	3.37	2.77	2.28	2.09
		II*	5.92	8.19	3.82	2.23	4.61	5.37
	Size (%)	GMM	5.24	6.42	6.66	10.76	24.62	52.45
		BC	2.48	2.28	2.21	3.06	3.96	4.49
		II	5.00	4.47	4.45	5.11	5.59	6.12
		BCS	8.19	20.43	24.40	26.68	24.70	21.94
		RML	6.55	6.44	6.29	6.42	8.26	15.06
		RMLc	9.29	10.64	11.66	12.51	13.28	13.95
		IIr		5.22	4.33	4.97	5.53	6.11
		BCSr		6.90	12.75	24.53	25.23	23.32
		RMLr	53.53	15.85	37.51	57.80	60.34	43.78
		RMLcr	93.56	6.96	19.16	34.21	43.30	46.48
i.d.N(0, t)	Bias	GMM	−0.16	−2.72	−4.44	−7.66	−11.08	−12.17
		BC	28.90	22.48	18.30	13.82	9.90	6.46
		II	28.90	22.48	18.30	13.82	9.90	6.46
		BCS	54.03	36.06	27.33	18.48	11.35	6.59
		RML	0.93	0.24	−0.04	−0.47	−1.27	−2.42
		RMLc	−0.02	−0.02	−0.06	−0.14	−0.15	−0.21
		IIr		2.96	1.59	0.47	0.13	−0.10
		BCSr		3.46	1.80	0.61	−0.02	−0.37
		RMLr	21.84	15.03	7.51	3.81	−10.76	−10.06
		RMLcr	7.30	−1.90	−1.39	−1.09	−1.14	−1.25
		II*	−13.51	−15.10	−9.68	−4.58	−0.93	0.22
	RMSE	GMM	14.49	10.85	9.80	10.59	12.49	12.96
		BC	29.80	23.12	18.86	14.37	10.98	8.73
		II	29.80	23.12	18.86	14.37	10.98	8.73
		BCS	56.37	37.15	28.15	19.25	12.70	9.01
		RML	10.18	6.58	5.38	4.51	4.94	5.14
		RMLc	7.30	5.56	4.76	4.13	3.71	3.59
		IIr		15.90	11.15	7.39	5.55	4.52
		BCSr		19.35	14.17	13.39	5.75	4.19
		RMLr	81.42	804.19	469.84	35.82	19.42	11.21
		RMLcr	147.99	7.94	6.20	5.14	4.23	3.88
		II*	15.66	15.48	10.18	5.49	3.38	3.49
	Size (%)	GMM	5.66	7.66	9.05	19.55	49.10	79.23
		BC	0.00	0.01	0.06	0.02	0.20	1.87
		II	0.00	0.00	0.00	0.04	0.46	2.01
		BCS	99.33	99.78	99.68	99.30	97.69	95.41
		RML	0.79	0.42	0.40	0.33	1.48	3.68
		RMLc	1.76	3.37	4.27	5.53	6.55	7.62
		IIr		5.50	4.55	4.66	4.97	5.44
		BCSr		11.82	6.80	23.89	27.70	25.86
		RMLr	41.03	27.33	44.22	44.36	45.42	33.56
		RMLcr	95.42	14.32	29.76	43.78	50.69	53.08

Notes: $\phi_0 = 0.8$, $\beta_0 = 1$, $\alpha_i \sim$ i.i.d.N(0, 1), $y_{i0} = \alpha_i/(1-\phi_0)$, $x_{it} = \rho_i\alpha_i + \xi_{it}$, $\rho_i \sim$ i.i.d.U[0, 1], and $\xi_{it} \sim$ i.i.d.N(0, 1). Bias and RMSE are multiplied by 100.

when the pseudo errors are simulated as such. However, under heteroskedasticity, with simulated homoskedastic errors, its performance can get worse.[22]

[22] Additional results, when $\phi_0 = 0.4$, 0.95, and 1, under homoskedastic and heteroskedastic errors and various values of y_{i0}, not reported here but available upon request from the corresponding author, reveal that the II estimator performs remarkably well across different model specifications and reasonably well even under heteroskedasticity. Its robust version performs very well under both homoskedasticity and heteroskedasticity.

3.4. Higher-order dynamic panels

This subsection provides simulation results for higher-order dynamic panel models. First considered is a second-order dynamic panel model with two exogenous regressors (with one of them correlated with the fixed effects),

$$\begin{aligned} &y_{it} = \alpha_i + \phi_1 y_{i,t-1} + \phi_2 y_{i,t-2} + \beta_1 x_{1,it} + \beta_2 x_{2,it} + u_{it},\ \alpha_i \sim \text{i.i.d.N}(0,1),\\ &x_{1,it} = 0.8 x_{1,it-1} + \xi_{1,it},\ x_{2,it} = \rho_i \alpha_i + \xi_{2,it},\\ &\xi_{1,it} \sim \text{i.i.d.N}(0,1),\ \xi_{2,it} \sim \text{i.i.d.N}(0,1),\quad \rho_i \sim \text{i.i.d.U}[0,1]. \end{aligned} \tag{37}$$

The values of β_1 and β_2 are both set at 1. The initial observations are simulated as $y_{is} = \alpha_i/(1-\phi_{01}-\phi_{02}) + u_{is}\sqrt{\text{Var}(z_s)} + \boldsymbol{x}_{is}'\boldsymbol{\beta}_0$, $s = 0, -1$, where z_s follows a stationary zero-mean second-order autoregressive (AR(2)) process with coefficients ϕ_{01} and ϕ_{02} and its shock term is a unit-variance normal white noise. The combinations of N and T satisfy $NT = 600$. For the GMM estimator, it is based on the first differenced equation $\Delta y_{it} = \phi_1 \Delta y_{i,t-1} + \phi_2 \Delta y_{i,t-2} + \beta_1 \Delta x_{1,it} + \beta_2 \Delta x_{2,it} + \Delta u_{it}$, leading to the following moment conditions: $\text{E}(y_{i,t-s}\Delta u_{it}) = 0,\ t = 2,\ldots,T,\ s = 2,\ldots,t$; $\text{E}(x_{1,is}\Delta u_{it}) = 0,\ t = 2,\ldots,T,\ s = 1,\ldots,T$; $\text{E}(x_{2,is}\Delta u_{it}) = 0,\ t = 2,\ldots,T,\ s = 1,\ldots,T$. In total, there are $T(T-1)/2 + 2T(T-1)$ moment conditions.

Table 5 presents the bias, RMSE, and empirical size results, where $\phi_{01} = 0.5$, $\phi_{02} = 0.3$, and the error term u_{it} is simulated from a normal distribution. In addition to the individual parameters ϕ_1 and ϕ_2, their sum $\phi_1 + \phi_2$ may also be of interest, which is needed for calculating the cumulative partial effects of past shocks. So in Table 5 statistics related to $\phi_1 + \phi_2$ are included as well. Results related to $\boldsymbol{\beta}$ are relegated to Supplementary Appendix H.

One can observe that the GMM estimator is almost unbiased when T is small. However, it becomes more biased when T goes up. This is not unexpected as the performance of GMM is related to the number of moment conditions used. Meanwhile, the bias of II is stable across T and is virtually zero for all (N, T) combinations considered here and in terms of RMSE, II is the best across the majority of all configurations and GMM is almost as good as II. In terms of the empirical size for testing ϕ_1 and $\phi_1 + \phi_2$, GMM performs well in small T but becomes more and more over-sized as T grows, while II delivers very good size performance in all cases. Interestingly, for testing the second-order autoregressive parameter ϕ_2, GMM performs reasonably well and is less sensitive to T. The robust BCS- and RML-based approaches produce relatively larger bias and higher RMSE in many cases and can give empirical rejection rates very off the nominal size.

Supplementary Appendix H contains results for DP(3). Available upon request are additional results for DP(2) and DP(3) under other error distributions and under different specifications when the cumulative partial effects of past shocks are formulated to be zero (e.g., $\phi_{01} = -0.1$ and $\phi_{02} = 0.1$). The conclusions one can draw are very similar to those from Table 5.

3.5. DP-τ

This subsection provides simulation results for dynamic panel models in convergence studies. Now the DGP considered is (24) with $\phi = 1 + \tau\rho$, $\alpha_i \sim$ i.i.d.N(0, 1), and $u_{it} \sim$ i.i.d.N(0, 1). The primary parameter of interest in the so-called convergence parameter $\rho = (\phi - 1)/\tau$. Table 6 presents the bias, RMSE, and empirical rejection rates (pertaining to ρ) from the GMM, OLS1, and II methods under $\tau = 5$ and $\phi_0 = 0.8, 0.5, 0.2, -0.2$. The combinations of N and T satisfy $N = \lfloor 600/(T-1) \rfloor$.[23] The WG estimator is included as well since there are cases of relatively large T.

When $T > 10$, the GMM estimator in Table 6 represents the τth-order difference GMM of Caselli et al. (1996). In terms of bias and RMSE, GMM performs, in general, the worst among the four estimators, especially when ϕ_0 is large (see, for example, cases where $T = 11$ and $\phi_0 = 0.8, 0.5$).[24] Equivalently, it means that when a negative converge parameter ρ is close to 0, GMM tends to produce very misleading results. The WG estimator does not perform too bad in terms of its bias and RMSE, which is not unexpected as the T considered here is not too small. In contrast, the II estimator proposed in this paper is the best in terms of bias and RMSE among the four estimators under all cases considered here. Compared with II, OLS1 is nearly as good in terms of bias, but it fares worse in terms of RMSE. Now turning to the size performance, one can see that GMM gets worse as T increases, which echoes the finding in the DP(1) experiments. WG also displays substantial upward size distortions. Both OLS1 and II are slightly over-sized, but II-based empirical rejection rate is typically closer to the nominal size.

