Prediction When Factors are Weak[*]

Stefano Giglio[†]
Yale School of Management
NBER and CEPR

Dacheng Xiu[‡]
Booth School of Business
University of Chicago

Dake Zhang[§]
Booth School of Business
University of Chicago

Abstract

In macroeconomic forecasting, principal component analysis (PCA) has been the most prevalent approach to the recovery of factors, which summarize information in a large set of macro predictors. Nevertheless, the theoretical justification of this approach often relies on a convenient and critical assumption that factors are pervasive. To incorporate information from weaker factors, we propose a new prediction procedure based on supervised PCA, which iterates over selection, PCA, and projection. The selection step finds a subset of predictors most correlated with the prediction target, whereas the projection step permits multiple weak factors of distinct strength. We justify our procedure in an asymptotic scheme where both the sample size and the cross-sectional dimension increase at potentially different rates. Our empirical analysis highlights the role of weak factors in predicting inflation, industrial production growth, and changes in unemployment.

Key words: Supervised PCA, PCA, PLS, weak factors, marginal screening

[*] We benefited tremendously from discussions with seminar and conference participants at Gregory Chow Seminar Series in Econometrics and Statistics, Glasgow University, University of Science and Technology of China, Nankai University, IAAE Invited Session at ASSA 2023, Conference on Big Data and Machine Learning in Econometrics, Finance, and Statistics at the University of Chicago, and 14th Annual SoFiE Conference.

[†] Address: 165 Whitney Avenue, New Haven, CT 06520, USA. Email: stefano.giglio@yale.edu.

[‡] Address: 5807 S Woodlawn Avenue, Chicago, IL 60637, USA. Email: dacheng.xiu@chicagobooth.edu.

[§] Address: 5807 S Woodlawn Avenue, Chicago, IL 60637, USA. Email: dkzhang@chicagobooth.edu.

1 Introduction

Starting from the seminal contribution of Stock and Watson (2002), factor models have played a prominent role in macroeconomic forecasting. Principal component analysis (PCA), advocated in that paper, has been the most prevalent approach to the recovery of factors that summarize the information contained in a large set of macroeconomic predictors, and reduce the dimensionality of the forecasting problem.

The theoretical justification for the PCA approach to factor analysis often relies on a convenient – but critical – assumption that factors are pervasive (*strong*), see for example Bai and Ng (2002) and Bai (2003). In that case, the common components of predictors can be extracted consistently by PCA and separated from the idiosyncratic components. Recently, Bai and Ng (2021) relax this condition, showing that PCA can consistently recover the underlying factors under weaker assumptions.

Nevertheless, PCA is an unsupervised approach, and by its nature, this poses some limits to its ability to find the most useful low-dimensional predictors in a forecasting context. Specifically, if the signal-to-noise ratio is sufficiently low, the factor space spanned by the principal components is inconsistent, or even nearly orthogonal to the space spanned by true factors, see Hoyle and Rattray (2004) and Johnstone and Lu (2009). In such instances, we refer to the underlying factors as *weak*.

In this paper we study a setting in which factors are sufficiently weak that PCA fails to recover them. We propose a new approach to dimension reduction for forecasting, based on *supervised PCA* (SPCA). The key idea of supervised PCA is to select a subset of predictors that are correlated with the prediction target before applying PCA. The concept of supervised PCA originated from a cancer diagnosis technique applied to DNA microarray data by Bair and Tibshirani (2004), and was later formalized by Bair et al. (2006) in a prediction framework, in which some predictors are not correlated with the latent factors that drive the outcome of interest. Bai and Ng (2008) generalize this selection procedure (i.e., a form of hard-thresholding) to what they call the use of targeted predictors (that include soft-thresholding as well), and find it helpful in a macroeconomic forecasting environment.

Unlike Bair et al. (2006), our supervised PCA proposal involves an additional projection step, and a subsequent iterative procedure over selection, PCA, and projection to extract latent factors. More specifically: we first select a subset of the predictors that correlate with the target, and extract a first factor from that subset using PCA. Then, we project the target and all the predictors (including those not selected) on the first factor, and take the residuals. We then repeat the selection step using these residuals, extract a second factor from the new subset using PCA, and then project again the residuals of the target and all predictors on this second factor. We keep iterating these steps until all factors are extracted, each from a different subset of predictors (or their residuals). We provide examples to illustrate that our iterative procedure is necessary in general settings where factors can grow at distinct rates (that is, they are of different strength) and factors are not necessarily marginally correlated with the target. The final step of our procedure is to make predictions with estimated factors via time-series regressions.

We justify our procedure in an asymptotic scheme where both the sample size and the cross-sectional dimension increase but at potentially different rates. We show that our iterative procedure delivers consistent prediction of the target. While our procedure can extract weak factors, we do not have asymptotic guarantee for recovery of the factor space that is orthogonal to the target. Importantly, this is irrelevant for consistency in prediction. Intuitively, using information about the correlation between each predictor and the target, we gain additional information useful to extract some of the factors even when they are weak. As a result, the factor space that we may fail to recover must be orthogonal to the target, and therefore missing it does not affect the consistency of the prediction.

The weak factor problem in our setting arises from the factor loading matrix, whose singular values increase but at a potentially slower rate than the cross-sectional dimension. The factors we consider are weaker than those discussed in Bai and Ng (2021); as we show in the paper, PCA cannot consistently recover them, and prediction via PCA is biased. Interestingly, in this setting even supervised procedures may in general fail to recover the relevant factors: specifically, we show that a widely used supervised procedure, partial least squares (PLS), is in fact subject to the same bias as PCA. That said, our procedure will miss factors that are *extremely weak*. These are the kind of factors studied by Onatski (2009) and Onatski (2010), cases in which the eigenvalues corresponding to the factor component are of the same order of magnitude as those of the idiosyncratic component. In this context, while it is possible to infer the number of factors, Onatski (2012) show that the factor space cannot be recovered consistently (and neither SPCA will be able to do so).

Finally, beyond consistency (which requires weaker assumptions), if we make an additional assumption that each of the latent factors is correlated with at least one of the variables in a multivariate target, we can obtain stronger results: we can estimate the number of weak factors consistently, recover the space spanned by all factors, as well as provide a valid prediction interval on the target. Our asymptotic result does not rely on a perfect recovery of the set of predictors that are correlated with the factors, unlike Bair et al. (2006). Moreover, our result accounts for potential errors accumulated over the iterative procedure.

Supplementary evidence, collected in Giglio et al. (2022) for space reasons, illustrates the use of SPCA with an empirical application to macroeconomic forecasting. There, we combine the standard Fred-Md dataset of 127 macroeconomic and financial variables with the Blue Chip Financial Forecasts dataset, that contains hundreds of forecasts of various variables (like interest rates and inflation) from professional forecasters, thus obtaining a large dataset of predictors. We then apply different prediction and dimension reduction methods to forecast quarterly inflation, industrial production growth, and changes in unemployment. We compare the results using SPCA to those obtained using PCA (as in Stock and Watson (2002)) and PLS (as in Kelly and Pruitt (2013)). We show that in a setting with a large number of (potentially noisy and/or redundant) predictors, SPCA performs well in forecasting macroeconomic quantities out of sample. We also investigate the selection that SPCA operates, and find that it isolates, for each target, a different group of useful predictors; it also focuses on a few financial forecasters, whose survey responses are selected particularly often. Finally, we illustrate the use of SPCA with multiple targets at the same time (macroeconomic variables forecasted at different horizons: 1, 2, 3, 6 and 12 months).

Our paper relates to several strands of the literature on forecasting and on dimension reduction. Within the context of forecasting using latent factors, it focuses on static approximate factor models. Dynamic factor models are developed in Forni et al. (2000), Forni and Lippi (2001), Forni et al. (2004), and Forni et al. (2009), in which the lagged values of the unobserved factors may also affect the observed predictors. It is possible to extend our approach to the dynamic factor setting, which is beyond the scope of this paper. Chao and Swanson (2022) study estimation and forecasting within a weak-factor-augmented VAR framework. They also use a pre-selection step since factors only have influence on a subset of predictors. A unique contribution of theirs is a self-normalized score statistics for selection in place of correlation screening as in supervised PCA, which ensures consistent selection of marginally correlated predictors with vanishing Type I and II errors. Similar to Bair et al. (2006), they assume all factors to have the same order of strength and all important predictors to be marginally correlated with the target, which our iterative procedure is designed to avoid.

Our paper is also related to a strand of the literature on spike covariance models defined in Johnstone (2001), where the largest few eigenvalues in the covariance matrix differ from the rest in population, yet are still bounded. In this setting, Bai and Silverstein (2006), Johnstone and Lu (2009) and Paul (2007) show that the largest sample eigenvalues and their corresponding eigenvectors are inconsistent unless the sample size grows at a faster rate than the increase of the cross-sectional dimension. Wang and Fan (2017) extend this setting to the case of diverging eigenvalue spikes, and characterize the limiting distribution of the extreme eigenvalues and certain entries of the eigenvectors in a regime where the sample size grows much slower than the dimension. All these papers shed light on the source of bias with the standard PCA procedure in various asymptotic settings.

Besides supervised PCA, an alternative route taken by an adjacent literature to resolving the inconsistency of PCA is sparse PCA, which imposes sparsity on population eigenvectors, see, e.g., Jolliffe et al. (2003), Zou et al. (2006), d'Aspremont et al. (2007), Johnstone and Lu (2009), and Amini and Wainwright (2009). Uematsu and Yamagata (2021) adopt a variant of the sparse PCA algorithm proposed in Uematsu et al. (2019) to estimate a sparsity-induced weak factor model. Bailey et al. (2020) and Freyaldenhoven (2022) adopt a similar framework for estimating factor strength and number of factors. Because sparsity is rotation dependent, such weak factor models require rotation-specific identification assumptions, whereas standard factor models do not. The weak factor models we consider, for instance, avoid such a sparsity assumption, which makes our approach distinct from the sparse PCA.

In a companion paper, Giglio et al. (2020) incorporate a similar supervised PCA algorithm into the two-pass cross-sectional regression and study its parameter inference in an asset pricing context. In contrast, this paper focuses on predictive inference and provides asymptotic guarantee on the convergence of extracted factors. Our approach also shares the spirit with Bai and Ng (2008) and Huang et al. (2021). The former suggests a hard or soft thresholding procedure to select "targeted" predictors to which PCA is then applied, without providing theoretical justification. The latter suggests scaling each predictor with its predictive slope on the prediction target before applying the PCA. Our procedure and its asymptotic justification are more involved because the eigenvalues of the factor loadings in our setting can grow at distinct and slower rates.

The rest of the paper is organized as follows. In Section 2 we introduce the model, provide examples to illustrate the impact of weak factors on prediction, and develop our supervised PCA procedure. In Section 3, we present our approach in general settings and provide asymptotic theory for our procedure. Section 4 provides Monte Carlo simulations demonstrating the finite-sample performance. Section 5 concludes. The appendix provides mathematical proofs of the main theorems in the paper. The online appendix presents details on the asymptotic variance estimation and proofs of propositions and technical lemmas.

2 Methodology

2.1 Notation

Throughout the paper, we use (A, B) to denote the concatenation (by columns) of two matrices A and B. For any time series of vectors $\{a_t\}_{t=1}^{T}$, we use the capital letter A to denote the matrix $(a_1, a_2, \cdots, a_T)$, $\overline{A}$ for $(a_{1+h}, a_{2+h}, \cdots, a_T)$, and $\underline{A}$ for $(a_1, a_2, \cdots, a_{T-h})$, for some h. We use $\langle N \rangle$ to denote the set of integers: $\{1, 2, \ldots, N\}$. For an index set $I \subset \langle N \rangle$, we use $|I|$ to denote its cardinality. We use $A_{[I]}$ to denote a submatrix of A whose rows are indexed in I.

We use $a \vee b$ to denote the max of a and b, and $a \wedge b$ as their min for any scalars a and b. We also use the notation $a \lesssim b$ to denote $a \leq Kb$ for some constant $K > 0$ and $a \lesssim_{\mathrm{P}} b$ to denote $a = O_{\mathrm{P}}(b)$. If $a \lesssim b$ and $b \lesssim a$, we write $a \asymp b$ for short. Similarly, we use $a \asymp_{\mathrm{P}} b$ if $a \lesssim_{\mathrm{P}} b$ and $b \lesssim_{\mathrm{P}} a$.

We use $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ to denote the minimum and maximum eigenvalues of A, and use $\lambda_i(A)$ to denote the i-th largest eigenvalue of A. Similarly, we use $\sigma_i(A)$ to denote the ith singular value of A. We use $\|A\|$ and $\|A\|_{\mathrm{F}}$ to denote the operator norm (or ℓ_2 norm), and the Frobenius norm of a matrix $A = (a_{ij})$, that is, $\sqrt{\lambda_{\max}(A'A)}$, and $\sqrt{\mathrm{Tr}(A'A)}$, respectively. We also use $\|A\|_{\mathrm{MAX}} = \max_{i,j} |a_{ij}|$ to denote the ℓ_∞ norm of A on the vector space. We use $\mathbb{P}_A = A(A'A)^{-1}A'$ and $\mathbb{M}_A = \mathbb{I}_d - \mathbb{P}_A$, for any matrix A with d rows and rank d, where $\mathbb{I}_d$ is a $d \times d$ identity matrix.

2.2 Model Setup

Our objective is to predict a $D \times 1$ vector of targets, y_{T+h}, h-step ahead from a set of N predictor variables x_t with a sample of size T.

We assume that x_t follows a linear factor model, that is,

$$x_t = \beta f_t + \beta_w w_t + u_t, \tag{1}$$

where f_t is a $K \times 1$ vector of latent factors, w_t is an $M \times 1$ vector of observed variables, u_t is an $N \times 1$ vector of idiosyncratic errors satisfying $\mathrm{E}(u_t) = 0$, $\mathrm{E}(f_t u_t') = 0$, and $\mathrm{E}(w_t u_t') = 0$. Without loss of generality, we also impose that $\mathrm{E}(f_t w_t') = 0$.[1]

We assume that the target variables in y are related to x through factors f in a predictive model:

$$y_{t+h} = \alpha f_t + \alpha_w w_t + z_{t+h}, \tag{2}$$

where z_{t+h} is a $D \times 1$ vector of prediction errors.

Using the aforementioned notation, we can rewrite the above two equations in their matrix

[1] Otherwise, we can define $\widetilde{f}_t = f_t - \mathrm{E}(f_t w_t')\mathrm{E}(w_t w_t')^{-1} w_t$ and $\widetilde{\beta}_w = \beta_w + \beta \mathrm{E}(f_t w_t')\mathrm{E}(w_t w_t')^{-1}$, then $\mathrm{E}(\widetilde{f}_t w_t') = 0$ and x_t satisfies a similar equation to (1): $x_t = \beta \widetilde{f}_t + \widetilde{\beta}_w w_t + u_t$.

form as

$$X = \beta F + \beta_w W + U,$$
$$\overline{Y} = \alpha \underline{F} + \alpha_w \underline{W} + \overline{Z}.$$

We now discuss assumptions that characterize the data generating processes (DGPs) of these variables. For clarity of the presentation, we use high-level assumptions, which can easily be verified by standard primitive conditions for i.i.d. or weakly dependent series. Our asymptotic analysis assumes that $N, T \to \infty$, whereas h, K, D, and M are fixed constants.

Assumption 1. *The factor F, the prediction error Z, and the observable regressor W, satisfy:*

$$\left\|T^{-1}\underline{F}\underline{F}' - \Sigma_f\right\| \lesssim_{\mathrm{P}} T^{-1/2}, \|F\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} (\log T)^{1/2}, \left\|T^{-1}\underline{W}\underline{W}' - \Sigma_w\right\| \lesssim_{\mathrm{P}} T^{-1/2},$$
$$\|WF'\| \lesssim_{\mathrm{P}} T^{1/2}, \|Z\| \lesssim_{\mathrm{P}} T^{1/2}, \|Z\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} (\log T)^{1/2}, \left\|\overline{Z}\underline{F}'\right\| \lesssim_{\mathrm{P}} T^{1/2}, \left\|\overline{Z}\underline{W}'\right\| \lesssim_{\mathrm{P}} T^{1/2},$$

where $\Sigma_f \in \mathbb{R}^{K \times K}$, $\Sigma_w \in \mathbb{R}^{M \times M}$ are positive-definite matrices with $\lambda_K(\Sigma_f) \gtrsim 1$, $\lambda_M(\Sigma_w) \gtrsim 1$, $\lambda_1(\Sigma_f) \lesssim 1$, and $\lambda_1(\Sigma_w) \lesssim 1$.

Assumption 1 imposes rather weak conditions on the time series behavior of f_t, z_t, and w_t. Since all of them are finite dimensional time series, the imposed inequalities hold if these processes are stationary, strong mixing, and satisfy sufficient moment conditions.

Moreover, Assumption 1 implies that the K left-singular values of F neither vanish nor explode. Therefore, it is the factor loadings that dictate the strength of factors in our setting. This is without loss of generality because F can always be normalized to satisfy this condition.

