Prediction When Factors are Weak

Stefano Giglio*
Yale University
Dacheng Xiu
University of Chicago
Dake Zhang
Shanghai Jiao Tong University

Abstract

In economic forecasting, principal component analysis (PCA) has been the most prevalent approach to the recovery of factors, which summarize information in a large set of 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.

Keywords: 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.

1 Introduction

Building on the seminal contribution of Stock and Watson (2002), factor models have become central to economic forecasting. In this context, PCA has become the predominant method for recovering latent factors from a large set of predictors and reducing dimensionality.

The theoretical justification for PCA typically relies on factor pervasiveness; see Bai and Ng (2002) and Bai (2003). Under this assumption, PCA effectively extracts the common components of the predictors while separating idiosyncratic noise. Recent work, including Bai and Ng (2023) and Choi and Yuan (2025), relaxes this requirement and shows that PCA can remain reliable under weaker conditions. Nevertheless, PCA is an unsupervised method and may fail to identify the most predictive low-dimensional features. When the signal-to-noise ratio is low, the space spanned by principal components can become inconsistent or nearly orthogonal to the true factors; see Hoyle and Rattray (2004) and Johnstone and Lu (2009). We refer to such underlying factors as *weak*.

This paper studies a setting in which factors are sufficiently weak that PCA cannot recover them. We propose a new dimension-reduction method based on *supervised PCA* (SPCA). The approach begins by selecting predictors correlated with the target before applying PCA, following the idea introduced in Bair and Tibshirani (2004) and formalized in Bair et al. (2006). Our SPCA procedure extends this framework by incorporating an additional projection step and an iterative algorithm: we select predictors correlated with the target, extract a factor via PCA, project the target and all predictors onto that factor, take residuals, and repeat the process to recover multiple factors—typically *from different predictor subsets*. Final predictions use these estimated factors in time-series regressions.

We justify our procedure in an asymptotic framework where both the sample size and the cross-sectional dimension increase, possibly at different rates. We show that the iterative algorithm yields consistent prediction of the target and asymptotically recovers all weak factors that are relevant for prediction. It does not, however, guarantee recovery of weak factors

orthogonal to the target—though this is immaterial for prediction consistency, as factors orthogonal to the target do not contribute to its forecasting.

The weak-factor problem in our setting arises from a factor loadings matrix whose singular values grow more slowly than the cross-sectional dimension. These factors are weaker than those considered in Bai and Ng (2023), making PCA inconsistent and causing its predictions to be biased. We show that standard supervised methods, such as Partial Least Squares (PLS), suffer from the same issue. Our procedure, however, is not intended to detect the extremely weak factors studied in Onatski (2009), Onatski (2010), and Onatski (2012), where the eigenvalues of the factor and idiosyncratic components are of comparable magnitude.

Beyond consistency, if each latent factor correlates with at least one component of a multivariate target, we obtain stronger results: consistent estimation of the number of weak factors, recovery of the full factor space, and valid prediction intervals. These results do not require perfect identification of the predictors correlated with the factors (in contrast to Bair et al. (2006)) and fully account for the errors accumulated over the iterative procedure.

Our empirical analysis applies SPCA to macroeconomic forecasting. We combine the Fred-Md dataset of 127 macro variables with Blue Chip Financial Forecasts containing hundreds of professional forecasts of the macroeconomic targets, creating a large predictor dataset. We forecast quarterly inflation, industrial production growth, and unemployment changes, comparing SPCA with PCA and PLS. We show that SPCA excels in out-of-sample forecasting with many potentially noisy predictors.

Our work contributes to the literature on forecasting and dimension reduction, sharing a core philosophy with methods that "prioritize" predictors by predictive power, such as the thresholding of Bai and Ng (2008) and the Scaled PCA of Huang et al. (2022), as well as their extensions to dynamic factor settings (Chao and Swanson (2022), Gao and Tsay (2024), and Huang and Tsay (2024)). A key limitation of this literature is the restrictive assumptions that all factors have a uniform order of strength and that a subset of predictors are pure noise.

