Consistent Estimation, Variable Selection, and Forecasting in Factor-Augmented VAR Models*

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

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

January 2022

Abstract

In the context of latent factor models that are widely used in economics, a common assumption made is one of factor pervasiveness, which implies that all available predictor or informative variables in a dataset, with the possible exception of a negligible number of them, load significantly on the underlying factors. In this paper, we analyze the more likely scenario where there is significant underlying heterogeneity in the sense that some of the variables load significantly on the underlying factors, while 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 data set. We show that, even in such a setting, consistent factor estimation can be achieved if one pre-screens the variables and successfully prunes out the irrelevant ones. To do so, we introduce a new variable selection procedure that, with probability approaching one, correctly distinguishes between relevant and irrelevant variables. We study this problem within a factor-augmented VAR (FAVAR) framework, and 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 the FAVAR model, we can obtain consistent estimates of the conditional mean function of said equations. In particular, our methodology allows the conditional mean function of a factor-augmented forecast equation to be 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 asymptotics, 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 Juhee Bae, Simon Freyaldenhoven, Yuan Liao, Minchul Shin, Xiye Yang, and seminar participants at the Federal Reserve Bank of Philadelphia, the University of Riverside, and the University of Glasgow for useful comments received on earlier versions of this paper. Chao thanks the University of Maryland for research support.

1 Introduction

As a result of the astounding rate at which raw information is currently being accumulated, there is a clear need for variable selection, dimension reduction and shrinkage techniques when analyzing big data using machine learning techniques. This has led to a profusion of novel research in areas ranging from the analysis of high dimensional and/or high frequency datasets to statistical learning methods. Needless to say, there are many critical unanswered questions in this burgeoning literature. One such question, which we address in this paper stems from 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 these papers, the authors develop methods for constructing forecasts based on factor-augmented regression models. An obvious appeal of using factor analytical methods for this problem is the capacity for dimension reduction, so that in terms of the specification of the forecasting equation, information embedded in a possibly high-dimensional vector of (predictor or information) variables can be represented parsimoniously. The question that we address in this context is how important the pervasiveness assumption is, when establishing factor consistency, and subsequently when constructing forecasts with models that include estimated factors. More specifically, consistent estimation of factors in the above papers makes an assumption that factors are pervasive, in the sense that all available predictor or information variables in a dataset, with the possible exception of a negligible number of them, load significantly on the underlying factors[1]. Given a particular dataset, this would seem to be a rather fortuitous situation. A more likely scenario might be 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 the first part of this paper, we show that if one estimates factors using all available variables, without regard as to whether they are relevant or not, then the factors may be estimated inconsistently, if a significant number of the variables do not actually load significantly on the underlying factors. This is clearly problematic in a number of empirical situations, such as when the objective is to estimate forecast functions that incorporate estimated factors. However, if one pre-screens the variables and successfully prunes out the irrelevant ones, then consistent estimation can be achieved, assuming that the number of relevant variables is sufficiently large, in a sense that will be made formal later. In light of this, a main contribution of this paper is the introduction of a new

[1] A more formal definition of factor pervasiveness in given in the next section of this paper.

variable selection procedure that allows empirical researchers to, with probability approaching one, correctly distinguish the relevant from the irrelevant variables, prior to factor estimation. We study this problem within a factor-augmented VAR (FAVAR) framework, thus allowing time series forecasts to be made using an information set much richer than that used in traditional VAR models. Our results 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, consistent estimates of the conditional mean function of said equations is achieved. This 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 violation of factor pervasiveness is such that consistent estimation is precluded in the absence of variable pre-screening. Moreover, there are also clear benefits to using our procedure in cases where weaker pervasiveness assumptions, such as that discussed in Bai and Ng (2021), characterize the data.

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 the model that they consider is very different from the type of dynamic vector time series model studied here.

In another important related 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, 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, 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. In this context, 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).

Finally, it is worth pointing out that our variable selection procedure differs substantially from the approach to variable/model selection taken in much of the traditional econometrics literature. In particular, we show that important moderate deviation results obtained recently by Chen, Shao, Wu, and Xu (2016) can be used to help control the probability of a Type I error, i.e., the error that an irrelevant variable which is not informative about the underlying factors is falsely selected as a relevant variable. This is so in situations where the number of irrelevant variables is very large, even if the tails of the underlying distributions do not satisfy the kind of sub-exponential behavior typically assumed by large deviation inequalities used in high-dimensional analysis. Hence, we are able to design a variable selection procedure where the probability of a Type I error goes to zero, as the sample sizes grows to infinity. This fact, taken together with the fact that the probability of a Type II error for our procedure also goes to zero asymptotically, allows us to establish that our variable selection procedure is completely consistent, in the sense that both Type I and Type II errors go to zero in the limit. This property of complete consistency is important because if we try to simply control the probability of a Type I error at some predetermined level, which is the typical approach in multiple hypothesis testing, then we will not in general be able to estimate the factors consistently, even up to an invertible matrix transformation, and in consequence, we will have fallen short of our ultimate goal of obtaining a consistent estimate of the conditional mean function of the factor-augmented forecasting equation.

The rest of the paper is organized as follows. In Section 2 , we provide our counterexample, stated formally as Theorem 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 and the assumptions that we impose on this model. We also describe our variable selection procedure and provide theoretical results establishing the complete consistency of the procedure. Section 4 provides 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 a promising Monte Carlo study on the finite sample performance of our variable selection method, and makes recommendations regarding calibration of the tuning parameter used in said method. Finally, Section 6 offers concluding remarks. The technical details of the paper are organized in four appendices. Appendix A provides proofs of the main theorems. Appendix B provides proofs of lemmas used in proving Theorem 1, and Appendix C contains proofs of supporting lemmas, used primarily in the proofs of Theorems 2 and 3. Finally, Appendix D contains the proofs of supporting lemmas used primarily in the proofs of Theorems 4 and 5.

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 $\lfloor \cdot \rfloor$ denote the floor function, so the $\lfloor x \rfloor$ 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 PZ_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 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,

[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.

