Journal of Asian Economics 28 (2013) 3–11

Contents lists available at SciVerse ScienceDirect

Journal of Asian Economics

On the implementation and use of factor-augmented regressions in panel data☆

Joakim Westerlund [a,*], Jean-Pierre Urbain [b]

[a] Deakin University, Australia
[b] Maastricht University, The Netherlands

ARTICLE INFO

Article history:
Received 10 December 2012
Received in revised form 13 January 2013
Accepted 7 February 2013
Available online 5 March 2013

JEL classification:
C12
C13
C33

Keywords:
Factor-augmented panel regressions
Principal components
Cross-sectional averages
China
Predictive regression

ABSTRACT

Practitioners are generally well aware of the fact that most standard approaches for estimation and inference in panel data regressions are based on assuming that the cross-sectional units are independent of each other, an assumption that is surely mistaken in applications, especially in macroeconomics and finance. Yet, applications involving anything but these standard approaches are very rare. The current paper can be seen as a reaction to this. The purpose is to point to some of the recent advances in the area of factor-augmented panel regressions, and to also provide some guidance regarding their implementation.

1. Introduction

Recently, there has been increased interest in analysis of panel data models in which the standard assumption that the regression errors are cross-sectionally uncorrelated is violated. When the regression errors are cross-sectionally correlated standard estimation methods do not necessarily produce consistent estimates of the coefficients of interest, and much effort has therefore gone into the development of robust methods. In particular, the use of factor-augmented regressions has recently become very popular.

The key assumption in factor-augmented regressions is that the cross-section dependence can be represented by means of a small number of common factors, which can then be included as additional regressors. For factors that are observed, such as interest rates or oil prices, this is of course very easy. However, in most of the applications the factors are unobserved and they lack good proxies. The most common approach to deal with the presence of such latent factors is to use estimates in

☆ A version of the paper was presented at the 27th International Conference of the American Committee for Asian Economic Studies (ACAES) hosted by Deakin University in Melbourne. The authors would like to thank conference participants, and in particular Paresh Narayan (the guest editor) for many valuable comments and suggestions. The second author would like to thank the Financial Econometrics Group of the School of Accounting, Economics and Finance at Deakin University for its hospitality during a visit in October 2012.

* Corresponding author at: Deakin University, Faculty of Business and Law, School of Accounting, Economics and Finance, Melbourne Burwood Campus, 221 Burwood Highway, VIC 3125, Australia. Tel.: +61 3 924 46973; fax: +61 3 924 46283.
E-mail address: j.westerlund@deakin.edu.au (J. Westerlund).

1049-0078/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.asieco.2013.02.002

their stead. In this paper we focus on two of the most popular approaches to factor-augmented regressions, namely, the principal components (PC) approach (see, for example, Bai, 2009; Greenaway-McGrevy, Han, & Sul, 2012), and the cross-sectional average (CA) approach (see Pesaran, 2006). Note that while the discussion presented in this paper will focus on (stationary) factor-augmented panel regressions, the PC and CA approaches could be useful also for other situations where unobserved factors are present, such as factor-augmented vector autoregressive (FAVAR) models (see, for example, Bernanke, Boivin, & Eliasz, 2005).

PC and CA were recently considered in details by Westerlund and Urbain (2012), who argue that while very popular in applied work, the relative merits of these approaches are not well understood. In fact, some practitioners seem to use them quite interchangeably, as though their properties were the same, and the little formal evidence that does exists is based exclusively on Monte Carlo simulation (see Chudik, Pesaran, & Tosetti, 2011; Kapetanios & Pesaran, 2005), which need not necessarily be informative of any real theoretical differences.

As a response to this, Westerlund and Urbain (2012) conduct a theoretical investigation of relative performance of the two estimation approaches. Their main finding is that if the number of time series dimension, T, is similar in magnitude to the cross-sectional dimension, N, then the properties of the two estimators can be quite different. Specifically, while both are (asymptotically) normal, there is a "second-order" bias effect that depends critically on the choice of estimator of the common factors (PC or CA). Thus, while still consistent, due to the miscentering of the asymptotic distribution, inference based on PC and CA will generally be affected, and in some cases even rendered invalid. The main implication for applied work is that any differences in conclusions obtained when applying both approaches cannot be taken as a lack of robustness, as PC and CA are expected to lead to different results in most cases. It also means that the choice of estimator is not innocuous.