Our approach relaxes these assumptions and is supported by asymptotic theory in settings where eigenvalues may grow at heterogeneous and slower rates. Although the weak-factor issue arises in both static and dynamic factor models—and our procedure applies in either context—we develop and analyze it within the static approximate factor model.

Our paper relates to spike covariance models (Johnstone (2001)), where the largest few eigenvalues differ from the others in population yet remain bounded. Bai and Silverstein (2009), Johnstone and Lu (2009), and Paul (2007) show that the largest sample eigenvalues and eigenvectors are inconsistent unless sample size grows faster than the cross-sectional dimension. Wang and Fan (2017) extend this framework to diverging eigenvalue spikes and derive limiting distributions under general high-dimensional regimes, allowing sample size to grow much slower than the cross-sectional dimension. Collectively, these papers shed light on the source of bias in standard PCA across different asymptotic settings.

Besides supervised PCA, another approach to resolving PCA's inconsistency is sparse PCA, which imposes eigenvector sparsity (Jolliffe et al. (2003), Zou et al. (2006), d'Aspremont et al. (2007), Johnstone and Lu (2009), Amini and Wainwright (2009)). Uematsu and Yamagata (2022) apply the sparse PCA method of Uematsu et al. (2019) to estimate a sparsity-induced weak factor model, and Bailey et al. (2021) and Freyaldenhoven (2022) adopt related frameworks. Since sparsity is rotation-dependent, such methods—unlike SPCA—require rotation-specific identification assumptions.

Our SPCA procedure is related in spirit to least angle regression Efron et al. (2004) and the orthogonal greedy algorithm Cai and Wang (2011), which also iteratively select covariates or linear combinations based on predictive relevance. However, the weak-factor setting we study differs fundamentally from the data-generating processes (DGPs) underlying these methods, rendering them ineffective in our environment despite the superficial algorithmic resemblance.

The paper is organized as follows. Section 2 introduces the weak factor setup and our supervised PCA methodology. Section 3 presents the asymptotic theory. Section 4 reports

Monte Carlo evidence, and Section 5 provides the empirical analysis. Section 6 concludes. The appendix contains additional empirical results and all proofs.

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 given 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 with rows indexed by 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 use the notation $x_n \lesssim y_n$ when there exists a constant C such that $x_n \leq C y_n$ holds for sufficiently large n. Similarly, we use $x_n \lesssim_{\mathrm{P}} y_n$ to denote $x_n = O_{\mathrm{P}}(y_n)$. If $x_n \lesssim y_n$ and $y_n \lesssim x_n$, we write $x_n \asymp y_n$ for short. Similarly, we use $x_n \asymp_{\mathrm{P}} y_n$ if $x_n \lesssim_{\mathrm{P}} y_n$ and $y_n \lesssim_{\mathrm{P}} x_n$.

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 its i-th largest eigenvalue. Similarly, we use $\sigma_i(A)$ to denote the ith singular value. We use $\|A\|$ and $\|A\|_{\mathrm{F}}$ to denote the operator 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 rank-$d$ matrix A with d rows, 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 predictors x_t with a sample of size T. We assume that x_t follows a linear factor model:

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

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 the target variables in y are related to x through 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 above notation, we can rewrite the above two equations in their matrix form as

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

We now discuss assumptions that characterize the 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, prediction error Z, and observable regressor W satisfy:

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

where $\Sigma_f \in \mathbb{R}^{K\times K}$, $\Sigma_w \in \mathbb{R}^{M\times M}$ are positive-definite 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 F's K left-singular values 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, since F can always be normalized to satisfy this condition. Next, we assume

[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): $x_t = \beta \widetilde{f}_t + \widetilde{\beta}_w w_t + u_t$.

