Research Program on Forecasting

QUASI MAXIMUM-LIKELIHOOD ESTIMATION OF DYNAMIC PANEL DATA MODELS FOR SHORT TIME SERIES

Robert F. Phillips
The George Washington University
rphil@gwu.edu

RPF Working Paper No. 2014-006
http://www.gwu.edu/~forcpgm/2014-006.pdf

September 23, 2014

RESEARCH PROGRAM ON FORECASTING
Center of Economic Research
Department of Economics
The George Washington University
Washington, DC 20052
http://www.gwu.edu/~forcpgm

QUASI MAXIMUM-LIKELIHOOD ESTIMATION OF DYNAMIC PANEL DATA MODELS FOR SHORT TIME SERIES

Robert F. Phillips*

Department of Economics, George Washington University

September 2014

Abstract

This paper establishes the almost sure convergence and asymptotic normality of quasi maximum-likelihood (QML) estimators of a dynamic panel data model when the time series for each cross section is short. The QML estimators are robust with respect to initial conditions and misspecification of the log-likelihood, and results are provided for a general specification of the error variance-covariance matrix. The paper also provides procedures for computing QML estimates that improve on computational methods previously recommended in the literature. Moreover, it compares the finite sample performance of several QML estimators, the differenced GMM estimator, and the system GMM estimator.

Keywords: random effects; fixed effects; differenced QML; augmented dynamic panel data model

JEL code: C23

* Address: 2115 G Street NW, Suite 340, Washington DC 20052; phone: 202-994-8619; fax: 202-994-6147; e-mail: rphil@gwu.edu

1 Introduction

Dynamic panel data models are often estimated with samples for which the number of cross sections (N) far exceeds the number of available time periods (T). When T is small, straightforward application of maximum-likelihood can yield unreliable estimates, a fact that has been known for over forty years (see, e.g., Nerlove 1971). Recently, however, papers by Phillips (2010) and Kruiniger (2013) have shown how quasi maximum-likelihood (QML) can produce consistent and asymptotically normal estimators when the model contains lagged dependent variable regressors and T is small. Indeed, these QML estimators can perform well in situations where generalized method of moments (GMM) estimators do not. For example, unlike the differenced GMM estimator introduced by Arellano and Bond (1991), the QML estimators described in Phillips (2010) and Kruiniger (2013) have little finite sample bias even for small T and significant persistence in the dependent variable (see Phillips 2010; Kruiniger 2013).

This paper makes several contributions to this nascent literature on QML estimation. I examine two approaches — levels and differenced QML — to estimating the parameters of a pth-order dynamic panel data model. The model studied includes p lags of the dependent variable as well as other explanatory variables. The QML estimators of the parameters of the model are robust with respect to initial conditions and misspecification of the log-likelihood, but they are not robust with respect to misspecification of the unconditional error variance-covariance matrix. Therefore, for the most general error variance-covariance matrices possible, I provide large-sample results for $N \to \infty$ with T fixed. By considering general error variance-covariance matrices, the analysis subsumes, as special cases, error models appearing in the literature as well as cases not yet examined.

The paper also shows how QML estimates can be calculated by iterated feasible generalized least squares (IFGLS) procedures, both when the error variance-covariance matrix is unstructured and when well-known structured error variance-covariance matrices are assumed. Phillips (2010) provided an IFGLS procedure for computing QML estimates while assuming a specialized error variance-covariance matrix. Unlike in Phillips (2010), however, here I provide IFGLS procedures based on the expectation-conditional maximization either (ECME) algorithm. Monte Carlo evidence illustrates that an ECME algorithm more reliably computes QML estimates than the IFGLS procedure described in Phillips (2010).

Finally, using simulated data, the finite sample behavior of several QML estimators — both differenced and in levels — are compared, and their finite sample behavior is compared to the differenced GMM esti-

mator, the system GMM estimator (Blundell and Bond 1998), and ordinary least squares (OLS) applied to an augmented model. Although OLS is inconsistent, it is included in the experiments to illustrate that, when most of the correlation in the regression errors is removed via a control function, its sampling performance can improve on GMM and QML estimators.

2 QML via Regression Augmentation

Since Anderson and Hsiao (1981) it has been known that whether or not application of maximum-likelihood to a dynamic panel data model will yield a consistent estimator, as $N \to \infty$ with T fixed, depends on initial conditions. Unfortunately, these initial conditions are rather restrictive. This section shows how QML estimation can be made free of initial condition restrictions through the use of a suitable control function.

To see which control function is useful and why, consider the pth-order dynamic panel data model

$$\boldsymbol{y}_i = \mathbf{Y}_i \delta_0 + \mathbf{X}_i \boldsymbol{\beta}_0 + \boldsymbol{e}_i \qquad (i = 1, \ldots, N)\,. \tag{1}$$

In this expression $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})'$, $\mathbf{Y}_i = \left(\boldsymbol{y}_{i,-1}, \ldots, \boldsymbol{y}_{i,-p}\right)$, $\boldsymbol{y}_{i,-j} = \left(y_{i,1-j}, \ldots, y_{i,T-j}\right)'$ ($j = 1, \ldots, p$), and $\mathbf{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT})'$, with $\boldsymbol{x}_{it}$ a $K \times 1$ vector of explanatory variables that vary with t (for at least some i). Moreover, $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{iT})'$ is a vector of regression errors. For notational convenience, the numbering of observed variables begins with $t = -p + 1$.

Now let $\boldsymbol{y}_i^o = \left(y_{i0}, \ldots, y_{i,-p+1}\right)'$; let $\boldsymbol{x}_i$ be a column vector consisting of all of the distinct elements of $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT}$; and set $\boldsymbol{z}_i = \left(\boldsymbol{x}_i', \boldsymbol{y}_i^{o\prime}\right)'$. Then, assuming $\boldsymbol{e}_i | \boldsymbol{z}_i \sim IIN\left(\mathbf{0}, \Omega_0^*\right)$, the log-likelihood is given by

$$-\frac{NT}{2} \ln(2\pi) - \frac{N}{2} \ln|\Omega^*| - \frac{1}{2} \sum_{i=1}^{N} \boldsymbol{e}_i(\boldsymbol{\varphi})' \, \Omega^{*-1} \boldsymbol{e}_i(\boldsymbol{\varphi})\,, \tag{2}$$

where $\boldsymbol{e}_i(\boldsymbol{\varphi}) = \boldsymbol{y}_i - \mathbf{Y}_i \delta - \mathbf{X}_i \boldsymbol{\beta}$, and $\boldsymbol{\varphi} = \left(\boldsymbol{\delta}', \boldsymbol{\beta}'\right)'$.

To see why maximizing the log-likelihood in (2) does not produce a reliable estimator for small T, observe that the consistency of a maximizer of the log-likelihood in (2) depends on the following moment restrictions $E\left(\mathbf{X}_i' \Omega_0^{*-1} \boldsymbol{e}_i\right) = \mathbf{0}$ and $E\left(\boldsymbol{y}_{i,-j}' \Omega_0^{*-1} \boldsymbol{e}_i\right) = \mathbf{0}$ ($j = 1, \ldots, p$). We have $E\left(\mathbf{X}_i' \Omega_0^{*-1} \boldsymbol{e}_i\right) = \mathbf{0}$ if the regressors in $\mathbf{X}_i$ are strictly exogenous with respect to the errors in $\boldsymbol{e}_i$. On the other hand, the moment restrictions $E\left(\boldsymbol{y}_{i,-j}' \Omega_0^{*-1} \boldsymbol{e}_i\right) = \mathbf{0}$ ($j = 1, \ldots, p$) depend on an even stronger assumption, which is summarized in Lemma 1.

Lemma 1. If $E\left(\boldsymbol{e}_i \boldsymbol{y}_i^{o\prime}\right) = \mathbf{0}$, $E\left(\boldsymbol{e}_i \boldsymbol{x}_i'\right) = \mathbf{0}$, and $E\left(\boldsymbol{e}_i \boldsymbol{e}_i'\right) = \Omega_0^*$, then $E\left(\boldsymbol{y}_{i,-j}' \Omega_0^{*-1} \boldsymbol{e}_i\right) = \mathbf{0}$ $(j = 1, \ldots, p)$.

Proof. See Appendix A.

According to Lemma 1, if the regressors in $\boldsymbol{x}_{it}$ are all uncorrelated with respect to the e_{it}s *and* the initial values of the dependent variable $y_{i0}, \ldots, y_{i,-p+1}$ are uncorrelated with the subsequent errors $e_{i1}, \ldots, e_{iT}$, then $E\left(\boldsymbol{y}_{i,-j}' \Omega_0^{*-1} \boldsymbol{e}_i\right) = \mathbf{0}$ $(j = 1, \ldots, p)$. Assuming the initial values of the dependent variable are uncorrelated with subsequent errors is quite restrictive. For example, suppose the errors are given by the error-components model

$$e_{it} = c_i + v_{it}. \tag{3}$$

If the v_{it}s are uncorrelated, we can take v_{it} to be uncorrelated with the elements of $\boldsymbol{y}_i^o$, for $t \geq 1$, but assuming the elements of $\boldsymbol{y}_i^o$ are also uncorrelated with c_i is a strong initial condition restriction.

Fortunately, we need make no such initial condition assumption if the model in (1) is augmented with a suitable control function. Nor need we assume the regressors in $\boldsymbol{x}_{it}$ are strictly exogenous with respect to the e_{it}s. The possible correlation between the elements in $\boldsymbol{e}_i$ and the elements in $\boldsymbol{z}_i$ can be controlled for by the linear projection of e_{it} on 1 and $\boldsymbol{z}_i$:

$$e_{it} = \mu_0 + \boldsymbol{z}_i' \boldsymbol{\theta}_0 + u_{it}, \qquad (t = 1, \ldots, T,\ i = 1, \ldots, N) \tag{4}$$

where $\boldsymbol{\theta}_0 = Var\left(\boldsymbol{z}_i\right)^{-1} Cov\left(\boldsymbol{z}_i, e_{it}\right)$ and $\mu_0 = E\left(e_{it}\right) - E\left(\boldsymbol{z}_i'\right)\boldsymbol{\theta}_0$. The linear projection parameters μ_0 and $\boldsymbol{\theta}_0$ exist and depend on neither i nor t if the moments determining them exist and do not depend on i and t.

The restriction that the linear projection parameters are independent of t is met if the errors have a one-way error-components structure given by (3) and v_{it} is a mean zero random variable that is uncorrelated with the elements of $\boldsymbol{z}_i$ for $t \geq 1$. Then $Cov\left(\boldsymbol{z}_i, e_{it}\right) = Cov\left(\boldsymbol{z}_i, c_i\right)$ and $E\left(e_{it}\right) = E\left(c_i\right)$ for $t \geq 1$.

For this case, the linear projection reduces to that considered in Phillips (2010). Specifically, we have

$$c_i = \mu_0 + \boldsymbol{z}_i' \boldsymbol{\theta}_0 + a_i \qquad (i = 1, \ldots, N) \tag{5}$$

(cf Phillips 2010, p. 411, Eq. (2)).[1] When the errors can be decomposed as in Eq. (3), $\mu_0 + \boldsymbol{z}_i' \boldsymbol{\theta}_0$ controls for

[1] See also Kruiniger (2013), who uses a linear projection of an individual effect on y_{i0}. The linear projection parameters used in that paper are implicitly assumed to be independent of i.

possible correlation between time-invariant unobservables, captured by c_i, and the elements of $\boldsymbol{z}_i$. However, restricting v_{it} to be uncorrelated with the elements of $\boldsymbol{z}_i$, for $t \geq 1$, of course, implies the regressors in $\boldsymbol{x}_{it}$ are uncorrelated with v_{it}, for $t \geq 1$.[2]

Replacing the right-hand side of (4) for e_{it} in (1) gives the augmented dynamic panel data model

$$\boldsymbol{y}_i = \boldsymbol{W}_i \gamma_0 + \boldsymbol{u}_i, \qquad (i = 1, \ldots, N), \tag{6}$$

where $\boldsymbol{W}_i = (\boldsymbol{Y}_i, \boldsymbol{Z}_i)$, $\boldsymbol{Z}_i = \left(\boldsymbol{X}_i, \boldsymbol{\iota}, \boldsymbol{\iota}\, \boldsymbol{z}_i'\right)$, $\boldsymbol{\iota}$ is a $T \times 1$ vector of ones, and $\gamma_0 = \left(\boldsymbol{\delta}_0', \boldsymbol{\beta}_0', \mu_0, \boldsymbol{\theta}_0'\right)'$. The errors in this augmented model — $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{iT})'$ — are now uncorrelated with the elements of $\boldsymbol{Z}_i$ by construction. Thus, upon letting $\Omega_0 = E\left(\boldsymbol{u}_i \boldsymbol{u}_i'\right)$, we have $E\left(\boldsymbol{Z}_i' \Omega_0^{-1} \boldsymbol{u}_i\right) = \boldsymbol{0}$. Moreover, because $E\left(\boldsymbol{u}_i \boldsymbol{y}_i^{o\prime}\right) = \boldsymbol{0}$ and $E\left(\boldsymbol{u}_i \boldsymbol{x}_i'\right) = \boldsymbol{0}$, it follows from Lemma 1 that $E\left(\boldsymbol{y}_{i,-j}' \Omega_0^{-1} \boldsymbol{u}_i\right) = \boldsymbol{0}$ $(j = 1, \ldots, p)$. Consequently, applying QML to the augmented model in (6) should now yield a consistent estimator under suitable conditions.

Theorems 1 and 2 provide sufficient conditions for the almost sure convergence of the QML estimator and its asymptotic normality (as $N \to \infty$, with T fixed). In order to state these theorems, let $\boldsymbol{\Omega}$ be a positive definite matrix and define the quasi log-likelihood that is maximized as $\sum_{i=1}^{N} l_i(\psi)$, where

$$l_i(\psi) = -\frac{T}{2} \ln(2\pi) - \frac{1}{2} \ln|\Omega| - \frac{1}{2} \boldsymbol{u}_i(\gamma)' \Omega^{-1} \boldsymbol{u}_i(\gamma),$$

$\boldsymbol{u}_i(\gamma) = \boldsymbol{y}_i - \boldsymbol{W}_i \gamma$, $\gamma = \left(\boldsymbol{\delta}', \boldsymbol{\beta}', \mu, \boldsymbol{\theta}'\right)'$, $\psi = \left(\gamma', \boldsymbol{\omega}'\right)'$, and $\boldsymbol{\omega} = \text{vech}(\boldsymbol{\Omega})$. Also, set $L_N(\psi) = N^{-1} \sum_{i=1}^{N} l_i(\psi)$ and $\boldsymbol{H}_N(\psi) = \partial^2 L_N(\psi) / \partial\psi \partial\psi'$; let x_{itk} denote the kth element of $\boldsymbol{x}_{it}$; and set $\Psi = \left\{\psi = \left(\gamma', \boldsymbol{\omega}'\right)' \in \mathbb{R}^m : \boldsymbol{\Omega} \text{ is positive definite}\right\}$.

Theorem 1. Assume the following conditions are satisfied:

C1: $E|y_{it}|^{2+\epsilon} < M$ and $E|x_{itk}|^{2+\epsilon} < M$ for all i, t, and k and some $\epsilon > 0$ and $M < \infty$;

C2: $Var(\boldsymbol{z}_i)$ is a positive definite matrix that is independent of i, $E(\boldsymbol{z}_i)$ is independent of i, and $E(e_{it})$ and $E(\boldsymbol{z}_i e_{it})$ are independent of i and t, for $t \geq 1$;

C3: $E\left(\boldsymbol{u}_i \boldsymbol{u}_i'\right) = \boldsymbol{\Omega}_0$ for all i, with $\boldsymbol{\Omega}_0$ a positive definite matrix;

[2] An even simpler case in which the linear projection parameters trivially depend on neither i nor t is when there are no individual specific effects, that is, the model in (3) does not hold, and instead the e_{it}s are uncorrelated among themselves and with the elements of $\boldsymbol{z}_i$, for $t \geq 1$. In this case, $\boldsymbol{\theta}_0 = \boldsymbol{0}$, and the linear projection in (4) simplifies to $e_{it} = \mu_0 + u_{it}$, assuming $E(e_{it}) = \mu_0$.

C4: the limits $\lim_{N\to\infty} N^{-1}\sum_i E\left(y_{is}y_{it}\right)$, $\lim_{N\to\infty} N^{-1}\sum_i E\left(y_{is}x_{itk}\right)$, and $\lim_{N\to\infty} N^{-1}\sum_i E\left(x_{isj}x_{itk}\right)$ exist for all s, t, j, and k; and

C5: the vectors $\left(\boldsymbol{z}_1', \boldsymbol{y}_1'\right)', \ldots, \left(\boldsymbol{z}_N', \boldsymbol{y}_N'\right)'$ are independent for all N.

Then, the limit $\boldsymbol{H}\left(\boldsymbol{\psi}\right) = \lim_{N\to\infty} E\left[\boldsymbol{H}_N\left(\boldsymbol{\psi}\right)\right]$ exists. Moreover, if $\boldsymbol{H}_0 = \boldsymbol{H}\left(\boldsymbol{\psi}_0\right)$ is negative definite, where $\boldsymbol{\psi}_0 = \left(\boldsymbol{\gamma}_0', \boldsymbol{\omega}_0'\right)'$ and $\boldsymbol{\omega}_0 = \text{vech}(\boldsymbol{\Omega}_0)$, then there is a compact subset, say $\overline{\Psi}$, of Ψ, with $\boldsymbol{\psi}_0$ in its interior, and there is a measurable maximizer, $\widehat{\boldsymbol{\psi}}$, of $L_N\left(\cdot\right)$ in $\overline{\Psi}$ such that $\widehat{\boldsymbol{\psi}} \xrightarrow{a.s.} \boldsymbol{\psi}_0$ ($N \to \infty$, T fixed).

Proof. See Appendix B.

Theorem 2. Assume Conditions C2–C5 are satisfied, $\boldsymbol{H}_0$ is negative definite, and the following conditions are satisfied:

C1′: $E\left|y_{it}\right|^{4+\epsilon} < M$ and $E\left|x_{itk}\right|^{4+\epsilon} < M$ for all i, t, and k and some $\epsilon > 0$ and $M < \infty$; and

C6: the limit $\mathcal{I}_0 = \lim_{N\to\infty} N^{-1}\sum_i E\left[\left(\partial l_i\left(\boldsymbol{\psi}_0\right)/\partial\boldsymbol{\psi}\right)\left(\partial l_i\left(\boldsymbol{\psi}_0\right)/\partial\boldsymbol{\psi}\right)'\right]$ exists and is positive definite.

Then $\sqrt{N}\left(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\right) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{H}_0^{-1}\mathcal{I}_0\boldsymbol{H}_0^{-1}\right)$ ($N \to \infty$, T fixed).

Proof. See Appendix C.

The conditions and conclusions of Theorems 1 and 2 reveal that the log-likelihood $\sum_{i=1}^N l_i\left(\boldsymbol{\psi}\right)$ may be misspecified. The log-likelihood is based on the assumption that $\boldsymbol{u}_i$ is normally distributed with mean vector $\boldsymbol{0}$ and variance-covariance matrix $\boldsymbol{\Omega}_0$, conditionally on $\boldsymbol{z}_i$. These conditions imply $\boldsymbol{u}_i$ and $\boldsymbol{z}_i$ are independent, but the conclusions of Theorems 1 and 2 do not require this restriction. In particular, the conditional variance-covariance matrix of $\boldsymbol{u}_i$ given $\boldsymbol{z}_i$ may depend on elements in $\boldsymbol{z}_i$ — specifically, the errors may be conditionally heteroskedastic so long as the unconditional variance-covariance matrix $\boldsymbol{\Omega}_0$ does not depend on i. Moreover, $\boldsymbol{u}_i$ need not be normally distributed.

