Large Dimensional Factor Analysis

Large Dimensional Factor Analysis

Jushan Bai

Department of Economics
New York University
New York, NY 10012
USA
Jushan.Bai@nyu.edu

Serena Ng

Economics Department
Columbia University
New York, NY 10027
USA
Serena.Ng@columbia.edu

the essence of knowledge

Boston – Delft

Foundations and Trends® in Econometrics

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is J. Bai and S. Ng, Large Dimensional Factor Analysis, Foundations and Trends® in Econometrics, vol 3, no 2, pp 89–168, 2008

ISBN: 978-1-60198-144-8

Foundations and Trends® in Econometrics

Volume 3 Issue 2, 2008

Editorial Board

Editor-in-Chief:
William H. Greene
Department of Economics
New York Univeristy
44 West Fourth Street, 7–78
New York, NY 10012
USA
wgreene@stern.nyu.edu

Editors

Full text available at: http://dx.doi.org/10.1561/0800000002

Editorial Scope

Foundations and Trends® in Econometrics will publish survey and tutorial articles in the following topics:

- Identification
- Model Choice and Specification Analysis
- Non-linear Regression Models
- Simultaneous Equation Models
- Estimation Frameworks
- Biased Estimation
- Computational Problems
- Microeconometrics
- Treatment Modeling
- Discrete Choice Modeling
- Models for Count Data
- Duration Models
- Limited Dependent Variables
- Panel Data
- Dynamic Specification
- Inference and Causality
- Continuous Time Stochastic Models
- Modeling Non-linear Time Series
- Unit Roots
- Cointegration
- Latent Variable Models
- Qualitative Response Models
- Hypothesis Testing
- Interactions-based Models
- Duration Models
- Financial Econometrics
- Measurement Error in Survey Data
- Productivity Measurement and Analysis
- Semiparametric and Nonparametric Estimation
- Bootstrap Methods
- Nonstationary Time Series
- Robust Estimation

Information for Librarians
Foundations and Trends® in Econometrics, 2008, Volume 3, 4 issues. ISSN paper version 1551-3076. ISSN online version 1551-3084. Also available as a combined paper and online subscription.

Foundations and Trends® in
Econometrics
Vol. 3, No. 2 (2008) 89–168

DOI: 10.1561/0800000002

now
the essence of knowledge

Large Dimensional Factor Analysis*

Jushan Bai[1] and Serena Ng[2]

[1] *Department of Economics, New York University, 19 West 4th St, New York, NY 10012, USA, Jushan.Bai@nyu.edu*
[2] *Economics Department, Columbia University, 1019 International Affairs Building, 420 West 118th St, New York, NY 10027, USA, Serena.Ng@columbia.edu*

Abstract

Econometric analysis of large dimensional factor models has been a heavily researched topic in recent years. This review surveys the main theoretical results that relate to static factor models or dynamic factor models that can be cast in a static framework. Among the topics covered are how to determine the number of factors, how to conduct inference when estimated factors are used in regressions, how to assess the adequacy of observed variables as proxies for latent factors, how to exploit the estimated factors to test unit root tests and common trends,

* We are grateful to an anonymous referee for careful reading and constructive comments. This survey was presented by the second author as an invited lecture at the 2007 annual meeting of the Society of Computational Economics in Montreal. Financial support from the National Science Foundation (SES055127 and SES-0549978) is gratefully acknowledged.

and how to estimate panel cointegration models. The fundamental result that justifies these analyses is that the method of asymptotic principal components consistently estimates the true factor space. We use simulations to better understand the conditions that can affect the precision of the factor estimates.

Contents

1

Introduction

An inevitable fact as we move forward in time and as information technology improves is that data will be available for many more series and over an increasingly long span. While the availability of more data provides the opportunity to understand economic phenomena and anomalies better, researchers can also suffer from an information overload without some way to organize the data into an easy to interpret manner. In recent years, the analysis of large dimensional data has received the attention of theoretical and empirical researchers alike. The early focus has primarily been on the use of factor models as a means of dimension reduction. But the volume of research, both at the empirical and theoretical levels, has grown substantially. Empirical researchers have found it useful to extract a few factors from a large number of series in many forecasting and policy exercises. Theoretical researchers have taken up the challenge to extend standard factor analysis to allow the size of both dimensions of a panel data set to increase. The theoretical implications of using estimated factors in both estimation and inference are now better understood. Factor analysis plays a role not just in forecasting. In recent years, the factor structure has been incorporated into regression analysis to deal with cross-sectionally correlated errors and endogeneity bias.