Next, we assume

Assumption 2. *The $N \times K$ factor loading matrix β satisfies*

$$\|\beta\|_{\mathrm{MAX}} \lesssim 1, \qquad \lambda_K(\beta_{[I_0]}'\beta_{[I_0]}) \gtrsim N_0,$$

for some index set $I_0 \subset \langle N \rangle$, where $N_0 = |I_0| \to \infty$.

Assumption 2 implies that there exists a subset, I_0, of predictors within which all latent factors are pervasive. This is a much weaker condition than requiring factors to be pervasive in the set of all predictors, in which case $\lambda_1(\beta'\beta) \asymp \ldots \asymp \lambda_K(\beta'\beta) \asymp N$. In contrast, Assumption 2 allows for distinct growth rates for these eigenvalues, in that no requirement is imposed on $\beta_{[I_0^c]}$. Moreover, these eigenvalues can grow at a slower rate than N, since N_0/N is allowed to vanish very rapidly. We will make precise statement about the relative magnitudes of these quantities when it comes to our asymptotic results.

Since the number of factors, K, is assumed finite, even if each factor is pervasive in some separate (and potentially non-overlapping) index set, it is possible to construct a common index set I_0 within which all factors are pervasive.[2] Assumption 2, nevertheless, rules out a somewhat extreme case where all entires of β are uniformly vanishing, i.e., $\sup_{I,|I|\to\infty} |I|^{-1}\lambda_K\left(\beta'_{[I]}\beta_{[I]}\right) = o_{\mathrm{P}}(1)$, to the extent that the desired subset I_0 does not exist.

Next, we need the following moment conditions on U.

Assumption 3. *The idiosyncratic component U satisfies:*

$$\|U\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} (\log T)^{1/2} + (\log N)^{1/2}.$$

In addition, for any given non-random subset $I \subset \langle N\rangle$,

$$\left\|U_{[I]}\right\| \lesssim_{\mathrm{P}} |I|^{1/2} + T^{1/2}.$$

Assumption 3 imposes restrictions on the time-series dependence and heteroskedasticity of u_t. The first inequality is a direct result of a large deviation theorem, see, e.g., Fan et al. (2011). The second inequality can be shown by random matrix theory, see Bai and Silverstein (2009), provided that u_t is i.i.d. both in time and in the cross-section. While it is tempting to impose a stronger inequality that bounds $\sup_{I\subset\langle N\rangle}\left\|U_{[I]}\right\|$ uniformly over all index sets of a given size $|I|$, the rate $|I|^{1/2} + T^{1/2}$ we desire may not hold. In fact, assuming $|I|$ is small, Cai et al. (2021) establish a uniform bound that differs from our non-uniform rate only by a log factor. When $|I|$ is large, the result on uniform bounds no longer exists to the best of our knowledge. We thereby avoid making any assumption on uniform bound over all index sets.

For the same reason, we make the following moment conditions with any given non-random set I. The conditions should hold under weak dependences among U, F, W, and β.

Assumption 4. *For any non-random subset $I \subset \langle N\rangle$, the factor loading $\beta_{[I]}$, and the idiosyncratic error $U_{[I]}$ satisfy the following conditions:*

(i) $\left\|\underline{U}_{[I]}A'\right\| \lesssim_{\mathrm{P}} |I|^{1/2}T^{1/2}$, $\left\|\underline{U}_{[I]}A'\right\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} (\log N)^{1/2}T^{1/2}$,

(ii) $\left\|\beta'_{[I]}U_{[I]}\right\| \lesssim_{\mathrm{P}} |I|^{1/2}T^{1/2}$, $\left\|\beta'_{[I]}U_{[I]}\right\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} |I|^{1/2}(\log T)^{1/2}$, $\left\|\beta'_{[I]}\underline{U}_{[I]}A'\right\| \lesssim_{\mathrm{P}} |I|^{1/2}T^{1/2}$,

[2] To see a concrete example, suppose that β has a block diagonal structure, such that its kth column β_k is supported on an index set J_k, and the intersection of all J_ks is empty. Suppose the non-zero entries of β follow standard normal. Then we can find $k^\star := \min_k |J_k|$, and build up I_0 from $J_{k^\star}$ (so that $|I_0| \geq |J_{k^\star}|$) by arbitrarily adding $|J_{k^\star}|$ number of predictors from each $J_k, k = 1, 2, \ldots, K, k \neq k^\star$. We can take a union of all such subsets of J_k. The resulting index set I_0 contains $K \times |J_{k^\star}|$ number of predictors, and all factors are pervasive within this common set.

(iii) $\left\|(u_T)'_{[I]}\underline{U}_{[I]}A'\right\| \lesssim_{\mathrm{P}} |I| + |I|^{1/2}T^{1/2}$, $\left\|\beta'_{[I]}\left(u_T\right)_{[I]}\right\| \lesssim_{\mathrm{P}} |I|^{1/2}$,

where A is either $\underline{F}$, $\underline{W}$ or $\overline{Z}$.

The ℓ_2-norm bounds in Assumption 4(i) and (ii) are results of Assumptions D, F2, F3 of Bai (2003) when $I = \langle N \rangle$, and Assumption 4(iii) is implied by Assumptions A, E1, F1 and C3 in Bai (2003), except that here we impose a stronger version which holds for any non-random subset $I \subset \langle N \rangle$. The MAX-norm results can be shown by some large deviation theorem as in Fan et al. (2011).

Assumptions 2 and 3 are the key identification conditions of the weak factor model we consider. It is helpful to compare these conditions with those spelled out by Chamberlain and Rothschild (1983). We do not require that u_t is stationary, but for the sake of comparison here, we assume that the covariance matrix of u_t exists, denoted by Σ_u and that $\beta_w = 0$. By model setup (1), we have $\Sigma := \mathrm{Cov}(x_t) = \beta\Sigma_f\beta' + \Sigma_u$. Chamberlain and Rothschild (1983) show that the model is identified if $\|\Sigma_u\| \lesssim 1$ and $\lambda_K \to \infty$, which guarantees the separation of the common and idiosyncratic components in the population model. To implement this strategy, Bai (2003) provides an alternative set of conditions (Assumption C therein) on the time-series and cross-sectional dependence of the idiosyncratic components that ensure the consistency of PCA, but in the case of pervasive factors, that is $\lambda_K(\beta'\beta) \gtrsim N$.

In fact, PCA can separate the factor and idiosyncratic components from the sample covariance matrix under much weaker conditions. To see this, note that from (1) and $\beta_w = 0$, we have $XX' = \beta FF'\beta' + UU' + \beta FU' + UF'\beta'$. Using random matrix theory from Bai and Silverstein (2009), $\lambda_1(UU') \lesssim_{\mathrm{P}} T + N$, if u_t is i.i.d. with $\|\Sigma_u\| \lesssim 1$. Since $T\lambda_K(\beta'\beta) \asymp_{\mathrm{P}} \lambda_K(\beta FF'\beta')$ and because of the weak dependence between U and F as in Assumption 4, the eigenvalues corresponding to the factor component $\beta FF'\beta'$ dominate the three remainder terms that are related to the idiosyncratic component U asymptotically, if $(T+N)/(T\lambda_K(\beta'\beta)) \to 0$, enabling the factor components to be identified from XX'. Wang and Fan (2017) and Bai and Ng (2021) study the setting $N/(T\lambda_K(\beta'\beta)) \to 0$, in which case PCA remains consistent despite the fact that factor exposures are not pervasive. Wang and Fan (2017) also study the borderline case $N \asymp T\lambda_K(\beta'\beta)$, and document a bias term in the estimated eigenvalues and eigenvectors associated with factors.

In this paper, we consider an even weaker factor setting in which $N/(T\lambda_K(\beta'\beta))$ may diverge. In this case, PCA generally fails to recover the underlying factors (except for the special case in which errors are homoscedastic). We will require, instead, the existence of a subset $I_0 \subset \langle N \rangle$, for which $|I_0|/(T\lambda_K(\beta'_{[I_0]}\beta_{[I_0]})) \to 0$, to ensure the identification of factors on this subset.[3] In what follows, we introduce our methodology to deal with this case.

2.3 Prediction via Supervised Principal Components

One potential solution to the weak factor problem was proposed by Bair and Tibshirani (2004), namely, supervised principal component analysis. Their proposal is to locate a subset, $\widehat{I}$, of predictors via marginal screening, keeping only those that have nontrivial exposure to the prediction target, before applying PCA. Intuitively, this procedure reduces the total number of predictors from N to $|\widehat{I}|$, while under certain assumptions it also guarantees that this subset of predictors has a strong factor structure, i.e., $\lambda_{\min}(\beta_{[\widehat{I}]}'\beta_{[\widehat{I}]}) \asymp |\widehat{I}|$. As a result, applying PCA on this subset leads to consistent recovery of factors.

We use a simple one factor example to illustrate the procedure, before explaining its caveats with the general multi-factor case. To illustrate the idea, we consider the case in which $D = K = 1$, $\alpha_w = 0$, and $\beta_w = 0$. We select a subset $\widehat{I}$ that satisfies:

$$\widehat{I} = \left\{ i \middle| T^{-1} |\underline{X}_{[i]} \overline{Y}'| \geq c \right\}, \tag{3}$$

where c is some threshold. Therefore, we keep predictors that covary sufficiently strongly (positively or negatively) with the target. This step involves a single tuning parameter, c, that effectively determines how many predictors we use to extract the factor. The fact that $\widehat{I}$ incorporates information from the target reflects the distinctive nature of a supervised procedure. Given the existence of I_0 by Assumption 2, there exists a choice of c such that predictors within the set $\widehat{I}$ have a strong factor structure. The rest of the procedure is a straightforward application of the principal component regression for prediction. Specifically, we apply PCA to extract factors $\{\widehat{f}_t\}_{t=1}^{T-h}$ from $\underline{X}_{[\widehat{I}]}$, which can be written as $\widehat{f}_t = \widehat{\zeta}' x_t$ for some loading matrix $\widehat{\zeta}$, then obtain $\widehat{\alpha}$ by regressing $\{y_t\}_{t=1+h}^{T}$ onto $\{\widehat{f}_t\}_{t=1}^{T-h}$ based on the predictive model (2). The resulting predictor for y_{T+h} is therefore given by: $\widehat{y}_{T+h} = \widehat{\alpha}\widehat{f}_T = \widehat{\alpha}\widehat{\zeta}' x_T$.

Bair et al. (2006)'s proposal proceeds in the same way when it comes to multiple factors, with the only exception that multiple factors are extracted in the PCA step. Yet, to ensure that marginal screening remains valid in the multi-factor setting, they assume that predictors are marginally correlated with the target *if and only if* they belong to a *uniquely* determined subset I_0, outside which predictors are assumed to have zero correlations with the prediction target, i.e., they are pure noise for prediction purpose. Given this condition, they show marginal

[3] The aforementioned settings all require $\lambda_K(\beta'\beta) \to \infty$, in contrast with the extremely weak factor model that imposes $\lambda_K(\beta'\beta) \lesssim 1$. As such, eigenvalues of factors and idiosyncratic components do not diverge as dimension increases. While Onatski (2009) and Onatski (2010) develop tests for the number of factors, Onatski (2012) shows that factors cannot be consistently recovered in this regime.

screening can consistently recover I_0, and all factors can thereby be extracted altogether with a single pass of PCA to this subset of predictors.

In contrast, we assume the existence of a set I_0 within which predictors have a strong factor structure, yet we do not make any assumptions on the correlation between the target and predictors outside this set I_0, nor on the strength of their factor structure. As a result, I_0 under our Assumption 2 needs not be unique, and we will show that the validity of the prediction procedure does not rely on consistent recovery of any pre-determined set I_0. More importantly, since marginal screening is based on marginal covariances between $\overline{Y}$ and $\underline{X}$, in a multi-factor model the condition that marginal screening can recover a subset within which all factors are pervasive (even if such a subset is uniquely defined as in Bair et al. (2006)) is rather strong. On the one hand, marginal screening can be misguided by the correlation induced by a strong factor to the extent that weak factors after screening remain unidentifiable. On the other hand, predictors eliminated by marginal screening can be instrumental or even essential for prediction. We illustrate these points using examples of two-factor models below.

Example 1: Suppose x_t and y_t satisfy the following dynamics:

$$x_t = \left[\begin{array}{c|c} \beta_{11} & \beta_{12} \\ \hline \beta_{21} & 0 \end{array}\right] f_t + u_t, \quad y_{t+h} = \begin{bmatrix} 1 & 1 \end{bmatrix} f_t, \tag{4}$$

where β_{11} and β_{12} are $N_0 \times 1$ vectors, β_{21} is an $(N - N_0) \times 1$ vector, satisfying $\|\beta_{12}\| \asymp N_0^{1/2}$ and $\|\beta_{21}\| \asymp (N - N_0)^{1/2}$, and N_0 is small relative to N.

In this example, the first factor is strong (all predictors are exposed to it) while the second factor is weak, since most exposures to it are zero. In addition, the target variable y is correlated with both factors and hence potentially with all predictors. As a result, the screening step described above may not eliminate any predictors: all predictors may correlate with the target (through the first factor). But because the second factor is weak, a single pass of PCA, extracting two factors from the entire universe of predictors, would fail to recover it: we can show that $\lambda_{\min}(\beta'\beta) \leq \|\beta_{12}\|^2 \lesssim N_0$, so that PCA would not recover the second factor consistently if $N/(N_0 T)$ does not vanish.

The issue highlighted with this example is that the (single) screening step does not eliminate any predictors, because their correlations with the target are (at least partially) induced by their exposure to the strong factor, and therefore PCA after screening cannot recover the weak factor. The assumptions proposed by Bair et al. (2006) rule this case out, but we can clearly locate an index set I_0 (say, top N_0 predictors), within which both factors are strong. In other words, our assumptions can accommodate this case.

We provide next another example, that shows that in some situations screening can eliminate *too many* predictors, making a strong factor model become weak or even rank-deficient.

Example 2: Suppose x_t and y_t satisfy the following dynamics:

$$x_t = \left[\begin{array}{c|c} \beta_{11} & \beta_{11} \\ \hline 0 & \beta_{22} \end{array}\right] f_t + u_t, \qquad y_{t+h} = \left[\begin{array}{cc} 1 & 0 \end{array}\right] f_t, \tag{5}$$

where β_{11} and β_{22} are $N/2 \times 1$ non-zero vectors satisfying $\|\beta_{11}\| \asymp \|\beta_{22}\| \asymp \sqrt{N}$ and f_{1t} and f_{2t} are uncorrelated.

In this example, there are two equal-sized groups of predictors, so that β is full-rank and both factors are strong and that I_0 can be the entire set $\langle N \rangle$ (therefore, a standard PCA procedure applied to all predictors will consistently recover both factors). But two features of this model will make supervised PCA fail, if the selection step based on marginal correlations is applied only once (as in the original procedure by Bair et al. (2006)). First, y_{t+h} is uncorrelated with the second half of predictors (since only the first group is useful for prediction). Second, the exposure of the first half of predictors to the first and second factors are the same (both equal to β_{11}).

After the screening step the second group of predictors would be eliminated, because they do not marginally correlate with y_{t+h}. But the remaining predictors (the first half) have perfectly correlated exposures to both factors, so that only one factor, $f_{1t}+f_{2t}$, can be recovered by PCA. Therefore, the one-step supervised PCA of Bair et al. (2006) would fail to recover the factor space consistently, resulting in inconsistent prediction. This example highlights an important point that marginally uncorrelated predictors (the second half) could be essential in recovering the factor space. Eliminating such predictors may lead to inconsistency in prediction.

Both examples demonstrate the failure of a one-step supervised PCA procedure in a general multi-factor setting. Such data generating processes are excluded by the model assumptions in Bair et al. (2006), whereas we do not rule them out. We thus propose below a new and more complete version of the supervised PCA (SPCA) procedure that can accommodate such cases.

2.4 Iterative Screening and Projection

To resolve the issue of weak factors in a general multi-factor setting, we propose a multi-step procedure that iteratively conducts selection and projection. The projection step eliminates the influence of the estimated factor, which ensures the success of the screening steps that occur over the following iterations. More specifically, a screening step can help identify one strong factor from a selected subset of predictors. Once we have recovered this factor, we project *all* predictors x_t (not just those selected at the first step) and y_{t+h} onto this factor, so that their residuals will not be correlated with this factor. Then we can repeat the same selection procedure with these residuals. This approach enables a continued discovery of factors, and guarantees that each new factor is orthogonal to the estimated factors in the previous steps, similar to the standard PCA.

It is straightforward to verify that this iterative screening and projection approach successfully addresses the issues with the aforementioned examples. Consider first Example 1. In this case, the first screening does not rule out any predictor, and the first PC will recover the strong factor f_1; after projecting both X and y onto f_1, the residuals for the first N_0 predictors still load on f_2, whereas the remaining $N - N_0$ predictors should have zero correlation with the residuals of y. Therefore, a second screening will eliminate these predictors, paving the way for PCA to recover the second factor f_2 based on the residuals of the first N_0 predictors. Similarly, for Example 2, the first screening step eliminates the second half of the predictors, so that the first pass of PCA will recover the only factor left over in the remaining predictors, namely, $f_1 + f_2$. The residuals of the first half of predictors consist of pure noise after the projection step, whereas the residuals of the second half of predictors are spanned by $f_1 - f_2$, which a second PCA step will recover. Therefore, the iterated supervised PCA will recover the entire factor space. This example illustrates that marginal screening can succeed as long as iteration and projection are also employed.

