Quasi maximum likelihood estimation of dynamic panel data models

Robert F. Phillips
Department of Economics, George Washington University, Washington, DC, USA

ABSTRACT
This article establishes the almost sure convergence and asymptotic normality of levels and differenced quasi maximum likelihood (QML) estimators of dynamic panel data models. The QML estimators are robust with respect to initial conditions, conditional and time-series heteroskedasticity, and misspecification of the log-likelihood. The article also provides an ECME algorithm for calculating levels QML estimates. Finally, it compares the finite-sample performance of levels QML estimators, the differenced generalized method of moments (GMM) estimator, and the system GMM estimator. The QML estimators usually have smaller—typically substantially smaller—bias and root mean squared errors than the panel data GMM estimators.

1. Introduction

Two prominent approaches to estimating a dynamic panel data model are generalized method of moments (GMM) and maximum likelihood (ML). Several authors have studied ML estimation of dynamic panel data models; see, for example, Alvarez and Arellano (2004), Anderson and Hsiao (1981), Hsiao, Pesaran, and Tahmiscioglu (2002), and Moral-Benito (2013), among others. As is well known, the consistency and asymptotic normality of an ML estimator follows from ML theory assuming the likelihood is correctly specified and standard regularity conditions are met. On the other hand, strong distributional assumptions are not required to establish the sampling behavior of a GMM estimator. This fact would appear to make GMM more attractive than ML. But GMM has its drawbacks as well. For example, some GMM estimators are known to suffer from severe finite-sample bias (see, e.g., Blundell and Bond 1998). Furthermore, some papers have shown that the maximizer of a log-likelihood for a panel data model can be consistent and asymptotically normal under assumptions that do not require normality. Binder, Hsiao, and Pesaran (2005), for example, considered quasi-ML (QML) estimation of vector panel autoregressions. Kruiniger (2013), on the other hand, studied QML estimation of a first-order autoregressive (AR(1)) panel data model. And Phillips (2010, 2015) examined QML estimation of a \( p \)th-order dynamic panel data model. These papers provide conditions under which the log-likelihood for a dynamic panel data model can be misspecified, and the maximizer of the quasi log-likelihood is nevertheless consistent and asymptotically normal.

This article makes several contributions to the literature on QML estimation. Like Phillips (2010, 2015), the model studied in this article includes \( p \) lags of the dependent variable as well
as other explanatory variables. Phillips (2010, 2015), however, focused on QML estimation without differencing the observations—i.e., levels QML (LQML)—while assuming the errors are unconditionally homoskedastic. The assumption of unconditional homoskedasticity is more general than it might first appear, for it allows for conditional heteroskedasticity. But it does not allow for time-series heteroskedasticity. Allowing for general forms of heteroskedasticity is important, for QML estimation, although robust with respect to initial conditions and misspecification of the log-likelihood, is not robust to misspecification of the unconditional error variance–covariance matrix (see also Alvarez and Arellano 2004). This article, therefore, provides large $N$, fixed $T$ asymptotics under more general conditions than those examined in Phillips (2010, 2015)—conditions that allow for time-series heteroskedasticity. Indeed, the error variance–covariance matrix can be of a general form.

Phillips (2010) provided a straightforward iterative feasible generalized least-squares (FGLS) algorithm for calculating QML estimates when the errors in the dynamic regression model have an error-component structure. However, that procedure is not easily extended to the case where the idiosyncratic errors are time-series heteroskedastic. Furthermore, derivative-based algorithms can produce negative fitted variance components when applied to error-component models if they are not substantially modified to avoid that outcome (see also Meng and van Dyk 1998). This article improves on these algorithms by providing an ECME (expectation-conditional maximization either) algorithm for calculating LQML estimates that allows for conditional and time-series heteroskedasticity. The ECME algorithm is straightforward and guarantees non-negative estimated variance components.

The article also examines QML estimation after differencing the observations (differenced QML). It shows that the ML estimator examined by Hsiao, Pesaran, and Tahmiscioglu (2002) is consistent and asymptotically normal under more general conditions than the conditions considered by Hsiao, Pesaran, and Tahmiscioglu (2002). For example, Hsiao, Pesaran, and Tahmiscioglu (2002) assumed normality. This article shows the estimator can be consistent and asymptotically normal even if the log-likelihood is misspecified. Moreover, restrictive initial conditions are not required, and the errors can be conditionally heteroskedastic.

Finally, using simulated data, the finite-sample behavior of levels and differenced QML estimators are compared, and their finite-sample behavior is compared to the differenced GMM (Arellano and Bond 1991) and the system GMM estimators (Blundell and Bond 1998). The Monte Carlo results show that, compared to GMM estimators, the QML estimators have negligible finite-sample bias, and consequently they have smaller—sometimes much smaller—root mean squared errors.

2. QML via regression augmentation

Since Anderson and Hsiao (1981), it has been known that whether or not application of ML estimation to a dynamic panel data model will yield a consistent estimator as $N \to \infty$, with $T$ fixed, depends on initial conditions. However, Phillips (2010) showed that, when QML estimation is based on observations in levels (henceforth LQML), it does not depend on the initial condition restrictions if the regression is augmented with a suitable control function. This section extends the results in Phillips (2010) by establishing the almost sure convergence and asymptotic normality of LQML estimation under weaker conditions than those used in Phillips (2010). For example, the results provided here allow for more general specifications of the error variance–covariance matrix. This generalization is important because QML estimation is inconsistent if the error variance–covariance matrix is misspecified.
The model examined in this article is the $p$th-order dynamic panel data model

$$y_i = Y_i \delta_0 + X_i \beta_0 + e_i \quad (i = 1, \ldots, N).$$

(1)

In this expression, $y_i = (y_{i1}, \ldots, y_{iT})'$, $Y_i = (y_{i,-1}, \ldots, y_{i,-p})$, $y_{i,j} = (y_{i1-j}, \ldots, y_{iT-j})'$ ($j = 1, \ldots, p$), and $X_i = (x_{i1}, \ldots, x_{iT})'$, with $x_{it}$ a $K \times 1$ vector of explanatory variables that vary with $t$ (for at least some $i$). Moreover, $e_i = (e_{i1}, \ldots, e_{iT})'$ is a vector of regression errors.

For notational convenience, the numbering of observed variables begins with $t = -p + 1$.

Straightforward ML estimation of the model in (1) will not generally yield a consistent estimator. To see why, let $y_i^p = (y_{i0}, \ldots, y_{i-p+1})'$; let $x_i$ be a column vector consisting of all of the distinct elements of $x_{i1}, \ldots, x_{iT}$; and set $z_i = (x_i', y_i^p)'.$ Then, assuming $e_i|z_i \sim \mathcal{N}(0, \Omega_0^p)$, the log-likelihood is given by

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

(2)

where $e_i(\varphi) = y_i - Y_i \delta - X_i \beta$, and $\varphi = (\delta', \beta')'$. If $\Omega_0^p$ were known, then maximizing the log-likelihood in (2) yields the generalized least-squares (GLS) estimator based on $\Omega_0^p$, and the consistency of that estimator requires $E(X_i' \Omega_0^{p^{-1}} e_i) = 0$ and $E(y_{i,j}' \Omega_0^{p^{-1}} e_i) = 0$ ($j = 1, \ldots, p$).

We have $E(X_i' \Omega_0^{p^{-1}} e_i) = 0$ if the regressors in $X_i$ are strictly exogenous with respect to the errors in $e_i$. But the moment restrictions $E(y_{i,j}' \Omega_0^{p^{-1}} e_i) = 0$ ($j = 1, \ldots, p$) depend on an even stronger assumption, which is summarized in Lemma 1.

**Lemma 1.** If $E(e_i y_i^p) = 0$, $E(e_i x_i') = 0$, and $E(e_i e_i') = \Omega_0^p$, then $E(y_{i,j}' \Omega_0^{p^{-1}} e_i) = 0$ ($j = 1, \ldots, p$).

The proof of Lemma 1 is available in the working paper version (Phillips 2017) of this paper.

