Estimation of Dynamic Panel Data Models with Cross-Sectional Dependence: Using Cluster Dependence for Efficiency

Valentin Verdier*

November 10, 2014

Abstract

This paper considers the estimation of dynamic panel data models when data are suspected to exhibit cross-sectional dependence. A new estimator is defined that uses cross-sectional dependence for efficiency while being robust to the misspecification of the form of the cross-sectional dependence. We show that using cross-sectional dependence for estimation is important to obtain an estimator that is more efficient than existing estimators. This new estimator also uses nuisance parameters parsimoniously so that it is exhibits good small and large sample properties even when the number of time periods is large. As an empirical application, we estimate the effect of attending private school on student achievement using a value added model.

Keywords: Panel Data, Dynamic Models, Cross-Sectional Dependence, Optimal Instruments, Value-Added Models of Student Achievement

*Department of Economics, University of North Carolina, Chapel Hill, NC 27599, United States. Tel.: +1 919-966-3962. E-mail address: vverdier@email.unc.edu.

1 Introduction

In some econometric studies of panel data, researchers want to account for the presence of feedback between the dependent variable and explanatory variables, i.e. for current values of the dependent variable to affect future values of the explanatory variables or even for both dependent and independent variables to be jointly determined. The simplest example of such models is the dynamic panel data model where lagged values of the dependent variable are used as covariates. In such cases, explanatory variables can not be treated as strictly exogenous. In virtually all panel data applications, researchers also want to control for unobserved heterogeneity that affects the dependent variable but might also be correlated with the covariates.

The presence of both non strictly exogenous covariates and unobserved heterogeneity in panel data models causes many estimation methods to be invalid (see for instance Wooldridge (2010)). In the context of cross-sectionally independent data, a valid estimator for dynamic panel data models that relies on first differencing and instrumental variables has been defined in early work by Anderson and Hsiao (1981). Additionally, an asymptotically efficient estimator is found in Arellano and Bond (1991)[1]. In the rest of the paper, we refer to this estimator as the AB estimator. These estimators often suffer from having a large variance because the instrumental variables that they use are weak.[2] In addition, inference for the AB estimator is often unsatisfactory when the number of time periods in the data set is relatively large because of problems due to using many moment conditions, as studied in Alvarez and Arellano (2003) or Windmeijer (2005) for the case of cross-sectional independence.

In this paper, we consider the estimation of panel data models with covariates that are not strictly exogenous when data also exhibit cross-sectional dependence. We will define a new

[1]With cross-sectionally independent data, the Arellano and Bond estimator is asymptotically efficient in the class of estimators using linear functions of the instruments.

[2]To address this problem, papers such as Ahn and Schmidt (1995), Arellano and Bover (1995), and Blundell and Bond (1998) considered using for estimation additional assumptions such as homoscedasticity, serial uncorrelation of the transitory shocks, or restrictions on initial conditions. Another approach to obtain efficiency gains by using additional assumptions can be found in the literature on First Difference Quasi-Maximum Likelihood estimation, as in Hsiao et al. (2002) or Han et al. (2014) for instance, which rely on assumptions of homoscedasticity and serial uncorrelation of the transitory shocks. We do not consider these estimators here since we are interested in estimators that are consistent under the only assumption of mean independence of the transitory shock, without any other assumption holding.

estimator that is more efficient than the AB estimator and for which inference is significantly better in small samples. The main reason why our estimator is more efficient than previous estimators that were defined for data with cross-sectional independence is that it makes use of cross-sectional dependence to obtain stronger instruments.

In order to obtain an estimator with not only good properties in terms of point estimation, but also good properties for inference, we use an auxiliary model for optimal instruments. Optimal instruments are instruments that, once interacted with corresponding moment functions, provide an optimal set of exactly identifying moment conditions so that the resulting estimator achieves the asymptotic efficiency bound for estimating unknown parameters from the assumption of mean independence of the transitory shocks. Optimal instruments for estimating dynamic panel data models without cross-sectional dependence are found in Chamberlain (1992) and they can be generalized to the case of cross-sectional dependence. In this paper, we propose auxiliary assumptions sufficient to model optimal instruments for panel data models with covariates that are not strictly exogenous and cross-sectional dependence. The advantage of such an approach is that it provides a systematic way of weighting many moment conditions while making use of few nuisance parameters. As a result, our estimator exhibits good small sample properties and inference while being robust to the misspecification of our model of optimal instruments.

Arellano (2003) and Alvarez and Arellano (2004) have previously considered modeling optimal instruments for dynamic panel data models in the special case of cross-sectional independence. We show that cross-sectional dependence can be particularly useful to obtain more accurate estimators. Previous work on dynamic panel data models that has considered cross-sectional dependence has not made use of this dependence to obtain stronger instruments. Mutl (2006), for instance, studied a GMM estimator based on the same moment conditions as in Anderson and Hsiao (1981) or Arellano and Bond (1991) and only uses an optimal weighting matrix based on a specific model of spatial dependence. Elhorst (2005) and Su and Yang (2013) generalized maximum likelihood estimators as in Hsiao et al. (2002) to the case of cross-sectional dependence but these estimators are not robust to heteroscedasticity, serial

correlation of the transitory shocks or misspecification of the cross-sectional dependence.

In Section 2, we present the simplest example of the models we consider, the dynamic panel data model without covariates for data with cross-sectional dependence, and characterize a general class of consistent and asymptotically normal estimators for this model under large n, fixed T asymptotics. In Section 3, we define our estimator and compare it to existing estimators. We also show that our estimator is consistent and asymptotically normal unbiased even when large n, large T asymptotics are used, i.e. that it is not subject to the issue of many instruments. Finally, we show how to generalize the estimator defined for the simple dynamic model without covariates to models with covariates and sequentially exogenous instruments. In Section 4, we present Monte-Carlo evidence that the efficiency gains from using cross-sectional dependence for estimation can be significant and that the estimator we propose has superior small sample properties compared to existing estimators. In Section 5, we apply our estimator to the estimation of the effect of attending private school on student achievement using a value-added model and taking into account the possibility that student achievements are correlated within schools.

2 Dynamic Panel Data Models with Cross-Sectional Dependence

2.1 The Model

Throughout the paper we will consider large n, fixed T asymptotics, except for Section 3.4, which derives the large n, large T asymptotic properties of the new estimator we define in the next section. Consider first the model for any observation i from a sample of n observations and any time period t from a fixed number T of time periods:

$$y_{it} = \rho_0 y_{it-1} + c_i + u_{it} \tag{2.1}$$

$$E(u_{it}|Y_{t-1}) = 0 \tag{2.2}$$

where $Y_t = [Y_{1t}^{'}, ..., Y_{nt}^{'}]^{'}$ and $Y_{it} = [y_{i0}, ..., y_{it}]^{'}$ are random vectors that stack values of y_{it} across time and observations and c_i are unobserved effects constant over time, also called unobserved heterogeneity. We also assume that $\rho_0 \neq 1$ so that ρ_0 is identified from differenced equations as seen in the next subsection.

In the case where there is no cross-sectional dependence, (2.1) and (2.2) correspond to the linear dynamic model for panel data as presented in Arellano and Bond (1991) for instance. When there is cross-sectional dependence, (2.1) and (2.2) impose the restriction that cross-sectional dependence does not cause Y_{t-1} to be endogenous. For instance if contemporaneous spatial lags were omitted variables in (2.1), then (2.2) would be violated. Some papers such as Cizek et al. (2011), Elhorst (2005), Su and Yang (2013) and Baltagi et al. (2014) have considered models with both dynamic effects and contemporaneous spatial lag effects. Since estimators for such models rely on correct specification of the form of cross-sectional dependence, we do not consider them here and concentrate on models where cross-sectional dependence of some unknown form is present in the residuals.[3] Lagged values of the dependent variable of neighboring observations could also be included in the model as covariates to control for dynamic cross-sectional effects. We will discuss models with covariates in Section 3.5.

The objective of the next section is to characterize estimators for ρ_0 that are consistent when (2.1) and (2.2) hold under general conditions on the form of cross-sectional dependence in c_i and u_{it}.

2.2 Consistent Estimation

The presence of unobserved heterogeneity rules out estimation of ρ_0 by a regression. Because (2.1) and (2.2) form a dynamic model, fixed effects estimation is also ruled out because explanatory variables are not strictly exogenous.

[3]We can also note that, with cross-sectional dependence, it is not likely for $E(u_{it}|Y_{it-1}) = 0$ to hold without (2.2) holding. If (2.2) is not satisfied, it is likely that both estimators for cross-sectionally independent data such as the Arellano and Bond estimator and the alternative estimator proposed in this paper will be inconsistent. For instance suppose for simplicity that $n = 2$ and $E(u_{1t}|Y_{t-1}) = \alpha + \beta_1 y_{1t-1} + \beta_2 y_{2t-1} \neq 0$ so that $\beta_1 \neq 0$ or $\beta_2 \neq 0$. Then $E(u_{1t}|Y_{1t-1}) = \alpha + \beta_1 y_{1t-1} + \beta_2 E(y_{2t-1}|Y_{1t-1})$ and it is likely that $E(y_{2t-1}|Y_{1t-1})$ is a function of $y_{10}, ..., y_{1t-2}$ in addition to y_{1t-1} so that, in general, $\alpha + \beta_1 y_{1t-1} \neq -\beta_2 E(y_{2t-1}|Y_{1t-1})$ and $E(u_{1t}|y_{1t-1}) \neq 0$.

To estimate ρ_0, we will consider a forward filtering transformation as in Arellano and Bover (1995).[4] Define:

$$m_{it}(\rho) = \tilde{y}_{it} - \rho \tilde{y}_{it,-1} \; \forall \, t = 1, ..., T-1 \tag{2.3}$$

where $\tilde{y}_{it} = y_{it} - \frac{1}{T-t} \sum_{s=t+1}^{T} y_{is}$ and $\tilde{y}_{it,-1} = y_{it-1} - \frac{1}{T-t} \sum_{s=t+1}^{T} y_{is-1}$. Therefore, $m_{it}(\rho_0) = \tilde{u}_{it} = u_{it} - \frac{1}{T-t} \sum_{s=t+1}^{T} u_{is}$ and (2.1) and (2.2) imply:

$$E(m_{it}(\rho_0)|Y_{t-1}) = 0 \; \forall \, t = 1, ..., T-1 \tag{2.4}$$

Define $m_i(\rho) = [m_{it}(\rho)]_{t=1,...,T-1}$ to be the column vector with $m_{it}(\rho)$ as its t^{th} element. We will also use the notation $\tilde{Y}_i = [\tilde{y}_{it}]_{t=1,...,T-1}$, $\tilde{U}_i = [\tilde{u}_{it}]_{t=1,...,T-1}$ and $\tilde{Y}_{-1,i} = [\tilde{y}_{it,-1}]_{t=1,...,T-1}$.

