Selecting the Relevant Variables for Factor Estimation in FAVAR Models, With An Application to Forecasting the Yield Curve*

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

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

July 28, 2023

Abstract

When specifying and estimating latent factor models, factor pervasiveness is often assumed, which requires that $\Gamma'\Gamma/N$ converges to a positive definite matrix, as $N \to \infty$, where Γ denotes the factor model loading matrix. We show that consistent factor estimation can be feasible, even under factor nonpervasiveness, if one prescreens the available variables. For this purpose, we introduce a variable selection procedure that, with probability approaching one, correctly distinguishes between relevant and irrelevant variables. This in turn enables consistent estimation of conditional mean functions of factor-augmented forecast equations, even in certain situations where the factor pervasiveness assumption is violated. Our procedure is designed to assess whether the usual strong factor assumption holds or not; and so does not rule out the possibility that the strong factor assumption actually holds for a given dataset. In addition, even if a particular dataset is such that most of the variables are actually relevant, so that inconsistent factor estimation does not occur, it may still be beneficial to use our procedure, since empirical researchers can prune out irrelevant variables and improve the finite sample performance of their factor estimator. Monte Carlo and empirical experiments are presented that indicate the empirical relevance of our procedure.

Keywords: Factor analysis, forecasting, variable selection.

* *Corresponding Author:* Norman R. Swanson, Department of Economics, 9500 Hamilton Street, Rutgers University, nswanson@econ.rutgers.edu.

John C. Chao, Department of Economics, 7343 Preinkert Drive, University of Maryland, jcchao@umd.edu. Kaiwen Qiu, Department of Economics, 9500 Hamilton Street, Rutgers University, kq60@econ.rutgers.edu.

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 methodologies. This has led to a profusion of novel research in areas ranging from the analysis of high dimensional and/or high frequency datasets to the development of new 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 the work due to Bai and Ng (2002), Stock and Watson (2002a,b), 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, employment of a factor structure allows the parsimonious representation of information embedded in a possibly high-dimensional vector of predictor variables.

Within this context, we note that a key assumption commonly used in the literature to obtain consistent factor estimation is the so-called factor pervasiveness assumption, which requires that $\Gamma'\Gamma/N$ converges to a positive definite matrix, as $N \to \infty$, where Γ denotes the loading matrix of the factor model. Since this assumption imposes certain conditions on how the variables in a given dataset load on the underlying latent factors, it is of interest to have statistical tools which allow researchers to check the empirical content of this assumption for the particular datasets they are using. Along these lines, our paper explores situations where the pervasiveness assumption may not hold because one is working with a dataset where some of the variables are irrelevant, in the sense that they do not load on the underlying latent factors. If a sufficient number of such irrelevant variables exist, inconsistency in factor estimation may result if one naively includes all available variables when estimating the underlying factors, without regard to whether they are relevant or not. See Chao, Liu, and Swanson (2023), for a particularly pathological example where an estimated factor, $\widehat{f}_t$, approaches 0 in probability, regardless of what the true value of f_t happens to be - a situation which can arise when the underlying factors are nonpervasive.[1] Not being able to obtain consistent estimates of the underlying factors will clearly cause problems for empirical researchers, such as when the objective is to estimate forecast functions that incorporate estimated factors. On the other hand, if one prescreens the variables and successfully prunes out the irrelevant ones, then consistent estimation can be achieved, under appropriate conditions. For this reason, a main contribution of this paper is to introduce a novel variable selection procedure which allows empirical researchers to correctly distinguish the relevant from the irrelevant variables prior to factor estimation, with probability approaching one. We study this problem within a factor-augmented VAR (FAVAR) framework - a setup which has the advantage that it allows time series forecasts to be made using information sets much richer than those used in traditional VAR models. While the present paper focuses on the development of a variable selection procedure and the analysis of its asymptotic properties; we show in Chao, Liu, and Swanson (2023) that the use of our methodology will allow the conditional mean function of a factor-augmented forecast equation to be consistently estimated 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. Overall, the results detailed in this paper can be viewed as adding to a nascent literature which considers the problem of factor estimation under various relaxations of the conventional factor pervasiveness assumption (see, for example, the interesting papers by Giglio, Xiu, and Zhang (2021), Freyaldenhoven (2021a,b), and Bai and Ng (2021)).

The variable selection procedure reported here is related to the well-known supervised prin-

[1] The example given in Chao, Liu, and Swanson (2023) is related to results that have been obtained in the statistics literature showing the inconsistency of sample eigenvectors as estimators of population eigenvectors in certain high dimensional situations (see, e.g. Paul (2007) and Johnstone and Paul (2018).

cipal components method proposed by Bair, Hastie, Paul, and Tibshirani (2006). Additionally, our procedure is related to 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 method 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 some economic and financial applications, where the distribution of the data is known to exhibit certain thick-tailed behavior. 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, thus, precluding complications associated with the introduction of dependence and non-normality. Giglio, Xiu, and Zhang (2021), on the other hand, make certain high-level assumptions which can accommodate some dependence both cross-sectionally and intertemporally, but the model that they consider is very different from the dynamic vector time series model studied in the sequel.[2]For all of the above reasons, the research reported in the sequel is meant to add to the suite of tools available to empirical researchers for variable selection in high dimensional data analysis.

It is also 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,

[2] Another interesting recent paper on factor estimation is Ahn and Bae (2022). This paper uses partial least squares instead of principal component methods to estimate a factor-based forecasting equation, and thus utilizes an approach that differs from the one taken in this paper. In addition, Ahn and Bae (2022) assume factor pervasiveness so that issues of variable selection, which are the main focus of this paper, do not arise in their paper.

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 even in situations where the number of irrelevant variables is very large and even if the underlying distribution does not satisfy the kind of sub-Gaussian tail behavior typically assumed in high-dimensional statistical 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 grow 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 the probabilities of both Type I and Type II errors go to zero in the limit. This property of complete consistency is important because if we try simply to control the probability of a Type I error at some predetermined non-zero 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.

In order to assess the practical usefulness of our method, we carry out a series of Monte Carlo experiments as well as an empirical application. In our Monte Carlo experiments, we show that the probability of a false positive and the probability of a false negative, when applying our methods using a number of data generating processes and admissible tuning parameter values, both approach zero, as expected, even for relatively small values of T and N. In our empirical application, we forecast the U.S. yield curve using a large macroeconomic dataset, with models constructed using our method, various other PCA, least absolute shrinkage operator (LASSO) and elastic net (EN), methods, a strawman autoregressive (AR), and the so-called dynamic Nelson-Siegel (DNS) models that is based on rational expectations. Interestingly, for 1-month ahead predictions of interest rates at 3-month and 1-year maturities, our method yields statistically superior forecasts relative to all other methods analyzed, and is approximately "tied" with PCA for longer maturities. Additionally, for 3-month ahead predictions, our method yields models with the lowest or second lowest mean-square forecast errors of any model, at all maturities. Finally, for our longest horizon forecasts of 1-year, the theoretically derived DNS model that includes no big data elements (i.e., only utilizes interest rates) yields superior forecasts. This suggests that markets may be relatively inefficient in the short-run, but are more efficient in the longer run.