According to Lemma 1, if the regressors in $x_{it}$ and the initial values of the dependent variable $y_{i0}, \ldots, y_{i-p+1}$ are uncorrelated with the errors $e_{i1}, \ldots, e_{iT}$, then $E(y_{i,j}' \Omega_0^{p^{-1}} e_i) = 0$ ($j = 1, \ldots, p$). However, assuming the initial values of the dependent variable are uncorrelated with subsequent errors is quite restrictive. For example, a commonly used model for the errors is the error-component model

$$e_{it} = c_i + v_{it}.$$  

(3)

If the $v_{it}$s are uncorrelated, we can take $v_{it}$ to be uncorrelated with the elements of $y_i^p$, for $t \geq 1$, but assuming the elements of $y_i^p$ 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 $x_{it}$ are strictly exogenous with respect to the $e_i$s. The possible correlation between the elements in $e_i$ and the elements in $z_i$ can be controlled for by the linear projection of $e_{it}$ on 1 and $z_i$:

$$e_{it} = \mu_0 + z_i' \theta_0 + u_{it}, \quad (t = 1, \ldots, T, \ i = 1, \ldots, N)$$

(4)

where $\theta_0 = Var(z_i)^{-1} \text{Cov}(z_i, e_{it})$ and $\mu_0 = E(e_{it}) - E(z_i)' \theta_0$.

The linear projection parameters $\mu_0$ and $\theta_0$ exist and depend on neither $i$ nor $t$ if $E(e_{it})$ and the moments in $\text{Cov}(z_i, e_{it})$ depend on neither $i$ nor $t$ and the moments in $Var(z_i)$ and $E(z_i)$ do not depend on $i$. The restriction that the linear projection parameters are independent of $t$ is met if the errors have a one-way error-component structure given by (3) and
$v_i$ is a mean zero random variable that is uncorrelated with the elements of $z_i$ for $t \geq 1$. Then, $\text{Cov}(z_i, e_t) = \text{Cov}(z_i, c_i)$ and $E(e_{it}) = E(c_i)$ for $t \geq 1$. For this case, the linear projection reduces to that considered in Phillips (2010, 2015). Specifically, we have

$$c_i = \mu_0 + z_i'\theta_0 + a_i \quad (i = 1, \ldots, N) \tag{5}$$

(cf., Phillips 2010, p. 411, Equation (2)).$^1$ If the errors can be decomposed as in Equation (3), then $\mu_0 + z_i'\theta_0$ controls for possible correlation between time-invariant unobservables, captured by $c_i$, and the elements of $z_i$.

Another, albeit trivial, case in which the linear projection parameters depend on neither $i$ nor $t$ is when there are no individual specific effects and the $e_{it}s$ are uncorrelated among themselves and with the elements of $z_i$, for $t \geq 1$. In this case, $\theta_0 = 0$, and the linear projection in (4) simplifies to $e_{it} = \mu_e + u_{it}$, where $E(e_{it}) = \mu_e$. This example illustrates that the necessity of adding the control function $\mu_0 + z_i'\theta_0$ follows from the presence of unobservable time-invariant omitted variables, which are captured by $c_i$.

Moreover, although it is obvious we must include $x_i$ in the control function when the regressors in $x_{it}$ are correlated with $c_i$, it is also true that we typically must do so even when all of the regressors in $x_{it}$ are uncorrelated with $c_i$, as in the random-effects model. To see this, consider the linear projection of $c_i$ on just 1 and $y_i$:

$$c_i = \mu_{yo} + y_i'\theta_{yo} + a_{yi} \quad (i = 1, \ldots, N), \tag{6}$$

where $\theta_{yo} = \text{Var}(y_i) - 1^{-1}\text{Cov}(y_i', c_i)$ and $\mu_{yo} = E(c_i) - E(y_i)\theta_{yo}$. If we augment the model in (1) with the control function $\mu_{yo} + y_i'\theta_{yo}$ rather than the control function $\mu_0 + z_i'\theta_0$, then the error term in the augmented model is $a_{yi} + u_{it}$ rather than $a_i + v_{it}$, and, in order for QML estimation of the augmented model to be consistent, we must have not just $\text{Cov}(y_i, a_{yi}) = 0$, which the linear projection in (6) ensures, but also $\text{Cov}(x_i, a_{yi}) = 0$, which the linear projection in (6) does not guarantee. Indeed, given $\text{Cov}(x_i, c_i) = 0$, the result $\text{Cov}(x_i, a_{yi}) = 0$ is not guaranteed unless $\text{Cov}(x_i, y_i')\theta_{yo} = 0$, which will not be satisfied in general assuming $\theta_{yo} \neq 0$.

This last example illustrates that results obtained for the AR(1) panel data model (see Kruiniger 2013) or the AR(p) panel data model (see Alvarez and Arellano 2004) do not extend in a straightforward manner to models with additional regressors even under the random-effects assumption that the elements of $x_{it}$ are uncorrelated with $c_i$. For example, in his treatment of the “random effects” case of the AR(1) panel data model, Kruiniger includes a linear projection of $c_i$ on the initial value $y_{0i}$ in a control function. However, such a control function will not suffice if there are additional regressors even when these additional regressors are uncorrelated with $c_i$.

Equations (1) and (4) imply the augmented dynamic panel data model

$$y_i = W_iy_0 + u_i \quad (i = 1, \ldots, N), \tag{7}$$

where $W_i = (Y_i, Z_i)$, $Z_i = (X_i, t \cdot t z_i')$, $t$ is a $T \times 1$ vector of ones, and $y_0 = (\delta_0', \beta_0', \mu_0, \theta_0')'$. The errors in this augmented model—$u_i = (u_{i1}, \ldots, u_{iT})'$—are now uncorrelated with the elements of $Z_i$ by construction. Thus, upon letting $\Omega_0 = E(u_i'u_i)$, we have $E(Z_i'\Omega_0^{-1}u_i) = 0$. Moreover, because $E(u_i'y_i') = 0$ and $E(u_i'x_i') = 0$, it follows from Lemma 1 that $E(y_{it-1}'\Omega_0^{-1}u_i) = 0$ (j = 1, . . . , p). The preceding shows $E(W_i'y_i'\Omega_0^{-1}u_i) = 0$.

$^1$ See also Chamberlain (1982, 1984) and Kruiniger (2013), who use a linear projection of an individual effect on $y_{it}$. The linear projection parameters used in Kruiniger (2013) are implicitly assumed to be independent of $i$.

$^2$ This conclusion follows from $\text{Cov}(x_i, c_i) = \text{Cov}(x_i, c_i - \mu_{yo} - y_i'\theta_{yo}) = -\text{Cov}(x_i, y_i'\theta_{yo})$ if $\text{Cov}(x_i, c_i) = 0$. 
Now consider the quasi log-likelihood for the augmented model in (7): \( \sum_{i=1}^{N} l_i(\psi) \), where
\[
l_i(\psi) = -\frac{T}{2} \ln (2\pi) - \frac{1}{2} \ln |\Omega| - \frac{1}{2} u_i(\psi)' \Omega^{-1} u_i(\psi),
\]
\( u_i(\psi) = y_i - W_i' \psi, \psi = (\delta, \beta', \mu, \theta)', \psi = (\gamma, \omega', \omega, \omega)' \), \( \omega = \text{vech}(\Omega) \), and \( \Omega \) is a positive definite matrix. For known \( \omega_0 = \text{vech}(\Omega_0) \), the maximizer of this log-likelihood is the GLS estimator \( \hat{\psi}_{\text{GLS}} = (\sum_{i=1}^{N} W_i' \Omega_0^{-1} W_i)^{-1} \sum_{i=1}^{N} W_i' \Omega_0^{-1} y_i \), and this estimator is consistent because \( E(W_i' \Omega_0^{-1} u_i) = 0 \). Moreover, if \( \Omega \) is a consistent estimator of \( \Omega_0 \), the feasible GLS (FGLS) estimator \( \hat{\psi}_{\text{FGLS}} = (\sum_{i=1}^{N} W_i' \hat{\Omega}_0^{-1} W_i)^{-1} \sum_{i=1}^{N} W_i' \hat{\Omega}_0^{-1} y_i \) is also consistent.

