# FIRST DIFFERENCE MAXIMUM LIKELIHOOD AND DYNAMIC PANEL ESTIMATION

**By**

**Chirok Han, Peter C.B. Phillips**

**COWLES FOUNDATION PAPER NO. 1379**


**COWLES FOUNDATION FOR RESEARCH IN ECONOMICS**
**YALE UNIVERSITY**
**Box 208281**
**New Haven, Connecticut 06520-8281**

**2013**

**http://cowles.econ.yale.edu/**

Journal of Econometrics 175 (2013) 35–45

Contents lists available at SciVerse ScienceDirect

Journal of Econometrics

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

# First difference maximum likelihood and dynamic panel estimation☆

Chirok Han[a,*], Peter C.B. Phillips[b,c,d,e]

[a] Korea University, Republic of Korea
[b] Yale University, United States
[c] University of Auckland, New Zealand
[d] University of Southampton, United Kingdom
[e] Singapore Management University, Singapore

ARTICLE INFO

*Article history:*
Received 3 June 2011
Received in revised form
11 February 2013
Accepted 11 March 2013
Available online 20 March 2013

*JEL classification:*
C22
C23

*Keywords:*
Asymptote
Bounded support
Dynamic panel
Efficiency
First difference MLE
Likelihood
Quartic equation
Restricted extremum estimator

ABSTRACT

First difference maximum likelihood (FDML) seems an attractive estimation methodology in dynamic panel data modeling because differencing eliminates fixed effects and, in the case of a unit root, differencing transforms the data to stationarity, thereby addressing both incidental parameter problems and the possible effects of nonstationarity. This paper draws attention to certain pathologies that arise in the use of FDML that have gone unnoticed in the literature and that affect both finite sample performance and asymptotics. FDML uses the Gaussian likelihood function for first differenced data and parameter estimation is based on the whole domain over which the log-likelihood is defined. However, extending the domain of the likelihood beyond the stationary region has certain consequences that have a major effect on finite sample and asymptotic performance. First, the extended likelihood is not the true likelihood even in the Gaussian case and it has a finite upper bound of definition. Second, it is often bimodal, and one of its peaks can be so peculiar that numerical maximization of the extended likelihood frequently fails to locate the global maximum. As a result of these pathologies, the FDML estimator is a restricted estimator, numerical implementation is not straightforward and asymptotics are hard to derive in cases where the peculiarity occurs with non-negligible probabilities. The peculiarities in the likelihood are found to be particularly marked in time series with a unit root. In this case, the asymptotic distribution of the FDMLE has bounded support and its density is infinite at the upper bound when the time series sample size $T \to \infty$. As the panel width $n \to \infty$ the pathology is removed and the limit theory is normal. This result applies even for $T$ fixed and we present an expression for the asymptotic distribution which does not depend on the time dimension. We also show how this limit theory depends on the form of the extended likelihood.



## 1. Introduction

Maximum likelihood estimation based on first-differenced data (FDML) has recently attracted attention as an alternative estimation methodology to conventional maximum likelihood (ML) and GMM approaches in dynamic panel models (Hsiao et al., 2002; Kruiniger, 2008). FDML appears to offer certain immediate advantages in dynamic panels with fixed effects. Unlike unconditional ML where fixed effects are treated as parameters to estimate, FDML is free from the incidental parameter problem (Neyman and Scott, 1948) because nuisance individual effects have already been eliminated before deriving the likelihood. In addition, the differenced data are stationary whether the original data are stationary or integrated, and hence the presence of a unit root does not appear to require any special treatment or modification of the likelihood function. This feature is deemed especially useful when panel data show a large degree of persistence.

These advantages, coupled with the computational convenience of modern numerical optimization, have spurred the use of FDMLE in applied research. The empirical literature dates back to MaCurdy (1982). But there has been little research on the method's properties or on certain of its peculiarities such as negative variance estimates that are known to arise in its implementation by numerical optimization. Most importantly, it seems not to have been recognized in the literature that FDMLE is *not* a maximum likelihood procedure because the 'likelihood' that is used in optimization is based on analytically extending the stationary likelihood outside the stationary region. The resulting function is *not* a

☆ CH acknowledges support from the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2011-332-B00026). PCBP acknowledges support from the NSF under Grant No. SES09-56687.

* Correspondence to: Department of Economics, Korea University, Anam-dong, Seongbuk-gu, Seoul, Republic of Korea. Tel.: +82 2 3290 2205.
*E-mail address:* chirokhan@korea.ac.kr (C. Han).

0304-4076/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jeconom.2013.03.003

true likelihood outside the stationary region even though it is well defined for certain nonstationary regions. This feature of FDMLE is subtle, which partly explains why it has gone unnoticed in the literature for so long. But it has significant implications and leads to further complications, including an upper bound restriction on the domain that affects both finite sample theory and asymptotic behavior. An investigator may, of course, choose *a priori* to restrict the domain of the autoregressive roots to the unit circle, but in this event an appropriate asymptotic theory that accounts for the restriction would need to be used in practical work.

Wilson (1988) provided an exact likelihood for the differenced data generated from a stationary AR(1) process based on Ansley's (1979) expression for ARMA(1, 1), and discovered in simulations that FDMLE outperforms the maximum likelihood (ML) estimator in terms of mean squared error for small samples. Hsiao et al. (2002); hereafter HPT) studied FDML in linear dynamic panel models with wide short panels – that is panels with large cross sectional dimension ($n$) and short time series length ($T$) – where conventional ML is inconsistent due to the effects of incidental parameters. The authors appealed to standard regularity conditions for the asymptotic theory of FDMLE, and used Newton–Raphson optimization in simulations to compute the FDMLE. Their simulations confirmed the superior performance of the FDMLE in terms of bias, root mean square error, test accuracy and power over a range of commonly used panel estimators. HPT do note that FDMLE "sometimes breaks down completely" giving negative variance estimates and estimates of the autoregressive coefficient greater than unity but they "skipped those replications altogether" and provided no analysis of these anomalies.

The present work will explain these anomalies and make it clear why standard asymptotic arguments do not apply to derive the limit theory of the FDMLE. The most closely related work to the present paper is Kruiniger (2008). Kruiniger derived asymptotics for the FDMLE in the panel AR(1) model with large $nT$ (i.e., for $n$ or $T$ large or both $n$ and $T$ large) for the stationary case, and with large $n$ and arbitrary $T$ for the unit root case. Though first differencing uses up one observation for each panel, there appears to be no serious information loss in comparison with other methods like ML because one degree of freedom is needed in conventional ML to identify each individual intercept. Curiously, the asymptotics that are now available speak to the opposite, although this has not so far been discussed in the literature. Indeed, for AR(1) panels with large $n$, large $T$ and a unit root, the LSDV estimator (which is the MLE under normality of the idiosyncratic error, conditional on initial observations and without any restriction of covariance stationarity) is known to have a $N(0, \frac{51}{5})$ limit distribution when the bias is corrected (Hahn and Kuersteiner, 2002). By contrast, the FDMLE is also asymptotically normal, has no asymptotic bias and its limit variance is 8 (Kruiniger, 2008), thereby producing an asymptotic gain in efficiency at unity over bias corrected LSDV. This reduction in asymptotic variance between the two ML approaches is partly explained by the fact that the FDMLE uses a stationarity condition for the differenced data in setting up the likelihood. Such a condition does not allow for the fact that differenced data is explosive when the AR coefficient exceeds unity, thereby leading to an implied restriction on the model and parameter space that affects both finite sample and asymptotic behavior.

Recent work by Han et al. (2011, forthcoming) shows that there are other estimators involving difference transformations that have performance superior to the bias corrected MLE in dynamic panels. These authors give a panel fully aggregated estimator (FAE) that aggregates the effects of a full set of differences in a simple linear regression framework. The panel FAE has a limiting $N(0, 9)$ distribution after centering and standardization, and like the FDMLE is more efficient asymptotically than the bias corrected MLE with no stationarity restriction imposed (i.e., the bias corrected LSDV) for the autoregressive coefficient in a vicinity of unity. There is much other recent work on dynamic panel models, but none of that work relates to the issues connected with the FDMLE procedure that are discussed in the present paper.

For all the attractive properties of FDMLE, some of its most important features have not been noted or studied in the literature. These features, as we demonstrate here, play a critical role in the asymptotic theory and in the finite sample performance of the estimator. First and most importantly, the 'likelihood' function considered in the panel literature that is used for numerical computation of the FDMLE is *not* in fact the correct likelihood function over the whole domain. As indicated above, it is a pseudo-likelihood based on extending the stationary likelihood outside its natural domain of definition to a bounded part of the nonstationary region. Second, this pseudo 'likelihood' function can behave so wildly that numerical maximization procedures can often fail to identify the global maximum. These two issues combine to make a careful analytical treatment of FDMLE very difficult. On the one hand, the asymptotic theory depends subtly on the (rapidly changing) form of the likelihood function near its natural upper boundary which arises from the extension of the stationary likelihood. On the other hand, the wild behavior of the likelihood itself often compromises the numerical evaluation of the FDMLE, giving rise to anomalous results such as those reported above.

The present paper explains these pathologies and their material impact on the finite sample distribution and limit distribution of the FDMLE. We also show how the effects of this anomaly diminish when the FDML is applied to dynamic panel data as the cross-sectional dimension increases.

The next section lays out the model, notation and discusses the FDML 'likelihood'. Section 3 examines the anomaly that arises when the data are persistent, considering in turn the time series ($n = 1$), panel ($n > 1$) and panel asymptotic case ($n \to \infty$). Section 4 concludes. Proofs are given in the Appendix and reference is made to the original version of this paper (Han and Phillips, 2010) for further technical details. Throughout the remainder of the paper it will be convenient to use the notation $T_m = T - m$ and $\tilde{T}_m = T + m$.

## 2. Model, notation and the FDMLE

We consider a Gaussian panel $y_{it}$ generated by the simple panel dynamic model $y_{it} = \eta_i(1 - \rho_0) + \rho_0 y_{it-1} + \varepsilon_{it}$, where $\varepsilon_{it} \sim iid\ N(0, \sigma_0^2)$ and $-1 < \rho_0 \leq 1$.[1] Suppose that $y_{it}$ is observed for $i = 1, \ldots, n$ and $t = 0, \ldots, T$.

The likelihood function is derived from the joint distribution of $\Delta y_i := (\Delta y_{i1}, \ldots, \Delta y_{iT})'$. Under the stationarity assumption for $\Delta y_{it}$, we have

$$\Delta y_i \sim N\left(0, \sigma_0^2 C_T(\rho_0)\right), \tag{1}$$

where $C_T(\rho_0)$ is a Toeplitz matrix whose leading row is formed from the elements $\frac{1}{1+\rho_0}\{2, -(1-\rho_0), -\rho_0(1-\rho_0), \ldots, -\rho_0^{T-2}(1-\rho_0)\}$. Direct evaluation leads to $\det C_T(\rho_0) = J_T(\rho_0)/(1 + \rho_0)$, where $J_T(\rho) = \tilde{T}_1 - T_1\rho$ (e.g., Galbraith and Galbraith, 1974; HPT, 2002; Kruiniger, 2008; Han, 2007). Thus, for $-1 < \rho \leq 1$ and $\sigma^2 > 0$, the log-likelihood function for $\Delta y_i$ is

$$\ln L(\rho, \sigma^2) = -\frac{nT}{2}\ln 2\pi - \frac{nT}{2}\ln\sigma^2 - \frac{n}{2}\ln\left[\frac{J_T(\rho)}{1+\rho}\right] - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\Delta y_i' C_T(\rho)^{-1}\Delta y_i. \tag{2}$$

[1] The analysis can be extended to the model where $y_{it}$ is replaced with $y_{it} - \beta' x_{it}$ and $x_{it}$ contains exogenous regressors. The focus in the present paper is on the estimation of $\rho$ and the peculiarities of its limit theory. Asymptotics for the corresponding estimates of $\beta$ may be derived in a standard way and are not discussed here.