The rest of the paper is organized as follows. In Section 2, 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 this procedure. Section 3 presents the results of a promising Monte Carlo study on the finite sample performance of our variable selection method. Section 4 presents the findings of our empirical application, and Section 5 offers some concluding remarks. Due to space considerations, proofs of the main theorems and all supporting lemmas as well as some additional technical details are provided in a not-for-publication online supplement (see Chao, Qiu, and Swanson (2023)).

Before proceeding, we first say a few words about some of the frequently used notation in this paper. Throughout, let $\lambda_{(j)}\left(A\right)$, $\lambda_{\max}\left(A\right)$, and $\lambda_{\min}\left(A\right)$ denote, respectively, the j^{th} largest eigenvalue, the maximal eigenvalue, and the minimal eigenvalue of a square matrix A. Similarly, let $\sigma_{(j)}\left(B\right)$, $\sigma_{\max}\left(B\right)$, and $\sigma_{\min}\left(B\right)$ 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 $\left\|a\right\|_2$ denote the usual Euclidean norm when applied to a (finite-dimensional) vector a. Also, for a matrix A, $\left\|A\right\|_2 \equiv \max\left\{\sqrt{\lambda\left(A'A\right)} : \lambda\left(A'A\right) \text{ is an eigenvalue of } A'A\right\}$ denotes the matrix spectral norm. For two sequences, $\{x_T\}$ and $\{y_T\}$, write $x_T \sim y_T$ if $x_T/y_T = O\left(1\right)$ and $y_T/x_T = O\left(1\right)$, as $T \rightarrow \infty$. Furthermore, let $|z|$ denote the absolute value or the modulus of the number z; let $\lfloor \cdot \rfloor$ denote the floor function, so 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, for a sequence of random variables $u_{i,t+m}, u_{i,t+m+1}, u_{i,t+m+2}, ...$; we let $\sigma\left(u_{i,t+m}, u_{i,t+m+1}, u_{i,t+m+2}, ...\right)$

denote the σ-field generated by this sequence of random variables.

2 Model, Assumptions, and Variable Selection

Consider the following p^{th}-order factor-augmented vector autoregression (FAVAR):

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

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

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

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, which will facilitate writing down assumptions and conditions used in the sequel. We thus briefly outline two alternative representations of the above model. It is easy to see that the system of equations given in (1) can be written as:

$$Y_{t+1} = \mu_Y + A_{YY}\underline{Y}_t + A_{YF}\underline{F}_t + \varepsilon^Y_{t+1}, \tag{2}$$

$$F_{t+1} = \mu_F + A_{FY}\underline{Y}_t + A_{FF}\underline{F}_t + \varepsilon^F_{t+1}, \tag{3}$$

where $\underset{d\times dp}{A_{YY}} = \begin{pmatrix} A_{YY,1} & A_{YY,2} & \cdots & A_{YY,p} \end{pmatrix}$, $\underset{d\times Kp}{A_{YF}} = \begin{pmatrix} A_{YF,1} & A_{YF,2} & \cdots & A_{YF,p} \end{pmatrix}$, $\underset{K\times dp}{A_{FY}} = \begin{pmatrix} A_{FY,1} & A_{FY,2} & \cdots & A_{FY,p} \end{pmatrix}$, $\underset{K\times Kp}{A_{FF}} = \begin{pmatrix} A_{FF,1} & A_{FF,2} & \cdots & A_{FF,p} \end{pmatrix}$, $\underset{dp\times 1}{\underline{Y}_t} = \begin{pmatrix} Y_t' & Y_{t-1}' & \cdots & Y_{t-p+1}' \end{pmatrix}'$, and $\underset{Kp\times 1}{\underline{F}_t} = \begin{pmatrix} F_t' & F_{t-1}' & \cdots & F_{t-p+1}' \end{pmatrix}'$. Another useful representation of the FAVAR model is the so-called companion form, wherein the p^{th}-order model given in expression (1) 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 $\underline{W}_t = (W_t', W_{t-1}', \cdots, W_{t-p+2}', W_{t-p+1}')'$, and where

$$\alpha = \begin{pmatrix} \mu \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, A = \begin{pmatrix} A_1 & A_2 & \cdots & A_{p-1} & A_p \\ I_{d+K} & 0 & \cdots & 0 & 0 \\ 0 & I_{d+K} & \ddots & \vdots & 0 \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & I_{d+K} & 0 \end{pmatrix}, \text{ and } E_t = \begin{pmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}. \tag{4}$$

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:

$$Z_t = \Gamma \underline{F}_t + u_t, \tag{5}$$

where $\underset{N\times 1}{Z_t} = (Z_{1t}, Z_{2t}, ..., Z_{Nt})'$. Assume, however, 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:

$$H = \{k \in \{1, ..., N\} : \gamma_k = 0\} \text{ and } H^c = \{k \in \{1, ..., N\} : \gamma_k \neq 0\}. \tag{6}$$

Now, let $\mathcal{P}$ be a permutation matrix which reorders the components of Z_t such that $\mathcal{P}Z_t = \begin{pmatrix} Z_t^{(1)\prime} & Z_t^{(2)\prime} \end{pmatrix}'$, where

$$\underset{N_1\times 1}{Z_t^{(1)}} = \Gamma_1 \underline{F}_t + u_t^{(1)} \tag{7}$$

$$\underset{N_2\times 1}{Z_t^{(2)}} = u_t^{(2)}. \tag{8}$$

The above representation suggests that the components of $Z_t^{(1)}$ can be interpreted as the relevant variables for the purpose of factor estimation, as the information that they supply will be helpful in estimating $\underline{F}_t$. On the other hand, the components of the subvector $Z_t^{(2)}$ are irrelevant variables (or 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 typically not have prior knowledge as to which variables are elements of $Z_t^{(1)}$ and which are elements of $Z_t^{(2)}$, it will be nice to have a variable selection procedure which will allow us to properly identify the components of $Z_t^{(1)}$ and to use only these variables when we try to estimate $\underline{F}_t$. On the other hand, if we unknowingly include too many components of $Z_t^{(2)}$ in the estimation process, then inconsistent factor estimation can arise. This is demonstrated in an example analyzed recently in Chao, Liu, and Swanson (2023) which considers a setting similar to the specification given in expressions (5)-(8) above, but for the case of a simple one-factor model. More precisely, an example is given in which, without variable pre-screening, the usual principal-component-based factor estimator $\widehat{f_t} \xrightarrow{p} 0$ regardless of the true value f_t under the additional rate condition that $N/\left(TN_1^{(1+\kappa)}\right) = c + o\left(N_1^{-1}\right)$, where c and κ are constants such that $0 < c < \infty$ and $0 < \kappa < 1$ and where N_1 is the number of relevant variables, N_2 is the number of irrelevant variables, and $N = N_1 + N_2$. This example shows the kind of severe inconsistency in factor estimation that could result if the commonly assumed condition of factor pervasiveness (which essentially requires that $N_1 \sim N$) does not hold.[3]

It should be noted that, in an important recent paper, Bai and Ng (2021) provide results which show that factors can still be estimated consistently in certain situations where factor loadings are weaker than implied by the conventional pervasiveness assumption; although, as might be expected, in such cases the rate of convergence of the factor estimator is slower and additional assumptions are needed. To understand the relationship between their results and

[3] The reason why we refer to the result given in Chao, Liu, and Swanson (2023) as a severe form of inconsistency in factor estimation is because inconsistency of this type will preclude the consistent estimation of the conditional mean function of a factor-augmented forecast equation. This is different from the case where the factors may be estimated consistently up to a non-zero scalar multiplication or, more generally, up to an invertible matrix transformation. In the latter case, consistent estimation of the conditional mean function of a factor-augmented forecast equation can still be attained.