However, the large \( N \) (fixed \( T \)) distribution of such a FGLS estimator depends on the first-round estimator of \( \psi_0 \) used to estimate \( \Omega_0 \) (see Phillips 2010). An alternative that does not depend on a first-round estimator is to estimate \( \psi_0 = (\gamma, \omega, \omega)' \) by maximizing the quasi log-likelihood \( \sum_{i=1}^{N} l_i(\psi) \).

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 the theorems, set \( L_N(\psi) = N^{-1} \sum_{i=1}^{N} l_i(\psi) \) and \( H_N(\psi) = \partial^2 L_N(\psi)/\partial \psi \partial \psi' \); let \( x_{it} \) denote the \( k \)th element of \( x_t \); and set \( \Omega = \{ \psi = (\gamma, \omega, \omega)' \in \mathbb{R}^m : \Omega \) is positive definite\}.

**Theorem 1.** Assume the following conditions are satisfied:

- **C1:** \( E[|y_i|^{2+\epsilon}] < M \) and \( E[|x_{it}|^{2+\epsilon}] < M \) for all \( i, t, \) and \( k \) and some \( \epsilon > 0 \) and \( M < \infty \);
- **C2:** \( \text{Var}(z_i) = \Omega_z \) for all \( i, \) with \( \Omega_z \) a positive definite matrix, \( E(z_i) = \mu_z \) for all \( i, \) and \( E(e_{ii}) = 0_z \) for all \( i, \) and \( t \);\n- **C3:** \( E(u_i u_i') = \Omega_u \) for all \( i, \) with \( \Omega_u \) a positive definite matrix;
- **C4:** the limits \( \lim_{N \to \infty} N^{-1} \sum_i E(y_{it} y_{it}') \), \( \lim_{N \to \infty} N^{-1} \sum_i E(y_{it} x_{it}) \), and \( \lim_{N \to \infty} N^{-1} \sum_i E(x_{it} x_{it}) \) exist for all \( s, t, j, \) and \( k \); and
- **C5:** the vectors \( (z_1', y_1'), \ldots, (z_N', y_N') \) are independent for all \( N \).

Then, \( E[\partial L_N(\psi_0)/\partial \psi] = 0 \) and the limit \( H(\psi) = \lim_{N \to \infty} E[H_N(\psi)] \) exists. Moreover, if \( H_0 = H(\psi_0) \) is negative definite, then there is a compact subset, say \( \bar{\Omega} \), of \( \Omega \), with \( \psi_0 \) in its interior, and there is a measurable maximizer, \( \hat{\psi} \), of \( L_N(\cdot) \) in \( \bar{\Omega} \) such that \( \hat{\psi} \overset{a.s.}{\to} \psi_0 \) (\( N \to \infty \), \( T \) fixed).

**Theorem 2.** Assume Conditions C2–C5 are satisfied, \( H_0 \) is negative definite, and the following conditions are satisfied:

- **C1′:** \( E[|y_i|^{4+\epsilon}] < M \) and \( E[|x_{it}|^{4+\epsilon}] < M \) for all \( i, t, \) and \( k \) and some \( \epsilon > 0 \) and \( M < \infty \); and
- **C6:** the limit \( \mathcal{I}_0 = \lim_{N \to \infty} N^{-1} \sum_i E[\partial^2 l_i(\psi_0)/\partial \psi \partial \psi'] \) exists and is positive definite.

Then \( \sqrt{N}(\hat{\psi} - \psi_0) \overset{d}{\to} N(0, H_0^{-1} \mathcal{I}_0 H_0^{-1}) \) (\( N \to \infty \), \( T \) fixed).

Proofs of Theorems 1 and 2 are provided in Phillips (2017).

In order for the QML estimator to be consistent and asymptotically normal, it must be the case that the true parameter vector, \( \psi_0 \), uniquely maximizes the expected log-likelihood, at least within a neighborhood of \( \psi_0 \). Conditions C1 through C3 are mild, and they suffice to guarantee that \( \psi_0 \) is indeed a stationary value of the expected log-likelihood. But the fact that \( \psi_0 \) is a stationary value is necessary but not sufficient to ensure it is a unique maximizer of the expected log-likelihood. The matrix \( H_0 \) must also be negative definite. If the log-likelihood \( \sum_{i=1}^{N} l_i(\psi) \) is correctly specified, that is, if \( u_i \) is normally distributed with mean vector 0 and variance–covariance matrix \( \Omega_0 \), conditionally on \( z_i \), then by well-known ML theory, we have \( H_0 = -\mathcal{I}_0 \) and \( H_0 \) exists and is negative definite by virtue of Condition C6. However, even when \( \sum_{i=1}^{N} l_i(\psi) \) is misspecified, \( H_0 \) can be shown to be negative definite in particular cases.
Phillips (2015), for example, provides an example in which $H_0$ is negative definite under conditions that do not include normality.

Moreover, $\Omega_0$ is the unconditional variance–covariance matrix of $u$, and, although it does not depend on $i$, the variance–covariance matrix of $u$ conditionally on $z_i$ may depend on $i$—for example, the errors may be conditionally heteroskedastic (see also Phillips 2010, 2015). The errors can also be unconditionally time-series heteroskedastic, for the diagonal elements of $\Omega_0$ can differ.

Furthermore, the conditions in Theorems 1 and 2 do not require the random vectors $(z'_{1}, y'_{1}), \ldots, (z'_{N}, y'_{N})'$ be drawn from a common distribution. On the other hand, Conditions C2 and C3 imply some homogeneity is required.

Estimators previously considered in the literature are covered by Theorems 1 and 2. Blundell and Bond (1998) considered a conditional GLS estimator of an AR(1) panel data model that relied on augmenting the regression model with the initial observation on the dependent variable. They argued that if the error components are homoskedastic across individuals and time, then restrictions on the initial conditions can be used to derive the GLS estimator. Theorems 1 and 2, however, show that these conditions are unnecessarily restrictive. The errors can be conditionally and time-series heteroskedastic. Moreover, initial condition restrictions are not needed. All that is required is that the moments defining the control function parameters exist and depend on neither $i$ nor $t$. Furthermore, the structured error variance–covariance matrices, such as those considered by Phillips (2010, 2015) and Kruiniger (2013), are special cases of $\Omega_0$, and, therefore, Theorems 1 and 2 cover those cases.

### 3. Fixed-effects QML

An alternative to first augmenting the regression model with a control function and then applying QML estimation to the model in levels is to instead first difference the observations and then apply QML estimation. In the literature, ML or QML estimation based on first differencing the observations has been referred to as fixed-effects ML estimation (e.g., Hsiao, Pesaran, and Tahmiscioglu 2002) or fixed-effects QML estimation (e.g., Kruiniger 2013). This description, however, should not lead one to interpret LQML estimation as random-effects QML, for the results in Section 2 make clear that LQML estimation is not restricted to random-effects models with regressors that are exogenous with respect to $c_i$.

Kruiniger (2013) studied differenced QML for an AR(1) panel data model. Hsiao, Pesaran, and Tahmiscioglu (2002), on the other hand, studied ML estimation, after differencing, and, like this article, considered a model with additional explanatory variables beyond a lagged dependent variable. This section shows that likelihood-based methods using differences are consistent and asymptotically normal under much weaker conditions than those assumed in Hsiao, Pesaran, and Tahmiscioglu (2002).

