Bias-corrected method of moments estimators for dynamic panel data models

Jörg Breitung* Kazuhiko Hayakawa† Sebastian Kripfganz‡

December 29, 2020

Abstract

In this paper, we propose a computationally simple bias correction for linear dynamic panel data models and study its asymptotic properties when the number of time periods is fixed or tends to infinity with the number of panel units. The approach can accommodate both fixed-effects and random-effects assumptions, heteroskedastic errors, as well as higher-order autoregressive models. Panel-corrected standard errors are proposed that allow for robust inference in dynamic models with cross-sectionally correlated errors. Monte Carlo experiments suggest that the bias-corrected method of moment estimator outperforms popular GMM estimators in terms of efficiency and correctly sized tests.

JEL Classification: C23; C33; C63

Keywords: Bias correction; Moment conditions; Autoregressive model; Panel data; Fixed effects; Random Effects

*University of Cologne, Institute of Econometrics and Statistics, Universitätsstraße 24, 50923 Cologne, Germany. E-mail: breitung@statistik.uni-koeln.de

†Hiroshima University, Department of Economics, 1-2-1 Kagamiyama, Higashi-Hiroshima, Hiroshima, 739-8525, Japan. E-mail: kazuhaya@hiroshima-u.ac.jp

‡University of Exeter Business School, Department of Economics, Streatham Court, Rennes Drive, Exeter, EX4 4PU, United Kingdom. E-mail: S.Kripfganz@exeter.ac.uk

1 Introduction

Dynamic panel data models are now used in a wide area of empirical applications. Since the work of Anderson and Hsiao (1981), instrumental variables and generalized method of moments (GMM) estimators have been extensively applied in the estimation of linear dynamic panel data models. However, it is known that the GMM estimator by Holtz-Eakin et al. (1988) and Arellano and Bond (1991) suffers from the weak-instruments problem when the persistency of the data is strong, as demonstrated by Blundell and Bond (1998). They also showed that the GMM estimator for models in levels with first-differenced instruments mitigates that problem and they proposed the so-called system GMM estimator that combines moment conditions for the models in first differences and in levels. Nowadays, the system GMM estimator is most frequently used in practice, albeit Bun and Windmeijer (2010) showed that it still suffers from the weak-instruments problem when the variance of the individual-specific effects is larger than that of the idiosyncratic errors.

As alternatives to the GMM approach, maximum likelihood (ML) estimators and bias-corrected within-groups (WG) estimators were proposed. Hsiao et al. (2002) suggested a transformed ML estimator that adapts the ML approach to the differenced variables. Hayakawa and Pesaran (2015) extended this transformed ML estimator to allow for cross-sectional heteroskedasticity and proposed robust standard errors. With regard to bias-corrected WG estimators, Kiviet (1995) and Judson and Owen (1999) demonstrate that they are attractive alternatives to GMM estimators. Although the bias-corrected WG estimator of Kiviet (1995) is based on a higher-order expansion of the bias term, the analytical results are based on the unknown parameters that have to be estimated by some consistent initial estimator. Accordingly, the asymptotic distribution of this estimator is unknown. Bun and Carree (2005) proposed an alternative bias-corrected WG estimator which iteratively solves a nonlinear equation with regard to unknown parameters. Dhaene and Jochmans (2016) obtain an adjusted profile likelihood function by integrating a bias-corrected profile score.

In this paper, we demonstrate that a bias-corrected estimator can be obtained as a method of moments estimator. The adjusted profile score is transformed into nonlinear moment conditions that can be easily solved with standard numerical methods. Asymptotic results are readily available. The underlying estimating equations are equivalent to those of the Dhaene and Jochmans (2016) estimator when we adopt a fixed-effects assumption for the exogenous regressors. For the first-order autoregressive model, they are also equivalent to those of Bun and Carree (2005). Furthermore, we show that the ML estimators of Hsiao et al. (2002) and Bai (2013) are based on a modified log-likelihood function that leads to an asymptotically equivalent bias correction of the first-order condition. Yet, they rely on additional assumptions about the initial observations that are not required for our approach.

Within our method of moments framework, we can easily differentiate between regressors that are correlated with the individual-specific effects and those that are uncorrelated with them. All it requires is a slight modification of the respective moment conditions. This allows for the estimation of dynamic fixed-effects and dynamic random-effects models, as well as hybrid versions. Under appropriate orthogonality assumptions, time-invariant regressors can be incorporated as well.

Moreover, the model allows for individual-specific heteroskedasticity in the large-N, fixed-T framework. When both N and T are large, the estimator is also robust to time series heteroskedasticity. Furthermore, we propose cluster-robust/panel-corrected standard errors that account for cross-sectional dependence, and we extend our bias-corrected method of moments approach to higher-order autoregressive models. Monte Carlo experiments suggest that these estimators perform well in terms of efficiency and correctly sized tests, relative to uncorrected WG and GMM approaches.

2 Bias-corrected method of moments estimation

To motivate the bias correction approach, we initially consider the pure first-order autoregressive model

$$y_{it} = \alpha y_{i,t-1} + \mu_i + u_{it}, \quad t = 1, 2, \ldots T, \quad i = 1, 2, \ldots, N,$$

where $u_{it} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. For estimating the parameter α in such an AR(1) model, Dhaene and Jochmans (2016) consider the profile likelihood function $\ell(\alpha)$ after profiling out the nuisance parameters μ_i and σ^2 from the log-likelihood function. This yields the profile score

$$s_{NT}(\alpha) = \frac{\partial \ell(\alpha)}{\partial \alpha} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} (y_{i,t-1} - \overline{y}_{-1,i}) e_{it}(\alpha)}{\sum_{i=1}^{N} \sum_{t=1}^{T} [e_{it}(\alpha) - \overline{e}_i(\alpha)] e_{it}(\alpha)},$$

where $\overline{y}_{-1,i} = T^{-1} \sum_{t=1}^{T} y_{i,t-1}$, $e_{it}(\alpha) = y_{it} - \alpha y_{i,t-1}$, and $\overline{e}_i(\alpha) = T^{-1} \sum_{t=1}^{T} e_{it}(\alpha)$. Their bias-adjusted profile likelihood estimator $\widehat{\alpha}_{al}$ results from correcting the inconsistency in the profile score by solving $s_{NT}(\widehat{\alpha}_{al}) - \operatorname{plim}_{N\to\infty} s_{NT}(\widehat{\alpha}_{al}) = 0$.

Instead, we simplify the bias correction task by initially assuming that the variance σ^2 is known such that the profile score (divided by N and T) becomes

$$\tilde{s}_{NT}(\alpha) = \frac{1}{\sigma^2 NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{i,t-1} - \overline{y}_{-1,i}) e_{it}(\alpha).$$

This allows us to restrict the focus on the numerator in the profile score. With

$$b_T(\alpha) = \mathbb{E}\left[\frac{1}{\sigma^2 T}\sum_{t=1}^{T}(y_{i,t-1} - \overline{y}_{-1,i})e_{it}(\alpha)\right],$$

which is a polynomial in α as presented further below, a bias-corrected method of moments estimator $\widehat{\alpha}_{bc}$ then solves the moment equation

$$m_{NT}(\alpha) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}(y_{i,t-1} - \overline{y}_{-1,i})e_{it}(\widehat{\alpha}_{bc}) - b_T(\widehat{\alpha}_{bc})\widehat{\sigma}_T^2(\widehat{\alpha}_{bc})\right) = 0,$$

where the unknown variance parameter σ^2 was replaced with an estimator $\widehat{\sigma}_T^2(\alpha)$.

Let us now look in detail at the first-order dynamic model with strictly exogenous regressors given by

$$y_{it} = \alpha_0 y_{i,t-1} + \boldsymbol{\beta}_0'\mathbf{x}_{it} + \mu_i + u_{it}, \quad t = 1, 2, \ldots, T, \quad i = 1, 2, \ldots, N,$$

where α_0 and the $K \times 1$ vector $\boldsymbol{\beta}_0$ denote the true values of the parameters of interest. For the $K \times 1$ vector of regressors $\mathbf{x}_{it}$ and the error term u_{it}, the following set of assumptions is imposed that is standard in the literature on dynamic panel data models:

Assumption 1: (i) The errors u_{it} are independent across i and t with $\mathbb{E}[u_{it}] = 0$ and $\mathbb{E}[u_{it}^2] = \sigma_i^2 < C$ for some constant $C < \infty$. (ii) The regressors are strictly exogenous with $\mathbb{E}[\mathbf{x}_{it}u_{is}] = \mathbf{0}$ and $\mathbb{E}||u_{it}u_{is}\mathbf{x}_{it}\mathbf{x}_{is}'|| < \infty$ for all $t, s \in \{1, 2, \ldots, T\}$ and $i \in \{1, 2, \ldots, N\}$. (iii) $\mathbb{E}|u_{it}|^{4+\delta} < \infty$ for all i and t and some $\delta > 0$. (iv) For the initial values, we assume $\mathbb{E}[y_{i0}^2] < \infty$ for all i and $\mathbb{E}[y_{i0}u_{it}] = 0$ for all i and $t \in \{1, 2, \ldots, T\}$.

Note that we do not impose any stationarity restriction on the initial values. The process is allowed to start at any fixed or random level in the finite past. If T is fixed, Assumption 1 does not rule out nonstationary regressors or unstable processes with $|\alpha_0| \geq 1$. However, if T tends to infinity, additional assumptions would be required for the limiting distribution of the estimator. It should also be noted that we allow for individual-specific heteroskedasticity. Robustness to time series heteroskedasticity can also be achieved when $T \to \infty$.