Define:

$$z_i = [z_{i1}, ..., z_{iT-1}] \tag{2.5}$$

to be a row vector of dimension $1 \times (T-1)$ containing instruments for each time period, so that z_{it} is a function of Y_{t-1}. Therefore we have $E(z_{it} m_{it}(\rho_0)) = 0$ and:[5]

$$E(z_i m_i(\rho_0)) = \sum_{t=1}^{T-1} E(z_{it} m_{it}(\rho_0)) = 0 \tag{2.6}$$

Define $\hat{\rho}$ to be the estimator obtained from solving the sample analogue of (2.6):

$$\sum_{i=1}^{n} z_i m_i(\hat{\rho}) = 0 \tag{2.7}$$

so that:

$$\hat{\rho} = \frac{\sum_{i=1}^{n} z_i \tilde{Y}_i}{\sum_{i=1}^{n} z_i \tilde{Y}_{-1,i}} \tag{2.8}$$

$$= \rho_0 + \frac{\sum_{i=1}^{n} z_i \tilde{U}_i}{\sum_{i=1}^{n} z_i \tilde{Y}_{-1,i}} \tag{2.9}$$

Note that defining $\hat{\rho}$ in this way is more general than first appears since z_i is free to be defined in any way. For instance, $\hat{\rho}$ is asymptotically equivalent to the GMM estimator of ρ_0 from the moment functions $m_i(\rho)$ with overidentifying instruments Z_i and weighting matrix

[4] In this paper, a first difference transformation instead of forward filtering would eventually lead to the same proposed estimator, but we use forward filtering because it will simplify notation in later sections.

[5] Note that we need to assume $\rho_0 \neq 1$ for $E(z_i m_i(\rho)) = 0$ to hold for $\rho = \rho_0$ only since if $\rho_0 = 1$ then $E(z_i m_i(\rho)) = 0 \; \forall \, \rho$.

Ξ when z_i is chosen to be $z_i = E(Z_i\tilde{Y}_{-1,i})'\Xi Z_i$.

Conditions that are sufficient for $\hat{\rho}$ to be consistent and asymptotically normal are:

1. The absolute value of $D_n = E(n^{-1}\sum_{i=1}^n z_i\tilde{Y}_{-1,i})$ is uniformly positive for n large enough, i.e. $|D_n| > c > 0,\ \forall\, n > N$ for some constants c and N.

2. $\sigma_n^2 = n^{-1}Var(\sum_{i=1}^n z_i\tilde{U}_i) > c > 0,\ \forall\, n > N$.

3. $plim(n^{-1}\sum_{i=1}^n z_i\tilde{Y}_{-1,i} - D_n) = 0$ as $n \to \infty$.

4. $plim(n^{-1}\sum_{i=1}^n z_i\tilde{U}_i) = 0$ as $n \to \infty$.

5. $\sigma_n^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n z_i\tilde{U}_i \xrightarrow{d} N(0,1)$ as $n \to \infty$.

Conditions 1 and 2 are commonly imposed regularity conditions. Condition 1 implies that the instruments z_i are useful for predicting the covariates $\tilde{Y}_{-1,i}$. If $D = lim_{n\to\infty}(D_n)$ exists, then Condition 1 can be replaced by $D \neq 0$. Condition 2 guarantees that a particular summand $z_i\tilde{U}_i$ is asymptotically ignorable, i.e. can't offset the information accumulated from other summands. If $\sigma^2 = lim_{n\to\infty}\sigma_n^2$ exists, then Condition 2 can be replaced by $\sigma^2 > 0$.

Since $E(z_i\tilde{U}_i) = 0$ under (2.2), Conditions 3, 4, and 5 state that weak laws of large numbers and central limit theorems hold. Because we allow for unknown forms of cross-sectional dependence, additional restrictions have to be imposed on the strength of this dependence for Conditions 3, 4, and 5 hold. Here we consider two examples of sets of restrictions that are sufficient for Conditions 3-5 to hold.

The first example is the case of cluster dependence, where the data can be split between several independent clusters, each with a fixed number of observations. While observations across clusters are independent, observations in the same cluster can be dependent in an arbitrary way. In this case, standard results on asymptotic properties with independent observations, found in White (2001) for instance, imply that weak laws of large numbers and central limit theorems hold for sample averages.

The second example is the case of spatial dependence. In this case, some measure of distance is available and cross-sectional dependence decays with distance. Conley (1999),

Jenish and Prucha (2009), Jenish and Prucha (2012) consider different sets of assumptions that guarantee that sample averages are consistent and asymptotically normal in the presence of spatial dependence. We choose the framework of Jenish and Prucha (2012) to apply to our model since it is the most general out of the three papers cited above.

Appendix A presents the formal assumptions that are sufficient for Conditions 3 to 5 to hold for either case, and proves that each set of assumptions is indeed sufficient.

As long as the Conditions 1 to 5 listed above hold, as $n \to \infty$, we have:

$$\hat{\rho} \xrightarrow{p} \rho$$
$$\frac{D_n}{\sigma_n}\sqrt{n}(\hat{\rho} - \rho) \xrightarrow{d} N(0,1)$$

In the next section, we consider a model of optimal instruments, so that the resulting estimator is efficient if our model of optimal instruments is correct. It will however be consistent and asymptotically normal as long as (2.1), (2.2) and Conditions 1-5 hold, independently of whether the auxiliary model of optimal instruments we specify is true or not.

3 Efficient Estimation under Cluster Dependence

In this section, we consider an auxiliary model for deriving optimal instruments that assumes that every observation belongs to one of a large number of clusters. Observations are treated as correlated within clusters but independent across clusters. While clustering only represents a specific form of cross-sectional dependence, it might be a good approximation for more general forms of dependence in many applications. This idea is present for instance in Bester et al. (2011a) which studies the use of clustered standard errors for inference in cases where cross-sectional dependence of unknown form is present. In addition, the method outlined in this section for the special case of clustering can easily be extended to other forms of cross-sectional dependence. Therefore we restrict our attention in this paper to auxiliary models that make use of the cluster dependence assumption. For simplicity we will consider in this section the case where each observation belongs to the same cluster across all time periods but the results in this section can be generalized to clusters changing over time as shown in

Section 5.

Previous work that estimated dynamic models of panel data with clustered sampling generally used estimators developed for i.i.d. data such as the ones found in Anderson and Hsiao (1981), Arellano and Bond (1991), or Ahn and Schmidt (1995), and adjusted inference by using clustered standard errors. Such an analysis can be found for instance in de Brauw and Giles (2008) where farming households are treated as clustered by village or Andrabi et al. (2011) where students are clustered by school. Topalova and Khandelwal (2010) and Balasubramanian and Sivadasan (2010) consider the case where firms are clustered by industry.

In this section, we show that there is much to gain in terms of efficiency by using a different estimator that takes into account correlation within cluster while being robust to the misspecification of the form of this correlation, or of the form of the cross-sectional dependence all together. The next subsection shows how to model optimal instruments for estimating (2.1) and (2.2) and how cross-sectional dependence can be leveraged to obtain stronger instruments.

3.1 Model for the Optimal Instruments

Our model of optimal instruments relies on three auxiliary assumptions. The first assumption of cluster dependence is:

Auxiliary Assumption 1: The data can be divided in a large number of clusters indexed by $g = 1, ..., G$, each with a fixed number of observations denoted n_g, where $0 < n_g < C$ for some constant C. Observations are independent across clusters.

In this section we will index observations by cluster so that for any i, g_i denotes the cluster to which observation i belongs and j_g denotes the j^{th} observation of cluster g so that for any observation i in g, there is j such that $j_g = i$ and $\{\{x_{j_g}\}_{j=1,...,n_g}\}_{g=1,...,G} = \{x_i\}_{i=1,...,n}$ for any sequence of variables $\{x_i\}_{i=1,...,n}$. Consider stacking all observations by cluster and define $m_t^g(\rho) = [m_{1_g,t}(\rho), ..., m_{n_{gg},t}(\rho)]'$, $m^g(\rho) = [m_1^{g'}(\rho), ..., m_{T-1}^{g'}(\rho)]'$, $\tilde{u}_t^g = m_t^g(\rho_0)$ and $\tilde{U}^g = m^g(\rho_0)$. Similarly, define $c^g = [c_{1_g}, ..., c_{n_{gg}}]'$, $y_t^g = [y_{1_g,t}, ..., y_{n_{gg},t}]'$, $\tilde{y}_t^g = [\tilde{y}_{1_g,t}, ..., \tilde{y}_{n_{gg},t}]'$, $\tilde{y}_{t,-1}^g = [\tilde{y}_{1_g,t,-1}, ..., \tilde{y}_{n_{gg},t,-1}]'$, $\tilde{Y}^g = [\tilde{y}_1^{g'}, ..., \tilde{y}_{T-1}^{g'}]'$, $\tilde{Y}_{-1}^g = [\tilde{y}_{1,-1}^{g'}, ..., \tilde{y}_{T-1,-1}^{g'}]'$,

$Y_t^g = [y_0^{g'}, ..., y_t^{g'}]$, $u_t^g = [u_{1_g t}, ..., u_{n_{gg} t}]$, $U^g = [u_1^{g'}, ..., u_T^{g'}]'$.

With Auxiliary Assumption 1, optimal instruments are a function of $Y_{t-1}^{g_i}$ only instead of Y_{t-1}. By generalizing the work on optimal instruments for cross-sectionally independent data found in Chamberlain (1992) to the case of cluster-sampling, Appendix B.1 shows that the optimal estimator for ρ_0 is defined by:

$$\sum_{g=1}^{G} Z_{opt}^g m^g(\hat{\rho}_{opt}) = 0 \tag{3.1}$$

where $Z_{opt}^g = L^{\star g'}(\Phi^g)^{-1/2}$ where $\Phi^g = [Cov(\tilde{u}_t^g, \tilde{u}_s^{g'} | Y_{max\{t,s\}-1}^g)]_{t=1,...,T-1}^{s=1,...,T-1}$, $(\Phi^g)^{-1/2}$ is the upper diagonal matrix such that $(\Phi^g)^{-1/2'}(\Phi^g)^{-1/2} = (\Phi^g)^{-1}$, $L^{\star g} = [L_t^{\star g'}]_{t=1,...,T-1}$ and $L_t^{\star g} = E((\Phi_t^g)^{-1/2}\tilde{Y}_{-1}^g | Y_{t-1}^g)$ where $(\Phi_t^g)^{-1/2}$ is the t^{th} $n_g \times n_g(T-1)$ matrix composing $(\Phi^g)^{-1/2}$.

One could estimate these optimal instruments non-parametrically by using series of instruments that include lagged values of the dependent variable for an observation but also lagged values of the dependent variable for neighboring observations. A similar estimator has been studied for the case of cross-sectionally independent data in Donald et al. (2009) for static models and Hahn (1997) for dynamic models. However such an approach would not be practical here since there are too many possible terms to consider as instruments. Also, it would involve using many nuisance parameters which can cause poor small sample properties for the estimator, as is discussed later.

