

**UNIFORM ASYMPTOTIC NORMALITY IN STATIONARY  
AND UNIT ROOT AUTOREGRESSION**

**By**

**Chirok Han, Peter C. B. Phillips and Donggyu Sul**

**January 2010**

**COWLES FOUNDATION DISCUSSION PAPER NO. 1746**



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

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

# Uniform Asymptotic Normality in Stationary and Unit Root Autoregression\*

Chirok Han

*University of Auckland*

Peter C. B. Phillips

*Yale University, University of Auckland,*

*University of York & Singapore Management University*

Donggyu Sul

*University of Texas at Dallas*

Completed August, 2008; Revision, January, 2010

## Abstract

While differencing transformations can eliminate nonstationarity, they typically reduce signal strength and correspondingly reduce rates of convergence in unit root autoregressions. The present paper shows that aggregating moment conditions that are formulated in differences provides an orderly mechanism for preserving information and signal strength in autoregressions with some very desirable properties. In first order autoregression, a *partially aggregated* estimator based on moment conditions in differences is shown to have a limiting normal distribution which holds uniformly in the autoregressive coefficient  $\rho$  including stationary and unit root cases. The rate of convergence is  $\sqrt{n}$  when  $|\rho| < 1$  and the limit distribution is the same as the Gaussian maximum likelihood estimator (MLE), but when  $\rho = 1$  the rate of convergence to the normal distribution is within a slowly varying factor of  $n$ . A *fully aggregated* estimator is shown to have the same limit behavior in the stationary case and to have nonstandard limit distributions in unit root and near integrated cases which reduce both the bias and the variance of the MLE. This result shows that it is possible to improve on the asymptotic behavior of the MLE without using an artificial shrinkage technique or otherwise accelerating convergence at unity at the cost of performance in the neighborhood of unity.

*Keywords:* Aggregating information, Asymptotic normality, Bias Reduction, Differencing, Efficiency, Full aggregation, Maximum likelihood estimation.

*JEL classification:* C22

---

\*Thanks go to three referees for comments on the earlier version. Research for the paper was supported under a Marsden grant. Phillips acknowledges partial support from a Kelly Fellowship and the NSF under Grant No. SES 06-47086.

# 1 Introduction

The model considered in this paper is the simple autoregression with intercept

$$(1) \quad y_t = \alpha + x_t, \quad x_t = \rho x_{t-1} + \varepsilon_t, \quad t = 1, 2, \dots, n,$$

where  $\varepsilon_t$  is *iid*  $(0, \sigma^2)$ , and the autoregressive process  $x_t$  is initialized at some random quantity  $x_0$  with  $Ex_0^2 < \infty$ , allowing for stationarity by setting  $x_0 =_d N(0, \sigma^2 / (1 - \rho^2))$  when  $|\rho| < 1$  and for random recent initializations of the form  $x_0 = \sum_{j=1}^{\kappa} \varepsilon_{-j}$  for some fixed  $\kappa$  when  $\rho = 1$  (e.g., Phillips and Magdalinos, 2008). While the autoregression (1) is simple, it is the kernel of most dynamic econometric models and its properties are fundamental to more complicated models. The asymptotic properties of estimators of the scalar autoregressive coefficient  $\rho$  have been extensively studied and are reflected in various ways in more complex models of higher order and dimension. Moreover, difficulties in the development of a general optimal theory of estimation and testing in this simple model are well known and these too carry over to more complex settings. In particular, the discontinuity in the asymptotic theory as  $\rho$  passes through unity has attracted much attention since the work of White (1958) and Anderson (1959), and nonstandard limit theory in the locality of  $\rho = 1$  has been a continuing obstacle to a theory of efficient estimation and testing. Bias correction in the estimation of  $\rho$  is also affected because the bias function of the least squares estimator is nonlinear in  $\rho$  and asymptotic approximations to the bias formula are also discontinuous.

In practice, these obstacles have intensified interest in the unit root case and focused attention on testing  $\rho = 1$ . Of course, unit root testing does not completely solve the issue because of the low discriminatory power in unit root tests and because the true  $\rho$  may actually lie in a vicinity of unity with a localizing coefficient of the form  $k_n(1 - \rho)$  for some unknown sequence  $k_n \rightarrow \infty$ , which presents further difficulties in confidence interval construction. These issues also manifest in cointegrating regressions when the integration properties of the system variables are imperfectly known. One way of addressing these issues is to transform the system to stationarity by first differencing. In the univariate setting (1) it is well known that first differencing can help to unify limit theory in estimation (e.g. Phillips and Han, 2008) but this unification comes at the cost of infinitely deficient estimation when  $\rho$  is in the vicinity of unity.

This paper provides a new solution to this problem by using a novel form of variable differencing and aggregation that stacks information from differences at all lags in order to maximize the information used in estimation. To the best of our knowledge, this method of variable differencing has never been used before in statistics.

The present paper focuses on the simple AR(1) model though the approach we develop also works in a general AR(p) environment. The reason for this focus is that the goal of the paper is to present the key idea of the approach, to show the effects of partial as well as full aggregation in unifying the limit theory, and to demonstrate the possibility of uniform improvement over least squares and maximum likelihood estimation in autoregression. A companion paper (Han, Phillips, and Sul, 2009) deals with the AR(p) case and shows that the approach explored here is useful in panel data modeling as well as time series estimation.

Of course, the present work is not the first attempt to attack the discontinuity issue. Recently, for instance, Phillips and Han (2008) showed that a simple differencing transformation of (1) leads to the model

$$(2) \quad 2\Delta y_t + \Delta y_{t-1} = \rho \Delta y_{t-1} + \eta_t,$$

and least squares estimation of  $\rho$  in (2) gives the estimator

$$(3) \quad \hat{\rho}_n = \frac{\sum_{t=2}^n \Delta y_{t-1} (2\Delta y_t + \Delta y_{t-1})}{\sum_{t=2}^n (\Delta y_{t-1})^2},$$

which has a Gaussian limit distribution which is continuous as  $\rho$  passes through unity. In particular, uniformly for  $\rho \in (-1, 1]$ ,  $\sqrt{n}(\hat{\rho}_n - \rho) \Rightarrow N(0, 2(1 + \rho))$ . The symmetry of the limit distribution carries over to very small samples and the estimator  $\hat{\rho}_n$  has very little bias. Moreover, differencing removes the effects of both the intercept and initial conditions on the limit theory. These properties are particularly useful in panel models (Han and Phillips, 2009). However, in the time series context, the estimator  $\hat{\rho}_n$  is clearly inefficient for  $|\rho| < 1$  and infinitely deficient at  $\rho = 1$ . In spite of these drawbacks, the results reveal that uniform asymptotic normality is possible and that bias elimination in the estimation of dynamic models can be straightforward using differencing transformations.

The limit theory for  $\hat{\rho}_n$  exemplifies the fact that differencing transformations eliminate nonstationarity but in doing so they typically reduce signal strength and attenuate rates of convergence in unit root autoregressions. The present paper shows that it is possible to retain many of the useful properties of differencing transformations without reducing signal strength and information loss in estimation. In particular, it is possible to retain optimal rates of convergence for both stationary and nonstationary cases, to retain asymptotic efficiency in estimation in the stationary case, and to reduce both asymptotic bias and variance in unit root and near integrated cases, all with the same estimation procedure.

The approach that is adopted in the current paper is to aggregate moment conditions that are formulated in differences in an orderly fashion in order to preserve information and signal strength in autoregressions. This aggregation has some very desirable effects in terms of both bias and efficiency in regression. The basic idea is simply explained as follows. Underlying the estimator (3) is the orthogonality condition

$$(4) \quad E[2\Delta y_{t-1}\Delta y_t + (1 - \rho)(\Delta y_{t-1})^2] = E\Delta y_{t-1}\eta_t = 0,$$

which involves differences and lagged differences of  $y_t$ . The orthogonality condition (4) makes use of the variance and autocovariance of  $\Delta y_t$ , viz.  $E[(\Delta y_{t-1})^2] = 2\sigma^2/(1 + \rho)$  and  $E\Delta y_{t-1}\Delta y_t = -(1 - \rho)\sigma^2/(1 + \rho)$  and the dependence of these moments on  $\rho$ .

The central idea of the present paper is to make this notion systematic by exploiting the full autocovariance sequence of the differences. In particular, it is shown that the following moment conditions hold for differences and higher order differences

$$(5) \quad E(y_{t-1} - y_{s+1})[(y_t - y_s) - \rho(y_{t-1} - y_{s+1})] = 0,$$

and that these conditions apply for all  $\rho$  ( $-1 < \rho \leq 1$ ) and for all  $s \leq t - 3$ . Observe that, when  $s = t - 3$ , condition (5) can be written as

$$(6) \quad E[\Delta y_{t-1} \{ \{\Delta y_t + \Delta y_{t-2}\} + (1 - \rho)\Delta y_{t-1} \}] = 0,$$

which is a temporally balanced version of (4) in which  $2\Delta y_t$  is replaced by the sum  $\Delta y_t + \Delta y_{t-2}$  of differences on either side of  $\Delta y_{t-1}$ . Interestingly, the moment conditions (5) may also be written as simple orthogonality conditions for the differences

$$(7) \quad E(y_{t-1} - y_{s+1})(\varepsilon_t - \varepsilon_s^*) = 0,$$

using residuals  $\varepsilon_t$  of the regression equation (1) and the reverse regression residuals  $\varepsilon_s^* = x_s - \rho x_{s+1}$ . Our approach makes systematic use of all of these conditions in the estimation of  $\rho$ .

The moment conditions (5) hold simply because of the Yule-Walker equations for the autocovariance function of the differences when  $|\rho| < 1$ . So, *prima facie*, there would seem to be nothing new in (5) and no reason for there to be any advantage in using these conditions. However, the conditions also apply when  $\rho = 1$  and in this case the process and its moments are nonstationary. In particular, the sample moment functions are no longer stationary when the difference  $t - s$  increases too fast. In this event, the moment conditions subtly carry information about the nonstationarity of the process while still operating in difference format. It is this facility to control the amount of information carried in the conditions that opens up new possibilities in estimation.

We mention three possibilities here. The simplest scenario is to use the moment condition (5) with  $s = t - 2 - \ell$  for some fixed  $\ell \geq 1$ . The resulting method of moments estimator will then be uniformly asymptotically normal, just as for the estimator  $\hat{\rho}_n$  which effectively corresponds with the case  $\ell = 1$  as shown above.

The next possibility is to combine the moment conditions (5) for  $\ell = 1, \dots, L$  and allow  $L \rightarrow \infty$  but at a slower rate than  $n$  so that  $L/n \rightarrow 0$ . The number of moment conditions is then a *small infinity* relative to  $n$ . The resulting unweighted GMM estimator is a *partially aggregated* moment condition estimator based on differences of the original series. This partially aggregated estimator is consistent and uniformly asymptotically normal for all  $-1 < \rho \leq 1$ . Importantly, this estimator uses information that includes the behavior of the differences  $y_t - y_{t-2-\ell}$  as  $\ell$  increases and, by doing so, achieves full efficiency in estimation for  $|\rho| < 1$  and raises the rate of convergence to  $\sqrt{nL}$  while retaining asymptotic normality for  $\rho = 1$ . The rate  $\sqrt{nL}$  can be made arbitrarily close to  $n$  by choosing  $L = n/L(n)$  for some suitable slowly varying function such as  $L(n) = \log n$ . To the authors' knowledge, this is the first demonstration that uniform asymptotic normality is achievable in autoregressive estimation for  $-1 < \rho \leq 1$  with a rate of convergence at  $\rho = 1$  that is arbitrarily close to  $n$ . The result also suggests, although that it is not proved, that in order to achieve an asymptotic normalizing transformation of the Gaussian MLE that holds uniformly in  $\rho$  (including  $\rho = 1$ ) it is necessary to give up at least a slowly varying factor in the rate of convergence at  $\rho = 1$ .

A final possibility is to combine all possible moment conditions in (5) using the differences  $y_t - y_{t-2-\ell}$  for all  $\ell = 1, \dots, t - 3$ . We call the unweighted GMM estimator in this case the *fully aggregated* estimator. The fully aggregated estimator (FAE) is consistent and now uses information in all the differences  $y_t - y_{t-2-\ell}$ . In this case, the number of moment conditions increases at the rate  $O(n)$ . Interestingly, as for the partially aggregated estimator, this *unweighted* GMM is  $\sqrt{n}$  convergent, asymptotically normal, and attains full asymptotic efficiency for  $|\rho| < 1$ . This efficiency result, like that for the partially aggregated estimator in this case, is somewhat unexpected because the moment functions are cross-correlated and no information about cross-correlation is used in estimation. When  $\rho = 1$ , the FAE is  $n$ -convergent and has a nonstandard limit distribution. Computations indicate that the limit distribution of this FAE has much less bias and has smaller variance than that of the Gaussian MLE. These gains carry over to finite samples and to stationary values of  $\rho$ .

When the intercept is known in (1), the model is effectively a levels autoregression through the origin. In that case, it is well known that higher lag moment conditions of the form  $E y_{t-\ell} \varepsilon_t = 0$  are redundant for all  $\ell > 1$  once the primary moment condition  $E y_{t-1} \varepsilon_t = 0$  that is implied by the martingale structure is used (see Kim, Qian and Schmidt, 1999). However, in the case of an

unknown intercept, the present results reveal that higher lag moment conditions for the differenced series  $y_t - y_{t-2-\ell}$  are not redundant for  $\ell > 1$  and play a major role in achieving asymptotic efficiency for  $|\rho| < 1$  and in improving the convergence rate for  $\rho = 1$ . In this case, the martingale structure is lost in the differences and there is information in moment conditions associated with long differences. Use of this additional information is important in attaining efficiency for  $|\rho| < 1$  and the optimal rate of convergence when  $\rho = 1$ .

The use of differences provides an automated mechanism for reducing bias in autoregression. One of the main reasons for bias in the maximum likelihood estimation of autoregressive models like (1) is the non-orthogonality induced by removal of the sample mean. While the use of differences complicates the form of the autocovariogram, orthogonality conditions such as (7) are retained in estimation and this provides a natural mechanism for bias removal in the new regression. Importantly, in the unit root case, this effect of bias removal persists in the limit distribution for both the partially aggregated and fully aggregated estimators. Accordingly, the use of additional information in the moment conditions for long differences serves to accelerate convergence in the unit root case from  $\sqrt{n}$  to  $n$  while retaining the advantages of bias correction from the use of differences and maintaining good performance in estimation for  $|\rho| < 1$ . These advantages might be expected to have even greater impact in the case of panel regressions, where maximum likelihood bias turns to inconsistency in the presence of incidental parameter problems induced by fixed effects. A companion paper (Han, Phillips and Sul, 2009) shows this to be so and considers extensions of the method presented here to higher order autoregressions and panel models.

The present paper is organized as follows. Section 2 develops the new moment conditions and discusses their implications. Section 3 develops moment based estimation procedures in the case of a single lag. Section 4 provides the limit theory for the fully aggregated estimator and studies the relation of this limit distribution to that of the Gaussian MLE. The partially aggregated estimator is studied in Section 5, uniform asymptotic normality is established, and continuity of the limit theory through roots that are local to unity is shown. Section 6 concludes. Technical derivations and proofs are given in the Appendix. Notation is standard.