This review provides a survey of the main theoretical results for large dimensional factor models, emphasizing results that have implications for empirical work. We focus on the development of the static factor models, which are to be distinguished from dynamic factor models in ways to be made precise. Key results concerning large dynamic factor models are given in Forni et al. (2000, 2004, 2005). Results concerning the use of factors in forecasting are discussed in Stock and Watson (2006), Banerjee et al. (2006), and Giannone et al. (2007). Here, our focus will be on the use of estimated factors in subsequent estimation and inference. While we survey many of the analytical results that are of use to empirical researchers, a survey of empirical applications of large factor models will not be included. Surveys with heavier empirical focus can be found in Breitung and Eickmeier (2005) and Reichlin (2003). Suffice it to say that factor models have been used in forecasting of the conditional mean by Stock and Watson (2002b), Cristadoro et al. (2001), Artis et al. (2005), Marcellino et al. (2003), Schumacher (2005), Forni et al. (2001), den Reijer (2005), and many others. Boivin and Ng (2005) compared the use of dynamic and static factors in forecasting. Anderson and Vahid (2007) used the factor model to forecast volatility with jump components. A non-exhaustive list of policy analyses that adopt a factor approach includes Bernanke and Boivin (2003), Giannone et al. (2005a,b), Favero et al. (2005), Stock and Watson (2005), and Forni et al. (2003). Use of factors as conditioning information is discussed in the conditional risk-return analysis of Ludvigson and Ng (2007), and term structure analysis of Ludvigson and Ng (2005).

This survey, drawing heavily from our previous work, is organized to serve three purposes. First, the results are presented under a coherent and general set of assumptions. Situations that require stronger assumptions will be made clear as we go along. Second, results for stationary and non-stationary data are discussed separately, as they involve different assumptions and are used in different contexts. Third, consistent estimation of the factor space is fundamental to many of the results. We use simulations to study what are the main aspects of the data that affect the precision of the factor estimates.

References

Amengual, D. and M. Watson (2007), 'Consistent estimation of the number of dynamic factors in large N and T panel'. *Journal of Business and Economic Statistics* **25**(1), 91–96.

Anderson, H. and F. Vahid (2007), 'Forecasting the volatility of australian stock returns: Do common factors help?'. *Journal of the American Statistical Association* **25**(1), 75–90.

Anderson, T. W. (1984), *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

Anderson, T. W. and H. Rubin (1956), 'Statistical inference in factor analysis'. In: J. Neyman (ed.): *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. V. Berkeley: University of California Press, pp. 114–150.

Artis, M., A. Banerjee, and M. Marcellino (2005), 'Factor forecasts for the UK'. *Journal of Forecasting* **24**, 279–298.

Bai, J. (2003), 'Inferential theory for factor models of large dimensions'. *Econometrica* **71**(1), 135–172.

Bai, J. (2004), 'Estimating cross-section common stochastic trends in non-stationary panel data'. *Journal of Econometrics* **122**, 137–183.

Bai, J. (2005), 'Panel data models with interactive fixed effects'. Department of Economics, New York University, Unpublished Manuscript.

Bai, J. and J. L. Carrion-i-Silvestre (2004), 'Structural changes, common stochastic trends, and unit roots in panel data'. Unpublished Manuscript, revised 2007.

Bai, J. and J. L. Carrion-i-Silvestre (2005), 'Testing panel cointegration with unobservable dynamic common factors'. Department of Economics, New York University, Unpublished Manuscript.

Bai, J., C. Kao, and S. Ng (2006), 'Panel cointegration with global stochastic trends'. Department of Economics, University of Michigan, Unpublished Manuscript, http://www.columbia.edu/sn2294/papers/bkn.pdf.

Bai, J. and S. Ng (2002), 'Determining the number of factors in approximate factor models'. *Econometrica* **70**(1), 191–221.

Bai, J. and S. Ng (2004), 'A PANIC attack on unit roots and cointegration'. *Econometrica* **72**(4), 1127–1177.

Bai, J. and S. Ng (2006a), 'Confidence intervals for diffusion index forecasts and inference with factor-augmented regressions'. *Econometrica* **74**(4), 1133–1150.

