Mathematical Statistics
Stockholm University

Extended Factor-Augmented Vector Autoregression: macroeconomic forecasting with the Lasso

Ying Pang

Research Report 2016:8

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se

Mathematical Statistics
Stockholm University
Research Report **2016:8**,
http://www.math.su.se

Extended Factor-Augmented Vector Autoregression: macroeconomic forecasting with the Lasso

Ying Pang

April 2016

Abstract

This paper considers macroeconomic forecasting with a large number of predictor variables. We propose an extended factor-augmented vector autoregressive model (EFAVAR), that describes the joint dynamics of macro variables, latent factors and high-dimensional predictors. Furthermore, We employ a factor model to obtain a small number of principal-component-based factor estimates which can represent common movement of informational data. Meanwhile, we utilize the least absolute shrinkage and selection operator (Lasso) to select a few of the most relevant observed predictors, which can capture idiosyncratic fluctuation of data. Then, the multi-step-ahead forecasts can be constructed using a handful of estimated factors and selected predictors. In addition, we investigate the consistency of Lasso estimate and forecasting accuracy in the theoretical perspective. Also, we examine the predictive performance by a small Monte Carlo study and an empirical analysis, and conclude that EFAVAR shows some improvements in comparison to other model candidates.

Keywords: Macroeconomic Time Series, Factor-augmented Forecasting, Principal Component, Large Vector Autoregression, Lasso, Predictive Performance

Extended Factor-Augmented Vector Autoregression macroeconomic forecasting with the Lasso

Ying Pang [*]

1 Introduction

1.1 Background

Over past decades, high-dimensional data has drawn more and more attention in the field of economics and finance. Usually, high-dimensional data comes with complex structures and features, and it is quite different from traditional data where the sample size is greater than the dimension of variables. For instance, a macroeconomic data set can include literally hundreds of time series, in which variables are non-stationary over time, cyclically-fluctuating, serially correlated and cross-sectionally dependent. Moreover, observations are usually grouped or clustered, which implies that the heterogeneity might be exhibited. On the methodology side, when a large number of parameters need to be estimated simultaneously, it could result in estimation errors accumulating. Thus, the noise accumulation is also considered as a main feature of high-dimensional data, which can be overcome by a sparse model and model selection (Donoho et al. (2000); Bühlmann and Van De Geer (2011)). Furthermore, when the dimension of data is high, there might be spurious correlation between variables. In other words, uncorrelated variables might have high sample correlations, which would lead to the false inference. Additionally, economic and financial data sets are collected from various sources, and observations are recorded and reported at different time points over a span of many years. And, that possibly causes the experimental biases, statistical instability and heavy computational costs.

There are many applications for investigating the effects of monetary policy, business cycles, portfolio management and so forth (Sims (1980);

[*] Department of Mathematics, Stockholm University, SE-106 91 Stockholm, Sweden, ypang@math.su.se.

Bernanke and Blinder (1992); Leeper et al. (1996); Christiano et al. (2001)). In addition, many researchers, economists and policy makers are more concerned about gaining insight into macroeconomic data and developing effective methods for prediction. And, it is without doubt that most classical models and standard approaches, such as ordinary least squares (OLS), perform poorly with the increase of dimensions. Therefore, numerous challenges to the statistical theory and methodology has been posed, when the number of parameters involved is of a much larger magnitude than the sample size.

In terms of statistical accuracy and computational efficiency, it is of crucial importance to reduce the dimension of data before going further. And, there are two main approaches for dimension reduction. One approach assumes that high-dimensional data is characterized by some common features that can be represented by a handful of latent variables/factors. This is often relative to the factor models, together with principal component analysis which plays a vital role in the dimension reduction techniques. For instance, the factor models are widely developed and implemented in order to analyze macroeconomic data (Connor and Korajczyk (1993); Forni et al. (2000); Bai and Ng (2002); Stock and Watson (2001, 2002a,b, 2005); Forni et al. (2003); Bernanke et al. (2004); Belviso and Milani (2006)). As an alternative approach, dimension reduction can be achieved by selecting a small subset of variables from high-dimensional set. This requires the usage of regularization theory and shrinkage methods, in which Lasso is extensively employed and improved, in order to successfully select a few of important predictor variables and enhance the prediction accuracy in computationally effective ways (Zou (2006); Song and Bickel (2011); Kock and Callot (2015)).

1.2 Motivation

The vector autoregression (VAR) is one of the central pillars for multivariate analysis, and Bernanke et al. (2004) propose the factor-augmented VAR (FAVAR), that describes the joint dynamics of several macro variables Y_t and latent factors F_t as follows,

$$\begin{bmatrix} F_t \\ Y_t \end{bmatrix} = \Phi(L) \begin{bmatrix} F_{t-1} \\ y_{t-1} \end{bmatrix} + v_t \ ,$$

where the matrix $\Phi(L)$ is a matrix conformable polynomial of lag operator L, and v_t is a vector of errors with zero mean and finite covariance. In addition, Bernanke et al. (2004) suggest that a small number of factors can be obtained throughout the relation,

$$X_t = \Lambda^f F_t + \Lambda^y Y_t + e_t \ , \tag{1}$$

where X_t is a vector of high-dimensional time series that are relevant to forecasting, Λ^f is a matrix of factor loadings, and e_t is a vector of errors.

We agree with that the a large amount of the variation of data X_t can be satisfactorily represented by a few factors, however, we believe that there should be some information that cannot be captured by common factors, which is important and valuable for explaining and/or forecasting macro variables. And, it can be interpreted as idiosyncrasy or individual fluctuation affected by a small number of variables in X_t. Therefore, in order to avoid leaving out potentially useful information, we consider an extended FAVAR (EFAVAR), which models the joint dynamics of macro variable, latent factors, and informational variables over a common time period.

Unfortunately, it is unknown which of the variables in X_t are relevant, thus, there is no way to determine a small subset of variables in advance. In other words, involving a large number of observables X_t has embodied the high-dimensionality into the EFAVAR, which makes the standard VAR difficult to employ[1]. Furthermore, because we assume that there is a small number of observables in X_t relevant for prediction, it indicates the sparse pattern of coefficient matrix for X_t in the EFAVAR. Consequently, we adopt Lasso to select relevant variables by shrinking the coefficients of irrelevant variables to be exact zeros. Simultaneously, the resulting nonzero values are the coefficient estimates of retained variables in X_t.

1.3 Goals and contributions

We have two goals in this paper. The first is to investigate the consistency of Lasso estimate and forecasting accuracy in our context. which the forecasts can be constructed using principal-component-based factor estimated as augmented predictors based on the EFAVAR. The second goal is to examine the resulting predictive performance. Based on an empirical analysis, the EFAVAR can make some contributions for improving predictive performance in comparison to the methods proposed by Stock and Watson (2002a) and Song and Bickel (2011). However, the extent of improvement differs about the forecasting time horizon and the variable to be forecast of our interest. Besides this, a small Monte Carlo experiment provides the evidence for theoretical results with finite samples.

[1] The VAR model rarely employs more than six to eight variables (Bernanke et al. (2004)). Moreover, Leeper et al. (1996) apply Bayesian priors to increase the number of variables included; however, VAR models still cannot contain more than twenty variables.

1.4 Layout

The rest of the paper is organized as follows. In section 2, we describe the EFAVAR with model assumptions, estimation procedures regarding factors and coefficients, and theoretical properties of resulting estimates. Additionally, section 3 examines the predictive performance of EFAVAR using a Monte Carlo experiment and an empirical study for macroeconomic forecasting, Moreover, section 4 concludes. Besides, the appendix provides proofs and more results of both Monte Carlo and empirical studies.

2 Economic framework

2.1 Models

Let y_t be a scalar macro variable assumed to have pervasive effects throughout the economy, and X_t be an N-dimensional vector of informational variables for $t = 1, \dots, T$. Then, the EFAVAR models the joint dynamics of y_t, F_t and X_t through the following

$$\begin{bmatrix} F_t \\ X_t \\ y_t \end{bmatrix} = \Pi(L) \begin{bmatrix} F_{t-1} \\ X_{t-1} \\ y_{t-1} \end{bmatrix} + \zeta_t \ , \tag{2}$$

where

$$\Pi(L) = \begin{bmatrix} \Theta & 0 & 0 \\ \Lambda\Theta & \Gamma(L) & 0 \\ A' & B^x(L) & B^y(L) \end{bmatrix} \quad \text{and} \quad \zeta_t = \begin{bmatrix} \eta_t \\ \mu_t \\ \varepsilon_t \end{bmatrix}.$$

F_t is an r-dimensional vector of static factors, which means that there is no lag of factor involved in the model. And, ζ_t is a vector of idiosyncratic terms with conditional mean zero, that is $E(\zeta_t | F_{t-1}, X_{t-1}, y_{t-1}, F_{t-2}, X_{t-2}, y_{t-2} \dots) = 0$, meanwhile, η_t, μ_t and ε_t are mutually independent. In addition, the parameter Θ is a $r \times r$ matrix of VAR(1) coefficients of factors, Λ is a $N \times r$ matrix and A is a $r \times 1$ vector. Furthermore, $\Gamma(L)$ is a matrix lag polynomial with finite power, that is specifically defined by

$$\Gamma(L) = \begin{bmatrix} \Gamma_{11}(L) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Gamma_{NN}(L) \end{bmatrix},$$

where $\Gamma_{ii}(L) = \sum_{j=1}^{q} \gamma_{ij} L^{j-1}$, in which the AR coefficient γ_{ij} satisfying that $|\gamma_{ij}| < 1$ for all i and j. Furthermore, assuming that both lag polynomials have a same finite order p, we can write that $B^x(L) = \sum_{i=1}^{p} B_i^x L^{i-1}$ and

$B^y(L) = \sum_{i=1}^{p} B_i^y L^{i-1}$, where B_i^x is a row vector of coefficients for X_{t-i+1} and B_i^y is a coefficient of y_{t-i+1} for $i = 1, \ldots, p$.

Now, we decompose model (2) into two parts. One provides the foundation to estimate factors, which is considered as a VAR form of dynamic factor model (DFM) (Stock and Watson (2005)),

$$\begin{bmatrix} F_t \\ X_t \end{bmatrix} = \begin{bmatrix} \Theta & 0 \\ \Lambda\Theta & \Gamma(L) \end{bmatrix} \begin{bmatrix} F_{t-1} \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} \eta_t \\ \mu_t \end{bmatrix} . \tag{3}$$

Meanwhile, the other delivers a preliminary forecast model,

$$y_t = A' F_{t-1} + B^x(L) X_{t-1} + B^y(L) y_{t-1} + \varepsilon_t \ . \tag{4}$$

Next, in order to pose model assumptions and derive consistency of estimates, we will represent equations (3) and (4) in more concise ways.

2.1.1 Factor models

Before rewriting model (3), we would like to explain how it reflects our initial thoughts with more details. Suppose that the process of dynamic factors f_t can be described by

$$f_t = \tilde{\Theta}(L) f_{t-1} + \tilde{\eta}_t \ , \tag{5}$$

where $\tilde{\Theta}(L)$ is a matrix lag polynomial, and $\tilde{\eta}_t$ is a vector of idiosyncratic terms. Moreover, dynamic factors refer to the fact that lags of factors appear in the model.

Originally, we consider the complex structure of informational variables X_t, which can be expressed by

$$X_t = \tilde{\Lambda}(L) f_t + \Gamma(L) X_{t-1} + \epsilon_t \ , \tag{6}$$

where f_t is an $\tilde{r}$-dimensional vector of dynamic factors, and both $\tilde{\Lambda}(L)$ and $\Gamma(L)$ are matrix lag polynomials. According to model (6), the commonality is represented by the distributed lags of a handful of f_t, a part of dependence is explained by the linear interdependencies among X_t and the past values, and the idiosyncratic disturbances is expressed by ϵ_t, in which there possibly exists a cross-sectional correlation. Additionally, assume that f_t and ϵ_t are mutually uncorrelated at all leads and lags, that is $E(f_{it} u_{js}) = 0$ for all i, j, t, s.

Forni et al. (2000) derive dynamic factor estimation based on frequency domain using two-sided filtering, which cannot be directly utilized for predictions. Therefore, we adapt the static factors and obtain factor estimates in time domain, which are applicable to be used in forecasting. Suppose that $\tilde{\Lambda}(L)$ has a finite lags order l, and let $F_t = (f_t', f_{t-1}', \ldots, f_{t-l+1}')'$ or its subset if not all f_t comes with l lags. In addition, the dimension of F_t is then bounded by $\tilde{r} \le r \le \tilde{r}l$. Next, we rewrite model (6) in a static form,

$$X_t = \Lambda F_t + \Gamma(L) X_{t-1} + \epsilon_t \ , \tag{7}$$

where the ith row of Λ is composed by coefficients of $\tilde{\Lambda}_i(L)$ and zeros. We actually prefer estimating factors based on model (7), because F_t and X_t are presented concurrently.