The conditions in Theorems 1 and 2 do not require the random vectors $\left(\boldsymbol{z}_1', \boldsymbol{y}_1'\right)', \ldots, \left(\boldsymbol{z}_N', \boldsymbol{y}_N'\right)'$ be drawn from a common distribution. On the other hand, some homogeneity is assumed. For example, Condition C3 requires $\boldsymbol{\Omega}_0$ be the same across i, and C1 and C2 ensure the linear projection parameters μ_0 and $\boldsymbol{\theta}_0$ are defined and depend on neither i nor t.

Theorems 1 and 2 cover models previously considered in the literature. For example, structured error variance-covariance matrices, such as those considered by Phillips (2010) and Kruiniger (2013), are special

cases of Ω_0, and, therefore, Theorems 1 and 2 apply to those cases. Specifically, if e_{it} can be decomposed as in (3), with $E(v_{it}) = 0$ and $Cov(\boldsymbol{z}_i, v_{it}) = \boldsymbol{0}$ for $t \geq 1$, then, as already noted, the linear projection in (4) reduces to the linear projection of c_i on 1 and $\boldsymbol{z}_i$ given by (5). In this case, $\boldsymbol{u}_i = \boldsymbol{\iota}\, a_i + \boldsymbol{v}_i$, with $\boldsymbol{v}_i = (v_{i1}, \ldots, v_{iT})'$. Also, for this case, $\Omega_0 = \sigma_{a0}^2 \boldsymbol{\iota}\boldsymbol{\iota}' + \Sigma_0$, where $\sigma_{a0}^2 = var(a_i)$ and $\Sigma_0 = E\left(\boldsymbol{v}_i \boldsymbol{v}_i'\right)$. Both Phillips (2010) and Kruiniger (2013) studied QML estimation for special cases of Σ_0. Phillips (2010), for example, examined estimation of a dynamic panel data model while assuming $\Sigma_0 = \sigma_0^2 \boldsymbol{I}$. Kruiniger (2013), on the other hand, studied estimation of a first-order autoregressive (AR(1)) panel data model with $\Sigma_0 = \text{diag}(\sigma_{01}^2, \ldots, \sigma_{0T}^2)$.

Finally, although Alvarez and Arellano (2003) and Kruiniger (2013) describe QML estimation of an AR(1) panel data model as random-effects QML when c_i is taken to be random, this description is eschewed here, for the label random-effects is misleading when $\boldsymbol{x}_{it}$ is included in the model. This is because a random-effects model is typically identified as a case in which c_i is not only random but, more importantly, assumed to be uncorrelated with the elements of $\boldsymbol{x}_{it}$. But the linear projection in (5) allows for c_i to be correlated with the elements of $\boldsymbol{x}_{it}$ in an arbitrary fashion.

Moreover, assuming random effects provides no advantages in terms of estimation, for it provides no simplications. In particular, even when c_i is uncorrelated with the elements of $\boldsymbol{x}_{it}$, for all t, this does not imply we can drop $\boldsymbol{x}_i$ from the control function $\mu_0 + \boldsymbol{z}_i'\boldsymbol{\theta}_0$. To see this, consider the linear projection of c_i on just 1 and $\boldsymbol{y}_i^o$:

$$c_i = \mu_{y0} + \boldsymbol{y}_i^{o\prime}\boldsymbol{\theta}_{y0} + a_{yi} \qquad (i = 1, \ldots, N), \tag{7}$$

where $\boldsymbol{\theta}_{y0} = Var\left(\boldsymbol{y}_i^o\right)^{-1} Cov\left(\boldsymbol{y}_i^o, c_i\right)$ and $\mu_{y0} = E(c_i) - E\left(\boldsymbol{y}_i^{o\prime}\right)\boldsymbol{\theta}_{y0}$. If we augment the model in (1) with the control function $\mu_{y0} + \boldsymbol{y}_i^{o\prime}\boldsymbol{\theta}_{y0}$ rather than the control function $\mu_0 + \boldsymbol{z}_i'\boldsymbol{\theta}_0$, then the error term in the augmented model is $a_{yi} + v_{it}$ rather than $a_i + v_{it}$, and, in order for QML estimation of the augmented model to be consistent, we must have not just $Cov\left(\boldsymbol{y}_i^o, a_{yi}\right) = \boldsymbol{0}$, which the linear projection in (7) ensures, but also $Cov\left(\boldsymbol{x}_i, a_{yi}\right) = \boldsymbol{0}$, which the linear projection in (7) does not guarantee. Indeed, given $Cov(\boldsymbol{x}_i, c_i) = \boldsymbol{0}$, the result $Cov\left(\boldsymbol{x}_i, a_{yi}\right) = \boldsymbol{0}$ is not guaranteed unless $Cov\left(\boldsymbol{x}_i, \boldsymbol{y}_i^{o\prime}\boldsymbol{\theta}_{y0}\right) = \boldsymbol{0}$,[3] which will not be satisfied in general assuming $\boldsymbol{\theta}_{y0} \neq \boldsymbol{0}$. Of course, this observation does not mean some elements of $\boldsymbol{\theta}_0$ cannot be small or even zero in some applications, in which case the control function can be simplified. But this simplification is not implied by the random-effects model.

[3] This conclusion follows from $Cov\left(\boldsymbol{x}_i, a_{yi}\right) = Cov(\boldsymbol{x}_i, c_i - \mu_{y0} - \boldsymbol{y}_i^{o\prime}\boldsymbol{\theta}_{y0}) = -Cov(\boldsymbol{x}_i, \boldsymbol{y}_i^{o\prime}\boldsymbol{\theta}_{y0})$ if $Cov(\boldsymbol{x}_i, c_i) = \boldsymbol{0}$.

3 Differenced QML

In this section an alternative QML estimation approach — differenced or fixed-effects QML — is examined. Kruiniger (2013) studied fixed-effects QML for an AR(1) panel data model. Hsiao et al. (2002), on the other hand, studied fixed-effects maximum-likelihood estimation, after differencing, and, like this paper, considered a model with additional explanatory variables beyond a lagged dependent variable.

Differencing observations and then applying QML estimation has two advantages over QML estimation based on using observations in levels. First, if the error-components structure in (3) is assumed, differencing eliminates c_i, and the elimination of c_i allows for not only the possibility the elements of $\boldsymbol{x}_{it}$ are correlated with c_i in an arbitrary fashion, should the c_is be treated as random, but also for the possibility the c_is are fixed parameters. Furthermore, even if the c_is are treated as random, differenced QML does not require imposing any moment homogeneity assumptions on the c_is, whereas QML estimation in levels does.

The second advantage of differencing is it allows for a broader range of possible stochastic models for the error terms. This is because differenced QML does not require a control function involving $\boldsymbol{y}_i^o$. To see that using a control function involving $\boldsymbol{y}_i^o$ imposes restrictions on the error generating process, observe that Condition C2 in Theorem 1 implies $Cov\left(\boldsymbol{y}_i^o, e_{it}\right)$ does not depend on t (≥ 1). Sufficient conditions for this restriction are the error-components model in (3) with $Cov\left(\boldsymbol{y}_i^o, v_{it}\right) = \mathbf{0}$, for $t \geq 1$, for then $Cov\left(\boldsymbol{y}_i^o, e_{it}\right) = Cov\left(\boldsymbol{y}_i^o, c_i\right)$, for $t \geq 1$. But, unfortunately, the restriction $Cov\left(\boldsymbol{y}_i^o, v_{it}\right) = \mathbf{0}$, for $t \geq 1$, rules out some time-series processes for the v_{it}s. For example, if the stochastic processes for the cross sections began in the past, then, although the condition $Cov\left(\boldsymbol{y}_i^o, v_{it}\right) = \mathbf{0}$, for $t \geq 1$, does not rule out correlation between v_{it} and v_{is} for $t \geq 1$ and $s \geq 1$, it does rule out correlation between v_{it} and v_{is} for $t \geq 1$ and $s < 1$. Hence, it rules out, for example, a first-order moving average (MA(1)) process for the v_{it}s, for $Cov\left(\boldsymbol{y}_i^o, v_{it}\right) = \mathbf{0}$ fails for $t = 1$ given MA(1) v_{it}s. On the other hand, differenced QML estimation does not rely on a control function involving $\boldsymbol{y}_i^o$, and, therefore, even when the errors are generated by the error-components process in (3), it does not require $Cov\left(\boldsymbol{y}_i^o, v_{it}\right) = \mathbf{0}$ for $t \geq 1$. Hence, it does not rule out processes for which v_{it} and v_{is} are correlated for $t \geq 1$ and $s < 1$.

On the other hand, for differenced QML, instead of using a control function that involves $\boldsymbol{y}_i^o$ we must estimate a system of equations that includes a separate linear projection for each initial difference $\Delta y_{i,-p+2}, \ldots, \Delta y_{i1}$, where $\Delta y_{it} = y_{it} - y_{i,t-1}$. Specifically, suppose $Var\left(\boldsymbol{x}_i\right)$ is positive definite, and set $\boldsymbol{\theta}_{0,p+1-j} = Var\left(\boldsymbol{x}_i\right)^{-1} Cov\left(\boldsymbol{x}_i, \Delta y_{i,-j+2}\right)$ and $\mu_{0,p+1-j} = E\left(\Delta y_{i,-j+2}\right) - E\left(\boldsymbol{x}_i'\right)\boldsymbol{\theta}_{0,p+1-j}$

($j = 1, \ldots, p$). Then, differenced QML relies on the linear projections

$$\Delta y_{i,-j+2} = \mu_{0,p+1-j} + \boldsymbol{x}_i'\boldsymbol{\theta}_{0,p+1-j} + r_{i,p+1-j} \qquad (j = 1, \ldots, p). \tag{8}$$

Here $r_{i,p+1-j}$ is a linear projection residual, which is, by construction, uncorrelated with all of the elements of $\boldsymbol{x}_i$. Note that because the linear projection in (8) does not specify how $\Delta y_{i,-j+2}$ was generated — it only allows for the possibility $\Delta y_{i,-j+2}$ is correlated with elements of $\boldsymbol{x}_i$ — it does not depend on initial condition restrictions.

For differenced QML, in addition to the linear projection equations in (8) we estimate the differenced equation:

$$\Delta \boldsymbol{y}_i = \Delta \boldsymbol{Y}_i\boldsymbol{\delta}_0 + \Delta \boldsymbol{X}_i\boldsymbol{\beta}_0 + \Delta \boldsymbol{e}_i \qquad (i = 1, \ldots, N), \tag{9}$$

where $\Delta \boldsymbol{y}_i = (\Delta y_{i2}, \ldots, \Delta y_{iT})'$, $\Delta \boldsymbol{Y}_i = \left(\Delta \boldsymbol{y}_{i,-1}, \ldots, \Delta \boldsymbol{y}_{i,-p}\right)$, and $\Delta \boldsymbol{y}_{i,-j} = \left(\Delta y_{i,-j+2}, \ldots \Delta y_{i,T-j}\right)'$ ($j = 1, \ldots, p$). Moreover, $\Delta \boldsymbol{X}_i = (\Delta \boldsymbol{x}_{i2}, \ldots, \Delta \boldsymbol{x}_{iT})'$, $\Delta \boldsymbol{x}_{it} = \boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1}$, and $\Delta \boldsymbol{e}_i = (\Delta e_{i2}, \ldots, \Delta e_{iT})'$, with $\Delta e_{it} = e_{it} - e_{i,t-1}$.

If the error-components model in (3) is assumed and the elements in $\boldsymbol{x}_i$ are strictly exogenous with respect to the v_{it}s, then, because c_i is eliminated by differencing, the elements of $\Delta \boldsymbol{X}_i$ are uncorrelated with the elements of $\Delta \boldsymbol{e}_i$, and the model in (9) need not be augmented with a control function. Therefore, for this case, the equations in (8) and (9) can be estimated as a system of equations given by

$$\widetilde{\boldsymbol{y}}_i = \widetilde{\boldsymbol{W}}_i\boldsymbol{\eta}_0 + \widetilde{\boldsymbol{u}}_i \qquad (i = 1, \ldots, N), \tag{10}$$

with $\widetilde{\boldsymbol{y}}_i = \left(\Delta y_{i,-p+2}, \ldots, \Delta y_{i1}, \Delta \boldsymbol{y}_i'\right)'$, $\widetilde{\boldsymbol{u}}_i = \left(r_{i1}, \ldots, r_{ip}, \Delta \boldsymbol{e}_i'\right)'$,

$$\widetilde{\boldsymbol{W}}_i = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_p \otimes \left(1, \boldsymbol{x}_i'\right) \\ \Delta \boldsymbol{Y}_i & \Delta \boldsymbol{X}_i & \boldsymbol{0} \end{pmatrix},$$

and $\boldsymbol{\eta}_0 = \left(\boldsymbol{\delta}_0', \boldsymbol{\beta}_0', \mu_{01}, \boldsymbol{\theta}_{01}', \mu_{02}, \boldsymbol{\theta}_{02}', \ldots, \mu_{0p}, \boldsymbol{\theta}_{0p}'\right)'$.

If $\widetilde{\boldsymbol{u}}_i$ is multivariate normal with mean vector $\boldsymbol{0}$ and variance-covariance matrix Υ_0 conditional on $\boldsymbol{x}_i$,

then the log-likelihood for the system in (10) is $\sum_{i=1}^{N} \widetilde{l}_i(\boldsymbol{\lambda})$, where

$$\widetilde{l}_i(\boldsymbol{\lambda}) = -\frac{(T+p-1)}{2}\ln(2\pi) - \frac{1}{2}\ln|\Upsilon| - \frac{1}{2}\widetilde{\boldsymbol{u}}_i(\boldsymbol{\eta})'\,\Upsilon^{-1}\widetilde{\boldsymbol{u}}_i(\boldsymbol{\eta}),$$

$\widetilde{\boldsymbol{u}}_i(\boldsymbol{\eta}) = \widetilde{\boldsymbol{y}}_i - \widetilde{\boldsymbol{W}}_i\boldsymbol{\eta}$, $\boldsymbol{\eta} = \left(\boldsymbol{\delta}', \beta', \mu_1, \boldsymbol{\theta}_1', \mu_2, \boldsymbol{\theta}_2', \ldots, \mu_p, \boldsymbol{\theta}_p'\right)'$, $\boldsymbol{\lambda} = \left(\boldsymbol{\eta}', \boldsymbol{\upsilon}'\right)'$, and $\boldsymbol{\upsilon} = \text{vech}(\Upsilon)$. Also, set $\widetilde{L}_N(\boldsymbol{\lambda}) = N^{-1}\sum_{i=1}^{N}\widetilde{l}_i(\boldsymbol{\lambda})$, $\widetilde{\boldsymbol{H}}_N(\boldsymbol{\lambda}) = \partial^2\widetilde{L}_N(\boldsymbol{\lambda})/\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'$, and $\Lambda = \left\{\boldsymbol{\lambda} = \left(\boldsymbol{\eta}', \boldsymbol{\upsilon}'\right)' \in \mathbb{R}^n : \Upsilon \text{ is positive definite}\right\}$.

The maximizer of $\sum_{i=1}^{N}\widetilde{l}_i(\cdot)$ is a maximum-likelihood estimator given normality, but even if the log-likelihood is misspecified — that is, the errors are not normally distributed given $\boldsymbol{x}_i$, nor are they necessarily conditionally homoskedastic — maximizing $\sum_{i=1}^{N}\widetilde{l}_i(\cdot)$ will still yield a consistent and asymptotically normal estimator under suitable conditions. Sufficient conditions are provided in Theorems 3 and 4.

Theorem 3. Suppose C1, C4, and C5 are satisfied. Further assume:

C2′: $Var(\boldsymbol{x}_i)$ is positive definite and does not depend on i, $E(\boldsymbol{x}_i)$ is independent of i, $E\left(\Delta y_{i,-j+2}\right)$ and $E\left(\boldsymbol{x}_i\,\Delta y_{i,-j+2}\right)$ are independent of i $(j = 1, \ldots, p)$, and $Cov(\boldsymbol{x}_i, \Delta\boldsymbol{e}_i) = \boldsymbol{0}$; also,

C3′: $E\left(\widetilde{\boldsymbol{u}}_i\widetilde{\boldsymbol{u}}_i'\right) = \Upsilon_0$ for all i and Υ_0 is a positive definite matrix.

Then the limit $\widetilde{\boldsymbol{H}}(\boldsymbol{\lambda}) = \lim_{N\to\infty}\widetilde{\boldsymbol{H}}_N(\boldsymbol{\lambda})$ exists, and, if $\widetilde{\boldsymbol{H}}_0 = \widetilde{\boldsymbol{H}}(\boldsymbol{\lambda}_0)$ is negative definite, where $\boldsymbol{\lambda}_0 = \left(\boldsymbol{\eta}_0', \boldsymbol{\upsilon}_0'\right)'$ and $\boldsymbol{\upsilon}_0 = \text{vech}(\Upsilon_0)$, there is a compact subset, say $\overline{\Lambda}$, of Λ, with $\boldsymbol{\lambda}_0$ in its interior, and there is a measurable maximizer, $\widehat{\boldsymbol{\lambda}}$, of $\widetilde{L}_N(\cdot)$ in $\overline{\Lambda}$ such that $\widehat{\boldsymbol{\lambda}} \overset{a.s.}{\to} \boldsymbol{\lambda}_0$ $(N \to \infty, T \text{ fixed})$.

Theorem 4. Suppose C1′–C3′, C4, and C5 are satisfied and $\widetilde{\boldsymbol{H}}_0$ is negative definite. Further assume the following condition is met:

C6′: the limit $\widetilde{\mathcal{I}}_0 = \lim_{N\to\infty} N^{-1}\sum_i E\left[\left(\partial\widetilde{l}_i(\boldsymbol{\lambda}_0)/\partial\boldsymbol{\lambda}\right)\left(\partial\widetilde{l}_i(\boldsymbol{\lambda}_0)/\partial\boldsymbol{\lambda}\right)'\right]$ exists and is positive definite.

Then $\sqrt{N}\left(\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0\right) \overset{d}{\to} \mathcal{N}\left(\boldsymbol{0}, \widetilde{\boldsymbol{H}}_0^{-1}\widetilde{\mathcal{I}}_0\widetilde{\boldsymbol{H}}_0^{-1}\right)$ $(N \to \infty, T \text{ fixed})$.

Proof. For proofs of Theorems 3 and 4, see Appendix D.