When $T = 7$, corresponding to the case of $T = \tau + 2$, as analyzed before, the WG, OLS1, and II estimators are all consistent and yield numerically the same result. The GMM estimator in Table 6 is based on the first differenced equation with y_{i1} and y_{i2} as instruments. Under the weighting matrix adopted in Arellano and Bond (1991), it is the same as the

[23] Keep in mind that in this subsection, the effective sample size over time is T and it is assumed that pre-time observations are not available. Such a choice of (N, T) is used to make data at $t = 1, \tau+1, 2\tau+1, \ldots, \lfloor T/\tau \rfloor \tau + 1$ available (when $T > 10$) and one can use the τth-order difference GMM estimator of Caselli et al. (1996).

[24] When $T = 11$ and $\tau = 5$, the τth-order difference GMM of Caselli et al. (1996) uses, for each cross-sectional unit, observations from $t = 1, 6, 11$ only and there is only one orthogonality moment condition, see equation (4) in Bao and Dhongde (2009). In contrast, the OLS1 estimator uses observations from $t = 6, \ldots, 11$ and the II estimator uses observations from $t = 1, \ldots, 11$. Thus the performance of the τth-order difference GMM estimator can be really bad under this $N-T$-combination. This was also documented in the Monte Carlo study of Bao and Dhongde (2009).

Table 5
Bias, RMSE, and rejection rates (%) of 5% t tests from various estimator in DP(2).

		(N, T)	(200,3)	(100,6)	(40,15)	(200,3)	(100,6)	(40,15)	(200,3)	(100,6)	(40,15)
u_{it}			ϕ_1			ϕ_2			$\phi_1+\phi_2$		
i.i.d.N(0, 1)	Bias	GMM	−0.59	−1.20	−2.05	−0.09	−0.31	−0.52	−0.67	−1.50	−2.57
		II	0.07	0.04	−0.01	0.07	−0.01	−0.02	0.14	0.03	−0.04
		BCS	0.06	0.02	−0.02	0.07	−0.02	−0.02	0.13	−0.00	−0.05
		RML	−1.32	−0.78	−0.16	−1.16	−1.32	−0.74	−2.49	−2.10	−0.90
		RMLc	−1.60	−0.75	0.06	−1.22	−1.32	−0.70	−2.82	−2.07	−0.64
		IIr	0.14	0.05	−0.01	0.06	−0.01	−0.02	0.20	0.04	−0.04
		BCSr	0.14	0.10	0.00	0.09	0.02	−0.05	0.23	0.12	−0.04
		RMLr	−11.75	−6.71	−2.04	−3.96	−2.33	−0.64	−15.71	−9.04	−2.68
		RMLcr	−1.88	−0.81	0.02	−1.28	−1.33	−0.72	−3.16	−2.14	−0.69
	RMSE	GMM	3.86	3.08	3.12	3.94	2.82	2.35	5.10	3.16	3.02
		II	3.70	2.63	2.35	2.90	2.43	2.27	5.14	2.84	1.59
		BCS	3.69	2.63	2.35	2.90	2.43	2.27	5.12	2.84	1.59
		RML	3.96	2.54	2.43	3.03	2.67	2.35	5.10	3.02	1.76
		RMLc	3.22	2.48	2.34	2.89	2.66	2.34	4.66	2.96	1.49
		IIr	4.06	2.65	2.35	3.05	2.44	2.27	5.72	2.89	1.59
		BCSr	4.08	2.69	2.43	3.07	2.47	2.33	5.75	2.94	1.66
		RMLr	16.48	7.25	3.16	8.17	3.41	2.43	21.80	9.52	3.13
		RMLcr	3.40	2.53	2.40	2.94	2.69	2.40	4.90	3.03	1.55
	Size (%)	GMM	5.76	7.55	14.33	5.39	5.25	5.50	5.66	9.06	39.13
		II	5.11	5.47	5.26	5.53	5.31	4.99	5.63	5.75	5.64
		BCS	9.24	9.13	7.01	5.51	5.06	5.89	8.13	10.41	10.15
		RML	10.85	7.28	7.43	8.18	9.50	7.68	13.87	16.73	11.42
		RMLc	13.08	8.63	6.96	9.20	10.27	8.01	17.56	25.41	14.09
		IIr	5.06	5.39	5.23	5.55	5.27	4.97	5.24	5.81	5.54
		BCSr	3.37	9.71	7.83	3.48	5.51	6.78	2.47	10.15	11.34
		RMLr	68.34	58.40	13.88	38.89	15.14	7.78	68.89	72.82	28.70
		RMLcr	0.47	5.93	6.82	1.31	8.65	8.38	0.47	12.82	23.79
i.d.N(0, t)	Bias	GMM	−1.93	−4.82	−7.66	−0.46	−1.63	−4.25	−2.39	−6.45	−11.91
		II	17.92	8.73	2.57	8.65	5.48	2.53	26.57	14.21	5.10
		BCS	18.77	8.98	2.70	8.87	5.59	2.60	27.64	14.56	5.30
		RML	0.34	−0.69	−2.28	−0.52	−1.01	−2.13	−0.18	−1.70	−4.41
		RMLc	−0.41	−0.01	3.25	−1.00	−1.06	−2.87	−1.42	−1.07	0.37
		IIr	1.25	0.22	−0.02	0.55	0.03	−0.14	1.79	0.25	−0.17
		BCSr	1.28	0.20	−0.30	0.47	−0.04	−0.50	1.75	0.16	−0.80
		RMLr	25.13	−10.34	−6.03	−5.03	−5.68	−3.30	20.10	−16.02	−9.33
		RMLcr	−2.28	−1.01	−0.67	−1.41	−1.17	−1.09	−3.69	−2.19	−1.76
	RMSE	GMM	7.33	7.17	8.76	6.67	5.12	6.01	10.15	8.97	12.59
		II	21.95	11.62	5.66	11.21	7.61	5.33	32.20	17.72	8.01
		BCS	23.43	12.09	5.89	11.55	7.78	5.43	34.05	18.40	8.44
		RML	6.59	4.49	83.65	4.73	4.28	78.18	8.45	5.22	9.65
		RMLc	7.58	6.69	18.22	6.05	6.00	14.75	7.23	4.69	5.08
		IIr	9.89	5.20	4.39	6.36	4.51	4.32	14.68	6.73	4.40
		BCSr	12.32	7.20	4.21	7.00	4.46	4.10	17.67	8.62	4.19
		RMLr	997.56	42.98	7.67	679.57	42.35	5.67	684.61	60.20	10.47
		RMLcr	6.58	4.34	4.08	5.21	4.20	4.11	9.70	5.28	3.92
	Size (%)	GMM	6.29	16.72	50.29	6.12	7.24	20.33	6.60	19.94	90.64
		II	0.92	3.07	4.10	4.94	7.72	5.27	1.46	4.35	2.90
		BCS	94.40	91.83	89.03	86.07	85.41	87.55	94.06	95.13	94.17
		RML	2.09	3.07	9.10	3.79	4.70	7.59	1.81	2.00	11.79
		RMLc	5.90	6.87	9.28	7.10	7.81	9.73	5.98	9.73	14.48
		IIr	4.42	5.30	5.13	4.32	5.14	4.98	4.43	4.89	5.72
		BCSr	3.11	15.61	10.76	2.04	5.40	7.74	2.76	14.46	21.14
		RMLr	9.95	30.67	17.43	12.32	30.41	10.51	9.26	48.01	26.09
		RMLcr	2.20	4.89	8.44	1.95	5.44	9.14	2.03	6.40	22.46

Notes: See (37) for simulation design; $\phi_{01} = 0.5$, $\phi_{02} = 0.3$, and $\beta_{01} = \beta_{02} = 1$. Bias and RMSE are multiplied by 100.

2SLS estimator and it generates exactly the same result as OLS1.[25] Their rejection rates are all close to the nominal size and II's size performance is slightly worse than WG, given that II tries to bias correct the consistent WG estimator. When $T = 8$, WG is still the same as II, but OLS1 and GMM (based on the differenced equation with instruments y_{i1} and y_{i2} for $t = \tau + 2$ and instruments y_{i1}, y_{i2}, and y_{i3} for $t = \tau + 3$) are different.[26] As in the previous case, since the bias correction in II is redundant, the associated t test may perform worse than that from WG.

Supplementary Appendix H gives simulation results under a different set-up, where $T - 1$ is fixed at 30, N is either 30 or 60, the test horizon τ takes on four values (5, 10, 15, 20), and the convergence parameter ρ takes on five values (−0.005, −0.01, −0.03, −0.05, −0.08), largely motivated by the empirical results from the next section. Overall, the II method performs the best with virtually zero bias, the lowest RMSE, and generally the best empirical rejection rate over the four test horizons under different degrees of convergence. The GMM performs the worst with the highest RMSE and substantial size distortions. The OLS1 estimator performs somewhere in between. Additional results under various error

25 This is because in the first stage, the regression of $y_{i2} - y_{i1}$ on y_{i1} and y_{i2} creates a perfect fit and thus the second-stage regression is the same as OLS1.

26 Though it is not so evident in Table 6 due to numerical rounding.

Table 6
Finite-sample performance of WG, GMM, OLS1, and II in DP-τ.