Instead of augmenting the regression with a control function that involves $y^p$, differenced QML requires the estimation of a system of equations that includes a separate linear projection for each initial difference $\Delta y_{i,-p+2}, \ldots, \Delta y_{i,t}$, where $\Delta y_{it} = y_{it} - y_{i,t-1}$. Specifically, suppose $Var(x_i)$ is positive definite, and set $\theta_{0,p+1-j} = Var(x_i)^{-1}Cov(x_i, \Delta y_{i,-j+2})$ and $\mu_{0,p+1-j} = E(\Delta y_{i,-j+2}) - E(x_i')\theta_{0,p+1-j}$ $(j = 1, \ldots, p)$. Then, system differenced QML relies on estimating the linear projections

$$\Delta y_{i,-j+2} = \mu_{0,p+1-j} + x_i'\theta_{0,p+1-j} + r_{i,p+1-j} \quad \text{ (} j = 1, \ldots, p \text{).} \tag{8}$$

Here $r_{i,p+1-j}$ is a linear projection residual, which is, by construction, uncorrelated with all of the elements of $x_i$. Note that because the linear projection in (8) does not specify how $\Delta y_{i,-j+2}$
was generated, it does not depend on initial condition restrictions. In addition to the linear projection equations in (8), we also estimate the differenced equation:

$$\Delta y_i = \Delta Y_i \delta_0 + \Delta X_i \beta_0 + \Delta e_i \quad (i = 1, \ldots, N),$$

(9)

where $\Delta y_i = (\Delta y_{i,2}, \ldots, \Delta y_{i,T})'$, $\Delta Y_i = (\Delta y_{i,-1}, \ldots, \Delta y_{i,-p})$, and $\Delta y_{i,j} = (\Delta y_{i,j+2}, \ldots, \Delta y_{i,T-1})'$ ($j = 1, \ldots, p$). Moreover, $\Delta X_i = (\Delta x_{i,2}, \ldots, \Delta x_{i,T})'$, $\Delta x_{it} = x_{it} - x_{i,t-1}$, and $\Delta e_i = (\Delta e_{i,2}, \ldots, \Delta e_{iT})'$, with $\Delta e_t = e_t - e_{t-1}$. For differenced QML, the equations in (8) and (9) are estimated as a system given by

$$\tilde{y}_i = \tilde{W}_i \eta_0 + \tilde{u}_i \quad (i = 1, \ldots, N),$$

(10)

with $\tilde{y}_i = (\Delta y_{i,-p+2}, \ldots, \Delta y_{i1}, \Delta y_i)', \tilde{u}_i = (r_{i1}, \ldots, r_{ip}, \Delta e_i)',$

$$\tilde{W}_i = \left( \begin{array}{ccc} 0 & 0 & I_p \otimes (1, x_i') \\ \Delta Y_i & \Delta X_i & 0 \end{array} \right),$$

and $\eta_0 = (\delta_0', \beta_0', \mu_{01}, \theta_{01}', \mu_{02}, \theta_{02}', \ldots, \mu_{0p}, \theta_{0p}')'. $

If $\tilde{u}_i$ is multivariate normal with mean vector $0$ and variance–covariance matrix $\Upsilon_0$ conditional on $x_i$, then the log-likelihood for the system in (10) is $\sum_{i=1}^{N} I_i(\lambda)$, where

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

and $\tilde{u}_i(\eta) = \tilde{y}_i - \tilde{W}_i \eta, \eta = (\delta', \beta', \mu, \theta_1, \mu_2, \theta_2', \ldots, \mu_p, \theta_p)', \lambda = (\eta', \nu')', \nu = \text{vech}(\Upsilon).$

Also, set $\tilde{I}_N(\lambda) = N^{-1} \sum_{i=1}^{N} I_i(\lambda), \tilde{H}_N(\lambda) = \partial^2 \tilde{I}_N(\lambda)/\partial \lambda \partial \lambda'$, and $\Lambda = \{\lambda = (\eta', \nu')' \in \mathbb{R}^n : \Upsilon \text{ is positive definite}\}.$

The maximizer of $\sum_{i=1}^{N} I_i(\cdot)$ is a ML estimator given normality, but even if the log-likelihood is misspecified—that is, the errors are not normally distributed given $x_i$, nor are they necessarily conditionally homoskedastic—maximizing $\sum_{i=1}^{N} I_i(\cdot)$ 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($x_i$) = $\Xi_{x x}$ for all i, with $\Xi_{x x}$ positive definite, $E(x_i) = \mu_x$ for all i, $E(\Delta y_{i,j+2}) = \mu_{\Delta y_j}$ and $E(x_i \Delta y_{i,j+2}) = \varphi_{x \Delta y_j}$ for all i ($j = 1, \ldots, p$), and Cov($x_i$, $\Delta e_i) = 0$; also, $E(\tilde{u}_i, \tilde{u}_i') = \Upsilon_0$ for all i, with $\Upsilon_0$ a positive definite matrix.

- Then, $E[\partial \tilde{I}_N(\lambda_0)/\partial \lambda] = 0$, where $\lambda_0 = (\eta_0', \nu_0')$ and $\nu_0 = \text{vech}(\Upsilon_0)$. Furthermore, the limit $\tilde{H}(\lambda) = \lim_{N \to \infty} \tilde{H}_N(\lambda)$ exists. Moreover, if $\tilde{H}_0 = \tilde{H}(\lambda_0)$ is negative definite, there is a compact subset, say $\overline{\Lambda}$, of $\Lambda$, with $\lambda_0$ in its interior, and there is a measurable maximizer, $\hat{\lambda}$, of $\tilde{I}_N(\cdot)$ in $\overline{\Lambda}$ such that $\hat{\lambda} \overset{d}{\to} \lambda_0 \quad (N \to \infty, T \text{ fixed}).$

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

- **C6':** the limit $\tilde{I}_0 = \lim_{N \to \infty} N^{-1} \sum_i E[(\partial \tilde{I}_i(\lambda_0)/\partial \lambda)(\partial \tilde{I}_i(\lambda_0)/\partial \lambda)']$ exists and is positive definite.

Then, $\sqrt{N}(\hat{\lambda} - \lambda_0) \overset{d}{\to} N(0, \tilde{H}_0^{-1} \tilde{I}_0 \tilde{H}_0^{-1}) \quad (N \to \infty, T \text{ fixed}).$

For proofs of Theorems 3 and 4, see Phillips (2017).
The linear projection of $\Delta y_{i,-j+2}$ on 1 and $x_i$ guarantees the residual in this linear projection is uncorrelated with the elements of $\Delta X_i$. 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 $x_i$, where $\Delta x_i$ is a vector consisting of the distinct elements of $\Delta X_i$. The latter approach generalizes an estimator studied by Hsiao, Pesaran, and Tahmiscioglu (2002). Hsiao, Pesaran, and Tahmiscioglu (2002) studied differenced ML estimation of a dynamic panel data model while assuming $p = 1$, individual specific effects, and uncorrelated and conditionally homoskedastic $v_{it}$s. Moreover, Hsiao, Pesaran, and Tahmiscioglu (2002) also imposed restrictions on how the regressors are generated. Furthermore, Hsiao, Pesaran, and Tahmiscioglu (2002) noted that the likelihood satisfies standard regularity conditions, and therefore the ML estimator is consistent and asymptotically normal. However, that conclusion follows from ML theory assuming the log-likelihood is correctly specified. The analysis in this section provides weaker conditions that imply the differenced ML estimator proposed by Hsiao, Pesaran, and Tahmiscioglu (2002) is consistent and asymptotically normal (for $N \to \infty$, $T$ fixed). Specifically, the log-likelihood can be misspecified and the $v_{it}$s can be conditionally heteroskedastic. Moreover, all that is required of the elements of $x_{it}$ is that they be uncorrelated with the $v_{it}$s and that the linear projection of $\Delta y_{it}$ on 1 and $\Delta x_i$ does not depend on $i$.