Assumption 2. For some $0 \leq \nu < 1$, the $N \times K$ factor loading matrix β satisfies: (i) $\|\beta\|_{\text{MAX}} \lesssim N^{-\nu/2}$; (ii) $\lambda_K(\beta'_{[I_0]}\beta_{[I_0]}) \asymp N_0/N^\nu$, for some index set I_0, where $N_0 = |I_0| \to \infty$.

Assumption 2 generalizes the traditional pervasive condition in two key directions. First, the parameter ν accommodates weaker loadings, following Bai and Ng (2023). Second, and more importantly, we require only the existence of a subset I_0 of predictors for which the factors are strong enough to be recovered by PCA. This is far weaker than assuming uniform factor strength across all predictors, where $\lambda_1(\beta'\beta) \asymp \ldots \asymp \lambda_K(\beta'\beta) \asymp N^{1-\nu}$. As a result, Assumption 2—which extends Assumption A in Bai and Ng (2023) (the special case $I_0 = \langle N \rangle$)—allows these eigenvalues to grow at heterogeneous rates, since no restriction is imposed on $\beta_{[I_0^c]}$. We will make precise statement about the relative magnitudes of N_0, N^ν, and T, when it comes to our asymptotic results.

Since the number of factors K is finite, even if each factor is strong only within its own (possibly non-overlapping) index set, one can still construct a common index set I_0 on which all factors are sufficiently strong.[2] Assumption 2 nevertheless excludes the extreme case in which all entries of β uniformly and rapidly vanish—namely, when $\sup_{\{I||I|\geq \breve{N}\}} |I|^{-1}\lambda_K\left(\beta'_{[I]}\beta_{[I]}\right) = o(N^{-\nu})$ for any $\breve{N} \to \infty$, so that no subset I_0 with sufficiently strong factors can exist.

Next, we need the following moment conditions on U.

Assumption 3. The idiosyncratic component U satisfies: $\|U\|_{\text{MAX}} \lesssim_{\text{P}} (\log T)^{1/2} + (\log N)^{1/2}$. In addition, for any given non-random subset $I \subset \langle N \rangle$ with $|I| \to \infty$, $\left\|U_{[I]}\right\| \lesssim_{\text{P}} |I|^{1/2} + T^{1/2}$.

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),

[2] To illustrate with a concrete example in the case $\nu = 0$, suppose β has a block diagonal structure, with the kth column β_k supported on an index set J_k and with $\bigcap_k J_k = \varnothing$. Suppose the non-zero entries of β are standard normal. Let $k^\star := \arg\min_k |J_k|$. Form I_0 by starting with $J_{k^\star}$ (so that $|I_0| \geq |J_{k^\star}|$) and then adding $|J_{k^\star}|$ arbitrary elements from each J_k for $k \neq k^\star$. Taking the union across all such subsets yields an index set I_0 of size $K \times |J_{k^\star}|$, within which all factors are pervasive.

provided that u_t is i.i.d. both in time and in the cross-section. While it may be tempting to impose a stronger condition that uniformly bounds $\sup_{I\subset\langle N\rangle}\|U_{[I]}\|$ over all index sets of a given size $|I|$, the desirable rate of $|I|^{1/2}+T^{1/2}$ may not hold in general. In fact, when $|I|$ is small, Cai et al. (2021) establish a uniform bound that deviates from our non-uniform rate only by a logarithmic factor. However, for large $|I|$, no such uniform result exists to the best of our knowledge. We therefore refrain from imposing any assumptions on uniform bounds over all index sets. Instead, we impose assumptions only on arbitrary non-random index sets. Given the tuning parameters of our algorithm (introduced below), each index set it selects admits a population counterpart that is a deterministic function of the DGP parameters and the tuning parameters. This allows us to focus exclusively on these non-random index sets in population; a formal characterization of these sets is provided in Section 3.1.