The linear projection $\Delta y_{i,-j+2}$ on 1 and $\boldsymbol{x}_i$ guarantees that the residual of this linear projection is uncorrelated with the elements of $\Delta\boldsymbol{X}_i$. This is a critical condition for consistent differenced QML estimation. But this condition is also met if we instead used the linear projection of $\Delta y_{i,-j+2}$ on 1 and $\Delta\boldsymbol{x}_i$, where

$\Delta \boldsymbol{x}_i$ is a vector consisting of the distinct elements of $\Delta \boldsymbol{X}_i$. This latter approach generalizes an estimator studied by Hsiao et al. (2002). Hsiao et al. (2002) studied differenced maximum-likelihood estimation of a dynamic panel data model while assuming $p = 1$, individual specific effects, and uncorrelated and conditionally homoskedastic v_{it}s. Moreover, Hsiao et al. (2002) also imposed restrictions on the regressors that were sufficiently strong to guarantee that the conditional mean of Δy_{i1} given $\Delta \boldsymbol{x}_i$ is linear in $\Delta \boldsymbol{x}_i$. The analysis in this section shows that the differenced maximum-likelihood estimator proposed by Hsiao et al. (2002) is consistent and asymptotically normal under much weaker assumptions. It is consistent and asymptotically normal even if the log-likelihood is misspecified and the v_{it}s are conditionally heteroskedastic but unconditionally homoskedastic. Moreover, all that is required of the elements of $\boldsymbol{x}_{it}$ is that they be uncorrelated with the v_{it}s and that the linear projection of Δy_{i1} on 1 and $\Delta \boldsymbol{x}_i$ does not depend on i. On the other hand, the maximum-likelihood estimator Hsiao et al. (2002) proposed will be inconsistent if the v_{it}s are correlated or unconditionally heteroskedastic. The results in this section show that their proposed estimator can be extended to such cases, provided Υ_0 is unrestricted or the restrictions that are imposed on Υ_0 are valid.[4]

4 Computation

4.1 IFGLS procedures

If the error variance-covariance matrix is unrestricted, QML estimates can be easily computed using IFGLS. Consider, for example, calculating QML estimates of the elements of Ω_0 and γ_0. These estimates can be calculated by iterating back and forth between fitting Ω_0 and fitting γ_0. Specifically, $L_N(\cdot)$ is maximized with respect to the elements of Ω, conditional on the current fit of the regression parameters, say γ^c, by the fit $\Omega^+ = \sum_{i=1}^N \boldsymbol{u}_i(\gamma^c)\,\boldsymbol{u}_i(\gamma^c)'/N$. And, after Ω^+ is obtained, $L_N(\cdot)$ is then maximized with respect to γ, conditional on $\Omega = \Omega^+$, which gives the feasible generalized least squares (FGLS) fit:

$$\gamma^+ = \left(\sum_{i=1}^N \boldsymbol{W}_i' \left(\Omega^+\right)^{-1} \boldsymbol{W}_i \right)^{-1} \sum_{i=1}^N \boldsymbol{W}_i' \left(\Omega^+\right)^{-1} \boldsymbol{y}_i. \tag{11}$$

This fit is then made the current fit, γ^c, and new fits Ω^+ and γ^+ are calculated again, and so on, until the sequence of fitted values converges. Calculating QML estimates of λ_0 and Υ_0, based on differenced

[4] For example, the methods described in this section can be specialized to the case where the v_{it}s follow a moving average. Examination of that specific model is beyond the scope of this paper.

observations, is similar when Υ_0 is unrestricted.

IFGLS can also be applied when restrictions on the error variance-covariance matrix are incorporated. Phillips (2010), for example, recommends an IFGLS method for QML estimation of $\boldsymbol{\gamma}_0$. The IFGLS procedure described in Phillips (2010) is applicable when the error-components model in (3) is assumed and $\boldsymbol{\Omega}_0 = \sigma_{a0}^2 \boldsymbol{\iota}\boldsymbol{\iota}' + \sigma_0^2 \boldsymbol{I}$. Implementation is as follows: Given a current fit, $\boldsymbol{\gamma}^c$, the new fit of σ_0^2 is $\left(\sigma^2\right)^+ = \sum_{i=1}^N \boldsymbol{u}_i\left(\boldsymbol{\gamma}^c\right)' \boldsymbol{Q}\boldsymbol{u}_i\left(\boldsymbol{\gamma}^c\right) / \left[(T-1)N\right]$, where $\boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{\iota}\,\boldsymbol{\iota}'/T$, and the new fit of $\sigma_{\bar{u}0}^2 = \sigma_{a0}^2 + \sigma_0^2/T$ is $\left(\sigma_{\bar{u}}^2\right)^+ = \sum_{i=1}^N \boldsymbol{u}_i\left(\boldsymbol{\gamma}^c\right)' \boldsymbol{\iota}\,\boldsymbol{\iota}'\boldsymbol{u}_i\left(\boldsymbol{\gamma}^c\right) / \left(T^2 N\right)$. Given $\left(\sigma^2\right)^+$ and $\left(\sigma_{\bar{u}}^2\right)^+$, a new fit of $\boldsymbol{\Omega}$ is calculated as $\boldsymbol{\Omega}^+ = \boldsymbol{Q}/\left(\sigma^2\right)^+ + \boldsymbol{\iota}\,\boldsymbol{\iota}'/\left[T^2\left(\sigma_{\bar{u}}^2\right)^+\right]$ (see Phillips 2010). Once this variance-covariance matrix fit is in hand, a new fit for $\boldsymbol{\gamma}_0$ is calculated, as in (11), and so on until convergence.

Notice that this algorithm does not rely on calculating an estimate of σ_{a0}^2, and consequently it does not incorporate the restriction $\sigma_{a0}^2 \geq 0$. Therefore, although the estimates of σ_{a0}^2 and $\sigma_{\bar{u}0}^2$ are guaranteed to be non-negative, the implied estimate of $\sigma_{a0}^2 = \sigma_{\bar{u}0}^2 - \sigma_0^2/T$ can be negative. If $\sigma_{a0}^2 > 0$, the probability the implied estimate of σ_{a0}^2 is greater than zero goes to one, as $N \to \infty$, but in small samples a negative estimate of σ_{a0}^2 can occur, and this outcome can lead to poorer sampling performance than if the constraint $\sigma_{a0}^2 \geq 0$ is imposed (see Section 5.3).

An alternative computational strategy, which will not produce negative estimated variance components, is the ECME algorithm. The ECME algorithm is an extension of the expectation-maximization (EM) algorithm. Unlike the EM algorithm, the ECME algorithm relies on conditional or constrained maximization (CM) of either an imputed log-likelihood — constructed using augmented data — or the log-likelihood based on the observed data. In the present application, the observed or "incomplete data" is $\boldsymbol{y} = \left(\boldsymbol{y}_1', \ldots, \boldsymbol{y}_N'\right)'$, while the augmented or "complete data" consists of $\boldsymbol{y}$ and $\boldsymbol{a} = (a_1, \ldots, a_N)'$.[5] The imputed log-likelihood is built during the expectation (E) step by taking the conditional expectation of the log-likelihood for the complete data given the incomplete data, while treating the current fit of the parameters $\boldsymbol{\psi}^c$ as the parameters of the conditional distribution.[6]

Applying the ECME algorithm to an error-components model for which $\boldsymbol{\Omega}_0 = \sigma_{a0}^2 \boldsymbol{\iota}\boldsymbol{\iota}' + \Sigma_0$, with $\Sigma_0 = \operatorname{diag}\left(\sigma_{01}^2, \ldots, \sigma_{0T}^2\right)$, leads to the following E and CM steps:

E-step: Let $\left(\sigma_a^2\right)^c$, $\boldsymbol{\gamma}^c$, and $\boldsymbol{\Omega}^c = \left(\sigma_a^2\right)^c \boldsymbol{\iota}\boldsymbol{\iota}' + \Sigma^c$, with $\Sigma^c = \operatorname{diag}\left(\left(\sigma_1^2\right)^c, \ldots, \left(\sigma_T^2\right)^c\right)$, denote the current

[5] For the purposes of deriving the imputed log-likelihood and the actual log-likelihood, the variables in $\boldsymbol{z} = \left(\boldsymbol{z}_1', \ldots, \boldsymbol{z}_N'\right)'$ are treated as fixed.

[6] Liu and Rubin (1994) describe the properties of the ECME algorithm. For applications of it to panel data see Phillips (2004, 2012).

fits of σ_{a0}^2, γ_0, and Ω_0. Compute the conditional mean and variance of a_i given $\boldsymbol{y}_i$ evaluated at the current fit of the parameters. These are $a_i^c = \left(\sigma_a^2\right)^c \boldsymbol{\iota}' \left(\Omega^c\right)^{-1} \boldsymbol{u}_i\left(\gamma^c\right)$ and $v_a^c = \left(\sigma_a^2\right)^c \left[1 - \left(\sigma_a^2\right)^c \boldsymbol{\iota}' \left(\Omega^c\right)^{-1} \boldsymbol{\iota}\right]$, respectively (see, e.g., Greene 2012, Theorem B.7, pp. 1041-1042). Then the imputed log-likelihood is

$$\begin{aligned} Q\left(\psi\,;\psi^c\right) &= const - \frac{N}{2}\left(\ln\sigma_a^2 + \sum_{t=1}^{T}\ln\sigma_t^2\right) - \frac{1}{2\sigma_a^2}\sum_{i=1}^{N}\left(a_i^c\right)^2 - \frac{N}{2\sigma_a^2}v_a^c \\ &\quad -\frac{1}{2}\sum_{i=1}^{N}\left[\boldsymbol{u}_i\left(\gamma\right) - \boldsymbol{\iota}\, a_i^c\right]' \Sigma^{-1}\left[\boldsymbol{u}_i\left(\gamma\right) - \boldsymbol{\iota}\, a_i^c\right] - \frac{N}{2}\boldsymbol{\iota}'\Sigma^{-1}\boldsymbol{\iota}\, v_a^c. \end{aligned}$$

CM-step 1: Maximize $Q\left(\cdot;\psi^c\right)$ with respect to $\boldsymbol{\omega} = \left(\sigma_a^2, \sigma_1^2, \ldots, \sigma_T^2\right)'$ subject to the constraint $\gamma = \gamma^c$. This step yields $\left(\sigma_a^2\right)^+ = v_a^c + \sum_{i=1}^{N}\left(a_i^c\right)^2/N$ and

$$\left(\sigma_t^2\right)^+ = v_a^c + \frac{1}{N}\sum_{i=1}^{N}\left[u_{it}\left(\gamma^c\right) - a_i^c\right]^2 \qquad t = 1, \ldots, T. \tag{12}$$

CM-step 2: Maximize the actual log-likelihood $\sum_{i=1}^{N} l_i\left(\cdot\right)$ with respect to γ subject to the constraint $\boldsymbol{\omega} = \boldsymbol{\omega}^+$, where $\boldsymbol{\omega}^+ = \left(\left(\sigma_a^2\right)^+, \left(\sigma_1^2\right)^+, \ldots, \left(\sigma_T^2\right)^+\right)'$. This step gives the FGLS fit in Eq. (11) with $\Omega^+ = \left(\sigma_a^2\right)^+ \boldsymbol{\iota}\boldsymbol{\iota}' + \Sigma^+$ and $\Sigma^+ = \mathrm{diag}\left(\left(\sigma_1^2\right)^+, \ldots, \left(\sigma_T^2\right)^+\right)$.

After the new fits of the parameters are obtained, they become the current fits, and the preceding steps are repeated, until convergence.

If the restrictions $\sigma_{0t}^2 = \sigma_0^2$ $(t = 1, \ldots, T)$ are imposed, as in Phillips (2010), then the equations in (12) are replaced with

$$\left(\sigma^2\right)^+ = v_a^c + \frac{1}{NT}\sum_{i=1}^{N}\left[\boldsymbol{u}_i\left(\gamma^c\right) - \boldsymbol{\iota}\, a_i^c\right]'\left[\boldsymbol{u}_i\left(\gamma^c\right) - \boldsymbol{\iota}\, a_i^c\right]. \tag{13}$$

Moreover, Ω^c and Ω^+ are calculated using the structured form $\Omega = \sigma_a^2\boldsymbol{\iota}\boldsymbol{\iota}' + \sigma^2\boldsymbol{I}$.

Unlike some other algorithms, the ECME fitted values for the error variance components are guaranteed to be non-negative. But this advantage can lead to another complication. Specifically, EM-like algorithms — including the ECME algorithm — can be excruciatingly slow to converge, and, when calculating estimates of error-components models, the rate of convergence can slow when the sequence of one of the fitted variance components gets close to zero. Moreover, there is always the possibility that the error-components model in (3) is inappropriate; specifically, there may be no individual-specific effects. In this case, we have $\sigma_{c0}^2 = 0$,

where $\sigma_{c0}^2 = var(c_i)$, and $\sigma_{a0}^2 = 0$, and consequently the sequence of fitted values for σ_{a0}^2 can approach zero. Furthermore, even if σ_{c0}^2 is positive and large, σ_{a0}^2 can be small, for the control function $\mu_0 + \boldsymbol{z}_i'\boldsymbol{\theta}_0$ is the best linear predictor of c_i based on $\boldsymbol{z}_i$, and if that predictor is accurate, then σ_{a0}^2 can be near zero. If so, the sequence of fitted values for σ_{a0}^2 can get close to zero.

As a practical matter, however, if $\boldsymbol{\Omega}_0 = \sigma_{a0}^2 \boldsymbol{\iota}\boldsymbol{\iota}' + \Sigma_0$, with $\Sigma_0 = \text{diag}\left(\sigma_{01}^2, \ldots, \sigma_{0T}^2\right)$, then, when a fitted value for σ_{a0}^2 is near zero, the fitted value $\boldsymbol{\gamma}^+$ in (11) differs little from the weighted least squares fit $\left(\sum_{i=1}^N \boldsymbol{W}_i' \left(\Sigma^+\right)^{-1} \boldsymbol{W}_i\right)^{-1} \sum_{i=1}^N \boldsymbol{W}_i' \left(\Sigma^+\right)^{-1} \boldsymbol{y}_i$, which is obtained by setting $\left(\sigma_a^2\right)^+ = 0$. Furthermore, once $\left(\sigma_a^2\right)^+$ is set to zero, all subsequent fitted values for σ_{a0}^2 will be zero. Also, when $\left(\sigma_a^2\right)^c = 0$, Eq. (12) simplifies to $\left(\sigma_t^2\right)^+ = \sum_{i=1}^N u_{it} \left(\boldsymbol{\gamma}^c\right)^2 / N$, whereas, if we use the ECME algorithm based on the assumption $\sigma_{0t}^2 = \sigma_0^2$ $(t = 1, \ldots, T)$, Eq. (13) simplifies to $\left(\sigma^2\right)^+ = \sum_{i=1}^N \boldsymbol{u}_i \left(\boldsymbol{\gamma}^c\right)' \boldsymbol{u}_i \left(\boldsymbol{\gamma}^c\right) / (NT)$ when $\left(\sigma_a^2\right)^c = 0$. Regardless of whether or not the constraints $\sigma_{0t}^2 = \sigma_0^2$ $(t = 1, \ldots, T)$ are imposed, if $\left(\sigma_a^2\right)^+$ is set to zero, convergence is rapid. Consequently, the ECME algorithm for computing QML estimates based on augmenting the regression model with $\mu_0 + \boldsymbol{z}_i'\boldsymbol{\theta}_0$ will generally converge at a robust rate if, as part of the convergence criterion, the size of the fitted value for σ_{a0}^2 is evaluated and $\left(\sigma_a^2\right)^+$ is set to zero should it become sufficiently small.

Finally, note that if we assume an error-components model with unconditionally homoskedastic v_{it}s and $\left(\sigma_a^2\right)^+$ is set to zero, then the fit of $\boldsymbol{\gamma}_0$ becomes the OLS estimator $\left(\sum_{i=1}^N \boldsymbol{W}_i' \boldsymbol{W}_i\right)^{-1} \sum_{i=1}^N \boldsymbol{W}_i' \boldsymbol{y}_i$. There are two observations to make about this outcome. First, if $\sigma_{a0}^2 > 0$, then the probability of this outcome occurring goes to zero as $N \to \infty$. And, second, even when σ_{a0}^2 is not zero, if it is close to zero, using the OLS estimator of $\boldsymbol{\gamma}_0$ is not such a bad idea. This is because the smaller σ_{a0}^2 is, the smaller the correlation in the errors, and correlation in the errors is the source of the inconsistency in the OLS estimator. Hence, the smaller σ_{a0}^2 is, the more the inconsistency in OLS is mitigated.

4.2 Computation of restricted differenced QML estimates

The ECME algorithm is well-suited to calculating QML estimates for error-components models. But if we assume the error-components model in (3) and difference observations, then the differenced errors no longer have an error-components structure. In this case we must resort to other iterative algorithms; typically, gradient methods.

If a gradient method is used, a reparameterization may be helpful. This is because, although EM-like methods guarantee fitted variance-covariance matrices that are always non-negative definite, gradient

methods do not always have this property. A reparameterization, however, can often make a gradient method better behaved. For example, suppose $p = 1$, e_{it} is given by the error-components model in (3), the v_{it}s are uncorrelated and unconditionally homoskedastic, and the regressors in $\boldsymbol{x}_{it}$ are strictly exogenous with respect to the v_{it}s. Furthermore, assume Δy_{i1} is generated by the same process generating Δy_{it} for $t \geq 2$. Under these conditions, it is easy to show that the error variance-covariance matrix is $\Upsilon_0 = \sigma_0^2 \Phi_0$, with

$$\Phi_0 = \begin{pmatrix} \phi_0 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ 0 & -1 & 2 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & -1 \\ 0 & 0 & \cdots & -1 & 2 \end{pmatrix} \tag{14}$$

(cf Hsiao et al. 2002, p. 110, Eq. (3.2)). From the determinant $\left|\sigma_0^2 \Phi_0\right| = \sigma_0^{2T} \left[1 + T\left(\phi_0 - 1\right)\right]$ (see, e.g., Hsiao et al 2002, p. 111, Eq. (3.7)) we see that, in order to ensure a positive definite fitted value for $\sigma_0^2 \Phi_0$, we must search over values of ϕ satisfying $\phi > 1 - 1/T$. This restriction is guaranteed if we set $\varpi = \ln\left(\phi - 1 + 1/T\right)$ and maximize the log-likelihood

$$const - \frac{NT}{2} \ln\left(\sigma^2\right) - \frac{N\varpi}{2} - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \widetilde{\boldsymbol{u}}_i\left(\boldsymbol{\eta}\right)' \Phi^{-1} \widetilde{\boldsymbol{u}}_i\left(\boldsymbol{\eta}\right)$$

with respect to $\boldsymbol{\eta}$, σ^2, and ϖ. Here Φ has $\exp\left(\varpi\right) + 1 - 1/T$ in its first row, first column and everywhere else is the same as Φ_0 in (14).

5 Monte Carlo Experiments

5.1 Design

In order to assess the finite sampling properties of QML estimators described in Section 4, Monte Carlo experiments were conducted. For all of the experiments, observations on the dependent variable y_{it} were generated according to the model

$$y_{it} = \delta_0 y_{i,t-1} + 0.5 x_{it} + c_i + v_{it} \qquad \left(t = -t_0 + 1, \ldots, T, \;\; i = 1, \ldots, N\right),$$

with $y_{i,-t_0} = 0$. The values for δ_0 considered were 0.4 and 0.9. Moreover, the x_{it}s were generated according to the autoregressive process

$$x_{it} = 0.5 + 0.5x_{i,t-1} + \xi_{it} \qquad (t = -t_0 + 1, \ldots, T, \;\; i = 1, \ldots, N).$$

The starting value $x_{i,-t_0}$ was set equal to $5 + 10\xi_{i,-t_0}$ and the ξ_{it}s were generated as independent uniform random variates with mean zero and variance one.

As for the v_{it}s, they were generated as $v_{it} = x_{it}^{\kappa}(\epsilon_{it} - 5)/\sqrt{10}$, with ϵ_{it} a chi-square random variate with five degress of freedom. The variate $(\epsilon_{it} - 5)/\sqrt{10}$ has an asymmetric distribution about zero with a variance of one. Moreover, because the ϵ_{it}s were generated independently of one another and of the x_{it}s, the v_{it}s were uncorrelated but conditionally heteroskedastic depending on the value of κ. The values of κ considered were zero, for conditional homoskedasticity, and one, in which case the idiosyncratic errors were conditionally heteroskedastic.

On the other hand, the heterogeneity component, c_i, was generated as $c_i = \sum_{t=0}^{T} \ln|x_{it}|/(T+1) + \sigma_\zeta \zeta_i$, with ζ_i a standard normal variate and σ_ζ set to either one or two. This specification for c_i induced correlation between c_i and the x_{it}s.

