Econometric Analysis of Large Factor Models

Jushan Bai* and Peng Wang†

August 2015

Abstract

Large factor models use a few latent factors to characterize the co-movement of economic variables in a high dimensional data set. High dimensionality brings challenge as well as new insight into the advancement of econometric theory. Due to its ability to effectively summarize information in large data sets, factor models have become increasingly popular in economics and finance. The factors, being estimated from the high dimensional data, can help to improve forecast, provide efficient instruments, control for nonlinear unobserved heterogeneity, etc. This article reviews the theory on estimation and statistical inference of large factor models. It also discusses important applications and highlights future directions.

Key words: high dimensional data, factor-augmented regression, FAVAR, number of factors, interactive effects, principal components, regularization, Bayesian estimation

Prepared for the Annual Review of Economics

*Department of Economics, Columbia University, New York, NY, USA. jb3064@columbia.edu.
†Department of Economics, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong. Email: pwang@ust.hk.
## Contents

1 Introduction ........................................ 3

2 Large Factor Models ................................... 4

3 Determining the Number of Factors .................. 5
   3.1 Determining the number of dynamic factors ............ 7

4 Estimating the Large Factor Model ................. 9
   4.1 The Principal Components Method .................. 10
   4.2 The Generalized Principal Components .............. 10
   4.3 The Maximum Likelihood Method .................. 12

5 Factor-Augmented Regressions ....................... 13

6 Factor-Augmented Vector-Autoregression (FAVAR) .... 15

7 IV Estimation with Many Instrumental Variables ... 16

8 Structural Changes in Large Factor Models ......... 18

9 Panel Data Models with Interactive Fixed Effects ... 20

10 Non-stationary Panel ................................ 23
    10.1 Estimating nonstationary factors .................. 26

11 Factor Models with Structural Restrictions ......... 26
    11.1 Factor Models with Block Structure ................ 27
    11.2 Linear Restrictions on Factor Loadings ............ 28
    11.3 SVAR and Restricted Dynamic Factor Models ....... 29

12 High Dimensional Covariance Estimation .......... 30

13 Bayesian Method to Large Factor Models .......... 31

14 Concluding Remarks ................................ 33

References ........................................... 34
1 Introduction

With the rapid development of econometric theory and methodologies on large factor models over the last decade, researchers are equipped with useful tools to analyze high dimensional data sets, which become increasingly available due to advancement of data collection technique and information technology. High dimensional data sets in economics and finance are typically characterized by both a large cross-sectional dimension $N$ and a large time dimension $T$. Conventionally, researchers rely heavily on factor models with observed factors to analyze such a data set, such as the capital asset pricing model (CAPM) and the Fama-French factor model for asset returns, and the affine models for bond yields. In reality, however, not all factors are observed. This poses both theoretical and empirical challenges to researchers. The development of modern theory on large factor models greatly broadens the scope of factor analysis, especially when we have high dimensional data and the factors are latent. Such a framework has helped substantially in the literature of diffusion-index forecast, business cycle co-movement analysis, economic linkage between countries, as well as improved causal inference through factor-augmented VAR, and providing more efficient instrumental variables. Stock & Watson (2002) find that if they first extract a few factors from the large data and then use the factors to augment an autoregressive model, the model has much better forecast performance than alternative univariate or multivariate models. Giannoni et al. (2008) apply the factor model to conduct now-casting, which combines data of different frequencies and forms a forecast for real GDP. Bernanke et al. (2004) find that using a few factors to augment a VAR helps to better summarize the information from different economic sectors and produces more credible impulse responses than conventional VAR model. Boivin & Giannoni (2006) combine factor analysis with DSGE models, providing a framework for estimating dynamic economic models using large data set. Such a methodology helps to mitigate the measurement error problem as well as the omitted variables problem during estimation. Using large dynamic factor models, Ng & Ludvigson (2009) have identified important linkages between bond returns and macroeconomic fundamentals. Fan et al. (2011) base their high dimensional covariance estimation on large approximate factor models, allowing sparse error covariance matrix after taking out common factors.

Apart from its wide applications, the large factor model also brings new insight
to our understanding of non-stationary data. For example, the link between cointegration and common trend is broken in the setup of large \((N,T)\). More importantly, the common trend can be consistently estimated regardless whether individual errors are stationary or integrated processes. That is, the number of unit roots can exceed the number of series, yet common trends are well defined and can be alienated from the data.

This review aims to introduce the theory and various applications of large factor models. In particular, this review examines basic issues related to high dimensional factors models, including determination of the number of factors, estimation of the model via the principal components as well as the maximum likelihood method, factor augmented regression and its application, factor-augmented vector autoregression (FAVAR), structural changes in large factor models, panel data models with interactive fixed effects. This review also examines how the framework of factor models can deal with the many instrumental variables problem, help to study non-stationarity in panel data, incorporate restrictions to identify structural shocks, facilitate high dimensional covariance estimation. In addition, this review provides a discussion of the Bayesian approach to large factor models as well as its possible extensions. We conclude with a few potential topics that deserve future research.

## 2 Large Factor Models

Suppose we observe \(x_{it}\) for the \(i\)-th cross-section unit at period \(t\). The large factor model for \(x_{it}\) is given by

\[
x_{it} = \lambda_i'F_t + e_{it}, \quad i = 1, \ldots, N, \ t = 1, \ldots, T,
\]

where the \(r \times 1\) vector of factors \(F_t\) is latent and the associated factor loadings \(\lambda_i\) is unknown. Model (1) can also be represented in vector form,

\[
X_t = \Lambda F_t + e_t,
\]

where \(X_t = [x_{1t}, \ldots, x_{Nt}]'\) is \(N \times 1\), \(\Lambda = [\lambda_1, \ldots, \lambda_N]'\) is \(N \times r\), and \(e_t = [e_{1t}, \ldots, e_{Nt}]'\) is \(N \times 1\). Unlike short panel data study (large \(N\), fixed \(T\)) and multivariate time series models (fixed \(N\), large \(T\)), the large factor model is characterized by both large \(N\) and large \(T\). The estimation and statistical inference are thus based on double asymptotic theory, in which both \(N\) and \(T\) converge to infinity. Such a large
dimensional framework greatly expands the applicability of factor models to more realistic economic environment. For example, weak correlations are allowed along both the time dimension and the cross-section dimension for $e_{it}$ without affecting the main properties of factor estimates. Such weak correlations in errors give rise to what Chamberlain & Rothschild (1983) called the approximate factor structure. Weak correlations between the factors and the idiosyncratic errors are also allowed. The high dimensional framework also brings new insight into double asymptotics. For example, it can be shown that it is $C_{NT} = \min \{N^{1/2}, T^{1/2}\}$ that determines the rate of convergence for the factor and factor loading estimates under large $N$ and large $T$.

3 Determining the Number of Factors

The number of factors is usually unknown. Alternative methods are available for estimation. We will mainly focus on two types of methods. One is based on information criteria and the other is based on the distribution of eigenvalues.

Bai & Ng (2002) treated this as a model selection problem, and proposed a procedure which can consistently estimate the number of factors when $N$ and $T$ simultaneously converge to infinity. Let $\hat{\lambda}_i^k$ and $\hat{F}_t^k$ be the principal component estimators assuming that the number of factors is $k$. We may treat the sum of squared residuals (divided by $NT$) as a function of $k$

$$V(k) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left( x_{it} - \hat{\lambda}_i^k \hat{F}_t^k \right)^2.$$ 

Define the following loss function

$$IC(k) = \ln(V(k)) + kg(N,T),$$

where the penalty function $g(N,T)$ satisfies two conditions: (i) $g(N,T) \to 0$, and (ii) $\min \{N^{1/2}, T^{1/2}\} \cdot g(N,T) \to \infty$, as $N,T \to \infty$. Define the estimator for the number of factors as $\hat{k}_{IC} = \arg\min_{0 \leq k \leq k_{max}} IC(k)$, where $k_{max}$ is the upper bound of the true number of factors $r$. Then consistency can be established under standard conditions on the factor model (Bai & Ng 2002; Bai 2003): $\hat{k}_{IC} \overset{p}{\to} r$, as $N,T \to \infty$. Bai & Ng (2002) considered six formulations of the information criteria, which are shown to have good finite sample performance. We list three here,

$$IC_1(k) = \ln(V(k)) + k \left( \frac{N + T}{NT} \right) \ln \left( \frac{NT}{N + T} \right),$$
\[ IC_2 (k) = \ln (V (k)) + k \left( \frac{N + T}{NT} \right) \ln \left( C_{NT}^2 \right), \]
\[ IC_3 (k) = \ln (V (k)) + k \left( \frac{\ln (C_{NT}^2)}{C_{NT}^2} \right). \]

The logarithm transformation in \( IC \) could be practically desirable, which avoids the need for a scaling factor in alternative criteria. Monte Carlo simulations show that all criteria perform well when both \( N \) and \( T \) are large. For the cases where either \( N \) or \( T \) is small, and if errors are uncorrelated across units and time, the preferred criteria tend to be \( IC_1 \) and \( IC_2 \). Weak serial and cross-section correlation in errors does not alter the result, however, the relative performance of each criterion depends on specific form of the correlation.

Some desirable features of the above method are worth mentioning. Firstly, the consistency is established without any restriction between \( N \) and \( T \), and it does not rely on sequential limits. Secondly, the results hold under heteroskedasticity in both the time and the cross-section dimensions, as well as under weak serial and cross-section correlation.

Based on large random matrix theory, Onatski (2009) established a test of \( k_0 \) factors against the alternative that the number of factors is between \( k_0 \) and \( k_1 \) \((k_0 < k \leq k_1)\). The test statistic is given by
\[ R = \max_{k_0 < k \leq k_1} \frac{\gamma_k - \gamma_{k+1}}{\gamma_{k+1} - \gamma_{k+2}}, \]
where \( \gamma_k \) is the \( k \)-th largest eigenvalue of the sample spectral density of data at a given frequency. For macroeconomic data, the frequency could be chosen at the business cycle frequency. The basic idea of this approach is that under the null of \( k_0 \) factors, the first leading \( k_0 \) eigenvalues will be unbounded, while the remaining eigenvalues are all bounded. As a result, \( R \) will be bounded under the null, while explode under the alternative, making \( R \) asymptotically pivotal. The limiting distribution of \( R \) is derived under the assumption that \( T \) grows sufficiently faster than \( N \), which turns out to be a function of the Tracy-Widom distribution.

