Consistent Factor Estimation and Forecasting in Factor-Augmented VAR Models*

John C. Chao[1], Yang Liu[2] and Norman R. Swanson[2]

[1]University of Maryland and [2]Rutgers University

July 19, 2023

preliminary and incomplete

Abstract

In this paper we establish that conditional mean functions associated with h-step ahead forecasting equations implied by a factor augmented vector autoregressions (FAVARs) can be consistently estimated in the case where factor pervasiveness does not hold. In particular, we begin by stating a common assumption of factor pervasiveness in which all available predictor variables (excepting a negligible subset) load significantly on the underlying factors. We then establish that even when this assumption is relaxed, consistent factor estimation can be achieved if one pre-screens the variables and successfully prunes out the irrelevant ones. Furthermore, use of factors estimated in this manner when constructing h-step ahead forecasting equations implied by FAVAR models enables the consistent estimation of the conditional mean function of said equations, and conditional mean functions constructed used our precedure are consistently estimable in a wide range of situations, including cases where violation of factor pervasiveness is such that consistent estimation is precluded in the absence of variable pre-screening.

Keywords: Factor analysis, factor augmented vector autoregression, forecasting, moderate deviation, principal components, self-normalization, variable selection.

JEL Classification: C32, C33, C38, C52, C53, C55.

**Corresponding Author:* John C. Chao, Department of Economics, 7343 Preinkert Drive, University of Maryland, chao@econ.umd.edu.

Norman R. Swanson, Department of Economics, 9500 Hamilton Street, Rutgers University, nswanson@econ.rutgers.edu. The authors are grateful to Simon Freyaldenhoven, Yuan Liao, Minchul Shin, Xiye Yang, and seminar participants at the Federal Reserve Bank of Pihadelphia for useful comments received on earlier versions of this paper. Chao thanks the University of Maryland for research support.

1 Introduction

In economics, three of the key areas in machine learning that have drawn considerable attention in recent years include variable selection, dimension reduction and shrinkage. One reason for this is the availability of new high frequency and high dimensional datasets that are being analyzed in areas ranging from targeted marketing and customer segmentation to forecasting and maroeconomic policy making. This has in turn led to numerous theoretical advances in the areas of estimation, implementation, and inference using techinques such as the least absolute shrinkage operator (lasso) and principal components analysis (PCA). In this paper, we build on pathbreaking work due to Bai and Ng (2002), Stock and Watson (2002a,b), Bai (2003), Forni, Hallin, Lippi, and Reichlin (2005), and Bai and Ng (2008), in which methods for constructing forecasts based on factor-augmented regression models are developed and analyzed. In particular, we establish that latent factors that are critical to the estimation of factor augmented vector autoregressions (FAVARs) can be consistently estimated in cases where factor pervasiveness does not hold, where by factor pervasivness we mean that (almost) all available predictors load significantly on a set of factors that we wish to estimate. To do so, we draw on results of Chao and Swanson (CS: 2022a), where a completely consistent variable selection procedure useful for specifying FAVAR models is developed. We then establish that the conditional mean of the infeasible h-step ahead forecasting equation implied by an FAVAR can be consistently estimated.

As discussed above, a key assumption commonly used in the factor analysis literature to show consistent factor estimation is that of factor pervasiveness. This assumption presupposes that all available predictor variables in a dataset, with the possible exception of a negligible number of them, load significantly on the underlying factors. Needless to say, this assumption may not be satisfied by many datasets that are available for empirical research. Indeed, a likely scenario is that there is significant underlying heterogeneity, so that some of the available variables are relevant in the sense that they load significantly on the underlying factors, whereas others are irrelevant, in the sense that they do not share any common dynamic structure with each other or with the relevant variables in the dataset. In this paper, we begin by establishing that, under failure of factor pervasiveness in a stylized model with one factor, consistency cannot be achieved, and indeed $\widehat{f}_t \xrightarrow{p} 0$, as $N, T \to \infty$, where f_t is a latent factor, N is the number of variables in the dataset being modelled, and T is the number of time series observations. Findings such as this are the impetus for the work of Chao and Swanson (CS: 2022a), where a variable selection procedure is developed for pre-selecting relevant predictor variables for use in the consistent estimation of

latent factors in an FAVAR model. Their variable selection procedure is based on the use of easy to construct self normalized statistics measuring the covariation between target variables to be predicted and possible predictor variables to be used in factor estimation. CS (2022a) show that for their procedure, the probability of Type I and Type II errors goes to zero, asymptotically, implying that the procedure is completely consistent. This property turns out to be important because if one tries to simply control the probability of a Type I error at some predetermined level, which is the typical approach used in multiple hypothesis testing, then one will not in general be able to estimate factors consistently, even up to an invertible matrix transformation. A main result of the current paper is to show that factors estimated using predictor variables selected using the procedure of CS (2022a) are consistent, up to a rotation. With these results in hand, we then show that by using variables selected via our pre-screening procedure to estimate the underlying factors, and then inserting these factor estimates into h-step ahead forecasting equations implied by a FAVAR model, we can consistently estimate the conditional mean function of the said equations. Importantly, we argue that this result allows the conditional mean function of a factor-augmented forecasting equation to be consistently estimable in a wide range of situations, and in particular in situations where there are violations of factor pervasiveness.

Finally, in order to illustrate the methods discussed in this paper, we analyze a large dataset. This part of the paper is to be completed.

Some of the research reported here is related to the well-known supervised principal components method proposed by Bair, Hastie, Paul, and Tibshirani (2006). Additionally, our research is related to some interesting recent work by Giglio, Xiu, and Zhang (2021), who propose a method for selecting test assets, with the objective of estimating risk premia in a Fama-MacBeth type framework. A crucial difference between the variable selection procedure proposed in our paper and those proposed in these papers is that we use a score statistic that is self-nomalized, whereas the aforementioned papers do not make use of statistics that involve self-normalization. An important advantage of self-normalized statistics is their ability to accommodate a much wider range of possible tail behavior in the underlying distributions, relative to their non-self-normalized counterparts. This makes self-normalized statistics better suited for various types of economic and financial applications, where the data are known not to exhibit the type of exponentially decaying tail behavior assumed in much of the statistics literature on high-dimensional models. In addition, the type of models studied in Bair, Hastie, Paul, and Tibshirani (2006) and Giglio, Xiu, and Zhang (2021) differ significantly from the FAVAR model studied here. In particular, Bair, Hastie, Paul, and Tibshirani (2006) study a one-factor model in an *i.i.d.* Gaussian framework so that complications introduced by dependence and non-normality of distribution are not considered in their paper. Giglio, Xiu, and Zhang (2021) do make certain high-level assumptions which may potentially accommodate some

dependence both cross-sectionally and intertemporally, but they do not consider the implications of variable selection and factor estimation for forecasting, and the model that they consider is very different from the type of dynamic vector time series model studied here.

Our research is also closely related to the work of Bai and Ng (2021), who provide results which show that factors can still be estimated consistently in certain situations where the factor loadings are weaker than that implied by the conventional pervasiveness assumption, although in such cases the rate of convergence of the factor estimator is slower and additional assumptions are needed. As discussed in the next section of this paper, their factor consistency result relies on a key condition, and the appropriateness of this condition depends on how severely the condition of factor pervasiveness is violated, which is ultimately an empirical issue.[1]

The rest of the paper is organized as follows. In Section 2 , we provide our counterexample, stated formally as Theorem 2.1, which shows that a latent factor may be inconsistently estimated when the standard assumption of factor pervasiveness does not hold. In Section 3, we discuss the FAVAR model, the variable selection procedure of CS (2022a), and the assumptions that are required in the sequel. Section 4 gathers our theoretical results on the consistent estimation of latent factors, up to an invertible matrix transformation, as well as results on the consistent estimation of the h-step ahead predictor, based on the FAVAR model. Section 5 presents the results of an empirical illustration where our forecasting approach is compared with related approaches in the literature. Finally, Section 6 offers concluding remarks. Proofs of the main theorems and supporting lemmas are given in an appendix as well as a separate online technical supplement (see Chao and Swanson (2022c).

Before proceeding, we first say a few words about some of the frequently used notation in this paper. Throughout, let $\lambda_{(j)}(A)$, $\lambda_{\max}(A)$, $\lambda_{\min}(A)$, and $tr(A)$ denote, respectively, the j^{th} largest eigenvalue, the maximal eigenvalue, the minimal eigenvalue, and the trace of a square matrix A. Similarly, let $\sigma_{(j)}(B)$, $\sigma_{\max}(B)$, and $\sigma_{\min}(B)$ denote, respectively, the j^{th} largest singular value, the maximal singular value, and the minimal singular value of a matrix B, which is not restricted to be a square matrix. In addition, let $\|a\|_2$ denote the usual Euclidean norm when applied to a (finite-dimensional) vector a. Also, for a matrix A, $\|A\|_2 \equiv \max\left\{\sqrt{\lambda(A'A)} : \lambda(A'A) \text{ is an eigenvalue of } A'A\right\}$ denotes the matrix spectral norm, and $\|A\|_F \equiv \sqrt{tr\{A'A\}}$ denotes the Frobenius norm. For two random variables X and Y, write $X \sim Y$, if $X/Y = O_p(1)$ and $Y/X = O_p(1)$. Furthermore, let $|z|$ denote the absolute value or the modulus of the number z; let $\lfloor \cdot \rfloor$ denote the floor function, so

[1] Various authors have documented cases in economics-related research where empirical results suggest that the underlying factors may be quite weak, so that the rate condition given in Bai and Ng (2021) may not be appropriate. See, for example, the discussions in Jagannathan and Wang (1998), Kan and Zhang (1999), Harding (2008), Kleibergen (2009), Ontaski (2012), Bryzgalova (2016), Burnside (2016), Gospodinov, Kan, and Robotti (2017), Anatolyev and Mikusheva (2021), and Freyaldenhoven (2021a,b).

that $\lfloor x \rfloor$ gives the integer part of the real number x, and let $\iota_p = (1, 1, ..., 1)'$ denote a $p \times 1$ vector of ones. Finally, the abbreviation w.p.a.1 stands for "with probability approaching one".

2 Inconsistency in High-Dimensional Factor Estimation

To provide some motivation for the problem we will be studying in this paper, consider the following simple, stylized one-factor model:

$$\underset{N\times 1}{Z_t} = \underset{N\times 1}{\gamma}\ \underset{1\times 1}{f_t} + \underset{N\times 1}{u_t}\ , t = 1, ..., T \tag{1}$$

for which we make the following assumption.

Assumption 2-1: (a) $\{u_t\} \equiv i.i.d.N(0, I_N)$;(b) $\{f_t\} \equiv i.i.d.N(0, 1)$;and (c) u_s and f_t are independent for all t, s.