Instead, we can use two additional auxiliary assumptions to impose parametric restrictions on the model for optimal instruments and drastically reduce the number of nuisance parameters needed. Auxiliary Assumption 2 is an assumption of conditional homoscedasticity, conditional serial uncorrelation, and conditional equi-correlation within clusters:[6]

[6]This Assumption is admittedly strong but presents the advantage of introducing only two nuisance parameters, which results in a parsimonious estimator. We show in this section that the resulting estimator is consistent and asymptotically normal as long as (2.1), (2.2) and Conditions 1-5 hold, independently of whether heteroscedasticity is present or not. In addition, Monte Carlo results in Section 4 show that, for the specific data generating processes considered, the presence of heteroscedasticity does not result in large losses in efficiency for the estimator we propose. However, with large datasets, one could use an estimator that is efficient under more general conditions by modeling heteroscedasticity as, for instance: for any $i, j \in g$, $t, s = 1, ..., T$,

Auxiliary Assumption 2: For any $i, j \in g$, $t, s = 1, ..., T$, $t \geq s$:

$$\begin{aligned} Cov(u_{it}, u_{js}|c^g, Y^g_{t-1}) &= \sigma^2_u \; if \; i = j, \; t = s \\ &= \tau_u \sigma^2_u \; if \; i \neq j, \; t = s \\ &= 0 \; if \; t > s \end{aligned}$$

Under Auxiliary Assumption 2, Appendix B.2 shows that the optimal instrument for $m^g_t(\rho)$, $z^g_{opt,t}$, is a linear function of $E(\tilde{y}^g_{t,-1}|Y^g_{t-1})$, i.e. that optimal instruments are a linear function of the best prediction of the covariates in (2.4), $\tilde{y}_{it-1}$, based on the instruments available, Y^g_{t-1}.

From (2.1) and (2.2):

$$E(\tilde{y}^g_{t,-1}|Y^g_{t-1}) = y^g_{t-1} - \frac{1}{T-t} \sum_{s=t+1}^{T} E(y^g_{s-1}|Y^g_{t-1}) \tag{3.2}$$

$$= y^g_{t-1} - \frac{1}{T-t} \sum_{s=t+1}^{T} (\rho_0^{s-t} y^g_{t-1} + \sum_{r=0}^{s-t-1} \rho_0^r E(c^g|Y^g_{t-1})) \tag{3.3}$$

$$= y^g_{t-1}(1 - \frac{1}{T-t}\rho_0 \frac{1-\rho_0^{T-t}}{1-\rho_0}) + E(c^g|Y^g_{t-1})(-\frac{1}{1-\rho_0} + \frac{1}{T-t}\frac{\rho_0}{1-\rho_0}\frac{1-\rho_0^{T-t}}{1-\rho_0}) \tag{3.4}$$

Hence we see that the quality of prediction of $\tilde{y}^g_{t-1}$ based on Y^g_{t-1}, and hence the strength of the instruments and the efficiency of the resulting estimator, will depend on the quality of the prediction of c^g based on Y^g_{t-1}. In many applications, it is very likely that agents that belong to the same cluster will have levels of unobserved heterogeneity that are related. For instance, farmers that live in the same village might farm plots with similar soil quality or develop similar farming practices over time. Firms that operate in the same industry

---

$t \geq s$,

$$\begin{aligned} Cov(u_{it}, u_{js}|c^g, Y^g_{t-1}) &= \alpha + \beta u^2_{it-1} \; if \; i = j, \; t = s \\ &= \eta + \gamma u_{it-1} u_{jt-1} \; if \; i \neq j, \; t = s \\ &= 0 \; if \; t \neq s \end{aligned}$$

From this model, let $w_{it-1} = y_{it-1} - \rho_0 y_{it-2}$, we obtain: $Var(u_{it}|Y^g_{t-1}) = \alpha + \beta E((y_{it-1} - \rho_0 y_{it-2} - c_i)^2|Y^g_{t-1}) = \alpha + \beta((w_{it-1})^2 - 2w_{it-1}E(c_{it}|Y^g_{t-1}) + E(c^2_i|Y^g_{t-1})$ and $Cov(u_{it}, u_{jt}|Y^g_{t-1}) = \eta + \gamma(w_{it-1}w_{jt-1} - w_{it-1}E(c_j|Y^g_{t-1}) - w_{jt-1}E(c_i|Y^g_{t-1}) + E(c_i c_j|Y^g_{t-1})$. Hence, in order to obtain a model for Φ^g, models for $E(c^2_i|Y^g_{t-1})$ and $E(c_i c_j|Y^g_{t-1})$ are needed in addition to a model for $E(c_i|Y^g_{t-1})$. This can result in introducing many new nuisance parameters and complicates the estimator significantly since it also complicates the derivation of a model for $L^{\star g}$.

might also face similar constraints such as for instance regulation or access to skilled labor force. Similarly, households that live in the same district might have been selected based on common characteristics such as wealth, income, family status or values. As a result, in many applications, we can expect that using information from other observations in the same cluster in addition to one's own previous outcomes can provide a better predictor for one's level of unobserved heterogeneity, and consequently stronger instruments as well as a more efficient estimator.

Auxiliary Assumption 3 enables us to derive a model for the mean of unobserved heterogeneity conditional on lagged values of the dependent variable:

Auxiliary Assumption 3: Suppose that for any cluster $g = 1, ..., G$:

$$\begin{bmatrix} c^g \\ y_0^g \\ u^g \end{bmatrix} \sim N(\begin{bmatrix} \mu_c \iota_{n_g} \\ \frac{1}{1-\rho_0}\mu_c \iota_{n_g} \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_c^g & & \\ \frac{1}{1-\rho_0}\Sigma_c^g & \frac{1}{(1-\rho_0)^2}\Sigma_c^g + \frac{1}{1-\rho_0^2}\Sigma_u^g & \\ 0 & 0 & I_T \otimes \Sigma_u^g \end{bmatrix}) \tag{3.5}$$

where $\Sigma_u^g = \sigma_u^2 \begin{bmatrix} 1 & & & \\ \tau_u & 1 & & \\ ... & & ... & \\ \tau_u & ... & \tau_u & 1 \end{bmatrix}$, $\Sigma_c^g = \sigma_c^2 \begin{bmatrix} 1 & & & \\ \tau_c & 1 & & \\ ... & & ... & \\ \tau_c & ... & \tau_c & 1 \end{bmatrix}$, and ι_{n_g} is a column vector of ones of dimension $n_g \times 1$.

In addition to joint normal distribution of c^g and $\{u_t^g\}_{t=1,...,T}$, Auxiliary Assumption 3 assumes that the initial distribution of y_{it} is stationary, i.e.:

$$y_0^g = \frac{c^g}{1-\rho_0} + \dot{u}_0^g \tag{3.6}$$

where $\dot{u}_0^g$ is independent of c^g and $\{u_t^g\}_{t=1,...,T}$, follows normal distribution with zero mean, variance equal to $\sigma_u^2/(1-\rho_0^2)$, and has a within cluster correlation of τ_u.[7]

From Auxiliary Assumption 3, we can derive the distribution of $[c^{g'}, y_0^{g'}, c^{g'} + u_1^{g'}, ..., c^{g'} + u_T^{g'}]'$ as a multivariate normal with mean 0 and variance V^g. Therefore, using the properties

[7]The auxiliary assumption of stationary initial conditions can easily be generalized, at the expense of

of the multivariate normal distribution, under Auxiliary Assumptions 1-3, $E(c^g|Y_t)$ can be obtained as a linear function of $y_0^g, c^g + u_1^g, ..., c^g + u_t^g$ with coefficients given by the elements of V^g (note that $c^g + u_t^g = y_t^g - \rho_0 y_{t-1}^g$, so that it is a function of observed variables and the estimable parameter ρ_0). The exact form of $E(c^g|Y_t)$ under Auxiliary Assumptions 1, 2, 3 is given in Appendix B.3.

3.2 Definition of a New Estimator and Asymptotic Distribution

Only five nuisance parameters compose V^g: σ_u^2, τ_u, μ_c, σ_c^2, τ_c. Preliminary estimators of ρ_0 and of these five nuisance parameters can be estimated from the non-linear regression model given by (3.4):

$$E(\tilde{y}_{it,-1}|Y_{t-1}^g) = a_t(\rho_0)y_{it-1} + b_t(\rho_0)M(c_i|Y_{t-1}^g, \theta_0)$$
$$a_t(\rho) = (1 - \frac{1}{T-t}\rho\frac{1-\rho^{T-t}}{1-\rho})$$
$$b_t(\rho) = (-\frac{1}{1-\rho} + \frac{1}{T-t}\frac{\rho}{1-\rho}\frac{1-\rho^{T-t}}{1-\rho})$$

where $\theta_0 = \{\rho_0, \sigma_u^2, \tau_u, \mu_c, \sigma_c^2.\tau_c\}$ and $M(c_i|Y_{t-1}^g, \theta_0)$ is the model for $E(c_i|Y_{t-1}^g)$ as a function of Y_{t-1}^g and the nuisance parameters θ_0 obtained in Appendix B.3. A preliminary estimator $\hat{\theta}$ can be defined from the pooled non-linear least-squares regression:

$$\hat{\theta} = argmin_\theta \sum_{i=1}^{n}\sum_{t=1}^{T}(\tilde{y}_{it,-1} - a_t(\rho)y_{it-1} - b_t(\rho)M(c_i|Y_{t-1}^g, \theta))^2 \tag{3.7}$$

This non-linear regression resembles a pooled "first-stage regression", where covariates are regressed on instruments. The difference here is that the parametric model of optimal instruments developed in the previous section is used to drastically reduce the number of nuisance parameters present in the first-stage regression. If Auxiliary Assumption 1-3 hold, θ_0 will be estimated consistently from this non-linear regression. In general, Appendix B.4

introducing three additional nuisance parameters, by assuming:

$$y_0^g = \alpha + \beta c^g + \dot{u}_0^g$$
$$\dot{u}_0^g|c^g \sim N(0, \tilde{\Sigma}_0)$$
$$Var(\dot{u}_{i0}) = \dot{\sigma}_0$$
$$Corr(\dot{u}_{i0}, \dot{u}_{j0}) = \tau_u \, if \, i \neq j \, but \, g_i = g_j$$

shows that $\hat{\theta}$ will converge to $\ddot{\theta}$ defined by:

$$\ddot{\theta} = argmin_\theta \sum_{t=1}^{T} E((\tilde{y}_{it,-1} - a_t(\rho)y_{it-1} - b_t(\rho)M(c_i|Y_{t-1}^g, \theta))^2) \tag{3.8}$$

where we have assumed that $E((\tilde{y}_{it,-1} - a_t(\rho)y_{it-1} - b_t(\rho)M(c_i|Y_{t-1}^g, \theta))^2)$ is constant across i for simplicity.

Let $\hat{\Phi}^g$ be the estimator for the variance-covariance matrix $\Phi^g = Var(\tilde{U}^g)$ composed of $\hat{\sigma}_u$ and $\hat{\tau}_u$ from the formula derived in Appendix B.2. Let $\hat{\mu}_t^{gc} = [M(c_i|Y_{t-1}^g, \hat{\theta})]_{i\in g}$ be the estimator of $E(c^g|Y_t)$ from the model $M(x_i|Y_t^g, \theta)$ given in Appendix B.3.