## 2 New Moment Conditions and Information Aggregation

Observed data  $y_t$  are assumed to be generated by (1) for some  $\rho$  satisfying  $-1 < \rho \leq 1$  and unknown  $\alpha$ . Define  $\varepsilon_s^* = x_s - \rho x_{s+1}$ , so that  $x_s$  satisfies the reverse regression  $x_s = \rho x_{s+1} + \varepsilon_s^*$ . Subtracting this reversed regression from the original autoregression for  $x_t$  gives the differenced equation  $x_t - x_s = \rho(x_{t-1} - x_{s+1}) + (\varepsilon_t - \varepsilon_s^*)$ . Using the fact that  $x_t - x_s \equiv y_t - y_s$  and letting  $s = t - 2 - \ell$  yields

$$(8) \quad y_t - y_{t-2-\ell} = \rho(y_{t-1} - y_{t-1-\ell}) + (\varepsilon_t - \varepsilon_{t-2-\ell}^*).$$

Somewhat unexpectedly, the regressor and the regression error are uncorrelated in (8), producing the key orthogonality condition (5) discussed above.

**Lemma 1** *For any  $\ell \geq 1$ ,  $E(y_{t-1} - y_{t-1-\ell})(\varepsilon_t - \varepsilon_{t-2-\ell}^*) = 0$ .*

To the best of our knowledge, the moment conditions given in Lemma 1 have not been noticed before in the literature, although previous authors (e.g., Fuller, 1976; Park and Fuller, 1995) have

noted properties of the reverse regression  $x_s = \rho x_{s+1} + \varepsilon_s^*$ . As shown in the proof, covariance stationarity of  $x_t$  is central to this moment condition when  $\rho < 1$ . But, interestingly, the condition still holds in the unit root case. A few other points are worth making.

First, equation (8) involves no intercept. So estimators based on (8) will not suffer from the bias that is typically induced by demeaning procedures to remove the constant term. This bias can be significant and it is well known to be of substantial importance in panel models with fixed effects, especially when the cross section sample size is large in relation to the time series sample. Our companion paper (Han, Phillips and Sul, 2009) systematically investigates the use of aggregated differenced based estimators for bias elimination in dynamic panel models.

As discussed in the introduction, the special case where  $\ell = 1$  relates to the moment condition (4) used in Phillips and Han (2008). When  $\ell = 1$ , the moment condition in Lemma 1 can be written as in (6) so that Phillips and Han's moment  $2E\Delta y_{t-1}\Delta y_t$  is replaced by  $E\Delta y_{t-1}\Delta y_t + E\Delta y_{t-2}\Delta y_{t-1}$ . So the moment condition in Lemma 1 for  $\ell = 1$  is a temporally balanced version of Phillips and Han's (2008) moment condition, as indicated in the Introduction.

In the levels autoregression  $x_t = \rho x_{t-1} + \varepsilon_t$  where  $x_t$  is observable, higher order orthogonality conditions such as  $E x_{t-k} \varepsilon_t = 0$  for  $k \geq 2$  do not add efficiency once the first lag-order moment condition  $E x_{t-1} \varepsilon_t = 0$  is used (see Kim, Qian and Schmidt, 1999, Theorem 2). However, higher  $\ell$  values in the transformed equation (8) add information in both the stationary and unit root cases. As indicated earlier, this information accumulation is due to the fact that the martingale condition does not hold in the differenced model. So, the presence of an intercept makes an important difference in terms of the relevance/redundancy of moment conditions. As shown below, the effects of the information accumulation are particularly dramatic in the unit root case. We first investigate how information from a single lagged difference can be used and then consider the full and partial usage of all the lagged difference moment conditions.

### 3 Partial Information Using a Single-Lag Difference

Let  $y_t$  be observed for  $t = 1, \dots, n$ . From Lemma 1, simple least squares (OLS) regression applied to (8) yields a consistent *single-lag* estimator

$$(9) \quad \tilde{\rho}_\ell = \frac{\sum_{t=3+\ell}^n (y_{t-1} - y_{t-1-\ell})(y_t - y_{t-2-\ell})}{\sum_{t=3+\ell}^n (y_{t-1} - y_{t-1-\ell})^2} = \rho + \frac{\sum_{t=3+\ell}^n (x_{t-1} - x_{t-1-\ell})(\varepsilon_t - \varepsilon_{t-2-\ell}^*)}{\sum_{t=3+\ell}^n (x_{t-1} - x_{t-1-\ell})^2},$$

for any fixed  $\ell \geq 1$ . In particular, for any  $\ell$ , we have the following result.

**Theorem 2** Let  $S_\ell(\rho) = \sum_{j=0}^{\ell-1} \rho^j$ . Let  $n_\ell = n - 2 - \ell$ . For all  $\rho \in (-1, 1]$ , we have

$$(10) \quad n_\ell^{1/2}(\tilde{\rho}_\ell - \rho) \Rightarrow N(0, \Omega_\ell), \quad \Omega_\ell = 2(1 + \rho)S_\ell(\rho)^{-1},$$

as  $n \rightarrow \infty$ .

The limit distribution (10) is uniformly valid for all stationary  $\rho$  values and for  $\rho = 1$ .

As shown in the proof of Theorem 2, the inverse of  $\Omega_\ell$ , which is proportional to  $S_\ell(\rho)/(1+\rho)$ , is the probability limit of  $n^{-1}$  times the denominator in (9), and represents the amount of information

contained in the moment condition in Lemma 1 for a given  $\ell$ . When  $\rho = 0$ , it is clear that  $S_\ell(\rho)$  is identical for all  $\ell$ , and so in this case asymptotic efficiency remains unaffected by the choice of  $\ell$ , although some initial observations are obviously lost in finite samples. On the other hand, if  $\rho$  is close to unity, then the amount of information strictly increases with the lag difference  $\ell$ , and the resulting estimator becomes correspondingly more efficient as  $\ell$  increases. This reasoning is confirmed by the fact that  $\Omega_\ell$  decreases as  $\ell$  increases if  $\rho > 0$ . Clearly more observations are lost by increasing  $\ell$  so the finite sample variance (and asymptotic variance  $\Omega_\ell / (n - 2 - \ell)$ ) does not always decrease as  $\ell$  increases. But if  $n$  is large, then the efficiency loss from losing the first few observations would be small relative to the fall in  $\Omega_\ell$ , and the asymptotic variance of  $\tilde{\rho}_\ell$  will fall as  $\ell$  increases.

## 4 Full Information Aggregation using All Lag Differences

Information contained in the form of the moment conditions in Lemma 1 for a single  $\ell$  is obviously limited. We can make use of the conditions for all  $\ell$  by stacking the equations as

$$(11) \quad y_t - y_{t-2-\ell} = \rho(y_{t-1} - y_{t-1-\ell}) + (\varepsilon_t - \varepsilon_{t-2-\ell}^*), \quad \ell = 1, 2, \dots, t-3,$$

and using all of them in estimation. This procedure is equivalent to unweighted GMM estimation based on Lemma 1 for all possible  $\ell$  values. The resulting estimator  $\hat{\rho}_{fa}$ , which we call the *full aggregation estimator* (FAE), is

$$(12) \quad \hat{\rho}_{fa} = \frac{\sum_{\ell=1}^{n-3} \sum_{t=3+\ell}^n (y_{t-1} - y_{t-1-\ell})(y_t - y_{t-2-\ell})}{\sum_{\ell=1}^{n-3} \sum_{t=3+\ell}^n (y_{t-1} - y_{t-1-\ell})^2},$$

and simply involves pooling the information in the system (11). The following equivalence relating  $\hat{\rho}_{fa}$  to the usual least squares regression estimator is useful.

**Theorem 3** *Let  $\tilde{y}_{t-1} = y_{t-1} - n_0^{-1} \sum_{s=3}^n y_{s-1}$  and  $\tilde{y}_t = y_t - n_0^{-1} \sum_{s=3}^n y_s$  (at the risk of some notational confusion). Let  $\hat{\rho}_{ols} = (\sum_{t=3}^n \tilde{y}_{t-1}^2)^{-1} \sum_{t=3}^n \tilde{y}_{t-1} \tilde{y}_t$  be the OLS coefficient estimator from the regression of  $y_t$  on  $y_{t-1}$  and an intercept using the observations  $\{y_2, \dots, y_n\}$ . Then*

$$(13) \quad \hat{\rho}_{fa} \equiv \hat{\rho}_{ols} + \frac{n_0^{-1} \sum_{t=3}^n x_{t-1}^2}{\sum_{t=3}^n \tilde{x}_{t-1}^2} + \frac{x_1 x_2 - n_0^{-1} (x_1 + x_2) \sum_{t=3}^n x_{t-1}}{\sum_{t=3}^n \tilde{x}_{t-1}^2},$$

where  $n_0 = n - 2$ .

When  $|\rho| < 1$ , the second and third terms on the right hand side of (13) are  $O_p(n^{-1})$ , so  $n^{1/2}(\hat{\rho}_{fa} - \hat{\rho}_{ols}) = O_p(n^{-1/2})$ , implying that the full aggregation estimator and OLS estimator have the same asymptotic distribution in the stationary case. Since  $\hat{\rho}_{ols}$  is asymptotically equivalent to (and hence as efficient as) the Gaussian MLE using  $y_2, \dots, y_n$  when  $|\rho| < 1$ , it follows that no information is lost asymptotically by the differencing and information aggregation procedure in this case. However, the first observation  $y_1$  is lost, which is not important asymptotically but does affect performance in very small samples.



By contrast, when  $\rho = 1$ , the second term of (13) is  $O_p(n^{-1})$  and the third term is  $O_p(n^{-2})$ . So the limit distribution of  $n(\hat{\rho}_{fa} - \rho)$  differs from the distribution of  $n(\hat{\rho}_{ols} - \rho)$  by virtue of the second term. Note that the second term is positive, so that the term always provides an upward adjustment to the MLE<sup>1</sup>, delivering a built-in bias reduction. Importantly, this bias reduction is retained and plays a significant role in the form of the asymptotic distribution. In fact, as we report below, the full aggregation (FA) procedure substantially reduces the bias in ML estimation and also produces a reduction in variance, with both improvements holding in the limit distribution.

The limit theory is straightforward and is detailed as follows.

**Corollary 4** (i) If  $|\rho| < 1$ , then  $n^{1/2}(\hat{\rho}_{fa} - \rho) \Rightarrow N(0, 1 - \rho^2)$ . (ii) If  $\rho = 1$ , then

$$(14) \quad n(\hat{\rho}_{fa} - 1) \Rightarrow \frac{\int_0^1 \tilde{B}_r dB_r + \int_0^1 B_r^2 dr}{\int_0^1 \tilde{B}_r^2 dr}$$

where  $B_r$  is standard Brownian motion and  $\tilde{B}_r$  is the corresponding demeaned process, i.e.,  $\tilde{B}_r = B_r - \int_0^1 B_s ds$ . (iii) If  $\rho = 1 - c/n$  for some constant  $c$ , then

$$(15) \quad n(\hat{\rho}_{fa} - \rho) \Rightarrow \frac{\int_0^1 \tilde{J}_r dB_r + \int_0^1 J_r^2 dr}{\int_0^1 \tilde{J}_r^2 dr},$$

where  $J_r = \int_0^r e^{c(r-s)} dB_s$  and  $\tilde{J}_r = J_r - \int_0^1 J_s ds$ .

As shown in the Appendix, an alternative form of the limit distribution given in (14) is

$$(16) \quad n(\hat{\rho}_{fa} - 1) \Rightarrow \frac{\int_0^1 \int_0^r (1 + p - r) dB_p dB_r}{\int_0^1 \int_0^r (B_r - B_s)^2 ds dr}.$$

The denominator and the double summations in (16) more closely resemble the form of the corresponding expressions in (12) for the estimator  $\hat{\rho}_{fa}$ . The FAE estimator pools moment conditions involving long differences in the time series and these differences are reflected in the limit formula (16).

When  $\rho = 1$ , the MLE has the limit distribution  $(\int_0^1 \tilde{B}_r^2 dr)^{-1} \int_0^1 \tilde{B}_r dB_r$ , and it follows that  $n(\hat{\rho}_{fa} - \hat{\rho}_{mle}) \Rightarrow (\int_0^1 \tilde{B}_r^2 dr)^{-1} \int_0^1 B_r^2 dr > 0$ . The FAE therefore provides an upward correction to the MLE asymptotically, thereby adjusting the well-known downward bias of the MLE. In a similar way,  $n(\hat{\rho}_{fa} - \hat{\rho}_{mle}) \Rightarrow (\int_0^1 \tilde{J}_r^2 dr)^{-1} \int_0^1 J_r^2 dr > 0$  when  $\rho = 1 + c/n$ , so the same upward correction applies in local neighborhoods of unity. Figure 1 illustrates the finite sample effects of this upward correction to the MLE when  $n = 500$ . As is apparent from the cross plots of  $(\hat{\rho}_{fa}, \hat{\rho}_{mle})$  given in these figures in relation to the 45° line, the greatest correction occurs in the unit root case. But the correction is active in all other cases. At  $\rho = 0$  the correction is clearly small but is still effective in reducing bias as confirmed in simulations reported below.

Importantly, the numerator of  $\hat{\rho}_{fa} - \rho$  has no bias whatsoever for all  $n$  and for all  $\rho$ , unlike the corresponding numerator of  $\hat{\rho}_{mle} - \rho$ . This property turns out to have a large impact on the finite sample performance of the estimator  $\hat{\rho}_{fa}$  for all  $\rho$  and its limit behavior for  $\rho = 1$ . Also,

<sup>1</sup>This is a conditional MLE or OLS estimator.

