# X-DIFFERENCING AND DYNAMIC PANEL MODEL ESTIMATION

**By**

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

**January 2010**

**COWLES FOUNDATION DISCUSSION PAPER NO. 1747**


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

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

# X-Differencing and Dynamic Panel Model Estimation*

Chirok Han
*Korea University*

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

Donggyu Sul
*University of Texas Dallas*

October 2009

**Abstract**

This paper introduces a new estimation method for dynamic panel models with fixed effects and AR($p$) idiosyncratic errors. The proposed estimator uses a novel form of systematic differencing, called X-differencing, that eliminates fixed effects and retains information and signal strength in cases where there is a root at or near unity. The resulting "panel fully aggregated" estimator (PFAE) is obtained by pooled least squares on the system of X-differenced equations. The method is simple to implement, free from bias for all parameter values, including unit root cases, and has strong asymptotic and finite sample performance characteristics that dominate other procedures, such as bias corrected least squares, GMM and system GMM methods. The asymptotic theory holds as long as the cross section ($n$) or time series ($T$) sample size is large, regardless of the $n/T$ ratio, which makes the approach appealing for practical work. In the time series AR(1) case ($n = 1$), the FAE estimator has a limit distribution with smaller bias and variance than the maximum likelihood estimator (MLE) when the autoregressive coefficient is at or near unity and the same limit distribution as the MLE in the stationary case, so the advantages of the approach continue to hold for fixed and even small $n$. For panel data modeling purposes, a general-to-specific selection rule is suggested for choosing the lag parameter $p$ and the procedure works in a standard manner, aiding practical implementation. The PFAE estimation method is also applicable to dynamic panel models with exogenous regressors. Some simulation results are reported giving comparisons with other dynamic panel estimation methods.

*Keywords:* GMM, Panel full aggregation, Stacked and pooled least squares, Panel unit root, X-Differencing.

*JEL classification:* C22, C23

*Aspects of this research were presented in seminars at Kyoto University and University College London in April and May, 2009. Phillips gratefully acknowledges support from a Kelly Fellowship and the NSF under Grant No. SES 06-47086. Han thanks Korea University for research support under Grant No. K0823571.

# 1 Introduction

There is now a vast empirical literature on dynamic panel regressions covering a wide arena of data sets and applications that extend beyond economics across the social sciences. Much of the appeal of panel data stems from its potential to address general socio-economic issues involving decision making over time, so that dynamics play an important role in model formulation and estimation. To the extent that there is commonality in dynamic behavior across individuals, it is natural to expect that pooling cross section data will be advantageous in regression. However, since Nickell (1981) pointed to the incidental-parameter-induced bias effects in pooled least squares regression, there has been an ongoing search for improved statistical procedures.

Prominent among these alternative methods is GMM estimation, which is now the most common approach in practical empirical work with dynamic panel regression. The popularity of GMM is manifest in the extensive citation of articles such as Arellano and Bond (1991) which developed a general GMM approach to dynamic panel estimation. GMM is convenient to implement in empirical research and its widespread availability in packaged software enhances the useability of this methodology. On the other hand, it is now well understood that the original first difference IV (Anderson and Hsiao, 1982) and more general GMM approaches to the estimation of autoregressive parameters in dynamic panels often suffer from problems of inefficiency and substantial bias, especially when there is weak instrumentation as in the commonly occuring case of persistent or near unit root dynamics. Solutions to the weak instrument problem have followed several directions. One approach focuses on the levels equation, where there is no loss of signal in the unit root case, combined with the use of differenced lagged variables as instruments under the assumption that the fixed effects are uncorrelated with the idiosyncratic errors, as developed by Arellano and Bover (1995) and Blundell and Bond (1998). Another approach corrects for the bias of least squares estimators based on parametric assumptions, leading to improved estimation procedures. For example, Kiviet (1995) proposed a bias correction that is based on Nickell's (1981) bias calculations for the panel AR(1); and Hahn and Kuersteiner (2002) modified the pooled least squares (LSDV) method to remove bias up to order $O(T^{-1})$, where $T$ is the time dimension. Other recent work suggests alternative methods of bias-free parametric estimation. For instance, Hsiao, Pesaran and Tahmiscioglu (2002) and Kruiniger (2008) propose the use of quasi-maximum likelihood on differenced data under some parametric assumptions on the distribution of the idiosyncratic errors, which appears to reduce bias without making an explicit bias correction. Han and Phillips (2009) suggest a simple least squares procedure applied to a difference-transformed panel model

that effectively reduces bias in the panel AR(1) case and leads to an asymptotic theory that is continuous as the autoregressive coefficient passes through unity. While the first approach makes moment assumptions on the unobservable individual effects, the other approaches effectively make parametric assumptions on the idiosyncratic error process.

The methods developed in the present paper belong to the second category above but they introduce a novel technique of systematic differencing, which we call "X-differencing", that eliminates fixed effects while retaining information and signal strength in cases of practical importance where there is an autoregressive root at or near unity. The resulting "panel fully aggregated" estimator (PFAE) is obtained by applying least squares regression to the full system of X-differenced equations. The method is simple to implement, is free from bias for all parameter values and has higher asymptotic efficiency than bias-corrected LSDV estimation, thereby retaining signal strength in the unit root case and resolving many of the difficulties associated with weak instrumentation and dynamic panel regression bias. The general model considered here is a linear dynamic panel model with AR($p$) idiosyncratic errors and exogenous variables, so the framework is well suited to a wide range of models used in applied work.

Unlike the Hahn and Kuersteiner (2002) bias corrected LSDV estimator, the PFAE method does not require large $T$ for consistency. The PFAE procedure also supercedes the Han and Phillips (2009) least squares method by generalizing it to AR($p$) models and by considerably improving its efficiency both in stationary and unit root cases. Since the PFAE is a least squares estimator, there is no dependence on distributional assumptions and none of the computational burden and potential singularities that exist in numerical procedures such first difference MLE (Hsiao et al, 2002; Kruiniger, 2008). Moreover, since X-differencing eliminates fixed effects, the asymptotic distribution of the PFAE estimator does not depend on the distribution of the individual effects, whereas GMM in levels (Arellano and Bover, 1995) and system GMM (Blundell and Bond, 1998) are both known to suffer from this problem (Hayakawa, 2008). Finally, because the autoregressive coefficients are consistently estimated, it is straightforward to implement parametric panel GLS estimation in a second stage regression (e.g., Bhargava et al, 1982, for the panel AR(1) model).

The current paper relates to a companion work by the authors (Han, Phillips and Sul, 2009; HPS hereafter), which introduced the 'time-reversal' technology used here to design the X-differencing transformations that eliminate fixed effects and correct for autoregressive estimation bias. Using this methodology, the companion paper developed a new "fully aggregated" estimator (FAE) specifically for the time series AR(1) model. That paper focused on the process of information aggregation in X-differenced equation systems to enhance efficiency in time series regression and

to retain asymptotic normality for inference purposes, while the current paper emphasizes bias removal and efficiency improvement in the panel context. The present paper also extends the HPS technology to AR(p) panel regressions and to models with exogenous variables.

The remainder of the paper is organized as follows. Section 2 provides the key motivating ideas and some heuristics that explain the X-differencing process and how the new estimation method works in the simple panel AR(1) model. Section 3 extends the methodology to the panel AR(p) model, develops the X-differenced equation system, verifies orthogonality, and discusses implementation of the PFAE procedure. Section 4 presents the limit theory of the PFAE and provides comparisions with other methods such as bias corrected LSDV and first difference MLE (FDMLE). This section also discusses issues of lag length selection in the context of dynamic panels with unknown lag length. Section 5 reports some simulation results which compare the finite sample performance of the new procedure with existing estimators. Section 6 concludes. Some more general limit theory, proofs, and supporting technical material are given in the Appendices.

# 2 Key Ideas and X-Differencing

We start by developing some key ideas and provide intuition for the new procedure using the simple panel AR(1) model with fixed effects

$$(1) \quad y_{it} = a_i + u_{it}, \text{ with } u_{it} = \rho u_{it-1} + \varepsilon_{it}, \quad t = 1, ..., T; \ i = 1, ..., n,$$

where the innovations $\varepsilon_{it}$ are $iid\ (0, \sigma^2)$ over $i$ and $t$. The model can be written in alternative form as

$$(2) \quad y_{it} = \alpha_i + \rho y_{it-1} + \varepsilon_{it}, \quad \alpha_i = a_i(1-\rho),$$

which corresponds to the conventional dynamic panel AR(1) model $y_{it} = \alpha_i + \rho y_{it-1} + \varepsilon_{it}$ when $|\rho| < 1$. When $\rho = 1$, the individual effects are eliminated by differencing and both (1) and (2) reduce to $\Delta y_{it} = \varepsilon_{it}$. The AR(1) specification is used only for expository purposes and is replaced by AR($p$) dynamics in the rest of the paper, where we also relax the conditions on the innovations $\varepsilon_{it}$. Initial conditions are conventionally set in the infinite past in the stable case $|\rho| < 1$ and at $t = 0$ with some $O_p(1)$ initialization when $\rho = 1$, although various other settings, while not our concern here, are possible and can be treated as in Phillips and Magdalinos (2009). Observe that there is no restriction on $\rho$ in (1), whereas in (2) $\rho$ is effectively restricted to the region $-1 < \rho \leq 1$ because for $\rho > 1$, $\alpha_i = a_i(1-\rho) \neq 0$ in which case the system has a deterministic explosive

component in contrast to (1). This implicit restriction in (2) is not commonly recognised in the literature but, as mentioned later in the paper, it is important in comparing different estimation procedures where some may be restricted in terms of their support but not others.

No distributional assumptions are placed on the individual effects $\alpha_i$. So the model corresponds to a fixed effects environment where the incidental parameters need to be estimated or eliminated. Various approaches have been developed in the literature, including the within-group (regression) transformation, first differencing, recursive mean adjustment, forward filtering, and long-differencing. However, all of these methods lead to final estimating equations for $\rho$ in which the transformed (dynamic) regressor is correlated with the transformed error. In the simple time series case, where the intercept is fitted in least squares regression leading to a demeaning transformation, the effects of bias in the estimation of $\rho$ have long been known to be exacerbated by the demeaning (e.g., Orcutt and Winokur, 1969) and in the panel case these bias effects persist asymptotically as $n \to \infty$ for $T$ fixed (Nickell, 1981). Accordingly, various estimation methods have been proposed to address the difficulty such as instrumental variable and GMM methods, direct bias correction methods, and the various transformation and quasi-likelihood methods discussed in the Introduction.

The essence of the technique introduced in the present paper is a novel differencing procedure that successfully eliminates the individual effects (like conventional differencing) while at the same time making the regressor and the error uncorrelated after the transformation (which other methods fail to do). A key advantage is that the new approach does not suffer from the weak identification and instrumentation problems that bedevil IV/GMM methods based on first differenced (or forward filtered) equations when the dynamics are persistent. This failure of GMM in unit root and near unit root cases produces some undesirable performance characteristics in the GMM estimator and poor approximation by the usual asymptotic theory[1]. At the same time, because the $\alpha_i$ are eliminated, the new method is unaffected by the relative variance ratio between the individual effects $\alpha_i$ and the idiosyncratic errors $\varepsilon_{it}$, which, if large, makes the system GMM estimator (Blundell and Bond, 1998) perform poorly (see Hayakawa, 2008). Hence, we expect that the new procedure should offer substantial gains over both GMM and system GMM methods, while still having the advantage of easy computation.

The new procedure begins by combining (2) with the implied forward looking regression equa-

[1]For instance, the finite sample variance of the first difference GMM estimator in the stationary case *increases* rather than decreases as $\rho$ increases (see, Alvarez and Arellano, 2003; Hayakawa, 2008) in contrast to the prediction of asymptotic theory.

tion

(3) $y_{is} = \alpha_i + \rho y_{is+1} + \varepsilon^*_{is}$, with $\varepsilon^*_{is} = \varepsilon_{is} - \rho(y_{is+1} - y_{is-1})$,

and where the 'future' variable is on the right hand side, as opposed to the original 'backward looking' equation (2). Importantly in both the backward and the forward looking equations, the regressors are uncorrelated with the corresponding regression errors. That is, $Ey_{it-1}\varepsilon_{it} = 0$ in (2) and

(4) $Ey_{is+1}\varepsilon^*_{is} = Ey_{is+1}\varepsilon_{is} - \rho E\left[y_{is+1}\left(y_{is+1} - y_{is-1}\right)\right] = \rho\sigma^2_\varepsilon - \rho\sigma^2_\varepsilon = 0,$

in (3), under the following conditions: (i) $E\alpha_i\varepsilon_{it} = 0$ for all $t$ (a condition that is not actually required in our subsequent development because the $\alpha_i$ are eliminated - see equation (6) below); (ii) $\varepsilon_{it}$ is white noise over $t$; and (iii) $|\rho| < 1$. The proof of (4) is given in Appendix A. If $\rho = 1$, then the last equality of (4) is not true, but this restriction is removed in the final transformation (see (7) below). The orthogonality (4) is a critical element in the development of the new estimation procedure involving systematic differencing.

Importantly, the orthogonality (4) still holds if we replace $s+1$ with any $t > s$, i.e., $Ey_{it}\varepsilon^*_{is} = 0$ for any $t > s$. The implication is that the original backward looking regressor $y_{it-1}$ is uncorrelated with the forward looking regression errors $\varepsilon^*_{is}$ as long as $t - 1 > s$. That is, under the conditions that $E\alpha_i\varepsilon_{it} = 0$, $\varepsilon_{it}$ is white-noise over $t$, and $|\rho| < 1$, we have

(5) $Ey_{it-1}\varepsilon^*_{is} = -\rho E\left[y_{it-1}\left(y_{is+1} - y_{is-1}\right)\right] + Ey_{it-1}\varepsilon_{is} = 0$ for any $t > s + 1$.

Again the condition that $|\rho| < 1$ is not required in the final transformation step shown below in (7).

Results (4) and (5) can be used to eliminate the fixed effects. By simply subtracting (3) from (2), we get the new regression equation

(6) $y_{it} - y_{is} = \rho\left(y_{it-1} - y_{is+1}\right) + \left(\varepsilon_{it} - \varepsilon^*_{is}\right),$

where the regressor $y_{it-1} - y_{is+1}$ is uncorrelated with the error $\varepsilon_{it} - \varepsilon^*_{is}$ as long as $s < t - 1$ for all $-1 < \rho \leq 1$. Note that we now allow for the unit root case $\rho = 1$ and this relaxation is justified in Lemma 1 below. Thus, for model (2), if $\varepsilon_{it}$ is white-noise over $t$, then the key orthogonality condition

(7) $E\left(y_{it-1} - y_{is+1}\right)\left(\varepsilon_{it} - \varepsilon^*_{is}\right) = 0$ for all $s < t - 1$ and $-1 < \rho \leq 1$,

holds for model (6), thereby validating the use of pooled least squares regression techniques.

We call the data transformation involved in setting up the regression equation (6) "X-differencing". Observe that the dependent variable $y_{it} - y_{is}$ is $X = t - s$ differenced whereas the regressor $y_{it-1} - y_{is+1}$ is $X = t - s - 2$ differenced. So, the regression equation is structured with variable differencing: the differencing varies in a systematic and critical way between the dependent variable and the regressor. Further, we want to allow for the differencing rate $X$ itself to change, so $X$ is a variable. Hence, the terminology X-differencing.

The simple X-differencing transformation that leads to (6) eliminates the nuisance parameters $\alpha_i$, just like ordinary differencing, but it has the additional advantage that the regression equation satisifies a fundamental orthogonality condition: there is no correlation between the regressor and the error in (6). As a result, X-differencing is very different from existing differencing methods that have been used in the literature. In one way it is fundamentally simpler – because of the appealing orthogonality property satisfied by (6). In another way it is more complete – because the differencing rate $X$ is variable, so that it is possible to think of (6) as a system of equations over $s < t - 1$, each equation of which carries useful information about the autoregressive coefficient $\rho$.

It is interesting to compare (6) with other differencing transformations that have been used in the literature. First, it is different from long differencing (Hahn, Hausman and Kuersteiner, 2007), which transforms equation (2) to $y_{it} - y_{i2} = \rho(y_{it-1} - y_{i1}) + (\varepsilon_{it} - \varepsilon_{i2})$, whereas our method (when $s = 1$) yields $y_{it} - y_{i1} = \rho(y_{it-1} - y_{i2}) + (\varepsilon_{it} - \varepsilon^*_{i1})$, so the positions of $y_{i1}$ and $y_{i2}$ are switched, the equation error is different and our approach allows $s$ to vary. Second, X-differencing (when $s = t - 3$) is also distinguished from simple first differencing, which gives the equation $y_{it} - y_{it-1} = \rho(y_{it-1} - y_{it-2}) + (\varepsilon_{it} - \varepsilon_{it-1})$. In our model, we replace $y_{it-1}$ on the left hand side with $y_{it-3}$, the equation error is different, and again we allow for higher order differences.

Third, when $s = t - 3$, the transformed equation (6) in our model can be written as

$$\Delta y_{it} + \Delta y_{it-1} + \Delta y_{it-2} = \rho \Delta y_{it-1} + (\varepsilon_{it} - \varepsilon^*_{it-3}), \tag{8}$$

where $\Delta y_{it} = y_{it} - y_{it-1}$. This equation can usefully be compared with the AR(1) bias-correction transformation model

$$2\Delta y_{it} + \Delta y_{it-1} = \rho \Delta y_{it-1} + \text{error}_{it} \tag{9}$$

that was used in Phillips and Han (2008) and Han and Phillips (2009). In the new X-differencing approach, the present method replaces the term $2\Delta y_{it}$ in model (9) with $\Delta y_{it} + \Delta y_{it-2}$. This "temporal balancing" around the lagged difference $\Delta y_{it-1}$ is a subtle but important breakthrough that

leads to the variable X-differencing generalization of (9) and, as we shall see, leads to considerable efficiency gains and further allows for convenient generalization from AR(1) to AR($p$) models.

Importantly, any $s$ values such that $s < t-1$ satisfy (7) under the stated regularity, so that the new regression equation (6) is valid across all these values. To make full use of all this information, we propose to stack the regression equations (6) for all possible $s$ values. But we exclude $s = t-2$ because in this case the corresponding regressor in (6) is zeroed out. Thus, we propose to use equation (6) for $s = 1, 2, \ldots, t-3$. The resulting stacked and pooled least squares estimator has the following simple form

$$\hat{\rho} = \frac{\sum_{i=1}^{n}\sum_{t=4}^{T}\sum_{s=1}^{t-3}(y_{it-1} - y_{is+1})(y_{it} - y_{is})}{\sum_{i=1}^{n}\sum_{t=4}^{T}\sum_{s=1}^{t-3}(y_{it-1} - y_{is+1})^2}$$

and is the panel fully aggregated estimator (PFAE) of $\rho$ in the panel AR(1) model (2). In the time series case where $n = 1$, $\hat{\rho}$ reduces to the FAE estimator introduced in HPS (2009).

The estimator $\hat{\rho}$ has virtually no bias for all $\rho$ values, as might be expected in view of the prevailing orthogonality (7), the simple no intercept form of (6) and the differenced form of the regressor. In the limit, consistency holds provided the total number of observations tends to infinity—irrespective of the $n/T$ ratio—indicating that the estimator will be useful in short and long panels, as well as narrow and wide panels, making it appealing in both microeconometric and macroeconometric data sets. This result, together with the asymptotic distribution theory and associated tools for inference, will be developed in the following sections in the context of the general AR($p$) panel model.

# 3 The Panel AR($p$) Model with Fixed Effects

This section extends the above ideas on X-differencing and fully aggregated estimation to the general case of a dynamic panel AR($p$) model. Our primary concern is the estimation of the common autoregressive parameters $\{\rho_j : j = 1, \ldots, p\}$ in the following panel model with fixed effects and autoregressive errors

$$y_{it} = a_i + u_{it}, \quad \rho(L)\, u_{it} = \varepsilon_{it}, \quad t = 1, \ldots, T;\ i = 1, \ldots, n, \tag{10}$$

$$\rho(L) = 1 - \rho_1 L - \cdots - \rho_p L^p, \tag{11}$$

where $\varepsilon_{it}$ is, for each $i$, a martingale difference sequence (mds) under the natural filtration with $E\varepsilon_{it} = 0$, and $E\varepsilon_{it}^2 = \sigma_i^2$. As in the AR(1) case we have the equivalent specification (at least in the

stationary and unit root cases, c.f. the discussion following (2) above)

$$(12)\quad y_{it} = \alpha_i + \rho_1 y_{it-1} + \cdots + \rho_p y_{it-p} + \varepsilon_{it}, \quad \alpha_i = a_i(1 - \rho_1 - \cdots - \rho_p).$$

We maintain the assumption that $u_{it}$ has at most one unit root. When $u_{it}$ is $I(1)$, the long run AR coefficient is $\rho_{lr} = \sum_{j=1}^{p} \rho_j = 1$, and we write $\rho(L) = (1-L)\rho^*(L)$ where the roots of $\rho^*(L) = 0$ are outside the unit circle. In this event, $\alpha_i = 0$ in (12) and there is no drift in the process. Initial conditions for $u_{it}$ may be set in the infinite past in the stationary case. In the unit root case, we can write $\Delta u_{it} = \frac{1}{\rho^*(L)}\varepsilon_{it} := u^*_{it}$ and set the initial conditions for the stationary AR(p-1) process $u^*_{it}$ in the infinite past. Since our estimation procedure relies only on X-differenced data, it is not necessary to be explicit about initial conditions for $u_{it}$. In fact, our results will hold for distant and infinitely distant initializations (where $u_{i0}$ can be $O_p\left(\sqrt{T\kappa_T}\right)$ for some $\kappa_T$ which may tend to infinity with $T$) as well as $O_p(1)$ initializations (see Phillips and Magdalinos, 2009, for discussion of these initial conditions).

Following the same motivation as in the AR(1) case, to construct the X-differenced equation system we rewrite (12) in forward looking format as

$$y_{is} = \alpha_i + \rho_1 y_{is+1} + \cdots + \rho_p y_{is+p} + \varepsilon^*_{is},$$

where $\varepsilon^*_{is} = \varepsilon_{is} - \sum_{j=1}^{p} \rho_j (y_{is+j} - y_{is-j})$. Then, by subtracting this equation from the original backward looking equation (12), we construct the X-differenced equation system

$$(13)\quad y_{it} - y_{is} = \rho_1(y_{it-1} - y_{is+1}) + \cdots + \rho_p(y_{it-p} - y_{is+p}) + (\varepsilon_{it} - \varepsilon^*_{is}),$$

just as in the AR(1) case. The system may also be written as

$$u_{it} - u_{is} = \rho_1(u_{it-1} - u_{is+1}) + \cdots + \rho_p(u_{it-p} - u_{is+p}) + (\varepsilon_{it} - \varepsilon^*_{is}),$$

and is free of fixed effects.

Observe that the variables appearing in (13) involve $X = t - s - 2k$ differences for $k = 0, .., p$. The regressors in (13) are all uncorrelated with the regression error in the equation, as shown in Lemma 1 below. Importantly, this orthogonality condition holds for the full system of equations given in (13)—that is for all $t - s \geq p + 2$.

**Lemma 1** $E(y_{it-k} - y_{is+k})(\varepsilon_{it} - \varepsilon^*_{is}) = 0$ *for all* $s \leq t - p - 2$, *for all* $k = 1, \ldots, p$.

In stacking the system (13) for estimation purposes, we use all possible $s$ values up to $s = t - 2p - 1$. This setting avoids the collinearity (or zeroing out) of the regressors that occurs when

the system includes $s$ values within the range $t-2p \leq s < t-p-1$. To express the estimator in a concise form, let $\widetilde{Z}_{it,s} = (y_{it-1}, y_{it-2}, \ldots, y_{it-p})' - (y_{is+1}, y_{is+2}, \ldots, y_{is+p})'$, $\widetilde{y}_{it,s} = y_{it} - y_{is}$, $\widetilde{\varepsilon}_{it,s} = \varepsilon_{it} - \varepsilon^*_{is}$, and $\rho = (\rho_1, \ldots, \rho_p)'$. Then, (13) can be expressed as

$$\text{(14)} \quad \widetilde{y}_{it,s} = \rho' \widetilde{Z}_{it,s} + \widetilde{\varepsilon}_{it,s}.$$

The PFAE for $\rho$ is simply the least squares estimator based on the stacked (over $s$) and pooled (over $i$ and $t$) system (14), viz.,

$$\text{(15)} \quad \hat{\rho} = \left( \sum_{i=1}^{n} \sum_{t=2p+2}^{T} \sum_{s=1}^{t-2p-1} \widetilde{Z}_{it,s} \widetilde{Z}'_{it,s} \right)^{-1} \sum_{i=1}^{n} \sum_{t=2p+2}^{T} \sum_{s=1}^{t-2p-1} \widetilde{Z}_{it,s} \widetilde{y}_{it,s}.$$

Clearly, the degrees of freedom condition $T \geq 2p+2$ is required for the existence of $\hat{\rho}$.