Bai, J. and S. Ng (2006b), 'Evaluating latent and observed factors in macroeconomics and finance'. *Journal of Econometrics* **113**(1–2), 507–537.

Bai, J. and S. Ng (2006c), 'Instrumental variables in a data rich environment'. Department of Economics, Columbia University, Unpublished Manuscript, http://www.columbia.edu/sn2294/ papers/iv.pdf.

Bai, J. and S. Ng (2006d), 'Panel unit root tests with cross-section dependence'. Department of Economics, Columbia University, Unpublished Manuscript, http://www.columbia.edu/sn2294/papers/newpanic.pdf.

Bai, J. and S. Ng (2007), 'Determining the number of primitive shocks'. *Journal of Business and Economic Statistics* **25**(1), 52–60.

Bai, J. and S. Ng (2008), 'Extremum estimation when the predictors are estimated from large panels'. Department of Economics, Columbia University, Unpublished Manuscript, http://www.columbia.edu/ sn2294/papers/probit.pdf.

Banerjee, A., M. Marcellino, and C. Osbat (2005), 'Testing for PPP: Should we use panel methods?'. *Empirical Economics* **30**, 77–91.

Banerjee, A., I. Masten, and M. Massimiliano (2006), 'Forecasting macroeconomic variables using diffusion indexes in short samples with structural change'. *Forecasting in the Presence of Structural Breaks and Model Uncertainty*, Elsevier.

Bernanke, B. and J. Boivin (2003), 'Monetary policy in a data rich environment'. *Journal of Monetary Economics* **50**(3), 525–546.

Boivin, J. and S. Ng (2005), 'Undertanding and comparing factor based forecasts'. *International Journal of Central Banking* **1**(3), 117–152.

Boivin, J. and S. Ng (2006), 'Are more data always better for factor analysis'. *Journal of Econometrics* **132**, 169–194.

Breitung, J. and S. Das (2007), 'Testing for unit root in panels with a factor structure'. *Econometric Theory* **24**, 88–108.

Breitung, J. and S. Eickmeier (2005), 'Dynamic factor models'. Deutsche Bundesbank Discussion Paper 38/2005.

Breitung, J. and H. M. Pesaran (2007), 'Unit roots and cointegration in panels'. In: L. Matyas and P. Sevestre (eds.): *The Econometrics of Panel Data,* Kluver Academic Press.

Chamberlain, G. and M. Rothschild (1983), 'Arbitrage, factor structure and mean-variance analysis in large asset markets'. *Econometrica* **51**(5), 1281–1304.

Connor, G., M. Hagmann, and O. Linton (2007), 'Efficient estimation of a semiparametric characteristic-based factor model of security returns'. Department of Economics, London School of Economics, Unpublished Manuscript.

Connor, G. and R. Korajzcyk (1986), 'Performance measurement with the arbitrage pricing theory: A new framework for analysis'. *Journal of Financial Economics* **15**, 373–394.

Connor, G. and R. Korajzcyk (1998), 'Risk and return in an equilibrium APT application of a new test methodology'. *Journal of Financial Economics* **21**, 225–289.

Cristadoro, R., M. Forni, L. Reichlin, and V. Giovanni (2001), 'A core inflation index for the euro area'. Unpublished Manuscript, www.dynfactor.org.

den Reijer, A. H. J. (2005), 'Forecasting dutch GDP using large scale factor models'. DNB Working Papers 028, http://ideas.repec.org/p/dnb/dnbwpp/028.html.

Doz, C., D. Giannone, and L. Reichlin (2007), 'A quasi-maximum likelihood approach for large approximate dynamic factor models'. European Central Bank Working Paper Series 674, http://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp674.pdf.

Engle, R. F. and C. Granger (1987), 'Cointegration and error-correction: Representation, estimation, and testing'. *Econometrica* **55**(2), 251–276.

Favero, C., M. Marcellino, and F. Neglia (2005), 'Principal components at work: The empirical analysis of monetary policy with large datasets'. *Journal of Applied Econometrics* **20**, 603–620.

Forni, M., D. Giannone, M. Lippi, and L. Reichlin (2003), 'Opening the black box: Identifying shocks and propagation mechanisms in VAR and factor models'. Unpublished Manuscript.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000), 'The generalized dynamic factor model: Identification and estimation'. *Review of Economics and Statistics* **82**(4), 540–554.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2001), 'Do financial variables help in forecasting inflation and real activity in the euro area'. Manuscript, http://homepages.ulb.ac.be/lreichli/financial.pdf.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2004), 'The generalized factor model: Consistency and rates'. *Journal of Econometrics* **119**, 231–255.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2005), 'The generalized dynamic factor model, one sided estimation and forecasting'. *Journal of the American Statistical Association* **100**, 830–840.