This log-likelihood is valid for $\rho \in (-1, 1]$. If the true $\rho$ is strictly smaller than 1 and if the parameter space (for $\rho$) is limited to $(-1, 1]$, then the asymptotic theory for the FDMLE can be derived by invoking generic theories for MLE under the condition that the log-likelihood (2) behaves regularly. However, if the true persistence parameter is $\rho_0 = 1$ and if the parameter space for $\rho$ is limited to $(-1, 1]$, then the true parameter lies on the boundary of the parameter space and nonstandard results (both for time series and for panels) are to be expected. In that case the limit distribution involves a positive probability mass at the boundary. (See Geyer, 1994; Andrews, 1999, 2001.)

Rather than limiting the domain of $\rho$ to $(-1, 1]$, one can analytically extend the function (2) to the whole domain over which the criterion $\ln L(\rho, \sigma^2)$ is defined. This is the approach taken (either explicitly or implicitly) in recent work by HPT (2002) and Kruiniger (2008). This domain for $(\rho, \sigma^2)$ is $(-1, \frac{T+1}{T-1}) \times (0, \infty)$ (cf., Kruiniger, 2008), which contains $\rho = 1$ in its interior. By means of this analytic extension, HPT (2002) and Kruiniger (2008) proceed to deduce asymptotic normality for the FDMLE as $n \to \infty$ for all $\rho$ in $(-1, 1]$. However (2) is the correct log-likelihood function only for $\rho \in (-1, 1]$, but *not* for $\rho \in (1, \frac{T+1}{T-1})$ because (1) does not hold for $\rho_0 > 1$.[2] Thus, maximizing (2) over the whole domain does not yield an MLE but rather a restricted estimator that depends on an extension of the stationary likelihood beyond its natural domain of definition. In consequence, deriving asymptotics using standard regularity properties and stationary limit theory for the MLE and "information matrix" calculations to obtain the variance is not justified when the true value of $\rho$ is unity.

A related issue stems from the boundary behavior of (2) as $\rho \to \frac{T+1}{T-1}$. Though $\ln L(\rho, \sigma^2)$ is differentiable on $(-1, \frac{T+1}{T-1}) \times (0, \infty)$ as Kruiniger (2008), Lemma 7, finds, the behavior of the 'log-likelihood' function may be very violent especially for small $n$. Fig. 1 shows a sample path generated with $\rho_0 = 1$, $\sigma_0^2 = 1$, $n = 1$ and $T = 101$, in which case the upper bound of the extended domain is $\frac{102}{100} = 1.02$ for the $\rho$ parameter. When the profile 'log-likelihood' criterion function $\ln L^*(\rho) \equiv \max_{\sigma^2>0} \ln L(\rho, \sigma^2)$ is plotted over the whole domain $(-1, 1.02)$, we obtain the curve shown in the left graphic, and numerical optimization (using the 'optimize' function of R) finds a maximizer at 0.99 (the vertical line of alternating dots and dashes). However, when the profile criterion is plotted on very fine grids near the upper bound, we obtain the dramatically different curve shown in the right graphic of Fig. 1. This curve reveals that the profile criterion behaves with a violent fluctuation as $\rho$ approaches the upper bound 1.02 and that 0.99 is only a local maximizer. In particular, the criterion rises sharply and then rapidly falls for $\rho$ values close to the upper bound. (The sharp peak is smooth and differentiable as the inset graph shows.) This anomaly in the criterion function will not be detected unless the graph is drawn very carefully and, for the considered sample path, the global maximum (the vertical dashed lines) which is attained in this region may be missed entirely, as it usually is with standard optimization algorithms unless they are carefully tuned. For other sample paths the profile criterion may lack such sharp peaks and be unimodal, while yet other sample paths may produce bimodal profile criteria for which the global maxima are attained at the other peak for a smaller (stationary) value of $\rho$. Fine grid searches combined with other optimization tools may help in finding global maxima in particular situations but raise difficulties in usability because extremely fine grids of unknown precision may be required for some sample paths and will not be known *a priori*, as is evident from Fig. 1. In sum, the criterion function (2) has the potential for unstable, rapidly fluctuating behavior in a small region close to the upper bound of the extended domain of definition. This instability affects both the numerical evaluation of the FDMLE and its limit theory.

Because (2) is not a proper log-likelihood for the domain $(1, \frac{T+1}{T-1}) \times (0, \infty)$, general results on MLE for stationary time series cannot be employed to derive asymptotic results for the FDMLE even though (2) is differentiable infinitely many times over the full domain $(-1, \frac{T+1}{T-1}) \times (0, \infty)$, a property which Kruiniger (2008) notes. Furthermore, due to the described peculiarity of the profile 'log-likelihood' criterion near the upper bound $\frac{T+1}{T-1}$, we cannot expect numerical studies based on simulations conducted with standard numerical maximization methods to provide reliable results. Also, in order to apply a general theory for extremum estimators (which usually involves the use of a quadratic approximation), some basic properties of (2) should be known so that the existence of the extremum estimator is verified and the global maximum (rather than a local one) is characterized and used. It is therefore necessary to examine the criterion function itself carefully rather than the first order conditions. The fact that the upper bound depends on the sample size $T$ provides a further source of complication if $T \to \infty$ because the upper limit of the support shrinks to unity.

We handle these issues by using a tractable algebraic expression for the criterion function which allows a direct treatment for asymptotic analysis and numerical calculation. The unit root limit theory for the FDMLE developed in the present paper takes an interesting and revealing form. In particular, the FDMLE is shown to have an asymptote with infinite density at the upper limit of its support when $n = 1$, a new feature that is the result of the anomalies in the criterion function and the fact that the FDMLE is a restricted estimator. This peculiarity diminishes and then disappears as the objective function is averaged across a large number of cross sections.[3] Simulations are done using an explicit solution to the profile objective function by finding the roots of a quartic equation which avoids problems of numerical optimization, and these corroborate the new asymptotic theory.

The remainder of this section provides an explicit expression for the criterion function in (2), shows the existence of the global maximizer, and presents a method to compute the FDMLE which avoids the numerical difficulties associated with peculiarities of the type shown in Fig. 1.

The determinant and inverse of $C_T(\rho)$ can be analytically evaluated (see, e.g., Kruiniger, 2008; Han and Phillips, 2010). Specifically,

$$Q_{iT}(\rho) := \Delta y_i' C_T(\rho)^{-1} \Delta y_i = \sum_{t=1}^{T} u_{it}(\rho)^2 - \frac{1-\rho}{J_T(\rho)} \left[ \sum_{t=1}^{T} u_{it}(\rho) \right]^2,$$

[2] This is most easily seen by noting that $\Delta y_{it}$ is explosive when $\rho_0 > 1$ and its second order moments depend on $t$, so the process is not stationary. In particular

$$\Delta y_{it} = \Delta u_{it} = \varepsilon_{it} + (\rho_0 - 1) \sum_{j=1}^{t-1} \rho_0^{j-1} \varepsilon_{it-j} + \rho_0^{t-1} (\rho_0 - 1) u_{i0},$$

so that

$$\mathrm{Var}(\Delta u_{it}) = \left[ 1 + (\rho_0 - 1)^2 \left( \frac{\rho_0^{2(t-1)} - 1}{\rho_0^2 - 1} \right) \right] \sigma_0^2 + (\rho_0 - 1)^2 \rho_0^{2(t-1)} \mathrm{Var}(u_{i0}), \quad (3)$$

which depends on $t$ explosively when $\rho_0 > 1$. Moreover, when $\rho_0 > 1$ and $\sigma_0^2 > 0$, $\mathrm{Var}(\Delta u_{it})$ is the same over $t$ only if $\frac{\rho_0^{2(t-1)}}{\rho_0^2 - 1} \sigma_0^2 + \rho_0^{2(t-1)} \mathrm{Var}(u_{i0}) = 0$, i.e., if $\mathrm{Var}(u_{i0}) = \sigma_0^2/(1 - \rho_0^2) < 0$, which is impossible. Note also that $\sigma_0^2 C_T(\rho_0)$ is Toeplitz whereas the covariance matrix when $\rho_0 > 1$ is *not* Toeplitz as it is not lead diagonal constant — as shown in (3). So $\sigma_0^2 C_T(\rho_0)$ is not the covariance matrix of $\Delta y_i$ when $\rho_0 > 1$. It follows that the FDML 'likelihood' (2) is not the true likelihood between 1 and $(T + 1)/(T - 1)$.

[3] Simulations conducted for $T = 30$ and $n = 1$ with 5000 replications exhibited a 30% probability of the estimator from a standard numerical optimization (using R's optimize function with the parameter space restricted to $[-1 + 10^{-12}, 1 + \frac{2}{29} - 10^{-12}]$) being different from the global maximizer. This probability falls to approximately 7.8%, 1.4% and 0.08% for $n = 5$, $n = 15$ and $n = 30$, respectively.

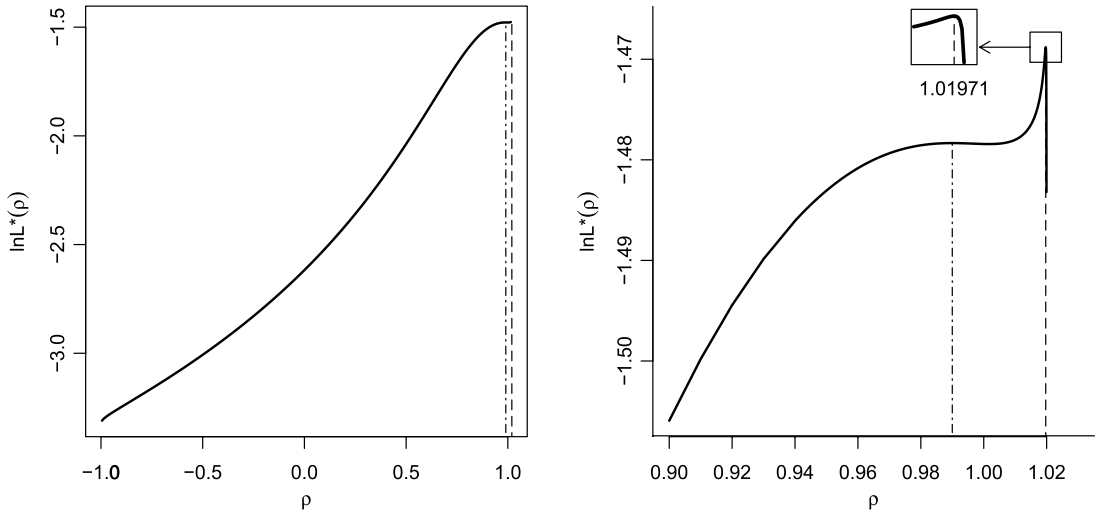


**Fig. 1.** Multimodal average profile 'log-likelihood' for a sample path where numerical optimization finds a local maximum (the lines of alternating dots and dashes) instead of the global maximum (the dashed lines). The left graph drawn over the whole domain $(-1, \frac{T+1}{T-1})$ fails to reveal the real shape of the criterion near the upper bound, while the right panel illustrates the violent upshoot and rapid decline as $\rho$ approaches the upper bound.

where $u_{it}(\rho) = z_{it} - \rho z_{it-1}$, $z_{it} = y_{it} - y_{i0}$ and $J_T(\rho) = \tilde{T}_1 - T_1\rho$ as before. Note that $Q_{iT}(\rho)$ is strictly positive if $-1 < \rho < \frac{T+1}{T-1}$ and $\Delta y_i \neq 0$.