Finally, two values for t_0 were considered: one and 50. And, after a sample was generated, the start up observations were discarded so that QML estimation was based on $(x_{i1}, y_{i1}), \ldots, (x_{iT}, y_{iT})$ and y_{i0} $(i = 1, \ldots, N)$, while GMM estimation was based on $(x_{i0}, y_{i0}), \ldots, (x_{iT}, y_{iT})$. Furthermore, T was set to five, and N was set to 100 or 500. And, for each combination of parameters, 5,000 independent samples were generated.

Table 1 lists the sample designs.

(Table 1 here)

5.2 Estimators

The finite sample properties of several QML estimators were compared to each other and two well-known GMM estimators. The GMM estimators considered were the differenced GMM estimator proposed by Arellano and Bond (1991) (denoted DGMM) and the system GMM estimator suggested by Blundell and Bond (1998) (SGMM).

The sampling performance of three unrestricted QML estimators was also investigated. One of these unrestricted QML estimators was the levels QML estimator based on regression augmentation that imposes no restrictions on the error variance-covariance matrix. The results for this estimator are denoted by QML. The other two unrestricted QML estimators were differenced QML estimators that impose no restrictions on the error variance-covariance matrix. For differenced QML, we can exploit the linear projection of Δy_{i1} on 1 and either $\Delta \boldsymbol{x}_i$ or $\boldsymbol{x}_i$. Results for both cases are provided. The results using $\Delta \boldsymbol{x}_i$ are denoted by $\text{DQML}_{\Delta \boldsymbol{x}}$, while those for $\boldsymbol{x}_i$ are denoted by $\text{DQML}_{\boldsymbol{x}}$.

Results are also provided for levels QML estimation while relying on the structured variance-covariance matrix $\boldsymbol{\Omega}_0 = \sigma_{a0}^2 \boldsymbol{\iota}\boldsymbol{\iota}' + \Sigma_0$ with $\Sigma_0 = \text{diag}\big(\sigma_{01}^2, \ldots, \sigma_{0T}^2\big)$. For this case, estimates were calculated with the ECME algorithm (see Section 4.1), and, therefore, the results for this estimator are denoted by ECME_{he}.

The tables also provide results for levels QML estimation while imposing the restriction $\boldsymbol{\Omega}_0 = \sigma_{a0}^2 \boldsymbol{\iota}\boldsymbol{\iota}' + \sigma_0^2 \boldsymbol{I}$. The QML estimates for this case were calculated two ways: with the IFGLS procedure described in Phillips (2010) and the EMCE algorithm for unconditionally homoskedastic v_{it}s (see Section 4.1). In the tables, estimates calculated with the IFGLS procedure described in Phillips (2010) are indicated as IFGLS_{ho}, while estimates calculated with the ECME algorithm are denoted as ECME_{ho}.

I also calculated differenced QML estimates that restrict the v_{it}s to be uncorrelated and homoskedastic (see Section 4.2). Because we can use either a linear projection of Δy_{i1} on 1 and $\Delta \boldsymbol{x}_i$ or a linear projection of Δy_{i1} on 1 and $\boldsymbol{x}_i$, results for both choices are reported and are denoted by $\text{DQML}_{\Delta \boldsymbol{x},\text{ho}}$ and $\text{DQML}_{\boldsymbol{x},\text{ho}}$.

Finally, as noted in Section 4.1, when σ_{a0}^2 is near zero, the bias in the OLS estimator of $\boldsymbol{\gamma}_0$ may be small. To investigate this possibility, results for OLS applied to the augmented model in (6) are also reported for some Monte Carlo experiments.

5.3 Results

5.3.1 Stationary Designs

Table 2 provides estimates of finite sample bias and root mean squared error for Designs 1 through 8. For these designs the generated variables were approximately stationary ($t_0 = 50$). The table also provides some statistics.

(Table 2 here)

Table 2 provides information about the average correlation coefficient $\rho = 2\sum_{s=1}^{T-1}\sum_{t>s}^{T}\rho_{st}/[T(T-1)]$, where $\rho_{st} = \sigma_{a0}^2/\left[\left(\sigma_{a0}^2+\sigma_{0s}^2\right)\left(\sigma_{a0}^2+\sigma_{0t}^2\right)\right]^{1/2}$. Estimates of the average correlation coefficient ρ are of interest because the closer ρ is to zero, that is, the less correlation in the augmented regression errors, then the better OLS — when applied to the augmented regression — will perform in terms of finite sample bias (see Section 4). For each sample, ρ was estimated with $\widehat{\rho} = 2\sum_{s=1}^{T-1}\sum_{t>s}^{T}\widehat{\rho}_{st}/[T(T-1)]$, where $\widehat{\rho}_{st} = \widehat{\sigma}_{a0}^2/\left[\left(\widehat{\sigma}_{a0}^2+\widehat{\sigma}_{0s}^2\right)\left(\widehat{\sigma}_{a0}^2+\widehat{\sigma}_{0t}^2\right)\right]^{1/2}$. The estimates $\widehat{\sigma}_{a0}^2$ and $\widehat{\sigma}_{0t}^2$ $(t=1,\ldots,T)$ were obtained using the ECME_{he} algorithm (see Section 5.2). Table 2 provides, for each design, the average of the $\widehat{\rho}$ estimates over 5,000 samples.[7]

Table 2 also provides information about how much of the variation in the individual effect is removed by the linear projection $\mu_0 + \boldsymbol{z}_i'\boldsymbol{\theta}_0$. Specifically, if the errors are generated by an error-components model, we can define a pseudo R^2 given by $R_c^2 = 1 - \widehat{\sigma}_{a0}^2/\widehat{\sigma}_{c0}^2$, where $\widehat{\sigma}_{a0}^2$ is an estimator of σ_{a0}^2, $\widehat{\sigma}_{c0}^2 = \widehat{\theta}'N^{-1}\sum_{i=1}^{N}(\boldsymbol{z}_i-\overline{\boldsymbol{z}})(\boldsymbol{z}_i-\overline{\boldsymbol{z}})'\widehat{\theta}+\widehat{\sigma}_{a0}^2$, and $\overline{\boldsymbol{z}} = \sum_{i=1}^{N}\boldsymbol{z}_i/N$. Note that $\sigma_{c0}^2 = \theta_0'Var(\boldsymbol{z}_i)\theta_0+\sigma_{a0}^2$, and thus $\widehat{\sigma}_{c0}^2$ is a consistent estimator of $\widehat{\sigma}_{c0}^2$ if $\widehat{\theta}$ and $\widehat{\sigma}_{a0}^2$ are consistent estimators. Furthermore, we have $\widehat{\sigma}_{c0}^2 \geq \widehat{\sigma}_{a0}^2 \geq 0$. Hence, $R_c^2 = 1-\widehat{\sigma}_{a0}^2/\widehat{\sigma}_{c0}^2$ is always between 0 and 1, and, therefore, it can be interpreted as an estimate of the proportion of the variation in c_i that is controlled for by the control function $\mu_0+\boldsymbol{z}_i'\boldsymbol{\theta}_0$. Table 2 gives, for each design, the average of the R_c^2s over the samples, where $\widehat{\sigma}_{a0}^2$ and $\widehat{\theta}$ were calculated using the ECME_{he} algorithm.

The evidence in Table 2 shows that the unrestricted QML estimators — QML, $\text{DQML}_{\Delta\boldsymbol{x}}$, and $\text{DQML}_{\boldsymbol{x}}$ — generally have neglible finite sample bias. As for root mean squared error, there is an unambiguous ranking among these three estimators. Specifically, for each design, the levels QML estimator typically has a smaller standard deviation, and hence root mean squared error, than both of the differenced QML estimators. Moreover, when comparing the differenced QML estimators, the $\text{DQML}_{\boldsymbol{x}}$ estimator usually has the smaller — sometimes much smaller — standard deviation and root mean squared error. For Designs 5, 6, 7, and 8, for example, the root mean squared error of the $\text{DQML}_{\Delta\boldsymbol{x}}$ estimator is about 74, 76, 50, and 42 percent larger than that of the $\text{DQML}_{\boldsymbol{x}}$ estimator.

The GMM estimators have more finite sample bias than the unrestricted QML estimators; in some cases, by several orders of magnitude. For example, the finite sample biases of the differenced GMM estimator are so large when $\delta_0 = 0.9$ (Designs 5–8) that it has larger root mean squared errors than the QML and $\text{DQML}_{\boldsymbol{x}}$

[7] In Section 4.1 it was noted that ECME iterations can be slow to converge if the fitted values for σ_{a0}^2 become small. The estimate of ρ was used to measure the relative size of σ_{a0}^2. When the estimate of ρ became less than 0.01, the fit of σ_{a0}^2 was set to zero (see Section 4.1).

estimators. On the other hand, all of the unrestricted QML estimators are, in terms of root mean squared errors, inferior to the differenced GMM estimator when $\delta_0 = 0.4$. Furthermore, all of the unrestricted QML estimators are dominated by the system GMM estimator in terms of sampling efficiency.

The unrestricted QML estimators do not exploit as much information as the GMM estimators; specifically, they do not exploit the restriction that the v_{it}s are uncorrelated whereas the GMM estimators rely on this restriction. The evidence in Table 2 shows that exploiting valid restrictions on Σ_0 can yield substantial reductions in the sampling variability of QML estimators. The restricted QML estimators — ECME_{he}, ECME_{ho}, IFGLS_{ho}, $\text{DQML}_{\Delta\boldsymbol{x},\text{ho}}$, and $\text{DQML}_{\boldsymbol{x},\text{ho}}$ — all exploit the fact that the v_{it}s are uncorrelated, and these estimators generally have significantly smaller root mean squared errors than the unrestricted QML estimators.

A comparison of the restricted QML estimators leads to rankings that are similar to how the unrestricted QML estimators are ranked. For example, just as the $\text{DQML}_{\boldsymbol{x}}$ estimator generally has smaller root mean squared error than the $\text{DQML}_{\Delta\boldsymbol{x}}$ estimator, the $\text{DQML}_{\boldsymbol{x},\text{ho}}$ estimator is often more efficient than the $\text{DQML}_{\Delta\boldsymbol{x},\text{ho}}$ estimator. They have comparable root mean squared errors for $\delta_0 = 0.4$, but $\text{DQML}_{\boldsymbol{x},\text{ho}}$ is more efficient when $\delta_0 = 0.9$. Moreover, for $\delta_0 = 0.9$, the restricted levels QML estimators, calculated with the ECME algorithm (ECME_{he} and ECME_{ho}), often provide substantial reductions in root mean squared error compared to the restricted differenced QML estimators ($\text{DQML}_{\Delta\boldsymbol{x},\text{ho}}$ and $\text{DQML}_{\boldsymbol{x},\text{ho}}$).

The restricted QML estimators often compare favorably to the GMM estimators. For all designs in Table 2, all of the restricted QML estimators — in levels and differences — have smaller root mean squared errors than the differenced GMM estimator. Furthermore, they also have smaller root mean squared errors than the system GMM estimator when $\delta_0 = 0.4$. On the other hand, for $\delta_0 = 0.9$, the relative ranking between the restricted QML estimators and the system GMM estimator is not so clear cut. For $\delta_0 = 0.9$, the system GMM estimator dominates the IFGLS_{ho}, $\text{DQML}_{\boldsymbol{x},\text{ho}}$, and $\text{DQML}_{\Delta\boldsymbol{x},\text{ho}}$ estimators in terms of root mean squared error, but not the ECME_{he} and ECME_{ho} estimators. The latter estimators have smaller root mean squared errors than the system GMM estimator for Designs 5 and 6, while the system GMM estimator has the smaller root mean squared errors for Designs 7 and 8.

Possibly more surprising is the relative ranking between the GMM estimators and OLS when the latter is applied to the augmented model. For Designs 5 through 8 the ranking is unambiguous: the OLS estimator dominates both GMM estimators in terms of root mean squared error. (For Designs 1 through 4, this ranking is reversed, however.) Indeed, for Designs 7 and 8, the OLS estimator has the smallest root mean squared

error among all estimators considered. This result is due to the low level of correlation among the errors in the augmented model for Designs 5 through 8. For these designs, the average value for $\widehat{\rho}$ is not greater than 0.10. Note also the high values for R_c^2 for these designs. These R_c^2s indicate the control function eliminates much of the variation in the individual effect, and because the remaining variation is small relative to the variation in the idiosyncratic errors, the v_{it}s, the control function effectively eliminates most of the correlation in the augmented model regression errors.

Whereas the sampling performance of OLS applied to the augmented model improves when $\widehat{\rho}$ is near zero, the sampling performance of the IFGLS_{ho} estimator proposed in Phillips (2010) deteriorates. This is because when $\widehat{\rho}$ is near zero, σ_{a0}^2 is small, and when σ_{a0}^2 is small, the IFGLS algorithm proposed in Phillips (2010) can wander outside of the relevant parameter space — that is, it can converge to a fit implying a negative estimate of σ_{a0}^2. A consequence of this fact is a deterioration in sampling performance. For example, for Design 2, when $\widehat{\rho}$ is largest, the root mean squared errors of the IFGLS_{ho} and ECME_{ho} estimators are nearly the same — a result one would expect when the two algorithms are generally locating the same maximizer of the quasi log-likelihood. On the other hand, for Designs 5 through 8, for which $\widehat{\rho}$ is small, the root mean squared errors of the IFGLS_{he} estimator are much larger than those of the ECME_{ho} estimator.

Table 3 provides estimates of finite sample bias and root mean squared error for Designs 1 through 8 for a larger number of cross sections ($N = 500$). The evidence in this table indicates that many of the relative rankings of the estimators are unaffected by this increase in sample size. Specifically, QML estimation using observations in levels rather than differences generally leads to an improvement in sampling precision. Compare, for example, QML to $\text{DQML}_{\boldsymbol{x}}$ and $\text{DQML}_{\Delta\boldsymbol{x}}$, and compare ECME_{he}, ECME_{ho}, and IFGLS_{ho} to $\text{DQML}_{\boldsymbol{x},\text{ho}}$ and $\text{DQML}_{\Delta\boldsymbol{x},\text{ho}}$. Second, upon comparing $\text{DQML}_{\boldsymbol{x}}$ to $\text{DQML}_{\Delta\boldsymbol{x}}$ and $\text{DQML}_{\boldsymbol{x},\text{ho}}$ to $\text{DQML}_{\Delta\boldsymbol{x},\text{ho}}$, we see that if differences are used, then estimating a system of equations that includes the linear projection of Δy_{i1} on 1 and $\boldsymbol{x}_i$ is rarely worse and often better than using the linear projection of Δy_{i1} on 1 and $\Delta\boldsymbol{x}_i$. Third, imposing valid restrictions on the error variance-covariance matrix leads to often significant improvements in relative efficiency. Compare ECME_{he}, ECME_{ho}, and IFGLS_{ho} to QML, compare $\text{DQML}_{\boldsymbol{x},\text{ho}}$ to $\text{DQML}_{\boldsymbol{x}}$, and compare $\text{DQML}_{\Delta\boldsymbol{x},\text{ho}}$ to $\text{DQML}_{\Delta\boldsymbol{x}}$. Fourth, the unrestricted QML and $\text{DQML}_{\boldsymbol{x}}$ estimators are still more accurate, in terms of root mean squared error, than the differenced GMM estimator when $\delta_0 = 0.9$. Fifth, the restricted QML estimators generally, though not always, have smaller root mean squared errors than the GMM estimators. And, finally, system GMM and OLS applied to the augmented model still perform comparably in terms of root mean squared error when $\widehat{\rho}$ is near zero

($\delta_0 = 0.9$).

(Table 3 here)

On the other hand, increasing the number of cross sections sampled does affect some relative rankings. For example, given the OLS estimator is inconsistent, it is not surprising that it no longer dominates all of the estimators in terms of root mean squared error for Designs 7 and 8. That distinction now goes to the ECME_{he} estimator. Moreover, although estimates calculated with the ECME_{ho} algorithm still improve on those calculated with the IFGLS procedure recommended in Phillips (2010) when $\delta_0 = 0.9$, the sampling performance of these two estimators is quite similar for $\delta_0 = 0.4$. Furthermore, system GMM no longer dominates all of the unrestricted QML estimators. For $\delta_0 = 0.4$, it still has smaller root mean squared errors, but, for $\delta_0 = 0.9$, it no longer always has a smaller root mean squared error than the unrestricted QML estimators. For example, the QML estimator has a smaller root mean squared error for Designs 5, 6, and 8. Finally, although differenced GMM has significantly larger root mean squared errors than system GMM for all designs for which $\delta_0 = 0.9$ when $N = 100$, when the number of cross sections is increased to $N = 500$, the finite sample bias of the differenced GMM estimator falls in magnitude relative to that of the system GMM estimator by enough that the latter estimator has significantly larger root mean squared errors for Designs 5 and 6 when $N = 500$.

5.3.2 Nonstationary Designs

Tables 4 and 5 provide finite sample bias and root mean squared error estimates for Designs 9 through 16. For these designs $t_0 = 1$, and, therefore, for each cross section, the time series began in the immediate past. Consequently, none of the time series are stationary.

(Tables 4 and 5 here)

Tables 4 and 5 drop two estimators from the analysis: OLS and system GMM. The system GMM estimator is omitted, because, given $t_0 = 1$, the designs violate the initial condition restriction upon which the system GMM estimator relies. In particular, not all of the instruments used in constructing the estimator are valid for these designs. Moreover, OLS is also inconsistent for all designs in Tables 4 and 5, and for no design is $\widehat{\rho}$ close to zero. Consequently, OLS performed, as expected, poorly for all of the designs reported on in Tables 4 and 5.

Several other estimators are also inconsistent for some of the designs in Tables 4 and 5. Specifically, the estimators ECME_{ho}, IFGLS_{ho}, $\text{DQML}_{\boldsymbol{x},\text{ho}}$, and $\text{DQML}_{\Delta\boldsymbol{x},\text{ho}}$ are inconsistent for Designs 11, 12, 15, and 16. This is because these estimators rely on the assumption the errors are unconditionally homoskedastic, but, for Designs 11, 12, 15, and 16, the errors are conditionally heteroskedastic and given they are also nonstationary, they are also unconditionally heteroskedastic. On the other hand, for Designs 9, 10, 13, and 14, the errors are conditionally homoskedastic, and hence unconditionally homoskedastic, and, therefore, the estimators ECME_{ho}, IFGLS_{ho}, $\text{DQML}_{\boldsymbol{x},\text{ho}}$, and $\text{DQML}_{\Delta\boldsymbol{x},\text{ho}}$ are consistent for these designs. In order to see how these estimators perform for these designs and how their performance deteroriates when the errors are unconditionally heteroskedastic, they were included in the experiments.

The main message of Tables 4 and 5 is conveyed by the performance of the ECME_{he} estimator. In terms of sampling efficiency, this estimator arguably outperforms all of the other estimators considered. Like the differenced GMM estimator, it incorporates valid restrictions on the error generating process (e.g., uncorrelated v_{it}s), while not incorporating invalid restrictions (e.g., unconditional homoskedasticity). But it has much less finite sample bias than the differenced GMM estimator. The ECME_{ho}, IFGLS_{ho}, $\text{DQML}_{\boldsymbol{x},\text{ho}}$, and $\text{DQML}_{\Delta\boldsymbol{x},\text{ho}}$ estimators all have among the smallest root mean squared errors for Designs 9, 10, 13, and 14, designs for which the errors are unconditionally homoskedastic, but, not surprisingly, their sampling performance deteriorates when the errors are unconditionally heteroskedastic (Designs 11, 12, 15, and 16).

6 Conclusion

This paper established the almost sure convergence and asymptotic normality of levels and differenced QML estimators of the parameters of a pth-order dynamic panel data model. The consistency and asymptotic normality of the estimators do not depend on initial conditions and the log-likelihood can be misspecified. Furthermore, the error variance-covariance matrix can be of a general form. Models with structured error variance-covariance matrices are special cases of the model examined here.