It is sometimes convenient to obtain the PFAE as follows. For a given lag $\ell$, let $\hat{\rho}_{(\ell)}$ be the X-differencing estimator based on the equation

$$y_{it} - y_{it-2p-\ell} = \sum_{j=1}^{p} \rho_j (y_{it-j} - y_{it-2p-\ell+j}) + (\varepsilon_{it} - \varepsilon^*_{it-2p-\ell}).$$

Here, $\ell = 1$ is the minimum lag allowed in PFA estimation (to avoid perfect collinearity), and $\ell = T-2p-1$ is the maximum lag. Let $\widetilde{\mathbf{Z}}_\ell$ be the regressor matrix for this lag (for all $i$ and for all possible $t$) and let $\widetilde{\mathbf{y}}_\ell$ be the corresponding regressand vector. When we regress $\widetilde{\mathbf{y}}_\ell$ on $\widetilde{\mathbf{Z}}_\ell$, we get the *lag-*$\ell$ estimator $\hat{\rho}_{(\ell)} = (\widetilde{\mathbf{Z}}'_\ell \widetilde{\mathbf{Z}}_\ell)^{-1} \widetilde{\mathbf{Z}}'_\ell \widetilde{\mathbf{y}}_\ell$. Then the PFAE is

$$\text{(16)} \quad \hat{\rho} = \left( \sum_{\ell=1}^{T-2p-1} \widetilde{\mathbf{Z}}'_\ell \widetilde{\mathbf{Z}}_\ell \right)^{-1} \sum_{\ell=1}^{T-2p-1} \widetilde{\mathbf{Z}}'_\ell \widetilde{\mathbf{y}}_\ell = \left( \sum_{\ell=1}^{T-2p-1} \widetilde{\mathbf{Z}}'_\ell \widetilde{\mathbf{Z}}_\ell \right)^{-1} \sum_{\ell=1}^{T-2p-1} \widetilde{\mathbf{Z}}'_\ell \widetilde{\mathbf{Z}}_\ell \hat{\rho}_{(\ell)},$$

which is a weighted average of all *lag-*$\ell$ estimators, where the weights are assigned according to the magnitude of the *lag-*$\ell$ signal matrix $\widetilde{\mathbf{Z}}'_\ell \widetilde{\mathbf{Z}}_\ell$. Note that all single *lag-*$\ell$ estimators are themselves individually consistent as the sample size increases.

The weighted regression formulation (16) offers some computational advantages in practical implementation and it is used in some of the simulations (undertaken in Stata and Gauss) that are reported in Section 5.

The orthogonality condition in Lemma 1 holds if $\varepsilon_{it}$ is white noise for each $i$. However, the development of an asymptotic theory for $\hat{\rho}$ requires stronger regularity conditions that validate laws of large numbers (LLNs), central limit theorems (CLTs) and functional CLTs as $n$ and $T$ pass to infinity. Our theory includes both fixed $T$ and fixed $n$ cases. For these developments, we assume the following.

**Condition A** *(i)* $\varepsilon_{it} = \sigma_i \varepsilon_{it}^{\circ}$ *with* $\inf_i \sigma_i > 0$ *and* $\sup_i \sigma_i < \infty$*, where* $\varepsilon_{it}^{\circ}$ *is* $iid$ *across* $i$ *with* $E[(\varepsilon_{it}^{\circ})^{4+\delta}] \leq M$ *for all* $t$ *and some* $M < \infty$ *and* $\delta > 0$*; (ii)* $\varepsilon_{it}^{\circ}$ *is a stationary and ergodic martingale difference sequence (mds) over* $t$ *for all* $i$ *such that* $E(\varepsilon_{it}^{\circ}|\varepsilon_{it-1}^{\circ}, \varepsilon_{it-2}^{\circ}, \ldots) = 0$, $E(\varepsilon_{it}^{\circ}|\varepsilon_{it+1}^{\circ}, \varepsilon_{it+2}^{\circ}, \ldots) = 0$, *and with unit conditional variances*

$$E(\varepsilon_{it}^{\circ 2}|\varepsilon_{it-1}^{\circ}, \varepsilon_{it-2}^{\circ}, \ldots) = E(\varepsilon_{it}^{\circ 2}|\varepsilon_{it+1}^{\circ}, \varepsilon_{it+2}^{\circ}, \ldots) = 1 \ \ a.s.\,;$$

*(iii)* $n^{-1}\sum_{i=1}^{n} \sigma_i^2$ *and* $n^{-1}\sum_{i=1}^{n} \sigma_i^4$ *converge to finite limits as* $n \to \infty$.

**Remarks.**

1. We allow cross-section heterogeneity in (i) by considering a scaled version $\varepsilon_{it} = \sigma_i \varepsilon_{it}^{\circ}$ of an $mds$ random sequence $(\varepsilon_{it}^{\circ})$ for each $t$. This assumption is not crucial but it simplifies the analysis considerably. Generalization to non-identically distributed (across $i$) innovations is possible but involves further technicalities, including some explicit conditions for third and fourth moments and the Lindeberg condition.

2. Condition (ii) is a bidirectional $mds$ condition and corresponds to a conventional white noise assumption. This condition is weaker than requiring independence in $\varepsilon_{it}^{\circ}$ over $t$, but is stronger than a unidirectional $mds$ condition.

3. Conditional heteroskedasticity or higher order serial dependence (over $t$) may be allowed as long as Condition A(ii) is satisfied. If $T$ is fixed and $n$ is large, no conditions on the serial dependence of $\varepsilon_{it}$ are required other than $E\varepsilon_{it} = 0$, $E\varepsilon_{it}^2 = \sigma_i^2$ and $E\varepsilon_{it}\varepsilon_{is} = 0$ for all $t$ and $s \neq t$.

4. Condition A(iii) seems quite weak, although it is not implied by Condition A(i). When A(iii) holds, the average moments converge to finite positive limits in view of Condition A(i).

When $T$ is fixed and $n \to \infty$, we require the following regularity for the standardized error sequence $\varepsilon_{it}/\sigma_i$ so we may establish standard asymptotics for the PFAE.

**Condition B** *For any given* $T$*, (i)* $E(\eta_{iT}^{\circ}\eta_{iT}^{\circ\prime})$ *is nonsingular, where*

$$\eta_{iT}^{\circ} = \sum_{t=2p+2}^{T}\sum_{s=1}^{t-2p-1} \widetilde{Z}_{it,s}\widetilde{\varepsilon}_{it,s}/\sigma_i^2$$

*and* $\widetilde{Z}_{it,s}$ *and* $\widetilde{\varepsilon}_{it,s}$ *are defined in (14); (ii)* $n^{-1}\sum_{i=1}^{n}(\eta_{iT}\eta_{iT}' - E\eta_{iT}\eta_{iT}') \to_p 0$.

**Remark.** In developing a CLT for the numerator of a centred form of (15), only Condition A is required. Condition B (i) is relevant for establishing the standard normal limit given in Theorem 2 below. Condition B (ii) is useful for the estimation of the variance-covariance matrix of the limit distribution. When $\varepsilon_{it}$ is independent and possibly heterogeneous across $i$, a sufficient condition for B (ii) is given in Phillips and Solo (1992, Theorem 2.3). ■

When $T \to \infty$, the temporal dependence structure matters and affects the limit theory and rates of convergence. In the general AR($p$) model with a unit root, there is an asymptotic singularity in the sample moment matrix because of the stronger signal in the data in the unit root direction, just as in the time series case (Park and Phillips, 1988). Singularities are treated by rotating the regressor space and reparameterization as detailed in Appendix A.

# 4 Asymptotic Theory

This section develops an asymptotic theory for the PFAE $\hat{\rho}$. Technical derivations and a general theory are given in Appendix A. To make the results of the paper more accessible, only the main findings that are useful for empirical research are reported here. We start with the following notation

$$(17)\quad V_{iT} = \frac{1}{T}\sum_{t=2p+2}^{T}\sum_{s=1}^{t-2p-1}\widetilde{Z}_{it,s}\widetilde{Z}'_{it,s} \text{ and } \eta_{iT} = \frac{1}{T}\sum_{t=2p+2}^{T}\sum_{s=1}^{t-2p-1}\widetilde{Z}_{it,s}\widetilde{\varepsilon}_{it,s},$$

so that $\hat{\boldsymbol{\rho}} = \boldsymbol{\rho} + \left(\sum_{i=1}^{n} V_{iT}\right)^{-1}\sum_{i=1}^{n}\eta_{iT}$.

Because $E\eta_{iT} = 0$ for all $T$ by Lemma 1, we can expect the panel estimator $\hat{\boldsymbol{\rho}}$ to be consistent and asymptotically normal under regularity conditions that ensure suitable behavior for the sample components $\left(\sum_{i=1}^{n} V_{iT}, \sum_{i=1}^{n}\eta_{iT}\right)$ of $\hat{\boldsymbol{\rho}}$. In particular, if $y_{it}$ is stationary, then consistency and asymptotic normality will hold, provided the total number of observations in the regression is large, i.e. if $N = n(T - 2p - 1) \to \infty$. So, no condition on the behavior of the ratio $n/T$ is required in the limit theory. If $y_{it}$ is persistent (so that the long run AR coefficient $\rho_{lr} := \sum_{j=1}^{p}\rho_j$ is unity) and $T$ is finite, then large-$n$ asymptotics are again standard because any special behavior in the components (e.g. nonstandard convergence rates and limit behavior associated with nonstationarity) occurs only when $T \to \infty$. Next, if $y_{it}$ is persistent and $T \to \infty$, the estimator $\hat{\boldsymbol{\rho}}$ is consistent and still asymptotically normal when $n \to \infty$, again irrespective of the $n/T$ ratio. In this case, the corresponding estimate of the long run AR coefficient $\rho_{lr}$ (which, because of persistence, is $\rho_{lr} = 1$) has a faster convergence rate $O_p(n^{1/2}T)$ stemming from the stronger signal in the nonstationary component of the data, thereby producing a singularity in the joint asymptotic

normal distribution of $\hat{\boldsymbol{\rho}}$ with one component (in the direction $\hat{\rho}_{lr} = \sum_{j=1}^{p} \hat{\rho}_j$) converging faster to its normal distribution than the other components. When $n$ is fixed and $T \to \infty$ in the persistent case, then the limit distribution of $\hat{\boldsymbol{\rho}}$ is again singular normal (when $p > 1$) but there is a faster rate of convergence in the direction $\hat{\rho}_{lr}$ and the limit distribution is nonstandard in that direction. The latter result is related to the limit theory of the time series FAE estimator given in HPS (2009) for the special case where $n = 1$.

Theorem 5 in Appendix A provides a complete statement for interested readers of this limit theory, covering the general panel AR($p$) case in a uniform way for large $T$ and $n$, as well as both fixed $T$ and fixed $n$ cases. The remainder of this section focuses on practical aspects of this limit theory and the useability of the PFAE in applied work.

## 4.1 Limiting Distribution of the PFAE

For inference and practical implementation, Theorem 2 presents a feasible version of the main part of Theorem 5 in Appendix A that holds uniformly for all $\rho$ values including both stationary and unit root cases. For convenience, we use the model (1) formulation in which $y_{it} = a_i + u_{it}$, where $u_{it}$ is an AR($p$) process as defined in (10).

**Theorem 2** *Suppose $u_{it}$ is AR(p) as defined in (10). Under Condition A,*

(18) $B_{nT} \left( \sum_{i=1}^{n} V_{iT} \right) (\hat{\rho} - \rho) \Rightarrow N(0, I_p),$

*for any $B_{nT}$ such that $B_{nT}(\sum_{i=1}^{n} \eta_{iT}\eta_{iT}')B_{nT}' = I_p$, where $V_{iT}$ and $\eta_{iT}$ are defined in (17). The convergence (18) holds as $nT \to \infty$ if $\rho_{lr} := \sum_{j=1}^{p} \rho_j < 1$, and as $n \to \infty$ in all cases (that is, for any $T$, either finite or increasing to infinity, no matter how fast). The limit distribution of $\hat{\rho}$ when $n$ is fixed, $T \to \infty$ and $\rho_{lr} = 1$ is partly normal and partly nonstandard. It is given in Theorem 5(d) in Appendix A.*

**Remarks.**

1. Note that cross section heterogeneity is permitted in Theorem 2 under Condition A. The matrices $\sum_{i=1}^{n} V_{iT}$ and $\sum_{i=1}^{n} \eta_{iT}\eta_{iT}'$ in the theorem are designed to be heteroskedasticity robust so that (18) provides a central limit theorem suitable for implementation upon estimation of $\sum_{i=1}^{n} \eta_{iT}\eta_{iT}'$ as discussed below. The asymptotic form of the standardization matrix $B_{nT}$ in (18) is given in (53) in Appendix A and shows explicitly the convergence rates in terms of $n$ and $T$ as well as the transformation matrix involved in arranging directions of faster and slower convergence when there is a unit root in the system.

2. For statistical testing, it is necessary to replace $\eta_{iT}$ by a feasible statistic. In view of (17) and the consistency of $\hat{\rho}$, we can use the residuals

$$(19)\quad \widehat{\eta}_{iT} = \sum_{t=2p+2}^{T} \sum_{s=1}^{t-2p-1} \widetilde{Z}_{it,s}(\widetilde{y}_{it,s} - \widetilde{Z}'_{it,s}\hat{\rho}),$$

in place of $\eta_{iT}$. The asymptotic covariance matrix estimate $[\sum_i V_{iT}]^{-1} \sum_i \widehat{\eta}_{iT}\widehat{\eta}'_{iT}[\sum_i V_{iT}]^{-1}$ may then be used in inference. Simulations show that this choice works well when $n$ is large. If $n$ is not so large, inferences based on this method still show reasonable performance and may be improved by modification of the limit distribution of the associated (scalar) test statistics to a Student $t$ distribution with $n-1$ degrees of freedom as proposed in Hansen (2007) if the random variables are $iid$ across $i$.

3. For practical work, it may be useful to provide estimates of the remaining (non dynamic) parameters in the model (10). Consistent estimation of the autoregressive coefficients in (10) enables estimation of the fixed effects, the variance of the fixed effects and that of the random innovations in a standard way. For example, the transformed fixed effects $\alpha_i := a_i(1-\rho_{lr})$ can be estimated by the individual sample mean, $\hat{\alpha}_i$, of the residuals $\hat{e}_{it} := y_{it} - \sum_{j=1}^{p} \hat{\rho}_j y_{it-1}$, and the random idiosyncratic innovations $\varepsilon_{it}$ can be estimated by the quantity $\hat{e}_{it} - \hat{\alpha}_i$. The average variances of $\alpha_i$ and $\varepsilon_{it}$ can then be estimated by the sample variances of $\hat{\alpha}_i$ (across $i$) and $\hat{e}_{it} - \hat{\alpha}_i$ (across $i$ and $t$ after the degrees of freedom correction), respectively. Asymptotics for these additional estimates follow in a standard way from the usual limit theory for sample moments and the consistency of the fitted autoregressive coefficients.

We now provide some further discussion of efficiency. At present there is no general theory of asymptotic efficiency for panel data models that applies for multi-index asymptotics and possible nonstationarity. The usual Hájek-Le Cam representation theory (Hájek, 1972; Le Cam, 1972) holds for locally asymptotically normal (LAN) families and regular estimators in the context of single index and $\sqrt{n}$ asymptotics. Panel LAN asymptotics were developed for the stationary Gaussian AR(1) case by Hahn and Kuersteiner (2002) allowing for fixed effects under certain rate conditions on $n$ and $T$ passing to infinity. But their result does not apply when there is a unit root in the system. Any such further extension of existing optimality theory would require that $n \to \infty$ because for fixed $n$ (and in particular $n = 1$) the likelihood does not belong to the LAN family but is of the locally asymptotically Brownian functional family (Phillips, 1989; Jeganathan, 1995), for which there is no present theory of optimal estimation or asymptotic efficiency. Moreover, it is now

known from the results of HPS (2009) that improvements in both bias and variance over the MLE and bias corrected MLE are possible in local neighborhoods of unity in the time series case ($n = 1$).

For the purposes of the present study, we undertake a more limited investigation of efficiency and consider the simple panel AR(1) model (1) with Gaussian errors. Normality is not needed for the limit theory but only for the discussion of optimality in the stationary case (c.f. Hahn and Kuersteiner, 2002). For this model, the following result holds and sheds light on the relative efficiency properties of the PFAE procedure, including both the stationary and unit root cases, in relation to the MLE.

**Theorem 3** *Suppose that* $\varepsilon_{it} = u_{it} - \rho u_{it-1}$ *is* $iid\ N(0, \sigma^2)$ *for some* $\rho \in (-1, 1]$. *Then*

(20) $(nT)^{1/2}(\hat{\rho} - \rho) \Rightarrow N(0, 1 - \rho^2)$, *as* $T \to \infty$ *if* $|\rho| < 1$,

(21) $n^{1/2}T(\hat{\rho} - 1) \Rightarrow N(0, 9)$, *as* $n, T \to \infty$ *if* $\rho = 1$.

**Remarks.**

1. Asymptotics for the stationary case (20) hold as $T \to \infty$ regardless of the cross sectional dimension $n$. We further note that asymptotic normality does not require large $T$. However, the form of the asymptotic variance given in (20) does require $T \to \infty$. In this case, LAN asymptotics apply as $T \to \infty$ and the variance attains the Cramér Rao bound, which is the same as in the stationary time series ($n = 1$) case. So, when $|\rho| < 1$, the PFAE is asymptotically efficient as $T \to \infty$. This result corresponds to the finding in Hahn and Kuersteiner (2002, theorem 3) that the bias corrected MLE attains the (semiparametric) efficiency bound for the estimation of the common autoregressive coefficient in the presence of fixed effects under the rate condition $0 < \lim_{n,T\to\infty} \frac{n}{T} < \infty$. However, the efficiency bound is attained for the PFAE without this rate condition and holds even for fixed $n$.

2. Hahn and Kuersteiner (2002, theorem 4) show that when $\rho = 1$ and $n, T \to \infty$, the (bias corrected) LSDV estimator $\widehat{\rho}_{lsdv}$ is asymptotically distributed as

   (22) $n^{1/2}T\left(\hat{\rho}_{lsdv} - 1 + \dfrac{3}{T+1}\right) \Rightarrow N\left(0, \dfrac{51}{5}\right)$.

   Thus, the PFAE estimator has smaller asymptotic variance than the bias-corrected LSDV estimator and the PFAE requires no bias correction. Observe that the LSDV estimator is the Gaussian MLE corrected for its asymptotic bias. So, the improvement of the PFAE over the bias corrected LSDV estimator at $\rho = 1$ is analogous to the improvement of the FAE

estimator over the MLE in the time series unit root case shown in HPS (2009). In that case, correcting for the bias by re-centering the MLE estimator about its mean does not reduce variation, whereas HPS (2009) show that the FAE estimator reduces both the asymptotic bias and the variance of the MLE not only at $\rho = 1$ but also in the vicinity of unity, while having the same limit theory in the stationary case. The limit result (21) reveals that the improvement of the FAE over the (levels) MLE at unity in the time series case carries over to the panel case where $n \to \infty$.

3. The improvement of the PFAE over the bias corrected LSDV estimator might be considered counterintuitive because differencing is usually regarded as inferior in terms of efficiency to levels estimation and the use of a within-group transformation to eliminate individual effects (unless GLS or maximum likelihod is applied to the differenced data). However, the considerable advantage of the PFAE technique is that it removes individual effects by systematic X-differencing and, in addition, because long differences are included in the stacked system estimation, any strong signal information in the data is retained by virtue of the full aggregation that is built into the estimator. The result is improved estimation in terms of both bias and efficiency over regression-based demeaning of the levels data and bias-correction in ML estimation.

4. Similarly, for the AR($p$) panel model, when $u_{it}$ is stationary, the PFAE is approximately equivalent to the bias-corrected OLS estimator. In this case bias rapidly disappears as the total sample size increases. When $u_{it}$ has a unit root, the PFAE has substantially smaller bias and no efficiency loss compared with the OLS estimator.

5. When $\rho = 1$, there is a simple relationship between the PFAE and the bias corrected MLE or LSDV estimator. In particular, as shown in Appendix D, when $\rho = 1$ and $\frac{\sqrt{n}}{T} \to 0$, we have

$$\sqrt{n}T\left(\hat{\rho} - 1\right) = \sqrt{n}T\left(\hat{\rho}_{lsdv} - 1 + \frac{3}{T}\right) + \sqrt{n}\frac{3\sum_i T^{-1}\left(\sum_{t=3}^{T} y_{it-1}\right)^2 - 2\sum_i \sum_{t=3}^{T} y_{it-1}^2}{\sum_i \sum_{t=3}^{T} \breve{y}_{it-1}^2} + o_p(1), \tag{23}$$

where $\breve{y}_{it-1} := y_{it-1} - T_2^{-1}\sum_{s=3}^{T} y_{is-1}$. According to (23), $\hat{\rho}$ may be interpreted as a modified version of the bias corrected form of $\hat{\rho}_{lsdv}$. The modification is important because the second term of (23) contributes to the limit distribution and leads to a reduction in the limiting variance of the LSDV estimator. In particular, it is the (negative) cor-

relation of the second term with the first term of (23) that reduces the asymptotic variance of LSDV, Avar$\{\sqrt{n}T\left(\hat{\rho}_{ls}-1+\frac{3}{T}\right)\}=51/5$, to the asymptotic variance of PFAE, Avar$\{\sqrt{n}T\left(\hat{\rho}_{fa}-1\right)\}=9$. In fact, this negative correlation makes it possible to lower the asymptotic variance further, as shown in Appendix D at least for $\rho=1$.

6. For the panel AR(1) model when $\rho=1$, using sequential limits as $n\to\infty$ followed by $T\to\infty$, Kruiniger (2008) showed that the first difference Gaussian quasi-MLE (called FDMLE; see also Hsiao et al., 2002) has the asymptotic distribution $n^{1/2}T(\hat{\rho}_{fdml}-1)\Rightarrow N(0,8)$. The limit distribution of the FDMLE for $|\rho|<1$ is $(nT)^{1/2}\left(\hat{\rho}_{fdml}-\rho\right)\Rightarrow N(0,1-\rho^2)$, comparable to (20). But when $\rho=1$ the variance of the limit distribution of the FDMLE is smaller than that of the PFAE. This reduction in variance is explained by the fact that the FDMLE is a *restricted* maximum likelihood estimator. The FDMLE is computed using a quasi-likelihod that is defined only for $\rho<1+\frac{2}{T-1}$ (see Kruiniger, 2008). So $\rho$ is restricted by the upper bound of this region at which point the quasi-likelihood becomes undefined. We use the term "quasi-likelihood" in describing the FDMLE because it is *not* the true likelihood. In fact, no data generating mechanism is given in Kruiniger (2008) for the case $\rho>1$ and the quasi likelihood is constructed over that region simply by taking an analytic extension to the region $\rho\in[1,1+\frac{2}{T-1})$ of the Gaussian likelihood based on the density of the differenced data over the stationary region $|\rho|<1$. The consequential restriction in domain, and hence in estimation, plays a key role in the variance reduction of the FDMLE. This reduction is borne out in simulations. For example, simulations with $n=200$, $T=50$ and $\rho=1$ show the variance of FDMLE to be approximately 87% of the variance of PFAE, which corresponds well with the limit theory variance ratio of $8/9\simeq 88.9\%$. Also, in view of the singularity in the quasi likelihood at the upper limit of the domain of definition, numerical maximization of the log-likelihood frequently encounters convergence difficulties in the computation of the FDMLE. Numerical optimization can fail if $\rho\simeq 1$ and $n$ is not large. For example, in simulations with $n=10$, $T=50$ and $\rho=1$, we found that a total 32 out of 1000 iterations failed to converge to a local optimizer. These restricted domain and convergence issues associated with the FDMLE procedure are discussed more fully in separate work (Han and Phillips, 2009b).

7. Asymptotics for the FDMLE procedure are developed in Kruiniger (2008) only for the panel AR(1) model and computation is much more difficult in the case of the panel AR(p) model. These limitations make it desirable to have a simple unrestricted estimator like PFAE with

good finite sample and asymptotic properties that can be easily implemented in general panel AR(p) models.

8. In the unit root case with $\rho = 1$, the limit distribution (21) holds for both $n, T \to \infty$, but no condition is required on the $n/T$ ratio. For $n = 1$, we know from the results in HPS (2009) that the (time series) MLE based on levels is not efficient and that remains true even when we bias correct the MLE. In fact, as shown in HPS (2009), the FAE is superior to the MLE in the whole vicinity of unity when $n = 1$. So, we can at least conclude that the PFAE is superior to the MLE for $n = 1$. We expect but do not prove that this conclusion holds for all fixed $n$.

