Selecting the Relevant Variables for Factor Estimation in FAVAR Models*

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

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

October 6, 2023

Abstract

When specifying and estimating latent factor models, a common assumption made is one of factor pervasiveness, 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. This paper builds on the recent nascent literature that examines how to relax this assumption (see e.g., Giglio, Xiu, and Zhang (2021), Freyaldenhoven (2021a,b), and Bai and Ng (2021)) and analyzes the 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. Consistent factor estimation turns out to be feasible, even under factor nonpervasiveness, if one first prescreens all available variables and prunes out the irrelevant ones. For this purpose, we introduce, within a factor-augmented VAR framework, a novel variable selection procedure that, with probability approaching one, correctly distinguishes between relevant and irrelevant variables. Our methodology enables the consistent estimation of conditional mean functions of factor-augmented forecast equations, even when the conventional assumption of factor pervasiveness is violated.

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

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

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

Norman R. Swanson, Department of Economics, 9500 Hamilton Street, Rutgers University, nswanson@econ.rutgers.edu.
The authors are grateful to Simon Freyaldenhoven, Yuan Liao, Minchul Shin, Jim Stock, Timothy Vogelsang, Endong Wang, Xiye Yang, Bo Zhou and seminar participants at the University of Glasgow, the University of California Riverside, the Federal Reserve Bank of Pihadelphia, the 2022 Summer Econometrics Society Meetings, the 2022 International Association of Applied Econometrics Association meetings, the DC-MD-VA Economtrics Workshop, and the 2022 NBER-NSF Time Series Conference 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 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 the work of Bai and Ng (2002), Stock and Watson (2002a,b), and Forni, Hallin, Lippi, and Reichlin (2005). 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[1].

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 econometric 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 (2022a), 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 nonperva-

[1] In addition to the greater variety of data that are being collected now, an important source of high dimensionality in economic datasets is the use of disaggregate, as opposed to aggregate data (see e.g. Qiu and Qu (2021)). Disaggregate data may be more informative than aggregate data in situations where there is information loss in the process of aggregation.

sive. 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 pre-screens 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, Qiu, 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.[2] 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 principal 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 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 mod-

[2] See Theorem 4.2 of Chao, Liu, and Swanson (2022a). A proof of Theorem 4.2 can be found in the Technical Appendix to that paper (see Chao, Liu, and Swanson (2022b)).

els 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.[3]

Before continuing, it should be stressed that our variable selection method is not designed to pre-select variables that have predictive content for a target variable, for later use in the construction of factors. Rather, we are interested in consistent factor estimation under conditions weaker than those posited in the extant literature. We argue that extant methods that are used to pre-select a group of variables for subsequent use in factor estimation and forecasting may yield inconsistent factor estimates. This is true because variables in the uderlying factor model that are relevant for estimation of the underlying factors may not have predictive content for the target variable of interest. This is somewhat analogous to a ommitted variables problem leading to inconsistent factor estimation. Consider the following example, which illustrates how the methodology developed in this paper might be misinterpreted. To give a concrete example of what's at stake, consider the following. Assume that there are two uncorrelated factors, one loads on variable Z1 and one loads on variable Z2, but only Z2 is directly useful for predicting a target variable, say Y. Our method will not just select Z2. Rather, both Z1 and Z2 will be selected because because they both provide valuable information for factor estimation. Our method ensures consistent factor estimation, and guaranetees that subsequent prediction using these factors in a factor-augmented regression with not be affected by the possible inclusion of inconsistently estimated factors in the model. Of course, the proverbial proof is in the pudding, and to this end we show via Monte Carlo experiments that our method works as expected in finite samples, and that it is easy to construct an empirical example where use of our method to estimate factors yields appreciably better predictions in factor augmented autoregressions than when standard PCA or t-statistic based thresholding methods are used in factor construction. Interestingly, our empirical illustration uses a relatively

[3] 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.

small real-time macroeconomic dataset with only around 100 variables. In this sense, our pervasivness assumption, as well as our assumption regarding the rate at which $N/T \to \infty$ delivers asymptotic results that are relevant even in empirical settings where N is not very large, realtive to T.

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 even in situations where the number of irrelevant variables is very large and 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 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.

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, and makes recommendations regarding the calibration of the tuning parameter used in the said method. Section 4 offers some concluding remarks. Proofs of the main theorems and of two key supporting lemmas are provided in the Appendix to this paper. In addition, some further technical results are reported in an Online Appendix, Chao and Swanson (2022).

Before proceeding, we first say a few words about some of the frequently used notation in this paper. Throughout, let $\lambda_{(j)}(A)$, $\lambda_{\max}(A)$, and $\lambda_{\min}(A)$ denote, respectively, the j^{th} largest eigenvalue, the maximal eigenvalue, and the minimal eigenvalue 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. For two sequences, $\{x_T\}$ and $\{y_T\}$, write $x_T \sim y_T$ if $x_T/y_T = O(1)$ and $y_T/x_T = O(1)$, as $T \to \infty$. 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 $\sigma(u_{i,t+m}, u_{i,t+m+1}, u_{i,t+m+2}, ...)$ denote the σ-field generated by this sequence of random variables.

2 Model, Assumptions, and Variable Selection in High Dimensions

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

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

where

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

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

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. First, it is easy to see that the system of equations given in (1) can be written in the form:

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

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

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

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

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:

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

$$H^c = \{k \in \{1, ..., N\} : \gamma_k \neq 0\}. \tag{7}$$

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

$$\underset{N_1 \times 1}{Z_t^{(1)}} = \Gamma_1 \underline{F}_t + u_t^{(1)} \tag{8}$$

$$\underset{N_2 \times 1}{Z_t^{(2)}} = u_t^{(2)}. \tag{9}$$

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 (2022a) which considers a setting similar to the specification given in expressions (5)-(9) above, but for the case of a simple one-factor model. More precisely, Chao, Liu, and Swanson (2022a) give an example which shows that, in this situation 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[4].

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

[4]The reason why we refer to the result given in Chao, Liu, and Swanson (2022a) 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.

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 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 Chao, Liu, and Swanson (2022a), we add to the results given in Bai and Ng (2021) by giving a result (Theorem 4.1 of Chao, Liu, and Swanson (2022a)) 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), Kan and Zhang (1999), 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:

$$\det\left\{I_{(d+K)} - A_1 z - \cdots - A_p z^p\right\} = 0, \text{ implies that } |z| > 1. \tag{10}$$

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 $M < \infty$, $\sup_t \int \left|g_{\varepsilon_t}(v-u) - g_{\varepsilon_t}(v)\right| dv \leq M\|u\|$, whenever $\|u\| \leq \overline{\kappa}$ 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 $\mathcal{F}_{i,-\infty}^t = \sigma\left(\ldots., u_{i,t-2}, u_{i,t-1}, u_t\right)$, $\mathcal{F}_{i,t+m}^{\infty} = \sigma\left(u_{i,t+m}, u_{i,t+m+1}, u_{i,t+m+2}, \ldots.\right)$, 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

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

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 $\|\mu\|_2 \leq \overline{C} < \infty$, where $\mu = \left(\mu_Y', \mu_F'\right)'$.

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

$$\sigma_{\min}\left(I_{(d+K)p} - A\right) \geq \underline{C}\,\overline{\phi}_{\min} \tag{11}$$

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

$$\sigma_{\max}\left(A^j\right) \leq \overline{C} \max\left\{\left|\lambda_{\max}\left(A^j\right)\right|, \left|\lambda_{\min}\left(A^j\right)\right|\right\}. \tag{12}$$

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 (11) and (12) 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 $\mathcal{A}_{-\infty}^t = \sigma(..., W_{t-2}, W_{t-1}, W_t)$ and $\mathcal{A}_{t+m}^{\infty} = \sigma(W_{t+m}, W_{t+m+1}, W_{t+m+2}, ...)$[5]. Note, in addition, 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

[5] This can be shown by applying Theorem 2.1 of Pham and Tran (1985). A proof of this result is also given in Chao and Swanson (2022). See, in particular, Lemma OA-11 and its proof in Chao and Swanson (2022).

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 assumption of Gaussianity or of sub-exponential tail behavior which has been made in papers studying high-dimensional factor models and/or high-dimensional covariance matrices, without employing statistics that are self-normalized[6].

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:

$$\max_{1\leq \ell\leq d} |S_{i,\ell,T}| = \max_{1\leq \ell\leq d} \left| \frac{\overline{S}_{i,\ell,T}}{\sqrt{\overline{V}_{i,\ell,T}}} \right|, \tag{13}$$

where

$$\overline{S}_{i,\ell,T} = \sum_{r=1}^{q} \sum_{t=(r-1)\tau+p}^{(r-1)\tau+\tau_1+p-1} Z_{it} y_{\ell,t+1} \text{ and } \overline{V}_{i,\ell,T} = \sum_{r=1}^{q} \left[\sum_{t=(r-1)\tau+p}^{(r-1)\tau+\tau_1+p-1} Z_{it} y_{\ell,t+1} \right]^2 . \tag{14}$$

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 (13) 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 (14) 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

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

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:

$$\begin{aligned}\overline{S}_{i,\ell,T} &= \sum_{t=p}^{\tau_1+p-1} Z_{it}y_{\ell,t+1} + \sum_{t=\tau+p}^{\tau+\tau_1+p-1} Z_{it}y_{\ell,t+1} \\ &\quad + \sum_{t=2\tau+p}^{2\tau+\tau_1+p-1} Z_{it}y_{\ell,t+1} + \cdots + \sum_{t=(q-1)\tau+p}^{(q-1)\tau+\tau_1+p-1} Z_{it}y_{\ell,t+1} \qquad (15)\end{aligned}$$

Comparing the first term and the second term on the right-hand side of expression (15), 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, $T \to \infty$[7]. 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

[7]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.

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 $\sum_{\ell=1}^{d} \varpi_{\ell} |S_{i,\ell,T}|$.[8] Our variable selection procedure is based on the decision rule:

$$i \in \begin{cases} \widehat{H}^c & \text{if } \mathbb{S}^{+}_{i,T} \geq \Phi^{-1}\left(1-\frac{\varphi}{2N}\right) \\ \widehat{H} & \text{if } \mathbb{S}^{+}_{i,T} < \Phi^{-1}\left(1-\frac{\varphi}{2N}\right) \end{cases}, \tag{16}$$

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:

(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 a slight modification of the arguments given in the proof of Lemma A2[9], we can show that, as $T \to \infty$

$$P(|S_{i,\ell,T}| \geq z) = 2[1-\Phi(z)](1+o(1)), \tag{17}$$

which holds for all i and ℓ and for all z such that
$0 \leq z \leq c_0 \min\left\{T^{(1-\alpha_1)/6}/L(T), T^{\alpha_2/2}\right\}$, where $L(T)$ denotes a slowly varying function such that $L(T) \to \infty$ as $T \to \infty$. In view of expression (17), 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, expression (17) 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:

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

so that the probability of a Type I error is controlled at the desired level α asymptotically.

[8] 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 (2002a, 2002b), 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.

[9] The statement and proof of Lemma A2 are provided below in the Appendix to this paper.