Let $\ln L^*(\rho)$ denote the profile log likelihood function $\ln L^*(\rho) = \max_{\sigma^2>0} \ln L(\rho, \sigma^2)$. For given $\rho$, $\ln L(\rho, \sigma^2)$ is differentiable with respect to $\sigma^2$ and is globally concave in $\sigma^2$, so the maximizer of $\ln L(\rho, \sigma^2)$ for given $\rho$ satisfies the first order condition $\frac{\partial \ln L(\rho,\sigma^2)}{\partial \sigma^2} = 0$. Simple algebra shows that the maximizer is $\sigma^2 = (nT)^{-1}\sum_{i=1}^n Q_{iT}(\rho)$ for given $\rho$. Profiling gives

$$\ln L^*(\rho) = -\frac{nT}{2}\ln 2\pi - \frac{nT}{2}\ln\left[\frac{1}{nT}\sum_{i=1}^{n} Q_{iT}(\rho)\right] - \frac{n}{2}\ln\left[\frac{J_T(\rho)}{1+\rho}\right] - \frac{nT}{2}. \tag{4}$$

The FDMLE $\hat{\rho}$ maximizes the profile 'likelihood' criterion function (4), which is defined for $-1 < \rho < \frac{T+1}{T-1}$. It is clear that $\ln L^*(\rho) \to -\infty$ as $\rho \to -1$ or $\rho \to \frac{T+1}{T-1}$ and thus $\hat{\rho}$ exists in the interval $(-1, \frac{T+1}{T-1})$ almost surely. The first order conditions are $\frac{\partial}{\partial \rho}\ln L(\hat{\rho}, \hat{\sigma}^2) = 0$ and $\frac{\partial}{\partial \sigma^2}\ln L(\hat{\rho}, \hat{\sigma}^2)$, and by transforming them Kruiniger (2008) derives a quartic equation

$$a_0 + a_1\hat{\rho} + a_2\hat{\rho}^2 + a_3\hat{\rho}^3 + a_4\hat{\rho}^4 = 0, \tag{5}$$

where some lengthy algebra gives $a_0 = \tilde{T}_1 c_0 + \tilde{T}_1^2 c_1 - 2d_0 - \tilde{T}_1 d_1$, $a_1 = -T_1 c_0 - \tilde{T}_1 T_1 c_1 - \tilde{T}_1^2 c_2 + \tilde{T}_3 d_1 + \tilde{T}_1 d_2$, $a_2 = -T_1\tilde{T}_1 c_1 + \tilde{T}_1 T_2 c_2 + \tilde{T}_1 d_1 - \tilde{T}_1 d_2$, $a_3 = T_1^2 c_1 + T_1\tilde{T}_2 c_2 - T_1 d_1 - \tilde{T}_1 d_2$, and $a_4 = -T_1^2 c_2 + T_1 d_2$.[4] This equation can be solved directly, for example by Euler's method (see Appendix B of Han and Phillips, 2010 for details), and $\hat{\sigma}^2$ is obtained by $\hat{\sigma}^2 = \frac{1}{nT}\sum_{i=1}^n Q_{iT}(\hat{\rho})$. Eq. (5) removes the singularity that occurs in the criterion $\ln L(\rho, \sigma^2)$ at $\rho = -1$ and $\rho = \frac{T+1}{T-1}$ so its solutions can lie outside of the domain $(-1, \frac{T+1}{T-1})$. Thus, for optimization it is important to check that $\hat{\rho}$ falls in the domain $(-1, \frac{T+1}{T-1})$. If there are multiple solutions of (5) in the domain $(-1, \frac{T+1}{T-1})$, then the $\ln L(\rho, \sigma^2)$ values are compared in order to maximize the criterion. Simulations in the present paper have been conducted using this optimization routine.

The following Section 3 establishes asymptotics for time series and for panels when $\rho_0 = 1$. The time series unit root case clarifies the impact of the criterion function peculiarity and shows its asymptotic effects. Although the panel asymptotic case has already been studied in Kruiniger (2008), it is reconsidered here in the second part of Section 3 because the unusual behavior of the objective function requires special treatment which has been overlooked in the literature.

## 3. Asymptotic anomalies for persistent data

For the model $y_{it} = \eta_i(1-\rho_0) + \rho_0 y_{it-1} + \varepsilon_{it}$ with $\varepsilon_{it} \sim iid\ N(0, \sigma_0^2)$, the asymptotic theory for the FDMLE is known in the stationary case $|\rho_0| < 1$, is equivalent to that of the MLE if $T \to \infty$, and follows by standard arguments. We here establish asymptotics for the case $\rho_0 = 1$ which turn out to be very different from the usual unit root theory for the MLE.

Following the standard approach to deriving asymptotics, we reparameterize $\rho$ as $r_{nT}(\rho - 1)$ for some appropriate convergence rate $r_{nT}$. We may reasonably conjecture (and confirm below) that $r_{nT} = O(\sqrt{n}T)$, where the usual $O_p(\sqrt{n})$ rate is obtained from cross sectional aggregation and the fast $O_p(T)$ rate is common for unit root time series asymptotics. Given $r_{nT}$ and following the usual procedure (e.g., Geyer, 1994; Knight, 2003) for extremum asymptotic theory, we consider the reparameterized objective function $f_{nT}(\theta) := 2[\ln L^*(1 + r_{nT}^{-1}\theta) - \ln L^*(1)]$ obtained by letting $\theta = r_{nT}(\rho - 1)$, which is maximized at $r_{nT}(\hat{\rho} - 1)$. It is notationally convenient to let $r_{nT} = \sqrt{n}T_1$. Using the notations $\tilde{\sigma}^2 = \frac{1}{nT}\sum_{i=1}^n \sum_{t=1}^T \varepsilon_{it}^2$, $V_{0i,T} = \frac{1}{\tilde{\sigma}\sqrt{T}}\sum_{t=1}^T \varepsilon_{it}$, $V_{1i,T} = \frac{1}{\tilde{\sigma}T^{3/2}} z_{it-1}$, $V_{2i,T} = \frac{1}{\tilde{\sigma}^2 T^2}\sum_{t=1}^T z_{it-1}^2$, $W_{i,T} = \frac{1}{\tilde{\sigma}^2 T}\sum_{t=1}^T z_{it-1}\varepsilon_{it}$, $V_{0i,T}^* = (\frac{T}{T_1})^{1/2} V_{0i,T}$, $V_{1i,T}^* = (\frac{T}{T_1})^{3/2} V_{1i,T}$, $V_{2i,T}^* = (\frac{T}{T_1})^2 V_{2i,T}$, and $W_{i,T}^* = \frac{T}{T_1} W_{i,T}$, we have

$$f_{nT}(\theta) = -nT\ln\left[1 + \frac{g_{nT}(\theta)}{nT}\right] + n\ln\left[1 + \left(\frac{T}{T_1}\right)\frac{\theta/\sqrt{n}}{2 - \theta/\sqrt{n}}\right], \tag{6}$$

[4] Earlier versions of this paper suggested an iterative procedure where the first order condition for $\rho$ is expressed as a quartic equation for each $\sigma^2$. The authors thank a referee for referring to Kruiniger (2008) who expresses the concentrated first order condition as a quartic form after some transformation. The equations used to construct the quartic equation are $(1+\hat{\rho})\hat{\sigma}^2 J_T(\hat{\rho})^2 \frac{\partial}{\partial \rho}\ln L(\hat{\rho}, \hat{\sigma}^2) = 0$ and $\hat{\sigma}^2 = \frac{1}{nT}\sum_{i=1}^n Q_{iT}(\hat{\rho})$.

$$g_{nT}(\theta) = \frac{\theta^2}{n}\sum_{i=1}^{n} V^*_{2i,T} - \frac{2\theta}{\sqrt{n}}\sum_{i=1}^{n} W^*_{i,T} + \frac{\theta/\sqrt{n}}{2-\theta/\sqrt{n}}\sum_{i=1}^{n}\left(V^*_{0i,T} - \frac{\theta V^*_{1i,T}}{\sqrt{n}}\right)^2. \quad (7)$$

(The algebra is lengthy but straightforward. Details of the derivation are given in Han and Phillips, 2010.) Then the limit distribution of $\sqrt{n}T_1(\hat{\rho}-1)$ can be characterized in terms of the maximizer of the limit of $f_{nT}(\theta)$ by a suitable argmax theorem once the conditions are checked.

The remainder of this section considers separately the two cases where $n$ is fixed and where $n \to \infty$.

### 3.1. Large T asymptotics

We start by deriving large-$T$ asymptotics, where $n$ is fixed and $T \to \infty$. In this case the peculiarity of the criterion function noted earlier is a prominent characteristic and must be addressed in the asymptotics together with its impact on the distribution of the FDMLE. The technicalities are challenging and of some independent interest.

For $n$ fixed, $g_{nT}(\theta)$ is stochastically bounded for each $\theta$ and we are first interested in the pointwise weak limit of $g_{nT}(\theta)$ as $T \to \infty$. For the components of $g_{nT}(\theta)$, the limits follow from standard weak convergence theory for unit root time series (Phillips, 1987). That is, $V^*_{0i,T} \Rightarrow V_{0i} := B_i(1)$, $V^*_{1i,T} \Rightarrow V_{1i} := \int B_i$, $V^*_{2i,T} \Rightarrow V_{2i} := \int B_i^2$, and $W^*_{i,T} \Rightarrow W_i := \int B_i dB$, where the $B_i$ are standard Brownian motions independent over $i$.

Because $\lim nT \ln[1+(nT)^{-1}g_{nT}] = \lim g_{nT}$ for bounded $g_{nT}$, as $T \to \infty$, for every $\theta$ we have

$$f_{nT}(\theta) \Rightarrow f_n(\theta) := -g_n(\theta) + n\ln\frac{2}{2-\theta/\sqrt{n}}, \quad (8)$$

and

$$\begin{aligned} g_n(\theta) &= \theta^2\left(\frac{1}{n}\sum_{i=1}^{n} V_{2i}\right) - 2\theta\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i\right) + \frac{n^{-1/2}\theta}{2-n^{-1/2}\theta}\sum_{i=1}^{n}\left(V_{0i} - \frac{\theta}{\sqrt{n}}V_{1i}\right)^2 \\ &= -\sum_{i=1}^{n} V_{0i}^2 - \frac{2\theta}{\sqrt{n}}\sum_{i=1}^{n}\tilde{W}_i + \frac{\theta^2}{n}\sum_{i=1}^{n}\tilde{V}_{2i} + \frac{2}{2-n^{-1/2}\theta}\sum_{i=1}^{n}\left(V_{0i} - \frac{\theta}{\sqrt{n}}V_{1i}\right)^2, \end{aligned} \quad (9)$$

where $\tilde{V}_{2i} = V_{2i} - V_{1i}^2 = \int B_i^2 - (\int B_i)^2$, and $\tilde{W}_i = W_i - V_{0i}V_{1i} = \int B_i dB_i - B_i(1)\int B_i$.

We first provide technical results which hold as $T \to \infty$ for fixed $n$.

**Lemma 1.** *As $T \to \infty$,* (i) $\sqrt{n}T_1(\hat{\rho}-1) = O_p(1)$[5]*;* (ii) *in every compact subset of $(-\infty, 2\sqrt{n})$, $f_{nT}(\theta) \Rightarrow f_n(\theta)$ uniformly in $\theta$;* (iii) $f_n(\theta) \to -\infty$ *as* $\theta \to -\infty$ *or* $\theta \uparrow 2\sqrt{n}$ *for almost all sample paths; and* (iv) *almost surely, the global maximizer $\tilde{\theta}$ of $f_n(\theta)$ exists and is in $(-\infty, 2\sqrt{n})$.*

The implication is the following asymptotic theory for the FDMLE.

**Theorem 2.** $\sqrt{n}T_1(\hat{\rho}-1) \to_d \arg\max_{\theta<2\sqrt{n}} f_n(\theta)$.

It is worth noting that the argmax theorem requires the stochastic boundedness of the rescaled and centered estimator, which is satisfied for the explosive domain for fixed $n$ because $\sqrt{n}T_1(\hat{\rho}-1) < 2\sqrt{n}$.

We will examine the asymptotic distribution for $n = 1$ and $n > 1$ separately below. The single time series case ($n = 1$) illuminates the peculiarity at the upper bound, and the multiple time series case ($n > 1$) reveals how this peculiarity diminishes with cross section averaging as $n$ increases. The limit theory as $n \to \infty$ is treated separately later.