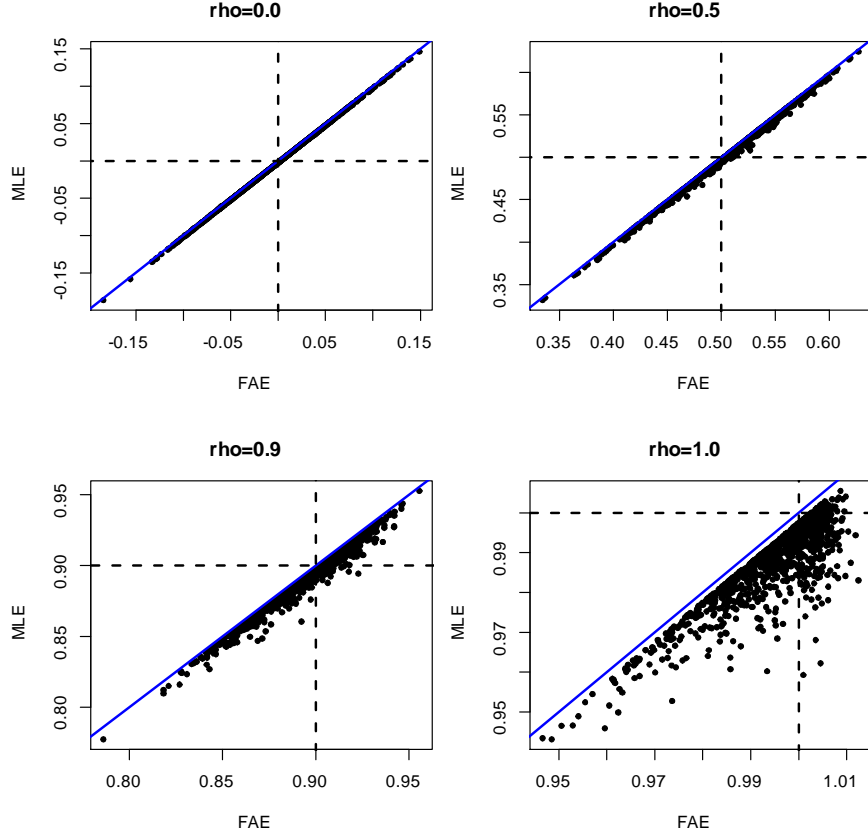


Figure 1: Cross plots of  $\hat{\rho}_{MLE}$  and  $\hat{\rho}_{fa}$  for  $\rho = 0, 0.5, 0.9, 1.0$  with  $n = 400$  against a  $45^\circ$  line.

the bias reduction in the estimator  $\hat{\rho}_{fa}$  is achieved automatically for any error distribution without calculating a bias formula (e.g. Andrews, 1993) or using simulation based methods like indirect inference (Gouriéroux, Phillips and Yu, 2006). The method is therefore very convenient to use in practical work. It also has very useful applications in dynamic panel data models with fixed effects.

When  $|\rho| < 1$ , the FAE is asymptotically efficient and has less finite sample bias than the MLE. When  $\rho = 1$ , the asymptotic distribution of the FA estimator has less bias and turns out to have less variance also. Table 1 reports detailed results from 100,000 simulations of (1) with standard normal errors  $\varepsilon_t$  and for  $\rho = 0, 0.5, 0.9, 0.95, 0.99, 1.0$ . The table shows the bias, standard deviation, and RMSE measures, and the ratio of the standard deviations of the FAE and MLE estimators for these values of  $\rho$  and for  $n = 100, 500, 5000$ . Bias reduction is clear in all cases and for all parameter values, even for  $\rho = 0$  and very large sample sizes. For  $\rho = 1$ , the bias reduction is substantial—greater than 50% for all sample sizes, including  $n = 5,000$ . Strikingly, the standard deviation is also smaller for the FAE than the MLE when  $\rho \geq 0.9$  in the table. Although not reported here, the asymptotic simulation standard errors for the standard deviation ratio were found to be very small, and the null hypothesis that the standard deviation ratio is unity is rejected for all  $n$  and  $\rho$  values. The reduction for  $\rho = 1$  is observed for all simulations conducted for  $n > 100$  and is sustained in very large samples up to  $n = 10,000$ . Corresponding to the reductions in both bias and variance,

RMSE is smaller for the FAE than the MLE for all parameter settings except  $\rho = 0$ .

Figure 3 illustrates simulated densities for the MLE and FAE from 10,000 replications with standard normal disturbances and  $n = 500$ . We observe that the distribution of the FAE is well centered at the true parameter (though the bias is not completely removed, as is clear from the table, and there is some asymmetry in the distribution) and that its density at the mode is slightly higher than the highest density of the MLE, corresponding to the smaller simulated variances.

Remarkably, the calculations for large  $n$  indicate that the FAE has smaller asymptotic variance than the MLE when  $\rho = 1$  as well as smaller bias. Of course, standard optimality theory for ML estimation does not apply in autoregressions which include nonstationary cases because of the change in the convergence rate as  $\rho$  passes through unity and the presence of nonstandard limiting distribution theory with skewness and bias in the asymptotic distribution at  $\rho = 1$ . We may therefore expect that it is possible to “beat” the MLE at unity while retaining its good behavior elsewhere and to do so without resorting to “superefficient” estimators of the Hodges or Bayesian type. For example, it is known (Phillips, 1993, 1995) that standard FM regression methods lead to accelerated convergence in estimation with rates that exceed  $O(n)$  at  $\rho = 1$ , while retaining the usual Gaussian limit theory of the MLE for  $|\rho| < 1$ . While FM methods of estimation are not pathological (in the sense that they are not intentionally designed to provide faster convergence rates at a single point in the parameter space), these methods do have the unsatisfactory property that they have similar behavior to the unit root case in a nontrivial locality of unity. By contrast, as shown in Corollary 4, the FAE has discriminatory asymptotic behavior (different from that at unity) in a local neighborhood of unity, while showing improved performance at unity and at neighboring points.

According to unreported simulations for various  $n$  (ranging from 50 to 10,000), for all values of  $1 \geq \rho \geq 0.70$  and with the exception of a few small  $n$  values, the FAE reduces variance in estimation relative to the MLE. The gains for  $\rho \geq 0.9$  are clear and in the unit root case are dramatic because they are sustained in the limit as  $n \rightarrow \infty$ . For  $n = 10,000$  the standard deviation ratio of FAE to MLE is 0.98, indicating a 2% reduction in dispersion in the unit root case. Gains in terms of variance reduction also occur when  $\rho$  is local to unity. Figure 2 plots the ratio of the standard deviations of the FAE and MLE for  $\rho = 1 - c/n$  for the same values of  $n$  and for various values of the localizing coefficient  $c$ . The reductions in dispersion are clear across all values of  $c$  and follow a similar pattern to the gains in the unit root case. The values at  $n = 10,000$  may be interpreted as estimates of the reduction in the asymptotic variance. As in the unit root case, these are significant. Interestingly, it appears that the asymptotic variance reduction in the FAE over the MLE is greater for small  $c \neq 0$  than it is  $c = 0$ .<sup>2</sup> The maximum gain seems to be around  $c = 3$  where FAE achieves more than a 4% reduction in asymptotic standard deviation over the MLE.

Note that the bias correction delivered by full aggregation is not complete because of the correlation between the numerator and the denominator. Thus, a mean or median unbiased estimator (e.g., Andrews, 1993; Gouriéroux, Phillips and Yu, 2006) can be expected to have better bias correction performance because those methods use the bias function (computed by simulation) of the ML estimator in constructing the bias corrected estimator. According to a simulation of 10,000 replications, the simulation-based indirect inference method (Gouriéroux, Phillips and Yu, 2006) has smaller bias (approximately 1/2 of the bias of the FAE) when  $\rho = 1$ . Figure 3 gives the density

---

<sup>2</sup>Phillips and Lee (1996) found a similar result in terms of the maximum efficiency gains of generalized least squares relative to ordinary least squares in trend removal in regression with near integrated errors.

Table 1: Performance Characteristics of the MLE and FAE from 100,000 Simulations

$\rho$	$n$	Bias $\times 100$		$\sqrt{n} \times$ SD		SD	RMSE	
		MLE	FAE	MLE	FAE	ratio	MLE	FAE
0.00	100	-0.9957	0.0150	1.0008	1.0112	1.0104	0.1006	0.1011
0.00	500	-0.2113	-0.0109	1.0012	1.0032	1.0020	0.0448	0.0449
0.00	5000	-0.0237	-0.0037	1.0033	1.0035	1.0002	0.0142	0.0142
0.50	100	-2.5221	-1.0006	0.8928	0.9002	1.0082	0.0928	0.0906
0.50	500	-0.5153	-0.2148	0.8732	0.8744	1.0013	0.0394	0.0392
0.50	5000	-0.0528	-0.0228	0.8686	0.8687	1.0002	0.0123	0.0123
0.90	100	-3.9992	-1.8850	0.5684	0.5640	0.9923	0.0695	0.0595
0.90	500	-0.7621	-0.3729	0.4668	0.4642	0.9944	0.0222	0.0211
0.90	5000	-0.0751	-0.0371	0.4385	0.4382	0.9994	0.0062	0.0062
0.95	100	-4.3771	-2.0286	0.4997	0.4941	0.9889	0.0664	0.0534
0.95	500	-0.8044	-0.3942	0.3583	0.3540	0.9881	0.0179	0.0163
0.95	5000	-0.0780	-0.0389	0.3166	0.3161	0.9984	0.0045	0.0045
0.99	100	-4.9770	-2.1986	0.4445	0.4430	0.9965	0.0667	0.0495
0.99	500	-0.9115	-0.4274	0.2339	0.2273	0.9720	0.0139	0.0110
0.99	5000	-0.0810	-0.0401	0.1522	0.1509	0.9913	0.0023	0.0022
1.00	100	-5.2544	-2.3045	0.4324	0.4355	1.0070	0.0681	0.0493
1.00	500	-1.0711	-0.4731	0.1997	0.1976	0.9895	0.0139	0.0100
1.00	5000	-0.1078	-0.0477	0.0641	0.0627	0.9785	0.0014	0.0010

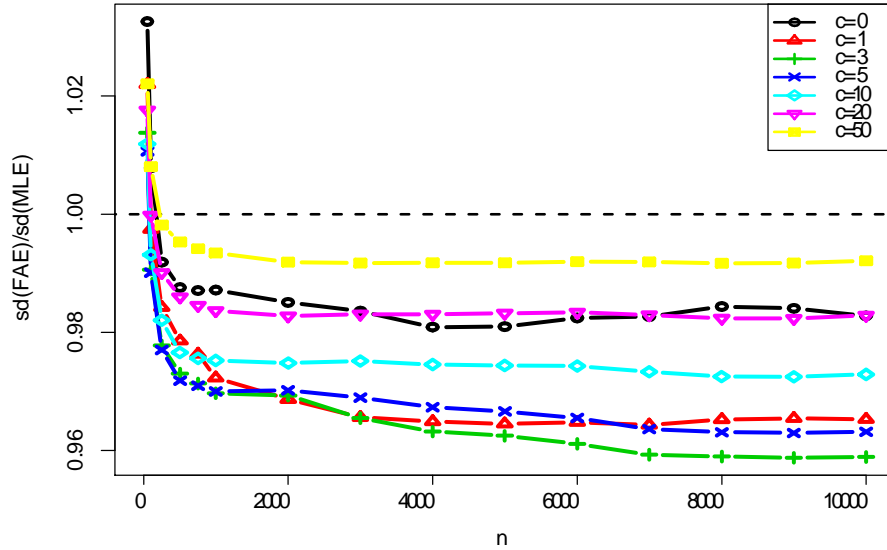


Figure 2: Ratios of standard deviations of FAE and MLE estimators for  $\rho = 1 - c/n$ , various  $c$ , and  $n = 100, 500, 1000, 2000, \dots, 10,000$  computed from 100,000 simulations.

of the indirect inference estimator from this simulation. Though the FAE has larger bias, the mode of the simulated distribution is a lot closer to the true value than the indirect inference estimator.

The performance of FAE may also be compared with bias correction estimators in the literature such as the weighted symmetric estimator (WSE) of Park and Fuller (1995), the restricted MLE (RML) of Cheang and Reinsel (2000), and bias correction by Roy and Fuller (2001). According to simulations (see Table 2) conducted for  $n = 100$  and 20,000 replications, the FAE has a smaller bias and a slightly bigger variance than the WSE and the RML, and the mean squared error (MSE) of the FAE is smaller for large  $\rho$  and larger for small  $\rho$  than the WSE and the RML. (Other  $n$  values have also been examined, and the behavior is largely the same.) The Roy-Fuller bias correction estimator (RF) looks better than the FAE around unity, which is partly explained by the fact that RF approximately corrects for bias by a piecewise function that employs an approximate analytical bias formula for the OLS estimator adjusted to do well at unity where it is median unbiased. The RF adjustments change according to the value of  $n$ , while the FAE uses the same formula for all values of  $\rho$  and  $n$ . It is also notable that RF uses the estimator when it exceeds unity (depending on a choice parameter), which affects the MSE favorably. However, MSE comparisons favor the FAE procedure for  $\rho \leq 0.90$ , bias reduction in FAE is achieved by a simple least squares method without any special devices or adjustments, FAE has the same form in both stationary and unit root cases, and FAE estimation does not require knowledge of a bias formula or simulations of the bias. These properties are especially useful in application of the FAE procedure in panel models, as explored in other work by the authors (Han, Phillips and Sul, 2009).

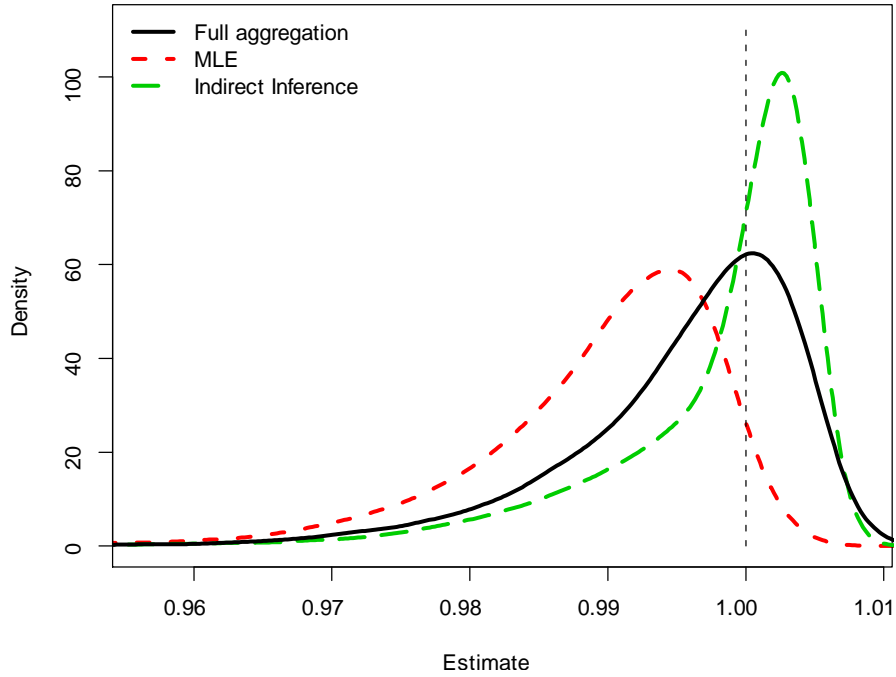


Figure 3: Densities of MLE, FAE and Indirect Inference Estimators ( $n = 500, \rho = 1$ )

Table 2: Comparison of OLS, FAE, Park-Fuller, Roy-Fuller, and Cheang-Reinsel for  $n = 100$

$\rho$	Mean					$n \times \text{MSE}$				
	OLS	FAE	WSE	RF	RML	OLS	FAE	WSE	RF	RML
0.00	-.010	-.000	-.010	-.000	-.000	1.015	1.025	1.015	1.066	1.024
0.20	.184	.196	.184	.200	.196	1.000	.996	1.000	1.035	.994
0.40	.378	.392	.378	.400	.391	.921	.895	.919	.927	.893
0.60	.572	.588	.572	.600	.587	.780	.724	.774	.747	.722
0.80	.765	.784	.766	.800	.781	.585	.487	.570	.505	.487
0.90	.860	.881	.862	.900	.876	.479	.351	.453	.371	.354
0.95	.906	.930	.910	.949	.922	.437	.283	.396	.289	.292
0.99	.940	.968	.948	.980	.956	.439	.239	.361	.213	.263
1.00	.947	.977	.957	.986	.963	.461	.243	.372	.206	.286

## 5 Partial Aggregation and Uniform Asymptotic Normality

When  $|\rho| < 1$ , both the single-lag estimator  $\tilde{\rho}_\ell$  (using a single  $\ell$ ) and the FAE  $\hat{\rho}_{fa}$  have a  $\sqrt{n}$  convergence rate and the limit distribution is normal. But if  $\rho = 1$ , then  $\tilde{\rho}_\ell$  has a  $\sqrt{n}$ -rate of convergence with a Gaussian limit, while  $\hat{\rho}_{fa}$  has the faster  $O(n)$  convergence rate and a non-Gaussian limit distribution. The asymptotic distribution of the FAE therefore has a discontinuity at  $\rho = 1$  analogous to the MLE. Since there are many intermediate choices of moment conditions underlying the estimators  $\tilde{\rho}_\ell$  and  $\hat{\rho}_{fa}$ , it is natural to reason that there may exist a *partial aggregation* method that embodies some of the advantages of both  $\tilde{\rho}_\ell$  and  $\hat{\rho}_{fa}$ . In particular, it might be anticipated that such a procedure might yield a uniform Gaussian limit distribution and at the same time a faster convergence rate at  $\rho = 1$  than  $\hat{\rho}_\ell$ .