A consistent estimator for the optimal instrument for $m^g(\rho)$ under (2.1) and (2.2) and Auxiliary Assumptions 1, 2, 3 is:

$$\hat{Z}_{opt}^g = [(a_1(\hat{\rho}_p)y_0^g + b_1(\hat{\rho}_p)\hat{\mu}_1^{gc})', ..., (a_{T-1}(\hat{\rho}_p)y_{T-2}^g + b_{T-1}(\hat{\rho}_p)\hat{\mu}_{T-1}^{gc})'](\hat{\Phi}^g)^{-1} \tag{3.9}$$

where $\hat{\rho}_p$ is the estimator of ρ_0 found as the first element of $\hat{\theta}$. The estimator obtained from using this instrument is $\hat{\rho}^\star$ defined by:

$$\sum_{g=1}^{G} \hat{Z}_{opt}^g m^g(\hat{\rho}^\star) = 0 \tag{3.10}$$

So that:

$$\hat{\rho}^\star = \frac{\sum_{g=1}^G \hat{Z}_{opt}^g \tilde{Y}^g}{\sum_{g=1}^G \hat{Z}_{opt}^g \tilde{Y}_{-1}^g} \tag{3.11}$$

$$= \rho_0 + \frac{\sum_{g=1}^G \hat{Z}_{opt}^g \tilde{U}^g}{\sum_{g=1}^G \hat{Z}_{opt}^g \tilde{Y}_{-1}^g} \tag{3.12}$$

Let $\ddot{Z}_{opt}^g$ be the random vector defined as in (3.9) but where $\hat{\theta}$ is replaced by $\ddot{\theta}$. Appendix B.4 shows that $\hat{\rho}^\star$ is asymptotically equivalent to:

$$\ddot{\rho}^\star = \rho_0 + \frac{\sum_{g=1}^G \ddot{Z}_{opt}^g \tilde{U}^g}{\sum_{g=1}^G \ddot{Z}_{opt}^g \tilde{Y}_{-1}^g} \tag{3.13}$$

When (2.1), (2.2) hold and Conditions 1-5 of Section 2.2 hold for $[z_i]^{i\in g} = \ddot{Z}_{opt}^g$, $\ddot{\rho}^\star$, and

consequently $\hat{\rho}^\star$, are consistent for ρ_0 and asymptotically normal:

$$\sqrt{G}(\hat{\rho}^\star - \rho_0) \xrightarrow{d} N(0, V_\rho) \tag{3.14}$$

$$V_\rho = D^{-2}\sigma^2 \tag{3.15}$$

$$D = lim_{G\to\infty}(\frac{1}{G}\sum_{g=1}^{G} E(\ddot{Z}^g_{opt}\tilde{Y}^g_{-1})) \tag{3.16}$$

$$\sigma^2 = lim_{G\to\infty}(\frac{1}{G}Var(\sum_{g=1}^{G} \ddot{Z}^g_{opt}\tilde{U}^g)) \tag{3.17}$$

where here we assumed that D and σ^2 exist for notational simplicity.[8]

For inference, if we have cluster sampling (Auxiliary Assumption 1), then:

$$\sigma^2 = lim_{g\to\infty}(\frac{1}{G}\sum_{g=1}^{G} Var(\ddot{Z}^g_{opt}\tilde{U}^g)) \tag{3.18}$$

In this case, standard errors for $\hat{\rho}^\star$ that are consistent as long as (2.1), (2.2) and Assumption 1 of cluster dependence hold are given by:

$$s.e. = ((\sum_{g=1}^{G} \hat{Z}^g_{opt}\tilde{Y}^g_{-1})^{-2}\sum_{g=1}^{G}(\hat{Z}^g_{opt}m^g(\hat{\rho}^\star))^2)^{1/2} \tag{3.19}$$

If the assumption of cluster sampling is violated and instead we have some measure of distance and can assume that cross-sectional dependence decays with distance as in Assumption 3 of Appendix A, one can use non-parametric estimators for σ^2. Statistical tests with general forms of spatial dependence are available and have been discussed in Conley (1999), Kelejian and Prucha (2007), Bester et al. (2011b), Kim and Sun (2011) and Bester et al. (2011a).

3.3 Comparison to Existing Estimators

The estimator defined by (3.10) can be rewritten as $\hat{\rho}^*$ that satisfies the equation:

$$\sum_{g=1}^{G} w^{g*}(\hat{\theta})Z^g m^g(\hat{\rho}^*) = 0 \tag{3.20}$$

[8]When Auxiliary Assumptions 1, 2, 3 also hold, $D = \sigma^2 = E(L^{\star g'}(\Phi^g)^{-1}L^{\star g})$, so that $Avar(\sqrt{G}(\hat{\rho}^\star - \rho_0)) = E(L^{\star g'}(\Phi^g)^{-1}L^{\star g})^{-1}$, and $\hat{\rho}^\star$ indeed achieves the efficiency bound for estimating ρ_0 from (2.1) and (2.2) derived in Appendix B.1.

where Z^g is the matrix containing all valid instruments for m^g, $Z^g = diag(\{I_{ng} \otimes Y_t^g\}_{t=0,...,T-2})$, and $w^{g*}(.)$ is the deterministic row vector function such that $w^{g*}(\hat{\theta})Z^g = \hat{Z}^g_{opt}$.

The Arellano and Bond estimator[9] can also be written as exactly identified from:

$$\sum_{g=1}^{G} \hat{w}^g_{AB} Z^g m^g(\hat{\rho}_{AB}) = 0 \tag{3.21}$$

where:

$$\hat{w}^g_{AB} = \sum_{i=1}^{n} (\tilde{Y}'_{-1,i} Z_i')(\sum_{i=1}^{n} Z_i m_i(\tilde{\rho}) m_i(\tilde{\rho})' Z_i')^{-1} S^g \tag{3.22}$$

where $\tilde{\rho}$ is a preliminary consistent estimator and S^g is the matrix of zeros and ones such that $S^g Z^g m^g(\rho) = \sum_{i \in g} Z_i m_i(\rho)$ where $Z_i = diag(\{Y_{it}\}_{t=0,...,T-2})$.

In the presence of cross-sectional dependence, it is likely that our estimator will perform better than the Arellano and Bond estimator even when some of the Auxiliary Assumptions 1, 2, 3 are violated because our estimator gives non-zero weights to moment conditions obtained from using instruments from neighboring observations, $E(Z_j m_i(\rho_0)) = 0$, $i \neq j$. As discussed in the previous section, these instruments may have significant predictive power for the covariates in the forward filtered equations, so that these additional moment conditions can be useful to improve the accuracy of the estimator.

In addition, our estimator relies on the estimation of only six nuisance parameters to compute weights for all $n_g^2 \times T \times (T-1)/2$ moment conditions available per cluster, whereas the Arellano and Bond estimator relies on the estimation of $T \times (T-1)/2$ weights. As a result, Alvarez and Arellano (2003) showed that the Arellano and Bond estimator is asymptotically biased under large n, large T asymptotics in the context of cross-sectionally independent data. In addition, Windmeijer (2005) showed that inference for the Arellano and Bond estimator is very poor when T is relatively large. Because our estimator makes use of few nuisance parameters, it will have good properties even when T is relatively large. As evidence of

[9]So-called system GMM estimators presented in Ahn and Schmidt (1995), Arellano and Bover (1995), and Blundell and Bond (1998) are similar to the Arellano and Bond estimator but use additional moment conditions based on additional assumptions of homoscedasticity, absence of serial correlation or stationary initial conditions. Quasi-maximum likelihood estimators such as in Hsiao et al. (2002) or Han et al. (2014) also rely on these additional assumptions for consistency. Since our estimator is consistent under the only assumption of mean independence of transitory shocks conditional on past outcomes, it is more robust than these estimators.

the superior properties of our estimator when T is relatively large, Section 3.4 shows that $\hat{\rho}^{\star}$ is consistent and asymptotically normal unbiased under large n, large T asymptotics, independently of the relative rates at which T and n grow. In addition, Monte Carlo simulation results presented in Section 4 show that our estimator has significantly better small sample properties than the Arellano and Bond estimator in terms of bias, efficiency and quality of inference, even when our model of optimal instruments does not correspond to the true data generating process or when there is no cross-sectional dependence.

3.4 Large n, Large T Asymptotics

The advantage of using a parsimonious model for optimal instruments instead of an unrestricted optimal GMM estimation method can be formalized by looking at the asymptotic distribution of our estimator under large n, large T asymptotics.

Appendix C derives the asymptotic distribution of $\hat{\rho}^{\star}$ under asymptotics where n and T both grow unboundedly. It shows that $\hat{\rho}^{\star}$ is consistent and asymptotically normal unbiased under large n, large T asymptotics, independently of the relative rates at which n and T grow, i.e. our estimator does not suffer from the problem of many instruments unlike optimal GMM estimators such as the Arellano and Bond estimator.

When deriving the asymptotic distribution of $\hat{\rho}^{\star}$ as n and T grow unboundedly, we need to impose $|\rho_0| < 1$ instead of $\rho_0 \neq 1$, so that explosive time series behaviors are ruled out. We also strengthen (2.2) to:

$$E(u_{it}|Y_{t-1}, c) = 0 \tag{3.23}$$

where $c = \{c_i\}_{i=1,...,n}$. This assumption implies that u_{it} is mean independent of U_{t-1} in addition to Y_{t-1} since $u_{it-1} = y_{it-1} - \rho_0 y_{it-2} - c_i$. We also assume:

$$Cov(u_{it}u_{jt}|Y_{t-1}, c) = \sigma_{ij}^{u} \tag{3.24}$$

We also assume that the process has reached its stationary distribution:

$$y_{it} = \frac{c_i}{1-\rho_0} + \sum_{s=0}^{\infty} \rho_0^s u_{it-s} \tag{3.25}$$

These three assumptions simplify derivations significantly and are in line with the assumptions made in other papers deriving the large n large T asymptotic properties of estimators of dynamic panel data models in the context of cross-sectional independence, such as for instance Hahn and Kuersteiner (2002), Alvarez and Arellano (2003), or Hayakawa (2009).