*The case $n = 1$*

Let $n = 1$ and omit the $i$ and $n$ subscripts from all notation for the analysis of this case. From (8) and (9) with $n = 1$, we deduce the following limit behavior and form of the limit function.

When $n = 1$, the function $f_n(\theta)$ in (8) reduces to

$$f(\theta) = V_0^2 + 2\tilde{W}\theta - \tilde{V}_2\theta^2 - \frac{2}{2-\theta}(V_0 - V_1\theta)^2 + \ln\frac{2}{2-\theta}. \quad (10)$$

Importantly, the peculiarity that is manifest in Fig. 1 carries over to the limit criterion function $f(\theta)$, yielding a function with similar potential characteristics to those of Fig. 2. Brownian motion trajectories giving rise to a limit function $f(\theta)$ similar to Fig. 2 are not rare. Note again that the sharp peak close to the upper bound is smooth in this graph, just as it is in the finite sample case, although it is not immediately apparent on the scale shown.

The global maximizer $\tilde{\theta}$ of $f(\theta)$ can be found by evaluating the first order condition, which is validated by Lemma 1(iii). According to straightforward algebra, $\tilde{\theta}$ solves the cubic equation $\sum_{j=0}^{3} b_j\theta^j = 0$, where $b_0 = 4W - V_0^2 + 1$, $b_1 = -4V_2 - 4\tilde{W} - \frac{1}{2}$, $b_2 = 4\tilde{V}_2 + \tilde{W} + V_1^2$, and $b_3 = -\tilde{V}_2$.[6] In the above $V_2 := \tilde{V}_2 + V_1^2 = \int B^2$, and $W := \tilde{W} + V_0V_1 = \int BdB$.

Simulations of 10,000 replications were conducted with $\sigma_0^2 = 1.3$ for $T = 50, 100, 500, 1000$. (Scaling the data by considering different $\sigma_0^2$ values does not affect the $\hat{\rho}$ value.) For the asymptotic expression, the components $b_j$ were computed using the finite sample formulas ($V_{0,T}$, $V_{1,T}$, $\tilde{V}_{2,T}$ and $\tilde{W}_T$) with $T = 5000$. The empirical distribution functions are plotted in Fig. 3, where the asymptotic expression is simulated by independently generating $T = 5000$ observations for each replication. The finite sample distribution is well approximated by the limit theory even for $T = 50$ and convergence to the asymptotic is manifest as $T$ increases. The asymptotic (centered) density has a mode at a negative value and an asymptote at 2. As seen in Fig. 3, the weak limit $\tilde{\theta}$ of $T_1(\hat{\rho}-1)$ is not median unbiased. The median of $\tilde{\theta}$ is approximately $-0.5$, and $P\{\tilde{\theta} \le 0\} \simeq 56.5\%$ according to simulations with $T = 5000$. The simulated mean of $\tilde{\theta}$ is approximately $-1.88$. While $\tilde{\theta} < 2$ with probability 1, we have the successive probabilities $P(\tilde{\theta} > 1) \simeq 33.8\%$, $P(\tilde{\theta} > 1.9) \simeq 20.2\%$, $P(\tilde{\theta} > 1.99) \simeq 8.6\%$, and even $P(\tilde{\theta} > 1.999) \simeq 3.1\%$. This means a considerable probability mass is placed in a range very close to the upper bound, implying that cases similar to Figs. 1 and 2 are far from being rare.

While $T_1(\hat{\rho}-1)$ never reaches the upper bound 2 for finite $T$, the simulated distributions (both finite sample and asymptotic) of $T_1(\hat{\rho}-1)$ are all highly peaked near 2. From Fig. 3, which shows the asymptotic distribution based on simulations with $T = 5000$, it appears that the density of the limit distribution of $T_1(\hat{\rho}-1)$ is

[5] Kruiniger (2008) establishes consistency for large $T$ and fixed $n$ by evaluating the limit of the quartic first order condition.

[6] A referee pointed out that this is given in Kruiniger (2008) as an intermediate step.

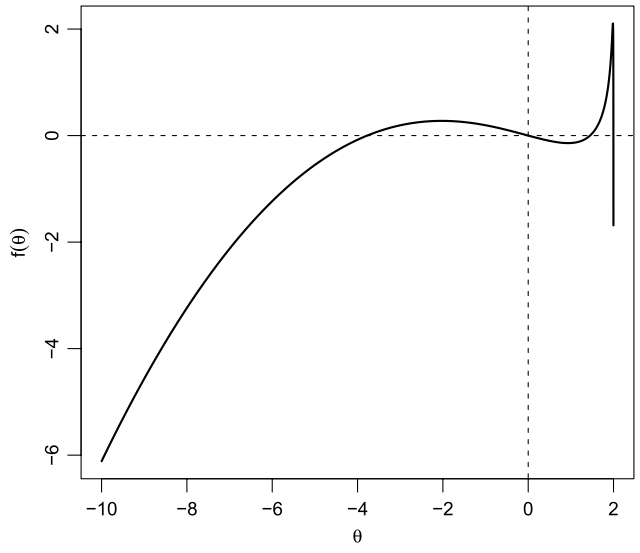


**Fig. 2.** Reparameterized limit 'log-likelihood' criterion $f(\theta)$ exhibiting violent behavior near the upper bound $\theta = 2$.

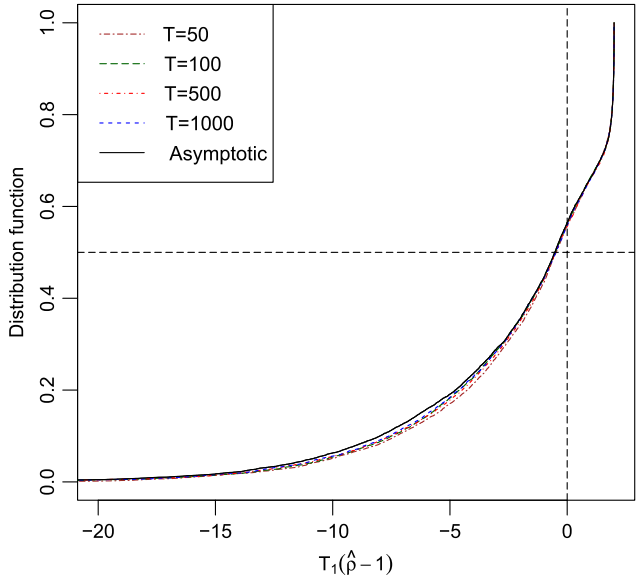


**Fig. 3.** Simulated finite sample and asymptotic CDFs for $T_1\,(\hat{\rho} - 1)$.

infinite at the boundary. The following theorem establishes that fact, showing that although there is no probability mass at the boundary in the limit, the density of $\tilde{\theta}$ escapes at 2.

**Theorem 3.** (i) $P(\tilde{\theta} > 2 - \epsilon) = O(\epsilon^{1/2})$ *as* $\epsilon \to 0$, *and* (ii) $\lim_{\epsilon \to 0} P(\tilde{\theta} > 2 - \epsilon)/\sqrt{\epsilon} > 0$.

According to the first part of the theorem, there is no probability mass at boundary 2, which is to be expected because $\tilde{\theta}$ can never attain the boundary. However, the second part of Theorem 3 implies that the density of $\tilde{\theta}$ is infinite at 2 because the density, which is the limit of $P(\tilde{\theta} > 2 - \epsilon)/\epsilon$, diverges at an $\epsilon^{-1/2}$ rate as $\epsilon \to 0$. Simulations of 10,000 replications show the results of Table 1 for different values of $\epsilon$ indicating that $P(\tilde{\theta} > 2 - \epsilon)$ diminishes at a rate no faster than $\sqrt{\epsilon}$, corroborating the finding of Theorem 3. As a result, $P(\tilde{\theta} > 2 - \epsilon)/\epsilon$ diverges, which implies that the density is infinite at the upper bound.

**Table 1**
Simulated right tail probabilities for the FDMLE.

| $\epsilon$ | 0.1 | 0.01 | 0.001 | 0.0001 | ⋯ | $\to 0$ |
|---|---|---|---|---|---|---|
| $P(\tilde{\theta} > 2 - \epsilon)$ | 0.2017 | 0.0862 | 0.0306 | 0.0107 | ⋯ | $\to 0$ |
| $P(\tilde{\theta} > 2 - \epsilon)/\sqrt{\epsilon}$ | 0.6378 | 0.862 | 0.9677 | 1.07 | ⋯ | $>0$ |
| $P(\tilde{\theta} > 2 - \epsilon)/\epsilon$ | 2.017 | 8.62 | 30.6 | 107 | ⋯ | $\to \infty$ |

The last two terms of the limit criterion (10), viz.,

$$-\frac{2}{2-\theta}(V_0 - V_1\theta)^2 + \ln\frac{2}{2-\theta} \tag{11}$$

are responsible for the limit distribution having an infinite density at the boundary. The factor $\frac{2}{2-\theta}$ diverges to infinity as $\theta \to 2$, so $\frac{2}{2-\theta}(V_0 - V_1\theta)^2$ eventually dominates $\ln\frac{2}{2-\theta}$ for almost all sample paths and Lemma 1(ii) holds. However, even for a $\theta$ value very close to 2 and thus for a very large value of $\ln\frac{2}{2-\theta}$, there is still a nonnegligible probability that $V_0 - 2V_1$ is very close to zero with the effect that $\frac{2}{2-\theta}(V_0 - V_1\theta)^2$ is dominated by $\ln\frac{2}{2-\theta}$ giving a maximum of the criterion at a value in an extremely tight (left hand) neighborhood of 2. Theorem 3 shows that this probability shrinks to zero at a rate slower than $\theta$ approaches to 2 so the density is infinite at 2. Note that the $2 - \theta$ term that appears in the denominator of (11) is $J_T(\rho) = T_1(1 + \frac{2}{T_1} - \rho) = 2 - \theta$. Thus, the source of the abnormal behavior of the limit criterion and the distribution around 2 is that for $\theta$ in a shrinking neighborhood of the upper bound 2 the component (11) of the limit criterion cannot be approximated uniformly by a quadratic in $\theta$. The limit function is therefore not locally asymptotic quadratic (LAQ) and the limit distribution is correspondingly very different from that of the unit root MLE.

*The case* $n > 1$

We next examine the case where $n > 1$ but fixed and $T \to \infty$. From (8) and (9), we have $f_{nT}(\theta) \Rightarrow f_n(\theta)$, where

$$f_n(\theta) = \sum_{i=1}^{n} V_{0i}^2 + \frac{2}{\sqrt{n}}\sum_{i=1}^{n}\tilde{W}_i\theta - \frac{1}{n}\sum_{i=1}^{n}\tilde{V}_{2i}\theta^2 - \frac{2}{2 - n^{-1/2}\theta}\sum_{i=1}^{n}(V_{0i} - n^{-1/2}V_{1i}\theta)^2 + n\ln\frac{2}{2 - n^{-1/2}\theta}.$$

We now have the restriction $\sqrt{n}T_1(\rho - 1) < 2\sqrt{n}$ or $2 - n^{-1/2}\theta > 0$, which is clearly much less restrictive for large $n$. However, for all finite $n$, $f_n(\theta)$ is still not LAQ and the global maximizer of $f_n(\theta)$ is still nonstandard — both non-normal and non-unit root class. The simulated cumulative distribution functions are drawn in Fig. 4, obtained from 5000 replications with $T = 500$. For small $n$ values, the limit distribution is far from normality, but the simulated distribution for $n = 100$ is quite close to normal and, in particular, to the $N(0, 8)$ distribution.

Theorem 3 established that the probability $P(\tilde{\theta} > 2 - \epsilon)$ is $O(\epsilon^{1/2})$ as $\epsilon \to 0$, where $\tilde{\theta}$ has the limit distribution of $T_1(\hat{\rho} - 1)$ for the case with $n = 1$. When $n > 1$, the probability of the limit distribution being close to the upper bound is much smaller as the following result shows.