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 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 (16), 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 Belloni, Chen, Chernozhukov, and Hansen (2012) and further studied 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 the two above papers is quite different from that investigated in our paper. In particular, Belloni, Chenozhukov, and Hansen (2014) 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. As noted in Belloni, Chernozhukov, and Hansen (2014), 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 those in Belloni, Chen, Chernozhukov, and Hansen (2012) and 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 $\tau_1 \geq 1$:

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

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 2-9 imposes the condition that there exists a positive constant, $\underline{c}$, such that, for all N_1, N_2, and T sufficiently large:

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

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) above implies that the ℓ^{th} component of Y_{t+1} will have the representation $y_{\ell,t+1} = \mu_{Y,\ell} + \underline{Y}_t'\alpha_{YY,\ell} + \underline{F}_t'\alpha_{YF,\ell} + \varepsilon^Y_{\ell,t+1} = \varepsilon^Y_{\ell,t+1}$, 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)
$$P\left(\max_{i \in H} \max_{1 \leq \ell \leq d} \left|S_{i,\ell,T}\right| \geq \Phi^{-1}\left(1 - \frac{\varphi}{2N}\right)\right) = O\left(\frac{N_2 \varphi}{N}\right) = o(1).$$

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:*

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

(b)

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

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 $N_1, N_2, T \to \infty$. 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 $N_1, N_2, T \to \infty$. Together, these two theorems show that our procedure is (completely) consistent in the sense that the probability of committing a misclassification error vanishes as $N_1, N_2, T \to \infty$.

(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 (2022b), 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 tri-variate FAVAR(1) process:

$$W_t = \mu + AW_{t-1} + \varepsilon_t, \quad (18)$$

$$Z_t = \gamma F_t + u_t, \quad (19)$$

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 (18), we take $\{\varepsilon_t\} \equiv i.i.d.N(0, \Sigma_\varepsilon)$, 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 (19) 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 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. 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.

Looking across each row of Tables 1, note that FPRs decrease when moving from left to right, whereas FNRs increase. This result remains the same when comparing Tables 2-7, and is not surprising, because moving from $\varphi = N^{-0.2}$ to $\varphi = N^{-0.7}$ 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, our 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, 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 $\vartheta = 0.6$ or 0.7 can lead to high FNRs, 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 $\varphi = N^{-0.4}$. As discussed in Remark 4.1 of Chao, Liu, and Swanson (2022a), this choice of the tuning parameter allows a rate condition needed for consistent factor estimation using the selected variables to be automatically satisfied as long as $N_1 \to \infty$, so that there is no need to further impose a condition on the rate at which N_1 grows. See Theorem 4.1 and Remark 4.1 of Chao, Liu, and Swanson (2022a) for additional discussion.

Looking down the columns of each 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 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.

4 Empirical Illustration

In this illustration, we forecast eight target variables from the monthly real-time macroeconomic FRED-MD dataset maintained by St. Louis Federal Reserve Bank. We follow the data cleaning methods outlined in on the FRED-MD data website, as well as removing all discontinued series, when pre-processing the data, yielding a dataset, $\mathbf{X}$, containing of $N = 97$ variables for the period 1973:3 to 2022:9. The full list of all macroeconomic variables and their transformations is available upon request from the authors.

Of note is that the dataset used here is "truly" real-time, in the sense that a "vintage" of data is available at calendar date in our sample period. Consider the value of industrial production for January 2020. In February 2020, the government reported a "first release" value for January. In March 2020, however, they further updated their "estimate" of industrial production for January. Namely, they reported a "second release" for January. This process of revision continues indefinitely. Namely, as the government changes data collection and processing methodology, collects new data and/or revises definitions of variables, new releases are reported. A "vintage" of data is a date, say February 2020. For industrial production, there is a whole vector of truly real-time data that includes different releases, all available in February 2020. For example, this vintage includes a 1st release value for January 2020, a 2nd release value for December 2019 is included in this vintage, a 3rd release value for November 2019, etc. In this sense, there is an entirely unique vintage of industrial production available each month, and the values of the calendar dated observations in each vintage change because the government updates historical values of the variable every month. Using this type of data allows the practitioner to truly simulate a forecasting environment in which models are updated at each point in time using data that were actually available at that time. Were we to simply collect industrial production data from a website today, calendar dated observations in our dataset from 2020 would reflect revisions that occurred after 2020. For variables that are revised regularly, this means that forecasting experiments of the type carried out in this paper would be invalid, in the sense that they would be utilizing "future data" as explanatory variables when estimating forecasting model regressions. For further discussion of the structure of real-time datasets, as well as methods for real-time forecasting, refer to Swanson (1996), Swanson, Ghysels, and Callan (1999), and Kim and Swanson (2018). For variables that are subject to revision, this means that forecasting experiments of the type carried out in this paper would be invalid, in the sense that they would be utilizing "future data" as explanatory variables when estimating forecasting model regressions, were the correct vintage of data not used at each point in time when estimating models and constructing forecasts.

The 8 target variables for which we construct predictions are: Industrial Production (INDPRO), Civilian Unemployment Rate (UNRATE), Housing Starts: new, privately owned (HOUST), Housing Permits: new, privately owned (PERMIT), Real M2 Money Stock (M2REAL), 10-Year Government Treasury Bond Rate (R10), CPI - All Items (CPI), S&P Common Stock Price Index - Composite (S&P 500). Our experiments compare variable selection and dimension reduction methods when used to estimate and/or select factors and observable variables for inclusion in forecasting models of the form:

$$y_{t+h} = \alpha + \beta_h(L)y_t + \gamma_h(L)\boldsymbol{F_t} + \epsilon_{t+h}, \tag{20}$$

where y_t is a scalar target variable to be predicted, $\beta_h(L)$ and $\gamma_h(L)$ are finite order lag polynomials, $\boldsymbol{Z_t}$ contains latent factors (set $\boldsymbol{Z_t} = \boldsymbol{F_t}$) and/or observable variables, and ϵ_t is a stochastic disturbance term. Lags in this model are selected using the Schwarz Information Crierion (SIC), and our benchmark model sets the coefficients in $\gamma_h(L) = 0$. In the sequel, we carry out variable selection and dimension reduction using seven different methods.

Principal Components Analysis (PCA): Excluding the target variable, apply PCA to $\mathbf{X}$ and estimate latent factors, $\boldsymbol{F_t}$, with the number of factors determined using the PC_{p2} criterion in Bai and Ng (2002). The maximum number of the factors is set equal to eight, following the findings of McCracken and Ng (2016), who introduce and examine the dataset that we utilize in our experiments.

t-Statistic Thresholding: For each variable in $\mathbf{X}$, and forecast horizon, h, perform a regression of y_{t+h} on lags of y_t and on $X_{i,t}$, where $X_{i,t}$ is a scalar variable in $\mathbf{X}$, for $i = 1, ..., N$, and lags of y_t are selected using the SIC. Let t_i denote the t statistic associated with X_{it-h} in the regression, and select variables, X_{it} if $|t_i| > 1.28$. If the number of selected variables is greater than 20, utilize PCA to estimate factors for inclusion in the above forecasting equation, otherwise use the AR(SIC) model. As models are re-estimated at each point in time, this approach is a hybrid, in the sense that some models may include factors as regressors, while others may be simple AR(SIC) models. Note that in our experiments, less than 10% of the total number of forecasting periods involved replacing the thresholding model with our AR(SIC) benchmark.

Chao-Swanson Variable Selection: Use the variable selection method introduced in this paper to select variables. Then, use PCA to estimate factors for inclusion in the forecasting equation. There are three tuning parameters in the CS method, including: τ, τ_1, and φ.

We set $\{\tau = 5, \tau_1 = 3, 5\}$ and $\{\tau = 10, \tau_1 = 6, 8\}$ and consider the following values for φ:

$$\varphi = \begin{cases} (lnlnN)^{-0.1} & (lnlnN)^{-0.6} & (lnN)^{-0.1} & (lnN)^{-0.6} & N^{-0.1} & N^{-0.6} \\ (lnlnN)^{-0.2} & (lnlnN)^{-0.7} & (lnN)^{-0.2} & (lnN)^{-0.7} & N^{-0.2} & N^{-0.7} \\ (lnlnN)^{-0.3} & (lnlnN)^{-0.8} & (lnN)^{-0.3} & (lnN)^{-0.8} & N^{-0.3} & N^{-0.8} \\ (lnlnN)^{-0.4} & (lnlnN)^{-0.9} & (lnN)^{-0.4} & (lnN)^{-0.9} & N^{-0.4} & N^{-0.9} \\ (lnlnN)^{-0.5} & (lnlnN)^{-1} & (lnN)^{-0.5} & (lnN)^{-1} & N^{-0.5} & N^{-1} \end{cases}$$

Different tuning parameters select different numbers of variables and we exclude tuning parameter permutations that select less than 25 variables for use in factor construction. In this method, the tuning parameter used for each value of h and target variable is selected selected by partitioning a “training dataset” consisting of the first 10 years on data in our sample into an in-sample period of 7 years and an out-of-sample period of 3 years. The tuning parameter is set equal to that yielding the smallest mean square forecast error (MSFE) after constructing real-time predictions based on models estimated at each point in time prior to the construction of each new prediction for the out-of-sample period. Note that in our experiments, less than 10% of the total number of forecasting periods involved replacing the selected variables with those from the previous period.

In summary, we carry out truly real-time h-month ahead predictions using monthly data, with h =1, 3, 6, and 12. Our “full sample forecasting period” is 2000:1-2022:9 (when reporting results for this period, we omit predictions for 2008:1-2008:12 and 2020:1-2020:12, in order to mitigate the influence of predictions made during the 2008 Financial Crisis and the Covid-19 period on our results). However, even though some predictions are omitted in our “full-sample”, data from these extraordinary periods in history still affects the estimated models used when predicting other periods. For this reason, this first set of results appears only in the online appendix, where it is clearly seen that the impact of these periods on estimated models is severe, in the sense that predictions are adversely affected. In the sequel, we thus instead report results for the out-of-sample period 2000:1-2007:12 in the online supplement. All factors and forecasting equations are re-estimated at each point in time, prior to the construction of each new forecast, using rolling windows of length 120 observations. Additionally, in-sample estimation periods used when constructing our h =3, 6, and 12-step ahead forecasts are adjusted so that the forecast period remains the same regardless of forecast horizon.