4. Computation

If the error variance–covariance matrix is unrestricted, QML estimates can be easily computed using iterated FGLS. Consider, for example, calculating QML estimates of the elements of $\Omega_0$ and $y_0$. These estimates can be calculated by iterating back and forth between fitting $\Omega_0$ and fitting $y_0$. Specifically, $L_N(\cdot)$ is maximized with respect to the elements of $\Omega$, conditional on the current fit of the regression parameters, say $y^*$, by the fit $\Omega^+ = \sum_{i=1}^{N} u_i(y^*)u_i(y^*)' / N$. And, after $\Omega^+$ is obtained, $L_N(\cdot)$ is then maximized with respect to $y$, conditional on $\Omega = \Omega^+$, which gives the FGLS fit:

$$y^+ = \left( \sum_{i=1}^{N} W_i(\Omega^+)^{-1} W_i \right)^{-1} \sum_{i=1}^{N} W_i(\Omega^+)^{-1} y_i.$$  (11)

This fit is then made the current fit, $y^*$, and new fits $\Omega^+$ and $y^+$ are calculated again, and so on, until the sequence of fitted values converges. Calculating QML estimates of $\lambda_0$ and $\Upsilon_0$, based on differenced observations, is similar when $\Upsilon_0$ is unrestricted.

Although it is easy to calculate estimates by iterating back and forth between fitting $\Omega_0$ and fitting $y_0$, or between fitting $\lambda_0$ and $\Upsilon_0$, this approach implies that the number of free parameters being fitted in either $\Omega_0$ or $\Upsilon_0$ increases with $T$ as the rate $T^2$ increases. This fact, in turn, suggests that, if $T$ is not quite small, the sampling performance of a QML estimator that does not impose valid restrictions on $\Omega_0$ or $\Upsilon_0$ will be poor compared to that of a QML estimator that does rely on valid restrictions.

Unfortunately, maximizing the likelihood for differenced observations when restrictions on $\Upsilon_0$ are imposed is tractable only for a specialized case. Specifically, we must assume $p = 1$, $e_{it}$ is given by the error-component model in (3), the $v_{it}$s are uncorrelated and unconditionally homoskedastic, and the regressors in $x_{it}$ are strictly exogenous with respect to the $v_{it}$s. Further, assume $\Delta y_{it}$ is generated by the same process generating $\Delta y_{it}$ for $t \geq 2$. Then, it is easy to show
that the error variance–covariance matrix is $\gamma_0 = \sigma_0^2 \Phi_0$,

$$\Phi_0 = \begin{pmatrix} \phi_0 & -1 & 0 & \ldots & 0 \\ -1 & 2 & -1 & \ldots & 0 \\ 0 & -1 & 2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & \ldots & -1 & 2 \end{pmatrix}$$  \hspace{1cm} (12)

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

$$\text{const} - \frac{NT}{2} \ln (\sigma^2) - \frac{N\sigma}{2} - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \tilde{u}_i (\eta)' \Phi^{-1} \tilde{u}_i (\eta)$$

with respect to $\eta$, $\sigma^2$, and $\sigma$. Here, $\Phi$ has $\exp (\sigma) + 1 - 1/T$ in its first row, first column and everywhere else is the same as $\Phi_0$ in (12).

Maximizing the log-likelihood for differenced QML estimation becomes much more complicated if the $v_n$s are time-series heteroskedastic or $p > 1$. On the other hand, the ease with which QML estimates can be calculated is not affected by the size of $p$ nor by whether or not the $v_n$s are time-series heteroskedastic. The remainder of this section is devoted to describing an ECME algorithm that can be applied to calculate QML estimates for arbitrary $p$ and for an error variance–covariance matrix given by $\Omega_0 = \sigma_0^2 \Omega' + \Sigma_0$, with $\Sigma_0 = \text{diag}(\sigma_{01}^2, \ldots, \sigma_{0T}^2)$.

The ECME algorithm relies on conditional or constrained maximization (CM) of either an imputed log-likelihood, based on augmented data, or the log-likelihood based on the observed data. In the present application, the observed data are $y = (y_1', \ldots, y_N')'$, while the augmented data consists of $y$ and $a = (a_1, \ldots, a_N)'$. The imputed log-likelihood is built during the expectation (E) step by taking the conditional expectation of the log-likelihood for the augmented data given the observed data, while treating the current fit of the parameters $\psi^c$ as the parameters of the conditional distribution.

Applying the ECME algorithm to an error-component model for which $\Omega_0 = \sigma_0^2 \Omega' + \Sigma_0$, with $\Sigma_0 = \text{diag}(\sigma_{01}^2, \ldots, \sigma_{0T}^2)$, leads to the following E and CM steps:

\textit{E-step:} Let $(\sigma_a^2)^c$, $\psi'$, and $\Omega^c = (\sigma_a^2)^c \Omega' + \Sigma^c$, with $\Sigma^c = \text{diag}((\sigma_{01}^2)^c, \ldots, (\sigma_{0T}^2)^c)$, denote the current fits of $\sigma_0^2$, $\gamma_0$, and $\Omega_0$. Compute the conditional mean and variance of $a_i$ given $y_t$ evaluated at the current fit of the parameters. These are $a_t^c = (\sigma_a^2)^c \Omega^c (\Omega^c)^{-1} u_t (\psi')$ and $\sigma_a^c = (\sigma_a^2)^c [1 - (\sigma_a^2)^c \Omega^c (\Omega^c)^{-1} t]$, respectively (see, e.g., Greene 2012, Theorem B.7, pp. 1041–1042). Then, the imputed log-likelihood is

$$Q(\psi'; \psi^c) = \text{const} - \frac{N}{2} \left( \ln \sigma_a^2 + \sum_{t=1}^{T} \ln \sigma_t^2 \right) - \frac{1}{2\sigma_a^2} \sum_{i=1}^{N} (a_t^c)^2 - \frac{N}{2\sigma_a^2} v_a^c$$

$$- \frac{1}{2} \sum_{i=1}^{N} \left[ u_t (\psi') - t a_t^c \right]' \Sigma^{-1} \left[ u_t (\psi') - t a_t^c \right] - \frac{N}{2} t' \Sigma^{-1} t v_a^c.$$

\footnote{For the purposes of deriving the imputed log-likelihood and the actual log-likelihood, the variables in $z = (z_1', \ldots, z_N')'$ are treated as fixed.}

\footnote{Liu and Rubin (1994) describe the properties of the ECME algorithm. For applications of it to panel data, see Phillips (2004, 2012).}
CM-step 1: Maximize $Q(\cdot; \Psi)$ with respect to $\omega = (\sigma_\alpha^2, \sigma_\gamma^2, \ldots, \sigma_T^2)'$ subject to the constraint $\gamma = \gamma'$. This step yields $(\sigma_\alpha^2)^+ = v_\alpha^c + \sum_{i=1}^N (a_i^c)^2/N$ and

$$(\sigma_i^2)^+ = v_i^c + \frac{1}{N} \sum_{i=1}^N [u_{it}(\gamma_i^c - a_i^c)]^2 \quad t = 1, \ldots, T. \quad (13)$$

CM-step 2: Maximize the actual log-likelihood $\sum_{i=1}^N l_i(\cdot)$ with respect to $\gamma$ subject to the constraint $\omega = \omega^+$, where $\omega^+ = (\sigma_\alpha^2)^+, (\sigma_\gamma^2)^+, \ldots, (\sigma_T^2)^+)'$. This step gives the FGLS fit in Equation (11) with $\Omega^+ = (\sigma_\alpha^2)^+ \mathbf{u}' + \Sigma^+$ and $\Sigma^+ = \text{diag}((\sigma_\gamma^2)^+, \ldots, (\sigma_T^2)^+)$. After the new fits of the parameters are obtained, they become the current fits, and the preceding steps are repeated, until convergence.

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 the estimates of error-component models, the rate of convergence can slow when the sequence of the fitted variance of the individual-specific effect gets close to zero (see Meng and van Dyk 1998). Moreover, there is always the possibility that the error-component model in (3) is inappropriate; specifically, there may be no individual-specific effects. In this case, we have $\sigma_0^2 = 0$, where $\sigma_0^2 = \text{var}(c_t)$, and $\sigma_\alpha^2 = 0$, and consequently the sequence of fitted values for $\sigma_\alpha^2$ can approach zero. Furthermore, even if $\sigma_\alpha^2$ is positive and large, $\sigma_\alpha^2$ can be small, for the control function $\mu_0 + \mathbf{z}'\theta_0$ is the best linear predictor of $c_t$ based on $\mathbf{z}_t$, and if that predictor is accurate, then $\sigma_\alpha^2$ can be near zero. If so, the sequence of fitted values for $\sigma_\alpha^2$ can get close to zero.