It is not difficult to realize that we can obtain model (7) by combing two sub equations of VAR (3), under the condition $\mu_t = \Lambda\eta_t + \epsilon_t$. Moreover, when $\tilde{\Theta}(L)$ also has a finite order l, equation (5) can be rewritten as

$$F_t = \Theta F_{t-1} + \eta_t \ , \tag{8}$$

where Θ incorporates coefficients of $\tilde{\Theta}(L)$ and zeros, and $\eta_t = H\tilde{\eta}_t$ in which H is a $r \times \tilde{r}$ matrix. Furthermore, note that equation (8) is the exact expression of F_t from VAR (3).

2.1.2 Forecasting models

Define $\omega_t = (X_t', y_t)'$ and $W_t = (\omega_t', \omega_{t-1}', \ldots, \omega_{t-p+1}')'$. Then, equation (4) can be rewritten as

$$y_t = A' F_{t-1} + B' W_{t-1} + \varepsilon_t \ , \tag{9}$$

where $B = (B_1', B_2', \ldots, B_p')'$ in which $B_i = (B_i^x, B_i^y)'$ for $i = 1, \ldots, p$. As the coefficient matrix of observables, B is considered to be sparse. In other words, a small number of nonzero values are coefficients for variables that should be retained in the model. Alternatively, a more concise form of model (9) can be expressed by

$$y_t = \Omega' Z_{t-1} + \varepsilon_t \ , \tag{10}$$

where $\Omega = (A', B')'$ and Z_t is an $r + (N+1)p$ dimensional vector of all predictor variables, $Z_t = (F_t', W_t')'$.

Moreover, suppose that the data is available up to time T, and write $y = (y_T, y_{T-1}, \ldots)'$, $F = (F_{T-1}, F_{T-2}, \ldots)'$ and $W = (W_{T-1}, W_{T-2}, \ldots)'$. Then, the stacked form of model (9) and (10) are given by

$$y = FA + WB + \varepsilon \ , \tag{11}$$

and

$$y = Z\Omega + \varepsilon \ , \tag{12}$$

respectively, where $\varepsilon = (\varepsilon_T, \varepsilon_{T-1}, \ldots)'$ and $Z = (F, W)'$.

To be clear, the static factor model (7) and forecast model (10) serve as the foundation to pose assumptions of models and parameters, as well as derive consistent factor and forecast estimations. Moreover, both models (11) and (12) provide the basis to examine theoretical properties of Lasso estimation and prediction errors when T and N increase.

2.2 Assumptions

Before making assumptions and deriving estimates, we introduce some notations. Let $\| \bullet \|_1$ and $\| \bullet \|$ be L_1 and L_2 matrix norm[2], respectively. Denote $J_i = J(B_i) = \{j : B_{ji} \neq 0\}$ as the set of indices of nonzero elements in B_i, and $s_i = |J_i|$ as the cardinality of J_i for $i = 1, \ldots, p$. Furthermore, let $s = \sum_{i=1}^{p} s_i$ and $J = J(B) = \bigcup_{i=1}^{p} J_i \subseteq \{1, \ldots, (N+1)p\}$, which J has the cardinality s at most. Next, define $\sigma_{i,y}$ as the variance of Z_{it} and σ_ε as the variance of ε_t. Let $\sigma_T = \max\{\sigma_Z, \sigma_\varepsilon\}$, where σ_Z is a supreme of the set containing $\sigma_{i,y}$. In addition, let $\Psi_W = W'W/T$ and $\Psi_Z = Z'Z/T$ be scaled Gramian matrix of W and Z, respectively.

Some classical model assumptions have to be modified for high dimensional framework. Supposing that $N, T \to \infty$ jointly, the following assumptions are made according to Stock and Watson (2002a).

Assumptions 1 (A.1)

(a) $E(\epsilon_t' \epsilon_{t+u}/N) = \gamma_{Nt}(u)$, and $\lim_{N\to\infty} \sup_t \sum_{u=-\infty}^{\infty} |\gamma_{Nt}(u)| < \infty$;

(b) $E(\epsilon_{it}\epsilon_{jt}) = \tau_{ijt}$, and $\lim_{N\to\infty} \sup_t N^{-1} \sum_{i=1}^{N} \sum_{j=1}^{N} |\tau_{ijt}| < \infty$;

(c) $\lim_{N\to\infty} \sup_{t,s} N^{-1} \sum_{i=1}^{N} \sum_{j=1}^{N} |\text{cov}(\epsilon_{is}\epsilon_{it}, \epsilon_{js}\epsilon_{jt})| < \infty$.

Assumptions 2 (A.2)

(a) $E(F_t F_t') = \Sigma^F$ which is a diagonal matrix of entries $\Sigma_{ii}^F > \Sigma_{jj}^F > 0$ for $i < j$;

(b) $T^{-1} \sum_t F_t F_t' \xrightarrow{p} \Sigma^F$;

(c) $\Lambda' \Lambda / N \to I_r$ where I_r is a $r \times r$ identity matrix;

(d) $|\bar{\Lambda}| < \infty$ where $\bar{\Lambda} = \max_{i,j} \{\Lambda_{ij}\}$.

[2] For a $m \times n$ matrix $\mathbf{x}$, L_1 and L_2 matrix norms are defined as $\|\mathbf{x}\|_1 = \sum_{i=1}^{m} \sum_{j=1}^{n} |x_{ij}|$ and $\|\mathbf{x}\| = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij}^2\right)^{1/2}$. The special case is that $\mathbf{x}$ is a vector where $n = 1$. Then obtain $\|\mathbf{x}\|_1 = \sum_{i=1}^{m} |x_i|$ and $\|\mathbf{x}\| = \left(\sum_{i=1}^{m} x_i^2\right)^{1/2}$.

Assumptions 3 (A.3)

(a) $T^{-1}\sum_t \varepsilon_t^2 \xrightarrow{p} \sigma^{\varepsilon}$;

(b) $T^{-1}\sum_t Z_t \varepsilon_{t+h} \xrightarrow{p} 0$;

(c) $\|\Omega\|_1 < \infty$.

A.1 pose restrictions on the moments of ϵ_t in factor model (7), in which *A.1(a)* and *A.1(b)* allow for weakly serial and cross-sectional correlation among ϵ_t, respectively. And, *A.1(c)* limits the size of four moments for ϵ_t processes, instead of normality assumption. Furthermore, *A.2* restricts the factors F_t and factor loadings Λ, which makes sure factors can be identified. In addition, *A.3* is useful for showing the consistency of the Lasso estimates and resulting forecast errors, when unobserved factors are replaced by factor estimates in the regression.

2.3 Estimates

Generally speaking, we consider a stepwise approach. Firstly, factor estimates are obtained using principal components based on factor model, which we shall substitute for true factors as regressors. Secondly, forecast estimates can be constructed using Lasso estimates obtained by regressing the variable to be forecast onto estimated factors and observed predictors.

2.3.1 Factor estimates and consistency

Suppose that both *A.1* and *A.2* hold. We consider a least squares approach where the estimates of F and Λ can solve the following

$$\min_{F_1,\dots,F_T,\Lambda,\Gamma(L)} (NT)^{-1}\sum_{t=1}^{T}[(I-\Gamma(L)L)X_t - \Lambda F_t)]'[(I-\Gamma(L)L)X_t - \Lambda F_t] \quad . \tag{13}$$

The minimization of (13) can be conveniently achieved using the following iteration. As long as the mean squares errors (MSE) in the target function monotonically decreases, a (local) minimum of (13) can be approached (Stock and Watson (2005)). Let $\hat{r}$ be the number of estimated factors $\widehat{F}_t$, and $\hat{q}$ be the lags order (at most) to $\widehat{\Gamma}(L)$. The iteration is to estimate factors, which shall be obtained by a $T \times \hat{r}$ matrix $\widehat{F} = (\widehat{F}_1,\dots,\widehat{F}_T)'$, based on data $\{X_{it}\}_{i=1,t=1}^{N,T}$. Moreover, the algorithm can be described as follows

1. Let $\text{MSE}^{[0]} = 0$ and $\widehat{\Gamma}^{[0]}(L) = 0$ with diagonal lag polynomial $\widehat{\Gamma}_{ii}(L) = 0$ for $i = 1,\dots,N$.

2. Iterate for $j = 1, 2, \ldots$
   (a) Produce $\widetilde{X}^{[j]}$ in which the entry $\widetilde{X}_{it}^{[j]} = X_{it} - \widehat{\Gamma}_{ii}^{[j-1]}(L)X_{it-1}$ for $i = 1, \ldots, N$ and $t = 1, \ldots, T$, and standardize it to let each column (variable) have zero mean and unit standard deviation.
   (b) Obtain $\widehat{F}^{[j]} = \sqrt{T}V^{[j]}$, where $V^{[j]}$ consists of eigenvectors corresponding to the $\hat{r}$ largest eigenvalues of matrix $\widetilde{X}^{[j]}\widetilde{X}^{[j]\prime}/NT$.
   (c) For each i, produce the ith row of estimated factor loadings $\widehat{\Lambda}_i^{[j]}$ and coefficient $\widehat{\Gamma}_{ii}^{[j]}(L)$ by linearly regressing $\widetilde{X}_{it}^{[j]}$ onto $\widehat{F}_t^{[j]}$ and lags of $\widetilde{X}_{it}^{[j]}$ for all available t.
   (d) Compute resulting residuals $\hat{e}_{it}^{[j]} = X_{it}^{[j]} - \widehat{\Lambda}_i^{[j]}\widehat{F}_t^{[j]} - \widehat{\Gamma}_{ii}^{[j]}(L)\widetilde{X}_{it-1}^{[j]}$ for all i and t, and calculate $\text{MSE}^{[j]} = \sum_{i=1}^{N}\sum_{t=1}^{T}\hat{e}_{it}^{[j]2}/NT$ and $\text{Dev}^{[j]} = |\text{MSE}^{[j]} - \text{MSE}^{[j-1]}|$.
   (e) Given a criteria level CL, if $\text{Dev}^{[j]} \leq CL$, break the loop and jump to step 3; otherwise, let $j = j + 1$ and continue to iterate.
3. Obtain factor estimates $\widehat{F} = \widehat{F}^{[j]}$.

Now, we state the properties of factor estimates $\widehat{F}$ that solve the minimization (13) using the iterative algorithm, which is adapted from Stock and Watson (2002a).

Theorem 1. Suppose that *A.1* and *A.2* hold. Let r and $\hat{r}$ be the number of true factors F_t and estimated factors $\widehat{F}_t$. As $N, T \to \infty$, S_j can be determined so that the followings hold for $t = 1, \ldots, T$, where S_j is a sign variable with a value of either $+1$ or -1,

(a) For $j = 1, 2, \ldots, r$, $S_j\widehat{F}_{jt} \xrightarrow{p} F_{jt}$;

(b) For $j = 1, 2, \ldots, r$, $T^{-1}\sum_{t=1}^{T}(S_j\widehat{F}_{jt} - F_{jt})^2 \xrightarrow{p} 0$;

(c) For $j = r + 1, \ldots, \hat{r}$, $T^{-1}\sum_{t=1}^{T}\widehat{F}_{jt}^2 \xrightarrow{p} 0$.

2.3.2 Forecast estimates and accuracy

Given that observations are available up to time T, the one-step-ahead out-of-sample forecast is defined by $y_{T+1|T}$. In theory, an optimal forecast based on equation (10) is the expected value,

$$\begin{aligned} y_{T+1|T} &= E(y_{T+1}|Z_T, Z_{T-1}\ldots) \\ &= E(\Omega' Z_T + \varepsilon_{T+1}|Z_T, Z_{T-1}\ldots) \\ &= \Omega' Z_T + E(\varepsilon_{T+1}|Z_T, Z_{T-1}\ldots) \\ &= \Omega' Z_T \ , \end{aligned} \tag{14}$$

where the last equality holds under the assumption that ζ_t has conditional mean zero, that is $E(\zeta_t|F_{t-1}, X_{t-1}, y_{t-1}, F_{t-2}, X_{t-2}, y_{t-2}\ldots) = 0$. However, (14) cannot be produced, simply because of involving unobserved factors. Instead, a feasible one-step-ahead forecast can be constructed by

$$\hat{y}_{T+1|T} = \widehat{\Omega}'\widehat{Z}_T \ , \tag{15}$$

where $\widehat{\Omega}$ can be obtained by fitting model (10) in which Z_t is replaced by $\widehat{Z}_t = (\widehat{F}_t', W_t')'$ for $t = 1, \ldots, T$.