Forecasting performance is evaluated using point mean squared forecast errors (MS-

FEs), where MSFE$=\frac{1}{P}\sum_{t=1}^{T}(y_{j,t}-\hat{y}_{j,t})^2$, and $\hat{y}_{j,t}$ denotes the prediction for target variable y_j that is made using data that are truly available in real-time at period t. In our tabulated results, MSFEs, relative to that of the benchmark AR(SIC) model are reported. Additionally, we report the results of Giacomini and White (GW) tests (see Giacomini and White (2006)), which can be viewed as conditional Diebold-Mariano (DM) predictive accuracy tests (see Diebold and Mariano (1995)). Recall that the null hypothesis of the DM test when formulated using the conditioning approach of Giacomini and White is: $H_0 : \mathrm{E}[L(\hat{\epsilon}_{t+h}^{(1)})|G_t] - \mathrm{E}[L(\hat{\epsilon}_{t+h}^{(2)})|G_t] = 0$, where the $\hat{\epsilon}_{t+h}^{(i)}$ are prediction errors associated with model i, for $i = 1, 2$, and G_t denotes the conditioning set, which includes the model and estimated parameters. Here, $L(\cdot)$ is a quadratic loss function, and the test statistic is $\mathrm{DM}_P = P^{-1}\sum\limits_{t=1}^{P}\frac{d_{t+h}}{\hat{\sigma}_{\bar{d}}}$, where $d_{t+h} = [\hat{\epsilon}_{t+h}^{(1)}]^2 - [\hat{\epsilon}_{t+h}^{(2)}]^2$, $\bar{d}$ denotes the mean of d_{t+h}, $\hat{\sigma}_{\bar{d}}$ is a heteroskedasticity and autocorrelation consistent estimate of the standard deviation of $\bar{d}$, and P denotes the number of ex-ante predictions used to construct the test statistic.[10] If the statistic is "significantly negative", then Model 1 is preferred to Model 2, and in our context, where we report relative MSFEs, rejection indicates that the benchmark AR(SIC) model is preferred if the relative MSFE is greater than one, and the converse if it is less than one.

Turning to our empirical findings, note first that a summary of the target variables in our experiments is contained in Table 9. Table 10 contains results for our full sample forecasting period. In this table, all entries are relative MSFEs, as discussed above. Additionally, bolded entries indicate the "MSFE-best" method for a particular target variable and forecast horizon. Since relative MSFEs are reported, however, if the lowest relative MSFE is greater than 1, it is not bolded, as this means that the AR(SIC) benchmark yields the MSFE-best predictions. Starred entries denote rejection of the null hypothesis of equal forecast accuracy when comparing the model associated with a given method against the AR(SIC) benchmark. Turning to the results in this table, a number of conclusions can be made.

First the compare only PCA with our variable selection method (called CS hereafter). PCA yields a MSFE that is less than unity (hence "beating" the AR(SIC) benchmark) and lower than that associated with CS in 3 of 32 variable and forecast horizon permutations. On the other hand, CS "wins" in 15 of 32 cases. Thus, even for this small dataset, there are substantial gains to using CS. Second, when hard thresholding (HT) is brought into the mix, PCA wins only once, THRESH wins 5 times, and CS wins 13 times.[11] Thus, while we

[10] In this paper, we report test results for the Wald version of this test statistic (see Giacomini and White (2006) for further details).

[11] The reason why the sum of "wins" is now 19 instead of 18 is that TRESH and CS "tie" in one case. In

might expect CS to outperform PCA, which is handily does, CS also yield superior results to THRESH, where variables are pre-selected using thresholding. Our theory suggests that this should be so in certain settings, such as when pervasiveness is substantially relaxed. Finally, notice that the AR(SIC) benchmark "wins" in 14 of 32 caes, or 44% of the time. This speaks to the approximate nature of the factor model that we are using, particularly when forecating variables like interest rates, which we see is usually most accuratley forecast using the AR(SIC) model. Additionally, when the $h = 1$ case is ommitted from the tally, the AR(SIC) model "wins" in 9 of 24 cases, or 35% of the time. Thus, big data reamins useful at longer horizons when forecasting the macroeconomic variables in our dataset.

5 Conclusion

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 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. Moreover, we present empirical evidence suggesting that use of our procedure yields factors that improve the forecasting performance of factor augmented regressions, relative to the case when principal component analysis or hard thresholding methods are used to construct factors. It is also worth noting that in a Chao, Qiu, 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 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. In sum, the collective body of results discussed in this paper indicates that the variable selection methodology introduced in this paper can be useful to empirical researchers as they

this case, each are assigned a "win".

engage in the important tasks of factor estimation and the construction of point forecasts based on factor-augmented forecasting equations.

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

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

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

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

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

[10] 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.

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

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

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

[14] Chao, J. C., Y. Liu, and N. R. Swanson (2022a): “Consistent Factor Estimation and Forecasting in Factor-Augmented VAR Models,” Working Paper, Rutgers University and University of Maryland.

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

[16] Chao, J. C. and N. R. Swanson (2022): Online Appendix to “Selecting the Relevant Variables for Factor Estimation in VAR Models,” Working Paper, Rutgers University and University of Maryland.

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

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

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

[20] 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.

[21] 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.

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

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

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

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

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

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

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

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

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

[31] Kiefer, N. M. and T. J. Vogelsang (2002b): "Heteroskedasticity-Autocorrelation Robust Testing Using Bandwidth Equal to Sample Size," *Econometric Theory*, 18, 1350-1366.

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

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

[35] Qiu, A. and Z. Qu (2021): "Modeling Regime Switching in High-Dimensional Data with Applications to U.S. Business Cycles," Working Paper, Boston University.

[36] 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.

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

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

6 Appendix: Proofs of Theorems

This appendix contains the proofs of the main results of the paper: Theorems 1 and 2. In addition, two key supporting lemmas, Lemmas A1and A2, along with their proofs are also given here. Additional technical results are available in an Online Appendix, Chao and Swanson (2022).

Proof of Theorem 1: To show part (a), first set $z = \Phi^{-1}\left(1 - \frac{\varphi}{2N}\right)$, where $N = N_1 + N_2$. Note that, under Assumption 2-10, we can easily show that $\Phi^{-1}\left(1 - \frac{\varphi}{2N}\right) \leq \sqrt{2\left(1+a\right)}\sqrt{\ln N}$, for all N_1, N_2 sufficiently large.[12] By part (a) of Assumption 2-9, $\sqrt{\ln N}/\min\left\{T^{(1-\alpha_1)/6}, T^{\alpha_2/2}\right\} \to 0$ as $N_1, N_2, T \to \infty$; this, in turn, implies that, for some positive constant c_0, $\Phi^{-1}\left(1 - \frac{\varphi}{2N}\right)$ satisfies the inequality constraint $0 \leq \Phi^{-1}\left(1 - \frac{\varphi}{2N}\right) \leq c_0 \min\left\{T^{(1-\alpha_1)/6}, T^{\alpha_2/2}\right\}$ for all N_1, N_2, T sufficiently large, so that $\Phi^{-1}\left(1 - \frac{\varphi}{2N}\right)$ lies within the range of values of z for which the moderate deviation inequality given in Lemma A2 holds. Thus, plugging $\Phi^{-1}\left(1 - \frac{\varphi}{2N}\right)$ into the moderate deviation inequality (21) given in Lemma A2 below, we see that there exists a positive constant A such that:

$$
\begin{aligned}
&P\left(\left|S_{i,\ell,T}\right| \geq \Phi^{-1}\left(1 - \tfrac{\varphi}{2N}\right)\right) \\
&\leq 2\left[1 - \Phi\left(\Phi^{-1}\left(1 - \tfrac{\varphi}{2N}\right)\right)\right]\left\{1 + A\left[1 + \Phi^{-1}\left(1 - \tfrac{\varphi}{2N}\right)\right]^3 T^{-\frac{1-\alpha_1}{2}}\right\} \\
&= 2\left[1 - \left(1 - \tfrac{\varphi}{2N}\right)\right]\left\{1 + A\left[1 + \Phi^{-1}\left(1 - \tfrac{\varphi}{2N}\right)\right]^3 T^{-\frac{1-\alpha_1}{2}}\right\} \\
&= \tfrac{\varphi}{N}\left\{1 + A\left[1 + \Phi^{-1}\left(1 - \tfrac{\varphi}{2N}\right)\right]^3 T^{-\frac{1-\alpha_1}{2}}\right\},
\end{aligned}
$$

for $\ell \in \{1, ..., d\}$, for $i \in H = \{k \in \{1, ..., N\} : \gamma_k = 0\}$, and for all N_1, N_2, T sufficiently large. Next, note that:

$$
\begin{aligned}
&P\left(\max_{i \in H} \sum_{\ell=1}^{d} \varpi_\ell \left|S_{i,\ell,T}\right| \geq \Phi^{-1}\left(1 - \tfrac{\varphi}{2N}\right)\right) \\
&\leq P\left(\bigcup_{i \in H} \bigcup_{1 \leq \ell \leq d} \left\{\left|S_{i,\ell,T}\right| \geq \Phi^{-1}\left(1 - \tfrac{\varphi}{2N}\right)\right\}\right) \quad \left(\text{since } 0 \leq \varpi_\ell \leq 1 \text{ and } \sum_{\ell=1}^{d} \varpi_\ell = 1\right) \\
&\leq \sum_{i \in H} \sum_{\ell=1}^{d} P\left(\left|S_{i,\ell,T}\right| \geq \Phi^{-1}\left(1 - \tfrac{\varphi}{2N}\right)\right) \quad \text{(by union bound)} \\
&\leq \sum_{i \in H} \sum_{\ell=1}^{d} \tfrac{\varphi}{N}\left\{1 + A\left[1 + \Phi^{-1}\left(1 - \tfrac{\varphi}{2N}\right)\right]^3 T^{-(1-\alpha_1)\frac{1}{2}}\right\} \\
&= d\tfrac{N_2\varphi}{N}\left\{1 + A\left[1 + \Phi^{-1}\left(1 - \tfrac{\varphi}{2N}\right)\right]^3 T^{-(1-\alpha_1)\frac{1}{2}}\right\}
\end{aligned}
$$

Using the inequality $\Phi^{-1}\left(1 - \frac{\varphi}{2N}\right) \leq \sqrt{2\left(1+a\right)}\sqrt{\ln N}$ discussed above, we further obtain, for all N_1, N_2, T sufficiently large:

[12] An explicit proof of this result is given in Chao and Swanson (2022). In particular, this inequality is shown in part (b) of Lemma OA-15 in Chao and Swanson (2022).

$$P\left(\max_{i\in H}\sum_{\ell=1}^{d}\varpi_\ell\left|S_{i,\ell,T}\right|\geq\Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right)\leq\frac{dN_2\varphi}{N}\left\{1+\frac{A}{T^{(1-\alpha_1)/2}}\left[1+\Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right]^3\right\}$$
$$\leq\frac{dN_2\varphi}{N}\left\{1+2^2AT^{-\frac{(1-\alpha_1)}{2}}+2^2A\left[\Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right]^3T^{-\frac{(1-\alpha_1)}{2}}\right\}$$
$$\left(\text{by the inequality }\left|\sum_{i=1}^{m}a_i\right|^r\leq c_r\sum_{i=1}^{m}\left|a_i\right|^r\text{ where }c_r=m^{r-1}\text{ for }r\geq1\right)$$
$$\leq\frac{dN_2\varphi}{N}\left\{1+4AT^{-\frac{(1-\alpha_1)}{2}}+4A\left[\sqrt{2\left(1+a\right)}\sqrt{\ln N}\right]^3T^{-\frac{(1-\alpha_1)}{2}}\right\}$$
$$=\frac{dN_2\varphi}{N}\left\{1+4AT^{-\frac{(1-\alpha_1)}{2}}+2^{\frac{7}{2}}A\left(1+a\right)^{\frac{3}{2}}\frac{\left(\ln N\right)^{\frac{3}{2}}}{T^{\frac{1-\alpha_1}{2}}}\right\}.$$