our setup, note that a key condition for the consistency result given in their paper, when expressed in terms of our setup, is the assumption that $N/(TN_1) \to 0$. When violation of the factor pervasiveness condition is more severe than that characterized by this rate condition (i.e., if $N/(TN_1) \to c_1$, for some positive constant c_1 or if $N/(TN_1) \to \infty$), then factors will be estimated inconsistently unless there is some method which can correctly identify the relevant variables, and only these variables are used to estimate the factors. Indeed, in Thoerem 4.1 of Chao, Liu, and Swanson (2023), we add to the results given in Bai and Ng (2021) by giving a result which shows that if one pre-screens variables using the variable selection method proposed below, then consistent factor estimation can be achieved, even if the rate condition that $N/(TN_1) \to 0$ is not satisfied. In general, knowledge about the severity with which the conventional factor pervasiveness assumption may be violated must ultimately be gathered on a case-by-case basis, and depends on the dataset used for a particular study. Along these lines, various authors have already documented cases where the empirical evidence shows that the underlying factors are quite weak, suggesting that there may be rather severe violation of the assumption of factor pervasiveness. For example, see Jagannathan and Wang (1998), Harding (2008), Kleibergen (2009), Onatski (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 weakness in loadings is not uniform across all variables, but rather is due to the fact that only a fraction of the Z_{it} variables loads significantly on the underlying factors. Furthermore, even if the empirical situation of interest is one where, strictly speaking, the condition $N/(TN_1) \to 0$ 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 $N/(TN_1) \to 0$ 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 may add unwanted noise to the estimation process. For these reasons, we believe that there is a need to develop methods which 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. In summary, our paper aims to build on the results developed by Bai and Ng (2021) and others by introducing additional tools for situations where factor estimator properties may be impacted by failure of the conventional pervasiveness assumption.

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

Assumption 2-1: Suppose that: $\det\left\{I_{(d+K)} - A_1 z - \cdots - A_p z^p\right\} = 0,$ implies that $|z| > 1$.

Assumption 2-2: Let ε_t satisfy the following set of conditions: (a) $\{\varepsilon_t\}$ is an independent sequence of random vectors with $E\left[\varepsilon_t\right] = 0\ \forall t$; (b) there exists a positive constant C such that $\sup_t E\left\|\varepsilon_t\right\|_2^6 \leq C < \infty$; and (c) ε_t admits a density g_{ε_t} such that, for some positive constant $M < \infty$, $\sup_t \displaystyle\int \left|g_{\varepsilon_t}\left(v - u\right) - g_{\varepsilon_t}\left(v\right)\right| dv \leq M \left\|u\right\|$, whenever $\left\|u\right\| \leq \overline{\kappa}$ for some constant $\overline{\kappa} > 0$.

Assumption 2-3: Let $u_{i,t}$ be the i^{th} element of the error vector u_t in expression (5), and we assume that it satisfies the following conditions: (a) $E\left[u_{i,t}\right] = 0$ for all i and t; (b) there exists a positive constant $\overline{C}$ such that $\sup_{i,t} E\left|u_{i,t}\right|^7 \leq \overline{C} < \infty$, and there exists a constant $\underline{C} > 0$ such that $\inf_{i,t} E\left[u_{i,t}^2\right] \geq \underline{C}$; and (c) define $\mathcal{F}_{i,-\infty}^{t} = \sigma\left(...., u_{i,t-2}, u_{i,t-1}, u_t\right)$, $\mathcal{F}_{i,t+m}^{\infty} = \sigma\left(u_{i,t+m}, u_{i,t+m+1}, u_{i,t+m+2}, ...\right)$, and $\beta_i\left(m\right) = \sup_t E\left[\sup\left\{\left|P\left(B|\mathcal{F}_{i,-\infty}^{t}\right) - P\left(B\right)\right| : B \in \mathcal{F}_{i,t+m}^{\infty}\right\}\right]$. Assume $\exists$ constants $a_1 > 0$ and $a_2 > 0$ such that $\beta_i\left(m\right) \leq a_1 \exp\left\{-a_2 m\right\}$, for all i.

Assumption 2-4: ε_t and $u_{i,s}$ are independent, for all i, t, and s.

Assumption 2-5: There exists a positive constant $\overline{C}$, such that $\sup_{i \in H^c} \left\|\gamma_i\right\|_2 \leq \overline{C} < \infty$ and $\left\|\mu\right\|_2 \leq \overline{C} < \infty$, where $\mu = (\mu_Y', \mu_F')'$.

Assumption 2-6: Let A be as defined in expression (4) above, and let the modulus of 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:

$$\sigma_{\min}\left(I_{(d+K)p} - A\right) \geq \underline{C}\,\overline{\phi}_{\min} \tag{9}$$

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

$$\sigma_{\max}\left(A^j\right) \leq \overline{C} \max\left\{\left|\lambda_{\max}\left(A^j\right)\right|, \left|\lambda_{\min}\left(A^j\right)\right|\right\}. \tag{10}$$

Remark 2.1: **(a)** Note that Assumption 2-1 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 2-1, 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 (4). It follows that Assumption 2-1 is equivalent to the condition that $\det\left\{I_{(d+K)} - Az\right\} = 0$ implies that $|z| > 1$. In addition, Assumption 2-1 is also, of course, equivalent to the assumption that all eigenvalues of A have modulus less than 1. **(c)** Assumption 2-6 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 fact, it is easy to show that Assumption 2-6 holds automatically if the matrix A is diagonalizable, even if it is not symmetric. Assumptions 2-6, on the other hand, takes into account other situations where expressions (9) and (10) are valid even though the matrix A is not diagonalizable. **(d)** Note that Assumptions 2-1, 2-2, and 2-6 together imply that the process $\{W_t\}$ generated by the FAVAR model given in expression (1) 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
$\mathcal{A}_{-\infty}^t = \sigma\left(..., W_{t-2}, W_{t-1}, W_t\right)$ and $\mathcal{A}_{t+m}^{\infty} = \sigma\left(W_{t+m}, W_{t+m+1}, W_{t+m+2}, ...\right)$.[4] Note, in addition,

[4] This can be shown by applying Theorem 2.1 of Pham and Tran (1985). This result is given as Lemma OA-11 in Chao, Qiu, and Swanson (2023).

that Assumption 2-2 (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 2-2(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 2-2 and 2-3 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 of sub-exponential tail behavior, which has been made in various papers studying high-dimensional factor models and/or high-dimensional covariance matrices, without employing statistics that are self-normalized.[5]

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:

$$\max_{1\leq\ell\leq d}|S_{i,\ell,T}| = \max_{1\leq\ell\leq d}\left|\frac{\overline{S}_{i,\ell,T}}{\sqrt{\overline{V}_{i,\ell,T}}}\right|, \text{ where} \tag{11}$$

$$\overline{S}_{i,\ell,T} = \sum_{r=1}^{q}\sum_{t=(r-1)\tau+p}^{(r-1)\tau+\tau_1+p-1} Z_{it}y_{\ell,t+1} \text{ and } \overline{V}_{i,\ell,T} = \sum_{r=1}^{q}\left[\sum_{t=(r-1)\tau+p}^{(r-1)\tau+\tau_1+p-1} Z_{it}y_{\ell,t+1}\right]^2. \tag{12}$$

[5] See, for example, Bickel and Levina (2008) and Fan, Liao, and Mincheva (2011, 2013).

