

# EFFICIENT ESTIMATION AND INFERENCE FOR DIFFERENCE-IN- DIFFERENCE REGRESSIONS WITH PERSISTENT ERRORS

Ryan Greenaway-McGrevy<sup>a</sup>, Chirok Han<sup>b</sup> and  
Donggyu Sul<sup>c</sup>

<sup>a</sup>*Department of Economics, University of Auckland, Auckland, New Zealand*

<sup>b</sup>*Department of Economics, Korea University, Seoul, South Korea*

<sup>c</sup>*Department of Economics, University of Texas at Dallas, Richardson,  
TX, USA*

## ABSTRACT

*This paper is concerned with estimation and inference for difference-in-difference regressions with errors that exhibit high serial dependence, including near unit roots, unit roots, and linear trends. We propose a couple of solutions based on a parametric formulation of the error covariance. First stage estimates of autoregressive structures are obtained by using the Han, Phillips, and Sul (2011, 2013) X-differencing transformation. The X-differencing method is simple to implement and is unbiased in large N settings. Compared to similar parametric methods,*

---

Essays in Honor of Peter C. B. Phillips

Advances in Econometrics, Volume 33, 281–302

Copyright © 2014 by Emerald Group Publishing Limited

All rights of reproduction in any form reserved

ISSN: 0731-9053/doi:10.1108/S0731-905320140000033009

*the approach is computationally simple and requires fewer restrictions on the permissible parameter space of the error process. Simulations suggest that our methods perform well in the finite sample across a wide range of panel dimensions and dependence structures.*

**Keywords:** Treatment effects; difference-in-difference estimation; serial dependence; x-differencing; Cochrane-Orcutt transformation

**JEL classification:** C33

## 1. INTRODUCTION

This paper is concerned with estimating the following difference-in-difference (DD) regression with fixed effects and autoregressive errors of the form

$$y_{it} = \alpha_i + \lambda_t + \beta I_{it} + Z_{it}\gamma + \varepsilon_{it}, \quad \varepsilon_{it} = \sum_{j=1}^p \rho_j \varepsilon_{it-j} + u_{it} \quad (1)$$

where  $i=1, \dots, N$ ;  $t=1, \dots, T$ ,  $\alpha_i$  is a fixed effect,  $\lambda_t$  is a common time effect,  $I_{it}$  is the policy variable or treatment of interest (typically a binary or dummy variable), and  $Z_{it}$  is a vector of exogenous control variables. Within the DD literature focus typically centers on  $\beta$  – the coefficient on the dummy variable  $I_{it}$ . As emphasized in [Bertrand, Duflo, and Mullainathan \(2004, BDM\)](#), serial dependence in the error term  $\varepsilon_{it}$  generates problems for conducting valid statistical inference. Statistical tests may exhibit a size distortion if the serial dependence is not accounted for when constructing standard errors. By means of Monte Carlo simulations and pseudo-empirical applications, BDM show that inference based on the conventional OLS  $t$ -statistic over-rejects the null and suggest using robust covariance estimators to restore the size of the test for large  $N$ , small  $T$  panels.

BDM point out that much of the DD literature is potentially affected by this problem. Subsequent research has focused on HAC estimation with  $T \rightarrow \infty$  in order to accommodate panel datasets with large  $T$ . [Donald and Lang \(2007\)](#), [Hansen \(2007a\)](#), [Cameron, Miller, and Gelbach \(2008\)](#), [Bester, Conley, and Hansen \(2011\)](#) and [Miller, Cameron, and Gelbach \(2011\)](#) all study estimators under large  $T$  asymptotics, and suggest

various methods to obtain correctly sized tests for cases when  $T$  is larger than  $N$ .

Another strand of literature imposes a parametric serial dependence structure in order to simultaneously conduct valid inference and to gain more efficient point estimates. Bhargava, Franzini, and Narendranathan (1982) study a feasible generalized least squares (FGLS) method that relies on least squares (LS) estimation of an AR(1) process in the error. One potential drawback of the parametric approach is that it relies on accurate first stage estimates of the autoregressive process, and it is well-known that LS estimators exhibit  $O(T^{-1})$  bias and that the weak-instrument problem hampers instrumental variables (IV) methods in certain regions of the parameter space (Blundell & Bond, 1998). Although both Bhargava et al. (1982) and Hansen (2007b) correct the LS bias, their proposed method is computationally burdensome (requiring an iterative algorithm to correct for the bias exactly) and requires restrictions on the permissible parameter space in order for the inverted bias function to be unique (see Hansen, 2007b). For example, regions in the neighborhood of unity where the bias function is non-monotonic are excluded.

In this paper we provide some simple but effective methods to improve inference in DD regressions of the form of Eq. (1). Our suggested methods are based on the recently developed X-differencing estimator of autoregressive parameters proposed by Han, Phillips, and Sul (2011, 2014; HPS hereafter). We then transform the data using the fitted autoregressive parameters, using either an FGLS or Cochrane–Orcutt (CO) transformation, in order to remove serial dependence. Although the CO approach is less efficient than FGLS (due to the loss of initial observations in each time series), it is practically attractive and flexible, particularly if the data are highly persistent.

Three key features of our X-differencing-based approaches are appealing. First, the method is straightforward to apply and can be easily adopted by practitioners. Second, the method can be applied under fewer restrictions on the parameter space than bias-corrected LS or IV methods (e.g., it permits unit roots), and the efficiency gains relative to conventional LS can be substantial, particularly when the errors follow a unit root process and the CO transformation is used. This feature is particularly useful given that the typical dependent variable used in DD regressions – such as wages, consumption, employment, and healthcare expenditures (see BDM) – is highly persistent. Third, the estimator provides valid inference under a broad range of asymptotic sequences, requiring only that  $N(T-p) \rightarrow \infty$ .

Under large  $N$  frameworks (i.e., when  $N \rightarrow \infty$ ), the X-differencing method delivers a consistent estimator of the autoregressive parameters, regardless of the size of  $T$ . Under small  $N$  frameworks (i.e., when  $N$  is fixed in the asymptotics) the method significantly attenuates the bias of the LS estimator (Han et al., 2011). (In large  $T$ , small  $N$  settings however, this bias is less of a concern, of course.) These features permit application of the methods in a variety of sample size settings, including both “long” and “short” panels.

The rest of the paper is constructed as follows. The next section defines the serial dependence issue, and suggests a solution how to correct this statistical problem. A small Monte Carlo simulation results are reported in Section 3. Section 4 summarizes the findings of this paper. Sample Stata codes are provided in the [appendix](#).

## 2. ESTIMATION AND INFERENCE FOR AUTOREGRESSIVE ERROR STRUCTURES

In this section we consider models of the form of Eq. (1), where the innovation ( $u_{it}$ ) to the AR(p) process ( $\varepsilon_{it}$ ) is white noise, and the AR polynomial is permitted to have a unit root. There are two broad approaches considered in the literature for dealing with serial dependence: Correcting the standard errors of the LS estimator, or LS estimation based on transformed data (such as GLS).