Finally, note that the rate condition given in part (a) of Assumption 2-9 (i.e., $\sqrt{\ln N}/\min\left\{T^{(1-\alpha_1)/6},T^{\alpha_2/2}\right\}\to0$ as $N_1,N_2,T\to\infty$) implies that $\left(\ln N\right)^{\frac{3}{2}}/T^{\frac{1-\alpha_1}{2}}\to0$ as $N_1,N_2,T\to\infty$, from which it follows that:

$$P\left(\max_{i\in H}\sum_{\ell=1}^{d}\varpi_\ell\left|S_{i,\ell,T}\right|\geq\Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right)$$
$$\leq\frac{dN_2\varphi}{N}\left\{1+4AT^{-\frac{(1-\alpha_1)}{2}}+2^{\frac{7}{2}}A\left(1+a\right)^{\frac{3}{2}}\frac{\left(\ln N\right)^{\frac{3}{2}}}{T^{\frac{1-\alpha_1}{2}}}\right\}=\frac{dN_2\varphi}{N}\left[1+o\left(1\right)\right]=O\left(\frac{N_2\varphi}{N}\right)=o\left(1\right).$$

Next, to show part (b), note that, by a similar argument as that given for part (a) above, we have:

$$P\left(\max_{i\in H}\max_{1\leq\ell\leq d}\left|S_{i,\ell,T}\right|\geq\Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right)$$
$$=P\left(\bigcup_{i\in H}\bigcup_{1\leq\ell\leq d}\left\{\left|S_{i,\ell,T}\right|\geq\Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right\}\right)$$
$$\leq\frac{dN_2\varphi}{N}\left\{1+\frac{4A}{T^{(1-\alpha_1)/2}}+\frac{2^{\frac{7}{2}}A(1+a)^{\frac{3}{2}}(\ln N)^{\frac{3}{2}}}{T^{(1-\alpha_1)/2}}\right\}=\frac{dN_2\varphi}{N}\left[1+o\left(1\right)\right]=O\left(\frac{N_2\varphi}{N}\right)=o\left(1\right).\ \square$$

Proof of Theorem 2: To show part (a), let $\overline{S}_{i,\ell,T}$ and $\overline{V}_{i,\ell,T}$ be as defined in expression (14), and note that:

$$P\left(\min_{i\in H^c}\sum_{\ell=1}^{d}\varpi_\ell\left|S_{i,\ell,T}\right|\geq\Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right)$$
$$=P\left(\min_{i\in H^c}\sum_{\ell=1}^{d}\varpi_\ell\left|\frac{\overline{S}_{i,,\ell,T}-\mu_{i,\ell,T}}{\sqrt{\overline{V}_{i,\ell,T}}}+\frac{\mu_{i,,\ell,T}}{\sqrt{\overline{V}_{i,\ell,T}}}\right|\geq\Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right)$$
$$\geq P\left(\min_{i\in H^c}\sum_{\ell=1}^{d}\varpi_\ell\left\{\left|\frac{\mu_{i,\ell,T}}{\sqrt{\overline{V}_{i,\ell,T}}}\right|-\left|\frac{\overline{S}_{i,\ell,T}-\mu_{i,\ell,T}}{\sqrt{\overline{V}_{i,\ell,T}}}\right|\right\}\geq\Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right)$$
$$=P\left(\min_{i\in H^c}\sum_{\ell=1}^{d}\varpi_\ell\left\{\left|\frac{\mu_{i,\ell,T}}{\sqrt{\overline{V}_{i,\ell,T}}}\right|\left[1-\left|\frac{\sqrt{\overline{V}_{i,\ell,T}}}{\mu_{i,\ell,T}}\right|\left|\frac{\overline{S}_{i,\ell,T}-\mu_{i,\ell,T}}{\sqrt{\overline{V}_{i,\ell,T}}}\right|\right]\right\}\geq\Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right)$$
$$=P\left(\min_{i\in H^c}\sum_{\ell=1}^{d}\varpi_\ell\left\{\left|\frac{\mu_{i,\ell,T}}{\sqrt{\overline{V}_{i,\ell,T}}}\right|\left[1-\left|\frac{\overline{S}_{i,\ell,T}-\mu_{i,\ell,T}}{\mu_{i,\ell,T}}\right|\right]\right\}\geq\Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right),$$

where $\mu_{i,\ell,T}=\sum_{r=1}^{q}\sum_{t=b_1(r)}^{b_2(r)}\gamma_i'\left\{E\left[\underline{F}_t\right]\mu_{Y,\ell}+E\left[\underline{F}_t\underline{Y}_t'\right]\alpha_{YY,\ell}+E\left[\underline{F}_t\underline{F}_t'\right]\alpha_{YF,\ell}\right\}$,for

$b_1(r) = (r-1)\tau + p$ and $b_2(r) = b_1(r) + \tau_1 - 1$. Next, let $\pi_{i,\ell,T} = \sum_{r=1}^{q}\left(\sum_{t=b_1(r)}^{b_2(r)}\left\{\gamma_i' E[\underline{F}_t]\mu_{Y,\ell} + \gamma_i' E[\underline{F}_t\underline{Y}_t']\alpha_{YY,\ell} + \gamma_i' E[\underline{F}_t\underline{F}_t']\alpha_{YF,\ell}\right\}\right)^2$, and we see that, under Assumption 2-8, there exists a positive constant $\underline{c}$ such that for every $\ell \in \{1, ..., d\}$ and for all N_1, N_2, and T sufficiently large:

$$
\begin{aligned}
&\min_{i\in H^c}\left\{\pi_{i,\ell,T}/\left(q\tau_1^2\right)\right\} \\
&= \min_{i\in H^c}\frac{1}{q}\sum_{r=1}^{q}\left(\frac{1}{\tau_1}\sum_{t=b_1(r)}^{b_2(r)}\left\{\gamma_i' E[\underline{F}_t]\mu_{Y,\ell} + \gamma_i' E[\underline{F}_t\underline{Y}_t']\alpha_{YY,\ell} + \gamma_i' E[\underline{F}_t\underline{F}_t']\alpha_{YF,\ell}\right\}\right)^2 \\
&= \min_{i\in H^c}\frac{1}{q}\sum_{r=1}^{q}\left(\frac{1}{\tau_1}\sum_{t=b_1(r)}^{b_2(r)} E[\gamma_i'\underline{F}_t y_{\ell,t+1}]\right)^2 \\
&\geq \min_{i\in H^c}\left(\frac{1}{q}\sum_{r=1}^{q}\frac{1}{\tau_1}\sum_{t=b_1(r)}^{b_2(r)} E[\gamma_i'\underline{F}_t y_{\ell,t+1}]\right)^2 \quad \text{(by Jensen's inequality)} \\
&= \min_{i\in H^c}\left|\frac{1}{q}\sum_{r=1}^{q}\frac{1}{\tau_1}\sum_{t=b_1(r)}^{b_2(r)}\left\{\gamma_i' E[\underline{F}_t]\mu_{Y,\ell} + \gamma_i' E[\underline{F}_t\underline{Y}_t']\alpha_{YY,\ell} + \gamma_i' E[\underline{F}_t\underline{F}_t']\alpha_{YF,\ell}\right\}\right|^2 \\
&\geq \underline{c}^2 > 0 \quad \text{(in light of Assumption 2-8)}.
\end{aligned}
$$

It follows that for all N_1, N_2, and T sufficiently large:

$$
\begin{aligned}
&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) \\
&\geq P\left(\min_{i\in H^c}\sum_{\ell=1}^{d}\varpi_\ell\left\{\left|\frac{\mu_{i,\ell,T}}{\sqrt{\overline{V}_{i,\ell,T}}}\right|\left[1-\left|\frac{\overline{S}_{i,\ell,T}-\mu_{i,\ell,T}}{\mu_{i,\ell,T}}\right|\right]\right\} \geq \Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right) \\
&= P\left(\min_{i\in H^c}\sum_{\ell=1}^{d}\varpi_\ell\left\{\left|\frac{\sqrt{q}\left[\mu_{i,\ell,T}/(q\tau_1)\right]}{\sqrt{\pi_{i,\ell,T}/\left(q\tau_1^2\right)}}\right|\left|\frac{\sqrt{\pi_{i,\ell,T}/\left(q\tau_1^2\right)}}{\sqrt{\pi_{i,\ell,T}/\left(q\tau_1^2\right)}+\sqrt{\overline{V}_{i,\ell,T}/\left(q\tau_1^2\right)}-\sqrt{\pi_{i,\ell,T}/\left(q\tau_1^2\right)}}\right|\right.\right. \\
&\qquad\qquad \left.\left.\times\left[1-\left|\frac{\overline{S}_{i,\ell,T}-\mu_{i,\ell,T}}{\mu_{i,\ell,T}}\right|\right]\right\} \geq \Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right) \\
&= P\left(\min_{i\in H^c}\sum_{\ell=1}^{d}\varpi_\ell\left\{\left|\frac{\sqrt{q}\left[\mu_{i,\ell,T}/(q\tau_1)\right]}{\sqrt{\pi_{i,\ell,T}/\left(q\tau_1^2\right)}}\right|\left|\frac{1}{1+\left(\sqrt{\overline{V}_{i,\ell,T}}-\sqrt{\pi_{i,\ell,T}}\right)/\sqrt{\pi_{i,\ell,T}}}\right|\right.\right. \\
&\qquad\qquad \left.\left.\times\left[1-\left|\frac{\overline{S}_{i,\ell,T}-\mu_{i,\ell,T}}{\mu_{i,\ell,T}}\right|\right]\right\} \geq \Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right) \\
&\geq P\left(\min_{i\in H^c}\sum_{\ell=1}^{d}\varpi_\ell\left\{\left|\frac{\sqrt{q}\left(\mu_{i,\ell,T}/(q\tau_1)\right)}{\sqrt{\pi_{i,\ell,T}/\left(q\tau_1^2\right)}}\right|\frac{1}{1+\max_{k\in H^c}\left|\sqrt{\overline{V}_{k,\ell,T}}-\sqrt{\pi_{k,\ell,T}}\right|/\sqrt{\pi_{k,\ell,T}}}\right.\right. \\
&\qquad\qquad \left.\left.\times\left[1-\max_{k\in H^c}\left|\frac{\overline{S}_{k,\ell,T}-\mu_{k,\ell,T}}{\mu_{k,\ell,T}}\right|\right]\right\} \geq \Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right) \\
&\geq P\left(\min_{i\in H^c}\sum_{\ell=1}^{d}\varpi_\ell\left\{\left|\frac{\sqrt{q}\left[\mu_{i,\ell,T}/(q\tau_1)\right]}{\sqrt{\pi_{i,\ell,T}/\left(q\tau_1^2\right)}}\right|\frac{1}{1+\max_{k\in H^c}\sqrt{\left|\overline{V}_{k,\ell,T}-\pi_{k,\ell,T}\right|/\pi_{k,\ell,T}}}\right.\right.
\end{aligned}
$$

$$\times \left[1 - \max_{k\in H^c}\left|\frac{\overline{S}_{k,\ell,T}-\mu_{k,\ell,T}}{\mu_{k,\ell,T}}\right|\right]\right\} \geq \Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\Bigg)$$