A natural approach to consider is to use only part of the information in the moment conditions (5) by aggregating a small infinity (say  $L \rightarrow \infty$  with  $\frac{L}{n} \rightarrow 0$ ) of these conditions rather than all of them. Accordingly, we define the *partial aggregation estimator* (PAE) as

$$(17) \quad \hat{\rho}_{pa} = \frac{\sum_{\ell=1}^L \sum_{t=3+\ell}^n (y_{t-1} - y_{t-1-\ell})(y_t - y_{t-2-\ell})}{\sum_{\ell=1}^L \sum_{t=3+\ell}^n (y_{t-1} - y_{t-1-\ell})^2}.$$

When  $\frac{1}{L} + \frac{L}{n} \rightarrow 0$ , uniform asymptotic normality is obtained and the rate of convergence at  $\rho = 1$  depends on the expansion rate of  $L$ . A general form of the limit theory for the PAE is as follows.

**Theorem 5** Let  $S_\ell(\rho) = \sum_{j=1}^{\ell} \rho^{j-1}$ , and  $\bar{S}_L(\rho) = L^{-1} \sum_{\ell=1}^L S_\ell(\rho)$ . Let

$$V_L(\rho) = \frac{1}{L} + \frac{1}{L} \sum_{\ell=1}^{L-1} \left[ 1 - \frac{\ell}{L} + \frac{S_{L-\ell}(\rho)}{L} \right] \frac{S_\ell(\rho)}{\bar{S}_L(\rho)}.$$

For all  $\rho$ , we have

$$(18) \quad \left[ \frac{n \bar{S}_L(\rho)}{2(1+\rho)V_L(\rho)} \right]^{1/2} (\hat{\rho}_{pa} - \rho) \Rightarrow N(0, 1)$$

as  $L/n \rightarrow 0$ .

The CLT in (18) holds for all  $\rho$  and for any  $L$  as long as  $L/n \rightarrow 0$ , and in that sense (18) presents a unified asymptotic result for the estimator  $\hat{\rho}_{pa}$ .

### Remarks.

1. The expression simplifies considerably when  $L \rightarrow \infty$ . If  $\rho$  is fixed and  $|\rho| < 1$ , then  $\bar{S}_L(\rho) \rightarrow 1/(1-\rho)$ ,  $S_{L-\ell}/L \rightarrow 0$  and  $V_L(\rho) \rightarrow 1/2$  by dominated convergence, leading to

$$n^{1/2}(\hat{\rho}_{pa} - \rho) \Rightarrow N(0, 1 - \rho^2), \quad |\rho| < 1, \quad \text{as } \frac{1}{L} + \frac{L}{n} \rightarrow 0.$$

If  $\rho = 1$ , then  $S_\ell = \ell$ ,  $L^{-1}\bar{S}_L \rightarrow 1/2$  and  $V_L \rightarrow \int_0^1 2(1-x) \cdot 2x dx = 2/3$ . In this case

$$(19) \quad (nL)^{1/2}(\hat{\rho}_{pa} - 1) \Rightarrow N(0, 16/3), \quad \rho = 1, \quad \text{as } \frac{1}{L} + \frac{L}{n} \rightarrow 0,$$

and the PAE estimator is  $\sqrt{nL}$  convergent to unity with an asymptotic normal distribution.

2. By establishing local asymptotics, we can verify that the asymptotic distributions are continuous at  $\rho = 1$ . To do so, let  $\rho \nearrow 1$  in such a way that  $1 - \rho = c/L$  with  $0 \leq c < \infty$ . This formulation corresponds to the usual local to unity framework with localizing coefficient  $c$ , but here uses a region around unity measured in units of  $1/L$  rather than  $1/n$ . For the calculation, let  $\ell = [Lx]$ , the integer part of  $Lx$  for some  $x \in [0, 1]$ , and let  $L \rightarrow \infty$ . Then, for given  $x > 0$ , we have  $\rho^\ell \rightarrow e^{-cx}$  as  $L \rightarrow \infty$ , and  $L^{-1}S_\ell \rightarrow c^{-1}(1 - e^{-cx})$ . Furthermore,

$$L^{-1}\bar{S}_L = \frac{1}{L} \sum_{\ell=1}^L L^{-1}S_\ell \rightarrow \int_0^1 c^{-1}(1 - e^{-cx})dx = c^{-2}(c + e^{-c} - 1) =: h_0(c),$$

and

$$V_L \rightarrow \frac{c^2}{c + e^{-c} - 1} \int_0^1 \left[ 1 - x + \frac{1 - e^{-c(1-x)}}{c} \right] \left( \frac{1 - e^{-cx}}{c} \right) dx =: h_1(c).$$

So the convergence rate of the PAE  $\hat{\rho}_{pa}$  is  $(nL)^{1/2}$  and its limit distribution in this local to unity case is given by

$$(20) \quad (nL)^{1/2}(\hat{\rho}_{pa} - \rho) \Rightarrow N(0, 4h_1(c)/h_0(c)).$$

Note that  $1 + \rho \rightarrow 2$  in this case. The case where  $\rho \nearrow 1$  faster than  $O(L)$  corresponds to  $c = 0$ , and can be analyzed directly from the limit (20) using L'Hôpital's rule. More specifically, we have  $h_0(0) = 1/2$  and  $h_1(0) = 2 \int_0^1 2(1-x)xdx = 2/3$ , so that  $(nL)^{1/2}(\hat{\rho}_{pa} - \rho) \Rightarrow N(0, 16/3)$ , corresponding to (19) for the case where  $\rho = 1$ . If  $\rho$  reaches 1 at a slower rate than  $O(L)$  so  $L(1 - \rho) \rightarrow \infty$ , then  $\rho^\ell \rightarrow 0$  as  $L \rightarrow \infty$  along the sequence  $\ell = [Lx]$  for  $x > 0$ . Hence,  $(1 - \rho)S_\ell \rightarrow 1$  along all such paths for each  $x > 0$  as is easily calculated. Thus,  $(1 - \rho)\bar{S}_L \rightarrow 1$  as well, and  $V_L \rightarrow 1/2$ . In this event, we have the limit theory

$$n^{1/2}(1 - \rho^2)^{-1/2}(\hat{\rho}_{pa} - \rho) \Rightarrow N(0, 1),$$

which also includes the case of fixed  $\rho$ . It follows that local to unity asymptotics for the PAE estimator are continuous in both directions, with normal limit theory towards unity as well as towards fixed  $|\rho| < 1$  and stationarity. In this sense, the results are quite different from usual local to unity asymptotics (Phillips, 1987).

Figure 4 presents simulated densities of the rescaled and centered partial aggregation estimates from 10,000 replications for  $n = 400$  and  $\rho = 0, 0.5, 0.9, 1$  for various  $L$  values, where the rescaling factors are determined by (18) of Theorem 5. The figures show that the distributions are well approximated by standard normal asymptotics for all  $L$  values for  $\rho < 1$ . When  $\rho = 1$  (the lower right figure), the rescaled PAE is seen to be approximately normally distributed for  $L \leq 20 = n^{1/2}$  (the top three densities in this figure), but as  $L$  increases, the distribution departs further from normality and moves towards the distribution of the FAE. In all cases, the distributions are well centred at the true parameter. Note that a more dispersed distribution does not mean a larger variance because the scaling factor increases with  $L$ . On the contrary, when  $\rho = 1$ , the distribution of the PAE for larger  $L$  is a lot more condensed around the true parameter than it is for a smaller  $L$ . (See the scales in the figures.)



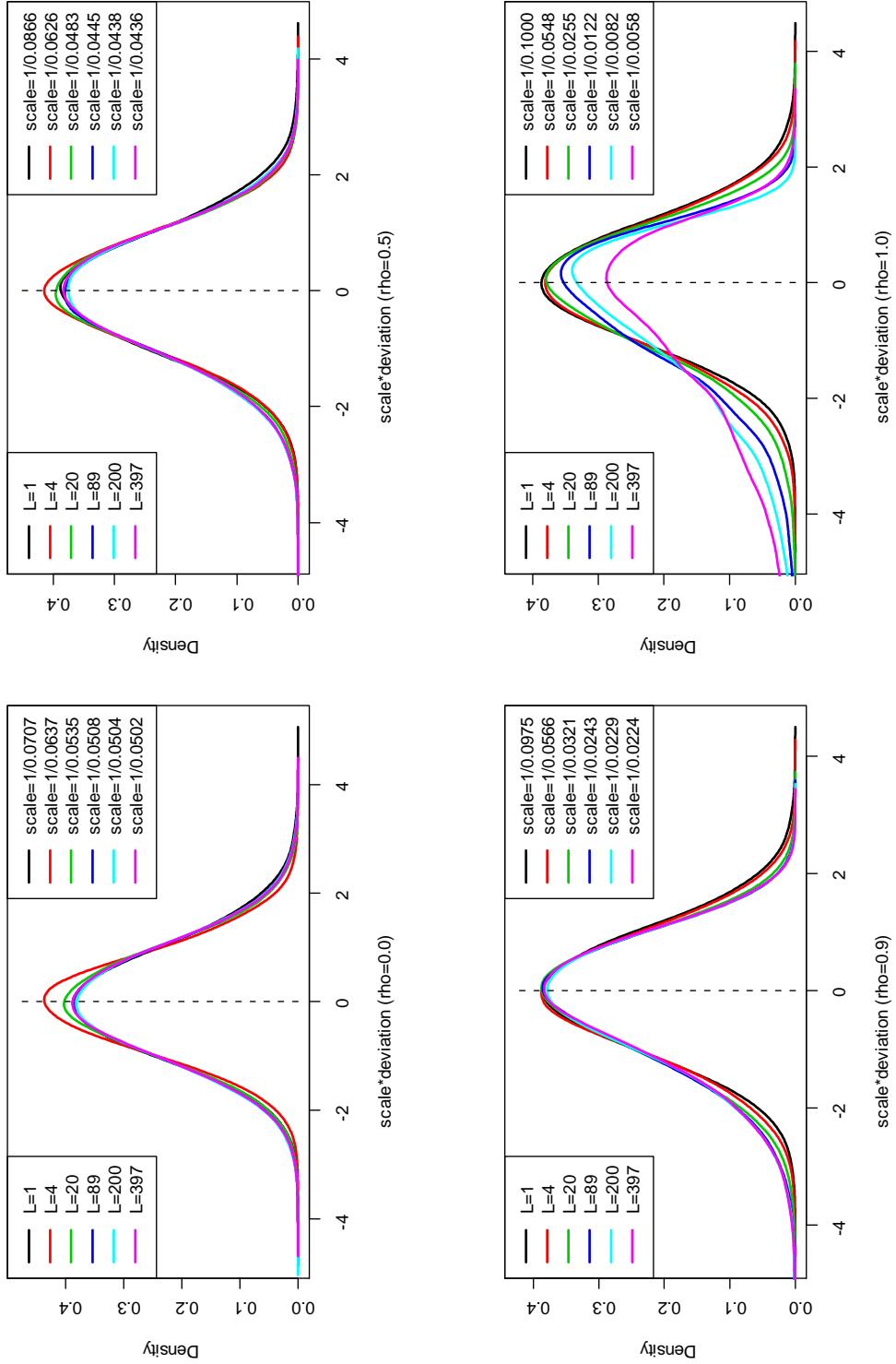


Figure 4: Densities of Partial Aggregation estimators (10,000 replications,  $n = 400$ )

## 6 Conclusion

This paper develops some new moment conditions involving short and long differences for autoregression. The moment conditions, which are asymptotically infinite in number, lead to unweighted GMM estimators which have some interesting and desirable properties. Partially aggregating  $L$  moment conditions and allowing  $L$  to pass to infinity at a slower rate than the sample size leads to an estimator with a limiting normal distribution that holds uniformly in the autoregressive coefficient  $\rho$  including stationary and unit root cases. The rate of convergence of this estimator is  $\sqrt{n}$  when  $|\rho| < 1$  and the limit distribution is the same as the Gaussian MLE, but when  $\rho = 1$  the rate of convergence to the normal distribution is  $\sqrt{nL}$  and can be within a slowly varying factor of  $n$ . This result shows that uniform asymptotic normality in autoregression is possible with a small sacrifice in the rate of convergence at  $\rho = 1$ .

The fully aggregated estimator, which uses all of the moment conditions, has the same limiting normal distribution in the stationary case as the Gaussian MLE and nonstandard limit distributions in the unit root and local to unity cases which differ from those of the MLE. This estimator has less bias than the MLE and achieves a reduction in variance, showing that improvements on the MLE are possible without using "superefficient" methods and that these gains hold across a range of  $\rho$  values in the vicinity of unity and persist in the limit distributions as  $n \rightarrow \infty$ . So, there are uniform asymptotic gains in efficiency in the neighborhood of unity.

The ideas and results presented here are primarily of theoretical interest because of the simplicity of the dynamic model (1). The practical importance of the work lies in extensions of the ideas and methods of the current paper to models that are of greater interest in applications but where related problems arise. In particular, the problems involved in fitting an intercept in dynamic regressions that are considered here are well known to be exacerbated in dynamic panel regressions because of the presence of potentially large numbers of fixed effects. A companion paper (Han, Phillips and Sul, 2009, explores those problems from the perspective of the moment conditions given in the current paper and develops new procedures of estimation based on similar techniques of aggregating moment conditions that allow for higher order dynamics and the presence of fixed effects in panel models.