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 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

Consider 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

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

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

Here, Y_t denotes the vector of observable economic variables, and F_t is a vector of unobserved (latent) factors. In our analysis of this model, it will often be convenient to rewrite the FAVAR in several alternative forms, such as when making assumptions used in the sequel. We thus briefly outline two alternative representations of the above model. First, it is easy to see that the system of equations given in (2) can be written in the form:

$$\begin{aligned} Y_{t+1} &= \mu_Y + A_{YY}\underline{Y}_t + A_{YF}\underline{F}_t + \varepsilon^Y_{\cdot t+1}, &(3)\\ F_{t+1} &= \mu_F + A_{FY}\underline{Y}_t + A_{FF}\underline{F}_t + \varepsilon^F_{\cdot t+1}, &(4)\end{aligned}$$

where

$$\begin{aligned} \underset{d\times dp}{A_{YY}} &= \begin{pmatrix} A_{YY,1} & A_{YY,2} & \cdots & A_{YY,p} \end{pmatrix},\ \underset{d\times Kp}{A_{YF}} = \begin{pmatrix} A_{YF,1} & A_{YF,2} & \cdots & A_{YF,p} \end{pmatrix},\\ \underset{K\times dp}{A_{FY}} &= \begin{pmatrix} A_{FY,1} & A_{FY,2} & \cdots & A_{FY,p} \end{pmatrix},\ \underset{K\times Kp}{A_{FF}} = \begin{pmatrix} A_{FF,1} & A_{FF,2} & \cdots & A_{FF,p} \end{pmatrix},\\ \underset{dp\times 1}{\underline{Y}_t} &= \begin{pmatrix} Y_t \\ Y_{t-1} \\ \vdots \\ Y_{t-p+1} \end{pmatrix}, \text{ and } \underset{Kp\times 1}{\underline{F}_t} = \begin{pmatrix} F_t \\ F_{t-1} \\ \vdots \\ F_{t-p+1} \end{pmatrix}. \qquad (5)\end{aligned}$$

Another useful representation of the FAVAR model is the so-called companion form, wherein the p^{th}-order model given in expression (2) is written in terms of a first-order model:

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

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

$$\alpha = \begin{pmatrix} \mu \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, A = \begin{pmatrix} A_1 & A_1 & \cdots & A_{p-1} & A_p \\ I_d & 0 & \cdots & 0 & 0 \\ 0 & I_d & \ddots & \vdots & 0 \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & I_d & 0 \end{pmatrix}, \text{ and } E_t = \begin{pmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}. \qquad (6)$$

This companion form 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 Appendix C.

In addition to observations on Y_t, suppose that the data set available to researchers includes a vector of time series variables which are related to the unobserved factors in the following manner:

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

where the properties of u_t are given in Assumptions 3-3 and 3-4, below. Now, assume that not all components of Z_t provide useful information for estimating the unobserved vector, $\underline{F}_t$, so 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:

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

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

Hence, similar to what has been discussed in the previous section, 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

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

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

The above representation suggests that the components of $Z_t^{(1)}$ can be interpreted as some sort of "information" variables, as the information that they supply will be helpful in 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 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.

To provide a variable selection procedure with provable guarantees, we must first specify a number of conditions on the FAVAR model defined above.

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 $M < \infty, \sup_t \displaystyle\int |g_{\varepsilon_t}(\upsilon - u) - g_{\varepsilon_t}(\upsilon)| \, d\varepsilon \leq M|u|$, 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

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

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,

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

Remark 3.1:

(a) Note that Assumption 3-1 is the stability condition that one typically assumes for a stationary VAR process. One difference is that 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.

(b) It is well known that $\det\left\{I_{(d+K)} - Az\right\} = \det\left\{I_{(d+K)} - A_1 z - \cdots - A_p z^p\right\}$, where A is the coefficient matrix of the companion form given in expression (6). See, for example, page 16 of Lütkepohl (2005). It follows that Assumption 3-1 is equivalent to the condition that

$$\det\left\{I_{(d+K)} - Az\right\} = 0 \text{ implies that } |z| > 1. \tag{15}$$

In addition, Assumption 3-1 is also, of course, equivalent to the assumption that all eigenvalues of A have modulus less than 1.

(c) 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 $N_1/N \to 0$.

(d) 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. More primitive conditions for such a relationship between the singular values and the eigenvalues of a (not necessarily symmetric) matrix have been studied in the linear algebra literature. In Appendix C of this paper, we prove one such result which extends a well-known result by Ruhe (1975). More specifically, we state and prove the following lemma:

Lemma C-9: *Let A be an $n \times n$ square matrix with (ordered) singular values given by:*

$$\sigma_{(1)}(A) \geq \sigma_{(2)}(A) \geq \cdots \geq \sigma_{(n)}(A) \geq 0.$$

Suppose that A is diagonalizable, i.e., $A = S\Lambda S^{-1}$, where Λ is diagonal matrix whose diagonal *elements are the eigenvalues of A. Let the modulus of these eigenvalues be ordered as follows:*

$$\left|\lambda_{(1)}(A)\right| \geq \left|\lambda_{(2)}(A)\right| \geq \cdots \geq \left|\lambda_{(n)}(A)\right|.$$

Then, for $k \in \{1, ..., n\}$ and for any positive integer j, we have that:

$$\chi(S)^{-1}\left|\lambda_{(k)}\left(A^j\right)\right| \leq \sigma_{(k)}\left(A^j\right) \leq \chi(S)\left|\lambda_{(k)}\left(A^j\right)\right|$$

where

$$\chi(S) = \sigma_{(1)}(S)\,\sigma_{(1)}\left(S^{-1}\right).$$

Note that in the special case where the matrices A and $I_{(d+K)p} - A$ are diagonalizable, the inequalities given in expressions (13) and (14) are a direct consequence of this lemma. On the other hand, Assumption 3-7 takes into account other situations where expressions (13) and (14) are valid even though the matrices A and $I_{(d+K)p} - A$ are not diagonalizable.

(e) Assumptions 3-1, 3-2(a)-(c), and 3-7 together allow us to show in Lemma C-11 of Appendix C 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.

Our variable selection procedure is based on a self-normalized statistic and makes use of some pathbreaking moderate deviation results for weakly dependent processes recently obtained by Chen, Shao, Wu, and Xu (2016). An advantage of using a self-normalized statistic is that doing so allows us to impose much weaker moment conditions, even when N is much larger than T. In particular, as can be seen from Assumptions 3-2 and 3-3 above, we only make moment conditions that are of a polynomial order on the errors processes $\{\varepsilon_t\}$ and $\{u_{it}\}$. Such conditions are substantially weaker than assumptions of Gaussianity or sub-Gaussianity which have been made in papers studying high-dimensional factor models and/or high-dimensional covariance matrices, without employing statistics that are self-normalized[4].

