# FIRST DIFFERENCE MLE AND DYNAMIC PANEL ESTIMATION

**By**

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

**January 2011**

**COWLES FOUNDATION DISCUSSION PAPER NO. 1780**

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

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

# First Difference MLE and Dynamic Panel Estimation

Chirok Han
*Korea University*

Peter C. B. Phillips
*Yale University, University of Auckland*
*University of Southampton & Singapore Management University*

December 2010

**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 peformance 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. We investigate these problems, provide a convenient new expression for the likelihood and a new algorithm to maximize it. 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. When $n, T \to \infty$, the FDMLE has smaller asymptotic variance than that of the bias corrected MLE, an outcome that is explained by the restricted nature of the FDMLE.

*Key Words*: Asymptote, Bounded support, Dynamic panel, Efficiency, First difference MLE, Likelihood, Quartic equation, Restricted extremum estimator.
*JEL classification*: C22, C23

# 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. 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. A second advantage of FDML is that the differenced data are stationary whether the original data are stationary or integrated. 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 McCurdy (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. Also, 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 extending the stationary likelihood outside the stationary region. This extension leads to further complications, including the fact that the estimator is restricted by an upper bound which affects both finite sample theory and asymptotic behavior.

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, Pesaran and Tahmiscioglu (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 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.

Most recently, Kruiniger (2008) derived asymptotics for 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. Indeed, for AR(1) panels with large $n,$ large $T$ and a unit root, the MLE is known to have a $N(0, \frac{51}{5})$ limit distribution when the bias of the MLE is corrected (Hahn and Kuersteiner, 2001). By contrast, the FDMLE is also asymptotically normal, has no asymptotic bias and its variance is $8$ (Kruiniger, 2008), thereby producing an asymptotic gain in efficiency at unity over bias corrected MLE[1]. This reduction in asymptotic variance between the two ML approaches is left unexplained. The reasons for this variance reduction over the traditional MLE will be explained 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.

[1] It has further been shown in recent work by Han, Phillips and Sul (2010, 2011) 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 $N\left(0,9\right)$ distribution after centering and standardization, and is also more efficient asymptotically than the bias corrected MLE for the autoregressive coefficient in a vicinity of unity.

## 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 $-1 < \rho_0 \leq 1$. 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

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

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 the following formula for the determinant of $C_T(\rho_0)$

$$\det C_T(\rho_0) = \frac{J_T(\rho_0)}{1 + \rho_0}, \quad J_T(\rho) \equiv (T + 1) - (T - 1)\rho$$

(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

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

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 $\rho = 1$ 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 (for time series and for panels) are to be expected and 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 the recent work by HPT (2002) and Kruiniger (2008). This domain for $(\rho, \sigma^2)$ is $(-1, 1 + \frac{2}{T-1}) \times (0, \infty)$ (Kruiniger, 2008), which contains $\rho = 1$ in its interior. By means of this analytic extension, HPT (2002) and Kruiniger (2008) 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 $\rho \in (1, 1 + \frac{2}{T-1})$ because (1) does not hold for $\rho > 1$. Thus, maximizing (2) over the whole domain does not yield an ML estimator 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 $\rho \to 1 + \frac{2}{T-1}$. Though $\ln L(\rho, \sigma^2)$ is differentiable on $(-1, 1 + \frac{2}{T-1}) \times (0, \infty)$ as Kruiniger (2008, Lemma 7) finds, the behavior of the 'log-likelihood' function may be very violent especially for small $n$. Figure 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 $1 + \frac{2}{100} = 1.02$ 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 Figure 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 may 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. 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 $\rho$. 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.

As we show later, this peculiarity happens with a non-negligible probability when $n$ is small and the true autoregressive parameter is unity. Especially, when $n = 1$, the asymptotic distribution of the FDMLE is quite unusual: its density shows one small mode at a value below the true value and an infinite asymptote at the upper bound. This pecularity disappears in probability as $n$ increases or when $\rho_0$ is well inside the stationary region, in which case asymptotic Gaussianity is attained.

Because (2) is not a proper log-likelihood for the domain $(1, 1 + \frac{2}{T-1}) \times (0, \infty)$, 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 $(-1, 1 + \frac{2}{T-1}) \times (0, \infty)$

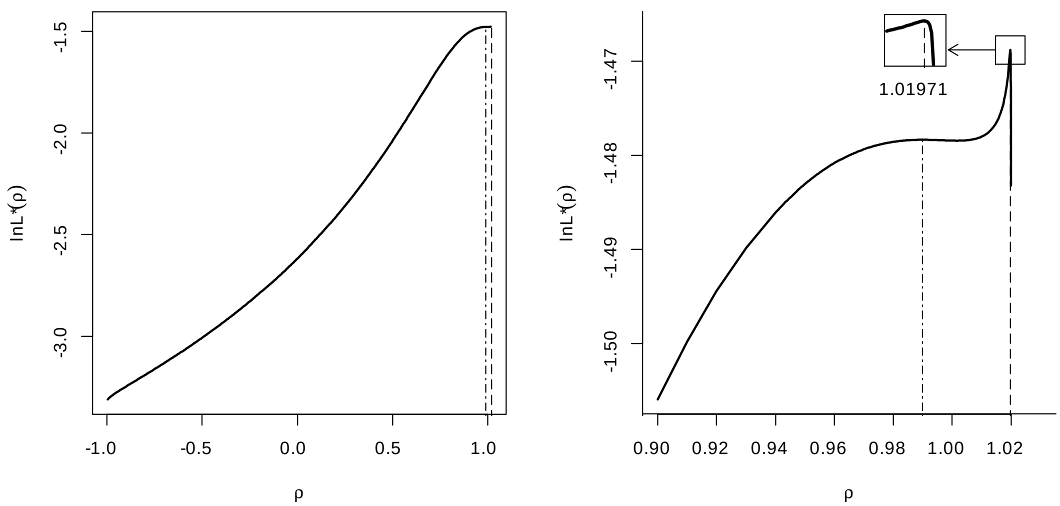


Figure 1: Multimodal average profile 'log-likelihood' for a sample path where numerical optimization finds a local maximum (the lines of alternating dots and dashes) instead of the global maximum (the dashed lines). The left graph drawn over the whole domain $(-1, 1 + \frac{2}{T-1})$ fails to reveal 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.

as Kruiniger (2008) points out. Furthermore, due to the described peculiarity of the profile 'log-likelihood' criterion near the upper bound $1 + \frac{2}{T-1}$, 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 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 providing a new expression for the criterion function which allows a direct treatment for asymptotic analysis and numerical calculation. This expression is based on the long-differenced variables ($y_{it} - y_{i0}$) and is different from those obtained by Wilson (1988) or Kruiniger (2008). As a result we establish some new unit root limit theory for the FDMLE that takes a particularly interesting and revealing form. In particular, the FDMLE is shown to have an asymptote with infinite density at the upper limit of its support, a new feature that is the result of the anomalies in the criterion function and the fact that the FDMLE is a restricted estimator. We also provide an explicit solution for the FDMLE in terms of the roots of a quartic equation which avoids problems of numerical optimization. Simulations are done using this numerical method and these corroborate the new asymptotic theory.

The rest of the paper is organized as follows. Section 3 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 Figure 1. Section 4 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, the most remarkable of which is that the density of the limit distribution has an infinite asymptote at the upper bound. Although the panel asymptotic case has already been studied in Kruiniger (2008), it is reconsidered here in the second part of Section 4. This section demonstrates how the peculiarity of the time series asymptotics is removed eventually as $n$ increases and explains why in the unit root case the asymptotic $N(0, 8)$ limit theory for the FDMLE improves on the $N(0, 51/5)$ limit theory for the bias corrected MLE. Standard asymptotics apply in the stationary case and are not considered here - we refer readers to Kruiniger (2008). Section 5 concludes. Proofs and some supplementary technical material involving the algebraic solutions of quartic and cubic equations are given in the Appendix. Throughout the remainder of the paper we use the notation $T_m = T - m$ and $\tilde{T}_m = T + m$ for convenience.

# 3 A Closed-form Expression for FDML

## 3.1 The Log-likelihood Criterion

We start by simplifying the expression for the 'log likelihood' criterion (2). In particular, the term involving $\Delta y_i' C_T(\rho)^{-1}\Delta y_i$ in (2) can be simplified considerably, as we now show.

Let $z_{it} = y_{it} - y_{i0}$ and $z_i = (z_{i1}, \ldots, z_{iT})'$. Then $z_i = H_T \Delta y_i$, where $H_T$ is the $T \times T$ lower triangular partial sum matrix (whose diagonal and lower diagonal elements are all one). Also let $u_{it}(\rho) = z_{it} - \rho z_{it-1}$ and $u_i(\rho) = [u_{i1}(\rho), \ldots, u_{iT}(\rho)]'$. Then $u_i(\rho) = D_T(\rho) z_i = D_T(\rho) H_T \Delta y_i$, where $D_T(\rho)$ is the $T \times T$ matrix whose diagonal elements are unity and whose first lower off-diagonal elements are $-\rho$. We thus have $\Delta y_i' C_T(\rho)^{-1}\Delta y_i = u_{it}(\rho)'\tilde{C}_T(\rho)^{-1} u_{it}(\rho)$, where $\tilde{C}_T(\rho) = D_T(\rho) H_T C_T(\rho) H_T' D_T(\rho)'$. Now, as shown in Appendix A (where both algebraic and statistical proofs are given), we have the explicit form

(3) $\tilde{C}_T(\rho) = I_T + \frac{1-\rho}{1+\rho} 1_T 1_T',$

where $1_T$ is the $T$-vector of ones. The inverse of $\tilde{C}_T(\rho)$ is then

$$\tilde{C}_T(\rho)^{-1} = I_T - \tfrac{1-\rho}{J_T(\rho)} 1_T 1_T', \quad J_T(\rho) = \tilde{T}_1 - T_1 \rho.$$