ϕ_0		(N, T)	(100,7)	(85,8)	(60,11)	(40,16)	(30,21)	(24,26)	(20,31)
0.8	Bias	WG	−0.1	−0.0	−0.3	−0.5	−0.6	−0.6	−0.6
		GMM	−0.1	−0.0	−26.8	−11.1	−10.7	−8.9	−8.1
		OLS1	−0.1	−0.0	−0.0	−0.0	−0.0	−0.0	−0.0
		II	−0.1	−0.0	−0.0	−0.0	−0.0	−0.0	−0.0
	RMSE	WG	1.3	0.9	0.8	0.8	0.8	0.8	0.8
		GMM	1.3	1.0	1718.7	19.9	13.9	10.9	9.6
		OLS1	1.3	1.0	0.8	0.7	0.7	0.7	0.6
		II	1.3	0.9	0.7	0.7	0.6	0.6	0.6
	Size (%)	WG	6.6	5.5	6.4	13.7	15.7	16.5	15.7
		GMM	7.0	6.3	7.6	19.5	30.1	33.3	40.8
		OLS1	7.0	6.8	5.5	6.3	5.8	6.8	7.5
		II	6.8	6.5	5.1	5.9	5.9	6.1	5.6
0.5	Bias	WG	0.0	0.0	−0.6	−0.9	−1.0	−0.9	−0.8
		GMM	0.0	−0.0	5.6	−2.1	−3.2	−3.6	−3.7
		OLS1	0.0	−0.0	−0.1	0.0	−0.0	−0.1	−0.0
		II	0.0	0.0	−0.1	0.0	−0.0	−0.1	−0.0
	RMSE	WG	1.8	1.3	1.2	1.3	1.3	1.2	1.2
		GMM	1.8	1.4	133.2	8.5	6.4	5.7	5.1
		OLS1	1.8	1.4	1.2	1.0	1.0	0.9	1.0
		II	1.8	1.3	1.1	0.9	0.8	0.8	0.8
	Size (%)	WG	5.1	5.0	9.5	18.2	19.2	19.5	18.4
		GMM	5.9	4.4	6.3	9.2	13.4	20.9	19.4
		OLS1	5.9	5.1	5.8	5.7	6.1	7.4	9.2
		II	5.9	5.5	6.4	5.4	4.9	6.0	7.7
0.2	Bias	WG	0.0	0.0	−0.6	−1.1	−1.1	−0.9	−0.8
		GMM	0.0	0.0	5.2	−0.6	−1.5	−1.5	−1.7
		OLS1	0.0	0.0	−0.0	−0.0	−0.1	−0.0	−0.0
		II	0.0	0.0	0.0	−0.0	−0.1	−0.0	−0.0
	RMSE	WG	2.0	1.5	1.3	1.5	1.4	1.3	1.2
		GMM	2.0	1.6	88.0	5.1	4.1	3.7	3.3
		OLS1	2.0	1.7	1.4	1.2	1.1	1.1	1.1
		II	2.0	1.5	1.1	1.0	0.9	1.0	0.9
	Size (%)	WG	5.6	3.9	8.3	22.1	19.7	16.6	14.7
		GMM	5.6	5.2	5.3	7.6	10.7	11.5	13.6
		OLS1	5.6	6.3	7.2	7.9	6.0	8.0	7.6
		II	5.5	4.5	5.2	7.2	5.8	6.1	6.6
−0.2	Bias	WG	0.1	−0.0	−0.7	−0.9	−0.8	−0.6	−0.6
		GMM	0.1	−0.0	0.4	−0.2	−0.4	−0.6	−0.7
		OLS1	0.1	−0.0	−0.0	0.1	−0.0	0.1	−0.0
		II	0.1	−0.0	−0.0	0.0	−0.0	0.1	−0.0
	RMSE	WG	2.0	1.5	1.3	1.4	1.2	1.1	1.0
		GMM	2.0	1.6	4.2	3.1	2.8	2.5	2.4
		OLS1	2.0	1.6	1.4	1.2	1.1	1.1	1.1
		II	2.0	1.5	1.1	1.0	1.0	0.9	0.9
	Size (%)	WG	5.1	4.2	8.5	16.8	14.4	9.7	9.0
		GMM	5.8	5.7	5.5	7.0	9.8	10.0	12.1
		OLS1	5.8	5.3	5.5	6.6	7.1	6.9	8.8
		II	5.6	5.2	5.4	6.4	8.0	7.3	6.2

Notes: $\tau = 5$, $\alpha_i \sim$ i.i.d.N(0, 1), $y_{is} = \alpha_i/(1-\phi_0) + u_{is}/\sqrt{1-\phi_0^2}$, $s = 0, \ldots, -(\tau-1)$, and $u_{it} \sim$ i.i.d.N(0, 1). The bias, RMSE, and size results are pertaining to the convergence parameter $\rho = (\phi - 1)/\tau$. For $T > 10$, GMM is Caselli et al.'s (1996) estimator and for $T = 7, 8$ it is based on the first differenced equation with lagged variables in level as instruments. Bias and RMSE are multiplied by 100.

distributions and time-series heteroskedasticity from the BCS and RML procedures of Alvarez and Arellano (2022) are available upon request. Notably, when T is moderate, it is extremely difficult for one to use the robust BCS and RML and the associated inferences can be very misleading.[27] In contrast, the II estimator and its robust version perform very well.

3.6. The sample Jacobian

Recall that in general one cannot check nonsingularity of $\boldsymbol{G}$, but one can always check numerically the identification condition for a given sample $\boldsymbol{W}$ by examining the determinant of $\boldsymbol{G}_N(\boldsymbol{\theta})$ over a grid of values of $\boldsymbol{\theta}$. For all the simulations in the previous subsections, a singular sample Jacobian has never happened and the II procedure (as well as the robust version) never fails.

[27] This is because the robust BCS and RML need to estimate T variance parameters. It takes substantially longer, ranging from about ten ($T = 7$) to hundreds of times ($T = 31$) longer than the other estimates, to get the robust BCS and RML estimates. For example, when $(N, T) = (20, 31)$ and $\tau = 5$, on a PC with Intel i9-9900k, it takes on average 0.018, 9.481, 10.174, and 9.531 seconds, respectively, for robust II, BCS, RML, and concentrated RML to estimate one simulated sample when $u_{it} \sim$ i.i.d.N(0, 1). Thus the number of simulations used for DP-τ is kept at 1000 in this subsection. There are also occasions where the robust BCS and RML fail when T is not small.

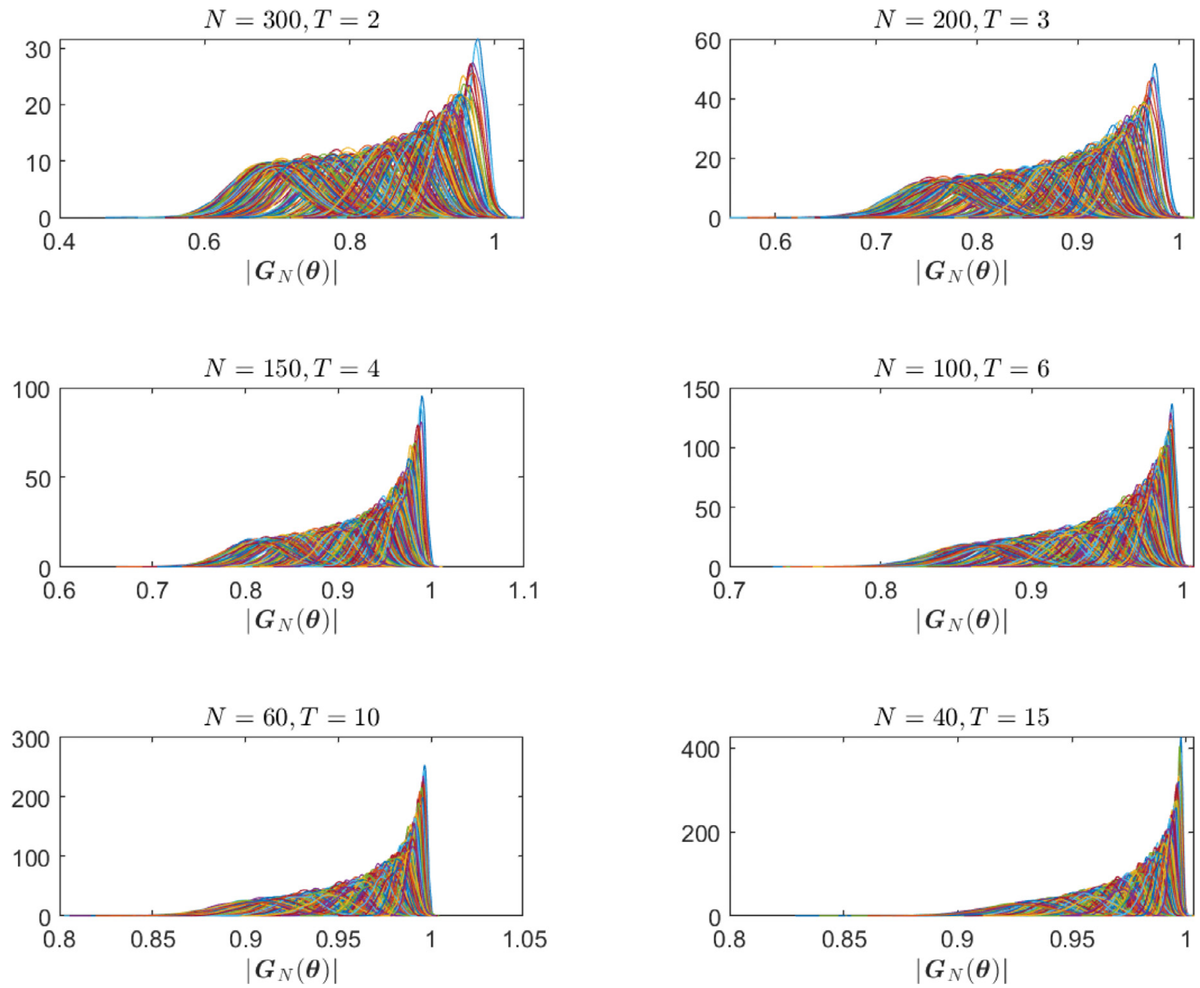

Fig. 2. Distribution of $|\boldsymbol{G}_N(\boldsymbol{\theta})|$ in DP(2).

Fig. 2 plots the distribution of $|\boldsymbol{G}_N(\boldsymbol{\theta})|$, where $\boldsymbol{G}_N(\boldsymbol{\theta})$ is calculated from (10), in DP(2) simulated as in (37) with $u_{it} \sim$ i.i.d.N(0, 1), $\phi_{01} = -0.9, -0.8, \ldots, 0.9$, and $\phi_{02} = -0.9, -0.8, \ldots, 0.9$.[28] Each line plots the kernel density from 1000 simulations of $|\boldsymbol{G}_N(\boldsymbol{\theta})|$ out of a particular combination of ϕ_{01} and ϕ_{02}.[29] So in Fig. 2, each panel contains 361 ($= 19 \times 19$) lines of densities across different combinations of ϕ_{01} and ϕ_{02}. It is obvious that for all combinations of ϕ_{01} and ϕ_{02}, it never happens that $\boldsymbol{G}_N(\boldsymbol{\theta})$ becomes singular. Thus, one can straightforwardly implement the II procedure. Supplementary Appendix G also contains density plots of $|\boldsymbol{G}_N(\boldsymbol{\theta})|$, where $\boldsymbol{G}_N(\boldsymbol{\theta})$ is calculated from (32), in DP(2) under heteroskedasticity, as well as those for DP(3). Again, for all combinations of the autoregressive parameters, $\boldsymbol{G}_N(\boldsymbol{\theta})$ is nonsingular.