When N is small and fixed, $\widehat{\Omega}$ is the OLS estimate, and we know that

$$\hat{y}^{\text{ols}}_{T+1|T} - y_{T+1|T} \xrightarrow{p} 0 \ , \tag{16}$$

where $\hat{y}^{\text{ols}}_{T+1|T} = \widehat{\Omega}'_{\text{ols}}\widehat{Z}_T$ according to Stock and Watson (2002a)[3]. In other words, the forecast estimate $\hat{y}^{\text{ols}}_{T+1|T}$ converges to the optimal infeasible forecast $y_{T+1|T}$ of (14) in probability as $T\rightarrow\infty$.

Consider that N is large, where $N > T$ or even $N \gg T$. Under the sparsity assumption, the matrix ψ_W is singular. In other words, the minimal eigenvalue of Ψ_W is zero, which, for any vector Δ, we have that

$$\min\left\{\frac{\Delta'\Psi_W\Delta}{\|\Delta\|^2} : \Delta \in \mathbb{R}^{(N+1)p}\backslash\{0\}\right\} = 0 \ . \tag{17}$$

However, the OLS requires a positive definite Gram matrix, in which all eigenvalues are positive,

$$\min\left\{\frac{\Delta'\Psi_W\Delta}{\|\Delta\|^2} : \Delta \in \mathbb{R}^{(N+1)p}\backslash\{0\}\right\} > 0 \ . \tag{18}$$

Therefore, the OLS method does not work in this case.

As well known, the Lasso is considered as one of the computationally effective methods to deal with sparse models, which asks for a very weak assumption on the Gram matrix (Bickel et al. (2009)). In order to guarantee the nice statistical properties of Lasso, we use the following condition.

Restricted Eigenvalues of the Gram matrix

$$\kappa^2_W(d) = \min\left\{\frac{\Delta'\Psi_W\Delta}{\|\Delta_D\|^2} : \Delta \in \mathbb{R}^{(N+1)p}\backslash\{0\}, |D| \leq d, \|\Delta_{D^c}\|_1 \leq 3\|\Delta_D\|_1\right\} > 0 \ , \tag{19}$$

where D is a set of indices, $D \subseteq \{1, \ldots, (N+1)p\}$, $|D|$ is its cardinality, D^c denotes the complement of the set D, Δ_D and Δ_{D^c} define vectors formed by

[3] The convergence holds with the assumption that $T^{-1}\sum_t Z_t Z_t' \xrightarrow{p} \Sigma^Z$ where $\Sigma^Z = E(Z_t Z_t')$, which is a positive definite matrix in the case of small and fixed N.

the coordinates of Δ with respect to the index set D and D^c. The minimum in (18) can be replaced by the minimum over a restricted set of Δ, and the L_2 norm $\|\Delta\|$ in the denominator can be substituted by $\|\Delta_D\|$. Consequently, this is seen as "restricted" positive definiteness for the Gram matrix, which is valid only for those vectors Δ satisfying that $\|\Delta_{D^c}\|_1 \leq 3\|\Delta_D\|_1$. In other words, the condition makes a restriction on the eigenvalues of the Gram matrix Ψ_W as a function of the sparsity d. Furthermore, the minimum in (19) is equivalently as follows,

$$\kappa_W(d) = \min\left\{\frac{\|W\Delta\|}{\sqrt{T}\|\Delta_D\|} : \Delta \in \mathbb{R}^{(N+1)p}\backslash\{0\}, |D| \leq d, \|\Delta_{D^c}\|_1 \leq 3\|\Delta_D\|_1\right\} > 0 \ . \tag{20}$$

We expect a small subset of observed predictors retained in the model by shrinking coefficients of irrelevant variables to be zero with Lasso. Therefore, only coefficient B is regularized subject to a constraint in L_1 norm. Furthermore, substitute factor estimate $\widehat{F}_t$ for F_t in forecast model (11). And, the Lasso estimates satisfies that

$$(\widehat{A}, \widehat{B}) = \arg\min_{A,B} T^{-1}\|y - (\widehat{F}A + WB)\|^2 + 2\lambda_T\|B\|_1 \ , \tag{21}$$

where λ_T is a data dependent tuning parameter that controls the amount of shrinkage.

Let $\widehat{\Omega} = (\widehat{A}', \widehat{B}')'$ and $\widehat{Z} = (\widehat{F}, W)$. We now state the asymptotic properties of prediction errors and the accuracy of Lasso estimate under the condition for *Restricted Eigenvalues of the Gram matrix*, which builds on the work of Callot (2012).

Theorem 2. Let $\lambda_T = \sqrt{8\ln(1+T)^5\ln(1+N)^4\ln(1+p)^2\ln(N^2p)\sigma_T^4/T}$, $N, p \in O(e^{T^a})$ and $s \in O(T^b)$, and assume that $7a + 2b < 1$ for $a, b \geq 0$. Suppose *Theorem 1* and *A.3* hold. Then, if $\sup\sigma_T < \infty$, the following holds as $T \to \infty$,

(a) $T^{-1}\|\widehat{Z}\widehat{\Omega} - Z\Omega\|^2 \xrightarrow{p} 0$,

(b) $\|\widehat{\Omega} - \Omega\|_1 \xrightarrow{p} 0$.

These two statements testify that mean squared forecasting errors (MSFE) converge to zero in probability, and the Lasso estimate $\widehat{\Omega}$ is L_1 consistent. The proof is given in the appendix.

3 Results

3.1 Forecast models comparison

For the comparison of predictive performance, we consider a few model candidates, and use a superscript i for the model index. Given the data available up to time T, we construct h-step-ahead forecast estimate $\hat{y}^{[i]}_{T+h|T}$ based on forecast model i. Furthermore, all the forecasts are constructed directly, which indicates the model should be re-estimated for different h. In addition, the argument for direct forecasts is that they are sensitive to the choices of h and reasonably robust at long forecast horizons (Callot (2012)).

3.1.1 Model 0

The benchmark model is random walk with drift, and it offers the forecast estimate at time $T + h$ as

$$\hat{y}^{[0]}_{T+h|T} = y_T + \frac{1}{T-h} \sum_{t=h+1}^{T} (y_t - y_{t-h}) \ . \tag{22}$$

3.1.2 Model 1 (EFAVAR)

Note that our theoretical results are particularly for one-step-ahead forecast, and the estimate can be constructed by (15). Furthermore, we can extend model (10) for multi-step-ahead forecast variables as follows,

$$y_{t+h} = \Omega'_h Z_t + \varepsilon_{t+h} \ , \tag{23}$$

where the subscript h of coefficient highlights the fact that separate models should be estimated regarding various forecast time horizons. Then, the h-step-ahead forecast estimate can be constructed by

$$\hat{y}^{[1]}_{T+h|T} = \widehat{\Omega}'_h \widehat{Z}_T \ , \tag{24}$$

where the Lasso estimate $\widehat{\Omega}_h$ can be obtained by fitting model (23) in which Z_t is replaced by $\widehat{Z}_t$ for $t = 1, \ldots, T-h$. In addition, 10-fold cross validation is used for choosing tuning parameter and model validation.

3.1.3 Model 2

Song and Bickel (2011) propose a large VAR, in which lags of the variable to be forecast are much more important than lags of predictors, and the distant lags have less influence on forecasting at current time than the recent lags. Therefore, observed variables can be weighted such that, for $i = 1, \ldots, p$,

$$\widetilde{\omega}_{t-i+1,j} = \begin{cases} i^{-\alpha} \cdot \omega_{t-i+1,j} & \text{if } j\text{th variable is } y \,, \\ \varphi i^{-\alpha} \cdot \omega_{t-i+1,j} & \text{otherwise} \,, \end{cases} \tag{25}$$

where $\alpha > 1$ and $0 < \varphi < 1$. Then, let $\widetilde{W}_t = (\widetilde{\omega}'_t, \widetilde{\omega}'_{t-1}, \ldots, \widetilde{\omega}'_{t-p+1})'$. And, the forecasts can be estimated as follows,

$$\hat{y}^{[2]}_{T+h|T} = \widehat{B}'_h \widetilde{W}_T \ , \tag{26}$$

where $\widehat{B}_h$ is a Lasso estimate, which the estimation procedure involves 10-fold cross validation as well.

3.1.4 MODEL 3

Forecasting with principal-component-based factors is proposed by Stock and Watson (2002a),

$$\hat{y}^{[3]}_{T+h|T} = \widehat{A}'_h \widehat{F}_T + \widehat{B}_h(L) y_T \ , \tag{27}$$

where $\widehat{A}_h$ and $\widehat{B}_h(L)$ can be obtained using OLS and Akaike information criterion (AIC) because of the low-dimensional predictors. In addition, $\widehat{B}_h(L) = \sum_{i=1}^{\hat{p}} \hat{\beta}_i L^{i-1}$ where $\hat{\beta}_i$ is a AR coefficient estimate for y_{t-i+1} for $i = 1, \ldots, \hat{p}$.

3.2 Simulation studies

Two small Monte Carlo experiments are conducted in order to explore how much forecasting improvement can be achieved in finite samples, if there is any, regarding the h-step-ahead forecasts, for $h = 1, 3, 6$ and 12. And, for each Monte Carlo replication, the data generating processes are described as follows.

Assume that $q = 1$, which corresponds to $\Gamma(L) = \Gamma$ with the diagonal element $\Gamma_{ii}(L) = \Gamma_{ii} = \gamma_{i1}$ for $i = 1, \ldots, N$. Then, the data $\{X_t\}_{t=1}^{T}$ can be generated based on static factor model (7) for each i,

$$X_{it} = \Lambda_i F_t + \gamma_{i1} X_{it-1} + \epsilon_{it} \ . \tag{28}$$

F_t and ϵ_t are simulated from multivariate normal distribution family, that are $F_t \sim \mathsf{N}(0, I_r)$ and $\epsilon_t \sim 0.1 \cdot \mathsf{N}(0, I_N)$. In addition, elements of Λ_i are uniformly distributed on $[0.1, 0.9]$, and AR(1) coefficient γ_{i1} is also uniformly distributed on $[0.5, 0.9]$ for $i = 1, \ldots, N$.

Suppose that $p = 1$, which results in $B = B_1 = (B_1^x, B_1^y)'$. The sparsity of B is determined as 2% of the dimension, and the index of nonzero parameters is chosen randomly. Then, the variable $\{y_t\}_{t=1}^{T+h}$ can be simulated by following model (23), which can be rewritten as follows,

$$y_{t+h} = A'F_t + B_1^x X_t + B_1^y y_t + \varepsilon_{t+h} \ , \tag{29}$$

where $\varepsilon_t \sim 0.1 \cdot \mathsf{N}(0,1)$, and elements of A and nonzero parameters of B are uniformly distributed on $[0.1, 0.9]$.

For each replication j, we compare the real observation $y_{T+h}(j)$ and its forecast estimate $\hat{y}_{T+h|T}^{[i]}(j)$ constructed by Model i. Additionally, the predictive performance of Model i is assessed over all the replications,

$$\text{root MSFE}^{[i]} = \sqrt{\frac{1}{MC}\sum_{j=1}^{MC}\left[\hat{y}_{T+h|T}^{[i]}(j) - y_{T+h}(j)\right]^2} \ , \tag{30}$$

where MC is the number of Monte Carlo replications.

3.2.1 Experiment A

This experiment is to examine how EFAVAR performs as changes of dimensions N and sample size T in comparison to other models. Given $r = 1$, the samples are simulated with different combinations of cross section $N = 100$, $250, 500$, $T = 20, 50, 100$. In addition, we specify the number of estimated factors and lags order of variables: $\hat{r} = \hat{p} = 1$ for Model 1 and Model 3; $\hat{p} = 1$ for Model 2[4].

Table 1 contains the results of root MSFE over 1000 Monte Carlo replications regarding various model candidates, in which there are four panels for different forecasting time horizons h. At first, we focus on the results of root MSFE$^{[1]}$. Within each sub-block, along with the increase of the dimension N and sample size T, the values of the root MSFE$^{[1]}$ have been in decline as expected. Besides that, the values of root MSFE$^{[1]}$ are greater in the case of larger h, which indicates the predictive performance becomes progressively worse as the value of h increases, however, the deterioration of performance can be weakened when N and T grow.