The first approach is concerned with constructing covariance estimators that are robust to serial dependence. Nonparametric covariance estimators such as those based on clustering can often provide statistical tests with asymptotically correct sizes without relying on parametric assumptions on the error processes. The cost of this generality is that these require large  $N$ . When the number of groups (or clusters) is small, alternative methods are required. The nonparametric estimators proposed by Donald and Lang (2007), Hansen (2007a), Cameron et al. (2008), Bester et al. (2011) and Miller et al. (2011) permit both serial dependence and  $T$  to grow in the asymptotics. For example, Hansen (2007a) shows that under fixed  $N$ , large  $T$  asymptotics, using the clustered covariance matrix of Arellano (1987) to estimate the standard errors leads to a  $t$ -statistic that is asymptotically distributed as  $t_{N-1}$ . An additional problem under this approach is a potential efficiency loss if the errors are highly persistent. The convergence

rate of the OLS estimator of  $\beta$  (in levels) becomes  $\sqrt{N}$  rather than  $\sqrt{NT}$  when either stochastic or non-stochastic trends are present in the error term, since identification essentially only comes from between group variation.

The second approach transforms the data in order to remove serial dependence before applying LS. This approach includes FGLS, weighted least squares, and the CO type correction. Wooldridge (2003) discusses random and fixed effects FGLS when the covariance matrix is estimated non-parametrically. It is worth noting that the (nonparametric) random and fixed effects FGLS estimators can be constructed as long as  $N$  is much larger than  $T$ . For each element of the  $T \times T$  covariance matrix (or  $(T - 1) \times (T - 1)$  matrix if the data have been differenced to eliminate the fixed effects) can be estimated by using cross-sectional average. To be specific, the  $st$ th,  $t$ th element of the covariance matrix,  $\omega_{st}$ , of  $E\varepsilon_i\varepsilon_i' = \Omega_{T \times T}$  can be estimated consistently by taking the cross-sectional averages of the product of the regression residuals. As  $N \rightarrow \infty$  with  $T$  fixed, it is easy to show that

$$\hat{\omega}_{st} = \frac{1}{N} \sum_{i=1}^N \hat{\varepsilon}_{is} \hat{\varepsilon}_{it} \rightarrow^p \omega_{st}$$

for each  $t$  and  $s$ . (For the fixed effects FGLS, the sample covariances of the differenced residuals are to be calculated.) However this nonparametric FGLS method would require the ratio condition of  $T^2/N \rightarrow 0$  as  $N, T \rightarrow \infty$  since the number of unknown parameters in the  $\Omega$  matrix is  $O(T^2)$  but the available information is just  $N$ .

Parametric assumptions are typically required if the restrictions on the asymptotic sequence are relaxed. The AR(p) structure embedded in the error of (1) is an example of such a parametric structure that has received attention in the extant literature (see, e.g., Bhargava et al., 1982; Hansen, 2007b; Kiefer, 1980). However it is well-known that the LS estimator of the  $\rho$ 's are inconsistent (or biased in the sense that the inconsistency goes away as  $T \rightarrow \infty$ ). If this inconsistency is not corrected, then naturally the resulting covariance estimator is inconsistent. Bhargava et al. (1982) reduced the bias by utilizing the panel Durbin–Watson statistic. Meanwhile Hansen (2007b) proposes an exact mean unbiased (EMU) estimation by using an iterative estimation. The existence of EMU estimator requires that the binding function be monotonic, where the binding function,  $B(\rho, T)$ , is defined for AR(1) case as

$$B(\rho, T) = \rho + E(\hat{\rho} - \rho)$$

It is well-known that the binding function is not monotonic when  $\rho$  is near unity. Also in the AR(p) case, monotonicity does not generally hold when the sum of  $\rho_j$  is near unity.

The CO type correction also requires a bias correction. To be specific, the regression model in Eq. (1), ignoring the  $Z_{it}\gamma$  and  $\lambda_t$  term, can be rewritten as

$$y_{it} - \sum_{j=1}^p \rho_j y_{it-j} = \alpha_i^* + \beta \left( I_{it} - \sum_{j=1}^p \rho_j I_{it-j} \right) + u_{it} \quad (2)$$

If  $u_{it}$  is free of serial dependence the limiting distribution of the standard OLS  $t$ -statistic becomes asymptotically standard normal. CO suggest using the regression residuals to estimate the unknown  $\rho_j$ s. However, as discussed above, accurate estimation requires either a bias correction of the LS estimator or use of IV methods.

Another weakness of the above parametric approach is that the lag order is unknown and must be estimated. This can prove difficult in the presence of fixed effects when  $T$  is small (Lee, 2012). Moreover, as shown by Lee (2006), using a bias correction when the lag order is misspecified can exacerbate the bias of the LS estimator, rather than attenuate it. We discuss some consistent lag order selection methods in the following section.

### 2.1. FGLS with X-Differencing

We also adopt a parametric approach to dealing with the problem of serial dependence. However, we differ with the extant literature in that we make explicit use of the “X-differencing” method developed by Han et al. (2011, 2014) to estimate the AR(p) structure. The X-differencing method is a systematic differencing where the time lags on both sides of the equation are different, and proves to be effective in removing the nuisance fixed effects while also preserving the regressor-error uncorrelatedness. Specifically, the panel AR(p) equation  $y_{it} = \alpha_i + \sum_{j=1}^p \rho_j y_{it-j} + u_{it}$  is transformed to  $y_{it} - y_{is} = \sum_{j=1}^p \rho_j (y_{it-j} - y_{is+j}) + \text{error}_{it,s}$ , and then for all  $t$  and  $s$  such that  $t - s > p$ , the transformed regressors and the transformed error are exactly uncorrelated under the assumption that  $\Delta y_{it}$  is covariance stationary and

the AR(p) structure is correct. Now using this transformation, the autoregressive coefficients are consistently estimated regardless of the relative sizes of  $N$  and  $T$ . The estimators are consistent over a broad parameter space that includes unit roots in the autoregressive polynomial and there are no issues of non-monotonicity near the unit root.

Han, Phillips, and Sul (2013) show that the typical panel BIC estimation fails to identify the correct lag order. They suggest a couple of modified BICs and also the general to specific method with a data dependent  $p$ -value, and show that the suggested methods consistently estimate the correct lag order. The simplest of the modified BIC is as follows:

$$IC(k) = \hat{\sigma}^2(k) + k \frac{\ln(\sqrt{N}(T - k))}{\sqrt{N}(T - k)}$$

where  $\hat{\sigma}^2(k)$  denotes the mean squared residuals from the X-differencing regression with  $k$  lags of the dependent variable. Alternatively, the usual BIC may be applied to the sample truncated so that the same numbers of observations are used for all  $k$ .