Hence, (2) may be rewritten as

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

$$Q_{iT}(\rho) \equiv \sum_{t=1}^T u_{it}(\rho)^2 - \frac{1-\rho}{J_T(\rho)}\left[\sum_{t=1}^T u_{it}(\rho)\right]^2, \tag{5}$$

where $u_{it}(\rho) = z_{it} - \rho z_{it-1}$ with $z_{it} = y_{it} - y_{i0}$. Note that $Q_{iT}(\rho) = \Delta y_i' C_T(\rho)^{-1}\Delta y_i$, which is strictly positive if $-1 < \rho < 1 + \frac{2}{T-1}$ and $\Delta y_i \neq 0$. Defining

$$c_0 = \sum_{i=1}^n \sum_{t=1}^T z_{it}^2, \;\; c_1 = \sum_{i=1}^n \sum_{t=1}^T z_{it-1} z_{it}, \;\; c_2 = \sum_{i=1}^n \sum_{t=1}^T z_{it-1}^2,$$
$$d_0 = \sum_{i=1}^n \left(\sum_{t=1}^T z_{it}\right)^2, \;\; d_1 = \sum_{i=1}^n \left(\sum_{t=1}^T z_{it-1}\right)\left(\sum_{t=1}^T z_{it}\right), \;\; d_2 = \sum_{i=1}^n \left(\sum_{t=1}^T z_{it-1}\right)^2,$$

we obtain the following simple form

$$\sum_{i=1}^n Q_{iT}(\rho) = (c_0 - 2c_1\rho + c_2\rho^2) - \frac{1-\rho}{J_T(\rho)}(d_0 - 2d_1\rho + d_2\rho^2).$$

### 3.2 Existence of the FDMLE

Let $\ln L^*(\rho)$ denote the profile log likelihood function $\ln L^*(\rho) = \max_{\sigma^2} \ln L(\rho, \sigma^2)$. 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

$$\frac{\partial \ln L(\rho, \sigma^2)}{\sigma^2} = -\frac{nT}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} Q_{iT}(\rho).$$

Thus, the maximizer is $\sigma^2 = (nT)^{-1} \sum_{i=1}^{n} Q_{iT}(\rho)$ for given $\rho$, and then

$$\text{(6)} \quad \ln L^*(\rho) = -\frac{nT}{2} \ln 2\pi - \frac{nT}{2} \ln \left[ \frac{1}{nT} \sum_{i=1}^{n} Q_{iT}(\rho) \right] - \frac{n}{2} \ln \left[ \frac{J_T(\rho)}{1+\rho} \right] - \frac{nT}{2}.$$

The FDMLE $\hat{\rho}$ maximizes the profile 'likelihood' criterion function (6), which is defined for $-1 < \rho < 1 + \frac{2}{T-1}$ if $\Delta y_i \neq 0$. The following is true.

**Proposition 1** *Almost surely, (i)* $\ln L^*(\rho) \to -\infty$ *as* $\rho \to -1$ *and* $\rho \to 1 + \frac{2}{T-1}$*, (ii)* $\hat{\rho}$ *exists and* $-1 < \hat{\rho} < 1 + \frac{2}{T-1}$*, (iii)* $(1+\hat{\rho}) J_T(\hat{\rho})^2 \frac{\partial}{\partial \rho} \ln L(\hat{\rho}, \hat{\sigma}^2) = 0$*, and (iv)* $\hat{\sigma}^2 = \frac{1}{nT} \sum_{i=1}^{n} Q_{iT}(\hat{\rho})$.

Proposition 1(ii) confirms that the FDMLE exists in the open interval $(-1, 1 + \frac{2}{T-1})$ almost surely.

### 3.3 Computation of the FDMLE

We can compute the FDMLE while involving minimal use of numerical procedures. This simplification is important because direct numerical maximization of $\ln L(\rho, \sigma^2)$ can be highly inaccurate as the example in Figure 1 of the Introduction demonstrates. The problem occurs in estimating $\rho$. Below we provide a method that entirely avoids numerical optimization with respect to $\rho$, given $\sigma^2$.

Using the notation $T_m = T - m$ and $\tilde{T}_m = T + m$ for any integer $m$, the derivative of the criterion with respect to $\rho$ is

$$\begin{aligned}
\frac{\partial \ln L(\rho, \sigma^2)}{\partial \rho} &= \left[ \frac{nT_1}{2J_T(\rho)} + \frac{n}{2(1+\rho)} \right] + \frac{1}{\sigma^2}(c_1 - c_2\rho) \\
&\quad + \frac{h'_T(\rho)}{2\sigma^2}(d_0 - 2d_1\rho + d_2\rho^2) - \frac{h_T(\rho)}{\sigma^2}(d_1 - d_2\rho) \\
&= \frac{nT}{(1+\rho)J_T(\rho)} + \frac{c_1 - c_2\rho}{\sigma^2} - \frac{d_0 - 2d_1\rho + d_2\rho^2}{\sigma^2 J_T(\rho)^2} - \frac{(1-\rho)(d_1 - d_2\rho)}{\sigma^2 J_T(\rho)},
\end{aligned}$$

where $h'_T(\rho) = -2/J_T(\rho)^2$. By Proposition 1, $\hat{\rho}$ and $\hat{\sigma}^2$ satisfy $\hat{\sigma}^2 = (nT)^{-1} \sum_{i=1}^{n} Q_{iT}(\hat{\rho})$ and

$$\text{(7)} \quad a_0 + a_1\hat{\rho} + a_2\hat{\rho}^2 + a_3\hat{\rho}^3 + a_4\hat{\rho}^4 = 0,$$

where

$$a_0 = nT\tilde{T}_1\hat{\sigma}^2 + \tilde{T}_1^2 c_1 - d_0 - \tilde{T}_1 d_1,$$
$$a_1 = -nTT_1\hat{\sigma}^2 - \tilde{T}_1 T_3 c_1 - \tilde{T}_1^2 c_2 - d_0 + \tilde{T}_1 d_1 + \tilde{T}_1 d_2,$$
$$a_2 = -T_1\tilde{T}_3 c_1 + \tilde{T}_1 T_3 c_2 + \tilde{T}_3 d_1 - T d_2,$$
$$a_3 = T_1^2 c_1 + T_1\tilde{T}_3 c_2 - T_1 d_1 - \tilde{T}_2 d_2,$$
$$a_4 = -T_1^2 c_2 + T_1 d_2.$$

In the above, note that $\hat{\sigma}^2(1+\hat{\rho})J_T(\hat{\rho})^2\frac{\partial}{\partial\theta}\ln L(\hat{\rho},\hat{\sigma}^2) = a_0 + a_1\hat{\rho} + a_2\hat{\rho}^2 + a_3\hat{\rho}^3 + a_4\hat{\rho}^4$ and that $a_0$ and $a_1$ depend on $\hat{\sigma}^2$.

For given $\hat{\sigma}^2$, the quartic equation (7) can be solved directly, for example by Euler's method (see Appendix B for details), and $\hat{\sigma}^2$ is obtained by Proposition 1(iv) for given $\hat{\rho}$. The FDMLE $\hat{\rho}$ and $\hat{\sigma}^2$ can then be obtained by successive iteration. This iteration converges quickly and involves no singularity. In particular, equation (7) removes the singularity that occurs in the criterion $\ln L(\rho,\sigma^2)$ at $\rho = -1$ and $\rho = 1 + \frac{2}{T-1}$ so its solutions can lie outside of the domain $(-1, 1+\frac{2}{T-1})$. Thus, for optimization it is important to check that $\hat{\rho}$ falls in the domain $(-1, 1+\frac{2}{T-1})$. If there are multiple solutions of (7) in the domain $(-1, 1+\frac{2}{T-1})$, then the $\ln L(\rho,\sigma^2)$ values are compared in order to maximize the criterion. By virtue of Proposition 1, there should exist at least one solution of (7) in $(-1, 1+\frac{2}{T-1})$ almost surely.

In sum, $\hat{\rho}$ and $\hat{\sigma}^2$ can be found by the following procedure:

1. Let $\sigma^2 = 1$ (or use $\sigma^2 = (nT)^{-1}\sum_{i=1}^n\sum_{t=1}^T(\Delta y_{it})^2$).
2. For given $\sigma^2$, find the $\rho$ value which maximizes $\ln L(\rho,\sigma^2)$ by solving $\sum_{j=0}^4 a_j\rho^j = 0$.
3. Update $\sigma^2 = (nT)^{-1}\sum_{i=1}^n Q_{iT}(\rho)$.
4. Repeat steps 2–3 until convergence.

We have found this iteration to be a highly effective and efficient procedure for finding the FDMLE $(\hat{\rho},\hat{\sigma}^2)$.

# 4 Unit-Root Asymptotics

We consider the model $y_{it} = \eta_i + v_{it}$, where $v_{it} = \rho_0 v_{it-1} + \varepsilon_{it}$ and $\varepsilon_{it} \sim iid\, N(0,\sigma_0^2)$. As discussed above, the asymptotic theory for the FDMLE is well known for the stationary case $|\rho_0| < 1$ and is equivalent to that of the MLE. 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. The following lemma gives an identity for the profile 'log likelihood' criterion $\ln L^*(\rho)$ that is useful in deriving the limit theory.

**Lemma 2** *Let* $F_{nT}(\delta) = 2[\ln L^*(1+\delta) - \ln L^*(1)]$, $\tilde{\sigma}^2 = (nT)^{-1}\sum_{i=1}^n \sum_{t=1}^T \varepsilon_{it}^2$,

$$V_{0i,T} = \frac{1}{\tilde{\sigma}\sqrt{T}} \sum_{t=1}^T \varepsilon_{it}, V_{1i,T} = \frac{1}{\tilde{\sigma}T^{3/2}} \sum_{t=1}^T z_{it-1}, V_{2i,T} = \frac{1}{\tilde{\sigma}^2 T^2} \sum_{t=1}^T z_{it-1}^2,$$

*and*

$$W_{i,T} = \frac{1}{\tilde{\sigma}^2 T} \sum_{t=1}^T z_{it-1}\varepsilon_{it}.$$

*Then*

$$F_{nT}(\delta) = -nT \ln\left[1 + \frac{1}{nT}\sum_{i=1}^n G_{i,nT}(\delta)\right] + n\ln\left(\frac{2+\delta}{2-T_1\delta}\right), \quad (8)$$

*where*

$$\begin{aligned} G_{i,nT}(\delta) &= \frac{Q_{iT}(1+\delta) - Q_{iT}(1)}{(nT)^{-1}\sum_{i=1}^n Q_{iT}(1)} \\ &= (T\delta)^2 V_{2i,T} - 2(T\delta)W_{i,T} + \frac{T\delta}{2-T_1\delta}\left[V_{0i,T} - (T\delta)V_{1i,T}\right]^2. \end{aligned} \quad (9)$$

The profile criterion $\ln L^*(\rho)$ is maximized at $\hat{\rho}$, and thus $F_{nT}(\delta)$ is maximized at $\hat{\delta} = \hat{\rho} - 1$. Our approach to deriving asymptotics is to reparametrize $\rho$ as $r_{nT}(\rho - 1)$ for some appropriate convergence rate $r_{nT}$. When $T \to \infty$, the $T\delta$ terms in (9) suggest that the convergence rate is $O_p(T)$, and because the random variables are $iid$ across $i$, an extra $O_p(\sqrt{n})$ rate is obtained from cross sectional aggregation. Given $r_{nT}$ and following the usual procedure (e.g., Geyer, 1994, and Knight, 2003) for extremum asymptotic theory, we consider the reparametrized objective function $f_{nT}(\theta) := F_{nT}(r_{nT}^{-1}\theta) := 2[\ln L^*(1 + r_{nT}^{-1}\theta) - \ln L^*(1)]$, which is maximized at $r_{nT}(\hat{\rho} - 1)$. Then the limit distribution of $r_{nT}(\hat{\rho} - 1)$ 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.