4. An empirical study

The Standardized World Income Inequality Database (SWIID) by Solt (2009, 2016, 2020) is used in this section to study whether inequality, measured by the Gini index in deviation from period mean, converged during 1985–2005 ($T = 31$) over different horizons for 63 developed and developing countries. As described in Solt (2020), the SWIID maximizes the comparability of available income inequality data for the broadest possible sample of countries and years. Recent empirical examples that utilize this data set include Berman et al. (2017) and Heathcote et al. (2017). The newest version of this data set and corresponding descriptions can be downloaded from the website https://fsolt.org/swiid/. Out of the 63 countries in

[28] When a particular combination of ϕ_{01} and ϕ_{02} creates a dynamically nonstable process, the initial values are created as $y_{is} = \alpha_i + u_{is} + \boldsymbol{x}_{is}'\boldsymbol{\beta}_0$, $s = 0, -1$ for DP(2). The same strategy is used for DP(3) in Supplementary Appendix G.

[29] The default options in Matlab's `ksdensity` are used to generate the plots.

Table 7
Estimated convergence parameters.

		GMM	OLS1	II	IIr
All	$\tau = 5$	−0.021	−0.160	−0.035	−0.035
		(1.034)	(13.861)	(3.247)	(3.249)
	$\tau = 10$	−0.002	−0.088	−0.059	−0.059
		(0.064)	(15.377)	(5.435)	(5.399)
	$\tau = 15$	0.019	−0.068	−0.060	−0.059
		(0.616)	(18.145)	(7.075)	(7.039)
	$\tau = 20$		−0.049	−0.046	−0.046
			(14.624)	(8.480)	(8.480)
Developed	$\tau = 5$	0.019	−0.193	−0.079	−0.080
		(0.542)	(16.304)	(4.729)	(4.651)
	$\tau = 10$	0.043	−0.098	−0.087	−0.088
		(0.788)	(13.536)	(7.592)	(7.531)
	$\tau = 15$	0.069	−0.072	−0.066	−0.066
		(0.846)	(26.507)	(10.430)	(10.403)
	$\tau = 20$		−0.050	−0.050	−0.050
			(13.252)	(7.372)	(7.372)
Developing	$\tau = 5$	−0.053	−0.120	−0.012	−0.012
		(2.215)	(7.936)	(0.810)	(0.858)
	$\tau = 10$	−0.052	−0.073	−0.038	−0.039
		(1.871)	(8.385)	(2.639)	(2.635)
	$\tau = 15$	−0.013	−0.059	−0.052	−0.052
		(0.413)	(6.749)	(3.291)	(3.268)
	$\tau = 20$		−0.047	−0.044	−0.044
			(7.427)	(4.667)	(4.667)

Notes: t-ratio (in absolute value) inside parentheses. GMM is Caselli et al.'s (1996) estimator.

the sample, 33 are developed countries and the other 30 are developing countries.[30] The list of countries in this study and summary statistics of Gini indices are provided in Supplementary Appendix I. It can be observed that the Gini index is on average higher in developing countries compared to developed countries, in line with the well-documented fact of higher inequality in developing countries. The standard deviation (across all the countries) of the Gini index goes down over time. It is also the case for the group of developed countries, but not so obvious for the group of developing countries.[31]

Table 7 presents the estimated convergence parameter, namely, $\hat{\rho} = (\hat{\phi} - 1)/\tau$, as well as the associated t-ratio (in absolute value, inside parentheses) for each group of countries with different convergence test horizons.[32] Reading Table 7, one can make the following observations. First, the GMM estimator of Caselli et al. (1996) delivers somewhat unexpected results. For the whole sample, it gives insignificant estimates and for the group of developed countries, it gives positive (though insignificant) estimated values of ρ.[33] When the convergence parameter is positive, it implies an explosive dynamic process and thus divergent inequality measure, which seems to be very implausible. Recall that with $T = 31$, when $\tau = 15$, for example, the GMM approach of Caselli et al. (1996) runs a regression $y_{i,31} - y_{i,16} = \phi(y_{i,16} - y_{i,1}) + u_{i,31} - u_{i,16}$ with instrument $Z_i = y_{i,1}$, which uses sample observations at $t = 1, 16, 31$ only. So the unexpected results from the τth-order difference GMM estimator of Caselli et al. (1996) are not really surprising. Second, the II, together with its robust version, and OLS1 estimates appear to provide more plausible results, indicating strong evidence of convergence in inequality among developed and developing countries and all the countries as a group over the longer horizons ($\tau = 10, 15, 20$). They also indicate that inequality converges faster in developed counties than developing countries. Yet, when $\tau = 5$, for developing countries, while the OLS1 estimate is still significant (and also of a very high magnitude), the II and IIr estimates are in fact very insignificant. The estimated convergence parameters from II and IIr, ranging from −0.012 to −0.052 for developing countries and −0.050 to −0.088 for developed countries, are in fact consistent with estimates based on cross-country studies in the literature, see Chambers and Dhongde (2016) and references therein. The estimated $\hat{\rho}$ values from OLS1 when $\tau = 5$ appear to be too high. Given the better and more reliable performance of II and IIr relative to OLS1 in the Monte Carol experiments, it is more reasonable for one to conclude that there is little evidence of convergence in inequality among developing countries over a 5-year horizon.

5. Concluding remarks and discussions

This paper proposes an estimation strategy for the classical higher-order dynamic panels. Instead of seeking instruments based on lagged or differenced dependent variable, it is built on the simple WG estimator and solves numerically a sample binding function that links model parameters and the sample data. The resulting estimator is shown to be

30 Based on the most recent World Bank country classification, developed countries are those with a GNI per capita of \$12, 536 or more, while developing countries are upper-middle, lower-middle and low income countries with a GNI per capita less than \$12, 536. For details, see https://datahelpdesk.worldbank.org/knowledgebase/articles/906519.

31 The reduction of cross-country standard deviation over time signals the so-called sigma-convergence. Note that beta-convergence is necessary but not sufficient for sigma-convergence (e.g., Furceri, 2005).

32 Given the poor performance of the BCS and RML estimators when T is moderate, see the previous section, they are not included for consideration.

33 The White-type variance estimators are used for GMM and OLS1 to construct the t-ratios.

consistent, asymptotically normal, robust to the error distribution, and does not rely on some restrictive assumptions on the initial observations. For the special case of DP(1), the differences in the literature regarding the inference results are explained. Its good finite-sample properties, in terms of bias, RMSE, and empirical size for hypothesis testing, are demonstrated by Monte Carlo simulations. This new estimator is used to study empirically whether inequality converges among 63 countries during the period 1985–2015. It gives strong evidence of convergence over the longer 10-year, 15-year, and 20-year horizons. However, there is much weaker evidence of convergence among developing countries over a shorter 5-year horizon.

Note that how the initial conditions are generated does not directly matter for the II estimation procedure, as long as Assumption 5.iii is maintained, which is really innocuous. This is because $\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})$ is a function of $\boldsymbol{\phi}_0$ (or is the same as the expectation of $\hat{\mathrm{E}}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})$ in the case of time-series heteroskedasticity), but does not depend on the initial conditions. Thus given the data $\boldsymbol{W}$, one can solve for the II estimator. Of course, they do have impact on the asymptotic variance of the II estimator. Simulation results available upon request show that the performance of the II estimator is quite robust across different initial values. It should also be emphasized that the II approach in this paper does not restrict the panel to be dynamically stable.

Lee (2012) considered the case when DP(q) is fitted to DP(p) when $q \leq p$. This gives rise to omitted variable bias on top of the fixed-effects bias. The misspecification bias does not disappear even as $T \to \infty$. Thus Lee (2012) suggested either using penalized likelihood approach or conducting order selection before bias-correction. The essential idea behind the first approach is that omitted lags introduce serial correlation in the error term and also in the score function, so an autocorrelation-consistent estimator of the variance is needed. Such an approach, as one can expect, requires that $T \to \infty$. With serial correlation in the error term due to this kind of misspecification, $\mathrm{E}(\boldsymbol{u}_i'\boldsymbol{M}\boldsymbol{\Phi}_p^{-1}\boldsymbol{L}^t\boldsymbol{u}_i) = \mathrm{tr}(\boldsymbol{M}\boldsymbol{\Phi}_p^{-1}\boldsymbol{L}^t\boldsymbol{\Sigma}_i)$, where the $T \times T$ matrix $\boldsymbol{\Sigma}_i$ is not diagonal. How to arrive at some auxiliary statistic that is purely a function of data and model parameters (but not the nuisance $\boldsymbol{\Sigma}_i$) such that it has the same expectation as $\boldsymbol{u}_i'\boldsymbol{M}\boldsymbol{\Phi}_p^{-1}\boldsymbol{L}^t\boldsymbol{u}_i$ and is robust to any non-diagonal $\boldsymbol{\Sigma}_i$ (so that one can follow the II strategy as in Section 2.6) is an open (and challenging) question. The iterated GMM and related inference as discussed in Hansen and Lee (2021) may be worthwhile to explore. One can also follow the second approach of Lee (2012) by using a model selection criterion (Lee and Phillips, 2015) and applying the II procedure to the model picked up by the model selection criterion.[34] The model selection criterion in Lee and Phillips (2015) is consistent only if $T \to \infty$ and it is left for future research to study the asymptotic distribution of the II estimator under large T.