## A Proofs

### A.1 Subsidiary Lemmas

Throughout this section, we assume that  $0 \leq \rho \leq 1$  to simplify arguments, although the main results continue to hold for  $-1 < \rho \leq 1$  as discussed below. The following quantities will be used frequently:

$$(21) \quad S_k(\rho) = \sum_{j=1}^k \rho^{j-1}, \quad \bar{S}_k(\rho) = \frac{1}{k} \sum_{j=1}^k S_j(\rho).$$

Note that  $S_k(\rho)$  appears in the case of estimation with a single lag and  $\bar{S}_k(\cdot)$  is a recursive average of  $S_j(\cdot)$  over  $j = 1, \dots, k$ . The quantities  $S_k(\rho)$  and  $\bar{S}_k(\rho)$  are useful for handling the stationary

and unit root cases in a continuous way. If  $\rho < 0$ , we can let  $S_k(\rho) \equiv 1$  and  $\bar{S}_k(\rho) \equiv 1$ , then all the necessary analysis can be done with no significant changes.

For all  $\rho \leq 1$ , we have the following facts.

**Lemma A.1** (i)  $(1 - \rho^k)S_m = (1 - \rho^m)S_k$ ; (ii)  $S_k(\rho^2) \leq S_k(\rho)$ ; (iii)  $\bar{S}_k(\rho) \leq S_k(\rho) \leq 2\bar{S}_k(\rho)$ .

**Proof.** (i) Obvious. (ii) Obvious because  $\rho^2 \leq \rho$  when  $\rho \geq 0$ . (iii) Because  $S_j(\rho)$  is increasing in  $j$ , we have  $\bar{S}_k(\rho) = k^{-1} \sum_{j=1}^k S_j(\rho) \leq S_k(\rho)$ . Next, because  $S_j(\rho) + S_{k-j+1}(\rho) \geq S_k(\rho)$ , we have

$$2\bar{S}_k(\rho) = \frac{1}{k} \sum_{j=1}^k [S_j(\rho) + S_{k-j+1}(\rho)] \geq \frac{1}{k} \sum_{j=1}^k S_k(\rho) = S_k(\rho)$$

as claimed. ■

We also use the following fact.

**Lemma A.2** Let  $j \leq k$ . Then

$$E(x_t - x_{t-j})(x_t - x_{t-k}) = \frac{\sigma^2(1 + \rho^{k-j})S_j(\rho)}{1 + \rho}.$$

**Proof.** Straightforward for both  $|\rho| < 1$  and  $\rho = 1$ . ■

We are now prove Lemma 1.

**Proof of Lemma 1.** By definition  $\varepsilon_s^* = x_s - \rho x_{s+1}$  and so the moment conditions

$$E(x_{t-1} - x_{s+1})[(x_t - x_s) - \rho(x_{t-1} - x_{s+1})] = 0$$

are equivalent to  $E(x_{t-1} - x_{s+1})(\varepsilon_t - \varepsilon_s^*) = 0$ . These conditions hold trivially because  $E(x_{t-1} - x_{s+1})\varepsilon_t = 0$  and  $E(x_{t-1} - x_{s+1})\varepsilon_s^* = 0$  by direct calculation for both  $\rho = 1$  and  $|\rho| < 1$ . ■

## A.2 Fixed Lag Estimation

First note that

$$(22) \quad x_{t-1} - x_{t-1-\ell} = \sum_{j=0}^{\ell-1} \rho^j \varepsilon_{t-1-j} - (1 - \rho^\ell)x_{t-1-\ell}.$$

We consider the numerator and denominator of the fixed lag estimator

$$\tilde{\rho}_\ell = \rho + \frac{\sum_{t=3+\ell}^n (x_{t-1} - x_{t-1-\ell})(\varepsilon_t - \varepsilon_{t-2-\ell}^*)}{\sum_{t=3+\ell}^n (x_{t-1} - x_{t-1-\ell})^2}$$

given in (9).

*Denominator:* Let  $C_n = \sum_{t=3+\ell}^n (x_{t-1} - x_{t-1-\ell})^2$ . We will establish a law of large numbers (LLN) for  $n_\ell^{-1}C_n$  where  $n_\ell = n - 2 - \ell$ . Because  $x_s = \sum_{j=0}^{s-1} \rho^j \varepsilon_{s-j} + \rho^s x_0$ , (22) implies that

$$x_{t-1} - x_{t-1-\ell} = \sum_{j=0}^{\ell-1} \rho^j \varepsilon_{t-1-j} - (1 - \rho^\ell) \sum_{j=\ell}^{t-2} \rho^{j-\ell} \varepsilon_{t-1-j} - (1 - \rho^\ell) \rho^{t-1-\ell} x_0.$$

We can write the first two terms of this expression as  $\sum_{j=0}^{\infty} c_j \varepsilon_{t-1-j}$ . Then, whether  $\rho < 1$  or  $\rho = 1$ , the  $c_j$ 's satisfy conditions for a LLN for  $n_\ell^{-1} \sum_{t=3+\ell}^n (x_{t-1} - x_{t-1-\ell})^2$  by means of the Phillips and Solo (1992) device. For the limit, we have

$$(23) \quad n_\ell^{-1} E(C_n) = \frac{1}{n_\ell} \sum_{t=3+\ell}^n \left( \frac{2\sigma^2}{1+\rho} \right) S_\ell(\rho) = \left( \frac{2\sigma^2}{1+\rho} \right) S_\ell(\rho),$$

by Lemma A.2. This is also the probability limit of  $n_\ell^{-1} C_n$ .

*Numerator:* Let  $D_n = \sum_{t=3+\ell}^n (x_{t-1} - x_{t-1-\ell})(\varepsilon_t - \varepsilon_{t-2-\ell}^*)$ . Because  $\varepsilon_{t-2-\ell}^* = x_{t-2-\ell} - \rho x_{t-1-\ell}$ , we have

$$\begin{aligned} D_n &= \sum_{t=3+\ell}^n (x_{t-1} - x_{t-1-\ell})\varepsilon_t - \sum_{t=2+\ell}^{n-1} (x_t - x_{t-\ell})x_{t-1-\ell} + \rho \sum_{t=3+\ell}^n (x_{t-1} - x_{t-1-\ell})x_{t-1-\ell} \\ &= \sum_{t=3+\ell}^n (x_{t-1} - x_{t-1-\ell})\varepsilon_t - \sum_{t=3+\ell}^n [(x_t - x_{t-\ell}) - \rho(x_{t-1} - x_{t-1-\ell})]x_{t-1-\ell} + \eta_n, \end{aligned}$$

where  $\eta_n = (x_n - x_{n-\ell})x_{n-1-\ell} - (x_{2+\ell} - x_2)x_1$ . Because  $x_t - \rho x_{t-1} = \varepsilon_t$ , we have

$$\begin{aligned} D_n &= \sum_{t=3+\ell}^n (x_{t-1} - x_{t-1-\ell})\varepsilon_t - \sum_{t=3+\ell}^n x_{t-1-\ell}(\varepsilon_t - \varepsilon_{t-\ell}) + \eta_n \\ &= 2 \sum_{t=3+\ell}^n (x_{t-1} - x_{t-1-\ell})\varepsilon_t + R_n, \quad R_n = \sum_{t=3}^{2+\ell} x_{t-1}\varepsilon_t - \sum_{t=n-\ell+1}^n x_{t-1}\varepsilon_t + \eta_n. \end{aligned}$$

Upon suitable normalization, the first term ( $\tilde{D}_n$ , say) should follow a martingale difference central limit theorem (CLT) and  $R_n$  is negligible. Each summand of  $\tilde{D}_n$  has variance  $8\sigma^4 S_\ell(\rho)/(1+\rho)$ , so the Lindeberg condition (see Theorem 2.3 of McLeish, 1974) holds, and

$$n_\ell^{-1/2} \tilde{D}_n \Rightarrow N \left( 0, \frac{8\sigma^4 S_\ell(\rho)}{1+\rho} \right).$$

Next, we have

$$\begin{aligned} R_n &= \sum_{t=3}^{2+\ell} x_{t-1}\varepsilon_t - \sum_{t=n-\ell+1}^n x_{t-1}\varepsilon_t - (x_{2+\ell} - x_2)x_1 + (x_n - x_{n-\ell})x_{n-1-\ell} \\ &= \left[ (x_n - x_{n-\ell})x_{n-1-\ell} - \sum_{j=0}^{\ell-1} x_{n-1-j}\varepsilon_{n-j} \right] - \left[ (x_{2+\ell} - x_2)x_1 - \sum_{j=0}^{\ell-1} x_{2+j}\varepsilon_{3+j} \right] \\ &= \Psi_n - \Psi_{\ell+2}. \end{aligned}$$

Upon further manipulation of  $\Psi_k$  for general  $k$ , we have

$$\begin{aligned} \Psi_k &= (x_k - x_{k-\ell})x_{k-1-\ell} - \sum_{j=0}^{\ell-1} x_{k-1-j}\varepsilon_{k-j} \\ &= \left( \sum_{j=0}^{\ell-1} \rho^j \varepsilon_{k-j} + \rho^\ell x_{k-\ell} - x_{k-\ell} \right) x_{k-1-\ell} - \sum_{j=0}^{\ell-1} x_{k-1-j}\varepsilon_{k-j} \end{aligned}$$

$$\begin{aligned}
&= \sum_{j=0}^{\ell-1} \varepsilon_{k-j} (\rho^j x_{k-1-\ell} - x_{k-1-j}) - (1 - \rho^\ell) x_{k-\ell} x_{k-1-\ell} \\
&= \sum_{j=0}^{\ell-1} \varepsilon_{k-j} (\rho^j x_{k-1-\ell} - x_{k-1-j}) - (1 - \rho^\ell) \varepsilon_{k-\ell} x_{k-1-\ell} - \rho(1 - \rho^\ell) x_{k-1-\ell}^2 \\
&= \Psi_{k,a} - \Psi_{k,b} - \Psi_{k,c}.
\end{aligned}$$

First  $E\Psi_{k,a}^2 = \sigma^2 \sum_{j=0}^{\ell-1} E(\rho^j x_{k-1-\ell} - x_{k-1-j})^2$ . We have the following fact.

**Lemma A.3** *Let  $\omega_0 = (1 - \rho)Ex_0^2$ . Then, for all  $k$ , we have*

$$E(x_{k-j} - \rho^j x_{k-\ell})^2 \leq 4 \left[ \omega_0 + \frac{4\sigma^2}{1 + \rho} \right] S_\ell(\rho).$$

**Proof.** Using  $x_{k-j} - \rho^j x_{k-\ell} = (1 - \rho^j)x_0 + (1 - \rho^j)(x_{k-j} - x_0) + \rho^j(x_{k-j} - x_{k-\ell})$ , we have

$$\begin{aligned}
\frac{1}{4}E(x_{k-j} - \rho^j x_{k-\ell})^2 &\leq (1 - \rho^j)^2 Ex_0^2 + (1 - \rho^j)^2 E(x_{k-j} - x_0)^2 + \rho^{2j} E(x_{k-j} - x_{k-\ell})^2 \\
&= (1 - \rho^j)^2 Ex_0^2 + \sigma^2 (1 - \rho^j)^2 (1 + \rho^{k-j})(1 + \rho)^{-1} S_{k-j}(\rho) \\
&\quad + \sigma^2 \rho^{2j} (1 + \rho^{\ell-j})(1 + \rho)^{-1} S_{\ell-j}(\rho) \\
&\leq (1 - \rho^j) \omega_0 S_j + 2\sigma^2 \rho_*^{-1} (1 - \rho^j) S_j + 2\sigma^2 \rho^{2j} \rho_*^{-1} S_\ell
\end{aligned}$$

where  $\rho_* = 1 + \rho$  and  $S_j = S_j(\rho)$  for notation brevity. In the above display, the equality holds by Lemma A.2, and the final inequality holds due to Lemma A.1. The result follows because  $S_j \leq S_\ell$  for  $j \leq \ell$ . ■

Lemma A.3 implies that  $E\Psi_{k,a}^2$  is uniformly bounded, so that  $\Psi_{\ell+2,a} = O_p(1)$  and  $\Psi_{n,a} = O_p(1)$ . Next, it is simple to show that  $E\Psi_{k,b}^2$  is also uniformly bounded. To see this, note that

$$E\Psi_{k,b}^2 = (1 - \rho^\ell)^2 \sigma^2 Ex_{k-1-\ell}^2.$$

Since  $x_t = \sum_{j=0}^{t-1} \rho^j \varepsilon_{t-j} + \rho^t x_0$ , we have  $Ex_t^2 = \sigma^2 S_t + \rho^{2t} Ex_0^2$ , and so

$$E\Psi_{k,b}^2 = \sigma^2 (1 - \rho^\ell)^2 [\sigma^2 S_{k-1-\ell} + \rho^{2(k-1-\ell)} Ex_0^2] \leq \sigma^2 (\sigma^2 + \omega_0) S_\ell.$$

Finally,  $E\Psi_{k,c} = \rho(1 - \rho^\ell) [\sigma^2 S_{k-1-\ell}(\rho^2) + \rho^{2(k-1-\ell)} Ex_0^2]$ , which is also bounded by  $(\sigma^2 + \omega_0) S_\ell$ . The fact that the first or second moments of  $\Psi_{k,a}$ ,  $\Psi_{k,b}$  and  $\Psi_{k,c}$  are uniformly bounded implies that  $\Psi_n = O_p(1)$  and  $\Psi_{\ell+2} = O_p(1)$ , so that  $R_n = O_p(1)$ . Hence,  $n_\ell^{-1/2} R_n \rightarrow_p 0$ . Thus the limit distribution of  $n_\ell^{-1/2} D_n$  is the same as that of  $n_\ell^{-1/2} \tilde{D}_n$ , and we have

$$(24) \quad n_\ell^{-1/2} D_n \Rightarrow N \left( 0, \frac{8\sigma^4 S_\ell(\rho)}{1 + \rho} \right).$$

**Proof of Theorem 2.** Using (23) and (24), we get

$$n_\ell^{1/2} (\tilde{\rho}_\ell - \rho) \Rightarrow N \left( 0, \frac{2(1 + \rho)}{S_\ell(\rho)} \right),$$

where  $n_\ell = n - 2 - \ell$  as before. ■

### A.3 Full Aggregation

**Proof of Theorem 3.** The proof proceeds by rearranging the terms in the double summations. Let  $C_n = \sum_{\ell=1}^{n-3} \sum_{t=3+\ell}^n (x_{t-1} - x_{t-1-\ell})^2 = \sum_{t=4}^n \sum_{s=1}^{t-3} (x_{t-1} - x_{s+1})^2$ . Then

$$\begin{aligned} C_n &= \sum_{t=3}^{n-1} \sum_{s=2}^{t-1} (x_t - x_s)^2 = \sum_{t=3}^{n-1} \sum_{s=2}^{t-1} (x_t^2 + x_s^2) - 2 \sum_{t=3}^{n-1} \sum_{s=2}^{t-1} x_t x_s \\ &= (n_0 - 1) \sum_{t=2}^{n-1} x_t^2 - \left[ \left( \sum_{t=2}^{n-1} x_t \right)^2 - \sum_{t=2}^{n-1} x_t^2 \right] = n_0 \sum_{t=3}^n x_{t-1}^2 - \left( \sum_{t=3}^n x_{t-1} \right)^2 \\ &= n_0 \sum_{t=3}^n \tilde{x}_{t-1}^2, \quad \tilde{x}_{t-1} = x_{t-1} - \frac{1}{n_0} \sum_{s=3}^n x_{s-1}. \end{aligned}$$