Much of the literature on factor analysis focuses on the case where the factors are pervasive. In the special case of the simple one factor model given in expression (1) above, pervasiveness means that:

$$\frac{\|\gamma\|_2^2}{N} \to c,$$

for some constant c such that $0 < c < \infty$, where $\|\gamma\|_2 = \sqrt{\gamma'\gamma}$. In practice, however, one may have a high-dimensional data vector Z_t such that not all of the components of Z_t load significantly on the underlying factor, f_t. In particular, let $\mathcal{P}$ be a permutation matrix which reorders the components of Z_t, so that $\mathcal{P}Z_t$ can be partitioned as follows:

$$\mathcal{P}Z_t = \begin{pmatrix} \underset{N_1\times 1}{Z_t^{(1)}} \\ \underset{N_2\times 1}{Z_t^{(2)}} \end{pmatrix},$$

where $Z_t^{(1)} = \gamma^{(1)} f_t + u_t^{(1)}$ and $Z_t^{(2)} = u_t^{(2)}$ and where all components of the $N_1 \times 1$ vector $\gamma^{(1)}$ are different from zero, so that the components of $Z_t^{(1)}$ all load significantly on f_t, whereas the components of $Z_t^{(2)}$ do not. Of course, an empirical researcher will not typically have à priori knowledge as to which components of Z_t will load significantly on f_t and which will not. The following result shows that if one proceeds with factor estimation assuming that the factor is pervasive, then the usual estimator of a factor based on principal component methods may be inconsistent and may, in fact, behave in a rather pathological manner in large samples. To consider this possibility, assume the following condition, which implies a violation of the pervasiveness assumption.

Assumption 2-2: As $N, T \to \infty$, let $\|\gamma\|_2 \to \infty$ such that:

$$\frac{N}{T\|\gamma\|_2^{2(1+\kappa)}} = c + o\left(\frac{1}{\|\gamma\|_2^2}\right),$$

for some constant c, such that $0 < c < \infty$, and for some constant κ, such that $0 < \kappa < 1$. Note that under Assumption 2-2:

$$\frac{\|\gamma\|_2^2}{N} \sim (TN^{\kappa})^{-\frac{1}{(1+\kappa)}} \to 0 \text{ as } N, T \to \infty,$$

so that the factor does not satisfy the pervasiveness assumption. This can, of course, occur if a significant proportion of the components of γ are zero or are very small. Next, let $\widehat{\pi}_1 / \|\widehat{\pi}_1\|_2$ denote the (normalized) eigenvector associated with the largest eigenvalue of the sample covariance matrix, $\widehat{\Sigma}_Z = \mathbf{Z}'\mathbf{Z}/T$, where $\mathbf{Z} = (Z_1, ..., Z_T)'$. Then, the usual principal component estimator of f_t is given by:

$$\widehat{f}_t = \frac{\langle \widehat{\pi}_1, Z_t \rangle}{\sqrt{N}\|\widehat{\pi}_1\|_2}.$$

The following theorem characterizes the asymptotic behavior of this estimator under the assumptions given above.

Theorem 2.1: *Suppose that Assumptions 2-1 and 2-2 hold. Then, for all t:* $\widehat{f}_t \xrightarrow{p} 0$, as $N, T \to \infty$.

It is well-known that without further identifying assumptions, such as those given in Assumption F1 of Stock and Watson (2002a), factors can only be estimated consistently up to an invertible matrix transformation. However, even in cases where we are not willing to specify enough conditions so as to fully identify the factors, estimating the factors consistently up to an invertible matrix transformation will often suffice for many purposes. One such case is when we are trying to forecast using a factor-augmented vector autoregression (FAVAR). As we will show in results given in Section 4 of this paper, point forecasts constructed using factors which are estimated consistently up to an invertible matrix transformation will nevertheless converge in probability to the desired infeasible forecast (i.e., the conditional mean of the FAVAR), that in turn depends on the true unobserved factors. On the other hand, the problem illustrated by the result given in Theorem 1 is different and is in some sense more problematic and pathological. The estimated factor in Theorem 1 converges to zero regardless of what happens to be the realized value of the true latent factor. In this case, one clearly cannot consistently estimate the conditional mean of the FAVAR.

Theorem 1 is related to results previously given in the statistics literature showing the possible inconsistency of sample eigenvectors as estimators of population eigenvectors in high dimensional situations. See, for example, Paul (2007), Johnstone and Lu (2009), Shen, Shen, Zhu, and Marron

(2016), and Johnstone and Paul (2018). However, most of the results in the statistics literature are not explicitly framed in the setting of a factor model, but are instead derived for the related spiked covariance model. Theorem 1 is intended to give an inconsistency result of this type, but in a context that may be more familiar to researchers in economics.

It should also be noted that, in an interesting and thought-provoking recent paper, Bai and Ng (2021) provide results which show that factors can still be estimated consistently in certain situations where the factor loadings are weaker than that implied by the conventional pervasiveness assumption, but that in such cases the rate of convergence is slower and additional assumptions are needed. To understand the relationship between their results and the example given above, note that a key condition for the consistency result given in their paper, when expressed in terms of our notation, is the assumption that $N/\left(T\left\|\gamma\right\|_2^2\right) \to 0$[2]. On the other hand, if $N/\left(T\left\|\gamma\right\|_2^2\right) \to c_1$, for some positive constant c_1, or even worse, if $N/\left(T\left\|\gamma\right\|_2^2\right) \to \infty$, which is essentially what is specified in Assumption 2-2 above, then consistent factor estimation cannot be achieved[3]. Hence, whether or not consistent factor estimation can be attained depends on how nonpervasive the factors are, which is ultimately an empirical question, and which depends on the application and on the dataset employed. Moreover, various authors have now documented cases where empirical results suggest that the underlying factors may be quite weak, so that the rate condition given in Bai and Ng (2021) may not be appropriate, at least for some of the situations for which factor modeling is of interest. For example, see Jagannathan and Wang (1998), Kan and Zhang (1999), Harding (2008), Kleibergen (2009), Ontaski (2012), Bryzgalova (2016), Burnside (2016), Gospodinov, Kan, and Robotti (2017), Anatolyev and Mikusheva (2021), and Freyaldenhoven (2021a,b). In such cases, it is of interest to explore the possibility that the weakness in the loadings is not uniform across all variables, but rather is due to the fact that only a small percentage of the variables loads significantly on the underlying factors. Furthermore, even if the empirical situation of interest is one where, strictly speaking, the condition $N/\left(T\left\|\gamma\right\|_2^2\right) \to 0$ does hold, it may still be beneficial in some such instances to do variable pre-screening. This is particularly true in situations where the condition $N/\left(T\left\|\gamma\right\|_2^2\right) \to 0$ is "barely" satisfied, in which case one would expect to pay a rather hefty finite sample price for not pruning out variables that do not load significantly on the underlying factors, since these variables will add unwanted noise to the estimation process. For all these reasons, there is a clear need to develop methods that will enable empirical researchers

[2] See Assumption A4 of Bai and Ng (2021). Note that Bai and Ng (2021) state this condition in the form $N/\left(TN^{\alpha}\right) \to 0$, for some $\alpha \in (0,1]$, but since part (ii) of their Assumption A2, when specialized to the one factor model studied here, simplifies to the condition that $\lim_{N\to\infty} \left\|\gamma\right\|_2^2 / N^{\alpha} = \sigma_{\Lambda} > 0$, it is easy to see that their Assumption A4 is equivalent to the condition that $N/\left(T\left\|\gamma\right\|_2^2\right) \to 0$.

[3] Note that Assumption 2-2 is actually stronger than required in order to show inconsistency, but that we impose this condition to highlight the fact that, in this case, not only is the estimator of the factor inconsistent but it actually converges to zero.

to pre-screen the components of Z_t, so that variables which are informative and helpful to the estimation process can be properly identified.

3 Model, Assumptions, and Variable Selection in High Dimensions

Following CS (2022a), we begin by considering the following p^{th}-order factor-augmented vector autoregression (FAVAR):

$$W_{t+1} = \mu + A_1 W_t + \cdots + A_p W_{t-p+1} + \varepsilon_{t+1}, \tag{2}$$

where

$$\underset{(d+K)\times 1}{W_{t+1}} = \begin{pmatrix} \underset{d\times 1}{Y_{t+1}} \\ \underset{K\times 1}{F_{t+1}} \end{pmatrix}, \ \underset{(d+K)\times 1}{\varepsilon_{t+1}} = \begin{pmatrix} \underset{d\times 1}{\varepsilon^Y_{t+1}} \\ \underset{K\times 1}{\varepsilon^F_{t+1}} \end{pmatrix}, \ \underset{(d+K)\times 1}{\mu} = \begin{pmatrix} \underset{d\times 1}{\mu_Y} \\ \underset{K\times 1}{\mu_F} \end{pmatrix}, \text{ and}$$

$$\underset{(d+K)\times(d+K)}{A_g} = \begin{pmatrix} \underset{d\times d}{A_{YY,g}} & \underset{d\times K}{A_{YF,g}} \\ \underset{K\times d}{A_{FY,g}} & \underset{K\times K}{A_{FF,g}} \end{pmatrix}, \text{ for } g = 1, ..., p.$$

This system of equations, where Y_t denotes the vector of observable economic variables, and F_t is a vector of unobserved (latent) factors can also be written is several alternative ways, the following two of which are are variously used throughout this paper. Namely:

$$Y_{t+1} = \mu_Y + A_{YY}\underline{Y}_t + A_{YF}\underline{F}_t + \varepsilon^Y_{t+1}, \tag{3}$$

$$F_{t+1} = \mu_F + A_{FY}\underline{Y}_t + A_{FF}\underline{F}_t + \varepsilon^F_{t+1}, \tag{4}$$

where

$$\underset{d\times dp}{A_{YY}} = \begin{pmatrix} A_{YY,1} & A_{YY,2} & \cdots & A_{YY,p} \end{pmatrix}, \ \underset{d\times Kp}{A_{YF}} = \begin{pmatrix} A_{YF,1} & A_{YF,2} & \cdots & A_{YF,p} \end{pmatrix},$$

$$\underset{K\times dp}{A_{FY}} = \begin{pmatrix} A_{FY,1} & A_{FY,2} & \cdots & A_{FY,p} \end{pmatrix}, \ \underset{K\times Kp}{A_{FF}} = \begin{pmatrix} A_{FF,1} & A_{FF,2} & \cdots & A_{FF,p} \end{pmatrix},$$

$$\underset{dp\times 1}{\underline{Y}_t} = \begin{pmatrix} Y_t \\ Y_{t-1} \\ \vdots \\ Y_{t-p+1} \end{pmatrix}, \text{ and } \underset{Kp\times 1}{\underline{F}_t} = \begin{pmatrix} F_t \\ F_{t-1} \\ \vdots \\ F_{t-p+1} \end{pmatrix}, \tag{5}$$

and

$$\underset{(d+K)p\times 1}{\underline{W}_t} = \alpha + A\underline{W}_{t-1} + E_t,$$

where $\underline{W}_t = \begin{pmatrix} W_t' & W_{t-1}' & \cdots & W_{t-p+2}' & W_{t-p+1}' \end{pmatrix}'$ and where

$$\alpha = \begin{pmatrix} \mu \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, A = \begin{pmatrix} A_1 & A_2 & \cdots & A_{p-1} & A_p \\ I_{d+K} & 0 & \cdots & 0 & 0 \\ 0 & I_{d+K} & \ddots & \vdots & 0 \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & I_{d+K} & 0 \end{pmatrix}, \text{ and } E_t = \begin{pmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}. \tag{6}$$