Acknowledgements

The authors would like to thank Serena Ng (managing editor), an associate editor, and two anonymous referees for their constructive feedback. The first author is grateful to Xiaohong Chen, Peter Phillips, Yundong Tu, Aman Ullah, Jihai Yu, Jun Yu, conference participants in Workshop on Advanced Quantitative Methods in Social Sciences (Macau, 2017), the 3rd Dongbei Econometrics Workshop (Dalian, 2017), the 2nd South Lake Innovation Forum (Wuhan, 2017), the 16th International Symposium on Econometric Theory and Applications (2022), and seminar participants at Peking University for their helpful comments.

Appendix A

The proofs of the distribution of the recentered estimator under time-series heteroskedasticity and Theorem 2 are very similar to the case of the classical model and they are omitted.

Proof of distribution of recentered estimator

Supplementary Appendix C shows that $\mathrm{Var}(T^{-\eta}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u}) = O(N)$ and $\mathrm{E}(T^{-(\eta-1)}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u}) = -\sigma^2 T^{-\eta} N\boldsymbol{r} = O(N)$. Further, $T^{-\eta}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W} - \mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})) = O_P(\sqrt{N})$ and $T^{-\eta}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W}) = O(N)$. By a Nagar-type (Nagar, 1959) expansion, one has

$$\begin{aligned}
&(T^{-\eta}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}\\
&= \{\boldsymbol{I} + [T^{-\eta}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}[T^{-\eta}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W} - T^{-\eta}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]\}^{-1}[T^{-\eta}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}\\
&= [T^{-\eta}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}\\
&- [T^{-\eta}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}[T^{-\eta}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W} - T^{-\eta}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})][T^{-\eta}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1} + \cdots\\
&= [T^{-\eta}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1} + O_P(N^{-3/2}).
\end{aligned} \tag{38}$$

Accordingly,

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \boldsymbol{\delta}) = \sqrt{N}(T^{-\eta}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}T^{-\eta}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} - \sqrt{N}[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})$$

[34] As shown in Lee's (2012) simulations, the model-selection based bias-correction procedure works better than the penalized likelihood function approach.

$$
\begin{aligned}
&= \sqrt{N}\{[\mathrm{E}(T^{-\eta}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1} + O_P(N^{-3/2})\}T^{-\eta}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} \\
&\quad - \sqrt{N}[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u}) + o_P(1) \\
&= \sqrt{N}[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}[\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} - \mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})] \\
&\quad + \sqrt{N}O_P(N^{-3/2})T^{-\eta}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} + o_P(1) \\
&= \sqrt{N}[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}[\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} - \mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})] + O_P(1). \qquad (39)
\end{aligned}
$$

One sees that $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \boldsymbol{\delta})$ and $\sqrt{N}[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}[\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} - \mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})]$ are not asymptotically equivalent, due to the $O_P(1)$ term in (39), namely, in view of (38), the term introduced by $\sqrt{N}\{-[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}[\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W} - \mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})][\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}\}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u}$. With the $O_P(1)$ term

$$
\begin{aligned}
&- \sqrt{N}[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}[\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W} - \mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})][\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} \\
&= -\sqrt{N}[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}[\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W} - \mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})][\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u}) + o_P(1) \\
&= -\sqrt{N}[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W}[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u}) + \sqrt{N}[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u}) + o_P(1) \\
&= \sqrt{N}[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u}) - \boldsymbol{W}'\boldsymbol{A}\boldsymbol{W}\boldsymbol{\delta}] + o_P(1)
\end{aligned}
$$

explicitly *included*, (39) becomes

$$
\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \boldsymbol{\delta}) = \sqrt{N}[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} - \boldsymbol{W}'\boldsymbol{A}\boldsymbol{W}\boldsymbol{\delta}) + o_P(1). \qquad (40)
$$

For any $\boldsymbol{\iota} \in \mathbb{R}^m$, $\{\boldsymbol{\iota}'(\boldsymbol{W}_i\boldsymbol{M}\boldsymbol{u}_i - \boldsymbol{W}_i'\boldsymbol{M}\boldsymbol{W}_i\boldsymbol{\delta})\}_{i=1}^N$ is a sequence of independent random variables under Assumptions 1 to 5 and Lyapunov's central limit theorem can be invoked to show that

$$
\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \boldsymbol{\delta}) \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Delta}), \qquad (41)
$$

where

$$
\boldsymbol{\Delta} = \lim_{N\to\infty} N[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}\mathrm{Var}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} - \boldsymbol{W}'\boldsymbol{A}\boldsymbol{W}\boldsymbol{\delta})[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}. \qquad (42)
$$

When the $O_P(1)$ is *excluded* from (39),

$$
\sqrt{N}[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}[\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} - \mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})] \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Omega}), \qquad (43)
$$

where

$$
\boldsymbol{\Omega} = \lim_{N\to\infty} N[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}\mathrm{Var}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}. \qquad (44)
$$

Supplementary Appendices B and C discuss how to evaluate $\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})$ and $\mathrm{Var}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})$, respectively. The $O_P(1)$ term appears to be ignored in BC when they presented the asymptotic distribution of $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \boldsymbol{\delta})$ for DP(1).[35]

If $\boldsymbol{d}_N = (\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})$ is used as the recentering term, then

$$
\begin{aligned}
\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \boldsymbol{d}_N) &= \sqrt{N}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}[\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} - \mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})] \\
&= \sqrt{N}[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}[\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} - \mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})] + o_P(1) \\
&\xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Omega}), \qquad (45)
\end{aligned}
$$

and thus

$$
\begin{aligned}
\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \boldsymbol{d}_N^{\dagger}) &= \sqrt{N}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} + \boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}\boldsymbol{h}), \\
&= \sqrt{N}[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} + \boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}\boldsymbol{h}) + o_P(1) \\
&\xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \boldsymbol{V}), \qquad (46)
\end{aligned}
$$

where

$$
\boldsymbol{V} = \lim_{N\to\infty} [N^{-1}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}N^{-1}\mathrm{Var}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} + \boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}\boldsymbol{h})[N^{-1}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}, \qquad (47)
$$

in which $\mathrm{Var}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} + \boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}\boldsymbol{h})$ is given by (D.3) in Supplementary Appendix D.

---

35 See their Eqs. (25) and (26), where the "$\boldsymbol{V_X}$" corresponds to the $\boldsymbol{\Omega}$ here. Their $\boldsymbol{V}_X = \sigma^2\boldsymbol{\Sigma}_{\boldsymbol{WAW}}^{-1} + \sigma^4\mathrm{tr}(\boldsymbol{MCMC})\boldsymbol{\Sigma}_{\boldsymbol{WAW}}^{-1}\boldsymbol{e}_{k+1,1}\boldsymbol{e}_{k+1,1}'\boldsymbol{\Sigma}_{\boldsymbol{WAW}}^{-1}$, where $\boldsymbol{\Sigma}_{\boldsymbol{WAW}} = \mathrm{plim}_{N\to\infty}N^{-1}\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W}$ and $\boldsymbol{C} = (\boldsymbol{I} - \phi_0\boldsymbol{L})^{-1}\boldsymbol{L}$, was based on Bun and Kiviet (2001) and Kiviet (1995). The $\boldsymbol{V_X}$ was derived under the assumption of normality and the condition on the initial latent variable $v_{i0} = y_{i0} - (1-\phi_0)^{-1}\alpha_i$ such that $v_{i0} = \mathrm{E}(v_{i0})$ (see the line proceeding equation (32) of Kiviet (1995)). Under these conditions, $\mathrm{Var}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})$ can be shown to be $\sigma^2[\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})] + \sigma^4N\mathrm{tr}(\boldsymbol{MCMC})\boldsymbol{e}_{k+1,1}\boldsymbol{e}_{k+1,1}'$ and thus $\boldsymbol{\Omega}$ can be simplified to $\boldsymbol{V}_X$, see Supplementary Appendix E.1 and discussions in Section 2.4.1.

Proof of Theorem 1. By a first-order expansion, $\boldsymbol{b}_N(\check{\boldsymbol{\theta}}) = \boldsymbol{b}_N(\boldsymbol{\theta}_0) + [\partial \boldsymbol{b}_N(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}'](\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_P(N^{-1/2}) = \boldsymbol{b}_N(\boldsymbol{\theta}_0) + \hat{\boldsymbol{G}}_N(\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_P(N^{-1/2})$, where $\hat{\boldsymbol{G}}_N = \boldsymbol{G}_N(\check{\boldsymbol{\theta}})$. It follows that

$$\begin{aligned}\sqrt{N}(\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) &= \hat{\boldsymbol{G}}_N^{-1}\sqrt{N}[\boldsymbol{b}_N(\check{\boldsymbol{\theta}}) - \boldsymbol{b}_N(\boldsymbol{\theta}_0)] + o_P(1)\\ &= \hat{\boldsymbol{G}}_N^{-1}\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \boldsymbol{d}_N^{\dagger}) + o_P(1)\\ &\xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \boldsymbol{V}_{II}),\end{aligned} \tag{48}$$

where $\boldsymbol{V}_{II} = \boldsymbol{G}^{-1}\boldsymbol{V}\boldsymbol{G}^{-1\prime}$ with $\boldsymbol{G} = \operatorname{plim}_{N\to\infty}\hat{\boldsymbol{G}}_N = \operatorname{plim}_{N\to\infty}\boldsymbol{G}_N(\boldsymbol{\theta}_0)$, provided that one can verify a technical condition, namely, $\partial \boldsymbol{b}_N^{-1}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$ is asymptotically locally equicontinuous at $\boldsymbol{\theta}_0$, see Phillips (2012). For this, the following condition is sufficient: for $\varepsilon > 0$, there exists a sequence $s_N \to \infty$ such that $s_N/\sqrt{N} \to 0$,

$$\sup_{\|s_N(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|<\varepsilon} \|\boldsymbol{G}_N(\boldsymbol{\theta}_0)[\boldsymbol{G}_N(\boldsymbol{\theta})^{-1} - \boldsymbol{G}_N(\boldsymbol{\theta}_0)^{-1}]\| \xrightarrow{a.s.} 0.$$