However, robustness with respect to the specification of the error variance-covariance matrix comes at a price: efficiency losses. The Monte Carlo results provided in Section 5.3 indicate that imposing valid restrictions on the error variance-covariance matrix can lead to substantial improvements in finite sample estimation efficiency. Conversely, exploiting invalid restrictions on the error variance-covariance matrix produces an inconsistent QML estimator.

These results were not surprising. But the Monte Carlo experiments illustrated other lessons as well, some of which were unexpected. For example, if the control function exploited by levels QML estimation controls for much of the variation in unobserved cross-sectional effects so that the remaining correlation in the augmented regression errors is small, then OLS applied to the augmented regression model can produce estimates that compare favorably, in terms of precision, to both QML and GMM estimates. Furthermore, if idiosyncratic errors are uncorrelated and one does not have a good reason to believe the individual effects are fixed, little appears to be gained by differencing and much may be lost. In particular, for the sample designs reported on here, QML estimation based on regression augmentation never performed significantly worse than differenced QML estimation and sometimes performed much better. Furthermore, if differenced QML estimation is used, the Monte Carlo evidence provided here indicates estimating a system of equations that includes a linear projection of Δy_{i1} on 1 and $\boldsymbol{x}_i$ is preferable to using a linear projection of Δy_{i1} on 1 and $\Delta \boldsymbol{x}_i$.

Appendix A: Lemma 1 Proof

In order to establish $E\left(\boldsymbol{y}_{i,-j}'\Omega_0^{*-1}\boldsymbol{e}_i\right)=0$, I first use an analysis similar to that in Hamilton (1994, pp. 7-9). Let $\boldsymbol{\xi}_{it}=\left(y_{it},y_{i,t-1},\ldots,y_{i,t-p+1}\right)'$, $\boldsymbol{\varsigma}_{it}=\left(\boldsymbol{x}_{it}'\boldsymbol{\beta}_0+e_{it},0,\ldots,0\right)'$, and

$$\boldsymbol{F}=\begin{pmatrix}\delta_{01} & \delta_{02} & \cdots & \delta_{0,p-1} & \delta_{0p}\\ 1 & 0 & \cdots & 0 & 0\\ 0 & 1 & \cdots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \cdots & 1 & 0\end{pmatrix}, \tag{15}$$

where $\boldsymbol{\delta}_0=\left(\delta_{01},\ldots,\delta_{0p}\right)'$. Then $\boldsymbol{\xi}_{it}=\boldsymbol{F}\boldsymbol{\xi}_{i,t-1}+\boldsymbol{\varsigma}_{it}$. Hence, $\boldsymbol{\xi}_{i1}=\boldsymbol{F}\boldsymbol{\xi}_{i0}+\boldsymbol{\varsigma}_{i1}$, and, for $t>1$, by repeated substitutions we get $\boldsymbol{\xi}_{it}=\boldsymbol{F}^t\boldsymbol{\xi}_{i0}+\boldsymbol{F}^{t-1}\boldsymbol{\varsigma}_{i1}+\boldsymbol{F}^{t-2}\boldsymbol{\varsigma}_{i2}+\cdots+\boldsymbol{F}\boldsymbol{\varsigma}_{i,t-1}+\boldsymbol{\varsigma}_{it}$. Writing this last expression out in full, we have

$$\begin{aligned}\begin{pmatrix}y_{it}\\ y_{i,t-1}\\ \vdots\\ y_{i,t-p+1}\end{pmatrix} &= \boldsymbol{F}^t\begin{pmatrix}y_{i0}\\ y_{i,-1}\\ \vdots\\ y_{i,-p+1}\end{pmatrix}+\boldsymbol{F}^{t-1}\begin{pmatrix}\boldsymbol{x}_{i1}'\boldsymbol{\beta}_0+e_{i1}\\ 0\\ \vdots\\ 0\end{pmatrix}+\boldsymbol{F}^{t-2}\begin{pmatrix}\boldsymbol{x}_{i2}'\boldsymbol{\beta}_0+e_{i2}\\ 0\\ \vdots\\ 0\end{pmatrix}\\ &\quad+\cdots+\boldsymbol{F}\begin{pmatrix}\boldsymbol{x}_{i,t-1}'\boldsymbol{\beta}_0+e_{i,t-1}\\ 0\\ \vdots\\ 0\end{pmatrix}+\begin{pmatrix}\boldsymbol{x}_{it}'\boldsymbol{\beta}_0+e_{it}\\ 0\\ \vdots\\ 0\end{pmatrix}.\end{aligned} \tag{16}$$

Next let $f_{rs}^{(t)}$ denote the (r,s)th element of $\boldsymbol{F}^t$. Then $y_{i1}=f_{11}^{(1)}y_{i0}+f_{12}^{(1)}y_{i,-1}+\cdots+f_{1p}^{(1)}y_{i,-p+1}+\boldsymbol{x}_{i1}'\boldsymbol{\beta}_0+e_{i1}$, and, for $t>1$, from the first equation in (16) we see that

$$\begin{aligned}y_{it} &= f_{11}^{(t)}y_{i0}+f_{12}^{(t)}y_{i,-1}+\cdots+f_{1p}^{(t)}y_{i,-p+1}+f_{11}^{(t-1)}\left(\boldsymbol{x}_{i1}'\boldsymbol{\beta}_0+e_{i1}\right)\\ &\quad+f_{11}^{(t-2)}\left(\boldsymbol{x}_{i2}'\boldsymbol{\beta}_0+e_{i2}\right)+\cdots+f_{11}^{(1)}\left(\boldsymbol{x}_{i,t-1}'\boldsymbol{\beta}_0+e_{i,t-1}\right)+\boldsymbol{x}_{it}'\boldsymbol{\beta}_0+e_{it}.\end{aligned} \tag{17}$$

Using the expression for y_{it} in Eq. (17), we can write $\boldsymbol{y}_{i,-j}$ in terms of $\boldsymbol{y}_i^o$, $\boldsymbol{X}_i$, and $\boldsymbol{e}_i$. To that end, let

$\boldsymbol{A}_j$ and $\boldsymbol{B}_j$ be $T \times p$ and $T \times T$ matrices given by

$$\boldsymbol{A}_j = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ f_{11}^{(1)} & f_{12}^{(1)} & \cdots & f_{1,j-1}^{(1)} & f_{1j}^{(1)} & f_{1,j+1}^{(1)} & \cdots & f_{1p}^{(1)} \\ f_{11}^{(2)} & f_{12}^{(2)} & \cdots & f_{1,j-1}^{(2)} & f_{1j}^{(2)} & f_{1,j+1}^{(2)} & \cdots & f_{1p}^{(2)} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ f_{11}^{(T-j)} & f_{12}^{(T-j)} & \cdots & f_{1,j-1}^{(T-j)} & f_{1j}^{(T-j)} & f_{1,j+1}^{(T-j)} & \cdots & f_{1p}^{(T-j)} \end{pmatrix} \tag{18}$$

$$\boldsymbol{B}_j = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ f_{11}^{(1)} & 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ f_{11}^{(2)} & f_{11}^{(1)} & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ f_{11}^{(T-j-1)} & f_{11}^{(T-j-2)} & f_{11}^{(T-j-3)} & \cdots & f_{11}^{(1)} & 1 & 0 & \cdots & 0 \end{pmatrix}. \tag{19}$$

Given these definitions, we have $\boldsymbol{y}_{i,-j} = \boldsymbol{A}_j \boldsymbol{y}_i^o + \boldsymbol{B}_j \left(\boldsymbol{X}_i \boldsymbol{\beta}_0 + \boldsymbol{e}_i\right)$.

Therefore, $E\left(\boldsymbol{y}_{i,-j}' \Omega_0^{*-1} \boldsymbol{e}_i\right) = E\left(\boldsymbol{y}_i^{o\prime} \boldsymbol{A}_j' \Omega_0^{*-1} \boldsymbol{e}_i\right) + E\left(\boldsymbol{\beta}_0' \boldsymbol{X}_i' \boldsymbol{B}_j' \Omega_0^{*-1} \boldsymbol{e}_i\right) + E\left(\boldsymbol{e}_i' \boldsymbol{B}_j' \Omega_0^{*-1} \boldsymbol{e}_i\right)$. Note that $E\left(\boldsymbol{e}_i' \boldsymbol{B}_j' \Omega_0^{*-1} \boldsymbol{e}_i\right) = E\left[\text{tr}\left(\Omega_0^{*-1} \boldsymbol{e}_i \boldsymbol{e}_i' \boldsymbol{B}_j'\right)\right] = \text{tr}\left[\Omega_0^{*-1} E\left(\boldsymbol{e}_i \boldsymbol{e}_i'\right) \boldsymbol{B}_j'\right] = \text{tr}\left(\boldsymbol{B}_j'\right) = 0$, where the last equality follows from the fact that $\boldsymbol{B}_j$ is a square matrix with zeros down the main diagonal. Moreover, if $E\left(\boldsymbol{e}_i \boldsymbol{x}_i'\right) = \boldsymbol{0}$, then $E\left(\boldsymbol{\beta}_0' \boldsymbol{X}_i' \boldsymbol{B}_j' \Omega_0^{*-1} \boldsymbol{e}_i\right) = 0$. And $E\left(\boldsymbol{y}_i^{o\prime} \boldsymbol{A}_j' \Omega_0^{*-1} \boldsymbol{e}_i\right) = \text{tr}\left[E\left(\boldsymbol{e}_i \boldsymbol{y}_i^{o\prime}\right) \boldsymbol{A}_j' \Omega_0^{*-1}\right] = 0$ given $E\left(\boldsymbol{e}_i \boldsymbol{y}_i^{o\prime}\right) = \boldsymbol{0}$. The preceding proves $E\left(\boldsymbol{y}_{i,-j}' \Omega_0^{*-1} \boldsymbol{e}_i\right) = 0$.

Appendix B: Theorem 1 Proof

The proof of Theorem 1 relies on verifying several preliminary results, which are provided as Lemmas B.1 through B.4. Throughout convergence is with respect to $N \to \infty$, with T fixed. Moreover, in the

sequel, M denotes a sufficiently large finite number.

Lemma B.1. Suppose $E\left(x_{itk}^2\right) < \infty$ and $E\left(y_{it}^2\right) < \infty$, for each i, t, and k, and Conditions C2 and C4 are satisfied. Then the linear projection in (4) exists. Furthermore, the limits $L(\boldsymbol{\psi}) = \lim_{N\to\infty} E[L_N(\boldsymbol{\psi})]$ and $\boldsymbol{H}(\boldsymbol{\psi}) = \lim_{N\to\infty} E[\boldsymbol{H}_N(\boldsymbol{\psi})]$ exist, and $L(\boldsymbol{\psi})$ and the elements of $\boldsymbol{H}(\boldsymbol{\psi})$ are continuous functions of $\boldsymbol{\psi}$.

Proof. The conditions $E\left(x_{itk}^2\right) < \infty$ and $E\left(y_{it}^2\right) < \infty$, for each i, t, and k, and C2 imply the existence of the linear projection in (4) (see, e.g., Wooldridge, 2010, pp. 25-26).

Also, $E[L_N(\boldsymbol{\psi})]$ is finite if $E\left[\boldsymbol{u}_i(\boldsymbol{\gamma})'\boldsymbol{\Omega}^{-1}\boldsymbol{u}_i(\boldsymbol{\gamma})\right]$ is finite, and the latter is finite if x_{itk} and y_{it} have finite second-order moments, for all i, t, and k.

The matrix $E[\boldsymbol{H}_N(\boldsymbol{\psi})]$ has finite elements as well. To see this, first let $\boldsymbol{W}_{i\cdot j}$ denote the jth column of $\boldsymbol{W}_i$, and let $\boldsymbol{S}_{\cdot j}$ denote the jth column of $\partial\text{vec}(\boldsymbol{\Omega})/\partial\boldsymbol{\omega}'$, where recall that $\boldsymbol{\omega} = \text{vech}(\boldsymbol{\Omega})$. Then, $\partial^2 l_i(\boldsymbol{\psi})/\partial\gamma_j\partial\gamma_k = -\boldsymbol{W}_{i\cdot j}'\boldsymbol{\Omega}^{-1}\boldsymbol{W}_{i\cdot k}$, $\partial^2 l_i(\boldsymbol{\psi})/\partial\gamma_j\partial\omega_k = -\boldsymbol{W}_{i\cdot j}'\boldsymbol{\Omega}^{-1}(\partial\boldsymbol{\Omega}/\partial\omega_k)\boldsymbol{\Omega}^{-1}\boldsymbol{u}_i(\boldsymbol{\gamma})$, and

$$\frac{\partial^2 l_i(\boldsymbol{\psi})}{\partial\omega_j\partial\omega_k} = \frac{1}{2}\boldsymbol{S}_{\cdot j}'\left(\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1}\right)\boldsymbol{S}_{\cdot k} - \frac{1}{2}s_{ijk}^{(1)}(\boldsymbol{\psi}) - \frac{1}{2}s_{ijk}^{(2)}(\boldsymbol{\psi}), \tag{20}$$

where $s_{ijk}^{(1)}(\boldsymbol{\psi}) = \boldsymbol{S}_{\cdot j}'\left(\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1}\boldsymbol{u}_i(\boldsymbol{\gamma})\,\boldsymbol{u}_i(\boldsymbol{\gamma})'\boldsymbol{\Omega}^{-1}\right)\boldsymbol{S}_{\cdot k}$ and $s_{ijk}^{(2)}(\boldsymbol{\psi}) = \boldsymbol{S}_{\cdot j}'\left(\boldsymbol{\Omega}^{-1}\boldsymbol{u}_i(\boldsymbol{\gamma})\,\boldsymbol{u}_i(\boldsymbol{\gamma})'\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1}\right)\boldsymbol{S}_{\cdot k}$ (see Ruud 2000, p. 930). From the preceding second-order partial derivatives we see that the condition $E\left(x_{itk}^2\right) < \infty$ and $E\left(y_{it}^2\right) < \infty$, for each i, t, and k, implies $E[\boldsymbol{H}_N(\boldsymbol{\psi})]$ has finite elements.

Inspection of $E[L_N(\boldsymbol{\psi})]$ and the elements of $E[\boldsymbol{H}_N(\boldsymbol{\psi})]$ reveals $E[L_N(\boldsymbol{\psi})]$ and the elements of $E[\boldsymbol{H}_N(\boldsymbol{\psi})]$ are functions of $\boldsymbol{\psi}$ and terms of the form $N^{-1}\sum_i E(y_{is}y_{it})$, $N^{-1}\sum_i E(y_{is}x_{itk})$, and $N^{-1}\sum_i E\left(x_{isj}x_{itk}\right)$. Therefore, if the limits of these averages exist (as $N\to\infty$), then the limits $L(\boldsymbol{\psi}) = \lim_{N\to\infty} E[L_N(\boldsymbol{\psi})]$ and $\boldsymbol{H}(\boldsymbol{\psi}) = \lim_{N\to\infty} E[\boldsymbol{H}_N(\boldsymbol{\psi})]$ exist, where $L(\boldsymbol{\psi})$ and the elements of $\boldsymbol{H}(\boldsymbol{\psi})$ are functions of $\boldsymbol{\psi}$ and terms involving limits of the form $\lim_{N\to\infty} N^{-1}\sum_i E(y_{is}y_{it})$, $\lim_{N\to\infty} N^{-1}\sum_i E(y_{is}x_{itk})$, and $\lim_{N\to\infty} N^{-1}\sum_i E\left(x_{isj}x_{itk}\right)$. And, inspection of $L(\boldsymbol{\psi})$ and the elements of $\boldsymbol{H}(\boldsymbol{\psi})$ reveals $L(\boldsymbol{\psi})$ and the elements of $\boldsymbol{H}(\boldsymbol{\psi})$ are continuous functions of $\boldsymbol{\psi}$.

Lemma B.2. Suppose $E\left(x_{itk}^2\right) < \infty$ and $E\left(y_{it}^2\right) < \infty$, for each i, t, and k, and Conditions C2 and C3 are satisfied. Then $E\left(\boldsymbol{W}_i\boldsymbol{\Omega}_0^{-1}\boldsymbol{u}_i\right) = \boldsymbol{0}$ for all i.

Proof. We have $E\left(\boldsymbol{Z}_i'\boldsymbol{\Omega}_0^{-1}\boldsymbol{u}_i\right) = \boldsymbol{0}$ because all of the elements of $\boldsymbol{u}_i$ are uncorrelated with all of the elements

of $\boldsymbol{Z}_i$ by construction. Moreover, the conditions of Lemma B.2 imply $E\left(\boldsymbol{u}_i \boldsymbol{y}_i^{o\prime}\right) = \mathbf{0}$, $E\left(\boldsymbol{u}_i \boldsymbol{x}_i'\right) = \mathbf{0}$, and $E\left(\boldsymbol{u}_i \boldsymbol{u}_i'\right) = \Omega_0$. Thus, the conditions of Lemma 1 hold for the augmented regression in (6). Hence, by Lemma 1, we have $E\left(\boldsymbol{y}_{i,-j}' \Omega_0^{-1} \boldsymbol{u}_i\right) = 0$ $(j = 1, \ldots, p)$. This proves $E\left(\boldsymbol{W}_i \Omega_0^{-1} \boldsymbol{u}_i\right) = \mathbf{0}$.

Lemma B.3. Let $\overline{\Psi}$ denote a compact subset of Ψ. Suppose C1, C4, and C5 are satisfied. Then $L_N(\cdot) \xrightarrow{a.s.} L(\cdot)$ uniformly on $\overline{\Psi}$.

Proof. Let ω^{st} denote the (s,t)th element of Ω^{-1}; let γ_k denote the kth element of $\boldsymbol{\gamma}$; recall that $\boldsymbol{W}_{i\cdot j}$ is the jth column of $\boldsymbol{W}_i$; and let W_{itj} denote the tth element of $\boldsymbol{W}_{i\cdot j}$. Also, let $S_{y_s y_t,N} = N^{-1}\sum_i \left[y_{is}y_{it} - E\left(y_{is}y_{it}\right)\right]$, $S_{y_s W_{tj},N} = N^{-1}\sum_i \left[y_{is}W_{itj} - E\left(y_{is}W_{itj}\right)\right]$, and $S_{W_{sj}W_{tk},N} = N^{-1}\sum_i \left[W_{isj}W_{itk} - E\left(W_{isj}W_{itk}\right)\right]$. Then $L_N(\boldsymbol{\psi}) - E\left[L_N(\boldsymbol{\psi})\right] = -\sum_s\sum_t \omega^{st} S_{y_s y_t,N}/2 + \sum_s\sum_t \omega^{st}\sum_j \gamma_j S_{y_s W_{tj},N} - \sum_s\sum_t \omega^{st}\sum_j\sum_k \gamma_j\gamma_k S_{W_{sj}W_{tk},N}/2$. Therefore, by an obvious inequality, we have $\left|L_N(\boldsymbol{\psi}) - E\left[L_N(\boldsymbol{\psi})\right]\right| \leq \sum_s\sum_t \left|\omega^{st}\right|\left|S_{y_s y_t,N}\right|/2 + \sum_s\sum_t \left|\omega^{st}\right|\sum_k \left|\gamma_k\right|\left|S_{y_s W_{tk},N}\right| + \sum_s\sum_t \left|\omega^{st}\right|\sum_j\sum_k \left|\gamma_j\gamma_k\right|\left|S_{W_{sj}W_{tk},N}\right|/2$. Given ω^{st} and γ_k are bounded for $\boldsymbol{\psi} \in \overline{\Psi}$, it follows that

$$\begin{aligned}\sup_{\boldsymbol{\psi}\in\overline{\Psi}} \left|L_N(\boldsymbol{\psi}) - E\left[L_N(\boldsymbol{\psi})\right]\right| \leq\ & M\sum_s\sum_t \left|S_{y_s y_t,N}\right| + M\sum_s\sum_t\sum_k \left|S_{y_s W_{tk},N}\right| \\ & + M\sum_s\sum_t\sum_j\sum_k \left|S_{W_{sj}W_{tk},N}\right|. \end{aligned} \tag{21}$$

Hence, $L_N(\cdot) - E\left[L_N(\cdot)\right] \xrightarrow{a.s.} 0$ uniformly on $\overline{\Psi}$ if $S_{y_s y_t,N} \xrightarrow{a.s.} 0$, $S_{y_s W_{tk},N} \xrightarrow{a.s.} 0$, and $S_{W_{sj}W_{tk},N} \xrightarrow{a.s.} 0$ for each s, t, j, and k.