Once the lag order and the AR coefficients are consistently estimated, the subsequent error covariance estimation and FGLS are straightforward. Following Hansen (2007b), we assume that the idiosyncratic error term  $\varepsilon_{it}$  of Eq. (1) is AR(p), i.e.,

$$\varepsilon_{it} = \sum_{j=1}^p \rho_j \varepsilon_{it-j} + u_{it}, \quad u_{i,t} \sim i.i.d.(0, \sigma^2) \tag{3}$$

The two procedures we consider are a fixed effects FGLS and a Cochrane–Orcutt (1949) type version, for which we need first to estimate  $\rho_j$  consistently. For this, we first get the residuals  $y_{it} - \hat{\beta}I_{it} - Z_{it}\hat{\gamma}$  for some consistent  $\hat{\beta}$  and  $\hat{\gamma}$  such as least squares and first difference estimators,<sup>1</sup> and then cross sectionally de-mean them in order to remove the common time effects. Let these de-meanded residuals be denoted by  $\hat{v}_{it}$ . Run the pooled OLS regression on the X-differenced equation

$$\hat{v}_{it} - \hat{v}_{is} = \sum_{j=1}^p \rho_j (\hat{v}_{it-j} - \hat{v}_{is+j}) + \text{error}_{it,s}$$

stacked for all  $i, t$ , and  $s$  for  $t - s > p$ . Let  $\hat{\rho}_j$  denote the resulting estimator. We have the following result.

**Theorem 1.** Suppose that  $y_{it}$  is generated according to Eq. (1) and  $\varepsilon_{it}$  is generated by Eq. (3). We assume that all regressors in Eq. (1) are strictly exogenous, and that  $\varepsilon_{it}$  is either (i) covariance stationary, or (ii)  $\varepsilon_{it} = \varepsilon_{it-1} + w_{it}$ , where  $w_{it}$  is covariance stationary. Then  $\hat{\rho}_j$  is consistent for  $\rho_j$  for each  $j$  as  $N(T - p - 1) \rightarrow \infty$ .

See the appendix for a short proof of Theorem 1. Note that the homogeneity of  $\beta$  and  $\gamma$  in Eq. (1) is required for the consistency of the X-differencing estimator.

The theorem permits both covariance stationary and unit root processes in the regression errors. Thus the X-differencing estimation of the AR(p) error structure is applicable under a wider range of processes than the bias-corrected methods proposed by Bhargava et al. (1982) and Hansen (2007b) in at least two respects: First, it permits the case of a unit root. It is well-known that the LS bias function is discontinuous at unity, making the bias correction dependent on knowledge of whether the true process is stationary or integrated. Second, the restrictions necessary to map from the LS parameters to the true parameter values are not required (see Proposition 1 in Hansen, 2007b). In particular, it is well-known that the bias function becomes non-monotonic in the neighborhood of unity, meaning that there is no unique mapping for highly persistent but stationary processes.

Although the X-differencing estimator can be applied in the case of unit root errors, GLS transformation cannot. Define  $\rho$  as the sum of AR coefficients, i.e.,  $\rho = \sum_j^p \rho_j$ . The GLS estimator does not exist when  $\rho = 1$  since the covariance matrix is not invertible. Also White's correction may not work well unless the cross-sectional dimension is large. The CO approach provides a practical alternative in this situation.

## 2.2. Cochrane–Orcutt Transformation with X-Differencing

Drop the  $Z_{it}\gamma$  and  $\lambda_t$  components from Eq. (1) for simplicity. The CO correction with X-differencing method becomes straightforward by replacing the  $\rho_j$ 's by the X-differenced estimator  $\hat{\rho}_{j,x}$  in Eq. (2). That is, the CO transformed regression with X-differencing is

$$y_{it} - \sum_{j=1}^p \hat{\rho}_{j,x} y_{it-j} = \tilde{\alpha}_i^* + \beta \left( I_{it} - \sum_{j=1}^p \hat{\rho}_{j,x} I_{it-j} \right) + u_{it}^+, \quad t = p + 1, \dots, T \quad (4)$$



where  $u_{it}^+$  is the implied regression error and, importantly,  $\tilde{\alpha}_i^* = \left(1 - \sum_{j=1}^p \hat{\rho}_{j,x}\right) \alpha_i$  is common for all  $t \geq p + 1$ . By abandoning the first  $p$  observations, efficiency is lost on the one hand, but the fixed effects remain time-constant and the simple LSDV method can be used on the other hand. This second point is important because it allows us to use the same method for persistent data as well, in which case the first  $p$  observations are difficult to handle. Note that the conventional  $t$ -statistics have an  $N(0,1)$  limiting distribution like in standard FGLS estimation due to the consistency of the X-differencing estimators (Han et al., 2014). The conventional  $t$ -statistic is defined as

$$t_{\hat{\beta}} = \frac{\hat{\beta}}{\sqrt{\hat{\sigma}_u^2 (\tilde{X}' \tilde{X})^{-1}}}, \text{ for } \hat{\sigma}_u^2 = \frac{1}{(T-p-1)N} \sum_{i=1}^N \sum_{t=p+1}^T (\hat{u}_{it}^+ - \bar{u}_i^+)^2$$

where  $\tilde{X}$  is the pooled matrix of the within-group transformed regressor,  $\hat{u}_{it}^+ = y_{it}^+ - \hat{\beta} \tilde{x}_{it}^+$ , and  $\bar{u}_i^+ = (T-p)^{-1} \sum_{t=p+1}^T \hat{u}_{it}^+$ . Therefore panel robust covariance estimation needs not be used here.

When  $T$  is small, the CO type correction is less efficient compared to FGLS since the CO regressions do not use the first  $p$  observations. Due to this efficiency gain, the FGLS method has been popularly used and studied in the broader panel data literature (e.g., Baltagi & Li, 1991, 1992). Of course, if  $T$  is large enough, the efficiency loss by discarding the first  $p$  observations from each  $i$  under the CO method is negligible. In addition, a strength of the CO approach is that it can be used in cases when the GLS transformation matrix is not well-defined (i.e., when  $\hat{\rho}_x = \sum_{j=1}^p \hat{\rho}_{j,x}$  is equal to or exceeds one).

### 2.3. Estimation and Inference for Alternative Dependence Structures

In the previous section, we discussed how to obtain a more accurate and powerful test statistic by utilizing the X-differencing method. However the asymptotic theory developed in the previous section is based on the crucial assumption that the error terms follow the data generating process given in Eq. (1). In this section, we consider regression errors that do not follow Eq. (1). We consider three cases.