It is sufficient to consider $\|\boldsymbol{G}_N(\boldsymbol{\theta}_0)[\boldsymbol{G}_N(\boldsymbol{\theta})^{-1} - \boldsymbol{G}_N(\boldsymbol{\theta}_0)^{-1}]\|$ when the norm is sub-multiplicative (say, $\|\cdot\|_2$). Then

$$\begin{aligned}\|\boldsymbol{G}_N(\boldsymbol{\theta}_0)[\boldsymbol{G}_N(\boldsymbol{\theta})^{-1} - \boldsymbol{G}_N(\boldsymbol{\theta}_0)^{-1}]\| &= \|\boldsymbol{G}_N(\boldsymbol{\theta}_0)\boldsymbol{G}_N(\boldsymbol{\theta})^{-1}[\boldsymbol{G}_N(\boldsymbol{\theta}_0) - \boldsymbol{G}_N(\boldsymbol{\theta})]\boldsymbol{G}_N(\boldsymbol{\theta}_0)^{-1}\|\\ &\le \|\boldsymbol{G}_N(\boldsymbol{\theta}_0)\| \cdot \|\boldsymbol{G}_N(\boldsymbol{\theta})^{-1}\| \cdot \|\boldsymbol{G}_N(\boldsymbol{\theta}_0) - \boldsymbol{G}_N(\boldsymbol{\theta})\| \cdot \|\boldsymbol{G}_N(\boldsymbol{\theta}_0)^{-1}\|.\end{aligned}$$

Note that

$$\|\boldsymbol{G}_N(\boldsymbol{\theta}_0) - \boldsymbol{G}_N(\boldsymbol{\theta})\| \le \left[\sup_{\tilde{\boldsymbol{\theta}}} \left\|\boldsymbol{J}_N(\tilde{\boldsymbol{\theta}})\right\|\right] \|(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \otimes \boldsymbol{I}_m\|,$$

where $\tilde{\boldsymbol{\theta}}$ lies between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$ and $\boldsymbol{J}_N(\boldsymbol{\theta}) = \partial \boldsymbol{G}_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$ denote the $m \times m^2$matrix of derivative of $\boldsymbol{G}_N(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}'$, consisting of horizontally stacked m blocks of $m \times m$ matrices $\partial \boldsymbol{G}_N(\boldsymbol{\theta})/\partial\theta_j$, $j = 1, \ldots, m$. Rewrite $\boldsymbol{G}_N(\boldsymbol{\theta})$ as

$$\begin{aligned}\boldsymbol{G}_N(\boldsymbol{\theta}) = \boldsymbol{I}_m &+ \frac{2}{T(T-1)}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}\begin{pmatrix}\boldsymbol{r}_p(\boldsymbol{\phi})\\ \boldsymbol{0}_k\end{pmatrix}(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\theta})'\boldsymbol{A}\boldsymbol{W}\\ &- \frac{(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\theta})'\boldsymbol{A}(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\theta})}{T(T-1)}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}\begin{pmatrix}\boldsymbol{R}(\boldsymbol{\phi}) & \boldsymbol{0}_{p\times k}\\ \boldsymbol{0}_{k\times p} & \boldsymbol{0}_k\end{pmatrix}.\end{aligned}$$

Then

$$\begin{aligned}\frac{\partial \boldsymbol{G}_N(\boldsymbol{\theta})}{\partial\theta_j} &= \frac{2}{T(T-1)}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}\left[\begin{pmatrix}\frac{\partial \boldsymbol{r}_p(\boldsymbol{\phi})}{\partial\theta_j}\\ \boldsymbol{0}_k\end{pmatrix}(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\theta})' - \begin{pmatrix}\boldsymbol{r}_p(\boldsymbol{\phi})\\ \boldsymbol{0}_k\end{pmatrix}(\boldsymbol{W}\boldsymbol{e}_{m,j})'\right]\boldsymbol{A}\boldsymbol{W}\\ &+ \frac{2(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\theta})'\boldsymbol{A}(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{e}_{m,j})}{T(T-1)}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}\begin{pmatrix}\boldsymbol{R}(\boldsymbol{\phi}) & \boldsymbol{0}_{p\times k}\\ \boldsymbol{0}_{k\times p} & \boldsymbol{0}_k\end{pmatrix}\\ &- \frac{(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\theta})'\boldsymbol{A}(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\theta})}{T(T-1)}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}\begin{pmatrix}\frac{\partial \boldsymbol{R}(\boldsymbol{\phi})}{\partial\theta_j} & \boldsymbol{0}_{p\times k}\\ \boldsymbol{0}_{k\times p} & \boldsymbol{0}_k\end{pmatrix},\end{aligned}$$

where $\partial \boldsymbol{r}_p(\boldsymbol{\phi})/\partial\theta_j$ is the same as the jth column of $\boldsymbol{R}(\boldsymbol{\phi})$ for $j = 1, \ldots, p$ and equal to $\boldsymbol{0}_p$ for $j = p+1, \ldots, m$, $\partial \boldsymbol{R}(\boldsymbol{\phi})/\partial\theta_j$ is a $p \times p$ matrix consisting of $\boldsymbol{1}'\boldsymbol{\Phi}_p^{-1}(\boldsymbol{\phi})\boldsymbol{L}^j\boldsymbol{\Phi}_p^{-1}(\boldsymbol{\phi})\boldsymbol{L}^{j_2}\boldsymbol{\Phi}_p^{-1}(\boldsymbol{\phi})\boldsymbol{L}^{j_1}\boldsymbol{1} + \boldsymbol{1}'\boldsymbol{\Phi}_p^{-1}(\boldsymbol{\phi})\boldsymbol{L}^{j_2}\boldsymbol{\Phi}_p^{-1}(\boldsymbol{\phi})\boldsymbol{L}^j\boldsymbol{\Phi}_p^{-1}(\boldsymbol{\phi})\boldsymbol{L}^{j_1}\boldsymbol{1}$ in its (j_1, j_2) position, $j_1, j_2 = 1, \ldots, p$, for $j = 1, \ldots, p$ and equal to $\boldsymbol{0}_p$ for $j = p+1, \ldots, m$. Given Assumptions 1 to 5, one can show that at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and for $\boldsymbol{\theta}$ such that $\|s_N(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\| < \varepsilon$, all the elements of $\boldsymbol{G}_N(\boldsymbol{\theta})$ and $\boldsymbol{J}_N(\boldsymbol{\theta})$ are bounded almost surely. Thus

$$\begin{aligned}\sup_{\|s_N(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|<\varepsilon}\|\boldsymbol{G}_N(\boldsymbol{\theta}_0) - \boldsymbol{G}_N(\boldsymbol{\theta})\| &\le \sup_{\|s_N(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|<\varepsilon}\left[\sup_{\tilde{\boldsymbol{\theta}}}\left\|\boldsymbol{J}_N(\tilde{\boldsymbol{\theta}})\right\|\right]\|(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\otimes\boldsymbol{I}_m\|\\ &\le \left[\sup_{\tilde{\boldsymbol{\theta}}}\left\|\boldsymbol{J}_N(\tilde{\boldsymbol{\theta}})\right\|\right]\left|\frac{\varepsilon}{s_N}\right|\\ &\xrightarrow{a.s.} 0,\end{aligned}$$

which is sufficient to show $\sup_{\|s_N(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|<\varepsilon}\|\boldsymbol{G}_N(\boldsymbol{\theta}_0)[\boldsymbol{G}_N(\boldsymbol{\theta})^{-1} - \boldsymbol{G}_N(\boldsymbol{\theta}_0)^{-1}]\| \xrightarrow{a.s.} 0$.

Proof of consistency of $\hat{\boldsymbol{V}}_{II}$

By writing $\boldsymbol{M}(\boldsymbol{y}_i - \boldsymbol{W}_i\check{\boldsymbol{\theta}}) = \boldsymbol{M}[\boldsymbol{u}_i + \boldsymbol{W}_i(\boldsymbol{\theta}_0 - \check{\boldsymbol{\theta}})]$, one has, for $\hat{\boldsymbol{v}}_i$ defined in (12),

$$\hat{\boldsymbol{v}}_i = \boldsymbol{W}_i'\boldsymbol{M}\boldsymbol{u}_i + \boldsymbol{W}_i'\boldsymbol{M}\boldsymbol{W}_i(\boldsymbol{\theta}_0 - \check{\boldsymbol{\theta}})$$

$$+ [\boldsymbol{u}_i'\boldsymbol{M}\boldsymbol{u}_i + (\boldsymbol{\theta}_0 - \check{\boldsymbol{\theta}})'\boldsymbol{W}_i'\boldsymbol{M}\boldsymbol{W}_i(\boldsymbol{\theta}_0 - \check{\boldsymbol{\theta}}) + 2(\boldsymbol{\theta}_0 - \check{\boldsymbol{\theta}})'\boldsymbol{W}_i'\boldsymbol{M}\boldsymbol{u}_i]\hat{\boldsymbol{h}}$$
$$\equiv \boldsymbol{v}_i + \hat{\boldsymbol{v}}_{1i} + \hat{\boldsymbol{v}}_{2i} + \hat{\boldsymbol{v}}_{3i},$$