**Theorem 4.** *Let* $\xi_{2i} = \tilde{W}_i + V_{1i}^2$. *Then* $P(\tilde{\theta} > 2\sqrt{n} - \epsilon) \le 2\left(3e\epsilon/\sqrt{n}\right)^{n/2} + \left(4\epsilon^2/n^2\right)E(\xi_{2i}^2)$.

The density of the limit distribution of $n^{1/2}T_1(\hat{\rho} - 1)$ at the upper bound $2\sqrt{n}$ is finite for $n = 2$ and zero for $n \ge 3$, which can be verified by differentiating $P(\tilde{\theta} > 2\sqrt{n} - \epsilon)$ and evaluating at $\epsilon = 0$. Note that Theorem 4 covers cases with fixed $n$, although it is suggestive that the upper bound becomes unimportant as $n$ increases. For large $n$, we have an asymptotic normal result, as presented in the following section.

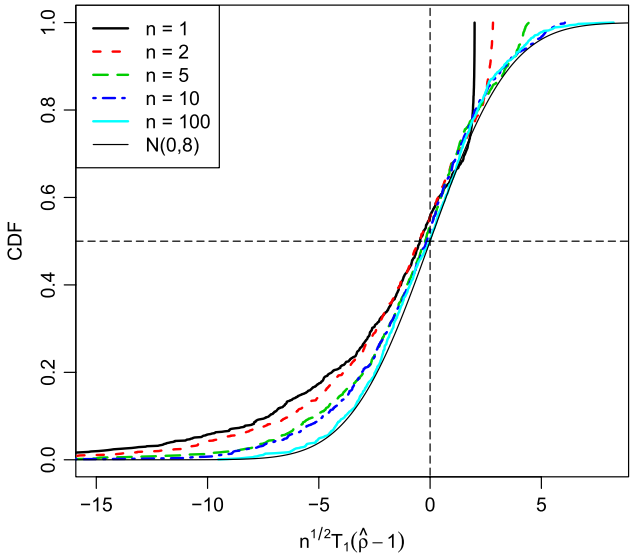


**Fig. 4.** Empirical distribution functions for $\sqrt{n}T_1(\hat{\rho}-1)$ for $n = 1, 2, 5, 10, 50, 100$ and $T = 500$.

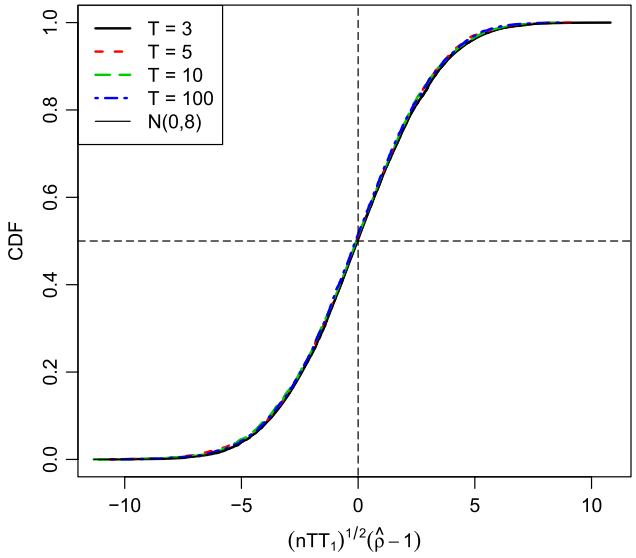


**Fig. 5.** Empirical distribution functions for $\sqrt{nTT_1}(\hat{\rho}-1)$ for $n = 500$ and $T = 3, 5, 10, 100$.

### 3.2. Large-n asymptotics

In this subsection, we let $n \to \infty$. Kruiniger (2008) established consistency and asymptotics for this case using general arguments on MLE by Newey and McFadden (1994). But the validity of this application has yet to be verified because the FDMLE is not an MLE, as noted earlier, and the conditions for application of their result do not hold. Specifically, the parameter space is not compact and $\ln L(\rho, \sigma^2)$ in (2) does not necessarily converge uniformly when rescaled appropriately. Moreover, the objective function is not concave, so pointwise convergence (e.g., Theorem 2.7 of Newey and McFadden (1994)) does not suffice. In fact, general results of this type for consistency are not applicable to the FDMLE. Instead, the objective function must be carefully investigated in order to establish the asymptotic behavior of the FDMLE. The task is nontrivial and is related with the particular form of the extended likelihood that is used in FDML estimation. This form produces a particular limit distribution which, while Gaussian, is dependent on the nature of the extension of the likelihood function to the region where $\rho > 1$.

Once these technical details are fully considered, we have the following consistency result as $n \to \infty$ for the case $\rho_0 = 1$. Here consistency is established in terms of the limiting behavior of the probability function of $T_1(\hat{\rho} - 1)$. The arguments present a nontrivial technical challenge and may be useful in other cases where usual asymptotic results fail.

**Theorem 5.** *If* $\rho_0 = 1$, $\lim_{n\to\infty} \sup_T P\{T_1|\hat{\rho} - 1| \geq \epsilon\} = 0$ *for all* $\epsilon > 0$.

From (6) and (7), we get

$$f_{nT}(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( \frac{1}{T_1} \sum_{t=1}^{T} \frac{z_{it-1}\varepsilon_{it}}{\sigma^2} \right) \theta - \left( \frac{T}{8T_1} \right) \theta^2 + o_p(1),$$

where the $o_p(1)$ term converges to zero pointwise in $\theta$ and uniformly on every compact set as well. (See Han and Phillips (2010) for a detailed derivation with $\theta$ reparameterized to $\frac{T}{T_1}\theta$.) However, the local uniformity of convergence is not sufficient for the argmax theorem (e.g., Theorem 3.2.2 of Van der Vaart and Wellner, 1996), and we also need the $O_p(\sqrt{n}T)$ convergence rate for $\hat{\rho}$. This important aspect of the proof has gone unnoticed in the literature. Unlike the case of fixed $n$, the tightness of $\hat{\theta} \equiv \sqrt{n}T_1(\hat{\rho} - 1)$ is not at all obvious as the upper bound $(2\sqrt{n})$ expands with $n$. Theorem 5 is particularly important for the convergence rate because it prevents $\hat{\rho}$ from approaching the upper bound (which depends on $T$). As a result the following property holds as $n \to \infty$ (irrespective of the size of $T$).

**Theorem 6.** *As* $n \to \infty$, $\sqrt{n}T_1(\hat{\rho} - 1) = O_p(1)$.

In view of Theorem 6, we are now able to invoke the argmax theorem to establish asymptotics for $\sqrt{n}T_1(\hat{\rho}-1)$ as the maximizer of the weak limit of $f_{nT}(\theta)$. By further letting $\theta_* = \sqrt{T/T_1}\theta$ so that $\theta = \sqrt{T_1/T}\theta_*$, we have

$$f_{nT}(\sqrt{T_1/T}\theta_*) = W_{nT}\theta_* - \frac{\theta_*^2}{8} + o_p(1),$$

$$W_{nT} = \left(\frac{T_1}{T}\right)^{1/2} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( \frac{1}{T_1} \sum_{t=1}^{T} \frac{z_{it-1}\varepsilon_{it}}{\sigma^2} \right),$$

$$\Rightarrow f_*(\theta_*) \equiv W\theta_* - \frac{\theta_*^2}{8}, \quad W \sim N\left(0, \frac{1}{2}\right),$$

where the variance of $W_{nT}$ is $\frac{1}{2}$ for all $T$, and so is that of $W$. The limit is maximized at $4W \sim N(0, 8)$. Noting that $f_{nT}(\sqrt{T_1/T}\theta_*)$ is maximized at $\sqrt{nTT_1}(\hat{\rho} - 1)$, we have the final result that

$$\sqrt{nTT_1}(\hat{\rho} - 1) \to_d N(0, 8) \tag{12}$$

as $n \to \infty$ by virtue of the argmax continuous mapping theorem regardless of the size of $T$. This limit theory provides a rigorous proof of the result in Kruiniger (2008). The standardization in (12) confirms the $\sqrt{nT}$ convergence rate as $n, T \to \infty$.

Asymptotic normality results from the fact that $J_T(1 + \theta/(\sqrt{n}T_1))$ converges to a constant and higher order terms become negligible as $n \to \infty$. The $N(0, 8)$ limit distribution for $\sqrt{nTT_1}(\hat{\rho} - 1)$ is valid for all $T$, whether small or large, as long as $n \to \infty$. The same limit is obtained as $T \to \infty$ and then $n \to \infty$ sequentially as well.

Simulated cumulative distribution functions from 5000 replications for $n = 500$ and $T = 3, 5, 10, 100$ are drawn in Fig. 5 and confirm the accuracy of this large $n$ limit theory even for small $T \geq 3$.

We finally note that the $N(0, 8)$ asymptotic distribution (12) for the unit root case with large $n$ is obtained because the log-likelihood for $\rho \leq 1$ is extended to the explosive region in a particular way by FDML estimation. In fact, the analytical extension

of $\ln L(\theta, \sigma^2)$ to the mildly explosive domain is so smooth at unity that the asymptotic distribution is symmetric on both sides of the origin. If $\ln L(\rho, \sigma^2)$ is extended to the explosive region in a different way, the limit distribution for large $n$ may not even be Gaussian. For example, if the objective function is $\ln L(\rho, \sigma^2) \cdot \mathbf{1}\{-1 < \rho \leq 1\} - \infty \cdot \mathbf{1}\{\rho > 1\}$, which is equivalent to restricting the parameter space to $(-1, 1]$, then the limit distribution of $\sqrt{nTT_1}(\hat{\rho} - 1)$ is $4W\{W \leq 0\}$ for $W \sim N(0, \frac{1}{2})$. In this case the distribution is one sided with a point mass at the origin and overall variance less than the limit variate of the FDML estimator in (12). As another example, if the objective function is $\psi(\rho, \sigma^2) \equiv \ln L(\rho, \sigma^2) \cdot \mathbf{1}\{-1 < \rho \leq 1\} + \ln L(2\rho, \sigma^2) \cdot \mathbf{1}\{\rho > 1\}$, then the concentrated localized objective function $f_{nT}(\theta) \equiv 2[\psi^*(\rho_{nT}) - \psi^*(1)]$, where $\psi^*(\rho) \equiv \max_{\sigma^2>0} \psi(\rho, \sigma^2)$ and $\rho_{nT} = 1 + \theta/(\sqrt{n}T)$, has the following asymptotic form

$$f_{nT}(\sqrt{T/T_1}\theta_*) \Rightarrow \left(W\theta_* - \frac{\theta_*^2}{8}\right)\{\theta_* \leq 0\} + 2\left(W\theta_* - \frac{\theta_*^2}{4}\right)\{\theta_* > 0\},$$

where $W \sim N(0, \frac{1}{2})$. The maximizer of the right hand side is $4W\{W \leq 0\} + 2W\{W > 0\}$, which is non-normal. Other extensions (e.g., using $\sqrt{\rho}$ in place of $2\rho$ for the explosive domain in the $\psi(\rho, \sigma^2)$ function above) can even make the convergence rate slower than the $\sqrt{nTT_1}$ rate. These examples reveal that the shape of the limit distribution, its variance, and the rate of convergence are all contingent on the form of the extension used for the likelihood function to the region $\rho > 1$.

## 4. Conclusion

As argued in earlier work by HPT (2002), transforming the likelihood offers certain key advantages in dynamic panel data modeling and estimation. The removal of incidental parameters and the transformation to stationarity by differencing when there is a unit autoregressive root make the FDMLE approach particularly appealing. There also appeared to be efficiency gains in the use of FDMLE over conventional and bias corrected MLE (applied without stationarity conditions), even in the limit theory as $n \to \infty$.