Next, we compare predictive performance between models. For each scenario specified T, N and h, the values of root MSFE$^{[1]}$ are smaller than

[4] Song and Bickel (2011) concludes that the predictive performance can be very robust for the choice of lags order, which primarily benefits from the re-weighting over lags as aforementioned. Thus, it is enough to use lagged variables having the order one.

Table 1: Results of Experiment A:
root MSFE of h-step-ahead forecasts over 1000 replications

	root MSFE$^{[1]}$			root MSFE$^{[2]}$			root MSFE$^{[3]}$		
T	20	50	100	20	50	100	20	50	100
$h = 1$									
N=100	0.155	0.122	0.112	0.161	0.127	0.115	0.333	0.266	0.229
N=250	0.100	0.078	0.066	0.107	0.078	0.071	0.293	0.223	0.173
N=500	0.072	0.059	0.055	0.100	0.067	0.052	0.256	0.181	0.139
$h = 3$									
N=100	0.366	0.254	0.217	0.385	0.275	0.226	0.443	0.347	0.319
N=250	0.217	0.146	0.115	0.244	0.158	0.124	0.358	0.260	0.211
N=500	0.137	0.093	0.085	0.252	0.113	0.087	0.354	0.214	0.165
$h = 6$									
N=100	0.418	0.271	0.231	0.444	0.293	0.248	0.524	0.402	0.356
N=250	0.277	0.174	0.132	0.347	0.178	0.143	0.491	0.299	0.234
N=500	0.188	0.116	0.096	0.378	0.157	0.099	0.501	0.258	0.186
$h = 12$									
N=100	0.660	0.290	0.239	0.703	0.318	0.246	0.727	0.425	0.373
N=250	0.484	0.211	0.144	0.613	0.216	0.159	0.736	0.345	0.263
N=500	0.371	0.138	0.115	0.661	0.187	0.129	0.696	0.321	0.222

root MSFE$^{[2]}$ and root MSFE$^{[3]}$, which suggests that EFAVAR performs better than other two models. However, the extent of improvement depends on choices of h and the development of T and N, or the ratio T/N. In addition, more effective amelioration from EFAVAR shows up when the value of T/N is small and h is large. For instance, when $h = 12$, $T = 20$ and $N = 500$, the values of root MSFE are 0.371, 0.661 and 0.696, respectively. And, it suggests that EFAVAR improves the predictive performance by reducing almost 50% of values for root MSFE$^{[2]}$ and root MSFE$^{[3]}$. Furthermore, when $h = 12$, $T = 100$ and $N = 500$, the values are 0.115, 0.129 and 0.222, thus EFAVAR performs slightly better than other two models.

3.2.2 Experiment B

This experiment is to investigate whether more predictive improvement can be obtained by using more factors as augmented predictors for EFAVAR. The dimension and sample size are pre-fixed to be large, $N = 500$ and $T = 100$. And we consider two scenarios regarding the number of factors generated, which are $r = 3$ and $r = 6$. Furthermore, we specify the number of estimated factors $\hat{r} = 1, \ldots, r$ and lags order $\hat{p} = 1$. Table 2 contains the results of root MSFE$^{[1]}$ for h-step-ahead forecasts over 1000 replications. It is not surprising that, by looking at each column, the general deterioration can be found for the long forecasting time horizon. Additionally, by looking at each row, the values are very robust for the choice of $\hat{r}$ under both scenarios. Therefore, we suggest that it seems enough to use one, or possibly two, estimated factor(s) to construct forecasts, although there are actually more true factors. Moreover, it provides some useful evidence that the precision of forecasts may not benefit from including more factors as augmented predictors.

Table 2: Results of Experiment B (N=500,T=100) :
root MSFE[1] of h-step-ahead forecasts over 1000 replications

	Scenario 1: $r = 3$			Scenario 2: $r = 6$					
	$\hat{r} = 1$	$\hat{r} = 2$	$\hat{r} = 3$	$\hat{r} = 1$	$\hat{r} = 2$	$\hat{r} = 3$	$\hat{r} = 4$	$\hat{r} = 5$	$\hat{r} = 6$
h = 1	0.044	0.044	0.045	0.042	0.042	0.042	0.041	0.041	0.041
h = 3	0.062	0.062	0.064	0.059	0.059	0.058	0.058	0.058	0.058
h = 6	0.076	0.077	0.077	0.076	0.076	0.075	0.075	0.076	0.077
h = 12	0.087	0.088	0.089	0.084	0.084	0.085	0.084	0.085	0.087

3.3 Empirical analysis

It is of our interest to examine the empirical performance of EFAVAR in terms of macroeconomic forecasting. We use a data set from Stock and Watson (2005) which contains 132 U.S. monthly time series for economics and finance from January 1959 to December 2003, $t = 1959:1,\ldots,2003:12$. In addition, the economic categories of variables include: real output and income; consumption; real retail, manufacturing and trade sales; employment and hours; housing starts and sales; real inventories and orders; money and credit quantity aggregates; stock prices; interest rates and spreads; exchange rates; price indexes; average hourly earnings; and miscellaneous.

All time series variables are transformed to be stationary, which can be achieved by taking logarithms and/or differencing once or twice. Generally speaking, first differences of logarithms are applied onto real quantity and activity measures, second differences of logarithms are employed for price series, first differences are used for nominal interest rates, and others remain at original levels. Additionally, more details about data description and pre-treatment can be found in Stock and Watson (2005). Furthermore, the transformed variables are examined for outliers and adjusted[5]. On

[5] Let $\mathbf{x} = \{x_i\}_{i=1}^{n}$ and consider x_i as a outlier if $|x_i - MED(\mathbf{x})| > 6 \cdot IQR(\mathbf{x})$, where $MED(\mathbf{x})$ and $IQR(\mathbf{x})$ are median and interquartile range of $\mathbf{x}$. Moreover, the outliers-adjusted observation can be obtained by the median value of previous five observations, that is $MED(x_{i-1},\ldots,x_{i-5})$.

one hand, the outliers-adjusted series are further standardized which will be employed for factor estimation. On the other hand, the outliers-unadjusted variables should be standardized as well which will be utilized to produce the Lasso estimates and construct forecasts.

The variables to be forecast of our interests are: industrial production index - total index (IP); consumer price index - all items (CPI); employees on nonfarm payrolls - total private ($EMPL$); and interest rate - Federal funds (FFR). For instance with $y = IP$, the transformed variable is $y_t = \ln(IP_t/IP_{t-1})$ and h-month-ahead forecast variable is $y_{t+h} = \ln(IP_{t+h}/IP_t)$. For each variable to be forecast, the h-month-ahead forecasts can be constructed monthly starting from $1970:1$ and ending at $2003:12$ for $h = 1, 3, 6$ and 12. To be specific, set an example of a 12-month-ahead forecast at $1970:1$. The factor estimate $\widehat{F}_t$ can be formed based on the data $\{X_t\}_{t=1959:1}^{1969:1}$. Next, the Lasso estimate $\widehat{\Omega}_{12}$ can be obtained by fitting model (23) in which Z_t is replaced by $\widehat{Z}_t$ for $t = 1959:1, \ldots, 1969:1$. Then, the forecast can be constructed by $\hat{y}_{1970:1|1969:1} = \widehat{\Omega}_{12}\widehat{Z}_{1969:1}$. Furthermore, this procedure is repeated for the rest $T = 1970:2, \ldots, 2003:12$.

The predictive performance can be assessed by (empirical) MSFE, which averages forecasting errors over a predictive time period as follows,

$$\text{MSFE}_h^{[i]} = \frac{1}{T_1 - T_0 + 1} \sum_{T=T_0-h}^{T_1-h} \left(\hat{y}_{T+h|T}^{[i]} - y_{T+h}\right)^2 , \tag{31}$$

where $T_0 = 1959:1$ and $T_1 = 2003:12$. Moreover, we shall report the empirical results using relative MSFE (RMSFE),

$$\text{RMSFE}_h^{[i]} = \text{MSFE}_h^{[i]} \;/\; \text{MSFE}_h^{[0]} , \tag{32}$$

where $\text{MSFE}_h^{[0]}$ is an benchmark produced by MODEL 0. In addition, Table A.1 lists the values of $\text{MSFE}_h^{[0]}$ regarding four macro variables to be forecast.

Table 3 shows the results of RMSFE, in which the bold type is the minimum of values in a row. Clearly, all the values are below one, which indicates that all estimated models perform better than benchmark model. The first block shows the values of $\text{RMSFE}_h^{[1]}$, which are calculated based on EFAVAR with various choices of the number of estimates factors $\hat{r}$ and lags order $\hat{p}$. And, we find that the smallest values of $\text{RMSFE}_h^{[1]}$ is obtained when $\hat{r} = 1$ and $\hat{p} = 1$. Moreover, by comparing the first column with the fifth, we can conclude that including more factors cannot better prediction, which the simulation experiment B provides supportive evidence. Furthermore, as changes of $\hat{p}$, the results seem to be robust, but the deterioration can be detected, which indicates that using more distant lagged variables cannot improve the forecasting accuracy.

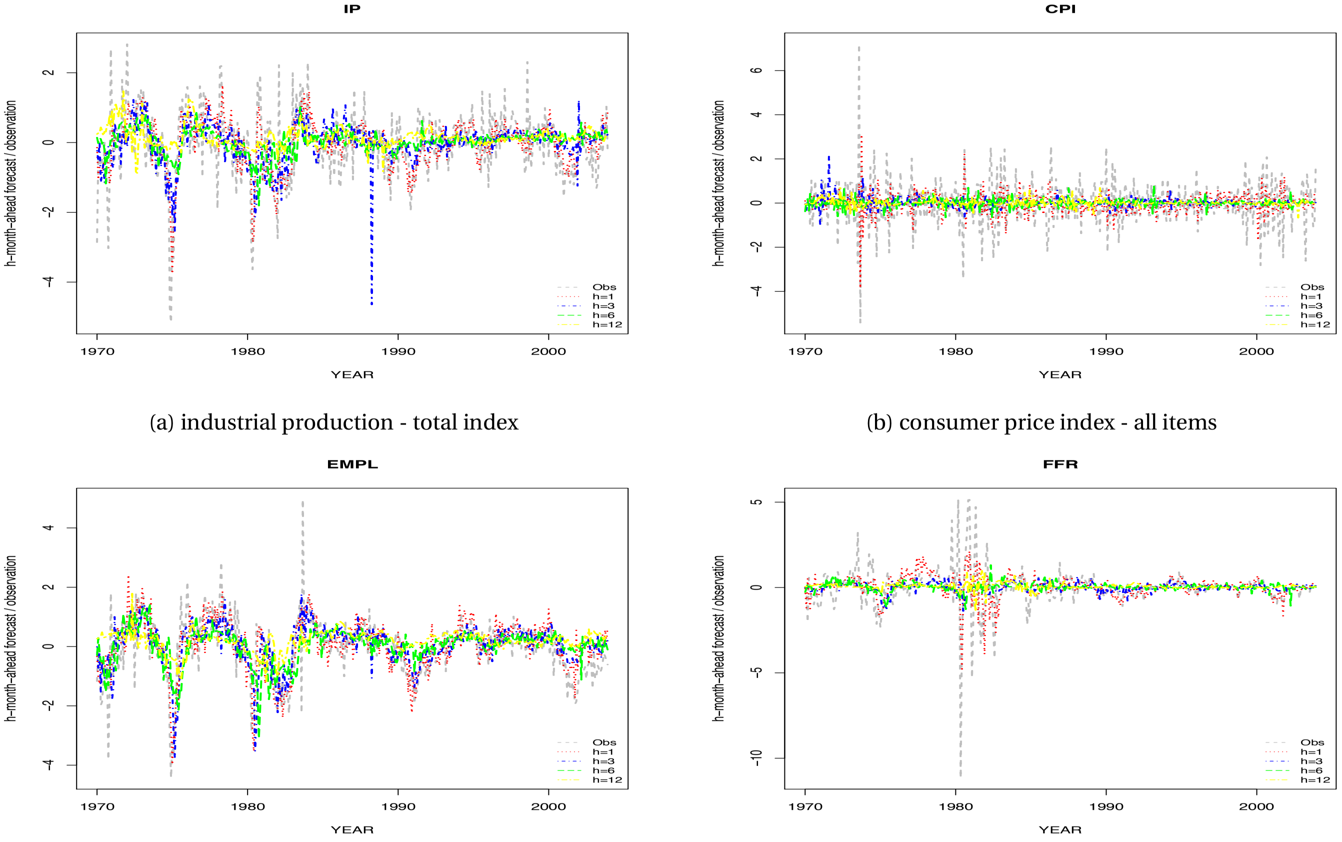
In order to compare EFAVAR with MODEL 2, we use the data weighted by (25), According to the results, we find that the predictive performance of EFAVAR and MODEL 2 are similar. In addition, the results of $\text{RMSFE}_h^{[3]}$ are produced based on MODEL 3 with $\hat{r} = 1$ and AIC-selected lags order $\hat{p}_{AIC}$. For the macro variable CPI in particular, all the minimum values locate the column $\text{RMSFE}_h^{[3]}$, which implies that MODEL 3 performs best. Except for CPI, the minimums show in the first column of $\text{RMSFE}_h^{[1]}$, which indicates that the optimal forecasts are constructed by EFAVAR. Therefore, the EFAVAR can successfully improve the predictive performance, although the improvement is not substantial in comparison with models proposed by Song and Bickel (2011) and Stock and Watson (2002a). Moreover, the extent of amelioration differs with the forecasting time horizon and macro variables to be forecast. In addition, Figure 1 illustrates the development of observations and optimal forecasts constructed by EFAVAR regarding four (stationary) macro variables.