Geweke, J. (1977), 'The dynamic factor analysis of economic time series'. In: D. J. Aigner and A. Goldberger (eds.): *Latent Variables in Socio Economic Models*. North Holland: Amsterdam.

Ghahramani, Z. and G. E. Hinton (1996), 'The EM algorithm for mixtures of factor analyzers'. Technical Report CRG-TR-96-1, Department of Computer Science, University of Toronto.

Giannone, D., L. Reichlin, and L. Sala (2005a), 'Monetary policy in real time'. *Macroeconomic Annual* **19**, 161–200.

Giannone, D., L. Reichlin, and L. Sala (2005b), 'VARs, common factors and the empirical validation of equilibrium business cycle models'. *Journal of Econometrics* **127**(1), 257–279.

Giannone, D., L. Reichlin, and D. Small (2007), 'Nowcasting: The real-time informational content of macroeconomic data'. Forthcoming in *Journal of Monetary Economics*.

Hallin, M. and R. Liska (2007), 'Determining the number of factors in the general dynamic factor model'. *Journal of the American Statistical Association* **102**, 603–617.

Heaton, C. and V. Solo (2006), 'Estimation of approximate factor models: Is it important to have a large number of variables'. Department of Economics, Macquarie University, Research Papers series 0605, http://www.econ.mq.edu.au/research/2006/HeatonEstimtnOfApproxFactorModels.pdf.

Jones, C. (2001), 'Extracting factors from heteroskedsatic asset returns'. *Journal of Financial Economics* **62**(2), 293–325.

Kao, C., L. Trapani, and G. Urga (2007), 'Modelling and testing for structural breaks in panels with common stochastic trends'. Cass Business School, Unpublished Manuscript.

Kapetanios, G. (2007), 'A testing procedure for determining the number of factors in approximate factor models with large datasets'. forthcoming in *Journal of Business and Economic Statistics*.

Kapetanios, G. and M. Marcellino (2006a), 'Factor-GMM estimation with large sets of possibly weak instruments'. Unpublished Manuscript.

Kapetanios, G. and M. Marcellino (2006b), 'A parametric estimation methods for dynamic factor models of large dimensions'. CEPR Working Paper, 5620.

Kim, C. and C. Nelson (2000), *State Space Models with Regime Switching*. Cambridge, MA: MIT Press.

Lawley, D. N. and A. E. Maxwell (1971), *Factor Analysis in a Statistical Method*. London: Butterworth.

Lehmann, B. N. and D. Modest (1988), 'The empirical foundations of the arbitrage pricing theory'. *Journal of Financial Economics* **21**, 213–254.

Ludvigson, S. and S. Ng (2005), 'Macro factors in bond risk premia'. NBER Working Paper 11703.

Ludvigson, S. and S. Ng (2007), 'The empirical risk return relation: A factor analysis approach'. *Journal of Financial Economics* **83**, 171–222.

Marcellino, M., J. H. Stock, and M. Watson (2003), 'Macroeconomic forecasting in the Euro area: Country specific versus Euro wide information'. *European Economic Review* **47**, 1–18.

Moon, R. and B. Perron (2004), 'Testing for a unit root in panels with dynamic factors'. *Journal of Econometrics* **122**(1), 81–126.

Moon, R., B. Perron, and P. Phillips (2007), 'Incidental trends and the power of panel unit root tests'. *Journal of Econometrics* **141**(2), 416–459.

Newey, W. K. and K. D. West (1987), 'A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix'. *Econometrica* **55**, 703–708.

Onatski, A. (2005), 'Determining the number of factors from empirical distribution of eigenvalues'. Department of Economics, Columbia University, Discussion Paper 0405-19, http://www.columbia.edu/cu/economics/discpapr/DP0405-19.pdf.

Onatski, A. (2006a), 'Asymptotic distribution of the principal components estimator of large factor models when factors are relatively weak'. Department of Economics, Columbia University, Unpublished Manuscript, http://www.columbia.edu/ ao2027/inference33.pdf.

Onatski, A. (2006b), 'A formal statistical test for the number of factors in approximate factor models'. Department of Economics, Columbia University, Unpublished Manuscript.