The present paper provides a rigorous justification for these heuristics and develops a complete asymptotic theory that covers the unit root case when $T \to \infty$ and $n$ is finite or tends to infinity. As shown here, the FDML criterion function combines the Gaussian likelihood over the stationary part of the domain of definition with an analytic extension of that likelihood into the nonstationary region where it is not the true likelihood. When $n$ is finite, the restrictions in the FDMLE are binding and affect the support and the form of the distribution. The restrictions even bound the domain of the limit distribution when $T \to \infty$ for finite $n$. But as $n$ increases, the bounds are much less restrictive. And when $n \to \infty$, the limit distribution is normal and normality holds even for fixed $T$ and when the autoregressive root is unity. Thus, analytically extending the likelihood in the unit root case beyond its natural domain of definition for a stationary panel is not restrictive in terms of the limit theory (and preserves asymptotic normality) provided $n$ increases and the extension is smooth. The parameter space widens as $n$ increases and the support of the limit distribution as $n \to \infty$ is the whole real line. Nonetheless, the effects of the domain restrictions and the implied stationarity condition on the differences (from the extension of the stationary likelihood) result in a reduction of the limit variance in comparison with the unrestricted MLE (or bias corrected LSDV).

For all practical purposes, at least when $n$ is large and a smooth extension of the likelihood function is used, the limit normal distribution appears to be a good approximation of finite sample behavior. Only when $n$ is small do the restrictions produce severe irregularities in the criterion function. These irregularities seriously affect the reliability of conventional numerical optimization in the persistent case and they even manifest in the large $T$ limit distribution which is neither normal nor a standard unit root type and has an unusual asymptote at the upper limit of the domain of definition, which reflects the importance of the domain restriction and the shape of the extended likelihood.

## Appendix A. Proofs

We first prove Lemma 1. The expressions for $f_{nT}(\theta)$ and $f_n(\theta)$ are given in (6) and (8), respectively.

**Proof of Lemma 1.** (i) Because $\hat{\rho} = \hat{\rho}\mathbf{1}\{\hat{\rho} \leq 1\} + \hat{\rho}\mathbf{1}\{\hat{\rho} > 1\}$, it suffices to show that $\sqrt{n}T_1(\hat{\rho} - 1)\mathbf{1}\{\hat{\rho} \leq 1\} = O_p(1)$ and $\sqrt{n}T_1(\hat{\rho} - 1)\mathbf{1}\{\hat{\rho} > 1\} = O_p(1)$. The first part is standard because $\ln L^*(\rho)$ is the true concentrated log-likelihood for $\rho \in (-1, 1]$. The second term is in $(0, 2\sqrt{n})$ so it is obviously bounded when $n$ is fixed.

(ii) Fix a compact subset $K$ of $(-\infty, 2\sqrt{n})$. For given $T$, $f_{nT}(\theta)$ is defined on $(-2\sqrt{n}T_1, 2\sqrt{n})$, so $K \subset (-2\sqrt{n}T_1, 2\sqrt{n})$ for all large enough $T$. Thus, $f_{nT}(\theta)$ converges weakly to $f_n(\theta)$ pointwise. The weak convergence is also uniform over all $\theta \in K$ because in $K$, $f_{nT}(\theta)$ is uniformly continuous for all large enough $T$ and finite almost surely.

(iii) Almost surely $\tilde{V}_{2i}$ and $V_{1i}^2$ are strictly positive, so $g_n(\theta) \to \infty$ almost surely as $\theta \to -\infty$. Also $\ln \frac{2}{2-\theta/\sqrt{n}} \to -\infty$ as $\theta \to -\infty$. Thus, $f_n(\theta) = -g_n(\theta) + n \ln \frac{2}{2-\theta/\sqrt{n}} \to -\infty$ almost surely as $\theta \to -\infty$. Next, for the case $\theta \uparrow 2\sqrt{n}$, we have

$$\begin{aligned} f_n(\theta) &= \left[\sum_{i=1}^n V_{0i}^2 + \frac{2\theta}{\sqrt{n}}\sum_{i=1}^n \tilde{W}_i - \frac{\theta^2}{n}\sum_{i=1}^n \tilde{V}_2\right] \\ &\quad + \left[n \ln \frac{2}{2-\theta/\sqrt{n}} - \frac{2}{2-\theta/\sqrt{n}}\sum_{i=1}^n (V_{0i} - V_{1i}\theta)^2\right] \\ &= f_n^{(1)}(\theta) + f_n^{(2)}(\theta). \end{aligned}$$

As $\theta \uparrow 2\sqrt{n}$, $f_n^{(1)}(\theta)$ converges to a tight random variable, and with probability 1, $\lim_{\theta\uparrow 2\sqrt{n}}(V_{0i} - V_{1i}\theta)^2 > 0$ for all $i$, implying that $\lim_{\theta\uparrow 2\sqrt{n}} f_n^{(2)}(\theta) = -\infty$ almost surely as claimed.

(iv) The global maximizer $\tilde{\theta}$ is in $(-\infty, 2\sqrt{n})$ by (iii) and the continuity of $f_n(\theta)$. Also the differentiability of $f_n(\theta)$ implies that $f_n'(\tilde{\theta}) = 0$. □

**Proof of Theorem 2.** By Lemma 1(ii), as $T \to \infty$, $f_{nT}(\theta) \Rightarrow f_n(\theta)$ uniformly in every compact subset of $(-\infty, 2\sqrt{n})$. The limit process $f_n(\theta)$ has continuous sample paths, and by Lemma 1(iii), the global maximizer of $f_n(\theta)$ exists. Lemma 1(i) verifies that $T_1(\hat{\rho} - 1)$ is tight. The probability of $f_n(\theta)$ having multiple global maxima is zero, and the result follows from a standard argmax theorem (e.g., Corollary 5.58 of Van der Vaart, 1998). □

For the next proofs we need the following preliminaries. First

$$-\frac{(V_{0i} - n^{-1/2}V_{i1}\theta)^2}{2 - n^{-1/2}\theta} = n^{-1/2}V_{1i}^2\theta + 2(V_{1i}^2 - V_{0i}V_{1i}) + \frac{(V_{0i} - 2V_{1i})^2}{2 - n^{-1/2}\theta}.$$

Thus, letting $\xi_{1i} = V_{0i} - 2V_{1i}$ and $\xi_{2i} = \tilde{W}_i + V_{1i}^2$, we have

$$f_n(\theta) = f_{an}(\theta) + f_{bn}(\theta), \tag{13}$$

where $f_{an}(\theta) = \sum_{i=1}^n \xi_{1i}^2 + \frac{2}{\sqrt{n}} \sum_{i=1}^n \xi_{2i}\theta - \frac{1}{n}\sum_{i=1}^n \tilde{V}_{2i}\theta^2$ and $f_{bn}(\theta) = n \ln \frac{2}{2-n^{-1/2}\theta} - \frac{2}{2-n^{-1/2}\theta}\sum_{i=1}^n \xi_{1i}^2$. The first derivatives are

$$f'_{an}(\theta) = 2\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_{2i} - \frac{1}{n}\sum_{i=1}^n \tilde{V}_{2i}\theta\right),$$

$$f'_{bn}(\theta) = \frac{2\sqrt{n}}{(2-n^{-1/2}\theta)^2}\left(\frac{2-n^{-1/2}\theta}{2} - \frac{1}{n}\sum_{i=1}^n \xi_{1i}^2\right). \tag{14}$$

**Proof of Theorem 3.** (i) This a special case of Theorem 4 with $n = 1$.

(ii) As $n = 1$, we omit the $i$ and $n$ subscripts. Let $\xi_1 = V_0 - 2V_1$ and $\xi_2 = \tilde{W} + V_1^2$. From (13), we have $f(\theta) = f_a(\theta) + f_b(\theta)$, where $f_a(\theta) = \xi_1^2 + 2\xi_2\theta - \tilde{V}_2\theta^2$ and $f_b(\theta) = \ln\frac{2}{2-\theta} - \frac{2}{2-\theta}\xi_1^2$. Fix $\theta_0$. Let $\tilde{\theta}$ be the global maximizer of $f(\theta)$. Almost surely, we have

$$\begin{aligned}
\left\{\tilde{\theta} > \theta_0\right\} &\Leftarrow \left\{\sup_{\theta_0<\theta<2} f(\theta) > \sup_{\theta\le\theta_0} f(\theta)\right\} \\
&\Leftarrow \left\{\sup_{\theta_0<\theta<2} f(\theta) > \sup_{\theta\le\theta_0} f_a(\theta) + \sup_{\theta\le\theta_0} f_b(\theta)\right\} \\
&\Leftarrow \left\{\inf_{\theta_0<\theta<2} f_a(\theta) + \sup_{\theta_0<\theta<2} f_b(\theta)\right. \\
&\quad \left. > \sup_{\theta\le\theta_0} f_a(\theta) + \sup_{\theta\le\theta_0} f_b(\theta)\right\} \\
&\Leftarrow \left\{\sup_{\theta_0<\theta<2} f_b(\theta) - \sup_{\theta\le\theta_0} f_b(\theta) > \sup_{\theta\le\theta_0} f_a(\theta)\right. \\
&\quad \left. - \inf_{\theta_0<\theta<2} f_a(\theta) =: \eta\right\},
\end{aligned}$$

where $\eta \ge 0$ because $f_a(\theta)$ is unimodal and continuous. Because $f_a(\theta)$ is unimodal and is globally maximized at $\theta = \tilde{V}_2^{-1}\xi_2$, we have $\eta = 0$ if $\tilde{V}_2^{-1}\xi_2 \ge 2$ (i.e., if $2\tilde{V}_2 \le \xi_2$). Thus, we have $\tilde{\theta} > \theta_0$ if (i) $2\tilde{V}_2 \le \xi_2$ and (ii) $\sup_{\theta_0<\theta<2} f_b(\theta) > \sup_{\theta\le\theta_0} f_b(\theta)$. But (ii) happens only if (iii) $f_b'(\theta_0) > 0$ (otherwise the left hand side is zero), i.e., $\frac{2}{2-\theta_0}\xi_1^2 < 1$, so (i) and (ii) are equivalent to (i)–(iii). When (iii) is true, we have $\sup_{\theta_0<\theta<2} f_b(\theta) = -\ln\xi_1^2 - 1$ and $\sup_{\theta\le\theta_0} f_b(\theta) = f_b(\theta_0) = \ln\frac{2}{2-\theta_0} - \frac{2}{2-\theta_0}\xi_1^2$. Under (iii), writing (ii) as $\frac{2}{2-\theta_0}\xi_1^2 - \ln\frac{2}{2-\theta_0}\xi_1^2 > 1$, we see that (ii) is implied by (iii) almost surely. Thus, (i)–(iii) are almost surely equivalent to (i) and (iii). Thus far, we have established that $\tilde{\theta} > \theta_0$ if $2\tilde{V}_2 \le \xi_2$ and $\frac{2}{2-\theta_0}\xi_1^2 < 1$ almost surely, implying that

$$P(\tilde{\theta} > \theta_0) \ge P(\xi_1^2 < \epsilon_0, 2\tilde{V}_2 \le \xi_2), \quad \epsilon_0 = 1 - \theta_0/2.$$

But we have $2\tilde{V}_2 \le \xi_2$ if and only if $\xi_1^2 \ge 4\tilde{V}_2 - 2V_0V_1 + 2V_1^2 + 1$ (which can be shown by using the fact that $W = \frac{1}{2}(V_0^2 - 1)$ almost surely), so the probability on the right hand side is

$$P(4\tilde{V}_2 - 2V_0V_1 + 2V_1^2 + 1 \le \xi_1^2 < \epsilon_0),$$

which is greater than or equal to $P(-\sqrt{\epsilon_0} < \xi_1 < \sqrt{\epsilon_0}, 4\tilde{V}_2 - 2V_0V_1 + 2V_1^2 + 1 \le 0)$.