Let  $G_n$  denote the numerator, i.e.,  $G_n = \sum_{\ell=1}^{n-3} \sum_{t=3+\ell}^n (x_{t-1} - x_{t-1-\ell})(x_t - x_{t-2-\ell})$ . Then

$$\begin{aligned} G_n &= \sum_{n=4}^n \sum_{s=1}^{t-3} (x_{t-1} - x_{s+1})(x_t - x_s) \\ &= \sum_{t=4}^n \sum_{s=1}^{t-3} (x_{t-1}x_t + x_sx_{s+1}) - \sum_{t=4}^n \sum_{s=1}^{t-3} (x_sx_{t-1} + x_{s+1}x_t) = G_{n1} - G_{n2}. \end{aligned}$$

By rearranging the terms, we get

$$G_{n1} = (n_0 - 2) \sum_{t=2}^n x_{t-1}x_t + x_1x_2 + x_{n-1}x_n.$$

For  $G_{n2}$ , we have

$$\begin{aligned} \sum_{t=4}^n \sum_{s=1}^{t-3} x_sx_{t-1} &= \frac{1}{2} \left( \sum_{t=1}^{n-1} x_t \right)^2 - \frac{1}{2} \sum_{t=1}^{n-1} x_t^2 - \sum_{t=2}^{n-1} x_{t-1}x_t, \\ \sum_{t=4}^n \sum_{s=1}^{t-3} x_{s+1}x_t &= \frac{1}{2} \left( \sum_{t=2}^n x_t \right)^2 - \frac{1}{2} \sum_{t=2}^n x_t^2 - \sum_{t=3}^n x_{t-1}x_t, \end{aligned}$$

so

$$G_{n2} = \left( \sum_{t=2}^{n-1} x_t \right)^2 + (x_1 + x_n) \sum_{t=2}^{n-1} x_t - \sum_{t=2}^{n-1} x_t^2 - 2 \sum_{t=2}^n x_{t-1}x_t + x_{n-1}x_n + x_1x_2.$$

Thus

$$G_n = G_{n1} - G_{n2} = n_0 \sum_{t=2}^n x_{t-1}x_t + \sum_{t=2}^{n-1} x_t^2 - \left( \sum_{t=2}^{n-1} x_t \right)^2 - (x_1 + x_n) \sum_{t=2}^{n-1} x_t.$$

Because  $\hat{\rho}_{fa} = G_n/C_n = \rho + (G_n - \rho C_n)/C_n$ , we now evaluate  $D_n := G_n - \rho C_n$ , which is

$$\begin{aligned} D_n &= n_0 \sum_{t=3}^n x_{t-1} \varepsilon_t - \left( \sum_{t=3}^n x_{t-1} \right) \sum_{t=3}^n \varepsilon_t + \sum_{t=3}^n x_{t-1}^2 + n_0 x_1 x_2 - (x_1 + x_2) \sum_{t=3}^n x_{t-1} \\ &= n_0 \sum_{t=3}^n \tilde{u}_{t-1} \tilde{\varepsilon}_t + \sum_{t=3}^n x_{t-1}^2 + n_0 x_1 x_2 - (x_1 + x_2) \sum_{t=3}^n x_{t-1}. \end{aligned}$$

The stated result now follows straightforwardly because

$$\hat{\rho}_{fa} = \rho + \frac{n_0^{-1} D_n}{n_0^{-1} C_n} = \rho + \frac{\sum_{t=3}^n \tilde{x}_{t-1} \tilde{\varepsilon}_t}{\sum_{t=3}^n \tilde{x}_{t-1}^2} + \frac{n_0^{-1} \sum_{t=3}^n x_{t-1}^2}{\sum_{t=3}^n \tilde{x}_{t-1}^2} + \eta_n,$$

where  $\eta_n = x_1 x_2 - n_0^{-1} (x_1 + x_2) \sum_{t=3}^n x_{t-1} / \sum_{t=3}^n \tilde{x}_{t-1}^2$ . Note that the sum of the first two terms on the right hand side of the above displayed equation is  $\hat{\rho}_{ols}$ . ■

**Proof of Corollary 4.** Lemma 3 and standard weak convergence arguments give the stated results in a straightforward manner. ■

**Remark.** When  $\rho = 1$ , the limit distribution of the full aggregation estimator can be expressed as the following ratio

$$(25) \quad n(\hat{\rho}_{fa} - 1) \Rightarrow \frac{\int_0^1 \int_0^r (1+p-r) dB_p dB_r}{\int_0^1 \int_0^r (B_r - B_s)^2 ds dr} := \frac{Y}{X}.$$

We can show that this expression is equivalent to (14), i.e.,  $X = \int_0^1 \tilde{B}_r^2 dr$  and  $Y = \int_0^1 \tilde{B}_r dB_r + \int_0^1 B_r^2 dr$ . First, we have  $X = \int_0^1 \int_0^r (B_r - B_s)^2 ds dr = \int_0^1 \int_r^1 (B_r - B_s)^2 ds dr$ , implying that

$$\begin{aligned} X &= \frac{1}{2} \int_0^1 \int_0^1 (B_r - B_p)^2 dp dr = \frac{1}{2} \int_0^1 \int_0^1 (B_r^2 - 2B_r B_s + B_s^2) dp dr \\ &= \int_0^1 B_r^2 dr - \left( \int_0^1 B_r dr \right)^2 = \int_0^1 \tilde{B}_r^2 dr. \end{aligned}$$

For the numerator, we have  $Y = \int_0^1 [(1-r)B_r + \int_0^r p dB_p] dB_r$ . But  $\int_0^r p dB_p = rB_r - \int_0^r B_p dp$  using integration by parts which is valid by smoothness of the integrand. Hence,

$$Y = \int_0^1 B_r dB_r - \int_0^1 \left( \int_0^r B_p dp \right) dB_r = \int_0^1 B_r dB_r - \int_0^1 B_r^* dB_r, \quad B_r^* = \int_0^r B_p dp.$$

Because  $B_r^*$  is differentiable, we again use integration by parts giving  $\int_0^1 B_r^* dB_r = B_1^* B_1 - \int_0^1 B_r^2 dr$ . Collecting terms gives

$$Y = \int_0^1 B_r dB_r - B_1 \int_0^1 B_r dr + \int_0^1 B_r^2 dr = \int_0^1 \tilde{B}_r dB_r + \int_0^1 B_r^2 dr,$$

as in the numerator of (14). ■

## A.4 Partial Aggregation

Results from the full aggregation case are also useful, especially the following.

**Lemma A.4** *Let  $C_n^L = \sum_{\ell=1}^L \sum_{t=3+\ell}^n (x_{t-1} - x_{t-1-\ell})^2$  and  $D_n^L = \sum_{\ell=1}^L \sum_{t=3+\ell}^n (x_{t-1} - x_{t-1-\ell})(\varepsilon_t - \varepsilon_{t-2-\ell}^*)$ . Then  $C_n^{n-3} = O_p(n^2 \bar{S}_n)$  and  $D_n^{n-3} = O_p(n \sqrt{n \bar{S}_n})$ .*

**Proof.** Straightforward from the proof of Theorem 3. ■

We first derive the probability limit of the denominator of the partially aggregated estimator and then consider the limit behavior of the numerator.

### (i) Denominator

Let  $C_n^L = \sum_{\ell=1}^L \sum_{t=3+\ell}^n (x_{t-1} - x_{t-1-\ell})^2$ , which is the denominator of the partial aggregation estimator. We begin with the following simple fact.

**Lemma A.5** *Let  $S_\ell = S_\ell(\rho)$ . Then*

$$(n_0 \bar{S}_L L)^{-1} EC_n^L = \frac{2\sigma^2}{1+\rho} \left[ 1 - \left( \frac{L}{n_0} \right) \frac{1}{L} \sum_{\ell=1}^L \frac{\ell S_\ell}{L S_L} \right], \quad n_0 = n - 2.$$

**Proof.** Use (23). ■

Because  $\ell S_\ell \leq L S_L$ , we have  $(n \bar{S}_L L)^{-1} EC_n^L \rightarrow 2\sigma^2/(1+\rho)$  when  $L/n \rightarrow 0$ . When  $L$  is fixed, therefore, we may proceed as in the fixed lag case to obtain

$$(26) \quad (n \bar{S}_L L)^{-1} C_n^L \rightarrow_p 2\sigma^2/(1+\rho), \quad \rho \in (-1, 1],$$

as  $n \rightarrow \infty$ .

To handle the case with  $L \rightarrow \infty$ , write

$$(27) \quad C_n^L = C_{L+2}^{L-1} + \tilde{C}_n^L, \quad \tilde{C}_n^L = \sum_{\ell=1}^L \sum_{t=L+3}^n (x_{t-1} - x_{t-1-\ell})^2.$$

By Lemma A.4, we have  $(L^2 \bar{S}_L)^{-1} C_{L+2}^{L-1} = O_p(1)$  and, so  $(n \bar{S}_L L)^{-1} C_{L+2}^{L-1} = O_p(L/n) = o_p(1)$  for all  $\rho$ . Also,  $(n^2 \bar{S}_n)^{-1} E(C_{L+2}^{L-1}) = o(1)$ . Hence, the term involving  $C_{L+2}^{L-1}$  is negligible provided  $L/n \rightarrow 0$ . We therefore focus on  $\tilde{C}_n^L$ .

Using  $x_t = \sum_{j=0}^{t-1} \rho^j \varepsilon_{t-j} + \rho^t x_0$ , we get

$$(28) \quad \begin{aligned} x_{t-1} - x_{t-1-\ell} &= \sum_{j=0}^{t-2} \rho^j \varepsilon_{t-1-j} + \rho^{t-1} x_0 - \sum_{j=0}^{t-2-\ell} \rho^j \varepsilon_{t-1-\ell-j} - \rho^{t-1-\ell} x_0 \\ &= \sum_{j=0}^{\ell-1} \rho^j \varepsilon_{t-1-j} - (1-\rho^\ell) \sum_{j=\ell}^{t-2} \rho^{j-\ell} \varepsilon_{t-1-j} - \rho^{t-1-\ell} (1-\rho^\ell) x_0 \\ &= \sum_{j=0}^{t-2} c_j^\ell \varepsilon_{t-1-j} - \rho^{t-1-\ell} (1-\rho^\ell) x_0, \end{aligned}$$



where  $c_j^\ell = \rho^j_{\{j < \ell\}} - (1 - \rho^\ell)\rho^{j-\ell}_{\{j \geq \ell\}}$ . So

$$(29) \quad \begin{aligned} \tilde{C}_n^L &= \sum_{\ell \times t} \left( \sum_{j=0}^{t-2} c_j^\ell \varepsilon_{t-1-j} \right)^2 - 2 \sum_{\ell \times t} \left( \sum_{j=0}^{t-2} c_j^\ell \varepsilon_{t-1-j} \right) \rho^{t-1-\ell} (1 - \rho^\ell) x_0 \\ &+ \sum_{\ell \times t} \rho^{2(t-1-\ell)} (1 - \rho^\ell)^2 x_0^2 = \tilde{C}_{n,a}^L - 2\tilde{C}_{n,b}^L + \tilde{C}_{n,c}^L, \end{aligned}$$

where  $\sum_{\ell \times t}$  denotes  $\sum_{\ell=1}^L \sum_{t=L+3}^n$ . As indicated above and in Lemma A.5, the standardization we will be using for  $\tilde{C}_n^L$  is given by  $n\bar{S}_L L$ , so terms in (29) are assessed against this standardization.

The  $\tilde{C}_{n,c}^L$  term can be handled by the following lemma.

**Lemma A.6**  $\sum_{\ell=1}^L \sum_{t=L+3}^n \rho^{2(t-1-\ell)} (1 - \rho^\ell)^2 \leq \rho(1 - \rho^L) S_L^2$ .

**Proof.** It is obvious if  $\rho = 1$ . If  $\rho < 1$ , then

$$\begin{aligned} \sum_{\ell=1}^L \sum_{t=L+3}^n \rho^{2(t-1-\ell)} (1 - \rho^\ell)^2 &\leq \sum_{\ell=1}^L \sum_{t=L+3}^n \rho^{2(t-L-4)+2(L-\ell)+6} (1 - \rho^L)^2 \\ &\leq \rho^6 S_n(\rho^2) S_L(\rho^2) (1 - \rho^L)^2 \leq \rho^6 (1 - \rho^L) S_L^2 \end{aligned}$$

due to Lemma A.1, where  $S_L = S_L(\rho)$ . ■

Lemma A.6 implies that  $\tilde{C}_{n,c}^L \leq (1 - \rho^L) S_L^2 x_0^2 = (1 - \rho) x_0^2 S_L^3$ . Thus, when  $(1 - \rho) x_0^2 = O_p(1)$ , which holds by the assumptions on  $x_0$ , we have  $(n\bar{S}_L L)^{-1} \tilde{C}_{n,c}^L = (S_L/n)(S_L/L) O_p(1)$ , which converges in probability to zero when  $L/n \rightarrow 0$  because  $S_L \leq L$ . (Note  $S_L$  and  $\bar{S}_L$  have the same order by Lemma A.1.) And  $\tilde{C}_{n,b}^L$ , divided by  $n\bar{S}_L L$ , can also be ignored by Hölder's inequality. So it remains to establish a LLN for  $(n\bar{S}_L L)^{-1} \tilde{C}_{n,a}^L$ . The algebra for this is mechanical and a little tedious.