The limit theory for the (restricted domain) FDMLE estimator at $\rho = 1$ indicates that there may be scope for improving estimation efficiency at $\rho = 1$ and possibly in the immediate neighborhood of unity. This issue is complex and, as indicated earlier, there is currently no general optimal estimation theory that can be applied to study this problem. In Appendix D we prove that a small modification to the PFAE procedure can indeed reduce variance for the case $\rho = 1$. The modification is of some independent interest because it makes use of the relationship (23) between PFAE and the bias-corrected LSDV estimator of Hahn and Kuersteiner (2002). In particular, in the simple panel AR(1) model (1), the modified estimator is obtained by taking the following linear combination for some scalar weight $\gamma$

$$\hat{\rho}^{+} = \gamma\hat{\rho} + (1-\gamma)(\hat{\rho}_{lsdv} + \tfrac{3}{T}) = \hat{\rho} - (1-\gamma)(\hat{\rho} - \hat{\rho}_{lsdv} - \tfrac{3}{T}), \tag{24}$$

so that the centred and scaled estimator has the form

$$n^{1/2}T(\hat{\rho}^{+} - 1) = n^{1/2}T(\hat{\rho}_{lsdv} - 1 + \tfrac{3}{T}) + n^{1/2}T\gamma(\hat{\rho} - \hat{\rho}_{lsdv} - \tfrac{3}{T}). \tag{25}$$

The PFAE corresponds to $\gamma = 1$. In this case, the (negative) correlation of the second term with the first term of (25) reduces the asymptotic variance of $n^{1/2}T(\hat{\rho}_{lsdv} - 1 + 3/T)$, which is $51/5$, down to the asymptotic variance of $n^{1/2}T(\hat{\rho} - 1)$, which is 9. The variance can be lowered further by choosing an optimal $\gamma$. According to the calculations shown in Appendix D, $\gamma = 5/8$ gives $n^{1/2}T(\hat{\rho}^{+} - 1) \Rightarrow N(0, 8.325)$, which is the minimal variance attainable by adjusting $\gamma$ in the relationship (25).

The modified estimator $\hat{\rho}^{+}$ can also be understood as a GMM estimator based on the two moment conditions $Eg_{1i}(\rho) = 0$ and $Eg_{2i}(\rho) \to 0$ at $\rho = 1$, where $g_{1i}(\rho)$ identifies $\widehat{\rho}$ and $g_{2i}(\rho)$

identifies $\hat{\rho}_{lsdv} + \frac{3}{T}$, i.e.,

$$g_{1i}(\rho) = \frac{1}{T_2^3} \sum_{t=4}^{T} \sum_{s=1}^{t-3} (y_{it-1} - y_{s+1}) \Big[ (y_{it} - y_{is}) - \rho (y_{it-1} - y_{is+1}) \Big],$$
$$g_{2i}(\rho) = \frac{1}{T_2^2} \sum_{t=3}^{T} \widetilde{y}_{it-1} \left[ \widetilde{y}_{it} - \left( \rho - \tfrac{3}{T} \right) \widetilde{y}_{it-1} \right],$$

with $\widetilde{y}_{it-1} = y_{it-1} - T_2^{-1} \sum_{s=3}^{T} y_{is-1}$, $\widetilde{y}_{it} = y_{it} - T_2^{-1} \sum_{s=3}^{T} y_{is}$, and $T_2 = T - 2$. Note that the first observations are ignored in $g_{2i}(\rho)$ for algebraic simplicity and their effect is asymptotically negligible when $T \to \infty$. In view of the identity (see HPS, 2009)

$$T_2^{-1} \sum_{t=4}^{T} \sum_{s=1}^{t-3} (y_{it-1} - y_{is+1})^2 = \sum_{t=3}^{T} \widetilde{y}_{it-1}^2$$

any weighted GMM estimator can be expressed in the form $\gamma \hat{\rho} + (1 - \gamma)(\hat{\rho}_{ls} + \frac{3}{T})$ for some $\gamma$, thereby leading back to the original formulation (24).

The modified PFAE $\hat{\rho}^+$ with $\gamma = 5/8$ attains an efficiency level of $8/8.325 = 0.96096$ (i.e., 96% efficiency) relative to the restricted FDMLE. However, this argument cannot be used for general $\rho$ values because $\hat{\rho}_{lsdv} + \frac{3}{T}$ does not correct the bias if $|\rho| < 1$ unless $n/T \to 0$. This is evident from the fact that

$$\sqrt{nT}(\hat{\rho}_{lsdv} + \tfrac{3}{T} - \rho) = \sqrt{nT}\,(\hat{\rho}_{hk} - \rho) + \frac{\sqrt{nT}}{T+1}(2 + \frac{3}{T} - \hat{\rho}_{hk}),$$

where $\hat{\rho}_{hk}$ is the bias corrected estimator proposed by Hahn and Kuersteiner (2002, p. 1645) for the stationary case, i.e., $\hat{\rho}_{hk} = \frac{T+1}{T}\,\hat{\rho}_{lsdv} + \frac{1}{T}$ such that $(nT)^{1/2}(\hat{\rho}_{hk} - \rho) \Rightarrow N(0, 1 - \rho^2)$ when $|\rho| < 1$ and $\lim n/T \in (0, \infty)$. Of course, when $n/T \to 0$ we also have $\sqrt{nT}(\hat{\rho}_{lsdv} + \frac{3}{T} - \rho) = \sqrt{nT}(\hat{\rho}_{lsdv} - \rho) + o_p(1)$, so in this event the bias is small because $T \to \infty$ so fast.

## 4.2 Lag Length Selection

When $T$ is large, lag length can be determined for each individual panel using conventional time series model selection methods. While this method is inevitably inefficient because it fails to take advantage of the panel structure and the pooling of information, it is still a consistent selection method when $T \to \infty$. When $T$ is small and $n$ is large as in microeconometric panels, this individual selection method is no longer available. In that case, the Sargan test combined with GMM methods (e.g., Arellano and Bond's GMM method) is often used instead. But the effectiveness of this method deteriorates when panel data manifests high persistence and there is substantial

individual heterogeneity because in such cases the AR coefficients are poorly (and possibly inconsistently) estimated. This section therefore proposes methods for consistent lag length selection which take advantage of the good asymptotic properties of the PFAE procedure.

### 4.2.1 Panel BIC Order Selection

Consistent estimation of $\rho$ by X-differencing and full aggregation allows us to construct information criteria to estimate the lag order in a panel model such as (10). In what follows, we will consider the panel BIC criterion and show some of its asymptotic properties.

For Gaussian autoregressive time series models, the BIC criterion takes the form

$$BIC\,(k) = \log \hat{\sigma}_k^2 + k(\log m)/m,$$

where $\hat{\sigma}_k^2$ is a error variance estimate calculated allowing for $k$ lags and $m$ is the sample size. Our proposed version of panel BIC uses this same formula for some deliberately designed $\hat{\sigma}_k^2$ and particularly chosen $m$. To be more precise, let $\tilde{\rho}^k = (\tilde{\rho}_1^k, \ldots, \tilde{\rho}_p^k)'$ denote the full aggregation estimator based on the equation

$$y_{it} - y_{is} = \sum_{j=1}^{k} \rho_j (y_{it-j} - y_{is+j}) + (\varepsilon_{it} - \varepsilon_{is}^*) \tag{26}$$

for $i = 1, \ldots, n$, $t = 2k_{\max} + 2, \ldots, T$, and $s = 1, \ldots, t - 2k_{\max} - 1$. The exact formula of $\tilde{\rho}^k$ is given in (64). It is important that (26) is aggregated as if we had $k_{\max}$ lags for all $k$ values. We call this operation 'full-aggregation after $k_{\max}$-truncation'. Thus, a total of $T_*(T_*+1)/2$ equations are aggregated (over $t$ and $s$) for each $i$, where $T_* = T - 2k_{\max} - 1$, and $\hat{\sigma}_k^2$ is defined as the residual sum of squares from PFAE estimation of (26) after $k_{\max}$-truncation, divided by $nT_*(T_*+1)/2$. The effective 'sample size' is $nT_*$ when $k_{\max}$ lags are allowed for, so we set $m = nT_*$. Hence, the information criterion is defined as

$$BIC(k) = \log \hat{\sigma}_k^2 + k(nT_*)^{-1} \log(nT_*), \tag{27}$$

where

$$\hat{\sigma}_k^2 = \frac{1}{nT_*(T_*+1)/2} \sum_{i=1}^{n} \sum_{t=2k_{\max}+2}^{T} \sum_{s=1}^{t-2k_{\max}-1} \left( \tilde{y}_{it,s} - \sum_{j=1}^{k} \tilde{\rho}_j^k \tilde{y}_{it-j,s+j} \right)^2, \tag{28}$$

using the notation $\tilde{y}_{it,s} = y_{it} - y_{is}$ defined earlier. The lag length (i.e., the maximal $p$ such that $\rho_p \neq 0$) is then consistently estimated by minimizing $BIC(k)$ over $k = 0, 1, \ldots, k_{\max}$ for some

finite $k_{\max}$, as shown in Appendix C. Note that this version of panel BIC loses the usual Gaussian log-likelihood interpretation.

In the above formulation of (27) and (28), the $k_{\max}$-truncation is important and is a key element in the consistency of the BIC approach especially when $T$ is small. This is true even for a simple panel AR model without fixed effects, where pooled ordinary least squares (OLS) estimates the autoregressive coefficients consistently. The following example of a panel AR(1) explains why the BIC method fails for small $T$ if the $k_{\max}$-truncation procedure is not implemented.

**Example: Panel BIC Lag Selection without $k_{\max}$- Truncation.**

Let $y_{it} = \rho_1 y_{it-1} + \varepsilon_{it}$ where $\varepsilon_{it}$ is $iid\ N(0, \sigma^2)$ and $\rho_1 \neq 0$, so that the true AR order is $p = 1$. Let $(\tilde{\rho}_{k,1}, \ldots, \tilde{\rho}_{k,k})'$ be the pooled OLS estimate of the coefficients in a fitted panel AR($k$) without the $k_{\max}$-truncation, i.e., $\tilde{\rho}_{1,1}$ is based on the pooled OLS regression of $y_{it}$ on $y_{it-1}$, $(\tilde{\rho}_{2,1}, \tilde{\rho}_{2,2})'$ is the coefficient in the pooled OLS regression of $y_{it}$ on $(y_{it-1}, y_{it-2})'$, and so on, using all available observations. Let the pooled sample error variance be $\tilde{\sigma}_k^2 = [n(T-k)]^{-1} \sum_{i=1}^{n} \sum_{t=k+1}^{T} \tilde{\varepsilon}_{k,it}^2$, where $\tilde{\varepsilon}_{k,it} = y_{it} - \sum_{j=1}^{k} \tilde{\rho}_{k,j} y_{it-j}$, so that $\tilde{\sigma}_k^2$ is consistent for $\sigma^2$ for all $k \geq p = 1$. Consider minimizing the information criterion

$$IC(k) = \log \tilde{\sigma}_k^2 + k[n(T-k)]^{-1} \log[n(T-k)].$$

Let $T = 3$ and $k_{\max} = 2$ for the purpose of illustration. Direct calculation for $k = 1, 2$ gives

$$IC(2) - IC(1) = \log(\tilde{\sigma}_2^2/\tilde{\sigma}_1^2) + (2n)^{-1}(3 \log n - \log 2), \tag{29}$$

where

$$\log(\tilde{\sigma}_2^2/\tilde{\sigma}_1^2) = \log\left(1 - \frac{\tilde{\sigma}_1^2 - \tilde{\sigma}_2^2}{\tilde{\sigma}_1^2}\right) = -\frac{\tilde{\sigma}_1^2 - \tilde{\sigma}_2^2}{\tilde{\sigma}_1^2} + o_p\left(\frac{\tilde{\sigma}_1^2 - \tilde{\sigma}_2^2}{\tilde{\sigma}_1^2}\right),$$

with

$$\frac{\tilde{\sigma}_1^2 - \tilde{\sigma}_2^2}{\tilde{\sigma}_1^2} = \frac{1}{n\tilde{\sigma}_1^2} \sum_{i=1}^{n} \left[\tfrac{1}{2}(\tilde{\varepsilon}_{1,i2}^2 - \tilde{\varepsilon}_{2,i3}^2) + \tfrac{1}{2}(\tilde{\varepsilon}_{1,i3}^2 - \tilde{\varepsilon}_{2,i3}^2)\right] := A_n.$$

Now $\tilde{\sigma}_1^2 \to_p \sigma^2$, $\tilde{\varepsilon}_{k,it} = \varepsilon_{it} + \sum_{j=1}^{k} (\rho_j - \tilde{\rho}_{k,j})\, y_{it-j}$ for both $k = 1, 2$, and $\varepsilon_{i2}^2 - \varepsilon_{i3}^2$ is $iid$ with zero mean $E\{\varepsilon_{i2}^2 - \varepsilon_{i3}^2\} = 0$ and finite variance $E\{\varepsilon_{i2}^2 - \varepsilon_{i3}^2\}^2 = 4\sigma^4$. It

follows that

$$n^{1/2}A_n = \frac{1}{\tilde{\sigma}_1^2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tfrac{1}{2}(\varepsilon_{i2}^2 - \varepsilon_{i3}^2) + o_p(1) \Rightarrow N(0,1).$$

Collecting terms and scaling (29) yields

$$\begin{aligned} n^{1/2}[IC(2) - IC(1)] &= -n^{1/2}A_n + o_p(1) + (3/2)n^{-1/2}\log n - O(n^{-1/2}) \\ &= -n^{1/2}A_n + o_p(1), \end{aligned}$$

which is asymptotically *negative* with a nontrivial probability of $1/2$ as $n \to \infty$. Thus, minimizing $IC(k)$ overestimates the lag order with a positive probability asymptotically as $n \to \infty$. ■

The $k_{\max}$-truncation resolves this problem and estimates $p$ consistently by minimizing BIC. (See Appendix C for a proof.) However this approach involves considerable data loss if $T$ is small and, in consequence, the finite sample performance of the $k_{\max}$-truncation BIC is not impressive. For example, when $k_{\max} = 4$ and $T = 10$, the effective time dimension after $k_{\max}$-truncation is $T_* = T - 2k_{\max} - 1 = 1$. In that case the BIC method reduces to a cross-sectional regression and requires large $n$ for reasonable performance. Table 3 provides some confirmatory simulation findings.

A possible solution to this data loss problem is to set a smaller $k_{\max}$ when $T$ is small. For example, reducing $k_{\max}$ by one increases the effective time dimension by two and thus the total number of effective observations increases by $2n$. So the performance of the $k_{\max}$ BIC selection method can be substantially influenced by the choice of the maximal lag length unless $T$ itself is large. This sensitivity is unnatural and undesirable because $k_{\max}$ is usually given only a purely nominal setting in applied work. As another solution, we might consider modifying the penalty or extending the analysis to other information criteria such as the posterior odds criterion (PIC; Phillips and Ploberger, 1994) for panel models. These possibilities are worth considering and can be pursued in separate work. Another option is to use general-to-specific selection methods, as we do in the following section.

### 4.2.2 General-to-Specific Significance Testing

An obvious alternative approach that avoids data loss is a general-to-specific sequential modeling procedure. This selection procedure can be implemented in the usual way. The sequence begins

by estimating the largest model—the panel AR($k_{\max}$) model for some given $k_{\max}$—and tests the significance of $\hat{\rho}_{k_{\max}}$. If the null hypothesis that $\rho_{k_{\max}} = 0$ is accepted at the chosen level, then the panel AR($k_{\max} - 1$) model is fitted and the null hypothesis $\rho_{k_{\max}-1} = 0$ is tested. This sequential process of estimating and testing is continued until the null hypothesis is rejected, and $\hat{p}$ is defined as the largest $k$ value such that the regressor $y_{it-k}$ is significant.

In this process, all available time series units are fully used. That is, for $k = 4, 3, 2, 1$, the numbers of the time series sample used in the regression are $T - 9, T - 7, T - 5, T - 3$, respectively. Hence, even when $T = 10$ and $k_{\max} = 4$, the general-to-specific approach will use more observations than the BIC procedure above for $k < k_{\max}$, and the resulting lag length estimate will generally be more accurate when $T$ is small. In implementing the sequential asymptotic tests, the asymptotic variances are estimated by the generalized heteroskedasticity-robust formula (Arellano, 1987; Kezdi, 2002) $Q_Z^{-1}\widehat{Q}_V Q_Z^{-1}$, where $Q_Z$ is defined in Theorem 2 and $\widehat{Q}_V$ is found in (19).

The general-to-specific methodology applies conventional statistical tests. So if the significance level for the tests is fixed, then the order estimator inevitably allows for a nonzero probability of overestimation. Furthermore, as is typical in sequential tests, this overestimation probability is bigger than the significance level when there are multiple steps between $k_{\max}$ and $p$ because the probability of false rejection accumulates as $k$ step downs from $k_{\max}$ to $p$.

These problems can be mitigated (and overcome at least asymptotically) by letting the level of the test be dependent on the sample size. More precisely, following Bauer, Pötscher and Hackl (1988), we can set the critical value $c_{nT}$ in such a way that (i) $c_{nT} \to \infty$, and (ii) $r_{nT}^{-1} c_{nT} \to 0$ as $n, T \to \infty$, where $r_{nT}$ is again the convergence rate of the full aggregation estimator. (Here, condition (i) prevents overestimation and condition (ii) prevents underestimation.) The critical value corresponds to the standard normal critical value for the significance level $\alpha_{nT} = 1 - \Phi(c_{nT})$, where $\Phi(\cdot)$ is the standard normal c.d.f. Conditions (i) and (ii) are equivalent to the requirement that the significance level $\alpha_{nT} \to 0$ and $-r_{nT}^{-1} \log \alpha_{nT} \to 0$ (proved in equation (22) of Pötscher, 1983).

If the significance level is too high, then test size increases and the lag length is usually overestimated. On the other hand, too small a level causes the model to be underfitted. Since parameters in a generous model are still consistently estimated while an underspecified model leads to some inconsistent coefficient estimates, practitioners are recommended to use a significance level that is not too small. A 1% level seems a reasonable choice in cases where the sample sizes are moderate to large. Simulations may be used to explore the performance of various significance level choices in relation to the cross section and time series sample sizes. Such experiments would need to be

extensive and to cover many different models to be valuable beyond a simple procedure such as the 1% rule. They would form a useful subsequent research project.

# 5 Simulations

This section reports simulations which shed light on the finite sample properties of our procedures in relation to existing methods of dynamic panel estimation. In particular, we compare the PFAE procedure with existing estimators such as Arellano and Bond's (1991) difference GMM estimator and Blundell and Bond's (1998) system GMM estimator for a panel AR(2) model. (The FDMLE method is not included because of computational difficulties with this procedure and the fact that it is a restricted estimator, as discussed earlier.) We then compare the performance of two alternative lag-length selection methods – the $k_{\max}$-truncated BIC procedure and general-to-specific testing.

**I. Comparison of bias and efficiency: AR(1).** We first compare the properties of the PFAE with the LSDV estimator (which is inconsistent), Hahn and Kuersteiner's bias-corrected LSDV estimator (HK), the one-step first difference GMM (GMM1/DIF), and the two-step system GMM (GMM2/SYS), for the panel AR(1) model. The model is $y_{it} = a_i + u_{it}$, $u_{it} = \rho u_{it-1} + \varepsilon_{it}$, where $\varepsilon_{it}$ is $iid$ standard normal variables and $a_i$ is also normal with $E(a_i)$ arbitrarily set to 2. When generating the data, the processes are initialized at $t = -100$ such that $u_{i,-100} := 0$, and then observations for $t \leq 0$ are discarded. The normal variates are generated using the `rnormal` function of Stata. The difference GMM and the system GMM are estimated by the 'xtabond' and the 'xtdpdsys' commands of Stata respectively, and the PFAE is obtained by direct calculation using formula (16).

Table 1 reports the simulated means of the estimators from 1,000 replications. The LSDV estimator is obviously biased downward, as per Nickell (1981). The (small sample) biases of the first difference and system GMM estimators depend on the distribution of $\alpha_i$. On the other hand, PFAE shows very little bias for all parameter values and is considerably superior to HK.

Table 1 also presents simulated variances of the estimators. When $T$ is small ($T = 10$), PFAE is less efficient than the bias-corrected LSDV estimator (HK), but when $T$ is larger ($T = 20$) and $\rho$ is large, PFAE is as efficient or more efficient than HK. With larger $T$ values, PFAE attains the asymptotic variance $(nT)^{-1}(1-\rho^2)$, as does the HK estimator. For $T = 20$, we notice that PFAE appears less efficient than HK at $\rho = 1$, which looks contrary to the asymptotic finding that $n^{1/2}T(\hat{\rho}_{hk} - 1) \Rightarrow N(0, 51/5)$ and $n^{1/2}T(\hat{\rho}_{fa} - 1) \Rightarrow N(0, 9)$ with $\hat{\rho}_{hk}$ and $\hat{\rho}_{fa}$ respectively

denoting the HK and PFAE estimators. This outcome occurs because $T = 20$ is not large enough for the asymptotics to be accurate without a degrees of freedom adjustment. For $\rho = 1$, the asymptotic variance of $\hat{\rho}_{fa}$ is $9/n(T-2)^2$, which is approximately $0.27\dot{7} \times 10^{-3}$ with $n = 100$ and $T = 20$. This theoretical value is close to the simulated variance $0.273 \times 10^{-3}$. As $T$ increases further, so that $T^2/(T-2)^2$ is close to 1, we expect the higher asymptotic efficiency of PFAE relative to HK to become evident in simulations. Table 2 reveals that this expected improvement occurs for $T \geq 80$ for all values of $n$.

The performance of the GMM estimators differs as $sd(a_i)$ changes. Comparing PFAE and GMM, PFAE performs uniformly better than the GMM estimators in our simulations except for $\rho = 1$ with $T = 10$. It is however worth noting that the GMM estimators are based on moment conditions different from those used by PFAE and LSDV, and that the performance of the GMM estimators also depends on the inital cross sectional variance of the idiosyncratic errors.

**II. Comparison of bias and efficiency: AR(2).** We next consider an AR(2) dynamic panel model (i.e., $y_{it} = a_i + u_{it}$, $u_{it} = \rho_1 u_{it-1} + \rho_2 u_{it-2} + \varepsilon_{it}$). Except for $u_{it}$ being AR(2), all other settings are the same as in the previous simulation. We set $\rho_2 = -0.2$, and $\rho_1 = 0.2, 0.5, 0.7, 0.9, 1.1$ and 1.2. The panels are stationary when $\rho_1 < 1.2$, and are integrated when $\rho_1 = 1.2$.

Table 3 reports the simulated means and varinaces of the estimates of $\rho_1$. Note that Hahn and Kuersteiner's (2002) estimator is not examined because the model is not AR(1), so one of their assumptions is violated. The LSDV estimator is again biased downward, and the PFAE exhibits very low finite sample bias. The GMM estimator performance depends on the variance of $a_i$. Again, LSDV and PFAE are free from the effects of the $a_i$, while the two GMM estimators are not. The PFAE performs well in all considered cases. As remarked in the discussion of the AR(1) simulations, it is noteworthy that the accuracy of the GMM estimators depends on the variance of the initial idiosyncratic errors as well.

**III. Inference.** We next investigate the properties of the estimated variance $Q_Z^{-1}\widehat{Q}_V Q_Z^{-1}$ of the PFAE, where

$$Q_Z = \sum_{i=1}^{n} \sum_{t=2p+2}^{T} \sum_{s=1}^{t-2p-1} \widetilde{Z}_{it,s}\widetilde{Z}'_{it,s} \text{ and } \widehat{Q}_V = \sum_{i=1}^{n} \widehat{V}_{iT}\widehat{V}'_{iT},$$

with $\widetilde{Z}_{it,s}$ defined right after Lemma 1 and $\widehat{V}_{iT}$ found in (19).

Because all the statistics are free from individual effects, we can eliminate $a_i$ from the data generation process. Model evaluation for AR($p$) models with $p > 1$ will be considered later while

simulating lag length selection. So here we focus on the panel AR(1) model $y_{it} = \rho y_{it-1} + \varepsilon_{it}$, where $\varepsilon_{it} \sim N(0, \sigma^2)$ with $\sigma^2 = 1$. We test (i) $H_0 : \rho = 0$ and (ii) $H_0 : \rho = 1$. We present test sizes for the null hypothesis that the $\rho$ parameter is the same as the true parameter used in the data generation. Gauss was used for the simulations. We use the $t_{n-1}$ critical values in testing, as recommended by Hansen's (2007).

Table 4 reports the empirical sizes from a simulation of 5,000 replications. Except for a slight over-rejection in small samples with high $\rho$, size performance is reasonably good. The simulated powers for the null hypotheses $H_0 : \rho = 0$ (left) and $H_0 : \rho = 1$ (right) are presented in Table 5. This part of the simulation is intended to be illustrative as its main purpose is to exhibit general performance characteristics of inference with the PFAE procedure. Thorough comparisons with other estimators would require a more systematic simulation study.

**IV. Lag length selection.** Table 6 reports the finite sample performance of the BIC and general-to-specific (GS) approaches to model selection, based on the PFAE methodology, as explained in Section 4. The BIC method is seen to be consistent as $n, T \to \infty$, while the probability of over-estimation by the GS method does not diminish to zero for a given significance level (and this probability is larger than the significance level because of the accumulation of the type I errors). But as the significance level shrinks to zero as described in Section 4.2, the GS lag order estimator is consistent. In small samples, the performance of the GS method is generally much better than the BIC method.