$\Big($making use of the inequality $\left|\sqrt{x}-\sqrt{y}\right| \leq \sqrt{|x-y|}$ for $x \geq 0$ and $y \geq 0$ $\Big)$

$$= P\left(\min_{i\in H^c}\sum_{\ell=1}^{d}\varpi_\ell\left\{\left|\frac{\sqrt{q}\left[\mu_{i,\ell,T}/(q\tau_1)\right]}{\sqrt{\pi_{i,\ell,T}/\left(q\tau_1^2\right)}}\right|\frac{1-\max_{k\in H^c}\left|\mathcal{E}_{k,\ell,T}\right|}{1+\max_{k\in H^c}\sqrt{\left|\mathcal{V}_{k,\ell,T}\right|}}\right\} \geq \Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right),$$

where $\mathcal{E}_{k,\ell,T} = \left(\overline{S}_{k,\ell,T}-\mu_{k,\ell,T}\right)/\mu_{k,\ell,T}$ and $\mathcal{V}_{k,\ell,T} = \left(\overline{V}_{k,\ell,T}-\pi_{k,\ell,T}\right)/\pi_{k,\ell,T}$. By part (a) of Lemma QA-16 (given in the Online Appendix, Chao and Swanson, 2022), there exists a sequence of positive numbers $\{\epsilon_T\}$ such that, as $T \to \infty$, $\epsilon_T \to 0$ and $P\left(\max_{1\leq\ell\leq d}\max_{k\in H^c}\left|\mathcal{E}_{k,\ell,T}\right| \geq \epsilon_T\right) \to 0$. In addition, by the result of part (b) of Lemma QA-16, there exists a sequence of positive numbers $\{\epsilon_T^*\}$ such that, as $T\to\infty$, $\epsilon_T^* \to 0$ and $P\left(\max_{1\leq\ell\leq d}\max_{k\in H^c}\left|\mathcal{V}_{k,\ell,T}\right| \geq \epsilon_T^*\right) \to 0$. Further define $\overline{\mathbb{E}}_T = \max_{1\leq\ell\leq d}\max_{k\in H^c}\left|\mathcal{E}_{k,\ell,T}\right|$ and $\overline{\mathbb{V}}_T = \max_{1\leq\ell\leq d}\max_{k\in H^c}\left|\mathcal{V}_{k,\ell,T}\right|$; and note that, for all N_1, N_2, and T sufficiently large,

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

$$\geq P\left(\min_{i\in H^c}\sum_{\ell=1}^{d}\varpi_\ell\left\{\left|\frac{\sqrt{q}\left[\mu_{i,\ell,T}/(q\tau_1)\right]}{\sqrt{\pi_{i,\ell,T}/\left(q\tau_1^2\right)}}\right|\frac{1-\max_{k\in H^c}\left|\mathcal{E}_{k,\ell,T}\right|}{1+\max_{k\in H^c}\sqrt{\left|\mathcal{V}_{k,\ell,T}\right|}}\right\} \geq \Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right)$$

$$\geq P\left(\frac{1-\max_{1\leq\ell\leq d}\max_{k\in H^c}\left|\mathcal{E}_{k,\ell,T}\right|}{1+\max_{1\leq\ell\leq d}\max_{k\in H^c}\sqrt{\left|\mathcal{V}_{k,\ell,T}\right|}}\min_{i\in H^c}\sum_{\ell=1}^{d}\varpi_\ell\left|\frac{\sqrt{q}\left[\mu_{i,\ell,T}/(q\tau_1)\right]}{\sqrt{\pi_{i,\ell,T}/\left(q\tau_1^2\right)}}\right| \geq \Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right)$$

$$= P\left(\frac{1-\overline{\mathbb{E}}_T}{1+\sqrt{\overline{\mathbb{V}}_T}}\min_{i\in H^c}\sum_{\ell=1}^{d}\varpi_\ell\left|\frac{\mu_{i,\ell,T}}{\sqrt{\pi_{i,\ell,T}}}\right| \geq \Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right)$$

$$\geq P\left(\left\{\left|\frac{1-\epsilon_T}{1+\sqrt{\epsilon_T^*}}\right|\min_{i\in H^c}\sum_{\ell=1}^{d}\varpi_\ell\left|\frac{\mu_{i,\ell,T}}{\sqrt{\pi_{i,\ell,T}}}\right| \geq \Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right\}\cap\left\{\overline{\mathbb{E}}_T < \epsilon_T\right\}\cap\left\{\overline{\mathbb{V}}_T < \epsilon_T^*\right\}\right)$$

$$+P\left(\left\{\frac{1-\overline{\mathbb{E}}_T}{1+\sqrt{\overline{\mathbb{V}}_T}}\min_{i\in H}\sum_{\ell=1}^{d}\varpi_\ell\left|\frac{\mu_{i,\ell,T}}{\sqrt{\pi_{i,\ell,T}}}\right| \geq \Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right\}\cap\left\{\overline{\mathbb{E}}_T \geq \epsilon_T \cup \overline{\mathbb{V}}_T \geq \epsilon_T^*\right\}\right)$$

$$\geq P\left(\left\{\left|\frac{1-\epsilon_T}{1+\sqrt{\epsilon_T^*}}\right|\min_{i\in H^c}\sum_{\ell=1}^{d}\varpi_\ell\left|\frac{\mu_{i,\ell,T}}{\sqrt{\pi_{i,\ell,T}}}\right| \geq \Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right\}\cap\left\{\overline{\mathbb{E}}_T < \epsilon_T\right\}\cap\left\{\overline{\mathbb{V}}_T < \epsilon_T^*\right\}\right)$$

$$+P\left(\left\{\frac{1-\overline{\mathbb{E}}_T}{1+\sqrt{\overline{\mathbb{V}}_T}}\min_{i\in H}\sum_{\ell=1}^{d}\varpi_\ell\left|\frac{\mu_{i,\ell,T}}{\sqrt{\pi_{i,\ell,T}}}\right| \geq \Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right\}\cap\left\{\overline{\mathbb{E}}_T \geq \epsilon_T\right\}\right)$$

$$= P\left(\left\{\left|\frac{1-\epsilon_T}{1+\sqrt{\epsilon_T^*}}\right|\min_{i\in H^c}\sum_{\ell=1}^{d}\varpi_\ell\left|\frac{\mu_{i,\ell,T}}{\sqrt{\pi_{i,\ell,T}}}\right| \geq \Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right\}\cap\left\{\overline{\mathbb{E}}_T < \epsilon_T\right\}\cap\left\{\overline{\mathbb{V}}_T < \epsilon_T^*\right\}\right)$$

$$+o(1).$$

where the last equality above follows from the fact that

$$P\left(\left\{\frac{1-\overline{\mathbb{E}}_T}{1+\sqrt{\overline{\mathbb{V}}_T}}\min_{i\in H}\sum_{\ell=1}^{d}\varpi_\ell\left|\frac{\mu_{i,\ell,T}}{\sqrt{\pi_{i,\ell,T}}}\right| \geq \Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right\}\cap\left\{\overline{\mathbb{E}}_T \geq \epsilon_T\right\}\right)$$

$$\leq P\left(\overline{\mathbb{E}}_T \geq \epsilon_T\right) = o(1)$$

Moreover, making use of Assumption 2-8, the result given in Lemma A1, and the fact that $q = \lfloor T_0/\tau \rfloor \sim T^{1-\alpha_1}$, we see that, there exists positive constants $\underline{c}$ and $\overline{C}$ such that:
$\min_{i\in H^c} \sum_{\ell=1}^{d} \varpi_\ell \left| \frac{\mu_{i,\ell,T}}{\sqrt{\pi_{i,\ell,T}}} \right| = \min_{i\in H^c} \sum_{\ell=1}^{d} \varpi_\ell \frac{\sqrt{q}\left|\mu_{i,\ell,T}/(q\tau_1)\right|}{\sqrt{\pi_{i,\ell,T}/\left(q\tau_1^2\right)}}$
$\geq \sqrt{q} \sum_{\ell=1}^{d} \varpi_\ell \frac{\min_{i\in H^c}\left|\mu_{i,\ell,T}/(q\tau_1)\right|}{\sqrt{\max_{i\in H^c} \pi_{i,\ell,T}/\left(q\tau_1^2\right)}} \geq \sqrt{q} \sum_{\ell=1}^{d} \varpi_\ell \frac{\underline{c}}{\sqrt{\overline{C}}} = \sqrt{q} \frac{\underline{c}}{\sqrt{\overline{C}}} \sim \sqrt{q} \sim \sqrt{\frac{T_0}{\tau}} \sim T^{(1-\alpha_1)/2}$.
On the other hand, applying the inequality
$\Phi^{-1}\left(1 - \frac{\varphi}{2N}\right) \leq \sqrt{2\left(1+a\right)}\sqrt{\ln N} \sim \sqrt{\ln N}$,[13] we further deduce that,
as $N_1, N_2, T \to \infty$,

$$\frac{1}{\Phi^{-1}\left(1-\frac{\varphi}{2N}\right)} \min_{i\in H^c} \sum_{\ell=1}^{d} \varpi_\ell \left| \frac{\mu_{i,\ell,T}}{\sqrt{\pi_{i,\ell,T}}} \right| \geq \frac{\underline{c}}{\sqrt{\overline{C}}} \sqrt{\frac{q}{2\left(1+a\right)\ln N}} \sim \sqrt{\frac{T^{(1-\alpha_1)}}{\ln N}} \to \infty.$$