Ahn & Horenstein (2013) proposed two estimators, the Eigenvalue Ratio (ER) estimator and the Growth Ratio (GR) estimator, based on simple calculation of eigenvalues. For example, the ER estimator is defined as maximizing the ratio of two adjacent eigenvalues in decreasing order. The intuition of these estimators is similar to Onatski (2009, 2010), though their properties are derived under slightly different model assumptions. Li et al. (2013) consider an increasing number of factors
3.1 Determining the number of dynamic factors

So far we have only considered the static factor model, where the relationship between \( x_{it} \) and \( F_t \) is static. The dynamic factor model considers the case in which lags of factors also directly affect \( x_{it} \). The methods for static factor models can be readily extended to estimate the number of dynamic factors. Consider

\[
x_{it} = \lambda_{i0} f_t + \lambda_{i1} f_{t-1} + \cdots + \lambda_{is} f_{t-s} + \epsilon_{it} = \lambda_i (L)^s f_t + \epsilon_{it},
\]

where \( f_t \) is \( q \times 1 \) and \( \lambda_i (L) = \lambda_{i0} + \lambda_{i1} L + \cdots + \lambda_{is} L^s \). While Forni et al. (2000, 2004, 2005) and a number of their subsequent studies considered the case with \( s \to \infty \), we will focus on the case with a fixed \( s \). Model (3) can be represented as a static factor model with \( r = q(s + 1) \) static factors,

\[
x_{it} = \lambda_i' F_t + \epsilon_{it},
\]

where

\[
\lambda_i = \begin{bmatrix} \lambda_{i0} \\ \lambda_{i1} \\ \vdots \\ \lambda_{is} \end{bmatrix} \quad \text{and} \quad F_t = \begin{bmatrix} f_t \\ f_{t-1} \\ \vdots \\ f_{t-s} \end{bmatrix}.
\]

We will refer to \( f_t \) as the dynamic factors and \( F_t \) as the static factors. Regarding the dynamic process for \( f_t \), we may use a finite-order VAR to approximate its dynamics. For example, \( f_t \) can follow VAR(h),

\[
\Phi(L) f_t = \epsilon_t,
\]

where \( \Phi(L) = I_q - \Phi_1 L - \cdots - \Phi_h L^h \). Then we may form the VAR(k) representation of the static factor \( F_t \), where \( k = \max\{h, s + 1\} \),

\[
\Phi_F(L) F_t = u_t,
\]

\[
u_t = R \epsilon_t,
\]

where \( \Phi_F(L) = I_q - \Phi_{F,1} L - \cdots - \Phi_{F,k} L^k \), and the \( q(s + 1) \times q \) matrix \( R \) are given by \( R = [I_q, 0, ..., 0]' \).

We may see from the VAR representation that the spectrum of the static factors has rank \( q \) instead of \( r = q(s + 1) \). Given that \( \Phi_F(L) F_t = R \epsilon_t \), the spectrum of \( F \) at frequency \( \omega \) is

\[
S_F(\omega) = \Phi_F(e^{-i\omega})^{-1} R S_\epsilon(\omega) R' \Phi_F(e^{i\omega})^{-1},
\]
whose rank is \( q \) if \( S_\varepsilon (\omega) \) has rank \( q \) for \( |\omega| \leq \pi \). This implies \( S_F (\omega) \) has only \( q \) nonzero eigenvalues. Bai & Ng (2007) refer to \( q \) as the number of primitive shocks. Hallin & Liska (2007) estimate the rank of this matrix to determine the number of dynamic factors. Onatski (2009) also considers estimating \( q \) using the sample estimates of \( S_F (\omega) \).

Alternatively, we may first estimate a static factor model using Bai & Ng (2002) to obtain \( \hat{F}_t \). Next, we may estimate a VAR\((p)\) for \( \hat{F}_t \) to obtain the residuals \( \hat{u}_t \). Let \( \hat{\Sigma}_u = \frac{1}{T} \sum_{t=1}^{T} \hat{u}_t \hat{u}_t', \) which is semipositive definite. Note that the theoretical moments \( E (u_t u_t') \) has rank \( q \). We expect that we may estimate \( q \) using the information about the rank of \( \hat{\Sigma}_u \). The formal procedure was provided in Bai & Ng (2007).

Using a different approach, Stock & Watson (2005) considered a richer dynamics in error terms, and transformed the model such that the residual of the transformed model has a static factor representation with \( q \) factors. Bai & Ng (2002)’s information criteria can then be directly applied to estimate \( q \).

Amengual & Watson (2007) considered a similar transformation as Stock & Watson (2005) and derived the corresponding econometric theory for estimating \( q \). They started from the static factor model (2), \( X_t = \Lambda F_t + e_t \), and considered a VAR\((p)\) for \( F_t \),

\[
F_t = \sum_{i=1}^{p} \Phi_i F_{t-i} + \varepsilon_t, \\
\varepsilon_t = G \eta_t,
\]

where \( G \) is \( r \times q \) with full column rank and \( \eta_t \) is a sequence of shocks with mean 0 and variance \( I_q \). The shock \( \eta_t \) is called the dynamic factor shock, whose dimension is called the number of dynamic factors. Let \( Y_t = X_t - \sum_{i=1}^{p} \Lambda \Phi_i F_{t-i} \) and \( \Gamma = \Lambda G \), then \( Y_t \) has a static factor representation with \( q \) factors,

\[
Y_t = \Gamma \eta_t + e_t.
\]

If \( Y_t \) is observed, \( q \) can be directly estimated using Bai & Ng (2002)’s information criteria. In practice, \( Y_t \) needs to be estimated. Let \( \hat{Y}_t = X_t - \sum_{i=1}^{p} \hat{\Lambda} \hat{\Phi}_i \hat{F}_{t-i} \), where \( \hat{\Lambda} \) and \( \hat{F}_t \) are principal components estimators from \( X_t \), and \( \hat{\Phi}_i \) is obtained by VAR\((p)\) regression of \( \hat{F}_t \). Amengual & Watson (2007) showed that Bai & Ng (2002)’s information criteria, when applied to \( \hat{Y}_t \), can consistently estimate the number of dynamic factors \( q \).
4 Estimating the Large Factor Model

The large factor model can be estimated using either the time-domain approach (mainly for static factor models) or frequency-domain approach (for dynamic factors). In this section, we focus on the time domain approach. In particular, we will mainly consider the principal component methods and the maximum likelihood methods. Examples of the frequency domain approach is provided by Forni et al. (2000, 2004, 2005). Throughout the remaining part, we assume the number of factors $r$ is known. If $r$ is unknown, it can be replaced by $\hat{k}$ using any of the information criteria discussed earlier without affecting the asymptotic properties of the estimators.

Before estimating the model, we need to impose normalizations on the factors and factor loadings to pin down the rotational indeterminacy. This is due to the fact that $\lambda_i'F_t = (A^{-1}\lambda_i)' (A'F_t)$ for any $r \times r$ full-rank matrix $A$. Because an arbitrary $r \times r$ matrix has $r^2$ degrees of freedom, we need to impose at least $r^2$ restrictions (order condition) to remove the indeterminacy. Let $F = (F_1, ..., F_T)'$. Three commonly applied normalizations are PC1, PC2, and PC3 as follows.

PC1: $\frac{1}{T}F'F = I_r$, $\Lambda'\Lambda$ is diagonal with distinct entries.

PC2: $\frac{1}{T}F'F = I_r$, the upper $r \times r$ block of $\Lambda$ is lower triangular with nonzero diagonal entries.

PC3: the upper $r \times r$ block of $\Lambda$ is given by $I_r$.

PC1 is often imposed by the maximum likelihood estimation in classical factor analysis, see Anderson & Rubin (1956). PC2 is analogous to a recursive system of simultaneous equations. PC3 is linked with the measurement error problem such that the first observable variable $x_{1t}$ is equal to the first factor $f_{1t}$ plus a measurement error $e_{1t}$, and the second observable variable $x_{2t}$ is equal to the second factor $f_{2t}$ plus a measurement error $e_{2t}$, and so on. Bai & Ng (2013) give a more detailed discussion on these restrictions.

Each preceding set of normalizations yields $r^2$ restrictions, meeting the order condition for identification (eliminating the rotational indeterminacy). These restrictions also satisfy the rank condition for identification (Bai & Wang 2014). The common components $\lambda_i'F_t$ have no rotational indeterminacy, and are identifiable without restrictions.
4.1 The Principal Components Method

The principal components estimators for factors and factor loadings can be treated as outcomes of a least squares problem under normalization PC1. Estimators under normalizations PC2 and PC3 can be obtained by properly rotating the principal components estimators. Consider minimizing the sum of squares residuals under PC1,

\[ \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - \lambda_i' F_t)^2 = \text{tr}[(X - F\Lambda')(X - F\Lambda')'] \]

where \( X = (X_1, X_2, ..., X_T)' \), a \( T \times N \) data matrix. It can be shown that \( \hat{F} \) is given by the first \( r \) leading eigenvectors\(^1\) of \( XX' \) multiplied by \( T^{1/2} / 2 \), and \( \hat{\Lambda} = X'\hat{F} / T \); see, for example, Connor & Korajczyk (1986) and Stock & Watson (1998). Bai (2003) studied the asymptotic properties under large \( N \) and \( T \). Under the general condition \( \sqrt{N}/T \to 0 \), \( \sqrt{N}(\hat{F}_t - HF_t^0) \xrightarrow{d} N(0, V_F) \), where \( H \) is the rotation matrix, \( F_t^0 \) is the true factor, and \( V_F \) is the estimable asymptotic variance. By symmetry, when \( \sqrt{T}/N \to 0 \), \( \hat{\lambda}_i \) is also asymptotically normal. Specifically, \( \sqrt{T}(\hat{\lambda}_i - H^{-1} \lambda_i^0) \xrightarrow{d} N(0, V_\Lambda) \), for some \( V_\Lambda > 0 \). The standard principal components estimates can be rotated to obtain estimates satisfying PC2 or PC3. The limiting distributions under PC2 and PC3 are obtained by Bai & Ng (2013). As for the common component \( \lambda_i' F_t \), its limiting distribution requires no restriction on the relationship between \( N \) and \( T \), which is always normal. In particular, there exists a sequence of \( b_{NT} = \min\{N^{1/2}, T^{1/2}\} \) such that

\[ b_{NT} (\hat{\lambda}_i' \hat{F}_t - \lambda_i' F_t) \xrightarrow{d} N(0, 1), \quad \text{as } N, T \to \infty. \]

So the convergence rate for the estimated common components is \( \min\{N^{1/2}, T^{1/2}\} \). This is the best rate possible.