# 6 Conclusion

The estimation method introduced in this paper for linear dynamic panel models uses a new differencing procedure called X-differencing to eliminate fixed effects and a simple technique of stacked and pooled least squares on the full system of X-differenced equations. The method is therefore straightforward to implement in practical work. It is also free from bias for all parameter values and avoids weak instrumentation problems in unit root and near unit root cases. The asymptotic theory shows gains in efficiency in the unit root case over bias-corrected maximum likelihood and equivalent efficiency in the stationary case but the new method has no need for bias correction. The asymptotics also apply irrespective of the $n/T$ ratio as $n, T \to \infty$. These advantages make the new estimation procedure attractive for empirical research, especially in cases of data persistence and dispersed individual effects where other methods can perform poorly.

The findings of the present paper point the way to further research. First, there is a need for a theory of optimal estimation in panel models which allows for roots in the vicinity of unity and dual index asymptotics. While there is, as yet, no optimal estimation theory in time series autoregression that includes the unit root case, the process of cross section averaging in panel estimation leads to important simplifications in the limit theory that make such an optimality theory feasible. In particular, the limit theory belongs to an asymptotically normal (as distinct from a nonstandard distribution) family when $n \to \infty$. But the limit distribution can also be degenerate with a singularity in the covariance structure and a change in the convergence rate when there is an autoregressive unit root. These features of the limit theory and their impact on optimality in estimation deserve detailed study. As indicated earlier, there is also scope for further work on model selection in dynamic panels, including an extensive numerical study of sequential testing rules and a further analysis of the asymptotic behavior of various information criteria.

Second, consistent estimation of panel autoregressions using X-differencing and PFAE methods is useful in the estimation of more general panel models with additional regressors. For example, in parametric models with exogenous regressors and AR($p$) errors such as $y_{it} = a_i + \beta' x_{it} + u_{it}$, with $u_{it} = \sum_{j=1}^{p} \rho_j u_{it-j} + \varepsilon_{it}$, we can consistently estimate $\rho = (\rho_1, \ldots, \rho_p)'$ using PFAE and residuals based on a preliminary consistent estimate of $\beta$. Then, a parametric feasible GLS estimate can be conducted as a natural extension of Bhargava, Franzini and Narendranathan's (1982) treatment of the AR(1). Such stepwise estimation of $\beta$ and $\rho$ may be iterated until convergence, combining moment conditions for $\beta$ based on assumed exogeneity of $x_{it}$ and the moment conditions implied by Lemma 1 using $y_{it} - \beta' x_{it}$ for given $\beta$.

Finally, direct treatment of dynamic models with exogenous regressors of the form $y_{it} = \alpha_i + \sum_{j=1}^{p} \rho_j y_{it-j} + \beta' x_{it} + \varepsilon_{it}$ is also possible using the methods of this paper. Transforming $x_{it} \mapsto x^*_{it}(\rho)$ where $x^*_{it}(\rho)$ is defined by $x_{it} \equiv x^*_{it}(\rho) - \sum_{j=1}^{p} \rho_j x^*_{it-j}(\rho)$ enables the model to be rewritten in latent form as $y_{it} = a_i + x^*_{it}(\rho)'\beta + u_{it}$, where $u_{it} = \sum_{j=1}^{p} \rho_j u_{it-j} + \varepsilon_{it}$. Then, the parameters $\beta$ and $\rho$ are identified by the exogeneity of $x^*_{it}(\rho)$ and the moment conditions of Lemma 1 for $y_{it} - x^*_{it}(\rho)'\beta$. The parameters may then be jointly estimated by an extended non-linear in parameters version of the PFAE approach[2]. Full exploration of this extension is an important future research topic.

[2]More specifically, the moment conditions $E\{\tilde{U}_{it-k,s+k}(\beta,\rho)\ [\tilde{U}_{it,s}(\beta,\rho) - \sum_{j=1}^{p} \tilde{U}_{it-j,s+j}(\beta,\rho)\ ]\ \} = 0$for $s < t - 2p$ and for all $k = 1, \ldots, p$, where $\tilde{U}_{it,s}(\beta,\rho) := [y_{it} - x^*_{it}(\rho)'\beta] - [y_{is} - x^*_{is}(\rho)'\beta]$, identify $\rho$ for given $\beta$ (using the PFAE method), and an exogeneity condition such as $Ex_{is}\varepsilon_{it} = 0$ for $s \leq t$ implies that $E[x_{is}(\Delta y_{it} - \sum_{j=1}^{p} \rho_j \Delta y_{it-j} - \beta' \Delta x_{it})] = 0$ for $s < t$, which identifies $\beta$ given $\rho$ if $x_{it-1}$ and $\Delta x_{it}$ are correlated.

# Appendix A: Technical Results and Proofs

**Proof of (4).** Because $\varepsilon_{is}^* = y_{is} - \alpha_i - \rho_1 y_{is+1}$, we have

$$Ey_{is+1}\varepsilon_{is}^* = Ey_{is+1}y_{is} - Ey_{is+1}\alpha_i - \rho_1 Ey_{is+1}^2.$$

Replacing the first $y_{is+1}$ on the right hand side with $\alpha_i + \rho_1 y_{is} + \varepsilon_{is+1}$, we get

$$Ey_{is+1}\varepsilon_{is}^* = Ey_{is}\alpha_i + \rho_1 Ey_{is}^2 - Ey_{is+1}\alpha_i - \rho_1 Ey_{is+1}^2$$

Because $Ey_{it}\alpha_i$ is the same for all $t$ and $Ey_{is}^2 = Ey_{is+1}^2$, we have $Ey_{is+1}\varepsilon_{is}^* = 0$. ■

**Proof of (7).** It is simpler to work with $u_{it} = y_{it} - a_i$, where $u_{it} = \rho_1 u_{it-1} + \varepsilon_{it}$. We shall show that $A := E(u_{it-1} - u_{is+1})(\varepsilon_{it} - \varepsilon_{is}^*) = 0$. For $s+1 < t$, we have

$$\begin{aligned} A &= E(u_{it-1} - u_{is+1})\varepsilon_{it} - E(u_{it-1} - u_{is+1})(u_{is} - \rho_1 u_{is+1}) \\ &= -E(u_{it-1} - u_{is+1})(u_{is} - \rho_1 u_{is+1}) \\ &= -Eu_{it-1}u_{is} + \rho_1 Eu_{it-1}u_{is+1} + Eu_{is}u_{is+1} - \rho_1 Eu_{is+1}^2 \\ &= -\rho_1 Eu_{it-2}u_{is} + \rho_1 Eu_{it-1}u_{is+1} + \rho_1 Eu_{is}^2 - \rho_1 Eu_{is+1}^2, \end{aligned}$$

where the last equality is derived by expanding $u_{it-1} = \rho_1 u_{it-2} + \varepsilon_{it-1}$ and $u_{is+1} = \rho_1 u_{is} + \varepsilon_{is+1}$. When $|\rho_1| < 1$, $u_{it}$ is stationary, so $A$ is obviously zero. If $\rho_1 = 1$, then $Eu_{it}u_{is} = Eu_{is}^2$ for $s \leq t$, so when $s \leq t-2$, we have

$$A = -\rho_1 Eu_{is}^2 + \rho_1 Eu_{is+1}^2 + \rho_1 Eu_{is}^2 - \rho_1 Eu_{is+1}^2 = 0$$

as claimed. ■

We prove Lemma 1 using $u_{it} = y_{it} - a_i$. Note that $u_{it} = \sum_{j=1}^p \rho_j u_{it-j} + \varepsilon_{it}$ where $\varepsilon_{it}$ is white noise $(0, \sigma_i^2)$. We also have $\varepsilon_{is}^* = u_{is} - \sum_{j=1}^p \rho_j u_{is+j}$. We first establish the following general lemma.

**Lemma 4** *Let $u_{it}$ be a panel AR(p) process such that $\Delta^m u_{it}$ is stationary AR(p−m) for some non-negative integer $m \leq p$, where $\Delta := 1 - L$. Then for all $t$ and $s$ such that $t > s$, $E\varepsilon_{is-m}^* \Delta^m u_{it} = 0$.*

**Proof.** First consider the case where $u_{it}$ is covariance stationary AR($p$), i.e., $m = 0$. Let $\gamma_j = Eu_{it}u_{it-j}/\sigma_i^2$. Let $\rho(L) = 1 - \rho_1 L - \cdots - \rho_p L^p$. We have

$$Eu_{it}\varepsilon_{is}^* = Eu_{it}\left(u_{is} - \textstyle\sum_{j=1}^p \rho_j u_{is+j}\right) = \sigma_i^2\left(\gamma_{t-s} - \textstyle\sum_{j=1}^p \rho_j \gamma_{t-s-j}\right) = 0$$

by the Yule-Walker equations when $t > s$ as claimed. Now for general $m \leq p$, we have $\rho(L) = (1-L)^m \rho^*(L)$, where $\rho^*(L) = 1 - \rho_1^* L - \cdots - \rho_{p-m}^* L^{p-m}$ and the roots of $\rho^*(L) = 0$ are outside the unit circle. First note that $\varepsilon_{is}^* = \rho(L^{-1})u_{is}$, so using $(1-L^{-1})^m u_{is-m} = (-1)^m (1-L)^m u_{is}$ and $\Delta := 1 - L$, we have

$$\varepsilon_{is-m}^* = \rho^*(L^{-1})(1-L^{-1})^m u_{is-m} = (-1)^m \rho^*(L^{-1}) \Delta^m u_{is} =: \tilde{\varepsilon}_{is}^*.$$

That is, $\rho^*(L^{-1})\tilde{u}_{is} = (-1)^m \tilde{\varepsilon}_{is}^*$, where $\tilde{u}_{is} = \Delta^m u_{is}$. Furthermore, $\tilde{u}_{it}$ is stationary AR($p-m$) by assumption, and by the result for the stationary case, we have $E(-1)^m \tilde{\varepsilon}_{is}^* \tilde{u}_{it} = 0$ for all $s < t$. The result follows by writing $\tilde{\varepsilon}_{is}^* = \varepsilon_{is-m}^*$ and $\tilde{u}_{it} = \Delta^m u_{it}$. ■

Lemma 1 is now straightforward.

**Proof of Lemma 1.** Because $u_{it} - u_{is} = y_{it} - y_{is}$ for all $s$ and $t$, we shall prove that $E(u_{it-k} - u_{is+k})(\varepsilon_{it} - \varepsilon_{is}^*) = 0$ for all $s < t - p$. Because $E(u_{it-k} - u_{is+k})\varepsilon_{it} = 0$ for all $s < t - p$ and $1 \leq k \leq p$, it suffices to show that $E(u_{it-k} - u_{is+k})\varepsilon_{is}^* = 0$ for such $s$ and $k$. If $u_{it}$ is stationary AR($p$), then this holds because of Lemma 4 with $m = 0$. If $u_{it}$ is $I(1)$ and $\Delta u_{it}$ is stationary AR($p-1$), then the result follows from Lemma 4 for $m = 1$ because $y_{it-k} - y_{is+k} = u_{it-k} - u_{is+k} = \Delta u_{it-k} + \cdots + \Delta u_{is+k+1}$. ■

Next we prove Theorem 2.

We first introduce some useful notation and transformations that facilitate analysis of the unit root case.

Let $V_{iT} = \frac{1}{T} \sum_{t=2p+2}^{T} \sum_{s=1}^{t-2p-1} \widetilde{Z}_{it,s} \widetilde{Z}_{it,s}'$ and $\eta_{iT} = \frac{1}{T} \sum_{t=2p+2}^{T} \sum_{s=1}^{t-2p-1} \widetilde{Z}_{it,s} \widetilde{\varepsilon}_{it,s}$, where $\widetilde{Z}_{it,s}$ and $\widetilde{\varepsilon}_{it,s}$ are defined in (14). Define the $p \times p$ transformation matrix $F$ and its inverse $F^{-1}$ as follows

$$F = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}, \quad F^{-1} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 1 & \cdots & 1 \\ 0 & 0 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{30}$$

Note that $F'z = (z_1, z_2 - z_1, \ldots, z_p - z_{p-1})'$ for any $z = (z_1, \ldots, z_p)'$, and $F^{-1}\rho = (\sum_{j=1}^{p} \rho_j, \sum_{j=2}^{p} \rho_j, \ldots, \rho_p)'$ for any $\rho = (\rho_1, \ldots, \rho_p)'$. These transformation matrices are needed for the unit root case. Also let

$$D_T = \begin{cases} T^{1/2} I & \text{if } u_{it} \sim I(0), \\ \operatorname{diag}(T, T^{1/2}, \ldots, T^{1/2}) & \text{if } u_{it} \sim I(1) \text{ and } \Delta u_{it} \sim I(0). \end{cases} \tag{31}$$

For a uniform development of the asymptotic theory, we derive the limit distribution of the standardized and centered quantity $n^{1/2}D_TF^{-1}(\hat{\rho}-\rho)$ in what follows. Note that

$$(32)\quad n^{1/2}D_TF^{-1}(\hat{\rho}-\rho)=A_{nT}^{-1}b_{nT},$$

where

$$(33)\quad A_{nT}=\frac{1}{n}\sum_{i=1}^{n}D_T^{-1}F'V_{iT}FD_T^{-1}\ \text{ and }\ b_{nT}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}D_T^{-1}F'\eta_{iT}.$$

Let $C_{nT}=n^{-1}\sum_{i=1}^{n}D_T^{-1}F'\eta_{iT}\eta_{iT}'FD_T^{-1}$.

**Theorem 5** *If $u_{it}$ is stationary AR(p) or if $u_{it}\sim I(1)$ and $\Delta u_{it-1}$ is stationary AR($p-1$), then under Conditions A and B(i), the following results hold:*

*(a) If $n\to\infty$ and $T$ is fixed,*

$$n^{1/2}D_TF^{-1}(\hat{\rho}-\rho)\Rightarrow N(0,A_T^{-1}C_TA_T^{-1}),$$

*where $A_T:=\lim_{n\to\infty}EA_{nT}=\operatorname{plim}_{n\to\infty}A_{nT}$ and $C_T:=\lim_{n\to\infty}EC_{nT}=\operatorname{plim}_{n\to\infty}C_{nT}$.*

*(b) If $n,T\to\infty$ jointly*

$$n^{1/2}D_TF^{-1}(\hat{\rho}-\rho)\Rightarrow N(0,A^{-1}CA^{-1}),$$

*where $A=\lim_{T\to\infty}A_T=\lim_{n,T\to\infty}EA_{nT}$, and $C=\lim_{T\to\infty}C_T=\lim_{n,T\to\infty}EC_{nT}$.*

*(c) If $T\to\infty$ and $n\geq 1$ is fixed, and if $u_{it}$ is stationary AR(p)*

$$n^{1/2}D_T^{-1}F'(\hat{\rho}-\rho)\Rightarrow N\big(0,\lambda_n^2(F'\Gamma F)^{-1}\big),\quad \Gamma=\sigma_i^{-2}E(X_{it-1}X_{it-1}'),$$

*where $X_{it-1}=(u_{it-1},\ldots,u_{it-p})'$, and $\lambda_n^2=\sum_{i=1}^{n}\sigma_i^4/(\sum_{i=1}^{n}\sigma_i^2)^2$.*

*(d) If $T\to\infty$ and $n\geq 1$ is fixed, and if $u_{it}\sim I(1)$ and $\Delta u_{it-1}$ is stationary AR($p-1$)*

$$n^{1/2}D_TF^{-1}(\hat{\rho}-\rho)\Rightarrow\left[\frac{\sqrt{n}(\pi'\rho)\sum_{i=1}^{n}\sigma_i^2Y_{bi}}{\sum_{i=1}^{n}\sigma_i^2Y_{ai}},\ Z_n'\right]',$$

*with*

$$Y_{ai}=\int_0^1 W_i(r)^2dr-\left[\int_0^1 W_i(r)dr\right]^2,$$
$$Y_{bi}=\int_0^1 W_i(r)dW_i(r)-\int_0^1 W_i(r)[1-W_i(r)]dr,$$

*where $W_i(\cdot)$ are independent standard Brownian motions, $Z_n\sim N(0,\lambda_n^2\Omega^{-1})$, $\Omega$ is the variance-covariance matrix of $(\Delta u_{it-1},\ldots,\Delta u_{it-p+1})'$, and $W_i(\cdot)$ and $Z_n$ are independent.*

The proof of (a) is straightforward and is given first. Let $E(\sigma_i^k) := \lim_{n\to\infty} n^{-1}\sum_{i=1}^n \sigma_i^k$.

**Proof of Theorem 5 (a).** We consider the numerator and denominator of (32) separately.

(i) Denominator: Note that $EV_{iT}^\circ := EV_{iT}/\sigma_i^2$ is identical for all $i$. Also $EV_{iT}^\circ$ is finite due to the uniformly finite fourth moment assumption for $\varepsilon_{it}/\sigma_i$. So

$$\frac{1}{n}\sum_{i=1}^n EV_{iT} = \left(\frac{1}{n}\sum_{i=1}^n \sigma_i^2\right) EV_{1T}^\circ \to E(\sigma_i^2)EV_{1T}^\circ := A_T^o, \tag{34}$$

where $E(\sigma_i^2) := \lim_{n\to\infty} n^{-1}\sum_{i=1}^n \sigma_i^2$ and $D_T^{-1}F'A_T^oFD_T^{-1} = A_T$**.** The uniform boundedness of $E\varepsilon_{it}^4$ implies that $E[V_{iT}(j,k)^2]$ is bounded uniformly over all $i$ for all $j$ and $k$, where $V_{iT}(j,k)$ is the $(j,k)$ element of $V_{iT}$, so

$$\text{var}\left[\frac{1}{n}\sum_{i=1}^n V_{iT}(j,k)\right] \le \frac{1}{n^2}\sum_{i=1}^n E[V_{iT}(j,k)^2] = O(n^{-1}).$$

Thus the denominator converges to the right hand side of (34) in mean and therefore in probability. The equivalence of $A_T$ and $\text{plim}_{n\to\infty} A_{nT}$ is also implied straightforwardly.

(ii) Numerator: We have $E\eta_{iT} = 0$ by Lemma 1**.** Condition A implies the convergence of $n^{-1}\sum_{i=1}^n E\eta_{iT}\eta_{iT}'$. The Lindeberg condition holds since $\sigma_i^{-2}\eta_{iT}$ is $iid$ and $\sigma_i^2$ is bounded under the uniform finite fourth moment condition. Thus $n^{-1/2}\sum_{i=1}^n \eta_{iT} \Rightarrow N(0, C_T^o)$, where $C_T^o := \lim_{n\to\infty} n^{-1}\sum_{i=1}^n E\eta_{iT}\eta_{iT}'$ and $C_T = D_T^{-1}F'C_T^oFD_T^{-1}$. The result for $b_{nT}$ follows immediately. That $C_T = \text{plim}_{n\to\infty} C_{nT}$ is implied by Condition B(ii). ■

The remaining parts of Theorem 5 involve $T \to \infty$, and we proceed by approximating the components of $\hat\rho - \rho$ by simpler terms. Let $X_{it-1} = (u_{it-1}, \ldots, u_{it-p})'$, $X_i = (X_{i0}, \ldots, X_{iT-1})'$, and $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iT})'$. Let $M_1 = I_T - T^{-1}1_T1_T'$, where $1_T$ is a $T$-vector with unit elements. Let $F$ and $D_T$ be defined by (30) and (31), respectively. Let $\Pi = \text{diag}(1, 2, \ldots, p)$. Also let

$$\psi_{iT}^{(j,k)} = \frac{1}{T}\sum_{t=2p+2}^{T}\sum_{s=1}^{t-2p-1}(u_{it-j} - u_{is+j})(u_{it-k} - u_{is+k}), \quad j,k = 0,1,\ldots,p,$$

so $\hat\rho = (\sum_{i=1}^n \Psi_{iT}^{den})^{-1}\sum_{i=1}^n \Psi_{iT}^{num}$, where

$$\Psi_{iT}^{den} = \begin{bmatrix} \psi_{iT}^{(1,1)} & \cdots & \psi_{iT}^{(1,p)} \\ \vdots & & \vdots \\ \psi_{iT}^{(p,1)} & \cdots & \psi_{iT}^{(p,p)} \end{bmatrix} \text{ and } \Psi_{iT}^{num} = \begin{bmatrix} \psi_{iT}^{(1,0)} \\ \vdots \\ \psi_{iT}^{(p,0)} \end{bmatrix}. \tag{35}$$

(Thus we have $V_{iT} = \Psi_{iT}^{den}$ and $\eta_{iT} = \Psi_{iT}^{num} - \Psi_{iT}^{den}\rho$.) Let $T_m = T - m$ for notational brevity. We first approximate $\psi_{iT}^{(j,k)}$ as shown in the following result:

$$\begin{aligned}(36)\quad \psi_{iT}^{(j,k)} &= \frac{T_{j+k}}{T}\sum_{t=j+k+1}^{T} u_{it-j}u_{it-k} - \frac{1}{T}\sum_{s=j+k+1}^{T} u_{is-j}\sum_{t=j+k+1}^{T} u_{it-k} + R_{1,iT}^{(j,k)},\\ &= \frac{T_{j+k}}{T}\sum_{t=1}^{T} u_{it-j}u_{it-k} - \frac{1}{T}\sum_{s=1}^{T} u_{is-j}\sum_{t=1}^{T} u_{it-k} + R_{1,iT}^{(j,k)} + R_{2,iT}^{(j,k)} + R_{3,iT}^{(j,k)},\end{aligned}$$

where

$$\begin{aligned}(37)\quad R_{1,iT}^{(j,k)} &= \frac{1}{2T}\sum_{t=j+k+1}^{T}(u_{it-j} - u_{it-k})^2 - \frac{1}{2T}\left[\sum_{t=j+k+1}^{T}(u_{it-j} - u_{it-k})\right]^2\\ &\quad - \frac{1}{T}\sum_{\ell=1}^{2p-j-k}\sum_{t=j+k+1+\ell}^{T}(u_{it-j} - u_{it-k+\ell})(u_{it-k} - u_{it-j+\ell}),\end{aligned}$$

$$(38)\quad R_{2,iT}^{(j,k)} = -\frac{T_{j+k}}{T}\sum_{t=1}^{j+k} u_{it-j}u_{it-k} - \frac{1}{T}\sum_{s=1}^{j+k} u_{is-j}\sum_{t=1}^{j+k} u_{it-k},$$

$$(39)\quad R_{3,iT}^{(j,k)} = \frac{1}{T}\sum_{t=1}^{T} u_{it-j}\sum_{s=1}^{j+k} u_{is-k} + \sum_{s=1}^{j+k} u_{is-j}\frac{1}{T}\sum_{t=1}^{T} u_{it-k}.$$

**Proof of (36).** Let $j \leq k$. Let $r = k - j$. We derive the first line of (36) for given $j$ and $r$. Let $f_{ts}^r = (u_{it} - u_{is})(u_{it-r} - u_{is+r})$ omitting the $i$ subscript. We have

$$\begin{aligned}T\psi_{iT}^{(j,j+r)} &= \sum_{t=2p+2}^{T}\sum_{s=1}^{t-2p-1} f_{t-j,s+j}^r = \sum_{t=2p+2-j}^{T-j}\sum_{s=1}^{t-2p+j-1} f_{t,s+j}^r = \sum_{t=2p+2-j}^{T-j}\sum_{s=j+r+1}^{t-2p+2j+r-1} f_{t,s-r}^r\\ &= \sum_{s=j+r+1}^{T+j+r-2p-1}\sum_{t=s+2p-2j-r+1}^{T-j} f_{t,s-r}^r = \sum_{t=j+r+1}^{T+j+r-2p-1}\sum_{s=t+2p-2j-r+1}^{T-j} f_{t,s-r}^r.\end{aligned}$$

The second and third identities above are obtained by letting $t' = t - j$ and $s' = s + j + r$, respectively, and then removing the dashes. The first identity of the second line is obtained by rearranging terms, and the last identity is obtained by swapping $t$ and $s$ and then noting $f_{s,t-r}^r = f_{t,s-r}^r$. The right hand side on the first line and the right hand side term on the second line together yield

$$\begin{aligned}2T\psi_{iT}^{(j,j+r)} &= \sum_{t=2p+2-j}^{T-j}\sum_{s=j+r+1}^{t-2p+2j+r-1} f_{t,s-r}^r + \sum_{t=j+r+1}^{T+j+r-2p-1}\sum_{s=t+2p-2j-r+1}^{T-j} f_{t,s-r}^r\\ &= \sum_{t=j+r+1}^{T-j}\sum_{s=j+r+1}^{T-j} f_{t,s-r}^r - \sum_{t=j+r+1}^{T-j} f_{t,t-r}^r - 2\sum_{\ell=1}^{2p-2j-r}\sum_{t=j+r+1+\ell}^{T-j} f_{t,t-\ell-r}^r.\end{aligned}$$

Note that $f^r_{t,t-\ell-r} = f^r_{t-\ell,t-r}$. Transforming by $t' = t + j$ and $s' = s + j$, then removing the dashes from $t'$ and $s'$, we get

$$2T\psi_{iT}^{(j,k)} = \sum_{t=m+1}^{T}\sum_{s=m+1}^{T} f^r_{t-j,s-k} - \sum_{t=m+1}^{T} f^r_{t-j,t-k} - 2\sum_{\ell=1}^{2p-m}\sum_{t=m+1+\ell}^{T} f^r_{t-j,t-k-\ell},$$

where $k = j + r$ and $m = 2j + r = j + k$. We have

$$\begin{aligned}
f^r_{t-j,s-k} &= (u_{it-j} - u_{is-k})(u_{it-k} - u_{is-j}) \\
&= u_{it-j}u_{it-k} + u_{is-j}u_{is-k} - u_{it-j}u_{is-j} - u_{it-k}u_{is-k}, \\
f^r_{t-j,t-k} &= (u_{it-j} - u_{it-k})(u_{it-k} - u_{it-j}) = -(u_{it-j} - u_{it-k})^2, \\
f^r_{t-j,t-k-\ell} &= (u_{it-j} - u_{it-k-\ell})(u_{it-k} - u_{it-j-\ell}).
\end{aligned}$$

Thus

$$\begin{aligned}
2T\psi_{iT}^{(j,k)} = 2T_m &\sum_{t=m+1}^{T} u_{it-j}u_{it-k} - \left(\sum_{t=m+1}^{T} u_{it-j}\right)^2 - \left(\sum_{t=m+1}^{T} u_{it-k}\right)^2 \\
&+ \sum_{t=m+1}^{T}(u_{it-j} - u_{it-k})^2 - 2\sum_{\ell=1}^{2p-m}\sum_{t=m+1+\ell}^{T}(u_{it-j} - u_{it-k-\ell})(u_{it-k} - u_{it-j-\ell}).
\end{aligned}$$

Result (36) is obtained by subtracting and adding $2(\sum_{t=m+1}^{T} u_{it-j})(\sum_{t=m+1}^{T} u_{it-k})$ and then dividing through by $2T$. The identity holds for $j > k$ as well because $\psi_{iT}^{(j,k)} = \psi_{iT}^{(k,j)}$. Finally, the second line of (36) is derived by means of the identity $\sum_{t=j+k+1}^{T} a_t = \sum_{t=1}^{T} a_t - \sum_{t=1}^{j+k} a_t$. ■

All the $R_{h,iT}^{(j,k)}$ terms in (36) turn out to be negligible compared with the other terms when considering either time series or panel asymptotics with large $T$. More precisely, the denominator $A_{nT}$ and numerator $b_{nT}$ in (33) above may be approximated as shown in the following lemma, where the approximation holds both for stationary and integrated $u_{it}$.