4 Conclusions

This paper is concerned about forecasting for high-dimensional macroeconomic time series. We propose EFAVAR which models joint dynamics of macro variables to be forecast, a handful of latent factors, and a large number of observed predictors. Furthermore, we investigate the consistency of Lasso estimates and the accuracy of multi-step-ahead forecasts constructed by EFAVAR. In addition, we examine the predictive performance by conducting two small Monte Carlo experiments and an empirical study. Moreover, we conclude that EFAVAR can produce more precise forecasts in comparison to other forecast model candidates considered in this paper, and the precise forecasts can be constructed using one, or two, factor(s) as augmented predictors together with lagged variables of order one at most. Besides, the extent of improvement in predictive performance differs with forecasting time horizon and macro variable to be forecast.

Table 3: Empirical results for models comparison:
RMSFE of h-month-ahead forecasts

		$\text{RMSFE}_h^{[1]}$						$\text{RMSFE}_h^{[2]}$	$\text{RMSFE}_h^{[3]}$
		original data					weighted data	weighted data	original data
		$\hat{r} = 1$	$\hat{r} = 1$	$\hat{r} = 1$	$\hat{r} = 1$	$\hat{r} = 3$	$\hat{r} = 1$	no factor	$\hat{r} = 1$
		$\hat{p} = 1$	$\hat{p} = 4$	$\hat{p} = 7$	$\hat{p} = 13$	$\hat{p} = 1$	$\hat{p} = 1$	$\hat{p} = 1$	$\hat{p}_{AIC}$
$h = 1$	IP	**0.576**	0.594	0.605	0.622	0.888	0.583	0.588	0.623
	CPI	0.309	0.309	0.324	0.352	0.355	0.313	0.310	**0.307**
	$EMPL$	**0.613**	0.621	0.630	0.641	1.312	0.750	0.724	0.693
	FFR	**0.633**	0.661	0.649	0.668	1.085	0.675	0.639	0.672
$h = 3$	IP	**0.583**	0.586	0.592	0.607	0.661	0.601	0.607	0.609
	CPI	0.519	0.568	0.555	0.641	0.546	0.521	0.519	**0.502**
	$EMPL$	**0.644**	0.675	0.697	0.720	0.788	0.732	0.768	0.712
	FFR	**0.414**	0.505	0.445	0.483	0.506	0.451	0.457	0.432
$h = 6$	IP	**0.509**	0.522	0.533	0.517	0.521	0.510	0.512	0.544
	CPI	0.494	0.529	0.580	0.687	0.509	0.498	0.499	**0.494**
	$EMPL$	**0.594**	0.608	0.632	0.680	0.679	0.646	0.648	0.649
	FFR	**0.458**	0.481	0.488	0.491	0.482	0.479	0.462	0.496
$h = 12$	IP	**0.487**	0.492	0.493	0.496	0.525	0.508	0.495	0.508
	CPI	0.479	0.579	0.571	0.576	0.496	0.484	0.481	**0.476**
	$EMPL$	**0.526**	0.559	0.534	0.592	0.570	0.536	0.615	0.576
	FFR	**0.420**	0.457	0.456	0.471	0.485	0.457	0.457	0.482



(a) industrial production - total index

(b) consumer price index - all items

(c) employees on nonfarm payrolls - total private

(d) interest rates - Federal funds

Figure 1: Development of the observations and optimal *h*-month-ahead forecasts constructed by EFAVAR regarding four macro variables to be forecast of our interest (transformed)

Appendices

A Tables

Table A.1: Results of empirical benchmarks $\text{MSFE}_h^{[0]}$

	IP	CPI	EMP	FFR
h = 1	1.103	2.923	0.794	1.475
h = 3	1.399	2.125	0.890	2.762
h = 6	1.736	2.178	1.305	2.621
h = 12	1.933	2.262	1.834	2.853

B Proofs

Recall that we use $\tilde{X}_t$ as predictors to estimate factors, where $\tilde{X}_t = (I - \Gamma(L)L)X_t$, and then rewrite factor model (7) as

$$\tilde{X}_t = \Lambda F_t + \epsilon_t \ .$$

The simplest case is that $\Gamma(L) = 0$, which gives the static form of the DFM proposed by Stock and Watson (2002a). Note that $\tilde{X}_t$ and X_t share the same assumptions and restrictions, because the linear transformation between them does not influence theoretical properties.

Before proving the main theorems, we introduce several preparatory lemmas, which are adapted from Stock and Watson (2002a)) and then rearranged. At first, let v be an N-dimensional vector. Define $\Upsilon = \{v|v'v/N = 1\}$, $R(v) = N^{-2}T^{-1}v'\sum_t \tilde{X}_t\tilde{X}_t'v$, and $R^*(v) = N^{-2}T^{-1}v'\sum_t \Lambda F_t F_t'\Lambda'v$.

Lemma 1 $|\sup_{v\in\Upsilon} R(v) - \sup_{v\in\Upsilon} R^*(v)| \xrightarrow{p} 0$.

Proof: We can derive the following,

$$\begin{aligned} |\sup_{v\in\Upsilon} R(v) - \sup_{v\in\Upsilon} R^*(v)| &\leq \sup_{v\in\Upsilon} |R(v) - R^*(v)| \\ &= \sup_{v\in\Upsilon} |N^{-2}T^{-1}v'\epsilon'\epsilon v + 2N^{-2}T^{-1}v'\Lambda F'\epsilon v| \\ &\leq \sup_{v\in\Upsilon} N^{-2}T^{-1}|v'\epsilon'\epsilon v| + 2\sup_{v\in\Upsilon} N^{-2}T^{-1}|v'\Lambda F'\epsilon v| \ . \end{aligned}$$

Now, we want the two parts on the right hand side (RHS) of the inequality converge to 0 in probability so that *Lemma 1* can be proved.

(1). Check $\sup_{v\in\Upsilon} N^{-2}T^{-1}|v'\epsilon'\epsilon v|$. Consider that

$$\begin{aligned}
N^{-2}T^{-1}v'\epsilon'\epsilon v &= N^{-2}T^{-1}\sum_t\sum_i\sum_j v_i v_j \epsilon_{it}\epsilon_{jt} \\
&= N^{-2}\sum_i\sum_j v_i v_j (T^{-1}\sum_t \epsilon_{it}\epsilon_{jt}) \\
&\leq N^{-2}[\sum_i\sum_j (v_i v_j)^2 \times \sum_i\sum_j (T^{-1}\sum_t \epsilon_{it}\epsilon_{jt})^2]^{1/2} \\
&= (N^{-2}\sum_i\sum_j v_i^2 v_j^2)^{1/2} \times [N^{-2}\sum_i\sum_j (T^{-1}\sum_t \epsilon_{it}\epsilon_{jt})^2]^{1/2} \quad .
\end{aligned}$$

For all $v \in \Upsilon$, we have that $v'v/N = 1$ and $N^{-2}\sum_i\sum_j v_i^2 v_j^2 = (v'v/N)^2 = 1$. Thus, we obtain that

$$\begin{aligned}
\sup_{v\in\Upsilon} N^{-2}T^{-2}|v'\epsilon'\epsilon v| &\leq [N^{-2}\sum_i\sum_j (T^{-1}\sum_t \epsilon_{it}\epsilon_{jt})^2]^{1/2} \\
&= \underbrace{(N^{-2}T^{-2}\sum_i\sum_j\sum_t\sum_s \epsilon_{it}\epsilon_{is}\epsilon_{jt}\epsilon_{js})^{1/2}}_{\text{(i)}} \quad .
\end{aligned}$$

Also, denote $\gamma_{its} = E(\epsilon_{it}\epsilon_{is})$ and $\gamma_{jts} = E(\epsilon_{jt}\epsilon_{js})$. Then, we can derive the expectation as

$$\begin{aligned}
E\text{(i)} &= E(N^{-2}T^{-2}\sum_i\sum_j\sum_t\sum_s \epsilon_{it}\epsilon_{is}\epsilon_{jt}\epsilon_{js}) \\
&= \underbrace{N^{-2}T^{-2}\sum_i\sum_j\sum_t\sum_s \gamma_{its}\gamma_{jts}}_{\text{(ii)}} + \underbrace{N^{-2}T^{-2}\sum_i\sum_j\sum_t\sum_s E[(\epsilon_{it}\epsilon_{is}-\gamma_{its})(\epsilon_{jt}\epsilon_{js}-\gamma_{jts})]}_{\text{(iii)}} \quad .
\end{aligned}$$

On one hand, let $s = t + u$. Then the term (ii) is rewritten by

$$\begin{aligned}
\text{(ii)} &= T^{-2}\sum_t\sum_u (N^{-1}\sum_i \gamma_{it(t+u)})(N^{-1}\sum_j \gamma_{jt(t+u)}) \\
&= T^{-2}\sum_t\sum_u \gamma_{Nt}(u)^2 \quad ,
\end{aligned}$$

in which $\gamma_{Nt}(u)$ is defined by *A.1(a)* and further expressed as follows,

$$\gamma_{Nt}(u) = E(\epsilon_t'\epsilon_{t+u}/N) = N^{-1}\sum_i E[\epsilon_{it}\epsilon_{i(t+u)}] = N^{-1}\sum_i \gamma_{it(t+u)} \quad .$$

Moreover, *A.1(a)* provides the property of absolute summability, which leads to the square summability, $\lim_{N\to\infty}\sup_t\sum_{u=-\infty}^{\infty}\gamma_{N,t}(u)^2 < \infty$. And, it implies that the term (ii) converges to zero. On the other hand, the term (iii) is represented by

$$\begin{aligned}
\text{(iii)} &= N^{-2}T^{-2}\sum_i\sum_j\sum_t\sum_s \text{cov}\,(\epsilon_{it}\epsilon_{is}, \epsilon_{jt}\epsilon_{js}) \\
&\leq N^{-2}\sup_{t,s}\sum_i\sum_j |\text{cov}\,\epsilon_{it}\epsilon_{is}, \epsilon_{jt}\epsilon_{js}| \to 0 \quad ,
\end{aligned}$$

where the convergence holds under assumption *A.1(c)*. Therefore, both (ii) and (iii) towards zero gives E(i) = (ii) + (iii) → 0, and we conclude that

$$\sup_{v\in\Upsilon} N^{-2}T^{-1}|v'\epsilon'\epsilon v| \le (\mathrm{i})^{\frac{1}{2}} \xrightarrow{p} 0 \ .$$