The companion form given in equation (6) is convenient for establishing certain moment conditions on $\underline{Y}_t$ and $\underline{F}_t$, given a moment condition on ε_t, and for establishing certain mixing properties of the FAVAR model, as shown in the proofs of Lemmas C-5 and Lemma C-11 given in Chao and Swanson (2022b). It remains to define the relationship between the F_t and the variables used to extract these factors. To do this, we assume that:

$$\underset{N\times 1}{Z_t} = \underset{N\times Kp}{\Gamma}\underline{F}_t + u_t, \tag{7}$$

where the properties of u_t are given in Assumptions 3-3 and 3-4, below. Following Chao and Swanson (2022a), we assume that not all components of Z_t provide useful information for estimating $\underline{F}_t$, implying that the $N \times Kp$ parameter matrix Γ may have some rows whose elements are all zero. More precisely, let the $1 \times Kp$ vector, γ_i', denote the i^{th} row of Γ, and assume that the rows of the matrix Γ can be divided into two classes:

$$H = \{k \in \{1, ..., N\} : \gamma_k = 0\} \text{ and} \tag{8}$$

$$H^c = \{k \in \{1, ..., N\} : \gamma_k \neq 0\}. \tag{9}$$

Thus, there exists a permutation matrix $\mathcal{P}$ such that $\mathcal{P}Z_t = \begin{pmatrix} Z_t^{(1)\prime} & Z_t^{(2)\prime} \end{pmatrix}'$, where

$$\underset{N_1\times 1}{Z_t^{(1)}} = \Gamma_1\underline{F}_t + u_t^{(1)} \tag{10}$$

$$\underset{N_2\times 1}{Z_t^{(2)}} = u_t^{(2)}. \tag{11}$$

In this way, the components of $Z_t^{(1)}$ can be interpreted as "information" variables that are useful for estimating $\underline{F}_t$. On the other hand, for the purpose of factor estimation, the components of the subvector $Z_t^{(2)}$ are pure "noise" variables, as they do not load on the underlying factors and only add noise if they are included in the factor estimation process. Given that an empirical researcher will often not have prior knowledge as to which variables are elements of $Z_t^{(1)}$ and which are elements of $Z_t^{(2)}$, Theorem 2.1 suggests the need for a variable selection procedure which will allow us to properly identify the components of of $Z_t^{(1)}$ and to use only these variables when we try to estimate $\underline{F}_t$; for, if we unknowingly include too many components of $Z_t^{(2)}$ in the estimation process, then inconsistent estimation in the sense described in the previous section can result.[4] As discussed in CS (2022a), there is an important related paper by Bai and Ng (2021) that establishes factor estimator consistency for cases where $N/(TN_1) \to 0$. For cases where $N/(TN_1) \to c$, or $N/(TN_1) \to \infty$, where c is a constant, their result does not hold. In this paper, we establish that consistency can be achieved in our context even if $N/(TN_1) \nrightarrow 0$, if one pre-screens variables using the self-normalized statistics outlined below. This is important because the degree of factor pervasiveness is ultimately data dependent, and one way to estimate N_1 invloves utilizing the variable screening statistic that is discussed in the sequel.

In the sequel, we require the following assumptions.

Assumption 3-1: Suppose that:

$$\det\left\{I_{(d+K)} - A_1 z - \cdots - A_p z^p\right\} = 0, \text{ implies that } |z| > 1. \tag{12}$$

Assumption 3-2: Let ε_t satisfy the following set of conditions: (a) $\{\varepsilon_t\}$ is an independent sequence of random vectors with $E[\varepsilon_t] = 0 \ \forall t$; (b) there exists a positive constant C such that $\sup_t E\|\varepsilon_t\|_2^6 \leq C < \infty$; (c) ε_t admits a density g_{ε_t} such that, for some positive constant $M < \infty, \sup_t \int |g_{\varepsilon_t}(\upsilon - u) - g_{\varepsilon_t}(\upsilon)| \, d\varepsilon \leq M|u|$, whenever $|u| \leq \overline{\kappa}$ for some constant $\overline{\kappa} > 0$; and (d) there exists a constant $\underline{C} > 0$ such that $\inf_t \lambda_{\min}\{E[\varepsilon_t \varepsilon_t']\} \geq \underline{C} > 0$.

Assumption 3-3: Let $u_{i,t}$ be the i^{th} element of the error vector u_t in expression (7), and we assume that it satisfies the following conditions: (a) $E[u_{i,t}] = 0$ for all i and t; (b) there exists a positive constant $\overline{C}$ such that $\sup_{i,t} E|u_{i,t}|^7 \leq \overline{C} < \infty$, and there exists a constant $\underline{C} > 0$ such that $\inf_{i,t} E\left[u_{i,t}^2\right] \geq \underline{C}$; (c) define $\mathcal{F}_{i,-\infty}^t = \sigma(...., u_{i,t-2}, u_{i,t-1}, u_t)$, $\mathcal{F}_{i,t+m}^{\infty} = \sigma(u_{i,t+m}, u_{i,t+m+1}, u_{i,t+m+2}, ...)$, and

$$\beta_i(m) = \sup_t E\left[\sup\left\{\left|P\left(B|\mathcal{F}_{i,-\infty}^t\right) - P(B)\right| : B \in \mathcal{F}_{i,t+m}^{\infty}\right\}\right].$$

[4] In the statistics literature, there is a growing literature on the potential inconsistency of sample eigenvectors in high dimensional problems, as discussed in Paul (2007), Johnstone and Lu (2009), Shen Shen, Zhu, and Marron (2016), and Johnstone and Paul (2018).

Assume that there exist constants $a_1 > 0$ and $a_2 > 0$ such that

$$\beta_i(m) \leq a_1 \exp\{-a_2 m\}, \text{ for all } i;$$

and (d) there exists a positive constant C such that $\sup_t \left(\frac{1}{N_1} \sum_{i \in H^c} \sum_{k \in H^c} |E[u_{i,t} u_{k,t}]| \right) \leq C < \infty$ for every positive integer N_1, where H^c is defined in expression (9) above.

Assumption 3-4: ε_t and $u_{i,s}$ are independent, for all i, t, and s.

Assumption 3-5: There exists a positive constant $\overline{C}$, such that $\sup_{i \in H^c} \|\gamma_i\|_2 \leq \overline{C} < \infty$ and $\|\mu\|_2 \leq \overline{C} < \infty$, where $\mu = (\mu_Y', \mu_F')'$.

Assumption 3-6: There exists a positive constant $\overline{C}$, such that:

$$0 < \frac{1}{\overline{C}} \leq \lambda_{\min}\left(\frac{\Gamma'\Gamma}{N_1}\right) \leq \lambda_{\max}\left(\frac{\Gamma'\Gamma}{N_1}\right) \leq \overline{C} < \infty \text{ for all } N_1, N_2 \text{ sufficiently large,}$$

where N_1 is the number of components of the subvector $Z_t^{(1)}$ and N_2 is the number of components of the subvector $Z_t^{(2)}$, as previously defined in expressions (10) and (11).

Assumption 3-7: Let A be as defined in expression (6) above, and let the eigenvalues of the matrix $I_{(d+K)p} - A$ be sorted so that:

$$\left|\lambda_{(1)}\left(I_{(d+K)p} - A\right)\right| \geq \left|\lambda_{(2)}\left(I_{(d+K)p} - A\right)\right| \geq \cdots \geq \left|\lambda_{((d+K)p)}\left(I_{(d+K)p} - A\right)\right| = \overline{\phi}_{\min}.$$

Suppose that there is a constant $\underline{C} > 0$ such that

$$\sigma_{\min}\left(I_{(d+K)p} - A\right) \geq \underline{C}\overline{\phi}_{\min} \tag{13}$$

In addition, there exists a positive constant $\overline{C} < \infty$ such that, for all positive integer j,

$$\sigma_{\max}\left(A^j\right) \leq \overline{C} \max\left\{\left|\lambda_{\max}\left(A^j\right)\right|, \left|\lambda_{\min}\left(A^j\right)\right|\right\}. \tag{14}$$

Assumption 3-1 is the stability condition that one typically assumes for a stationary VAR process, although we allow for possible heterogeneity in the distribution of ε_t across time, so that our FAVAR process is not necessarily a strictly stationary process. Under Assumption 3-1, there exists a vector moving average representation for the FAVAR process. Assumption 3-1 is a well known assumption that is equivalent to the condition that $\det\left\{I_{(d+K)} - Az\right\} = 0$ implies that $|z| > 1$.

Since the factor loading matrix Γ is an $N \times Kp$ matrix, where $N = N_1 + N_2$, the matrix $\Gamma'\Gamma$ will have order of magnitude equal to N if the factors are pervasive. Much of the factor analysis

literature in both econometrics and statistics has studied the case where factors are pervasive in this sense. For example, see Bai and Ng (2002), Stock and Watson (2002a), Bai (2003), and Fan, Liao, and Mincheva (2011, 2013). Assumption 3-6 allows for possible violations of this conventional pervasiveness assumption, which will occur in our setup when $N_1/N \to 0$.
Finally, Assumption 3-7 imposes a condition whereby the extreme singular values of the matrices A^j and $I_{(d+K)p} - A$ have bounds that depend on the extreme eigenvalues of these matrices. For further discussion of this Assumption, see CS (2022a).

Note that Assumptions 3-1, 3-2(a)-(c), and 3-7 are sufficient to prove Lemma C-11 of Chao and Swanson (2022c)[5], which states that the process $\{W_t\}$ generated by the FAVAR model given in expression (2) is a β-mixing process with β-mixing coefficient satisfying:

$$\beta_W(m) \leq a_1 \exp\{-a_2 m\},$$

for some positive constants a_1 and a_2, with

$$\beta_W(m) = \sup_t E\left[\sup\left\{\left|P\left(B|\mathcal{A}_{-\infty}^t\right) - P(B)\right| : B \in \mathcal{A}_{t+m}^{\infty}\right\}\right],$$

and with $\mathcal{A}_{-\infty}^t = \sigma(..., W_{t-2}, W_{t-1}, W_t)$ and $\mathcal{A}_{t+m}^{\infty} = \sigma(W_{t+m}, W_{t+m+1}, W_{t+m+2}, ...)$. Note that Assumption 3-2 (c) rules out situations such as that given in the famous counterexample presented by Andrews (1984) which shows that a first-order autoregression with errors having a discrete Bernoulli distribution is not α-mixing, even if it satisfies the stability condition. Conditions similar to Assumption 3-2(c) have also appeared in previous papers, such as Gorodetskii (1977) and Pham and Tran (1985), which seek to provide sufficient conditions for establishing the α or β mixing properties of linear time series processes.