**Lemma 6** *Under Condition A, we have*

$$(40)\quad A_{nT} = \frac{1}{n}\sum_{i=1}^{n} D_T^{-1}F'X_i'M_1X_iFD_T^{-1} + \xi_{nT}^A,$$

*and*

$$(41)\quad b_{nT} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_T^{-1}F'(\zeta_i - E\zeta_i) + \xi_{nT}^b,$$

*where* $\zeta_i = X_i' M_1 \varepsilon_i + T^{-1} X_i' X_i \Pi \rho$*, and* $A_{nT}$ *and* $b_{nT}$ *are defined in (33), for some* $\xi_{nT}^A$ *and* $\xi_{nT}^b$ *such that*

$$\text{(42)} \quad \lim_{T\to\infty} \sup_n E\|\xi_{nT}^A\| = 0 \ \textit{and} \ \lim_{T\to\infty} \sup_n E\left[\xi_{nT}^b \xi_{nT}^{b\prime}\right] = 0,$$

*as given in (45) and (46) below.*

**Proof.** Let

$$\text{(43)} \quad R_{h,iT}^{den} = \begin{bmatrix} R_{h,iT}^{(1,1)} & \cdots & R_{h,iT}^{(1,p)} \\ \vdots & & \vdots \\ R_{h,iT}^{(p,1)} & \cdots & R_{h,iT}^{(p,p)} \end{bmatrix} \text{ and } R_{h,iT}^{num} = \begin{bmatrix} R_{h,iT}^{(1,0)} \\ \vdots \\ R_{h,iT}^{(p,0)} \end{bmatrix},$$

where $R_{h,iT}^{(j,k)}$ are defined in (36).

(i) Denominator: For (40), the second line of (36) implies

$$\text{(44)} \quad V_{iT} = \Psi_{iT}^{den} = X_i' M_1 X_i - T^{-1}(\pi 1_p' + 1_p \pi') \odot X_i' X_i + \textstyle\sum_{h=1}^3 R_{h,iT}^{den},$$

where $\pi = (1, \ldots, p)'$ and $\odot$ stands for the Hadamard (element-wise) product. Because $\pi 1_p' \odot X_i' X_i = \Pi X_i' X_i$ and $1_p \pi' \odot X_i' X_i = X_i' X_i \Pi$ with $\Pi = \mathrm{diag}(\pi)$, we have

$$\text{(45)} \quad \xi_{nT}^A = -\frac{1}{nT} \textstyle\sum_{i=1}^n D_T^{-1} F' (\Pi X_i' X_i + X_i' X_i \Pi) F D_T^{-1} + \sum_{h=1}^3 \frac{1}{n} \sum_{i=1}^n D_T^{-1} F' R_{h,iT}^{den} F D_T^{-1}.$$

The expectation of the absolute value of the first term is $O(T^{-1})$, which can be obtained by writing $D_T^{-1} F' \Pi X_i' X_i F D_T^{-1}$ as $D_T^{-1} F' \Pi F'^{-1} D_T \cdot D_T^{-1} F' X_i' X_i F D_T^{-1}$ and noting that $n^{-1} \sum_{i=1}^n D_T^{-1} F' X_i' X_i F D_T^{-1}$ has a uniformly bounded first moment. We can also show that $E\|\sigma_i^{-2} D_T' F' R_{h,iT}^{den} F D_T^{-1}\| \to 0$ as $T \to \infty$ for all $h$ by Lemma 9 in Appendix B. Thus (40) and the first part of (42) follow.

(ii) Numerator: For (41) and the second part of (42), we use (35) and the second line of (36) again, giving

$$\Psi_{iT}^{num} = X_i' M_1 u_i - T^{-1} \pi \odot X_i' u_i + \sum_{h=1}^3 R_{h,iT}^{num},$$

where $u_i = (u_{i1}, \ldots, u_{iT})'$. This last expression and (44) imply that

$$\eta_{iT} := \Psi_{iT}^{num} - \Psi_{iT}^{den} \rho = X_i' M_1 \varepsilon_i - T^{-1} \pi \odot X_i' u_i + T^{-1}[(\pi 1_p' + 1_p \pi') \odot X_i' X_i] \rho + \sum_{h=1}^3 (R_{h,iT}^{num} - R_{h,iT}^{den} \rho).$$

Since $1_p\pi' \odot X_i'X_i = X_i'X_i\Pi$, we have $\zeta_{iT} = X_i'M_1\varepsilon_i + T^{-1}(1_p\pi' \odot X_i'X_i)\rho$. Using $\pi \odot X_i'u_i = \Pi X_i'u_i$ and $\pi 1_p' \odot X_i'X_i = \Pi X_i'X_i$, it follows that

$$\text{(46)} \quad \xi_{nT}^b = -\frac{1}{n^{1/2}T}\sum_{i=1}^{n} D_T^{-1}F'\Pi X_i'\varepsilon_i + \sum_{h=1}^{3}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_T^{-1}F'(R_{h,iT}^* - ER_{h,iT}^*),$$

where $R_{h,iT}^* = R_{h,iT}^{num} - R_{h,iT}^{den}\rho$. (Note that subtracting means is valid because $E\eta_{iT} = 0$.) Lemma 10 shows that the variance-covariance matrix of the last term on the right hand side is $o(1)$, and the first term is $-T^{-1} \cdot D_T^{-1}F'\Pi F^{-1\prime}D_T \cdot n^{-1/2}\sum_{i=1}^{n} D_T^{-1}F'X_i'\varepsilon_i$, where the second moment of $\sigma_i^{-2}D_T^{-1}F'X_i'\varepsilon_i$ is bounded. The result follows. ■

With these results in hand, the proof of Theorem 5(c) for the stationary case with large $T$ and small $n$ is now straightforward.

**Proof of Theorem 5 (c).** In this case, note that $n$ is fixed, $T \to \infty$, $u_{it}$ is stationary (over $t$), and $D_T = T^{1/2}I_p$. Under Condition A, we have $T^{-1}X_i'M_1X_i = T^{-1}X_i'X_i + o_p(1) \to_p \sigma_i^2\Gamma$ for each $i$, where $\Gamma = \sigma_i^{-2}E(X_{it-1}X_{it-1}')$ is independent of $i$ in view of Condition A(i). From this result and (40), we have

$$\operatorname*{plim}_{T\to\infty} A_{nT} = \left(\frac{1}{n}\textstyle\sum_{i=1}^{n}\sigma_i^2\right)F'\Gamma F$$

(see Phillips and Solo, 1992, Theorem 3.16). Also $T^{-1/2}X_i'M_1\varepsilon_i = T^{-1/2}X_i'\varepsilon_i + o_p(1) \Rightarrow N(0, \sigma_i^4\Gamma)$, which together with (41) implies that

$$b_{nT} \Rightarrow N\left(0, \left[\frac{1}{n}\textstyle\sum_{i=1}^{n}\sigma_i^4\right]F'\Gamma F\right).$$

The result follows immediately. ■

In the unit root case with large $T$, we use the standardization matrix $D_T = \operatorname{diag}(T, T^{1/2}, \ldots, T^{1/2})$ and coordinate transformation

$$\text{(47)} \quad F'X_{it-1} = (u_{it-1}, -\Delta u_{it-1}, \ldots, -\Delta u_{it-p+1})'.$$

The denominator can be handled using (40). For the numerator, we have

$$\text{(48)} \quad b_{nT} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\varphi_{iT} - E\varphi_{iT}) + \xi_{nT}^c, \text{ where } E[\|\xi_{nT}^c\|^2] = o(1),$$

and $\varphi_{iT} = (\varphi_{1,iT}, \varphi_{2,iT}, \ldots, \varphi_{p,iT})'$ with

$$\text{(49)} \quad \varphi_{1,iT} = \frac{\rho^*(1)^{-1}}{T}\sum_{t=1}^T v_{it-1}\varepsilon_{it} - \frac{\rho^*(1)^{-1}}{T^2}\sum_{t=1}^T v_{it-1}\sum_{t=1}^T \varepsilon_{it} + \frac{\rho^*(1)^{-1}}{T^2}\sum_{t=1}^T v_{it-1}^2,$$

$$\text{(50)} \quad \varphi_{j,iT} = \frac{1}{\sqrt{T}}\sum_{t=1}^T \Delta u_{it-j+1}\varepsilon_{it}, \quad j = 2, \ldots, p,$$

due to Lemma 11. The large $T$ asymptotics (for small $n$ or large $n$) are obtained by evaluating $n^{-1/2}\sum_{i=1}^n \varphi_{iT}$ because $E\varphi_{1,iT} \to 0$ as $T \to \infty$.

**Proof of Theorem 5 (d).** Note that $u_{i0} := 0$ without loss of generality because the estimator is expressed in terms of differences. Otherwise we could simply replace $u_{it}$ with $u_{it} - u_{i0}$..

(i) Denominator: The first diagonal element of $D_T^{-1}F'X_i'M_1X_iFD_T^{-1}$ is

$$\frac{1}{T^2}\sum_{t=1}^T \left[u_{it-1} - \frac{1}{T}\sum_{s=1}^T u_{is-1}\right]^2 \Rightarrow \frac{\sigma_i^2}{\rho^*(1)^2}\int_0^1 \tilde{W}_i(r)^2 dr, \quad \tilde{W}_i(r) := W_i(r) - \int_0^1 W_i(r)dr,$$

where the $W_i(r)$ are independent standard Brownian motions. (See Phillips, 1987, Theorem 3.1, or use the BN decomposition in (55) below.) The other elements of the first row (and the first column) are $-T^{3/2}\sum_{t=1}^T u_{it-1}\Delta u_{it-j}$ for $j = 1, \ldots, p-1$, which are $O_p(T^{-1/2})$ and thus converge to zero as $T \to \infty$. The remaining elements of the $D_T^{-1}F'X_i'M_1X_iFD_T^{-1}$ matrix correspond to the stationary series $\{-\Delta u_{it-j}\}_{j=1,\ldots,p-1}$ and this matrix converges in probability to $\sigma_i^2\Omega$, where $\Omega$ is the variance-covariance matrix of $\sigma_i^{-1}(\Delta u_{it-1}, \ldots, \Delta u_{it-p+1})'$. We therefore have

$$\text{(51)} \quad D_T^{-1}F'X_i'M_1X_iFD_T^{-1} \Rightarrow \sigma_i^2 \text{diag}\left\{(\pi'\rho)^{-2}Y_{ai},\ \Omega\right\}, \quad Y_{ai} = \int_0^1 \tilde{W}_i(r)^2 dr,$$

for each $i$, where the coefficient $(\pi'\rho)^{-2}$ appears in the limit because of Lemma 7 below.

(ii) Numerator: Due to (49) and Lemma 7, we have

$$\varphi_{1,iT} \Rightarrow \frac{\sigma_i^2}{\rho^*(1)}\left[\int_0^1 W_i(r)dW_i(r) - W_i(1)\int_0^1 W_i(r)dr + \int_0^1 W_i(r)^2 dr\right] := \frac{\sigma_i^2 Y_{bi}}{\rho^*(1)},$$

which is also the weak limit of the first element of $D_T^{-1}F'\zeta_{iT}$. From (48) and (50), the vector of the second to last elements of $D_T^{-1}F'\zeta_{iT}$, denoted by $d_{2,iT}$ (a notation used only in this proof), is

$$d_{2,iT} = T^{-1/2}\Delta\ddot{X}_i'\varepsilon_i + O_p(T^{-1/2}) \Rightarrow \sigma_i^2 Z_{2i}, \quad Z_{2i} \sim N(0, \Omega),$$

where $\Delta\ddot{X}_i$ denotes the first $p-1$ columns of $\Delta X_i$, $\Omega = E\Delta\ddot{X}_{it-1}\Delta\ddot{X}_{it-1}'$, and $\Delta\ddot{X}_{it-1}$ denotes the first $p-1$ elements of $\Delta X_{it-1}$. Thus, $D_T^{-1}F'\zeta_{iT} \Rightarrow [\sigma_i^2(\pi'\rho)^{-1}Y_{bi}, \sigma_i^2 Z_{2i}']'$.

Finally, to see the relationship between the limits of $\varphi_{1,iT}$ and $d_{2,iT}$, we note that the sample random function corresponding to $W_i(r)$ is $T^{-1/2}\sum_{t=1}^{[Tr]}\varepsilon_{it}$ and the $j$th element of $d_{2,iT}$ is

$-T^{-1/2}\sum_{t=1}^{T}\Delta u_{it-j}\varepsilon_{it}$. The joint Gaussianity of $(\varphi_{1,iT}, d'_{2,iT})'$ is straightforward, and the covariance between $\varphi_{1,iT}$ and $d_{2,iT}$ is zero under the bi-directional martingale difference assumption. So $Y_{bi}$ and $Z_{2i}$ are independent.

Combining these results with (51) and (41), and noting that $EY_{bi} = 0$, $EZ_{2i} = 0$, we get the stated result. ■

Next we prove the panel limit theory where $n \to \infty$. Here the LLN and CLT are established using variation across $i$.

**Proof of Theorem 5 (b).** Let $E(\sigma_i^2) := \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\sigma_i^2$ as before and $E(\sigma_i^4) := \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\sigma_i^4$.
(i) Stationary case: We have $D_T = T^{1/2}I$. For the denominator, we have $E\bar{u}_i^2 = \sigma_i^2 O(T^{-1})$, where $\bar{u}_i = T^{-1}\sum_{t=1}^{T} u_{it}$**,** thus

$$\frac{1}{nT}\sum_{i=1}^{n} X_i'M_1X_i = \frac{1}{nT}\sum_{i=1}^{n} X_i'X_i + O_p(T^{-1}) \to_p E(\sigma_i^2)\Gamma.$$

For the numerator, by the martingale CLT we have

$$\frac{1}{\sqrt{nT}}\sum_{i=1}^{n}(\zeta_{iT} - E\zeta_{it}) = \frac{1}{\sqrt{nT}}\sum_{i=1}^{n} X_i'\varepsilon_i + o_p(1) \Rightarrow N\big(0, E(\sigma_i^4)\Gamma\big).$$

The result follows straightforwardly as $n, T \to \infty$.

(ii) Integrated case: We work with the rotated variables. For the denominator, let $A^*_{nT}(j,k)$ be the $(j,k)$ element of $A^*_{nT} := n^{-1}\sum_{i=1}^{n} D_T^{-1}F'X_i'M_1X_iFD_T^{-1}$, which is the leading term of $A_{nT}$ in (40). Then

$$A^*_{nT}(1,1) = \frac{1}{nT^2}\sum_{i=1}^{n}\sum_{t=1}^{T}u_{it-1}^2 - \frac{1}{nT^3}\sum_{i=1}^{n}\left[\sum_{t=1}^{T}u_{it-1}\right]^2 \to_p \frac{E(\sigma_i^2)}{6\rho^*(1)^2},$$

because $\lim_{n,T\to\infty} E[A^*_{nT}(1,1)] = \rho^*(1)^{-2}E(\sigma_i^2)/6$ and its variance is $O(n^{-1})$ by Lemma 8 below. So $A^*_{nT}(1,1) \to_p \rho^*(1)^{-2}/6$. This is also the probability limit of the $(1,1)$ element of $A_{nT}$ by Lemma 6.

The remaining elements in the first row (and the first column) of the denominator matrix are

$$A_{nT}(1,j) = \frac{1}{nT^{3/2}}\sum_{i=1}^{n}\sum_{t=1}^{T} u_{it-1}\Delta u_{it-j+1}, \quad j = 2,\ldots,p,$$

whose first moment is $O(T^{-1/2})$ by Lemma 8(iii) and second moment is $O(n^{-1}T^{-1})$ by Lemma 8(vii). So $A_{nT}(1,j) \to_p 0$ for all $j = 2,\ldots,p$, which is $\lim_{n,T\to\infty} E\left[A_{nT}(1,j)\right]$. Finally, for

$j \geq 2, k \geq 2,$

$$A_{nT}(j,k) = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} \Delta u_{it-j+1}\Delta u_{it-k+1} \to_p E(\sigma_i^2)\omega_{|j-k|},$$

which is $\lim_{n,T\to\infty} EA_{nT}(j,k)$, by evaluating the mean and the variance again. So $A_{nT} \to_p \lim_{n,T\to\infty} EA_{nT}$, where the limit is taken as $n, T \to \infty$.

For the numerator, we use (48), (49) and (50). Lemma 12 shows that the variance of the first element of $b_{nT}$ converges, and its limit is the same as the variance of the corresponding weak limit obtained in Theorem 5(d). The variance of the remaining terms of $b_{nT}$ and the covariances are also straightforwardly shown to converge to the limit variance and covariance of the corresponding weak limits in Theorem 5(d). Convergence of the variance and the boundedness of $\sigma_i^2$ imply the Lindeberg condition

$$(52)\quad \frac{1}{n}\sum_{i=1}^{n} E\Big[(\lambda' d_{iT})^2 \mathbf{1}\{(\lambda' d_{iT})^2 > Tc\}\Big] \to 0 \quad \forall\, c > 0, \quad d_{iT} = D_T^{-1}F'\eta_{iT},$$

for all $p \times 1$ vectors $\lambda$, which ensures the CLT for $b_{nT}$.

These arguments justify joint limits as $n, T \to \infty$, as discussed in Phillips and Moon (1999) general treatment of panel asymptotics. ■

**Proof of Theorem 2.** Theorem 5(a)–(c) imply that

$$n^{1/2} D_T F^{-1}(\hat{\rho} - \rho) \Rightarrow N\left(0, \operatorname{plim} A_{nT}^{-1} C_{nT} A_{nT}^{-1}\right),$$

where

$$A_{nT} = \frac{1}{n}\sum_{i=1}^{n} D_T^{-1}F'V_{iT}FD_T^{-1}, \quad C_{nT} = \frac{1}{n}\sum_{i=1}^{n} D_T^{-1}F'\eta_{iT}\eta_{iT}'FD_T^{-1},$$

and the probability limits are taken as $n(T - 2p - 1) \to \infty$ when $u_{it} \sim I(0)$, or as $n \to \infty$ (and for any $T$ sequence) when $u_{it} \sim I(1)$. Thus,

$$n^{1/2} A_{nT} D_T F^{-1}(\hat{\rho} - \rho) = n^{-1/2} D_T^{-1} F' Q_Z (\hat{\rho} - \rho) \Rightarrow N(0, \operatorname{plim} C_{nT}),$$

where $Q_Z := \sum_{i=1}^{n} V_{iT}$. For any $G_{nT}$ such that $G_{nT}C_{nT}G_{nT}' = I$, i.e., such that

$$n^{-1} G_{nT} D_T^{-1} F' Q_\eta F D_T^{-1} G_{nT}' = I, \text{ where } Q_\eta = \sum_{i=1}^{n} \eta_{iT}\eta_{iT}',$$

we have

$$n^{-1/2} G_{nT} D_T^{-1} F' Q_Z (\hat{\rho} - \rho) \Rightarrow N(0, I).$$

(Here we used the Lyapunov condition A(i) and the high level condition B(ii). See Phillips and Solo, 1992, for the convergence of $C_{nT}$.) The result follows by letting

(53) $B_{nT} := n^{-1/2} G_{nT} D_T^{-1} F'$.

■

**Proof of Theorem 3.** The first result is immediate from Corollary 4(i) of HPS (2009). The second result follows from the direct evaluation of the mean of the denominator and the variance of the expression in the numerator of Corollary 4(ii) of HPS (2009). ■

# Appendix B: Supplementary Lemmas

This section gathers together some technical lemmas. Since $\sigma_i^{-1}\varepsilon_{it}$ is $iid$, the $\sigma_i$ are uniformly bounded, and the quantities $n^{-1}\sum_1^n \sigma_i^2$ and $n^{-1}\sum_1^n \sigma_i^4$ are convergent,the heteroskedasticity may be ignored in the calculations given here. Hence, instead of introducing new notation for the standardized quantities $\sigma_i^{-1}u_{it}$, $\sigma_i^{-1}X_i$, $\sigma_i^{-2}V_{iT}$, we simply let

(54) $\sigma_i^2 := 1 \ \forall i$,

so that the component random variables are $iid$ across $i$. We also maintain Conditions A and B throughout, and assume that $u_{i0} := 0$ without loss of generality if $u_{it} \sim I(1)$; otherwise we could simply replace all the $u_{it}$ in the proofs with $u_{it} - u_{i0}$. This translation is justified by that fact that the PFAE is expressed in terms of differences.

We frequently use the following BN decomposition (Phillips and Solo, 1992, Lemma 2.1; Phillips and Moon, 1999, Lemma 2): Let $G(L) = \sum_0^\infty g_j L^j$. Then

$$G(L) = G(1) - (1-L)\ddot{G}(L),$$

where $\ddot{G}(L) = \sum_0^\infty \ddot{g}_j L^j$, $\ddot{g}_j = \sum_{j+1}^\infty g_k$. In the AR($p$) case, $G(L) = \rho(L)^{-1}$, where $\rho(L) := 1 - \rho_1 L - \cdots - \rho_p L^p$, so $\sum_1^\infty j^k |g_j|^k < \infty$ for any $k \geq 1$, thus $\sum_0^\infty |\ddot{g}_j|^k < \infty$ for any $k \geq 1$ and $|G(1)| < \infty$ (Phillips and Solo, 1992). Therefore,

(55) $$u_{it} = \begin{cases} \rho(1)^{-1}\varepsilon_{it} + \ddot{\varepsilon}_{it-1} - \ddot{\varepsilon}_{it} & \text{if } u_{it} \sim I(0), \\ \rho^*(1)^{-1}\sum_{s=1}^t \varepsilon_{is} + \ddot{\varepsilon}_{i0} - \ddot{\varepsilon}_{it} & \text{if } u_{it} \sim I(1), \end{cases}$$

where $\rho^*(L) = \rho(L)/(1-L)$, and

(56) $\sum_{t=1}^T u_{it-1} = \rho^*(1)^{-1}\sum_{t=1}^{T-1}(T-t)\varepsilon_{it} + T\ddot{\varepsilon}_{i0} - \sum_{t=1}^T \ddot{\varepsilon}_{it-1}$ if $u_{it} \sim I(1)$.

Note that $\ddot{\varepsilon}_{it}$ for the stationary case has a different meaning than the same notation for the $I(1)$ case. This duplicated usage of one notation will not cause any confusion because these terms do not appear together.

For $\rho^*(1)$, the following is true.

**Lemma 7** *If* $1_p'\rho = 1$, $\rho^*(1) = \pi'\rho$, *where* $\rho^*(L) = \rho(L)/(1-L)$ *and* $\pi = (1, \ldots, p)'$.