Formally, we present our algorithm for the general model given by (1) and (2):

Algorithm 1 (Prediction via SPCA)**.**
Inputs: $\overline{Y}$, $\underline{X}$, $\underline{W}$, x_T, *and* w_T. *Initialization:* $Y_{(1)} := \overline{Y}\mathbb{M}_{\underline{W}'}$, $X_{(1)} := \underline{X}\mathbb{M}_{\underline{W}'}$.

S1. For $k = 1, 2, \ldots$ *iterate the following steps using* $X_{(k)}$ *and* $Y_{(k)}$*:*

- *a. Select an appropriate subset* $\widehat{I}_k \subset \langle N \rangle$ *via marginal screening.*
- *b. Estimate the kth factor* $\widehat{\underline{F}}_{(k)} = \widehat{\varsigma}'_{(k)} \left(X_{(k)}\right)_{[\widehat{I}_k]}$ *via SVD, where* $\widehat{\varsigma}_{(k)}$ *is the first left singular vector of* $\left(X_{(k)}\right)_{[\widehat{I}_k]}$*.* $\widehat{\underline{F}}_{(k)}$ *can also be rewritten as* $\widehat{\underline{F}}_{(k)} = \widehat{\zeta}'_{(k)}\underline{X}\mathbb{M}_{\underline{W}'}$*, where* $\widehat{\zeta}_{(k)} = \left(\mathbb{I}_N - \sum_{i=1}^{k-1} \widehat{\beta}_{(i)}\widehat{\zeta}'_{(i)}\right)'_{[\widehat{I}_k]} \widehat{\varsigma}_{(k)}$ *is constructed recursively using* $\widehat{\beta}_{(k-1)}$ *(defined in*

c.).

c. *Estimate the coefficients* $\widehat{\alpha}_{(k)} = Y_{(k)}\underline{\widehat{F}}'_{(k)}(\underline{\widehat{F}}_{(k)}\underline{\widehat{F}}'_{(k)})^{-1}$ *and* $\widehat{\beta}_{(k)} = X_{(k)}\underline{\widehat{F}}'_{(k)}(\underline{\widehat{F}}_{(k)}\underline{\widehat{F}}'_{(k)})^{-1}$.

d. *Obtain residuals* $Y_{(k+1)} = Y_{(k)} - \widehat{\alpha}_{(k)}\underline{\widehat{F}}_{(k)}$ *and* $X_{(k+1)} = X_{(k)} - \widehat{\beta}_{(k)}\underline{\widehat{F}}_{(k)}$.

Stop at $k = \widehat{K}$, *where* $\widehat{K}$ *is chosen based on some proper stopping rule.*

S2. Obtain $\widehat{f}_T = \widehat{\zeta}'(x_T - \widehat{\beta}_w w_T)$, *where* $\widehat{\zeta} := (\widehat{\zeta}_{(1)}, \ldots, \widehat{\zeta}_{(\widehat{K})})$ *and* $\widehat{\beta}_w = \underline{XW'}(\underline{WW'})^{-1}$, *and the prediction* $\widehat{y}_{T+h} = \widehat{\alpha}\widehat{f}_T + \widehat{\alpha}_w w_T = \widehat{\gamma}x_T + (\widehat{\alpha}_w - \widehat{\gamma}\widehat{\beta}_w)w_T$, *where* $\widehat{\alpha} := (\widehat{\alpha}_{(1)}, \widehat{\alpha}_{(2)}, \ldots, \widehat{\alpha}_{(\widehat{K})})$, $\widehat{\gamma} = \widehat{\alpha}\widehat{\zeta}'$, *and* $\widehat{\alpha}_w = \overline{Y}\underline{W}'(\underline{WW'})^{-1}$.

Outputs: the prediction $\widehat{y}_{T+h}$, *the factors* $\underline{\widehat{F}} := (\underline{\widehat{F}}'_{(1)}, \ldots, \underline{\widehat{F}}'_{(\widehat{K})})'$, *their loadings,* $\widehat{\beta} := (\widehat{\beta}_{(1)}, \ldots, \widehat{\beta}_{(\widehat{K})})$, *and the coefficient estimates* $\widehat{\alpha}$, $\widehat{\zeta}$, $\widehat{\alpha}_w$, $\widehat{\beta}_w$, *and* $\widehat{\gamma}$.

We discuss the details of the algorithm below.

Step S1. of Algorithm 1 requires an appropriate choice of $\widehat{I}_k$ and a stopping rule. One possible choice for $\widehat{I}_k$ is:[4]

$$\widehat{I}_k = \left\{ i \middle| T^{-1} \left\| (X_{(k)})_{[i]} Y'_{(k)} \right\|_{\mathrm{MAX}} \geq \widehat{c}^{(k)}_{qN} \right\},$$
$$\text{where } \widehat{c}^{(k)}_{qN} \text{ is the } (1-q)th\text{-quantile of } \left\{ T^{-1} \left\| (X_{(k)})_{[i]} Y'_{(k)} \right\|_{\mathrm{MAX}} \right\}_{i=1,\ldots,N}. \tag{6}$$

The reason we suggest using the top qN predictors based on the magnitude of the covariances between $X_{(k)}$ and $Y_{(k)}$ is that the factor estimates tend to be more stable and less sensitive to this tuning parameter q, compared to a conventional hard threshold parameter adopted in a marginal screening procedure. Moreover, at each step, a subset of a *fixed* number of predictors are selected, which substantially simplifies the notation and the proof.

Correspondingly, the algorithm terminates as soon as

$$\widehat{c}^{(k+1)}_{qN} < c, \quad \text{for some threshold } c. \tag{7}$$

Thus, the resulting number of factors is set as $\widehat{K} = k$. As a result, the tuning parameter, c, effectively determines the number of factors extracted out of our procedure.

[4] Using covariance for screening allows us to replace all $Y_{(k)}$ in the definition of $\widehat{I}_k$ and Algorithm 1 by $Y_{(1)}$, that is, only the projection of $X_{(k)}$ is needed, because this replacement would not affect the covariance between $Y_{(k)}$ and $X_{(k)}$. We use this fact in the proofs, which simplifies the notation. We can also use correlation instead of covariance in constructing $\widehat{I}_k$, which does not affect the asymptotic analysis. That said, we find correlation screening performs better in finite samples when the scale of the predictors differs.

For any given tuning parameters, q and c, we select predictors that have predictive power for (at least one variable in) y_{t+h} at each stage of the iteration. With a good choice of tuning parameters, q and c, the iteration stops as soon as most of the rows of the projected residuals of predictors appear uncorrelated with the projected residuals of y_{t+h}, which implies that the factors left over, if any, are uncorrelated with y_{t+h}.

The last step of the algorithm needs more explanations. Step S1. provides a set of factor estimates, $\underline{\widehat{F}}$, on the basis of $\overline{Y}$ and $\underline{X}$. Moreover, a time series regression of $\overline{Y}$ on $\underline{\widehat{F}}$ and $\underline{W}$ yields an estimator of α_w (coefficient defined in (2)). That is, $\widehat{\alpha}_w = \overline{Y}\mathbb{M}_{\underline{\widehat{F}}'}\underline{W}'\left(\underline{W}\mathbb{M}_{\underline{\widehat{F}}'}\underline{W}'\right)^{-1} = \overline{Y}\underline{W}'(\underline{WW}')^{-1}$, since $\mathbb{M}_{\underline{\widehat{F}}'}\underline{W}' = \underline{W}'$ by construction, which explains the formula for $\widehat{\alpha}_w$ in Step S2.. Finally, with $\widehat{\alpha}$, $\widehat{\alpha}_w$, and $\widehat{f}_T$, it is sufficient to construct the predicted value of y_{T+h} by combining $\widehat{\alpha}\widehat{f}_T$ with $\widehat{\alpha}_w w_T$, which yields the final prediction formula for $\widehat{y}_{T+h}$, a projection on observables, x_T and w_T.

3 Asymptotic Theory

We now examine the asymptotic properties of SPCA. The analysis is more involved than those of Bair et al. (2006) because of the iterative nature of our new SPCA procedure and the general weak factor setting we consider.

3.1 Consistency in Prediction

To establish the consistency of SPCA for prediction, we first investigate the consistency of factor estimation. In the strong factor case, e.g., Stock and Watson (2002), all factors are recovered consistently via PCA, which is a prerequisite for the consistency of prediction. In our setup of weak factors, we show that the consistency of prediction only relies on consistent recovery of factors that are relevant for the prediction target.

Recall that in Algorithm 1, we denote the selected subsets in the SPCA procedure as $\widehat{I}_k$, $k = 1, 2, \ldots$. We now construct their population counterparts iteratively, for any given choice of c and q. This step is critical to characterize the exact factor space recovered by SPCA. For simplicity in notation and without loss of generality, we consider the case $\Sigma_f = \mathbb{I}_K$ here, because in the general case, we can simply replace β and α by $\beta^* = \beta\Sigma_f^{1/2}$ and $\alpha^* = \alpha\Sigma_f^{1/2}$ in the following construction.

In detail, we start with $a_i^{(1)} := \left\|\beta_{[i]}\alpha'\right\|_{\text{MAX}}$ and define $I_1 := \{i | a_i^{(1)} \geq c_{qN}^{(1)}\}$, where $c_{qN}^{(1)}$ is the $\lfloor qN \rfloor$th largest value in $\left\{a_i^{(1)}\right\}_{i=1,\ldots,N}$. Then, we denote the largest singular value of $\beta_{(1)} := \beta_{[I_1]}$ by $\lambda_{(1)}^{1/2}$ and the corresponding left and right singular vectors by $\varsigma_{(1)}$ and $b_{(1)}$.

For $k > 1$, we obtain $a_i^{(k)} := \left\| \beta_{[i]} \prod_{j<k} \mathbb{M}_{b_{(j)}} \alpha' \right\|_{\mathrm{MAX}}$, $I_k := \{i | a_i^{(k)} \geq c_{qN}^{(k)}\}$, and $\lambda_{(k)}^{1/2}$, $\varsigma_{(k)}$, $b_{(k)}$ are the leading singular value, left and right singular vectors of $\beta_{(k)} := \beta_{[I_k]} \prod_{j<k} \mathbb{M}_{b_{(j)}}$. This procedure is stopped at step $\tilde{K}$ (for some $\tilde{K}$ that is not necessarily equal to K or $\widehat{K}$) if $c_{qN}^{(\tilde{K}+1)} < c$. In a nutshell, I_k's are what we will select if we do SPCA directly on $\beta \in \mathbb{R}^{N \times K}$ and $\alpha \in \mathbb{R}^{D \times K}$ and they are deterministically defined by $\alpha, \beta, \Sigma_f, c, q$, and N, whereas $\widehat{I}_k$'s are random, obtained by SPCA on $\underline{X} \in \mathbb{R}^{N \times T}$ and $\overline{Y} \in \mathbb{R}^{D \times T}$.

To ensure that the singular vectors $b_{(j)}$'s are well defined and identifiable, we need that the top two singular values of $\beta_{(k)}$ are distinct at each stage k. We also need distinct values of $c_{qN}^{(k)}$ to ensure that I_k's are identifiable. More precisely, we say that two sequences of variables a_N and b_N are asymptotically distinct if there exists a constant $\delta > 0$ such that $|a_N - b_N| \geq \delta |b_N|$ for sufficiently large N. In light of the above discussion, we make the following assumption:

Assumption 5. *For any given k, the following three pairs of sequences of variables, $\sigma_1(\beta_{(k)})$ and $\sigma_2(\beta_{(k)})$, $c_{qN}^{(k)}$ and $c_{qN+1}^{(k)}$, and $c_{qN}^{(\tilde{K}+1)}$ and c are asymptotically distinct, as $N \to \infty$.*

This assumption is rather mild as it only rules out corner cases, despite the fact that this is not very explicit. Excluding such corner cases is common in the literature on high dimensional PCA, see, e.g., Assumption 2.1 of Wang and Fan (2017). Assumption 5 is closely tied to our choice of the number of predictors qN and the parameter c in the stopping rule. In particular, the current algorithm adopts a strategy where the same number of predictors is selected at each step, representing one version of SPCA. An alternative approach may involve selecting predictors based on a predetermined threshold for their covariances and stopping the selection process when $|I_k|$ becomes smaller than another threshold. By allowing for the flexibility of using varying numbers of predictors at each step, this alternative approach can be particularly useful in addressing certain corner cases ruled out by the current version of Assumption 5.[5] Similar asymptotic results, akin to those presented in Theorem 1 through 3 below, can be derived with more intricate conditions regarding the rate of convergence, etc. However, the current version of SPCA, with its more concise theorems and superior performance in simulation, is the primary focus of our discussion in the main text. We now are ready to present the consistency of the estimated factors by SPCA:

Theorem 1. *Suppose that x_t follows (1) and y_t satisfies (2), and that Assumptions 1-5 hold. If $\log(NT)(N_0^{-1} + T^{-1}) \to 0$, then for any tuning parameters c and q that satisfy*

$$c \to 0, \quad c^{-1}(\log NT)^{1/2}(q^{-1/2}N^{-1/2} + T^{-1/2}) \to 0, \quad qN/N_0 \to 0, \tag{8}$$

[5] A concrete example may be the case where all $a_i^{(1)}$s defined above are identical, resulting in $c_{qN}^{(1)} = c_{qN+1}^{(1)}$. By adopting the alternative algorithm, we only need an assumption on a non-vanishing lower bound of $a_i^{(1)}$, i.e., $a_i^{(1)} > c > 0$. Correspondingly, this alternative procedure will select all predictors in this iteration.

we have $\tilde{K} \leq K$, $\mathrm{P}(\widehat{I}_k = I_k) \to 1$, *for any* $1 \leq k \leq \tilde{K}$, *and* $\mathrm{P}(\widehat{K} = \tilde{K}) \to 1$. *Moreover, the factors recovered by SPCA are consistent. That is, for any* $1 \leq k \leq \tilde{K}$,

$$\left\|\underline{\widehat{F}}_{(k)}\right\|^{-1}\left\|\underline{\widehat{F}}_{(k)} - \underline{\widehat{F}}_{(k)}\mathbb{P}_{\underline{F}'}\right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1}. \tag{9}$$

We make a few observations regarding this result. First, the assumptions in Theorem 1 do not guarantee a consistent estimate of the number of factors, K, because the SPCA procedure cannot guarantee to recover factors that are uninformative about y. At the same time, the factors recovered by SPCA are not necessarily useful for prediction, because it is possible that some strong factors with no predictive power are also recovered by SPCA. Ultimately, the factor space recoverable is determined by β, α, Σ_f, c, q, and N. For this reason, we have consistency of factor estimates up to the first $\tilde{K}$ factors. Moreover, $\widehat{K}$ is a consistent estimator of $\tilde{K}$, which we prove satisfies $\tilde{K} \leq K$. That is, SPCA omits $K - \tilde{K}$ factors. Also, the inequality (9) has a clear geometric interpretation. The left-hand-side is exactly equal to $\sin(\widehat{\Theta}_{(k)})$, where $\widehat{\Theta}_{(k)}$ is the angle between the estimated factor at each stage k and the factor space spanned by the true factors, $\mathbb{P}_{\underline{F}'}$. (9) shows that this angle vanishes asymptotically.

Second, with respect to the tuning parameters, the condition (8) implies that $c \to 0$, $c\sqrt{T} \to \infty$, and $c\sqrt{qN} \to \infty$. On the one hand, the threshold c needs be sufficiently small so that the iteration procedure continues until selected predictors have asymptotically vanishing predictive power; on the other hand, c needs be large enough that dominates error in the covariance estimates from the screening step. The estimation error consists of the usual error in the construction of the sample covariances between $X_{(1)}$ and $Y_{(1)}$, which introduces an error of order $T^{-1/2}$, as well as the construction of residuals in the projection step, $X_{(k)}$ and $Y_{(k)}$, for $k > 1$, as soon as multiple factors are involved (i.e., $\tilde{K} > 1$). As we show next, the factor estimation error is of order $(qN)^{-1/2} + T^{-1}$, which pollutes the residuals and hence affects screening. Taking these two points into consideration, the choice of c needs dominate $T^{-1/2} + (qN)^{-1/2}$. In terms of q, it appears that the maximal number of selected predictors, $\lfloor qN \rfloor$, allowed for should be of the same order as N_0. Nevertheless, since N_0 given by Assumption 2 is not precisely defined, in the sense that the assumption holds if N_0 is scaled by any non-zero constant, we require $qN/N_0 \to 0$ to ensure that the scaling constant of N_0 does not matter for the choice of q and that the selected $\lfloor qN \rfloor$ predictors are within the subset of N_0 predictors that guarantee a strong factor structure.

