Journal of Econometrics 174 (2013) 107–126

Contents lists available at SciVerse ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Limit theory for panel data models with cross sectional dependence and sequential exogeneity

Guido M. Kuersteiner[a,1], Ingmar R. Prucha[b,*]

[a] Department of Economics, Georgetown University, Washington, DC 20057, United States

[b] Department of Economics, University of Maryland, College Park, MD 20742, United States

ARTICLE INFO

Article history:
Received 5 October 2010
Received in revised form
5 February 2013
Accepted 12 February 2013
Available online 24 February 2013

Keywords:
Cross-sectional dependence
Spatial martingale difference sequence
Central limit theorem
Spatial
Panel
GMM
MLE
Multinomial choice
Social interaction

ABSTRACT

The paper derives a general Central Limit Theorem (CLT) and asymptotic distributions for sample moments related to panel data models with large *n*. The results allow for the data to be cross sectionally dependent, while at the same time allowing the regressors to be only sequentially rather than strictly exogenous. The setup is sufficiently general to accommodate situations where cross sectional dependence stems from spatial interactions and/or from the presence of common factors. The latter leads to the need for random norming. The limit theorem for sample moments is derived by showing that the moment conditions can be recast such that a martingale difference array central limit theorem can be applied. We prove such a central limit theorem by first extending results for stable convergence in Hall and Heyde (1980) to non-nested martingale arrays relevant for our applications. We illustrate our result by establishing a generalized estimation theory for GMM estimators of a fixed effect panel model without imposing i.i.d. or strict exogeneity conditions. We also discuss a class of Maximum Likelihood (ML) estimators that can be analyzed using our CLT.

1. Introduction

In this paper we develop a central limit theory for data sets with cross-sectional dependence. Importantly, the theory is sufficiently general to cover panel data sets, allowing the data to be cross sectionally dependent, while at the same time allowing for regressors that are only sequentially (rather than strictly) exogenous. The paper considers cases where the time series dimension T is fixed. Our results also cover purely cross-sectional data sets.

At the center of our results lies a cross-sectional conditional moment restriction that avoids the assumption of cross-sectional independence. The paper proves a central limit theorem for the corresponding sample moment vector by extending results of Hall and Heyde (1980) for stable convergence of martingale difference arrays to a situation of non-nested information sets arising in cross-sections and panel data sets. We then show that by judiciously constructing information sets in a way that preserves a martingale structure for the moment vector in the cross-section our martingale array central limit theorem is applicable to cross-sectionally dependent panel and spatial models.

The classical literature on dynamic panel data has generally assumed that the observations, including observations on the exogenous variables, which were predominantly treated as sequentially exogenous, are cross sectionally independent. The assumption of cross sectional independence will be satisfied in many settings where the cross sectional units correspond to individuals, firms, etc., and decisions are not interdependent or when dependent units are sampled at random as discussed in Andrews (2005). However in many other settings the assumption of cross-sectional independence may be violated. Examples where it seems appropriate to allow for cross sectional dependence in the exogenous variables may be situations where regressors constitute weighted averages of data that include neighboring units (as is common in spatial analysis or in social interaction models), situations where the cross sectional units refer to counties, states, countries or industries, and situations where random sampling from the population is not feasible.

A popular approach to model cross sectional dependence is through common factors; see, e.g., Phillips and Sul (2007), Bai and Ng (2006a,b), Pesaran (2006), and Andrews (2005) for recent contributions. This represents an important class of models, however they are not geared towards modeling cross sectional interactions.[2] Our approach allows for factor structures in addition to

* Corresponding author. Tel.: +1 301 405 3499.
E-mail addresses: gk232@georgetown.edu (G.M. Kuersteiner), prucha@econ.umd.edu (I.R. Prucha).

1 Tel.: +1 202 687 0956.

2 Bai and Ng (2006a,b) allow for cross sectional correlation in the idiosyncratic disturbances, but assume that the disturbance process is independent of the factors and loadings. The setups considered in the other papers imply that the observations are independent in the cross sectional dimension conditional on the common factors.

0304-4076/$ – see front matter
http://dx.doi.org/10.1016/j.jeconom.2013.02.004

general, unmodeled (through covariates) cross-sectional dependence of the observed sample. Using the GMM estimator for a linear panel model as an example, we illustrate that conventional inference methods remain valid under the conditions of our central limit theory when samples are not i.i.d. in the cross-section. These results extend findings in Andrews (2005) to situations where samples are not i.i.d. even after conditioning on a common factor. Given that our assumptions allow for factor structures, our limit theory involves and accommodates random norming. Technically this is achieved by establishing stable convergence and not just convergence in distribution for the underlying vector of sample moments. We prove a martingale central limit theorem for stable convergence by extending results of Hall and Heyde (1980) to allow for non-nested σ-fields that naturally arise in our setting.

Another popular approach to model cross sectional dependence is to allow for spatial interactions in terms of spatial lags as is done in Cliff and Ord (1981) type models. Dynamic panel data models with spatial interactions have recently been considered by, e.g., Mutl (2006), and Yu et al. (2008, 2012). All of those papers assume that the exogenous variables are fixed constants and thus maintain strict exogeneity. The methodology developed in this paper should be helpful in developing estimation theory for Cliff–Ord type spatial dynamic panel data models with sequentially exogenous regressors.