**Proof.** When $1_p'\rho = 1$, we have $\rho(L) = (1-L)\rho^*(L)$. So $\rho'(L) = -\rho^*(L) + (1-L)\rho^{*\prime}(L)$, implying that $\rho^*(1) = -\rho'(1) = \sum_{j=1}^p j\rho_j = \pi'\rho$ because $\rho(L) = 1 - \sum_{j=1}^p \rho_j L^j$. ■

Some results for the unit root case are provided next. These are useful in analyzing terms when $u_{it} \sim I(1)$**.**

**Lemma 8** *Under (54), if* $u_{i0} = 0$ *and* $u_{it} \sim I(1)$, *then*

*(i)* $T^{-2}\sum_{t=1}^T Eu_{it-1}^2 \to (1/2)\rho^*(1)^{-2}$;

*(ii)* $T^{-3}E[(\sum_{t=1}^T u_{it-1})^2] \to (1/3)\rho^*(1)^{-2}$;

*(iii)* $E(\sum_{t=1}^T u_{it-1}\Delta u_{it-j}) = O(T)$ *for all* $j$;

*(iv)* $T^{-1}Eu_{iT}^2 \to \rho^*(1)^{-2}$;

*(v)* $Eu_{iT}^4 = O(T^2)$;

*(vi)* $E[(\sum_{t=1}^T u_{t-1}^2)^2] = O(T^4)$;

*(vii)* $E[(\sum_{t=1}^T u_{t-1}\Delta u_{t-j})^2] = O(T^2)$ *for all* $j$.

**Proof.** (i) From (55), we have

$$\frac{1}{T^2}\sum_{t=1}^T Eu_{it-1}^2 = \frac{\rho^*(1)^{-2}}{T^2}\sum_{t=1}^T (t-1)^2 + O(T^{-1}) \to \frac{1}{2}\rho^*(1)^{-2}.$$

(ii) From (56), we have

$$\frac{1}{T^3}E\left[\left(\sum_{t=1}^T u_{it-1}\right)^2\right] = \frac{\rho^*(1)^{-2}}{T^3}\sum_{t=1}^{T-1}(T-t)^2 + O(T^{-1}) \to \frac{1}{3}\rho^*(1)^{-2}.$$

(iii) We have $u_{it-1} = \sum_{s=1}^{t-1}\Delta u_{is}$, so

$$E(u_{it-1}\Delta u_{it-j}) = \sum_{s=1}^{t-1} E(\Delta u_{is}\Delta u_{it-j}) = \sum_{s=1}^{t-1}\omega_{|t-j-s|} \le \sum_{k=0}^{\infty}|\omega_k| < \infty,$$

where $\omega_k = E\Delta u_{it}\Delta u_{it-k}$. So $T^{-1}\sum_{t=1}^{T} Eu_{it-1}\Delta u_{it-j} \leq T^{-1}\sum_{t=1}^{T}\sum_{0}^{\infty}|\omega_k| = \sum_{0}^{\infty}|\omega_k| < \infty$ for all $T$.

(iv) and (v): By (56), $u_{iT} = \rho^*(1)^{-1}\sum_{1}^{T}\varepsilon_{it} + \ddot{\varepsilon}_{i0} - \ddot{\varepsilon}_{iT}$. So

$$T^{-1}Eu_{iT}^2 = \frac{\rho^*(1)^{-2}}{T}E\left[\left(\sum_{t=1}^{T}\varepsilon_{it}\right)^2\right] + o(1) = \rho^*(1)^{-2} + o(1) \to \rho^*(1)^{-2},$$

and

$$u_{iT}^4 \leq 8\rho^*(1)^{-4}\left(\sum_{t=1}^{T}\varepsilon_{it}\right)^4 + 8(\ddot{\varepsilon}_{i0} - \ddot{\varepsilon}_{iT})^4,$$

implying that $T^{-2}E(u_{it}^4) = O(1)$.

(vi) We have

$$\left(\sum_{t=1}^{T}u_{it-1}^2\right)^2 = \sum_{t=1}^{T}u_{it-1}^4 + 2\sum_{t=2}^{T}\sum_{s=1}^{t-1}u_{is-1}^2u_{it-1}^2.$$

And $Eu_{it-1}^4 \leq Mt^2$ for some uniformly finite constant $M$. Thus the expectation of the above displayed equation is $O(T^3) + O(T^4)$. (For the second term, use the Cauchy-Schwarz inequality.)

(vii) We have

$$\left(\sum_{t=1}^{T}u_{it-1}\Delta u_{it-j}\right)^2 = \sum_{t=1}^{T}u_{it-1}^2(\Delta u_{it-j})^2 + 2\sum_{s<t}u_{it-1}u_{is-1}\Delta u_{it-j}\Delta u_{is-j}.$$

But $E[u_{it-1}^2(\Delta u_{it-j})^2] \leq Mt$ for some finite $M$ and the result follows. (For the second term, use the Cauchy-Schwarz inequality.) ■

Now we show that the remainder terms $R_{h,iT}^{den}$ in the denominator are negligible under large $T$ asymptotics (whether $n$ is large or small).

**Lemma 9** *Under (54),* $\lim_{T\to\infty}E\|D_T^{-1}F'R_{h,iT}^{den}FD_T^{-1}\| = 0$ *for* $h = 1, 2, 3$*, where* $F$ *and* $D_T$ *are defined in (30) and (31) and* $R_{h,iT}^{den}$ *are defined in (43).*

**Proof.** We will show that $E|R_{h,iT}^{(j,k)}| = O(1)$ for $h = 1, 2$ and $E|R_{3,iT}^{(j,k)}| = O(T^{1/2})$ at most for all $j, k = 1, \ldots, p$, where $R_{h,iT}^{(j,k)}$ are defined in (37)–(39).

(i) $h = 1$: Let the three components of $R_{1,iT}^{(j,k)}$ be denoted by $R_{1a,iT}^{(j,k)}$, $R_{1b,iT}^{(j,k)}$ and $R_{1c,iT}^{(j,k)}$, so $R_{1,iT}^{(j,k)} = R_{1a,iT}^{(j,k)} + R_{1b,iT}^{(j,k)} + R_{1c,iT}^{(j,k)}$ as written in (37). For $R_{1a,iT}^{(j,k)}$, $j \leq k$, we have

$$\sum_{t=j+k+1}^{T}(u_{it-j} - u_{it-k}) = \sum_{t=j+k+1}^{T}\sum_{r=0}^{k-j-1}\Delta u_{it-j-r} = \sum_{r=0}^{k-j-1}(u_{iT-j-r} - u_{ik-r}),$$

so

$$0 \leq R_{1a,iT}^{(j,k)} = \frac{1}{2T}\left[\sum_{r=0}^{k-j-1}(u_{iT-j-r} - u_{ik-r})\right]^2 \leq \frac{k-j}{2T}\sum_{r=0}^{k-j-1}(u_{iT-j-r} - u_{ik-r})^2.$$

Taking expectations and averaging across $i$ yields

$$0 \leq ER_{1a,iT}^{(j,k)} \leq \frac{k-j}{2}\sum_{r=0}^{k-j-1} E\Big[T^{-1}(u_{iT-j-r} - u_{ik-r})^2\Big] = O(1)$$

at most by Lemma 8(iv). For $R_{1b,iT}^{(j,k)}$ and $R_{1c,iT}^{(j,k)}$, consider

$$(57)\quad d_{iT,\ell} := \frac{1}{T}\sum_{t=j+k+1}^{T}\theta_{it,\ell}^{(j,k)}, \quad \theta_{it,\ell}^{(j,k)} := (u_{it-j} - u_{it-k+\ell})(u_{it-k} - u_{it-j+\ell}).$$

(The $d_{iT,\ell}$ notation is used only in this part of the proof.) Because of the inequality

$$\frac{1}{T}\textstyle\sum_{t=1}^{T}E|X_tY_t| \leq \frac{1}{T}\sum_{t=1}^{T}(EX_t^2EY_t^2)^{1/2} \leq \left[\frac{1}{T}\sum_{t=1}^{T}EX_t^2 \cdot \frac{1}{T}\sum_{t=1}^{T}EY_t^2\right]^{1/2},$$

we have

$$\Big(E|d_{iT,\ell}|\Big)^2 \leq \frac{1}{T}\textstyle\sum_{t=1}^{T}E\Big[(u_{it-j} - u_{it-k-\ell})^2\Big] \cdot \frac{1}{T}\sum_{t=1}^{T}E\Big[(u_{it-j} - u_{it-k-\ell})^2\Big] = O(1).$$

Because this bound holds for any $\ell$, we have $E|R_{1b,iT}^{(j,k)}| = O(1)$ and $E|R_{1c,iT}^{(j,k)}| = O(1)$.

(ii) $h = 2$: This case is clear because $t$ runs from 1 to $j + k$.

(iii) $h = 3$: We first show that $E|T^{-1}\sum_{t=1}^{T}u_{it-j}u_{ik}| = O(T^{1/2})$ for given $j$ and $k$, which is true because

$$E\left|\frac{1}{T}\textstyle\sum_{t=1}^{T}u_{it-j}u_{ik}\right| \leq \frac{1}{T}\sum_{t=1}^{T}(Eu_{it-j}^2)^{1/2}(Eu_{ik}^2)^{1/2} \leq \left[\frac{1}{T}\sum_{t=1}^{T}Eu_{it-j}^2\right]^{1/2}(Eu_{ik}^2)^{1/2} = O(T^{1/2}),$$

where we used the fact that $T^{-1}\sum_{t=1}^{T}Eu_{it-j}^2$ is $O(1)$ if $u_{it} \sim I(0)$ and $O(T)$ if $u_{it} \sim I(1)$ by Lemma 8(i). The result follows because $n^{-1}\sum_{i=1}^{n}D_T^{-1}F'E[R_{3,iT}^{den}]FD_T^{-1} = D_T^{-1}F'O(T^{1/2})FD_T^{-1} = O(T^{-1/2})$, where $D_T^{-1} = O(T^{-1/2})$. ■

We derive similar results for the numerator. Here, the remainder terms disappear in $L_2$.

**Lemma 10** $\lim_{T\to\infty}E[\|D_T^{-1}F'(R_{h,iT}^{num} - R_{h,iT}^{den}\rho)\|^2] = 0\ \forall h.$

**Proof.** For $h = 1, 2$, we will get $E[(R_{h,iT}^{(j,k)} - ER_{h,iT}^{(j,k)})^2] \leq ER_{h,iT}^{(j,k)2} = O(1)$ because then $E[D_T^{-1}F'(R_{h,iT}^{num} - R_{h,iT}^{den}\rho)(R_{h,iT}^{num} - R_{h,iT}^{den}\rho)'FD_T^{-1}] = O(D_T^{-2}) = O(T^{-1})$. For $h = 3$, we will establish a sharper boundary for the rotated and rescaled remainder $D_T^{-1}F[R_{3,iT}^{num} - R_{3,iT}^{den}\rho]$.

(i) $h = 1$: Again note that $R_{1,iT}^{(j,k)} = R_{1a,iT}^{(j,k)} + R_{1b,iT}^{(j,k)} + R_{1c,iT}^{(j,k)}$ as in the proof of Lemma 9. For $R_{1a,iT}^{(j,k)}$, we have

$$ER_{1a,iT}^{(j,k)\,2} \leq \frac{(k-j)^3}{4} \sum_{r=0}^{k-j-1} E\Big[T^{-2}(u_{iT-j-r} - u_{ik-r})^4\Big] = O(1), \tag{58}$$

by Lemma 8(v). For $R_{1b,iT}^{(j,k)}$, we have

$$ER_{1b,iT}^{(j,k)\,2} = \frac{1}{4T^2}\sum_{t=1}^{T} E\big[(u_{it-j} - u_{it-k})^2\big] + \frac{1}{4T^2}\sum_{s<t}^{T} E\big[(u_{it-j} - u_{it-k})^2(u_{is-j} - u_{is-k})^2\big],$$

which is $O(1)$ for given $j$ and $k$ (small) because $u_{it-j} - u_{it-k}$ is a finite sum of stationary terms for given $j$ and $k$ irrespective of the existence of the unit root, so its fourth moments are uniformly (over $t$) bounded. $R_{1c,iT}^{(j,k)}$ is similarly handled.

(ii) $h = 2$: This case is straightforward because $j + k$ is fixed and small.

(iii) $h = 3$: We have

$$R_{3,iT}^{den} = \bar{X}_i 1_p' \odot G_i + (\bar{X}_i 1_p' \odot G_i)' \text{ and } R_{3,iT}^{num} = \bar{X}_i \odot \dot{v}_i + \ddot{v}_i \bar{u}_i,$$

where $G_i$ is the $p \times p$ matrix whose $(j,k)$ element is $\sum_{s=1}^{j+k} u_{is-k}$, $\dot{v}_i$ is the $p \times 1$ vector whose $j$th element is $\sum_{t=1}^{j} u_{it}$, $\ddot{v}_i$ is the $p \times 1$ vector whose $j$th element is $\sum_{t=1}^{j} u_{it-j}$, and $\odot$ is the Hadamard product. Because $\dot{v}_i(j) + \ddot{v}_i(k) = u_{i1-k} + u_{i2-k} + \cdots + u_{ij} = \sum_{t=1}^{j+k} u_{it-k} = G_i(j,k)$, where $\dot{v}_i(j)$ is the $j$th element of $\dot{v}_i$, $\ddot{v}_i(k)$ is the $k$th element of $\ddot{v}_i$ and $G_i(j,k)$ is the $(j,k)$ element of $G_i$, we have $G_i = \dot{v}_i 1' + 1\ddot{v}_i'$. So

$$\begin{aligned}
R_{3,iT}^{num} - R_{3,iT}^{den}\rho &= \bar{X}_i \odot \dot{v}_i - (\bar{X}_i 1_p' \odot G_i)\rho + \ddot{v}_i\bar{u}_i - (1_p\bar{X}_i' \odot G_i')\rho \\
&= \bar{X}_i \odot \dot{v}_i - (\bar{X}_i 1_p' \odot \dot{v}_i 1_p')\rho - (\bar{X}_i 1_p' \odot 1_p\ddot{v}_i')\rho + \ddot{v}_i\bar{u}_i \\
&\quad - (1_p\bar{X}_i' \odot 1_p\dot{v}_i')\rho - (1_p\bar{X}_i' \odot \ddot{v}_i 1_p')\rho \\
&= (\bar{X}_i \odot \dot{v}_i)(1 - 1'\rho) - \bar{X}_i\ddot{v}_i'\rho + \ddot{v}_i(\bar{u}_i - \bar{X}_i'\rho) - 1_p(\bar{X}_i \odot \dot{v}_i)'\rho,
\end{aligned}$$

where we use the relation $ab' \odot cd' = (a \odot c)(b \odot d)'$ for column vectors $a$, $b$, $c$ and $d$. Because $\bar{u}_i - \bar{X}_i'\rho = \bar{\varepsilon}_i$, $\ddot{v}_i = F^{-1\prime}X_{i0}$ and $F'1 = e_1$, where $e_1$ is the first column of $I_p$, we have

$$\begin{aligned}
D_T^{-1}F'(R_{3,iT}^{num} - R_{3,iT}^{den}\rho) &= D_T^{-1}F'(\bar{X}_i \odot \dot{v}_i)(1 - 1'\rho) - D_T^{-1}F'\bar{X}_i\ddot{v}_i'\rho \\
&\quad + D_T^{-1}X_{i0}\bar{\varepsilon}_i - D_T^{-1}e_1(\bar{X}_i \odot \dot{v}_i)'\rho.
\end{aligned} \tag{59}$$

If $u_{it} \sim I(0)$, then all the terms in (59) are easy to handle: the variances disappear as $T \to \infty$ because the variance of $\bar{X}_i$ and $\bar{\varepsilon}_i$ disappear at an $O(T^{-1})$ rate. Now let $u_{it} \sim I(1)$. The first term of (59) is null because $1'\rho = 1$. For the second term of (59), we have

$$D_T^{-1} F' \bar{X}_i \ddot{v}_i' \rho = D_T^{-1} F' \bar{X}_i X_{i0}' F^{-1} \rho,$$

where $D_T^{-1} F' \bar{X}_i = (T^{-1} \sum_t u_{it-1}, -T^{-1/2} \sum_t \Delta u_{it-1}, \ldots, -T^{-1/2} \sum_t \Delta u_{it-p+1})'$. So the $(1,k)$ element of $D_T^{-1} F' \bar{X}_i X_{i0}'$ is $T^{-2} \sum_t u_{it-1} u_{i1-k}$ and satisfies

$$(60) \quad E\left[\frac{1}{T^2}\sum_{t=1}^T (u_{it-1}u_{i1-k} - Eu_{it-1}u_{i1-k})\right]^2 \le \frac{1}{T^4} E\left[\left(\sum_{t=1}^T u_{it-1}u_{i1-k}\right)^2\right] = O(T^{-1}),$$

where the last order can be obtained using (56). The $(j,k)$ elements of $D_T^{-1} F' \bar{X}_i X_{i0}'$ for $j > 1$ are easily handled because they involve only differences (which are stationary) and initial values. The variance of the third term on the right hand side of (59) is $O(T^{-2})$. The last term of (59) contains only one nonzero element, which is the first element equal to $T^{-1}(\bar{X}_i \odot \dot{v}_i)'\rho$. Its variance is $O(T^{-1})$, as shown in (60). ■

Next we approximate $D_T^{-1} F' \zeta_i$ when $u_{it} \sim I(1)$. The first element of $D_T^{-1} F' \zeta_i$ is $T^{-1} \sum_{t=1}^T u_{it-1}\varepsilon_{it} - T^{-2}(\sum_1^T u_{it-1})(\sum_1^T \varepsilon_{it}) + \sum_{j=1}^p T^{-2} \sum_{t=1}^T u_{it-1}u_{it-j} j\rho_j$. Of these terms, the $u_{it-j}$ terms in the last term can be replaced by $u_{it-1}$ in the sense that

$$\sum_{j=1}^p \frac{1}{T^2}\sum_{t=1}^T u_{it-1}u_{it-j} j\rho_j = \frac{1}{T^2}\sum_{t=1}^T u_{it-1}^2 \sum_{j=1}^p j\rho_j + o_p(1),$$

where the last $o_p(1)$ term is negligible in the $L_2$ sense, and all the $u_{it-1}$ terms can be replaced with the leading term of (55), i.e., with $\rho^*(1)^{-1} \sum_{s=1}^{t-1} \varepsilon_{is}$. Also, the vector of the second to last elements of $D_T^{-1} F' \zeta_i$ is approximated by $-T^{-1/2}[\Delta u_{it-1}, \ldots, \Delta u_{it-p+1}]'$ because the remaining terms are negligible in the $L_2$ sense as shown later. Thus, we have the following result:

**Lemma 11** *Let $u_{it} \sim I(1)$. Then $D_T^{-1} F' \zeta_i = \varphi_{iT} + \delta_{iT}$ with $\varphi_{iT} = (\varphi_{1,iT}, \varphi_{2,iT}')'$, where*

$$\varphi_{1,iT} = \frac{1}{\rho^*(1)}\left[\frac{1}{T}\sum_{t=1}^T v_{it-1}\varepsilon_{it} - \frac{1}{T^2}\left(\sum_{t=1}^T v_{it-1}\right)\left(\sum_{t=1}^T \varepsilon_{it}\right) + \frac{1}{T^2}\sum_{t=1}^T v_{it-1}^2\right],$$
$$\varphi_{2,iT} = -\frac{1}{\sqrt{T}}\sum_{t=1}^T \Big[\Delta u_{it-1}, \ldots, \Delta u_{it-p+1}\Big]' \varepsilon_{it},$$

$v_{it} = \sum_1^t \varepsilon_{is}$, *and* $\lim_{T\to\infty} E\delta_{iT}\delta_{iT}' = 0$.

**Proof.** Let

$$\tilde{\varphi}_{1,iT} = \frac{1}{T}\sum_{t=1}^T u_{it-1}\varepsilon_{it} - \frac{1}{T^2}\left(\sum_{t=1}^T u_{it-1}\right)\left(\sum_{t=1}^T \varepsilon_{it}\right) + \sum_{j=1}^p \frac{1}{T^2}\sum_{t=1}^T u_{it-1}^2 j\rho_j,$$

and $\tilde{\varphi}_{iT} = (\tilde{\varphi}_{1,iT}, \varphi'_{2,iT})'$. We first show that $D_T^{-1}F'\zeta_i = \tilde{\varphi}_{iT} + \tilde{\delta}_{iT}$, where $E\tilde{\delta}_{iT}\tilde{\delta}'_{iT} = o(1)$. Let $\tilde{\delta}_{1,iT}$ be the first element of $\tilde{\delta}_{iT}$ and $\delta_{2,iT}$ the remaining elements, so that $\tilde{\delta}_{iT} = (\tilde{\delta}_{1,iT}, \delta'_{2,iT})'$. Then

$$\tilde{\delta}_{1,iT} = \sum_{j=1}^{p} \frac{1}{T^2} \sum_{t=1}^{T} u_{it-1}(u_{it-j} - u_{it-1}) j\rho_j.$$

Because $u_{it-j} - u_{it-1} = -\sum_{k=1}^{j-1} \Delta u_{it-k}$, we have

$$\tilde{\delta}_{1,iT} = \sum_{j=1}^{p} \sum_{k=1}^{j-1} \left[ \frac{1}{T^2} \sum_{t=1}^{T} u_{it-1} \Delta u_{it-k} \right] j\rho_j := \sum_{j=1}^{p} \sum_{k=1}^{j-1} d_{iT}(k) j\rho_j.$$

(This $d_{iT}(k)$ notation is used only in this proof.) But

$$E\left[d_{iT}(k)^2\right] = \frac{1}{T^4} \sum_{t=1}^{T} E\left[u_{it-1}^2 (\Delta u_{it-k})^2\right] + \frac{2}{T^4} \sum_{t=2}^{T} \sum_{s=1}^{t-1} E\left[u_{it-1} u_{is-1} \Delta u_{it-k} \Delta u_{is-k}\right].$$

Using the BN decomposition (55), we can approximate $u_{it-1}$ by $\rho^*(1)^{-1} \sum_{s=1}^{t-1} \varepsilon_{is}$ and $\Delta u_{it}$ by $\rho^*(1)\varepsilon_{it}$. Then the first term on the right hand side of the last expression is $O(T^{-2})$ and the second term is also $O(T^{-2})$. Because $\tilde{\delta}_{1,iT}$ is a finite sum of $d_{iT}(k)$, we have shown that $E\tilde{\delta}_{1,iT}^2 = o(1)$. Next, we have

$$\delta_{2,iT} = -\frac{1}{T^{3/2}} \left[ \sum_{t=1}^{T} \Delta \ddot{X}_{it-1} \right] \sum_{t=1}^{T} \varepsilon_{it} + \frac{1}{T^{3/2}} \sum_{t=1}^{T} \Delta \ddot{X}_{it-1} X'_{it-1} \Pi\rho,$$

where $\Delta \ddot{X}_{it-1}$ is the first $p-1$ elements of $\Delta X_{it-1}$. Because $\Delta X_{it-1}$ is stationary, the variance of the first term of $\delta_{2,iT}$ is $O(T^{-1})$ and the second term also has an $O(T^{-1})$ variance-covariance matrix, which can be shown using (55). The covariance also disappears due to Hölder's inequality.