In the remainder of the section, we consider separately the two cases where $n$ is fixed and where $n \to \infty$. It is notationally convenient to set $r_{nT} = \sqrt{n}T_1$ when $n$ is fixed, and $r_{nT} = \sqrt{nTT_1}$ when $n \to \infty$.

## 4.1 Time Series Asymptotics

We start by deriving time series 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.

To simplify presentation, let $V_{0i,T}^{*}=(\frac{T}{T_1})^{1/2}V_{0i,T}$, $V_{1i,T}^{*}=(\frac{T}{T_1})^{3/2}V_{1i,T}$, $V_{2i,T}^{*}=(\frac{T}{T_1})^{2}V_{2i,T}$, and $W_{i,T}^{*}=\frac{T}{T_1}W_{i,T}$. Then from (9), $G_{i,nT}(\delta)$ is expressed as

(10) $$G_{i,nT}(\delta)=(T_1\delta)^2V_{2i,T}^{*}-2(T_1\delta)W_{i,T}^{*}+\frac{T_1\delta}{2-T_1\delta}\left[V_{0i,T}^{*}-(T_1\delta)V_{1i,T}^{*}\right]^2.$$

To derive asymptotics, we fix $\theta$ and examine the limit behavior of $f_{nT}(\theta)=F_{nT}(n^{-1/2}T_1^{-1}\theta)=2[\ln L^*(1+n^{-1/2}T_1^{-1}\theta)-\ln L^*(1)]$, which is maximized at $\theta=\sqrt{n}T_1(\hat{\rho}-1)$. We substitute $n^{-1/2}T_1^{-1}\theta$ for $\delta$ in Lemma 2, and get

(11) $$f_{nT}(\theta):=F_{nT}(n^{-1/2}T_1^{-1}\delta)=-nT\ln[1+(nT)^{-1}g_{nT}(\theta)]+n\ln\left(\frac{2+n^{-1/2}T_1^{-1}\theta}{2-n^{-1/2}\theta}\right),$$

where by (10),

(12) $$\begin{aligned}g_{nT}(\theta) &= \sum_{i=1}^{n}G_{i,nT}(n^{-1/2}T_1^{-1}\theta)\\ &= \theta^2\left(\frac{1}{n}\sum_{i=1}^{n}V_{2i,T}^{*}\right)-2\theta\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}W_{i,T}^{*}\right)+\frac{n^{-1/2}\theta}{2-n^{-1/2}\theta}\sum_{i=1}^{n}\left(V_{0i,T}^{*}-\frac{\theta}{\sqrt{n}}V_{1i,T}^{*}\right)^2.\end{aligned}$$

For $n$ fixed, $g_{nT}(\theta)$ is stochastically bounded and we are 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) and are given here for ease of reference.

**Lemma 3** *Let $B_i(r)$ be standard Brownian motions independent across $i$ and $\tilde{B}_i=B_i-\int B_i$. We have*

$$\begin{aligned}V_{0i,T}^{*} &\Rightarrow V_{0i}=B_i(1),\ \ V_{1i,T}^{*}\Rightarrow V_{1i}=\int B_i,\\ V_{2i,T}^{*} &\Rightarrow V_{2i}=\int B_i^2,\ \ W_{i,T}^{*}\Rightarrow W_i=\int B_i dB_i.\end{aligned}$$

Using $\lim nT\ln[1+(nT)^{-1}g_{nT}]=\lim g_{nT}$, we have

(13) $$f_{nT}(\theta)\Rightarrow f_n(\theta):=-g_n(\theta)+n\ln\left(\tfrac{2}{2-n^{-1/2}\theta}\right),$$

where

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

We will consider the case with $n=1$ 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 disappears with cross section averaging as $n$ increases. The limit theory as $n\to\infty$ is treated separately later.

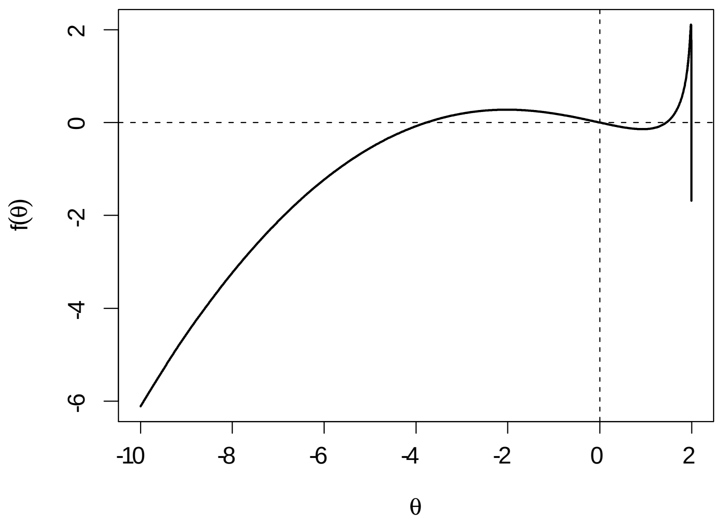


Figure 2: Reparametrized limit 'log-likelihood' criterion $f(\theta)$ exhibiting violent behavior near the upper bound $\theta = 2$.

**The Case** $n = 1$

Let $n = 1$ and omit the $i$ and $n$ subscripts from all notation for the analysis of this case. From (13) and (14) with $n = 1$, we deduce the following limit behavior and form of the limit function.

**Lemma 4** *(i) In every compact subset of* $(-\infty, 2)$, $f_T(\theta) \Rightarrow f(\theta)$ *uniformly in* $\theta$, *where*

(15) $$f(\theta) = V_0^2 + 2\tilde{W}\theta - \tilde{V}_2\theta^2 - \tfrac{2}{2-\theta}(V_0 - V_1\theta)^2 + \ln \tfrac{2}{2-\theta},$$

*uniformly in* $\theta$*; (ii)* $f(\theta) \to -\infty$ *as* $\theta \to -\infty$ *or* $\theta \uparrow 2$ *for almost all sample paths; (iii) Almost surely, the global maximizer* $\tilde{\theta}$ *of* $f(\theta)$ *exists, is in* $(-\infty, 2)$, *and satisfies* $f'(\tilde{\theta}) = 0$.

The implication is the following result for the FDMLE.

**Theorem 5** $T_1(\hat{\rho} - 1) \to_d \arg\max_{\theta<2} f(\theta)$.

Importantly, the peculiarity that is manifest in Figure 1 carries over to the limit criterion function $f(\theta)$, yielding a function with similar potential characteristics as those in Figure 2. Brownian motion trajectories giving rise to a limit function $f(\theta)$ similar to Figure 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.

We next find an expression for the global maximizer $\tilde{\theta}$ of $f(\theta)$, by evaluating the first order condition, which is validated by Lemma 4(iii). We have the following result.

**Proposition 6** *The global maximizer of $f(\theta)$ solves the cubic equation $\sum_{j=0}^{3} b_j\theta^j = 0$, where*

$$\begin{aligned} b_0 &= 4W - V_0^2 + 1, & b_1 &= -4V_2 - 4\tilde{W} - \tfrac{1}{2}, \\ b_2 &= 4\tilde{V}_2 + \tilde{W} + V_1^2, & b_3 &= -\tilde{V}_2. \end{aligned}$$

*In the above $V_2 := \tilde{V}_2 + V_1^2 = \int B^2$, and $W := \tilde{W} + V_0V_1 = \int BdB$.*

The first order condition $\sum_{j=0}^{3} b_j\theta^j = 0$ for the maximization of $f(\theta)$ is the same as the limit of the first order condition for the maximization of $f_T(\theta)$ as follows.

**Proposition 7** *We have*

$$\frac{1}{T_1}\sum_{j=0}^{4} a_j(1+\theta/T_1)^j \Rightarrow \sum_{j=0}^{3} b_j\theta^j$$

*for all $\theta$, where the $a_j$ appear in (7).*

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 formulae ($V_{0,T}$, $V_{1,T}$, $\tilde{V}_{2,T}$ and $\tilde{W}_T$) given in Lemma 2 with $T = 5000$. The empirical distribution functions are plotted in Figure 3 and the estimated densities are shown in Figure 4, 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. Evidence of bimodality and high density around the upper bound (i.e., 2) is visible in the graph of the densities shown in Figure 4. In fact, the asymptotic (centred) density has a mode at a negative value and an asymptote at 2, which is confirmed analytically below.

Let $\tilde{\theta}$ denote the limit distribution of $T_1(\hat{\rho}-1)$. As seen in Figure 3, $\tilde{\theta}$ is not median unbiased. The median of $\tilde{\theta}$ is approximately $-0.5$, and $P\{\tilde{\theta} \leq 0\} \simeq 56.5\%$ 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 Figures 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 Figures 3 and 4, which show the asymptotic distribution based on simulations with $T = 5000$, it appears that the

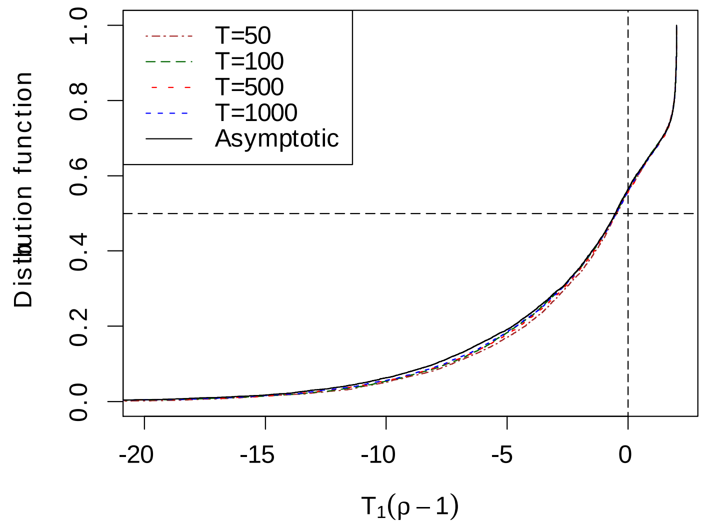


Figure 3: Simulated finite sample and asymptotic CDFs for $T_1\left(\hat{\rho}-1\right)$.