As n, T, grow to infinity, under (2.1), (3.23), (3.24), (3.25), under restrictions of $L_{2+\delta}$ integrability as well as of cross-sectional and serial near epoch dependence shown in Assumptions 2 and 4 of the Appendix, Appendix C shows that we have:

$$\sqrt{nT}(\hat{\rho}^{\star} - \rho_0) \xrightarrow{d} N(0, (D^{\infty})^{-2}\sigma_{\infty}^2)$$

$$D^{\infty} = lim_{n,T\to\infty}(\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} E(w_{it}\ddot{w}_{it}))$$

$$w_{it} = \sum_{s_1=0}^{\infty} \rho_0^{s_1} u_{it-1-s_1}$$

$$\ddot{w}_{it} = a_{\Sigma^{-1}}^{g_i} w_{it} + b_{\Sigma^{-1}}^{g_i} \bar{w}_t^{g,-i}$$

$$\sigma_{\infty}^2 = lim_{n,T\to\infty}(\frac{1}{nT}\sum_{t} Var(\sum_{i} w_{it}(a_{\Sigma^{-1}}^{g_i} u_{it} + b_{\Sigma^{-1}}^{g_i} \bar{u}_t^{g,-i})))$$

where $\bar{x}^{g,-i} = \frac{1}{n_{g_i}-1}\sum_{j\in g_i, j\neq i} x_j$ for any variable x_i, a_{Σ}^{g} and b_{Σ}^{g} are scalars defined in Appendix C by $(\ddot{\sigma}_u^2 \begin{bmatrix} 1 & & & \\ \ddot{\tau}_u & 1 & & \\ ... & & ... & \\ \ddot{\tau}_u & ... & \ddot{\tau}_u & 1 \end{bmatrix})^{-1} = \begin{bmatrix} a_{\Sigma^{-1}}^{g} & & & \\ \frac{1}{n_g-1}b_{\Sigma^{-1}}^{g} & a_{\Sigma^{-1}}^{g} & & \\ ... & & ... & \\ \frac{1}{n_g-1}b_{\Sigma^{-1}}^{g} & ... & \frac{1}{n_g-1}b_{\Sigma^{-1}}^{g} & a_{\Sigma^{-1}}^{g} \end{bmatrix}$, $\ddot{\sigma}_u$, $\ddot{\tau}_u$ being the elements of $\ddot{\theta} = lim_{n,T\to\infty} argmin_{\theta} \sum_{i=1}^{n}\sum_{t=1}^{T} E(\tilde{y}_{it,-1} - a_t(\rho)y_{it-1} - b_t(\rho)M(c_i|Y_{t-1}^{g}, \theta))^2$ corresponding to σ_u and τ_u in θ_0, and where we have assumed that D^{∞}, σ_{∞}^2 exist for notational simplicity.

Hence the estimator defined in this paper is indeed asymptotically normal unbiased under large n, large T asymptotics.

For inference with unknown forms of cross-sectional and serial dependence, one can use a non-parametric estimator for σ_{∞}^2 and obtain critical values for the resulting tests as in Kim and Sun (2013).[10]

[10] Kim and Sun (2013)consider fixed effects IV estimation so that the estimator they study resembles ours

3.5 Model with Covariates

The estimator defined in the previous sections can be extended to a more general model with covariates:

$$y_{it} = x_{it}\beta_0 + c_i + u_{it} \tag{3.26}$$

$$E(u_{it}|Z_t) = 0 \tag{3.27}$$

where $Z_t = [Z_{1t}^{'}, ..., Z_{nt}^{'}]^{'}$, $Z_{it} = [z_{i1}^{'}, ..., z_{it}^{'}]^{'}$. Z_t defined in this way implies that it is a set of sequentially exogenous instruments, i.e. that the set of instruments increases with time. (2.1) and (2.2) form a special case of (3.26) and (3.27) with $x_{it} = y_{it-1}$ and $z_{it} = y_{it-1}$.

The same operation of forward filtering that was used in the previous sections can be applied to (3.26) and (3.27). Similarly, Auxiliary Assumptions 1 and 2 can be generalized to this model with covariates without any modification except that in Auxiliary Assumption 2, $Cov(u_{it}, u_{js}|c^g, Z_t^g)$ is considered instead of $Cov(u_{it}, u_{js}|c^g, Y_{t-1}^g)$.

In order to model optimal instruments, in addition to Auxiliary Assumption 1 and 2, we can impose the following structure on the first stage model linking z_{it} to x_{it}:

$$x_{it} = z_{it}\gamma + d_i + v_{it} \tag{3.28}$$

$$E(v_{it}|Z_t) = 0 \tag{3.29}$$

Then, as in Appendix B.2, we can show that the optimal instruments for estimating β_0 from the forward filtered moment functions obtained from (3.26) and (3.27) are linear functions of

$$E(\tilde{x}_{it}|Z_t) = E(\tilde{z}_{it}|Z_t)\gamma \tag{3.30}$$

Hence, we see that to obtain a model of optimal instruments in the presence of covariates, one should specify a model for $E(\tilde{z}_{it}|Z_t)$, i.e. for the dynamics in z_{it}. For the simple dynamic model without covariates of the previous sections, the model for the dynamics in the instruments was the same model describing the relationship between dependent variable and

closely but makes use of strictly exogenous instruments. One could work out the details of generalizing their approach the case of sequentially exogenous instruments but this is left for future work.

covariates. In the presence of covariates, we need an additional auxiliary model, which we can specify as:[11]

$$z_{it} = z_{it-1}\lambda + e_i + w_{it} \tag{3.31}$$

$$E(w_{it}|z_{it-1}) = 0 \tag{3.32}$$

With this additional model, we can show as in the previous sections that $E(\tilde{z}_{it}|Z_t)$ is a function of z_{it} and $E(e_i|Z_t)$, and that $E(e_i|Z_t)$ can be modeled in a parsimonious way by generalizing Auxiliary Assumptions 1, 2 and 3 to (3.31) and (3.32). Consequently, an estimator of the nuisance parameters needed to calculate $E(\tilde{z}_{it}|Z_t)$ can be obtained as in Section 3.2 from non-linear regressions of $\tilde{z}_{it}$ on z_{it} and the model obtained for $E(e_i|Z_t)$. A preliminary estimator for γ can be obtained from regressing $\tilde{x}_{it}$ on the estimated $E(\tilde{z}_{it}|Z_t)$. Under Auxiliary Assumption 2 generalized to (3.26) and (3.27), only two nuisance parameters compose $Var(\tilde{u}_t^g|Z_t)$, σ_u^2 and τ_u, which can be estimated from any consistent preliminary estimator of β_0, $\check{\beta}$. Indeed a consistent estimator of $\tilde{u}_{it}$ can be obtained as $\check{\tilde{u}}_{it} = \tilde{y}_{it} - \tilde{x}_{it}\check{\beta}$ and under Auxiliary Assumption 2 generalized to (3.26) and (3.27), we have $\sigma_u^2 = \frac{T-t}{T-t+1}Var(\tilde{u}_{it})$, $\tau_u\sigma_u^2 = \frac{T-t}{T-t+1}Cov(\tilde{u}_{it}, \tilde{u}_{jt})$, $g_i = g_j$, $i \neq j$, which can be estimated consistently from sample variances and covariances of $\tilde{u}_{it}$.

Hence the asymptotically efficient estimator for β_0 from (3.26) and (3.27) can be defined in the same way that an efficient estimator for the simple dynamic model was defined in Section 3.2.[12]

[11] Note that this dynamic model for the instruments allows for the presence of strictly exogenous instruments, i.e. instruments z_{it}^s such that $z_{it}^s = z_{is}^s \ \forall\, t, s = 1, ..., T$, since in such a case, (3.31) and (3.32) will hold with $\lambda[s] = I[s]$, $e_i[s] = 0$, $w_{it}[s] = 0$, where $A[s]$ denotes the s^{th} vector of matrix A.

[12] The estimator we propose in this paper leverages cross-sectional dependence to obtain stronger instruments, so that it is less likely to suffer from a weak instruments problem than estimators designed for cross-sectionally independent data. However, in specific applications, one might still be interested in testing whether the instruments used are weak. Because the estimator defined in this paper ultimately takes the form of an exactly identified IV estimator, the weak IV test developed in Olea and Pflueger (2013) can be applied to our model and estimator. The framework used by Olea and Pflueger (2013) allows for cross-sectional dependence, serial dependence, and heteroscedasticity. In order to compute their test, one has to estimate consistently long-run variance covariance matrices of sample averages, which is done easily in the presence of cluster dependence. In the presence of spatial dependence of unknown form, one can use non-parametric estimators of long-run variances discussed in Section 3.2.

4 Monte Carlo Simulations

In this section, we present results from Monte Carlo simulations that study the small sample performance of the new estimator defined in the previous section for data generating processes where the true form of cross-sectional dependence is cluster dependence (Tables 1 and 2) and others where it is spatial dependence (Table 3). The first main result is that, when the number of time periods is large compared to the size of the cross-sectional sample, and for any data generating process considered here, modeling optimal instruments parsimoniously results in estimators with virtually no bias compared to an unrestricted optimal GMM estimator. This corresponds to the formal results in Section 3.4 that shows that our estimator is asymptotically unbiased under large n, large T asymptotics, i.e. is not subject to the many instruments problem. The second main result is that, in the presence of cross-sectional dependence, using the auxiliary assumption of cluster sampling to model optimal instruments results in significant gains in efficiency when the assumption of cluster sampling is a good approximation for the true form of cross-sectional dependence. The third main result is that inference for our estimator is also not subject to the problem of many instruments, i.e. tests have correct size even when the number of time periods is relatively large, which corresponds to the formal results in Section 3.4 that show that the asymptotic distribution of our estimator under large n large T asymptotics is the limit of the asymptotic distribution derived under large n fixed T asymptotics.

4.1 Results for Data Generating Processes with Cluster Dependence

As in Section 3, we index observations by cluster so that for any i, g_i denotes the cluster to which observation i belongs and j_g denotes the j^{th} observation of cluster g, so that for any observation i in g, there is j such that $j_g = i$ and $\{\{x_{j_g}\}_{j=1,...,n_g}\}_{g=1,...,G} = \{x_i\}_{i=1,...,n}$ for any sequence of variables $\{x_i\}_{i=1,...,n}$, where n_g is the number of observations in cluster g and G is the number of clusters. We stack all observations by cluster and define: $x_t^g = [x'_{1_g,t}, ..., x'_{n_{gg},t}]'$ for any vector of variables x_{it}.

In this section we consider three data generating processes for a model with cluster corre-

lation and without covariates that can be defined by:

$$
\begin{aligned}
c^g &\sim F_c \\
y_0^g &= \frac{c^g}{1-\rho_0} + \dot{u}_0^g \\
\dot{u}_0^g &\sim Normal(0, \dot{\Sigma}_0^g) \\
y_t^g &\sim \rho_0 y_{t-1}^g + c^g + u_t^g \, \forall \, t = 1, ..., T \\
u_t^g &\sim Normal(0, \Sigma_u^g(u_{t-1}^g)) \, \forall \, t = 1, ..., T
\end{aligned}
$$

We study the small sample properties of estimators in three different scenarios: ideal conditions where auxiliary assumptions 1-3 are satisfied, cross-sectional independence, and general within cluster correlation where auxiliary assumptions 2 and 3 are violated.

More precisely, the ideal conditions scenario uses the following parameterization: $F_c = Normal(0, \begin{bmatrix} 1 & & & \\ 0.5 & 1 & & \\ ... & & ... & \\ 0.5 & ... & 0.5 & 1 \end{bmatrix})$, $\dot{\Sigma}_0^g = \frac{1}{1-\rho_0^2} \begin{bmatrix} 1 & & & \\ 0.5 & 1 & & \\ ... & & ... & \\ 0.5 & ... & 0.5 & 1 \end{bmatrix}$, $\Sigma_u^g(u_{t-1}^g) = \begin{bmatrix} 1 & & & \\ 0.5 & 1 & & \\ ... & & ... & \\ 0.5 & ... & 0.5 & 1 \end{bmatrix}$.