This is true because the condition $\sqrt{\ln N}/\min\left\{T^{(1-\alpha_1)/6}, T^{\alpha_2/2}\right\} \to 0$ as $N_1, N_2, T \to \infty$ (as specified in Assumption 2-9 part (a)) implies that $\ln N/T^{(1-\alpha_1)} \to 0$ as $N_1, N_2, T \to \infty$. Hence, there exists a natural number M such that, for all $N_1 \geq M, N_2 \geq M$, and $T \geq M$, we have $\left|\frac{1-\epsilon_T}{1+\sqrt{\epsilon_T^*}}\right| \min_{i\in H^c} \sum_{\ell=1}^{d} \varpi_\ell \left|\frac{\mu_{i,\ell,T}}{\sqrt{\pi_{i,\ell,T}}}\right| \geq \Phi^{-1}\left(1 - \frac{\varphi}{2N}\right)$ so that:

$$\begin{aligned}
&P\left(\min_{i\in H^c} \sum_{\ell=1}^{d} \varpi_\ell \left|S_{i,\ell,T}\right| \geq \Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right) \\
&\geq P\left(\left\{\left|\frac{1-\epsilon_T}{1+\sqrt{\epsilon_T^*}}\right| \min_{i\in H^c} \sum_{\ell=1}^{d} \varpi_\ell \left|\frac{\mu_{i,\ell,T}}{\sqrt{\pi_{i,\ell,T}}}\right| \geq \Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right\} \cap \left\{\overline{\mathbb{E}}_T < \epsilon_T\right\} \cap \left\{\overline{\mathbb{V}}_T < \epsilon_T^*\right\}\right) \\
&\quad + o\left(1\right) \\
&= P\left(\left\{\overline{\mathbb{E}}_T < \epsilon_T\right\} \cap \left\{\overline{\mathbb{V}}_T < \epsilon_T^*\right\}\right) + o\left(1\right) \\
&\quad \text{(for all } N_1 \geq M, N_2 \geq M, \text{and } T \geq M) \\
&\geq P\left(\overline{\mathbb{E}}_T < \epsilon_T\right) + P\left(\overline{\mathbb{V}}_T < \epsilon_T^*\right) - 1 + o\left(1\right) \text{ (using the inequality} \\
&\quad P\left\{\bigcap_{i=1}^{m} A_i\right\} \geq \sum_{i=1}^{m} P\left(A_i\right) - \left(m-1\right) \text{ in Chao and Swanson (2022) Lemma OA-14}\Big) \\
&= 1 - P\left(\overline{\mathbb{E}}_T \geq \epsilon_T\right) + 1 - P\left(\overline{\mathbb{V}}_T \geq \epsilon_T^*\right) - 1 + o\left(1\right) \\
&= 1 - P\left(\overline{\mathbb{E}}_T \geq \epsilon_T\right) - P\left(\overline{\mathbb{V}}_T \geq \epsilon_T^*\right) + o\left(1\right) \\
&= 1 + o\left(1\right).
\end{aligned}$$

Next, to show part (b), note that, by applying the result in part (a), we have that:
$P\left(\min_{i\in H^c} \max_{1\leq \ell\leq d} \left|S_{i,\ell,T}\right| \geq \Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right)$

[13] As noted previously, an explicit proof of this result is given in Chao and Swanson (2022). In particular, this inequality is shown in part (b) of Lemma QA-15 in Chao and Swanson (2022).

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

Lemma A1: Let $\underline{Y}_t=\begin{pmatrix}Y_t' & Y_{t-1}' & \cdots & Y_{t-p+1}'\end{pmatrix}'$ and $\underline{F}_t=\begin{pmatrix}F_t' & F_{t-1}' & \cdots & F_{t-p+1}'\end{pmatrix}'$, and define $b_1\left(r\right)=\left(r-1\right)\tau+p$ and $b_2\left(r\right)=b_1\left(r\right)+\tau_1-1$. Under Assumptions 2-1, 2-2, 2-5, 2-6, and 2-9(b); there exists a positive constant C such that:

$$\max_{1\leq\ell\leq d,\ i\in H^c}\left(\frac{\pi_{i,\ell,T}}{q\tau_1^2}\right)$$
$$=\max_{1\leq\ell\leq d,\ i\in H^c}\frac{1}{q}\sum_{r=1}^{q}\left(\frac{1}{\tau_1}\sum_{t=b_1(r)}^{b_2(r)}\gamma_i'\left\{E\left[\underline{F}_t\right]\mu_{Y,\ell}+E\left[\underline{F}_t\underline{Y}_t'\right]\alpha_{YY,\ell}+E\left[\underline{F}_t\underline{F}_t'\right]\alpha_{YF,\ell}\right\}\right)^2$$
$\leq C<\infty$, for all N_1,N_2,T sufficiently large.

Proof of Lemma A1: To proceed, let $\phi_{\max}=\max\left\{\left|\lambda_{\max}\left(A\right)\right|,\left|\lambda_{\min}\left(A\right)\right|\right\}$ and, for $\ell\in\{1,...,d\}$, let $e_{\ell,d}$ denote a $d\times1$ elementary vector whose ℓ^{th} component is 1 and all other components are 0. Now, note that:

$$\max_{1\leq\ell\leq d,\ i\in H^c}\left\{\pi_{i,\ell,T}/\left(q\tau_1^2\right)\right\}$$
$$=\max_{1\leq\ell\leq d,\ i\in H^c}\frac{1}{q}\sum_{r=1}^{q}\left(\frac{1}{\tau_1}\sum_{t=b_1(r)}^{b_2(r)}\gamma_i'\left\{E\left[\underline{F}_t\right]\mu_{Y,\ell}+E\left[\underline{F}_t\underline{Y}_t'\right]\alpha_{YY,\ell}+E\left[\underline{F}_t\underline{F}_t'\right]\alpha_{YF,\ell}\right\}\right)^2$$
$$\leq\max_{1\leq\ell\leq d,\ i\in H^c}\frac{1}{q}\sum_{r=1}^{q}\left(\frac{1}{\tau_1}\sum_{t=b_1(r)}^{b_2(r)}\left\{E\left[\left|\gamma_i'\underline{F}_t\right|\right]\left|\mu_{Y,\ell}\right|+E\left[\left|\gamma_i'\underline{F}_t\underline{Y}_t'A_{YY}'e_{\ell,d}\right|\right]\right.\right.$$
$$\left.\left.+E\left[\left|\gamma_i'\underline{F}_t\underline{F}_t'A_{YF}'e_{\ell,d}\right|\right]\right\}\right)^2\text{ (by triangle and Jensen's inequalities)}$$
$$\leq\max_{i\in H^c}\frac{1}{q}\sum_{r=1}^{q}\left(\frac{1}{\tau_1}\sum_{t=b_1(r)}^{b_2(r)}\left\{\sqrt{\left\|\gamma_i\right\|_2^2}\sqrt{E\left\|\underline{F}_t\right\|_2^2}\max_{1\leq\ell\leq d}\left|\mu_{Y,\ell}\right|\right.\right.$$
$$+\sqrt{\gamma_i'E\left[\underline{F}_t\underline{F}_t'\right]\gamma_i}\sqrt{\max_{1\leq\ell\leq d}e_{\ell,d}'A_{YY}E\left[\underline{Y}_t\underline{Y}_t'\right]A_{YY}'e_{\ell,d}}$$
$$\left.\left.+\sqrt{\gamma_i'E\left[\underline{F}_t\underline{F}_t'\right]\gamma_i}\sqrt{\max_{1\leq\ell\leq d}e_{\ell,d}'A_{YF}E\left[\underline{F}_t\underline{F}_t'\right]A_{YF}'e_{\ell,d}}\right\}\right)^2$$
$$\leq\left(\max_{i\in H^c}\left\|\gamma_i\right\|_2^2\right)\frac{1}{q}\sum_{r=1}^{q}\left(\frac{1}{\tau_1}\sum_{t=b_1(r)}^{b_2(r)}\left\{\sqrt{E\left\|\underline{F}_t\right\|_2^2}\max_{1\leq\ell\leq d}\left|\mu_{Y,\ell}\right|\right.\right.$$
$$+\sqrt{E\left\|\underline{F}_t\right\|_2^2}\sqrt{E\left\|\underline{Y}_t\right\|_2^2}\sqrt{\max_{1\leq\ell\leq d}e_{\ell,d}'A_{YY}A_{YY}'e_{\ell,d}}$$
$$\left.\left.+E\left\|\underline{F}_t\right\|_2^2\sqrt{\max_{1\leq\ell\leq d}e_{\ell,d}'A_{YF}A_{YF}'e_{\ell,d}}\right\}\right)^2$$
$$\leq\left(\max_{i\in H^c}\left\|\gamma_i\right\|_2^2\right)\frac{1}{q}\sum_{r=1}^{q}\left(\frac{1}{\tau_1}\sum_{t=b_1(r)}^{b_2(r)}\left\{\sqrt{E\left\|\underline{F}_t\right\|_2^2}\max_{1\leq\ell\leq d}\left|\mu_{Y,\ell}\right|\right.\right.$$
$$\left.\left.+\sqrt{E\left\|\underline{F}_t\right\|_2^2}\sqrt{E\left\|\underline{Y}_t\right\|_2^2}C^{\dagger}\phi_{\max}+E\left\|\underline{F}_t\right\|_2^2C^{\dagger}\phi_{\max}\right\}\right)^2,$$

where the last inequality follows from the fact that, by making use of Assumption 2-6, it is easy to show that there exists a constant $C^{\dagger}>0$ such that

$\sqrt{\max_{1\leq\ell\leq d} e_{\ell,d}' A_{YY} A_{YY}' e_{\ell,d}} \leq \|A_{YY}\|_2 \sqrt{\max_{1\leq\ell\leq d} e_{\ell,d}' e_{\ell,d}} = \|A_{YY}\|_2 \leq C^{\dagger}\phi_{\max}$ and, similarly, $\sqrt{\max_{1\leq\ell\leq d} e_{\ell,d}' A_{YF} A_{YF}' e_{\ell,d}} \leq \|A_{YF}\|_2 \leq C^{\dagger}\phi_{\max}$.[14] Hence,

$$
\begin{aligned}
&\max_{1\leq\ell\leq d} \max_{k\in H^c} \left\{\pi_{i,\ell,T}/\left(q\tau_1^2\right)\right\} \\
&\leq \left(\max_{i\in H^c} \|\gamma_i\|_2^2\right) \frac{1}{q}\sum_{r=1}^{q}\left(\frac{1}{\tau_1}\sum_{t=b_1(r)}^{b_2(r)}\left\{\sqrt{E\|\underline{F}_t\|_2^2}\max_{1\leq\ell\leq d}\left|\mu_{Y,\ell}\right|\right.\right. \\
&\quad \left.\left. +\sqrt{E\|\underline{F}_t\|_2^2}\sqrt{E\|\underline{Y}_t\|_2^2}C^{\dagger}\phi_{\max} + E\|\underline{F}_t\|_2^2 C^{\dagger}\phi_{\max}\right\}\right)^2 \\
&\leq \left(\max_{i\in H^c} \|\gamma_i\|_2^2\right) \frac{1}{q}\sum_{r=1}^{q}\frac{1}{\tau_1}\sum_{t=b_1(r)}^{b_2(r)} E\|\underline{F}_t\|_2^2\left(\|\mu_Y\|_2^2 + \left[\sqrt{E\|\underline{Y}_t\|_2^2} + \sqrt{E\|\underline{F}_t\|_2^2}\right]C^{\dagger}\phi_{\max}\right)^2 \\
&\leq C < \infty,
\end{aligned}
$$

for some positive constant C such that $C \geq \left(\max_{i\in H^c} \|\gamma_i\|_2^2\right) E\|\underline{F}_t\|_2^2\left(\|\mu_Y\|_2^2 + \left[\sqrt{E\|\underline{Y}_t\|_2^2} + \sqrt{E\|\underline{F}_t\|_2^2}\right]C^{\dagger}\phi_{\max}\right)^2$, where such a constant exists because $\max_{i\in H^c}\|\gamma_i\|_2^2$ and $\|\mu_Y\|_2^2$ are both bounded given Assumption 2-5; because $0 < \phi_{\max} < 1$ given Assumption 2-1; and because, under Assumptions 2-1, 2-2(a)-(b), 2-5, and 2-6; one can easily show that there exists a constant $C^* > 0$ such that $E\|\underline{F}_t\|_2^2 \leq C^*$ and $E\|\underline{Y}_t\|_2^2 \leq \left(E\|\underline{Y}_t\|_2^6\right)^{1/3} \leq C^*$.[15] $\square$

Lemma A2: Suppose that Assumptions 2-1, 2-2, 2-3, 2-4, 2-5, 2-6, and 2-7 hold. Let $\Phi(\cdot)$ denote the cumulative distribution function of the standard normal random variable. Then, there exists a positive constant A such that

$$
P\left(\left|S_{i,\ell,T}\right| \geq z\right) \leq 2\left[1-\Phi(z)\right]\left\{1 + A(1+z)^3 T^{-(1-\alpha_1)\frac{1}{2}}\right\} \tag{21}
$$

for $i \in H = \{k \in \{1, ..., N\} : \gamma_k = 0\}$, for $\ell \in \{1, ..., d\}$, for T sufficiently large, and for all z such that $0 \leq z \leq c_0 \min\left\{T^{(1-\alpha_1)/6}, T^{\alpha_2/2}\right\}$ with c_0 being a positive constant.