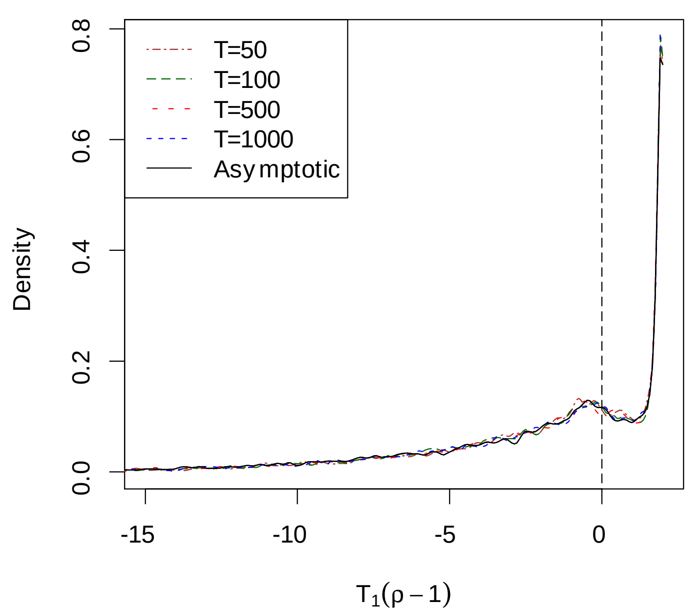


Figure 4: Estimated densities of $T_1(\hat{\rho}-1)$ corresponding to the cdfs shown in Figure 3 (bandwidth = 0.1).

density of the limit distribution of $T_1(\hat{\rho} - 1)$ is infinite. 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 8** *(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 the boundary 2, which is natural because $\tilde{\theta}$ can never attain the boundary. However, the second part of Theorem 8 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 following results for different values of $\epsilon$

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

indicating that $P(\tilde{\theta} > 2 - \epsilon)$ diminishes at a rate no faster than $\sqrt{\epsilon}$, corroborating the finding of Theorem 8. As a result, $P(\tilde{\theta} > 2 - \epsilon)/\epsilon$ diverges, which implies that the density is infinite at the upper bound.

The last two terms of the limit criterion (15), viz.,

$$(16) \quad -\tfrac{2}{2-\theta}(V_0 - V_1\theta)^2 + \ln \tfrac{2}{2-\theta}$$

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 4(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 8 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 (16) is $J_T(\rho) = T_1\left(1 + \frac{2}{T_1} - \rho\right) = 2 - \theta$. 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 (16) 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.

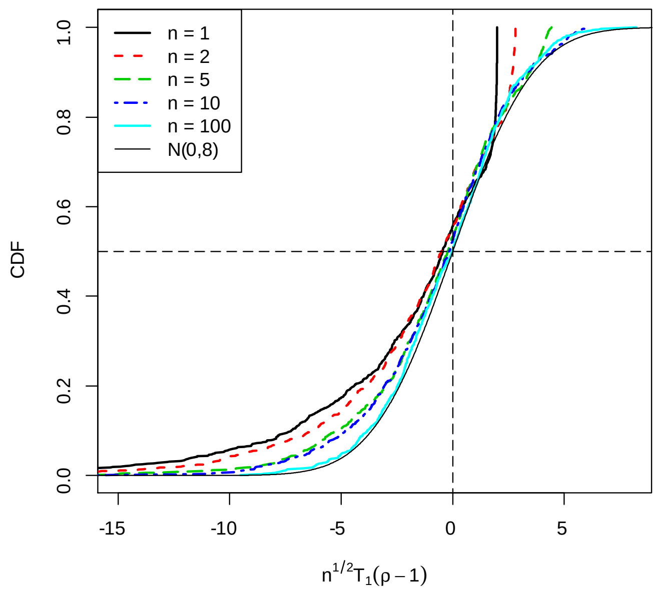


Figure 5: Empirical distribution functions for $\sqrt{n}T_1(\hat{\rho}-1)$ for $n = 1, 2, 5, 10, 50, 100$ and $T = 500$.

**The Case** $n > 1$

We next examine the case where $n > 1$ but fixed and $T \to \infty$. From (13) and (14), we have $f_{nT}(\theta) \Rightarrow f_n(\theta)$, where

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

We now have the restriction $\sqrt{n}T_1(\hat{\rho}-1) < 2\sqrt{n}$ 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 Figure 5, obtained from 5,000 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, $N(0, 8)$.

Theorem 8 established that the probability $P(\tilde{\theta} > 2 - \epsilon)$ is $O(\epsilon^{1/2})$ as $\epsilon \to 0$, where $\tilde{\theta}$ is 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 9** *Let $\xi_{2i} = \tilde{W}_i + V_{1i}^2$. Then*

$$P(\tilde{\theta}_n > 2\sqrt{n} - \epsilon) \leq 2\left(\frac{3e\epsilon}{\sqrt{n}}\right)^{n/2} + \left(\frac{4\epsilon^2}{n^2}\right) E(\xi_{2i}^2).$$

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 $n \geq 3$. Note that Theorem 9 serves 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.

## 4.2 Large-$n$ Asymptotics

In this subsection, we let $n \to \infty$. Kruiniger (2008) has already established asymptotics for this case, but we reproduce the result here with a new proof in order to to examine how the peculiarity described in the Introduction and analyzed above for finite $n$ is removed as $n \to \infty$.

Let $\rho_0 = 1$ and let $n \to \infty$. It is a little more convenient in this case to work with $\sqrt{n}T(\hat{\rho} - 1)$ instead of $\sqrt{n}T_1(\hat{\rho} - 1)$. For a given $\theta$, we thus let $\rho_{nT} = 1 + \theta/(\sqrt{n}T)$ instead of $\rho_{nT} = 1 + \theta/(\sqrt{n}T_1)$. As before, define $f_{nT}(\theta) := 2[\ln L^*(\rho_{nT}) - \ln L^*(1)] = F_{nT}(\theta/(\sqrt{n}T))$, which is minimized by $\sqrt{n}T(\hat{\rho} - 1)$. Here $F_{nT}(\cdot)$ is defined as in Lemma 2. We have

$$\text{(17)}\quad \begin{aligned} f_{nT}(\theta) &= -nT\ln[1 + (nT)^{-1}g_{nT}(\theta)] + n\ln\left[\frac{2 + n^{-1/2}T^{-1}\theta}{J_T(\rho_{nT})}\right] \\ &= -nT\ln[1 + (nT)^{-1}g_{nT}(\theta)] + n\ln\left[1 + \frac{n^{-1/2}\theta}{J_T(\rho_{nT})}\right], \end{aligned}$$

where $J_T(\rho_{nT}) = 2 - T_1(\rho_{nT} - 1) = 2 - n^{-1/2}(T_1/T)\theta$. From (9),

$$g_{nT}(\theta) = \frac{1}{n}\sum_{i=1}^n V_{2i,T}\theta^2 - \frac{2}{\sqrt{n}}\sum_{i=1}^n W_{i,T}\theta + \frac{n^{-1/2}\theta}{J_T(\rho_{nT})}\sum_{i=1}^n \left(V_{0i,T} - n^{-1/2}V_{1i,T}\theta\right)^2,$$

and thus

$$\text{(18)}\quad T^{-1}g_{nT}(\theta) = T^{-1}V_{2,nT}\theta^2 - 2T^{-1}W_{nT}\theta + \frac{\theta}{T\,J_T(\rho_{nT})}(\sqrt{n}A_{nT}^{00} - 2A_{nT}^{01}\theta + n^{-1/2}A_{nT}^{11}\theta^2),$$

where $V_{2,nT} = n^{-1}\sum_{i=1}^n V_{2i,T}$, $W_{nT} = n^{-1/2}\sum_{i=1}^n W_{i,T}$, and $A_{nT}^{jk} = n^{-1}\sum_{i=1}^n V_{ji,T}V_{ki,T}$.

Notice that $(nT)^{-1}g_{nT}(\theta)$ is $O_p(n^{-1/2}T^{-1})$, while the component $n^{-1/2}\theta/J_T(\rho_{nT})$ in the logarithm in the second term on the second line of (17) is $O(n^{-1/2})$. Thus, to approximate

$$n\ln\left[1 + \frac{n^{-1/2}\theta}{J_T(\rho_{nT})}\right]$$

in $f_{nT}(\theta)$, we need a second order Taylor development as in the following lemma.

**Lemma 10** *Let $a_n$ and $b_n$ be bounded. Then*

$$n\ln(1+n^{-1/2}a_n+n^{-1}b_n) = n^{1/2}a_n - a_n^2/2 + b_n + O(n^{-1/2}).$$

Using Lemma 10, from (18) we get

$$\begin{aligned} Tn\ln[1+n^{-1}T^{-1}g_{nT}(\theta)] &= -2W_{nT}\theta + V_{2,nT}\theta^2 + \frac{\sqrt{n}A_{nT}^{00}\theta}{J_T(\rho_{nT})} \\ &\quad -\frac{(A_{nT}^{00})^2\theta^2}{2TJ_T(\rho_{nT})^2} - \frac{2A_{nT}^{01}\theta^2}{J_T(\rho_{nT})} + O_p(n^{-1/2}) \end{aligned}$$

for the first term of (17), and for the second term of (17), we have

$$n\ln\left[1+\frac{n^{-1/2}\theta}{J_T(\rho_{nT})}\right] = \frac{\sqrt{n}\theta}{J_T(\rho_{nT})} - \frac{\theta^2}{2J_T(\rho_{nT})^2} + O(n^{-1/2}).$$

Thus,

$$\begin{aligned} f_{nT}(\theta) &= 2W_{nT}\theta - V_{2,nT}\theta^2 + \frac{(A_{nT}^{00})^2\theta^2}{2TJ_T(\rho_{nT})^2} + \frac{2A_{nT}^{01}\theta^2}{J_T(\rho_{nT})} - \frac{\theta^2}{2J_T(\rho_{nT})^2} \\ &\quad -\frac{\sqrt{n}(A_{nT}^{00}-1)\theta}{J_T(\rho_{nT})} + O_p(n^{-1/2}) \\ &= \left[2W_{nT} - \frac{\sqrt{n}(A_{nT}^{00}-1)}{J_T(\rho_{nT})}\right]\theta - \left[V_{2,nT} - \frac{2A_{nT}^{01}}{J_T(\rho_{nT})} + \frac{T-(A_{nT}^{00})^2}{2TJ_T(\rho_{nT})^2}\right]\theta^2 + o_p(1). \end{aligned}$$