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 (11) 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: $\sum_{\ell=1}^{d} \varpi_\ell \left| S_{i,\ell,T} \right| = \sum_{\ell=1}^{d} \varpi_\ell \left| \frac{\overline{S}_{i,\ell,T}}{\sqrt{\overline{V}_{i,\ell,T}}} \right|$, where $\overline{S}_{i,\ell,T}$ and $\overline{V}_{i,\ell,T}$ are as defined in (12) 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 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:

$$\overline{S}_{i,\ell,T} = \sum_{t=p}^{\tau_1+p-1} Z_{it} y_{\ell,t+1} + \sum_{t=\tau+p}^{\tau+\tau_1+p-1} Z_{it} y_{\ell,t+1} + \sum_{t=2\tau+p}^{2\tau+\tau_1+p-1} Z_{it} y_{\ell,t+1} + \cdots + \sum_{t=(q-1)\tau+p}^{(q-1)\tau+\tau_1+p-1} Z_{it} y_{\ell,t+1} \quad (13)$$

Comparing the first term and the second term on the right-hand side of expression (13), we see that the observations $Z_{it} y_{\ell,t+1}$, for $t = \tau_1 + p, ..., \tau + p - 1$, 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 is in some sense more akin to a model selection problem rather than a multiple hypothesis testing problem. In order to consistently estimate the factors (at least 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, $T \to \infty$.[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.

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 $\sum_{\ell=1}^{d} \varpi_\ell |S_{i,\ell,T}|$.[7] Our variable selection procedure is based on the decision rule:

$$i \in \begin{cases} \widehat{H}^c & \text{if } \mathbb{S}^+_{i,T} \geq \Phi^{-1}\left(1 - \frac{\varphi}{2N}\right) \\ \widehat{H} & \text{if } \mathbb{S}^+_{i,T} < \Phi^{-1}\left(1 - \frac{\varphi}{2N}\right) \end{cases}, \tag{14}$$

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 Assumption 2-10 below.

Remark 2.2: 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 a slight modifi-

[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.

[7] It should be noted that the denominator of the statistic $S_{i,\ell,T} = \overline{S}_{i,\ell,T}/\sqrt{\overline{V}_{i,\ell,T}}$ does not correspond to the use of an HAR standard error constructed using the fixed b (or fixed smoothing) approach pioneered by Kiefer and Vogelsang (2002), even in the case without any truncation. Hence, our statistic differs from the usual Studentized statistic that is normalized by an HAR estimator. This can be shown by straightforward calculations for the case of the Bartlett kernel, for example. For interesting discussions of different approaches to self-normalization in the statistics and probability literature, refer to Z. Zhou and X. Shao (2013), X. Chen, Q-M. Shao, W.B. Wu, and L. Xu (2016), and the references cited therein.

cation of the arguments given in the proof of Lemma A2[8], we can show that, as $T \to \infty$, $P\left(|S_{i,\ell,T}| \geq z\right) = 2\left[1-\Phi\left(z\right)\right]\left(1+o\left(1\right)\right)$, which holds for all i and ℓ and for all z such that: $0 \leq z \leq c_0 \min\left\{T^{(1-\alpha_1)/6}/L\left(T\right), T^{\alpha_2/2}\right\}$, where $L\left(T\right)$ denotes a slowly varying function such that $L\left(T\right) \to \infty$, as $T \to \infty$. In view of the above expression, 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. 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 significance level α. Then, the above expression suggests that a natural way to do this is to set $z = \Phi^{-1}\left(1-\alpha/2\right)$. This is because, given that the quantile function $\Phi^{-1}\left(\cdot\right)$ is, by definition, the inverse function of the cdf $\Phi\left(\cdot\right)$, we have that:

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

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 some 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

[8] The statement and proof of Lemma A2 are provided in the online supplement to this paper.

the factors consistently, even up to an invertible matrix transformation. We show in Theorem 1 below that these objectives can all be accomplished using the threshold function specified in expression (14), since a threshold function of this form makes it easy for us to properly control the probability of a false positive in large samples. The threshold function used here is reminiscent of the one employed in Belloni, Chernozhukov, and Hansen (2014). The latter paper focuses on developing a variable screening methodology for a partially linear treatment effects model. In that paper, a threshold function that is similar to ours is used to set the penalty level for a lasso-based procedure for selecting the terms in a series expansion of the nonlinear component of their model under conditions of sparsity. In spite of the similarity in the form of the threshold function used, the nature of the variable selection problem studied in Belloni, Chernozhukov, and Hansen (2014) is quite different from that investigated in this paper. In particular, these authors do not require their variable selection procedure to be completely consistent, nor do they provide a result showing that the probability of both Type I and Type II error vanishes asymptotically as sample sizes approach infinity. They also stress that perfect variable selection is not needed in the type of regression settings considered in their paper if the goal is to approximate the nonlinear functions in their model sufficiently well so that the post-selection estimators of the treatment effect parameter will have good asymptotic properties. Here, we instead argue that having a variable selection procedure that is completely consistent is quite useful given our objective of ensuring that good factor estimates can be obtained in a high-dimensional latent factor model. This is because, as noted earlier, if the probability of a Type I error is only controlled at some fixed nonzero level asymptotically, then consistent factor estimation may not be possible. In addition, the precision with which the latent factors are estimated will be reduced if we have a variable selection procedure where the probability of a Type II error does not go to zero. As a result of these differences in setup and objectives, the conditions that we specify for setting the tuning parameter φ will also be quite different from that in Belloni, Chernozhukov, and Hansen (2014).

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. $P\left(i \in \widehat{H}^c | i \in H\right)$, and the probability of a false negative, i.e., $P\left(i \in \widehat{H} | i \in H^c\right)$, approach zero as $N_1, N_2, T \to \infty$. To show this result, we must first state a number of additional assumptions.

Assumption 2-7: There exists a positive constant $\underline{c}$ such that for all $\tau \geq 1$ and $\tau_1 \geq 1$:

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

where, as defined earlier, $\tau_1 = \lfloor T_0^{\alpha_1} \rfloor$, $\tau_2 = \lfloor T_0^{\alpha_2} \rfloor$ for $1 > \alpha_1 \geq \alpha_2 > 0$ and $q = \left\lfloor \frac{T_0}{\tau_1 + \tau_2} \right\rfloor$, and $T_0 = T - p + 1$.

Assumption 2-8: 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 2-9: Suppose that, as N_1, N_2, and $T \to \infty$, the following rate conditions hold:

(a) $\sqrt{\ln N} / \min\left\{T^{(1-\alpha_1)/6}, T^{\alpha_2/2}\right\} \to 0$, where $1 > \alpha_1 \geq \alpha_2 > 0$ and $N = N_1 + N_2$.

(b) $N_1 / T^{3\alpha_1} \to 0$ where α_1 is as defined in part (a) above.

Assumption 2-10: 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 $\varphi \geq 1/N^a$, for all N_1, N_2 sufficiently large.