4.2 The Generalized Principal Components

The principal components estimator is a least squares estimator (OLS) and is efficient if \( \Sigma_e \) is a scalar multiple of an \( N \times N \) identity matrix, that is, \( \Sigma_e = cI_N \) for a constant \( c > 0 \). This is hardly true in practice, therefore a generalized least squares (GLS) will give more efficient estimation. Consider the GLS objective function, assuming \( \Sigma_e \) is known,

\(^1\)If there is an intercept in the model \( X_t = \mu + \Lambda f_t + e_t \), then the matrix \( X \) is replaced by its demeaned version \( X - (1, 1, ..., 1) \otimes \bar{X} = (X_1 - \bar{X}, X_2 - \bar{X}, ..., X_T - \bar{X}) \), where \( \bar{X} = T^{-1} \sum_{t=1}^{T} X_t \).
\[
\min_{\Lambda, F} \text{tr} \left[ (X - F\Lambda')\Sigma_e^{-1}(X - F\Lambda')' \right]
\]

Then the solution for \( F \) is given by the first \( r \) eigenvectors of the matrix \( X\Sigma_e^{-1}X' \), multiplied by \( T^{1/2} \), and the solution for \( \Lambda \) is equal to \( X'\hat{F}/T \). The latter has the same expression as the standard principal components estimator.

This is the generalized principal components estimator (GPCE) considered by Choi (2012). He showed that the GPCE of the common component has smaller variance than the principal component estimator. Using the GPCE-based factor estimates will also produce smaller variance of the forecasting error.

In practice \( \Sigma_e \) is unknown, and needs to be replaced by an estimate. The usual covariance matrix estimator \( \frac{1}{T} \sum_{t=1}^{T} \hat{e}_t \hat{e}_t' \) based on the standard principal components residuals \( \hat{e}_t \) is not applicable since it is not a full rank matrix, regardless of the magnitude of \( N \) and \( T \). Even if every element of \( \Sigma_e \) can be consistently estimated by some way and the resulting estimator \( \Sigma_e \) is of full rank, the GPCE is not necessarily more accurate than the standard PC estimator. Unlike standard inference with a finite dimensional weighting matrix (such as GMM), a mere consistency of \( \hat{\Sigma}_e \) is insufficient to obtain the limiting distribution of the GPCE. Even an optimally estimated \( \hat{\Sigma}_e \) in the sense of Cai & Zhou (2012) is not enough to establish the asymptotically equivalence between the feasible and infeasible estimators. So the high dimensionality of \( \Sigma_e \) makes a fundamental difference in terms of inference. Bai & Liao (2013) show that the true matrix \( \Sigma_e \) has to be sparse and its estimates should take into account the sparsity assumption. Sparsity does not require many zero elements, but many elements must be sufficiently small. Under the sparsity assumption, a shrinkage estimator of \( \Sigma_e \) (for example, a hard thresholding method based on the residuals from the standard PC method) will give a consistent estimation of \( \Sigma_e \). Bai & Liao (2013) derive the conditions under which the estimated \( \Sigma_e \) can be treated as known.

Other related methods include Breitung & Tenhofen (2011), who proposed a two-step estimation procedure, which allows for heteroskedastic (diagonal \( \Sigma_e \)) and serially correlated errors. They showed that the feasible two-step estimator has the same limiting distribution as the GLS estimator. In finite samples, the GLS estimators tend to be more efficient than the usual principal components estimators. An iterated version of the two-step estimation method is also proposed, which is shown to further improve efficiency in finite sample.
For the factor model $X_t = \mu + \Lambda f_t + e_t$, under the assumption that $e_t$ is iid normal $N(0, \Sigma_e)$ and $f_t$ is iid normal $N(0, I_r)$, then $X_t$ is normal $N(\mu, \Omega)$, where $\Omega = \Lambda \Lambda' + \Sigma_e$, it follows that the likelihood function is

$$L(\Lambda, \Sigma_e) = -\frac{1}{N} \log |\Lambda \Lambda' + \Sigma_e| - \frac{1}{N} \text{tr}(S(\Lambda \Lambda' + \Sigma_e)^{-1})$$

where $S = \frac{1}{T} \sum_{t=1}^{T} (X_t - \bar{X})(X_t - \bar{X})'$ is the sample covariance matrix. The classical inferential theory of the MLE is developed assuming $N$ is fixed, and the sample size $T$ goes to infinity, see Anderson (2013), Anderson & Rubin (1956), and Lawley & Maxwell (1971). The basic assumption in classical factor analysis is that $\sqrt{T}(S - \Omega)$ is asymptotically normal. This basic premise does not hold when $N$ also goes to infinity.

A new approach is required to obtain consistency and the limiting distribution under large $N$. Bai & Li (2012a) derive the inferential theory assuming $\Sigma_e$ is diagonal under various identification restrictions. Normality assumption is inessential. The preceding likelihood is considered as the quasi-likelihood under non-normality. It is useful to point out that MLE is consistent even if $N$ is fixed because the fixed $N$ case falls within the classical framework. In contrast, the principal components method is inconsistent under fixed $N$ unless $\Sigma_e$ is proportional to an identity matrix. The GPCE is hardly consistent when using residuals to estimate $\Sigma_e$ because the residual $\hat{e}_{it} = x_{it} - \hat{\mu}_i - \hat{\lambda}_i'\hat{f}_t$ is not consistent for $e_{it}$. This follows because $\hat{f}_t$ is not consistent for $f_t$ under fixed $N$. The MLE treats $\Sigma_e$ as a parameter, which is jointly estimated with $\Lambda$. The MLE does not rely on residuals to estimate $\Sigma_e$.

The maximum likelihood estimation for non-diagonal $\Sigma_e$ is considered by Bai & Liao (2012). They assume $\Sigma_e$ is sparse and use the regularization method to jointly estimate $\Lambda$ and $\Sigma_e$. Consistency is established for the estimated $\Lambda$ and $\Sigma_e$, but the limiting distributions remain unsolved, though the limiting distributions are conjectured to be the same as the two-step feasible GLS estimator in Bai & Liao (2013) under large $N$ and $T$.

Given the MLE for $\Lambda$ and $\Sigma_e$, the estimator for $f_t$ is $\hat{f}_t = (\hat{\Lambda}'\hat{\Sigma}_e^{-1}\hat{\Lambda})^{-1}\hat{\Lambda}'\hat{\Sigma}_e^{-1}(X_t - \bar{X})$ for $t = 1, 2, ..., T$. This is a feasible GLS estimator of $f_t$ in the model $X_t = \mu + \Lambda f_t + e_t$. The estimated factor loadings have the same asymptotic distributions for the three different estimation methods (PC, GPCE, MLE) under large $N$ and large $T$. But the estimated factors are more efficient under GPCE and MLE than standard
If time series heteroskedasticity is of more concern, and especially when \( T \) is relatively small, then the role of \( F \) and \( \Lambda \) (also \( T \) and \( N \)) should be switched. Bai & Li (2012) considered the likelihood function for this setting.

The preceding discussion assumes the factors \( f_t \) are iid. Doz et al. (2011, 2012) explicitly considered a finite-order VAR specification for \( f_t \), and then proposed a two-step method or a quasi-maximum likelihood estimation procedure. The method is similar to the maximum likelihood estimation of a linear state space model. The main difference is that they initialize the estimation by using properly rotated principal component estimators. The factor estimates are obtained as either the Kalman filter or the Kalman smoother. They showed that estimation under independent Gaussian errors still lead to consistent estimators for the large approximate factor model, even when the true model has cross-sectional and time series correlation in the idiosyncratic errors. Bai & Li (2012b) studied related issues for dynamic factors and cross-sectionally and serially correlated errors estimated by the maximum likelihood method.

5 Factor-Augmented Regressions

One of the popular applications of large factor model is the factor-augmented regressions. For example, Stock & Watson (1999, 2002a) added a single factor to standard univariate autoregressive models and found that they provided the most accurate forecasts of macroeconomic time series such as inflation and industrial production among a large set of models. Bai & Ng (2006) developed the econometric theory for such factor-augmented regressions so that inference can be conducted.