As $n\to\infty$, we have $J_T(\rho_{nT})\to 2$, $V_{2,nT}\to_p \frac{1}{2}$, $A_{nT}^{01}\to_p \frac{1}{2}$, $A_{nT}^{00}\to_p 1$, and hence,

$$f_{nT}(\theta) = Z_{nT}(\theta)\theta - \frac{T_1\theta^2}{8T} + o_p(1), \quad \text{with } Z_{nT}(\theta) = 2W_{nT} - \frac{\sqrt{n}(A_{nT}^{00}-1)}{J_T(\rho_{nT})}.$$

But

$$\begin{aligned} Z_{nT}(\theta) &= J_T(\rho_{nT})^{-1}[2W_{nT}J_T(\rho_{nT}) - \sqrt{n}(A_{nT}^{00}-1)] \\ &= \frac{1}{J_T(\rho_{nT})}\cdot\frac{1}{\sqrt{n}}\sum_{i=1}^n(4W_{i,T} - V_{0i,T}^2+1) - \frac{2(T_1/T)W_{nT}\theta}{\sqrt{n}J_T(\rho_{nT})} \\ &= \frac{1}{J_T(\rho_{nT})\tilde{\sigma}^2}\cdot\frac{1}{\sqrt{n}}\sum_{i=1}^n(4W_{i,T}\tilde{\sigma}^2 - V_{0i,T}^2\tilde{\sigma}^2+\tilde{\sigma}^2) + O_p(n^{-1/2}) \\ &= \frac{1}{J_T(\rho_{nT})\tilde{\sigma}^2}\cdot\frac{1}{\sqrt{n}}\sum_{i=1}^n\left[\frac{4}{T}\sum_{t=1}^T z_{it-1}\varepsilon_{it} - \left(\frac{1}{\sqrt{T}}\sum_{t=1}^T\varepsilon_{it}\right)^2 + \frac{1}{T}\sum_{t=1}^T\varepsilon_{it}^2\right] + o_p(1) \\ &= \frac{1}{J_T(\rho_{nT})}\left(\frac{\sigma^2}{\tilde{\sigma}^2}\right)\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\frac{2}{T}\sum_{t=1}^T\frac{z_{it-1}\varepsilon_{it}}{\sigma^2}\right) + o_p(1), \end{aligned}$$

where we used $z_{i0}\equiv 0$ for the last identity. We thus have

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

where the $o_p(1)$ term converges to zero pointwise in $\theta$ and uniformly on every compact set as well. Finally, let $\theta_* = \sqrt{T_1/T}\theta$ so $\theta = \sqrt{T/T_1}\theta_*$. Then we have

$$f_{nT}(\sqrt{T/T_1}\theta_*) = \left(\frac{T}{T_1}\right)^{1/2} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{1}{T}\sum_{t=1}^{T} \frac{z_{it-1}\varepsilon_{it}}{\sigma^2}\right)\theta_* - \frac{\theta_*^2}{8} + o_p(1)$$
$$\Rightarrow \quad f_*(\theta_*) \equiv W\theta_* - \frac{\theta_*^2}{8}, \quad W \sim N(0, \frac{1}{2}),$$

where the limit (the second line) is maximized at $4W \sim N(0,8)$. Noting that the weak convergence of $f_{nT}(\sqrt{T/T_1}\theta_*)$ to $f_*(\theta_*)$ is not only pointwise but also uniform over $\theta_*$ in every compact set, and noting that $f_{nT}(\sqrt{T/T_1}\theta_*)$ is maximized at $\sqrt{nTT_1}(\hat{\rho}-1)$, we have the final result that

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

as $n \to \infty$ by virtue of the argmax continuous mapping theorem regardless of the size of $T$. This limit theory matches the result in Kruiniger (2008), and the expression in (19) has the $\sqrt{n}T$ convergence rate as $n, T \to \infty$.

Asymptotic normality results from the fact that $J_T(\rho_{nT})$ 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.

Simulated cumulative distribution functions from 5,000 replications for $n = 500$ and $T = 3, 5, 10, 100$ are drawn in Figure 6 and confirm the accuracy of this large $n$ limit theory even for small $T \geq 3$.

# 5 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 approach particularly appealing. There also appear to be unexpected efficiency gains in the use of FDMLE over conventional and bias corrected MLE, even in the limit theory as $n \to \infty$ (Kruiniger, 2008).

As shown here, these advantages come from the fact that FDMLE is a restricted extremum estimator for which the 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

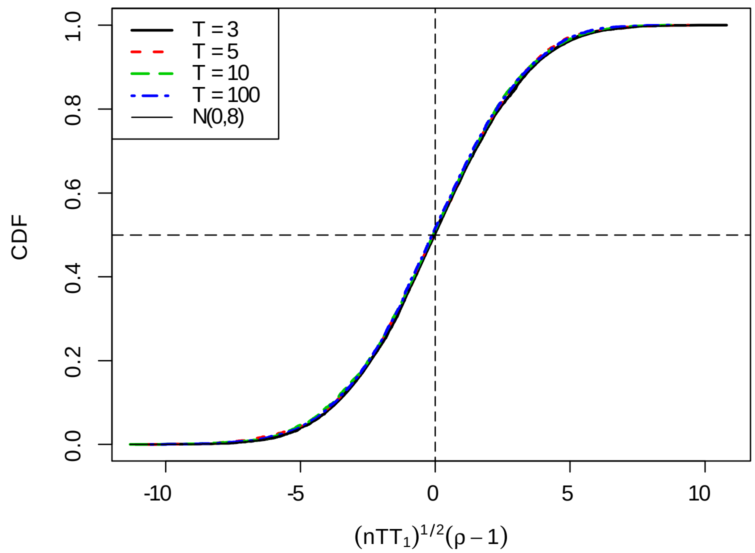


Figure 6: Empirical distribution functions for $\sqrt{nTT_1}(\hat{\rho}-1)$ for $n=500$ and $T=3,5,10,100$.

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 provided $n$ increases. 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 restrictions that are inherent in FDML estimation persist in the limit. They manifest in the efficiency gain of the FDMLE in the unit root case (where the restrictions in finite samples are most binding) and in the fact that the limit distribution is normal when $n\to\infty$. For all practical purposes, at least when $n$ is large, 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.

## A Proofs

**Proof of (3).** *Method 1 (Induction)*: We omit the $\rho$ argument in what follows. For $T=1$, we have $C_T = 2/(1+\rho)$, $H_T = 1$, and $D_T = 1$, thus $\tilde{C}_T = \frac{2}{1+\rho} = 1 + \frac{1-\rho}{1+\rho}$ as (3) suggests. Next, suppose that (3) is true for a given $T$. Because

$$C_{T+1} = \begin{pmatrix} C_T & -a\xi_T \\ -a\xi_T' & \frac{2}{1+\rho} \end{pmatrix}, \quad H_{T+1} = \begin{pmatrix} H_T & 0 \\ 1_T' & 1 \end{pmatrix}, \quad D_{T+1} = \begin{pmatrix} D_T & 0 \\ -\rho e_T' & 1 \end{pmatrix},$$

where $a = \frac{1-\rho}{1+\rho}$, $\xi_T = (1, \rho, \ldots, \rho^{T-1})'$, $1_T = (1, \ldots, 1)'$ and $e_T = (0, \ldots, 0, 1)'$, we have (using $e_T' H_T = 1_T'$)

$$D_{T+1} H_{T+1} C_{T+1} H_{T+1}' D_{T+1}' = \begin{pmatrix} \eta_{11} & \eta_{12} \\ \eta_{12}' & \eta_{22} \end{pmatrix},$$

where

$$\begin{aligned} \eta_{11} &= D_T H_T C_T H_T' D_T' = \tilde{C}_T = I_T + a 1_T 1_T', \\ \eta_{12} &= (1-\rho) D_T H_T C_T 1_T - a D_T H_T \xi_T, \\ \eta_{22} &= \tfrac{2}{1+\rho} + (1-\rho)^2 1_T' C_T 1_T - 2a(1-\rho) 1_T' \xi_T. \end{aligned}$$

Because $-a\xi_T = \rho C_T e_T - e_T$, $1_T = H_T' e_T$, $H_T e_T = e_T$ and $D_T e_T = e_T$, we have

$$\begin{aligned} \eta_{12} &= (1-\rho) D_T H_T C_T H_T' e + \rho D_T H_T C_T e_T - D_T H_T e_T] \\ &= D_T H_T C_T H_T' [(1-\rho) I_T + \rho H_T^{-1\prime}] e_T - e_T. \end{aligned}$$

But $H_T^{-1} = D_T(1)$ so $(1-\rho) I_T + \rho H_T^{-1} = I_T - \rho(I_T - H_T^{-1}) = D_T$, which implies that

$$\eta_{12} = D_T H_T C_T H_T' D_T' e_T - e_T = \tilde{C}_T e_T - e_T = (I_T + a 1_T 1_T') e_T - e_T = a 1_T.$$

For $\eta_{22}$, we have

$$\begin{aligned} \eta_{22} - \tfrac{2}{1+\rho} &= (1-\rho)^2 1_T' C_T 1_T + 2(1-\rho) 1_T' (\rho C_T e_T - e_T) \\ &= (1-\rho)[(1-\rho) 1_T' C_T 1_T + 2\rho 1_T' C_T e_T - 2], \end{aligned}$$

where we use $1_T' e_T = 1$ for the last term. But

$$\begin{aligned} 1_T' C_T 1_T &= \frac{1}{1+\rho} \left[ 2T - 2(1-\rho) \sum_{t=1}^{T-1} (T-t) \rho^{t-1} \right] = \frac{2}{1+\rho} \sum_{t=1}^{T} \rho^{t-1}, \\ 1_T' C_T e_T &= \frac{1}{1+\rho} \left[ 2 - (1-\rho) \sum_{t=1}^{T-1} \rho^{t-1} \right], \end{aligned}$$

and thus

$$\begin{aligned}\eta_{22} - \frac{2}{1+\rho} &= \frac{2(1-\rho)}{1+\rho}\sum_{t=1}^{T}\rho^{t-1} + \frac{4\rho}{1+\rho} - \frac{2\rho(1-\rho)}{1+\rho}\sum_{t=1}^{T-1}\rho^{t-1} - 2 \\ &= \frac{2(1-\rho)}{1+\rho} + \frac{4\rho}{1+\rho} - 2 = 0.\end{aligned}$$

We have shown that $\eta_{11} = I_T + a1_T1_T'$, $\eta_{12} = a1_T$ and $\eta_{22} = \frac{2}{1+\rho} = 1 + a$. Therefore, $D_{T+1}H_{T+1}C_{T+1}H_{T+1}'D_{T+1}' = I_{T+1} + a1_{T+1}1_{T+1}' = \tilde{C}_{T+1}$.