where $v_i = \boldsymbol{W}_i'\boldsymbol{M}\boldsymbol{u}_i + \boldsymbol{u}_i'\boldsymbol{M}\boldsymbol{u}_i\boldsymbol{h}$, $\hat{\boldsymbol{v}}_{1i} = \boldsymbol{W}_i'\boldsymbol{M}\boldsymbol{W}_i(\boldsymbol{\theta}_0 - \check{\boldsymbol{\theta}})$, $\hat{\boldsymbol{v}}_{2i} = [(\boldsymbol{\theta}_0 - \check{\boldsymbol{\theta}})'\boldsymbol{W}_i'\boldsymbol{M}\boldsymbol{W}_i(\boldsymbol{\theta}_0 - \check{\boldsymbol{\theta}}) + 2(\boldsymbol{\theta}_0 - \check{\boldsymbol{\theta}})'\boldsymbol{W}_i'\boldsymbol{M}\boldsymbol{u}_i]\hat{\boldsymbol{h}}$, and $\hat{\boldsymbol{v}}_{3i} = \boldsymbol{u}_i'\boldsymbol{M}\boldsymbol{u}_i(\hat{\boldsymbol{h}} - \boldsymbol{h})$. Note that $\boldsymbol{\theta}_0 - \check{\boldsymbol{\theta}} = O_P(N^{-1/2})$, $T^{-\eta}\boldsymbol{W}_i'\boldsymbol{M}\boldsymbol{W}_i = O_P(1)$ (from Supplementary Appendix B), $T^{-(\eta-1)}\boldsymbol{W}_i'\boldsymbol{M}\boldsymbol{u}_i = O_P(1)$ (from Supplementary Appendix C), and $\hat{\boldsymbol{h}} - \boldsymbol{h} = (-(\boldsymbol{\phi}_0 - \check{\boldsymbol{\phi}})'\boldsymbol{R}(\tilde{\boldsymbol{\phi}})', \boldsymbol{0}_k')'/[T(T-1)]$, where $\tilde{\boldsymbol{\phi}}$ lies between $\boldsymbol{\phi}_0$ and $\check{\boldsymbol{\phi}}$. Recall that $\boldsymbol{R}(\boldsymbol{\phi}) = \partial \boldsymbol{r}_p(\boldsymbol{\phi})/\partial\boldsymbol{\phi}'$ is a $p \times p$ matrix with $\boldsymbol{1}'\boldsymbol{\Phi}_p^{-1}(\boldsymbol{\phi})\boldsymbol{L}^{j_2}\boldsymbol{\Phi}_p^{-1}(\boldsymbol{\phi})\boldsymbol{L}^{j_1}\boldsymbol{1}$ in its (j_1, j_2) position, $j_1, j_2 = 1, \ldots, p$, and it can be verified that $T^{-\eta}\boldsymbol{R}(\tilde{\boldsymbol{\phi}}) = O(1)$. Further, $\mathrm{E}(\boldsymbol{u}_i'\boldsymbol{M}\boldsymbol{u}_i) = \sigma^2(T-1)$ and $\mathrm{Var}(\boldsymbol{u}_i'\boldsymbol{M}\boldsymbol{u}_i) = \sigma^4[2(T-1) + \gamma_2(T - 2 + T^{-1})]$. Then $T^{-\eta}\hat{\boldsymbol{v}}_{1i} = O_P(N^{-1/2})$, $T^{-(\eta-3)}\hat{\boldsymbol{v}}_{2i} = O_P(N^{-1/2}) + T^{-1}O_P(N^{-1})$, and $T^{-(\eta-1)}\hat{\boldsymbol{v}}_{3i} = O_P(N^{-1/2})$. Thus it follows that $N^{-1}\sum_{i=1}^N \hat{\boldsymbol{v}}_i\hat{\boldsymbol{v}}_i' = N^{-1}\sum_{i=1}^N \boldsymbol{v}_i\boldsymbol{v}_i' + o_P(1)$, $N^{-1}\sum_{i=1}^N \hat{\boldsymbol{v}}_i = N^{-1}\sum_{i=1}^N \boldsymbol{v}_i + O_P(N^{-1/2})$, $\boldsymbol{v}_i - N^{-1}\sum_{i=1}^N \boldsymbol{v}_i = \hat{\boldsymbol{v}}_i - N^{-1}\sum_{i=1}^N \hat{\boldsymbol{v}}_i + O_P(N^{-1/2})$, and $N^{-1}\sum_{i=1}^N(\hat{\boldsymbol{v}}_i - \bar{\hat{\boldsymbol{v}}})(\hat{\boldsymbol{v}}_i - \bar{\hat{\boldsymbol{v}}})' = N^{-1}\sum_{i=1}^N(\boldsymbol{v}_i - \bar{\boldsymbol{v}})(\boldsymbol{v}_i - \bar{\boldsymbol{v}})' + o_P(1)$, where $\bar{\hat{\boldsymbol{v}}} = N^{-1}\sum_{i=1}^N \hat{\boldsymbol{v}}_i = O_P(N^{-1/2})$ and $\bar{\boldsymbol{v}} = N^{-1}\sum_{i=1}^N \boldsymbol{v}_i$. Then under Assumptions 1 to 5, $N^{-1}\mathrm{Var}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} + \boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}\boldsymbol{h}) = N^{-1}\mathrm{Var}(\sum_{i=1}^N \boldsymbol{v}_i)$, which can be consistently estimated by $N^{-1}\sum_{i=1}^N(\boldsymbol{v}_i - \bar{\boldsymbol{v}})(\hat{\boldsymbol{v}}_i - \bar{\boldsymbol{v}})'$ and also by $N^{-1}\sum_{i=1}^N(\hat{\boldsymbol{v}}_i - \bar{\hat{\boldsymbol{v}}})(\hat{\boldsymbol{v}}_i - \bar{\hat{\boldsymbol{v}}})'$ and $N^{-1}\sum_{i=1}^N \hat{\boldsymbol{v}}_i\hat{\boldsymbol{v}}_i'$. Therefore, one can claim that $\hat{\boldsymbol{V}} = N(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}(\sum_{i=1}^N \hat{\boldsymbol{v}}_i\hat{\boldsymbol{v}}_i')(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})^{-1}$ consistently estimates $\boldsymbol{V} = \lim_{N\to\infty}[N^{-1}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}N^{-1}\mathrm{Var}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} + \boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}\boldsymbol{h})[N^{-1}\mathrm{E}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W})]^{-1}$. With all the elements of $\boldsymbol{J}(\boldsymbol{\theta})$ bounded almost surely for $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$ (see Proof of Theorem 1), $\hat{\boldsymbol{G}} = \boldsymbol{G}_N(\check{\boldsymbol{\theta}})$ is obviously consistent for $\boldsymbol{G}$. Thus it follows that $\hat{\boldsymbol{V}}_{II} = \hat{\boldsymbol{G}}^{-1}\hat{\boldsymbol{V}}\hat{\boldsymbol{G}}^{-1\prime}$ is consistent for $\boldsymbol{V}_{II} = \boldsymbol{G}^{-1}\boldsymbol{V}\boldsymbol{G}^{-1\prime}$.

For the case of $\hat{\boldsymbol{v}}_i$ defined in (35) when there is time-series heteroskedasticity, one can follow a similar procedure to decompose $\hat{\boldsymbol{v}}_i$ as

$$\hat{\boldsymbol{v}}_i = \boldsymbol{W}_i'\boldsymbol{M}\boldsymbol{u}_i - \begin{pmatrix} (\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\theta}_0)'\boldsymbol{M}\boldsymbol{E}_1\boldsymbol{M}(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\theta}_0) \\ \vdots \\ (\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\theta}_0)'\boldsymbol{M}\boldsymbol{E}_p\boldsymbol{M}(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\theta}_0) \\ \boldsymbol{0}_k \end{pmatrix}$$
$$+ \boldsymbol{W}_i'\boldsymbol{M}\boldsymbol{W}_i(\boldsymbol{\theta}_0 - \check{\boldsymbol{\theta}}) - \begin{pmatrix} (\boldsymbol{\theta}_0 - \check{\boldsymbol{\theta}})'\boldsymbol{W}_i'\boldsymbol{M}\boldsymbol{E}_1\boldsymbol{M}\boldsymbol{W}_i(\boldsymbol{\theta}_0 - \check{\boldsymbol{\theta}}) - 2(\boldsymbol{\theta}_0 - \check{\boldsymbol{\theta}})'\boldsymbol{W}_i'\boldsymbol{M}\boldsymbol{E}_1\boldsymbol{M}\boldsymbol{u}_i \\ \vdots \\ (\boldsymbol{\theta}_0 - \check{\boldsymbol{\theta}})'\boldsymbol{W}_i'\boldsymbol{M}\boldsymbol{E}_p\boldsymbol{M}\boldsymbol{W}_i(\boldsymbol{\theta}_0 - \check{\boldsymbol{\theta}}) - 2(\boldsymbol{\theta}_0 - \check{\boldsymbol{\theta}})'\boldsymbol{W}_i'\boldsymbol{M}\boldsymbol{E}_p\boldsymbol{M}\boldsymbol{u}_i \\ \boldsymbol{0}_k \end{pmatrix},$$

where for finite T, each $\boldsymbol{E}_t$, $t = 1, \ldots, p$, is diagonal with $O(1)$ elements. Thus replacing $\boldsymbol{M}\boldsymbol{E}_t\boldsymbol{M}$ everywhere with $\boldsymbol{M}$ does not affect the orders of magnitudes of various terms in the above decomposition. Based on the arguments in the previous paragraph, one can claim that $N^{-1}\sum_{i=1}^N \hat{\boldsymbol{v}}_i\hat{\boldsymbol{v}}_i'$ consistently estimates $N^{-1}\mathrm{Var}[\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u} - \hat{\mathrm{E}}(\boldsymbol{W}'\boldsymbol{A}\boldsymbol{u})]$. Further, one can show that all the elements $\boldsymbol{G}_N(\boldsymbol{\theta})$ as defined in (32) and those of the derivative of $\boldsymbol{G}_N(\boldsymbol{\theta})$ are bounded almost surely for $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$. Then $\hat{\boldsymbol{G}}^{-1}\hat{\boldsymbol{V}}\hat{\boldsymbol{G}}^{-1\prime}$ is consistent for $\boldsymbol{G}^{-1}\boldsymbol{V}\boldsymbol{G}^{-1\prime}$.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2022.09.003.

References

Ahn, S.C., Schmidt, P., 1995. Efficient estimation of models for dynamic panel data. J. Econometrics 68 (1), 5–27. http://dx.doi.org/10.1016/0304-4076(94)01641-C.

Alvarez, J., Arellano, M., 2003. The time series and cross-section asymptotics of dynamic panel data estimators. Econometrica 71 (4), 1121–1159. http://dx.doi.org/10.1111/1468-0262.00441.

Alvarez, J., Arellano, M., 2022. Robust likelihood estimation of dynamic panel data models. J. Econometrics 226 (1), 21–61. http://dx.doi.org/10.1016/j.jeconom.2021.03.005.