To see that $S_{y_s y_t,N} \xrightarrow{a.s.} 0$, note that, by the Cauchy-Schwarz inequality and C1, we get $E\left|y_{is}y_{it}\right|^{1+\epsilon/2} \leq \left(E\left|y_{is}\right|^{2+\epsilon} E\left|y_{it}\right|^{2+\epsilon}\right)^{1/2} < M$ for some $\epsilon > 0$ and all i, s, and t. This conclusion and C5 imply $S_{y_s y_t,N} \xrightarrow{a.s.} 0$ (see White 2001, p. 35, Corollary 3.9). By similar arguments, we also have $S_{y_s W_{tk},N} \xrightarrow{a.s.} 0$ and $S_{W_{sj}W_{tk},N} \xrightarrow{a.s.} 0$. Hence, $L_N(\cdot) - E\left[L_N(\cdot)\right] \xrightarrow{a.s.} 0$ uniformly on $\overline{\Psi}$.

Given C4, the following expressions are defined: $A_{y_s y_t,N} = N^{-1}\sum_i E\left(y_{is}y_{it}\right) - \lim_{N\to\infty} N^{-1}\sum_i E\left(y_{is}y_{it}\right)$, $A_{y_s W_{tj},N} = N^{-1}\sum_i E\left(y_{is}W_{itj}\right) - \lim_{N\to\infty} N^{-1}\sum_i E\left(y_{is}W_{itj}\right)$, and $A_{W_{sj}W_{tk},N} = N^{-1}\sum_i E\left(W_{isj}W_{itk}\right) - \lim_{N\to\infty} N^{-1}\sum_i E\left(W_{isj}W_{itk}\right)$. And, by arguments analogous to those leading to the inequality in (21), one can show $\sup_{\boldsymbol{\psi}\in\overline{\Psi}} \left|E\left[L_N(\boldsymbol{\psi})\right] - L(\boldsymbol{\psi})\right| \leq M\sum_s\sum_t \left|A_{y_s y_t,N}\right| + M\sum_s\sum_t\sum_j \left|A_{y_s W_{tj},N}\right| + M\sum_s\sum_t\sum_j\sum_k \left|A_{W_{sj}W_{tk},N}\right|$. Because $A_{y_s y_t,N}$, $A_{y_s W_{tj},N}$, and $A_{W_{sj}W_{tk},N}$ all $\to 0$, we have $E\left[L_N(\cdot)\right] \to L(\cdot)$ uniformly on $\overline{\Psi}$.

The conclusions of the last two paragraphs imply $L_N(\cdot) \xrightarrow{a.s.} L(\cdot)$ uniformly on $\overline{\Psi}$.

Lemma B.4. Suppose Conditions C1–C4 are satisfied and $\boldsymbol{H}_0$ is negative definite. Then there is a compact subset $\overline{\Psi}$ of Ψ, with $\boldsymbol{\psi}_0$ in its interior, such that $L(\boldsymbol{\psi}) < L(\boldsymbol{\psi}_0)$ if $\boldsymbol{\psi} \in \overline{\Psi}$ and $\boldsymbol{\psi} \neq \boldsymbol{\psi}_0$.

Proof. First

$$E\left[\partial l_i(\boldsymbol{\psi}_0)/\partial \boldsymbol{\psi}\right] = \mathbf{0} \tag{22}$$

is established. By well known results, $\partial l_i(\boldsymbol{\psi})/\partial \boldsymbol{\gamma} = \boldsymbol{W}_i' \boldsymbol{\Omega}^{-1} \boldsymbol{u}_i(\boldsymbol{\gamma})$ and

$$\frac{\partial l_i(\boldsymbol{\psi})}{\partial \boldsymbol{\omega}} = -\frac{1}{2}\text{vech}\left(\boldsymbol{\Omega}^{-1} - \boldsymbol{\Omega}^{-1}\boldsymbol{u}_i(\boldsymbol{\gamma})\,\boldsymbol{u}_i(\boldsymbol{\gamma})'\,\boldsymbol{\Omega}^{-1}\right) \tag{23}$$

(see, e.g., Ruud, 2000, pp. 928-930). We have $E\left[\partial l_i(\boldsymbol{\psi}_0)/\partial \boldsymbol{\gamma}\right] = \mathbf{0}$ (Lemma B.2), and from Eq. (23), it is clear that, because $E\left(\boldsymbol{u}_i \boldsymbol{u}_i'\right) = \boldsymbol{\Omega}_0$, we have $E\left[\partial l_i(\boldsymbol{\psi}_0)/\partial \boldsymbol{\omega}\right] = \mathbf{0}$.

Next, a Taylor series expansion gives

$$L_N(\boldsymbol{\psi}) = L_N(\boldsymbol{\psi}_0) + (\boldsymbol{\psi} - \boldsymbol{\psi}_0)'\,\boldsymbol{g}_N(\boldsymbol{\psi}_0) + (\boldsymbol{\psi} - \boldsymbol{\psi}_0)'\,\boldsymbol{H}_N(\boldsymbol{\psi}^*)(\boldsymbol{\psi} - \boldsymbol{\psi}_0)/2, \tag{24}$$

where $\boldsymbol{g}_N(\boldsymbol{\psi}_0) = \partial L_N(\boldsymbol{\psi}_0)/\partial \boldsymbol{\psi}$, and $\boldsymbol{\psi}^*$ satisfies $\|\boldsymbol{\psi} - \boldsymbol{\psi}^*\| \leq \|\boldsymbol{\psi} - \boldsymbol{\psi}_0\|$. Given Eq. (22) and Lemma B.1, taking the expectation of the left and right-hand sides of (24) and then letting $N \to \infty$ gives $L(\boldsymbol{\psi}) = L(\boldsymbol{\psi}_0) + (\boldsymbol{\psi} - \boldsymbol{\psi}_0)'\,\boldsymbol{H}(\boldsymbol{\psi}^*)(\boldsymbol{\psi} - \boldsymbol{\psi}_0)/2$.

Let $h_{jk}(\boldsymbol{\psi})$ denote the (j,k)th element of $\boldsymbol{H}(\boldsymbol{\psi})$, and define determinants

$$d_j(\boldsymbol{\psi}) = \begin{vmatrix} h_{11}(\boldsymbol{\psi}) & \cdots & h_{1j}(\boldsymbol{\psi}) \\ \vdots & \ddots & \vdots \\ h_{j1}(\boldsymbol{\psi}) & \cdots & h_{jj}(\boldsymbol{\psi}) \end{vmatrix} \qquad (j = 1, \ldots, m).$$

A necessary and sufficient condition for $\boldsymbol{H}(\boldsymbol{\psi})$ to be negative definite is $d_1(\boldsymbol{\psi}) < 0$, $d_2(\boldsymbol{\psi}) > 0$, $d_3(\boldsymbol{\psi}) < 0, \ldots$ (see Rao 1973, p. 37). By assumption, $\boldsymbol{H}_0 = \boldsymbol{H}(\boldsymbol{\psi}_0)$ is negative definite, and thus $d_1(\boldsymbol{\psi}_0) < 0$, $d_2(\boldsymbol{\psi}_0) > 0$, $d_3(\boldsymbol{\psi}_0) < 0, \ldots$. Moreover, the determinant $d_j(\cdot)$ is continuous in $h_{11}(\cdot), h_{12}(\cdot), \ldots$, which are, in turn, continuous in $\boldsymbol{\psi}$ (see Lemma B.1). Hence, $d_j(\cdot)$ is continuous in $\boldsymbol{\psi}$. It follows that there is a $r > 0$ such that for the closed ball in $\mathbb{R}^m$, centered at $\boldsymbol{\psi}_0$, with radius r, we have $d_1(\boldsymbol{\psi}) < 0$, $d_2(\boldsymbol{\psi}) > 0$, $d_3(\boldsymbol{\psi}) < 0, \ldots$ for $\boldsymbol{\psi}$ in the ball. Let $\overline{\Psi}$ denote the ball (a compact subset of $\mathbb{R}^m$). Then $\boldsymbol{H}(\boldsymbol{\psi})$ is negative definite for $\boldsymbol{\psi} \in \overline{\Psi}$. Therefore, for $\boldsymbol{\psi} \neq \boldsymbol{\psi}_0$ and $\boldsymbol{\psi} \in \overline{\Psi}$, we must have $(\boldsymbol{\psi} - \boldsymbol{\psi}_0)'\,\boldsymbol{H}_N(\boldsymbol{\psi}^*)(\boldsymbol{\psi} - \boldsymbol{\psi}_0) < 0$,

because $\psi \in \overline{\Psi}$ implies $\psi^* \in \overline{\Psi}$ and, therefore, $\boldsymbol{H}\left(\psi^*\right)$ is negative definite. Hence, $L(\psi) < L\left(\psi_0\right)$ if $\psi \in \overline{\Psi}$ and $\psi \neq \psi_0$.

Proof of Theorem 1: The conclusions of Lemmas B.3 and and B.4 imply there is a measurable maximizer, $\widehat{\psi}$, in $\overline{\Psi}$ and $\widehat{\psi} \xrightarrow{a.s.} \psi_0$ (see, e.g., Amemiya, 1985, Theorem 4.1.1, and his footnote 1 on p. 107).

Appendix C: Theorem 2 Proof

Theorem 2 is proven by establishing several lemmas. The first result is an elementary inequality, which is applied repeatedly in the sequel.

Lemma C.1. For $r > 0$, $\left|\sum_{j=1}^m a_j\right|^r \leq b_r \sum_{j=1}^m \left|a_j\right|^r$ where $b_r = 1$ or $2^{(r-1)(m-1)}$ according as $r \leq 1$ or $r \geq 1$.

Proof. By repeated application of the inequality $|a+b|^r \leq c_r |a|^r + c_r |b|^r$, $r > 0$, where $c_r = 1$ or 2^{r-1} according as $r \leq 1$ or $r \geq 1$ (see Loève 1977, p. 157), we have $\left|\sum_{j=1}^m a_j\right|^r \leq c_r |a_1|^r + c_r \left|\sum_{j=2}^m a_j\right|^r \leq c_r |a_1|^r + c_r^2 |a_2|^r + c_r^2 \left|\sum_{j=3}^m a_j\right|^r \leq \sum_{j=1}^{m-1} c_r^j \left|a_j\right|^r + c_r^{m-1} |a_m|^r$. Also, $\sum_{j=1}^{m-1} c_r^j \left|a_j\right|^r + c_r^{m-1} |a_m|^r \leq b_r \sum_{j=1}^m \left|a_j\right|^r$ for $b_r = c_r^{m-1}$.

Lemma C.2. Suppose C1′, C2, C3, C5, and C6 are satisfied. Then $\sqrt{N}\boldsymbol{g}_N\left(\psi_0\right) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \mathcal{I}_0)$.

Proof. Let $\boldsymbol{\mu}$ be a $m \times 1$ vector of constants such that $\boldsymbol{\mu} \neq \boldsymbol{0}$. We have $\boldsymbol{\mu}'\sqrt{N}\boldsymbol{g}_N\left(\psi_0\right) = N^{-1/2}\sum_i \mathcal{Z}_i$ for $\mathcal{Z}_i = \boldsymbol{\mu}'\left(\partial l_i\left(\psi_0\right)/\partial\psi\right)$. And $\sqrt{N}\boldsymbol{g}_N\left(\psi_0\right) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \mathcal{I}_0)$ if $N^{-1/2}\sum_i \mathcal{Z}_i \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{\mu}'\mathcal{I}_0\boldsymbol{\mu}\right)$ (see Amemiya 1985, Theorem 3.3.8).

To verify $N^{-1/2}\sum_i \mathcal{Z}_i \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{\mu}'\mathcal{I}_0\boldsymbol{\mu}\right)$, let $v_i^2 = var\left(\mathcal{Z}_i\right) = \boldsymbol{\mu}' E\left[\left(\partial l_i\left(\psi_0\right)/\partial\psi\right)\left(\partial l_i\left(\psi_0\right)/\partial\psi\right)'\right]\boldsymbol{\mu}$, and $\overline{v}_N^2 = N^{-1}\sum_i v_i^2$. Because $\lim_{N\to\infty}\overline{v}_N^2 = \boldsymbol{\mu}'\mathcal{I}_0\boldsymbol{\mu}$ (by C6), we have $N^{-1/2}\sum_i \mathcal{Z}_i \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{\mu}'\mathcal{I}_0\boldsymbol{\mu}\right)$ if $N^{-1/2}\sum_i \mathcal{Z}_i/\overline{v}_N \xrightarrow{d} \mathcal{N}(0,1)$. Moreover, $N^{-1/2}\sum_i \mathcal{Z}_i/\overline{v}_N \xrightarrow{d} \mathcal{N}(0,1)$ if $E\left(\mathcal{Z}_i\right) = 0$, $\overline{v}_N^2 > \epsilon' > 0$ for all N sufficiently large, and $E\left|\mathcal{Z}_i\right|^{2+\epsilon/2} < M$ for all i and some $\epsilon/2 > 0$ (see White 2001, Theorem 5.10). Therefore, Lemma C.2 is proven upon proving $E\left(\mathcal{Z}_i\right) = 0$, $\overline{v}_N^2 > \epsilon' > 0$ for all N sufficiently large, and $E\left|\mathcal{Z}_i\right|^{2+\epsilon/2} < M$ for all i and some $\epsilon/2 > 0$.

We can verify $E\left(\mathcal{Z}_i\right) = 0$ and $\overline{v}_N^2 > \epsilon' > 0$ for all N sufficiently large easily. In particular, Eq. (22) implies $E\left(\mathcal{Z}_i\right) = 0$. Moreover, given C6, we have $\lim_{N\to\infty}\overline{v}_N^2 = \boldsymbol{\mu}'\mathcal{I}_0\boldsymbol{\mu}$, and, because $\mathcal{I}_0$ is positive definite, we can find an $\epsilon' > 0$ such that $\overline{v}_N^2 > \epsilon'$ for all N sufficiently large.

To verify $E\left|\mathcal{Z}_i\right|^{2+\epsilon/2} < M$ for all i and some $\epsilon/2 > 0$, first let μ_j and ψ_j denote the jth elements of $\boldsymbol{\mu}$ and $\boldsymbol{\psi}$. Then $\mathcal{Z}_i = \sum_j \mu_j \partial l_i\left(\boldsymbol{\psi}_0\right)/\partial\psi_j$. Hence, by Lemma C.1, we have $E\left|\mathcal{Z}_i\right|^{2+\epsilon/2} < M$ for all i if $E\left|\partial l_i\left(\boldsymbol{\psi}_0\right)/\partial\psi_j\right|^{2+\epsilon/2} < M$ for all i and j. Next, recall $\partial l_i\left(\boldsymbol{\psi}_0\right)/\partial\boldsymbol{\gamma} = \boldsymbol{W}_i'\boldsymbol{\Omega}_0^{-1}\boldsymbol{u}_i$ while $\partial l_i\left(\boldsymbol{\psi}_0\right)/\partial\boldsymbol{\omega} = -\text{vech}\left(\boldsymbol{\Omega}_0^{-1} - \boldsymbol{\Omega}_0^{-1}\boldsymbol{u}_i\boldsymbol{u}_i'\boldsymbol{\Omega}_0^{-1}\right)/2$. Moreover, upon letting ω_0^{st} denote the (s,t)th element of $\boldsymbol{\Omega}_0^{-1}$ and recalling W_{isj} denotes the (s,j)th element of $\boldsymbol{W}_i$, the elements of $\boldsymbol{W}_i'\boldsymbol{\Omega}_0^{-1}\boldsymbol{u}_i$ are of the form $\sum_s\sum_t \omega_0^{st}W_{isj}u_{it}$ while the elements of $\text{vech}\left(\boldsymbol{\Omega}_0^{-1} - \boldsymbol{\Omega}_0^{-1}\boldsymbol{u}_i\boldsymbol{u}_i'\boldsymbol{\Omega}_0^{-1}\right)$ are of the form $\omega_0^{jk} - \sum_s\sum_t \omega_0^{js}\omega_0^{kt}u_{is}u_{it}$. These observations and another application of Lemma C.1 implies $E\left|\partial l_i\left(\boldsymbol{\psi}_0\right)/\partial\psi_j\right|^{2+\epsilon/2} < M$ for all i and j if $E\left|W_{isj}u_{it}\right|^{2+\epsilon/2} < M$ and $E\left|u_{is}u_{it}\right|^{2+\epsilon/2} < M$ for all i, j, s, and t. But $E\left|W_{isj}u_{it}\right|^{2+\epsilon/2} \leq \left(E\left|W_{isj}\right|^{4+\epsilon}E\left|u_{it}\right|^{4+\epsilon}\right)^{1/2}$ by the Cauchy-Schwarz inequality. Moreover, for a suitable choice of $\epsilon > 0$, we have $E\left|W_{isj}\right|^{4+\epsilon} < M$ for all i, s, and j by C1′. Condition C1′ also implies $E\left|u_{it}\right|^{4+\epsilon} < M$ for all i and t. Hence, $E\left|W_{isj}u_{it}\right|^{2+\epsilon/2} < M$ for all i, j, s, and t. Similar arguments give $E\left|u_{is}u_{it}\right|^{2+\epsilon/2} < M$ for all i, s, and t. It follows that $E\left|\mathcal{Z}_i\right|^{2+\epsilon/2} < M$ for all i and some $\epsilon/2 > 0$.

Lemma C.3. Let $\overline{\Psi}$ be a compact subset of Ψ. Suppose C1, C4, and C5 are satisfied. Then $\boldsymbol{H}_N(\cdot) \xrightarrow{a.s.} \boldsymbol{H}(\cdot)$ uniformly on $\overline{\Psi}$.

Proof. Let $h_{\gamma_j\gamma_k}(\boldsymbol{\psi}) = \lim_{N\to\infty} E\left[\partial^2 L_N(\boldsymbol{\psi})/\partial\gamma_j\partial\gamma_k\right]$. Then $\left|\partial^2 L_N(\boldsymbol{\psi})/\partial\gamma_j\partial\gamma - h_{\gamma_j\gamma_k}(\boldsymbol{\psi})\right| = \left|\sum_s\sum_t \omega^{st}\left(S_{W_{sj}W_{tk},N} + A_{W_{sj}W_{tk},N}\right)\right| \leq \sum_s\sum_t \left|\omega^{st}\right|\left(\left|S_{W_{sj}W_{tk},N}\right| + \left|A_{W_{sj}W_{tk},N}\right|\right)$. (For the definitions of $S_{W_{sj}W_{tk},N}$ and $A_{W_{sj}W_{tk},N}$, see the proof of Lemma B.3.) Given ω^{st} is bounded for $\boldsymbol{\psi} \in \overline{\Psi}$, we have $\sup_{\boldsymbol{\psi}\in\overline{\Psi}}\left|\partial^2 L_N(\boldsymbol{\psi})/\partial\gamma_j\partial\gamma_k - h_{\gamma_j\gamma_k}(\boldsymbol{\psi})\right| \leq M\sum_s\sum_t\left(\left|S_{W_{sj}W_{tk},N}\right| + \left|A_{W_{sj}W_{tk},N}\right|\right)$. Recall that $S_{W_{sj}W_{tk},N} \xrightarrow{a.s.} 0$ (see the proof of Lemma B.3), and $A_{W_{sj}W_{tk},N} \to 0$. Therefore, $\partial^2 L_N(\cdot)/\partial\gamma_j\partial\gamma_k \xrightarrow{a.s.} h_{\gamma_j\gamma_k}(\cdot)$ uniformly on $\overline{\Psi}$.