Third, the estimation error of factors are bounded from the above by $q^{-1/2}N^{-1/2} + T^{-1}$. Recall that in the strong factor case, the factor space can be recovered at the rate of $N^{-1/2} + T^{-1}$, see, e.g., Bai (2003). In our result, qN plays the same role as N in the strong factor case. Nevertheless, our Assumption 2 does not require all factors to have the same strength.

It is possible that some factors could be recovered with a higher convergence rate, should we select a different number of predictors for each factor based on its strength. In fact, an alternative choice of $\widehat{I}_k$ based on (3) allows different numbers of predictors to be selected at each stage, since the threshold itself is a fixed level. While this approach may achieve a faster rate for relatively stronger factors, the prediction error rate is ultimately determined by the estimation error of the weakest factor. Yet, we find that the approach based on (6) offers more stable prediction out of sample, whereas prediction based on (3) can be sensitive to the tuning parameters. Given that our ultimate goal is about prediction rather than factor recovery, we prefer a more stable procedure and thereby focus our analysis on the former approach.

With no relevant factors omitted, our prediction $\widehat{y}_{T+h}$ is consistent, as we show next.

Theorem 2. *Under the same assumptions as in Theorem 1, we have* $\widehat{\alpha}_w - \alpha_w \xrightarrow{\mathrm{P}} 0$, $\|\widehat{\gamma}\beta - \alpha\| \xrightarrow{\mathrm{P}} 0$, *and consequently,* $\widehat{y}_{T+h} \xrightarrow{\mathrm{P}} \mathrm{E}_T(y_{T+h}) = \alpha f_T + \alpha_w w_T$.

Theorem 2 first analyzes the parameter estimation "error" measured as $\widehat{\alpha}_w - \alpha_w$ and $\widehat{\gamma}\beta - \alpha$. The reason the latter quantity matters is that there exists a matrix H such that $\widehat{\gamma}\beta = \widehat{\alpha}H$. In other words, the first statement of the theorem implies that we can consistently estimate α, up to a matrix H. This extra adjustment matrix H exists due to the fundamental indeterminacy of latent factor models. In fact, we can define $H \in \mathbb{R}^{\widehat{K}\times K}$ as $\widehat{\zeta}'\beta$, where $\widehat{\zeta}$ is given by Algorithm 1. Then, it is straightforward to see from the definition of $\widehat{\gamma}$ that

$$\widehat{\gamma}\beta = \widehat{\alpha}H, \quad \text{so that by Theorem 2} \quad \|\widehat{\alpha}H - \alpha\| = o_{\mathrm{P}}(1). \tag{10}$$

On the other hand, the proof of Theorem 1 also establishes that for $k \leq \tilde{K}$:

$$\left\|\widehat{\underline{F}}_{(k)}\right\|^{-1}\left\|\widehat{\underline{F}}_{(k)} - h_k\underline{F}\right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1}, \tag{11}$$

where h_k is the kth row of H. Therefore, $\widehat{\alpha}\widehat{\underline{F}} \overset{\text{by(11)}}{\approx} \widehat{\alpha}H\underline{F} \overset{\text{by(10)}}{\approx} \alpha\underline{F}$, which, together with $\widehat{\alpha}_w - \alpha_w = o_{\mathrm{P}}(1)$, leads to the consistency of prediction.

The consistency result in Theorem 2 does not require a full recovery of all factors. In other words, $\widehat{K}$ is not necessarily equal to K. On the one hand, factors omitted by SPCA are guaranteed to be uncorrelated with y_{t+h}; on the other hand, some factors not useful for prediction may be recovered by SPCA. Obviously, missing any uncorrelated factors or having extra useless factors (for prediction purposes) do not affect the consistency of $\widehat{y}_{T+h}$.

Moreover, this result does not rely on normally distributed error nor on the assumption that all factors share the same strength with respect to all predictors. The assumption on the relative size of N and T is also quite flexible, in contrast with existing results in the literature in which N cannot grow faster than a certain polynomial rate of T, e.g., Bai and Ng (2021), Huang et al. (2021).

3.2 Recovery of All Factors

In this section we develop the asymptotic distribution of $\widehat{y}_{T+h}$ from Algorithm 1. Not surprisingly, the conditions in Theorem 2 are inadequate to guarantee that $\widehat{y}_{T+h}$ converges to $\mathrm{E}_T(y_{T+h})$ at the desirable rate $T^{-1/2}$. The major obstacle lies in the recovery of all factors, which we will illustrate with a one-factor example.

Example 3: Suppose that x_t follows a single-factor model with sparse β:

$$x_t = \begin{bmatrix} \beta_1 \\ \hline \\ 0 \\ \\ \end{bmatrix} f_t + u_t, \quad y_{t+h} = \alpha f_t + z_{t+h},$$

where β_1 is the first N_0 entries of β with $\|\beta_1\| \asymp N_0^{1/2}$ and $\alpha \asymp T^{-1/2}$.

Recall that we use the sample covariance between x_t and y_{t+h} to screen predictors. Even if y_{t+h} is independent of x_t, their sample covariance can be as large as $T^{-1/2}(\log N)^{1/2}$. Therefore, the threshold c needs be strictly greater than $T^{-1/2}(\log N)^{1/2}$ to control Type I error in screening. However, the signal-to-noise ratio in this example is rather low, i.e., $\alpha \asymp T^{-1/2}$, that is, y_{t+h} is not too different from random noise. Consequently, screening will terminate right away because the covariances between y_{t+h} and x_t are at best of order $T^{-1/2}(\log N)^{1/2} < c$, which in turn leads to no discovery of factors. Our procedure thereby gives $\widehat{y}_{T+h} = 0$, which is certainly consistent as the bias $|\mathrm{E}_T(y_{T+h}) - 0| \asymp T^{-1/2}$, but the usual central limit theorem (CLT) fails.

Generally speaking, this issue arises because of the potential failure to recover all factors in the DGP. As long as all factors are found, the bias is negligible and the central limit theorem holds regardless of the magnitude of α. So to go beyond consistency and make valid inference we need a stronger assumption that rules out cases like this, in order to insure against a higher order omitted factor bias that impedes the CLT even if it does not affect consistency. It turns out that as long as $\alpha \in \mathbb{R}^{D\times K}$ satisfies $\lambda_{\min}(\alpha'\alpha) \gtrsim 1$, we can rule out the possibility of missing factors asymptotically. On the one hand, in this case the dimension of target variables, D, must be no smaller than the dimension of the factors, K; and for each factor there exist at least one target variable in y that is correlated with the factor; together they guarantee that no factors would be omitted. On the other hand, our algorithm will not select more factors than needed asymptotically, because the iteration is terminated as soon as all covariances vanish. With a consistent estimator of the number of factors, we can recover the factor space as well as conduct inference on the prediction targets.

The inference theory on strong factor models also relies on a consistent estimator of the count of (strong) factors, e.g., Bai and Ng (2006). Our assumptions here are substantially weaker than the pervasive factor assumption adopted in the literature. That said, in a finite sample, a perfect recovery of the number of factors may be a stretch. In Section 3.4, we show that our version of the PCA regression is more robust than the procedure of Stock and Watson (2002) with respect to the error due to overestimating the number of factors. We also provide simulation evidence on the finite sample performance of our estimator of the number of factors.

The next theorem summarizes a set of stronger asymptotic results under conditions that guarantee perfect recovery of all factors:

Theorem 3. *Under the same assumptions as Theorem 2, if we further have $\lambda_{\min}(\alpha'\alpha) \gtrsim 1$, then for any tuning parameters c and q in (6) and (7) satisfying*

$$c \to 0, \quad c^{-1}(\log NT)^{1/2}(q^{-1/2}N^{-1/2} + T^{-1/2}) \to 0, \quad qN/N_0 \to 0,$$

we have

(i) $\widehat{K}$ defined in Algorithm 1 satisfies: $\mathrm{P}(\widehat{K} = K) \to 1$.

(ii) The factor space is consistently recovered in the sense that

$$\left\|\mathbb{P}_{\underline{\widehat{F}}'} - \mathbb{P}_{\underline{F}'}\right\| = O_{\mathrm{P}}\left(q^{-1/2}N^{-1/2} + T^{-1}\right).$$

(iii) The estimator $\widehat{\gamma}$ constructed via Algorithm 1 satisfies

$$\left\|\widehat{\gamma}\beta - \alpha - T^{-1}\overline{Z}\underline{F}'\Sigma_f^{-1}\right\| = O_{\mathrm{P}}(q^{-1}N^{-1} + T^{-1}).$$

Theorem 3 extends the strong factor case of Bai and Ng (2002) and Bai (2003). In particular, (i) shows that our procedure can recover the true number of factors asymptotically, which extends Bai and Ng (2002) to the case of weak factors. Combining this result with Theorem 1(i) suggests that $\tilde{K} = K$ under the strengthened set of assumptions. We thereby do not need distinguish $\tilde{K}$ with K below. Our setting is distinct from that of Onatski (2010), and as a result we can also recover the space spanned by weak factors, as shown by (ii). This result also suggests that the convergence rate for factor estimation is of order $(qN)^{1/2} \wedge T$, as opposed to $N^{1/2} \wedge T$ given by Theorem 1 of Bai (2003). (iii) extends the result of Theorem 2, replacing the target α by $\alpha + T^{-1}\overline{Z}\underline{F}'\Sigma_f^{-1}$. Note that the latter is precisely a regression estimator of α if F were observable. (iii) thereby points out that the error due to latent factor estimation is no larger than $O_{\mathrm{P}}(q^{-1}N^{-1} + T^{-1})$.

3.3 Inference on the Prediction Target

In the case without observable regressors w, the prediction error can be written as $\widehat{y}_{T+h} - \mathrm{E}_T(y_{T+h}) = (\widehat{\gamma}\beta - \alpha)f_T + \widehat{\gamma}u_T$, where the second term $\widehat{\gamma}u_T$ is of order $(qN)^{-1/2}$. In light of Theorem 3(iii), if $q^{-1}N^{-1}T \to 0$, then the second term is asymptotically negligible (i.e., $o_{\mathrm{P}}(T^{-1/2})$) compared to the first term, $(\widehat{\gamma}\beta - \alpha)f_T = T^{-1}\overline{Z}\underline{F}'\Sigma_f^{-1}f_T + O_{\mathrm{P}}(T^{-1})$, in which case we can achieve root-T inference on $\mathrm{E}_T(y_{T+h})$. Nevertheless, we strive to achieve a better approximation to the finite sample performance by taking into account both terms of the prediction error altogether, without imposing additional restriction on the relative magnitude of qN and T.

To do so, we impose the following assumption:

Assumption 6. *As $N, T \to \infty$, $T^{-1/2}\overline{Z}\underline{F}'$, $T^{-1/2}\overline{Z}\underline{W}'$, and $(qN)^{-1/2}\Psi u_T$ are jointly asymptotically normally distributed, satisfying:*

$$\begin{pmatrix} \mathrm{vec}(T^{-1/2}\overline{Z}\underline{F}') \\ \mathrm{vec}(T^{-1/2}\overline{Z}\underline{W}') \\ (qN)^{-1/2}\Psi u_T \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Pi = \begin{pmatrix} \Pi_{11} & \Pi_{12} & 0 \\ \Pi_{12}' & \Pi_{22} & 0 \\ 0 & 0 & \Pi_{33} \end{pmatrix}\right),$$

where Ψ is a $K \times N$ matrix whose kth row is equal to $b_{(k)}'\beta_{[I_k]}'(\mathbb{I}_N)_{[I_k]}$ and $b_{(k)}$ is the first right singular vector of $\beta_{(k)} = \beta_{[I_k]}\prod_{j<k}\mathbb{M}_{b_{(j)}}$ as defined in Section 3.1.

Assumption 6 characterizes the joint asymptotic distribution of $\overline{Z}\underline{F}'$, $\overline{Z}\underline{W}'$ and Ψu_T. For the first two components, as the dimensions of these random processes are finite, this CLT is a direct result of a large-T central limit theory for mixing processes. With respect to Ψu_T, its large-N asymptotic distribution is assumed normal, asymptotically independent of the distribution of the other two components. This holds trivially if u_{iT}'s are cross-sectionally i.i.d., independent of z_t, w_t, and f_t for $t < T$, so that the kth row of Ψu_T, $b_{(k)}'\beta_{[I_k]}'(u_T)_{[I_k]}$, is a weighted average of u_{iT} for $i \in I_k$. The convergence rate $(qN)^{1/2}$ for Ψu_T arises naturally because $|I_k| = qN$.

Before we present the CLT next, we need define a $K \times K$ matrix $\Omega = (\omega_1, \ldots, \omega_K)$ with $\omega_1 = e_1$ and $\omega_k = e_k - \sum_{i=1}^{k-1} \lambda_{(i)}^{-1} b_{(k)}'\beta_{[I_k]}'\beta_{[I_k]}b_{(i)}\omega_i$, where e_k is a K-dimensional unit vector with 1 on the kth entry and 0 elsewhere.

Theorem 4. *Suppose the same assumptions as in Theorem 3 hold. If in addition, Assumption 6 holds, we have*

$$\Phi^{-1/2}(\widehat{y}_{T+h} - \mathrm{E}_T(y_{T+h})) \xrightarrow{d} \mathcal{N}(0, \mathbb{I}_D),$$

where $\Phi = T^{-1}\Phi_1 + q^{-1}N^{-1}\Phi_2$ and Φ_1 and Φ_2 are given by

$$\begin{aligned}\Phi_1 &= \left((f'_T, w'_T)\Sigma_{f,w}^{-1} \otimes \mathbb{I}_D\right) \begin{pmatrix} \Pi_{11} & \Pi_{12} \\ \Pi'_{12} & \Pi_{22} \end{pmatrix} \left(\Sigma_{f,w}^{-1}(f'_T, w'_T)' \otimes \mathbb{I}_D\right), \\ \Phi_2 &= \alpha B(\Lambda/qN)^{-1}\Omega'\Pi_{33}\Omega(\Lambda/qN)^{-1}B'\alpha',\end{aligned}$$

Π_{ij} is specified by Assumption 6, $\Sigma_{f,w} = \mathrm{diag}(\Sigma_f, \Sigma_w)$, $\Lambda = \mathrm{diag}(\lambda_{(1)}, \ldots, \lambda_{(K)})$, and B is a $K \times K$ matrix whose kth column is given by $b_{(k)}$, where $\lambda_{(k)}^{1/2}$ is the largest singular value of $\beta_{(k)}$ and $b_{(k)}$ is the corresponding right singular vector as defined in Section 3.1.

The convergence rate of $\widehat{y}_{T+h}$ depends on the relative magnitudes of T and qN. For inference, we need construct estimators for each component of Φ_1 and Φ_2. Estimating Φ_1 is straightforward based on its sample analog, constructed from the outputs of Algorithm 1. Estimating Φ_2 is more involved, in that Π_{33} depends on the large covariance matrix of u_T. We leave the details to the appendix.

Algorithm 1 (Step S2.) makes predictions by exploiting the projection of y_{T+h} onto x_T and w_T, with loadings given by γ and $\alpha_w - \gamma\beta_w$. This is convenient and easily extendable out of sample, as both x_T and w_T are directly observable, unlike latent factors. The next section investigates potential issues with plain PCA and PLS, as well as an alternative algorithm based on Stock and Watson (2002), which does not involve the projection parameter γ.

3.4 Alternative Procedures

In this section, we at first discuss the failure of PCA and PLS in the presence of weak factors. To illustrate the issue, it is sufficient to consider a one-factor model example:

Example 4: Suppose that x_t follows a single-factor model with sparse β:

$$x_t = \begin{bmatrix} \beta_1 \\ \hline \\ 0 \\ \\ \end{bmatrix} f_t + u_t, \quad y_{t+h} = \alpha f_t,$$

where β_1 is the first N_0 entries of β with $\|\beta_1\| \asymp N_0^{1/2}$. Moreover, $f_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and $U = \epsilon A$, where ϵ is an $N \times T$ matrix with i.i.d. $\mathcal{N}(0,1)$ entries and A is a $T \times T$ matrix satisfying $\|A\| \lesssim 1$.

3.4.1 Principal Component Regression

Formally, we present the algorithm below:

Algorithm 2 (PCA Regression)**.**
Inputs: $\overline{Y}$, $\underline{X}$, $\underline{W}$, x_T, *and* w_T.

S1. Apply SVD on $\underline{X}\mathbb{M}_{\underline{W}'}$ *and obtain the estimated factors* $\widehat{\underline{F}}_{PCA} = \widehat{\varsigma}'\underline{X}\mathbb{M}_{\underline{W}'}$, *where* $\widehat{\varsigma} \in \mathbb{R}^{N\times K}$ *are the first* K *left singular vectors of* $\underline{X}\mathbb{M}_{\underline{W}'}$. *Estimate the coefficients* $\widehat{\alpha} = \overline{Y}\widehat{\underline{F}}_{PCA}'\left(\widehat{\underline{F}}_{PCA}\widehat{\underline{F}}_{PCA}'\right)^{-1}$.