To accommodate data dependence, we 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. Two such statistics are proposed in this paper. The first of these statistics has the form of an ℓ_∞ norm and is given by:

$$\max_{1\leq\ell\leq d}|S_{i,\ell,T}| = \max_{1\leq\ell\leq d}\left|\frac{\overline{S}_{i,\ell,T}}{\sqrt{\overline{V}_{i,\ell,T}}}\right|, \tag{16}$$

where

$$\overline{S}_{i,\ell,T} = \sum_{r=1}^{q}\sum_{t=(r-1)\tau+p}^{(r-1)\tau+\tau_1+p-1} Z_{it}y_{\ell,t+1} \text{ and} \tag{17}$$

$$\overline{V}_{i,\ell,T} = \sum_{r=1}^{q}\left[\sum_{t=(r-1)\tau+p}^{(r-1)\tau+\tau_1+p-1} Z_{it}y_{\ell,t+1}\right]^2. \tag{18}$$

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 (16) 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}. Our second statistic has the form of a pseudo-L_1 norm and is given by:

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

where $\overline{S}_{i,\ell,T}$ and $\overline{V}_{i,\ell,T}$ are as defined in expressions (17) and (18) 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 $\sum_{\ell=1}^{d}\varpi_\ell = 1$. Both of these statistics employ a blocking scheme similar to that proposed in Chen, Shao, Wu, and Xu (2016), where, in order to keep the effects of dependence under control, the construction of these

[4] For example, see Bickel and Levina (2008) and Fan, Liao, and Mincheva (2013).

statistics is based only on observations in every other block. To see this, note that if we write out the "numerator" term $\overline{S}_{i,\ell,T}$ in greater detail, we have that:

$$\begin{aligned}\overline{S}_{i,\ell,T} &= \sum_{t=p}^{\tau_1+p-1} Z_{it}y_{\ell,t+1} + \sum_{t=\tau+p}^{\tau+\tau_1+p-1} Z_{it}y_{\ell,t+1} \\ &\quad + \sum_{t=2\tau+p}^{2\tau+\tau_1+p-1} Z_{it}y_{\ell,t+1} + \cdots + \sum_{t=(q-1)\tau+p}^{(q-1)\tau+\tau_1+p-1} Z_{it}y_{\ell,t+1} \qquad (19)\end{aligned}$$

Comparing the first term and the second terms on the right-hand side of expression (19), we see that the observations $Z_{it}y_{\ell,t+1}$, for $t = \tau_1 + p, ..., \tau - p$, have not been included in the construction of the sum. Similar observations hold when comparing the second and the third terms, and so on.

It should also be pointed out that although we make use of some of their fundamental results on moderate deviation, both the model studied in our paper and the objective of our paper are very different from that of Chen, Shao, Wu, and Xu (2016). Whereas Chen, Shao, Wu, and Xu **(2016)** focus their analysis on problems of testing and inference for the mean of a scalar weakly dependent time series using self-normalized Student-type test statistics, our paper applies the self-normalization approach to a variable selection problem in a FAVAR setting. Indeed, the problem which we study here is in some sense more akin to a classification (or model selection) problem rather than a multiple hypothesis testing problem. In order to consistently estimate the factors up to an invertible matrix transformation, we need to develop a variable selection procedure whereby both the probability of a false positive and the probability of a false negative converge to zero as N_1, N_2, $T \to \infty$[5]. 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.

To determine whether the i^{th} component of Z_t is a relevant variable for the purpose of factor estimation, we propose the following procedure. Define $i \in \widehat{H}^c$ to indicate that the procedure has classified Z_{it} to be a relevant variable for the purpose of factor estimation. **Similarly,** define $i \in \widehat{H}$ to indicate that the procedure has classified Z_{it} to be an irrelevant variable. Now, let $\mathbb{S}^+_{i,T}$ denote either the statistic $\max_{1\leq\ell\leq d} |S_{i,\ell,T}|$ or the statistic $\sum_{\ell=1}^{d} \varpi_\ell |S_{i,\ell,T}|$. Our variable selection

[5] 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.

procedure is based on the decision rule:

$$i \in \begin{cases} \widehat{H}^c & \text{if } \mathbb{S}_{i,T}^{+} > \Phi^{-1}\left(1 - \frac{\varphi}{2N}\right) \\ \widehat{H} & \text{if } \mathbb{S}_{i,T}^{+} \leq \Phi^{-1}\left(1 - \frac{\varphi}{2N}\right) \end{cases}, \tag{20}$$

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.

Remarks 3.2:

(a) To understand why using the quantile function of the standard normal as the threshold function for our procedure is a natural choice, note first that, by applying Lemma C-17 given in Appendix C below, we obtain that:

$$P\left(|S_{i,\ell,T}| \geq z\right) \leq 2\left[1 - \Phi(z)\right](1 + o(1)), \tag{21}$$

which holds for all i and ℓ, for all T sufficiently large, and for all z such that:

$$0 \leq z \leq c_0 \min\left\{T^{(1-\alpha_1)\frac{1}{6}}, T^{\frac{\alpha_2}{2}}\right\}.$$

In view of expression (21), we can interpret moderate deviation as providing an asymptotic approximation of the (two-sided) tail behavior of the statistic, $S_{i,\ell,T}$, based on the tails of the standard normal distribution. Indeed, the inequality in (21) can be easily made into an equality by imposing a mild additional restriction on the range for z[6]. Now, suppose initially that we wish simply to control the probability of a Type I error for testing the null hypothesis $H_0 : \gamma_i = 0$ (i.e., the i^{th} variable does not load on the underlying factors) at some fixed level α. Then, expression (21) suggests that a natural way to do this is to set $z = \Phi^{-1}(1 - \alpha/2)$. This is because, given that the quantile function $\Phi^{-1}(\cdot)$ is, by definition, the inverse function of the cdf $\Phi(\cdot)$, we have that:

$$\begin{aligned} P\left(|S_{i,\ell,T}| \geq \Phi^{-1}(1 - \alpha/2)\right) &\leq 2\left[1 - \Phi\left(\Phi^{-1}(1 - \alpha/2)\right)\right](1 + o(1)) \\ &= \alpha(1 + o(1)), \end{aligned}$$