Treating y_{i0} as a fixed constant, the first-order conditions of the (quasi-)ML or least-squares WG estimator $\widehat{\boldsymbol{\theta}}_{ml} = (\widehat{\alpha}_{ml}, \widehat{\boldsymbol{\beta}}_{ml}')'$ are

$$\mathbf{g}_{NT}(\widehat{\boldsymbol{\theta}}_{ml}) = \begin{pmatrix} g_{\alpha,NT}(\widehat{\boldsymbol{\theta}}_{ml}) \\ \mathbf{g}_{\beta,NT}(\widehat{\boldsymbol{\theta}}_{ml}) \end{pmatrix} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\begin{pmatrix} y_{i,t-1} - \overline{y}_{-1,i} \\ \mathbf{x}_{it} - \overline{\mathbf{x}}_i \end{pmatrix} e_{it}(\widehat{\boldsymbol{\theta}}_{ml}) = \mathbf{0},$$

where $\overline{\mathbf{x}}_i = T^{-1}\sum_{t=1}^{T}\mathbf{x}_{it}$, $e_{it}(\boldsymbol{\theta}) = y_{i,t-1} - \alpha y_{i,t-1} - \boldsymbol{\beta}'\mathbf{x}_{it}$, and $\overline{e}_i(\boldsymbol{\theta}) = T^{-1}\sum_{t=1}^{T}e_{it}(\boldsymbol{\theta})$.

Following Nickell (1981) and Moon et al. (2015), we obtain

$$\begin{aligned}\mathbb{E}[g_{\alpha,NT}(\boldsymbol{\theta}_0)] &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(y_{i,t-1}-\overline{y}_{-1,i})e_{it}(\boldsymbol{\theta}_0)\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(-\frac{\sigma_i^2}{T^2}\left[(T-1)+(T-2)\alpha_0+(T-3)\alpha_0^2+\ldots+2\alpha_0^{T-3}+\alpha_0^{T-2}\right]\right) \\ &= b_T(\alpha_0)\frac{1}{N}\sum_{i=1}^{N}\sigma_i^2,\end{aligned}$$

where $b_T(\alpha) = -T^{-2}\sum_{t=0}^{T-2}\sum_{s=0}^{t}\alpha^s$ is a negative-valued monotonously decreasing function that simplifies to

$$b_T(\alpha) = \begin{cases} -\frac{1}{(1-\alpha)T}\left(1-\frac{1-\alpha^T}{T(1-\alpha)}\right) = -\frac{1}{(1-\alpha)T}+O(T^{-2}) & \text{for } |\alpha|<1, \\ -\frac{1}{2}+\frac{1}{2T} & \text{for } \alpha = 1. \end{cases}$$

Since $\operatorname{plim}_{N\to\infty} g_{\alpha,NT}(\boldsymbol{\theta}_0) = b_T(\alpha_0)\sigma^2$, where $\sigma^2 = \lim_{N\to\infty} N^{-1}\sum_{i=1}^{N}\sigma_i^2$, the WG estimator is inconsistent for fixed T. Using

$$\widehat{\sigma}_{Ti}^2(\boldsymbol{\theta}) = \frac{1}{T-1}\sum_{t=1}^{T}[e_{it}(\boldsymbol{\theta})-\overline{e}_i(\boldsymbol{\theta})]e_{it}(\boldsymbol{\theta}) \tag{1}$$

such that $\mathbb{E}[\widehat{\sigma}_{Ti}^2(\boldsymbol{\theta}_0)] = \sigma_i^2$, we obtain the moment conditions

$$\mathbb{E}[\mathbf{m}_{Ti}(\boldsymbol{\theta}_0)] = \mathbb{E}\left[\begin{pmatrix} m_{\alpha,Ti}(\boldsymbol{\theta}_0) \\ \mathbf{m}_{\beta,Ti}(\boldsymbol{\theta}_0)\end{pmatrix}\right] = \mathbf{0}, \tag{2}$$

where

$$\begin{aligned} m_{\alpha,Ti}(\boldsymbol{\theta}) &= \frac{1}{T}\sum_{t=1}^{T}(y_{i,t-1}-\overline{y}_{-1,i})e_{it}(\boldsymbol{\theta}) - b_T(\alpha)\widehat{\sigma}_{Ti}^2(\boldsymbol{\theta}), \\ \mathbf{m}_{\beta,Ti}(\boldsymbol{\theta}) &= \frac{1}{T}\sum_{t=1}^{T}(\mathbf{x}_{it}-\overline{\mathbf{x}}_i)e_{it}(\boldsymbol{\theta}).\end{aligned}$$

The bias-corrected method of moments estimator $\widehat{\boldsymbol{\theta}}_{bc}$ is obtained by solving the sample moment conditions $\mathbf{m}_{NT}(\widehat{\boldsymbol{\theta}}_{bc}) = N^{-1}\sum_{i=1}^{N}\mathbf{m}_{Ti}(\widehat{\boldsymbol{\theta}}_{bc}) = \mathbf{0}$.

Notice that the adjusted profile likelihood estimator of Dhaene and Jochmans (2016) is based on the estimating equations

$$\left(\frac{T}{T-1}\right)\frac{\mathbf{m}_{NT}(\widehat{\boldsymbol{\theta}}_{al})}{\widehat{\sigma}^2_{NT}(\widehat{\boldsymbol{\theta}}_{al})} = \mathbf{0},$$

with $\widehat{\sigma}^2_{NT}(\boldsymbol{\theta}) = N^{-1}\sum_{i=1}^N \widehat{\sigma}^2_{Ti}(\boldsymbol{\theta})$. Consequently, the estimators $\widehat{\boldsymbol{\theta}}_{bc}$ and $\widehat{\boldsymbol{\theta}}_{al}$ are equivalent but $\widehat{\boldsymbol{\theta}}_{bc}$ is simpler to compute. We can furthermore demonstrate that also the iterative estimator proposed by Bun and Carree (2005) is based on equivalent estimating equations.[1] Define

$$s_{11} = \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T (y_{i,t-1} - \overline{y}_{-1,i})y_{i,t-1}, \qquad s_{10} = \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T (y_{i,t-1} - \overline{y}_{-1,i})y_{it},$$

$$\boldsymbol{s}_{x1} = \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T (\mathbf{x}_{it} - \overline{\mathbf{x}}_i)y_{i,t-1}, \qquad \boldsymbol{s}_{x0} = \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T (\mathbf{x}_{it} - \overline{\mathbf{x}}_i)y_{it},$$

$$\boldsymbol{S}_{xx} = \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T (\mathbf{x}_{it} - \overline{\mathbf{x}}_i)\mathbf{x}'_{it}.$$

From $\mathbf{m}_{\beta,NT}(\widehat{\boldsymbol{\theta}}_{bc}) = N^{-1}\sum_{i=1}^N \mathbf{m}_{\beta,Ti}(\widehat{\boldsymbol{\theta}}_{bc}) = \mathbf{0}$, we obtain the closed-form solution

$$\widehat{\boldsymbol{\beta}}_{bc} = \boldsymbol{S}_{xx}^{-1}(\boldsymbol{s}_{x0} - \widehat{\alpha}_{bc}\boldsymbol{s}_{x1}) = \widehat{\boldsymbol{\beta}}_{ml} - (\widehat{\alpha}_{bc} - \widehat{\alpha}_{ml})\boldsymbol{S}_{xx}^{-1}\boldsymbol{s}_{x1}, \tag{3}$$

with the uncorrected WG estimators

$$\widehat{\alpha}_{ml} = \frac{s_{10} - \boldsymbol{s}'_{x0}\boldsymbol{S}_{xx}^{-1}\boldsymbol{s}_{x1}}{s_{11}^2 - \boldsymbol{s}'_{x1}\boldsymbol{S}_{xx}^{-1}\boldsymbol{s}_{x1}}, \qquad \widehat{\boldsymbol{\beta}}_{ml} = \boldsymbol{S}_{xx}^{-1}(\boldsymbol{s}_{x0} - \widehat{\alpha}_{ml}\boldsymbol{s}_{x1}).$$

Using (3), after a few algebraic manipulations of $m_{\alpha,NT}(\widehat{\boldsymbol{\theta}}_{bc}) = N^{-1}\sum_{i=1}^N m_{\alpha,Ti}(\widehat{\boldsymbol{\theta}}_{bc}) = 0$ we obtain

$$\widehat{\alpha}_{bc} = \widehat{\alpha}_{ml} - \frac{b_T(\widehat{\alpha}_{bc})\widehat{\sigma}^2_{NT}(\widehat{\boldsymbol{\theta}}_{bc})}{s_{11} - \boldsymbol{s}'_{x1}\boldsymbol{S}_{xx}^{-1}\boldsymbol{s}_{x1}}. \tag{4}$$

Equations (3) and (4) are identical to those in Bun and Carree (2005, equation 20). Our formulation as a method of moments estimator has the advantage that standard numerical optimization procedures are readily applicable and the derivation of the asymptotic properties is straightforward. Moreover, as we will outline further below, extending the estimator to higher-order autoregressive models or to a dynamic random-effects model is easily done within our framework by adjusting the moment conditions accordingly.

Due to the nonlinearity of the moment function $m_{\alpha,NT}(\boldsymbol{\theta})$, the solution to $\mathbf{m}_{NT}(\widehat{\boldsymbol{\theta}}_{bc}) = \mathbf{0}$ needs to be obtained numerically. This can be done with the recursive Gauss-Newton

[1] In addition to this observation, Dhaene and Jochmans (2016) highlight further equivalence results.

algorithm, solving $\widehat{\boldsymbol{\theta}}_{bc} = \arg\min_{\boldsymbol{\theta}} \mathbf{m}_{NT}(\boldsymbol{\theta})'\mathbf{m}_{NT}(\boldsymbol{\theta})$. Starting with an initial guess $\boldsymbol{\theta}^{(0)}$, $\widehat{\boldsymbol{\theta}}_{bc}$ results upon convergence using the generic iteration step

$$\boldsymbol{\theta}^{(s+1)} = \boldsymbol{\theta}^{(s)} - \left(\nabla \mathbf{m}_{NT}(\boldsymbol{\theta}^{(s)})\right)^{-1} \mathbf{m}_{NT}(\boldsymbol{\theta}^{(s)}),$$

with the gradient

$$\boldsymbol{\nabla}\mathbf{m}_{NT}(\boldsymbol{\theta}^{(s)}) = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\nabla}\mathbf{m}_{Ti}(\boldsymbol{\theta}^{(s)}) = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \mathbf{m}_{Ti}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(s)}}.$$

The detailed entries of $\boldsymbol{\nabla}\mathbf{m}_{Ti}(\boldsymbol{\theta})$ are provided in Appendix A.1. The numerical optimization can be simplified by profiling out $\widehat{\boldsymbol{\beta}}_{bc} = \widehat{\boldsymbol{\beta}}(\widehat{\alpha}_{bc})$ using the closed-form solution (3), thus solving $\widehat{\alpha}_{bc} = \arg\min_{\alpha} \tilde{m}_{NT}(\alpha)^2$, where $\tilde{m}_{NT}(\alpha) = m_{\alpha,NT}\left((\alpha, \widehat{\boldsymbol{\beta}}(\alpha))'\right)$.