S2. Obtain $\widehat{\gamma} = \widehat{\alpha}\widehat{\varsigma}'$ *and output the prediction* $\widehat{y}_{T+h}^{PCA} = \widehat{\gamma}x_T + (\widehat{\alpha}_w - \widehat{\gamma}\widehat{\beta}_w)w_T$, *where* $\widehat{\alpha}_w = \overline{Y}\underline{W}'(\underline{WW}')^{-1}$ *and* $\widehat{\beta}_w = \underline{XW}'(\underline{WW}')^{-1}$.

Outputs: $\widehat{y}_{T+h}^{PCA}$, $\widehat{\underline{F}}_{PCA}$, $\widehat{\alpha}$, $\widehat{\alpha}_w$, $\widehat{\beta}_w$, *and* $\widehat{\gamma}$.

Proposition 1. *In Example 4, suppose that* $N/(N_0T) \to \delta \geq 0$ *and* $\|\beta\| \to \infty$ *and define* M *as* $M := T^{-1}\underline{F}'\underline{F} + \delta A_1'A_1$, *where* A_1 *is the first* $T-h$ *columns of* A. *Then, if the two leading eignvalues of* M *are distinct in the sense that* $(\lambda_1(M) - \lambda_2(M))/\lambda_1(M) \gtrsim_{\mathrm{P}} 1$, *the estimated factor* $\widehat{\underline{F}}_{PCA}$ *satisfies*

$$\left\|\mathbb{P}_{\widehat{\underline{F}}_{PCA}'} - \mathbb{P}_{\eta_{PCA}}\right\| \xrightarrow{\mathrm{P}} 0,$$

where η_{PCA} *is the first eigenvector of* M. *In the special case that* $A_1'A_1 = \mathbb{I}_{T-h}$, *it satisfies that*

$$\left\|\mathbb{P}_{\widehat{\underline{F}}_{PCA}'} - \mathbb{P}_{\underline{F}'}\right\| \xrightarrow{\mathrm{P}} 0.$$

Proposition 1 first shows that even if the number of factors is known to be 1, the factor estimated by PCA is in general inconsistent, because the eigenvector η_{PCA} deviates from that of $T^{-1}\underline{F}'\underline{F}$, as the latter is polluted by A. In the special case where error is homoskedastic and has no serial correlation, i.e., $A_1'A_1 = \mathbb{I}_{T-h}$, the estimated factor becomes consistent, in that $\delta A_1'A_1$ in M does not change the eigenvectors of $T^{-1}\underline{F}'\underline{F}$. This result echoes a similar result in Section 4 of Bai (2003), who established the consistency of factors with homoskedasticity and serially independent error even when T is fixed. That said, while factors can be estimated consistently in this special case, the prediction of y_{T+h} based on Algorithm 2 is not consistent.

Proposition 2. *Under the same assumptions as in Proposition 1, if we further assume* $A_1'A_1 = \mathbb{I}_{T-h}$*, then we have* $\widehat{y}^{PCA}_{T+h} \xrightarrow{\text{P}} (1+\delta)^{-1}\mathrm{E}_T(y_{T+h})$.

The reason behind the inconsistency is that even though $\widehat{\underline{F}}_{PCA}$, (effectively the right singular vector of $\underline{X}$) is consistent in the special case, the left singular vector, $\widehat{\varsigma}$ and the singular values are not consistent, which lead to a biased prediction. This result demonstrates the limitation of PC regressions in the presence of weak factor structure.

3.4.2 Partial Least Squares

PCA is an unsupervised approach, in that the PCs are obtained without any information from the prediction target. Therefore, it might be misled by large idiosyncratic errors in x_t when the signal is not sufficiently strong. In contrast with PCA, partial least squares (PLS) is another supervised technique for prediction, which has been shown to work better than PCA in other settings, see, e.g., Kelly and Pruitt (2013). Unlike PCA, PLS uses the information of the response variable when estimating factors. Ahn and Bae (2022) develop its asymptotic properties for prediction in the case of strong factors. We now investigate its asymptotic performance in the same setting above.

The PLS regression algorithm is formulated below:

Algorithm 3 (PLS)**.** *The estimator proceeds as follows:*
Inputs: $\overline{Y}$, $\underline{X}$, $\underline{W}$, x_T, *and* w_T. *Initialization:* $Y_{(1)} := \overline{Y}\mathbb{M}_{\underline{W}'}$, $X_{(1)} := \underline{X}\mathbb{M}_{\underline{W}'}$.

S1. For $k = 1, 2, \cdots, K$, *repeat the following steps using* $X_{(k)}$.

- *a. Obtain the weight vector* $\widehat{\varsigma}_{(k)}$ *from the largest left singular vector of* $X_{(k)}Y'_{(k)}$.
- *b. Estimate the* k*th factor as* $\widehat{\underline{F}}_{(k)} = \widehat{\varsigma}'_{(k)}X_{(k)}$.
- *c. Estimate coefficients* $\widehat{\alpha}_{(k)} = Y_{(k)}\widehat{\underline{F}}'_{(k)}\left(\widehat{\underline{F}}_{(k)}\widehat{\underline{F}}'_{(k)}\right)^{-1}$ *and* $\widehat{\beta}_{(k)} = X_{(k)}\widehat{\underline{F}}'_{(k)}\left(\widehat{\underline{F}}_{(k)}\widehat{\underline{F}}'_{(k)}\right)^{-1}$.
- *e. Remove* $\widehat{\underline{F}}_{(k)}$ *to obtain residuals for the next step:* $X_{(k+1)} = X_{(k)} - \widehat{\beta}_{(k)}\widehat{\underline{F}}_{(k)}$ *and* $Y_{(k+1)} = Y_{(k)} - \widehat{\alpha}_{(k)}\widehat{\underline{F}}_{(k)}$.

S2. Obtain $\widehat{\gamma} = \widehat{\alpha}\widehat{\varsigma}'$ *and the prediction* $\widehat{y}^{PLS}_{T+h} = \widehat{\gamma}x_T + (\widehat{\alpha}_w - \widehat{\gamma}\widehat{\beta}_w)w_T$, *where* $\widehat{\alpha}_w = \overline{Y}\underline{W}'(\underline{W}\underline{W}')^{-1}$ *and* $\widehat{\beta}_w = \underline{X}\underline{W}'(\underline{W}\underline{W}')^{-1}$.

Outputs: $\widehat{y}^{PLS}_{T+h}$, $\widehat{\underline{F}}_{PLS} := (\widehat{\underline{F}}'_{(1)}, \dots, \widehat{\underline{F}}'_{(\widehat{K})})'$, $\widehat{\alpha}$, $\widehat{\alpha}_w$, $\widehat{\beta}_w$, *and* $\widehat{\gamma}$.

The PLS estimator has a closed-form formula if Y is a $1 \times T$ vector and a single factor model is estimated ($K = 1$):

$$\widehat{y}^{PLS}_{T+h} = \left\| \overline{Y}\underline{X}'\underline{X} \right\|^{-2} \overline{Y}\underline{X}'\underline{X}\overline{Y}'\overline{Y}\underline{X}' x_T.$$

While the PLS procedure is intuitively appealing, the next propositions show that this approach produces biased prediction results in the presence of weak factors.

Proposition 3. *In Example 4, suppose that $N/(N_0 T) \to \delta \geq 0$ and $\|\beta\| \to \infty$, then the estimated factor $\widehat{\underline{F}}_{PLS}$ satisfies*

$$\left\| \mathbb{P}_{\widehat{\underline{F}}'_{PLS}} - \mathbb{P}_{\eta_{PLS}} \right\| \xrightarrow{\mathrm{P}} 0,$$

where $\eta_{PLS} = (\mathbb{I}_{T-h} + \delta A_1' A_1)\underline{F}'$. In the special case that $A_1' A_1 = \mathbb{I}_{T-h}$, it satisfies

$$\left\| \mathbb{P}_{\widehat{\underline{F}}'_{PLS}} - \mathbb{P}_{\underline{F}'} \right\| \xrightarrow{\mathrm{P}} 0.$$

Proposition 4. *Under the assumptions of Proposition 3, if we further assume that $A_1' A_1 = \mathbb{I}_{T-h}$, then we have $\widehat{y}^{PLS}_{T+h} \xrightarrow{\mathrm{P}} (1 + \delta)^{-1} \mathrm{E}_T(y_{T+h})$.*

Therefore, the consistency of the PLS factor also depends on the homoskedasticity assumption $A_1' A_1 = \mathbb{I}_{T-h}$ and the forecasting performance of PLS regression is similar to PCA in our weak factor setting. The reason is that the information about the covariance between $\underline{X}$ and $\overline{Y}$ used by PLS is dominated by the noise component of $\underline{X}$, hence PLS does not resolve the issue of weak factors, despite it being a supervised predictor.

Finally, before we conclude the analysis on PLS, we demonstrate a potential issue of PLS due to "overfitting." It turns out that PLS can severely overfit the in-sample data and perform badly out of sample, because PLS overuses information on y to construct its predictor. We illustrate this issue with the following example:

Example 5: Suppose x_t and y_{t+h} follow a "0-factor" model:

$$x_t = u_t, \quad y_{t+h} = z_{t+h},$$

where u_ts follow i.i.d. $\mathcal{N}(0, \mathbb{I}_N)$ and z_ts follow i.i.d. $\mathcal{N}(0, 1)$.

Proposition 5. *In Example 5, if we use $\widehat{K} = 1$, then we have $\widehat{y}^{PLS}_{T+h} \gtrsim_{\mathrm{P}} N^{3/2} T^{1/2}/(N^2 + T^2)$ while $\widehat{y}^{PCA}_{T+h} \lesssim_{\mathrm{P}} 1/(N^{1/2} + T^{1/2})$. Specifically, in the case of $N \asymp T$, $\widehat{y}^{PLS}_{T+h} \gtrsim_{\mathrm{P}} 1$ and $\widehat{y}^{PCA}_{T+h} \lesssim_{\mathrm{P}} N^{-1/2}$.*

The conditional expectation of y_{T+h} is 0 in this example, but $\widehat{y}^{PLS}_{T+h}$ can be bounded away from 0 when using more factors than necessary. In contrast, $\widehat{y}^{PCA}_{T+h}$ remains consistent. The failure of PLS is precisely due to that it selects a component in x that appears correlated with y, despite the fact that there is no correlation between them in this DGP. While SPCA's behavior is difficult to pin down in this example, intuitively, it falls in between these two cases. When q is very large, SPCA resembles PCA as it uses a large number of predictors in x to obtain components. When q is too small, SPCA is prone to overfitting like PLS. With a good choice of q by cross-validation, SPCA can also avoid overfitting.

3.4.3 PCA Regression of Stock and Watson (2002)

Stock and Watson (2002) adopt an alternative version of the PCA regression algorithm (hereafter SW-PCA) to what we have presented in Algorithm 2. The key difference is that SW-PCA conducts PCA on the entire X instead of $\underline{X}$. Therefore, they can obtain $\widehat{f}_T$ directly from this step, instead of reconstructing it using the estimated weights in-sample. While our focus is not on PCA, the PCA algorithm is part of our SPCA procedure. Given the popularity of SW-PCA, we explain why we prefer our version of PCA regression given by Algorithm 2.

Formally, we present their algorithm below:

Algorithm 4 (SW-PCA)**.**
Inputs: $\overline{Y}$*,* X*, and* W*.*

- *S1. Apply SVD on* X*, and obtain the estimated factors* $\widehat{F}_{SW} = \widehat{\varsigma}_*' X \mathbb{M}_{W'}$*, where* $\widehat{\varsigma}_* \in \mathbb{R}^{N\times K}$ *are the first* K *left singular vectors of* X*. Estimate the coefficients by time-series regression:* $\widehat{\alpha} = \overline{Y}\mathbb{M}_{\underline{W}'}\underline{\widehat{F}}'_{SW}\left(\underline{\widehat{F}}_{SW}\mathbb{M}_{\underline{W}'}\underline{\widehat{F}}'_{SW}\right)^{-1}$[6] *and* $\widehat{\alpha}_w = \overline{Y}M_{\underline{\widehat{F}}'_{SW}}\underline{W}'\left(\underline{W}\mathbb{M}_{\underline{\widehat{F}}'_{SW}}\underline{W}'\right)^{-1}$*.*
- *S2. Obtain the prediction* $\widehat{y}^{SW}_{T+h} = \widehat{\alpha}\widehat{f}_T + \widehat{\alpha}_w w_T$*, where* $\widehat{f}_T$ *is the last column of* $\widehat{F}_{SW}$ *and* $\widehat{\alpha}_w = \overline{Y}\underline{W}'(\underline{W}\underline{W}')^{-1}$*.*

Outputs: $\widehat{y}^{SW}_{T+h}$*,* $\widehat{F}_{SW}$*,* $\widehat{\alpha}$*, and* $\widehat{\alpha}_w$*.*

The advantage of SW-PCA is that the consistency of factors is sufficient for the consistency of the prediction, unlike PCA as shown by Proposition 2. In other words, even though this is not true in general, $\widehat{y}^{SW}_{T+h}$ can be consistent in the special case $A'A = \mathbb{I}_T$. Additionally, SW-PCA is more efficient for factor estimation in that it uses the entire data matrices X and W.

Nevertheless, the negative side of the SW-PCA is that it can be unstable because it is more prone to overfitting. We illustrated this issue using the example below.

[6] Unlike Algorithm 1, $\underline{\widehat{F}}_{SW}$ is not orthogonal to $\underline{W}$.

Example 6: Suppose x_t and y_{t+h} follow a "0-factor" model:

$$x_t = u_t, \quad y_{t+h} = z_{t+h},$$

where u_ts are generated from mean zero normal distributions independently with $\mathrm{Cov}(u_t) = \mathbb{I}_N$ for $t < T$ and $\mathrm{Var}(u_T) = (1+\epsilon)\mathbb{I}_N$ for some constant $\epsilon > 0$, and z_ts follow i.i.d. $\mathcal{N}(0,1)$.

Proposition 6. *In Example 6, suppose that $T/N \to 0$, if we use $\widehat{K} = 1$, then we have $\mathrm{Var}(\widehat{y}^{SW}_{T+h}) \to \infty$ and $\widehat{y}^{PCA}_{T+h} \xrightarrow{\mathrm{P}} 0$.*

Intuitively, SW-PCA uses in-sample estimates of the eigenvectors based on data up to T as factors for prediction, whereas PCA uses out-of-sample estimates of the factors, constructed at time T but based on weights estimated up to $T-h$. Because of this, SW-PCA may suffer more from "overfitting" compared to PCA, if the statistical properties of the data differ from $T-h$ to T. Example 6 investigates the case with heteroskedastic u_T in the scenario of overfitting $\widehat{K} = 1 > K = 0$, in which case SW-PCA could perform rather wildly. This example appears contrived, but in practice macroeconomic data are often heterogenous and the number of factors is difficult to pin down. Such an issue is thereby relevant and we hence advocate Algorithms 2 for robustness.

3.5 Tuning Parameter Selection

Along with the gain in robustness to weak factors comes the cost of an extra tuning parameter. To implement the SPCA estimator, we need to select two tuning parameters, q and c. The parameter q dictates the size of the subset used for PCA construction, whereas the parameter c determines the stopping rule, and in turn the number of factors, K. By comparison, PCA and PLS, effectively, only require selecting K. We have established in Theorem 3 that we can consistently recover K, provided q and c satisfy certain conditions.

In practice, we may as well directly tune K instead of c, given that K is more interpretable, that K can only take integer values, and that the scree plot is informative about reasonable ranges of K. Moon and Weidner (2015) demonstrate that, within the context of linear panel regression with interactive fixed effect, the inference on regression coefficients remains robust even with the inclusion of noise as factors. With respect to q, a larger choice of q renders the performance of SPCA resemble that of PCA, and hence becomes less robust to weak factors. Smaller values of q elevate the risk of overfitting, because the selected predictors are more prone to overfit y. We suggest tuning $\lfloor qN \rfloor$ instead of q, because the former can only take integer values, and that multiple choices of the latter may lead to the same integer values of the former.

In our applications, we select tuning parameters based on 3-fold cross-validation that proceeds as follows. We split the entire sample into 3 consecutive folds. Because of the time series dependence, we do not create these folds randomly. We then use each of the three folds, in turn, for validation while the other two are used for training. We select the optimal tuning parameters according to the average R^2 in the validation folds. With these selected parameters, we refit the model using the entire data before making predictions. We conduct a thorough investigation of the effect of tuning on the finite sample performance of all procedures below.

4 Simulations

In this section, we study the finite sample performance of our SPCA procedure using Monte Carlo simulations.

Specifically, we consider a 3-factor DGP as given by equation (1) with two strong factors f_{1t}, f_{2t} and one potentially weak factor f_{3t}. For strong factors f_{1t} and f_{2t}, we generate exposure to them independently from $\mathcal{N}(0,1)$. To simulate a weak factor f_{3t}, we generate exposure to it from a Gaussian mixture distribution, drawing values with probability a from $\mathcal{N}(0,1)$ and $1-a$ from $\mathcal{N}(0,0.1^2)$. The parameter a determines the strength of the third factor and it ranges from $\{0.5, 0.1, 0.05\}$ in the simulations.