The cross-sectional independence scenario uses: $F_c = Normal(0, \begin{bmatrix} 1 & & & \\ 0 & 1 & & \\ ... & & ... & \\ 0 & ... & 0 & 1 \end{bmatrix})$, $\dot{\Sigma}_0^g = \frac{1}{1-\rho_0^2} \begin{bmatrix} 1 & & & \\ 0 & 1 & & \\ ... & & ... & \\ 0 & ... & 0 & 1 \end{bmatrix}$, $\Sigma_u^g(u_{t-1}^g) = \begin{bmatrix} 1 & & & \\ 0 & 1 & & \\ ... & & ... & \\ 0 & ... & 0 & 1 \end{bmatrix}$.

The non-ideal conditions scenario uses: $F_c = LN(0, \begin{bmatrix} 1 & & & \\ 0.5 & 1 & & \\ ... & & ... & \\ 0.5 & ... & 0.5 & 1 \end{bmatrix})$, $\dot{\Sigma}_0^g = \frac{1}{1-\rho_0^2} \begin{bmatrix} 1 & & & \\ 0.5 & 1 & & \\ ... & & ... & \\ 0.5 & ... & 0.5 & 1 \end{bmatrix}$,

$\Sigma_u^g(u_{t-1}^g) = \begin{bmatrix} \sigma_{i_1t}^2 & & & \\ \sigma_{i_1i_2t} & \sigma_{i_2t}^2 & & \\ ... & & ... & \\ \sigma_{i_1i_{n_g}t} & ... & \sigma_{i_{n_g-1}i_{n_g}t} & \sigma_{i_{n_g}t}^2 \end{bmatrix}$, where $LN(\mu, \Sigma)$ is the distribution with mean μ and variance Σ obtained as a linear function of a vector of independent $Lognormal(0,1)$ random variables, $\sigma_{it}^2 = \frac{1}{2} + \frac{1}{2}u_{it-1}^2$, and $\sigma_{ijt} = \frac{1}{3} + \frac{1}{3}u_{it-1}u_{jt-1}$.

We compare the properties of three estimators of ρ_0: The estimator defined in Arellano and Bond (1991) which we call the AB estimator, the estimator defined by (3.10), which we call Estimator 1 and the estimator defined by (3.10) but with the estimated within-cluster correlations set to zero and the rest of the nuisance parameters estimated with this restriction holding, which we call Estimator 2.[13] As a benchmark for comparison, we also show the results from using an unfeasible optimal estimator (UO) which is asymptotically optimal in the class of estimators that use linear functions of the instruments. This estimator weights optimally all available moment conditions that use linear instruments using the true unobserved optimal weights so that it is defined by:

$$\hat{\rho}_{UO} = \frac{\sum_{g=1}^{G} w^g Z^g \tilde{Y}^g}{\sum_{g=1}^{G} w^g Z^g \tilde{Y}_{-1}^g} \tag{4.1}$$

$$w^g = \Delta^{g'}(W^g)^{-1} \tag{4.2}$$

$$\Delta^g = E(Z^g \tilde{Y}_{-1}^g) \tag{4.3}$$

$$W^g = E(Z^g \tilde{U}^g \tilde{U}^{g'} Z^{g'}) \tag{4.4}$$

where $\tilde{U}^g = [\tilde{u}_1^{g'}, ..., \tilde{u}_{T-1}^{g'}]'$, $\tilde{Y}^g = [\tilde{y}_1^{g'}, ..., \tilde{y}_{T-1}^{g'}]'$, $\tilde{Y}_{-1}^g = [\tilde{y}_{1,-1}^{g'}, ..., \tilde{y}_{T-1,-1}^{g'}]'$, and Z^g is the matrix containing all valid instruments for $\tilde{U}^g$ defined in Section 3.3.

In the ideal conditions scenario, Auxiliary Assumptions 1-3 hold so that the UO estimator and Estimator 1 are asymptotically equivalent and efficient whereas the AB estimator and Estimator 2 are not asymptotically efficient. In the cross-sectional independence scenario,

[13] In the two first scenarios we simulate, transitory shocks will be homoscedastic, serially uncorrelated and the dependent variable will be stationary so that additional moment conditions presented in Arellano and Bover (1995), Ahn and Schmidt (1995) or Blundell and Bond (1998) hold. We do not include estimators that use these moment conditions however since we are interested in studying the properties of estimators that are robust to these moment conditions being false.

Auxiliary Assumptions 1-3 also hold so that the UO estimator, the AB estimator, and Estimators 1 and 2 are all asymptotically equivalent and efficient. In the non-ideal conditions scenario, none of the feasible estimators are asymptotically efficient, but we expect Estimator 1 to perform better than the AB estimator or Estimator 2 since Estimator 1 makes use of instruments from other observations in a cluster so that it uses a weighted sum of moment conditions that should be closer to optimal than the ones used by the AB estimator or Estimator 2.

For inference for the AB estimator, we will consider GMM robust standard errors with clustered standard errors with and without the finite sample correction proposed by Windmeijer (2005). For inference for Estimators 1 and 2, we use the standard errors defined by (3.19) that only require (2.1), (2.2) and Auxiliary Assumption 1 to hold in order to be consistent.

All Monte Carlo results were obtained using 1,000 replications. We show here results for $n_g = 5$, $G = 100$, and $T = 5$ or $T = 15$, but results for more values of these parameters which lead to the same conclusions are presented at the end of the Appendix. Table 1 presents the results for point estimation in terms of bias, standard deviation and root MSE, Table 2 presents the results for inference in terms of bias in standard errors (captured by the ratio of the mean of the standard errors over the standard deviations of the estimators), coverage of the 95% confidence interval and average length of 95% confidence intervals. These two tables exhibit the main conclusions stated at the beginning of this section, i.e. that modeling optimal instruments results in estimators with very small bias compared to optimal GMM estimation methods (AB estimator), that accounting for cross-sectional dependence when modeling optimal instruments results in significant gains in efficiency in the presence of cluster dependence, and that modeling optimal instruments results in inference that is correctly sized compared to optimal GMM estimation methods, independently of whether cross-sectional dependence is present.

4.2 Results for Data Generating Processes with Spatial Dependence

In this section, we use data generating processes for which the true form of cross-sectional dependence is a spatial moving average process. As such, Auxiliary Assumption 1 is violated.

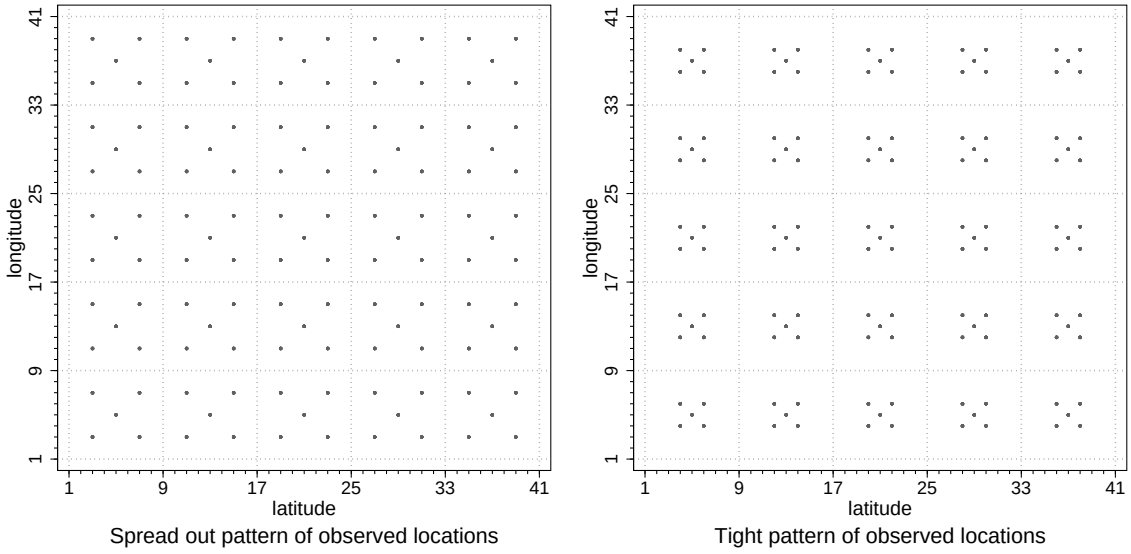
The two data generating processes we consider in this section can be described by first defining a squared lattice $D \subset \mathbb{R}^2$ with 41×41 evenly spaced locations. For any $i, j \in D$, define the distance measure $d(i,j) = max_{l=1,2}(|i_l - j_l)$ where i_l is the l^{th} element of i, and normalize the distance between neighboring locations to one. For each location $i \in D$, each time period $t = 1, ..., T$, we have:

$$\begin{aligned}
\epsilon_{it} &\overset{i.i.d.}{\sim} N(0,1) \\
u_{it} &= \sum_{d=1}^{4} \frac{1}{d} \sum_{j \in D} \epsilon_{jt} 1(d(i,j) = d) \\
\nu_i &\overset{i.i.d.}{\sim} N(0,1) \ \forall \, t = 1, ..., T \\
c_i &= \sum_{d=1}^{4} \frac{1}{d} \sum_{j \in D} \nu_i 1(d(i,j) = d) \\
\dot{\epsilon}_{i0} &\overset{i.i.d.}{\sim} N(0, \frac{1}{1-\rho_0^2}) \\
\dot{u}_{i0} &= \sum_{d=1}^{4} \frac{1}{d} \sum_{j \in D} \dot{\epsilon}_{i0} 1(d(i,j) = d) \\
y_{i0} &= \frac{c_i}{1-\rho_0} + \dot{u}_{i0} \\
y_{it} &= \rho_0 y_{it-1} + c_i + u_{it}
\end{aligned}$$

where $1(.)$ is the indicative function. In addition, we define 25 clusters, which are each squares of length eight. Finally, only 125 observations are observed. The variation in the location of these observed locations will determine how good of an approximation the assumption of cluster dependence is. Figure 4.1 shows the exact disposition of the observed locations with the boundaries of the 25 clusters in the two scenarios we consider in this section. If the observed locations are spread uniformly across the lattice, the assumption of cluster dependence will not be a very good approximation for the true form of cross-sectional dependence since the correlation of an observation at the boundary of a cluster will be stronger with observations

in the neighboring clusters than with some of the observations within its own clusters. If the observed locations are more concentrated towards the center of a cluster, the assumption of cluster dependence will be a decent approximation since the correlation between observations in the same clusters will be stronger than the correlation with observations in other clusters.