Consider the following forecasting model for \( y_t \),

\[
y_{t+h} = \alpha' F_t + \beta' W_t + \varepsilon_{t+h},
\]

where \( W_t \) is the vector of a small number of observables including lags of \( y_t \), and \( F_t \) is unobservable. Suppose there is a large number of series \( x_{it}, i = 1, ..., N, t = 1, ..., T \), which has a large factor representation as (1),

\[
x_{it} = \lambda_i' F_t + e_{it}.
\]

When \( y_t \) is a scalar, (5) and (1) become the diffusion index forecasting model of Stock & Watson (2002b). Clearly, each \( x_{it} \) is a noisy predictor for \( y_{t+h} \). Because \( F_t \)
is large relative to \( N \). Let \( \delta \) and then regress \( y_{t+h} \) on \( \hat{F}_t \) and \( W_t \) to obtain \( \hat{\alpha} \) and \( \hat{\beta} \). The feasible prediction for \( y_{T+h|T} \equiv E (y_{T+h}|\Omega_T) \), where \( \Omega_T = [F_T, W_T, F_{T-1}, W_{T-1}, \ldots] \), is given by

\[
\hat{y}_{T+h|T} = \hat{\alpha}' \hat{F}_T + \hat{\beta} W_T.
\]

Let \( \delta \equiv (\alpha'H^{-1}, \beta')' \) and \( \varepsilon_{T+h|T} \equiv y_{T+h} - y_{T+h|T} \), Bai & Ng (2006) showed that when \( N \) is large relative to \( T \) (i.e., \( \sqrt{T}/N \to 0 \)), \( \hat{\delta} \) will be \( \sqrt{T} \)-consistent and asymptotically normal. \( \hat{y}_{T+h|T} \) and \( \hat{\varepsilon}_{T+h|T} \) are min \( \{N^{1/2}, T^{1/2}\} \)-consistent and asymptotically normal. For all cases, inference needs to take into account the estimated factors, except for the special case \( T/N \to 0 \). In particular, under standard assumptions for large approximate factor model as in Bai & Ng (2002), when \( \sqrt{T}/N \to 0 \), we have

\[
\hat{\delta} - \delta \xrightarrow{d} N(0, \Sigma_\delta).
\]

Let \( z_t = [F_t', W_t']' \), \( \hat{z}_t = [\hat{F}_t', \hat{W}_t']' \), and \( \hat{\varepsilon}_{t+h} = y_{t+h} - \hat{y}_{t+h|t} \), a heteroskedasticity-consistent estimator for \( \Sigma_\delta \) is given by

\[
\hat{\Sigma}_\delta = \left( \frac{1}{T} \sum_{t=1}^{T-h} \hat{z}_t \hat{z}_t' \right)^{-1} \left( \frac{1}{T} \sum_{t=1}^{T-h} \hat{\varepsilon}_{t+h} \hat{z}_t \hat{z}_t' \right) \left( \frac{1}{T} \sum_{t=1}^{T-h} \hat{z}_t \hat{z}_t' \right)^{-1}.
\]

The key requirement of the theory is \( \sqrt{T}/N \to 0 \), which puts discipline on when estimated factors can be applied in the diffusion index forecasting models as well as the FAVAR to be considered in the next section.

If in addition, we assume \( \sqrt{N}/T \to 0 \), then

\[
\frac{\hat{y}_{T+h|T} - y_{T+h|T}}{\sqrt{\text{var} (\hat{y}_{T+h|T})}} \xrightarrow{d} N(0, 1),
\]

where

\[
\text{var} (\hat{y}_{T+h|T}) = \frac{1}{N} \hat{\varepsilon}_T \hat{\text{Avar}} \left( \hat{\delta} \right) \hat{\varepsilon}_T + \frac{1}{N} \hat{\alpha}' \hat{\text{Avar}} \left( \hat{F}_T \right) \hat{\alpha}.
\]

An estimator for \( \hat{\text{Avar}} \left( \hat{F}_T \right) \) is provided in Bai (2003). A notable feature of the limiting distribution of the forecast is that the overall convergence rate is given by min \( \{N^{1/2}, T^{1/2}\} \). Given that \( \hat{\varepsilon}_{T+h} = \hat{y}_{T+h|T} - y_{T+h} = \hat{y}_{T+h|T} - y_{T+h|T} + \varepsilon_{T+h} \), if we further assume that \( \varepsilon_t \) is normal with variance \( \sigma_\varepsilon^2 \), then the forecasting error also becomes approximately normal

\[
\hat{\varepsilon}_{T+h} \sim N \left( 0, \sigma_\varepsilon^2 + \text{var} (\hat{y}_{T+h|T}) \right),
\]

so that confidence intervals can be constructed for the forecasts.
6  Factor-Augmented Vector-Autoregression (FAVAR)

VAR models have been widely applied in macroeconomic analysis. A central question of using VAR is how to identify structural shocks, which in turn depends on what variables to include in the VAR system. A small VAR usually cannot fully capture the structural shocks. In the meantime, including more variables in the VAR system could be problematic due to either the degree of freedom problem or the variable selection problem. It has been a challenging job to determine which variables to be included in the system. There have been a number of ways to overcome such difficulties, such as Bayesian VAR, initially considered by Doan et al. (1984), Litterman (1986), and Sims (1993), and Global VAR of Pesaran et al. (2004). This section will focus on another popular solution, the factor-augmented vector autoregressions (FAVAR), originally proposed by Bernanke et al. (2005). The FAVAR assumes that a large number of economic variables are driven by a small VAR, which can include both latent and observed variables. The dimension of structural shocks can be estimated instead of being assumed to be known and fixed.

Consider the case where both the unobserved factors $F_t$ and the observed factors $W_t$ affect a large number of observed variables $x_{it}$,

$$x_{it} = \lambda_i' F_t + \gamma_i' W_t + e_{it}, \tag{6}$$

and that the vector $H_t = [F_t', W_t']'$ follows a VAR of finite order,

$$\Phi(L) H_t = u_t.$$

Bernanke et al. (2005) proposed two ways to analyze the FAVAR. The first is based on a two-step principal components method, where in the first step method of principal components is employed to form estimates of the space spanned by both $F_t$ and $W_t$. In the second step, various identification schemes can be applied to obtain estimates of latent factors $\hat{F}_t$, which is treated as observed when conduct VAR analysis of $[\hat{F}_t', W_t']'$. Bai et al. (2015) show that, under suitable identification conditions, inferential theory can be developed for such a two-step estimator, which differs from a standard large factor model. Confidence bands for the impulse responses can be readily constructed using the theory therein. The second method involves a one-step likelihood approach, implemented by Gibbs sampling, which leads to joint estimation of both the latent factors and the impulse responses. The two methods can be complement of each other, with the first one being computationally simple, and the
second providing possibly better inference in finite sample at the cost of increased computational cost.

A useful feature of the FAVAR is that the impulse response function of all variables to the fundamental shocks can be readily calculated. For example, the impulse response of the observable \( x_{i,t+h} \) with respect to the structural shock \( u_t \) is,

\[
\frac{\partial x_{i,t+h}}{\partial u_t} = (\lambda'_i, \gamma'_i) C_h, \tag{7}
\]

where \( C_h \) is the coefficient matrix for \( u_{t-h} \) in the vector moving average (VMA) representation of \( H_t \),

\[
H_t = \Phi(L)^{-1} u_t = C_0 u_t + C_1 u_{t-1} + C_2 u_{t-2} + \cdots.
\]

Theory of estimation and inference for (7) is provided in Bai et al. (2015). Forni et al. (2009) explored the structural implication of the factors and developed the corresponding econometric theory. Stock and Watson (2010) provide a survey on the application of dynamic factor models.

### 7 IV Estimation with Many Instrumental Variables

The IV method is fundamental in econometrics practice. It is useful when one or more explanatory variables are correlated with the error terms in a regression model, known as endogenous regressors. In this case, standard methods such as the OLS are inconsistent. With the availability of instrumental variables, which are correlated with regressors but uncorrelated with errors, consistent estimation is achievable. Consider a standard setup

\[
y_t = x'_t \beta + u_t, \quad t = 1, 2, ..., T, \tag{8}
\]

where \( x_t \) is correlated with the error term \( u_t \). Suppose there are \( N \) instrumental variables labeled as \( z_{it} \) for \( i = 1, 2, ..., N \). Consider the two-stage least squares (2SLS), a special IV method. In the first stage, the endogenous regressor \( x_t \) is regressed on the IVs

\[
x_t = c_0 + c_1 z_{1t} + \cdots + c_N z_{Nt} + \eta_t \tag{9}
\]

the fitted value \( \hat{x}_t \) is used as the regressor in the second stage, and the resulting estimator is consistent for \( \beta \) for a small \( N \). It is known that for large \( N \), the 2SLS
can be severely biased. In fact, if $N \geq T$, then $\hat{x}_t \equiv x_t$, and the 2SLS method coincides with the OLS, which is inconsistent. The problem lies in the overfitting in the first stage regression. The theory of many instruments bias has been extensively studied in the econometric literature, for example, Hansen et al. (2008) and Hausman et al. (2010). Within the GMM context, inaccurate estimation of a high dimensional optimal weighting matrix is not the cause for many-moments bias. Bai and Ng (2010) show that even if the true optimal weighting matrix is used, inconsistency is still obtained under large number of moments. In fact, with many moments, sparse weighting matrix such as an identity matrix will give consistent estimation, as is shown by Meng et al. (2011).

One solution for the many IV problem is to assume that many of the coefficients in (9) are zero (sparse) so that regularization method such as LASSO can be used in the first stage regression. Penalization prevents over fitting, and picks up the relevant instruments (non-zero coefficients). These methods are considered by Ng & Bai (2009) and Belloni et al. (2012). In fact, any machine learning method that prevents in-sample overfitting in the first stage regression will work.

The principal components method is an alternative solution, and can be more advantageous than the regularization method (Bai & Ng 2010; Kapetanios & Marcellino 2010). It is well known that the PC method is that of dimension reduction. The principal components are linear combinations of $z_{1t}, \ldots, z_{Nt}$. The high dimension of the IVs can be reduced into a smaller dimension via the PC method. If $z_{1t}, \ldots, z_{NT}$ are valid instruments, then any linear combination is also a valid IV, so are the principal components. Interestingly, the PC method does not require the $z$’s to be valid IVs to begin with. Suppose that both the regressors $x_t$ and $z_{1t}, \ldots, z_{NT}$ are driven by some common factors $F_t$ such that

\begin{align*}
x_t &= \phi'F_t + e_{xt}, \\
z_{it} &= \lambda_{i}'F_t + e_t.
\end{align*}

Provided that the common shocks $F_t$ are uncorrelated with errors $u_t$ in equation (8), then $F_t$ is a valid IV because $F_t$ also drives $x_t$. Although $F_t$ is unobservable, it can be estimated from $\{z_{it}\}$ via the principal components. Even though the $e_t$ can be correlated with $u_t$, so that $\{z_{it}\}$ are not valid IV, the principal components are valid IVs. This is the advantage of principal components method. The example in Meng et al. (2011) is instructive. Consider estimating the beta of an asset with respect
to the market portfolio, which is unobservable. The market index, as a proxy, is measured with errors. But other assets’ returns can be used as instrument variables because all assets’s returns are linked with the market portfolio. So there exists a large number of instruments. Each individual asset can be a weak IV because the idiosyncratic returns can be large. But the principal components method will wash out the idiosyncratic errors, giving rise to a more effective IV.