Proof of Lemma A2: Note first that, for any i such that $i \in H = \{k \in \{1, ..., N\} : \gamma_k = 0\}$, the formula for $S_{i,\ell,T}$ reduces to:

$$
S_{i,\ell,T} = \left(\sum\nolimits_{r=1}^{q}\left[\sum\nolimits_{t=(r-1)\tau+p}^{(r-1)\tau+\tau_1+p-1} y_{\ell,t+1}u_{it}\right]^2\right)^{-\frac{1}{2}} \sum\nolimits_{r=1}^{q}\sum\nolimits_{t=(r-1)\tau+p}^{(r-1)\tau+\tau_1+p-1} y_{\ell,t+1}u_{it}.
$$

Hence, to verify the conditions of Theorem 4.1 of Chen, Shao, Wu, and Xu (2016), we set $X_{it} = u_{it}y_{\ell,t+1}$, and note that $E\left[X_{it}\right] = E\left[u_{it}y_{\ell,t+1}\right] = E_Y\left[E\left[u_{it}\right]y_{\ell,t+1}\right] = 0$, where the

[14] Explicit proofs of these two inequalities are given in Chao and Swanson (2022). In particular, these inequalities are shown in parts (a) and (b) of Lemma OA-7 in Chao and Swanson (2022).

[15] An explicit proof that, under Assumptions 2-1, 2-2(a)-(b), 2-5, and 2-6; there exists some positive constant $C^{\#}$ such that $E\|\underline{F}_t\|_2^6 \leq C^{\#}$ and $E\|\underline{Y}_t\|_2^6 \leq C^{\#}$ is given in Chao and Swanson (2022). See Lemma OA-5 in Chao and Swanson (2022).

second equality follows by the law of iterated expectations given that Assumption 2-4 implies the independence of u_{it} and $y_{\ell,t+1}$ and where the third equality follows by Assumption 2-3(a). Hence, the first part of condition (4.1) of Chen, Shao, Wu, and Xu (2016) is fulfilled. Moreover, in light of Assumption 2-3(b) and in light of the fact that, under Assumptions 2-1, 2-2(a)-(b), 2-5, and 2-6; one can show by straightforward calculations that there exists a positive constant $\overline{C}$ such that $E\left\|\underline{Y}_t\right\|_2^6 \leq \overline{C}$; we see that there exists some positive constant c_1 such that, for every $\ell \in \{1..., d\}$,

$$\begin{aligned} E\left[\left|X_{it}\right|^{\frac{31}{10}}\right] &= E\left[\left|u_{it}y_{\ell,t+1}\right|^{\frac{31}{10}}\right] \leq \left(E\left|u_{it}\right|^{\frac{186}{29}}\right)^{\frac{29}{60}}\left(E\left|y_{\ell,t+1}\right|^{6}\right)^{\frac{31}{60}} \\ &\leq \left[\left(E\left|u_{it}\right|^{\frac{186}{29}}\right)^{\frac{29}{186}}\right]^{\frac{31}{10}}\left[E\left(\sum_{k=1}^{d}\sum_{j=0}^{p-1}y_{k,t+1-j}^{2}\right)^{3}\right]^{\frac{31}{60}} \\ &\leq \left[\left(E\left|u_{it}\right|^{7}\right)^{\frac{1}{7}}\right]^{\frac{31}{10}}\left[\left(E\left\|\underline{Y}_{t+1}\right\|_2^6\right)^{\frac{1}{6}}\right]^{\frac{31}{10}} \leq c_1^{\frac{31}{10}}, \end{aligned}$$

where the first and third inequalities above follow, respectively, by Hölder's and Liapunov's inequalities. Hence, the second part of condition (4.1) of Chen, Shao, Wu, and Xu (2016) is also fulfilled with $r = \frac{31}{10} > 2$. Moreover, note that, by Assumption 2-7, for all $r \geq 1$ and $\tau_1 \geq 1$:

$$E\left\{\left[\sum_{t=(r-1)\tau+p}^{(r-1)\tau+\tau_1+p-1} X_{it}\right]^2\right\} = \tau_1 E\left\{\left[\frac{1}{\sqrt{\tau_1}}\sum_{t=(r-1)\tau+p}^{(r-1)\tau+\tau_1+p-1} y_{\ell,t+1}u_{it}\right]^2\right\} \geq \tau_1 \underline{c},$$

so that condition (4.2) of Chen, Shao, Wu, and Xu (2016) is satisfied here. Now, making use of Assumption 2-3(c) and Assumption 2-4 and applying Theorem 2.1 of Pham and Tran (1985), it can be shown that $\left\{(y_{\ell,t+1}, u_{it})'\right\}$ is β mixing with β mixing coefficient satisfying $\beta(m) \leq \overline{a}_1 \exp\{-a_2 m\}$ for some constants $\overline{a}_1 > 0$ and $a_2 > 0$. Next, define $X_{it} = y_{\ell,t+1}u_{it}$, and note that $\{X_{it}\}$ is a β-mixing process with β-mixing coefficient $\beta_{X,m}$ satisfying the condition $\beta_{X,m} \leq a_1 \exp\{-a_2 m\}$ for some constant $a_1 > 0$ and for all m sufficiently large, given that measurable functions of a finite number of β-mixing random variables are also β-mixing, with β-mixing coefficients having the same order of magnitude[16]. It follows that $\{X_{it}\}$ satisfies the β mixing condition (2.1) stipulated in Chen, Shao, Wu, and Xu (2016) for all $i \in H$. Hence, by applying Theorem 4.1 of Chen, Shao, Wu, and Xu (2016) for the

[16] For α-mixing and ϕ-mixing, this result is given in Theorem 14.1 of Davidson (1994). However, using essentially the same argument as that given in the proof of Theorem 14.1, one can also prove a similar result for β-mixing. For an explicit proof of this result, see Lemma OA-2 part (a) in Chao and Swanson (2022).

case where $\delta = 1$[17], we obtain the Cramér-type moderate deviation result

$$\frac{P\left\{\overline{S}_{i,\ell,T}/\sqrt{\overline{V}_{i,\ell,T}} \geq z\right\}}{1-\Phi(z)} = 1 + O(1)(1+z)^3 T^{-(1-\alpha_1)\frac{1}{2}}, \tag{22}$$

which holds for all $0 \leq z \leq c_0 \min\left\{T^{(1-\alpha_1)/6}, T^{\alpha_2/2}\right\}$ and for $|O(1)| \leq A$, where A is an absolute constant and where $\overline{S}_{i,\ell,T}$ and $\overline{V}_{i,\ell,T}$ are as defined in expression (14).

Next, consider obtaining a moderate deviation result for $P\left\{-\overline{S}_{i,\ell,T}/\sqrt{\overline{V}_{i,\ell,T}} \geq z\right\}/[1-\Phi(z)]$. As $\overline{S}_{i,\ell,T} = \sum_{r=1}^{q}\sum_{t=(r-1)\tau+p}^{(r-1)\tau+\tau_1+p-1}(-u_{it}y_{\ell,t+1})$, we can take $X_{it} = -u_{it}y_{\ell,t+1}$, and note that, by calculations similar to those given above, we have $E[X_{it}] = E[-u_{it}y_{\ell,t+1}] = 0$, $E\left[|X_{it}|^{\frac{31}{10}}\right] = E\left[|-u_{it}y_{\ell,t+1}|^{\frac{31}{10}}\right] = E\left[|u_{it}y_{\ell,t+1}|^{\frac{31}{10}}\right] \leq c_1^{\frac{31}{10}}$, and

$$E\left\{\left[\sum_{t=(r-1)\tau+p}^{(r-1)\tau+\tau_1+p-1} X_{it}\right]^2\right\} = E\left\{\left[\sum_{t=(r-1)\tau+p}^{(r-1)\tau+\tau_1+p-1}(-u_{it}y_{\ell,t+1})\right]^2\right\} \geq \underline{c}\tau_1.$$

Moreover, it is easily seen that $\{X_{it}\}$ (with $X_{it} = -u_{it}y_{\ell,t+1}$) also satisfies the β mixing condition (2.1) stipulated in Chen, Shao, Wu, and Xu (2016) for every i. Thus, by applying Theorem 4.1 of Chen, Shao, Wu, and Xu (2016), we also obtain the Cramér-type moderate deviation result

$$\frac{P\left\{-\overline{S}_{i,\ell,T}/\sqrt{\overline{V}_{i,\ell,T}} \geq z\right\}}{1-\Phi(z)} = 1 + O(1)(1+z)^3 T^{-(1-\alpha_1)\frac{1}{2}}, \tag{23}$$

which holds for all $0 \leq z \leq c_0 \min\left\{T^{(1-\alpha_1)/6}, T^{\alpha_2/2}\right\}$ and for $|O(1)| \leq A$ with A being an absolute constant. Next, note that:

$$\begin{aligned}
\left|\frac{P\left(\left|S_{i,\ell,T}\right| \geq z\right)}{2[1-\Phi(z)]} - 1\right| &= \left|\frac{P\left(\left|\overline{S}_{i,\ell,T}/\sqrt{\overline{V}_{i,\ell,T}}\right| \geq z\right)}{2[1-\Phi(z)]} - 1\right| \\
&= \left|\frac{P\left(\left\{\overline{S}_{i,\ell,T}/\sqrt{\overline{V}_{i,\ell,T}} \geq z\right\} \cup \left\{-\overline{S}_{i,\ell,T}/\sqrt{\overline{V}_{i,\ell,T}} \geq z\right\}\right)}{2[1-\Phi(z)]} - 1\right| \\
&= \left|\frac{P\left(\overline{S}_{i,\ell,T}/\sqrt{\overline{V}_{i,\ell,T}} \geq z\right) + P\left(-\overline{S}_{i,\ell,T}/\sqrt{\overline{V}_{i,\ell,T}} \geq z\right)}{2[1-\Phi(z)]} - 1\right| \\
&\quad \left(\text{since } \left\{\overline{S}_{i,\ell,T}/\sqrt{\overline{V}_{i,\ell,T}} \geq z\right\} \cap \left\{-\overline{S}_{i,\ell,T}/\sqrt{\overline{V}_{i,\ell,T}} \geq z\right\} = \varnothing \text{ w.p.1}\right)
\end{aligned}$$