Figure 4.1: Different locations of observations



Note: Dotted lines show the boundaries of the clusters

Table 3 shows the performance of the AB estimator, Estimator 1 and Estimator 2 from 1,000 replications. The first result is that, as with cluster dependence, Estimators 1 and 2 are virtually unbiased compared to the Arellano and Bond estimator. As seen previously, this follows from Estimators 1 and 2 making use of few nuisance parameters compared to the Arellano and Bond estimator, independently of whether the form of cross-sectional dependence used to model for optimal instruments for Estimators 1 and 2 is correct or not. Secondly, we see that when the observed locations are distributed uniformly across the lattice D, i.e. when cluster dependence is not a very good approximation, then the standard deviation for Estimator 1 is not much smaller than that of Estimator 2. On the other hand, when observations are concentrated around the center of the cluster, Estimator 1 has a standard deviation that is significantly lower than Estimator 2. This corresponds to our intuition that using cluster dependence to model optimal instruments will result in efficiency gain when

cluster dependence is a good approximation of the true form of the optimal instruments.

The evidence in this section supports the use of the estimator defined in this paper when researchers suspect that the cross-sectional dependence in their data can be well approximated by cluster dependence, for instance when observations are distributed across space but tend to be more concentrated around cities. When cluster dependence is not a good approximation, for instance for farms uniformly distributed across the countryside, one should model optimal instruments using a better suited model of cross-sectional dependence, such as a model of spatial dependence. Studying the properties of an estimator similar to the one defined in this paper but using spatial dependence instead of cluster dependence to model optimal instruments is left for future work.

5 Application: Estimation of the Effect of Private School Attendance on Student Achievement

In this section we will use the data analyzed in Andrabi et al. (2011) to estimate the effect of attending private schools on student achievement in three districts of the Punjab province in Pakistan. The main covariate of interest is private school attendance. The other covariates included are wealth and variables indicating whether each parent lives with the student.

Let $y_{it}^{j\star}$ denote the achievement of student i in year t in subject $j = English, Urdu, Mathematics$, denoted E, U, M. As in Andrabi et al. (2011), we use the covariates $x_{it} = [p_{it}, w_{it}]$, where p_{it} indexes whether a student attends private school in year t, and w_{it} is a row vector containing parents' wealth and indicators for whether each parents lives under the same roof as the student. Let d_t^j denote time specific intercepts, with d_1^j normalized to zero, and u_{it}^j denote unobserved transitory shocks to achievement. Let $y_{it}^j = y_{it}^{j\star} + \epsilon_{it}^j$ denote the grade of a student in subject j, where ϵ_{it}^j is measurement error. Let $Y_{t-1} = \{y_{is}^j\}_{i=1,...,n,s=0,...,t-1,j=E,U,M}$, $P_{t-1} = \{p_{is}\}_{i=1,...,n,s=0,...,t-1}$, $W = \{w_{is}\}_{i=1,...,n,s=0,...,T}$, and $Y_t^{-j} = \{y_{is}^l\}_{i=1,...,n,s=0,...,t,l\neq j}$.

We use a value-added model similar to the one found in Andrabi et al. (2011):

$$y_{it}^j = d_t^j + x_{it}\beta_0^j + \rho^j y_{it-1}^j + c_i^j + u_{it}^j + \epsilon_{it}^j - \rho^j \epsilon_{it-1}^j \tag{5.1}$$

$$E(u_{it}^j | Y_{t-1}, P_{t-1}, W) = 0 \tag{5.2}$$

$$E(\epsilon_{it}^j | Y_t^{-j}, P_t, W) = 0 \tag{5.3}$$

This model contains four important features: 1) It includes past achievement as a covariate in order to control for the effect of past educational inputs on current achievement. 2) It accounts for unobserved heterogeneity c_i^j that can affect both achievement and private school attendance. 3) (5.2) requires that only past levels of private school attendance are exogenous with respect to the shocks to achievement, u_{it}^j. This assumption is relatively weak since it allows for current achievement to affect future school attendance or even for unobserved shocks to affect achievement and school attendance simultaneously. 4) There is measurement error in students' achievement which we assume to be classical in structure (i.e. independent across subjects and of x_{it} and $y_{it}^{j\star}$), so that past grades in other subjects can be used as an instrument for past grade in a given subject. A more thorough discussion of value-added models can be found in Andrabi et al. (2011) and of this particular model in Appendix D.1.

For estimation, one can get rid of c_t^j by forward filtering as in the previous sections. Since in the data we use, only three time periods are observed, the only equations available for estimation are, for $j = E, U, M$:

$$\Delta y_{i2}^j = d_2^j + \Delta x_{i2}\beta_0^j + \rho^j \Delta y_{i1}^j + \Delta u_{i2}^j + \Delta \epsilon_{i2}^j - \rho^j \Delta \epsilon_{i1}^j \tag{5.4}$$

$$E(\Delta u_{i2}^j + \Delta \epsilon_{i2}^j - \rho^j \Delta \epsilon_{i1}^j | P_0, W, Y_0^{-j}) = 0 \tag{5.5}$$

In order to estimate β_0^j and ρ^j from (5.4) and (5.5), one should build instruments for the covariates Δx_{i2} and Δy_{i1}^j.

Let $\phi_0^j = [d_2^j, \beta_0^{j'}, \rho^j]'$, $m_i^j(\phi) = \Delta y_{i2}^j - (d + \Delta x_{i2}\beta + \rho \Delta y_{i1}^j)$ and $m_i^j = m_i^j(\phi_0^j)$. The Arellano and Bond estimator for this model is defined by:

$$\hat{\phi}_{AB}^j = argmin_{\phi^j} (\sum_{i=1}^n Z_i^{jAB} m_i^j(\phi^j))' (\sum_{i=1}^n Z_i^{jAB} m_i^j(\tilde{\phi}^j) m_i^j(\tilde{\phi}^j)' Z_i^{jAB'})^{-1} (\sum_{i=1}^n Z_i^{jAB} m_i^j(\phi^j)) \tag{5.6}$$

where $\tilde{\phi}^j$ is a preliminary estimator of ϕ_0^j and $Z_i^{jAB} = [1, y_{i0}^{-j'}, x_{i0}, w_i']'$.

This estimator is inefficient because it ignores cross-sectional dependence[14]. Indeed in this application it is likely that transitory shocks are correlated within schools since there are school or class-level unobserved shocks, such as changes in infrastructure, staff or teachers, that will affect all students within a school or class. The data-set we use collected data from between 0 and 25 students per school in each year with most schools being represented by less than 10 students, which is too small to estimate time-varying school fixed effects accurately. Instead, we prefer treating u_{it} as cross-sectionally correlated within schools.

It is also likely that unobserved heterogeneity is correlated across students within schools since students might attend specific schools based on unobserved characteristics, such as residential location, socio-economic characteristics or past achievements, that relate to their performance.

As discussed in the previous sections, one can obtain stronger instruments, and hence a more precise estimator, by using not only Z_i^{jAB} as an instrument for Δx_{i2} and Δy_{i1}^j, but also Z_l^{jAB}, $l \in g_{i0}$.[15]

Appendix D.2 generalizes the auxiliary assumptions presented in Section 3.2 to the model defined by (5.1), (5.2), and (5.3), and derives a model and an estimator for optimal instruments. Denote by $\hat{Z}_i^{j,opt}$ the estimated optimal instruments from Appendix D.2. Our proposed estimator takes the form:

$$\sum_{i=1}^{n} \hat{Z}_i^{j,opt} m_i^j(\hat{\phi}_{opt}^j) = 0 \tag{5.9}$$

[14]Without measurement error, it would also be possible to use correlation of transitory shocks across outcomes to obtain an efficient joint estimator of $\{\phi^j\}_{j=U,E,M}$. However because of measurement error, the sets of instruments across subjects are non-overlapping, so that optimal instruments cannot be derived. Since there is no restriction in the parameters across equations, weighting of optimally weighted moment conditions or minimum distance methods cannot be used either.

[15]In this application, clusters (school membership) are not constant over time and, as pointed out previously, only past school attendance is exogenous. Therefore it is possible that:

$$E(u_{it}|g_{it}, X_{t-1}, Y_{t-1}, W) \neq 0 \tag{5.7}$$

even though:

$$E(u_{it}|g_{it-1}, X_{t-1}, Y_{t-1}, W) = 0 \tag{5.8}$$

Hence we can use as instruments lagged values of achievements of students from schools where an observation was previously enrolled but not from schools where it is currently enrolled.

Let $M_i^j = [\Delta y_1^j, \Delta p_{i2}, \Delta w_{i2}]$. Both the Arellano and Bond estimator and our optimal estimator can be written as:

$$\hat{\phi}^j = (\sum_{i=1}^{n} Z_i^j M_i^j)^{-1} \sum_{i=1}^{n} Z_i^j \Delta y_{i2}^j \tag{5.10}$$

where for the Arellano and Bond estimator, $Z_i^j = (\sum_{l=1}^{n} M_l^{j'} Z_l^{jAB'})\hat{\Theta}^{jAB-1} Z_i^{jAB}$ with $\hat{\Theta}^{jAB} = \sum_{i=1}^{n} Z_i^{jAB} m_i^j(\tilde{\phi}^j) m_i^j(\tilde{\phi}^j)' Z_i^{jAB'}$. For our optimal estimator, Z_i^j is simply $\hat{Z}_{opt}^j$.

Under the assumption that transitory shocks are independent across schools, $\hat{\phi}^j$ is consistent for ϕ_0^j and asymptotically normal. The asymptotic variance-covariance matrix of both estimators can be written as[16]:

$$AVar(\hat{\phi}) = A' B A \tag{5.11}$$

$$A = lim_{n\to\infty}(\frac{1}{n}\sum_{i=1}^{n} E(Z_i^j M_i^j))^{-1} \tag{5.12}$$

$$B = lim_{n\to\infty}(\frac{1}{n} Var(\sum_{i=1}^{n} Z_i^j m_i^j)) \tag{5.13}$$

$$= lim_{n\to\infty}(\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{n} 1[\{g_{it} = g_{ls}\}_{t,s=1,2}] E(Z_i^j m_i^j m_l^{j'} Z_l^{j'})) \tag{5.14}$$

which can be estimated consistently since there is a small number of observations in each school.

The students' achievement in each subject was measured by the results obtained by students on a test administered by the authors of Andrabi et al. (2011) and graded using the Item Response Theory so that scores can be compared across students and years. Table 4 shows the average and standard deviations of scores by subject, grade, and type of school attended. Table 5 reports the estimated degree of persistence and the estimated effect of attending private schools on performance for the three subjects considered. We also show the associated standard errors and 95% confidence intervals. Similarly as in Andrabi et al. (2011), we find that attending private school has a significant positive effect on student achievement in all subjects but Mathematics. The optimal estimator we defined in this section yields significantly smaller

[16]Note that clustering standard errors by the first school attended, which is used in Andrabi et al. (2011), is not justified since transitory shocks should be correlated within a school that a child is currently attending and not only across students who attended the same school in the first time period.

standard errors compared to the Arellano and Bond estimator both for estimating persistence in student achievements and for estimating the effect of attending private school, which corresponds to the idea developed in the rest of the paper that cross-sectional dependence can be used to estimate dynamic panel data models more efficiently.