The first case is that the error term does not follow a finite-order AR( $p$ ) but an AR( $\infty$ ), for example, an ARMA( $p,q$ ) process. In this case, all parametric correction methods including Hansen (2007b) and X-differencing do

not work “exactly.” However, both parametric corrections can be used to approximate the serial dependence structure by permitting  $p \rightarrow \infty$  as  $T \rightarrow \infty$ . Under this setting, Lee, Okui, and Shintani (2013) consider the inconsistency of various estimators. Particularly they show that the total bias of the LS estimator can be decomposed into truncation and fundamental biases, where the truncation bias arises due to the lag truncation and the fundamental bias is the Nickell (1981) bias (which disappears as  $T \rightarrow \infty$  but not as  $N \rightarrow \infty$ ). (GMM/IV estimators do not suffer from the fundamental bias. Hence as long as the truncation bias goes away as  $T, p \rightarrow \infty$  jointly, the GMM/IV estimators are consistent.) For the X-differencing estimator, Han et al. (2014, Eq. (32)) provides a formula that relates the full aggregation of X-differences to the sum of cross-products (such as shown in the LS estimation) and the remainder for bias correction. It would thus be natural that Lee et al.’s (2013) results for bias-corrected LS should hold.

To examine the case in more detail, suppose that  $y_{it} = a_i + z_{it}$ ,  $z_{it} = \sum_{j=1}^{\infty} \rho_j z_{it-j} + \varepsilon_{it}$ , but an AR(p) model is fitted by the X-differencing method. That is,  $\tilde{y}_{it,s} \equiv y_{it} - y_{is}$  is regressed on  $\tilde{y}_{it-1,s+1}, \dots, \tilde{y}_{it-p,s+p}$  (with pooling over all  $t$  and  $s$  such that  $t \geq s + p + 1$ ). Let  $\tilde{\varepsilon}_{it,s}$  denote the associated error term, i.e.,  $\tilde{\varepsilon}_{it,s} = \tilde{y}_{it,s} - \sum_{j=1}^p \rho_j \tilde{y}_{it-j,s+j}$ . Let  $\gamma_j = E(z_{it} z_{it-j})$ , which does not depend on  $t$  due to the maintained stationarity. (We discuss here the case with stationary  $z_{it}$  only.) Then the Yule–Walker equations imply that  $\gamma_k = \sum_{j=1}^{\infty} \rho_j \gamma_{|k-j|}$ . Because  $\tilde{y}_{it,s} = z_{it} - z_{is}$ , the expected cross-product of the  $k$ th regressor and the regression error is

$$\begin{aligned}
 E(\tilde{y}_{it-k,s+k} \tilde{\varepsilon}_{it,s}) &= 2 \left[ \left( \gamma_k - \sum_{j=1}^p \rho_j \gamma_{|k-j|} \right) - \left( \gamma_{t-s-k} - \sum_{j=1}^p \rho_j \gamma_{|t-s-k-j|} \right) \right] \\
 &= 2 \sum_{j=p+1}^{\infty} \rho_j (\gamma_{|k-j|} - \gamma_{|t-s-k-j|})
 \end{aligned} \tag{5}$$

for  $k = 1, \dots, p$ . By the Markov inequality and the fact that  $(a - b)^2 \leq 2(a^2 + b^2)$ , we have

$$|E(\tilde{y}_{it-k,s+k} \tilde{\varepsilon}_{it,s})| \leq 4c_{2p} \left( \sum_{j=0}^{\infty} \gamma_j^2 \right)^{1/2}$$

for all  $k$  and  $p$ , where  $c_{2p} = \left( \sum_{j=p+1}^{\infty} \rho_j^2 \right)^{1/2}$ . Let  $\hat{\rho}_1, \dots, \hat{\rho}_p$  be the X-differencing estimator, and let  $\hat{\rho} = \sum_{j=1}^p \hat{\rho}_j$ . Also let  $\rho^{(p)} = \sum_{j=1}^p \rho_j$  and  $\rho = \rho^{(\infty)}$ . Then

$$\hat{\rho} - \rho^{(p)} = \mathbf{1}'_p \hat{\Gamma}_p^{-1} (NT_p^2)^{-1} \sum_{i=1}^N \sum_{t=p+2}^T \sum_{s=1}^{t-p-1} \tilde{w}_{it,s,p} \tilde{\varepsilon}_{it,s}$$

where  $T_p = T - p$ ,  $\hat{\Gamma}_p$  is the denominator matrix divided by  $NT_p^2$ , and  $\tilde{w}_{it,s,p} = (\tilde{y}_{it-1,s+1}, \dots, \tilde{y}_{it-p,s+p})'$ . Under the regularity that the smallest eigenvalue of  $\hat{\Gamma}_p$  is bounded away from zero (technicality that is not dealt with here; see Lee et al., 2013), we have

$$\begin{aligned} |\hat{\rho} - \rho^{(p)}|^2 &\leq \frac{pC}{NT_p} \sum_{i=1}^N \sum_{t=p+2}^T \left( \frac{1}{T_p} \sum_{s=1}^{t-p-1} \tilde{w}_{it,s,p} \tilde{\varepsilon}_{it,s} \right)' \left( \frac{1}{T_p} \sum_{s=1}^{t-p-1} \tilde{w}_{it,s,p} \tilde{\varepsilon}_{it,s} \right) \\ &= pO(p c_{2p}^2) + pO_p(n_p^{-1}), \quad n_p \equiv N(T-p) \end{aligned}$$

as  $p \rightarrow \infty$  and  $n_p \rightarrow \infty$  for some universal constant  $C < \infty$ . Above the first bound on the second line is for the bias, and the second is for the variance. Thus,  $\hat{\rho} = \rho^{(p)} + O_p(p c_{2p}) + O_p(p^{1/2} n_p^{-1/2})$ . As  $|\rho - \rho^{(p)}| \leq c_{1p} \equiv \sum_{j=p+1}^{\infty} |\rho_j|$ , we have

$$|\hat{\rho} - \rho| = O(c_{1p}) + O(p c_{2p}) + O_p(p^{1/2} n_p^{-1/2})$$

which implies that  $\hat{\rho}$  is consistent for  $\rho$  as long as  $p \rightarrow \infty$  and  $p/n_p \rightarrow 0$  under the sufficient regularity conditions that  $\sum_{j=1}^{\infty} |\rho_j| < \infty$  and  $\sum_{j=1}^{\infty} j \rho_j^2 < \infty$ . These regularity conditions are satisfied for finite order ARMA processes because then  $\rho_j$  decays exponentially with  $j$ . Unbiased asymptotic distribution for  $(n_p/p)^{1/2}(\hat{\rho} - \rho)$  requires a bit more:  $(n_p/p)^{1/2} c_{1p} \rightarrow 0$  and  $(p n_p)^{1/2} c_{2p} \rightarrow 0$ . Usually  $c_{1p}$  and  $c_{2p}$  decay exponentially with  $p$ , thus these conditions are casually translated into that  $p$  is not too small compared to the total sample size  $n_p = N(T - p)$ . These conditions are much simpler than those for the estimators considered by Lee et al. (2013) because of the exact uncorrelatedness of the regressors and the regression error in the X-differenced equations for finite order panel AR models. More rigorous treatment would be called for in this regard but is not pursued here.