Dhaene and Jochmans (2016) point out that integrating the adjusted profile score leads to an adjusted likelihood function that is asymptotically unbounded from above. Moreover, while $\operatorname{plim}_{N\to\infty} \mathbf{m}_{NT}(\boldsymbol{\theta}_0) = \mathbf{0}$, the adjusted profile score has multiple zeros. However, they provide evidence that the adjusted likelihood function asymptotically has at most one local maximum for the most relevant situation of $\alpha_0 \geq -1$.[2] For practical purposes, this requires to verify that the numerical algorithm indeed converged to a local maximum, which requires $\nabla \tilde{m}_{NT}(\widehat{\alpha}_{bc}) < 0$. If this condition is violated, we suggest to re-initialize the search algorithm with a different initial guess $\alpha^{(0)}$. If necessary, the process should be repeated until a solution is found that satisfies the negativity condition for the gradient. In our simulations, this approach proves successful. In finite samples, we cannot exclude the possibility that there does not exist a unique local maximizer in the interior of $\Theta = \{\boldsymbol{\theta} : \alpha \in [-1, 1], \boldsymbol{\beta} \in \mathbb{R}^K\}$. In order to protect against a situation of multiple solutions, a grid search may be carried out over a reasonable range of $\alpha \in [-1, 1]$. To single out a point estimate, Dhaene and Jochmans (2016) suggest to choose the solution closest to $\widehat{\alpha}_{ml}$.

Assuming that a unique maximizer has been established, the limiting distribution of the bias-corrected method of moments estimator $\widehat{\boldsymbol{\theta}}_{bc}$ can be obtained in a straightforward way.

Theorem 1: (i) Under Assumption 1, the limiting distribution of $\widehat{\boldsymbol{\theta}}_{bc}$ for fixed T and as $N \to \infty$ is given by

$$\sqrt{N}(\widehat{\boldsymbol{\theta}}_{bc} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, [\boldsymbol{\Sigma}_T + \sigma^2 \mathbf{B}_T(\alpha_0)]^{-1} \mathbf{S}_T(\boldsymbol{\theta}_0) [\boldsymbol{\Sigma}_T + \sigma^2 \mathbf{B}_T(\alpha_0)]^{-1}\right),$$

[2] As discussed in detail by Dhaene and Jochmans (2016), the adjusted likelihood function may have an inflection point at $\boldsymbol{\theta}_0$ if $T = 2$ (plus a strong condition on the initial observations) or if $\alpha_0 = 1$.

with $\mathbf{S}_T(\boldsymbol{\theta}) = \text{plim}_{N\to\infty} N^{-1}\sum_{i=1}^N \mathbf{m}_{Ti}(\boldsymbol{\theta})\mathbf{m}_{Ti}(\boldsymbol{\theta})'$,

$$\boldsymbol{\Sigma}_T = \underset{N\to\infty}{\text{plim}} \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \begin{pmatrix} y_{i,t-1} - \overline{y}_{-1,i} \\ \mathbf{x}_{it} - \overline{\mathbf{x}}_i \end{pmatrix} \begin{pmatrix} y_{i,t-1} & \mathbf{x}_{it}' \end{pmatrix},$$

$$\mathbf{B}_T(\alpha) = \begin{pmatrix} \nabla_\alpha b_T(\alpha) - \frac{2T}{T-1} b_T(\alpha)^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

and $\nabla_\alpha b_T(\alpha) = -T^{-2}\sum_{t=1}^{T-2}\sum_{s=1}^t s\alpha^{s-1}$.

(ii) Under Assumption 1 and with $|\alpha_0| < 1$, the limiting distribution of $\widehat{\boldsymbol{\theta}}_{bc}$ as $N, T \to \infty$, $N/T \to \kappa$ for $0 \leq \kappa < \infty$, is given by

$$\sqrt{NT}(\widehat{\boldsymbol{\theta}}_{bc} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \sigma^2\boldsymbol{\Sigma}^{-1}\right),$$

with $\boldsymbol{\Sigma} = \lim_{T\to\infty} \boldsymbol{\Sigma}_T$.

A consistent fixed-T estimate of the covariance matrix of $\widehat{\boldsymbol{\theta}}_{bc}$ can be obtained with the finite-sample analog

$$\mathbf{V}_{NT}(\widehat{\boldsymbol{\theta}}_{bc}) = \frac{1}{N}\left(\nabla \mathbf{m}_{NT}(\widehat{\boldsymbol{\theta}}_{bc})\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^N \mathbf{m}_{Ti}(\widehat{\boldsymbol{\theta}}_{bc})\mathbf{m}_{Ti}(\widehat{\boldsymbol{\theta}}_{bc})'\right) \left(\nabla \mathbf{m}_{NT}(\widehat{\boldsymbol{\theta}}_{bc})'\right)^{-1}. \quad (5)$$

The usual test statistics can be employed for inference. For instance, we may examine the linear hypothesis $H_0 : \mathbf{R}\boldsymbol{\theta} = \mathbf{r}$ with the Wald statistic $(\mathbf{R}\widehat{\boldsymbol{\theta}}_{bc} - \mathbf{r})'\mathbf{V}_{NT}(\widehat{\boldsymbol{\theta}}_{bc})^{-1}(\mathbf{R}\widehat{\boldsymbol{\theta}}_{bc} - \mathbf{r})$.

Remark 1: As shown by Hahn and Kuersteiner (2002) and Bai (2013), the asymptotic variance in Theorem 1(ii) equals the lower variance bound, which is equivalent to the asymptotic variance in the case of no individual-specific effects μ_i. Therefore, the estimator is efficient when $T \to \infty$ at a rate not slower than N. Importantly, the estimator does not involve an asymptotic bias and, in contrast to the WG and GMM estimators, inference is valid for all values of κ. This finding suggests that the estimator is particularly attractive in macro panels, where N and T are of a similar magnitude (cf. Breitung, 2015).

Remark 2: Under Assumption 1, the bias-corrected method of moments estimator is robust against heteroskedasticity across panel units. When we relax the assumption of constant error variances over time within panel units, we notice that the unaccounted bias becomes less severe as T becomes large. Let $\mathbb{E}[u_{it}^2] = \sigma_{it}^2 < C$ for some $C < \infty$.

Then,

$$\mathbb{E}[g_{\alpha,NT}(\boldsymbol{\theta}_0)] = \frac{1}{N}\sum_{i=1}^{N}\left[-\frac{1}{T^2}\left(\sum_{t=1}^{T-1}\sigma_{it}^2 + \alpha_0\sum_{t=1}^{T-2}\sigma_{it}^2 + \ldots + \alpha^{T-3}(\sigma_{i1}^2+\sigma_{i2}^2) + \alpha^{T-2}\sigma_{i1}^2\right)\right]$$
$$= b_T(\alpha_0)\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sigma_{it}^2 - \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{T^2}\sum_{t=0}^{T-2}\left(\sigma_{i,T-t-1}^2 - \frac{1}{T}\sum_{s=1}^{T}\sigma_{is}^2\right)\sum_{s=0}^{t}\alpha_0^s\right]$$
$$= b_T(\alpha_0)\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sigma_{it}^2 - O(T^{-2}),$$

since $\lim_{T\to\infty} T^{-1}\sum_{t=0}^{T-2}(\sigma_{i,T-t-1}^2 - T^{-1}\sum_{s=1}^{T}\sigma_{is}^2) = 0$ and $\sum_{s=0}^{t}\alpha_0^s \leq \max\{1,(1-\alpha_0)^{-1}\}$, provided that $|\alpha_0| < 1$. Because the remainder term vanishes at a faster rate than $b_T(\alpha_0)$, the bias correction remains valid under temporal heteroskedasticity when T is large.

Remark 3: Assumption 1 remains silent about the relationship between the strictly exogenous regressors $\mathbf{x}_{it}$ and the individual-specific effects μ_i. The setup thus corresponds to a dynamic fixed-effects panel data model. An advantage of the method of moments estimator is its adaptability to a random-effects framework by imposing an additional assumption that the individual-specific intercept μ_i is a random variable with $\mathbb{E}[\mu_i] = 0$, $\mathbb{E}[\mu_i^2] = \sigma_\mu^2$, and $\mathbb{E}[\mathbf{x}_{it}\mu_i] = \mathbf{0}$ for all i and t.[3] We can then replace $\mathbf{m}_{\beta,Ti}(\boldsymbol{\theta})$ in the moment conditions (2) by

$$\mathbf{m}_{\beta,Ti}(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_{it}e_{it}(\boldsymbol{\theta}).$$

This yields a more efficient estimator. In practice, the random-effects assumption may be imposed on all or a subset of the regressors in $\mathbf{x}_{it}$. It also allows the regressors to be invariant over time. Dropping the mean adjustment in the definition of $\boldsymbol{\Sigma}_T$, the limiting distribution of $\widehat{\boldsymbol{\theta}}_{bc}$ for fixed T is a straightforward corollary of Theorem 1(i). The random-effects assumption can be tested with a conventional Hausman test by contrasting the bias-corrected random-effects estimates with their fixed-effects analogs.

3 Relationship to maximum likelihood estimation

In this section, we show that the ML estimation procedures proposed by Hsiao et al. (2002) and Bai (2013) can be seen in a similar light as our bias-corrected method of moments estimator. In particular, we demonstrate that their first-order conditions with respect to α can be decomposed into two terms, $g_{\alpha,NT}(\boldsymbol{\theta})$ and a bias-correction term. Accordingly, the main differences between those approaches are the assumptions on the

[3] In general, the mean zero assumption for μ_i requires to include a constant intercept in the regressor vector $\mathbf{x}_{it}$.

initial conditions that result in variations of the bias correction term. An important advantage of our method of moments approach is that we do not need to impose any assumption on the initial observations y_{i0} other than its finite variance and exogeneity with respect to all subsequent errors $u_{i1}, u_{i2}, \dots, u_{iT}$.