Otrok, C. and C. Whiteman (1998), 'Bayesian leading indicators: Measuring and predicting economic conditions in iowa'. *International Economic Review* **39**(4), 997–1014.

Pagan, A. (1984), 'Econometric issues in the analysis of regressions with generated regressors'. *International Economic Review* **25**, 221–247.

Perron, P. and S. Ng (1998), 'An autoregressive spectral density estimator at frequency zero for nonstationarity tests'. *Econometric Theory* **14**, 560–603.

Pesaran, M. H. (2006), 'Estimation and inference in large heterogeneous panels with a multifactor error structure'. *Econometrica* **74**, 967–1012.

Pesaran, M. H. (2007), 'A simple unit root test in the presence of cross-section dependence'. *Journal of Applied Economics* **22**(2), 265–312.

Phillips, P. C. B. and B. E. Hansen (1990), 'Statistical inference in instrumental variables regression with I(1) processes'. *Review of Economic Studies* **57**, 99–125.

Phillips, P. C. B. and W. Ploberger (2002), 'Optimal testing for unit roots in panel data'. Department of Economics, University of Rochester, Unpublished Manuscript.

Phillips, P. C. B. and D. Sul (2003), 'Dynamic panel estimation and homogeneity testing under cross-section dependence'. *Econometrics Journal* **6**(1), 217–259.

Quah, D. and T. Sargent (1992), 'A dynamic index model for large cross sections'. Federal Reserve Bank of Minneapolis, Discussion Paper 77.

Reichlin, L. (2003), 'Factor models in large cross sections of time series'. In: S. T. M. Dewatripoint and L. P. Hansen (eds.): *Advances in Economics and Econometrics: Theory and Applications,* Vol. 111, 8th World Congress of the Econometric Society. Cambridge: Cambridge University Press.

Ross, S. (1976), 'The arbitrage theory of capital asset pricing'. *Journal of Finance* **13**, 341–360.

Rubin, D. R. and D. T. Thayer (1982), 'EM algorithms for ML factor analysis'. *Psyhometrika* **47**, 69–76.

Sargan, J. D. and A. Bhargava (1983), 'Testing for residuals from least squares regression being generated by gaussian random walk'. *Econometrica* **51**, 153–174.

Sargent, T. and C. Sims (1977), 'Business cycle modelling without pretending to have too much a priori economic theory'. In: C. Sims (ed.): *New Methods in Business Cycle Research.* Minneapolis: Federal Reserve Bank of Minneapolis.

Schumacher, C. (2005), 'Forecasting german GDP using alternative factor models based on large datasets'. Bundesbank Discussion Paper 24-2005.

Shumway, R. and D. Stoffer (2000), *Time Series Analsyis and its Applications*. New York: Springer.

Stock, J. H. (1990), 'A class of tests for integration and cointegration'. Department of Economics, Harvard University, Unpublished Manuscript.

Stock, J. H. and M. W. Watson (1988), 'Testing for common trends'. *Journal of the American Statistical Association* **83**, 1097–1107.

Stock, J. H. and M. W. Watson (2002a), 'Forecasting using principal components from a large number of predictors'. *Journal of the American Statistical Association* **97**, 1167–1179.

Stock, J. H. and M. W. Watson (2002b), 'Macroeconomic forecasting using diffusion indexes'. *Journal of Business and Economic Statistics* **20**(2), 147–162.

Stock, J. H. and M. W. Watson (2005), 'Implications of dynamic factor models for VAR analysis'. NBER Working Paper 11467.

Stock, J. H. and M. W. Watson (2006), 'Forecasting with many predictors'. *Handbook of Economic Forecasting*. North Holland: Elsevier.

Watson, M. and R. Engle (1983), 'Alternative algorithms for the estimation of dynamic factor, MIMIC, and varying coefficient regression models'. *Journal of Econometrics* **23**, 385–400.

Westerlund, J. and D. Edgerton (2007), 'Simple tests for cointegration in dependent panels with structural breaks'. Working Paper No. 2006:13, Lund University Department of Economics, available at SSRN: http://ssrn.com/abstract=1080598.

Westerlund, J. and R. Larsson (2007), 'A note on the pooling of individual PANIC unit root tests'. Department of Economics, Lund University, Unpublished Manuscript, http://www.nek.lu.se/NEKfng/panic.pdf.