Let $\theta_0 \ge 0$ so $\epsilon_0 \le 1$. In the event that $\xi_1 > -\sqrt{\epsilon_0}$, i.e., when $V_0 > 2V_1 - \sqrt{\epsilon_0}$, we have $-2V_0V_1 < -4V_1^2 + 2V_1\sqrt{\epsilon_0} \le -4V_1^2 + 2V_1$, so the above displayed probability is at least as large as $P(-\sqrt{\epsilon_0} < V_0 - 2V_1 < \sqrt{\epsilon_0}, 4\tilde{V}_2 - 2V_1^2 + 2V_1 + 1 \le 0)$, where we used $\xi_1 = V_0 - 2V_1$. Conditional on $V_1$ and $\tilde{V}_2$, the density of $V_0$ is almost surely positive at $2V_1$, and $P(4\tilde{V}_2 - 2V_1^2 + 2V_1 + 1 \le 0) > 0$, so the last probability is of order $\sqrt{\epsilon_0}$. □

When we generalize the previous result to $n > 1$, the uniform probability bound

$$P(\chi_n^2 \le x) \le (ex/n)^{n/2} \tag{15}$$

is useful. This holds because

$$\begin{aligned}
P(\chi_n^2 \le x) &= \frac{1}{\Gamma(n/2)}\int_0^{x/2} z^{n/2-1}e^{-z}dz \le \frac{1}{\Gamma(n/2)}\int_0^{x/2} z^{n/2-1}dz \\
&= \frac{(x/2)^{n/2}}{(n/2)\Gamma(n/2)} = \frac{(n/2)^{n/2}}{(n/2)\Gamma(n/2)e^{n/2}}\cdot\frac{(ex/2)^{n/2}}{(n/2)^{n/2}} \\
&\le \left(\frac{ex}{n}\right)^{n/2},
\end{aligned}$$

where $\Gamma(s) = \int_0^\infty x^{s-1}e^{-x}dx$, and $(n/2)^{n/2}/[(n/2)\Gamma(n/2)e^{n/2}] \le 1$. (A proof of the latter inequality is provided in Han and Phillips, 2010).

**Proof of Theorem 4.** Let $\epsilon \le 2/(3e)$ be given. Let $c = 2\sqrt{n} - \epsilon$ so $\epsilon = 2\sqrt{n} - c = \sqrt{n}(2 - n^{-1/2}c)$. Let $\tilde{\theta}$ denote the global maximizer of $f_n(\theta)$ again. We have $\tilde{\theta} > c \Leftrightarrow \sup_{c<\theta<2\sqrt{n}} f_n(\theta) > \sup_{\theta\le c} f_n(\theta) \Rightarrow \sup_{c<\theta<2\sqrt{n}} f_n(\theta) > f_n(c)$. Let $A_n = \{\sup_{c<\theta<2\sqrt{n}} f_n(\theta) > f_n(c)\}$ and $B_n = \{f'_{bn}(c) < 0\}$, where $f_{bn}(\theta)$ is defined below (13). Because of (14), $B_n = \{\sum_{i=1}^n \xi_{1i}^2 > \frac{1}{2}\sqrt{n}\epsilon\}$, where $\xi_{1i} = V_{0i} - 2V_{1i} \sim N(0, \frac{1}{3})$. Clearly

$$P(A_n) = P(A_n \cap B_n^c) + P(A_n \cap B_n) \le P(B_n^c) + P(A_n \cap B_n). \tag{16}$$

For $P(B_n^c)$, because $3\sum_{i=1}^n \xi_{1i}^2 \sim \chi_n^2$, we have

$$P(B_n^c) = P\left(3\sum_{i=1}^n \xi_{1i}^2 \le \frac{3}{2}\sqrt{n}\epsilon\right) \le \left(\frac{3e\epsilon}{2\sqrt{n}}\right)^{n/2}, \tag{17}$$

for all $n$ by (15). Next, in the event $B_n$, because $f_{bn}(\theta)$ is unimodal, we have not only $f'_{bn}(c) < 0$ but also $f'_{bn}(\theta) < 0$ for all $\theta \in (c, 2\sqrt{n})$. But from (14), we have

$$\begin{aligned}
f''_{bn}(\theta) &= \frac{4n^{-1/2}}{(2-n^{-1/2}\theta)^3}\left(\sqrt{n} - \frac{\theta}{2} - \frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_{1i}^2\right) \\
&\quad - \frac{1}{(2-n^{-1/2}\theta)^2} \\
&= \frac{2}{(2-n^{-1/2}\theta)^2}\left[\frac{2n^{-1/2}}{2-n^{-1/2}\theta}\cdot f'_{bn}(\theta) - \frac{1}{2}\right],
\end{aligned}$$

which is strictly negative on $B_n$ for all $\theta \in (c, 2\sqrt{n})$ because $f'_{bn}(\theta) < 0$. Also $f''_{an}(\theta) < 0$ globally and thus for $\theta \in (c, 2\sqrt{n})$ as well. Hence, $f''_n(\theta) < 0$ for all $\theta \in (c, 2\sqrt{n})$ in the event of $B_n$. Thus, on $B_n$, $f_n(\theta) - f(c) \le (\theta - c)f'_n(c)$ for all $\theta \ge c$, implying that $\sup_{c<\theta<2\sqrt{n}} f_n(\theta) - f(c) \le (2\sqrt{n} - c)f'_n(c)$. But on $A_n$, the left hand side is positive, so, on $A_n \cap B_n$, the right hand side is also positive, i.e., $f'_n(c) > 0$, where

$$f'_n(c) = 2\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_{2i} - \frac{1}{n}\sum_{i=1}^n \tilde{V}_{2i}c\right) + \frac{n}{\epsilon} - \frac{2\sqrt{n}}{\epsilon^2}\sum_{i=1}^n \xi_{1i}^2,$$
$$\epsilon = 2\sqrt{n} - c.$$

Recall that $\xi_{2i} = \tilde{W}_i + V_{1i}^2$. We have

$$\begin{aligned}
P(A_n \cap B_n) \le P\left\{2\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_{2i} - \frac{1}{n}\sum_{i=1}^n \tilde{V}_{2i}c\right) + \frac{n}{\epsilon}\right. \\
\left. - \frac{2\sqrt{n}}{\epsilon^2}\sum_{i=1}^n \xi_{1i}^2 > 0\right\}
\end{aligned}$$

$$= P\left\{\sum_{i=1}^n \xi_{1i}^2 < \frac{\epsilon}{2}\sqrt{n} + \epsilon^2\left(\frac{1}{n}\sum_{i=1}^n \xi_{2i} - \frac{1}{n^{3/2}}\sum_{i=1}^n \tilde{V}_{2i}c\right)\right\}$$
$$\le P\left(\sum_{i=1}^n \xi_{1i}^2 < \frac{\epsilon}{2}\sqrt{n} + \epsilon^2\bar{\xi}_2\right), \qquad \bar{\xi}_2 := \frac{1}{n}\sum_{i=1}^n \xi_{2i},$$

where the last inequality holds because $\tilde{V}_{2i} \ge 0$ and $c > 0$. But

$$P(A_n \cap B_n) = P(A_n \cap B_n \cap \{\bar{\xi}_2 \le r\}) + P(A_n \cap B_n \cap \{\bar{\xi}_2 > r\})$$
$$\le P\left(\sum_{i=1}^n \xi_{1i}^2 < \frac{\epsilon}{2}\sqrt{n} + \epsilon^2 r\right) + P(\bar{\xi}_2 > r),$$

for any $r$. In particular, for $r = \sqrt{n}/(2\epsilon)$, we have

$$P(A_n \cap B_n) \le P\left(\sum_{i=1}^n \xi_{1i}^2 < \sqrt{n}\epsilon\right) + P\left(\bar{\xi}_2^2 > \frac{n}{4\epsilon^2}\right)$$
$$\le \left(\frac{3e\epsilon}{\sqrt{n}}\right)^{n/2} + \left(\frac{4\epsilon^2}{n^2}\right)E(\xi_{2i}^2), \tag{18}$$

where the first term of the right hand side is due to (15) and the second term by Chebyshev's inequality. The result now follows from (16) to (18). □

Now we prove that $T_1(\hat{\rho} - 1)$ is consistent when $n \to \infty$ regardless of the $T$ sequence, i.e., that $\lim_{n\to\infty} \sup_T P\{T_1|\hat{\rho} - 1| \ge \epsilon\} = 0$ for every $\epsilon > 0$.

Reparameterize $\rho$ to $\phi = T_1(\rho - 1)$ by letting $\rho_T = 1 + \frac{1}{T_1}\phi$ for given $\phi$. Then $J_T(\rho_T) = 2 - \phi$. Let $\hat{\ell}_{nT}(\phi) := \frac{2T_1}{nT}[\ln L^*(\rho_T) - \ln L^*(1)]$. Let $\tilde{\sigma}^2 = \frac{1}{nT}\sum_{i=1}^n\sum_{t=1}^T \varepsilon_{it}^2$, $\hat{a}_{nT}(\rho) = \frac{1}{nT\tilde{\sigma}^2}\sum_{i=1}^n\sum_{t=1}^T u_{it}(\rho)^2$ and $\hat{b}_{nT}(\rho) = \frac{1}{nT\tilde{\sigma}^2}\sum_{i=1}^n[\sum_{t=1}^T u_{it}(\rho)]^2$. Then

$$\hat{\ell}_{nT}(\phi) = -T_1 \ln[\hat{a}_{nT}(\rho_T) + \frac{\phi}{T_1(2-\phi)}\hat{b}_{nT}(\rho_T)] - \frac{T_1}{T}\ln(2-\phi) + \frac{T_1}{T}\ln(1+\rho_T). \tag{19}$$

The parameter space is $(-2T_1, 2)$, and the strict positivity of $\hat{a}_{nT}(\rho_T) + \frac{\phi}{T_1(2-\phi)}\hat{b}_{nT}(\rho_T)$ is verified by the fact that $\hat{a}_{nT}(\rho_T) + \frac{\phi}{T_1(2-\phi)}\hat{b}_{nT}(\rho_T) = [\hat{a}_{nT}(\rho_T) - \frac{1}{T}\hat{b}_{nT}(\rho_T)] + \frac{2T_1-\phi}{TT_1(2-\phi)}\hat{b}_{nT}(\rho_T)$ and that $\frac{2T_1-\phi}{TT_1(2-\phi)} > 0$ on the parameter space. The maximizer of $\hat{\ell}_{nT}(\phi)$ is $\hat{\phi} = T_1(\hat{\rho} - 1)$, where $\hat{\rho}$ is the FDMLE. As $u_{it}(\rho_T) = \varepsilon_{it} - \frac{\phi}{T_1}z_{it-1}$, we have

$$\hat{a}_{nT}(\rho_T) = 1 - \frac{2\phi}{nT_1\tilde{\sigma}^2}\sum_{i=1}^n\frac{1}{T}\sum_{t=1}^T z_{it-1}\varepsilon_{it} + \frac{\phi^2}{nT_1\tilde{\sigma}^2}\sum_{i=1}^n\frac{1}{TT_1}\sum_{t=1}^T z_{it-1}^2 \to_p 1 + \frac{\phi^2}{2T_1} =: a_T(\rho_T)$$

uniformly on every compact set of $\phi$ as $n \to \infty$ when $\rho_0 = 1$. Similarly,

$$\hat{b}_{nT}(\rho_T) = 1 + \frac{2}{n\tilde{\sigma}^2}\sum_{i=1}^n\frac{1}{T}\sum_{s=1}^{T-1}\sum_{t=s+1}^T \varepsilon_{is}\varepsilon_{it} - \frac{2\phi}{n\tilde{\sigma}^2}\sum_{i=1}^n\left(\frac{z_{iT}}{\sqrt{T}}\right)\left(\frac{1}{T_1}\sum_{t=1}^T\frac{z_{it-1}}{\sqrt{T}}\right) + \frac{\phi^2}{n\tilde{\sigma}^2}\sum_{i=1}^n\left(\frac{1}{T_1}\sum_{t=1}^T\frac{z_{it-1}}{\sqrt{T}}\right)^2$$
$$\to_p 1 - \phi + \frac{2T_1+1}{6T_1}\phi^2 =: b_T(\rho_T)$$

as $n \to \infty$ at the same mode. Also, after some further algebra, we have

$$\hat{\eta}_a(\rho_T) := \sqrt{n}T_1[\hat{a}_{nT}(\rho_T) - a_T(\rho_T)] = \phi O_p(1) + \phi^2 O_p(1), \tag{20}$$
$$\hat{\eta}_b(\rho_T) := \sqrt{n}[\hat{b}_{nT}(\rho_T) - b_T(\rho_T)] = O_p(1) + \phi O_p(1) + \phi^2 O_p(1), \tag{21}$$

where the $O_p(1)$ terms are stochastically bounded as $n \to \infty$ for all $T$ sequences whether fixed or diverging.