Remark 2.3: **(a)** Assumption 2-8 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 (2) implies that the ℓ^{th} component of Y_{t+1} will have the representation $y_{\ell,t+1}=\mu_{Y,\ell}+\underline{Y}_t'\alpha_{YY,\ell}+\underline{F}_t'\alpha_{YF,\ell}+\varepsilon_{\ell,t+1}^Y=\varepsilon_{\ell,t+1}^Y$, 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 a dependent process in this case. **(b)** Bai and Ng (2008) address the important issue of pre-selecting variables Z_{it} based on their predictability for Y_{t+1}. Our selection approach is related to theirs. However, it is worth stressing that for the FAVAR model considered here, whether Z_{it} helps predict 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 focuses on the first issue but not the second. Thus, we focus on obtaining factor estimates with desirable asymptotic properties before trying to assess which factor(s) 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, if we exclude a significant number of variables (at the variable selection stage) that load strongly on at least some of the factors, this can result in $\underline{F}_t$ 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 would like to use every data point Z_{it}, for which $\gamma_i' \neq 0$. Furthermore, 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 1: *Let $H = \{k \in \{1, ..., N\} : \gamma_k = 0\}$. Suppose that Assumptions 2-1, 2-2, 2-3, 2-4, 2-5, 2-6, 2-7, 2-9 (a) and 2-10 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:*

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

where $N = N_1 + N_2$.

(b)

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

Theorem 2: *Let $H^c=\{k\in\{1,....,N\}:\gamma_k\neq 0\}$. Suppose that Assumptions 2-1, 2-2, 2-3, 2-5, 2-6, 2-8, 2-9, and 2-10 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:*

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

(b)

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

Remark 2.4: (a) Theorem 1 shows that, for both of our statistics, the probability of a false positive approaches zero uniformly over all $i\in H$ as $N_1,N_2,T\rightarrow\infty$. The results of Theorem 2 further imply that, for both of our statistics, the probability of a false negative also approaches zero, uniformly over all $i\in H^c$ as $N_1,N_2,T\rightarrow\infty$. Together, these two theorems show that our procedure is (completely) consistent in the sense that the probability of committing a misclassification error vanishes as $N_1,N_2,T\rightarrow\infty$. **(b)** Note that our variable selection procedure also delivers a consistent estimate of N_1 (i.e., $\widehat{N}_1$); this is shown in Lemma D-15 part (a) of Chao, Liu, and Swanson (2023), where we establish that $\widehat{N}_1/N_1\xrightarrow{p}1$. The estimator $\widehat{N}_1$ is useful to applied researchers implementing the methodology developed in this paper, and also to empiricists interested in assessing the rate condition for consistent factor estimation, given in Assumption A4 of Bai and Ng (2021). This is another way in which the methods developed in this paper built upon the work of Bai and Ng (2021). **(c)** In addition, note that knowledge of the number of factors is not needed to implement our variable selection procedure. In the case where the number of factors needs to be determined empirically, an applied researcher can first use our procedure to 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.

3 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 trivariate FAVAR(1) process:

$$W_t = \mu + AW_{t-1} + \varepsilon_t, \tag{15}$$

$$Z_t = \gamma F_t + u_t, \tag{16}$$

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 below. For the error process in equation (15), 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 (16) is allowed to exhibit both temporal and cross-sectional dependence and also conditional heteroskedasticity. More specifically, we let $u_{it} = 0.8u_{it-1} + \zeta_{it}$, 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 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 consider specifications where $\varphi = (\ln\ln N)^{-\vartheta}$ for $\vartheta = 0.1, 0.5, 1$ and also $\varphi = N^{-\vartheta}$ for $\vartheta = 0.2, 0.4, 0.6$.[9] We also attempt to shed light on the effects of using 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. Due to space considerations, we only report Monte Carlo results for the statistic $\sum_{\ell=1}^{d} \varpi_\ell \left|S_{i,\ell,T}\right|$. Simulation results for the statistic $\max_{1\leq\ell\leq d}\left|S_{i,\ell,T}\right|$ have also been obtained by the authors and are qualitatively similar to the results reported here for $\sum_{\ell=1}^{d} \varpi_\ell \left|S_{i,\ell,T}\right|$. The results for $\max_{1\leq\ell\leq d}\left|S_{i,\ell,T}\right|$ are available from the authors upon request. In addition, since $d = 2$ in our Monte Carlo setup, we set $\varpi_1 = \varpi_2 = 1/2$. Results are gathered in in Table 1, where FPR denotes 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. FNR denotes 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.

Looking across each row of the table, note that FPRs decrease when moving from left to right, whereas FNRs increase. This is not surprising, because moving from $\varphi = (\ln\ln N)^{-0.1}$ to $\varphi = N^{-0.6}$ for a given N results in smaller values of the tuning parameter φ, and the specified threshold $\Phi^{-1}\left(1 - \frac{\varphi}{2N}\right)$ thus becomes larger. Overall, these results indicate that choosing φ in the range between $(\ln\ln N)^{-0.1}$ and $N^{-0.4}$ leads to very good performance, since within this range, 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. In contrast, choosing $\varphi = N^{-0.6}$ can lead to high FNRs, as such a choice of φ can set our threshold at such a high level that our procedure ends up having very little power.

[9] We have also obtained simulation results for the cases where $\varphi = (\ln N)^{-\vartheta}$ for $\vartheta = 0.1, 0.5, 1$ and where $\varphi = N^{-\vartheta}$ for $\vartheta = 0.3, 0.5, 0.7$. The results obtained for these cases are qualitatively similar to the results reported in this paper. Hence, due to space considerations, we do not report these results here, but they are available from the authors upon request.

Looking down the columns of the table, note that FPR tends to increase as τ_1 increases, whereas FNR tends to decrease as τ_1 increases. As 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 removed when constructing the 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, resulting in a lower FPR. However, removal of a larger number of observations can also lead to a reduction in power, 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 the last panel of the table, 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.

4 Empirical Application

In this section, we construct forecasts of 3-month, 1-year, 3-year, 5-year, and 10-year maturity interest rates, at h =1-month, 3-month, and 12-month ahead horizons. The interest rates that we analyze are contained in the so-called U.S. Treasury yield curve dataset, which is discussed in Gurkaynak, Sack, and Wright (GSW: 2007). In order to assess the empirical usefulness of our new variable selection method, we construct our forecasts using a variety of "big-data" as well as "small-data" models. Our big data models utilize the GSW dataset in conjunction with the so-called FRED-MD real-time macroeconomic dataset, which is available at the St. Louis federal reserve bank website, and includes a broad array of 130 economic variables (see GSW

(2007) for further discussion of this dataset). All of these time series variables are "real-time" in the sense that for each calendar date there may be multiple observations, corresponding to the "first-release" of an observation for that calendar date, a "second release", often a month or more later, for that same calendar date, and so on. Many macroeconomic variables are subject to these sorts of revisions, due to the data collection methodology used by the relevant government reporting agencies. These different releases are called vintages. In all of our experiments that utilize such real-time data, we ensure that only vintages truly available prior to the construction of each of our time series forecasts are actually used in model estimation and forecast construction. Needless to say, this requires us to re-estimate all models at each point in time, prior to the construction of each new forecast. In our analysis, we use rolling (fixed) windows of 120 months for all estimations. The sample period of our interest rate dataset, as well as our real-time macroeconomic dataset is August 1988 - December 2021. Our forecasting sample period is April 2001 - December 2021.