(2). Check $\sup_{v\in\Upsilon} N^{-2}T^{-1}|v'\Lambda F'\epsilon v|$. We consider the following,

$$\begin{aligned} |N^{-2}T^{-1}v'\Lambda F'\epsilon v| &= |\sum_j (v'\underline{\lambda}_j/N)\times T^{-1}\sum_t F_{jt}(N^{-1}\sum_i v_i\epsilon_{it})| \\ &\le \sum_j |v'\underline{\lambda}_j/N| \times |T^{-1}\sum_t F_{jt}(N^{-1}\sum_i v_i\epsilon_{it})| \ , \end{aligned}$$

where $\underline{\lambda}_j$ denotes the jth column of Λ. And we find that

$$\begin{aligned} & \sup_{v\in\Upsilon} N^{-2}T^{-1}|v'\Lambda F'\epsilon v| \\ \le\ & \max_j \sup_{v\in\Upsilon} |v'\underline{\lambda}_j/N| \times \sum_{j=1}^{r} \sup_{v\in\Upsilon} |T^{-1}\sum_t F_{jt}(N^{-1}\sum_i v_i\epsilon_{it})| \\ \le\ & \sup_{v\in\Upsilon}(v'v/N)^{\frac{1}{2}} \times \max_j(\underline{\lambda}_j'\underline{\lambda}_j/N)^{\frac{1}{2}} \times \sum_{j=1}^{r} \sup_{v\in\Upsilon} |T^{-1}\sum_t F_{jt}(N^{-1}\sum_i v_i\epsilon_{it})| \\ =\ & \sum_{j=1}^{r} \sup_{v\in\Upsilon} |T^{-1}\sum_t F_{jt}(N^{-1}\sum_i v_i\epsilon_{it})| \ , \end{aligned}$$

where the last line follows from $v'v/N = 1$ and $\Lambda'\Lambda/N \to I_r$, which are defined by *A.2(c)*. Moreover, we consider that, for $j = 1, \ldots, r$,

$$\sup_{v\in\Upsilon} |T^{-1}\sum_t F_{jt}(N^{-1}\sum_i v_i\epsilon_{it})| \le (T^{-1}\sum_t F_{jt}^2)^{\frac{1}{2}} \cdot [\sup_{v\in\Upsilon} T^{-1}\sum_t (N^{-1}\sum_i v_i\epsilon_{it})^2]^{\frac{1}{2}} \ ,$$

in which $T^{-1}\sum_t F_{jt}^2 \xrightarrow{p} \sigma_{jj}$ by *A.2(b)*, where $\sigma_{jj} = \Sigma_{jj}^F$ for brevity. Additionally, we have that

$$\begin{aligned} \sup_{v\in\Upsilon} T^{-1}\sum_t (N^{-1}\sum_i v_i\epsilon_{it})^2 &= \sup_{v\in\Upsilon} N^{-2}T^{-1}\sum_t\sum_i\sum_j v_iv_j\epsilon_{it}\epsilon_{jt} \\ &= \sup_{v\in\Upsilon} N^{-2}T^{-1}v'\epsilon'\epsilon v \\ &\le \sup_{v\in\Upsilon} N^{-2}T^{-1}|v'\epsilon'\epsilon v| \to 0 \ , \end{aligned}$$

in which the convergence has been proved earlier in (1). Therefore, we can obtain that

$$\sup_{v\in\Upsilon} N^{-2}T^{-1}|v'\Lambda F'\epsilon v| \to 0 \ .$$

□

Lemma 2 Let $\hat{\underline{\lambda}}_1 = \arg\sup_{v\in\Upsilon} R(v)$, then $R^*(\hat{\underline{\lambda}}_1) \xrightarrow{p} \sigma_{11}$.

Proof: Denote

$$v = \Lambda(\Lambda'\Lambda/N)^{-\frac{1}{2}}\delta + V \ ,$$

where $V'\Lambda = 0$. Also, we use that

$$v'v/N = \delta'\delta + V'V/N \ ,$$

in which v satisfies that $v'v/N = 1$ as defined earlier. It results in that $\delta'\delta \le 1$ while $V'V/N \le 1$ for all $v \in \Upsilon$. Then, we consider the following,

$$\sup_{v\in\Upsilon} R^*(v) = \sup_{\delta\in\{\delta|\delta'\delta\le 1\}} \delta' C_{NT}\delta \ ,$$

where $C_{NT} = (\Lambda'\Lambda/N)^{\frac{1}{2}\prime}(F'F/T)(\Lambda'\Lambda/N)^{\frac{1}{2}}$.

Recall that $(\Lambda'\Lambda/N)^{\frac{1}{2}} \to I_r$ and $F'F/T \xrightarrow{p} \Sigma^F$ where Σ^F is diagonal by assumptions *A.2(b)* & *(c)*, and we obtain that

$$C_{NT} \xrightarrow{p} \Sigma^F \ .$$

Additionally, eigenvalues of C_{NT} converge into diagonal entries of Σ^F in probability according to eigendecomposition. In other words, $\hat{\sigma}_{ii} \xrightarrow{p} \sigma_{ii}$ for $i = 1,\ldots,r$, where $\hat{\sigma}_{ii}$ are eigenvalues of C_{NT} in decreasing order. Therefore, we have that

$$\sup_{v\in\Upsilon} R^*(v) \xrightarrow{p} \sigma_{11} \ ,$$

and further

$$\sup_{v\in\Upsilon} R(v) \xrightarrow{p} \sigma_{11} \ ,$$

which follows from the statement of *Lemma 1*. Similarly, the definition of $\underline{\hat{\lambda}}_1$ gives that

$$R(\underline{\hat{\lambda}}_1) \xrightarrow{p} \sigma_{11} \ .$$

And, according to the proven result that $sup_{v\in\Upsilon}|R(v) - R^*(v)| \xrightarrow{p} 0$, we find that

$$R^*(\underline{\hat{\lambda}}_1) \xrightarrow{p} \sigma_{11} \ .$$

□

Lemma 3 Suppose that the $N \times r$ matrix $\widehat{\Lambda}$ is composed by the r ordered eigenvectors of $\tilde{X}'\tilde{X}$ normalized as $\widehat{\Lambda}'\widehat{\Lambda}/N = I_r$, in which the first column of $\widehat{\Lambda}$ is the eigenvector corresponding to the largest eigenvalue, etc. Denote $S = diag(sign(\widehat{\Lambda}'\Lambda))$. Then, $S\widehat{\Lambda}'\Lambda/N \xrightarrow{p} I$.

Proof: We start with the first column of $S\widehat{\Lambda}'\Lambda/N$. Denote $\underline{\hat{\lambda}}_1$ as the first column column of $\widehat{\Lambda}$ and $S_1 = sign(\underline{\hat{\lambda}}_1'\underline{\lambda}_1)$, where

$$sign(x) = \begin{cases} 1 & \text{if } x \ge 0 \,, \\ -1 & \text{if } x < 0 \,. \end{cases}$$

We can write $\underline{\hat{\lambda}}_1$ as a case of v, such that $\underline{\hat{\lambda}}_1 = \Lambda(\Lambda'\Lambda/N)^{-\frac{1}{2}}\hat{\delta}_1 + \widehat{V}_1$, for some values of $\hat{\delta}_1$ and $\widehat{V}_1$, where $\widehat{V}_1'\Lambda = 0$, $\hat{\delta}_1\hat{\delta}_1 \le 1$ and $\widehat{V}_1'\widehat{V}_1 \le 1$. Then, we consider that

$$\begin{aligned} R^*(\underline{\hat{\lambda}}_1) - \sigma_{11} &= \hat{\delta}_1' C_{NT}\hat{\delta}_1 - \sigma_{11} \\ &= \hat{\delta}_1'(C_{NT} - \Sigma^F)\hat{\delta}_1 + \hat{\delta}_1'\Sigma^F\hat{\delta}_1 - \sigma_{11} \\ &= \hat{\delta}_1'(C_{NT} - \Sigma^F)\hat{\delta}_1 + (\hat{\delta}_{11}^2 - 1)\sigma_{11} + \sum_{i=2}^{r} \hat{\delta}_{ii}^2\sigma_{ii} \ . \end{aligned}$$

The left hand side (LHS) converges towards zero in probability by *Lemma 2*. Hence, the RHS should go for zero as well. Note that C_{NT} is proved to converge to Σ^F in probability, and $\hat{\delta}_1$ is bounded by $\hat{\delta}_1\hat{\delta}_1 \leq 1$. Therefore, the following must be satisfied,

$$(\hat{\delta}_{11}^2 - 1)\sigma_{11} + \sum_{i=2}^{r} \hat{\delta}_{ii}^2 \sigma_{ii} \xrightarrow{p} 0 \ ,$$

where $\sigma_{ii} > 0$ for $i = 1, \cdots, r$ by *A.2(a)*. It results in that $\hat{\delta}_{11}^2 \xrightarrow{p} 1$ while $\hat{\delta}_{ii}^2 \xrightarrow{p} 0$ for $i > 2$, and is equivalent to $\hat{\delta}_1'\hat{\delta}_1 \xrightarrow{p} 1$.

Recalling that $\underline{\hat{\lambda}}_1$ as a form as v, we have that

$$\underline{\hat{\lambda}}_1'\underline{\hat{\lambda}}_1/N = \hat{\delta}_1'\hat{\delta}_1 + \widehat{V}_1'\widehat{V}_1 = 1 \ ,$$

which indicates that $\widehat{V}_1'\widehat{V}_1 \xrightarrow{p} 0$. Given $\Lambda'\Lambda/N \to I_r$, we know that

$$S_1\underline{\hat{\lambda}}_1'\Lambda/N \xrightarrow{p} (1\,0\cdots0) \ ,$$

which is of r dimensions with one as the first element and zeros as the rest. For other columns, the proof can be derived in similar ways. Within the orthonormal subspace of Υ, the jth column of $\widehat{\Lambda}$ can be written as

$$\underline{\hat{\lambda}}_j = \Lambda(\Lambda'\Lambda/N)^{-\frac{1}{2}}\hat{\delta}_j + \widehat{V}_j \ ,$$

where $\widehat{V}_j'\Lambda = 0$, $\widehat{V}_j'\widehat{V}_j \xrightarrow{p} 0$, $\hat{\delta}_{jj}^2 \xrightarrow{p} 1$ and $\hat{\delta}_{ij}^2 \xrightarrow{p} 0$ for $i \neq j$.

□

Theorem 1(a)

Proof: For $j = 1, 2, \ldots, r$, we derive the following,

$$\begin{aligned}
S_j\widehat{F}_{jt} - F_{jt} &= S_j(\underline{\hat{\lambda}}_j'\tilde{X}_t/N) - F_{jt} \\
&= S_j\underline{\hat{\lambda}}_j'(\Lambda F_t + \epsilon_t)/N - F_{jt} \\
&= S_j\underline{\hat{\lambda}}_j'\sum_i(\underline{\lambda}_i F_{it})/N + S_j\underline{\hat{\lambda}}_j'\epsilon_t/N - F_{jt} \\
&= \underbrace{(S_j\underline{\hat{\lambda}}_j'\underline{\lambda}_j/N - 1)F_{jt}}_{\text{(i)}} + \underbrace{\sum_{i\neq j} S_j\underline{\hat{\lambda}}_j'\underline{\lambda}_i F_{it}/N}_{\text{(ii)}} + \underbrace{S_j\underline{\hat{\lambda}}_j'\epsilon_t/N}_{\text{(iii)}} \ .
\end{aligned}$$

Lemma 3 says that $S\widehat{\Lambda}'\Lambda/N \xrightarrow{p} I_r$, which means $S_j\underline{\hat{\lambda}}_j'\underline{\lambda}_j/N \xrightarrow{p} 1$ and $\underline{\hat{\lambda}}_j'\underline{\lambda}_i \xrightarrow{p} 0$ for $i \neq j$. Additionally, the assumption *A.2(b)* indicates that $|F_t| \sim O(1)$, thus we can obtain that

$$\begin{aligned}
\text{(i)} &= (S_j\underline{\hat{\lambda}}_j'\underline{\lambda}_j/N - 1)F_{jt} \xrightarrow{p} 0 \ ; \\
\text{(ii)} &= \sum_{i\neq j} S_j\underline{\hat{\lambda}}_j'\underline{\lambda}_i F_{it}/N \xrightarrow{p} 0 \ .
\end{aligned}$$

Then, we rewrite the term (iii) with elements,

$$
\begin{aligned}
\text{(iii)} &= N^{-1}S_j\sum_i \hat{\lambda}_{ij}\epsilon_{it} \\
&= \underbrace{N^{-1}S_j\sum_i(\hat{\lambda}_{ij}-\lambda_{ij})\epsilon_{it}}_{\text{(iv)}} + \underbrace{N^{-1}S_j\sum_i \lambda_{ij}\epsilon_{it}}_{\text{(v)}} \; .
\end{aligned}
$$

Next, we will show both (iv) and (v) converge to zero in probability so that we can obtain that (iii) $\xrightarrow{p} 0$ and then complete the proof.