Let $\ell_T(\phi)$ be obtained by replacing $\tilde{a}_{nT}(\cdot)$ and $\hat{b}_{nT}(\cdot)$ in (19) with $\tilde{a}_T(\cdot)$ and $b_T(\cdot)$ respectively. The $n$-limit $\ell_T(\phi)$ is maximized at $\phi_0 = 0$ for all $T$ by the usual information inequality (e.g., Newey and McFadden, 1994, Lemma 2.2) even though $\hat{\ell}_{nT}(\phi)$ is not a correct profile log-likelihood for $\phi > 0$, i.e., for $\rho_T > 1$ (which is verified later). This identifiability of $\phi_0 = 0$ is uniform in $T$ as the following lemma shows.

**Lemma 7.** *If* $\rho_0 = 1$, *then* $\inf_T[\sup_{\phi:|\phi|\ge\delta}\ell_T(\phi) - \ell_T(0)] < 0$ *for all* $\delta > 0$.

**Proof.** We have,

$$\ell_T(\phi) = -T_1\ln\left[(2-\phi)a_T(\rho_T) + \frac{\phi}{T_1}b_T(\rho_T)\right] + \frac{T_1^2}{T}\ln(2-\phi) + \frac{T_1}{T}\ln(2+\frac{1}{T_1}\phi) = -T_1\ln\left(2 - \frac{T_2}{T_1}\phi - \frac{T_2}{6T_1^2}\phi^3\right) + \frac{T_1^2}{T}\ln(2-\phi) + \frac{T_1}{T}\ln\left(2+\frac{1}{T_1}\phi\right).$$

Some lengthy algebra gives

$$\ell_T'(\phi) = -\frac{\phi(T_2\phi^3 + 2T_2^2\phi^2 - 6T_1T_2\phi + 6T_1^2)}{3T_1^2\left(2 - \frac{T_2}{T_1}\phi - \frac{T_2}{6T_1^2}\phi^3\right)(2-\phi)\left(2+\frac{1}{T_1}\phi\right)},$$

where the $T_2\phi^3 + 2T_2^2\phi^2 - 6T_1T_2\phi + 6T_1^2$ term of the numerator is strictly positive by Lemma 8. Thus, for all $T \ge 2$, $\ell_T(\phi)$ is unimodal and attains its mode at 0 because $\ell_T'(\phi) > 0$ for $\phi < 0$, $\ell_T'(0) = 0$ and $\ell_T'(\phi) < 0$ for $\phi > 0$. Further algebra gives $\ell_T''(0) = -\frac{1}{4}$ for all $T$. The stated result follows from this fact and the local uniform continuous differentiability of $\ell_T''(\phi)$. □

**Lemma 8.** *Let* $\psi_T(\phi) = T_2\phi^3 + 2T_2^2\phi^2 - 6T_1T_2\phi + 6T_1^2$. *Then* $\min_{-2T_1\le\phi\le 2}\psi_T(\phi) \ge \min\{4T_1^3 + 2T_1^2, \frac{1}{27}(15T_1^2 + 115T_1 + 32)\}$ *for all* $T$.

**Proof.** The first derivative is $\psi_T'(\phi) = 3T_2\phi^2 + 4T_2^2\phi - 6T_1T_2$. We have $\psi_T'(-2T_1) = 4T_2^3 + 10T_2^2 + 6 > 0$, $\psi_T'(\frac{4}{3}) = -\frac{2}{3}T_2T_1 < 0$, and $\psi_T'(\frac{3}{2}) = \frac{3}{4}T_2 > 0$. Thus, $\psi_T(\phi)$ can be minimal either at $-2T_1$ or in the interval $(\frac{4}{3}, \frac{3}{2})$. But $\psi_T(-2T_1) = 4T_1^3 + 2T_1^2$, and for all $\phi \in (\frac{4}{3}, \frac{3}{2})$, $\psi_T(\phi) \ge (\frac{4}{3})^3T_2 + 2(\frac{4}{3})^2T_2^2 - 6(\frac{3}{2})T_1T_2 + 6T_1^2 = \frac{1}{27}(15T_1^2 + 115T_1 + 32)$. □

The remaining argument for proving consistency is to show that the variability of $\hat{\ell}_{nT}(\phi) - \ell_T(\phi)$ diminishes as $n \to \infty$ at a proper mode. We have $\hat{\ell}_{nT}(\phi) - \ell_T(\phi) = -T_1\ln[1 + \hat{\xi}(\phi)]$, where

$$\sqrt{n}T_1\hat{\xi}(\phi) = \frac{(2-\phi)\hat{\eta}_a(\rho_T) + \phi\hat{\eta}_b(\rho_T)}{(2-\phi)a_T(\rho_T) + \frac{\phi}{T_1}b_T(\rho_T)}, \tag{22}$$

and $\hat{\eta}_a(\rho_T)$ and $\hat{\eta}_b(\rho_T)$ are defined in (20) and (21), respectively, and are quadratic functions of $\phi$ with stochastically bounded coefficients. For fixed $T$, the denominator of (22) is strictly positive for all $\phi \in [-2T_1, 2]$, and the numerator is uniformly stochastically bounded. Thus, $\hat{\xi}(\phi) \Rightarrow 0$, and $\hat{\ell}_{nT}(\phi) - \ell_T(\phi) = -T_1O_p$

$(n^{-1/2}T_1^{-1}) = O_p(n^{-1/2})$ uniformly in $\phi$. Hence, the case with fixed $T$ is easily seen to satisfy the stated requirement.

However, if $T \to \infty$, then for $\phi \simeq 2$, $\lim_{\phi\to 2}\sqrt{n}T_1\hat{\xi}(\phi) = T_1\hat{\eta}_b(1+\frac{2}{T_1})/b_T(1+\frac{2}{T_1})$, which diverges as $T \to \infty$. Thus, as $T \to \infty$, for some $n/T$ ratio, the sampling variability of $\hat{\ell}_{nT}(\phi) - \ell_T(\phi)$ may explode for $\phi$ values sufficiently close to 2 (or approaching 2 as $T \to \infty$). However, the variability of $\hat{\ell}_{nT}(\phi) - \ell_T(\phi)$ near 2 or $-2T_1$ is minor in comparison to the value of $\ell_T(\phi)$, as shown in the proof below. Thus the probability of $\hat{\phi}$ (the maximizer of $\hat{\ell}_{nT}(\phi)$) being far away from the maximizer of $\ell_T(\phi)$ diminishes to zero.

**Proof of Theorem 5.** (i) Fixed $T$: We have $\hat{\ell}_{nT}(\phi) - \ell_T(\phi) = -T_1\ln[1+\hat{\xi}(\phi)]$, where $\sqrt{n}T_1\hat{\xi}(\phi)$ is stochastically bounded uniformly in $\phi \in [-2T_1, 2]$. Thus, $\sup_\phi|\hat{\ell}_{nT}(\phi) - \ell_T(\phi)| = O_p(n^{-1/2})$ by the linear approximation of the logarithm. Consistency follows from this uniform convergence of $\hat{\ell}_{nT}(\phi)$ to $\ell_T(\phi)$ and the identification by Lemma 7.

(ii) $T \to \infty$: For any given $c_1$ and $c_2$ with $-\infty < c_1 < 0 < c_2 < 2$, we have $\sqrt{n}T_1\hat{\xi}(\phi) = O_p(1)$ as $n \to \infty$ uniformly over $\phi \in [c_1, c_2]$ regardless of $T$. Thus on $[c_1, c_2]$, $\hat{\ell}_{nT}(\phi) - \ell_T(\phi) \to_p 0$ uniformly in $\phi$ regardless of $T$. For $\phi$ outside the interval $[c_1, c_2]$, we have

$$\frac{\hat{\ell}_{nT}(\phi) - \ell_T(\phi)}{\ell_T(\phi)} = \frac{\ln[(2-\phi)\hat{a}_{nT}(\rho_T) + \frac{\phi}{T_1}\hat{b}_{nT}(\rho_T)] - \ln[(2-\phi)a_T(\rho_T) + \frac{\phi}{T_1}b_T(\rho_T)]}{\ln[(2-\phi)a_T(\rho_T) + \frac{\phi}{T_1}b_T(\rho_T)] - \frac{T_1}{T}\ln(2-\phi) - \frac{1}{T}\ln(2+\frac{1}{T_1}\phi)}.$$

The numerator converges in probability to zero uniformly in $\phi$ (over the whole domain) regardless of $T$, and the denominator is far from zero outside $[c_1, c_2]$. Thus, given Lemma 7, the probability of $\hat{\ell}_{nT}(\phi)$ being maximized by a parameter outside $[c_1, c_2]$ disappears as $n \to \infty$ regardless of $T$. It thus suffices to consider only the domain $[c_1, c_2]$, and consistency follows straightforwardly. □

**Proof of Theorem 6.** Consistency (Theorem 5) leads to the convergence rate. As $\text{plim}_{n\to\infty}\hat{\phi} = 0$, where $\hat{\phi} = T_1(\hat{\rho} - 1)$, we have $|\hat{\phi}| \le c$ with arbitrarily high probability eventually as $n \to \infty$ for any $c > 0$. We can thus limit the parameter space for $\phi$ to $[-1, 1]$. With the parameter space restricted to this region, standard theory applies. □

## References


Andrews, D., 1999. Estimation when a parameter is on a boundary. Econometrica 67, 1341–1383.

Andrews, D., 2001. Testing when a parameter is on the boundary of the maintained hypothesis. Econometrica 69, 683–734.

Ansley, C.F., 1979. An algorithm for the exact likelihood of a mixed autoregressive-moving average process. Biometrika 66, 59–65.

Galbraith, R.F., Galbraith, J.I., 1974. On the inverses of some patterned matrices arising in the theory of stationary time series. Journal of Applied Probability 11, 63–71.

Geyer, C.J., 1994. On the asymptotics of constrained $M$-estimation. Annals of Statistics 22, 1993–2010.

Han, C., 2007. Determinants of covariance matrices of differenced AR(1) processes. Econometric Theory 23, 1248–1253.

Han, C., Phillips, P.C.B., 2010. First difference MLE and dynamic panel estimation, Cowles Foundation Discussion Paper No. 1780.

Han, C., Phillips, P.C.B., Sul, D., 2011. Uniform asymptotic normality in stationary and unit root autoregression. Econometric Theory 27, 1117–1151.

Han, C., Phillips, P.C.B., Sul, D., 2013. X-differencing and dynamic panel model estimation. Econometric Theory (forthcoming).

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

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

Knight, K., 2003. Epi-convergence in distribution and stochastic equi-semicontinuity. University of Toronto, Unpublished Manuscript.

Kruiniger, H., 2008. Maximum likelihood estimation and inference methods for the covariance stationary panel AR(1)/unit root model. Journal of Econometrics 144, 447–464.

MaCurdy, T., 1982. The use of time series processes to model the time structure of earnings in a longitudinal data analysis. Journal of Econometrics 18, 83–114.

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

Neyman, J., Scott, E., 1948. Consistent estimates based on partially consistent observations. Econometrica 16, 1–32.

Phillips, P.C.B., 1987. Time series regression with a unit root. Econometrica 55, 277–301.

Van der Vaart, A.W., 1998. Asymptotic Statistics. Cambridge University Press.

Van der Vaart, A.W., Wellner, J.A., 1996. Weak Convergence and Empirical Processes: With Applications to Statistics. Springer.

Wilson, P.D., 1988. Maximum likelihood estimation using differences in an autoregressive-1 process. Communications in Statistics–Theory and Methods 17, 17–26.