[6] For example, if we restrict z to lie in the range:

$$0 \leq z \leq c_0 \min\left\{T^{(1-\alpha_1)\frac{1}{6}}/L(T), T^{\frac{\alpha_2}{2}}\right\},$$

for some slowly varying function $L(T)$, such that $L(T) \to \infty$, then, we can show that:

$$P\left(|S_{i,\ell,T}| \geq z\right) = 2\left[1 - \Phi(z)\right](1 + o(1)).$$

so that the probability of a Type I error is controlled at the desired level, α, asymptotically. Note also that an advantage of moderate deviation theory is that it gives a characterization of the relative approximation error, as opposed to the absolute approximation error. As a result, the approximation given is useful and meaningful even when α is very small, which is of importance to us since we are interested in situations where we might want to let α go to zero, as sample size approaches infinity.

We give the above example to provide intuition concerning the form of the threshold function that we have specified. The variable selection problem that we actually consider is more complicated than what is illustrated by this example, since we need to control the probability of a Type I error (or of a false positive) not just for a single test involving the i^{th} variable but for all variables simultaneously. Moreover, as noted previously, we also need the probability of a false positive to go to zero asymptotically, if we want to be able to estimate factors consistently, even up to an invertible matrix transformation. We show in Theorem 2 below that these objectives can all be accomplished using the threshold function specified in expression (20), since a threshold function of this form makes it easy for us to properly control the probability of a false positive in large samples.
(b) The threshold function used here is reminiscent of the one employed in a celebrated paper by Belloni, Chen, Chernozhukov, and Hansen (2012). More specifically, Belloni, Chen, Chernozhukov, and Hansen (2012) use a similar threshold function to help set the penalty level for Lasso estimation of the first-stage equation of an IV regression model assuming *i.n.i.d.* data. In spite of the similarity in the form of the threshold function, the problem studied in that paper is very different from the one which we analyze here. In consequence, the conditions we will specify on the setting of the tuning parameter, φ, will also be quite different from what they recommend in their paper.

Under appropriate conditions, the variable selection procedure described above can be shown to be consistent, in the sense that both the probability of a false positive, i.e. $P\left(i \in \widehat{H}^c | \iota \in H\right)$, and the probability of a false negative, i.e., $P\left(i \in \widehat{H} | \iota \in H^c\right)$, approach zero as N, $T \to \infty$. To show this result, we must first state a number of additional 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.

Remarks 3.3:

(a) Assumption 3-9 imposes the condition that there exists a positive constant, $\underline{c}$, such that, for all N_1, N_2, and T sufficiently large:

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

This is a fairly mild condition which allows us to differentiate the alternative hypothesis, $i \in H^c$, from the null hypothesis, $i \in H$, since if $i \in H$, then it is clear that:

$$\frac{\mu_{i,\ell,T}}{q\tau_1} = \frac{1}{q} \sum_{r=1}^{q} \frac{1}{\tau_1} \sum_{t=(r-1)\tau+p}^{(r-1)\tau+\tau_1+p-1} \gamma_i' \left\{ E\left[\underline{F}_t\right] \mu_{Y,\ell} + E\left[\underline{F}_t \underline{Y}_t'\right] \alpha_{YY,\ell} + E\left[\underline{F}_t \underline{F}_t'\right] \alpha_{YF,\ell} \right\} = 0,$$

given that $\gamma_i = 0$. Note that this assumption does rule out certain specialized situations, such as the case when $\mu_{Y,\ell} = 0$, $\alpha_{YY,\ell} = 0$, and $\alpha_{YF,\ell} = 0$, for some $\ell \in \{1, ..., d\}$. However, we do not consider such cases to be of much practical interest since, for example, if $\mu_{Y,\ell} = 0$, $\alpha_{YY,\ell} = 0$, and $\alpha_{YF,\ell} = 0$ for some ℓ then expression (3) above implies that the ℓ^{th} component of Y_{t+1} will have the representation

$$\begin{aligned} y_{\ell,t+1} &= \mu_{Y,\ell} + \underline{Y}_t' \alpha_{YY,\ell} + \underline{F}_t' \alpha_{YF,\ell} + \varepsilon^Y_{\ell,t+1} \\ &= \varepsilon^Y_{\ell,t+1}, \end{aligned}$$

so that, in this case, $y_{\ell,t+1}$ depends neither on $\underline{Y}_t = \left(Y_t', Y_{t-1}', ..., Y_{t-p+1}'\right)'$ nor on $\underline{F}_t = \left(F_t', F_{t-1}', ..., F_{t-p+1}'\right)$. This is, of course, an unrealistic model for $y_{\ell,t+1}$ since it would not even be dependent.

(b) Bai and Ng (2008) address the important issue that one should choose the predictor variables Z_{it} based on their predictability for Y_{t+1}. While we agree with their viewpoint overall, it is worth stressing that for the FAVAR model considered here, whether Z_{it} helps to predict some future values of Y_t (say, 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}. Note that, for a given t, the precision with which $\underline{F}_t$ is estimated depends primarily on the size of the cross-sectional dimension, and the exclusion of any relevant Z_{it} (with $\gamma_i \neq 0$) will have the negative effect of reducing the sample size used for this estimation. More importantly, as we will discuss in greater details in Remark 4.2 below, 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. While the question of predictability is certainly an important one, the answer we get for this question can, in some situations, be at odds with the objective of achieving consistent factor estimation. This is because while $\gamma_i' = 0$ does imply that $Z_{i\cdot}$ will not be helpful for predicting future values of Y, the reverse is not necessarily true. On the other hand, to ensure consistent estimation of the factors, we want to use every data point Z_{it} for which the null hypothesis $\gamma_i' = 0$ is rejected. Moreover, if it is true that some of the factors load primarily on variables which are uninformative predictors for certain components of Y_{t+h}, then that will show up in the form of certain parameter restrictions on the forecasting equation, in which case the best way to address this problem is to perform hypothesis testing or model selection on the forecasting equation itself, after the unobserved factors have first been properly estimated.

The following two theorems give our main theoretical results on the variable selection procedure described above.