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

Assumption 4. For any non-random subset $I\subset\langle N\rangle$ with $|I|\to\infty$ and $\beta_{[I]}\neq 0$, the factor loading $\beta_{[I]}$, and the idiosyncratic error $U_{[I]}$ satisfy the following conditions:

$$
\begin{aligned}
&(i)\ \left\|\underline{U}_{[I]}A'\right\|\lesssim_{\mathrm{P}}|I|^{1/2}T^{1/2},\ \left\|\underline{U}_{[I]}A'\right\|_{\mathrm{MAX}}\lesssim_{\mathrm{P}}(\log N)^{1/2}T^{1/2},\\
&(ii)\ \left\|\breve{\beta}_{[I]}'U_{[I]}\right\|\lesssim_{\mathrm{P}}T^{1/2},\ \left\|\breve{\beta}_{[I]}'U_{[I]}\right\|_{\mathrm{MAX}}\lesssim_{\mathrm{P}}(\log T)^{1/2},\ \left\|\breve{\beta}_{[I]}'\underline{U}_{[I]}A'\right\|\lesssim_{\mathrm{P}}T^{1/2},\\
&(iii)\ \left\|\breve{e}'u_T\right\|\lesssim_{\mathrm{P}}1,
\end{aligned}
$$

where $\breve{\beta}_{[I]}:=\beta_{[I]}/\left\|\beta_{[I]}\right\|$ is the normalized loading, A is either $\underline{F}$, $\underline{W}$ or $\overline{Z}$, and $\breve{e}$ is any $N\times 1$ unit vector measurable with respect to the σ-algebra generated by $\{f_t,u_t,w_t,z_{t+h}\}_{t\leq T-h}$.

The ℓ_2-norm bounds in Assumption 4(i) and (ii) follow from Assumptions D, F2, and F3 in Bai (2003) when $I=\langle N\rangle$. The MAX-norm bounds can be established using large deviation results as in Fan et al. (2011). Although we strengthen these bounds to hold for any non-random subset $I\subset\langle N\rangle$, as long as $|I|\to\infty$, the same low-level conditions ensure that Assumption 4 is satisfied for arbitrary non-random subsets I. Finally, Assumption 4(iii)

imposes weak contemporaneous dependence between u_T and z_T, as well as weak lagged dependence between u_T and the remaining variables. This condition is mainly needed to establish the theoretical validity for forecasting, which again follow from large deviation arguments.

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 $\Sigma := \text{Cov}(x_t) = \beta\Sigma_f\beta' + \Sigma_u$. 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), $\lambda_1(UU') \lesssim_{\text{P}} T+N$, if u_t is i.i.d. with $\|\Sigma_u\| \lesssim 1$. Since $T\lambda_K(\beta'\beta) \asymp_{\text{P}} \lambda_K(\beta FF'\beta')$ 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 (2023) 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 such cases, PCA generally fails to recover the underlying factors (except in the special case of homoscedastic errors). Instead, we impose later a rate condition on the subset I_0 specified in Assumption 2, such that $|I_0|/(T\lambda_K(\beta'_{[I_0]}\beta_{[I_0]})) \to 0$, which guarantees that the factors remain identifiable on this subset.[3] We now turn to a description of the methodology.

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. The idea is to first identify a subset $\widehat{I}$ of predictors via marginal screening, retaining only those with nontrivial exposure to the prediction target, and then apply PCA. This reduces the dimension from N to $|\widehat{I}|$, and under suitable assumptions, guarantees that the selected predictors exhibit a sufficiently strong factor structure. Consequently, PCA on this subset yields consistent factor recovery.