Our aim is to predict y_{T+1}, or equivalently, estimate $\mathrm{E}_T(y_{T+1}) = \alpha f_T + \alpha_w w_T$, where w includes an intercept term and a lagged term of y. We consider two DGPs for y. In the first scenario, we set $\alpha_w = (0, 0.2)$ and $\alpha = (0,0,1)$, i.e., $y_{t+1} = f_{3t} + 0.2 y_t + z_{t+1}$. Since y is a univariate target, there is no guarantee that we can recover all factors. We thus examine the consistency of the prediction, as shown in Theorem 2, on the basis of MSE and $\|\widehat{\gamma}\beta - \alpha\|$. In the second scenario, we examine the quality of factor space recovery and inference. We thereby simulate a multivariate target with $\alpha = \mathbb{I}_3$ and $\alpha_w = (0_{3\times 1}, 0.2\mathbb{I}_3)$, i.e., $y_{i,t+1} = f_{it} + 0.2 y_{it} + z_{i,t+1}$, for $i = 1,2,3$.

We generate realizations of f_{it}, z_{it} independently from the standard normal distribution. To generate u_{it}, we first draw ϵ_{it}s from $\mathcal{N}(0,3)$ independently and construct the matrix $A = S\Gamma$, where S is a $(T+1)\times(T+1)$ diagonal matrix with elements drawn from $\mathrm{Unif}(0.5,1.5)$ and Γ is a $(T+1)\times(T+1)$ rotation matrix drawn uniformly from a unit sphere. Therefore, u_{it} as constructed by $U = \epsilon A$ features heteroskedasticity.

Table 1 compares the finite sample performance of SPCA, PCA, and PLS in the first scenario. In both panels, the sample size is $T = 60, 120$, and around $aN = 100$ predictors have exposure to the factor f_{3t}. We simulate $N = 200$ ($a = 0.5$) predictors in the upper panel, so that f_{3t} is exposed to half of them and is thereby strong, and set $N = 2,000$ ($a = 0.05$) in the lower panel, where f_{3t} becomes much weaker due to the large number of predictors that do not load on it.

To highlight the sensitivity of all estimators to the number of factors, we separately report results for each choice of K from 1 to 5 (not tuned), while only selecting the other tuning parameter q for SPCA via cross-validation. We also report results with both parameters tuned jointly for SPCA, and the single parameter K tuned for PCA and PLS, respectively.

The simulation results in Table 1 square well with our theoretical predictions. In the strong factor case (upper panel), PCA and SPCA perform similarly. They achieve minimum prediction error when K is set at the true value 3 in that the first two factors do not predict y. This suggests that tuning q does not worsen the performance of SPCA. PLS can also achieve desirable performance but typically with K smaller than 3. Interestingly, its performance deteriorates rapidly as K increases and surpasses the true value. The reason, as we explain in Proposition 5, is that PLS is more likely to overfit as it uses information about y to directly construct predictors. In contrast, PCA based approaches are more robust to noisy factors used in prediction.

As to the weak factor case (lower panel), SPCA outperforms both PLS and PCA as predicted by our theory. Moreover, SPCA tends to achieve optimal performance when $K = 2$. Recall that in this case, we do not have asymptotic guarantee that SPCA can recover the entire factor space. For this reason, it is possible that a third factor out of this procedure contributes more noise than signal, hence the performance of SPCA deteriorates with an additional factor.

Both panels show that tuning K in most cases slightly deteriorates the optimal prediction MSE and estimation error. That said, the resulting errors remain smaller than what the second best choice of K can achieve.

Furthermore, Table 2 reports the performance of SPCA, PCA, and PLS for each entry of y in the multi-target scenario. In this case we only report results with parameters tuned. As discussed previously, we expect the recovery of all factors using SPCA, because to each factor, at least one entry of y_t has exposure. We first report the distance between $\widehat{F}$ and the true factors F, defined by $d(\widehat{F}, F) = \left\| \mathbb{P}_{\widehat{F}'} - \mathbb{P}_{F'} \right\|$. We also report the MSE$_i$s for $\widehat{y}_{i,T+1}$, $i = 1, 2, 3$, where MSE$_3$ is based on $y_{3,T+1}$, which depends on the potentially weak factor f_{3T} by construction. Again, we vary the value of a and N, while maintaining $aN = 100$, so that the number of predictors with exposure to the third factor is fixed throughout.

The findings here are again consistent with our theory. In particular, as a varies from 0.5 to 0.05, the third factor becomes increasingly difficult to detect. Both PCA and PLS report

Table 1: Finite Sample Comparison of Predictors (Univariate y)

		MSE						$\|\widehat{\gamma}\beta-\alpha\|$					
	K	1	2	3	4	5	$\widehat{K}$	1	2	3	4	5	$\widehat{K}$
T		Panel A: $N=200$ $a=0.5$											
60	SPCA	0.91	0.52	**0.15**	0.17	0.17	0.16	0.92	0.59	**0.24**	0.25	0.25	0.25
	PCA	1.05	1.08	**0.15**	0.15	0.15	0.15	1.01	1.02	0.26	0.26	**0.25**	0.26
	PLS	0.34	**0.17**	0.37	0.51	0.70	0.21	0.50	**0.22**	0.28	0.27	0.27	0.25
120	SPCA	0.89	0.49	**0.09**	0.11	0.11	0.10	0.92	0.55	**0.17**	0.17	0.17	0.17
	PCA	1.04	1.06	**0.09**	0.09	0.09	0.09	1.00	1.01	**0.17**	0.17	0.17	0.17
	PLS	0.25	**0.10**	0.31	0.40	0.66	0.11	0.38	**0.16**	0.26	0.18	0.19	0.16
		Panel B: $N=2000$ $a=0.05$											
60	SPCA	0.75	**0.29**	0.41	0.52	0.58	0.36	0.78	**0.32**	0.42	0.45	0.47	0.36
	PCA	1.11	1.14	0.69	0.67	**0.65**	0.67	1.01	1.03	0.75	0.74	**0.73**	0.74
	PLS	1.14	0.55	**0.52**	0.67	0.75	0.55	1.00	0.56	0.50	0.49	**0.47**	0.51
120	SPCA	0.55	**0.13**	0.18	0.26	0.27	0.16	0.65	**0.19**	0.28	0.34	0.35	0.22
	PCA	1.05	1.08	**0.27**	0.27	0.27	0.27	1.01	1.02	**0.44**	0.44	0.44	0.44
	PLS	0.94	0.24	0.26	0.45	0.55	**0.23**	0.92	0.34	0.30	0.32	0.30	**0.29**

Notes: We evaluate the performance of SPCA, PCA, and PLS in terms of prediction MSE and $\|\widehat{\gamma}\beta-\alpha\|$. All numbers reported are based on averages over 1,000 Monte Carlo repetitions. We highlight the best values based on each criterion in bold.

a substantially larger distance $d(\widehat{F},F)$ than SPCA. In the mean-time, the distortion in the factor space translates to larger prediction errors for the third target y_3, in that it loads on the weak factor f_3 besides its own lag. Throughout this experiment, SPCA maintains almost the same level of performance as a varies, demonstrating its robustness to weak factors.

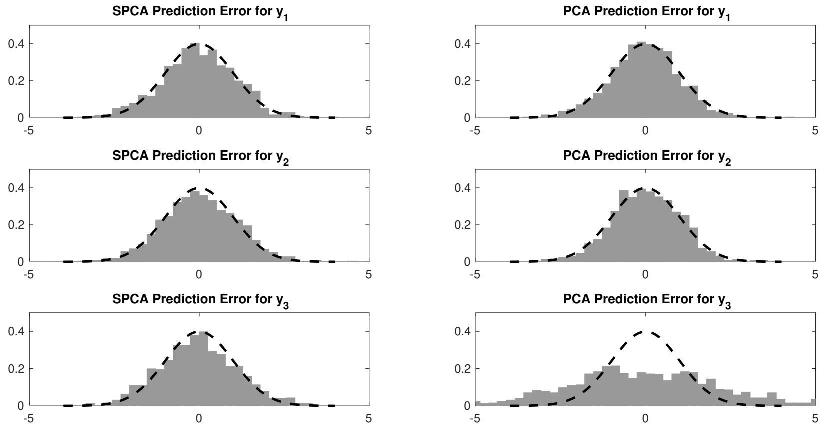
Table 2: Finite Sample Comparison of Predictors (Multivariate y)

	SPCA				PCA				PLS			
a	$d(\widehat{F},F)$	MSE$_1$	MSE$_2$	MSE$_3$	$d(\widehat{F},F)$	MSE$_1$	MSE$_2$	MSE$_3$	$d(\widehat{F},F)$	MSE$_1$	MSE$_2$	MSE$_3$
	$T=60$											
0.5	0.40	0.14	0.16	0.20	0.40	0.14	0.15	0.21	0.41	0.14	0.15	0.19
0.1	0.44	0.13	0.14	0.25	0.55	0.12	0.12	0.55	0.54	0.12	0.13	0.38
0.05	0.45	0.14	0.13	0.27	0.66	0.12	0.11	0.72	0.59	0.12	0.12	0.53
	$T=120$											
0.5	0.30	0.07	0.08	0.10	0.30	0.07	0.08	0.10	0.31	0.08	0.08	0.10
0.1	0.31	0.07	0.07	0.12	0.36	0.06	0.06	0.22	0.39	0.07	0.06	0.17
0.05	0.31	0.07	0.07	0.11	0.39	0.06	0.06	0.29	0.44	0.06	0.06	0.22

Notes: We evaluate the performance of SPCA, PCA, and PLS in terms of the distance between estimated factor space and the true factor space, $d(\widehat{F},F)=\left\|\mathbb{P}_{\widehat{F}'}-\mathbb{P}_{F'}\right\|$, as well as MSE$_i$ for predicting the ith entry of y. All numbers reported are based on averages over 1,000 Monte Carlo repetitions. We vary the value a takes, while fixing $aN=100$.

Last but not least, we report the histograms of the standardized prediction errors using the CLT of Theorem 4 in Figure 1. The setting is identical to that of Table 2 with $a=0.05$ and $T=120$. The histograms match well with the standard normal density for SPCA, and hence verifies the central limit result we derive. As to PCA, there is visible distortion to normality for y_3, due to the presence of the weak factor f_3.

Figure 1: Histograms of the Standardized Prediction Errors



Notes: We provide histograms of standardized prediction errors for each entry of y using SPCA and PCA, respectively, based on 1,000 Monte Carlo repetitions. The dashed curve on each plot corresponds to the standard normal density.

5 Conclusion

The problem of macroeconomic forecasting is central in both academic research as well as for designing policy. The availability of large datasets has spurred the development of methods, pioneered by Stock and Watson (2002), aimed at reducing the dimensionality of the predictors in order to preserve parsimony and achieve better out of sample predictions.

The existing methods that are typically applied to this problem aim to extract a common predictive signal from the large set of available predictors, separating it from the noise and reducing the problem's dimensionality. What our paper adds to this literature is the idea that the availability of a large number of predictors also allows us to *discard* predictors that are not sufficiently informative. That is, predictors that are mostly noise actually hurt the signal extraction because they contaminate the estimation of the common component contained in other, more informative, signals.

How can one know which predictors are noisy and which are useful? The key idea of SPCA is that one can discriminate between useful and noisy predictors by having the target itself guide the selection. This idea, first proposed in Bair and Tibshirani (2004), naturally leads to adding a screening step before factor extraction. But this original version of SPCA only works in very constrained environments that they can all be extracted via PCA from the same subset of predictors.

In practice, there is no guarantee for that to be the case. Whether a latent factor is strong or weak (and *how* strong) depends on how exposed the various predictors are to it – and each empirical applications could feature a different mix of strong and weak latent factors. Therefore, we propose a new SPCA approach that iterates a selection step, a factor extraction step, and a projection step. As we demonstrate in the paper, this procedure can consistently handle a whole range of latent factor strength. Our empirical analysis shows that indeed this procedure fares well in an application with a large number of potentially noisy macroeconomic predictors.

Two final points are worth noting. First, like any procedure, it will work best under some DGPs, and worse under others. In particular, the procedure will potentially miss factors that are extremely weak – no procedure can ever distinguish them from noise, because the exposures of the predictors to these factors are simply too small.

Second, our theory highlights an interesting tradeoff that emerges when working with weak factors. Detecting the weak factors using unsupervised methods (like PCA) is, by definition, difficult or impossible: there is a wide range of strength of factors that will be missed by these methods. Methods based on supervised selection can help extract additional signal, thanks to the guidance from the target. This ability comes at a cost: the possibility of missing factors that are not related to the target. Therefore, this procedure is most useful in applications, like forecasting, where omitting factors not related to the target does not bias the prediction. We leave to future work an additional exploration of other contexts in which SPCA can be useful.

A Mathematical Proofs

For notation simplicity, we use X, F, U, Y, Z in place of $\underline{X}$, $\underline{F}$, $\underline{U}$, $\overline{Y}$, and $\overline{Z}$, and use T_h for $T-h$. In addition, without loss of generality, we assume that $\Sigma_f = \mathbb{I}_K$ in the proof, in that we can always normalize the factors by $\Sigma_f^{-1/2}$ and redefine β in (1) and α in (2) accordingly.

A.1 Proof of Theorem 1

Proof. We start with the DGP without w_t first. Throughout the proof, we use $\widetilde{X}_{(k)} := \left(X_{(k)}\right)_{[\widehat{I}_k]}$ to denote the matrix on which we perform SVD in each step of Algorithm 1. The first left and right singular vectors of $\widetilde{X}_{(k)}$ are denoted by $\widehat{\varsigma}_{(k)}$ and $\widehat{\xi}_{(k)}$, while the largest singular value of $\widetilde{X}_{(k)}$ is denoted by $\sqrt{T_h\widehat{\lambda}_{(k)}}$. As a result, $\widehat{\lambda}_{(k)} = T_h^{-1}\left\|\widetilde{X}_{(k)}\right\|^2$. Moreover, by definition

$$\widehat{\varsigma}_{(k)} = T_h^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}\widetilde{X}_{(k)}\widehat{\xi}_{(k)}, \quad \widehat{\xi}_{(k)} = T_h^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}\widetilde{X}'_{(k)}\widehat{\varsigma}_{(k)}. \tag{12}$$

Therefore, our estimated factor at k-th step is $\widehat{F}_{(k)} = \widehat{\varsigma}'_{(k)}\widetilde{X}_{(k)} = T_h^{1/2}\widehat{\lambda}_{(k)}^{1/2}\widehat{\xi}'_{(k)}$. Consequently, the coefficients of regressing X and Y onto this factor are, respectively:

$$\widehat{\beta}_{(k)} = T_h^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}X_{(k)}\widehat{\xi}_{(k)} \quad \text{and} \quad \widehat{\alpha}_{(k)} = T_h^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}Y_{(k)}\widehat{\xi}_{(k)}. \tag{13}$$

Then we define $\widetilde{D}_{(k)} \in \mathbb{R}^{qN\times N}$ iteratively by

$$\widetilde{D}_{(k)} = (\mathbb{I}_N)_{[\widehat{I}_k]} - \sum_{i=1}^{k-1} T_h^{-1/2}\widehat{\lambda}_{(i)}^{-1/2}X_{[\widehat{I}_k]}\widehat{\xi}_{(i)}\widehat{\varsigma}'_{(i)}\widetilde{D}_{(i)},$$

with $\widetilde{D}_{(1)} = (\mathbb{I}_N)_{[\widehat{I}_1]}$. We can show by induction that $\widetilde{X}_{(k)} = \widetilde{D}_{(k)}X$. In fact, by Lemma 1, we have $\widehat{\xi}'_{(i)}\widehat{\xi}_{(j)} = 0$ for $i \neq j \leq \widehat{K}$ which suggests that $\widehat{F}_{(k)}$'s for all k are pairwise orthogonal. Using this property and the definition of $\widetilde{X}_{(k)}$, we have

$$\widetilde{X}_{(k)} = \big(X_{(k)}\big)_{[\widehat{I}_k]} = X_{[\widehat{I}_k]}\prod_{i=1}^{k-1}\mathbb{M}_{\widehat{F}'_{(i)}} = X_{[\widehat{I}_k]}\left(\mathbb{I}_{T_h} - \sum_{i=1}^{k-1}\widehat{\xi}_{(i)}\widehat{\xi}'_{(i)}\right), \tag{14}$$

for $k > 1$ and when $k = 1$,

$$\widetilde{X}_{(1)} = X_{[\widehat{I}_1]} = \beta_{[\widehat{I}_1]}F + U_{[\widehat{I}_1]}.$$