Prior to presenting the main theorems of this paper, we first summarize the variable selection procedure based on self-normalized statistics that is outlined in CS (2022a), and draws on path-breaking moderate deviation results from Chen, Shao, Wu, and Xu (2016). To accommodate data dependence, consider self-nomalized statistics that are constructed from observations which are first split into blocks in a manner similar to the kind of construction one would employ in implementing a block bootstrap or in proving a central limit theorem using the blocking technique. One such statistic has the form of an ℓ_∞ norm and is given by:

$$\max_{1 \leq \ell \leq d} |S_{i,\ell,T}| = \max_{1 \leq \ell \leq d} \left| \frac{\overline{S}_{i,\ell,T}}{\sqrt{\overline{V}_{i,\ell,T}}} \right|, \tag{15}$$

[5] A proof of Lemma C-11 was previously given in Chao and Swanson (2022b).

where

$$\overline{S}_{i,\ell,T} = \sum_{r=1}^{q} \sum_{t=(r-1)\tau+p}^{(r-1)\tau+\tau_1+p-1} Z_{it} y_{\ell,t+1} \text{ and} \tag{16}$$

$$\overline{V}_{i,\ell,T} = \sum_{r=1}^{q} \left[\sum_{t=(r-1)\tau+p}^{(r-1)\tau+\tau_1+p-1} Z_{it} y_{\ell,t+1} \right]^2 . \tag{17}$$

Here, Z_{it} denotes the i^{th} component of Z_t , $y_{\ell,t+1}$ denotes the ℓ^{th} component of Y_{t+1}, $\tau_1 = \lfloor T_0^{\alpha_1} \rfloor$, and $\tau_2 = \lfloor T_0^{\alpha_2} \rfloor$, where $1 > \alpha_1 \geq \alpha_2 > 0$, $\tau = \tau_1 + \tau_2$, $q = \lfloor T_0/\tau \rfloor$, and $T_0 = T - p + 1$. Note that the statistic given in expression (15) can be interpreted as the maximum of the (self-normalized) sample covariances between the i^{th} component of Z_t and the components of Y_{t+1}. A second statistic has the form of a pseudo-L_1 norm and is given by:

$$\sum_{\ell=1}^{d} \varpi_\ell \left| S_{i,\ell,T} \right| = \sum_{\ell=1}^{d} \varpi_\ell \left| \frac{\overline{S}_{i,\ell,T}}{\sqrt{\overline{V}_{i,\ell,T}}} \right| ,$$

where $\overline{S}_{i,\ell,T}$ and $\overline{V}_{i,\ell,T}$ are as defined in expressions (16) and (17) above and where $\{\varpi_\ell : \ell = 1, .., d\}$ denotes pre-specified weights, such that $\varpi_\ell \geq 0$, for every $\ell \in \{1, ..., d\}$ and $\sum_{\ell=1}^{d} \varpi_\ell = 1$. In order to keep the effects of dependence under control, the construction of these statistics is based only on observations in every other block. In order to consistently estimate the factors up to an invertible matrix transformation, the variable selection procedure here must be such that the the probability of a false positive and the probability of a false negative converge to zero as N_1, N_2, $T \to \infty$[6]. This is different from the typical multiple hypothesis testing approach whereby one tries to control the familywise error rate (or, alternatively, the false discovery rate), so that it is no greater than 0.05, say, but does not try to ensure that this probability goes to zero as the sample size grows.

In order to implement this procedure, it reamins only to determine whether the i^{th} component of Z_t is a relevant variable for the purpose of factor estimation. Define $i \in \widehat{H}^c$ to indicate that Z_{it} is a relevant variable and $i \in \widehat{H}$ to indicate that Z_{it} is an irrelevant variable, for factor estiamtion. Now, let $\mathbb{S}_{i,T}^{+}$ denote either the statistic $\max_{1 \leq \ell \leq d} |S_{i,\ell,T}|$ or the statistic $\sum_{\ell=1}^{d} \varpi_\ell |S_{i,\ell,T}|$. The variable selection procedure is based on the decision rule:

$$i \in \begin{cases} \widehat{H}^c & \text{if } \mathbb{S}_{i,T}^{+} \geq \Phi^{-1}\left(1 - \frac{\varphi}{2N}\right) \\ \widehat{H} & \text{if } \mathbb{S}_{i,T}^{+} < \Phi^{-1}\left(1 - \frac{\varphi}{2N}\right) \end{cases} , \tag{18}$$

[6]Here, a false positive refers to mis-classifying a variable, Z_{it}, as a relevant variable for the purpose of factor estimation when its factor loading $\gamma_i' = 0$, whereas a false negative refers to the opposite case, where $\gamma_i' \neq 0$, but the variable Z_{it} is mistakenly classified as irrelevant.

where $\Phi^{-1}(\cdot)$ denotes the quantile function or the inverse of the cumulative distribution function of the standard normal random variable, and where φ is a tuning parameter which may depend on N. Some conditions on φ will be given in Assumptions 3-11 and 3-11* below. For a discussion of the use of the quantile function of the standard normal as the threshold function, refer to CS (2022a), and note that the threshold function used here is related to the one employed in Belloni, Chen, Chernozhukov, and Hansen (2012).

In the sequel, we further require the following assumptions.

Assumption 3-8: There exists a positive constant, $\underline{c}$, such that for T sufficiently large:

$$\min_{1\leq \ell \leq d} \min_{i\in H} \min_{r\in\{1,...,q\}} E\left\{ \left[\frac{1}{\sqrt{\tau_1}} \sum_{t=(r-1)\tau+p}^{(r-1)\tau+\tau_1+p-1} y_{\ell,t+1} u_{it} \right]^2 \right\} \geq \underline{c},$$

where, as defined earlier,

$$\tau_1 = \lfloor T_0^{\alpha_1} \rfloor, \ \tau_2 = \lfloor T_0^{\alpha_2} \rfloor \text{ for } 1 > \alpha_1 \geq \alpha_2 > 0 \text{ and } q = \left\lfloor \frac{T_0}{\tau_1 + \tau_2} \right\rfloor,$$

and $T_0 = T - p + 1$.

Assumption 3-9: Let $i \in H^c = \{k \in \{1, ..., N\} : \gamma_k \neq 0\}$ .Suppose that there exists a positive constant, $\underline{c}$, such that, for all N_1, N_2,and T sufficiently large:

$$\begin{aligned} & \min_{1\leq \ell \leq d} \min_{i\in H^c} \left| \frac{\mu_{i,\ell,T}}{q\tau_1} \right| \\ = & \min_{1\leq \ell \leq d} \min_{i\in H^c} \left| \frac{1}{q} \sum_{r=1}^{q} \frac{1}{\tau_1} \sum_{t=(r-1)\tau+p}^{(r-1)\tau+\tau_1+p-1} \gamma_i' \left\{ E\left[\underline{F}_t\right] \mu_{Y,\ell} + E\left[\underline{F}_t \underline{Y}_t'\right] \alpha_{YY,\ell} + E\left[\underline{F}_t \underline{F}_t'\right] \alpha_{YF,\ell} \right\} \right| \\ \geq & \ \underline{c} > 0, \end{aligned}$$

where $\mu_{Y,\ell} = e_{\ell,d}' \mu_Y$, $\alpha_{YY,\ell} = A_{YY}' e_{\ell,d}$, and $\alpha_{YF,\ell} = A_{YF}' e_{\ell,d}$. Here, $e_{\ell,d}$ is a $d \times 1$ elementary vector whose ℓ^{th} component is 1 and all other components are 0.

Assumption 3-10: Suppose that, as N_1, N_2, and $T \to \infty$, the following rate conditions hold:

(a)

$$\frac{\sqrt{\ln N}}{T^{\min\left\{\frac{1-\alpha_1}{6}, \frac{\alpha_2}{2}\right\}}} \to 0$$

where $1 > \alpha_1 \geq \alpha_2 > 0$ and $N = N_1 + N_2$.

(b)

$$\frac{N_1}{T^{3\alpha_1}} \to 0 \text{ where } 1 > \alpha_1 > 0.$$

Assumption 3-11: Let φ satisfy the following two conditions: (a) $\varphi \to 0$ as $N_1, N_2 \to \infty$, and (b) there exists some constant $a > 0$, such that $\varphi \geq \frac{1}{N^a}$, for all N_1, N_2 sufficiently large.

Note that Assumption 3-9 is a fairly mild condition which allows us to differentiate the alternative hypothesis, $i \in H^c$, from the null hypothesis, $i \in H$. For further discussion of Assumptions 3-8 - 3-11, refer to CS (2022a). Given the above assumptions, Theorem 1 of CS (2022a) shows that the probability of a false positive, i.e., the probability that $i \in \widehat{H}^c$, even though $\gamma_i = 0$, approaches zero, as $N, T \to \infty$, and Theorem 2 of the same paper shows that the probability of a false negative, i.e., the probability that $i \in \widehat{H}$ even though $\gamma_i \neq 0$, also approaches zero, as $N, T \to \infty$. Together, these two theorems show that our variable selection procedure is (completely) consistent in the sense that the probability of committing a misclassification error vanishes as $N, T \to \infty$. CS (2022a) also note that the above variable selection procedure provides us with a consistent estimate $\widehat{N}_1$ of the unobserved quantity N_1, where the latter, in light of Assumption 3-6, can be interpreted as giving the order of magnitude of $\Gamma'\Gamma$ and is, thus, a measure of the overall pervasiveness of the factors in a given application. Finally, note that knowledge of the number of factors is not needed to implement the above variable selection procedure. Hence, in the case where the number of factors needs to be determined empirically, an applied researcher could first use our procedure to properly select the relevant variables and then apply an information criterion such as that proposed in Bai and Ng (2002) to estimate the number of factors.

Before presenting the main theoretical results proven in this paper, it is worth making a final comment about variable selection. In particualr, note that Bai and Ng (2008) address the important issue of choosing predictor variables Z_{it} based on their predictability for Y_{t+1}. While we agree with this viewpoint, it is worth stressing that in our setup, whether Z_{it} helps to predict Y_{t+h} depends on two things: (i) whether Z_{it} loads significantly on the underlying factors $\underline{F}_t$ (i.e., whether $\gamma_i \neq 0$ or not) and (ii) whether at least some components of $\underline{F}_t$ are helpful for predicting certain components of Y_{t+h}. The variable selection procedure which we propose here focuses on the first issue but not the second. This is because, in our view, it is important to first obtain good factor estimates with certain desirable asymptotic properties before trying to assess which factor may or may not be useful for predicting Y_{t+h}. It is important to distinguish between these two things because, if we try to do too much at the variable selection stage and end up excluding a significant number of (predictor) variables that load strongly on at least some of the factors, then, this can lead to the factor vector $\underline{F}_t$ being inconsistently estimated, and this is true even if the variables do not individually help to predict Y_{t+h}, but instead are crucial for the consistent estimation of the factor, which in turn is useful for predicting Y_{t+h}.

4 Consistent Estimation of Factors and the h-Step Ahead Predictor Based on the FAVAR Model

In this section, we provide our main theoretical results on factor estimation and on the estimation of the h-step predictor implied by the FAVAR model. To obtain these results, we need to impose a further rate condition on the tuning parameter, φ (see part (c) of Assumption 3-11*).