The forecasting models that we evaluate are summarized in Table 2, and all have the following function form:

$$y_{t+h}(\tau) = \alpha + \sum_{i=1}^{p} \beta_i y_{t-i+1}(\tau) + \gamma_1' W_t + \gamma_2' F_t + \epsilon_{t+h}, \tag{17}$$

where $y_{t+h}(\tau)$, is an annual yield interest rate, τ denotes the maturity of the bond (bill) being forecast, F_t is an $r-$dimensional vector of estimated factors, W_t includes additional variables from our real-time dataset, ϵ_{t+h} is a stochastic disturbance term, and p is the number of lags, selected using the Schwarz information criterion (SIC). All variables were differenced to stationarity, using stationarity test results reported in the documentation that accompanies the FRED-MD dataset. Additionally, we considered $r = \{1, 2, 3\}$. In all experiments, setting $r = 1$ yieled more precise predictions. Thus, results in the sequel are for the case where $r = 1$. In this setup, we consider a number of models with $\gamma_2 = 0$. These include: the simple AR(SIC) benchmark, where we additionally impose that $\gamma_1 = 0$; a model called AR+LASSO, where we

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

		$N = 100$	$N_1 = 50$	$T = 100$	$\tau = 5$		
		$\varphi = (\ln \ln N)^{-0.1}$	$\varphi = (\ln \ln N)^{-0.5}$	$\varphi = (\ln \ln N)^{-1}$	$\varphi = N^{-0.2}$	$\varphi = N^{-0.4}$	$\varphi = N^{-0.6}$
$\tau_1 = 2$	FPR	0.03916	0.03350	0.02678	0.01460	0.00382	0.00076
	FNR	0.00046	0.00068	0.00104	0.00284	0.01674	0.09412
$\tau_1 = 3$	FPR	0.04544	0.03902	0.03110	0.01810	0.00526	0.00092
	FNR	0.00022	0.00032	0.00052	0.00172	0.01100	0.06942
$\tau_1 = 4$	FPR	0.05408	0.04650	0.03756	0.02224	0.00702	0.00162
	FNR	0.00016	0.00024	0.00034	0.00118	0.00828	0.05194
$\tau_1 = 5$	FPR	0.06332	0.05462	0.04558	0.02796	0.00924	0.00232
	FNR	0.00014	0.00018	0.00034	0.00084	0.00574	0.03948
		$N = 200$	$N_1 = 100$	$T = 100$	$\tau = 5$		
$\tau_1 = 2$	FPR	0.01913	0.01470	0.01068	0.00486	0.00064	0.00002
	FNR	0.00206	0.00282	0.00449	0.01415	0.09966	0.48356
$\tau_1 = 3$	FPR	0.02341	0.01842	0.01365	0.00657	0.00098	0.00005
	FNR	0.00143	0.00190	0.00315	0.00921	0.07372	0.40894
$\tau_1 = 4$	FPR	0.02869	0.02306	0.01733	0.00841	0.00133	0.00004
	FNR	0.00111	0.00145	0.00224	0.00661	0.05564	0.34279
$\tau_1 = 5$	FPR	0.03506	0.02903	0.02194	0.01124	0.00213	0.00017
	FNR	0.00086	0.00112	0.00172	0.00477	0.04258	0.28620
		$N = 400$	$N_1 = 200$	$T = 200$	$\tau = 10$		
$\tau_1 = 5$	FPR	0.00214	0.00148	0.00090	0.00030	$2.5{\times}10^{-5}$	0.00000
	FNR	$7.5{\times}10^{-5}$	0.00016	0.00040	0.00231	0.06894	0.67266
$\tau_1 = 6$	FPR	0.00249	0.00166	0.00104	0.00034	0.00002	0.00000
	FNR	0.00004	0.00009	0.00025	0.00148	0.05058	0.60968
$\tau_1 = 8$	FPR	0.00337	0.00235	0.00142	0.00046	0.00004	0.00000
	FNR	0.00001	0.00002	0.00008	0.00068	0.02712	0.48133
$\tau_1 = 10$	FPR	0.00484	0.00350	0.00220	0.00079	$7.5{\times}10^{-5}$	$5.0{\times}10^{-6}$
	FNR	0.00001	0.00001	0.00002	0.00034	0.01535	0.36382
		$N = 1000$	$N_1 = 500$	$T = 600$	$\tau = 12$		
$\tau_1 = 6$	FPR	0.00155	0.00121	0.00086	0.00038	0.00006	0.00001
	FNR	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$\tau_1 = 8$	FPR	0.00201	0.00153	0.00106	0.00049	$8.2{\times}10^{-5}$	$1.4{\times}10^{-5}$
	FNR	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$\tau_1 = 10$	FPR	0.00274	0.00216	0.00155	0.00072	0.00016	$3.2{\times}10^{-5}$
	FNR	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$\tau_1 = 12$	FPR	0.00421	0.00332	0.00242	0.00115	0.00028	$6.0{\times}10^{-5}$
	FNR	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

Notes: False positive and negative rates are reported for various values of N, N_1, and T. Results are based on 1000 simulations. See Section 3 for complete details.

choose the elements of W_t using the LASSO; a model called AR+EN, where we choose the elements of W_t using the EN; and a model called AR+CS, where we choose the elements of W_t using the the "CS" method discussed in Section 2.[10] We also consider a number of models with $\gamma_1 = 0$. These include: AR+PCA, where factors are estimated using PCA applied to our entire real-time dataset; AR+LASSO+PCA, where factors are estimated using PCA applied to a subset of our real-time dataset that is selected using the LASSO; AR+EN+PCA, where factors are estimated using PCA applied to a subset of our real-time dataset that is selected using the EN; and AR+CS+PCA, where factors are estimated using PCA applied to a subset of our real-time dataset that is selected using the CS.[11]

In addition to our "small-data" AR(SIC) model, we include another parsimonious model that utilizes only interest rates when specifying forecasting models. This is the dynamic Nelson-Siegel (DNS) model that is widely used in industry and government for forecasting interest rates, as discussed in Diebold, Rudebusch and Aruoba (2006), and Swanson and Xiong (2018). The Nelson and Siegel (1987) model specifies the relationship between spot interest rates and instantaneous forward rates by applying rational expectation theory. Specifically, the yield of a bond with maturity m can be expressed by the averaged forward rates: $y_t(\tau) = \frac{1}{m}\int_0^m f_t(m)d\tau$, where $y_t(\tau)$ is the yield at time t for a bond with maturity τ, and $f_t(m)$ denotes the instantaneous forward rate at time t for a bond with time-to-maturity m. Based on this setup, the dynamic Nelson-Siegel model approximates the term structure of interest rates using a parsimonious three-factor model:

$$y_t(\tau) = \beta_{0,t} + \beta_{1,t}[\frac{1-exp(-\frac{\tau}{\theta_t})}{\frac{\tau}{\theta_t}}] + \beta_{2,t}[\frac{1-exp(-\frac{\tau}{\theta_t})}{\frac{\tau}{\theta_t}} - exp(-\frac{\tau}{\theta_t})]. \tag{18}$$

[10] Note that for this model we do not use the variables selected using the CS method to estimate factors. Rather, we simply use the $S^+_{i,T} = \sum_{\ell=1}^{d} \varpi_\ell \left|S_{i,\ell,T}\right|$ statistics to select a subset of variables for inclusion in the forecasting model. Also, we set τ, τ_1, and φ equal to 5, 4, and $(\ln\ln N)^{-0.1}$ in the sequel.