Using (12), if $\widetilde{X}_{(i)} = \widetilde{D}_{(i)}X$ for any $i < k$ we can write (14) as

$$\widetilde{X}_{(k)} = X_{[\widehat{I}_k]}\left(\mathbb{I}_{T_h} - \sum_{i=1}^{k-1}\widehat{\xi}_{(i)}\widehat{\xi}'_{(i)}\right) = X_{[\widehat{I}_k]} - \sum_{i=1}^{k-1}T_h^{-1/2}\widehat{\lambda}_{(i)}^{-1/2}X_{[\widehat{I}_k]}\widehat{\xi}_{(i)}\widehat{\varsigma}'_{(i)}\widetilde{X}_{(i)} = \widetilde{D}_{(k)}X.$$

Since $\widetilde{X}_{(1)} = X_{[\widehat{I}_1]} = \widetilde{D}_{(1)}X$ holds immediately by definition, we have $\widetilde{X}_{(k)} = \widetilde{D}_{(k)}X$ by induction. In light of this, the estimated factors satisfy

$$\widehat{F}_{(k)} = \widehat{\varsigma}'_{(k)}\widetilde{X}_{(k)} = \widehat{\varsigma}'_{(k)}\widetilde{D}_{(k)}X, \tag{15}$$

for all k, and by definition, we have $\widehat{\zeta}_{(k)} = (\widehat{\varsigma}'_{(k)}\widetilde{D}_{(k)})'$. Moreover, using (13) the estimated coefficient $\widehat{\gamma}$ can be written as

$$\widehat{\gamma} = \sum_{k=1}^{\widehat{K}}\widehat{\alpha}_{(k)}\widehat{\varsigma}'_{(k)}\widetilde{D}_{(k)} = \sum_{k=1}^{\widehat{K}}T_h^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}Y\widehat{\xi}_{(k)}\widehat{\varsigma}'_{(k)}\widetilde{D}_{(k)}. \tag{16}$$

We further define $\widetilde{\beta}_{(k)} = \widetilde{D}_{(k)}\beta$ and $\widetilde{U}_{(k)} = \widetilde{D}_{(k)}U$, then $\widetilde{X}_{(k)}$ can be written in the form of

$$\widetilde{X}_{(k)} = \widetilde{\beta}_{(k)}F + \widetilde{U}_{(k)}. \tag{17}$$

We also define the population analog of $\widetilde{D}_{(k)}$ for each k by

$$D_{(k)} = (\mathbb{I}_N)_{[I_k]} - \sum_{i=1}^{k-1} \lambda_{(i)}^{-1/2} \beta_{[I_k]} b_{(i)} \varsigma'_{(i)} D_{(i)}, \quad D_{(1)} = (\mathbb{I}_N)_{[I_1]},$$

where $\sqrt{\lambda_{(k)}}$ is the leading singular value of $\beta_{(k)}$, $\varsigma_{(k)}$ and $b_{(k)}$ are the corresponding left and right singular vectors of $\beta_{(k)}$. By a similar induction argument, we can show that

$$\beta_{(k)} = \beta_{[I_k]} \prod_{i<k} \mathbb{M}_{b_{(i)}} = D_{(k)}\beta.$$

Intuitively, $\widetilde{\beta}_{(k)}$ and $\widetilde{D}_{(k)}$ are sample analogs of $\beta_{(k)}$ and $D_{(k)}$.

Similar representations to (17) can be constructed for $Y_{(k)} := Y \prod_{i=1}^{k-1} \mathbb{M}_{\widehat{F}'_{(i)}}$ for each k. Specifically, we have

$$Y_{(k)} = Y \left(\mathbb{I}_{T_h} - \sum_{i=1}^{k-1} \widehat{\xi}_{(i)} \widehat{\xi}'_{(i)} \right) = \widetilde{\alpha}_{(k)} F + \widetilde{Z}_{(k)}, \tag{18}$$

where $\widetilde{\alpha}_{(k)} \in \mathbb{R}^{D \times K}$ and $\widetilde{Z}_{(k)} \in \mathbb{R}^{D \times T_h}$ are defined as

$$\widetilde{\alpha}_{(k)} := \alpha - \sum_{i=1}^{k-1} T_h^{-1/2} \widehat{\lambda}_{(i)}^{-1/2} Y \widehat{\xi}_{(i)} \widehat{\varsigma}'_{(i)} \widetilde{\beta}_{(i)} \text{ and } \widetilde{Z}_{(k)} := Z - \sum_{i=1}^{k-1} T_h^{-1/2} \widehat{\lambda}_{(i)}^{-1/2} Y \widehat{\xi}_{(i)} \widehat{\varsigma}'_{(i)} \widetilde{U}_{(i)}.$$

By Lemma 3, we have $\mathrm{P}(\widehat{I}_k = I_k) \to 1$ for $k \leq \tilde{K}$ and $\mathrm{P}(\widehat{K} = \tilde{K}) \to 1$. Thus, with probability approaching one, we can impose that $\widehat{I}_k = I_k$ for any k and $\widehat{K} = \tilde{K}$ in what follows.

To prove Theorem 1, using (17), the estimated factors can be written as

$$\widehat{F}_{(k)} = \widehat{\varsigma}'_{(k)} \widetilde{X}_{(k)} = \widehat{\varsigma}'_{(k)} \widetilde{\beta}_{(k)} F + \widehat{\varsigma}'_{(k)} \widetilde{U}_{(k)}.$$

Using Lemma 5(i), $\left\| \widehat{F}_{(k)} \right\| = \sqrt{T_h \widehat{\lambda}_{(k)}}$, and $\|\mathbb{M}_{F'}\| \leq 1$, we have

$$\left\| \widehat{F}_{(k)} \right\|^{-1} \left\| \widehat{F}_{(k)} \mathbb{M}_{F'} \right\| \leq \left\| \widehat{F}_{(k)} \right\|^{-1} \left\| \widehat{\varsigma}'_{(k)} \widetilde{U}_{(k)} \right\| \lesssim_{\mathrm{P}} q^{-1/2} N^{-1/2} + T^{-1}.$$

□

A.2 Proof of Theorem 2

Proof. By definition of $X_{(k)}$ in Algorithm 1, we have

$$X_{(k)} = X_{(k-1)}\mathbb{M}_{\widehat{F}'_{(k-1)}} = X\prod_{i=1}^{k-1}\mathbb{M}_{\widehat{F}'_{(i)}} = X\left(\mathbb{I}_{T_h} - \sum_{i=1}^{k-1}\widehat{\xi}_{(i)}\widehat{\xi}'_{(i)}\right).$$

Therefore, using (18), we have

$$X_{(k)}Y'_{(k)} = X\left(\mathbb{I}_{T_h} - \sum_{i=1}^{k-1}\widehat{\xi}_{(i)}\widehat{\xi}'_{(i)}\right)Y'_{(k)} = XY'_{(k)}$$

as $Y_{(k)}\widehat{\xi}_{(i)} = 0$ for $i < k$ by Lemma 1. Therefore, the covariance $\left(X_{(k)}\right)_{[i]}Y'_{(k)}$ for each predictor equals to $X_{[i]}Y'_{(k)}$. Based on the stopping rule, if our algorithm stops at $\tilde{K}$, there are at most $qN-1$ predictors among all satisfying $T_h^{-1}\left\|X_{[i]}Y'_{(\tilde{K}+1)}\right\|_{\mathrm{MAX}} \geq c$. Let S denote the set of these predictors. For $i \in S$, we have

$$\left\|T_h^{-1}X_{[i]}Y'_{(\tilde{K}+1)}\right\|_{\mathrm{F}}^2 \lesssim \left\|T_h^{-1}X_{[i]}Y'_{(\tilde{K}+1)}\right\|_{\mathrm{MAX}}^2 \lesssim_{\mathrm{P}} 1, \tag{19}$$

where we use $\|\beta\|_{\mathrm{MAX}} \lesssim 1$ from Assumption 2 and Lemma 3(vi) in the last step. On the other hand, in light of the set I_0 in Assumption 2, we have

$$\begin{aligned}\sum_{i\in I_0}\left\|T_h^{-1}X_{[i]}Y'_{(\tilde{K}+1)}\right\|_{\mathrm{F}}^2 &= \sum_{i\in I_0\cap S}\left\|T_h^{-1}X_{[i]}Y'_{(\tilde{K}+1)}\right\|_{\mathrm{F}}^2 + \sum_{i\in I_0\cap S^c}\left\|T_h^{-1}X_{[i]}Y'_{(\tilde{K}+1)}\right\|_{\mathrm{F}}^2 \\ &\lesssim_{\mathrm{P}} |I_0\cap S| + |I_0\cap S^c|c^2 \leq qN + c^2N_0 = o(N_0),\end{aligned} \tag{20}$$

where we use (19), $|S| \leq qN-1$, $c \to 0$, and $qN/N_0 \to 0$. Consequently, (20) leads to $\left\|Y_{(\tilde{K}+1)}X'_{[I_0]}\right\| = o_{\mathrm{P}}(TN_0^{1/2})$. Moreover, using (18) and that $X = \beta F + U$, we can decompose

$$Y_{(\tilde{K}+1)}X'_{[I_0]} = \widetilde{\alpha}_{(\tilde{K}+1)}FF'\beta'_{[I_0]} + \widetilde{\alpha}_{(\tilde{K}+1)}FU'_{[I_0]} + \widetilde{Z}_{(\tilde{K}+1)}F'\beta'_{[I_0]} + \widetilde{Z}_{(\tilde{K}+1)}U'_{[I_0]}. \tag{21}$$

Using (20), (21), Lemma 9(i)(ii), and the fact that $\left\|\beta_{[I_0]}\right\| \lesssim N_0^{1/2}$, we have

$$\left\|\widetilde{\alpha}_{(\tilde{K}+1)}\left(FF'\beta'_{[I_0]} + FU'_{[I_0]}\right)\right\| = o_{\mathrm{P}}\left(N_0^{1/2}T\right). \tag{22}$$

Also, using Assumption 4(i), Assumption 1(i) and Weyl's theorem, we have

$$
\begin{aligned}
|\sigma_K(FF'\beta'_{[I_0]} + FU'_{[I_0]}) - \sigma_K(T_h\beta_{[I_0]})| \leq \left\|FU'_{[I_0]}\right\| + \left\|T_h^{-1}FF' - \mathbb{I}_K\right\| \left\|T_h\beta_{[I_0]}\right\| \\
\lesssim_{\mathrm{P}} N_0^{1/2}T^{1/2}. \qquad (23)
\end{aligned}
$$

Since Assumption 2 implies that $\sigma_K(\beta_{[I_0]}) \asymp N_0^{1/2}$, we have $\sigma_K(FF'\beta'_{[I_0]} + FU'_{[I_0]}) \asymp N_0^{1/2}T$. Using this result, (22) and the inequality $\left\|\widetilde{\alpha}_{(\tilde{K}+1)}\left(FF'\beta'_{[I_0]} + FU'_{[I_0]}\right)\right\| \geq \sigma_K(FF'\beta_{[I_0]} + FU'_{[I_0]})\left\|\widetilde{\alpha}_{(\tilde{K}+1)}\right\|$, we have $\left\|\widetilde{\alpha}_{(\tilde{K}+1)}\right\| \xrightarrow{\mathrm{P}} 0$. That is, by definition of $\tilde{\alpha}_{(\tilde{K}+1)}$ in (18),

$$
\left\|\alpha - \sum_{i=1}^{\tilde{K}} Y\widehat{\xi}_{(i)}\frac{\widehat{\varsigma}'_{(i)}\widetilde{\beta}_{(i)}}{\sqrt{T_h\widehat{\lambda}_{(i)}}}\right\| = o_{\mathrm{P}}(1). \qquad (24)
$$

Next, (16) and $\widetilde{\beta}_{(k)} = \widetilde{D}_{(k)}\beta$ imply that

$$
\widehat{\gamma}\beta = \sum_{i=1}^{\tilde{K}} T_h^{-1/2}\widehat{\lambda}_{(i)}^{-1/2}Y\widehat{\xi}_{(i)}\widehat{\varsigma}'_{(i)}\widetilde{\beta}_{(i)}.
$$

Therefore, (24) is equivalent to $\|\widehat{\gamma}\beta - \alpha\| = o_{\mathrm{P}}(1)$.

As shown in Lemma 12, Assumptions 1, 3, and 4 hold when we replace F, Z and U by $F\mathbb{M}_{W'}$, $Z\mathbb{M}_{W'}$ and $U\mathbb{M}_{W'}$. Therefore all of the lemmas and the result $\|\widehat{\gamma}\beta - \alpha\| = o_{\mathrm{P}}(1)$ also hold when w_t is included. We write the prediction error of y_{T+h} as

$$
\begin{aligned}
&\widehat{y}_{T+h} - \mathrm{E}_T(y_{T+h}) = \widehat{\gamma}x_T + (\widehat{\alpha}_w - \widehat{\gamma}\widehat{\beta}_w)w_T - \alpha f_T - \alpha_w w_T \qquad (25)\\
=&(\widehat{\gamma}\beta - \alpha)\left(f_T - FW'(WW')^{-1}w_T\right) + \widehat{\gamma}(u_T - UW'(WW')^{-1}w_T) + ZW'(WW')^{-1}w_T.
\end{aligned}
$$

Using (16) and $\|Y\| \leq \|\alpha F\| + \|Z\| \lesssim_{\mathrm{P}} T^{1/2}$ by Assumption 1, we have

$$
\|\widehat{\gamma}u_T\| \leq \sum_{k\leq\tilde{K}} T_h^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}\|Y\|\left\|\widehat{\xi}_{(k)}\right\|\left\|\widehat{\varsigma}'_{(k)}\widetilde{D}_{(k)}u_T\right\| \lesssim_{\mathrm{P}} \sum_{k\leq\tilde{K}} \widehat{\lambda}_{(k)}^{-1/2}\left\|\widehat{\varsigma}'_{(k)}\widetilde{D}_{(k)}u_T\right\|, \qquad (26)
$$

and

$$
T_h^{-1}\|\widehat{\gamma}UW'\| \leq \sum_{k\leq\tilde{K}} T_h^{-3/2}\widehat{\lambda}_{(k)}^{-1/2}\|Y\|\left\|\widehat{\xi}_{(k)}\right\|\left\|\widehat{\varsigma}'_{(k)}\widetilde{D}_{(k)}UW'\right\| \lesssim_{\mathrm{P}} \sum_{k\leq\tilde{K}} T_h^{-1}\widehat{\lambda}_{(k)}^{-1/2}\left\|\widehat{\varsigma}'_{(k)}\widetilde{U}_{(k)}W'\right\|.
$$

(27)

Using $\widehat{\lambda}_{(k)} \asymp_{\mathrm{P}} qN$ from Lemma 3 and Lemma 5(ii)(iv), we have

$$T_h^{-1}\widehat{\lambda}_{(k)}^{-1/2}\left\|\widehat{\varsigma}_{(k)}\widetilde{U}_{(k)}W'\right\| \lesssim_{\mathrm{P}} q^{-1}N^{-1} + T^{-1}, \widehat{\lambda}_{(k)}^{-1/2}\left\|\widehat{\varsigma}'_{(k)}\widetilde{D}_{(k)}u_T\right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1/2}. \quad (28)$$

Therefore, $\|\widehat{\gamma}u_T\| = o_{\mathrm{P}}(1)$. Furthermore, with $\|(WW')^{-1}\| \lesssim_{\mathrm{P}} T^{-1}$ from Assumption 1, we have $\|\widehat{\gamma}UW'(WW')^{-1}\| = o_{\mathrm{P}}(1)$. Together with $\|FW'\| \lesssim_{\mathrm{P}} T^{1/2}$, $\|ZW'\| \lesssim_{\mathrm{P}} T^{1/2}$ from Assumption 1 and $\|\widehat{\gamma}\beta - \alpha\| = o_{\mathrm{P}}(1)$, we show that each term of (25) vanishes, and hence $\widehat{y}_{T+h} - \mathrm{E}_T[y_{T+h}] \xrightarrow{\mathrm{P}} 0$. □

A.3 Proof of Theorem 3

Proof. As in the proof of Theorem 1, we impose that $\widehat{K} = \tilde{K}$ and $\widehat{I}_k = I_k$, since Lemma 3 shows that both events occur with probability approaching 1. As shown in Lemma 2(iv), under the assumption that $\lambda_K(\alpha'\alpha) \gtrsim 1$, we have $\tilde{K} = K$. Together with $\mathrm{P}(\widehat{K} = \tilde{K}) \to 1$, we have obtained (i) of Theorem 3. Below we directly impose that $\widehat{K} = K$.