To simplify the discussion, we focus on the simple AR(1) process without exogenous regressors and with homoskedastic errors. The (Gaussian) log-likelihood function (conditional on the initial observations) is given by

$$\begin{aligned}
\ell(\alpha, \sigma^2, \boldsymbol{\mu}) &= -\frac{NT}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{t=1}^{T} u_{it}(\alpha, \boldsymbol{\mu})^2, \\
&= -\frac{NT}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{t=1}^{T}[u_{it}(\alpha, \boldsymbol{\mu}) - \overline{u}_i(\alpha, \boldsymbol{\mu})]^2 - \frac{T}{2\sigma^2}\sum_{i=1}^{N}\overline{u}_i(\alpha, \boldsymbol{\mu})^2, \\
&= -\frac{NT}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{t=1}^{T}[e_{it}(\alpha) - \overline{e}_i(\alpha)]^2 - \frac{T}{2\sigma^2}\sum_{i=1}^{N}\overline{u}_i(\alpha, \boldsymbol{\mu})^2, \qquad (6)
\end{aligned}$$

where $\overline{u}_i(\alpha, \boldsymbol{\mu}) = T^{-1}\sum_{t=1}^{T} u_{it}(\alpha, \boldsymbol{\mu})$. Profiling out $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_N)'$ is equivalent to dropping the last term in the log-likelihood function, which then results in the WG estimator that is known to be biased. The first derivative with respect to α (divided by N and T) is given by

$$\frac{1}{NT}\frac{\partial \ell(\alpha, \sigma^2, \boldsymbol{\mu})}{\partial \alpha} = \frac{1}{\sigma^2 NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{i,t-1} - \overline{y}_{-1,i})e_{it}(\alpha) + \frac{1}{\sigma^2 N}\sum_{i=1}^{N}\overline{y}_{-1,i}\overline{u}_i(\alpha, \boldsymbol{\mu}).$$

Since $\mathbb{E}[\overline{y}_{-1,i}\overline{u}_i] = -b_T(\alpha_0)\sigma^2$, replacing the last term with $-b_T(\alpha)$ yields the nonlinear moment conditions from the previous section.

Hsiao et al. (2002) treat the initial observations as a random variable with $\mathbb{E}[\Delta y_{i1}] = b$ and $\mathbb{E}[(\Delta y_{i1} - b)^2] = \omega\sigma^2$, where Δ is the first-difference operator. From $y_{i1} - y_{i0} = u_{i1} - u_{i0} + b$ and $\mathbb{E}[u_{i0}u_{i1}] = 0$, it follows that $\mathbb{E}[u_{i0}^2] = (\omega - 1)\sigma^2$. The log-likelihood function becomes[4]

$$\ell(\alpha, b, \sigma^2, \omega) = -\frac{NT}{2}\ln(2\pi\sigma^2) - \frac{N}{2}\ln(|\mathbf{\Omega}|) - \frac{1}{\sigma^2}\sum_{i=1}^{N}\mathbf{e}_i^{*\prime}\mathbf{D}'\mathbf{\Omega}^{-1}\mathbf{D}\mathbf{e}_i^*,$$

where $\mathbf{e}_i^* = (e_{i0}, e_{i1}, \dots, e_{iT})'$, $\mathbf{D} = (\mathbf{0}, \mathbf{I}_T) - (\mathbf{I}_T, \mathbf{0})$ is a $T \times (T+1)$ transformation matrix, $\mathbf{\Omega} = \mathbf{D}\mathbf{D}' - (2 - \omega)\boldsymbol{\varphi}\boldsymbol{\varphi}'$ is a $T \times T$ matrix with determinant $|\mathbf{\Omega}| = T(\omega - 1) + 1$, and $\boldsymbol{\varphi}$

[4]For notational convenience, we are using the short-hand notation $\mathbf{e}_i^* = \mathbf{e}_i^*(\alpha, b)$ and $\mathbf{\Omega} = \mathbf{\Omega}(\omega)$.

is the first column of the $T \times T$ identity matrix $\mathbf{I}_T$. With

$$\mathbf{\Omega}^{-1} = (\mathbf{D}\mathbf{D}')^{-1} + \frac{(2-\omega)(T+1)}{|\mathbf{\Omega}|}(\mathbf{D}\mathbf{D}')^{-1}\boldsymbol{\varphi}\boldsymbol{\varphi}'(\mathbf{D}\mathbf{D}')^{-1},$$

and $\mathbf{D}'(\mathbf{D}\mathbf{D}')^{-1}\mathbf{D} = \mathbf{I}_{T+1} + (T+1)^{-1}\boldsymbol{\iota}_{T+1}\boldsymbol{\iota}_{T+1}'$, where $\boldsymbol{\iota}_{T+1}$ is a $(T+1)\times 1$ vector of ones, the log-likelihood function can be rewritten as

$$\begin{aligned}
\ell(\alpha, b, \sigma^2, \omega) = &-\frac{NT}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{t=0}^{T}(u_{it} - \overline{u}_i^*)^2 \\
&-\frac{N}{2}\ln(|\mathbf{\Omega}|) - \frac{(2-\omega)(T+1)}{2\sigma^2|\mathbf{\Omega}|}\sum_{i=1}^{N}(u_{i0} - \overline{u}_i^*)^2 \\
= &-\frac{NT}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{t=1}^{T}(e_{it} - \overline{e}_i)^2 \\
&-\frac{N}{2}\ln(|\mathbf{\Omega}|) - \frac{T}{2\sigma^2|\mathbf{\Omega}|}\sum_{i=1}^{N}(u_{i0} - \overline{u}_i)^2, \qquad (7)
\end{aligned}$$

using the relationship $(T+1)(u_{i0} - \overline{u}_i^*) = T(u_{i0} - \overline{u}_i)$ between the within-group average $\overline{u}_i^* = (T+1)^{-1}\sum_{t=0}^{T} u_{it}$ that includes the initial value u_{i0} and $\overline{u}_i = T^{-1}\sum_{t=1}^{T} u_{it}$ that does not.

Hsiao et al. (2002) effectively replace the last term in the log-likelihood function (6) with the adjustment terms in the last line of equation (7). The first derivative with respect to α becomes

$$\frac{1}{NT}\frac{\partial \ell(\alpha, b, \sigma^2, \omega)}{\partial \alpha} = \frac{1}{\sigma^2 NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{i,t-1} - \overline{y}_{-1,i})e_{it} + \frac{1}{\sigma^2|\mathbf{\Omega}|N}\sum_{i=1}^{N}(y_{i0} - \overline{y}_{-1,i})(u_{i0} - \overline{u}_i).$$

In Appendix A.2, we show that $\mathbb{E}[(y_{i0} - \overline{y}_{-1,i})(u_{i0} - \overline{u}_i)] = -b_T(\alpha_0)\sigma^2|\mathbf{\Omega}|$. Thus, the bias corrections of the Hsiao et al. (2002) ML estimator and our method of moments estimator are asymptotically equivalent. However, the ML estimator has the important drawback that it requires estimation of the additional nuisance parameters b and ω, even though the bias does not involve these parameters. Morever, a violation of the initial-observations condition, for instance $\mathbb{E}[\Delta y_{i1}] = b_i \neq b$, would turn this ML estimator inconsistent for fixed T. Not surprisingly given the similarity of the bias correction, Bun et al. (2017) and Juodis (2018) find that this transformed ML approach also exhibits multiple solutions and the estimator may converge to a solution different from the global maximum. This issue can be non-negligible when T is small.

A similar logic applies to the ML framework of Bai (2013). In the AR(1) model without exogenous variables, the individual-specific effects μ_i can be treated as random effects with $\mathbb{E}[\mu_i] = \mu$ and $\mathbb{E}[(\mu_i - \mu)^2] = \zeta$. As an initial condition, Bai (2013) assumes $y_{i0} = 0$ for all i.[5] After profiling out μ, the first derivative of the Gaussian log-likelihood function results as

$$\frac{1}{NT}\frac{\partial \ell(\alpha, \sigma^2, \zeta)}{\partial \alpha} = \frac{1}{\sigma^2 NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{i,t-1} - \overline{y}_{-1,i})e_{it} + \frac{1}{(\zeta T + \sigma^2)N}\sum_{i=1}^{N}\overline{y}_{-1,i}\overline{e}_i.$$

Since $\mathbb{E}[\overline{y}_{-1,i}\overline{e}_i] = -b_T(\alpha_0)(\zeta T + \sigma^2)$, the second term can again be interpreted as an asymptotically equivalent bias correction term for the first-order condition, provided that the initial condition $y_{i0} = 0$ holds.

4 Higher-order dynamics

Similar to Dhaene and Jochmans (2016), we can extend the bias-corrected method of moments estimator to an autoregressive model of order p. To simplify the exposition, strictly exogenous regressors are initially neglected. The respective moment functions $\mathbf{m}_{\beta,Ti}(\boldsymbol{\theta})$ would remain the same as in the AR(1) model. Consider the AR(p) model

$$y_{it} = \alpha_1 y_{i,t-1} + \alpha_2 y_{i,t-2} + \ldots + \alpha_p y_{i,t-p} + \mu_i + u_{it}, \quad t = 1, 2, \ldots, T, \quad i = 1, 2, \ldots, N.$$

For notational convenience, we suppress the subscript 0 in denoting the true coefficient values in this section. Assumption 1 is slightly amended to account for the addition initial values:

Assumption 2: Assumption 1 (i)–(iii) continue to hold. (iv) For all $s \in \{1-p, \ldots, -1, 0\}$, we assume $\mathbb{E}[y_{is}^2] < \infty$ for all i and $\mathbb{E}[y_{is}u_{it}] = 0$ for all i and $t \in \{1, 2, \ldots, T\}$.