(1). Show that (iv) $\xrightarrow{p} 0$. Using Slutsky's theorem to show the convergence, we can write the following absolute value,

$$
|\text{(iv)}| \leq \underbrace{(N^{-1}S_j\sum_i(\hat{\lambda}_{ij}-\lambda_{ij})^2)^{\frac{1}{2}}}_{\text{(vi)}} \times \underbrace{(N^{-1}\sum_i \epsilon_{it}^2)^{\frac{1}{2}}}_{\text{(vii)}} \; .
$$

Firstly, we expand the squares (vi) as

$$
\begin{aligned}
\text{(vi)} &= N^{-1}\sum_i \hat{\lambda}_{ij}^2 + N^{-1}\sum_i \lambda_{ij}^2 - 2N^{-1}\sum_i S_j\hat{\lambda}_{ij}\lambda_{ij} \\
&= \underline{\hat{\lambda}}_j'\underline{\hat{\lambda}}_j/N + \underline{\lambda}_j'\underline{\lambda}_j/N - 2S_j\underline{\hat{\lambda}}_j'\underline{\lambda}_j \xrightarrow{p} 0 \; ,
\end{aligned}
$$

in which the convergence holds according to $\Lambda'\Lambda/N \xrightarrow{p} I_r$ in *A.2(c)*, $\widehat{\Lambda}'\widehat{\Lambda}/N = I_r$ and $S\widehat{\Lambda}'\Lambda/N \xrightarrow{p} I_r$ in *Lemma 3*. Secondly, *A.1(b)* gives that $\tau_{ij,t} = E(\epsilon_{it}\epsilon_{jt})$, and thus we derive the term (vii) as

$$
\text{(vii)} = \underbrace{N^{-1}\sum_i(\epsilon_{it}^2 - \tau_{ii,t})}_{\text{(viii)}} + \underbrace{N^{-1}\sum_i \tau_{ii,t}}_{\text{(ix)}} \; .
$$

Note that the fact (viii) $\xrightarrow{p} 0$ follows from

$$
\begin{aligned}
E\text{(viii)}^2 &= N^{-2}\sum_i\sum_j E(\epsilon_{it}^2 - \tau_{ii,t})(\epsilon_{jt}^2 - \tau_{jj,t}) \\
&= N^{-2}\sum_i\sum_j \text{cov}\,(\epsilon_{it}^2, \epsilon_{jt}^2) \\
&\leq N^{-2}\sum_i\sum_j |\text{cov}\,(\epsilon_{it}^2, \epsilon_{jt}^2)| \\
&\leq \sup_{t,s} N^{-2}\sum_i\sum_j |\text{cov}\,(\epsilon_{it}\epsilon_{is}, \epsilon_{jt}\epsilon_{js})| \to 0 \; ,
\end{aligned}
$$

in terms of *A.1(c)* that $\lim_{N\to\infty}\sup_{t,s} N^{-1}\sum_i\sum_j |\text{cov}\,(\epsilon_{it}\epsilon_{is}, \epsilon_{jt}\epsilon_{js})| < \infty$. Moreover, we obtain that (ix) $\leq N^{-1}\sum_i |\tau_{ii,t}| < \infty$ by *A.1(b)*. Therefore, with (vii) $\sim O_p(1)$ and (vi) $\xrightarrow{p} 0$, we can conclude that $|\text{(iv)}| \xrightarrow{p} 0$ by Slutsky's theorem, which leads to (iv) $\xrightarrow{p} 0$ as well.

(2). Show that (v) $\xrightarrow{p} 0$. We consider the mean squares of (v) as follows,

$$\begin{aligned} E(\mathrm{v})^2 &= N^{-2}S_j\sum_i\sum_m \lambda_{ij}\lambda_{mj}\tau_{im,t} \\ &\leq N^{-2}S_j\sum_i\sum_m \lambda_{ij}\lambda_{mj}|\tau_{im,t}| \\ &\leq \bar{\lambda}^2 N^{-2}S_j\sum_i\sum_m |\tau_{im,t}| \to 0 \ , \end{aligned}$$

in which the first inequality is based on the definition of τ in *A.1(b)*, the second inequality holds by introducing $\bar{\lambda}$ defined as the maximal element of Λ, and the convergence holds by *A.1(b)* as well. Thus, we conclude that (v) $\xrightarrow{p} 0$.

□

Lemma 4 Let g_t define a sequence of random variables, such that $T^{-1}\sum_t g_t^2 \xrightarrow{p} \sigma_g$ and $T^{-1}\sum_t F_t g_t \xrightarrow{p} \Sigma_{Fg}$. Then $T^{-1}\sum_t S\widehat{F}_t g_t \xrightarrow{p} \Sigma_{Fg}$.

Proof: We consider the following,

$$\begin{aligned} T^{-1}\sum_t S\widehat{F}_t g_t &= T^{-1}\sum_t S(\widehat{\Lambda}'\tilde{X}_t/N)g_t \\ &= (TN)^{-1}\sum_t S\widehat{\Lambda}'(\Lambda F_t+\epsilon_t)g_t \\ &= T^{-1}\sum_t(S\widehat{\Lambda}'\Lambda/N)F_t g_t + (TN)^{-1}\sum_t S\widehat{\Lambda}'\epsilon_t g_t \ , \end{aligned}$$

in which $T^{-1}\sum_t(S\widehat{\Lambda}'\Lambda/N)F_t g_t \xrightarrow{p} \Sigma_{Fg}$ under the result $S\widehat{\Lambda}'\Lambda/N \xrightarrow{p} I_r$ and the condition $T^{-1}\sum_t F_t g_t \xrightarrow{p} \Sigma_{Fg}$. In addition, recalling that $\hat{\underline{\lambda}}'_j \in \Upsilon$, we can obtain the following, for $j = 1,\ldots,r$,

$$\begin{aligned} |(TN)^{-1}\sum_t S_j\hat{\underline{\lambda}}'_j\epsilon_t g_t| &\leq (TN)^{-1}\sup_{v\in\Upsilon}\sum_t v'\epsilon_t g_t \\ &= \sup_{v\in\Upsilon} T^{-1}\sum_t(N^{-1}v'\epsilon_t)g_t \\ &\leq (T^{-1}\sum_t g_t^2)^{\frac{1}{2}}\cdot\big[\sup_{v\in\Upsilon}T^{-1}\sum_t(N^{-1}v'\epsilon_t)^2\big]^{\frac{1}{2}} \xrightarrow{p} 0 \ , \end{aligned}$$

where the convergence holds in terms of Slutsky's theorem under the condition $T^{-1}\sum_t g_t^2 \xrightarrow{p} \sigma_g$. Furthermore, we have that

$$\begin{aligned} \sup_{v\in\Upsilon}T^{-1}\sum_t(N^{-1}v'\epsilon_t)^2 &= \sup_{v\in\Upsilon}T^{-1}N^{-2}v'\epsilon'\epsilon v \\ &\leq \sup_{v\in\Upsilon}T^{-1}N^{-2}|v'\epsilon'\epsilon v| \xrightarrow{p} 0 \ , \end{aligned}$$

as shown in the proof of *Lemma 1*.

□

Theorem 1(b)

Proof: We expand the sum squares as follows,

$$T^{-1}\sum_t(S_i\widehat{F}_{jt}-F_{jt})^2 = T^{-1}\sum_t\widehat{F}_{jt}^2 - 2T^{-1}\sum_t S_i\widehat{F}_{jt}F_{jt} + T^{-1}\sum_t F_{jt}^2 \ ,$$

where the last term $T^{-1}\sum_t F_{jt}^2 \xrightarrow{p} \sigma_{jj}$ by *A.2(b)*. In order to complete the proof, we need to show that both $T^{-1}\sum_t \widehat{F}_{jt}^2$ and $T^{-1}\sum_t \widehat{F}_{jt}F_{jt}$ converge to σ_{jj} in probability. On one hand, let F_{jt} be the series of g_t from *Lemma 4*, for $j = 1,2,\ldots,r$. Additionally, *A.2(a)* states that $T^{-1}\sum_t F_{jt}^2 \xrightarrow{p} \sigma_{jj}$ and $T^{-1}\sum_t F_t F_{jt} \xrightarrow{p} (\Sigma^F)_j$. Then, according to *Lemma 4*, we obtain that, for $j = 1,2,\ldots,r$,

$$T^{-1}\sum_t S\widehat{F}_t F_{jt} \xrightarrow{p} (\Sigma^F)_j \ ,$$

or equivalently,

$$T^{-1}\sum_t S\widehat{F}_t F_t' \xrightarrow{p} \Sigma^F \ .$$

In other words, for any $j = 1,2,\ldots,r$, we know that

$$T^{-1}\sum_t S_j\widehat{F}_{jt}F_{jt} \xrightarrow{p} \sigma_{jj} \ .$$

On the other hand, set $g_t = S_j\widehat{F}_{jt}$, and the following holds on the basis of *Lemma 4*,

$$T^{-1}\sum_t \widehat{F}_t\widehat{F}_t' \xrightarrow{p} \Sigma^F,$$

which means, for $j = 1,2,\ldots,r$,

$$T^{-1}\sum_t \widehat{F}_{jt}^2 \xrightarrow{p} \sigma_{jj} \ .$$

□

Theorem 1(c)

Proof: Recalling the representation of v from *Lemma 2*, we let $\hat{v}$ be the ith ordered eigenvectors of $\tilde{X}'\tilde{X}$ for $i > r$, such that

$$\hat{v} = \Lambda(\Lambda'\Lambda/N)^{-\frac{1}{2}}\hat{\delta} + \widehat{V} \ ,$$

where $\widehat{V}'\Lambda = 0$, $\widehat{V}'\widehat{V}/N \le 1$ and $\hat{\delta}'\hat{\delta} \le 1$. Also, we normalize it by $\hat{v}'\hat{v} = 1$. By the definition, we obtain that $T^{-1}\sum_t \widehat{F}_{it}^2 = R(\hat{v})$. Then, we need to show that $R(\hat{v}) \xrightarrow{p} 0$ or $R^*(\hat{v}) \xrightarrow{p} 0$.

Recall that in the proof of *Lemma 3*, $\underline{\hat{\lambda}}_j$ defines the jth ordered eigenvectors of $\tilde{X}'\tilde{X}$ for $j = 1,\ldots,r$,

$$\underline{\hat{\lambda}}_j = \Lambda(\Lambda'\Lambda/N)^{-\frac{1}{2}}\hat{\delta}_j + \widehat{V}_j \ ,$$

where $\widehat{V}_j'\Lambda = 0$, $\widehat{V}_j'\widehat{V}_j \xrightarrow{p} 0$, $\hat{\delta}_{jj}^2 \xrightarrow{p} 1$ and $\hat{\delta}_{ij}^2 \xrightarrow{p} 0$ for $i \neq j$. Next, we consider the following,

$$\hat{v}\underline{\hat{\lambda}}_j/N = \hat{\delta}'\hat{\delta}_j + \widehat{V}'\widehat{V}_j/N \ ,$$

where $\hat{v}\underline{\hat{\lambda}}_j = 0$ in terms of the orthonormal eigen-basis by construction for $j = 1,\ldots,r$. The two conditions $\widehat{V}'\widehat{V}/N \le 1$ and $\widehat{V}_j'\widehat{V}_j \xrightarrow{p} 0$ together lead to that $\widehat{V}'\widehat{V}_j \xrightarrow{p} 0$.

Thus, it results in that $\hat{\delta}'\hat{\delta}_j \xrightarrow{p} 0$ for $j = 1,\dots,r$, or, equivalently, $\hat{\delta}'(\hat{\delta}_1\ \hat{\delta}_2 \cdots \hat{\delta}_r) \xrightarrow{p} \mathbf{0}$. Moreover, we know that $(\hat{\delta}_1\ \hat{\delta}_2 \cdots \hat{\delta}_r) \xrightarrow{p} I_r$, which makes $\hat{\delta} \xrightarrow{p} \vec{0}$. Finally, we obtain that

$$R(\hat{v})^* = \hat{\delta}' C_{NT} \hat{\delta} \xrightarrow{p} 0 \ ,$$

where $C_{NT} = (\Lambda'\Lambda/N)^{\frac{1}{2}'}(F'F/T)(\Lambda'\Lambda/N)^{\frac{1}{2}}$. Moreover, this implies that $R(\hat{v}) \xrightarrow{p} 0$ by *Lemma 1*.

□

Lemma 5 Let $\mathcal{G}_T = \{\max_{1\le i\le p}\max_{1\le j\le Np}|T^{-1}\sum_{t=1}^{T} W_{t-i,j}\varepsilon_t| \le \lambda_T/2\}$, then,

$$P(\mathcal{G}_T) \ge 1 - 2(N^2p)^{1-ln(1+T)} - 2(1+T)^{-g},$$

for $\lambda_T = \sqrt{8ln(1+T)^5 ln(1+N)^4 ln(1+p)^2 ln(N^2p)\sigma_T^4/T}$ and g as a positive constant.

Proof: This lemma is to bound the maximum of all cross products of variables and error terms, which is originally shown and proven by Kock and Callot (2015).