Again, following the same argument above (25), we only need analyze the case without w_t. As $\widehat{F}_{(k)} = T_h^{1/2}\widehat{\lambda}_{(k)}^{1/2}\widehat{\xi}'_{(k)}$, Theorem 1 implies $\left\|\widehat{\xi}'_{(k)}\mathbb{M}_{F'}\right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1}$ for $k \leq K$. Let v denote $F'(FF')^{-1/2}$, we have

$$\left\|\widehat{\xi} - \mathbb{P}_{F'}\widehat{\xi}\right\| = \left\|\widehat{\xi} - vv'\widehat{\xi}\right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1}, \quad (29)$$

where $\widehat{\xi}$ is a $T \times K$ matrix with each column equal to $\widehat{\xi}_{(k)}$. (29) implies that $\left\|\widehat{\xi}'vv'\widehat{\xi} - \mathbb{I}_K\right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1}$. By Weyl's inequality, $|\sigma_i(\widehat{\xi}'v) - 1| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1}$, for $1 \leq i \leq K$, and thus

$$\begin{aligned}\left\|v - \widehat{\xi}\widehat{\xi}'v\right\| \leq & \sigma_K^{-1}(v'\widehat{\xi})\left\|vv'\widehat{\xi} - \widehat{\xi}\widehat{\xi}'vv'\widehat{\xi}\right\| \lesssim_{\mathrm{P}} \left\|vv'\widehat{\xi} - \widehat{\xi}\right\| + \left\|\widehat{\xi}(\widehat{\xi}'vv'\widehat{\xi} - \mathbb{I}_K)\right\| \\ \lesssim_{\mathrm{P}} & q^{-1/2}N^{-1/2} + T^{-1}.\end{aligned}$$

Then, using this, (29), and the fact that $\|v\| = 1$ and $\left\|\widehat{\xi}\right\| = 1$, we have

$$\left\|\mathbb{P}_{\widehat{F}'} - \mathbb{P}_{F'}\right\| = \left\|\widehat{\xi}\widehat{\xi}' - vv'\right\| \leq \left\|\widehat{\xi}(\widehat{\xi} - vv'\widehat{\xi})'\right\| + \left\|(\widehat{\xi}\widehat{\xi}'v - v)v'\right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1}.$$

Next, we need a more intricate analysis of $\widehat{\gamma}$. Recall from the proof of Theorem 2 that

$$\widehat{\gamma}\beta = \sum_{k=1}^{\tilde{K}} T_h^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}Y\widehat{\xi}_{(k)}\widehat{\varsigma}'_{(k)}\widetilde{\beta}_{(k)}. \quad (30)$$

Denote $B_1 = (b_{11}, \ldots, b_{\widehat{K}1}) \in \mathbb{R}^{K \times \widehat{K}}$, $B_2 = (b_{12}, \ldots, b_{\widehat{K}2}) \in \mathbb{R}^{K \times \widehat{K}}$, where

$$b_{k1} = T^{-1/2} F \widehat{\xi}_{(k)}, \quad b_{k2} = \widehat{\lambda}_{(k)}^{-1/2} \widetilde{\beta}'_{(k)} \widehat{\varsigma}_{(k)}. \tag{31}$$

By Lemma 6,

$$\left\| T_h^{-1/2} Z \widehat{\xi}_{(k)} - T_h^{-1} Z F' b_{k2} \right\| \lesssim_{\mathrm{P}} T^{-1} + q^{-1} N^{-1}. \tag{32}$$

As we impose that $\widehat{K} = \tilde{K} = K$, combining (30), (31) and (32), with $\|B_1\| \lesssim_{\mathrm{P}} 1$, $\|B_2\| \lesssim_{\mathrm{P}} 1$ from Lemma 10, we have

$$\left\| \widehat{\gamma} \beta - \alpha B_1 B_2' - T_h^{-1} Z F' B_2 B_2' \right\| \lesssim_{\mathrm{P}} T^{-1} + q^{-1} N^{-1}. \tag{33}$$

Using Lemma 10(iv)(v), we obtain $\left\| \widehat{\gamma} \beta - \alpha - T_h^{-1} Z F' \right\| \lesssim_{\mathrm{P}} T^{-1} + q^{-1} N^{-1}$. □

A.4 Proof of Theorem 4

Proof. As in the proof of Theorem 2, we have $\|FW'(WW')^{-1}\| \lesssim_{\mathrm{P}} T^{-1/2}$ from Assumption 1 and $\|\widehat{\gamma} U W'(WW')^{-1}\| \lesssim_{\mathrm{P}} T^{-1} + q^{-1} N^{-1}$ as shown in (27) and (28). Together with $\left\| \widehat{\gamma} \beta - \alpha - T_h^{-1} Z F' \right\| \lesssim_{\mathrm{P}} T^{-1} + q^{-1} N^{-1}$, we can derive from (25) that:

$$\widehat{y}_{T+h} - \mathrm{E}_T(y_{T+h}) = T_h^{-1} Z F' f_T + Z W'(WW')^{-1} w_T + \widehat{\gamma} u_T + O_{\mathrm{P}}(T^{-1} + q^{-1} N^{-1}).$$

By Assumption 1, we have $|\lambda_i \left(T_h^{-1} \Sigma_w^{-1/2} W W' \Sigma_w^{-1/2} \right) - 1| \lesssim_{\mathrm{P}} T^{-1/2}$ and thus

$$\begin{aligned} & \left\| Z W'(WW')^{-1} w_T - T_h^{-1} Z W' \Sigma_w^{-1} w_T \right\| \leq T_h^{-1} \left\| Z W' \right\| \left\| (T_h^{-1} W W')^{-1} - \Sigma_w^{-1} \right\| \left\| w_T \right\| \\ \lesssim_{\mathrm{P}} & T_h^{-1/2} \left\| T_h^{-1} \Sigma_w^{-1/2} W W' \Sigma_w^{-1/2} - \mathbb{I}_D \right\| = T_h^{-1/2} \max_{i \leq D} |\lambda_i \left(T_h^{-1} \Sigma_w^{-1/2} W W' \Sigma_w^{-1/2} \right)^{-1} - 1| \\ \lesssim_{\mathrm{P}} & T^{-1}. \end{aligned} \tag{34}$$

For $\widehat{\gamma} u_T$, by (16), we have $\widehat{\gamma} u_T = \sum_{k=1}^{K} \widehat{\alpha}_{(k)} \widehat{\varsigma}'_{(k)} \widetilde{D}_{(k)} u_T$ and thus

$$\left\| \widehat{\gamma} u_T - \sum_{k=1}^{K} \lambda_{(k)}^{-1/2} \alpha b_{(k)} \varsigma'_{(k)} D_{(k)} u_T \right\| \leq \sum_{k=1}^{K} \left\| \widehat{\alpha}_{(k)} \widehat{\varsigma}'_{(k)} \widetilde{D}_{(k)} u_T - \lambda_{(k)}^{-1/2} \alpha b_{(k)} \varsigma'_{(k)} D_{(k)} u_T \right\|. \tag{35}$$

Lemma 8(vi) gives

$$q^{-1/2} N^{-1/2} |\widehat{\varsigma}'_{(k)} \widetilde{D}_{(k)} u_T - \varsigma'_{(k)} D_{(k)} u_T| \lesssim_{\mathrm{P}} T^{-1} + q^{-1} N^{-1}. \tag{36}$$

In addition, (13) and Lemma 1 give $\widehat{\lambda}_{(k)}^{1/2}\widehat{\alpha}_{(k)} = T_h^{-1/2}Y\widehat{\xi}_{(k)} = \alpha b_{k1} + T_h^{-1/2}Z\widehat{\xi}_{(k)}$. With (32), $\|ZF'\| \lesssim_{\mathrm{P}} T^{1/2}$ and $\|b_{k2}\| \lesssim_{\mathrm{P}} 1$ from Lemma 10(i), this equation leads to

$$\left\|\widehat{\lambda}_{(k)}^{1/2}\widehat{\alpha}_{(k)} - \alpha b_{k1}\right\| \leq \left\|T_h^{-1/2}Z\widehat{\xi}_{(k)} - T_h^{-1}ZF'b_{k2}\right\| + \left\|T_h^{-1}ZF'b_{k2}\right\| \lesssim_{\mathrm{P}} T^{-1/2} + q^{-1}N^{-1}.$$

Using $\left\|b_{k2} - b_{(k)}\right\| \lesssim_{\mathrm{P}} T^{-1/2} + q^{-1/2}N^{-1/2}$ implied by Lemma 10(iii) and $\widehat{\lambda}_{(k)} \asymp_{\mathrm{P}} qN$ from Lemma 3(iii), we have

$$\begin{aligned}\left\|\widehat{\alpha}_{(k)} - \widehat{\lambda}_{(k)}^{-1/2}\alpha b_{(k)}\right\| &\leq \left\|\widehat{\alpha}_{(k)} - \widehat{\lambda}_{(k)}^{-1/2}\alpha b_{k2}\right\| + \left\|\widehat{\lambda}_{(k)}^{-1/2}\alpha(b_{(k)} - b_{k2})\right\| \\ &\lesssim_{\mathrm{P}} T^{-1/2}q^{-1/2}N^{-1/2} + q^{-1}N^{-1}. \end{aligned} \tag{37}$$

Also, with Lemma 3(iii), we have

$$|\widehat{\lambda}_{(k)}^{-1/2} - \lambda_{(k)}^{-1/2}| \leq \widehat{\lambda}_{(k)}^{-1/2}|\widehat{\lambda}_{(k)}^{1/2}/\lambda_{(k)}^{1/2} - 1| \lesssim_{\mathrm{P}} T^{-1/2}q^{-1/2}N^{-1/2} + q^{-1}N^{-1}.$$

Since $\left\|b_{(k)}\right\| = 1$, the above two inequalities lead to

$$\left\|\widehat{\alpha}_{(k)} - \lambda_{(k)}^{-1/2}\alpha b_{(k)}\right\| \leq T^{-1/2}q^{-1/2}N^{-1/2} + q^{-1}N^{-1}. \tag{38}$$

For each term in the summation of (35), we have

$$\begin{aligned}&\left\|\widehat{\alpha}_{(k)}\widehat{\varsigma}'_{(k)}\widetilde{D}_{(k)}u_T - \lambda_{(k)}^{-1/2}\alpha b_{(k)}\varsigma'_{(k)}D_{(k)}u_T\right\| \\ \leq &\left\|\widehat{\alpha}_{(k)}(\widehat{\varsigma}'_{(k)}\widetilde{D}_{(k)}u_T - \varsigma'_{(k)}D_{(k)}u_T)\right\| + \left\|(\widehat{\alpha}_{(k)} - \lambda_{(k)}^{-1/2}\alpha b_{(k)})\varsigma'_{(k)}D_{(k)}u_T\right\|. \end{aligned} \tag{39}$$

Note that (37) also implies $\left\|\widehat{\alpha}_{(k)}\right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2}$ as $\widehat{\lambda}_{(k)} \asymp qN$, and that (36) implies the first term in (39) is $O_{\mathrm{P}}(T^{-1} + q^{-1}N^{-1})$. Furthermore, $|\varsigma'_{(k)}D_{(k)}u_T| \lesssim_{\mathrm{P}} 1$ from Lemma 5(iv) and (38) show that the second term in (39) is also $O_{\mathrm{P}}(T^{-1} + q^{-1}N^{-1})$. Given this, (35) becomes

$$\left\|\widehat{\gamma}u_T - \sum_{k=1}^{K}\lambda_{(k)}^{-1/2}\alpha b_{(k)}\varsigma'_{(k)}D_{(k)}u_T\right\| \lesssim_{\mathrm{P}} T^{-1} + q^{-1}N^{-1}. \tag{40}$$

To sum up, we have established that

$$\widehat{y}_{T+h} - \mathrm{E}_T(y_{T+h}) = \frac{ZF'}{T_h}f_T + \frac{ZW'}{T_h}\Sigma_w^{-1}w_T + \sum_{k=1}^{K}\lambda_{(k)}^{-1/2}\alpha b_{(k)}\varsigma'_{(k)}D_{(k)}u_T + O_{\mathrm{P}}\left(T^{-1} + q^{-1}N^{-1}\right).$$

In the general case that Σ_f may not be $\mathbb{I}_K$, the first term becomes $T_h^{-1}ZF'\Sigma_f^{-1}f_T$. Using the fact $\varsigma_{(k)} = \lambda_{(k)}^{-1/2}\beta_{(k)}b_{(k)} = \lambda_{(k)}^{-1/2}\beta_{[I_k]}b_{(k)}$ and the iterative definition of $D_{(k)}$, we can see that $\lambda_{(k)}^{-1/2}\varsigma_{(k)}' D_{(k)}u_T$ is exactly the kth row of $\Lambda^{-1}\Omega'\Psi u_T$ with Λ, Ω, and Ψ defined in Theorem 4. Using Delta method and Assumption 6, it is straightforward to obtain the desired CLT. □

References

Ahn, S. C. and J. Bae (2022). Forecasting with partial least squares when a large number of predictors are available. Technical report, Arizona State University and University of Glasgow.

Amini, A. A. and M. J. Wainwright (2009, October). High-dimensional analysis of semidefinite relaxations for sparse principal components. *Annals of Statistics 37*(5B), 2877–2921.

Bai, J. (2003). Inferential Theory for Factor Models of Large Dimensions. *Econometrica 71*(1), 135–171.

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

Bai, J. and S. Ng (2006). Determining the number of factors in approximate factor models, erata. http://www.columbia.edu/ sn2294/papers/correctionEcta2.pdf.

Bai, J. and S. Ng (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics 146*(2), 304–317.

Bai, J. and S. Ng (2021). Approximate factor models with weaker loading. Technical report, Columbia University.

Bai, Z. and J. W. Silverstein (2006). *Spectral Analysis of Large Dimensional Random Matrices*. Springer.

Bai, Z. and J. W. Silverstein (2009). *Spectral Analysis of Large Dimensional Random Matrices*. Springer.

Bailey, N., G. Kapetanios, and M. H. Pesaran (2020). Measurement of factor strenght: Theory and practice.

Bair, E., T. Hastie, D. Paul, and R. Tibshirani (2006). Prediction by supervised principal components. *Journal of the American Statistical Association 101*(473), 119–137.

Bair, E. and R. Tibshirani (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biology 2*(4), 511–522.

Cai, T. T., T. Jiang, and X. Li (2021). Asymptotic analysis for extreme eigenvalues of principal minors of random matrices. *The Annals of Applied Probability 31*(6), 2953–2990.

Chamberlain, G. and M. Rothschild (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica 51*, 1281–1304.

Chao, J. C. and N. R. Swanson (2022). Consistent estimation, variable selection, and forecasting in factor-augmented var models. Technical report, University of Maryland and Rutgers University.

d'Aspremont, A., L. E. Ghaoui, M. I. Jordan, and G. R. G. Lanckriet (2007, January). A direct formulation for sparse PCA using semidefinite programming. *SIAM Review 49*(3), 434–448.

Fan, J., Y. Liao, and M. Mincheva (2011). High-dimensional covariance matrix estimation in approximate factor models. *Annals of Statistics 39*(6), 3320–3356.

Forni, M., D. Giannone, M. Lippi, and L. Reichlin (2009). Opening the black box: Structural factor models with large cross sections. *Econometric Theory 25*, 1319–1347.

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

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2004, April). The generalized dynamic factor model: Consistency and rates. *Journal of Econometrics 119*(2), 231–255.

Forni, M. and M. Lippi (2001). The generalized dynamic factor model: Representation theory. *Econometric Theory 17*, 1113–1141.

Freyaldenhoven, S. (2022). Factor models with local factors - determining the number of relevant factors. *Journal of Econometrics 229*(1), 80–102.

Giglio, S., D. Xiu, and D. Zhang (2020). Test assets and weak factors. Technical report, Yale University and University of Chicago.

Giglio, S., D. Xiu, and D. Zhang (2022). Prediction when factors are weak: Supplementary evidence. Technical report, University of Chicago.

Hoyle, D. C. and M. Rattray (2004). Principal-component-analysis eigenvalue spectra from data with symmetry-breaking structure. *Physical Review E 69*(2), 026124.

Huang, D., F. Jiang, K. Li, G. Tong, and G. Zhou (2021). Scaled pca: A new approach to dimension reduction. *Management Science, forthcoming*.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics 29*, 295–327.

Johnstone, I. M. and A. Y. Lu (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association 104*(486), 682–693.

Jolliffe, I. T., N. T. Trendafilov, and M. Uddin (2003, September). A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics 12*(3), 531–547.

Kelly, B. and S. Pruitt (2013). Market expectations in the cross-section of present values. *The Journal of Finance 68*(5), 1721–1756.

Moon, H. R. and M. Weidner (2015). Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica 83*(4), 1543–1579.

Onatski, A. (2009). Testing hypotheses about the number of factors in large factor models. *Econometrica 77*(5), 1447–1479.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *Review of Economics and Statistics 92*, 1004–1016.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics 168*, 244–258.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistical Sinica 17*, 1617–1642.

Stock, J. H. and M. W. Watson (2002). Forecasting Using Principal Components from a Large Number of Predictors. *Journal of the American Statistical Association 97*(460), 1167–1179.

Uematsu, Y., Y. Fan, K. Chen, J. Lv, and W. Lin (2019). Sofar: Large-scale association network learning. *IEEE Transactions on Information Theory 65*(8), 4924–4939.

Uematsu, Y. and T. Yamagata (2021). Estimation of sparsity-induced weak factor models. *Journal of Business and Economic Statistics, forthcoming.*

Wang, W. and J. Fan (2017). Asymptotics of empirical eigenstructure for ultra-high dimensional spiked covariance model. *Annals of Statistics 45*, 1342–1374.

Zou, H., T. Hastie, and R. Tibshirani (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics 15*, 265–286.