A second prominent case to consider is a factor error component structure such as that considered in Pesaran (2006), Bai (2009), and Greenaway-McGrevy, Han, and Sul (2012). Error component structures can generate forms of serial dependence that are not encompassed within the set of AR(p) models. For example, the sum of two AR(1) processes can be

equivalently expressed as an ARMA(2,1) or an AR( $\infty$ ). Using a factor augmented estimator may help simplify the problem, since the factor component of the error is controlled for in estimation. See, for example, Hagedorn, Karahan, Manovskii, and Mitman (2013).

However, for approaches that control for factors to the explained and explanatory variables (Greenaway-McGrevy et al., 2012; Pesaran, 2006, etc.), it is unclear how to estimate the common factors to binary variables as the common component would not be additive. This issue is left for future research.

The third case to consider is important for micro panel data but also is a difficult issue for theoretical econometricians. Suppose that the serial correlation of the error term arises from the inclusion of a small trend or integrated component. Typical outcome variables of interest in the DD literature are wages, employment, consumption, or medical expenditures, as Bertrand, Duflo, and Mullainathan (2004) point out. All these variables may exhibit (either stochastic or non-stochastic) trending behavior over time. In this case, the error term exhibits serious serial correlation. For instructive purposes, consider the following example.

#### Example (Trend Nonstationary)

Assume that the dependent variable,  $y_{it}$ , has the following simple latent structure:

$$y_{it} = a_i + b_it + d_i\theta_t + e_{it} \quad (6)$$

where  $\theta_t = \theta_{t-1} + \eta_t$ , and all innovations are assumed to be  $e_{it} \sim \text{i.i.d.}(0, \sigma_e^2)$ ,  $b_i \sim \text{i.i.d.}(b, \sigma_b^2)$ , and  $d_i \sim \text{i.i.d.}(d, \sigma_d^2)$ . If  $d_i = 0$  for all  $i$ , then  $y_{it}$  is not serially correlated. Consider estimating the following simple dynamic regression:

$$y_{it} = c_i + \rho y_{it-1} + u_{it}$$

It is easy to show that the expectation of the within-group LS estimator is

$$E\hat{\rho}_T = 1 - \frac{12\sigma_e^2}{(b^2 + \sigma_b^2)T^2} + O(T^{-3})$$

so that as  $T$  increases, the estimated AR(1) coefficient approaches unity quite quickly. In applications however, the mean  $b$  and variance  $\sigma_b^2$  of the trend coefficients are sufficiently small so that the point estimate of  $\rho$  is not in the neighborhood of unity, even with large  $T$ .

Of course, as  $T$  increases, the dominant term becomes the deterministic trend component so that the trending behavior can be seen obviously. Also even when there is no deterministic trend component, as  $T$  increases, the integrated series  $\theta_t$  becomes the dominant term so that it becomes easy to detect such nonstationary behavior by using a typical panel unit root test. When  $T$  is not large, it is hard to rely on a formal statistical test to identify whether or not the dependent variable contains integrated components. However in this case, one can avoid this thorny issue by taking the first difference. That is,

$$\Delta y_{it} = b_i + \Delta \lambda_t + \beta \Delta I_{it} + \Delta u_{it}$$

where the fixed effects,  $b_i$ , capture the heterogeneous trend coefficients. Of course, the new regression error,  $\Delta u_{it}$ , may follow an ARMA(p,q) process rather than an AR(p). In this case, as we discussed before, one can increase the lag length for a large  $T$ .

FD removes the trend, so that GLS is feasible. For persistent but I(0) errors the induced (negative) serial correlation from over-differencing is corrected in the parametric FGLS or CO transformation. Although the FD errors do not in general follow a finite order AR(p) structure when the errors are AR(p), there should still be substantial efficiency gains from this approach. In this case lag estimation should be based on methods that find the best approximating model asymptotically (see [Lee et al., 2013](#)).

### 3. MONTE CARLO EXPERIMENTS

In this section we verify our asymptotic claims and investigate the finite sample performance of the suggested methods. The data generating process is given by

$$y_{it} = \beta I_{it} + \alpha_i + b_i t + d_i \theta_t + e_{it}$$

where

$$\theta_t = \theta_{t-1} + \eta_t, \quad e_{it} = \sum_{j=1}^p \rho_j e_{it-j} + u_{it}$$

We generate variables as follows:

$$\alpha_i \sim \text{i.i.d. } N(0,1), u_{it} \sim \text{i.i.d. } N(0,1), \eta_t \sim \text{i.i.d. } N\left(0, \sigma_\eta^2\right), b_i \sim \text{i.i.d. } N(0, \sigma_b^2),$$

$$\text{and } d_i \sim \text{i.i.d. } N(0, \sigma_d^2)$$

We set the initial observation of  $e_{i0}$  as i.i.d.  $N(0, 1/(1 - \rho^2))$  and  $\theta_0 = 0$ . The binary variable,  $I_{it}$ , is generated from

$$I_{it} = 1\{x_{it} \geq 0\} \quad \text{where } x_{it} = \phi x_{it-1} + v_{it}$$

Under the alternative, we set  $\beta = 0.1$ . We consider two sets of values of  $\sigma_b^2$  and  $\sigma_d^2$ . Under stationarity, both values are set to zero. Under I(1) errors, we set  $\sigma_b = 0.01$  and  $\sigma_d^2 = 1$  but vary  $\sigma_\eta$  from 0.05 to 0.2. These values are chosen based on PSID consumption expenditure data.

First we consider the case of stationary models. We consider  $\rho = 0.5, 0.8, 0.9$  and  $\phi = 0.8, 0.9$ . Since the over-rejection of the null hypothesis with different  $N/T$  ratios is of interest, broadly two sets of  $T$  and  $N$  are considered. We set  $T = 6$  or  $100$  and vary  $N$ . We report selected results here.

Table 1 shows the rejection rate under the null hypothesis. ‘‘GLS’’ refers to infeasible GLS where the true  $\rho$  value is known. Evidently, for small  $N$

**Table 1.** Finite Sample Performance of Suggested Estimators for Stationary Errors under the Null Hypothesis:  $\phi = 0.8, \rho = 0.8, \beta = 0$ .

N	T	Variance $\times 10^3$						Rejection Rates (Nominal: 5%)				
		X-differencing			Hansen			X-differencing			Hansen	
		WG	CO	FGLS	FGLS	GLS	WG	CO	FGLS	FGLS	GLS	
10	6	264.66	169.77	145.35	152.32	140.57	0.164	0.079	0.082	0.095	0.071	
25	6	95.35	55.99	48.83	50.25	48.15	0.087	0.064	0.060	0.067	0.059	
50	6	45.97	24.72	22.36	22.85	22.25	0.066	0.049	0.051	0.053	0.050	
100	6	23.23	13.44	11.84	11.98	11.79	0.069	0.062	0.063	0.063	0.062	
200	6	10.96	6.33	5.62	5.68	5.61	0.056	0.047	0.055	0.056	0.056	
400	6	5.47	3.19	2.84	2.87	2.84	0.049	0.051	0.053	0.058	0.053	
5	100	89.91	14.41	14.37	14.37	14.32	0.209	0.081	0.076	0.075	0.075	
6	100	74.91	12.43	12.32	12.33	12.26	0.181	0.086	0.083	0.082	0.083	
7	100	59.66	9.80	9.74	9.76	9.72	0.159	0.072	0.071	0.071	0.069	
8	100	52.39	7.86	7.80	7.81	7.79	0.131	0.061	0.058	0.059	0.059	
9	100	43.92	7.32	7.22	7.22	7.22	0.116	0.063	0.059	0.060	0.058	
10	100	39.94	6.21	6.17	6.17	6.18	0.115	0.057	0.055	0.055	0.055	