*Method 2 (Statistical)*: Let $\rho$ be given. Suppose that $\tilde{\varepsilon}_t \sim iid(0,1)$. Let $\tilde{v}_t = \rho\tilde{v}_{t-1} + \tilde{\varepsilon}_t$ and $\Delta\tilde{v}_t$ is covariance stationary. Let $\tilde{z}_t = \tilde{v}_t - \tilde{v}_0$, and let $\tilde{u}_t = \tilde{z}_t - \rho\tilde{z}_{t-1} = \tilde{\varepsilon}_t - (1-\rho)\tilde{v}_0$. Then $\tilde{C}_T$ is the covariance matrix of $(\tilde{u}_1, \ldots, \tilde{u}_T)'$. But $E(\tilde{u}_t\tilde{u}_s) = \mathbf{1}\{t=s\} + (1-\rho)^2/(1-\rho^2) = \mathbf{1}\{t=s\} + \frac{1-\rho}{1+\rho}$. ■

**Proof of Proposition 1.** (i) For a given sample, $\lim_{\rho\to-1}(nT)^{-1}\sum_{i=1}^n Q_{iT}(\rho)$ is strictly positive almost surely, $\lim_{\rho\to-1}\ln J_T(\rho) = \ln 2$, and $\lim_{\rho\to-1}\ln(1+\rho) = -\infty$, so $\lim_{\rho\to-1}\ln L^*(\rho) = -\infty$ almost surely. We also have

$$\text{(20)}\quad \ln L^*(\rho) = -\frac{nT}{2}\ln 2\pi - \frac{nT}{2}\ln\left(\frac{1}{nT}\sum_{i=1}^n Q_{iT}(\rho)J_T(\rho)\right) + \frac{nT_1}{2}\ln J_T(\rho) + \frac{n}{2}\ln(1+\rho) - \frac{nT}{2},$$

(which is obtained by subtracting and adding $-\frac{nT}{2}\ln J_T(\rho)$), and

$$\text{(21)}\quad \sum_{i=1}^n Q_{iT}(\rho)J_T(\rho) = J_T(\rho)\sum_{i=1}^n u_{it}(\rho)^2 - (1-\rho)\left[\sum_{i=1}^n u_{it}(\rho)\right]^2.$$

For all $\rho \in (-1, \frac{T+1}{T-1})$, $J_T(\rho) > 0$, and as $\rho \to \frac{T+1}{T-1}$, we have $-(1-\rho) \to \frac{2}{T_1}$, so the limit (as $\rho \to \frac{T+1}{T-1}$) of (21) is bigger than $\frac{1}{T_1}\left[\sum_{i=1}^n u_{it}(\rho)\right]^2$, which is almost surely strictly positive. Furthermore, $J_T(\rho) \to 0$ as $\rho \to \frac{T+1}{T-1}$, thus the third term in the right hand side of (20) diverges to $-\infty$, hence $\ln L^*(\rho) \to -\infty$ almost surely as $\rho \to \frac{T+1}{T-1}$.

(ii): This result holds because of (i), the continuity of $\ln L^*(\rho)$, and the fact that $\ln L^*(1)$ is finite.

(iii): True because $\ln L^*(\rho)$ is differentiable, $1 + \hat{\rho} \neq 0$ and $J_T(\hat{\rho}) \neq 0$ by (i).

(iv): This is the unique solution of the first order condition. ■

**Proof of Lemma 2.** From (6) and the definition of $F_{nT}(\delta) \equiv 2[\ln L^*(1+\delta) - \ln L^*(1)]$, we find

$$\text{(22)}\quad F_{nT}(\delta) = -nT\ln[1 + (nT)^{-1}G_{nT}(\delta)] + \ln\left(\frac{2+\delta}{2 - T_1\delta}\right),$$

where

$$G_{nT}(\delta) = \frac{\sum_{i=1}^n [Q_{iT}(1+\delta) - Q_{iT}(1)]}{(nT)^{-1}\sum_{i=1}^n Q_{iT}(1)}.$$

and where we use $J_T(1+\delta) = 2 - T_1\delta$ and $1+\rho = 2+\delta$ to give (22).

When $\rho_0 = 1$, we have $u_{it}(\rho) = \varepsilon_{it} - (1-\rho)z_{it-1}$, i.e., $u_{it}(1+\delta) = \varepsilon_{it} - \delta z_{it-1}$, hence

$$Q_{iT}(1+\delta) = \sum_{t=1}^{T}(\varepsilon_{it}^2 - 2\delta z_{it-1}\varepsilon_{it} + \delta^2 z_{it-1}^2) - \frac{\delta}{2-T_1\delta}\left[\sum_{t=1}^{T}(\varepsilon_{it} - \delta z_{it-1})\right]^2,$$

and $Q_{iT}(1) = \sum_{t=1}^{T}\varepsilon_{it}^2$. Thus, the denominator of $G_{nT}(\delta)$ is $\tilde{\sigma}^2$, and from (5), we have

$$Q_{iT}(1+\delta) - Q_{iT}(1) = -2\delta\sum_{t=1}^{T} z_{it-1}\varepsilon_{it} + \delta^2\sum_{t=1}^{T} z_{it-1}^2 - \frac{\delta}{2-T_1\delta}\left[\sum_{t=1}^{T}(\varepsilon_{it} - \delta z_{it-1})\right]^2.$$

The result follows straightforwardly. ■

We next prove Lemma 4. Recall that $f_T(\theta) = -T\ln(1+T_1^{-1}g_T(\theta)) - \ln(2-\theta) + \ln(2+T_1^{-1}\theta)$ and $g(\theta) = \tilde{V}_2\theta^2 - 2\tilde{W}\theta + \frac{2}{2-\theta}(V_0 - V_1\theta)^2 - V_0^2$.

**Proof of Lemma 4.** (i) Fix a compact subset $K$ of $(-\infty, 2)$. For given $T$, $f_T(\theta)$ is defined on $(-2T_1, 2)$, so $K \subset (-2T_1, 2)$ for all large enough $T$. Thus, $f_T(\theta)$ converges weakly to $f(\theta) = -g(\theta) - \ln(2-\theta) + \ln 2$ for every $\theta \in K$ because $g_T(\theta) \Rightarrow g(\theta)$, $T\ln(1+T_1^{-1}c) \to c$ pointwise, and $\ln(2+T_1^{-1}\theta) \to \ln 2$ pointwise. The weak convergence is also uniform over all $\theta \in K$ because in $K$, $f_T(\theta)$ is uniformly continuous and finite almost surely.

(ii) Almost surely $\tilde{V}_2$ and $V_1^2$ are strictly positive, so $g(\theta) \to \infty$ almost surely as $\theta \to -\infty$. Also $\ln\frac{2}{2-\theta} \to -\infty$ as $\theta \to -\infty$. Thus, $f(\theta) = -g(\theta) + \ln\frac{2}{2-\theta} \to -\infty$ almost surely as $\theta \to -\infty$. Next, for the case $\theta \uparrow 2$, we have

$$f(\theta) = \left[V_0^2 + 2\tilde{W}\theta - \tilde{V}_2\theta^2\right] + \left[\ln\tfrac{2}{2-\theta} - \tfrac{2}{2-\theta}(V_0 - V_1\theta)^2\right] = f_a(\theta) + f_b(\theta).$$

As $\theta \uparrow 2$, $f_a(\theta)$ converges to a tight random variable, and with probability 1, $\lim_{\theta\uparrow 2}(V_0 - V_1\theta)^2 > 0$, implying that $\lim_{\theta\uparrow 2} f_b(\theta) = -\infty$ almost surely as claimed.

(iii) The global maximizer $\tilde{\theta}$ is in $(-\infty, 2)$ by (ii) and the continuity of $f(\theta)$. Also the differentiability of $f(\theta)$ implies that $f'(\tilde{\theta}) = 0$. ■

**Proof of Theorem 5.** By Lemma 4(i), $f_T(\theta) \Rightarrow f(\theta)$ uniformly in every compact subset of $(-\infty, 2)$. The limit process $f(\theta)$ has continous sample paths, and by Lemma 4(ii), the global maximizer of $f(\theta)$ exists. The probability of $f(\theta)$ having multiple maxima is zero, and the result follows from a standard argmax theorem (e.g, Corollary 5.58 of van der Vaart, 1998). ■

**Proof of Proposition 6.** Let $\tilde{\theta}$ be the global maximizer of $f(\theta)$. By Lemma 4(iii), $\tilde{\theta}$ satisfies $\frac{1}{2}(2-\theta)^2 f'(\theta) = 0$. We have

$$f'(\theta) = -2\tilde{V}_2\theta + 2\tilde{W} - \frac{2(V_0 - V_1\theta)^2}{(2-\theta)^2} + \frac{4V_1(V_0 - V_1\theta)}{2-\theta} + \frac{1}{2-\theta},$$

so

$$\tfrac{1}{2}(2-\theta)^2 f'(\theta) = -\tilde{V}_2\theta(2-\theta)^2 + \tilde{W}(2-\theta)^2 - (V_0 - V_1\theta)^2 + 2V_1(V_0 - V_1\theta)(2-\theta) + \tfrac{1}{2}(2-\theta).$$

The right hand side equals $\sum_{j=0}^{3} b_j\theta^j$. ■

**Proof of Proposition 7.** Let $\phi = \rho - 1$, so $\theta = T_1\phi$. Then

$$\begin{aligned}\frac{1}{T_1}\sum_{j=0}^{4} a_j\rho^j &= \frac{1}{T_1}\sum_{j=0}^{4} a_j(1+\phi)^j = \frac{1}{T_1}\sum_{j=0}^{4} a_j\sum_{k=0}^{j}\binom{j}{k}\phi^j = \frac{1}{T_1}\sum_{k=0}^{4}\left[\sum_{j=k}^{4}\binom{j}{k}a_j\right]\phi^k \\ &= \sum_{k=0}^{4}\left[\frac{1}{T_1^{-k-1}}\sum_{j=k}^{4}\binom{j}{k}a_j\right]\theta^k \equiv \sum_{k=0}^{4} b_{k,T}\theta^k.\end{aligned}$$

The result is obtained by rather tedious algebra which shows that $b_{k,T} \Rightarrow b_k$ for each $k$ (and with $b_4 = 0$). We omit the details. ■

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

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

where

$$\begin{aligned}f_{an}(\theta) &= \sum_{i=1}^{n}\xi_{1i}^2 + \frac{2}{\sqrt{n}}\sum_{i=1}^{n}\xi_{2i}\theta - \frac{1}{n}\sum_{i=1}^{n}\tilde{V}_{2i}\theta^2, \\ f_{bn}(\theta) &= n\ln\frac{2}{2 - n^{-1/2}\theta} - \frac{2}{2 - n^{-1/2}\theta}\sum_{i=1}^{n}\xi_{1i}^2.\end{aligned}$$

The first derivatives are

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

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

**Proof of Theorem 8.** Recall that $n = 1$ and we omit the $i$ and $n$ subscripts. Let $\xi_1 = V_0 - 2V_1$ and $\xi_2 = \tilde{W} + V_1^2$. From (23), 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 $\tilde{\theta}$. Let $\theta_0$ be the global maximizer of $f(\theta)$.