[11] For the CS method, we again use the $S^+_{i,T} = \sum_{\ell=1}^{d} \varpi_\ell \left|S_{i,\ell,T}\right|$ statistic, and the tuning parameters for the LASSO and EN were estimated anew prior to the construction of each forecast, using 10-fold cross validation.

In this model, $\beta_{0,t}$, $\beta_{1,t}$, and $\beta_{2,t}$ are latent factors representing the level, slope, and curvature of the yield curve, respectively, and $\lambda = \frac{1}{\theta_t}$ is a decay parameter. For a complete discussion of this model, refer to Swanson and Xiong (2018), where the latent factors are modeled using a vector autoregression, and are estimated either solely using the GSW dataset (our "small data" DNS model), or by including extra variables and/or latent factors constructed via the FRED-MD dataset along with LASSO, EN, or CS methods (see Table 2 for details).

Our empirical findings are gathered in Table 3. Inspection of the mean square forecast errors (MSFEs) in this table reveal that many of our "big data" models are "MSFE-best" (i.e. exhibit lower point MSFEs), for shorter horizon forecasts of $h = 1$ and 3.[12] Interestingly, this is not the case for $h = 12$. For our longest horizon forecasts our "small data" DNS model "wins" at almost all maturities. Thus, big data appears to be particularly useful for shorter horizon forecasts. With regard to the performance of the CS method, note that for 1-month ahead predictions of interest rates at 3-month and 1-year maturities, our method yields statistically superior forecasts relative to all other methods analyzed, and is approximately "tied" with PCA for longer maturities. For 3-month ahead predictions, our method yields models with the lowest or second lowest mean-square forecast errors of any model, at all maturities. In this sense, our first evidence suggests that the CS method is useful when selecting variables prior to constructing factors for use in forecasting models.

5 Concluding Remarks

In this paper, we propose a new variable selection procedure based on two alternative self-normalized score statistics and provide asymptotic analyses showing that our procedure, based on either of these statistics, correctly identify the set of variables which load significantly on the underlying factors, with probability approaching one as the sample sizes go to infinity. Our

[12] Many "big data" models also deliver significantly superior predictions to the AR(SIC) model (as well as the DNS model), based on application of the Diebold-Mariano (1995) predictive accuracy test, as denoted by entries that are starred.

research is motivated by the observation that inconsistency in factor estimation could result in high dimensional settings when the conventional assumption of factor pervasiveness does not hold. Hence, in such settings, it is particularly important to pre-screen the variables in terms of their association with the underlying factors prior to estimation. We conduct a small Monte Carlo study which yields encouraging evidence about the finite sample properties of our variable selection procedure. It is also worth noting that in a companion paper (Chao, Liu, and Swanson, 2023), we prove that consistent estimation of factors (up to an invertible matrix transformation) can be achieved by estimating factors using only those variables selected by our method, and this is so even in certain situations where the standard pervasiveness assumption does not hold. In addition, in the same paper, we further show that by plugging factors estimated in such a manner into the factor-augmented forecasting equation implied by the FAVAR model, the conditional mean function of the forecasting equation can be consistently estimated, even for the case of multi-step ahead forecasts. Finally, we present a series Monte Carlo and empirical experiments underscoring the potential usefulness of our procedure in empirical settings. In sum, the collective body of results, including theoretical, Monte Carlo, and empirical discussed in this paper indicates that the proposed variable selection methodology can be useful to empirical researchers as they engage in the important tasks of factor estimation and the construction of point forecasts based on factor-augmented forecasting equations.

Supplementary Materials: Proofs of the main theorems of this paper as well as some supporting lemmas are reported in an Online Supplement (see Chao, Qiu, and Swanson (2023) at http://econweb.umd.edu/~chao/Research/research_files/variable_selection_factor_models -online-supplement-7-28-2023-jasa.pdf and at https://econweb.rutgers.edu/nswanson/papers.htm).

Acknowledgements: The authors are grateful to Matteo Barigozzi, Mehmet Caner, Bin Chen, Rong Chen, Simon Freyaldenhoven, Yuan Liao, Esther Ruiz, Minchul Shin, Jim Stock, Timothy Vogelsang, Endong Wang, Xiye Yang, Peter Zadrozny, Bo Zhou and seminar participants at the University of Glasgow, the University of California Riverside, the Federal Reserve Bank

Table 2: Models Used in Prediction Experiments

Model	Description
AR(SIC)	Autoregressive model with lag(s) selected by the Schwarz information criterion.
AR+LASSO	AR(SIC) model augmented to include a subset of variables selected from our real-time dataset using the LASSO.
AR+EN	AR(SIC) model augmented to include a subset of variables selected from our real-time dataset using the Elastic Net.
AR+CS	AR(SIC) model augmented to include a subset of variables selected from our real-time dataset using the "Chao-Swanson" (CS) variable selection method discussed in Section 2.
AR+PCA	AR(SIC) model augmented with PCA type factors constructed using our entire real-time dataset.
AR+LASSO+PCA	AR(SIC) model augmented with PCA type factors constructed using a subset of variables from our real-time dataset, with the subset selected using the LASSO.
AR+EN+PCA	AR(SIC) model augmented with PCA type factors constructed using a subset of variables from our real-time dataset, with the subset selected using the EN.
AR+CS+PCA	AR(SIC) model augmented with PCA type factors constructed using a subset of variables from our real-time dataset, with the subset selected using the CS method.
DNS	Dynamic Nelson-Siegel (DNS) model with underlying VAR(1) factors fitted using our twelve different maturity yield data, and with a static rate of decay parameter of $\lambda = 0.0609$
DNS+LASSO	DNS model augmented to include a subset of variables selected from our real-time dataset using the LASSO.
DNS+EN	DNS model augmented to include a subset of variables selected from our real-time dataset using the Elastic Net.
DNS+CS	DNS model augmented to include a subset of variables selected from our real-time dataset using the "Chao-Swanson" (CS) variable selection method.
DNS+PCA	DNS model augmented with PCA type factors constructed using our entire real-time dataset.
DNS+LASSO+PCA	DNS model augmented with PCA type factors constructed using a subset of variables from our real-time dataset, with the subset selected using the LASSO.
DNS+LASSO+EN	DNS model augmented with PCA type factors constructed using a subset of variables from our real-time dataset, with the subset selected using the EN.
DNS+LASSO+CS	DNS model augmented with PCA type factors constructed using a subset of variables from our real-time dataset, with the subset selected using the CS method.

Notes: This table includes brief descriptions of the 16 different forecasting models used in our empirical experiment in which we predict monthly interest rates for bonds of maturity 3-months, as well as 1, 3, 5, and 10-years, at 1, 3-, and 12-month ahead forecast horizons. See Section 5 for complete details.