We have  $(\sum_{j=0}^{t-2} c_j^\ell \varepsilon_{t-1-j})^2 = \sum_{j=0}^{t-2} (c_j^\ell)^2 \varepsilon_{t-1-j}^2 + 2 \sum_{j=1}^{t-2} \sum_{k=0}^{j-1} c_j^\ell c_k^\ell \varepsilon_{t-1-j} \varepsilon_{t-1-k}$ , thus

$$(30) \quad \begin{aligned} \tilde{C}_{n,a}^L &= \sum_{t=L+3}^n \sum_{j=0}^{t-2} \left[ \sum_{\ell=1}^L (c_j^\ell)^2 \right] \varepsilon_{t-1-j}^2 + 2 \sum_{t=L+3}^n \sum_{j=0}^{t-3} \sum_{k=j+1}^{t-2} \left[ \sum_{\ell=1}^L c_j^\ell c_k^\ell \right] \varepsilon_{t-1-j} \varepsilon_{t-1-k} \\ &:= \tilde{C}_{n,a1}^L + 2\tilde{C}_{n,a2}^L. \end{aligned}$$

We will show that, upon standardization,  $\tilde{C}_{n,a2}^L$  is negligible and  $\tilde{C}_{n,a1}^L$  satisfies an LLN. From the functional form of  $c_j^\ell$ , we have

$$\begin{aligned} (c_j^\ell)^2 &= \rho^{2j}_{\{j < \ell\}} + (1 - \rho^\ell)^2 \rho^{2(j-\ell)}_{\{j \geq \ell\}}, \\ c_j^\ell c_k^\ell &= \rho^{j+k}_{\{k < \ell\}} - (1 - \rho^\ell) \rho^{j+k-\ell}_{\{j < \ell \leq k\}} + (1 - \rho^\ell)^2 \rho^{j+k-2\ell}_{\{j \geq \ell\}}, \end{aligned}$$

for  $j < k$ . So

$$(31) \quad \ddot{c}_{jj} := \sum_{\ell=1}^L (c_j^\ell)^2 = \sum_{\ell=1}^{j \wedge L} (1 - \rho^\ell)^2 \rho^{2(j-\ell)} + \sum_{\ell=j+1}^L \rho^{2j},$$

$$(32) \quad \ddot{c}_{jk} := \sum_{\ell=1}^L c_j^\ell c_k^\ell = \sum_{\ell=1}^{j \wedge L} (1 - \rho^\ell)^2 \rho^{j+k-2\ell} - \sum_{\ell=j+1}^{k \wedge L} (1 - \rho^\ell) \rho^{j+k-\ell} + \sum_{\ell=k+1}^L \rho^{j+k}.$$

Now we can rearrange the terms for  $\tilde{C}_{n,a1}^L$  to get

$$\tilde{C}_{n,a1}^L = \sum_{t=L+3}^n \sum_{j=0}^{t-2} \ddot{c}_{jj} \varepsilon_{t-1-j}^2 = \sum_{t=L+3}^{n-1} \left( \sum_{j=\max\{L+2-t,0\}}^{n-1-t} \ddot{c}_{jj} \right) \varepsilon_t^2 = \sum_{t=L+3}^{n-1} c_{tj}^* \varepsilon_t^2,$$

where  $c_{tj}^* \leq \sum_{\ell=0}^{n-1} \ddot{c}_{jj}$ . To set a boundary for  $c_{tj}^*$ , note that

$$\begin{aligned} \ddot{c}_{jj} &\leq (1 - \rho^L)^2 \sum_{\ell=1}^{\min(j,L)} \rho^{2(j-\ell)} + \max(L - j, 0) \rho^{2j} \\ &= \max(L - j, 0) \rho^{2j} + (1 - \rho^L)^2 S_{\min(j,L)}(\rho^2) \rho^{2 \max(j-L,0)}, \end{aligned}$$

so by Lemma A.1,

$$\sum_{t=0}^{n-1} \ddot{c}_{jj} \leq LS_L(\rho^2) + (1 - \rho^L)^2 S_L(\rho^2) S_n(\rho^2) \leq LS_L(\rho) + (1 - \rho^L) S_L(\rho)^2 (1 - \rho^n),$$

which is bounded by  $4L\bar{S}_L$  because  $1 - \rho^L \leq 1$ ,  $1 - \rho^n \leq 1$ ,  $S_L \leq L$  and  $S_L \leq 2\bar{S}_L$ . Thus, we have  $(n\bar{S}_L L)^{-1} \tilde{C}_{n,a1}^L = n^{-1} \sum_{t=L+3}^{n-1} [(L\bar{S}_L)^{-1} c_{tj}^*] \varepsilon_t^2$ , where  $0 < (L\bar{S}_L)^{-1} c_{tj}^* \leq 4$ . Since  $[(L\bar{S}_L)^{-1} c_{tj}^*] \varepsilon_t^2$  is uniformly integrable and  $E \{[(L\bar{S}_L)^{-1} c_{tj}^*] (\varepsilon_t^2 - \sigma^2)\} = 0$ ,  $\{[(L\bar{S}_L)^{-1} c_{tj}^*] (\varepsilon_t^2 - \sigma^2)\}$  is a martingale difference array and satisfies an LLN for  $L_1$  mixingales (e.g. Andrews, 1988) so that  $n^{-1} \sum_{t=L+3}^{n-1} [(L\bar{S}_L)^{-1} c_{tj}^*] (\varepsilon_t^2 - \sigma^2) \rightarrow_p 0$ .

The  $\tilde{C}_{n,a2}^L$  term is more complicated. Since

$$\tilde{C}_{n,a2}^L = \sum_{t=L+3}^n \sum_{j=0}^{t-3} \varepsilon_{t-1-j} \tilde{\varepsilon}_{t-2-j}, \quad \tilde{\varepsilon}_{t-2-j} = \sum_{k=j+1}^{t-2} \ddot{c}_{jk} \varepsilon_{t-1-k},$$

where  $\ddot{c}_{jk}$  is defined in (32), we have  $\text{var}(\tilde{C}_{n,a2}^L) = \sigma^2 \sum_{t=L+3}^n \sum_{j=0}^{t-3} E \tilde{\varepsilon}_{t-2-j}^2$ . Now  $E \tilde{\varepsilon}_{t-2-j}^2 = \sigma^2 \sum_{k=j+1}^{t-2} \ddot{c}_{jk}^2$ , so

$$(33) \quad \text{var}(\tilde{C}_{n,a2}^L) = \sigma^4 \sum_{t=L+3}^n \sum_{j=0}^{t-3} \sum_{k=j+1}^{t-2} \ddot{c}_{jk}^2 \leq \sigma^4 \sum_{t=L+3}^n \sum_{j=0}^{n-2} \sum_{k=j+1}^{n-1} \ddot{c}_{jk}^2 = n\sigma^4 \sum_{j=0}^{n-2} \sum_{k=j+1}^{n-1} \ddot{c}_{jk}^2.$$

By (32), we have, for  $j < k$ ,

$$|\ddot{c}_{jk}| \leq \begin{cases} L\rho^{j+k} & \text{if } k < L, \\ L(1 - \rho^L)\rho^{j+k-L} & \text{if } j < L \leq k, \\ (1 - \rho^L)^2 \rho^{j+k-2L} & \text{if } L \leq j, \end{cases}$$

so

$$\sum_{k=j+1}^{n-1} \ddot{c}_{jk}^2 \leq \begin{cases} \sum_{k=j+1}^{L-1} L^2 \rho^{2(j+k)} + \sum_{k=L}^{n-1} (1 - \rho^L)^2 \rho^{2(j+k-L)} & \text{if } j < L, \\ \sum_{k=j+1}^{n-1} (1 - \rho^L)^4 \rho^{2(j+k-2L)} & \text{if } j \geq L. \end{cases}$$

Thus,  $\sum_{k=j+1}^{n-1} \ddot{c}_{jk}^2 \leq L^2 S_L(\rho^2) \rho^{4j} \mathbb{1}_{\{j < L\}} + (1 - \rho^L)^4 S_n(\rho^2) \rho^{2(j-L)} \mathbb{1}_{\{j \geq L\}}$ , which is again bounded by  $L^2 S_L(\rho) \rho^{4j} \mathbb{1}_{\{j < L\}} + (1 - \rho^L)^4 S_n(\rho) \rho^{2(j-L)} \mathbb{1}_{\{j \geq L\}}$ . According to Lemma A.1, this last expression equals  $L^2 S_L(\rho) \rho^{4j} \mathbb{1}_{\{j < L\}} + (1 - \rho^L)^3 S_L(\rho) \rho^{2(j-L)} \mathbb{1}_{\{j \geq L\}}$ . So

$$\sum_{j=0}^{n-2} \left( \sum_{k=j+1}^{n-1} \ddot{c}_{jk}^2 \right) \leq L^2 S_L(\rho) S_L(\rho^4) + (1 - \rho^L)^3 S_L(\rho) S_n(\rho^2) \leq L^2 S_L(\rho)^2 + S_L(\rho)^2,$$

by Lemma A.1 again. This last result, Lemma A.1 and (33) imply that  $\text{var}(\ddot{C}_{n2})$  is  $O(nL^2 \bar{S}_L^2)$ , so  $\tilde{C}_{n,a2}^L$  is  $O_p(n^{1/2} L \bar{S}_L)$  because  $E \tilde{C}_{n,a2}^L = 0$ . Thus  $(n \bar{S}_L L)^{-1} \tilde{C}_{n,a2}^L = O_p(n^{-1/2})$ . Taken together, these results now imply the following result.

**Theorem A.7**  $(n \bar{S}_L L)^{-1} C_n^L \rightarrow_p 2\sigma^2 / (1 + \rho)$  as  $L/n \rightarrow 0$ .

**Proof.** Because of Lemma A.5, we have  $(n \bar{S}_L L)^{-1} C_{L+2}^{L-1} = o_p(1)$  for all  $\rho$  and  $(n^2 \bar{S}_n)^{-1} E(C_{L+2}^{L-1}) = o(1)$ . Regarding (27), we have shown that  $(n \bar{S}_L L)^{-1} C_n^L = (n \bar{S}_L L)^{-1} \tilde{C}_n^L + o_p(1)$  and  $(n \bar{S}_L L)^{-1} E(C_n^L) = (n \bar{S}_L L)^{-1} E(\tilde{C}_n^L) + o(1)$ . For (29), we have shown that  $(n \bar{S}_L L)^{-1} \tilde{C}_n^L = (n \bar{S}_L L)^{-1} \tilde{C}_{n,a}^L + o_p(1)$  and we have  $(n \bar{S}_L L)^{-1} E(\tilde{C}_n^L) = (n \bar{S}_L L)^{-1} E(\tilde{C}_{n,a}^L) + o(1)$ . We have also shown that  $(n \bar{S}_L L)^{-1} \tilde{C}_{n,a}^L = (n \bar{S}_L L)^{-1} \tilde{C}_{n,a1}^L + o_p(1)$  where  $\tilde{C}_{n,a1}^L$  is defined in (30), and we have  $(n \bar{S}_L L)^{-1} E(\tilde{C}_{n,a}^L) = (n \bar{S}_L L)^{-1} E(\tilde{C}_{n,a1}^L) + o(1)$ . It has also been shown that  $n^{-1} \sum_{t=L+3}^{n-1} [(L \bar{S}_L)^{-1} c_{tj}^*] (\varepsilon_t^2 - \sigma^2) \rightarrow_p 0$ , so that  $(n \bar{S}_L L)^{-1} C_n^L$  converges in probability to the limit of its expectation. According to Lemma A.5,  $(n \bar{S}_L L)^{-1} E C_n^L \rightarrow 2\sigma^2 / (1 + \rho)$  as  $L/n \rightarrow 0$  and  $n \rightarrow \infty$ , giving the stated probability limit. ■

The following results show that Theorem A.7 is consistent with our previous findings based on direct calculation.

**Corollary A.8** As  $L \rightarrow \infty$  and  $L/n \rightarrow 0$ , we have (i)  $(nL)^{-1} C_n^L \rightarrow_p 2\sigma^2 / (1 - \rho^2)$  if  $|\rho| < 1$ , and (ii)  $(nL^2)^{-1} C_n^L \rightarrow_p \sigma^2 / 2$  if  $\rho = 1$ .

**Proof.** (i)  $|\rho| < 1$ : We have  $\bar{S}_L \rightarrow (1 - \rho)^{-1}$  as  $L \rightarrow \infty$ . The result follows from Theorem A.7. (ii)  $\rho = 1$ : We have  $S_\ell = \ell$ , so  $\bar{S}_L \rightarrow 1/2$  as  $L \rightarrow \infty$ . See Theorem A.7 for the rest. ■

## (ii) Numerator

Let  $D_n^L = \sum_{\ell=1}^L \sum_{t=3+\ell}^n (x_{t-1} - x_{t-1-\ell})(\varepsilon_t - \varepsilon_{t-2-\ell}^*)$ . Then  $D_n^L = D_{L+2}^{L-1} + \tilde{D}_n^L$ , where  $\tilde{D}_n^L = \sum_{\ell=1}^L \sum_{t=L+3}^n (x_{t-1} - x_{t-1-\ell})(\varepsilon_t - \varepsilon_{t-2-\ell}^*)$ . By Lemma A.4, we have  $D_{L+2}^{L-1} = O_p(L \sqrt{L \bar{S}_L})$ , thus  $(n \bar{S})^{-1/2} L^{-1} D_{L+2}^{L-1} = O_p(\sqrt{L/n})$ , which can be ignored when  $L/n \rightarrow 0$ . Thus, we focus on  $\tilde{D}_n^L$ .

Let  $z_{t\ell} = (x_{t-1} - x_{t-1-\ell})(\varepsilon_t - \varepsilon_{t-2-\ell}^*)$ . We have

$$\begin{aligned} z_{t\ell} &= (x_{t-1} - x_{t-1-\ell})(\varepsilon_t - x_{t-2-\ell} + \rho x_{t-1-\ell}) \\ &= (x_{t-1} - x_{t-1-\ell})\varepsilon_t - x_{t-1}(x_{t-2-\ell} - \rho x_{t-1-\ell}) + x_{t-1-\ell}(x_{t-2-\ell} - \rho x_{t-1-\ell}) \\ &:= z_{t\ell,1} - z_{t\ell,2} + z_{t\ell,3}. \end{aligned}$$

The  $z_{t\ell,1}$  term is a martingale difference and straightforward. For  $z_{t\ell,2}$ , we have

$$\begin{aligned}\sum_{t=L+3}^n \sum_{\ell=1}^L z_{t\ell,2} &= \sum_{t=L+2}^{n-1} \sum_{\ell=1}^L x_t x_{t-1-\ell} - \rho \sum_{t=L+3}^n \sum_{\ell=1}^L x_{t-1} x_{t-1-\ell} \\ &= \sum_{t=L+3}^n \sum_{\ell=1}^L \varepsilon_t x_{t-1-\ell} + x_{L+2} \sum_{\ell=1}^L x_{L+1-\ell} - x_n \sum_{\ell=1}^L x_{n-1-\ell}.\end{aligned}$$

Similarly for  $z_{t\ell,3}$ , we have

$$\begin{aligned}\sum_{t=L+3}^n \sum_{\ell=1}^L z_{t\ell,3} &= \sum_{t=L+2}^{n-1} \sum_{\ell=1}^L x_{t-\ell} x_{t-1-\ell} - \rho \sum_{t=L+3}^n \sum_{\ell=1}^L x_{t-1}^2 \\ &= \sum_{t=L+3}^n \sum_{\ell=1}^L \varepsilon_{t-\ell} x_{t-1-\ell} + \sum_{\ell=1}^L x_{L+2-\ell} x_{L+1-\ell} - \sum_{\ell=1}^L x_{n-\ell} x_{n-1-\ell} \\ &= \left( L \sum_{t=L+3}^n \varepsilon_t x_{t-1} + \sum_{\ell=1}^L \sum_{j=0}^{\ell-1} \varepsilon_{L+2-j} x_{L+1-j} - \sum_{\ell=1}^L \sum_{j=0}^{\ell-1} \varepsilon_{n-j} x_{n-1-j} \right) \\ &\quad + \sum_{\ell=1}^L x_{L+2-\ell} x_{L+1-\ell} - \sum_{\ell=1}^L x_{n-\ell} x_{n-1-\ell}.\end{aligned}$$

So

$$\begin{aligned}L^{-1} \tilde{D}_n^L &= \sum_{t=L+3}^n \varepsilon_t \left( x_{t-1} - \frac{1}{L} \sum_{\ell=1}^L x_{t-1-\ell} \right) - \sum_{t=L+3}^n \varepsilon_t \left( \frac{1}{L} \sum_{\ell=1}^L x_{t-1-\ell} \right) \\ &\quad + \sum_{t=L+3}^n \varepsilon_t x_{t-1} + (R_{n1} - R_{n2}) = 2 \sum_{t=L+3}^n \varepsilon_t \ddot{u}_{t-1} + (R_{n1} - R_{n2}),\end{aligned}$$

where  $\ddot{u}_{t-1} = x_{t-1} - L^{-1} \sum_{\ell=1}^L x_{t-1-\ell}$ ,  $R_{n1} = \Psi_n$  and  $R_{n2} = \Psi_{L+2}$ , with

$$\begin{aligned}\Psi_k &= \frac{1}{L} \sum_{\ell=1}^L \left( x_k x_{k-1-\ell} - x_{k-\ell} x_{k-1-\ell} - \sum_{j=0}^{\ell-1} \varepsilon_{k-j} x_{k-1-j} \right) \\ &= \frac{1}{L} \sum_{\ell=1}^L \left[ \sum_{j=0}^{\ell-1} \varepsilon_{k-j} (\rho^j x_{k-1-\ell} - x_{k-1-j}) - (1 - \rho^\ell) \varepsilon_{k-\ell} x_{k-1-\ell} - \rho(1 - \rho^\ell) x_{k-1-\ell}^2 \right] \\ &= \Psi_{k,a} - \Psi_{k,b} - \Psi_{k,c}.\end{aligned}$$

We have the following results (Lemmas A.9 and A.10). They generalize the analysis for the ‘‘single lag’’ case. (Readers can reference that part to help in the following derivations). Recall that  $(1 - \rho)Ex_0^2$  is uniformly bounded.