So far, we have approximated $D_T^{-1}F'\zeta_{iT}$ with $\tilde{\varphi}_{iT}$ (in the $L_2$ sense). Now we show that $\varphi_{iT} - \tilde{\varphi}_{iT} \to 0$ in $L_2$. This part can be done using (55) and Lemma 7. More precisely, because $\sum_1^p j\rho_j = \rho^*(1)$ by Lemma 7, we have

$$\begin{aligned} d_{1,iT} := \tilde{\varphi}_{1,iT} - \varphi_{1,iT} = {} & \frac{1}{T} \sum_{t=1}^{T} (\ddot{\varepsilon}_{it-2} - \ddot{\varepsilon}_{it-1}) \varepsilon_{it} - \frac{1}{T^2} \sum_{t=1}^{T} (\ddot{\varepsilon}_{i0} - \ddot{\varepsilon}_{it-1}) \cdot \sum_{t=1}^{T} \varepsilon_{it} \\ & + \frac{1}{T^2} \sum_{t=1}^{T} \left[ \rho^*(1) u_{it-1} + \sum_{s=1}^{t-1} \varepsilon_{is} \right] (\ddot{\varepsilon}_{i0} - \ddot{\varepsilon}_{it-1}). \end{aligned}$$

The second moments of the first and second terms are $O(T^{-1})$, and for the last term, we again apply (55) and show that its second moment is $O(T^{-1})$. ■

**Lemma 12** *If $u_{it} \sim I(1)$, under (54), $E\varphi_{1,iT} \to 0$ and $E\varphi_{1,iT}^2 \to (1/4)\rho^*(1)^{-2}$.*

**Proof.** Let $v_{it} = \sum_1^s \varepsilon_{is}$, $\bar{v}_i = T^{-1} \sum_1^T v_{it-1}$, and $\bar{\varepsilon}_i = T^{-1} \sum_1^T \varepsilon_{it}$. Then

$$\varphi_{1,iT} = \frac{1}{\rho^*(1)} \left[ \frac{1}{T} \sum_{t=1}^{T} v_{it-1} \varepsilon_{it} - \bar{v}_i \bar{\varepsilon}_i + \frac{1}{T^2} \sum_{t=1}^{T} v_{it-1}^2 \right]$$

(a notation used only in this proof). Using $\sum_{t=1}^{T} v_{it-1} = \sum_{t=1}^{T-1}(T-t)\varepsilon_{it}$, we have

$$E\varphi_{1,iT} \to \frac{1}{\rho^*(1)}\left[0 - \frac{1}{2} + \frac{1}{2}\right] = 0.$$

For the second moment, we have

$$E\left[\rho^*(1)^2\varphi_{1,iT}^2\right] = \frac{1}{T^2}\sum_{t=1}^{T} E v_{it-1}^2\varepsilon_{it}^2 + E\bar{v}_i^2\bar{\varepsilon}_i^2 + \frac{1}{T^4}E\left[\left(\sum_{t=1}^{T} v_{it-1}^2\right)^2\right]$$
$$- \frac{2}{T}E\left[\bar{v}_i\bar{\varepsilon}_i\sum_{t=1}^{T} v_{it-1}\varepsilon_{it}\right] + \frac{2}{T^3}E\left[\sum_{t=1}^{T} v_{it-1}\varepsilon_{it}\sum_{t=1}^{T} v_{it-1}^2\right] - \frac{2}{T^2}E\left[\bar{v}_i\bar{\varepsilon}_i\sum_{t=1}^{T} v_{it-1}^2\right]$$
$$= H_1 + H_2 + H_3 + H_4 + H_5 + H_6.$$

First, $H_1 \to 1/2$ because $Ev_{it-1}^2 = t-1$. For $H_2$, we have $\bar{v}_i = T^{-1}\sum_{t=1}^{T-1}(T-t)\varepsilon_{it}$, so

$$\left(\sum_{t=1}^{T} v_{it-1}\right)^2 = \sum_{t=1}^{T}(T-t)^2\varepsilon_{it}^2 + 2\sum_{t=2}^{T}\sum_{s=1}^{t-1}(T-t)(T-s)\varepsilon_{it}\varepsilon_{is},$$
$$\left(\sum_{t=1}^{T}\varepsilon_{it}\right)^2 = \sum_{t=1}^{T}\varepsilon_{it}^2 + 2\sum_{t=2}^{T}\sum_{s=1}^{t-1}\varepsilon_{it}\varepsilon_{is}.$$

Thus,

$$H_2 \to \int_0^1\int_0^r \left[(1-r)^2 + (1-s)^2\right] ds\, dr + 4\int_0^1\int_0^r (1-r)(1-s) ds\, dr = \frac{5}{6}.$$

For the rest, note that

$$\sum_{t=1}^{T} v_{it-1}\varepsilon_{it} = \sum_{t=2}^{T}\sum_{s=1}^{t-1}\varepsilon_{is}\varepsilon_{it}, \tag{61}$$

$$\sum_{t=1}^{T} v_{it-1}\cdot\sum_{t=1}^{T}\varepsilon_{it} = \sum_{t=1}^{T}(T-t)\varepsilon_{it}^2 + \sum_{t=2}^{T}\sum_{s=1}^{t-1}(2T-t-s)\varepsilon_{is}\varepsilon_{it}, \tag{62}$$

$$\sum_{t=1}^{T} v_{it-1}^2 = \sum_{t=1}^{T}(T-t)\varepsilon_{it}^2 + 2\sum_{t=2}^{T}\sum_{s=1}^{t-1}(T-t)\varepsilon_{is}\varepsilon_{it}, \tag{63}$$

where (61) is obvious, (62) uses $\sum_{t=1}^{T} v_{it-1} = \sum_{t=1}^{T}(T-t)\varepsilon_{it}$, and (63) is obtained by rearranging the terms after expanding $v_{it-1}^2$ to $\sum_{s=1}^{t-1}\varepsilon_{is}^2 + 2\sum_{s=2}^{t-1}\sum_{r=1}^{s-1}\varepsilon_{ir}\varepsilon_{is}$. Now, for $H_3$, from (63), we have

$$H_3 \to 2\int_0^1\int_0^r (1-r)(1-s) ds\, dr + 4\int_0^1\int_0^r (1-r)^2 ds\, dr = \frac{1}{4} + \frac{1}{3} = \frac{7}{12}.$$

Using (61) and (62), we have

$$H_4 = -\frac{2}{T^3}\sum_{t=2}^{T}\sum_{s=1}^{t-1}(2T-t-s) \to -2\int_0^1\int_0^r (2-r-s) ds\, dr = -1.$$

From (61) and (63), we have

$$H_5 = \frac{4}{T^3}\sum_{t=2}^{T}\sum_{s=1}^{t-1}(T-t) = \frac{4}{T^3}\sum_{t=2}^{T}(t-1)(T-t) \to 4\int_0^1 r(1-r) dr = \frac{2}{3}.$$

Finally, from (62) and (63), we have

$$H_6 \to -4\int_0^1\int_0^r (1-r)(1-s) ds\, dr - 4\int_0^1\int_0^r (2-r-s)(1-r) ds\, dr = -\frac{4}{3}.$$

So $E[\rho^*(1)^2\varphi_{1,iT}^2] = \frac{1}{2} + \frac{5}{6} + \frac{7}{12} - 1 + \frac{2}{3} - \frac{4}{3} = \frac{1}{4}$, which implies the result. ■

# Appendix C: Panel BIC

This Appendix proves consistency of lag length selection by minimizing the $k_{\max}$ BIC criterion proposed in Section 4.2 above. The analysis is based on the simple panel AR($p$) model $y_{it} = \alpha_i + u_{it}$ and $u_{it} = \sum_{j=1}^{p} \rho_j u_{it-j} + \varepsilon_{it}$, where $\varepsilon_{it}$ is $iid$ with zero mean and finite variance. Let $p$ denote the true lag length, i.e., the maximal $p$ such that $\rho_p \neq 0$. We calculate the sum of squared residuals (SSR).

**SSR from $k_{\max}$-Truncated Data** Let $k_{\max}$ be given and exceed $p$. Let $\widetilde{\rho}^k$ be a (modified) full aggregation estimator obtained allowing for $k$ lags after $k_{\max}$-truncation, i.e.,

$$\widetilde{\rho}^k = \left( \sum_{i=1}^{N} \sum_{t=2k_{\max}+2}^{T} \sum_{s=1}^{t-2k_{\max}-1} \tilde{z}_{it,s}^k \tilde{z}_{it,s}^{k\prime} \right)^{-1} \sum_{i=1}^{N} \sum_{t=2k_{\max}+2}^{T} \sum_{s=1}^{t-2k_{\max}-1} \tilde{z}_{it,s}^k \tilde{y}_{it,s}, \tag{64}$$

where $z_{it,s}^k = (y_{it-1} - y_{is+1}, \ldots, y_{it-k} - y_{is+k})'$. For $k < k_{\max}$, this estimator is not exactly the full-aggregation estimator because terms are summed over $t = 2k_{\max} + 2, \ldots, T$ and $s = 1, \ldots, t - 2k_{\max} - 1$ instead of $t = 2k + 2, \ldots, T$ and $s = 1, \ldots, t - 2k - 1$. Let

$$SSR(k) = \sum_{i=1}^{N} \sum_{t=2k_{\max}+2}^{T} \sum_{s=1}^{t-2k_{\max}-1} (\tilde{y}_{it,s} - \tilde{z}_{it,s}^{k\prime} \widetilde{\rho}^k)^2.$$

Let $T_* = T - 2k_{\max} - 1$. Let

$$\hat{\sigma}_k^2 = q_{NT}^{-1} SSR(k), \quad q_{nT} = nT_*(T_* + 1)/2.$$

We want to determine the behavior of $\hat{\sigma}_k^2 / \hat{\sigma}_p^2$ as $nT_* \to \infty$.

**Matrix Notation** It is simpler to use matrix algebra and we introduce the following notation. Let $Y$ be the row vector of $\tilde{y}_{it,s}$ for $s = 1, \ldots, t - 2k_{\max} - 1$, $t = 2k_{\max} + 2, \ldots, T$ and $i = 1, \ldots, n$; let $Z_k$ be the matrix of $\tilde{z}_{it,s}^k$ with $q_{nT}$ rows and $k$ columns arranged in the same order; and similarly let $\tilde{\varepsilon}$ denote the long row vector of $\tilde{\varepsilon}_{it,s}$. Then

$$SSR(k) = Y' M_{Z_k} Y,$$

where $M_A = I - P_A = I - A(A'A)^{-1}A'$ for any matrix $A$ such that $A'A$ is nonsingular. Let $Z = Z_p$. Let $V(k) = SSR(k) - SSR(p)$.

Let $C_{nT}$ and $R_{nT}$ be such that $C_{nT}^{-1} Z_k' Z_k$ converges to a nonsingular (almost surely) matrix and $R_{nT}^{-1} Z_k' \tilde{\varepsilon}$ converges weakly to a nondegenerate distribution. Then $C_{nT}$ and $R_{nT}$ are the stochastic

orders of the denominator and the numerator, respectively, of the centered full-aggregation estimator $\widetilde{\rho}^k - \rho$. Let $r_{nT} = C_{nT}/R_{nT}$, so that $r_{nT}$ is the convergence rate of the full aggregation estimator. If the panel is stationary, then we have $C_{nT} = nT_*^2$, $R_{nT} = n^{1/2}T_*^{3/2}$ and $r_{nT} = n^{1/2}T_*^{1/2}$; if the panel is integrated over $t$, then $C_{nT} = nT_*^3$, $R_{nT} = n^{1/2}T_*^2$, and $r_{nT} = n^{1/2}T_*$.

**Case 1:** $k < p$ Let $Z = (Z_k \vdots W_k)$. Then $M_{Z_k} = M_Z + P_{M_{Z_k}W_k}$, so $SSR(k) = SSR(p) + Y'P_{M_{Z_k}W_k}Y$, where $P_{M_{Z_k}W_k} = M_{Z_k}W_k(W_k'M_{Z_k}W_k)^{-1}W_k'M_{Z_k} = M_{Z_k}W_k \cdot Q_k^{-1} \cdot M_{Z_k}W_k$ with $Q_k = W_k'M_{Z_k}W_k$. (Note that $C_{nT}^{-1}Q_k$ converges to a nonsingular matrix.) Also

$$M_{Z_k}Y = M_{Z_k}(Z\rho + \tilde{\varepsilon}) = M_{Z_k}W_k\rho_* + M_{Z_k}\tilde{\varepsilon}, \quad \rho_* = (\rho_{k+1}, \ldots, \rho_p)',$$

and thus $W_k'M_{Z_k}Y = Q_k\rho_* + W_k'M_{Z_k}\tilde{\varepsilon}$. So

$$\begin{aligned} V(k) &= Y'P_{M_{Z_k}W_k}Y = Y'M_{Z_k}W_k \cdot Q_k^{-1} \cdot W_k'M_{Z_k}Y \\ &= \rho_*'Q_k\rho_* + 2\rho_*'W_k'M_{Z_k}\tilde{\varepsilon} + \tilde{\varepsilon}'M_{Z_k}W_kQ_k^{-1}W_k'M_{Z_k}\tilde{\varepsilon} \\ &= C_{nT}\Big[\rho_*'(C_{nT}^{-1}Q_k)\rho_* + O_p(r_{nT}^{-1}) + O_p(r_{nT}^{-2})\Big] \\ &= C_{nT}[h_{nT} + o_p(1)], \quad h_{nT} = \rho_*'(C_{nT}^{-1}Q_k)\rho_* \to_p h > 0. \end{aligned}$$

(To show that $h > 0$, remember that $\rho_p \neq 0$ and the limit of $C_{nT}^{-1}Q_k$ is strictly positive definite.) Now $C_{nT}^{-1}V(k) = \hat{\sigma}_k^2 - \hat{\sigma}_p^2$, so $\hat{\sigma}_k^2 = \hat{\sigma}_p^2 + h_{nT} + o_p(1)$, i.e.,

$$\hat{\sigma}_k^2/\hat{\sigma}_p^2 = 1 + h_{nT}/\hat{\sigma}_p^2 + o_p(1) = 1 + h/\sigma^2 + o_p(1), \quad k < p. \tag{65}$$

where $\sigma^2 := \operatorname{plim} \hat{\sigma}_p^2 \geq 0$ and $h/\sigma^2 > 0$.

**Case 2:** $k > p$ Now let $k > p$. This time, let $Z_k = (Z \vdots W_k)$ where $W_k$ has $k - p$ columns. Because $M_{Z_k}Y = M_{Z_k}\tilde{\varepsilon}$ for $k \geq p$, we have $SSR(k) = \tilde{\varepsilon}'M_{Z_k}\tilde{\varepsilon}$, $SSR(p) = \tilde{\varepsilon}'M_Z\tilde{\varepsilon}$ and $M_{Z_k} = M_Z - P_{M_ZW_k}$, so

$$V(k) = SSR(k) - SSR(p) = -\tilde{\varepsilon}'P_{M_ZW_k}\tilde{\varepsilon} = -\tilde{\varepsilon}'M_ZW_k(W_k'M_ZW_k)^{-1}W_k'M_Z\tilde{\varepsilon}.$$

We note that $(C_{nT}^{-1}W_k'M_ZW_k)^{-1}$ is the second diagonal block of the inverse of $C_{nT}^{-1}Z_k'Z_k$, which is asymptotically nonsingular. Noting that $W_k$ is the matrix (with $nT_*(T_*+1)/2$ rows) of $(y_{it-p-1} - y_{is+p+1}, \ldots, y_{it-k} - y_{is+k})$, we observe that $R_{nT}^{-1}W_k'M_Z\tilde{\varepsilon}$ converges to a nondegenerate random vector. So $C_{nT}R_{nT}^{-2}V(k)$ converges to a nondegenerate distribution. For notational brevity, let

$\xi_{nT} = -C_{nT}R_{nT}^{-2}V(k)$. Then $\xi_{nT} = O_p(1)$ if $k > p$, where $\xi$ is an almost surely positive nondegenerate random variable. Because $C_{nT}R_{nT}^{-2}V(k) = r_{nT}^2(\hat{\sigma}_k^2 - \hat{\sigma}_p^2) = -\xi_{nT}$, we have $\hat{\sigma}_k^2 = \hat{\sigma}_p^2 - r_{nT}^{-2}\xi_{nT}$, where $r_{nT} = C_{nT}/R_{nT}$ (the convergence rate of the full aggregation estimator) as before. That is,

$$\text{(66)} \quad \hat{\sigma}_k^2/\hat{\sigma}_p^2 = 1 - r_{nT}^{-2}\xi_{nT}/\hat{\sigma}_p^2, \text{ where } \xi_{nT} = O_p(1), \quad k > p.$$

Note that $r_{nT} \to \infty$.

**Consistency of BIC** Let $N = nT_*$, the number of available observations after truncation. Let

$$BIC(k) = \log \hat{\sigma}_k^2 + k\phi_n, \quad \phi_n = N^{-1}\log(N).$$

Then

$$BIC(k) - BIC(p) = \log(\hat{\sigma}_k^2/\hat{\sigma}_p^2) + (k-p)\phi_n.$$

If $k < p$, then by (65),

$$BIC(k) - BIC(p) = \log\left(1 + h/\sigma^2 + o_p(1)\right) - (p-k)\phi_n \to_p \log(1 + h/\sigma^2) > 0,$$

because $\phi_n \to 0$ and $h > 0$. So if $k < p$, then when $N$ is large enough (so $\phi_n$ is small enough), we have $BIC(k) > BIC(p)$. Next, if $k > p$, then by (66),

$$\begin{aligned} BIC(k) - BIC(p) &= \log\left(1 - r_{nT}^{-2}\xi_{nT}\right) + (k-p)\phi_n \\ &= -r_{nT}^{-2}\xi_{nT} + o_p(r_{nT}^{-2}) + (k-p)\phi_n, \end{aligned}$$

implying that

$$r_{nT}^2\Big[BIC(k) - BIC(p)\Big] = -\xi_{nT} + o_p(1) + (k-p)\log N \times \left(\frac{r_{nT}^2}{n}\right),$$

where $\xi_{nT} = O_p(1)$. Thus, $BIC(k) > BIC(p)$ asymptotically for $k > p$ under the sufficient condition that $\liminf r_{nT}^2/n > 0$. This condition holds whether the panel is stationary or integrated because $r_{nT}^2$ is at least $O(nT_*)$.

# Appendix D: Unit Root Asymptotics for a Modified PFAE

**Proof of (23).** Theorem 3 of HPS (2009) gives a representation of the FAE estimator in terms of the pooled OLS estimator. This relationship in the panel context gives the following relationship

between the PFAE estimator $\hat{\rho}$ and the LSDV estimator $\hat{\rho}_{lsdv}$:

$$\hat{\rho} = \hat{\rho}_{lsdv} + \frac{\sum_i T_2^{-1} \sum_{t=3}^T y_{it-1}^2}{\sum_i \sum_{t=3}^T \breve{y}_{it-1}^2} + \frac{\sum_i \left\{ y_{i1} y_{i2} - T_2^{-1} (y_{i1} + y_{i2}) \sum_{t=3}^T y_{it-1} \right\}}{\sum_i \sum_{t=3}^T \breve{y}_{it-1}^2},$$

where $\breve{y}_{it-1} = y_{it-1} - T_2^{-1} \sum_{s=3}^T y_{is-1}$, $T_2 = T - 2$, and where

$$\hat{\rho}_{lsdv} - \rho = \frac{\sum_i \sum_{t=3}^T \breve{y}_{t-1} \breve{u}_{it}}{\sum_i \sum_{t=3}^T \breve{y}_{it-1}^2},$$

with $\breve{u}_{it} := u_{it} - T_2^{-1} \sum_{s=3}^T u_{is}$**.** It follows that when $\rho = 1$ and $\frac{\sqrt{n}}{T} \to 0$

$$\begin{aligned}
&\sqrt{n}T \left( \hat{\rho} - 1 \right) \\
&= \sqrt{n}T \left( \hat{\rho}_{lsdv} - 1 \right) + \frac{\sqrt{n}T}{T_2} \frac{\sum_i \sum_{t=3}^T y_{it-1}^2}{\sum_i \sum_{t=3}^T \breve{y}_{it-1}^2} + \frac{\sqrt{n}T \sum_i \left\{ y_{i1} y_{i2} - T_2^{-1} (y_{i1} + y_{i2}) \sum_{t=3}^T y_{it-1} \right\}}{\sum_i \sum_{t=3}^T \breve{y}_{it-1}^2} \\
&= \sqrt{n}T \left( \hat{\rho}_{lsdv} - 1 + \frac{1}{T} \frac{\sum_i \sum_{t=3}^T y_{it-1}^2}{\sum_i \sum_{t=3}^T \breve{y}_{it-1}^2} \right) + O_p \left( \frac{\sqrt{n}}{T} \right) \\
&= \sqrt{n}T \left( \hat{\rho}_{lsdv} - 1 + \frac{1}{T} \frac{3 \sum_i \sum_{t=3}^T \breve{y}_{it-1}^2 + \left( \sum_i \sum_{t=3}^T y_{it-1}^2 - 3 \sum_i \sum_{t=3}^T \breve{y}_{it-1}^2 \right)}{\sum_i \sum_{t=3}^T \breve{y}_{it-1}^2} \right) + o_p(1) \\
&= \sqrt{n}T \left( \hat{\rho}_{lsdv} - 1 + \frac{3}{T} \right) + \sqrt{n} \frac{\sum_i \sum_{t=3}^T y_{it-1}^2 - 3 \sum_i \sum_{t=3}^T \breve{y}_{it-1}^2}{\sum_i \sum_{t=3}^T \breve{y}_{it-1}^2} + o_p(1) \\
&= \sqrt{n}T \left( \hat{\rho}_{lsdv} - 1 + \frac{3}{T} \right) + \sqrt{n} \frac{3 \sum_i T_2^{-1} \left( \sum_{t=3}^T y_{it-1} \right)^2 - 2 \sum_i \sum_{t=3}^T y_{it-1}^2}{\sum_i \sum_{t=3}^T \breve{y}_{it-1}^2} + o_p(1)
\end{aligned}$$

giving the stated relationship between the two estimators $\hat{\rho}$ and $\hat{\rho}_{lsdv}$. ■

We now proceed to derive asymptotics for the modified PFAE given by (24) as $n, T \to \infty$ when $\rho = 1$. Note that we can set $u_{i0} := 0$ without loss of generality when $\rho = 1$. Let $\widehat{Q} = n^{-1} T^{-2} \sum_{i=1}^n \sum_{t=3}^T \breve{u}_{it-1}^2$ where $\breve{u}_{it-1} := u_{it-1} - T_2^{-1} \sum_{t=3}^T u_{is-1}$**.** The first identity of (25) implies that

$$(67) \quad n^{1/2} T (\hat{\rho}^+ - 1) = n^{1/2} T (\hat{\rho}_{lsdv} - 1 + \tfrac{3}{T}) + n^{1/2} T \gamma (\hat{\rho} - \hat{\rho}_{lsdv} - \tfrac{3}{T}) = \widehat{G} + \gamma \widehat{H},$$

where

$$\widehat{G} = \widehat{Q}^{-1} \cdot \frac{1}{n^{1/2} T} \sum_{i=1}^n \sum_{t=3}^T \breve{u}_{it-1} \left[ \breve{\varepsilon}_{it} + \frac{3}{T} \breve{u}_{it-1} \right],$$

$\breve{\varepsilon}_{it} := \varepsilon_{it} - T_2^{-2}\sum_{s=3}^{T}\varepsilon_{is}$, and

$$\widehat{H} = \widehat{Q}^{-1}\cdot\frac{1}{n^{1/2}T}\sum_{i=1}^{n}\left[\frac{1}{T_2}\sum_{t=3}^{T}u_{it-1}^2 + u_{i1}u_{i2} - \frac{u_{i1}+u_{i2}}{T_2}\sum_{t=3}^{T}u_{it-1} - \frac{3}{T}\sum_{t=3}^{T}\breve{u}_{it-1}^2\right].$$

(For the expression for $\widehat{H}$, see HPS, 2009, Theorem 3.)

It is straightforward to show that $\widehat{Q} \to_p \sigma^2/6$. Next, Hahn and Kuersteiner (2002) show that $E\widehat{G} = 0$ and the asymptotic variance of $\widehat{Q}\widehat{G}$ is $51\sigma^4/180$. So the asymptotic variance of $\widehat{G}$ is $51/5$.

For the variance of $\widehat{H}$, we note that

$$\begin{aligned}\widehat{Q}\widehat{H} &= \frac{1}{n^{1/2}T}\sum_{i=1}^{n}\left[\frac{1}{T_2}\sum_{t=3}^{T}u_{it-1}^2 - \frac{3}{T}\sum_{t=3}^{T}\breve{u}_{it-1}^2\right] + O_p(T^{-1/2})\\ &= \frac{\sigma^2}{n^{1/2}}\sum_{i=1}^{n}(\xi_i - E\xi_i) + o_p(1), \quad \xi_i = -2\int_0^1 W_i(r)^2dr + 3\left[\int_0^1 W_i(r)dr\right]^2,\end{aligned}$$

where $W_i(r)$ are $iid$ standard Wiener processes. Note that $E\xi_i = 0$ and we need to calculate the variance of $\xi_i$, $E\xi_i^2$. First,

$$\begin{aligned}(68)\quad E\xi_i^2 &= 4E\left[\int_0^1 W_i(r)^2dr\right]^2 - 12E\left[\int_0^1 W_i(r)^2dr\left(\int_0^1 W_i(s)ds\right)^2\right]\\ &\quad + 9E\left[\int_0^1 EW_i(r)dr\right]^4.\end{aligned}$$

For the first term of (68), we have

$$\begin{aligned}E\left[\int_0^1 W_i(r)^2dr\right]^2 &= 2\int_0^1\int_0^r EW_i(r)^2W_i(s)^2dsdr, \quad W_i(r) = W_i(s) + [W_i(r) - W_i(s)],\\ &= 2\int_0^1\int_0^r \Big(EW_i(s)^4 + E[W_i(r) - W_i(s)]^2W_i(s)^2\Big)dsdr\\ &= 2\int_0^1\int_0^r \Big[3s^2 + (r-s)s\Big]dsdr = \frac{7}{12},\end{aligned}$$

by direct calculation, where the second identity holds because $E[W_i(r) - W_i(s)]W_i(r)^3 = 0$. For the second term of (68), after long and tedious algebra, we have

$$E\left[\int_0^1 W_i(r)^2dr\left(\int_0^1 W_i(s)ds\right)^2\right] = \frac{13}{30}.$$

For the third term of (68) we note that $\int_0^1 W_i(r)dr \sim N(0, 1/3)$, so that

$$E\left[\int_0^1 W_i(r)dr\right]^4 = \frac{1}{9}E\Big[N(0,1)^4\Big] = \frac{1}{9}\times 3 = \frac{1}{3}.$$

Thus, the asymptotic variance of $\widehat{Q}\widehat{H}$ is $\sigma^4$ times

$$4 \times \frac{7}{12} - 12 \times \frac{13}{30} + 9 \times \frac{1}{3} = \frac{24}{180},$$

implying that the asymptotic variance of $\widehat{H}$ is 24/5.