[17] Note that Theorem 4.1 of Chen, Shao, Wu and Xu (2016) requires that $0 < \delta \leq 1$ and $\delta < r - 2$. These conditions are satisfied here given that we choose $\delta = 1$ and $r = 31/10$.

$$\leq \tfrac{1}{2}\left|\frac{P\left(\overline{S}_{i,\ell,T}/\sqrt{\overline{V}_{i,\ell,T}}\geq z\right)}{1-\Phi(z)}-1\right|+\tfrac{1}{2}\left|\frac{P\left\{-\overline{S}_{i,\ell,T}/\sqrt{\overline{V}_{i,\ell,T}}\geq z\right\}}{1-\Phi(z)}-1\right|.$$

Thus, in light of expressions (22) and (23), we have that:

$$\begin{aligned}
&\left|\frac{P\left(\left|S_{i,\ell,T}\right|\geq z\right)}{2\left[1-\Phi\left(z\right)\right]}-1\right| \\
\leq\ & \frac{1}{2}\left|\frac{P\left(\overline{S}_{i,\ell,T}/\sqrt{\overline{V}_{i,\ell,T}}\geq z\right)}{1-\Phi\left(z\right)}-1\right|+\frac{1}{2}\left|\frac{P\left\{-\overline{S}_{i,\ell,T}/\sqrt{\overline{V}_{i,\ell,T}}\geq z\right\}}{1-\Phi\left(z\right)}-1\right| \\
\leq\ & \frac{A}{2}\left(1+z\right)^{3}T^{-(1-\alpha_1)\frac{1}{2}}+\frac{A}{2}\left(1+z\right)^{3}T^{-(1-\alpha_1)\frac{1}{2}}=A\left(1+z\right)^{3}T^{-(1-\alpha_1)\frac{1}{2}}
\end{aligned}$$

It then follows that:

$$-A\left(1+z\right)^{3}T^{-(1-\alpha_1)\frac{1}{2}}\leq\frac{P\left(\left|S_{i,\ell,T}\right|\geq z\right)}{2\left[1-\Phi\left(z\right)\right]}-1\leq A\left(1+z\right)^{3}T^{-(1-\alpha_1)\frac{1}{2}} \tag{24}$$

where $S_{i,\ell,T}=\overline{S}_{i,\ell,T}/\sqrt{\overline{V}_{i,\ell,T}}$. Focusing on the right-hand part of the inequality in (24), we have that:
$P\left(\left|S_{i,\ell,T}\right|\geq z\right)/\left(2\left[1-\Phi\left(z\right)\right]\right)-1\ \leq\ A\left(1+z\right)^{3}T^{-(1-\alpha_1)\frac{1}{2}}$. Simple rearrangement of this inequality then leads to the desired result:

$$P\left(\left|S_{i,\ell,T}\right|\geq z\right)\leq 2\left[1-\Phi\left(z\right)\right]\left\{1+A\left(1+z\right)^{3}T^{-(1-\alpha_1)\frac{1}{2}}\right\},$$

which holds for all $i\in H=\{k\in\{1,....,N\}:\gamma_k=0\}$, for every $\ell\in\{1,...,d\}$, for all T sufficiently large, and for all z such that $0\leq z\leq c_0\min\left\{T^{(1-\alpha_1)/6},T^{\alpha_2/2}\right\}$. $\square$

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

		$N = 100$	$N_1 = 50$	$T = 100$	$\tau = 5$		
		$\varphi = N^{-0.2}$	$\varphi = N^{-0.3}$	$\varphi = N^{-0.4}$	$\varphi = N^{-0.5}$	$\varphi = N^{-0.6}$	$\varphi = N^{-0.7}$
$\tau_1 = 2$	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: Monte Carlo Experiments - $\mathbb{S}^{+}_{i,T} = \sum_{\ell=1}^{d}\varpi_\ell |S_{i,\ell,T}|$

		$N = 100$	$N_1 = 50$	$T = 100$	$\tau = 5$		
		$\varphi = N^{-0.2}$	$\varphi = N^{-0.3}$	$\varphi = N^{-0.4}$	$\varphi = N^{-0.5}$	$\varphi = N^{-0.6}$	$\varphi = N^{-0.7}$
$\tau_1 = 2$	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: Monte Carlo Experiments - $\mathbb{S}^{+}_{i,T} = \max_{1\leq\ell\leq d}|S_{i,\ell,T}|$

		$N = 200$	$N_1 = 100$	$T = 100$	$\tau = 5$		
		$\varphi = N^{-0.2}$	$\varphi = N^{-0.3}$	$\varphi = N^{-0.4}$	$\varphi = N^{-0.5}$	$\varphi = N^{-0.6}$	$\varphi = N^{-0.7}$
$\tau_1 = 2$	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: Monte Carlo Experiments - $\mathbb{S}_{i,T}^{+} = \sum_{\ell=1}^{d} \varpi_\ell |S_{i,\ell,T}|$

		$N = 200$	$N_1 = 100$	$T = 100$	$\tau = 5$		
		$\varphi = N^{-0.2}$	$\varphi = N^{-0.3}$	$\varphi = N^{-0.4}$	$\varphi = N^{-0.5}$	$\varphi = N^{-0.6}$	$\varphi = N^{-0.7}$
$\tau_1 = 2$	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: Monte Carlo Experiments - $\mathbb{S}_{i,T}^{+} = \max_{1 \leq \ell \leq d} |S_{i,\ell,T}|$

		$N = 400$	$N_1 = 200$	$T = 200$	$\tau = 10$		
		$\varphi = N^{-0.2}$	$\varphi = N^{-0.3}$	$\varphi = N^{-0.4}$	$\varphi = N^{-0.5}$	$\varphi = N^{-0.6}$	$\varphi = N^{-0.7}$
$\tau_1 = 5$	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: Monte Carlo Experiments - $\mathbb{S}_{i,T}^{+} = \sum_{\ell=1}^{d} \varpi_\ell |S_{i,\ell,T}|$

		$N = 400$	$N_1 = 200$	$T = 200$	$\tau = 10$		
		$\varphi = N^{-0.2}$	$\varphi = N^{-0.3}$	$\varphi = N^{-0.4}$	$\varphi = N^{-0.5}$	$\varphi = N^{-0.6}$	$\varphi = N^{-0.7}$
$\tau_1 = 5$	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: Monte Carlo Experiments - $\mathbb{S}^{+}_{i,T} = \max_{1 \le \ell \le d} |S_{i,\ell,T}|$

		$N = 1000$	$N_1 = 500$	$T = 600$	$\tau = 12$		
		$\varphi = N^{-0.2}$	$\varphi = N^{-0.3}$	$\varphi = N^{-0.4}$	$\varphi = N^{-0.5}$	$\varphi = N^{-0.6}$	$\varphi = N^{-0.7}$
$\tau_1 = 6$	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: Monte Carlo Experiments - $\mathbb{S}^{+}_{i,T} = \sum_{\ell=1}^{d} \varpi_\ell |S_{i,\ell,T}|$

		$N = 1000$	$N_1 = 500$	$T = 600$	$\tau = 12$		
		$\varphi = N^{-0.2}$	$\varphi = N^{-0.3}$	$\varphi = N^{-0.4}$	$\varphi = N^{-0.5}$	$\varphi = N^{-0.6}$	$\varphi = N^{-0.7}$
$\tau_1 = 6$	FPR	0.00038	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.

Table 9: Empirical Illustration - Target Forecast Variables*

Target Variable	Abbreviation	Data Transformation
Industrial Production	INDPRO	$\Delta log(y_t)$
Civilian Unemployment Rate	UNRATE	y_t
Housing Starts (new, privately owned)	HOUST	$log(y_t)$
Housing Permits (new, privately owned)	PERMIT	$log(y_t)$
Real M2 Money Stock	M2REAL	$\Delta log(y_t)$
10-Year Government Treasury Bond Rate	R10	y_t
CPI (all items)	CPI	$\Delta log(y_t)$
S&P Common Stock Price Index (composite)	S&P500	$\Delta log(y_t)$

* Notes: This table lists the target forecast variables that are predicted in our empirical illustration, and associated data transformations.

Table 10: Empirical Illustration - Real-Time Predictive Accuracy Experiments*

		Factor Estimation Method		
	Target Variable	Principal Components Analysis	t-Statistic Thresholding	CS Variable Selection
h=1	INDPRO	**0.971**	1.025	1.092
	UNRATE	1.098	**0.966**	0.984
	HOUST	1.000	0.882 **	**0.877**
	PERMIT	1.009	1.005	1.042
	M2REAL	1.06	1.052 *	1.19 *
	R10	1.086	1.107	1.051
	CPI	1.08	1.125	1.161
	S&P500	1.142	1.104	1.118
h=3	INDPRO	1.042	1.042	1.083
	UNRATE	0.839	**0.691 ****	0.785 *
	HOUST	0.979	0.777	**0.766** **
	PERMIT	1.021	0.971	**0.895**
	M2REAL	0.979	**0.907**	1.035
	R10	1.187	1.225	**1.109**
	CPI	1.022	1.048 *	**0.993**
	S&P500	1.186	1.213	1.055
h=6	INDPRO	1.145	1.043	**0.965 ***
	UNRATE	0.561 *	**0.518 ****	0.617 *
	HOUST	0.818	0.696 *	**0.665 ***
	PERMIT	0.855	0.825	**0.744**
	M2REAL	1.031	1.049	**0.994**
	R10	1.513	1.046	1.064 **
	CPI	1.034	1.05	1.081 *
	S&P500	1.126 *	1.276 **	1.166
h=12	INDPRO	1.251 *	1.096 *	1.004
	UNRATE	0.632	0.489	**0.471**
	HOUST	0.693	0.605	**0.605**
	PERMIT	0.681	0.650	**0.621**
	M2REAL	1.087	1.036	**0.984**
	R10	0.992	0.641	**0.638 ***
	CPI	1.024	1.131	1.064 **
	S&P500	1.188	1.311 *	1.046

* Notes: See notes to Table 9. Tabulated entries are relative mean squared forecast error (MSFEs) for our 8 target variables, for forecast horizons of h=1,3,6, and 12 months ahead. The AR(SIC) benchmark model is in the denominator of the reported MSFEs, so that entries that are less than unity indicate that our factor and variable augmented forecast regressions yield lower MSFEs than those associated with the AR(SIC) benchmark. The forecast period is 2000:1-2007:12, and all models are estimated prior to the construction of each monthly forecast. In cases where at least one big data method outperfroms the AR(SIC) benchmark model, entries in bold denote the big data method yielding the lowest relative MSFE for a given target variable and forecast horizon. Starred entries indicate rejection of the null hypothesis of equal conditional predictive ability, at significance levels $p = 0.01$ $(***)$, $p = 0.05$, $(**)$, and $p = 0.10$ $(*)$. See Section 4 for complete details.