all estimators except for the within-group estimator (labeled WG) based on the panel robust covariance estimator reject the null hypothesis at a rate similar to GLS. Both FGLS and CO methods reject the null hypothesis that  $\beta=0$  slightly less than Hansen’s FGLS estimator when  $T$  is small, but the difference is negligible. It is important to report that the over-rejection rate of the WG estimator is not dependent on the degree of serial dependence. Even when  $\rho$  is 0.5, the rejection frequencies of the WG estimator are similar to those for smaller values of  $\rho$ . Hence the finding of the over-rejection of the null hypothesis by BDM was not mainly due to high serial dependence of the regression errors but the small- $N$  statistical properties of the panel robust covariance estimator which requires a large  $N$  for consistency. As long as  $N$  is large, the serial correlation has a negligible effect, as BDM report.

Table 2 reports the nominal power of the tests. Evidently, all estimators except the WG estimator with the panel robust covariance estimator perform very well. When  $T$  is small, the performance of CO with X-differencing is worse than GLS but better than WG in terms of variance for the considered data generating process. Of course, CO need not be more efficient than WG especially if the serial correlation in the error term is minor. In terms of power, the performance of CO with X-differencing is only slightly worse than FGLS estimators even when  $T$  is small.

**Table 2.** Comparison of Powers of Tests for Stationary Errors under the Alternative:  $\phi=0.8, \rho=0.8, \beta=0.1$ .

N	T	Variance $\times 10^3$					Rejection Rates (nominal 5%)				
		WG	X-differencing		Hansen		WG	X-differencing		Hansen	
			CO	FGLS	FGLS	GLS		CO	FGLS	FGLS	GLS
10	6	264.66	169.77	145.35	152.32	140.57	0.166	0.091	0.098	0.118	0.094
25	6	95.35	55.99	48.83	50.25	48.15	0.112	0.088	0.096	0.106	0.094
50	6	45.97	24.72	22.36	22.85	22.25	0.103	0.095	0.106	0.116	0.107
100	6	23.23	13.44	11.84	11.98	11.79	0.114	0.142	0.173	0.180	0.172
200	6	10.96	6.33	5.62	5.68	5.61	0.162	0.222	0.277	0.284	0.274
400	6	5.47	3.19	2.84	2.87	2.84	0.275	0.421	0.488	0.492	0.486
5	100	89.91	14.41	14.37	14.37	14.32	0.231	0.176	0.173	0.176	0.173
6	100	74.91	12.43	12.32	12.33	12.26	0.209	0.201	0.199	0.200	0.198
7	100	59.66	9.80	9.74	9.76	9.72	0.182	0.225	0.224	0.221	0.222
8	100	52.39	7.86	7.80	7.81	7.79	0.172	0.226	0.226	0.226	0.226
9	100	43.92	7.32	7.22	7.22	7.22	0.150	0.246	0.245	0.245	0.244
10	100	39.94	6.21	6.17	6.17	6.18	0.141	0.265	0.270	0.269	0.272

Lastly, Table 3 provides the results for I(1) errors. We consider only three estimators: the WG estimator, first differenced LS estimator (FD), and first differenced CO estimator (FD-CO). The panel robust covariance estimator is used to construct  $t$ -statistics for the first two estimators. Meanwhile the standard  $t$ -test is used for the FD-CO. The fourth column in Table 3 reports the estimated means of the AR(1) coefficient from the Monte Carlo simulation. These are dependent on the value of  $\sigma_d$  and  $T$ : as either  $\sigma_d$  or  $T$  increases, the expected value of  $\hat{\rho}$  increases. Note that the true  $\rho$  is set as 0.2. Evidently, the rejection rate of the null hypothesis of the WG estimator is reasonable when  $T$  is moderately large when the null is true ( $\beta=0$ ). It is a natural result since as long as  $N$  is large, the WG estimator provides accurate rejection rates. In terms of size distortion, FD-CO performs best. We do not report other FGLS estimators in Table 3 to save the space but their performances are very much similar to the FD-CO estimator. The power of the LS estimator does not improve with the size of  $T$ , because, as discussed above, the estimator converges at the  $\sqrt{N}$  rate. Meanwhile both FD and FD-CO estimators reject the null more as  $T$  increases.

**Table 3.** Comparison of the First-Differenced CO with X-Differencing to other Estimators:  $\rho = 0.2, \phi = 0.8$ .

$\beta=0$	$\sigma_d$	$T$	$N$	$E(\hat{\rho})$	Variance $\times 10^3$			Rejection Rates		
					WG	FD	FD-CO	WG	FD	FD-CO
	0.05	10	50	0.926	271.20	37.18	31.54	0.101	0.127	0.084
	0.05	25	50	0.940	102.12	12.44	10.33	0.075	0.065	0.057
	0.05	50	50	0.945	49.23	5.96	5.07	0.059	0.055	0.055
	0.05	100	50	0.948	25.92	3.02	2.54	0.065	0.055	0.065
	0.2	10	50	0.937	358.89	37.60	33.46	0.104	0.124	0.080
	0.2	25	50	0.950	137.38	14.04	12.30	0.075	0.076	0.069
	0.2	50	50	0.954	73.23	6.58	5.76	0.062	0.055	0.055
	0.2	100	50	0.954	34.04	3.27	2.96	0.059	0.054	0.067
$\beta=0.1$	0.05	10	50	0.926	266.38	34.10	29.98	0.117	0.173	0.138
	0.05	25	50	0.941	99.26	12.34	10.51	0.086	0.186	0.201
	0.05	50	50	0.946	48.67	6.23	5.30	0.085	0.261	0.325
	0.05	100	50	0.948	24.69	2.98	2.55	0.116	0.434	0.528
	0.2	10	50	0.937	374.72	37.07	32.34	0.117	0.164	0.118
	0.2	25	50	0.950	134.14	13.45	12.05	0.077	0.170	0.178
	0.2	50	50	0.953	67.71	6.75	5.91	0.085	0.252	0.286
	0.2	100	50	0.955	35.18	3.29	2.91	0.101	0.438	0.505



#### 4. CONCLUDING REMARKS

In this paper we propose a simple parametric transformation for LS estimation of DD regressions that exhibit serially dependent errors. First stage estimates of autoregressive structures in the error are obtained by using the Han et al. (2011, 2014) X-differencing transformation of the panel, before applying an FGLS or CO type estimator of the regression equation. The X-differencing method is simple to implement and is unbiased in large  $N$  settings, and can be applied to both stationary and unit root error processes.