Assumption 3-11*: Let φ satisfy the following three conditions: (a) $\varphi \to 0$ as $N_1, N_2 \to \infty$, (b) there exists some constant $a > 0$, such that $\varphi \geq \frac{1}{N^a}$ for all N_1, N_2 sufficiently large, and (c)

$$\max\left\{\frac{N^{\frac{2}{7}}\varphi^{\frac{5}{7}}}{N_1}, \frac{N^{\frac{1}{3}}\varphi}{N_1 T}\right\} \to 0 \text{ as } N_1, N_2, T \to \infty.$$

Remark 4.1: Note that the rate condition given in part (c) of Assumption 3-11* depends on N_1. However, if we choose φ so that:

$$\varphi N^{\frac{2}{5}} = O(1),$$

then

$$\frac{N^{\frac{2}{7}}\varphi^{\frac{5}{7}}}{N_1} = O\left(\frac{1}{N_1}\right) = o(1) \text{ and } \frac{N^{\frac{1}{3}}\varphi}{N_1 T} = O\left(\frac{1}{N_1 N^{\frac{1}{15}} T}\right) = o\left(\frac{1}{N_1}\right).$$

Hence, with this choice of φ, Assumption 3-11* part (c) will be satisfied as long as $N_1 \to \infty$, and there is no need to impose any further condition on the rate at which N_1 grows. Requiring that $N_1 \to \infty$ is a minimal condition, since if $N_1 \nrightarrow \infty$; then consistent factor estimation, even up to an invertible matrix transformation, is impossible. Additionally, Monte Carlo results reported in Section 3 of CS (2022a) show that the variable selection procedure discussed above performs very well in finite samples, under the tuning parameter choice $\varphi = N^{-\frac{2}{5}}$, both in terms of controlling the probability of a false positive (or Type I) error and in terms of controlling the probability of a false negative (or Type II) error.

Next, consider the post-variable-selection principal component estimator of $\underline{F}_t = \left(F_t', F_{t-1}', ..., F_{t-p+1}'\right)$:

$$\widehat{\underline{F}}_t = \frac{\widehat{\Gamma}' Z_{t,N}\left(\widehat{H^c}\right)}{\widehat{N}_1}, \tag{19}$$

where

$$Z_{t,N}\left(\widehat{H^c}\right) = \left[\begin{array}{cccc} Z_{1,t}\mathbb{I}\left\{1 \in \widehat{H^c}\right\} & Z_{2,t}\mathbb{I}\left\{2 \in \widehat{H^c}\right\} & \cdots & Z_{N,t}\mathbb{I}\left\{N \in \widehat{H^c}\right\} \end{array}\right]',$$

with

$$\mathbb{I}\left\{i \in \widehat{H^c}\right\} = \left\{\begin{array}{ll} 1 & \text{if } i \in \widehat{H^c}, \text{ i.e., if } \mathbb{S}_{i,T}^{+} > \Phi^{-1}\left(1 - \frac{\varphi}{2N}\right) \\ 0 & \text{if } i \in \widehat{H}, \text{ i.e., if } \mathbb{S}_{i,T}^{+} \leq \Phi^{-1}\left(1 - \frac{\varphi}{2N}\right) \end{array}\right.,$$

and where $\widehat{N}_1 = \#\left(\widehat{H^c}\right)$, i.e., the cardinality of the set $\widehat{H^c}$. Here, $\widehat{\Gamma}$ denotes the principal component estimator of the loading matrix Γ constructed from taking $\sqrt{\widehat{N}_1}$ times the matrix whose columns are the eigenvectors of the post-variable-selection sample covariance matrix $\widehat{\Sigma}\left(\widehat{H^c}\right)$ associated with the Kp largest eigenvalues of this matrix, where, in this case,

$$\widehat{\Sigma}\left(\widehat{H^c}\right) = \frac{Z\left(\widehat{H^c}\right)' Z\left(\widehat{H^c}\right)}{\widehat{N}_1 T_0} = \frac{1}{\widehat{N}_1 T_0} \sum_{t=p}^{T} Z_{t,N}\left(\widehat{H^c}\right) Z_{t,N}\left(\widehat{H^c}\right)',$$

with $T_0 = T - p + 1$.

Our next result shows that the estimator given in expression (19) consistently estimates the unobserved factors $\underline{F}_t$,up to an invertible $Kp \times Kp$ matrix transformation.

Theorem 4.1: *Suppose that Assumptions 3-1, 3-2, 3-3, 3-4, 3-5, 3-6, 3-7, 3-8, 3-9, and 3-10 hold. Let $\widehat{\underline{F}}_t$ be as defined in expression (19). Assume further that the specification of the tuning parameter, φ, in the decision rule (18) satisfies Assumption 3-11*. Then,*

$$\left\| \widehat{\underline{F}}_t - Q' \underline{F}_t \right\|_2 = o_p(1), \text{ for all fixed } t,$$

where

$$Q = \left(\frac{\Gamma'\Gamma}{N_1}\right)^{\frac{1}{2}} \Xi \widehat{V},$$

and where $\widehat{V}$ is the $Kp \times Kp$ orthogonal matrix given in Lemma D-14, and Ξ is a $Kp \times Kp$ orthogonal matrix whose columns are the eigenvectors of the matrix

$$M^*_{FF} = \left(\frac{\Gamma'\Gamma}{N_1}\right)^{1/2} M_{FF} \left(\frac{\Gamma'\Gamma}{N_1}\right)^{1/2} = \left(\frac{\Gamma'\Gamma}{N_1}\right)^{1/2} \frac{1}{T_0} \sum_{t=p}^{T} E\left[\underline{F}_t \underline{F}_t'\right] \left(\frac{\Gamma'\Gamma}{N_1}\right)^{1/2}.$$

If we examine the proof of Theorem 4.1 in the appendix and the supporting arguments given in the proof of Lemma D-15 of Appendix D of Chao and Swanson (2022c), we see that two of the key components of the proof involve showing that:

$$\left\| \frac{\Gamma\left(\widehat{H^c}\right) - \Gamma}{\sqrt{N_1}} \right\|_2 \xrightarrow{p} 0$$

and that

$$\frac{\widehat{N}_1 - N_1}{N_1} \xrightarrow{p} 0.$$

This is one of the reasons why we argue that initial variable selection should focus on determining

which variables load strongly on the factors without worrying specifically at that stage about the related issues of predictability or, for that matter, any other issue. By contrast, if we make our initial variable selection based on some more stringent criterion that takes into consideration not only variable relevance but also other concerns such as predictability, then, we may end up with a much smaller set $\widetilde{H}^c$ of selected variables relative to the set $\widehat{H^c}$ selected under our procedure. In particular, in this case, it may be possible that even in large samples a significant number of rows of $\Gamma\left(\widetilde{H^c}\right)$ may contain only zero elements even though the corresponding row of Γ is not a zero vector, so that the result:

$$\left\| \frac{\Gamma\left(\widetilde{H^c}\right) - \Gamma}{\sqrt{N_1}} \right\|_2 \xrightarrow{p} 0$$

may not hold. For the same reason, if we let $\widetilde{N}_1$ denote the cardinality of the set of selected indices based on an alternative, more stringent variable selection procedure, then, the result:

$$\frac{\widetilde{N}_1 - N_1}{N_1} \xrightarrow{p} 0$$

also may not hold, since, by definition, N_1 is the number of rows of Γ which have at least one non-zero element.

Although Theorem 4.1 shows that, without further identifying assumptions, we can only estimate the factors $\underline{F}_t$ consistently up to an invertible $Kp \times Kp$ matrix transformation, this result turns out to be sufficient for us to estimate the h-step ahead predictor consistently. More specifically, in Appendix D of Chao and Swanson (2022c) we show that for h-step ahead forecasts associated with the (infeasible) forecasting equation implied by the FAVAR model (2), we have the form

$$Y_{t+h} = \beta_0 + B_1' \underline{Y}_t + B_2' \underline{F}_t + \eta_{t+h}, \tag{20}$$

where $\underline{Y}_t$ and $\underline{F}_t$ are as defined in expression (5) above and where:

$$\begin{aligned} \beta_0 &= \sum_{j=0}^{h-1} J_d A^j \alpha,\ B_1' = J_d A^h \mathcal{P}'_{(d+K)p} S_d,\ B_2' = J_d A^h \mathcal{P}'_{(d+K)p} S_K \text{ and} \\ \eta_{t+h} &= \sum_{j=0}^{h-1} J_d A^j J'_{d+K} \varepsilon_{t+h-j}. \end{aligned} \tag{21}$$

Here, α and A are, respectively, the intercept (vector) and the coefficient matrix of the companion

form defined in expression (6) above, $\mathcal{P}_{(d+K)p}$ is a permutation matrix such that:

$$\mathcal{P}_{(d+K)p}\underline{W}_t = \begin{pmatrix} \underline{Y}_t \\ \underline{F}_t \end{pmatrix},$$

and

$$\begin{aligned} S_d &= \begin{pmatrix} I_{dp} \\ \underset{Kp\times dp}{0} \end{pmatrix}, S_K = \begin{pmatrix} \underset{dp\times Kp}{0} \\ I_{Kp} \end{pmatrix}, \underset{d\times(d+K)p}{J_d} = \begin{bmatrix} I_d & 0 & \cdots & 0 \end{bmatrix}, \text{ and} \\ \underset{(d+K)\times(d+K)p}{J_{d+K}} &= \begin{bmatrix} I_{d+K} & 0 & \cdots & 0 \end{bmatrix}. \end{aligned}$$

See the beginning of Appendix D of Chao and Swanson (2022c) for a derivation of the equation given in expression (20). The reason expression (20) is called an infeasible forecasting equation is, of course, because $\underline{F}_t$ is not observed, so to obtain a feasible version of this forecasting equation, we must replace $\underline{F}_t$ in equation (20) with the estimate $\widehat{\underline{F}}_t$ given in expression (19). Doing so, we arrive at a feasible h-step ahead forecasting equation of the form:

$$\begin{aligned} Y_{t+h} &= \beta_0 + \sum_{g=1}^{p} B_{1,g}' Y_{t-g+1} + \sum_{g=1}^{p} B_{2,g}' \widehat{F}_{t-g+1} + \widehat{\eta}_{t+h} \\ &= \beta_0 + B_1'\underline{Y}_t + B_2'\widehat{\underline{F}}_t + \widehat{\eta}_{t+h}, \end{aligned} \tag{22}$$

where $\widehat{\eta}_{t+h} = \eta_{t+h} - B_2'\left(\widehat{\underline{F}}_t - \underline{F}_t\right)$, with $\eta_{t+h} = \sum_{j=0}^{h-1} J_d A^j J_{d+K}' \varepsilon_{t+h-j}$.

One can interpret expression (22) as a "reduced form" formulation of the forecasting equation where the reduced form parameters β_0, B_1, and B_2 are nonlinear functions of the parameters $(\mu, A_1, ..., A_p)$ of the FAVAR model, in the case where $h > 1$. For forecasting purposes, while it is possible to estimate the conditional mean of the forecasting equation (22) by estimating the underlying parameters directly by nonlinear least squares, here we choose instead to estimate the conditional mean by estimating the reduced form parameters β_0, B_1, and B_2 via linear least squares. An important reason why we choose this latter approach is due to complications that arise both because we are forecasting with a FAVAR which contains unobserved factors that must first be estimated and because we do not make enough identifying assumptions so that the factors can only be estimated consistently up to an invertible $Kp \times Kp$ matrix transformation. In fact, it turns out that estimating the underlying parameters $\mu, A_1, ..., A_p$ by nonlinear least squares and constructing an estimator of the conditional mean of the forecasting equation based on these estimates will not lead to a consistently estimated h-step predictor, unless further identifying assumptions are made.

On the other hand, as we will show in Theorem 5 below, estimating the reduced form parameters β_0, B_1, and B_2 by linear least squares does allow us to construct a consistent estimator of the conditional mean, even in the absence of additional identifying assumptions.