As a practical matter, however, given $\Omega_0 = \sigma_{\alpha 0}^2 \mathbf{u}' + \Sigma_0$, with $\Sigma_0 = \text{diag}(\sigma_{01}^2, \ldots, \sigma_{0T}^2)$, then, when the fitted value for $\sigma_{\alpha 0}^2$ is near zero, the fitted value $\gamma^+$ in (11) differs little from the weighted least-squares fit $(\sum_{i=1}^N \mathbf{W}_i' / \sum_{i=1}^N \mathbf{W}_i / \sum_{i=1}^N \mathbf{W}_i')^{-1} \sum_{i=1}^N \mathbf{W}_i / \sum_{i=1}^N \mathbf{W}_i y_i$, which is obtained by setting $(\sigma_\alpha^2)^+ = 0$. Furthermore, once $(\sigma_\alpha^2)^+$ is set to zero, all subsequent fitted values for $\sigma_\alpha^2$ will be zero. Also, when $(\sigma_\alpha^2)^c = 0$, Equation (13) simplifies to $(\sigma_i^2)^+ = \sum_{i=1}^N u_{it}(\gamma_i^c)^2/N$. Thus, if $(\sigma_\alpha^2)^+$ is set to zero, convergence is rapid. Consequently, the ECME algorithm for computing LQML estimates will generally converge at a robust rate if, as part of the convergence criterion, the size of the fitted value for $\sigma_\alpha^2$ is evaluated and $(\sigma_\alpha^2)^+$ is set to zero should it become sufficiently small.\(^5\)

5. Monte Carlo experiments

5.1. Design

In order to assess the finite-sample 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

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

\(^5\) For example, the fitted value of $\sigma_\alpha^2$ might be set to zero when the fitted value for the average correlation coefficient, say $\overline{\rho}$, is small, where $\overline{\rho} = 2 \sum_{t=1}^{T-1} \sum_{t-1}^T \rho_{t_0} / (T(T-1))$, with $\rho_{t_0} = \sigma_\alpha^2 / \sum (\sigma_\alpha^2 + \sigma_\gamma^2)(\sigma_\alpha^2 + \sigma_\gamma^2)^{1/2}$. This criterion was used to obtain the results for the LQML estimator provided in Section 5.3. In particular, the fitted value of $\sigma_\alpha^2$ was set to zero when the fitted value of $\rho$ fell below 0.01.
with $y_{t-10} = 0$. The values for $\delta_0$ considered were 0, 0.2, 0.4, 0.6, 0.8, and 0.9. Moreover, the $x_{it}$s were generated according to the autoregressive process

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

The starting value $x_{i,-t_0}$ was set equal to $5 + 10\xi_{i,-t_0}$ and the $\xi_{it}$s were generated as independent uniform random variates with mean zero and variance one. Moreover, two values for $t_0$ were considered: $t_0 = 1$ and $t_0 = 50$. For $t_0 = 50$, the time series for $x_{it}$ and $y_{it}$ were essentially stationary, for $t \geq 1$, whereas for $t_0 = 1$ they were non-stationary.

As for the $v_{it}$s, they were generated as $v_{it} = x_{it}(\epsilon_{it} - 5)/\sqrt{T_0}$, with $\epsilon_{it}$ a Chi-square random variate with five degrees of freedom. The variate $(\epsilon_{it} - 5)/\sqrt{10}$ has an asymmetric distribution about zero with a variance of one. Moreover, because the $\epsilon_{it}$s were generated independently of one another and of the $x_{it}$s, the $v_{it}$s were uncorrelated but conditionally heteroskedastic. However, the $v_{it}$s were unconditionally homoskedastic for $t \geq 1$ when $t_0$ was set to 50, for in this case the $x_{it}$s were essentially stationary by the time $t = 1$. On the other hand, for $t_0 = 1$, the $x_{it}$s had insufficient time to become approximately stationary by the time $t = 1$. Hence, in this case, the $v_{it}$s were not only conditionally heteroskedastic, they were also unconditionally time-series heteroskedastic for $t \geq 1$.

The heterogeneity component, $c_i$, was generated as $c_i = \sum_{t=0}^{T} \ln |x_{it}|/(T + 1) + \sigma_\zeta (\zeta_i - 5)/\sqrt{T_0}$, with $\zeta_i$ a Chi-square random variate with five degrees of freedom. Furthermore, the parameter $\sigma_\zeta$ was set to either one or four. This specification for $c_i$ induced correlation between $c_i$ and the $x_{it}$s. Moreover, both $c_i$ and $v_{it}$, conditional on the $x_{it}$s, had non-normal asymmetric distributions, implying that, conditional on the $x_{it}$s, the error $e_{it} = c_i + v_{it}$ came from a non-normal asymmetric distribution.

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 10, and $N$ was set to 200. Finally, for each combination of parameters, 5,000 independent samples were generated.

### 5.2. Estimators

The finite-sample bias and root mean squared error of levels and differenced QML estimators were compared to each other and to well-known GMM estimators. For GMM, I considered differenced GMM (Arellano and Bond 1991) and system GMM (Blundell and Bond 1998). For differenced GMM, I used two-step differenced GMM. Moreover, results are provided for differenced GMM based on all available instruments—henceforth DGMM—and for differenced GMM based on a subset of the available instruments—henceforth DGMMs. System GMM was also based on all available instruments (SGMM) and a subset of the instruments (SGMMs).

Differenced GMM estimators rely on instrument matrices $Z_{di}$ ($i = 1, \ldots, N$), where $Z_{di}$ is a block-diagonal matrix with instrument vector $z_{dit}$ in its $t$th diagonal block ($t = 1, \ldots, T - 1$). The DGMM and DGMMs estimators differed in what was included in $z_{dit}$. Specifically, for the DGMM estimator, $z_{dit}$ was set equal to $(y_{i0}, y_{i1}, \ldots, y_{i,t-1}, x_{i})$ ($t = 1, \ldots, T - 1$). On the other hand, for the DGMMs estimator, I set $z_{dit}$ equal to $(y_{i0}, x_{i0}, x_{i1}, x_{i2})$ for the first diagonal block and $z_{dit}$ to $(y_{i,t-2}, y_{i,t-1}, x_{i,t-1}, x_{it}, x_{i,t+1})$ for the remaining diagonal blocks ($t = 2, \ldots, T - 1$).

Moreover, two-step differenced GMM estimators require one-step estimates first be calculated. For the one-step estimator, let $\Phi_N = N^{-1} \sum_{i=1}^{N} Z_{di}'GZ_{di}$, where $G$ is a $T - 1$ square matrix with twos running down its main diagonal, minus ones in its first subdiagonals,
and zeros everywhere else. Also, let $Z'_d = (Z_{d1}, \ldots, Z'_{dN})$; let $\Delta y' = (\Delta y'_1, \ldots, \Delta y'_N)$, where recall that $\Delta y_i = (\Delta y_{i,t}, \ldots, \Delta y_{i,T})$; and set $X'_d = (X'_{d1}, \ldots, X'_{dN})$, where $X_{d1}$ is a matrix with $(\Delta y_{it}, \Delta x_{i,t+1})$ in its $t$th row $(t = 1, \ldots, T - 1)$. Then, the one-step differenced GMM estimator is

$$\hat{\phi} = (X'_dZ_d\Phi_N^{-1}Z'_dX_d)^{-1}X'_dZ_d\Phi_N^{-1}Z'_d\Delta y. \quad (14)$$

The two-step differenced GMM estimator replaces $\Phi_N^{-1}$ in (14) with an asymptotically optimal weighting matrix. Specifically, $\Phi_N^{-1}$ is replaced with $V_N^{-1}$, where $V_N = N^{-1}\sum_{i=1}^N Z_{di}\Delta\hat{e}_i\Delta\hat{e}'_iZ_{di}$, and $\Delta\hat{e}_i = \Delta y_i - X_{di}\hat{\phi}$ (i.e., optimally weighting matrix. Specifically, let $\Phi_1 = (\Phi_1 \Phi_1, \ldots, \Phi_1 \Phi_1)$ be a block diagonal matrix with $\Phi_i$ for its first diagonal element and $\Phi_i$ in its remaining diagonal blocks $(i = 2, \ldots, T)$.