We use a simple one factor example to illustrate the procedure, before explaining its caveats with the general multi-factor case. Specifically, we consider the setting with $D = K = 1$, $\alpha_w = 0$, 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 a tuning parameter controlling the number of predictors retained. The fact that $\widehat{I}$ incorporates information from the target highlights the supervised nature of our procedure. Given the existence of I_0 under Assumption 2, one can choose c so that the predictors in $\widehat{I}$ exhibit a sufficiently strong factor structure. The remaining steps follow the standard principal component regression approach for prediction. Specifically, we apply PCA to $\underline{X}_{[\widehat{I}]}$ to extract factors $\{\widehat{f}_t\}_{t=1}^{T-h}$, which can be written as $\widehat{f}_t = \widehat{\zeta}'x_t$ for some loading matrix $\widehat{\zeta}$ supported on $\widehat{I}$. We then obtain $\widehat{\alpha}$ by regressing $\{y_t\}_{t=1+h}^{T}$ on $\{\widehat{f}_t\}_{t=1}^{T-h}$ using the predictive model (2). The resulting predictor for y_{T+h} is therefore given by: $\widehat{y}_{T+h} = \widehat{\alpha}\widehat{f}_T = \widehat{\alpha}\widehat{\zeta}'x_T$.

[3] The aforementioned settings all require $\lambda_K(\beta'\beta) \to \infty$, in contrast with the extremely weak factor model that imposes $\lambda_K(\beta'\beta) \lesssim 1$. 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.

Bair et al. (2006)'s proposal follows the same steps in the multi-factor case, except that multiple factors are extracted in the PCA stage. However, to ensure the validity of marginal screening in a 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, and that all predictors outside I_0 have zero correlation with the target—that is, they are pure noise for prediction. Given this condition, they show marginal screening can consistently recover I_0, allowing all factors to be extracted in one pass of PCA applied to this subset.

In contrast, we assume the existence of a set I_0 within which predictors exhibit a strong factor structure, but impose no restrictions on the correlation between the target and predictors outside this set, nor on the strength of their factor structure. Consequently, I_0 under Assumption 2 needs not be unique, and the validity of our prediction procedure does not hinge on consistently recovering this particular I_0. More importantly, requiring marginal screening to recover a subset of predictors that admits a factor structure with all factors sufficiently strong is a stringent condition (even when such a subset is uniquely defined, as in Bair et al. (2006)). Screening can be distorted by correlations induced by strong factors, leaving weak factors unidentifiable, while predictors discarded by screening may nevertheless be instrumental—or even essential—for prediction. We illustrate these issues below using two-factor examples.

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 while the second is weak. The target variable y is correlated with both factors and therefore potentially with all predictors. Consequently, the

screening step described above may either fail to eliminate any predictors or select them in an arbitrary way, since all predictors inherit correlation with the target through the strong factor. Because the second factor is weak, a single pass of PCA—extracting two factors from the full set or from an arbitrary subset of predictors—does not guarantee recovery of this component. Indeed, we can show that $\lambda_{\min}(\beta'\beta) \leq \|\beta_{12}\|^2 \lesssim N_0$, implying that PCA cannot consistently recover the second factor unless $N/(N_0T) \to 0$. The assumptions in Bair et al. (2006) rule out this case, but we can clearly locate an index set I_0 (e.g., the top N_0 predictors) within which both factors are strong. In other words, our assumptions 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 $N/2 \times 1$ vectors, $\|\beta_{11}\| \asymp \|\beta_{22}\| \asymp \sqrt{N}$, 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 marginal screening step is applied only once as in Bair et al. (2006). First, y_{t+h} is uncorrelated with the second half of predictors. Second, the exposure of the first half of predictors to the first and second factors are the same.

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 inconsistent prediction.