It is convenient to write the model in companion form

$$\mathbf{A}_T(\boldsymbol{\alpha})\mathbf{y}_i = \mathbf{C}_T(\boldsymbol{\alpha})\mathbf{y}_i^0 + \mu_i \boldsymbol{\iota}_T + \mathbf{u}_i,$$

[5] As discussed by Bai (2013, Supplementary Appendix), arbitrary initial conditions can be accommodated by rewriting the model in terms of deviations from the initial observation, $y_{it} - y_{i0}$. This resembles taking differences as in the approach by Hsiao et al. (2002).

where $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iT})'$, $\mathbf{y}_i^0 = (y_{i,1-p}, \ldots, y_{i,-1}, y_{i0})'$, $\mathbf{u}_i = (u_{i1}, u_{i2}, \ldots, u_{iT})'$, and

$$\mathbf{A}_T(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ -\alpha_1 & 1 & \ddots & & & & \vdots \\ \vdots & \vdots & \ddots & \ddots & & & \vdots \\ -\alpha_p & -\alpha_{p-1} & & \ddots & \ddots & & \vdots \\ 0 & -\alpha_p & & & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & & & 1 & 0 \\ 0 & \cdots & 0 & -\alpha_p & \cdots & -\alpha_1 & 1 \end{pmatrix}, \quad \mathbf{C}_T(\boldsymbol{\alpha}) = \begin{pmatrix} \alpha_p & \alpha_{p-1} & \cdots & \alpha_1 \\ 0 & \alpha_p & \ddots & \vdots \\ \vdots & \ddots & \ddots & \alpha_{p-1} \\ \vdots & & \ddots & \alpha_p \\ 0 & \cdots & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & \cdots & \cdots & 0 \end{pmatrix},$$

with $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_p)'$. The dimensions of the matrices $\mathbf{A}_T(\boldsymbol{\alpha})$ and $\mathbf{C}_T(\boldsymbol{\alpha})$ are $T \times T$ and $T \times p$, respectively. Accordingly,

$$\mathbf{y}_i = \mathbf{A}_T(\boldsymbol{\alpha})^{-1}\mathbf{C}_T(\boldsymbol{\alpha})\mathbf{y}_i^0 + \mu_i \mathbf{A}_T(\boldsymbol{\alpha})^{-1}\boldsymbol{\iota}_T + \mathbf{A}_T(\boldsymbol{\alpha})^{-1}\mathbf{u}_i.$$

With the $T \times T$ matrix

$$\mathbf{L}_T^{(l)} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{I}_{T-l} & \mathbf{0} \end{pmatrix},$$

we obtain for the $T \times 1$ vector of lags $\mathbf{y}_{-l,i} = \mathbf{L}_T^{(l)}\mathbf{y}_i = (0, \ldots, 0, y_{i1}, y_{i2}, \ldots, y_{i,T-l})'$ that

$$\mathbb{E}\left[\frac{1}{T}\mathbf{y}_{-l,i}'\mathbf{M}_T\mathbf{u}_i\right] = \frac{\sigma_i^2}{T}\,\mathrm{tr}\left(\mathbf{L}_T^{(l)}\mathbf{A}_T(\boldsymbol{\alpha})^{-1}\mathbf{M}_T\right) = b_T^{(l)}(\boldsymbol{\alpha})\sigma_i^2, \quad l \geq 1,$$

where $\mathbf{M}_T = \mathbf{I}_T - T^{-1}\boldsymbol{\iota}_T\boldsymbol{\iota}_T'$ is the $T \times T$ projection matrix that creates deviations from within-group means, and $b_T^{(l)}(\boldsymbol{\alpha}) = -T^{-2}\boldsymbol{\iota}_T'\mathbf{L}_T^{(l)}\mathbf{A}_T(\boldsymbol{\alpha})^{-1}\boldsymbol{\iota}_T$. For $|\sum_{j=1}^p \alpha_j| < 1$, $b_T^{(l)}(\boldsymbol{\alpha}) = -(T\sum_{j=1}^p \alpha_j)^{-1} + O(T^{-2})$. For the AR(1) model, it is easy to see that

$$\mathbf{A}_T(\alpha)^{-1} = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ \alpha & 1 & \ddots & & \vdots \\ \alpha^2 & \alpha & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \alpha^{T-1} & \alpha^{T-2} & \cdots & \alpha & 1 \end{pmatrix}, \quad \mathbf{L}_T^{(1)}\mathbf{A}_T(\alpha)^{-1} = \begin{pmatrix} 0 & 0 & \cdots & \cdots & 0 \\ 1 & 0 & & & \vdots \\ \alpha & 1 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \alpha^{T-2} & \alpha^{T-3} & \cdots & 1 & 0 \end{pmatrix},$$

such that $\boldsymbol{\iota}_T'\mathbf{L}_T^{(1)}\mathbf{A}_T(\boldsymbol{\alpha})^{-1}\boldsymbol{\iota}_T = \sum_{t=0}^{T-2}\sum_{s=0}^{t}\alpha^s$.

Remark 4: An interpretation of the elements of matrix $\mathbf{A}_T(\boldsymbol{\alpha})^{-1}$ can be obtained from the moving-average representation of the AR(p) model. Let

$$y_{it} - \mathbb{E}[y_{it}|\mu_i, \mathbf{x}_{it}, \mathbf{x}_{i,t-1}, \ldots] = \phi_0 u_{it} + \phi_1 u_{i,t-1} + \phi_2 u_{i,t-2} + \ldots,$$

with $\phi_0 = 1$. For $j \geq k$, the element in the j-th row and k-th column of $\mathbf{A}_T(\boldsymbol{\alpha})^{-1}$ equals ϕ_{j-k}. All other elements equal zero.

The moment functions for the bias-corrected estimator in the AR(p) model result as

$$m_{\alpha_l,Ti}(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}(y_{i,t-l} - \overline{y}_{-l,i})e_{it}(\boldsymbol{\theta}) - b_T^{(l)}(\boldsymbol{\alpha})\widehat{\sigma}_{Ti}^2(\boldsymbol{\theta}), \quad l = 1, 2, \ldots, p,$$

with $\widehat{\sigma}_{Ti}^2(\boldsymbol{\theta})$ given in equation (1). Let $\mathbf{m}_{\alpha,Ti}(\boldsymbol{\theta}) = (m_{\alpha_1,Ti}(\boldsymbol{\theta}), m_{\alpha_2,Ti}(\boldsymbol{\theta}), \ldots, m_{\alpha_p,Ti}(\boldsymbol{\theta}))'$ and $\mathbf{b}_T(\boldsymbol{\alpha}) = (b_T^{(1)}(\boldsymbol{\alpha}), b_T^{(2)}(\boldsymbol{\alpha}), \ldots, b_T^{(p)}(\boldsymbol{\alpha}))'$. For the model with strictly exogenous regressors $\mathbf{x}_{it}$, the complete set of moment conditions can be written compactly as

$$\mathbb{E}[\mathbf{m}_{Ti}(\boldsymbol{\theta}_0)] = \mathbb{E}\left[\begin{pmatrix}\mathbf{m}_{\alpha,Ti}(\boldsymbol{\theta}_0)\\ \mathbf{m}_{\beta,Ti}(\boldsymbol{\theta}_0)\end{pmatrix}\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}[\mathbf{z}_{it}(\boldsymbol{\theta}_0) - \overline{\mathbf{z}}_i(\boldsymbol{\theta}_0)]e_{it}(\boldsymbol{\theta}_0)\right] = \mathbf{0},$$

with

$$\mathbf{z}_{it}(\boldsymbol{\theta}) = \mathbf{w}_{it} - \frac{T}{T-1}\begin{pmatrix}\mathbf{b}_T(\boldsymbol{\alpha})\\ \mathbf{0}\end{pmatrix} e_{it}(\boldsymbol{\theta}),$$

$\overline{\mathbf{z}}_i(\boldsymbol{\theta}) = T^{-1}\sum_{t=1}^{T}\mathbf{z}_{it}(\boldsymbol{\theta})$, and $\mathbf{w}_{it} = (y_{i,t-1}, y_{i,t-2}, \ldots, y_{i,t-p}, \mathbf{x}_{it}')'$. The bias-corrected method of moments estimator solves $N^{-1}\sum_{i=1}^{N}\mathbf{m}_{Ti}(\widehat{\boldsymbol{\theta}}_{bc}) = \mathbf{0}$. The computational details are similar to the AR(1) case in Section 2.

5 Cross-sectional dependence

In many macroeconomic applications, it is reasonable to assume that the elements of the error vector $\mathbf{u}_t = (u_{1t}, u_{2t}, \ldots, u_{Nt})'$ are correlated:[6]

Assumption 3: (i) The errors u_{it} are independent across t but dependent across i with $\mathbb{E}[\mathbf{u}_t] = \mathbf{0}$ and $\mathbb{E}[\mathbf{u}_t\mathbf{u}_t'] = \boldsymbol{\Sigma}_{u,t}$ for all $t \in \{1, 2, \ldots, T\}$. The largest eigenvalue of the positive-definite matrices $\boldsymbol{\Sigma}_{u,1}, \boldsymbol{\Sigma}_{u,2}, \ldots, \boldsymbol{\Sigma}_{u,T}$ is bounded as $N \to \infty$. Assumption 1 (ii)–(iv) continue to hold.

Although the bias-corrected estimator remains consistent under cross correlation, the estimator of the covariance matrix in equation (5) is biased as, in general,

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{m}_{Ti}(\widehat{\boldsymbol{\theta}}_{bc})\sum_{j=1}^{N}\mathbf{m}_{Tj}(\widehat{\boldsymbol{\theta}}_{bc})'\right] \neq \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{m}_{Ti}(\widehat{\boldsymbol{\theta}}_{bc})\mathbf{m}_{Ti}(\widehat{\boldsymbol{\theta}}_{bc})'\right].$$

[6] For notational convenience, we focus on the AR(1) model. The results continue to hold for the AR(p) model.

We need to estimate the expression on the left-hand side consistently. We can rewrite the moment functions as

$$\mathbf{m}_{NT}(\boldsymbol{\theta}) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{z}_{it}(\boldsymbol{\theta})[u_{it}(\boldsymbol{\theta}) - \overline{u}_i(\boldsymbol{\theta})] = \frac{1}{NT}\sum_{t=1}^{T}\mathbf{Z}_t(\boldsymbol{\theta})'[\mathbf{u}_t(\boldsymbol{\theta}) - \overline{\mathbf{u}}(\boldsymbol{\theta})],$$

with $\mathbf{Z}_t(\boldsymbol{\theta}) = (\mathbf{z}_{1t}(\boldsymbol{\theta}), \mathbf{z}_{2t}(\boldsymbol{\theta}), \ldots, \mathbf{z}_{Nt}(\boldsymbol{\theta}))'$ and $\overline{\mathbf{u}}(\boldsymbol{\theta}) = (\overline{u}_1(\boldsymbol{\theta}), \overline{u}_2(\boldsymbol{\theta}), \ldots, \overline{u}_N(\boldsymbol{\theta}))'$. Furthermore, let $\mathbf{W}_t = (\mathbf{w}_{1t}, \mathbf{w}_{2t}, \ldots, \mathbf{w}_{Nt})'$. Since $\mathbf{u}_t$ is independent across t, we can obtain the following asymptotic result.