The preceding setup can be extended into panel data with endogenous regressors:

$$y_{it} = x_{it}' \beta + u_{it}, \quad t = 1, 2, \ldots, T; \quad i = 1, 2, \ldots, N,$$

where $x_{it}$ are correlated with the errors $u_{it}$. Suppose that $x_{it}$ are driven by the common shocks $F_t$ such that

$$x_{it} = \gamma_i' F_t + e_{it}$$

Provided that the common components $c_{it} = \gamma_i' F_t$ are uncorrelated with $u_{it}$, outside instrumental variables are not needed. We can extract the common components $c_{it}$ via the principal components estimation of $\gamma_i$ and $F_t$ to form $\hat{c}_{it} = \hat{\gamma}_i' \hat{F}_t$, and use $\hat{c}_{it}$ as IV. This method is considered by Bai & Ng (2010).

8 Structural Changes in Large Factor Models

In a lot of economic applications, researchers have to be cautious about the potential structural changes in the high dimensional data sets. Parameter instability has been a pervasive phenomenon in time series data (Stock & Watson 1996). Such instability could be due to technological changes, preference shifts of consumers, and policy regimes switching. Banerjee & Marcellino (2008) and Yamamoto (2014) provide simulation and empirical evidence that the forecasts based on estimated factors will be less accurate if the structural break in the factor loading matrix is ignored. In addition, evolution of an economy might introduce new factors, while conventional factors might phase out. Econometric analysis of the structural change in large factor models is challenging because the factors are unobserved and factor loadings have to be estimated. Structural change can happen to either the factor loadings or the dynamic process of factors or both. Most theoretical challenge comes from the break in factor loadings, given that factors can be consistently estimated by principal components even if their dynamic process is subject to a break while factor loadings are time-invariant. In this section, we will focus on the breaks in factor loadings.
Consider a time-varying version of (2),

\[ X_t = \Lambda_t F_t + e_t, \]

where the time-varying loading matrix \( \Lambda_t \) might assume different forms. Bates et al. (2013) considered the following dynamic equation

\[ \Lambda_t = \Lambda_0 + h_{NT} \xi_t, \]

where \( h_{NT} \) is a deterministic scalar that depends on \((N, T)\), and \( \xi_t \) is \( N \times r \) stochastic process. Three examples are considered for \( \xi_t \): white noise, random walk, and single break. Bates et al. (2013) then established conditions under which the changes in the loading matrix can be ignored in the estimation of factors. Intuitively, the estimation and inference of the factor estimates is not affected if the size of the break is small enough. If the size of the break is large, however, the principal components factor estimators will be inconsistent.

Recent years have witnessed fast development in this area. Tests for structural changes in factor loadings of a specific variable have been derived by Stock & Watson (2008), Breitung & Eickmeier (2011), and Tanaka & Yamamoto (2015). Chen et al. (2014) and Han & Inoue (2014) studied tests for structural changes in the overall factor loading matrix. Corradi & Swanson (2014) constructed joint tests for breaks in factor loadings and coefficients in factor-augmented regressions. Cheng et al. (2015) considered the determination of break date and introduced the shrinkage estimation method for factor models in the presence of structural changes. Current studies have not considered the case where break dates are possibly heterogeneous across variables and the number of break dates might increase with the sample size. It would be interesting to study the properties of the principal components factor estimators and the power of the structural change tests under such scenarios.

Another branch of methods considers Markov regime switching in factor loadings (Kim & Nelson 1999). The likelihood function can be constructed using various filters. Then one may use either maximum likelihood or Bayesian method to estimate factor loadings in different regimes, the regime probabilities, as well as the latent factors. Del Negro & Otrok (2008) developed a dynamic factor model with time-varying factor loadings and stochastic volatility in both the latent factors and idiosyncratic components. A Bayesian algorithm is developed to estimate the model, which is employed to study the evolution of international business cycles. The theoretical properties of such models remain to be studied under the large \( N \) large \( T \) setup.
9 Panel Data Models with Interactive Fixed Effects

There has been growing study in panel data models with interactive fixed effects. Conventional methods assume additive individual fixed effects and time fixed effects. The interactive fixed effects allows possible multiplicative effects. Such a methodology has important theoretical and empirical relevance. Consider the following large \( N \) large \( T \) panel data model

\[
y_{it} = X_{it}' \beta + u_{it},
\]

\[
u_{it} = \lambda_i' F_t + \varepsilon_{it}.
\]

We observe \( y_{it} \) and \( X_{it} \), but do not observe \( \lambda_i, F_t, \) and \( \varepsilon_{it} \). The coefficient of interest is \( \beta \). Note that such model nests conventional fixed effects panel data models as special cases due to the following simple transformation

\[
y_{it} = X_{it}' \beta + \alpha_i + \xi_t + \varepsilon_{it}
\]

\[
= X_{it}' \beta + \lambda_i' F_t + \varepsilon_{it},
\]

where \( \lambda_i = [1, \alpha_i]' \) and \( F_t = [\xi_t, 1]' \). In general, the interactive fixed effects allow a much richer form of unobserved heterogeneity. For example, \( F_t \) can represent a vector of macroeconomic common shocks and \( \lambda_i \) captures individual \( i \)'s heterogeneous response to such shocks.

The theoretic framework of Bai (2009) allows \( X_{it} \) to be correlated with \( \lambda_i, F_t, \) or both. Under the framework of large \( N \) and large \( T \), we may estimate the model by minimizing a least squares objective function

\[
SSR(\beta, F, \Lambda) = \sum_{i=1}^{N} (Y_i - X_i \beta - F \lambda_i)' (Y_i - X_i \beta - F \lambda_i)
\]

s.t. \( F'F/T = I_r, \ \Lambda' \Lambda \) is diagonal.

Although no closed-form solution is available, the estimators can be obtained by iterations. Firstly, consider some initial values \( \beta^{(0)} \), such as least squares estimators from regressing \( Y_i \) on \( X_i \). Then perform principal component analysis for the pseudo-data \( Y_i - X_i \beta^{(0)} \) to obtain \( F^{(1)} \) and \( \Lambda^{(1)} \). Next, regress \( Y_i - F^{(1)} \lambda_i^{(1)} \) on \( X_i \) to obtain \( \beta^{(1)} \). Iterate such steps until convergence is achieved. Bai (2009) showed that the resulting estimator \( \hat{\beta} \) is \( \sqrt{NT} \)-consistent. Given such results, the limiting distributions for \( \hat{F} \)
and $\hat{A}$ are the same as in Bai (2003) due to their slower convergence rates. The limiting distribution for $\hat{\beta}$ depends on specific assumptions on the error term $\varepsilon_{it}$ as well as on the ratio $T/N$. If $T/N \to 0$, then the limiting distribution of $\hat{\beta}$ will be centered around zero, given that $E(\varepsilon_{it}\varepsilon_{js}) = 0$ for $t \neq s$, and $E(\varepsilon_{it}\varepsilon_{jt}) = \sigma_{ij}$ for all $i, j, t$.

On the other hand, if $N$ and $T$ are comparable such that $T/N \to \rho > 0$, then the limiting distribution will not be centered around zero, which poses a challenge for inference. Bai (2009) provided a bias-corrected estimator for $\beta$, whose limiting distribution is centered around zero. In particular, the bias-corrected estimator allows for heteroskedasticity across both $N$ and $T$. Let $\hat{\beta}$ be the bias-corrected estimator, assume that $T/N^2 \to 0$ and $N/T^2 \to 0$, $E(\varepsilon_{it}^2) = \sigma_{it}^2$, and $E(\varepsilon_{it}\varepsilon_{js}) = 0$ for $i \neq j$ and $t \neq s$, then

$$\sqrt{NT}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Sigma_{\beta}),$$

where a consistent estimator for $\Sigma_{\beta}$ is also available in Bai (2009).

Ahn et al. (2001, 2013) studied model (10) under large $N$ but fixed $T$. They employed the GMM method that applies moments of zero correlation and homoskedasticity. Moon & Weidner (2014) considered the same model as (10), but allowed lagged dependent variable as regressors. They devised a quadratic approximation of the profile objective function to show the asymptotic theory for the least square estimators and various test statistics. Moon & Weidner (2015) further extended their study by allowing unknown number of factors. They showed that the limiting distribution of the least square estimator is not affected by the number of factors used in the estimation, as long as this number is no smaller than the true number of factors. Lu & Su (2015) proposed the method of adaptive group LASSO (least absolute shrinkage and selection operator), which can simultaneously select the proper regressors and determine the number of factors.

Pesaran (2006) considered a slightly different setup with individual-specific slopes,

$$y_{it} = \alpha_i' d_t + X_{it}' \beta_i + u_{it},$$
$$u_{it} = \lambda_i' F_t + \varepsilon_{it}, \quad (11)$$

where $d_t$ is observed common effects such as seasonal dummies. The regressor $X_{it}$ is allowed to be correlated with both $\lambda_i$ and $F_t$, which makes direct OLS inconsistent. The unobserved factors and the individual-specific errors are allowed to follow arbitrary stationary processes. Instead of estimating the factors and factor
loadings, Pesaran (2006) considered an auxiliary OLS regression. The proposed common correlated effects estimator (CCE) can be obtained by augmenting the model with additional regressors, which are the cross sectional averages of the dependent and independent variables, in an attempt to control for the common factors. Define $z_{it} = [y_{it}, X'_{it}]'$ as the collection of individual-specific observations. Consider weighted average of $z_{it}$ as

$$\tilde{z}_{\omega t} = \sum_{i=1}^{N} \omega_i z_{it},$$

with the weights that satisfy some very general conditions. For example, we may choose $\omega_i = 1/N$. Pesaran (2006) showed that the individual slope $\beta_i$ can be consistently estimated through the following OLS regression of $y_{it}$ on $d_t$, $X_{it}$, and $\tilde{z}_{\omega t}$ under the large $N$ and large $T$ framework,

$$y_{it} = \alpha_i' d_t + X_{it}' \beta_i + \tilde{z}_{\omega t}' \gamma_i + \varepsilon_{it}. \tag{12}$$

In the same time, the mean $\beta = E(\beta_i)$ can be consistently estimated by a pooled regression of $y_{it}$ on $d_t$, $X_{it}$, and $\tilde{z}_{\omega t}$ as long as $N$ tends to infinity, regardless of large or small $T$,

$$y_{it} = \alpha_i' d_t + X_{it}' \beta + \tilde{z}_{\omega t}' \gamma + v_{it}. \tag{13}$$

Alternatively, the mean $\beta$ can also be consistently estimated by taking simple average of individual estimators from (12), $\hat{\beta}_{MG} = \frac{1}{N} \sum_{i=1}^{N} \hat{\beta}_{CCE}^i$. Asymptotic normality for such CCE estimators can also be established and help conduct inference. A clear advantage of the CCE method is that the number of unobserved factors need not be estimated. In fact, the method is valid with a single or multiple unobserved factors and does not require the number of factors to be smaller than the number of observed cross-section averages. In addition, CCE is easy to compute as an outcome of OLS and no iteration is needed. Desirable small sample properties of CCE are also demonstrated.