Theorem 2: *Let $H = \{k \in \{1, ..., N\} : \gamma_k = 0\}$. Suppose that Assumptions 3-1, 3-2(a)-(c), 3-3(a)-(c) 3-4, 3-5, 3-7, 3-8, 3-10 (a) and 3-11 hold. Let $\Phi^{-1}(\cdot)$ denote the inverse of the cumulative distribution function of the standard normal random variable, or, alternatively, the quantile function of the standard normal distribution. Then the following statements are true:*

(a) *Let $\{\varpi_\ell : \ell = 1, .., d\}$ be pre-specified weights, such that $\varpi_\ell \geq 0$, for every $\ell \in \{1, ..., d\}$ and $\sum_{\ell=1}^{d} \varpi_\ell = 1$, then:*

$$P\left(\max_{i \in H} \sum_{\ell=1}^{d} \varpi_\ell \left|S_{i,\ell,T}\right| > \Phi^{-1}\left(1 - \frac{\varphi}{2N}\right)\right) = O\left(\frac{N_2 \varphi}{N}\right) = o(1),$$

where $N = N_1 + N_2$.

(b)

$$P\left(\max_{i \in H} \max_{1 \leq \ell \leq d} \left|S_{i,\ell,T}\right| > \Phi^{-1}\left(1 - \frac{\varphi}{2N}\right)\right) = O\left(\frac{N_2 \varphi}{N}\right) = o(1).$$

Theorem 3: *Let $H^c = \{k \in \{1, ..., N\} : \gamma_k \neq 0\}$. Suppose that Assumptions 3-1, 3-2(a)-(c), 3-3(a)-(c), 3-5, 3-7, 3-9, 3-10, and 3-11 hold. Then the following statements are true.*

(a) *Let $\{\varpi_\ell : \ell = 1, .., d\}$ be pre-specified weights, such that $\varpi_\ell \geq 0$, for every $\ell \in \{1, ..., d\}$ and $\sum_{\ell=1}^{d} \varpi_\ell = 1$, then:*

$$P\left(\min_{i \in H^c} \sum_{\ell=1}^{d} \varpi_\ell \left|S_{i,\ell,T}\right| > \Phi^{-1}\left(1 - \frac{\varphi}{2N}\right)\right) \to 1.$$

(b)

$$P\left(\min_{i \in H^c} \max_{1 \leq \ell \leq d} \left|S_{i,\ell,T}\right| > \Phi^{-1}\left(1 - \frac{\varphi}{2N}\right)\right) \to 1.$$

Remark 3.4:

(a) Theorem 2 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 $N, T \to \infty$. Theorem 3 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 $N, T \to \infty$.

(b) Note that as a by-product, our variable selection procedure provides us with an 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. As we show in part (a) of Lemma D-15 in Appendix D, $\widehat{N}_1$ is a consistent estimator of N_1, in the sense that $\widehat{N}_1/N_1 \xrightarrow{p} 1$. Moreover, note that the rate condition given in Bai and Ng (2021) for consistent factor estimation (i.e., Assumption A4 in their paper) can be restated in our setup as the assumption that $N/(TN_1) \to 0$. Now, because N_1 is not observed, this rate condition, by itself, only provides a rough guide to empirical researchers wishing to assess whether factors can be estimated accurately in the particular problem of interest to them. Viewed from this perspective, what we propose here actually builds on the work of Bai and Ng (2021), as our procedure helps to highlight the importance of the rate condition they have introduced and provides additional information that is useful to empirical researchers about the degree of pervasiveness of the underlying factors.

(c) Note, in addition, that knowledge of the number of factors is not needed to implement our 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. In ongoing research, we plan to show that doing so will lead to consistent estimation of the number of factors.

4 Consistent Estimation of 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)

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

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

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

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. Moreover, Monte Carlo results reported in Section 5 of this paper show that our variable selection procedure 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 $\underline{F}_t = \left(F_t', F_{t-1}', ..., F_{t-p+1}'\right)$:

$$\widehat{\underline{F}}_t = \frac{\widehat{\Gamma}' Z_{t,N}\left(\widehat{H^c}\right)}{\widehat{N}_1}, \tag{22}$$

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 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 (22) consistently estimates the unobserved factors $\underline{F}_t$,up to an invertible $Kp \times Kp$ matrix transformation.

Theorem 4: *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 as in expression (22). Assume further that the specification of the tuning* *parameter, φ, in the decision rule (20) 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

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

Remark 4.2:

If we examine the proof of Theorem 4 in Appendix A and the supporting arguments given in the proof of Lemma D-15 of Appendix D, 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 in Remark 3.3(b) above, 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 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, we show that, for h-step ahead forecast, the (infeasible) forecasting equation implied by the FAVAR model (2) has the form

$$Y_{t+h} = \beta_0 + B_1' \underline{Y}_t + B_2' \underline{F}_t + \eta_{t+h}, \tag{23}$$

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

$$\begin{aligned} \beta_0 &= \sum_{j=0}^{h-1} J_d A^j \alpha,\ B_1' = J_d A^h \mathcal{P}_{(d+K)p}' S_d,\ B_2' = J_d A^h \mathcal{P}_{(d+K)p}' S_K \text{ and} \\ \eta_{t+h} &= \sum_{j=0}^{h-1} J_d A^j J_{d+K}' \varepsilon_{t+h-j}. \end{aligned} \tag{24}$$

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 for a derivation of the equation given in expression (23). The reason expression (23) 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

(23) with the estimate $\underline{\widehat{F}}_t$ given in expression (22). Doing so, we arrive at a feasible h-step ahead forecasting equation of the form:

$$\begin{aligned} Y_{t+h} &= \beta_0 + \sum_{g=1}^{p} B_{1,g}' Y_{t-g+1} + \sum_{g=1}^{p} B_{2,g}' \widehat{F}_{t-g+1} + \widehat{\eta}_{t+h} \\ &= \beta_0 + B_1' \underline{Y}_t + B_2' \underline{\widehat{F}}_t + \widehat{\eta}_{t+h}, \end{aligned} \tag{25}$$

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

One can interpret expression (25) 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 (25) 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 $\underline{\widehat{F}}_t$ denotes the factor estimates given in expression (22). Our procedure minimizes the least squares criterion function:

$$\begin{aligned} Q(\beta_0, B_1, B_2) &= \sum_{t=p}^{T-h} \left\| Y_{t+h} - \beta_0 - B_1' \underline{Y}_t - B_2' \underline{\widehat{F}}_t \right\|_2^2 \\ &= \sum_{t=p}^{T-h} \left\| Y_{t+h} - \beta_0 - \sum_{g=1}^{p} B_{1,g}' Y_{t-g+1} - \sum_{g=1}^{p} B_{2,g}' \widehat{F}_{t-g+1} \right\|_2^2 \end{aligned} \tag{26}$$

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:

$$\widehat{Y}_{T+h} = \widehat{\beta}_0 + \widehat{B}_1' \underline{Y}_T + \widehat{B}_2' \widehat{\underline{F}}_T. \tag{27}$$

The following result shows that $\widehat{Y}_{T+h}$ is a consistent estimator of the conditional mean of the infeasible forecast equation (23).