As a natural by-product of the finding of bias, Westerlund and Urbain (2012) also mention the possibility of using bias-correction, although most of the details are left out of their paper. In the current paper we take this as our starting point. The purpose is to show how the bias-corrected PC and CA estimators can be implemented in practice. We believe this exercise to be important for at least two reasons. First, since the difference between the PC and CA approaches lies with the bias, the bias-corrected estimators should be asymptotically equivalent. Hence, while the original (uncorrected) approaches cannot be used interchangeably, the bias-corrected ones can, at least asymptotically. Second, by considering the case where N and T are similar in magnitude (formally equal), bias-correction enable inference in cases previously not possible. This last motivation is particularly relevant for practical work, especially in macroeconomics and finance, where N and T are typically of similar size, suggesting that bias-correction is appropriate. It also means that the bulk of existing empirical evidence based on these estimators needs to be reevaluated, as the possibility remains that the conclusions are biased.

The structure of the paper is as follows. In Section 2, we review the model considered in Westerlund and Urbain (2012), and show how it can be seen as a generalization of many of the more commonly used panel data models in the literature. Section 3 is devoted to a formal presentation of the PC and CA estimation approaches and their asymptotic properties. Sections 4–6 pertain to the implementation, and consider in turn the issues of bias-correction, selection of the number of common factors, and an empirical application to stock return predictability in China, respectively. Section 6 concludes.

2. The model

Consider the scalar and m-dimensional vector of observable panel data variables $y_{i,t}$ and $x_{i,t}$, where $i = 1, \ldots, N$ and $t = 1, \ldots, T$ index the cross-sectional and time series dimensions, respectively. The generating process (DGP) considered here is very similar to the ones considered by Bai (2009), Greenaway-McGrevy et al. (2012) and Pesaran (2006), and is given by

$$y_{i,t} = \beta' x_{i,t} + e_{i,t}, \tag{1}$$

$$e_{i,t} = \lambda_i' F_t + \epsilon_{i,t}, \tag{2}$$

where β is a m-dimensional vector of slope coefficients, $F_t = (F_{1,t}, \ldots, F_{r,t})'$ is a r-dimensional vector of unobserved common factors with $\lambda_i = (\lambda_{1,i}, \ldots, \lambda_{r,i})'$ being the associated r-vector of (non-random) factor loadings, such that $\lambda_i' F_t = \sum_{k=1}^{r} \lambda_{k,i} F_{k,t}$, and $\epsilon_{i,t}$ is a idiosyncratic error term that is assumed to be independent of F_t and $x_{i,t}$. Although $\epsilon_{i,t}$ can in principle be cross-section correlated to some extent (see Bai, 2009; Greenaway-McGrevy et al., 2012), in this paper we assume it to be independent with mean zero and variance $\sigma^2_{\epsilon,i}$. If $\epsilon_{i,t}$ is serially uncorrelated, then $\sigma^2_{\epsilon,i} = E(\epsilon^2_{i,t})$ is simply the contemporaneous variance, whereas if $\epsilon_{i,t}$ is serially correlated, then $\sigma^2_{\epsilon,i} = \sum_{s=-\infty}^{\infty} E(\epsilon_{i,t}\epsilon_{i,t-s})$ is the so-called "long-run" variance.

The above model can be interpreted in two ways. On way is to make (1) conditional on F_t, giving rise to the following factor-augmented regression model with idiosyncratic errors:

$$y_{i,t} = \beta' x_{i,t} + \lambda_i' F_t + \epsilon_{i,t}. \tag{3}$$

As pointed out by Bai (2009), in light of the interaction between λ_i and F_t, (3) can be interpreted as an "interactive effects" model that includes many of the more conventional panel data models as special cases. For example, the usual unit-specific fixed effects model takes the form

$$y_{i,t} = \beta' x_{i,t} + \alpha_i + \epsilon_{i,t}. \tag{4}$$