For SGMM estimates, the $Z_{di}$ block in $Z^+_i$ is the $Z_{di}$ matrix used for the DGMM estimator, whereas the $Z_{di}$ block in $Z^+_i$ is the $Z_{di}$ matrix used for the DGMMs estimator when SGMMs estimates are calculated. On the other hand, for both SGMM and SGMM, $Z_{di}$ is a block-diagonal matrix with $\Delta x_{i1}$ for its first diagonal element and $(\Delta y_{i,t-1}, \Delta x_{it})$ in its $t$th row $(t = 1, \ldots, T)$. Next, set $\hat{\phi} = (y^+_1, \ldots, y^+_N)'$, where $y^+_t = (\Delta y'_t, y'_t)'$. Moreover, let $\hat{\phi} = (X^+_dZ^+_d)^{-1}X^+_dZ^+_d\Delta y$. Then, for a system GMM estimator, replace $(Z^+_dZ^+_d)^{-1}$ in (15) with $(V_N^+)^{-1}$.

Three QML estimators were also considered. Results are provided for LQML while relying on the structured variance–covariance matrix $\Omega_0 = \sigma_{\theta_1}^2\Omega_0 + \Sigma_0$ with $\Sigma_0 = \text{diag}(\sigma_{\delta_1}^2, \ldots, \sigma_{\delta_N}^2)$. For this case, estimates were calculated with the ECME algorithm. Differences QML estimates were also calculated. As noted in Section 4, computing differenced QML estimates via gradient methods is complicated if we model the $v_{ils}$ as time-series heteroskedastic. For this reason, results are only provided for differenced QML estimates assuming the $v_{ils}$ to be uncorrelated and unconditionally homoskedastic. Because we can use either a linear projection of $\Delta y_{i1}$ on 1 and $\Delta x_i$ or a linear projection of $\Delta y_{i1}$ on 1 and $\Delta x_i$, results for both choices are reported and are denoted by DQML and DQML$_\delta$.

5.3. Results

5.3.1. Stationary designs

This section provides results for designs for which the generated variables were approximately stationary ($t_0 = 50$). Table 1 provides the estimates of finite-sample bias and root mean squared error for the panel data GMM and QML estimators for stationary designs with $\sigma_\varepsilon = 1$ and $\sigma_\delta = 4$.

The evidence in Table 1 shows that the QML estimators—LQML, DQML$_\delta$, and DQML$_\delta$—generally have negligible finite-sample bias, and, consequently, their root mean squared errors are typically significantly smaller than that of the GMM estimators, which have non-negligible
finite-sample bias. Moreover, for most designs, whether one uses DQML or DQMLΔx does not matter much; they have similar finite-sample bias and root mean squared error. The exception is when δ₀ = 0.9. For highly persistent designs, DQMLₓ outperforms DQMLΔx. But among the QML estimators, the LQML estimator is—in terms of root mean squared error—best.

The system GMM estimator was introduced as a response to the poor sampling performance of the differenced GMM estimator when δ₀ is near one. Blundell and Bond (1998) showed that the system GMM estimator will perform better than the differenced GMM estimator in this case. The SGMM estimator does indeed have smaller absolute bias and root mean squared error than the differenced GMM estimators for δ₀ = 0.9. However, surprisingly, the sampling performance of the SGMM estimator is sometimes worse—often much worse—than that of the differenced GMM estimators for δ₀ not near one. Moreover, although the bias of the SGMMs estimator is less than that of the SGMM estimator for σₓ = 1 and δ₀ < 0.9, its bias exceeds the bias of the SGMM estimator for other cases. Furthermore, when σₓ = 4, the system GMM estimators have substantial bias even when δ₀ is near one. Bun and Windmeijer (2010) provide an explanation for this result. They note that system GMM may suffer from a weak instrument problem when the variance of the individual-specific effect is large relative to the variance of the idiosyncratic error. On the other hand, the sampling performance of the QML estimators is unaffected by the relative size of the individual-specific effect variance versus the idiosyncratic error variance.
**Table 2.** Finite-sample characteristics of estimators of $\delta_0$ for $t_0 = 1.$

| $\sigma_\zeta$ | 0.0 | 0.2 | 0.4 | 0.6 | 0.8 | 0.9 |
|---------------|-----|-----|-----|-----|-----|-----|
| DGMMs         | 0.0293 | 0.0311 | 0.0312 | 0.0327 | 0.0458 | 0.0594 |
| DGMM          | 0.0329 | 0.0340 | 0.0333 | 0.0332 | 0.0408 | 0.0463 |
| SGMMs         | 0.0046 | 0.0058 | 0.0071 | 0.0130 | 0.0444 | 0.0789 |
| SGMM          | 0.0302 | 0.0315 | 0.0309 | 0.0305 | 0.0536 | 0.0814 |
| DQML $\Delta$ | 0.0328 | 0.0422 | 0.0439 | 0.0390 | 0.0317 | 0.0630 |
| DQML $\Delta x$ | 0.0021 | 0.0029 | 0.0068 | 0.0127 | 0.0291 | 0.0400 |
| LQML          | 0.0248 | 0.0250 | 0.0233 | 0.0218 | 0.0221 | 0.0226 |
| LQML          | 0.0005 | 0.0001 | -0.0012 | -0.0010 | -0.0010 | -0.0001 |

5.3.2. Non-stationary designs

Table 2 provides the finite-sample bias and root mean squared error estimates for non-stationary designs. For these designs, $t_0 = 1$, and, therefore, for each individual, the time series began in the immediate past.

In order for the system GMM estimators to be consistent (as $N \rightarrow \infty$), the stochastic process for each individual has to have had sufficient time to converge to its steady state by time $t = 1$ (see, e.g., Roodman 2009). However, given $t_0 = 1$, convergence to a steady state at time $t = 1$ has clearly not occurred. The effect of the failure of this initial condition restriction is most striking when $\sigma_\zeta = 4$. In this case, for many designs, the absolute bias and root mean squared error of the system GMM estimators are much larger than that of the other estimators.

Except for the condition that $y_{it}$ must be uncorrelated with $v_{it}$ for $t \geq 1$, the QML estimators are unaffected by initial conditions. However, the consistency (as $N \rightarrow \infty$) of the differenced QML estimators—DQML $\Delta x$ and DQML $\Delta x$—depends on the $v_{it}$s being unconditionally homoskedastic, and, when $t_0 = 1$, the $v_{it}$s are time-series heteroskedastic. Consequently, in Table 2, the differenced QML estimators no longer always dominate the differenced GMM estimators in terms of finite-sample bias. On the other hand, the LQML estimator is robust with respect to time-series heteroskedasticity, and therefore its finite-sample bias is still negligible for $t_0 = 1$.

5.3.3. Confidence intervals

Finally, as a check on how well the asymptotic results in Theorems 2 and 4 are approximated in finite samples, 95% confidence intervals were calculated for the QML estimators and the
fraction of times those intervals covered $\delta_0$ was recorded. Because the coverage of the 95% confidence intervals for the DQML$_x$ and DQML$_{\Delta x}$ estimators was similar, DQML$_{\Delta x}$ coverage results are omitted for the sake of brevity. This section provides the coverage results for the LQML and DQML$_x$ estimators.