Theorem 5: *Let $\widehat{Y}_{T+h}$ be as defined in expression (27). 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 Monte Carlo Study

In this section, we report some simulation results on the finite sample performance of our variable selection procedure. The model used in the Monte Carlo study is the following tri-variate FAVAR(1) process:

$$\begin{aligned} W_t &= \mu + A W_{t-1} + \varepsilon_t, && (28) \\ Z_t &= \gamma F_t + u_t, && (29) \end{aligned}$$

where

$$W_t = \begin{pmatrix} Y_{1t} \\ Y_{2t} \\ F_t \end{pmatrix}, \mu = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}, A = \begin{pmatrix} 0.9 & 0.3 & 0.5 \\ 0 & 0.7 & 0.1 \\ 0 & 0.6 & 0.7 \end{pmatrix}, \text{ and } \gamma = \begin{pmatrix} \iota_{N_1} \\ \underset{N_2 \times 1}{0} \end{pmatrix},$$

with ι_{N_1} denoting an $N_1 \times 1$ vector of ones. We consider different configurations of N, N_1, and T, as given in the tables below. For the error process in equation (28), we take $\{\varepsilon_t\} \equiv i.i.d.N\left(0, \Sigma_\varepsilon\right)$, where:

$$\Sigma_\varepsilon = \begin{pmatrix} 1.3 & 0.99 & 0.641 \\ 0.99 & 0.81 & 0.009 \\ 0.641 & 0.009 & 5.85 \end{pmatrix}.$$

The error process, $\{u_{it}\}$, in equation (29) is allowed to exhibit both temporal and cross-sectional dependence and also conditional heteroskedasticity. More specifically, we let:

$$u_{it} = 0.8 u_{it-1} + \zeta_{it},$$

and, following the approach for modeling cross-sectional dependence given in the Monte Carlo design of Stock and Watson (2002a), we specify:

$$\zeta_{it} = \left(1 + b^2\right)\eta_{it} + b\eta_{i+1,t} + b\eta_{i-1,t},$$

and in the experiments given below, we set $b = 1$. In addition, $\eta_{it} = \omega_{it}\xi_{it}$, with $\{\xi_{it}\} \equiv i.i.d.N\,(0,1)$ independent of $\{\varepsilon_t\}$, and ω_{it} follows a GARCH(1,1) process given by

$$\omega_{it}^2 = 1 + 0.9\omega_{it-1}^2 + 0.05\eta_{it-1}^2.$$

To study the effects of varying the tuning parameter, we let $\varphi = N^{-\vartheta}$, and consider six different values of ϑ, i.e., $\vartheta = 0.2$, 0.3, 0.4, 0.5, 0.6, and 0.7. We also attempt to shed light on the effects of forming blocks of different sizes on the performance of our procedure. To do this, for $T = 100$, we set $\tau_1 = 2$, 3, 4, and 5; for $T = 200$, we set $\tau_1 = 5$, 6, 8, and 10; and for $T = 600$, we set $\tau_1 = 6$, 8, 10, and 12. In addition, we present results for both statistics, i.e. $\max_{1\leq\ell\leq d}|S_{i,\ell,T}|$ and $\sum_{\ell=1}^{d}\varpi_\ell\,|S_{i,\ell,T}|$. Note that $d = 2$ in our setup; and, for the statistic $\sum_{\ell=1}^{2}\varpi_\ell\,|S_{i,\ell,T}|$, we set $\varpi_1 = \varpi_2 = 1/2$.

The results of our Monte Carlo study are reported in Tables 1-8 below. In these tables, we let FPR denote the "False Positive Rate" or the "Type I" error rate, i.e., the proportion of cases where an irrelevant variable Z_{it}, with associated coefficient $\gamma_i = 0$, is erroneously selected as a relevant variable. We let FNR denote the "False Negative Rate" or the "Type II" error rate, i.e., the proportion of cases where a relevant variable is erroneously identified as being irrelevant.

Table 1: $\mathbb{S}^{+}_{i,T} = \max_{1\leq\ell\leq d}|S_{i,\ell,T}|$

		$N = 100$	$N_1 = 50$	$T = 100$	$\tau = 5$		
		$\varphi = N^{-0.2}$	$\varphi = N^{-0.3}$	$\varphi = N^{-0.4}$	$\varphi = N^{-0.5}$	$\varphi = N^{-0.6}$	$\varphi = N^{-0.7}$
$\tau_1 = 2$	FPR	0.01690	0.00960	0.00464	0.00218	0.00096	0.00034
	FNR	0.00218	0.00548	0.01328	0.03204	0.07274	0.15890
$\tau_1 = 3$	FPR	0.02078	0.01156	0.00632	0.00288	0.00128	0.00048
	FNR	0.00126	0.00350	0.00866	0.02234	0.05374	0.12050
$\tau_1 = 4$	FPR	0.02544	0.01468	0.00826	0.00408	0.00194	0.00070
	FNR	0.00090	0.00228	0.00582	0.01582	0.04010	0.09362
$\tau_1 = 5$	FPR	0.03208	0.01980	0.01100	0.00584	0.00288	0.00122
	FNR	0.00052	0.00164	0.00430	0.01140	0.02988	0.07190