More precisely, let $\widehat{\underline{F}}_t$ denotes the factor estimates given in expression (19). Our procedure minimizes the least squares criterion function:

$$\begin{aligned} Q\left(\beta_0, B_1, B_2\right) &= \sum_{t=p}^{T-h} \left\| Y_{t+h} - \beta_0 - B_1'\underline{Y}_t - B_2'\widehat{\underline{F}}_t \right\|_2^2 \\ &= \sum_{t=p}^{T-h} \left\| Y_{t+h} - \beta_0 - \sum_{g=1}^{p} B_{1,g}' Y_{t-g+1} - \sum_{g=1}^{p} B_{2,g}'\widehat{F}_{t-g+1} \right\|_2^2 \end{aligned} \tag{23}$$

with respect to the parameters β_0, B_1, and B_2, and delivers the OLS estimates $\widehat{\beta}_0$, $\widehat{B}_1$, and $\widehat{B}_2$. We then forecast Y_{T+h} using the h-step predictor:

$$\widehat{Y}_{T+h} = \widehat{\beta}_0 + \widehat{B}_1'\underline{Y}_T + \widehat{B}_2'\widehat{\underline{F}}_T. \tag{24}$$

The following result shows that $\widehat{Y}_{T+h}$ is a consistent estimator of the conditional mean of the infeasible forecast equation (20).

Theorem 4.2: *Let $\widehat{Y}_{T+h}$ be as defined in expression (24). Suppose that Assumptions 3-1, 3-2, 3-3, 3-4, 3-5, 3-6, 3-7, 3-8, 3-9, 3-10, and 3-11* hold. Then,*

$$\widehat{Y}_{T+h} - \left(\beta_0 + B_1'\underline{Y}_T + B_2'\underline{F}_T\right) \xrightarrow{p} 0 \text{ as } N_1, N_2, T \to \infty.$$

5 Empirical Illustration

To be completed.

6 Conclusion

In this paper, we study the problem of consistently estimating the conditional mean of a factor-augmented forecasting equation based on the FAVAR model. When the underlying dynamic factor model generating the latent factors is high-dimensional, we show that it is important to pre-screen the variables in terms of their association with the underlying factors prior to estimation, particularly in cases where one suspects that the conventional assumption of factor pervasiveness may not

hold. For this purpose, we utiluize a new variable selection procedure based on a self-normalized score statistic (see Chao and Swanson (CS: 2022) that correctly identifies the set of variables which load significantly on the underlying factors, with probability approaching one, as the sample sizes go to infinity. Furthermore, given that CS(2022) show that estimating the factors using only those variables selected by their method allows factors to be consistently estimated, up to an invertible matrix transformation, even if the standard pervasiveness assumption does not hold, provided that the number of relevant variables is sufficiently large. Using the factors estimated in such a manner, we show that the conditional mean function of a factor-augmented forecasting equation can be consistently estimated, even for the case of multi-step ahead forecasts.

7 Appendix

All lemmas denoted C1-C17 and D1-D18 in this appendix are stated and proven in an accompanying online appendix (see Chao and Swanson (2022c)).

Proof of Theorem 2.1: The proof of this theorem is rather long, and is gathered in Appendix A of Chao and Swanson (2022c).

Proof of Theorem 4.1:

To proceed, note first that the principal component estimator of $\underline{F}_t$ can be written as

$$\widehat{\underline{F}}_t = \frac{\widehat{\Gamma}' Z_{t,N}\left(\widehat{H^c}\right)}{\widehat{N}_1}$$

where $\widehat{\Gamma} = \sqrt{\widehat{N}_1}\widehat{B}$ and where the columns of the matrix $\widehat{B}$ are the eigenvectors associated with the Kp largest eigenvalues of the (post-variable-selection) sample covariance matrix

$$\widehat{\Sigma}\left(\widehat{H^c}\right) = \frac{Z\left(\widehat{H^c}\right)' Z\left(\widehat{H^c}\right)}{\widehat{N}_1 T_0}.$$

Moreover, by the result of part (d) of Lemma D-14, the matrix $\widehat{B}$ has the representation

$$\widehat{B} = \widehat{G}_1 \widehat{V}$$

where $\widehat{G}_1$ is an $N \times Kp$ matrix, whose columns define an orthonormal basis for an invariant subspace of $\widehat{\Sigma}\left(\widehat{H^c}\right)$ and where $\widehat{V}$ is a $Kp \times Kp$ orthogonal matrix as defined in expression (**??**) in part (c) of Lemma D-14. (See Lemma D-14 and also Lemma D-13 for additional discussion on the origin of

this representation). Making use of this representation, we can further write

$$\begin{aligned}
\widehat{\underline{F}}_t - Q'\underline{F}_t &= \frac{\sqrt{\widehat{N}_1}\widehat{V}'\widehat{G}_1' Z_{t,N}\left(\widehat{H^c}\right)}{\widehat{N}_1} - Q'\underline{F}_t \\
&= \frac{\widehat{V}'\widehat{G}_1'\Gamma\left(\widehat{H^c}\right)\underline{F}_t}{\sqrt{\widehat{N}_1}} + \frac{\widehat{V}'\widehat{G}_1' U_{t,N}\left(\widehat{H^c}\right)}{\sqrt{\widehat{N}_1}} - Q'\underline{F}_t \\
&= \frac{\widehat{V}'\widehat{G}_1'\Gamma\left(\widehat{H^c}\right)\underline{F}_t}{\sqrt{\widehat{N}_1}} - Q'\underline{F}_t + \frac{\widehat{V}'\widehat{G}_1' U_{t,N}\left(\widehat{H^c}\right)}{\sqrt{\widehat{N}_1}} \\
&= \left(\frac{\widehat{V}'\widehat{G}_1'\Gamma\left(\widehat{H^c}\right)}{\sqrt{\widehat{N}_1}} - Q'\right)\underline{F}_t + \frac{\widehat{V}'\widehat{G}_1' U_{t,N}\left(\widehat{H^c}\right)}{\sqrt{\widehat{N}_1}}
\end{aligned}$$

Next, note that

$$\begin{aligned}
\frac{\widehat{V}'\widehat{G}_1'\Gamma}{\sqrt{\widehat{N}_1}} - Q' &= \frac{\widehat{V}'\widehat{G}_1'\Gamma}{\sqrt{N_1}\sqrt{\left(\widehat{N}_1 - N_1 + N_1\right)/N_1}} - Q' \\
&= \left(1 + \frac{\widehat{N}_1 - N_1}{N_1}\right)^{-\frac{1}{2}} \frac{\widehat{V}'\widehat{G}_1'\Gamma}{\sqrt{N_1}} - Q' \\
&= \left[\left(1 + \frac{\widehat{N}_1 - N_1}{N_1}\right)^{-\frac{1}{2}} - 1 + 1\right] \frac{\widehat{V}'\widehat{G}_1'\Gamma}{\sqrt{N_1}} - Q' \\
&= \left[\left(1 + \frac{\widehat{N}_1 - N_1}{N_1}\right)^{-\frac{1}{2}} - 1\right] \frac{\widehat{V}'\widehat{G}_1'\Gamma}{\sqrt{N_1}} + \frac{\widehat{V}'\widehat{G}_1'\Gamma}{\sqrt{N_1}} - Q'
\end{aligned}$$

and

$$\begin{aligned}
\frac{\Gamma\left(\widehat{H^c}\right) - \Gamma}{\sqrt{\widehat{N}_1}} &= \frac{\Gamma\left(\widehat{H^c}\right) - \Gamma}{\sqrt{N_1}\sqrt{\left(\widehat{N}_1 - N_1 + N_1\right)/N_1}} \\
&= \left(1 + \frac{\widehat{N}_1 - N_1}{N_1}\right)^{-\frac{1}{2}} \left(\frac{\Gamma\left(\widehat{H^c}\right) - \Gamma}{\sqrt{N_1}}\right)
\end{aligned}$$

so that

$$\begin{aligned}
&\frac{\widehat{V}'\widehat{G}_1'\Gamma\left(\widehat{H^c}\right)\underline{F}_t}{\sqrt{\widehat{N}_1}} \\
&= Q'\underline{F}_t + \left(\frac{\widehat{V}'\widehat{G}_1'\Gamma}{\sqrt{\widehat{N}_1}} - Q'\right)\underline{F}_t + \widehat{V}'\widehat{G}_1'\left(\frac{\Gamma\left(\widehat{H^c}\right)-\Gamma}{\sqrt{\widehat{N}_1}}\right)\underline{F}_t \\
&= Q'\underline{F}_t + \left(\frac{\widehat{V}'\widehat{G}_1'\Gamma}{\sqrt{N_1}} - Q'\right)\underline{F}_t + \left[\left(1+\frac{\widehat{N}_1-N_1}{N_1}\right)^{-\frac{1}{2}} - 1\right]\frac{\widehat{V}'\widehat{G}_1'\Gamma}{\sqrt{N_1}}\underline{F}_t \\
&\quad + \left[\left(1+\frac{\widehat{N}_1-N_1}{N_1}\right)^{-\frac{1}{2}}\right]\widehat{V}'\widehat{G}_1'\left(\frac{\Gamma\left(\widehat{H^c}\right)-\Gamma}{\sqrt{N}_1}\right)\underline{F}_t
\end{aligned}$$

It follows that

$$\begin{aligned}
\underline{\widehat{F}}_t - Q'\underline{F}_t &= \left(\frac{\widehat{V}'\widehat{G}_1'\Gamma\left(\widehat{H^c}\right)}{\sqrt{\widehat{N}_1}} - Q'\right)\underline{F}_t + \frac{\widehat{V}'\widehat{G}_1'U_{t,N}\left(\widehat{H^c}\right)}{\sqrt{\widehat{N}_1}} \\
&= \left(\frac{\widehat{V}'\widehat{G}_1'\Gamma}{\sqrt{N_1}} - Q'\right)\underline{F}_t + \left[\left(1+\frac{\widehat{N}_1-N_1}{N_1}\right)^{-\frac{1}{2}} - 1\right]\frac{\widehat{V}'\widehat{G}_1'\Gamma}{\sqrt{N_1}}\underline{F}_t \\
&\quad + \left[\left(1+\frac{\widehat{N}_1-N_1}{N_1}\right)^{-\frac{1}{2}}\right]\widehat{V}'\widehat{G}_1'\left(\frac{\Gamma\left(\widehat{H^c}\right)-\Gamma}{\sqrt{N}_1}\right)\underline{F}_t + \frac{\widehat{V}'\widehat{G}_1'U_{t,N}\left(\widehat{H^c}\right)}{\sqrt{\widehat{N}_1}}
\end{aligned}$$