Let $h_{\gamma_j\omega_k}(\boldsymbol{\psi}) = \lim_{N\to\infty} E\left[\partial^2 L_N(\boldsymbol{\psi})/\partial\gamma_j\partial\omega_k\right]$. Also, let $\vartheta_{k,st}$ denote the $(s,t)$the element of $\boldsymbol{\Omega}^{-1}\left(\partial\boldsymbol{\Omega}/\partial\omega_k\right)\boldsymbol{\Omega}^{-1}$. Then $\partial^2 L_N(\boldsymbol{\psi})/\partial\gamma_j\partial\omega_k - h_{\gamma_j\omega_k}(\boldsymbol{\psi}) = -\sum_s\sum_t \vartheta_{k,st}[S_{y_sW_{tj},N} + A_{y_sW_{tj},N}] + \sum_s\sum_t\sum_l \vartheta_{k,st}\gamma_l\left[S_{W_{sj}W_{tl},N} + A_{W_{sj}W_{tl},N}\right]$. (For the definitions of $S_{y_sW_{tj},N}$ and $A_{y_sW_{tj},N}$, see the proof of Lemma B.3.) Because $\vartheta_{k,st}$ is a continuous function on $\overline{\Psi}$, and, therefore, bounded on $\overline{\Psi}$, and γ_l is bounded for $\boldsymbol{\psi} \in \overline{\Psi}$, we have $\sup_{\boldsymbol{\psi}\in\overline{\Psi}}\left|\partial^2 L_N(\boldsymbol{\psi})/\partial\gamma_j\partial\omega_k - h_{\gamma_j\omega_k}(\boldsymbol{\psi})\right| \leq M\sum_s\sum_t\left(\left|S_{y_sW_{tj},N}\right| + \left|A_{y_sW_{tj},N}\right|\right) + M\sum_s\sum_t\sum_l\left(\left|S_{W_{sj}W_{tl},N}\right| + \left|A_{W_{sj}W_{tl},N}\right|\right)$. Given $S_{y_sW_{tj},N} \xrightarrow{a.s.} 0$, $S_{W_{sj}W_{tl},N} \xrightarrow{a.s.} 0$, $A_{y_sW_{tj},N} \to 0$, and $A_{W_{sj}W_{tl},N} \to 0$, we have $\partial^2 L_N(\cdot)/\partial\gamma_j\partial\omega_k \xrightarrow{a.s.} h_{\gamma_j\omega_k}(\cdot)$ uniformly on $\overline{\Psi}$.

Finally, from (20), we see that $\partial^2 L_N(\boldsymbol{\psi})/\partial\omega_j\partial\omega_k - E\left(\partial^2 L_N(\boldsymbol{\psi})/\partial\omega_j\partial\omega_k\right) =$

$-(2N)^{-1}\sum_i\left\{s_{ijk}^{(1)}(\psi)-E\left[s_{ijk}^{(1)}(\psi)\right]\right\}-(2N)^{-1}\sum_i\left\{s_{ijk}^{(2)}(\psi)-E\left[s_{ijk}^{(2)}(\psi)\right]\right\}$. Note that

$$\frac{1}{N}\sum_i\left\{s_{ijk}^{(1)}(\psi)-E\left[s_{ijk}^{(1)}(\psi)\right]\right\}=\boldsymbol{S}_{\cdot j}'\left(\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1}\boldsymbol{U}_N(\gamma)\,\boldsymbol{\Omega}^{-1}\right)\boldsymbol{S}_{\cdot k} \tag{25}$$

where $\boldsymbol{U}_N(\gamma)=N^{-1}\sum_i\left\{\boldsymbol{u}_i(\gamma)\,\boldsymbol{u}_i(\gamma)'-E\left[\boldsymbol{u}_i(\gamma)\,\boldsymbol{u}_i(\gamma)'\right]\right\}$. Because $\boldsymbol{S}_{\cdot j}$ is a vector of zeros and ones, we see that the right-hand side of (25) is a sum of the elements of $\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1}\boldsymbol{U}_N(\gamma)\,\boldsymbol{\Omega}^{-1}$. Therefore, if each element of this matrix converges almost surely to zero uniformly on $\overline{\Psi}$, then $N^{-1}\sum_i\left\{s_{ijk}^{(1)}(\cdot)-E\left[s_{ijk}^{(1)}(\cdot)\right]\right\}\overset{a.s.}{\to}0$ uniformly on $\overline{\Psi}$. Similar arguments can be used to show $N^{-1}\sum_i\left\{s_{ijk}^{(2)}(\cdot)-E\left[s_{ijk}^{(2)}(\cdot)\right]\right\}\overset{a.s.}{\to}0$ uniformly on $\overline{\Psi}$.

To see that each element of $\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1}\boldsymbol{U}_N(\gamma)\,\boldsymbol{\Omega}^{-1}$ converges almost surely to zero uniformly, note that the matrix $\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1}\boldsymbol{U}_N(\gamma)\,\boldsymbol{\Omega}^{-1}$ can be partitioned into $T\times T$ sub-matrices of the form $\omega^{lm}\boldsymbol{\Omega}^{-1}\boldsymbol{U}_N(\gamma)\,\boldsymbol{\Omega}^{-1}$ ($l=1,\ldots,T$, $m=1,\ldots,T$). Furthermore, the (j,k)th element of $\omega^{lm}\boldsymbol{\Omega}^{-1}\boldsymbol{U}_N(\gamma)\,\boldsymbol{\Omega}^{-1}$ is $\omega^{lm}\sum_s\sum_t\omega^{js}\omega^{kt}N^{-1}\sum_i\{u_{is}(\gamma)\,u_{it}(\gamma)-E[u_{is}(\gamma)\,u_{it}(\gamma)]\}$. And, by familiar arguments, we can show that the absolute value of this element is no greater than $M\sum_s\sum_t\left|N^{-1}\sum_i\{u_{is}(\gamma)\,u_{it}(\gamma)-E[u_{is}(\gamma)\,u_{it}(\gamma)]\}\right|$ for $\psi\in\overline{\Psi}$. Moreover, $N^{-1}\sum_i\{u_{is}(\gamma)\,u_{it}(\gamma)-E[u_{is}(\gamma)\,u_{it}(\gamma)]\}=S_{y_sy_t,N}-\sum_q\gamma_q(S_{y_sW_{tq},N}+S_{y_tW_{sq},N})+\sum_q\sum_r\gamma_q\gamma_rS_{W_{sq}W_{tr}N}$, and given γ is bounded for $\psi\in\overline{\Psi}$, we have

$$\begin{aligned}&\sup_{\psi\in\overline{\Psi}}\left|\frac{1}{N}\sum_i\{u_{is}(\gamma)\,u_{it}(\gamma)-E[u_{is}(\gamma)\,u_{it}(\gamma)]\}\right|\\ \leq\;&\left|S_{y_sy_t,N}\right|+M\sum_q\left(\left|S_{y_sW_{tq},N}\right|+\left|S_{y_tW_{sq},N}\right|+\sum_r\left|S_{W_{sq}W_{tr}N}\right|\right).\end{aligned} \tag{26}$$

Because the right-hand side (26) $\overset{a.s.}{\to}0$ (see the proof of Lemma B.3), we have $N^{-1}\sum_i\left\{s_{i,jk}^{(1)}(\cdot)-E\left[s_{i,jk}^{(1)}(\cdot)\right]\right\}\overset{a.s.}{\to}0$ uniformly on $\overline{\Psi}$. Simliar arguments establish $N^{-1}\sum_i\left\{s_{i,jk}^{(2)}(\cdot)-E\left[s_{i,jk}^{(2)}(\cdot)\right]\right\}\overset{a.s.}{\to}0$ uniformly on $\overline{\Psi}$. It follows that $\partial^2L_N(\cdot)/\partial\omega_j\partial\omega_k-E\left[\partial^2L_N(\cdot)/\partial\omega_j\partial\omega_k\right]\overset{a.s.}{\to}0$ uniformly on $\overline{\Psi}$.

Let $h_{\omega_j\omega_k}(\psi)=\lim_{N\to\infty}E\left[\partial^2L_N(\psi)/\partial\omega_j\partial\omega_k\right]$. We can establish $E\left[\partial^2L_N(\cdot)/\partial\omega_j\partial\omega_k\right]\to h_{\omega_j\omega_k}(\cdot)$ uniformly on $\overline{\Psi}$ by arguments paralleling those in the last two paragraphs. (For example, in the foregoing derivations, replace $N^{-1}\sum_iE[u_{is}(\gamma)\,u_{it}(\gamma)]$ with $\lim_{N\to\infty}N^{-1}\sum_iE[u_{is}(\gamma)\,u_{it}(\gamma)]$ and $N^{-1}\sum_iu_{is}(\gamma)\,u_{it}(\gamma)$ with $N^{-1}\sum_iE[u_{is}(\gamma)\,u_{it}(\gamma)]$. Also, replace $S_{y_sy_t,N}$, $S_{y_sW_{tq},N}$, $S_{y_tW_{sq},N}$, and

$S_{W_{sq}W_{tr}N}$ with $A_{y_s y_t,N}$, $A_{y_s W_{tq},N}$, $A_{y_t W_{sq},N}$, and $A_{W_{sq}W_{tr}N}$.)

From the foregoing, we have $\partial^2 L_N(\cdot)/\partial\omega_j\partial\omega_k \xrightarrow{a.s.} h_{\omega_j\omega_k}(\cdot)$ uniformly on $\overline{\Psi}$.

Proof of Theorem 2: The conclusions of Lemmas C.2 and C.3, the consistency of $\widehat{\boldsymbol{\psi}}$, the continuity of $\boldsymbol{H}(\cdot)$ at $\boldsymbol{\psi}_0$, and the nonsingularity of $\boldsymbol{H}_0 = \boldsymbol{H}(\boldsymbol{\psi}_0)$ imply $\sqrt{N}(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{H}_0^{-1}\mathcal{I}_0\boldsymbol{H}_0^{-1})$ (see Newey and McFadden 1994, Theorem 3.1).

Appendix D: Theorems 3 and 4

The proofs of Theorems 3 and 4 are similar to the proofs of Theorems 1 and 2. For example, Conditions C1 and C2′ ensure the linear projection parameters in (8) exist and do not depend on i and the errors in $\widetilde{\boldsymbol{u}}_i$ are uncorrelated with the regressors in $\boldsymbol{x}_i$. Furthermore, the quasi log-likelihood $\sum_{i=1}^N \widetilde{l}(\boldsymbol{\lambda}_0)$ is similar to the quasi log-likelihood $\sum_{i=1}^N l(\boldsymbol{\psi}_0)$, and, therefore, most of the technical details are the same as in Appendices B and C and need not be repeated.

However, the conlusions of Theorems 3 and 4 depend on $E\left(\widetilde{\boldsymbol{W}}_i'\Upsilon_0^{-1}\widetilde{\boldsymbol{u}}_i\right) = \mathbf{0}$ being true, and the proof of this result, though similar to the proof of Lemma 1, differs in some details. Therefore, the proof of $E\left(\widetilde{\boldsymbol{W}}_i'\Upsilon_0^{-1}\widetilde{\boldsymbol{u}}_i\right) = \mathbf{0}$ is provided in this appendix.

Lemma D.1. Suppose $E(x_{itk}^2) < \infty$ and $E(y_{it}^2) < \infty$, for each i, t, and k, and Conditions C2′ and C3′ are satisfied. Then $E\left(\widetilde{\boldsymbol{W}}_i'\Upsilon_0^{-1}\widetilde{\boldsymbol{u}}_i\right) = \mathbf{0}$.

Proof. Let

$$\widetilde{\boldsymbol{Z}}_i = \begin{pmatrix} \mathbf{0} & \boldsymbol{I}_p \otimes (1, \boldsymbol{x}_i') \\ \Delta\boldsymbol{X}_i & \mathbf{0} \end{pmatrix}.$$

Given this definition, showing $E\left(\widetilde{\boldsymbol{W}}_i'\Upsilon_0^{-1}\widetilde{\boldsymbol{u}}_i\right) = \mathbf{0}$ consists of showing $E\left(\widetilde{\boldsymbol{Z}}_i'\Upsilon_0^{-1}\widetilde{\boldsymbol{u}}_i\right) = \mathbf{0}$ and $E\left[\left(\mathbf{0}, \Delta\boldsymbol{y}_{i,-j}'\right)\Upsilon_0^{-1}\widetilde{\boldsymbol{u}}_i\right] = 0$ $(j = 1, \ldots, p)$. Under the conditions of the lemma, the elements of $\widetilde{\boldsymbol{Z}}_i$ are uncorrelated with the elements of $\widetilde{\boldsymbol{u}}_i$; hence, $E\left(\widetilde{\boldsymbol{Z}}_i'\Upsilon_0^{-1}\widetilde{\boldsymbol{u}}_i\right) = \mathbf{0}$. It remains to show $E\left[\left(\mathbf{0}, \Delta\boldsymbol{y}_{i,-j}'\right)\Upsilon_0^{-1}\widetilde{\boldsymbol{u}}_i\right] = 0$.

This result can be established by arguments similar to those used in the proof of Lemma 1. Specifically, let $\Delta\boldsymbol{\xi}_{it} = (\Delta y_{it}, \Delta y_{i,t-1}, \ldots, \Delta y_{i,t-p+1})'$, $\Delta\boldsymbol{\varsigma}_{it} = (\Delta\boldsymbol{x}_{it}'\boldsymbol{\beta}_0 + \Delta e_{it}, 0, \ldots, 0)'$ and let $\boldsymbol{F}$ be defined as in (15). Then we get $\Delta\boldsymbol{\xi}_{i2} = \boldsymbol{F}\Delta\boldsymbol{\xi}_{i1} + \Delta\boldsymbol{\varsigma}_{i2}$; and, for $t > 2$, we have $\Delta\boldsymbol{\xi}_{it} = \boldsymbol{F}^{t-1}\Delta\boldsymbol{\xi}_{i1} + \boldsymbol{F}^{t-2}\Delta\boldsymbol{\varsigma}_{i2} + \cdots + \boldsymbol{F}\Delta\boldsymbol{\varsigma}_{i,t-1} + \Delta\boldsymbol{\varsigma}_{it}$. Let $f_{rs}^{(t)}$ denote the (r,s)th element of $\boldsymbol{F}^t$. Then, the preceding implies $\Delta y_{i2} = f_{11}^{(1)}\Delta y_{i1} + f_{12}^{(1)}\Delta y_{i0} + \cdots + f_{1p}^{(1)}\Delta y_{i,-p+2} + \Delta\boldsymbol{x}_{i2}'\boldsymbol{\beta}_0 + \Delta e_{i2}$; and, for $t > 2$, we have $\Delta y_{it} = f_{11}^{(t-1)}\Delta y_{i1} + f_{12}^{(t-1)}\Delta y_{i0} +$

$\cdots + f_{1p}^{(t-1)} \Delta y_{i,-p+2} + f_{11}^{(t-2)} \left(\Delta \boldsymbol{x}_{i2}' \boldsymbol{\beta}_0 + \Delta e_{i2} \right) + \cdots + f_{11}^{(1)} \left(\Delta \boldsymbol{x}_{i,t-1}' \boldsymbol{\beta}_0 + \Delta e_{i,t-1} \right) + \Delta \boldsymbol{x}_{it}' \boldsymbol{\beta}_0 + \Delta e_{it}$ (see the proof of Lemma 1).

Using these equations we can write $\Delta \boldsymbol{y}_{i,-j}$ as $\Delta \boldsymbol{y}_{i,-j} = \widetilde{\boldsymbol{A}}_j \Delta \boldsymbol{\xi}_{i1} + \widetilde{\boldsymbol{B}}_j \left(\Delta \boldsymbol{X}_i \boldsymbol{\beta}_0 + \Delta \boldsymbol{e}_i \right)$, where $\widetilde{\boldsymbol{A}}_j$ is a $(T-1) \times p$ matrix consisting of the first $T-1$ rows of $\boldsymbol{A}_j$ (see Eq. (18)) and $\widetilde{\boldsymbol{B}}_j$ is a $(T-1) \times (T-1)$ matrix consisting of the first $T-1$ rows and first $T-1$ columns of $\boldsymbol{B}_j$ (see Eq. (19)). Recall $\left(\Delta y_{i,-p+2}, \ldots, \Delta y_{i1} \right)' = \left[\boldsymbol{I}_p \otimes \left(1, \boldsymbol{x}_i' \right) \right] \boldsymbol{\pi}_0 + \boldsymbol{r}_i$ for $\boldsymbol{\pi}_0 = \left(\mu_{01}, \boldsymbol{\theta}_{01}', \mu_{02}, \boldsymbol{\theta}_{02}', \ldots, \mu_{0,p}, \boldsymbol{\theta}_{0,p}' \right)'$ and $\boldsymbol{r}_i = \left(r_{i1}, \ldots, r_{ip} \right)'$ (see Eq. (10)). Moreover, note that $\Delta \boldsymbol{\xi}_{i1} = \boldsymbol{I}^* \left(\Delta y_{i,-p+2}, \ldots, \Delta y_{i1} \right)'$ for $p \times p$ matrix

$$\boldsymbol{I}^* = \begin{pmatrix} 0 & \cdots & 0 & 1 \\ 0 & \cdots & 1 & 0 \\ \vdots & & \vdots & \vdots \\ 1 & \cdots & 0 & 0 \end{pmatrix}.$$

Let

$$\boldsymbol{D}_j = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \widetilde{\boldsymbol{A}}_j \boldsymbol{I}^* & \widetilde{\boldsymbol{B}}_j \end{pmatrix}.$$

Then some straightforward calculations give

$$\left(\boldsymbol{0}, \quad \Delta \boldsymbol{y}_{i,-j}' \right) \Upsilon_0^{-1} \widetilde{\boldsymbol{u}}_i = \left(\boldsymbol{\beta}_0', \boldsymbol{\pi}_0' \right) \widetilde{\boldsymbol{Z}}_i' \boldsymbol{D}_j' \Upsilon_0^{-1} \widetilde{\boldsymbol{u}}_i + \widetilde{\boldsymbol{u}}_i' \boldsymbol{D}_j' \Upsilon_0^{-1} \widetilde{\boldsymbol{u}}_i. \tag{27}$$

Because the elements of $\widetilde{\boldsymbol{u}}_i$ are uncorrelated with the elements of $\widetilde{\widetilde{\boldsymbol{Z}}}_i$, we have $E\left[\left(\boldsymbol{\beta}_0', \boldsymbol{\pi}_0' \right) \widetilde{\boldsymbol{Z}}_i' \boldsymbol{D}_j' \Upsilon_0^{-1} \widetilde{\boldsymbol{u}}_i \right] = 0$. Also, $E\left(\widetilde{\boldsymbol{u}}_i' \boldsymbol{D}_j' \Upsilon_0^{-1} \widetilde{\boldsymbol{u}}_i \right) = \operatorname{tr}\left[\Upsilon_0^{-1} E\left(\widetilde{\boldsymbol{u}}_i \widetilde{\boldsymbol{u}}_i' \right) \boldsymbol{D}_j' \right] = \operatorname{tr}\left(\boldsymbol{D}_j' \right)$. But $\operatorname{tr}\left(\boldsymbol{D}_j' \right) = 0$, because the upper left-hand submatrix $\boldsymbol{0}$ in $\boldsymbol{D}_j$ is square with zeros down its main diagonal and $\widetilde{\boldsymbol{B}}_j$ is a square matrix with zeros down its main diagonal, and, therefore, $\boldsymbol{D}_j$ has zeros down its main diagonal. These observations and Eq. (27) prove $E\left[\left(\boldsymbol{0}, \quad \Delta \boldsymbol{y}_{i,-j}' \right) \Upsilon_0^{-1} \widetilde{\boldsymbol{u}}_i \right] = 0$.