We also consider the case where the error processes exhibits either a non-stochastic or a stochastic trend. In this case we suggest first differencing the data before applying the X-differencing transformation to account for any residual serial dependence in the first-differenced series. This method can account for a wider variety of dependence structures in the error term than X-differencing in levels, including possible linear trends. The suggested method performs well in Monte Carlo studies.

#### NOTE

1. If  $\varepsilon_{it}$  is stationary, then both the least squares dummy variables (LSDV) and the first difference least squares (FD) estimators are consistent as the product  $NT$  approaches infinity. If  $\varepsilon_{it} \sim I(1)$ , the convergence rate of the FD estimator is  $\sqrt{NT}$ , and that of the LSDV estimator is  $\sqrt{N}$ . The latter can be shown by evaluating the maximal eigenvalue of the covariance matrix of  $(\varepsilon_{i1}, \dots, \varepsilon_{iT})'$ , which is  $O(T^2)$  because the sum of eigenvalues is bounded by the trace (Lütkepohl, 1996, section 5.3.1).

#### ACKNOWLEDGMENT

This paper was written for the presentation at the 14th Advances in Econometrics Conference in honor of Peter C.B. Phillips at Southern Methodist University on November 1–3, 2014. We thank Tom Fomby for organizing the conference and Yoosoon Chang and Joon Y. Park for preparing the conference program. We also thank Badi Baltagi for helpful comments. Research by Han was supported by Korea University (K1222321).

#### REFERENCES

Arellano, M. (1987). Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics*, 49(4), 431–434.

- Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, 77(4), 1229–1279.
- Baltagi, B. H., & Li, Q. (1991). A transformation that will circumvent the problem of autocorrelation in an error component model. *Journal of Econometrics*, 48, 385–393.
- Baltagi, B. H., & Li, Q. (1992). Prediction in the one-way error component model with serial correlation. *Journal of Forecasting*, 11, 561–567.
- Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*, 119(1), 249–275.
- Bester, C. A., Conley, T. G., & Hansen, C. B. (2011). Inference with dependent data using cluster covariance estimators. *Journal of Econometrics*, 165(2), 137–151.
- Bhargava, A., Franzini, L., & Narendranathan, W. (1982). Serial correlation and the fixed effects model. *Review of Economic Studies*, 49, 533–549.
- Blundell, R., & Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, 87(1), 115–143.
- Cameron, A. C., Miller, D. L., & Gelbach, J. B. (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90(3), 414–427.
- Cochrane, D., & Orcutt, G. (1949). Application of least squares regression to relationships containing autocorrelated error terms. *Journal of the American Statistical Association*, 44(245), 32–61.
- Donald, S. G., & Lang, K. (2007). Inference with differences-in-differences and other panel data. *Journal of Business and Economic Statistics*, 89(2), 221–233.
- Greenaway-McGrevy, R., Han, C., & Sul, D. (2012). Asymptotic distribution of factor augmented estimators for panel regression. *Journal of Econometrics*, 169(1), 48–53.
- Hagedorn, M., Karahan, F., Manovskii, I., & Mitman, K. (2013). *Unemployment benefits and unemployment in the great recession: The role of macro effects*. Working Paper No. 19499. Retrieved from <http://www.nber.org/papers/w19499>
- Han, C., Phillips, P. C. B., & Sul, D. (2011). Uniform asymptotic normality in stationary and unit root autoregression. *Econometric Theory*, 27, 1117–1151.
- Han, C., Phillips, P. C. B., & Sul, D. (2013). *Lag length selection in panel autoregression*. Mimeo, University of Texas at Dallas.
- Han, C., Phillips, P. C. B., & Sul, D. (2014). X-differencing and dynamic panel model estimation. *Econometric Theory*, 30, 201–251.
- Hansen, C. B. (2007a). Asymptotic properties of a robust variance matrix when T is large. *Journal of Econometrics*, 141, 597–620.
- Hansen, C. B. (2007b). Generalized least squares inference in panel and multilevel models with serial correlation and fixed effects. *Journal of Econometrics*, 140, 670–694.
- Kiefer, N. M. (1980). Estimation of fixed effect models for time series of cross-sections with arbitrary intertemporal covariance. *Journal of Econometrics*, 14, 195–202.
- Lee, Y.-J., Okui, R., & Shintani, M. (2013). *Asymptotic inference for dynamic panel estimators of infinite order autoregressive processes*. Mimeo, Kyoto University.
- Lee, Y.-S. (2006). *A general approach to bias correction in dynamic panel models under time series misspecification*. Mimeo, Yale University.
- Lee, Y.-S. (2012). *Model selection in the presence of incidental parameters*. Manuscript, University of Michigan.
- Lütkepohl, H. (1996). *Handbook of matrices*. John Wiley & Sons, Chichester.
- Miller, D. L., Cameron, A. C., & Gelbach, J. B. (2011). Robust inference with multi-way clustering. *Journal of Business and Economic Statistics*, 29(2), 238–249.

- Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica*, 49, 1417–1426.
- Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74(4), 967–1012.
- Wooldridge, J. M. (2003). Cluster-sample methods in applied econometrics. *American Economic Review*, 93(2), 133–138.

### APPENDIX

**Proof of Theorem 1.** Let  $v_{it} = \alpha_i + \varepsilon_{it} + \lambda_i$ ; let  $\check{v}_{it} = v_{it} - N^{-1} \sum_{j=1}^N v_{jt}$  and similarly define  $\check{\alpha}_i$  and  $\check{\varepsilon}_{it}$  so  $\check{v}_{it} = \check{\alpha}_i + \check{\varepsilon}_{it}$ . We have  $\hat{v}_{it} = \check{v}_{it} - \check{X}'_{it}(\hat{\beta} - \beta)$ , where  $\check{X}_{it} = X_{it} - N^{-1} \sum_{k=1}^N X_{kt}$ . When  $\varepsilon_{it} = \sum_{j=1}^p \rho_j \varepsilon_{it-j} + u_{it}$ , we have  $\check{\varepsilon}_{it} = \sum_{j=1}^p \rho_j \check{\varepsilon}_{it-j} + \check{u}_{it}$ , where  $\check{u}_{it}$  is serially white noise. Thus, the moment conditions for X-differencing (Han et al., 2014, Eq. (7)) hold despite the presence of cross-sectional correlation in  $\check{\varepsilon}_{it}$ , and the consistency of the X-differencing estimator using  $\check{v}_{it}$  instead of  $\hat{v}_{it}$  follows from the convergences of  $\sum_{i=1}^N \sum_{t=4}^T \sum_{s=1}^{t-3} (\check{\varepsilon}_{it-1} - \check{\varepsilon}_{is+1})^2$  and  $\sum_{i=1}^N \sum_{t=4}^T \sum_{s=1}^{t-3} (\check{\varepsilon}_{it} - \check{\varepsilon}_{is})(\check{\varepsilon}_{it-1} - \check{\varepsilon}_{is+1})$  after appropriate normalization as  $NT \rightarrow \infty$ . (Note the convergence even with fixed  $N$  as long as  $NT \rightarrow \infty$ .) Now the result follows from the consistency of  $\hat{\beta}$ .