(i) This a special case of Theorem 9 with $n = 1$.

(ii) We note that $\tilde{\theta} \geq \theta_0$ if $\sup_{\theta_0 \leq \theta < 2} f(\theta) > \sup_{\theta < \theta_0} f(\theta)$, which in turn holds if

$$\sup_{\theta_0 \leq \theta < 2} f(\theta) > \sup_{\theta < \theta_0} f_a(\theta) + \sup_{\theta < \theta_0} f_b(\theta)$$

because $\sup(f_a + f_b) \leq \sup f_a + \sup f_b$. Applying the relationship $\sup(f_a + f_b) \geq \inf f_a + \sup f_b$ to the left hand side, we have the above displayed inequality implied by

$$\inf_{\theta_0 \leq \theta < 2} f_a(\theta) + \sup_{\theta_0 \leq \theta < 2} f_b(\theta) > \sup_{\theta < \theta_0} f_a(\theta) + \sup_{\theta < \theta_0} f_b(\theta),$$

i.e., by

$$\text{(25)} \quad \sup_{\theta \geq \theta_0} f_b(\theta) - \sup_{\theta < \theta_0} f_b(\theta) > \sup_{\theta < \theta_0} f_a(\theta) - \inf_{\theta_0 \leq \theta < 2} f_a(\theta) \equiv \eta.$$

This last event occurs only if $f_b(\theta)$ is maximized in $(\theta_0, 2)$ because otherwise the left hand side of (25) is nonpositive and the right hand side is nonnegative. We thus consider only the case where $\sup_{\theta < \theta_0} f_b(\theta) = f_b(\theta_0)$, which happens if $f_b'(\theta_0) > 0$, i.e., $\frac{2-\theta_0}{2} > \xi_1^2$. In that case we have

$$\sup_{\theta < \theta_0} f_b(\theta) = \ln \tfrac{2}{2-\theta_0} - \tfrac{2}{2-\theta_0}\xi_1^2,$$

and the global maximum of $f_b(\theta)$ is attained in $[\theta_0, 2)$, where $\sup_{\theta \geq \theta_0} f_b(\theta) = \sup_\theta f_b(\theta) = \ln \xi_1^{-2} - 1$. Thus (25) occurs if

$$\tfrac{2}{2-\theta_0} < \xi_1^{-2} \ \text{ and } \ \ln \xi_1^{-2} - 1 - \ln \tfrac{2}{2-\theta_0} + \tfrac{2}{2-\theta_0}\xi_1^2 > \eta,$$

which in turn is implied by the fact that

$$\tfrac{2}{2-\theta_0} < \xi_1^{-2} \ \text{ and } \ \ln \xi_1^{-2} - \eta > \ln \tfrac{2}{2-\theta_0} + 1.$$

Therefore,

$$\text{(26)} \quad P(\tilde{\theta} \geq \theta_0) \geq P(\ln \xi_1^{-2} - \eta > \ln \tfrac{2}{2-\theta_0} + 1, \xi_1^{-2} > \tfrac{2}{2-\theta_0}).$$

Because $\eta \geq 0$, $\ln \xi_1^{-2} - \eta > \ln \frac{2}{2-\theta_0} + 1$ only if $\xi_1^{-2} > \frac{2}{2-\theta_0}$, so the right hand side of (26) is greater than or equal to $P(\ln \xi_1^{-2} - \eta > \ln \frac{2}{2-\theta_0} + 1) = P(e^\eta \xi_1^2 < \epsilon_0)$, where $\epsilon_0 = \frac{2-\theta_0}{2e}$. We have so far established that

$$P(\tilde{\theta} \geq \theta_0) \geq P(e^\eta \xi_1^2 < \epsilon_0), \ \ \epsilon_0 = \tfrac{2-\theta_0}{2e}.$$

The $f_1(\theta)$ function is globally maximized at $\tilde{V}_2^{-1}\xi_2$, so if $\tilde{V}_2^{-1}\xi_2 \geq 2$ (i.e., if $2\tilde{V}_2 \leq \xi_2$), then $\sup_{\theta < \theta_0} f_a(\theta) = \inf_{\theta_0 \leq \theta < 2} f_a(\theta)$, i.e., $\eta = 0$. With this in mind, we note that

$$P(e^\eta \xi_1^2 < \epsilon_0) \geq P(e^\eta \xi_1^2 < \epsilon_0, 2\tilde{V}_2 \leq \xi_2) = P(\xi_1^2 < \epsilon_0, 2\tilde{V}_2 \leq \xi_2).$$

We have $2\tilde{V}_2 \leq \xi_2$ if and only if $\xi_1^2 \geq 4\tilde{V}_2 - 2V_0V_1 + 2V_1^2 + 1$ (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

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

which is greater than or equal to

$$P(-\sqrt{\epsilon_0} < \xi_1 < \sqrt{\epsilon_0}, 4\tilde{V}_2 - 2V_0V_1 + 2V_1^2 + 1 \leq 0).$$

Let $\epsilon_0 \leq 1$. In the event that $\xi_1 > -\sqrt{\epsilon_0}$, i.e., when $V_0 > 2V_1 - \sqrt{\epsilon_0}$, we have $-2V_0V_1 < -4V_1^2 + 2V_1\sqrt{\epsilon_0} \leq -4V_1^2 + 2V_1$, so the above displayed probability is at least as large as

$$P(-\sqrt{\epsilon_0} < V_0 - 2V_1 < \sqrt{\epsilon_0}, 4\tilde{V}_2 - 2V_1^2 + 2V_1 + 1 \leq 0),$$

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 $P(4\tilde{V}_2 - 2V_1^2 + 2V_1 + 1 \leq 0) > 0$, so the last probability is of order $\sqrt{\epsilon_0}$. ■

When we generalize the previous result to $n > 1$, the following lemma is useful.

**Lemma 11** *Let* $\Gamma(s) = \int_0^\infty x^{s-1}e^{-x}dx$. *Then*

$$\text{(27)} \quad \frac{(n/2)^{n/2}}{(n/2)\Gamma(n/2)e^{n/2}} \leq 1, \quad n = 1, 2, \ldots.$$

**Proof.** Let

$$q(m) = \frac{m^m}{m\Gamma(m)e^m}, \quad m = \tfrac{1}{2}, 1, \tfrac{3}{2}, \ldots.$$

so the left hand side of (27) is $q(\frac{n}{2})$. For all positive $m$, we have

$$\begin{aligned} \frac{q(m+1)}{q(m)} &= \frac{(m+1)^{m+1}}{m^m} \cdot \frac{m\Gamma(m)}{(m+1)\Gamma(m+1)} \cdot \frac{e^m}{e^{m+1}} = \frac{(m+1)^{m+1}}{e(m+1)m^m} \\ &= \frac{1}{e}\left(\frac{m+1}{m}\right)^m = \frac{1}{e}\left(1 + \frac{1}{m}\right)^m \leq 1, \end{aligned}$$

where we used the fact that $\Gamma(m+1) = m\Gamma(m)$. Substituting $\frac{n}{2}$ for $m$, we get $q(\frac{n}{2}+1) \leq q(\frac{n}{2})$ for all positive integer $n$. For the initial values $n = 1, 2$, we have

$$q(\tfrac{1}{2}) = \frac{(1/2)^{1/2}}{(1/2)\Gamma(1/2)e^{1/2}} = \sqrt{\frac{2}{\pi e}} \leq 1, \quad q(1) = \frac{1}{e} \leq 1.$$

Thus (27) holds for all $n$. ■

We also have the following uniform probability bound for chi-square distributions.

**Lemma 12** *Let* $X_n \sim \chi_n^2$. *We have* $P(X_n \leq x) \leq (ex/n)^{n/2}$ *for all* $n$.

**Proof.** Let $s = n/2$. We have

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

by Lemma 11. ■

**Proof of Theorem 9.** Let $\epsilon \le 2/(3e)$ be given. Let $\theta_0 = 2\sqrt{n} - \epsilon$ so $\epsilon = 2\sqrt{n} - \theta_0 = \sqrt{n}(2 - n^{-1/2}\theta_0)$. Let $\tilde{\theta}$ denote the global maximizer of $f_n(\theta)$ again. We have $\tilde{\theta} > \theta_0$ if and only if $\sup_{\theta_0 \le \theta < 2\sqrt{n}} f_n(\theta) > \sup_{\theta < \theta_0} f_n(\theta)$, which implies that $\sup_{\theta_0 \le \theta < 2\sqrt{n}} f_n(\theta) > f_n(\theta_0)$. Let $A_n = \{\sup_{\theta_0 \le \theta < 2\sqrt{n}} f_n(\theta) > f_n(\theta_0)\}$ and $B_n = \{f'_{bn}(\theta_0) < 0\}$, where $f_{bn}(\theta)$ is defined below (23). Because of (24), $B_n = \{\sum_{i=1}^n \xi_{1i}^2 > \frac{1}{2}\sqrt{n}\epsilon\}$, where $\xi_{1i} = V_{0i} - 2V_{1i} \sim N(0, \frac{1}{3})$. Clearly

(28) $P(A_n) = P(A_n \cap B_n^c) + P(A_n \cap B_n) \le P(B_n^c) + P(A_n \cap B_n).$

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

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

for all $n$ by Lemma 12.