References:

Alvarez, J. and M. Arellano, 2003, The time series and cross-section asymptotics of dynamic panel data estimators. Econometrica 71, 1121-1159.

Amemiya, T. 1985, Advanced econometrics, Harvard University Press, Cambridge, MA.

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

Arellano, M. and S. Bond, 1991, Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. The Review of Economic Studies 58, 277-297.

Blundell, R. and S. Bond, 1998, Initial conditions and moment restrictions in dynamic panel data models. Journal of Econometrics 87, 115-143.

Greene, W. H., 2012, Econometric analysis, 7th ed., Pearson Education, Upper Saddle River, NJ.

Hamilton, J. D., 1994, Time series analysis, Princeton University Press, Princeton, NJ.

Hsiao, C., Pesaran, H. M. and A. K. Tahmiscioglu, 2002, Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods. Journal of econometrics 109, 107-150.

Kruiniger, H., 2013, Quasi ML estimation of the panel AR(1) model with arbitrary initial conditions. Journal of Econometrics 173, 175-188.

Liu, C. and D. B. Rubin, 1994, The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. Biometrika 81, 633-648.

Loève, M., 1977, Probability theory I, 4th ed. Springer-Verlag, New York, NY.

Nerlove, M., 1971, Further evidence on the estimation of dynamic relations from a time series of cross sections. Econometrica 39, 359-382.

Newey, W. K. and D. McFadden, 1994, Large sample estimation and hypothesis testing, in: R. F. Engle and D. L. McFadden, (Eds.), Handbook of econometrics, Vol. 4 North Holland, Amsterdam, pp. 2111-2245.

Phillips, R. F., 2004, Estimation of a generalized random-effects model: Some ECME algorithms and Monte Carlo evidence. Journal of Economic Dynamics and Control 28, 1801-1824.

Phillips, R. F., 2010, Iterated feasible generalized least-squares estimation of augmented dynamic panel data models. Journal of Business and Economic Statistics 28, 410-422.

Phillips, R. F., 2012, On computing maximum-likelihood estimates of the unbalanced two-way random-effects model. Communications in Statistics – Simulation and Computation 41, 1921-1927.

Rao, C. R., 1973, Linear statistical inference and its applications, Wiley & Sons, New York, NY.

Ruud, P. A., 2000, An introduction to classical econometric theory, Oxford University Press, New York, NY.

White, H., 2001, Asymptotic theory for econometricians, Academic Press, New York.

Wooldridge, J. M., 2010, Econometric analysis of cross section and panel data, 2nd ed., MIT Press, Cambridge, MA.

Table 1: Simulation Designs

Design No.	Design Parameters t_0	δ_0	κ	σ_ζ	Design No.	Design Parameters t_0	δ_0	κ	σ_ζ
1	50	0.4	0	1.0	9	1	0.4	0	1.0
2	50	0.4	0	2.0	10	1	0.4	0	2.0
3	50	0.4	1	1.0	11	1	0.4	1	1.0
4	50	0.4	1	2.0	12	1	0.4	1	2.0
5	50	0.9	0	1.0	13	1	0.9	0	1.0
6	50	0.9	0	2.0	14	1	0.9	0	2.0
7	50	0.9	1	1.0	15	1	0.9	1	1.0
8	50	0.9	1	2.0	16	1	0.9	1	2.0

Table 2: Finite sample characteristics of estimators of δ_0 ($t_0 = 50$, $N = 100$)

	Design							
	$\delta_0 = 0.4$				$\delta_0 = 0.9$			
	homoskedastic[a]		heteroskedastic[b]		homoskedastic[a]		heteroskedastic[b]	
	1	2	3	4	5	6	7	8
Ave. $\widehat{\rho}$	0.26	0.32	0.18	0.26	0.09	0.09	0.09	0.10
Ave. R_c^2	*0.74*	*0.89*	*0.63*	*0.81*	*0.92*	*0.97*	*0.89*	*0.96*
QML								
bias	0.0007	0.0023	−0.0028	−0.0031	−0.0025	−0.0020	−0.0015	−0.0048
rmse	*0.0979*	*0.1080*	*0.1270*	*0.1590*	*0.0877*	*0.0899*	*0.1606*	*0.1797*
$\text{DQML}_{\boldsymbol{x}}$								
bias	−0.0006	0.0019	−0.0009	−0.0026	−0.0032	−0.0029	−0.0042	−0.0080
rmse	*0.1215*	*0.1195*	*0.2195*	*0.2199*	*0.0886*	*0.0893*	*0.1701*	*0.1793*
$\text{DQML}_{\Delta\boldsymbol{x}}$								
bias	−0.0010	0.0029	−0.0018	−0.0047	−0.0022	−0.0024	−0.0107	−0.0229
rmse	*0.1280*	*0.1272*	*0.2206*	*0.2176*	*0.1544*	*0.1568*	*0.2551*	*0.2547*
ECME_{he}								
bias	0.0000	0.0003	−0.0027	−0.0015	−0.0173	−0.0161	−0.0381	−0.0367
rmse	*0.0588*	*0.0593*	*0.0715*	*0.0742*	*0.0592*	*0.0587*	*0.0829*	*0.0845*
ECME_{ho}								
bias	−0.0005	−0.0007	−0.0030	−0.0017	−0.0186	−0.0176	−0.0495	−0.0458
rmse	*0.0579*	*0.0586*	*0.0740*	*0.0772*	*0.0596*	*0.0595*	*0.0981*	*0.0991*
IFGLS_{ho}								
bias	−0.0001	−0.0004	−0.0012	−0.0009	0.0031	0.0040	−0.0108	−0.0092
rmse	*0.0581*	*0.0587*	*0.0762*	*0.0785*	*0.0822*	*0.0825*	*0.1218*	*0.1220*
$\text{DQML}_{\boldsymbol{x},\text{ho}}$								
bias	−0.0001	−0.0005	0.0002	−0.0004	0.0016	0.0021	−0.0192	−0.0159
rmse	*0.0585*	*0.0589*	*0.0808*	*0.0801*	*0.0805*	*0.0799*	*0.1180*	*0.1183*
$\text{DQML}_{\Delta\boldsymbol{x},\text{ho}}$								
bias	−0.0001	−0.0004	0.0000	−0.0005	0.0091	0.0095	−0.0132	−0.0105
rmse	*0.0586*	*0.0591*	*0.0805*	*0.0795*	*0.0959*	*0.0953*	*0.1261*	*0.1253*

(Table 2 continued on next page)

Table 2 continued:

	$\delta_0 = 0.4$				$\delta_0 = 0.9$			
	homoskedastic[a]		heteroskedastic[b]		homoskedastic[a]		heteroskedastic[b]	
	1	2	3	4	5	6	7	8
DGMM								
bias	−0.0482	−0.0516	−0.0623	−0.0742	−0.1113	−0.1092	−0.2178	−0.2252
rmse	*0.0912*	*0.0946*	*0.1043*	*0.1173*	*0.1457*	*0.1444*	*0.2592*	*0.2671*
SGMM								
bias	−0.0025	0.0497	−0.0380	−0.0049	0.0585	0.0782	0.0299	0.0628
rmse	*0.0703*	*0.0965*	*0.0862*	*0.0853*	*0.0661*	*0.0814*	*0.0568*	*0.0718*
OLS								
bias	0.2186	0.2766	0.1601	0.2373	0.0526	0.0568	0.0370	0.0429
rmse	*0.2231*	*0.2804*	*0.1688*	*0.2453*	*0.0619*	*0.0655*	*0.0536*	*0.0579*

[a]For Designs 1, 2, 5, and 6, the errors are conditionally homoskedastic.
[b]For Designs 3, 4, 7, and 8, the errors are conditionally heteroskedastic, but unconditionally homoskedastic.

Table 3: Finite sample characteristics of estimators of δ_0 ($t_0 = 50$, $N = 500$)

	Design							
	$\delta_0 = 0.4$				$\delta_0 = 0.9$			
	homoskedastic[a]		heteroskedastic[b]		homoskedastic[a]		heteroskedastic[b]	
	1	2	3	4	5	6	7	8
Ave. $\widehat{\rho}$	0.27	0.34	0.19	0.28	0.09	0.09	0.08	0.08
Ave. R_c^2	*0.71*	*0.88*	*0.58*	*0.79*	*0.93*	*0.98*	*0.89*	*0.96*
QML								
bias	0.0006	0.0016	−0.0013	0.0006	0.0009	0.0007	−0.0015	0.0006
rmse	*0.0410*	*0.0451*	*0.0509*	*0.0598*	*0.0366*	*0.0373*	*0.0592*	*0.0620*
DQML$_{\boldsymbol{x}}$								
bias	0.0001	0.0014	−0.0007	−0.0010	0.0002	−0.0001	−0.0030	−0.0010
rmse	*0.0497*	*0.0489*	*0.0743*	*0.0742*	*0.0373*	*0.0373*	*0.0636*	*0.0634*
DQML$_{\Delta\boldsymbol{x}}$								
bias	0.0000	0.0018	−0.0006	−0.0022	0.0005	0.0008	−0.0051	−0.0038
rmse	*0.0529*	*0.0517*	*0.0766*	*0.0762*	*0.0586*	*0.0594*	*0.0890*	*0.0871*
ECME$_{\text{he}}$								
bias	0.0002	−0.0002	−0.0014	0.0001	−0.0055	−0.0054	−0.0182	−0.0159
rmse	*0.0257*	*0.0260*	*0.0325*	*0.0330*	*0.0289*	*0.0299*	*0.0412*	*0.0418*
ECME$_{\text{ho}}$								
bias	0.0002	−0.0002	−0.0012	0.0005	−0.0056	−0.0054	−0.0214	−0.0193
rmse	*0.0257*	*0.0258*	*0.0335*	*0.0347*	*0.0291*	*0.0299*	*0.0486*	*0.0496*
IFGLS$_{\text{ho}}$								
bias	0.0005	−0.0000	−0.0007	0.0008	0.0015	0.0012	−0.0031	−0.0011
rmse	*0.0257*	*0.0258*	*0.0337*	*0.0348*	*0.0348*	*0.0353*	*0.0602*	*0.0625*
DQML$_{\boldsymbol{x},\text{ho}}$								
bias	0.0005	−0.0000	−0.0003	0.0009	0.0013	0.0009	−0.0034	−0.0020
rmse	*0.0260*	*0.0259*	*0.0352*	*0.0354*	*0.0349*	*0.0352*	*0.0621*	*0.0623*
DQML$_{\Delta\boldsymbol{x},\text{ho}}$								
bias	0.0005	−0.0000	−0.0003	0.0009	0.0031	0.0032	0.0010	0.0014
rmse	*0.0260*	*0.0259*	*0.0351*	*0.0353*	*0.0412*	*0.0422*	*0.0715*	*0.0708*

(Table 3 continued on next page)

Table continued 3:

	Design							
	$\delta_0 = 0.4$				$\delta_0 = 0.9$			
	homoskedastic[a]		heteroskedastic[b]		homoskedastic[a]		heteroskedastic[b]	
	1	2	3	4	5	6	7	8
DGMM								
bias	−0.0095	−0.0109	−0.0136	−0.0156	−0.0244	−0.0244	−0.0597	−0.0621
rmse	*0.0352*	*0.0373*	*0.0414*	*0.0459*	*0.0482*	*0.0496*	*0.0889*	*0.0909*
SGMM								
bias	0.0117	0.0113	0.0059	0.0075	0.0637	0.0704	0.0489	0.0597
rmse	*0.0302*	*0.0317*	*0.0333*	*0.0351*	*0.0650*	*0.0721*	*0.0538*	*0.0644*
OLS								
bias	0.2351	0.2931	0.1787	0.2585	0.0645	0.0686	0.0505	0.0563
rmse	*0.2359*	*0.2938*	*0.1802*	*0.2600*	*0.0660*	*0.0700*	*0.0531*	*0.0587*

[a]For Designs 1, 2, 5, and 6, the errors are conditionally homoskedastic.
[b]For Designs 3, 4, 7, and 8, the errors are conditionally heteroskedastic, but unconditionally homoskedastic.

Table 4: Finite sample characteristics of estimators of δ_0 ($t_0 = 1$, $N = 100$)

	Design							
	$\delta_0 = 0.4$				$\delta_0 = 0.9$			
	homoskedastic[a]		heteroskedastic[b]		homoskedastic[a]		heteroskedastic[b]	
	9	10	11	12	13	14	15	16
Ave. $\widehat{\rho}$	0.39	0.54	0.21	0.45	0.39	0.54	0.23	0.43
Ave. R_c^2	*0.54*	*0.73*	*0.29*	*0.22*	*0.54*	*0.73*	*0.36*	*0.25*
QML								
bias	0.0000	0.0016	−0.0020	−0.0018	−0.0002	0.0000	−0.0020	0.0080
rmse	*0.0681*	*0.0742*	*0.0737*	*0.0802*	*0.0330*	*0.0213*	*0.1091*	*0.1384*
DQML$_{\boldsymbol{x}}$								
bias	0.0005	0.0015	−0.0009	0.0005	−0.0003	0.0005	−0.0013	0.0052
rmse	*0.0723*	*0.0756*	*0.1838*	*0.1840*	*0.0369*	*0.0423*	*0.1181*	*0.1303*
DQML$_{\Delta\boldsymbol{x}}$								
bias	0.0006	0.0015	−0.0035	−0.0026	0.0001	0.0026	−0.0081	−0.0030
rmse	*0.0729*	*0.0758*	*0.1740*	*0.1745*	*0.0585*	*0.0747*	*0.1610*	*0.1838*
ECME$_{\text{he}}$								
bias	−0.0012	0.0001	−0.0027	−0.0012	−0.0010	0.0001	−0.0126	0.0072
rmse	*0.0466*	*0.0437*	*0.0553*	*0.0569*	*0.0281*	*0.0181*	*0.0807*	*0.0715*
ECME$_{\text{ho}}$								
bias	−0.0010	0.0002	−0.0233	−0.0296	−0.0006	0.0002	−0.1415	−0.0854
rmse	*0.0458*	*0.0429*	*0.0723*	*0.0716*	*0.0293*	*0.0185*	*0.1763*	*0.1072*
IFGLS$_{\text{ho}}$								
bias	−0.0009	0.0002	−0.0221	−0.0295	−0.0005	0.0002	−0.1203	−0.0846
rmse	*0.0458*	*0.0429*	*0.0739*	*0.0716*	*0.0293*	*0.0185*	*0.1895*	*0.1083*
DQML$_{\boldsymbol{x},\text{ho}}$								
bias	−0.0009	0.0002	−0.0371	−0.0345	−0.0005	0.0002	−0.1424	−0.0853
rmse	*0.0459*	*0.0429*	*0.0758*	*0.0727*	*0.0296*	*0.0186*	*0.1808*	*0.1073*
DQML$_{\Delta\boldsymbol{x},\text{ho}}$								
bias	−0.0009	0.0002	−0.0371	−0.0345	−0.0005	0.0002	−0.1551	−0.0873
rmse	*0.0459*	*0.0429*	*0.0758*	*0.0727*	*0.0298*	*0.0186*	*0.1847*	*0.1083*
DGMM								
bias	−0.0305	−0.0267	−0.0198	−0.0260	−0.0132	−0.0054	−0.1018	−0.0721
rmse	*0.0665*	*0.0644*	*0.0629*	*0.0723*	*0.0341*	*0.0207*	*0.1372*	*0.1022*

[a]For Designs 9, 10, 13, and 14, the errors are unconditionally homoskedastic.

[b]For Designs 11, 12, 15, and 16, the errors are unconditionally heteroskedastic.

Table 5: Finite sample characteristics of estimators of δ_0 ($t_0 = 1$, $N = 500$)

	Design							
	$\delta_0 = 0.4$				$\delta_0 = 0.9$			
	homoskedastic[a]		heteroskedastic[b]		homoskedastic[a]		heteroskedastic[b]	
	9	10	11	12	13	14	15	16
Ave. $\widehat{\rho}$	0.40	0.55	0.21	0.46	0.41	0.55	0.22	0.46
Ave. R_c^2	*0.52*	*0.72*	*0.22*	*0.17*	*0.52*	*0.72*	*0.23*	*0.17*
QML								
bias	−0.0002	0.0000	−0.0008	0.0003	−0.0002	0.0001	0.0021	0.0010
rmse	*0.0295*	*0.0308*	*0.0319*	*0.0338*	*0.0142*	*0.0092*	*0.0437*	*0.0472*
DQML$_{\boldsymbol{x}}$								
bias	−0.0002	0.0000	−0.0009	0.0018	−0.0001	0.0005	0.0017	0.0002
rmse	*0.0312*	*0.0315*	*0.0583*	*0.0592*	*0.0158*	*0.0159*	*0.0472*	*0.0467*
DQML$_{\Delta\boldsymbol{x}}$								
bias	−0.0001	−0.0000	−0.0012	0.0015	−0.0001	0.0006	0.0004	−0.0008
rmse	*0.0313*	*0.0317*	*0.0584*	*0.0594*	*0.0225*	*0.0229*	*0.0552*	*0.0565*
ECME$_{\text{he}}$								
bias	−0.0003	−0.0002	−0.0011	−0.0006	−0.0004	0.0001	−0.0016	0.0007
rmse	*0.0209*	*0.0194*	*0.0253*	*0.0259*	*0.0121*	*0.0079*	*0.0359*	*0.0302*
ECME$_{\text{ho}}$								
bias	−0.0002	−0.0002	−0.0212	−0.0290	−0.0003	0.0001	−0.1445	−0.0874
rmse	*0.0205*	*0.0192*	*0.0378*	*0.0417*	*0.0128*	*0.0081*	*0.1522*	*0.0921*
IFGLS$_{\text{ho}}$								
bias	−0.0001	−0.0002	−0.0209	−0.0290	−0.0002	0.0001	−0.1427	−0.0874
rmse	*0.0205*	*0.0192*	*0.0377*	*0.0417*	*0.0128*	*0.0081*	*0.1531*	*0.0921*
DQML$_{\boldsymbol{x},\text{ho}}$								
bias	−0.0001	−0.0002	−0.0351	−0.0339	−0.0002	0.0001	−0.1481	−0.0875
rmse	*0.0205*	*0.0192*	*0.0462*	*0.0449*	*0.0129*	*0.0081*	*0.1553*	*0.0922*
DQML$_{\Delta\boldsymbol{x},\text{ho}}$								
bias	−0.0001	−0.0002	−0.0351	−0.0339	−0.0003	0.0001	−0.1573	−0.0892
rmse	*0.0205*	*0.0192*	*0.0462*	*0.0449*	*0.0130*	*0.0081*	*0.1634*	*0.0937*
DGMM								
bias	−0.0064	−0.0062	−0.0053	−0.0063	−0.0029	−0.0011	−0.0220	−0.0182
rmse	*0.0273*	*0.0261*	*0.0259*	*0.0291*	*0.0138*	*0.0086*	*0.0457*	*0.0382*

[a]For Designs 9, 10, 13, and 14, the errors are unconditionally homoskedastic.

[b]For Designs 11, 12, 15, and 16, the errors are unconditionally heteroskedastic.