**Construction of FGLS Estimator when  $p \geq 2$ .** With consistent estimates of  $\rho_1, \dots, \rho_p$  available, we can then construct an FGLS estimator. When  $\varepsilon_{it}$  is stationary and Eq. (3) is satisfied, let  $\gamma_j = E(\varepsilon_{it}\varepsilon_{it-j})/\sigma^2$ . Then by standard text book algebra  $\gamma_j$  are determined by  $\rho_j$  as follows:

$$\begin{aligned} \gamma_0 &= \rho_1\gamma_1 + \rho_2\gamma_2 + \dots + \rho_p\gamma_p + 1 \\ \gamma_1 &= \rho_1\gamma_0 + \rho_2\gamma_1 + \dots + \rho_p\gamma_{p-1} \\ &\vdots \\ \gamma_p &= \rho_1\gamma_{p-1} + \rho_2\gamma_{p-2} + \dots + \rho_p\gamma_0 \\ \gamma_j &= \rho_1\gamma_{j-1} + \rho_2\gamma_{j-2} + \dots + \rho_p\gamma_{j-p}, \quad j > p \end{aligned}$$

The first  $p + 1$  identities are written as  $\gamma_{0:p} = (A_1 + A_2)\gamma_{0:p} + (1, 0, \dots, 0)'$ , where  $\gamma_{0:p} = (\gamma_0, \gamma_1, \dots, \gamma_p)'$ ,  $A_1$  is the  $(p + 1) \times (p + 1)$  matrix whose  $i$ th row is  $(0, \rho_i, \rho_{i+1}, \dots, \rho_p, 0')$  with the first 0 being a scalar, and  $A_2$  is the  $(p + 1) \times (p + 1)$  matrix whose  $i$ th row is  $(\rho_{i-1}, \rho_{i-2}, \dots, \rho_1, 0')$ . Taking  $p=4$  as an illustrative example, we have

$$A_1 = \begin{pmatrix} 0 & \rho_1 & \rho_2 & \rho_3 & \rho_4 \\ 0 & \rho_2 & \rho_3 & \rho_4 & 0 \\ 0 & \rho_3 & \rho_4 & 0 & 0 \\ 0 & \rho_4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ \rho_1 & 0 & 0 & 0 & 0 \\ \rho_2 & \rho_1 & 0 & 0 & 0 \\ \rho_3 & \rho_2 & \rho_1 & 0 & 0 \\ \rho_4 & \rho_3 & \rho_2 & \rho_1 & 0 \end{pmatrix}$$

So  $(\gamma_0, \dots, \gamma_p)'$  is the first column of  $(I - A_1 - A_2)^{-1}$ , and the rest  $\gamma_j$  are obtained by recursion. One can then construct an estimate of  $\Omega = \text{Toeplitz}(\gamma_0, \gamma_1, \dots, \gamma_{T-1})$  using  $\hat{\rho}_j$ , and do the FGLS after first-differencing in order to eliminate the fixed effects. Specifically, letting  $y_i = (y_{i1}, \dots, y_{iT})'$ ,  $X_i = (X_{i1}, \dots, X_{iT})'$ , and  $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{iT})'$ , we have  $y_i = X_i\beta + 1\alpha_i + \varepsilon_i$ . Premultiplying the  $(T - 1) \times T$  first-differencing operator matrix  $\Delta$ ,

$$\Delta = \begin{pmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -1 & 1 \end{pmatrix}$$

so that the  $t$ th row of  $\Delta y_i$  is  $y_{it} - y_{it-1}$ , we have  $\Delta y_i = \Delta X_i + \Delta \varepsilon_i$ , where the fixed effects are eliminated. Then the error term  $\Delta \varepsilon_i$  has covariance  $\sigma^2 \Delta \Omega \Delta'$ , and the FGLS estimator is

$$\hat{\beta}_{\text{fglS}} = \left[ \sum_{i=1}^n X_i' \Delta' (\Delta \tilde{\Omega} \Delta')^{-1} \Delta X_i \right]^{-1} \sum_{i=1}^n X_i' \Delta' (\Delta \tilde{\Omega} \Delta')^{-1} \Delta y_i$$

where  $\tilde{\Omega}$  is obtained by replacing  $\rho_j$  with  $\hat{\rho}_j$  in the  $\gamma_j$  formulae. This estimator generalizes Bhargava et al.'s (1982) panel AR(1) FGLS to AR(p) and provides a simple alternative (requiring no numerical solutions and free from the non-monotonicity issue near unit root) to Hansen's (2007b) procedure. The  $\hat{\beta}_{\text{fglS}}$  estimator has all the properties of FGLS estimators under regularity if Eq. (3) is true and  $\varepsilon_{it}$  is stationary.

The FGLS estimator can also be obtained by the following procedure. The equations are first transformed to

$$\tilde{y}_{it} = \tilde{X}'_{it} \beta + (1 - \rho^*) \alpha_i + \tilde{\varepsilon}_{it} \tag{A.1}$$

for  $t = 1, \dots, T$ , so that  $\tilde{\varepsilon}_{it}$  are serially uncorrelated, where  $\rho^* = \sum_{j=1}^p \rho_j$ . Precisely, for  $t > p$ ,  $\tilde{y}_{it} = y_{it} - \sum_{j=1}^p \rho_j y_{it-j}$ ,  $\tilde{X}_{it} = X_{it} - \sum_{j=1}^p \rho_j X_{it-j}$ , and  $\tilde{\varepsilon}_{it} = u_{it}$  work, and for  $t \leq p$ , we can have  $\tilde{y}_{it} = [y_{it} - (y_{i1}, \dots, y_{it-1}) \delta_t] (1 - \rho^*) / (1 - 1' \delta_t)$  and  $\tilde{X}_{it}$  and  $\tilde{\varepsilon}_{it}$  are similarly obtained, where  $\delta_t = \Gamma_t^{-1} (\gamma_t, \gamma_{t-1}, \dots, \gamma_1)'$  and  $\Gamma_t = \text{Toeplitz}(\gamma_0, \gamma_1, \dots, \gamma_{t-1})$ . Then  $\tilde{\varepsilon}_{it}$  are serially uncorrelated but heteroskedastic. Especially,  $E \tilde{\varepsilon}_{it}^2 = \sigma^2 (1 - \rho^*)^2 h_t$ , where  $h_t = 1$  for

$t > p$ , and  $h_t = (-\delta'_t, 1)\Gamma_t(-\delta'_t, 1)' / (1 - 1'\delta_t)^2$  for  $t \leq p$ . This heteroskedasticity is unavoidable should the individual effects remain constant over time. Procedures as simple as LSDV would therefore not be available, and one should do GLS anyway after eliminating the fixed effects by first-differencing.