Theorem 2: Under Assumption 3 and with $|\alpha_0| < 1$, the limiting distribution of $\widehat{\boldsymbol{\theta}}_{bc}$ as $N, T \to \infty$, $N/T \to \kappa$ for $0 \leq \kappa < \infty$, is given by

$$\sqrt{NT}(\widehat{\boldsymbol{\theta}}_{bc} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}^{-1}\mathbf{S}(\boldsymbol{\theta}_0)\boldsymbol{\Sigma}^{-1}\right),$$

with $\boldsymbol{\Sigma}$ defined in Thereom 1 and

$$\mathbf{S}(\boldsymbol{\theta}_0) = \operatorname*{plim}_{N,T\to\infty}\frac{1}{NT}\sum_{t=1}^{T}\mathbf{Z}_t(\boldsymbol{\theta}_0)'\boldsymbol{\Sigma}_{u,t}\mathbf{Z}_t(\boldsymbol{\theta}_0) = \operatorname*{plim}_{N,T\to\infty}\frac{1}{NT}\sum_{t=1}^{T}\mathbf{W}_t'\boldsymbol{\Sigma}_{u,t}\mathbf{W}_t.$$

Under weakly cross-sectionally dependent errors and large-T asymptotics, $\mathbf{S}(\boldsymbol{\theta}_0)$ can be consistently estimated by the cluster-robust finite-sample analog

$$\mathbf{S}_{NT}(\widehat{\boldsymbol{\theta}}_{bc}) = \frac{1}{NT}\sum_{t=1}^{T}\mathbf{Z}_t(\widehat{\boldsymbol{\theta}}_{bc})'\left(\mathbf{e}_t(\widehat{\boldsymbol{\theta}}_{bc}) - \overline{\mathbf{e}}(\widehat{\boldsymbol{\theta}}_{bc})\right)\left(\mathbf{e}_t(\widehat{\boldsymbol{\theta}}_{bc}) - \overline{\mathbf{e}}(\widehat{\boldsymbol{\theta}}_{bc})\right)'\mathbf{Z}_t(\widehat{\boldsymbol{\theta}}_{bc}),$$

where $\mathbf{e}_t(\widehat{\boldsymbol{\theta}}_{bc}) - \overline{\mathbf{e}}(\widehat{\boldsymbol{\theta}}_{bc}) = \mathbf{u}_t(\widehat{\boldsymbol{\theta}}_{bc}) - \overline{\mathbf{u}}(\widehat{\boldsymbol{\theta}}_{bc})$ is an $N \times 1$ vector of mean-adjusted regression residuals.

Remark 5: The cluster-robust approach to the estimation of the covariance matrix runs into difficulties if the cross-sectional dependence is due to common factors. Assume that $u_{it} = \lambda_i f_t + \epsilon_{it}$, where f_t and ϵ_{it} are i.i.d. sequences with $\mathbb{E}[f_t^2] = \sigma_f^2$ and $\mathbb{E}[\epsilon_{it}^2] = \sigma_\epsilon^2$. Accordingly, $\boldsymbol{\Sigma}_{u,t} = \sigma_f^2\boldsymbol{\lambda}\boldsymbol{\lambda}' + \sigma_\epsilon^2\mathbf{I}_N$, where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_N)$. Since $(NT)^{-1}\sum_{t=1}^{T}\mathbf{Z}_t(\boldsymbol{\theta})'\boldsymbol{\lambda}\boldsymbol{\lambda}'\mathbf{Z}_t(\boldsymbol{\theta}) = O_p(N)$, in general, the estimator $\widehat{\boldsymbol{\theta}}_{bc}$ is no longer $\sqrt{NT}$-consistent (cf. Breitung, 2015, Sec. 15.4.3).

6 Small-sample properties

To assess the small-sample properties of the bias-corrected method of moments estimator in comparison to alternative estimators that have been suggested in the literature, we perform some Monte Carlo experiments. In the baseline scenario, the data are generated from a simplified homoskedastic version of the dynamic panel data model considered by Kiviet et al. (2017) in their simulation exercise:

$$y_{it} = \alpha y_{i,t-1} + \beta x_{it} + \sigma_\mu \mu_i + \sigma_u u_{it},$$
$$x_{it} = \gamma x_{i,t-1} + \pi_\mu \mu_i + \pi_\lambda \lambda_i + \sigma_\epsilon \epsilon_{it}.$$

The regressor x_{it} is strictly exogenous with respect to the idiosyncratic error term u_{it}. The errors u_{it} and ϵ_{it} and the individual-specific effects μ_i and λ_i are drawn from independent standard normal distributions. Following Kiviet et al. (2017), we choose the remaining free parameters to obtain a reasonable characterization of the data-generating process. Further details are relegated to the Supplementary Appendix.

We distinguish between a process with moderate persistence, $\alpha = 0.4$, and high persistence, $\alpha = 0.9$. The process is initialized at $t = -50$ with $y_{i,-50} = x_{i,-50} = 0$, and the first 50 observations are discarded. As a robustness check, we also consider the initialization without burn-in period, $y_{i0} = x_{i0} = 0$, which implies that the observed process starts off its stationary path.

For the experiments with higher-order dynamics, we modify the above data-generating process as follows:

$$y_{it} = \sum_{j=1}^{3} \alpha_j y_{i,t-j} + \beta x_{it} + \sigma_\mu \mu_i + \sigma_u u_{it},$$

and set $(\alpha_1, \alpha_2, \alpha_3) = (0.48, -0.2, 0.12)$ to achieve $\sum_{j=1}^{3} \alpha_j = 0.4$ and $(\alpha_1, \alpha_2, \alpha_3) = (1.08, -0.45, 0.27)$ to obtain $\sum_{j=1}^{3} \alpha_j = 0.9$. All other parameter values are left unchanged.

To analyze the estimators' performance under heteroskedasticity or cross-sectional error dependence, we consider the following modifications of the data-generating process, where the parameterizations ensure that $Var(u_{it}) = 1$ to keep the signal-to-noise ratio unaffected. A data-generating process with heteroskedasticity across both dimensions is obtained by replacing $u_{it} = \sqrt{3/4}\delta_i \tau_t v_{it}$. Uniform cross-sectional dependence is introduced by modifying $u_{it} = \sqrt{3/(4N)} \sum_{j=1}^{N} \omega_{ij} v_{jt}$, and interactive random effects are modeled by letting $u_{it} = \sqrt{3/7}(\delta_i \tau_t + v_{it})$. In all specifications, v_{it} and τ_t are independent standard normally distributed, and δ_i and ω_{ij} are uniformly distributed over the interval $(0, 2)$.

We compare the performance of the uncorrected within-groups estimator (WG), our bias-corrected estimator (BC), the one-step Arellano and Bond (1991) difference GMM estimator (AB-GMM), the two-step Ahn and Schmidt (1995) GMM estimator (AS-GMM) with additional nonlinear moment conditions that are valid under the absence of serial error correlation, the two-step Blundell and Bond (1998) system GMM estimator (BB-

GMM) with additional linear moment conditions valid under mean stationarity, and the Hsiao et al. (2002) QML estimator. In addition to the average bias and root mean square error (RMSE), we report the empirical size of Wald tests given a nominal size of 5%. We consider a fixed-T robust variance-covariance estimator clustered at the individual level and, for the WG and BC estimators in the simulation designs with cross-sectional dependence, a large-T robust variance-covariance estimator clustered at the time periods. For the AS-GMM and the BB-GMM estimators, the finite-sample Windmeijer (2005) correction is applied. Detailed information can be found in the Supplementary Appendix.

We consider all sample size combinations of $T \in \{5, 10, 25, 50\}$ and $N \in \{50, 200\}$. The results are based on 1,000 replications for each simulation design. As discussed in Section 2, for the BC estimator the numerical algorithm might converge to a local minimum of the adjusted likelihood function. We observe a small fraction of such incorrect solutions, especially when T is as small as 5 or 10. The problem disappears when T becomes large. If the negativity condition for the gradient is violated, we re-initialize the algorithm with a random draw for $\alpha^{(0)}$ from the uniform distribution over the interval $(0, 1)$. If necessary, we repeat this process until a solution is found that satisfies the negativity condition. In our experience, this procedure effectively prevents inappropriate solutions.

In the following, we sketch the main findings from our simulation exercise. For the AR(1) model with stationary initial conditions and i.i.d. errors (Tables 1–2 in the Supplementary Appendix), our BC estimator performs very similar in terms of bias and RMSE to the transformed QML estimator. This is not surprising given that the latter can be seen as an estimator that applies a similar bias correction to the first-order condition, as highlighted in Section 3. One should bear in mind, however, that the BC estimator does not require a specific assumption on the initial values of the dynamic process. Furthermore, our estimator is computationally much simpler and can be computed within a fraction of the computing time that is required for the QML estimator, in particular when T becomes large. All GMM estimators perform substantially worse under this baseline data-generating process.

Under higher-order dynamics with 3 lags of the dependent variable (Tables 3–4 in the Supplementary Appendix), the results suggest that the BC estimator effectively removes the bias and yields estimates with the lowest RMSE compared to the GMM estimators. The only exception is the case with high persistence and $T \leq 10$, where the BB-GMM estimator performs best.

When the data-generating process exhibits cross-sectional and time-dependent heteroskedasticity (Tables 5–6 in the Supplementary Appendix), our findings suggest that the BC estimator is robust under moderate persistence even if T is as small as 10. For processes with high persistence, a larger number of time periods, $T \geq 25$, is required to successfully remove the bias. Overall, the BC estimator again performs similar to the QML estimator, and both estimators clearly outperform the GMM estimators in most cases. The BB-GMM estimator again has some advantage under high persistence and small T.