□

Theorem 2

Proof: By the minimizing property of $(\widehat{A},\widehat{B})$ for the object function (21), we have that

$$T^{-1}\|y - (\widehat{F}\widehat{A} + W\widehat{B})\|^2 + 2\lambda_T\|\widehat{B}\|_1 \le T^{-1}\|y - (\widehat{F}A + WB)\|^2 + 2\lambda_T\|B\|_1 \ . \qquad \text{(B.1)}$$

Using model (11), we write LHS and RHS of the inequality (B.1) as

$$\begin{aligned} LHS \ &= \ T^{-1}\|(\widehat{F}\widehat{A} + W\widehat{B}) - (FA + WB)\|^2 - 2T^{-1}\varepsilon'[(\widehat{F}\widehat{A} + W\widehat{B}) - (FA + WB)] \\ &\quad + 2\lambda_T\|\widehat{B}\|_1 + T^{-1}\|\varepsilon\|^2 \ , \end{aligned}$$

and

$$RHS = T^{-1}\|(\widehat{F} - F)A\|^2 - 2T^{-1}\varepsilon'(\widehat{F}A - FA) + 2\lambda_T\|B\|_1 + T^{-1}\|\varepsilon\|^2 \ .$$

Note that $T^{-1}\|(\widehat{F}\widehat{A} + W\widehat{B}) - (FA + WB)\|^2 = T^{-1}\|\widehat{Z}\widehat{\Omega} - Z\Omega\|^2$ which is of our term of interest. Thus, we rearrange inequality (B.1) to obtain that

$$\begin{aligned} T^{-1}\|\widehat{Z}\widehat{\Omega} - Z\Omega\|^2 \ &\le \ \underbrace{T^{-1}\|(\widehat{F} - F)A\|^2}_{\text{(i)}} + \underbrace{2T^{-1}\varepsilon'\widehat{F}(\widehat{A} - A)}_{\text{(ii)}} + \underbrace{2T^{-1}\varepsilon' W(\widehat{B} - B)}_{\text{(iii)}} \\ &\quad + 2\lambda_T(\|B\|_1 - \|\widehat{B}\|_1) \ . \end{aligned}$$

First, we expand (i) as follows,

$$\begin{aligned} \text{(i)} \ &= \ T^{-1}\Big\{\sum_{j=1}^{r} A_j^2 \sum_{t=1}^{T}(S_j\widehat{F}_{jt} - F_{jt})^2 + \sum_{j=r+1}^{\hat{r}} A_j^2 \sum_{t=1}^{T}\widehat{F}_{jt}^2\Big\} \\ &\le \ \bar{A}^2\Big\{\underbrace{T^{-1}\sum_{j=1}^{r}\sum_{t=1}^{T}(S_j\widehat{F}_{jt} - F_{jt})^2}_{\to 0 \text{ by } \textit{Theorem 1(b)}} + \underbrace{T^{-1}\sum_{j=r+1}^{\hat{r}}\sum_{t=1}^{T}\widehat{F}_{jt}^2}_{\to 0 \text{ by } \textit{Theorem 1(c)}}\Big\} \to 0 \ , \end{aligned}$$

where $\bar{A} = \max_j A_j$ and the convergence holds according to *A.3(c)*. We then write $T^{-1}\varepsilon'\widehat{F} = T^{-1}\varepsilon'(\widehat{F} - F) + T^{-1}\varepsilon' F$, in which the first term converges to zero by *A.3(a)* and *Theorem 1(a)* and the second term converges to zero as well in terms of *A.3(b)*. Moreover, $\|\widehat{A} - A\|_1$ is finite under the condition *A.3(c)*, thus we obtain the following by Slutsky's theorem,

$$\text{(ii)} \leq 2T^{-1}\varepsilon'\widehat{F}\|\widehat{A} - A\|_1 \xrightarrow{p} 0 \ .$$

Now, on set $\mathscr{G}_T$, we have that

$$\text{(iii)} \leq 2\max_j |T^{-1}\sum_t \varepsilon_t W_{t-1,j}| \cdot \|\widehat{B} - B\|_1 \leq \lambda_T \|\widehat{B} - B\|_1 \ .$$

Therefore, we know that

$$T^{-1}\|\widehat{Z}\widehat{\Omega} - Z\Omega\|^2 \leq \lambda_T\|\widehat{B} - B\|_1 + 2\lambda_T(\|B\|_1 - \|\widehat{B}\|_1) + o_p(1) \ .$$

Then, adding the term $\lambda_T\|\widehat{B} - B\|_1$ on both sides yields that

$$T^{-1}\|\widehat{Z}\widehat{\Omega} - Z\Omega\|^2 + \lambda_T\|\widehat{B} - B\|_1 \leq 2\lambda_T(\underbrace{\|\widehat{B} - B\|_1 + \|B\|_1 - \|\widehat{B}\|_1}_{\text{(iv)}}) + o_p(1) \ .$$

Next, we derive that

$$\text{(iv)} = \|\widehat{B}_J - B_J\|_1 + \|B_J\|_1 - \|\widehat{B}_J\|_1 \leq 2\|\widehat{B}_J - B_J\|_1 \ ,$$

where the inequality holds because $\|B_J\|_1 - \|\widehat{B}_J\|_1 \leq \|\widehat{B}_J - B_J\|_1$. So far, we have that

$$T^{-1}\|\widehat{Z}\widehat{\Omega} - Z\Omega\|^2 + \lambda_T\|\widehat{B} - B\|_1 \leq 4\lambda_T\|\widehat{B}_J - B_J\|_1 + o_p(1) \ ,$$

in which $o_p(1)$ is asymptotic negligible. Thus, straightforwardly, we obtain the following two inequalities:

$$T^{-1}\|\widehat{Z}\widehat{\Omega} - Z\Omega\|^2 \leq 4\lambda_T\|\widehat{B}_J - B_J\|_1 \ ; \tag{B.2}$$

$$\|\widehat{B} - B\|_1 \leq 4\,\|\widehat{B}_J - B_J\|_1 \ . \tag{B.3}$$

Furthermore, inequality (B.3) is equivalent to

$$\|\widehat{B}_{J^C} - B_{J^C}\|_1 \leq 3\|\widehat{B}_J - B_J\|_1 \ ,$$

which satisfies the condition of *Restricted Eigenvalues*. Additionally, it also provides that

$$\kappa_W(s) \leq \frac{\|T^{-1}W(\widehat{B} - B)\|}{\|\widehat{B}_J - B_J\|} \leq \frac{\|T^{-1}(\widehat{Z}\widehat{\Omega} - Z\Omega)\|}{\|\widehat{B}_J - B_J\|} \ ,$$

or equivalently,

$$\|\widehat{B}_J - B_J\| \leq \frac{\|T^{-1}(\widehat{Z}\widehat{\Omega} - Z\Omega)\|}{\kappa_W(s)} \ .$$

Together with the fact that

$$\|\widehat{B}_J - B_J\|_1 \leq \sqrt{s}\|\widehat{B}_J - B_J\| \ ,$$

we can rewrite inequality (B.2) as

$$\begin{aligned} T^{-1}\|\widehat{Z}\widehat{\Omega}-Z\Omega\|^2 &\leq 4\sqrt{s}\lambda_T\|\widehat{B}_J-B_J\| \\ &\leq 4\sqrt{s}\lambda_T\frac{\|T^{-1}(\widehat{Z}\widehat{\Omega}-Z\Omega)\|}{\kappa_W(s)} \ . \end{aligned}$$

Supposing that $0<c<\min\{\kappa_W(s),\kappa_Z(s+r)\}$, we obtain the following inequality on $\mathscr{G}_{\mathscr{T}}$,

$$T^{-1}\|\widehat{Z}\widehat{\Omega}-Z\Omega\|^2 \leq 16\, s\lambda_T^2/c^2 \ . \tag{B.4}$$

Moreover, inequality (B.3) implies that

$$\|\widehat{\Omega}-\Omega\|_1 \leq 4\,\|\widehat{\Omega}_J-\Omega_J\|_1$$

and

$$\|\widehat{\Omega}_{J^C}-\Omega_{J^C}\|_1 \leq 3\,\|\widehat{\Omega}_J-\Omega_J\|_1 \ ,$$

which satisfies the condition of *Restricted Eigenvalues* for the Gram matrix Ψ_Z. Therefore, it gives that

$$\kappa_Z(s+r) \leq \frac{\|T^{-1}Z(\widehat{\Omega}-\Omega)\|}{\|\widehat{\Omega}_J-\Omega_J)\|} \ ,$$

or equivalently,

$$\|\widehat{\Omega}_J-\Omega_J)\,| \leq \frac{\|T^{-1}Z(\widehat{\Omega}-\Omega)\|}{\kappa_Z(s+r)} \ .$$

Furthermore, we know that

$$\begin{aligned} \|\widehat{\Omega}-\Omega\|_1 &\leq 4\sqrt{s+r}\|\widehat{\Omega}_J-\Omega_J\| \\ &\leq 4\sqrt{s+r}\frac{\|T^{-1}Z(\widehat{\Omega}-\Omega)\|}{\kappa_Z(s+r)} \ , \end{aligned}$$

where

$$\begin{aligned} \|T^{-1}Z(\widehat{\Omega}-\Omega)\| &\leq \|T^{-1}(\widehat{Z}\widehat{\Omega}-Z\Omega)\| + \|T^{-1}(\widehat{F}-F)A\| \\ &= \|T^{-1}(\widehat{Z}\widehat{\Omega}-Z\Omega)\| + o_p(1) \ , \end{aligned}$$

in terms of the proven result that (i) $\xrightarrow{p} 0$. Then, using the result (B.4), we obtain that

$$\|\widehat{\Omega}-\Omega\|_1 \leq \frac{16\sqrt{1+r/s}}{c^2}s\lambda_T + o_p(1) \ . \tag{B.5}$$

The assumption that $N,p\in O(e^{T^a})$ for some $a\geq 0$ suggests the following,

$$s^2\lambda_T^2 \in O(\ln(1+T)^5T^{7a+2b-1}) \sim o(1) \ ,$$

which implies that $s\lambda_T \to 0$. Therefore, we can conclude that $\|\widehat{\Omega}-\Omega\|_1 \xrightarrow{p} 0$ according to the inequality (B.5), which proves *Theorem 2(b)*. Moreover, the fact that $s\lambda_T \to 0$ can lead to $\lambda_T<1$ from a time point T in the future. It indicates that $s\lambda_T^2 \to 0$, and $T^{-1}\|\widehat{Z}\widehat{\Omega}-Z\Omega\|^2 \xrightarrow{p} 0$ in terms of the inequality (B.4), which shows the statement *Theorem 2(a)*.

□

References

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

Belviso, F. and Milani, F. (2006). Structural factor-augmented vars (sfavars) and the effects of monetary policy. *Topics in Macroeconomics*, 6(3).

Bernanke, B. S. and Blinder, A. S. (1992). The federal funds rate and the channels of monetary transmission. *The American Economic Review*, pages 901–921.

Bernanke, B. S., Boivin, J., and Eliasz, P. (2004). Measuring the effects of monetary policy: a factor-augmented vector autoregressive (favar) approach. Technical report, National Bureau of Economic Research.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Callot, A. F. L. (2012). Large panels and high-dimensional vector autoregressive models. Technical report.

Christiano, L. J., Eichenbaum, M., and Evans, C. (2001). Nominal rigidities and the dynamic effects of a shock to monetary policy. Technical report, National bureau of economic research.

Connor, G. and Korajczyk, R. A. (1993). A test for the number of factors in an approximate factor model. *The Journal of Finance*, 48(4):1263–1291.

Donoho, D. L. et al. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, pages 1–32.

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

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2003). Do financial variables help forecasting inflation and real activity in the euro area? *Journal of Monetary Economics*, 50(6):1243–1255.

Kock, A. B. and Callot, A. F. L. (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, 186(2):325–344.

Leeper, E. M., Sims, C. A., Zha, T., Hall, R. E., and Bernanke, B. S. (1996). What does monetary policy do? *Brookings papers on economic activity*, pages 1–78.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica: Journal of the Econometric Society*, pages 1–48.

Song, S. and Bickel, P. J. (2011). Large vector auto regressions. *arXiv preprint arXiv:1106.3915*.

Stock, J. H. and Watson, M. W. (2001). Vector autoregressions. *Journal of Economic perspectives*, pages 101–115.

Stock, J. H. and Watson, M. W. (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, 97(460):1167–1179.

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

Stock, J. H. and Watson, M. W. (2005). Implications of dynamic factor models for var analysis. Technical report, National Bureau of Economic Research.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.