Anderson, T.W., Hsiao, C., 1981. Estimation of dynamic models with error components. J. Amer. Statist. Assoc. 76 (375), 598–606. http://dx.doi.org/10.1080/01621459.1981.10477691.

Arellano, M., Bond, S., 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. Rev. Econom. Stud. 58 (2), 277–297. http://dx.doi.org/10.2307/2297968.

Arellano, M., Bover, O., 1995. Another look at the instrumental variable estimation of error-components models. J. Econometrics 68 (1), 29–51. http://dx.doi.org/10.1016/0304-4076(94)01642-D.

Ashley, R.A., Sun, X., 2016. Subset-continuous-updating GMM estimators for dynamic panel data models. Econometrics 4 (4), http://dx.doi.org/10.3390/econometrics4040047.

Bai, J., 2013. Fixed-effects dynamic panel models, a factor analytical method. Econometrica 81 (1), 285–314. http://dx.doi.org/10.3982/ECTA9409.

Baltagi, B.H., 2008. Econometric Analysis of Panel Data, fourth ed. John Wiley & Sons.

Bao, Y., Dhongde, S., 2009. Testing convergence in income distribution. Oxf. Bull. Econ. Stat. 71 (2), 295–302. http://dx.doi.org/10.1111/j.1468-0084.2008.00514.x.

Berman, N., Couttenier, M., Rohner, D., Thoenig, M., 2017. This mine is mine! How minerals fuel conflicts in Africa. Amer. Econ. Rev. 107 (6), 1564–1610. http://dx.doi.org/10.1257/aer.20150774.

Blundell, R., Bond, S., 1998. Initial conditions and moment restrictions in dynamic panel data models. J. Econometrics 87 (1), 115–143. http://dx.doi.org/10.1016/S0304-4076(98)00009-8.
Blundell, R., Bond, S., 2000. GMM estimation with persistent panel data: An application to production functions. Econometric Rev. 19 (3), 321–340. http://dx.doi.org/10.1080/07474930008800475.
Blundell, R., Smith, R.J., 1991. Initial conditions and efficient estimation in dynamic panel data models. Ann. Écon. Stat. 20/21, 109–123.
Bun, M.J.G., 2003. Bias correction in the dynamic panel data model with a nonscalar disturbance covariance matrix. Econometric Rev. 22 (1), 29–58. http://dx.doi.org/10.1081/ETC-120017973.
Bun, M.J.G., Carree, M.A., 2005. Bias-corrected estimation in dynamic panel data models. J. Bus. Econom. Statist. 23 (2), 200–210. http://dx.doi.org/10.1198/073500104000000532.
Bun, M.J.G., Kiviet, J.F., 2001. The Accuracy of Inference in Small Samples of Dynamic Panel Data Models. Discussion Paper 01-006/4, Tinbergen Institute.
Bun, M.J.G., Kiviet, J.F., 2006. The effects of dynamic feedbacks on LS and MM estimator accuracy in panel data models. J. Econometrics 132 (2), 409–444. http://dx.doi.org/10.1016/j.jeconom.2005.02.006.
Bun, M.J.G., Sarafidis, V., 2015. Dynamic panel data models. In: Baltagi, B.H. (Ed.), The Oxford Handbook of Panel Data. Oxford University Press, pp. 76–110. http://dx.doi.org/10.1093/oxfordhb/9780199940042.013.0003.
Caselli, F., Esquivel, G., Lefort, F., 1996. Reopening the convergence debate: A new look at cross-country growth empirics. J. Econ. Growth 1 (3), 363–389. http://dx.doi.org/10.1007/BF00141044.
Chambers, D., Dhongde, S., 2016. Convergence in income distributions: Evidence from a panel of countries. Econ. Model. 59, 262–270. http://dx.doi.org/10.1016/j.econmod.2016.07.019.
Chowdhury, G., 1987. A note on correcting biases in dynamic panel models. Appl. Econ. 19 (1), 31–37. http://dx.doi.org/10.1080/00036848700000055.
De Vos, I., Everaert, G., Ruyssen, I., 2015. Bootstrap-based bias correction and inference for dynamic panels with fixed effects. Stata J. 15 (4), 986–1018. http://dx.doi.org/10.1177/1536867X1501500404.
Everaert, G., Pozzi, L., 2007. Bootstrap-based bias correction for dynamic panels. J. Econom. Dynam. Control 31 (4), 1160–1184. http://dx.doi.org/10.1016/j.jedc.2006.04.006.
Forneron, J.-J., 2020. A sieve-SMM estimator for dynamic models. http://dx.doi.org/10.48550/ARXIV.1902.01456.
Forneron, J.-J., Ng, S., 2018. The ABC of simulation estimation with auxiliary statistics. J. Econometrics 205 (1), 112–139. http://dx.doi.org/10.1016/j.jeconom.2018.03.007.
Furceri, D., 2005. β and σ-convergence: A mathematical relation of causality. Econom. Lett. 89 (2), 212–215. http://dx.doi.org/10.1016/j.econlet.2005.05.026.
Gospodinov, N., Komunjer, I., Ng, S., 2017. Simulated minimum distance estimation of dynamic models with errors-in-variables. J. Econometrics 200 (2), 181–193. http://dx.doi.org/10.1016/j.jeconom.2017.06.004.
Gouriéroux, C., Monfort, A., Renault, E., 1993. Indirect inference. J. Appl. Econometrics 8 (S1), S85–S118. http://dx.doi.org/10.1002/jae.3950080507.
Gouriéroux, C., Phillips, P.C.B., Yu, J., 2010. Indirect inference for dynamic panel models. J. Econometrics 157 (1), 68–77. http://dx.doi.org/10.1016/j.jeconom.2009.10.024.
Hahn, J., 1997. Efficient estimation of panel data models with sequential moment restrictions. J. Econometrics 79 (1), 1–21. http://dx.doi.org/10.1016/S0304-4076(97)00012-2.
Hahn, J., Hausman, J., Kuersteiner, G., 2007. Long difference instrumental variables estimation for dynamic panel models with fixed effects. J. Econometrics 140 (2), 574–617. http://dx.doi.org/10.1016/j.jeconom.2006.07.005.
Hahn, J., Kuersteiner, G., 2002. Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and T are large. Econometrica 70 (4), 1639–1657. http://dx.doi.org/10.1111/1468-0262.00344.
Han, C., Phillips, P.C.B., 2010. GMM estimation for dynamic panels with fixed effects and strong instruments at unity. Econom. Theory 26 (1), 119–151. http://dx.doi.org/10.1017/S026646660909063X.
Hansen, B.E., Lee, S., 2021. Inference for iterated GMM under misspecification. Econometrica 89 (3), 1419–1447. http://dx.doi.org/10.3982/ECTA16274.
Heathcote, J., Storesletten, K., Violante, G.L., 2017. Optimal tax progressivity: An analytical framework. Q. J. Econ. 132 (4), 1693–1754. http://dx.doi.org/10.1093/qje/qjx018.
Holtz-Eakin, D., Newey, W., Rosen, H., 1988. Estimating vector autoregressions with panel data. Econometrica 56 (6), 1371–1395. http://dx.doi.org/10.2307/1913103.
Hsiao, C., Zhou, Q., 2017. First difference or forward demeaning: Implications for the method of moments estimators. Econometric Rev. 36 (6–9), 883–897. http://dx.doi.org/10.1080/07474938.2017.1307594.
Judson, R.A., Owen, A.L., 1999. Estimating dynamic panel data models: A guide for macroeconomists. Econom. Lett. 65 (1), 9–15. http://dx.doi.org/10.1016/S0165-1765(99)00130-5.
Kiviet, J.F., 1995. On bias, inconsistency, and efficiency of various estimators in dynamic panel data models. J. Econometrics 68 (1), 53–78. http://dx.doi.org/10.1016/0304-4076(94)01643-E.
Lancaster, T., 2002. Orthogonal parameters and panel data. Rev. Econom. Stud. 69 (3), 647–666. http://dx.doi.org/10.1111/1467-937X.t01-1-00025.
Lee, Y., 2012. Bias in dynamic panel models under time series misspecification. J. Econometrics 169 (1), 54–60. http://dx.doi.org/10.1016/j.jeconom.2012.01.009.
Lee, Y., Phillips, P.C., 2015. Model selection in the presence of incidental parameters. J. Econometrics 188 (2), 474–489. http://dx.doi.org/10.1016/j.jeconom.2015.03.012.
Nagar, A.L., 1959. The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. Econometrica 27 (4), 575–595. http://dx.doi.org/10.2307/1909352.
Nickell, S., 1981. Biases in dynamic models with fixed effects. Econometrica 49 (6), 1417–1426. http://dx.doi.org/10.2307/1911408.
Phillips, P.C.B., 2012. Folklore theorems, implicit maps, and indirect inference. Econometrica 80 (1), 425–454. http://dx.doi.org/10.3982/ECTA9350.
Smith, Jr., A.A., 1993. Estimating nonlinear time-series models using simulated vector autoregressions. J. Appl. Econometrics 8 (S1), S63–S84. http://dx.doi.org/10.1002/jae.3950080506.
Solt, F., 2009. Standardizing the world income inequality database. Soc. Sci. Q. 90 (2), 231–242. http://dx.doi.org/10.1111/j.1540-6237.2009.00614.x.
Solt, F., 2016. The standardized world income inequality database. Soc. Sci. Q. 97 (5), 1267–1281. http://dx.doi.org/10.1111/ssqu.12295.
Solt, F., 2020. Measuring income inequality across countries and over time: The standardized world income inequality database. Soc. Sci. Q. 101 (3), 1183–1199. http://dx.doi.org/10.1111/ssqu.12795.
Wansbeek, T., Bekker, P., 1996. On IV, GMM and ML in a dynamic panel data model. Econom. Lett. 51 (2), 145–152. http://dx.doi.org/10.1016/0165-1765(95)00785-7.
Ziliak, J.P., 1997. Efficient estimation with panel data when instruments are predetermined: An empirical comparison of moment-condition estimators. J. Bus. Econom. Statist. 15 (4), 419–431. http://dx.doi.org/10.1080/07350015.1997.10524720.