Next, in the event of $B_n$, because $f_{bn}(\theta)$ is unimodal, we have not only $f'_{bn}(\theta_0) < 0$ but also $f'_{bn}(\theta) < 0$ for all $\theta \in [\theta_0, 2\sqrt{n})$. But from (24), we have

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

which is strictly negative on $B_n$ for all $\theta \in [\theta_0, 2\sqrt{n})$ because then $f'_{bn}(\theta) < 0$. Also $f''_{an}(\theta) < 0$ globally and thus for all $\theta \in [\theta_0, 2\sqrt{n})$. Hence, $f''_n(\theta) < 0$ for all $\theta \in [\theta_0, 2\sqrt{n})$ in the event of $B_n$. This implies that $f_n(\theta_0) + (2\sqrt{n} - \theta_0)f'_n(\theta_0) \ge \sup_{\theta_0 \le \theta < 2\sqrt{n}} f_n(\theta)$ on $A_n \cap B_n$, and thus on $A_n \cap B_n$, $f_n(\theta_0) + (2\sqrt{n} - \theta_0)f'_n(\theta_0) > f_n(\theta_0)$, i.e., $(2\sqrt{n} - \theta_0)f'_n(\theta_0) > 0$, where

$$\begin{aligned} (2\sqrt{n} - \theta_0)f'_n(\theta_0) &= 2(2\sqrt{n} - \theta_0)\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_{2i} - \frac{1}{n}\sum_{i=1}^n \tilde{V}_{2i}\theta_0\right) \\ &\quad + n\left[1 - \frac{2}{2\sqrt{n} - \theta_0}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_{1i}^2\right)\right]. \end{aligned}$$

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

$$\begin{aligned} P(A_n \cap B_n) &\leq P\left\{2\epsilon\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_{2i} - \frac{1}{n}\sum_{i=1}^n \tilde{V}_{2i}\theta_0\right) + n\left[1 - \frac{2}{\epsilon}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_{1i}^2\right)\right] > 0\right\} \\ &= P\left\{\sum_{i=1}^n \xi_{1i}^2 < \frac{\epsilon}{2}\sqrt{n} + \epsilon^2\left(\frac{1}{n}\sum_{i=1}^n \xi_{2i} - \frac{1}{n^{3/2}}\sum_{i=1}^n \tilde{V}_{2i}\theta_0\right)\right\} \\ &\leq P\left(\sum_{i=1}^n \xi_{1i}^2 < \frac{\epsilon}{2}\sqrt{n} + \epsilon^2\bar{\xi}_2\right), \quad \bar{\xi}_2 := \frac{1}{n}\sum_{i=1}^n \xi_{2i}, \end{aligned}$$

where the last inequality holds because $\tilde{V}_{2i} \geq 0$. But

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

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

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

The result now follows from (28), (29) and (30). ■

**Proof of Lemma 10.** By Taylor expansion we have

$$\ln(1 + n^{-1/2}a_n + n^{-1}b_n) = n^{-1/2}a_n + n^{-1}b_n - \tfrac{1}{2}(n^{-1/2}a_n + n^{-1}b_n)^2 + R_n,$$

where $R_n = \frac{1}{3}(1 + n^{-1/2}\tilde{a}_n + n^{-1}\tilde{b}_n)^{-3}(n^{-1/2}a_n + n^{-1}b_n)^3$, $|\tilde{a}_n| \leq |a_n|$ and $|\tilde{b}_n| \leq |b_n|$. Because $a_n$ and $b_n$ are bounded, $R_n = O(n^{-3/2})$ and therefore

$$\begin{aligned} n\ln(1 + n^{-1/2}a_n + n^{-1}b_n) &= \sqrt{n}a_n + b_n - \tfrac{1}{2}a_n^2 + nR_n - n^{-1/2}a_n b_n - \tfrac{1}{2}n^{-1}b_n^2 \\ &= \sqrt{n}a_n + b_n - \tfrac{1}{2}a_n^2 + O(n^{-1/2}), \end{aligned}$$

as claimed. ■

# B Euler's Solution of Quartic Equations

In this appendix we present Euler's solution of the quartic equation

(31) $$x^4 + a_3x^3 + a_2x^2 + a_1x + a_0 = 0,$$

where the coefficient of $x^4$ is set to 1 without loss of generality. Transform by $x = y - a_3/4$. Then the original equation is written in terms of $y$ as in

(32) $$y^4 + b_2y^2 + b_1y + b_0 = 0,$$

where

$$(33)\quad b_2 = a_2 - \frac{3a_3^2}{8},\ \ b_1 = a_1 - \frac{a_3a_2}{2} + \frac{a_3^3}{8},\ \ b_0 = a_0 - \frac{a_3a_1}{4} + \frac{a_3^2a_2}{16} - \frac{3a_3^4}{256}.$$

If $b_1 = 0$, we can use the quadratic formula to solve the equation for $y^2$ and then recover $x$. If $b_0 = 0$, then $y = 0$ is a solution, and then the reduced cubic equation $y^3 + b_2y + b_1 = 0$ is to be solved. (These special cases happen with probability zero in our case.) In general, consider the auxiliary cubic equation

$$(34)\quad z^3 + \tfrac{1}{2}b_2z^2 + \tfrac{1}{16}(b_2^2 - 4b_0)z - \tfrac{1}{64}b_1^2 = 0,$$

which is called the *cubic resolvent* of equation (32). Let $z_1$, $z_2$ and $z_3$ denote the three roots of this cubic. (See Appendix C for a cubic solution.) Let $r_1$, $r_2$ and $r_3$ be such that $r_j^2 = z_j$, i.e., $r_j = \pm\sqrt{z_j}$. Choose signs such that $r_1r_2r_3 = -b_1$. (For example, choose $r_1 = \sqrt{z_1}$, $r_2 = \sqrt{z_2}$, and $r_3 = -b_1/(8r_1r_2)$.) Then the four roots of (32) are

$$\begin{aligned} y_1 &= +r_1 + r_2 + r_3, \quad y_2 = +r_1 - r_2 - r_3, \\ y_3 &= -r_1 + r_2 - r_3, \quad y_4 = -r_1 - r_2 + r_3. \end{aligned}$$

Finally, the roots of (31) are $x_j = y_j - a_3/4$.

When all roots of (34) are real and positive, (32) has four real roots. If (34) has one positive real root and two negative real roots, there are two pairs of complex conjugate roots of (32). Finally, if (34) has one positive real root and two complex conjugate roots, then (32) has two real roots and two complex conjugate roots. (These three cases are exhaustive because $z_1z_2z_3 = q^2 \geq 0$.) When $z_1$ is real (so is $r_1$) positive and $z_2$ and $z_3$ are complex, the two real roots are obtained as $r_1 \pm (r_2 + r_3)$ because $r_2 + r_3$ is real.

# C Real Roots of Cubic Equations

Birkhoff and MacLane (1996, *A Survey of Modern Algebra*, pp. 102–103) provide a trigonometric solution to find *real* roots of a cubic equation $a_3x^3 + a_2x^2 + a_1x_1 + a_0 = 0$ with $a_3 \neq 0$. Let $a_3 = 1$ without loss of generality. By the substitution $x = y - a_2/3$, we have

$$(35)\quad y^3 + py = q,\ \ p = a_1 - a_2^2/3,\ \ q = -a_0 + a_1a_2/3 - 2a_2^3/27.$$

If $p = 0$, the solution is $\text{sgn}(q)|q|^{1/3} - a_2/3$. Otherwise, let $z = y/h$, where $h = \sqrt{4|p|/3}$, and multiply $k \equiv 3/(h|p|)$ to both sides of (35). Then we have $k(hz)^3 + kp(hz) = kq$, i.e.,

$$4z^3 + 3\text{sgn}(p)z = C, \quad C = kq.$$

The solution depends on the sign of $p$.

**Case 1:** $p > 0$**.**

If $p > 0$, then the equation is $4z^3 + 3z = C$. To solve this equation, we use the trigonometric identity $\sinh 3\theta = 4\sinh^3\theta + 3\sinh\theta$, where $\sinh x = (e^x - e^{-x})/2$. Thus letting $z = \sinh\theta$, we have $\sinh 3\theta = C$, i.e., $\theta = (1/3)\sinh^{-1} C$, where $\sinh^{-1} x = \ln(x + \sqrt{x^2+1})$. So $z = \sinh(\frac{1}{3}\sinh^{-1} C)$. This is the unique real root of $z$.

**Case 2:** $p < 0$**.**

(i) If $p < 0$ and $C \geq 1$, then we use $\cosh 3\theta = 4\cosh^3\theta - 3\cosh\theta$, where $\cosh x = (e^x + e^{-x})/2$, to get $z = \cosh(\frac{1}{3}\cosh^{-1} C)$, where $\cosh^{-1} x = \ln(x + \sqrt{x^2-1})$. The is the unique real root. (ii) If $p < 0$ and $C \leq -1$, then the same method applies after changing the sign of $z$. Thus, we have $z = -\cosh(\frac{1}{3}\cosh^{-1} C)$. (iii) If $p < 0$ and $|C| < 1$, we use $\cos 3\theta = 4\cos^3\theta - 3\cos\theta$, to get $z = \cos(\frac{1}{3}\cos^{-1} C)$. There are three real roots in this case corresponding to $\cos^{-1} C = 2k\pi + \arccos C$ for $k = 0, 1, 2$.

Finally, $y = hz$ and $x = y - a_2/3$.

# References


Andrews, D. (1999). Estimation When a Parameter Is on a Boundary, *Econometrica*, 67(6), 1341–1383.

Andrews, D. (2001). Testing When a Parameter Is on the Boundary of the Maintained Hypothesis, *Econometrica*, 69(3), 683–734.

Ansley, C. F. (1979). An Algorithm for the Exact Likelihood of a Mixed Autoregressive-Moving Average Process, *Biometrika*, 66(1), 59–65.

Galbraith R. F., and J. I. Gailbrath (1974). On the inverses of some pattterned 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(4), 1993–2010.

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

Han, C., P. C. B. Phillips and D. Sul, (2010). "X-Differencing and Dynamic Panel Model Estimation" Yale University Working Paper.

Han, C., P. C. B. Phillips and D. Sul, (2011). "Uniform Asymptotic Normality in Stationary and Unit Root Autoregression," *Econometric Theory* (forthcoming).

Hahn, J., and G. Kuersteiner (2002). Asymptotically Unbiased Inference for a Dynamic Panel model with Fixed Effects When Both $n$ and $T$ Are Large, *Econometrica*, 70(4), 1639–1657.

Hsiao, C., M. H. Pesaran, and A. K. Tahmiscioglu (2002). Maximum Likelihood Estimation of Fixed Effects Dynamic Panel Data Models Covering Short Time Periods, *Journal of Econometrics*, 109, 107–150.

Knight, K. (2003). Epi-Convergence in Distribution and Stochastic Equi-Semicontinuity, Unpublished Manuscript, University of Toronto.

Kruiniger, H. (2008). Maximum Likelihood Estimation and Inference Methods for the Covariance Stationary Panel AR(1)/Unit Root Model, *Journal of Econometrics*, 2008, 447–464.

MacCurdy, 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.

Neyman, J. and E. Scott (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.

Wilson, P. D. (1988). Maximum Likelihood Estimation Using Differences in an Autoregressive-1 Process, *Communications in Statistics – Theory and Methods*, 17(1), 17–26.