In order to calculate 95% confidence intervals, standard errors are needed. Therefore, empirical counterparts to the asymptotic variance–covariance matrices of the QML estimators were constructed. For the LQML estimator, the variance–covariance matrix $H_0^{-1} I_0 H_0^{-1}$ (see Theorem 2) was estimated with $H_N(\hat{\psi})^{-1} I_N(\hat{\psi}) H_N(\hat{\psi})^{-1}$. In this expression, $\hat{\psi}$ is the QML estimator, $I_N(\hat{\psi}) = N^{-1} \sum_i (\partial l_i(\hat{\psi})/\partial \psi)(\partial l_i(\hat{\psi})/\partial \psi)'$, and $H_N(\hat{\psi}) = N^{-1} \sum_i \partial^2 l_i(\hat{\psi})/\partial \psi \partial \psi'$. Similarly, for the DQML$_x$ estimator, the matrix $\tilde{H}_0^{-1} \tilde{I}_0 \tilde{H}_0^{-1}$ (see Theorem 4) was estimated with $\tilde{H}_N(\hat{\lambda})^{-1} \tilde{I}_N(\hat{\lambda}) \tilde{H}_N(\hat{\lambda})^{-1}$, where $\hat{\lambda}$ is the QML estimator, $\tilde{I}_N(\hat{\lambda}) = N^{-1} \sum_i (\partial l_i(\hat{\lambda})/\partial \lambda)(\partial l_i(\hat{\lambda})/\partial \lambda)'$, and $\tilde{H}_N(\hat{\lambda}) = N^{-1} \sum_i \partial^2 l_i(\hat{\lambda})/\partial \lambda \partial \lambda'$.

The coverage of the 95% confidence intervals for the LQML and DQML$_x$ estimators are reported in Table 3. From Table 3, we see that the coverage of the 95% confidence intervals for both estimators is what we would expect in light of Theorems 2 and 4.

For $t_0 = 50$, the coverage of the 95% confidence intervals is fairly close to 95% for all values of $\delta_0$ except for $\delta_0 = 0.9$. For $\delta_0 = 0.9$, the coverage of the 95% confidence intervals for the DQML$_x$ estimator dips down to about 91% for $\sigma_x = 1$ and 92% for $\sigma_x = 4$. Nevertheless, this result is consistent with Theorem 4. This is because the theorem requires the true parameter vector be within the interior of the parameter space, and, although $\delta_0 = 0.9$ is within the parameter space interior, the closer $\delta_0$ gets to the edge of the parameter space, the larger the sample size required for large-sample theory to provide a good approximation.

Moreover, the coverage rates of the 95% confidence intervals for the DQML$_x$ estimator are poor approximations for $t_0 = 1$. This finding is also consistent with the large sample theory provided in Section 3. As already noted, the QML estimators examined in this article are robust to misspecification of the log-likelihood and initial conditions, but they are not robust to misspecification of the error variance–covariance matrix. When $t_0 = 1$, the $\nu_{ij}$s are both conditionally heteroskedastic and unconditionally time-series heteroskedastic. But the algorithm used for calculating DQML$_x$ estimates does not take into account the possibility of time-series heteroskedasticity.

On the other hand, the LQML estimator examined in this article is robust to both conditional heteroskedasticity and unconditional time-series heteroskedasticity. Therefore, although the confidence intervals for the LQML estimator appear to be a bit too short on average for all $\delta_0$ when $t_0 = 1$, they nevertheless approximate the desired coverage of 95% regardless of the value of $t_0$.

### Table 3. Percentage (%) of 95% confidence intervals that covered $\delta_0$.

| $t_0$ | $\sigma_x$ | Estimator  | 0.0  | 0.2  | 0.4  | 0.6  | 0.8  | 0.9  |
|-------|------------|------------|------|------|------|------|------|------|
| 50    | 1          | LQML       | 94.4 | 93.4 | 94.4 | 94.7 | 95.7 | 95.8 |
| 50    | 1          | DQML       | 94.3 | 93.7 | 93.9 | 94.3 | 95.0 | 91.4 |
| 50    | 4          | LQML       | 93.4 | 93.6 | 93.5 | 94.0 | 95.7 | 95.9 |
| 50    | 4          | DQML       | 93.9 | 94.1 | 93.8 | 95.4 | 95.1 | 92.0 |
| 1     | 1          | LQML       | 93.7 | 92.9 | 93.5 | 93.5 | 91.2 | 94.3 |
| 1     | 1          | DQML       | 92.9 | 91.9 | 92.1 | 90.2 | 82.0 | 70.7 |
| 1     | 4          | LQML       | 93.3 | 93.8 | 93.1 | 94.2 | 94.4 | 93.7 |
| 1     | 4          | DQML       | 92.0 | 92.6 | 92.1 | 92.6 | 89.9 | 88.0 |
6. Conclusions

This article established the almost sure convergence and asymptotic normality of levels and differenced QML estimators of the parameters of a $p$th-order dynamic panel data model. The almost sure convergence and asymptotic normality of the estimators do not depend on initial conditions, such as those required by the system GMM estimator. Moreover, the log-likelihood can be misspecified, and the errors can be conditionally and time-series heteroskedastic. However, the error variance–covariance matrix must be correctly specified, and only LQML estimates can be easily calculated when the errors are time-series heteroskedastic. The article provided an ECME algorithm for this case. Furthermore, the LQML estimator dominated all of the other estimators in terms of having the smallest root mean squared errors.

Acknowledgments

This article has benefited from helpful suggestions of an anonymous referee.

References

Alvarez, J., and M. Arellano. 2004. Robust likelihood estimation of dynamic panel data models. CEMFI Working Paper 0421.
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 58:277–97.
Binder, M., C. Hsiao, and M. H. Pesaran. 2005. Estimation and inference in short panel vector autoregressions with unit roots and cointegration. Econometric Theory 21:795–837.
Blundell, R., and S. Bond. 1998. Initial conditions and moment restrictions in dynamic panel data models. Journal of Econometrics 87:115–43.
Bun, M. J. G., and F. Windmeijer. 2010. The weak instrument problem of the system GMM estimator in dynamic panel data models. The Econometrics Journal 13:95–126.
Chamberlain, G. 1982. Multivariate regression models for panel data. Journal of Econometrics 18:5–46.
Chamberlain, G. 1984. Panel data. In Handbook of Econometrics, Vol. 2, eds. Z. Griliches and M. D. Intriligator. 1247–318. Amsterdam: North Holland.
Greene, W. H. 2012. Econometric analysis. 7th ed. Upper Saddle River, NJ: Prentice Hall.
Hsiao, C., H. M. Pesaran, and A. K. Tahmiscioglu. 2002. Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods. Journal of Econometrics 109:107–50.
Kruiniger, H. 2013. Quasi ML estimation of the panel AR(1) model with arbitrary initial conditions. Journal of Econometrics 173:175–88.
Liu, C., and D. B. Rubin. 1994. The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. Biometrika 81:633–48.
Meng, X.-L., and D. van Dyk. 1998. Fast EM-type implementations for mixed effects models. Journal of the Royal Statistical Society, Series B 60:559–78.
Moral-Benito, E. 2013. Likelihood-based estimation of dynamic panels with predetermined regressors. Journal of Business & Economic Statistics 31:451–72.
Phillips, R. F. 2004. Estimation of a generalized random-effects model: Some ECME algorithms and Monte Carlo evidence. Journal of Economic Dynamics & Control 28:1801–24.
Phillips, R. F. 2010. Iterated feasible generalized least-squares estimation of augmented dynamic panel data models. Journal of Business & Economic Statistics 28:410–22.
Phillips, R. F. 2012. On computing maximum-likelihood estimates of the unbalanced two-way random-effects model. Communications in Statistics—Simulation and Computation 41:1921–7.
Phillips, R. F. 2015. On quasi maximum-likelihood estimation of dynamic panel data models. *Economics Letters* 137:91–4.

Phillips, R. F. 2017. Quasi maximum-likelihood estimation of dynamic panel data models. [https://arxiv.org/abs/1702.00662](https://arxiv.org/abs/1702.00662).

Roodman, D. 2009. A note on the theme of too many instruments. *Oxford Bulletin of Economics and Statistics* 71:135–58.

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