**Lemma A.9**  $\sup_k E\Psi_{k,a}^2 = O(S_L^2)$ .

**Proof.** We have

$$E\Psi_{k,a}^2 = \frac{\sigma^2}{L^2} \sum_{\ell=1}^L \sum_{j=0}^{\ell-1} E(\rho^j x_{k-1-\ell} - x_{k-1-j})^2.$$

Using  $x_t = \sum_{j=0}^{t-1} \rho^j \varepsilon_{t-j} + \rho^t x_0$ , we get

$$(34) \quad E x_t^2 = \sigma^2 S_t(\rho^2) + \rho^{2t} E x_0^2, \quad \text{and} \quad E x_t x_s = \rho^{t-s} E x_s^2 = \sigma^2 \rho^{t-s} S_s(\rho^2) + \rho^{t+s} E x_0^2,$$

for  $s \leq t$ . Thus

$$\begin{aligned} \omega_{k,\ell,j} &:= E(\rho^j x_{k-1-\ell} - x_{k-1-j})^2 = \rho^{2j} [\sigma^2 S_{k-1-\ell}(\rho^2) + \rho^{2(k-1-\ell)} E x_0^2] \\ &\quad + [\sigma^2 S_{k-1-j}(\rho^2) + \rho^{2(k-1-j)} E x_0^2] - 2\rho^j [\sigma^2 \rho^{\ell-j} S_{k-1-\ell}(\rho^2) + \rho^{2(k-1)-(\ell+j)} E x_0^2]. \end{aligned}$$

We have

$$\begin{aligned} \omega_{k,\ell,j} &= \sigma^2 [\rho^{2j} S_{k-1-\ell}(\rho^2) + S_{k-1-j}(\rho^2) - 2\rho^\ell S_{k-1-\ell}(\rho^2)] \\ &\quad + [\rho^{2(k-1-\ell+j)} + \rho^{2(k-1-j)} - 2\rho^{2(k-1)-\ell}] E x_0^2. \end{aligned}$$

Using  $S_t(\rho^2) = S_s(\rho^2) + \rho^{2(t-s)} S_{t-s}(\rho^2)$ , we get

$$\begin{aligned} \omega_{k,\ell,j} &= \sigma^2 (1 - \rho^\ell)^2 S_{k-1-\ell}(\rho^2) + \sigma^2 \rho^{2j} [1 - \rho^{2(\ell-j)}] S_{k-1-\ell}(\rho^2) + \sigma^2 \rho^{2(\ell-j)} S_{\ell-j}(\rho^2) \\ &\quad + \rho^{2(k-1)+3\ell-2j} [\rho^\ell - \rho^{2(\ell-j)}] E x_0^2 + \rho^{2(k-1)-\ell-2j} (\rho^\ell - \rho^{2j}) E x_0^2. \end{aligned}$$

We now determine the orders of these five terms. First,

$$\frac{1}{L^2} \sum_{\ell=1}^L \sum_{j=0}^{\ell-1} (1 - \rho^\ell)^2 S_{k-1-\ell}(\rho^2) \leq \frac{1}{L} \sum_{\ell=1}^L (1 - \rho^\ell)^2 S_{k-1-\ell}(\rho^2) \leq (1 - \rho^L)^2 S_k(\rho^2) \leq S_L^2.$$

(Here we used Lemma A.1.) Second,

$$\begin{aligned} \frac{1}{L^2} \sum_{\ell=1}^L \sum_{j=0}^{\ell-1} \rho^{2j} [1 - \rho^{2(\ell-j)}] S_{k-1-\ell}(\rho^2) &\leq \frac{1 - \rho^{2L}}{L^2} \sum_{\ell=1}^L S_\ell(\rho^2) S_{k-1-\ell}(\rho^2) \\ &\leq L^{-2} (1 + \rho^L) S_k(\rho^2) (1 - \rho^L) L \bar{S}_L(\rho^2) \leq 2L^{-1} S_L^3. \end{aligned}$$

Third,

$$\frac{1}{L^2} \sum_{\ell=1}^L \sum_{j=0}^{\ell-1} \rho^{2(\ell-j)} S_{\ell-j}(\rho^2) \leq L^{-2} S_L^2 \leq 1.$$

Next,

$$\left| \frac{1}{L^2} \sum_{\ell=1}^L \sum_{j=0}^{\ell-1} \rho^{2(k-1)-3\ell+2j} [\rho^\ell - \rho^{2(\ell-j)}] E x_0^2 \right| \leq (1 - \rho^L) E x_0^2 = (1 - \rho) E x_0^2 S_L,$$

and finally,

$$\left| \frac{1}{L^2} \sum_{\ell=1}^L \sum_{j=0}^{\ell-1} \rho^{2(k-1)-\ell-2j} [\rho^\ell - \rho^{2j}] E x_0^2 \right| \leq (1 - \rho^L) E x_0^2 = (1 - \rho) E x_0^2 S_L.$$

Under the assumption that  $(1 - \rho) E x_0^2$  is finite, we have shown that  $E \Psi_{k,a}^2 = O(S_L^2)$ , where the order is uniform in  $k$ . (The boundaries do not depend on  $k$ .) ■

**Lemma A.10**  $\sup_k E \Psi_{k,b}^2 = O(1)$  and  $\sup_k E |\Psi_{k,c}| = O(S_L)$ .

**Proof.** We have

$$\begin{aligned} E \Psi_{k,b}^2 &= \frac{\sigma^2}{L^2} \sum_{\ell=1}^L (1 - \rho^\ell)^2 E x_{k-1-\ell}^2 \leq \frac{\sigma^2 (1 - \rho^L)^2}{L^2} \sum_{\ell=1}^L E x_{k-1-\ell}^2 \\ &\leq L^{-1} \sigma^2 (1 - \rho^L)^2 E x_k^2 \leq L^{-1} \sigma^2 (1 - \rho^L) S_L (\rho) (1 - \rho) E x_k^2 \\ &\leq \sigma^2 (1 - \rho) [\sigma^2 S_k(\rho^2) + \rho^{2k} E x_0^2] \leq \sigma^4 + \sigma^2 (1 - \rho) E x_0^2 \end{aligned}$$

by (34) and Lemma A.1, where the bound is uniform in  $k$ . Next,

$$\begin{aligned} E |\Psi_{k,c}| &= \frac{1}{L} \sum_{\ell=1}^L \rho (1 - \rho^\ell) E x_{k-1-\ell}^2 \leq (1 - \rho^L) E x_k^2 = S_L (\rho) (1 - \rho) E x_k^2 \\ &\leq S_L [\sigma^2 + (1 - \rho) E x_0^2], \end{aligned}$$

where the bound is again uniform. ■

Lemmas A.9 and A.10 show that

$$(35) \quad (n \bar{S}_L)^{-1/2} L^{-1} \tilde{D}_n^L = \frac{2}{(n \bar{S}_L)^{1/2}} \sum_{t=L+3}^n \varepsilon_t \ddot{u}_{t-1} + O_p(\sqrt{S_L/n}),$$

where  $S_L/n \leq L/n \rightarrow 0$ . So the limit distribution of  $(n \bar{S}_L)^{-1/2} L^{-1} D_n^L$  is the same as the limit distribution of  $Z_n := 2(n \bar{S}_L)^{-1/2} \sum_{t=L+3}^n \varepsilon_t \ddot{u}_{t-1}$ . To find the limit distribution of  $Z_n$ , we can apply the martingale CLT (Theorem 2.3 of McLeish, 1975). For that, we evaluate  $E \ddot{u}_t^2$ , which can be straightforwardly obtained using Lemma A.2. Because  $\ddot{u}_t = L^{-1} \sum_{\ell=1}^L (x_t - x_{t-\ell})$ , we have

$$E \ddot{u}_t^2 = \left( \frac{2\sigma^2}{1 + \rho} \right) \left[ \frac{1}{L^2} \sum_{\ell=1}^L S_\ell + \frac{1}{L^2} \sum_{\ell=1}^{L-1} \sum_{r=1}^{L-\ell} (1 + \rho^r) S_\ell \right].$$

Because  $\sum_{r=1}^{L-\ell} (1 + \rho^r) = L - \ell + \rho S_{L-\ell}$ , we get

$$(36) \quad \bar{S}_L^{-1} E \ddot{u}_t^2 = \left( \frac{2\sigma^2}{1 + \rho} \right) \left[ \frac{1}{L} + \frac{1}{L} \sum_{\ell=1}^{L-1} \left( 1 - \frac{\ell}{L} + \frac{S_{L-\ell}}{L} \right) \frac{S_\ell}{\bar{S}_L} \right] = \left( \frac{2\sigma^2}{1 + \rho} \right) V_L(\rho),$$

which is uniformly bounded by  $4\sigma^2/(1 + \rho)$ . So

$$E \left[ (n \bar{S}_L)^{-1/2} \varepsilon_t \ddot{u}_{t-1} \right]^2 \leq \frac{4\sigma^4}{n(1 + \rho)} \rightarrow 0,$$

implying the Lindeberg condition. It is also required that  $V_L(\rho)$  is bounded away from zero to apply the martingale difference CLT. This part follows because

$$V_L \geq \frac{1}{L} \sum_{\ell=1}^{L-1} \left(1 - \frac{\ell}{L}\right) \frac{S_\ell}{\bar{S}_L} = \frac{1}{L\bar{S}_L} \sum_{\ell=1}^{L-1} \bar{S}_\ell \geq \frac{\sum_{j=1}^{L-1} S_j}{2\sum_{j=1}^L S_j} = \frac{1}{2(1 + \rho^L / \sum_{j=1}^{L-1} S_j)} \geq \frac{1}{4}.$$

So asymptotic normality is obtained. The limit distribution of  $(n\bar{S}_L)^{-1/2}L^{-1}D_n^L$  now follows from the above arguments and (36).

**Theorem A.11**  $(n\bar{S}_L/V_L)^{-1/2}L^{-1}D_n^L \Rightarrow N(0, 8\sigma^4/(1 + \rho))$  as  $L/n \rightarrow 0$ .

**Proof.** The limit distribution of  $D_n^L$  after normalization is the same as that of  $\tilde{D}_n^L$  using the same normalization. The limit distribution of the latter is normal, and its variance is  $4\sigma^2$  (see (35) for the constant 4) times the expression calculated in (36) scaled by  $1/V_L(\rho)$ . ■

**Proof of Theorem 5.** The result follows from Theorems A.7 and A.11. ■

## B References

- Anderson, T.W. (1959). “On Asymptotic Distributions of Estimates of Parameters of Stochastic Difference Equations”, *Annals of Mathematical Statistics*, 30, 676–687.
- Andrews D.W.K. (1988). “Laws of Large Number for Dependent Non-Identically Distributed Random Variables,” *Econometric Theory*, 4, 458–467.
- Andrews, D. W. K. (1993). “Exactly Median-Unbiased Estimation of First Order Autoregressive/Unit Root Models.” *Econometrica*, 61, 139–165.
- Cheang, W. and G. C. Reinsel (2000). “Bias Reduction of Autoregressive Estimates in Time Series Regression Model through Restricted Maximum Likelihood,” *Journal of the American Statistical Association*, 95, 1173–1184.
- Fuller, W. A. (1976). *Introduction to Statistical Time Series*. Wiley.
- Gouriéroux, C., P. C. B. Phillips, and J. Yu (2006). “Indirect Inference for Dynamic Panel Models”, *Journal of Econometrics*, forthcoming.
- Han, C., P. C. B. Phillips and D. Sul (2009). “X-Differencing and Dynamic Panel Model Estimation”, Working Paper, University of Auckland.
- Kim Y., H. Qian and P. Schmidt (1999). “Efficient GMM and MD estimation of autoregressive models”. *Economics Letters*, 62, 265–270.
- McLeish, D.L. (1974). “Dependent central limit theorems and invariance principles,” *Annals of Probability*, 2, 620–628.

- Park, H. J. and W. A. Fuller (1995). "Alternative Estimators and Unit Root Tests for the Autoregressive Process," *Journal of Time Series Analysis*, 16, 415–429.
- Phillips, P. C. B. (1987). "Towards a Unified Asymptotic Theory for Autoregression," *Biometrika* 74, 535–547.
- Phillips, P. C. B. (1992). "Hyperconsistent estimation of a unit root in time series regression". Cowles Foundation Discussion Paper No. 1040.
- Phillips, P. C. B. (1995). "Fully modified least squares and vector autoregression," *Econometrica*, 63, 1023–1078.
- Phillips, P. C. B. and C. Han (2008). "Gaussian Inference in AR(1) Time Series with or without a Unit Root", *Econometric Theory*, 24, 631–650.
- Phillips, P. C. B. and C. C. Lee, (1996), "Efficiency Gains from Quasi-Differencing Under Non-stationarity." In P.M. Robinson and M. Rosenblatt (eds.), *Athens Conference on Applied Probability and Time Series: Essays in Memory of E.J. Hannan*, Springer-Verlag: New York.
- Phillips, P. C. B., and T. Magdalinos (2008). "Unit Root and Cointegrating Limit Theory when Initialization is in the Infinite Past", Cowles Foundation Discussion Paper No. 1655.
- Phillips, P. C. B. and V. Solo (1992). "Asymptotics for Linear Processes," *Annals of Statistics* 20, 971–1001.
- Roy, A. and W. A. Fuller (2001). "Estimation for Autoregressive Time Series with a Root Near 1," *Journal of Business and Economic Statistics*, 19, 482–493.
- White, J. S. (1958). "The Limiting Distribution of the Serial Correlation Coefficient in the Explosive Case," *Annals of Mathematical Statistics* 29, 1188–1197.