Results based on 1000 simulations

Table 2: $\mathbb{S}^{+}_{i,T} = \sum_{\ell=1}^{d}\varpi_\ell\,|S_{i,\ell,T}|$

		$N = 100$	$N_1 = 50$	$T = 100$	$\tau = 5$		
		$\varphi = N^{-0.2}$	$\varphi = N^{-0.3}$	$\varphi = N^{-0.4}$	$\varphi = N^{-0.5}$	$\varphi = N^{-0.6}$	$\varphi = N^{-0.7}$
$\tau_1 = 2$	FPR	0.01460	0.00810	0.00382	0.00174	0.00076	0.00028
	FNR	0.00284	0.00700	0.01674	0.04058	0.09412	0.19952
$\tau_1 = 3$	FPR	0.01810	0.00996	0.00526	0.00226	0.00092	0.00032
	FNR	0.00172	0.00450	0.01100	0.02860	0.06942	0.15378
$\tau_1 = 4$	FPR	0.02224	0.01276	0.00702	0.00338	0.00162	0.00044
	FNR	0.00118	0.00310	0.00828	0.02082	0.05194	0.12132
$\tau_1 = 5$	FPR	0.02796	0.01714	0.00924	0.00502	0.00232	0.00080
	FNR	0.00084	0.00222	0.00574	0.01508	0.03948	0.09456

Results based on 1000 simulations

Table 3: $\mathbb{S}^{+}_{i,T} = \max_{1\leq\ell\leq d} |S_{i,\ell,T}|$

		$N = 200$	$N_1 = 100$	$T = 100$	$\tau = 5$		
		$\varphi = N^{-0.2}$	$\varphi = N^{-0.3}$	$\varphi = N^{-0.4}$	$\varphi = N^{-0.5}$	$\varphi = N^{-0.6}$	$\varphi = N^{-0.7}$
$\tau_1 = 2$	FPR	0.00578	0.00239	0.00085	0.00020	0.00005	0.00000
	FNR	0.01074	0.02997	0.07812	0.18957	0.39889	0.68275
$\tau_1 = 3$	FPR	0.00775	0.00324	0.00126	0.00038	0.00006	0.00001
	FNR	0.00724	0.02088	0.05676	0.14547	0.32908	0.60780
$\tau_1 = 4$	FPR	0.00981	0.00457	0.00170	0.00057	0.00014	0.00002
	FNR	0.00517	0.01494	0.04224	0.11350	0.27048	0.53471
$\tau_1 = 5$	FPR	0.01334	0.00609	0.00266	0.00094	0.00023	0.00004
	FNR	0.00362	0.01133	0.03244	0.08901	0.22162	0.46424

Results based on 1000 simulations

Table 4: $\mathbb{S}^{+}_{i,T} = \sum_{\ell=1}^{d} \varpi_\ell |S_{i,\ell,T}|$

		$N = 200$	$N_1 = 100$	$T = 100$	$\tau = 5$		
		$\varphi = N^{-0.2}$	$\varphi = N^{-0.3}$	$\varphi = N^{-0.4}$	$\varphi = N^{-0.5}$	$\varphi = N^{-0.6}$	$\varphi = N^{-0.7}$
$\tau_1 = 2$	FPR	0.00486	0.00196	0.00064	0.00014	0.00002	0.00000
	FNR	0.01415	0.03813	0.09966	0.23933	0.48356	0.77511
$\tau_1 = 3$	FPR	0.00657	0.00268	0.00098	0.00024	0.00005	0.00001
	FNR	0.00921	0.02714	0.07372	0.18714	0.40894	0.70884
$\tau_1 = 4$	FPR	0.00841	0.00378	0.00133	0.00043	0.00004	0.00002
	FNR	0.00661	0.01975	0.05564	0.14734	0.34279	0.63906
$\tau_1 = 5$	FPR	0.01124	0.00509	0.00213	0.00069	0.00017	0.00002
	FNR	0.00477	0.01475	0.04258	0.11741	0.28620	0.56845

Results based on 1000 simulations

Table 5: $\mathbb{S}_{i,T}^{+} = \max_{1 \leq \ell \leq d} |S_{i,\ell,T}|$

		$N = 400$	$N_1 = 200$	$T = 200$	$\tau = 10$		
		$\varphi = N^{-0.2}$	$\varphi = N^{-0.3}$	$\varphi = N^{-0.4}$	$\varphi = N^{-0.5}$	$\varphi = N^{-0.6}$	$\varphi = N^{-0.7}$
$\tau_1 = 5$	FPR	0.00035	0.00009	0.00003	0.00001	0.00000	0.00000
	FNR	0.00200	0.01116	0.05764	0.23070	0.61173	0.94453
$\tau_1 = 6$	FPR	0.00040	0.00010	2.5×10^{-5}	5.0×10^{-6}	0.00000	0.00000
	FNR	0.00128	0.00740	0.04154	0.18482	0.54582	0.92176
$\tau_1 = 8$	FPR	0.00054	0.00015	0.00005	0.00001	0.00000	0.00000
	FNR	0.00054	0.00369	0.02191	0.11627	0.41851	0.85806
$\tau_1 = 10$	FPR	0.00093	0.00031	0.00008	1.5×10^{-5}	5.0×10^{-6}	0.00000
	FNR	0.00026	0.00194	0.01218	0.07226	0.30765	0.76833

Results based on 1000 simulations

Table 6: $\mathbb{S}_{i,T}^{+} = \sum_{\ell=1}^{d} \varpi_\ell |S_{i,\ell,T}|$

		$N = 400$	$N_1 = 200$	$T = 200$	$\tau = 10$		
		$\varphi = N^{-0.2}$	$\varphi = N^{-0.3}$	$\varphi = N^{-0.4}$	$\varphi = N^{-0.5}$	$\varphi = N^{-0.6}$	$\varphi = N^{-0.7}$
$\tau_1 = 5$	FPR	0.00030	8.5×10^{-5}	2.5×10^{-5}	5.0×10^{-6}	0.00000	0.00000
	FNR	0.00231	0.01355	0.06894	0.26683	0.67266	0.96749
$\tau_1 = 6$	FPR	0.00034	9.5×10^{-5}	0.00002	5.0×10^{-6}	0.00000	0.00000
	FNR	0.00148	0.00901	0.05058	0.21713	0.60968	0.95287
$\tau_1 = 8$	FPR	0.00046	0.00013	0.00004	0.00001	0.00000	0.00000
	FNR	0.00068	0.00448	0.02712	0.14045	0.48133	0.90649
$\tau_1 = 10$	FPR	0.00079	0.00026	7.5×10^{-5}	0.00001	5.0×10^{-6}	0.00000
	FNR	0.00034	0.00246	0.01535	0.08934	0.36382	0.83510