Hence, applying the triangle inequality as well as parts (a)-(c), (g), and (i) of Lemma D-15 along

with the Slutsky's theorem, we obtain

$$
\begin{aligned}
& \left\| \widehat{\underline{F}}_t - Q' \underline{F}_t \right\|_2 \\
\leq & \left\| \frac{\widehat{V}' \widehat{G}_1' \Gamma}{\sqrt{N_1}} - Q' \right\|_2 \left\| \underline{F}_t \right\|_2 + \left| \left(1 + \frac{\widehat{N}_1 - N_1}{N_1} \right)^{-\frac{1}{2}} - 1 \right| \left\| \frac{\widehat{V}' \widehat{G}_1' \Gamma}{\sqrt{N_1}} \right\|_2 \left\| \underline{F}_t \right\|_2 \\
& + \left| \left(1 + \frac{\widehat{N}_1 - N_1}{N_1} \right)^{-\frac{1}{2}} \right| \left\| \widehat{V}' \widehat{G}_1' \right\|_2 \left\| \frac{\Gamma \left(\widehat{H^c} \right) - \Gamma}{\sqrt{N_1}} \right\|_2 \left\| \underline{F}_t \right\|_2 + \left\| \frac{\widehat{V}' \widehat{G}_1' U_{t,N} \left(\widehat{H^c} \right)}{\sqrt{\widehat{N}_1}} \right\|_2 \\
= & \left\| \frac{\widehat{V}' \widehat{G}_1' \Gamma}{\sqrt{N_1}} - Q' \right\|_2 \left\| \underline{F}_t \right\|_2 + \left| \left(1 + \frac{\widehat{N}_1 - N_1}{N_1} \right)^{-\frac{1}{2}} - 1 \right| \left\| \frac{\widehat{V}' \widehat{G}_1' \Gamma}{\sqrt{N_1}} \right\|_2 \left\| \underline{F}_t \right\|_2 \\
& + \left| \left(1 + \frac{\widehat{N}_1 - N_1}{N_1} \right)^{-\frac{1}{2}} \right| \left\| \frac{\Gamma \left(\widehat{H^c} \right) - \Gamma}{\sqrt{N_1}} \right\|_2 \left\| \underline{F}_t \right\|_2 + \left\| \frac{\widehat{V}' \widehat{G}_1' U_{t,N} \left(\widehat{H^c} \right)}{\sqrt{\widehat{N}_1}} \right\|_2 \\
& \left(\text{since } \left\| \widehat{V}' \widehat{G}_1' \right\|_2 = \lambda_{\max} \left(\widehat{G}_1 \widehat{V} \widehat{V}' \widehat{G}_1' \right) = \lambda_{\max} \left(\widehat{V}' \widehat{G}_1' \widehat{G}_1 \widehat{V} \right) = \lambda_{\max} \left(I_{Kp} \right) = 1 \right) \\
= & o_p(1) O_p(1) + o_p(1) O_p(1) O_p(1) + O_p(1) o_p(1) O_p(1) + o_p(1) \\
= & o_p(1) . \ \square
\end{aligned}
$$

Proof of Theorem 4.2:

To proceed, note that for any $a \in \mathbb{R}^d$ such that $\|a\|_2 = 1$, we have

$$
\begin{aligned}
& \left| a' \widehat{Y}_{T+h} - a' \left(\beta_0 + B_1' \underline{Y}_T + B_2' \underline{F}_T \right) \right| \\
= & \left| a' \left(\widehat{\beta}_0 + \widehat{B}_1' \underline{Y}_T + \widehat{B}_2' \widehat{\underline{F}}_T \right) - a' \left(\beta_0 + B_1' \underline{Y}_T + B_2' \underline{F}_T \right) \right| \\
= & \left| a' \left(\widehat{\beta}_0 - \beta_0 \right) + a' \left(\widehat{B}_1 - B_1 \right)' \underline{Y}_T \right. \\
& \left. + a' \left(\widehat{B}_2 - Q^{-1} B_2 + Q^{-1} B_2 \right)' \left(\widehat{\underline{F}}_T - Q' \underline{F}_T + Q' \underline{F}_T \right) - a' B_2' \underline{F}_T \right| \\
\leq & \left| a' \left(\widehat{\beta}_0 - \beta_0 \right) \right| + \left| a' \left(\widehat{B}_1 - B_1 \right)' \underline{Y}_T \right| + \left| a' \left(\widehat{B}_2 - Q^{-1} B_2 \right)' \left(\widehat{\underline{F}}_T - Q' \underline{F}_T \right) \right| \\
& + \left| a' B_2' Q^{-1\prime} \left(\widehat{\underline{F}}_T - Q' \underline{F}_T \right) \right| + \left| a' \left(\widehat{B}_2 - Q^{-1} B_2 \right)' Q' \underline{F}_T \right| + \left| a' B_2' Q^{-1\prime} Q' \underline{F}_T - a' B_2' \underline{F}_T \right| \\
= & \left| a' \left(\widehat{\beta}_0 - \beta_0 \right) \right| + \left| a' \left(\widehat{B}_1 - B_1 \right)' \underline{Y}_T \right| + \left| a' \left(\widehat{B}_2 - Q^{-1} B_2 \right)' \left(\widehat{\underline{F}}_T - Q' \underline{F}_T \right) \right| \\
& + \left| a' B_2' Q^{-1\prime} \left(\widehat{\underline{F}}_T - Q' \underline{F}_T \right) \right| + \left| a' \left(\widehat{B}_2 - Q^{-1} B_2 \right)' Q' \underline{F}_T \right|
\end{aligned}
$$

Lemma D-18 and Slutsky's theorem directly imply that

$$\left|a'\left(\widehat{\beta}_0-\beta_0\right)\right|=o_p\left(1\right)$$

Now, applying the CS inequality, we obtain

$$\begin{aligned}\left|a'\left(\widehat{B}_1-B_1\right)'\underline{Y}_T\right| &\leq \sqrt{a'\left(\widehat{B}_1-B_1\right)'\left(\widehat{B}_1-B_1\right)a}\sqrt{\underline{Y}_T'\underline{Y}_T}\\ &= \sqrt{a'\left(\widehat{B}_1-B_1\right)'\left(\widehat{B}_1-B_1\right)a}\left\|\underline{Y}_T\right\|_2^2,\end{aligned}$$

and

$$\begin{aligned}&\left|a'\left(\widehat{B}_2-Q^{-1}B_2\right)'Q'\underline{F}_T\right|\\ \leq\ &\sqrt{a'\left(\widehat{B}_2-Q^{-1}B_2\right)'\left(\widehat{B}_2-Q^{-1}B_2\right)a}\sqrt{\underline{F}_T'QQ'\underline{F}_T}\\ =\ &\sqrt{a'\left(\widehat{B}_2-Q^{-1}B_2\right)'\left(\widehat{B}_2-Q^{-1}B_2\right)a}\sqrt{\underline{F}_T'\left(\frac{\Gamma'\Gamma}{N_1}\right)^{1/2}\Xi\widehat{V}\widehat{V}'\Xi'\left(\frac{\Gamma'\Gamma}{N_1}\right)^{1/2}\underline{F}_T}\\ =\ &\sqrt{a'\left(\widehat{B}_2-Q^{-1}B_2\right)'\left(\widehat{B}_2-Q^{-1}B_2\right)a}\sqrt{\underline{F}_T'\left(\frac{\Gamma'\Gamma}{N_1}\right)\underline{F}_T}\\ \leq\ &\sqrt{\lambda_{\max}\left(\frac{\Gamma'\Gamma}{N_1}\right)}\sqrt{a'\left(\widehat{B}_2-Q^{-1}B_2\right)'\left(\widehat{B}_2-Q^{-1}B_2\right)a}\left\|\underline{F}_T\right\|_2^2\\ \leq\ &\overline{C}\sqrt{a'\left(\widehat{B}_2-Q^{-1}B_2\right)'\left(\widehat{B}_2-Q^{-1}B_2\right)a}\left\|\underline{F}_T\right\|_2^2\end{aligned}$$

Moreover, note that

$$\begin{aligned}E\left[\left\|\underline{Y}_T\right\|_2^2\right] &\leq \left(E\left\|\underline{Y}_T\right\|_2^6\right)^{\frac{1}{3}} \quad \text{(by Liapunov's inequality)}\\ &\leq \overline{C}^{\frac{1}{3}}=C<\infty \quad \text{(by Lemma C-5)}\end{aligned}$$

and

$$\begin{aligned}E\left[\left\|\underline{F}_T\right\|_2^2\right] &\leq \left(E\left\|\underline{F}_T\right\|_2^6\right)^{\frac{1}{3}} \quad \text{(by Liapunov's inequality)}\\ &\leq \overline{C}^{\frac{1}{3}}=C<\infty \quad \text{(by Lemma C-5)}\end{aligned}$$

Hence, for any $\epsilon > 0$, set $C_\epsilon = \sqrt{C/\epsilon}$, and Markov's inequality then implies that, for all $T > p-1$,

$$\Pr\left\{\|\underline{Y}_T\|_2 \geq C_\epsilon\right\} = \Pr\left\{\|\underline{Y}_T\|_2^2 \geq C_\epsilon^2\right\} \leq \frac{E\left[\|\underline{Y}_T\|_2^2\right]}{C_\epsilon^2} = \frac{\epsilon E\left[\|\underline{Y}_T\|_2^2\right]}{C} \leq \epsilon$$

from which it follows that

$$\|\underline{Y}_T\|_2 = O_p(1).$$

In a similar way, we can also show that

$$\|\underline{F}_T\|_2 = O_p(1).$$

Application of the result given in Lemma D-18 then allows us to deduce that

$$\left|a'\left(\widehat{B}_1 - B_1\right)'\underline{Y}_T\right| \leq \sqrt{a'\left(\widehat{B}_1 - B_1\right)'\left(\widehat{B}_1 - B_1\right)a}\,\|\underline{Y}_T\|_2^2 = o_p(1)$$

and

$$\begin{aligned}
&\left|a'\left(\widehat{B}_2 - Q^{-1}B_2\right)'Q'\underline{F}_T\right| \\
\leq\ &\sqrt{a'\left(\widehat{B}_2 - Q^{-1}B_2\right)'\left(\widehat{B}_2 - Q^{-1}B_2\right)a}\sqrt{\underline{F}_T'QQ'\underline{F}_T} \\
\leq\ &\sqrt{a'\left(\widehat{B}_2 - Q^{-1}B_2\right)'\left(\widehat{B}_2 - Q^{-1}B_2\right)a}\sqrt{\lambda_{\max}\left(QQ'\right)}\,\|\underline{F}_T\|_2 \\
=\ &\sqrt{a'\left(\widehat{B}_2 - Q^{-1}B_2\right)'\left(\widehat{B}_2 - Q^{-1}B_2\right)a}\sqrt{\lambda_{\max}\left\{\left(\frac{\Gamma'\Gamma}{N_1}\right)^{\frac{1}{2}}\Xi\widehat{V}\widehat{V}'\Xi'\left(\frac{\Gamma'\Gamma}{N_1}\right)^{\frac{1}{2}}\right\}}\,\|\underline{F}_T\|_2 \\
=\ &\sqrt{a'\left(\widehat{B}_2 - Q^{-1}B_2\right)'\left(\widehat{B}_2 - Q^{-1}B_2\right)a}\sqrt{\lambda_{\max}\left\{\left(\frac{\Gamma'\Gamma}{N_1}\right)\right\}}\,\|\underline{F}_T\|_2 \\
&\left(\text{since } \widehat{V}\widehat{V}' = I_{Kp} \text{ and } \Xi\Xi' = I_{Kp}\right) \\
\leq\ &\sqrt{\overline{C}}\sqrt{a'\left(\widehat{B}_2 - Q^{-1}B_2\right)'\left(\widehat{B}_2 - Q^{-1}B_2\right)a}\,\|\underline{F}_T\|_2 \quad \text{(by Assumption 3-6)} \\
=\ &o_p(1)
\end{aligned}$$