6 Conclusion

We have presented an estimation method that used cross-sectional dependence to improve the accuracy with which dynamic models of panel data are estimated while making use of few nuisance parameters and being robust to the misspecification of the form of the cross-sectional dependence. This method can be generalized to models with covariates and sequentially exogenous instruments.

Monte Carlo simulations and an application to the estimation of a value-added model show that, when there is cross-sectional dependence, this method yields significant improvements in terms of efficiency and inference.

Extensions of this work that are the subject of ongoing research consider the generalization of the results in this paper to non-linear panel data models and the use of other forms of cross-sectional dependence than clustering to model optimal instruments.

References

Ahn, S. C. and Schmidt, P. (1995). Efficient estimation of models for dynamic panel data. *Journal of Econometrics*, 68(1):5–27.

Alvarez, J. and Arellano, M. (2003). The time series and cross-section asymptotics of dynamic panel data estimators. *Econometrica*, 71(4):1121–1159.

Alvarez, J. and Arellano, M. (2004). Robust likelihood estimation of dynamic panel data models. *CEMFI Working Paper 0421*.

Anderson, T. W. and Hsiao, C. (1981). Estimation of dynamic models with error components. *Journal of the American Statistical Association*, 76(375):598–606.

Andrabi, T., Das, J., Ijaz Khwaja, A., and Zajonc, T. (2011). Do value-added estimates add value? accounting for learning dynamics. *American Economic Journal. Applied Economics*, 3(3):29–54.

Arellano, M. (2003). Modelling optimal instrumental variables for dynamic panel data models. *CEMFI Working Paper*.

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The Review of Economic Studies*, 58(2):277–297.

Arellano, M. and Bover, O. (1995). Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics*, 68(1):29–51.

Balasubramanian, N. and Sivadasan, J. (2010). What happens when firms patent? new evidence from u.s. economic census data. *Review of Economics and Statistics*, 93(1):126–146.

Baltagi, B. H., Fingleton, B., and Pirotte, A. (2014). Estimating and forecasting with a dynamic spatial panel data model. *Oxford Bulletin of Economics and Statistics*, 76(1):112–138.

Bester, A. C., Conley, T. G., and Hansen, C. B. (2011a). Inference with dependent data using cluster covariance estimators. *Journal of Econometrics*, 165(2):137–151.

Bester, A. C., Conley, T. G., Hansen, C. B., and Vogelsang, T. J. (2011b). Fixed-b asymptotics for spatially dependent robust nonparametric covariance matrix estimators. *Working Paper*.

Blundell, R. and Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, 87(1):115–143.

Chamberlain, G. (1992). Comment: Sequential moment restrictions in panel data. *Journal of Business & Economic Statistics*, 10(1):20–26.

Cizek, P., Jacobs, J. P., Ligthart, J. E., and Vrijburg, H. (2011). GMM estimation of fixed effects dynamic panel data models with spatial lag and spatial errors. Discussion Paper 2011-134, Tilburg University, Center for Economic Research.

Conley, T. G. (1999). GMM estimation with cross sectional dependence. *Journal of Econometrics*, 92(1):1–45.

de Brauw, A. and Giles, J. (2008). Migrant labor markets and the welfare of rural households in the developing world: Evidence from china. 2008 Annual Meeting, July 27-29, 2008, Orlando, Florida 6085, American Agricultural Economics Association.

Donald, S. G., Imbens, G. W., and Newey, W. K. (2009). Choosing instrumental variables in conditional moment restriction models. *Journal of Econometrics*, 152(1):28–36.

Elhorst, P. J. (2005). Unconditional maximum likelihood estimation of linear and log-linear dynamic models for spatial panels. *Geographical Analysis*, 37(1):85–106.

Hahn, J. (1997). Efficient estimation of panel data models with sequential moment restrictions. *Journal of Econometrics*, 79(1):1–21.

Hahn, J. and Kuersteiner, G. (2002). Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and t are large. *Econometrica*, 70(4):1639 to 1657.

Han, C., Phillips, P. C. B., and Sul, D. (2014). X-differencing and dynamic panel model estimation. *Econometric Theory*, 30(01):201–251.

Hayakawa, K. (2009). A simple efficient instrumental variable estimator for panel AR(p) models when both n and t are large. *Econometric Theory*, 25(03):873–890.

Hsiao, C., Pesaran, H. M., and Tahmiscioglu, K. A. (2002). Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods. *Journal of Econometrics*, 109(1):107–150.

Jenish, N. and Prucha, I. R. (2009). Central limit theorems and uniform laws of large numbers for arrays of random fields. *Journal of Econometrics*, 150(1):86–98.

Jenish, N. and Prucha, I. R. (2012). On spatial processes and asymptotic inference under near-epoch dependence. *Journal of Econometrics*, 170(1):178–190.

Kelejian, H. H. and Prucha, I. R. (2007). HAC estimation in a spatial framework. *Journal of Econometrics*, 140(1):131–154.

Kim, M. S. and Sun, Y. (2011). Spatial heteroskedasticity and autocorrelation consistent estimation of covariance matrix. *Journal of Econometrics*, 160(2):349–371.

Kim, M. S. and Sun, Y. (2013). Heteroskedasticity and spatiotemporal dependence robust inference for linear panel models with fixed effects. *Journal of Econometrics*, 177(1):85–108.

Mutl, J. (2006). Dynamic panel data models with spatially correlated disturbances. *University of Maryland Theses and Dissertations*.

Olea, J. L. M. and Pflueger, C. (2013). A robust test for weak instruments. *Journal of Business & Economic Statistics*, 31(3):358–369.

Su, L. and Yang, Z. (2013). QML estimation of dynamic panel data models with spatial errors. *Research Collection School of Economics (Open Access)*.

Topalova, P. and Khandelwal, A. (2010). Trade liberalization and firm productivity: The case of india. *Review of Economics and Statistics*, 93(3):995–1009.

White, H. (2001). *Asymptotic theory for econometricians*. Academic Press, San Diego.

Windmeijer, F. (2005). A finite sample correction for the variance of linear efficient two-step GMM estimators. *Journal of Econometrics*, 126(1):25–51.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, second edition edition.

Table 1: Bias and RMSE, Cluster Dependence, $\rho = .8$, $n_g = 5$, $G = 100$

		Unfeasible Optimal Estimator	Arellano and Bond Estimator	Estimator 1	Estimator 2
		equicorrelation within clusters			
T=5	bias	0.012	−0.106	0.003	0.000
	sd	0.125	0.180	0.120	0.186
	rmse	0.125	0.208	0.120	0.185
T=15	bias	−0.001	−0.034	0.000	0.001
	sd	0.024	0.036	0.023	0.033
	rmse	0.024	0.049	0.023	0.034
		no correlation within clusters			
T=5	bias	0.012	−0.047	0.002	0.005
	sd	0.125	0.120	0.118	0.121
	rmse	0.125	0.129	0.118	0.121
T=15	bias	−0.001	−0.019	0.000	−0.000
	sd	0.024	0.026	0.023	0.024
	rmse	0.024	0.032	0.023	0.024
		heteroscedasticity and correlation within clusters			
T=5	bias	0.020	−0.150	0.002	−0.001
	sd	0.191	0.198	0.243	0.318
	rmse	0.192	0.249	0.243	0.318
T=15	bias	−0.003	−0.050	0.000	−0.000
	sd	0.035	0.044	0.043	0.046
	rmse	0.036	0.066	0.042	0.046

Table 2: Inference, Cluster Dependence, $\rho = .8$, $n_g = 5$, $G = 100$

		Unfeasible Optimal Estimator	Arellano and Bond Estimator	AB w/ Windmeijer correction	Estimator 1	Estimator 2
		equicorrelation within clusters				
T=5	ratio	1.001	0.850	0.973	0.991	0.935
	coverage	0.962	0.862	0.924	0.969	0.964
	length	0.495	0.606	0.708	0.472	0.689
T=15	ratio	0.997	0.611	1.007	1.005	0.986
	coverage	0.950	0.621	0.852	0.948	0.946
	length	0.094	0.087	0.130	0.093	0.131
		no correlation within clusters				
T=5	ratio	1.001	0.954	1.016	0.998	0.994
	coverage	0.962	0.928	0.951	0.960	0.966
	length	0.495	0.454	0.501	0.469	0.478
T=15	ratio	0.997	0.775	1.027	1.004	1.002
	coverage	0.950	0.783	0.902	0.946	0.951
	length	0.094	0.079	0.095	0.093	0.094
		heteroscedasticity and correlation within clusters				
T=5	ratio	0.937	0.781	1.002	0.983	1.015
	coverage	0.962	0.764	0.889	0.950	0.955
	length	0.712	0.615	0.790	0.948	1.282
T=15	ratio	0.975	0.495	0.963	0.923	0.962
	coverage	0.949	0.439	0.794	0.926	0.945
	length	0.137	0.086	0.174	0.156	0.177

Table 3: Bias and RMSE, Spatial Dependence, $\rho = .8$

	Spread out pattern of observed locations			Tight pattern of observed locations		
	Arellano and Bond Estimator	Estimator 1	Estimator 2	Arellano and Bond Estimator	Estimator 1	Estimator 2
bias	−0.133	0.007	0.006	−0.188	0.009	0.008
sd	0.102	0.088	0.100	0.123	0.087	0.130
rmse	0.167	0.088	0.100	0.224	0.088	0.130

Table 4: Averages and standard deviations of scores per subject, per grade and per school type

		English		Urdu		Math	
		Average	s.d.	Average	s.d.	Average	s.d.
Public School	Grade 3	-0.24	0.95	-0.14	0.98	-0.10	1.02
	Grade 4	0.11	0.99	0.21	1.05	0.18	1.09
	Grade 5	0.57	0.85	0.78	0.89	0.77	1.03
Private School	Grade 3	0.74	0.62	0.53	0.78	0.41	0.80
	Grade 4	0.94	0.60	0.89	0.79	0.78	0.81
	Grade 5	1.33	0.55	1.38	0.72	0.36	0.76

Table 5: Effects of Attending Private Schools on Student Achievement

	Optimal Estimator			Arellano and Bond Estimator		
	English	Urdu	Math	English	Urdu	Math
Persistence	0.14	0.16	-0.08	0.26	0.36	0.18
	(0.11)	(0.09)	(0.08)	(0.14)	(0.12)	(0.12)
	[-0.07,0.35]	[-0.01,0.33]	[-0.24,0.09]	[-0.01,0.53]	[0.12,0.59]	[-0.06,0.42]
Private School	0.64	0.48	0.26	0.40	0.81	0.26
	(0.23)	(0.24)	(0.23)	(0.43)	(0.45)	(0.45)
	[0.19,1.09]	[0.01,0.95]	[-0.19,0.71]	[-0.43,1.24]	[-0.06,1.68]	[-0.63,1.15]

Numbers in parenthesis are standard errors and intervals are 95% confidence intervals.