Heterogenous panel models with interactive effects are also studied by Ando & Bai (2014), where the number of regressors can be large and the regularization method is used to select relevant regressors. Su & Chen (2013) and Ando & Bai (2015) provided a formal test for homogenous coefficients in model (11).

The maximum likelihood estimation of model (10) is studied by Bai & Li (2014). They consider the case in which $X_{it}$ also follows a factor structure and is jointly modeled.
Other important topics includes the application of large factor models in non-stationary panel data, estimation and inference of dynamic factor models.

For non-stationary analysis, large factor model brings new prospective for the tests of unit roots. Consider the following data generating process for $x_{it}$,

$$
x_{it} = c_i + \beta_i t + \lambda_i' F_t + e_{it},
$$

(10)

$$
(1 - L) F_t = C(L) u_t,
$$

(11)

$$
(1 - \rho_i L) e_{it} = D_i(L) \epsilon_{it},
$$

where $C(L)$ and $D_i(L)$ are polynomials of lag operators. The $r \times 1$ factor $F_t$ has $r_0$ stationary factors and $r_1 I(1)$ components or common trends ($r = r_0 + r_1$). The idiosyncratic errors $e_{it}$ could be either $I(1)$ or $I(0)$, depending on whether $\rho_i = 1$ or $|\rho_i| < 1$. The Panel Analysis of Nonstationarity in the Idiosyncratic and Common components (PANIC) by Bai & Ng (2004) develops an econometric theory for determining $r_1$ and testing $\rho_i = 1$ when neither $F_t$ nor $e_{it}$ is observed. Model (14) has many important features that are of both theoretical interests and empirical relevance. For example, let $x_{it}$ denote the real output for country $i$. Then it may be determined by the global common trend $F_{1t}$, the global cyclical component $F_{2t}$, and an idiosyncratic component $e_{it}$ that could be either $I(0)$ or $I(1)$. PANIC provides a framework for the estimation and statistical inference for such components, which are all unobserved. While the conventional nonstationarity analysis looks at unit root in $x_{it}$ only, PANIC further explores whether possible unit roots are coming from common factors or idiosyncratic components or both. Another very important feature of PANIC is that it allows weak cross-section correlation in idiosyncratic errors.

The initial steps of PANIC include transformations of the data so that the deterministic trend part is removed, and then proceed with principal component analysis. In the case of no linear trend ($\beta_i = 0$ for all $i$), a simple first differencing will suffice. We will proceed with the example with a linear trend ($\beta_i \neq 0$ for some $i$). Let $\Delta x_{it} = x_{it} - x_{i,t-1}$, $\overline{\Delta x_i} = \frac{1}{T-1} \sum_{t=2}^{T} \Delta x_{it}$, $\overline{\Delta e_i} = \frac{1}{T-1} \sum_{t=2}^{T} \Delta e_{it}$, $\overline{\Delta F} = \frac{1}{T-1} \sum_{t=2}^{T} \Delta F_t$. After first differencing and remove the time average, model (14) becomes

$$
\Delta x_{it} - \overline{\Delta x_i} = \lambda_i' f_t + \Delta e_{it} - \overline{\Delta e_i},
$$

(15)
where \( f_t = \Delta F_t - \Delta F \). Let \( \hat{\lambda}_i \) and \( \hat{f}_t \) be the principal components estimators of (15). Let \( \hat{z}_{it} = \Delta x_{it} - \Delta x_{i} - \hat{\lambda}_i \hat{f}_t \), \( \hat{\epsilon}_{it} = \sum_{s=2}^{t} \hat{z}_{it} \), and \( \hat{F}_t = \sum_{s=2}^{t} \hat{f}_t \). Then test statistics for unit root in \( \hat{\epsilon}_{it} \) and \( \hat{F}_t \) can be constructed based on the estimates \( \hat{\epsilon}_{it} \) and \( \hat{F}_t \). Let \( ADF^\tau_{\hat{\epsilon}}(i) \) denote the Augmented Dickey-Fuller test using \( \hat{\epsilon}_{it} \). The limiting distributions and critical values are provided in Bai & Ng (2004).

A few important properties of PANIC are worth mentioning. First, the test on idiosyncratic errors \( \hat{\epsilon}_{it} \) can be conducted without knowing whether the factors are \( I(1) \) or \( I(0) \). Likewise, the test on factors is valid regardless whether \( \hat{\epsilon}_{it} \) is \( I(1) \) or \( I(0) \). Last but not the least, the test on \( \hat{\epsilon}_{it} \) is valid no matter whether \( \hat{\epsilon}_{jt} (j \neq i) \) is \( I(1) \) or \( I(0) \). In fact, the limiting distribution of \( ADF^\tau_{\hat{\epsilon}}(i) \) does not depend on the common factors. This helps to construct pooled panel unit root tests, which have improved power as compared to univariate unit root tests.

The literature on panel unit root tests has been growing fast. The early test in Quah (1994) requires strong homogeneous cross-sectional properties. Later tests of Levin et al. (2002) and Im et al. (2003) allow for heterogeneous intercepts and slopes, but assume cross-sectional independence. Such an assumption is restrictive, and tend to over reject the null hypothesis when violated. O’Connell (1998) provides a GLS solution to this problem under fixed \( N \). However, such a solution does not apply when \( N \) tends to \( \infty \), especially for the case where \( N > T \). The method of PANIC allows strong cross-section correlation due to common factors, as is widely observed in economic data, as well as weak cross-section correlation in \( \hat{\epsilon}_{it} \). In the same time, it allows heterogeneous intercepts and slopes. If we further assume that the idiosyncratic errors \( \hat{\epsilon}_{it} \) are independent across \( i \), and consider testing \( H_0 : \rho_i = 1 \) for all \( i \), against \( H_1 : \rho_i < 1 \) for some \( i \), a pool test statistic can be readily constructed.

Let \( p^\tau_{\hat{\epsilon}}(i) \) be the \( p \)-value associated with \( ADF^\tau_{\hat{\epsilon}}(i) \). Then

\[
P^\tau_{\hat{\epsilon}} = \frac{-2 \sum_{i=1}^{N} \log p^\tau_{\hat{\epsilon}}(i) - 2N}{\sqrt{4N}} \overset{d}{\rightarrow} N(0,1).
\]

The pooled test of the idiosyncratic errors can also be seen as a panel test of no cointegration, as the null hypothesis that \( \rho_i = 1 \) for all \( i \) holds only if no stationary combination of \( x_{it} \) can be formed.

Alternative methods to PANIC include Kapetanios et al. (2011), who derived the theoretical properties of the common correlated effects estimator (CCE) of Pesaran (2006) for panel regression with nonstationary common factors. The model slightly differs from PANIC in the sense that individual slopes can be studied. For example,
they consider
\[ y_{it} = \alpha_i't + X_{it}'\beta_i + \lambda'_iF_t + \varepsilon_{it}, \]
and the parameter of interest is \( \beta_i \). Another difference is that the individual error \( \varepsilon_{it} \) is assumed to be stationary. They show that the cross-sectional-average-based CCE estimator is robust to a wide variety of data generation processes, and not just restricted to stationary panel regression. Similar to Pesaran (2006), the method does not require the knowledge of the number of unobserved factors. The only requirement is that the number of unobserved factors remains fixed as the sample size grows. The main results of Pesaran (2006) continue to hold in the case of nonstationary panel. It is also shown that the CCE has lower biases than the alternative estimation methods.

While Kapetanios et al. (2011) does not focus on testing panel unit roots, the main idea of CCE can be applied for such a purpose. For a given individual \( i \), Pesaran (2007) augments the Dickey-Fuller (DF) regression of \( y_{it} \) with cross-section averages \( \bar{y}_{t-1} = \frac{1}{N} \sum_{i=1}^{N} y_{i,t-1} \) and \( \Delta \bar{y}_{t-1} = \bar{y}_{t-1} - \bar{y}_{t-2} \). Such auxiliary regressors help to take account of cross-section dependence in error terms. The regression can be further augmented with \( \Delta y_{i,t-s} \) and \( \Delta \bar{y}_{t-s} \) for \( s = 1, 2, ... \), to handle possible serial correlation in the errors. The resulting augmented DF statistic is referred to as \( CADF \) statistic for individual \( i \). The panel unit root test statistic is then computed as the average, \( CADF = \frac{1}{N} \sum_{i=1}^{N} CADF_i \). Due to correlation among \( CADF_i \), the limiting distribution of \( CADF \) is non-normal. The \( CADF \) is shown to have good finite sample performance while only a single common factor is present. However, it shows size distortions in case of multiple factors.