Results based on 1000 simulations

Table 7: $\mathbb{S}_{i,T}^{+} = \max_{1 \leq \ell \leq d} |S_{i,\ell,T}|$

		$N = 1000$	$N_1 = 500$	$T = 600$	$\tau = 12$		
		$\varphi = N^{-0.2}$	$\varphi = N^{-0.3}$	$\varphi = N^{-0.4}$	$\varphi = N^{-0.5}$	$\varphi = N^{-0.6}$	$\varphi = N^{-0.7}$
$\tau_1 = 6$	FPR	0.00044	0.00017	7.4×10^{-5}	2.8×10^{-5}	0.00001	2.0×10^{-6}
	FNR	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$\tau_1 = 8$	FPR	0.00054	0.00023	9.6×10^{-5}	4.2×10^{-5}	1.6×10^{-5}	8.0×10^{-6}
	FNR	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$\tau_1 = 10$	FPR	0.00080	0.00038	0.00018	0.00007	3.6×10^{-5}	2.0×10^{-5}
	FNR	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$\tau_1 = 12$	FPR	0.00127	0.00068	0.00031	0.00015	6.8×10^{-5}	3.0×10^{-5}
	FNR	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

Results based on 1000 simulations

Table 8: $\mathbb{S}^{+}_{i,T} = \sum_{\ell=1}^{d} \varpi_\ell \left|S_{i,\ell,T}\right|$

		$N = 1000$	$N_1 = 500$	$T = 600$	$\tau = 12$		
		$\varphi = N^{-0.2}$	$\varphi = N^{-0.3}$	$\varphi = N^{-0.4}$	$\varphi = N^{-0.5}$	$\varphi = N^{-0.6}$	$\varphi = N^{-0.7}$
$\tau_1 = 6$	FPR	0.00038	0.00015	0.00006	2.6×10^{-5}	0.00001	2.0×10^{-6}
	FNR	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$\tau_1 = 8$	FPR	0.00049	0.00020	8.2×10^{-5}	3.4×10^{-5}	1.4×10^{-5}	6.0×10^{-6}
	FNR	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$\tau_1 = 10$	FPR	0.00072	0.00033	0.00016	0.00006	3.2×10^{-5}	1.8×10^{-5}
	FNR	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$\tau_1 = 12$	FPR	0.00115	0.00062	0.00028	0.00014	6.0×10^{-5}	2.8×10^{-5}
	FNR	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

Results based on 1000 simulations

Looking across each row of each table given above, we see that, as we move from left to right, the FPR's are decreasing, whereas the FNR's are increasing. This is not surprising since, as we move from $\varphi = N^{-0.2}$ to $\varphi = N^{-0.7}$ for a given N, the size of the tuning parameter φ is becoming smaller which means that our specified threshold $\Phi^{-1}\left(1 - \frac{\varphi}{2N}\right)$ is becoming larger. Overall, our simulation results indicate that choosing $\varphi = N^{-\vartheta}$ with $\vartheta = 0.2$, 0.3, or 0.4 leads to very good performance, since with these choices of ϑ, neither FPR nor FNR exceeds 0.1 in any of the cases studied here. In fact, both are smaller than 0.05 in a vast majority of the cases. On the other hand, choosing $\vartheta = 0.6$ or 0.7 can lead to high values of FNR, as these values can set our threshold at such a high level that our procedure ends up having very little power. A particularly attractive choice of the tuning parameter is to take $\varphi = N^{-0.4}$, since as previously discussed in Section 4 above, this choice of the tuning parameter allows the rate condition in part (c) of Assumption 3.11* to be satisfied as long as $N_1 \to \infty$, without imposing further conditions on the rate at which N_1 grows.

Looking down the columns of each table, we see that FPR tends to increase as τ_1 increases, whereas FNR tends to decrease as τ_1 increases. To provide an explanation for this result, note first that the smaller is τ_1 relative to τ, the larger is τ_2 (since $\tau = \tau_1 + \tau_2$), and thus the larger is the number of observations that have been removed in constructing our self-normalized block sums. Intuitively, this can lead to better accommodation of the effects of dependence and better moderate deviation approximations under the null hypothesis, thus, resulting in a lower FPR. However, the removal of a larger number of observations can also lead to a reduction in the power of our procedure when the alternative hypothesis is correct, so that a negative consequence of having a smaller τ_1 relative to τ is that FNR will tend to be higher in this case. The opposite, of course, occurs when we try to specify a larger τ_1 relative to τ.

Our results also show that when the sample sizes are large enough such as the cases presented in Tables 7 and 8, where $T = 600$ and $N = 1000$, then both FPR and FNR are small for all of the cases that we consider. This is in accord with the results of our theoretical analysis, which shows that our variable selection procedure is completely consistent in the sense that both the probability of a false positive and the probability of a false negative approach zero, as the sample sizes go to infinity.

A final observation based on these Monte Carlo results is that there does not seem to be a great deal of difference in the performance of the statistic $\max_{1 \leq \ell \leq d} |S_{i,\ell,T}|$ vis-à-vis the statistic $\sum_{\ell=1}^{d} \varpi_\ell |S_{i,\ell,T}|$. Overall, the statistic $\max_{1 \leq \ell \leq d} |S_{i,\ell,T}|$ seems to be a bit better at controlling FNR, whereas the statistic $\sum_{\ell=1}^{d} \varpi_\ell |S_{i,\ell,T}|$ seems a bit better at controlling FPR.

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 propose a new variable selection procedure based on a self-normalized score statistic and provide asymptotic analyses showing that our procedure correctly identify the set of variables which load significantly on the underlying factors, with probability approaching one, as the sample sizes go to infinity. In addition, we show that estimating the factors using only those variables selected by our method allows the 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 then 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. Finally, we perform a small Monte Carlo study to examine the finite sample properties of our variable selection procedure. The results of this study yield insights on the range of choices for the tuning parameter for which our variable selection procedure exhibits good finite sample performance.

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

[8] 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.

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

[10] 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.

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

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

[13] 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.

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

[15] 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.

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

[17] 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.

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

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

[20] 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.

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

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

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

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

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

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

[27] 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.

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

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

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

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

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

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

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

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

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

[37] 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.

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

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

[40] 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.

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

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

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