The results under cross-sectional dependence (Tables 7–10 in the Supplementary Appendix) suggest that inference based on the i.i.d. assumption may be severely biased when it is violated. For the model with uniform cross-sectional dependence, the Wald tests with nominal size of 5% reject in more than 50% of the cases, and sometimes even more than 90%. On the other hand, the cluster-robust standard errors proposed in Section 5 effectively correct for cross-sectional dependence and yield empirical sizes close to the nominal size. The robustification even works well in models with a common-factor structure, although the underlying asymptotic theory underlying Theorem 2 does not apply to models with strong cross-sectional dependence.

We finally study the small-sample properties of the estimators when the initial condition affects the distribution of the dependent variable (Tables 11–12 in the Supplementary Appendix). It is well known that the BB-GMM estimator under a nonstationary initialization can be severely biased. This is confirmed by our simulations. All other estimators are virtually unbiased whenever N is sufficiently large. While the BC and QML estimators perform similarly under a stationary initialization, the BC estimator turns out to be more efficient than the QML estimator under the nonstationary initialization.

7 Conclusion

We proposed an estimator that is based on a simple set of moment conditions that can be easily solved with standard numerical optimization procedures. It is straightforward to generalize the estimator to higher-order autoregressive models or dynamic random-effects models. An estimator of the asymptotic covariance matrix is readily available, as are robust standard errors that effectively adjust for cross-sectional dependence, which is a relevant feature in the analysis of macroeconomic panel data.

Besides the added flexibility, the bias-corrected method of moments estimator is easier to implement than likelihood-based estimators that implicitly employ a similar bias correction. In our Monte Carlo simulations, our bias-corrected estimator also performs favorably compared to maximum likelihood and generalized method of moments estimators.

References

Ahn, S. C. and P. Schmidt (1995). Efficient estimation of models for dynamic panel data. *Journal of Econometrics 68*(1), 5–27.

Anderson, T. W. and C. Hsiao (1981). Estimation of dynamic models with error components. *Journal of the American Statistical Association 76*(375), 598–606.

Arellano, M. and S. R. Bond (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies 58*(2), 277–297.

Bai, J. (2013). Fixed-effects dynamic panel models, a factor analytical method. *Econometrica 81*(1), 285–314.

Blundell, R. and S. R. Bond (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics 87*(1), 115–143.

Breitung, J. (2015). The analysis of macroeconomic panel data. In B. H. Baltagi (Ed.), *The Oxford Handbook of Panel Data*, Chapter 15, pp. 453–493. Oxford: Oxford University Press.

Bun, M. J. G. and M. A. Carree (2005). Bias-corrected estimation in dynamic panel data models. *Journal of Business & Economic Statistics 23*(2), 200–210.

Bun, M. J. G., M. A. Carree, and A. Juodis (2017). On maximum likelihood estimation of dynamic panel data models. *Oxford Bulletin of Economics and Statistics 79*(4), 463–494.

Bun, M. J. G. and F. Windmeijer (2010). The weak instrument problem of the system GMM estimator in dynamic panel data models. *Econometrics Journal 13*(1), 95–126.

Dhaene, G. and K. Jochmans (2016). Likelihood inference in an autoregression with fixed effects. *Econometric Theory 32*(5), 1178–1215.

Hahn, J. and G. Kuersteiner (2002). Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and T are large. *Econometrica 70*(4), 1639–1657.

Hayakawa, K. and M. H. Pesaran (2015). Robust standard errors in transformed likelihood estimation of dynamic panel data models. *Journal of Econometrics 188*(1), 111–134.

Holtz-Eakin, D., W. Newey, and H. S. Rosen (1988). Estimating vector autoregressions with panel data. *Econometrica 56*(6), 1371–1395.

Hsiao, C., M. H. Pesaran, and A. K. Tahmiscioglu (2002). Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods. *Journal of Econometrics 109*(1), 107–150.

Judson, R. A. and A. L. Owen (1999). Estimating dynamic panel data models: a guide for macroeconomists. *Economics Letters 65*(1), 9–15.

Juodis, A. (2018). First difference transformation in panel VAR models: Robustness, estimation, and inference. *Econometric Reviews 37*(6), 650–693.

Kiviet, J. F. (1995). On bias, inconsistency, and efficiency of various estimators in dynamic panel data models. *Journal of Econometrics 68*(1), 53–78.

Kiviet, J. F., M. Pleus, and R. Poldermans (2017). Accuracy and efficiency of various GMM inference techniques in dynamic micro panel data models. *Econometrics 5*(1), 14.

Moon, H. R., B. Perron, and P. C. B. Phillips (2015). Incidental parameters and dynamic panel modeling. In B. H. Baltagi (Ed.), *The Oxford Handbook of Panel Data*, Chapter 4, pp. 111–148. Oxford: Oxford University Press.

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica 49*(6), 1417–1426.

Windmeijer, F. (2005). A finite sample correction for the variance of linear efficient two-step GMM estimators. *Journal of Econometrics 126*(1), 25–51.

Appendix A Computational details

A.1 Gradient of the moment function

For the bias-corrected method of moments estimator presented in Section 2, the gradient

$$\boldsymbol{\nabla}\mathbf{m}_{NT}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} \nabla_\alpha m_{\alpha,Ti}(\boldsymbol{\theta}) & \boldsymbol{\nabla}_\beta m_{\alpha,Ti}(\boldsymbol{\theta})' \\ \boldsymbol{\nabla}_\alpha \mathbf{m}_{\beta,Ti}(\boldsymbol{\theta}) & \boldsymbol{\nabla}_\beta \mathbf{m}_{\beta,Ti}(\boldsymbol{\theta}) \end{pmatrix}$$

has the elements

$$\nabla_\alpha m_{\alpha,Ti}(\boldsymbol{\theta}) = -\frac{1}{T}\sum_{t=1}^{T}(y_{i,t-1} - \overline{y}_{-1,i})y_{i,t-1} - \nabla_\alpha b_T(\alpha)\widehat{\sigma}^2_{Ti}(\boldsymbol{\theta}) - b_T(\alpha)\nabla_\alpha\widehat{\sigma}^2_{Ti}(\boldsymbol{\theta}),$$

$$\boldsymbol{\nabla}_\beta m_{\alpha,Ti}(\boldsymbol{\theta}) = -\frac{1}{T}\sum_{t=1}^{T}(\mathbf{x}_{it} - \overline{\mathbf{x}}_i)y_{i,t-1} - b_T(\alpha)\boldsymbol{\nabla}_\beta\widehat{\sigma}^2_{Ti}(\boldsymbol{\theta}),$$

$$\boldsymbol{\nabla}_\alpha m_{\beta,Ti}(\boldsymbol{\theta}) = -\frac{1}{T}\sum_{t=1}^{T}(\mathbf{x}_{it} - \overline{\mathbf{x}}_i)y_{i,t-1},$$

$$\boldsymbol{\nabla}_\beta m_{\beta,Ti}(\boldsymbol{\theta}) = -\frac{1}{T}\sum_{t=1}^{T}(\mathbf{x}_{it} - \overline{\mathbf{x}}_i)\mathbf{x}'_{it},$$

where $\nabla_\alpha b_T(\alpha) = -T^{-2}\sum_{t=1}^{T-2}\sum_{s=1}^{t} s\alpha^{s-1}$ and

$$\nabla_\alpha\widehat{\sigma}^2_{Ti}(\boldsymbol{\theta}) = -\frac{2}{T-1}\sum_{t=1}^{T}(y_{i,t-1} - \overline{y}_{-1,i})e_{it}(\boldsymbol{\theta}),$$

$$\boldsymbol{\nabla}_\beta\widehat{\sigma}^2_{Ti}(\boldsymbol{\theta}) = -\frac{2}{T-1}\sum_{t=1}^{T}(\mathbf{x}_{it} - \overline{\mathbf{x}}_i)e_{it}(\boldsymbol{\theta}).$$

Furthermore, $\mathbb{E}[\nabla_\alpha\widehat{\sigma}^2_{Ti}(\boldsymbol{\theta}_0)] = 2b_T(\alpha_0)\sigma_i^2 T/(T-1)$ and $\mathbb{E}[\boldsymbol{\nabla}_\beta\widehat{\sigma}^2_{Ti}(\boldsymbol{\theta}_0)] = \mathbf{0}$.

For the more general AR(p) model considered in Section 4, the gradient can be expressed compactly as

$$\boldsymbol{\nabla}\mathbf{m}_{NT}(\boldsymbol{\theta}) = -\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{w}_{it} - \overline{\mathbf{w}}_i)\mathbf{z}_{it}(\boldsymbol{\theta})' + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\boldsymbol{\nabla}\mathbf{z}_{it}(\boldsymbol{\theta})[e_{it}(\boldsymbol{\theta}) - \overline{e}_i(\boldsymbol{\theta})],$$

where

$$\boldsymbol{\nabla}\mathbf{z}_{it}(\boldsymbol{\theta}) = -\frac{T}{T-1}\begin{pmatrix}\boldsymbol{\nabla}_\alpha\mathbf{b}_T(\boldsymbol{\alpha}) & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix} e_{it}(\boldsymbol{\theta}) + \frac{T}{T-1}\begin{pmatrix}\mathbf{b}_T(\boldsymbol{\alpha})\\ \mathbf{0}\end{pmatrix}\mathbf{w}'_{it}.$$

The element in the j-th row and k-th column of the $p\times p$ matrix $\boldsymbol{\nabla}_\alpha\mathbf{b}_T(\boldsymbol{\alpha})$ equals $-T^{-2}\boldsymbol{\iota}'_T\mathbf{L}_T^{(k)}[\nabla_{\alpha_j}\mathbf{A}_T(\boldsymbol{\alpha})^{-1}]\boldsymbol{\iota}_T$, with $\nabla_{\alpha_j}\mathbf{A}_T(\boldsymbol{\alpha})^{-1} = -\mathbf{A}_T(\boldsymbol{\alpha})[\nabla_{\alpha_j}\mathbf{A}_T(\boldsymbol{\alpha})]\mathbf{A}_T(\boldsymbol{\alpha})$. The matrix $\nabla_{\alpha_j}\mathbf{A}_T(\boldsymbol{\alpha})$ has all elements on the j-th diagonal below the main diagonal equal to -1 and all other elements equal to 0. Eventually,

$$\mathbb{E}[\boldsymbol{\nabla}\mathbf{m}_{Ti}(\boldsymbol{\theta}_0)] = -\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(\mathbf{w}_{it} - \overline{\mathbf{w}}_i)\mathbf{w}'_{it}\right] - \sigma_i^2\begin{pmatrix}\boldsymbol{\nabla}_\alpha\mathbf{b}_T(\boldsymbol{\alpha}_0) - \frac{2T}{T-1}\mathbf{b}_T(\boldsymbol{\alpha}_0)\mathbf{b}_T(\boldsymbol{\alpha}_0)' & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}.$$