Pesaran et al. (2013) extends Pesaran (2007) to the case with multiple common factors. They proposed a new panel unit root test based on a simple average of cross-sectionally augmented Sargan-Bhargava statistics (CSB). The basic idea is similar to CCE of Pesaran (2006), which exploits information of the unobserved factors that are shared by all observed time series. They showed that the limit distribution of the tests is free from nuisance parameters given that the number of factors is no larger than the number of observed cross-section averages. The new test has the advantage that it does not require all the factors to be strong in the sense of Bailey et al. (2012). Monte Carlo simulations show that the proposed tests have the correct size for all combinations of \( N \) and \( T \) considered, with power rising with \( N \) and \( T \).
10.1 Estimating nonstationary factors

Studies in Bai & Ng (2002) and Bai (2003) assume the errors are \( I(0) \). The method of PANIC allows estimating factors under either \( I(1) \) or \( I(0) \) errors. To convey the main idea, consider the case without linear trend (\( \beta_i = 0 \) for all \( i \)). The factor model after differencing is

\[
\Delta x_{it} = \lambda_i' \Delta F_t + \Delta e_{it}.
\]

If \( e_{it} \) is \( I(1) \), \( \Delta e_{it} \) is \( I(0) \). If \( e_{it} \) is \( I(0) \), then \( \Delta e_{it} \) is still \( I(0) \), though over-differenced. Under the assumption of weak cross-section and serial correlation in \( \Delta e_{it} \), consistent estimators for \( \Delta F_t \) can be readily constructed.

It is worth mentioning that when \( e_{it} \) is \( I(0) \), estimating the original level equation already provides a consistent estimator for \( F_t \) (Bai & Ng 2002; Bai 2003). Although such estimators could be more efficient than the ones based on the differenced equations, they are not consistent when \( e_{it} \) is \( I(1) \). An advantage of estimation based on differenced equation is that the factors in levels can still be consistently estimated.

Define \( \hat{F}_t = \sum_{s=2}^{t} \hat{\Delta} F_s \) and \( \hat{e}_{it} = \sum_{s=2}^{t} \hat{\Delta} e_{is} \). Bai & Ng (2004) shows that \( \hat{F}_t \) and \( \hat{e}_{it} \) are consistent for \( F_t \) and \( e_{it} \) respectively. In particular, uniform convergence can be established (up to a location shift factor)\(^2\),

\[
\max_{1 \leq t \leq T} \left\| \hat{F}_t - HF_t + HF_1 \right\| = O_p \left( T^{1/2}N^{-1/2} \right) + O_p \left( T^{-1/4} \right).
\]

Such a result implies that even if each cross-section equation is a spurious regression, the common stochastic trends are well defined and can be consistently estimated, given their existence, a property that is not possible within the framework of fixed-\( N \) time series analysis.

11 Factor Models with Structural Restrictions

It is well known that factor models are only identified up to a rotation. Section 4 discussed three sets of restrictions, called PC1, PC2 and PC3. Each set provides \( r^2 \) restrictions, such that the static factor model is exactly identified. For dynamic factor models (3), Bai & Wang (2014, 2015) show that only \( q^2 \) restrictions are needed to identify the model, where \( q \) is the number of dynamic factors. For example, in order to identify the dynamic factor model (3), we only need to assume that \( var(\varepsilon_t) = I_q \),

\(^2\)The upper bound can be improved (smaller) if one assumes \( \Delta e_{it} \) have higher order finite moments than is assumed in Bai & Ng (2004).
and the $q \times q$ matrix $\Lambda_{01} = [\lambda_{10}, \ldots, \lambda_{q0}]$ is a lower-triangular matrix with strictly positive diagonal elements.

In a number of applications, there might be more restrictions so that the dynamic factor model is over-identified. Bai & Wang (2014) provide general rank conditions for identification linked with $q$ instead of $r$. In this section, we discuss some useful restrictions for both static and dynamic factor models.

### 11.1 Factor Models with Block Structure

The dynamic factor model with a multi-level factor structure has been increasingly applied to study the comovement of economic variables at different levels (see Gregory & Head 1999; Kose et al. 2003; Crucini et al. 2011; Moench et al. 2013, etc.). Such a model imposes a block structure on the factor model so as to attach economic meaning to factors. For example, Kose et al. (2003) and a number of subsequent papers consider a dynamic factor model with a multi-level factor structure to characterize the comovement of international business cycles on the global level, regional level, and country level, respectively. We will use an example with only two levels of factors, a world factor and a country-specific factor, to convey the main idea.

Consider $C$ countries, each having a $n_{c} \times 1$ vector of country variables $X_{ct}$, $t = 1, \ldots, T, c = 1, \ldots, C$. Assume $X_{ct}$ is affected by a world factor $F_{t}^{W}$ and a country factor $F_{ct}^{C}$, $c = 1, \ldots, C$, all factors being latent,

$$X_{ct} = \Lambda_{W}^{c}F_{t}^{W} + \Lambda_{C}^{c}F_{ct}^{C} + \epsilon_{ct}, \quad (16)$$

where $\Lambda_{W}^{c}, \Lambda_{C}^{c}$ are the matrices of factor loadings of country $c$, $\epsilon_{ct}$ is the vector of idiosyncratic error terms for country $c$’s variables. Let $F_{t}^{C}$ be a vector collecting all the country factors. We may assume that the factors follow a VAR specification,

$$\Phi (L) \begin{bmatrix} F_{t}^{W} \\ F_{t}^{C} \end{bmatrix} = \begin{bmatrix} u_{t}^{W} \\ U_{t}^{C} \end{bmatrix}, \quad (17)$$

where the innovation to factors $[u_{t}^{W}, U_{t}^{C}]$ is independent of $\epsilon_{ct}$ at all leads and lags and is i.i.d. normal. Given some sign restriction, this special VAR specification allows one to separately identify the factors at different levels. Wang (2012) and Bai & Wang (2015) provide detailed identification conditions for such models. Under a static factor model setup, Wang (2012) estimates model (16) using an iterated principal components method. The identification conditions therein assume that the world factors are orthogonal to all country factors while country factors can be
correlated with each other. Bai & Wang (2015) directly restrict the innovations of factors in (17) and allow lags of factors to enter equation (16). Such restrictions naturally allow all factors to be correlated with each other. A Bayesian method is then developed to jointly estimate model (16) and (17). Alternative joint estimation method is still not available in the literature, and would be an interesting topic to explore in the future.

11.2 Linear Restrictions on Factor Loadings

In general, there may be overidentifying restrictions in addition to the exact identification conditions (i)-(iii). For example, the multi-level factor model has many zero blocks. Cross-equation restrictions may also be present. Consider the static factor representation (2) for the dynamic factor model (3),

$$X_t = \Lambda F_t + e_t,$$

where $F_t = [f_t, f_{t-1}, ..., f_{t-s}]'$, and $\Lambda = [\Lambda_0, \cdots, \Lambda_s]$. Let $X$ be the $(T - s - 1) \times N$ data matrix, $E$ be the $(T - s - 1) \times N$ matrix of the idiosyncratic errors, $F$ be the $(T - s - 1) \times q(s + 1)$ matrix of the static factors, then we have a matrix representation of the factor model

$$X = FA' + E, \quad \text{or} \quad vec(X) = (I_N \otimes F) \lambda + vec(E),$$

where $\lambda = vec(\Lambda')$. Consider the following restriction on the factor loadings

$$\lambda = B\delta + C,$$

where $\delta$ is a vector of free parameters with $dim(\delta) \leq dim(\lambda)$. In general, $B$ and $C$ are known matrices and vectors that are defined by either identifying restrictions or other structural model restrictions. In view of (19), we may rewrite the restricted factor model (18) as

$$y = Z\delta + vec(E),$$

where $y = vec(X) - (I_N \otimes F) C$ and $Z = [(I_N \otimes F) B]$. If we impose some distributional assumptions on the error terms, for example, $vec(E|Z) \sim \mathcal{N}(0, R \otimes I_{T-s})$ for some $N \times N$ positive definite matrix $R$, such models can be estimated using the Bayesian algorithm from Bai & Wang (2015).
11.3 SVAR and Restricted Dynamic Factor Models

The dynamic factor models also bring new insight into the estimation of structural vector autoregression (SVAR) models with measurement errors. Consider a traditional SVAR given by

\[ A(L)Z_t = \alpha_t, \]

where \( Z_t \) is a \( q \times 1 \) vector of economic variables, and \( \alpha_t \) is the vector of structural shocks. Let

\[ A(L) = A_0 - A_1 L - \cdots - A_p L^p, \]

with \( A_0 \neq I_q \). The reduced form is given by

\[ Z_t = B_1 Z_{t-1} + \cdots + B_p Z_{t-p} + \varepsilon_t, \]

where \( \varepsilon_t = A_0^{-1} \alpha_t \) and \( B_j = A_0^{-1} A_j \). Assume that we do not directly observe \( Z_t \), but observe \( Y_t \)

\[ Y_t = Z_t + \eta_t, \]

where \( \eta_t \) is the \( q \times 1 \) measurement error. In this case, it is difficult to estimate the SVAR model based on \( Y_t \). Assume a large vector of other observed variables \( W_t \) are determined by

\[ W_t = \Gamma_0 Z_t + \cdots + \Gamma_s Z_{t-s} + e_{wt}. \]

Let

\[ X_t = \begin{bmatrix} Y_t \\ W_t \end{bmatrix}, \quad e_t = \begin{bmatrix} \eta_t \\ e_{wt} \end{bmatrix}, \quad f_t = Z_t, \]

\[ \Lambda_0 = \begin{bmatrix} I_q \\ \Gamma_0 \end{bmatrix}, \quad \Lambda_j = \begin{bmatrix} 0 \\ \Gamma_j \end{bmatrix}, \quad j \neq 0. \]

Then we have a structural dynamic factor model

\[ X_t = \Lambda_0 f_t + \cdots + \Lambda_s f_{t-s} + e_t, \quad (20) \]

\[ f_t = B_1 f_{t-1} + \cdots + B_p f_{t-p} + \varepsilon_t. \]

According to Bai & Wang (2015), because the upper \( q \times q \) block of \( \Lambda_0 \) is an identity matrix,

\[ \Lambda_0 = \begin{bmatrix} I_q \\ * \end{bmatrix}, \]

model (20) is identified and can be analyzed using a Bayesian approach. In particular, without further assumptions, we are able to estimate \( f_t = Z_t, B(L), \Lambda_i, \) and
$E(\varepsilon_t\varepsilon_t') = A_0^{-1}(A_0')^{-1}$. We may also incorporate additional structural restrictions (such as long-run restrictions) as in standard SVAR analysis to identify $A_0$. The impulse responses to structural shocks can be obtained as $\partial Y_{t+k}/\partial a_t = \partial Z_{t+k}/\partial a_t = \partial f_{t+k}/\partial a_t$ and

$$\frac{\partial W_{t+k}}{\partial a_t} = \begin{cases} 
\Gamma_0 \frac{\partial f_{t+k}}{\partial a_t} + \cdots + \Gamma_s \frac{\partial f_{t+s}}{\partial a_t}, & k \geq s, \\
\Gamma_0 \frac{\partial f_{t+k}}{\partial a_t} + \cdots + \Gamma_k \frac{\partial f_{t+k}}{\partial a_t}, & k < s,
\end{cases}$$

where the partial derivative is taken for each component of $a_t$ when $a_t$ is a vector.