In addition, we can apply the CS inequality to get

$$\begin{aligned}
&\left|a'\left(\widehat{B}_2 - Q^{-1}B_2\right)'\left(\widehat{\underline{F}}_T - Q'\underline{F}_T\right)\right| \\
\leq\ & \sqrt{a'\left(\widehat{B}_1 - B_1\right)'\left(\widehat{B}_1 - B_1\right)a}\sqrt{\left(\widehat{\underline{F}}_T - Q'\underline{F}_T\right)'\left(\widehat{\underline{F}}_T - Q'\underline{F}_T\right)} \\
\leq\ & \sqrt{a'\left(\widehat{B}_1 - B_1\right)'\left(\widehat{B}_1 - B_1\right)a}\left\|\widehat{\underline{F}}_T - Q'\underline{F}_T\right\|_2 \\
=\ & o_p(1) \quad \text{(by Lemma D-18 and part (j) of Lemma D-15 in Appendix D)}
\end{aligned}$$

and

$$\begin{aligned}
&\left|a'B_2'Q^{-1\prime}\left(\widehat{\underline{F}}_T - Q'\underline{F}_T\right)\right| \\
\leq\ & \sqrt{a'B_2'Q^{-1\prime}Q^{-1}B_2a}\sqrt{\left(\widehat{\underline{F}}_T - Q'\underline{F}_T\right)'\left(\widehat{\underline{F}}_T - Q'\underline{F}_T\right)} \\
=\ & \sqrt{a'B_2'Q^{-1\prime}Q^{-1}B_2a}\left\|\widehat{\underline{F}}_T - Q'\underline{F}_T\right\|_2 \\
\leq\ & \sqrt{\left[\lambda_{\min}\left(\frac{\Gamma'\Gamma}{N_1}\right)\right]^{-1}\lambda_{\max}\left(B_2'B_2\right)}\left\|\widehat{\underline{F}}_T - Q'\underline{F}_T\right\|_2 \\
\leq\ & \sqrt{C^*}\left\|\widehat{\underline{F}}_T - Q'\underline{F}_T\right\|_2 \quad \text{(for some positive constant } C^* \text{ as shown in} \\
& \text{expression (??) in Appendix D. See the proof of part (d) of Lemma D-17)} \\
=\ & o_p(1) \quad \text{(by part (j) of Lemma D-15)}
\end{aligned}$$

Putting everything together and applying Slutsky's theorem, we then obtain

$$\begin{aligned}
&\left|a'\widehat{Y}_{T+h} - a'\left(\beta_0 + B_1'\underline{Y}_T + B_2'\underline{F}_T\right)\right| \\
\leq\ & \left|a'\left(\widehat{\beta}_0 - \beta_0\right)\right| + \left|a'\left(\widehat{B}_1 - B_1\right)'\underline{Y}_T\right| + \left|a'\left(\widehat{B}_2 - Q^{-1}B_2\right)'\left(\widehat{\underline{F}}_T - Q'\underline{F}_T\right)\right| \\
& + \left|a'B_2'Q^{-1\prime}\left(\widehat{\underline{F}}_T - Q'\underline{F}_T\right)\right| + \left|a'\left(\widehat{B}_2 - Q^{-1}B_2\right)'Q'\underline{F}_T\right| \\
=\ & o_p(1).
\end{aligned}$$

Since the above argument holds for all $a \in \mathbb{R}^d$ such that $\|a\|_2 = 1$, we further deduce that

$$\widehat{Y}_{T+h} - \left(\beta_0 + B_1'\underline{Y}_T + B_2'\underline{F}_T\right) = o_p(1).$$

as required. □

References

[1] Anatolyev, S. and A. Mikusheva (2021): "Factor Models with Many Assets: Strong Factors, Weak Factors, and the Two-Pass Procedure," *Journal of Econometrics*, forthcoming.

[2] Andrews, D.W.K. (1984): "Non-strong Mixing Autoregressive Processes," *Journal of Applied Probability*, 21, 930-934.

[3] Bai, J. and S. Ng (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191-221.

[4] Bai, J. (2003): "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71, 135-171.

[5] Bai, J. and S. Ng (2008): "Forecasting Economic Time Series Using Targeted Predictors," *Journal of Econometrics*, 146, 304-317.

[6] Bai, J. and S. Ng (2021): "Approximate Factor Models with Weaker Loading," Working Paper, Columbia University.

[7] Bai, Z. D. and Y. Q. Yin (1993): "Limit of the Smallest Eigenvalue of a Large Dimensional Sample Covariance Matrix," *Annals of Probability*, 21, 1275-1294.

[8] Bair, E., T. Hastie, D. Paul, and R. Tibshirani (2006): "Prediction by Supervised Principal Components," *Journal of the American Statistical Association*, 101, 119-137.

[9] Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012): "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," *Econometrica*, 80, 2369-2429.

[10] Billingsley, P. (1995): *Probability and Measure*. New York: John Wiley & Sons.

[11] Borovkova, S., R. Burton, and H. Dehling (2001): "Limit Theorems for Functionals of Mixing Processes to U-Statistics and Dimension Estimation," *Transactions of the American Mathematical Society*, 353, 4261-4318.

[12] Bryzgalova, S. (2016): "Spurious Factors in Linear Asset Pricing Models," Working Paper, Stanford Graduate School of Business.

[13] Burnside, C. (2016): "Identification and Inference in Linear Stochastic Discount Factor Models with Excess Returns," *Journal of Financial Econometrics*, 14, 295-330.

[14] Chao, J. C. and N. R. Swanson (2022a): "Selecting the Relevant Variables for Factor Estimation in a Factor-Augmented VAR Model," Working Paper, Rutgers University and University of Maryland.

[15] Chao, J. C. and N. R. Swanson (2022b): Technical Appendix to "Consistent Estimation, Variable Selection, and Forecasting in Factor-Augmented VAR Models," Working Paper, Rutgers University and University of Maryland.

[16] Chao, J. C. and N. R. Swanson (2022c): Texchnical Appendix to "Consistent Factor Estimation and Forecasting in Factor-Augmented VAR Models," Working Paper, Rutgers University and University of Maryland.

[17] Chen, X., Q. Shao, W. B. Wu, and L. Xu (2016): "Self-normalized Cramér-type Moderate Deviations under Dependence," *Annals of Statistics*, 44, 1593-1617.

[18] Davidson. J. (1994): *Stochastic Limit Theory: An Introduction for Econometricians*. New York: Oxford University Press.

[19] Davidson, K. R. and S. J. Szarek (2001): "Local Operator Theory, Random Matrices and Banach Spaces." In *Handbook of the Geometry of Banach Spaces*, 1, 317-366. Amsterdam: North-Holland.

[20] Fan, J., Y. Liao, and M. Mincheva (2011): "High-dimensional Covariance Matrix Estimation in Approximate Factor Models," *Annals of Statistics*, 39, 3320-3356.

[21] Fan, J., Y. Liao, and M. Mincheva (2013): "Large Covariance Estimation by Thresholding Principal Orthogonal Complements," *Journal of the Royal Statistical Society, Series B*, 75, 603-680.

[22] Freyaldenhoven, S. (2021a): "Factor Models with Local Factors - Determining the Number of Relevant Factors," *Journal of Econometrics*, forthcoming.

[23] Freyaldenhoven, S. (2021b): "Identification through Sparsity in Factor Models: The ℓ_1-Rotation Criterion," Working Paper, Federal Reserve Bank of Philadelphia.

[24] Giglio, S., D. Xiu, and D. Zhang (2021): "Test Assets and Weak Factors," Working Paper, Yale School of Management and the Booth School of Business, University of Chicago.

[25] Golub, G. H. and C. F. van Loan (1996): *Matrix Computations*, 3rd Edition. Baltimore: The Johns Hopkins University Press.

[26] Goroketskii, V. V. (1977): "On the Strong Mixing Property for Linear Sequences," *Theory of Probability and Applications*, 22, 411-413.

[27] Gospodinov, N., R. Kan, and C. Robotti (2017): "Spurious Inference in Reduced-Rank Asset Pricing Models," *Econometrica*, 85, 1613-1628.

[28] Harding, M. C. (2008): "Explaining the Single Factor Bias of Arbitrage Pricing Models in Finite Samples," *Economics Letters*, 99, 85-88.

[29] Horn, R. and C. Johnson (1985): *Matrix Analysis*. Cambridge University Press.

[30] Jagannathan, R. and Z. Wang (1998): "An Asymptotic Theory for Estimating Beta-Pricing Models Using Cross-Sectional Regression," *Journal of Finance*, 53, 1285-1309.

[31] Johnstone, I. M. and A. Lu (2009): "On Consistency and Sparsity for Principal Components Analysis in High Dimensions," *Journal of the American Statistical Association*, 104, 682-697.

[32] Johnstone, I. M. and D. Paul (2018): "PCA in High Dimensions: An Orientation," *Proceedings of the IEEE*, 106, 1277-1292.

[33] Kan, R. and C. Zhang (1999): "Two-Pass Tests of Asset Pricing Models with Useless Factors," *Journal of Finance*, 54, 203-235.

[34] Kleibergen, F. (2009): "Tests of Risk Premia in Linear Factor Models," *Journal of Econometrics*, 149, 149-173.

[35] Lütkepohl, H. (2005): *New Introduction to Multiple Time Series Analysis*. New York: Springer.

[36] Nadler, B. (2008): "Finite Sample Approximation Results for Principal Component Analysis: A Matrix Perturbation Approach," *Annals of Statistics*, 36, 2791-2817.

[37] Onatski, A. (2012): "Asymptotics of the Principal Components Estimator of Large Factor Models with Weakly Influential Factors," *Journal of Econometrics*, 168, 244-258.

[38] Paul, D. (2007): "Asymptotics of Sample Eigenstructure for a Large Dimensional Spiked Covariance Model," *Statistica Sinica*, 17, 1617-1642.

[39] Pham, T. D. and L. T. Tran (1985): "Some Mixing Properties of Time Series Models," *Stochastic Processes and Their Applications*, 19, 297-303.

[40] Ruhe, A. (1975): "On the Closeness of Eigenvalues and Singular Values for Almost Normal Matrices," *Linear Algebra and Its Applications*, 11, 87-94.

[41] Shen, D., H. Shen, H. Zhu, J.S. Marron (2016): "The Statistics and Mathematics of High Dimension Low Sample Size Asymptotics," *Statistica Sinica*, 26, 1747-1770.

[42] Stewart, G.W. (1973): "Error and Perturbation Bounds for Subspaces Associated with Certain Eigenvalue Problems," *SIAM Review*, 15, 727-764.

[43] Stewart, G.W. and J. Sun (1990): *Matrix Perturbation Theory*. Boston: Academic Press.

[44] Stock, J. H. and M. W. Watson (2002a): "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167-1179.

[45] Stock, J. H. and M. W. Watson (2002b): "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business and Economic Statistics*, 20, 147-162.

[46] Vershynin, R. (2012): "Introduction to the Non-asymptotic Analysis of Random Matrices," In *Compressed Sensing, Theory and Applications,* 210-268. Cambridge University Press.

[47] Wang, W. and J. Fan (2017): "Asymptotics of Empirical Eigenstructure for High Dimensional Spiked Covariance," *Annals of Statistics*, 45, 1342-1374.