A.2 Bias-corrected maximum likelihood estimation

In Section 3, we argued that the transformed ML estimator of Hsiao et al. (2002) leads to a bias-corrected first-order condition similar to that of the bias-corrected method of moments estimator. This follows from the observation that for $|\alpha_0| < 1$ we obtain

$$\begin{aligned}\mathbb{E}[(y_{i0} - \overline{y}_{-1,i})(u_{i0} - \overline{u}_i)] &= \left[\frac{\omega - 1}{1 - \alpha_0} - \frac{(\omega - 1)(1 - \alpha_0^T)}{(1 - \alpha_0)^2 T} - b_T(\alpha_0)\right]\sigma^2 \\ &= \left[\frac{|\boldsymbol{\Omega}| - 1}{(1 - \alpha_0)T}\left(1 - \frac{1 - \alpha_0^T}{(1 - \alpha_0)T}\right) - b_T(\alpha_0)\right]\sigma^2 \\ &= \left[-b_T(\alpha_0)(|\boldsymbol{\Omega}| - 1) - b_T(\alpha_0)\right]\sigma^2 = -b_T(\alpha_0)\sigma^2|\boldsymbol{\Omega}|.\end{aligned}$$

Appendix B Proofs

B.1 Proof of Theorem 1

Let $\widehat{\boldsymbol{\theta}}_{bc} \in \Theta$ be the point-valued minimizer of the objective function $\mathbf{m}_{NT}(\boldsymbol{\theta})'\mathbf{m}_{NT}(\boldsymbol{\theta})$ such that $\text{plim}_{N\to\infty} \widehat{\boldsymbol{\theta}}_{bc} = \boldsymbol{\theta}_0$. By the mean value theorem,

$$\mathbf{m}_{NT}(\widehat{\boldsymbol{\theta}}_{bc}) = \mathbf{m}_{NT}(\boldsymbol{\theta}_0) + \boldsymbol{\nabla}\mathbf{m}_{NT}(\overline{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}_{bc} - \boldsymbol{\theta}_0),$$

with the mean value $\overline{\boldsymbol{\theta}}$. Due to continuity of the moment function, $\text{plim}_{N\to\infty} \overline{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$. Thus, with $\mathbf{m}_{NT}(\widehat{\boldsymbol{\theta}}_{bc}) = \mathbf{0}$,

$$\sqrt{N}(\widehat{\boldsymbol{\theta}}_{bc} - \boldsymbol{\theta}_0) = -\left(\boldsymbol{\nabla}\mathbf{m}_{NT}(\overline{\boldsymbol{\theta}})\right)^{-1}\sqrt{N}\mathbf{m}_{NT}(\boldsymbol{\theta}_0).$$

From the results in Appendix A.1 and with continuity of the gradient, it follows that $\text{plim}_{N\to\infty}[\boldsymbol{\nabla}\mathbf{m}_{NT}(\overline{\boldsymbol{\theta}})]^{-1} = -[\boldsymbol{\Sigma}_T + \sigma^2\mathbf{B}_T(\alpha_0)]^{-1}$. Due to the independence of $\mathbf{m}_{Ti}(\boldsymbol{\theta}_0)$ across i with $\mathbb{E}[\mathbf{m}_{Ti}(\boldsymbol{\theta}_0)] = \mathbf{0}$ and $\text{plim}_{N\to\infty} N^{-1}\sum_{i=1}^N \mathbf{m}_{Ti}(\boldsymbol{\theta}_0)\mathbf{m}_{Ti}(\boldsymbol{\theta}_0)' = \mathbf{S}_T(\boldsymbol{\theta}_0)$, the central limit theorem yields

$$\sqrt{N}\mathbf{m}_{NT}(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{S}_T(\boldsymbol{\theta}_0))$$

as $N \to \infty$. With Slutsky's theorem, the limiting distribution as in Theorem 1(i) follows.

Let $y_{it} = \xi_{it}(\boldsymbol{\theta}) + v_{it}(\boldsymbol{\theta})$, where $v_{it} = \sum_{j=0}^{t-1}\alpha^j u_{i,t-j}(\boldsymbol{\theta})$ and $\xi_{it}(\boldsymbol{\theta})$ is a function of current and past regressors $\mathbf{x}_{it}$, individual effects μ_i, and initial conditions y_{i0}. After some algebra, we can rewrite the elements of the moment function as

$$\begin{aligned}
m_{\alpha,NT}(\boldsymbol{\theta}) =& \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}[y_{i,t-1} - \overline{\xi}_{-1,i}(\boldsymbol{\theta})]u_{it}(\boldsymbol{\theta}) - \frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T} v_{i,t-1}(\boldsymbol{\theta})u_{is}(\boldsymbol{\theta}) \\
& - \frac{b_T(\alpha)}{N(T-1)}\sum_{i=1}^{N}\sum_{t=1}^{T}[u_{it}(\boldsymbol{\theta}) - \overline{u}_i(\boldsymbol{\theta})]u_{it}(\boldsymbol{\theta}), \\
\mathbf{m}_{\beta,NT}(\boldsymbol{\theta}) =& \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it} - \overline{\mathbf{x}}_i)u_{it}(\boldsymbol{\theta}),
\end{aligned}$$

where $\overline{\xi}_{-1,i}(\boldsymbol{\theta}) = T^{-1}\sum_{t=1}^{T}\xi_{i,t-1}(\boldsymbol{\theta})$. Moreover, $\mathbb{E}[\overline{\xi}_{-1,i}(\boldsymbol{\theta}_0)] = \mathbb{E}[\overline{y}_{-1,i}]$ and

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(y_{i,t-1} - \overline{\xi}_{-1,i}(\boldsymbol{\theta}_0))u_{it}(\boldsymbol{\theta}_0)\right] &= 0, \\
\mathbb{E}\left[\frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T} v_{i,t-1}(\boldsymbol{\theta}_0)u_{is}(\boldsymbol{\theta}_0)\right] &= -b_T(\alpha_0)\sigma_i^2, \\
\mathbb{E}\left[\frac{1}{T-1}\sum_{t=1}^{T}[u_{it}(\boldsymbol{\theta}_0) - \overline{u}_i(\boldsymbol{\theta}_0)]u_{it}(\boldsymbol{\theta}_0)\right] &= \sigma_i^2.
\end{aligned}$$

As $N, T \to \infty$, and with $|\alpha_0| < 1$, it follows that

$$\sqrt{NT}m_{\alpha,NT}(\boldsymbol{\theta}_0) = \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}[y_{i,t-1} - \overline{\xi}_{-1,i}(\boldsymbol{\theta}_0)]u_{it}(\boldsymbol{\theta}_0) + o_p(1),$$

and therefore

$$\sqrt{NT}\mathbf{m}_{NT}(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{\Sigma}).$$

Since $b_T(\alpha_0) = O(T^{-1})$, it follows with the results from Appendix A.1 that

$$\boldsymbol{\nabla}\mathbf{m}_{NT}(\boldsymbol{\theta}_0) = -\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{w}_{it} - \overline{\mathbf{w}}_i)\mathbf{w}_{it}' + O_p(T^{-1}),$$

where $\mathbf{w}_{it} = (y_{i,t-1}, \mathbf{x}_{it}')'$ and $\overline{\mathbf{w}}_i = T^{-1}\sum_{t=1}^{T}\mathbf{w}_{it}$. Consequently, by continuity of the gradient, $\operatorname{plim}_{N,T\to\infty}[\boldsymbol{\nabla}\mathbf{m}_{NT}(\overline{\boldsymbol{\theta}})]^{-1} = -\boldsymbol{\Sigma}^{-1}$, and the result in Theorem 1(ii) follows.

B.2 Proof of Theorem 2

Define $\mathbf{q}_{Nt}(\boldsymbol{\theta}) = N^{-1/2}\sum_{i=1}^{N}\mathbf{w}_{it}u_{it}(\boldsymbol{\theta})$. As $N, T \to \infty$ and with $|\alpha_0| < 1$,

$$\frac{1}{NT}\sum_{t=1}^{T}\mathbf{Z}_t(\boldsymbol{\theta})'[\mathbf{u}_t(\boldsymbol{\theta}) - \overline{\mathbf{u}}(\boldsymbol{\theta})][\mathbf{u}_t(\boldsymbol{\theta}) - \overline{\mathbf{u}}(\boldsymbol{\theta})]'\mathbf{Z}_t(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{q}_{Nt}(\boldsymbol{\theta})\mathbf{q}_{Nt}(\boldsymbol{\theta})' + o_p(1).$$

Since $u_{it}(\boldsymbol{\theta})$ is independent of $\mathbf{w}_{it}$, it follows that $\mathbb{E}[\mathbf{q}_{Nt}(\boldsymbol{\theta}_0)] = \mathbf{0}$. Furthermore, if the eigenvalues of $\boldsymbol{\Sigma}_{u,t}$ are bounded for all t, the norm of $\mathbb{E}[\mathbf{q}_{Nt}(\boldsymbol{\theta}_0)\mathbf{q}_{Nt}(\boldsymbol{\theta}_0)']$ is bounded as well and

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathbf{q}_{Nt}(\boldsymbol{\theta}_0)\mathbf{q}_{Nt}(\boldsymbol{\theta}_0)'] = \operatorname*{plim}_{N,T\to\infty}\frac{1}{NT}\sum_{t=1}^{T}\mathbf{W}_t'\boldsymbol{\Sigma}_{u,t}\mathbf{W}_t = \mathbf{S}(\boldsymbol{\theta}_0).$$

Then,

$$\sqrt{NT}\mathbf{m}_{NT}(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{S}(\boldsymbol{\theta}_0)).$$

The remainder of the proof follows along the lines of Appendix B.1.