While some of the classical literature on dynamic panel data models allowed for cross sectional correlation in the exogenous variables, this was, to the best of our knowledge, always combined with the assumption that the exogenous variables are strictly exogenous. This may stem from the fact that strict exogeneity conveniently allows the use of limit theorems conditional on all of the exogenous variables. There are many important cases where the strict exogeneity assumption does not hold, and regressors, apart from time-lagged endogenous variables, or other potential instruments are only sequentially exogenous. Examples given by Keane and Runkle (1992) include rational expectations models or models with predetermined choice variables as regressors. Other examples are the effects of children on the labor force participation of women considered by Arellano and Honoré (2001, p. 3237) or the relationship between patents and R&D expenditure studied by Hausman et al. (1984); see, e.g., Wooldridge (1997) for further commentary on strict vs. sequential exogeneity.

Motivated by the above, the main aim of our paper is to develop a general central limit theory for sample moments of a panel data set, where we allow for cross sectional dependence in the explanatory variables and disturbances (and thus in the dependent variable), while allowing for some of the explanatory variables to be sequentially exogenous. The setup will be sufficiently general to accommodate cross sectional dependence due to common factors and/or spatial interactions, both of which can affect the covariates. Our results are different from central limit theorems for spatial process such as Bolthausen (1982) and Jenish and Prucha (2009, 2012) because we do not impose a spatial structure on the cross-sectional dimension of the panel. As a result the high level conditions that need to be checked to apply our CLT are relatively simple compared to the spatial CLTs. On the other hand, the conditional moment restrictions we impose are often synonymous with correct specification of an underlying model which may not be required by CLTs for mixing processes as in Bolthausen (1982).

The paper is organized as follows. In Section 2 we formulate the moment conditions, and give our basic result concerning the limiting distribution of the normalized sample moments. The analysis establishes not only convergence in distribution but stable convergence. In Section 3 we illustrate how the central limit theory can be applied to efficient GMM estimators for linear panel models. We derive their limiting distribution, and give a consistent estimator for the limiting variance covariance matrix. In Section 4 we present regularity conditions for a class of maximum likelihood estimators (MLE) and show how our CLT can be applied. We give examples of specific multinomial choice models that fit our framework. Concluding remarks are given in Section 5. Basic results regarding stable convergence as well as all proofs are relegated to the appendices.

2. Central limit theory

2.1. Moment conditions

In the following we develop a central limit theory (CLT) for a vector of sample moments for panel data where n and T denote the cross section and time dimension, respectively. For the CLT developed in this section we assume that sample averages are taken over n, with n tending to infinity and T fixed. We allow for purely cross-sectional data sets by allowing for $T = 1$ in the CLT. However, this condition may need to be strengthened to $T > T_0$ for some $T_0 > 1$ for specific models and data transformations.

Our basic central limit theorem is stated for averages

$$\psi_{(n)} = n^{-1/2} \sum_{i=1}^{n} \psi_i, \tag{1}$$

over the cross-section of $p \times 1$ random vectors $\psi_i = (\psi_{i1}', \ldots, \psi_{iT}')'$.[3] The dimension of the sub-vectors ψ_{it} is $p_t \times 1$ and thus allowed to depend on t. The index i is an identifier for a particular unit, where units could be individuals, firms, industries, counties, etc. While units may refer to geographic entities, no spatial structure is explicitly imposed on ψ_i. On the other hand, the index t is given the conventional notion of sequential time.

In introducing our basic CLT the aim is to provide a convenient module that can be readily used to establish, in particular, a CLT for the sample moment vector associated with GMM estimators and the score of the log-likelihood function of ML estimators. For GMM estimators ψ_{it} will typically refer to the, say, p_t sample moments between a vector of instruments and some basic disturbances for unit i in period t. For ML estimation ψ_{it} will typically refer to the score of the log likelihood function corresponding to unit i and period t, with $p_t = d$, the dimension of the parameter vector of interest. In the following we set $p = \sum_{t=1}^{T} p_t$.

We next give some basic notational definitions used throughout the paper. All variables are assumed to be defined on a probability space $(\Omega, \mathcal{F}, P)$. With $y_{it}, x_{it}, z_{it}, \mu_i$ and u_{it} we denote, respectively, the dependent variable, the sequentially exogenous covariates, the strictly exogenous covariates, unit specific unobserved effects and idiosyncratic disturbances. The particular meaning of sequential and strict exogeneity will be made explicit below. Furthermore, it proves helpful to introduce the following notation: $y_i = (y_{i1}, \ldots, y_{iT}), x_i = (x_{i1} \ldots, x_{iT}), z_i = (z_{i1}, \ldots, z_{iT}), u_i = (u_{i1}, \ldots, u_{iT}), y_{it}^o = (y_{i1}, \ldots, y_{it}), x_{it}^o = (x_{i1}, \ldots, x_{it}), u_{it}^o = (u_{i1}, \ldots, u_{it})$, and $u_{-i,t} = (u_{1t}, \ldots, u_{i-1,t}, u_{i+1,t}, \ldots, u_{nt})$. Although not explicitly denoted, these random variables as well as the ψ_{it} are allowed to depend on the sample size n, i.e., to form triangular arrays.

Our setup is aimed at accommodating fairly general forms of cross-sectional dependence in the data. In particular, analogous to Andrews (2005), who considers static models, we allow in each period t for the possibility of regressors and disturbances (and thus for the dependent variable) to be affected by common shocks that are captured by a sigma field $\mathcal{C}_t \subset \mathcal{F}$. A special case arises when f_t denotes a vector of common shocks such that $\mathcal{C}_t = \sigma(f_t)$.

[3] With stronger assumptions than we impose in this paper it may be possible to prove a multivariate CLT for $\psi_{(n)}$ based on the martingale structure of ψ_i only, i.e. without regard to the time series nature of ψ_{it}. An example is the case when the random vectors ψ_i are exchangeable. Without such additional assumptions a detailed treatment of the time series structure of ψ_{it} is needed.