Table 3: Point MSFEs for Various Maturities of U.S. Interest Rates

Forecast Horizon/Model			Interest Rate Maturity		
h=1	3mo	1yr	3yr	5yr	10yr
AR(SIC)	0.0367	0.0320	0.0435	0.0477	0.0491
AR+LASSO	0.0303	0.0335	0.0467	0.0601**	0.0595**
AR+EN	0.0344*	0.0378	0.0652***	0.0730**	0.0686***
AR+CS	0.0374	0.0455	0.0978	0.1100	0.1095
AR+PCA	0.0356	0.0320	0.0438	0.0480	0.0494
AR+LASSO+PCA	0.0373	0.0313	0.0459	0.0492	0.0501
AR+EN+PCA	0.0357**	0.0324	0.0436***	0.0488	0.0501
AR+CS+PCA	0.0315	0.0294	0.0441	0.0488	0.0500
DNS	0.1680***	0.2895***	0.1416***	0.0955***	0.1887***
DNS+LASSO	0.0346***	0.0415***	0.0553***	0.0667***	0.0558***
DNS+EN	0.0376*	0.0508	0.0602	0.0880***	0.0576
DNS+CS	0.0401	0.0433	0.0527	0.0703	0.0715***
DNS+PCA	0.0493	0.0473	0.0532	0.0702	0.0559**
DNS+LASSO+PCA	0.0472	0.0461	0.0504	0.0651*	0.0570
DNS+EN+PCA	0.0436*	0.0442	0.0504	0.0663	0.0559
DNS+CS+PCA	0.0440	0.0343***	0.0481	0.0684	0.0546
h=3					
AR(SIC)	0.2209	0.2124	0.2301	0.2159	0.1909
AR+LASSO	0.1761*	0.2208	0.2977**	0.2995***	0.2259**
AR+EN	0.1821	0.2380	0.2972	0.3400	0.2766***
AR+CS	0.1870	0.2381	0.3547	0.3546	0.3164
AR+PCA	0.2096	0.2078	0.2423**	0.2351***	0.2055***
AR+LASSO+PCA	0.2363	0.2539**	0.2675	0.2318	0.2231
AR+EN+PCA	0.2242*	0.2390*	0.2608	0.2508**	0.2053*
AR+CS+PCA	0.1925	0.2172	0.2659	0.2453	0.2038
DNS	0.2186**	0.3444***	0.1975	0.1641	0.2704**
DNS+LASSO	0.1350***	0.1469***	0.1872***	0.2165	0.1654***
DNS+EN	0.1273	0.1546	0.2039	0.2274	0.1716
DNS+CS	0.1370	0.1448	0.1901	0.2092	0.1819
DNS+PCA	0.1927**	0.1795*	0.2116	0.2216	0.1767
DNS+LASSO+PCA	0.2056*	0.1901	0.2441*	0.2202	0.1791
DNS+EN+PCA	0.2108	0.1888	0.2126***	0.2325	0.1725
DNS+CS+PCA	0.1667***	0.1517***	0.2011	0.2159	0.1725
h=12					
AR(SIC)	2.7069	2.5697	1.9402	1.3847	0.8476
AR+LASSO	1.6320***	1.1118***	1.6501	1.5874	1.4823***
AR+EN	1.4464	1.0408	1.1346***	1.1303***	1.2279***
AR+CS	1.0182***	1.4160	1.6360*	1.2730	0.9431**
AR+PCA	3.5114***	2.9847***	1.9703	1.4047	0.9173
AR+LASSO+PCA	2.9536**	2.4207***	1.9940	1.4123	1.3234***
AR+EN+PCA	3.4974*	2.6272	1.9954	1.5801	1.3021
AR+CS+PCA	3.0519	3.3013**	2.5987***	1.7962*	1.1204
DNS	0.6586***	0.7340***	0.5146***	0.4720***	0.5409***
DNS+LASSO	1.7234***	1.3490***	0.9325***	0.7391***	0.5374***
DNS+EN	1.7606*	1.3172	0.9019	0.7612***	0.5265
DNS+CS	1.7807***	1.3727***	0.9808***	0.8312***	0.6337***
DNS+PCA	1.8072	1.4837***	1.0360	0.8345	0.5687***
DNS+LASSO+PCA	2.0002***	1.6740***	1.1609***	0.9063***	0.5864***
DNS+EN+PCA	2.0345	1.6647	1.1531	0.8863	0.5845
DNS+CS+PCA	1.7879***	1.4277***	0.9647***	0.7774***	0.5390**

Notes: See notes to Table 2. Entries are MSFEs for predictions of 3-month, 1, 3, 5, and 10-year maturity yields, at h =1, 3, and 12-month ahead horizons. The entire sample period of the yield and real-time macroeconomic datasets used is 1988:08-2021:12, while the forecast sample 2001:04-2021:12. Forecasting models, denoted in column one, are defined in Table 2. Entries superscripted with *,**,or *** denote rejection, based on application of the Diebold-Mariano (1995) predictive accuracy test, of the null hypothesis of equal predictive accuracy, relative to the AR(SIC) benchmark model, on average, at 10 %,5%,or 1% significance levels, respectively. See Section 5 for complete details.

of Pihadelphia as well as participants at the 2022 Summer Econometrics Society Meetings, the 2022 International Association of Applied Econometrics Association meetings, the DC-MD-VA Econometrics Workshop, the 2022 NBER-NSF Time Series Conference, the Spring 2023 Rochester Conference in Econometrics, and the 8th Annual Conference of the Society for Economic Measurement for useful comments received on earlier versions of this paper. Chao thanks the University of Maryland for research support.

References

[1] Ahn, S. C. and J. Bae (2022): "Forecasting with Partial Least Squares When a Large Number of Predictors Are Available," Working Paper, Arizona State University and University of Glasgow.

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

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

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

[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., V. Chernozhukov, and C. Hansen (2014): "Inference on Treatment Effects after Selection among High-Dimensional Controls," *Review of Economic Studies*, 81, 608-650.

[9] Bickel, P. J. and E. Levina (2008): "Covariance Regularization by Thresholding," *Annals of Statistics*, 36, 2577-2604.

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

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

[12] Chao, J. C., Y. Liu, and N. R. Swanson (2023): "Consistent Factor Estimation and Forecasting in Factor-Augmented VAR Models," Working Paper, Rutgers University and University of Maryland.

[13] Chao, J. C., K. Qiu, and N. R. Swanson (2023): Online Supplement to "Selecting the Relevant Variables for Factor Estimation in VAR Models, With An Application to Forecasting the Yield Curve" Working Paper, Rutgers University and University of Maryland.

[14] Diebold, F. X., G. D. Rudebusch, and S. B. Aruoba (2006): "The macroeconomy and the Yield Curve: A Dynamic Latent Factor Approach," *Journal of Econometrics*, 131, 309–338.

[15] 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.

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

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

[18] 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.

[19] Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2005): "The Generalized Dynamic Factor Model, One-Sided Estimation and Forecasting," *Journal of the American Statistical Association*, 100, 830-840.

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

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

[22] 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.

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

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

[25] Gurkaynak, R.S., B. Sack, and J.H. Wright(2007): "The US Treasury Yield Curve: 1961 to the Present,"*Journal of Monetary Economics*, 54, 2291-2304.

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

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

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

[29] Kiefer, N. M. and T. J. Vogelsang (2002): "Heteroskedasticity-Autocorrelation Robust Standard Errors Using the Bartlett Kernel without Truncation," *Econometrica*, 70, 2093-2095.

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

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

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

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

[34] 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.

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

[36] Swanson, N.R. and W. Xiong (2018): "Big Data Analytics in Economics: What Have We Learned So Far, and Where Should We Go From Here?" *Canadian Journal of Economics*, 51, 695–746.

[37] Zhou, Z. and X. Shao (2013): "Inference for Linear Models with Dependent Errors," *Journal of the Royal Statistical Society Series B*, 75, 323-343.