To recapitulate, what we have obtained so far is $\mathrm{Avar}(\widehat{G}) = 51/5$, and $\mathrm{Avar}(\widehat{H}) = 24/5$. We also have $\mathrm{Avar}\big(n^{1/2}T(\hat{\rho}_{fa} - 1)\big) = 9$ by Theorem 3, and

$$n^{1/2}T(\hat{\rho}_{fa} - 1) = \widehat{G} + \widehat{H}.$$

Thus,

$$\mathrm{Avar}\big(n^{1/2}T(\hat{\rho}_{fa} - 1)\big) = \mathrm{Avar}(\widehat{G}) + \mathrm{Avar}(\widehat{H}) + 2\,\mathrm{Acov}(\widehat{G}, \widehat{H}),$$

or $9 = 51/5 + 24/5 + 2\,\mathrm{Acov}(\widehat{G}, \widehat{H})$, implying that $\mathrm{Acov}(\widehat{G}, \widehat{H}) = -3$.

It therefore follows from (67) that

$$\begin{aligned}\mathrm{Avar}\Big(n^{1/2}T(\hat{\rho}^{+} - 1)\Big) &= \mathrm{Avar}(\widehat{G}) - 2\gamma\,\mathrm{Acov}(\widehat{G}, \widehat{H}) + \gamma^2\,\mathrm{Avar}(\widehat{H}) \\ &= \frac{51}{5} - 6\gamma + \frac{24}{5}\gamma^2.\end{aligned}$$

This asymptotic variance is minimized at $\gamma = 5/8$, where the minimum variance attained is $51/5 - 6 \times 5/8 + (24/5) \times (5/8)^2 = 333/40 = 8.325$.

# References


Ahn, S. C. and P. Schmidt (1995). Efficient Estimation of Models for Dynamic Panel Data. *Journal of Econometrics*, 68, 5–27.

Alvarez, J. and M. Arellano (2003). The Time Series and Cross-Section Asymptotics of Dynamic Panel Data Estimators. *Econometrica*, 71(4), 1121–1159.

Anderson, T. W. and C. Hsiao (1981). Estimation of Dynamic Models with Error Components. *Journal of American Statistical Association*, 76, 598–606.

Arellano, M. (1987). Computing Robust Standard Errors for Within-Groups Estimators. *Oxford Bulletin of Economics and Statistics*, 19, 431–434.

Arellano, M. and S. Bond (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *Review of Economic Studies*, 58, 277–297.

Arellano, M. and O. Bover (1995). Another Look at the Instrumental Variable Estimation of Error-components Models. *Journal of Econometrics*, 68, 29–51.

Bauer, P., Pötscher, B. M., and P. Hackl (1988). Model Selection by Multiple Test Procedures. *Statistics*, 19, 39–44.

Bertrand, M., E. Duflo, and S. Mullainathan (2004). How Much Should We Trust Differences-in-Differences Estimates? *Quarterly Journal of Economics*, 249–275.

Bhargava, A., L. Franzini and W. Narendranathan (1982). Serial Correlation and the Fixed Effects Model. *Review of Economic Studies*, 49(4), 533–549.

Blundell, R. and S. Bond (1998). Initial Conditions and Moment Restrictions in Dynamic Panel Data Models. *Journal of Econometrics*, 87, 115–143.

Davidson, J. (1994). *Stochastic Limit Theory*. Oxford University Press, New York.

Hahn, J., J. Hausman, and G. Kuersteiner (2007). Long Difference Instrumental Variables Estimation for Dynamic Panel Models with Fixed Effects. *Journal of Econometrics*, 140(2), 574–617.

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.

Hájek, J. (1972). "Local asymptotic minimax and admissibility in estimation," *Proceedings of Sixth Berkeley Symposium in Mathematical Statistics and Probability,* 1, 175–194.

Han, C. and P. C. B. Phillips (2006). GMM with Many Moment Conditions. *Econometrica*, 74, 147-192.

Han, C. and P. C. B. Phillips (2009). GMM Estimation for Dynamic Panels with Fixed Effects and Strong Instruments at Unity. *Econometric Theory*, 25, 1–33.

Han, C. and P. C. B. Phillips (2009b). Complications of FDMLE Estimation of Dynamic Panels with Fixed Effects. (under construction).

Han, C., P. C. B. Phillips, and D. Sul (2009). Uniform Asymptotic Normality in Stationary and Unit Root Autoregression. Unpublished manuscript. University of Auckland.

Hansen, C. B. (2007). Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data When $T$ is Large. *Journal of Econometrics*, 141, 597–620.

Hayakawa, K. (2007). A Simple Efficient Instrumental Variable Estimator in Panel AR($p$) Models. *Hi-Stat Discussion Paper Series d07-213*, Institute of Economic Research, Hitotsubashi University.

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.

Imbs, J., H. Mumtaz, M. O. Ravn, and H. Rey (2005). PPP Strikes Back: Aggregation and the Real Exchange Rate, *Quarterly Journal of Economics*, 120, 1–43.

Jeganathan, P. (1995). "Some aspects of asymptotic theory with applications to time series models," *Econometric Theory* 11, 818-867.

Keane, M., and D. Runkle (1992). On the Estimation of Panel-data Models with Serial-correlation When Instruments Are Not Strictly Exogenous. *Journal of Business & Economic Statistics*, 10(1), 1–9.

Kezdi, G. (2002). Robust Standard Error Estimation in Fixed-Effects Panel Models. Working Paper, University of Michigan.

Kiefer, N. M. (1980). Estimation of Fixed Effect Models for Time Series of Cross Section with Arbitrary Intertemporal Covariance. *Journal of Econometrics*, 14, 195–202.

Kiviet, J. F. (1995). On Bias, Inconsistency, and Efficiency of Various Etimators in Dynamic Panel Data Models. *Journal of Econometrics*, 68, 53–78.

Kruiniger, H. (2008). Maximum Likelihood Estimation and Inference Methods for the Covariance Stationary Panel AR(1)/Unit Root Model. *Journal of Econometrics*, 144, 447–464.

LeCam, L. (1972). "Limits of experiments." *Proceedings of the Sixth Berkeley Symposium in Mathematical Statistics and Probability*, 1, 245-261.

Nickell, S. (1981). Baises in Dynamic Models with Fixed Effects. *Econometrica*, 49, 1417–1426.

Orcutt, G. H. and H. S. Winokur (1969). "First order autoregression: inference, estimation and prediction". *Econometric*a, 37, 1-14.

Park, J. Y. and P. C. B. Phillips (1988). "Statistical Inference in Regressions With Integrated Processes: Part 2," *Econometric Theory* 5, 95-131.

Phillips, P. C. B. (1987). Time Series Regression with a Unit Root. *Econometrica*, 55, 277–301.

Phillips, P. C. B. (1989). "Partially identified econometric models," *Econometric Theory* 5, 181–240.

Phillips, P. C. B. and C. Han (2008). Gaussian Inference in AR(1) Times Series with or without Unit Root. *Econometric Theory*, 24, 631–650.

Phillips, P. C. B. and T. Magdalinos (2007). Limit theory for moderate deviations from a unit root, *Journal of Econometrics* 136, 115–130.

Phillips, P. C. B. and T. Magdalinos (2009). "Unit Root and Cointegrating Limit Theory when Initialization is in the Infinite Past" *Econometric Theor*y (forthcoming).

Phillips, P. C. B. and H. R. Moon (1999). Linear Regression Limit Theory for Nonstationary Panel Data. *Econometrica*, 67(5), 1057–1111.

Phillips, P. C. B. and W. Ploberger (1994). Posterior Odds for Testing for a Unit Root with Data-Based Model Selection. *Econometric Theory*, 10, 774–808.

Phillips, P. C. B. and V. Solo (1992). Asymptotics for Linear Processes, *Annals of Statistics*, 20(2), 971-1001.

Pötscher, B. M. (1983). Order Estimation in ARMA-models by Lagrangian Multiplier Tests. *Annals of Statistics*, 11, 872–885.

Qian, H., and P. Schmidt (2003). Partial GLS Regression. *Economics Letters*, 79, 385–392.

White, H. (1980). A heteroskedasticity-consistent covariance matrix and a direct test for heteroskedasticity. *Econometrica*, 48, 817–838.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge.

Table 1: Mean of $\hat{\rho}$ from AR(1), 1000 replications, $n = 100$
$y_{it} = a_i(1-\rho) + \rho y_{it-1} + \varepsilon_{it}$, $a_i \sim N(2, \sigma_a^2)$, $\varepsilon_{it} \sim iid\ N(0,1)$

Mean

| $\rho$ | $T$ | LSDV | HK | $\sigma_a = 1$ | | $\sigma_a = 3$ | | PFAE |
|---|---|---|---|---|---|---|---|---|
| | | | | GMM1 DIF | GMM2 SYS | GMM1 DIF | GMM2 SYS | |
| 0.0 | 10 | -0.1105 | -0.0215 | -0.0128 | 0.0022 | -0.0152 | 0.0487 | 0.0008 |
| 0.0 | 20 | -0.0533 | -0.0060 | -0.0121 | 0.0012 | -0.0127 | 0.0577 | -0.0007 |
| 0.3 | 10 | 0.1496 | 0.2646 | 0.2790 | 0.2994 | 0.2727 | 0.3407 | 0.2996 |
| 0.3 | 20 | 0.2291 | 0.2906 | 0.2821 | 0.2936 | 0.2805 | 0.3402 | 0.2989 |
| 0.5 | 10 | 0.3182 | 0.4501 | 0.4721 | 0.4963 | 0.4588 | 0.5309 | 0.4988 |
| 0.5 | 20 | 0.4160 | 0.4868 | 0.4770 | 0.4878 | 0.4733 | 0.5230 | 0.4987 |
| 0.7 | 10 | 0.4794 | 0.6273 | 0.6626 | 0.6901 | 0.6323 | 0.7145 | 0.6981 |
| 0.7 | 20 | 0.5997 | 0.6797 | 0.6704 | 0.6798 | 0.6601 | 0.7007 | 0.6988 |
| 0.9 | 10 | 0.6285 | 0.7914 | 0.8309 | 0.8785 | 0.7735 | 0.8830 | 0.8974 |
| 0.9 | 20 | 0.7729 | 0.8615 | 0.8556 | 0.8677 | 0.8293 | 0.8679 | 0.8991 |
| 1.0 | 10 | 0.6973 | 0.8859 | 0.5717 | 0.9877 | 0.5727 | 0.9878 | 0.9972 |
| 1.0 | 20 | 0.8493 | 0.9467 | 0.7801 | 0.9673 | 0.7798 | 0.9682 | 0.9986 |

Variance $\times 10^3$

| $\rho$ | $T$ | LSDV | HK | $\sigma_a = 1$ | | $\sigma_a = 3$ | | PFAE |
|---|---|---|---|---|---|---|---|---|
| | | | | GMM1 DIF | GMM2 SYS | GMM1 DIF | GMM2 SYS | |
| 0.0 | 10 | 1.159 | 1.403 | 2.278 | 2.103 | 2.500 | 3.430 | 1.503 |
| 0.0 | 20 | 0.497 | 0.548 | 0.722 | 0.827 | 0.768 | 1.494 | 0.557 |
| 0.3 | 10 | 1.213 | 1.468 | 2.825 | 2.364 | 3.401 | 3.267 | 1.593 |
| 0.3 | 20 | 0.492 | 0.542 | 0.769 | 0.867 | 0.857 | 1.242 | 0.551 |
| 0.5 | 10 | 1.174 | 1.421 | 3.124 | 2.485 | 4.300 | 3.250 | 1.545 |
| 0.5 | 20 | 0.460 | 0.507 | 0.752 | 0.885 | 0.900 | 1.123 | 0.512 |
| 0.7 | 10 | 1.084 | 1.311 | 3.410 | 2.414 | 5.934 | 3.270 | 1.442 |
| 0.7 | 20 | 0.401 | 0.442 | 0.705 | 0.763 | 0.977 | 0.984 | 0.439 |
| 0.9 | 10 | 0.973 | 1.177 | 4.940 | 2.261 | 10.11 | 2.712 | 1.367 |
| 0.9 | 20 | 0.315 | 0.348 | 0.797 | 0.691 | 1.345 | 0.882 | 0.345 |
| 1.0 | 10 | 0.921 | 1.138 | 30.54 | 0.769 | 30.37 | 0.760 | 1.369 |
| 1.0 | 20 | 0.252 | 0.279 | 4.177 | 0.681 | 4.214 | 0.682 | 0.273 |

* HK = LSDV $\times\, T/(T-1) + 1/(T-1)$

Table 2: $10^4 \times$Variance of LSDV and PFAE for AR(1) with $\rho = 1$, 10,000 replications
$y_{it} = y_{it-1} + \varepsilon_{it}, \varepsilon_{it} \sim iid\ N(0,1)$

| | $n = 50$ | | $n = 100$ | | $n = 200$ | |
|---|---|---|---|---|---|---|
| $T$ | LSDV | PFAE | LSDV | PFAE | LSDV | PFAE |
| 20 | 4.9100 | 5.5262 | 2.4475 | 2.7687 | 1.2312 | 1.3915 |
| 40 | 1.2455 | 1.2515 | 0.6375 | 0.6432 | 0.3096 | 0.3185 |
| 80 | 0.3275 | 0.3053 | 0.1591 | 0.1533 | 0.0784 | 0.0733 |
| 160 | 0.0802 | 0.0741 | 0.0402 | 0.0359 | 0.0196 | 0.0175 |

Note: The LSDV estimator is unbiased for $1 - 3/(T-1)$.

Table 3: Mean of $\hat{\rho}_1$ from AR(2), 1000 replications, $n = 100$
$y_{it} = a_i(1 - \rho_1 - \rho_2) + \rho_1 y_{it-1} + \rho_2 y_{it-2} + \varepsilon_{it}$, $\rho_2 = -0.2$
$a_i \sim N(2, \sigma_a^2)$, $\varepsilon_{it} \sim iid\ N(0,1)$, $\rho_2 = -0.2$

Mean

| $\rho_1$ | $T$ | LSDV | $\sigma_a = 1$ | | $\sigma_a = 3$ | | PFAE |
|---|---|---|---|---|---|---|---|
| | | | GMM1 DIF | GMM2 SYS | GMM1 DIF | GMM2 SYS | |
| 0.2 | 10 | 0.0865 | 0.1801 | 0.2033 | 0.1759 | 0.2976 | 0.2006 |
| 0.2 | 20 | 0.1524 | 0.1874 | 0.1998 | 0.1867 | 0.2825 | 0.1993 |
| 0.5 | 10 | 0.3748 | 0.4762 | 0.4980 | 0.4684 | 0.5601 | 0.4996 |
| 0.5 | 20 | 0.4500 | 0.4853 | 0.4919 | 0.4839 | 0.5428 | 0.4991 |
| 0.7 | 10 | 0.5596 | 0.6725 | 0.6943 | 0.6587 | 0.7380 | 0.6990 |
| 0.7 | 20 | 0.6469 | 0.6829 | 0.6875 | 0.6801 | 0.7193 | 0.6990 |
| 0.9 | 10 | 0.7296 | 0.8652 | 0.8895 | 0.8374 | 0.9136 | 0.8986 |
| 0.9 | 20 | 0.8401 | 0.8789 | 0.8828 | 0.8720 | 0.8980 | 0.8990 |
| 1.1 | 10 | 0.8638 | 1.0324 | 1.0812 | 0.9834 | 1.0838 | 1.0978 |
| 1.1 | 20 | 1.0147 | 1.0645 | 1.0750 | 1.0467 | 1.0729 | 1.0989 |
| 1.2 | 10 | 0.9010 | 0.7500 | 1.1925 | 0.7506 | 1.1925 | 1.1972 |
| 1.2 | 20 | 1.0659 | 0.9785 | 1.1767 | 0.9783 | 1.1774 | 1.1984 |

Variance $\times 10^3$

| $\rho_1$ | $T$ | LSDV | $\sigma_a = 1$ | | $\sigma_a = 3$ | | PFAE |
|---|---|---|---|---|---|---|---|
| | | | GMM1 DIF | GMM2 SYS | GMM1 DIF | GMM2 SYS | |
| 0.2 | 10 | 1.481 | 2.537 | 2.095 | 2.897 | 4.858 | 1.711 |
| 0.2 | 20 | 0.546 | 0.707 | 0.751 | 0.738 | 1.504 | 0.530 |
| 0.5 | 10 | 1.576 | 2.675 | 2.149 | 3.314 | 3.319 | 1.694 |
| 0.5 | 20 | 0.562 | 0.712 | 0.758 | 0.762 | 1.024 | 0.531 |
| 0.7 | 10 | 1.648 | 2.854 | 2.195 | 3.980 | 2.909 | 1.685 |
| 0.7 | 20 | 0.576 | 0.716 | 0.712 | 0.793 | 0.863 | 0.535 |
| 0.9 | 10 | 1.729 | 3.223 | 2.208 | 5.551 | 2.751 | 1.712 |
| 0.9 | 20 | 0.600 | 0.735 | 0.711 | 0.879 | 0.818 | 0.544 |
| 1.1 | 10 | 1.777 | 5.189 | 2.260 | 9.172 | 2.583 | 1.824 |
| 1.1 | 20 | 0.650 | 0.922 | 0.779 | 1.256 | 0.880 | 0.570 |
| 1.2 | 10 | 1.713 | 32.59 | 1.818 | 32.43 | 1.827 | 1.837 |
| 1.2 | 20 | 0.672 | 4.488 | 0.914 | 4.530 | 0.906 | 0.589 |

Table 4: Simulated sizes for AR(1), 5000 replications

| | | $\rho$, $H_0 : \rho =$ truth vs $H_1 : \rho \neq$ truth | | | | | |
|---|---|---|---|---|---|---|---|
| $n$ | $T$ | 0.0 | 0.3 | 0.5 | 0.7 | 0.9 | 1.0 |
| 25 | 10 | 0.0658 | 0.0652 | 0.0656 | 0.0672 | 0.0754 | 0.0770 |
| 25 | 20 | 0.0592 | 0.0628 | 0.0640 | 0.0650 | 0.0666 | 0.0726 |
| 25 | 40 | 0.0534 | 0.0534 | 0.0552 | 0.0572 | 0.0606 | 0.0710 |
| 50 | 10 | 0.0582 | 0.0590 | 0.0642 | 0.0638 | 0.0652 | 0.0630 |
| 50 | 20 | 0.0454 | 0.0468 | 0.0496 | 0.0530 | 0.0566 | 0.0628 |
| 50 | 40 | 0.0530 | 0.0504 | 0.0522 | 0.0540 | 0.0576 | 0.0618 |
| 100 | 10 | 0.0538 | 0.0520 | 0.0534 | 0.0512 | 0.0540 | 0.0522 |
| 100 | 20 | 0.0506 | 0.0532 | 0.0546 | 0.0534 | 0.0514 | 0.0614 |
| 100 | 40 | 0.0486 | 0.0510 | 0.0502 | 0.0558 | 0.0562 | 0.0610 |
| 200 | 10 | 0.0480 | 0.0498 | 0.0550 | 0.0558 | 0.0530 | 0.0556 |
| 200 | 20 | 0.0482 | 0.0502 | 0.0464 | 0.0504 | 0.0518 | 0.0522 |
| 200 | 40 | 0.0470 | 0.0498 | 0.0508 | 0.0466 | 0.0512 | 0.0514 |

Table 5: Simulated power for $H_0 : \rho = 0, 1$ for AR(1) model, 5000 replications

| | | $\rho$, $H_0 : \rho = 0$ vs $H_1 : \rho \neq 0$ | | | | $\rho$, $H_0 : \rho = 1$ vs $H_1 : \rho \neq 1$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| $n$ | $T$ | 0.000 | 0.025 | 0.050 | 0.075 | 0.925 | 0.950 | 0.975 | 1.000 |
| 25 | 10 | 0.0658 | 0.0742 | 0.1126 | 0.1768 | 0.2234 | 0.1440 | 0.0968 | 0.0770 |
| 25 | 20 | 0.0592 | 0.0874 | 0.1814 | 0.3380 | 0.6018 | 0.3340 | 0.1456 | 0.0726 |
| 25 | 40 | 0.0534 | 0.1214 | 0.3308 | 0.6156 | 0.9898 | 0.8466 | 0.3726 | 0.0710 |
| 50 | 10 | 0.0582 | 0.0790 | 0.1560 | 0.2748 | 0.3274 | 0.1794 | 0.0892 | 0.0630 |
| 50 | 20 | 0.0454 | 0.1134 | 0.3046 | 0.5822 | 0.8866 | 0.5760 | 0.2076 | 0.0628 |
| 50 | 40 | 0.0530 | 0.1796 | 0.5562 | 0.8888 | 1.0000 | 0.9916 | 0.5972 | 0.0618 |
| 100 | 10 | 0.0538 | 0.1006 | 0.2490 | 0.4826 | 0.5594 | 0.2964 | 0.1204 | 0.0522 |
| 100 | 20 | 0.0506 | 0.1838 | 0.5400 | 0.8734 | 0.9948 | 0.8598 | 0.3478 | 0.0614 |
| 100 | 40 | 0.0486 | 0.3320 | 0.8642 | 0.9932 | 1.0000 | 1.0000 | 0.8886 | 0.0610 |
| 200 | 10 | 0.0480 | 0.1478 | 0.4510 | 0.7910 | 0.8384 | 0.4952 | 0.1688 | 0.0556 |
| 200 | 20 | 0.0482 | 0.3108 | 0.8306 | 0.9916 | 1.0000 | 0.9936 | 0.5866 | 0.0522 |
| 200 | 40 | 0.0470 | 0.5724 | 0.9866 | 1.0000 | 1.0000 | 1.0000 | 0.9964 | 0.0514 |

Table 6: Lag Length Selection

$y_{it} = \alpha_i + u_{it}$, $u_{it} = \rho_1 u_{it-1} + \rho_2 u_{it-2} + \varepsilon_{it}$, $\varepsilon_{it} \sim N(0,1)$

$k_{\min} = 0$, $k_{\max} = 4$, 1000 replications

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\rho_1 = \rho_2 = 0.15$ | | | | | | | | | | | | | |
| | | BIC | | | General to Specific Method | | | | | | | | |
| | | | | | level=5% | | | level=2.5% | | | level=1% | | |
| $n$ | $T$ | $k$<2 | $k$=2 | $k$>2 | $k$<2 | $k$=2 | $k$>2 | $k$<2 | $k$=2 | $k$>2 | $k$<2 | $k$=2 | $k$>2 |
| 25 | 10 | 89 | 7 | 3 | 47 | 34 | 18 | 59 | 29 | 12 | 72 | 21 | 7 |
| 50 | 10 | 89 | 9 | 2 | 28 | 58 | 14 | 39 | 52 | 8 | 52 | 43 | 4 |
| 100 | 10 | 82 | 17 | 1 | 6 | 81 | 12 | 12 | 81 | 7 | 20 | 77 | 3 |
| 200 | 10 | 62 | 36 | 2 | 0 | 89 | 11 | 1 | 94 | 6 | 1 | 96 | 2 |
| 25 | 20 | 53 | 47 | 0 | 14 | 71 | 15 | 22 | 69 | 9 | 33 | 62 | 5 |
| 50 | 20 | 14 | 85 | 1 | 2 | 87 | 12 | 3 | 91 | 7 | 5 | 92 | 3 |
| 100 | 20 | 1 | 99 | 0 | 0 | 88 | 12 | 0 | 94 | 6 | 0 | 97 | 3 |
| 200 | 20 | 0 | 100 | 0 | 0 | 90 | 10 | 0 | 95 | 5 | 0 | 98 | 2 |
| $\rho_1 = \rho_2 = 0.5$ | | | | | | | | | | | | | |
| $n$ | $T$ | $k$<2 | $k$=2 | $k$>2 | $k$<2 | $k$=2 | $k$>2 | $k$<2 | $k$=2 | $k$>2 | $k$<2 | $k$=2 | $k$>2 |
| 25 | 10 | 27 | 62 | 10 | 0 | 81 | 19 | 0 | 87 | 12 | 1 | 92 | 7 |
| 50 | 10 | 10 | 84 | 7 | 0 | 85 | 15 | 0 | 91 | 9 | 0 | 96 | 4 |
| 100 | 10 | 1 | 95 | 5 | 0 | 87 | 13 | 0 | 92 | 8 | 0 | 96 | 4 |
| 200 | 10 | 0 | 98 | 2 | 0 | 89 | 11 | 0 | 94 | 6 | 0 | 98 | 2 |
| 25 | 20 | 0 | 99 | 1 | 0 | 85 | 15 | 0 | 91 | 9 | 0 | 96 | 5 |
| 50 | 20 | 0 | 100 | 0 | 0 | 88 | 12 | 0 | 93 | 7 | 0 | 97 | 3 |
| 100 | 20 | 0 | 100 | 0 | 0 | 89 | 11 | 0 | 94 | 6 | 0 | 98 | 2 |
| 200 | 20 | 0 | 100 | 0 | 0 | 89 | 11 | 0 | 95 | 5 | 0 | 98 | 2 |