## 12 High Dimensional Covariance Estimation

The variance-covariance matrix plays a key role in the inferential theories of high-dimensional factor models as well as in various applications in finance and economics. Using an observed factor model of large dimensions, Fan et al. (2008) examined the impacts of covariance matrix estimation on optimal portfolio allocation and portfolio risk assessment. Fan et al. (2011) further studied the case with unobserved factors. By assuming sparse error covariance matrix, they allow cross-sectional correlation in errors in the sense of an approximate factor model. An adaptive thresholding technique is employed to take into account the fact that the idiosyncratic components are unobserved. Consider the vector representation of the factor model (2),

$$X_t = \Lambda F_t + \epsilon_t,$$

which implies the following covariance structure

$$\Sigma_X = \Lambda \text{cov} (F_t) \Lambda' + \Sigma_e,$$

where $\Sigma_X$ and $\Sigma_e = (\sigma_{ij})_{N \times N}$ are covariance matrices of $X_t$ and $\epsilon_t$ respectively. Assume $\Sigma_e$ is sparse instead of diagonal, and define

$$m_T = \max_{i \leq N} \sum_{j \leq N} 1(\sigma_{ij} \neq 0).$$

The sparsity assumption puts an upper bound assumption on $m_T$ in the sense that

$$m_T^2 = o \left( \frac{T}{r^2 \log(N)} \right).$$

In this formulation, the number of factors $r$ is allowed to be large and grows with $T$. Using principal components estimators under the normalization $\frac{1}{T} \sum_{t=1}^T F_t F_t' = I_r$, 

30
the sample covariance of $X_t$ can be decomposed as

$$S_X = \hat{\Lambda} \hat{\Lambda}' + \sum_{i=r+1}^{N} \hat{\mu}_i \hat{\xi}_i \hat{\xi}_i',$$

where $\hat{\mu}_i$ and $\hat{\xi}_i$ are the $i$-th leading eigenvalues and eigenvectors of $S_X$ respectively. In the high dimensional setup, the sample covariance might be singular and provides a poor estimator for the population covariance. For example, when $N > T$, the rank of $S_X$ can never exceed $T$ while the theoretical covariance $\Sigma_X$ always has rank $N$. To overcome this problem, we may apply the thresholding technique to the component $\sum_{i=r+1}^{N} \hat{\mu}_i \hat{\xi}_i \hat{\xi}_i'$, which yields a consistent estimator of $\Sigma_e$, namely $\hat{\Sigma}_e$. Finally, the estimator for $\Sigma_X$ is defined as

$$\hat{\Sigma}_X = \hat{\Lambda} \hat{\Lambda}' + \hat{\Sigma}_e,$$

which is always of full rank and can be shown to be a consistent estimator for $\Sigma_X$.

The adaptive thresholding technique is easy to implement. Denote $\sum_{i=r+1}^{N} \hat{\mu}_i \hat{\xi}_i \hat{\xi}_i' = (\hat{\sigma}_{ij})_{N \times N}$ and $\hat{e}_{it} = x_{it} - \hat{\lambda}_i \hat{F}_t$. Define

$$\hat{\theta}_{ij} = \frac{1}{T} \sum_{t=1}^{T} (\hat{e}_{it} \hat{e}_{jt} - \hat{\sigma}_{ij})^2,$$

$$s_{ij} = \hat{\sigma}_{ij} 1 \left( |\hat{\sigma}_{ij}| \geq \sqrt{\hat{\theta}_{ij} \omega_T} \right),$$

where $\omega_T = C r \sqrt{\frac{\log N}{T}}$ for some positive constant $C$. In practice, alternative values of $C$ can be assumed to check the robustness of the outcome.

The estimated $\hat{\Sigma}_e$ can be used to obtain more efficient estimation of $\Lambda$ and $F$ based on the generalized principal components method (Bai & Liao, 2013). Alternatively, the unknown parameters of $\Lambda$ and $\Sigma_e$ can be jointly estimated by the maximum likelihood method (Bai & Liao, 2012). The resulting estimator $\hat{\Lambda} \hat{\Lambda}' + \hat{\Sigma}_e$ from the MLE is a direct estimator for the high dimensional covariance matrix. A survey on high dimensional covariance estimation is given by Bai & Shi (2011).

### 13 Bayesian Method to Large Factor Models

With the fast growing computing power, Markov Chain Monte Carlo (MCMC) methods are increasingly used in estimating large-scale models. Bayesian estimation method, facilitated by MCMC techniques, naturally incorporates identification
restrictions such as those from the structural VAR into the estimation of large factor model. The statistical inference on impulse responses can be readily constructed from the Bayesian estimation outcome.

The Bayesian approach has been considered by Kose et al. (2003), Bernanke et al. (2005), Del Negro & Otrok (2008), Crucini et al. (2011), and Moench et al. (2013), among others. The basic idea is to formulate the dynamic factor model as a state space system with structural restrictions. Initially, we setup the priors for factor loadings and the VAR parameters for factors. We also specify the prior distribution of factors for initial periods. Then Carter & Kohn (1994)'s multimove Gibbs-sampling algorithm can be adopted for estimation. One of the key steps is to use Kalman filter or other filters to form a conditional forecast for the latent factors. The main advantage of the Bayesian method is that researchers can incorporate prior information into the estimation of the large factor model, and the outcome is the joint distribution of both model parameters and latent factors. Inference for impulse responses is an easy by-product of the procedure. The computational intensity is largely related to the number of dynamic factors which is small, but only slightly affected by the dimension of the data.

The Bayesian approach can naturally incorporate structural restrictions. For example, consider the dynamic factor model with linear restrictions on the factor loadings, such as (18) and (19). We may rewrite the restricted factor model (18) as

$$y = Z\delta + \text{vec}(E), \quad \text{vec}(E|Z) \sim N(0, R \otimes I_{T-s}),$$

where $y = \text{vec}(X) - (I_N \otimes F)C$ and $Z = [(I_N \otimes F)B]$. Impose the Jeffreys prior for $\delta$ and $R$:

$$p(\delta, R) \propto |R|^{-(N+1)/2},$$

which implies the following conditional posterior:

$$\delta|R, X, F \sim N\left( B^\prime \left( B^\prime B \right)^{-1} \left( \text{vec} \left( \hat{\Lambda}^\prime \right) - C \right), \left( B^\prime \left( R^{-1} \otimes F^\prime F \right) B \right)^{-1} \right), \quad (21)$$

with $\hat{\Lambda} = X^\prime F (F^\prime F)^{-1}$. Thus we may draw $\delta$ according to (21) and construct the associated loading matrix $\Lambda(\delta)$. In the meantime, we may draw $R$ according to an inverse-Wishart distribution

$$R|X, F \sim \text{invWishart}(S, T-s + N + 1 - q(s+1)),$$
where \( S = (X - F\hat{\Lambda}')' (X - F\hat{\Lambda}') \).

There are still some challenges to the Bayesian approach. First, sometimes there is little guide as to how to choose the prior distribution. Bai & Wang (2015) employed the Jeffreys priors to account for the lack of a priori information about model parameters. It remains an open question how theoretical properties of the posterior distribution are affected by alternative choice of priors. Second, usually the number of restrictions for identification is small and fixed, while the number of parameters grows with sample size in both dimensions. This might lead to weak identification and poor inference. Some shrinkage method, such as the application of Minnesota-type priors, might help to mitigate such problems. One might also incorporate over-identifying restrictions to improve estimation efficiency. Third, model selection for the large factor model using Bayes factors is generally computationally intensive. Some simple-to-compute alternative model selection methods such as Li et al. (2014) might be considered.

14 Concluding Remarks

This review provides an introduction to some recent development in the theory and applications of large factor models. There are still lots of open and interesting issues which await future research. For example, almost all current studies focus on linear factor models and rely on information from covariance matrix for estimation. Introducing nonlinearities into large factor model could be relevant to a number of potential applications. Freyberger (2012) introduced interactive fixed effects into nonlinear panel regression and identified important differences between linear and nonlinear regression results. Su et al. (2015) developed a consistent nonparametric test for linearity versus nonlinear models in the presence of interactive effects. Similar area has been largely unexplored and could be potential future research topics. Other examples include theory on discrete choice models with factor error structure, quantile regression with interactive fixed effects (Ando & Tsay 2011; Harding & Lamarche 2014), factors from higher moments of the data, nonlinear functions of factors. In terms of estimation, one may also study alternative methods, such as the adaptive thresholding techniques of Fan et al. (2011), which is helpful in situations with many zero factor loadings. For the FAVAR model, one may consider one-step maximum likelihood estimators and compare their properties with the two-step estimators of
Bai et al. (2015). General inferential theory for large factor models with structural restrictions especially over-identification is also an important area to explore, which may help estimate macroeconomic models.

References


Ando T, Bai J. 2015. A simple new test for slope homogeneity in panel data models with interactive effects. Available at SSRN 2541467


Bai J, Liao Y. 2013. Statistical Inferences Using Large Estimated Covariances for Panel Data and Factor Models. Available at SSRN 2353396


Boivin J, Giannoni M. 2006. DSGE models in a data-rich environment. NBER working paper 12772


Freyberger J. 2012. Nonparametric panel data models with interactive fixed effects. Working paper


38


Li H, Li Q, Shi Y. 2013. Determining the number of factors when the number of factors can increase with sample size. Department of Economics, Texas A&M


Li Y, Zeng T, Yu J. 2013. Robust deviance information criterion for latent variable models. Working paper


Moon H, Weidner M. 2014. Dynamic linear panel regression models with interactive fixed effects. Working paper

Moon H, Weidner M. 2015. Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica* Forthcoming