Both examples demonstrate the failure of a one-step supervised PCA procedure in a general multi-factor setting. Similar to Bair et al. (2006), Huang et al. (2022) propose the Scaled PCA method, which also leverages the univariate predictive power of each predictor in a preprocessing step prior to applying PCA, aiming to address the challenge posed by weak factors. Rather than eliminating predictors with low univariate predictive power, Scaled PCA reweights each predictor according to its univariate strength before applying PCA. Ideally, this reweighting increases the influence of informative predictors, thereby amplifying weak factor signals. However, the method can also fail in a general multi-factor setting for the same reasons that limit the approach of Bair et al. (2006): in Example 1, the univariate covariances between all predictors and the target may be nearly random, preventing Scaled PCA from overweighting the top few informative predictors; in Example 2, the covariances between the second half of the predictors and the target are zero, leading Scaled PCA to discard them entirely. The DGPs in these examples are ruled out by the assumptions in both Bair et al. (2006) and Huang et al. (2022), but are explicitly allowed under our framework. 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

Our method employs a multi-step procedure that iteratively alternates between selection and projection. In each iteration, the projection step removes the influence of the previously estimated factor, thereby enhancing the effectiveness of the subsequent screening step. 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.

This iterative screening and projection approach resolves the issues in the preceding examples. In Example 1, the first screening retains a random subset of predictors, and the first PC recovers the strong factor f_1. After projecting X and y onto f_1, the residuals for the first N_0 predictors still load on f_2, while the remaining $N - N_0$ predictors are uncorrelated with the residuals of y. A second screening then eliminates these predictors, allowing PCA to recover f_2 from the residuals of the first N_0 predictors. In Example 2, the first screening eliminates the second half of the predictors, enabling PCA to recover $f_1 + f_2$ from the rest. After projection, the residuals of the first half contain only noise, while the second half are spanned by $f_1 - f_2$, which a second PCA step recovers. Thus, iterated supervised PCA recovers the full factor space, showing that marginal screening works when combined with iteration and projection.

Formally, we present our algorithm below for the general model given by (1) and (2). A key step in implementing the algorithm is the choice of $\widehat{I}_k$ and a suitable stopping rule. We suggest using the top qN predictors ranked by the magnitude of their covariances with $Y_{(k)}$:[4]

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

Selecting a fixed number, qN, of predictors at each step substantially simplifies both the notation and the proof. In addition, this choice yields factor estimates that are typically more stable and less sensitive to the tuning parameter q than those based on a hard threshold, as

[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. Alternatively, one can use correlation instead of covariance when constructing $\widehat{I}_k$. This modification does not alter the asymptotic analysis, provided that the variance of each predictor is uniformly bounded above and below. In practice, we find that correlation-based screening performs better in finite samples when predictors differ in scale. We therefore adopt correlation in both simulations and empirical analysis.

Algorithm 1 Prediction via SPCA

1: **Inputs**: $\overline{Y}$, $\underline{X}$, $\underline{W}$, x_T, and w_T.
2: **Initialization**: $Y_{(1)} := \overline{Y}\mathbb{M}_{\underline{W}'}$, $X_{(1)} := \underline{X}\mathbb{M}_{\underline{W}'}$, $k = 1$.
3: **repeat**
4: Select an appropriate subset $\widehat{I}_k \subset \langle N \rangle$ via marginal screening.
5: Estimate the kth factor $\underline{\widehat{F}}_{(k)} = \widehat{\varsigma}'_{(k)} \left(X_{(k)}\right)_{[\widehat{I}_k]}$ via SVD, where $\widehat{\varsigma}_{(k)}$ is the first left singular vector of $\left(X_{(k)}\right)_{[\widehat{I}_k]}$. $\underline{\widehat{F}}_{(k)}$ can also be rewritten as $\underline{\widehat{F}}_{(k)} = \widehat{\zeta}'_{(k)}\underline{X}\mathbb{M}_{\underline{W}'}$, 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 via $\widehat{\beta}_{(k-1)}$ (defined in Line 6.).
6: Estimate the coefficients $\widehat{\alpha}_{(k)} = Y_{(k)}\underline{\widehat{F}}'_{(k)}(\underline{\widehat{F}}_{(k)}\underline{\widehat{F}}'_{(k)})^{-1}$ and $\widehat{\beta}_{(k)} = X_{(k)}\underline{\widehat{F}}'_{(k)}(\underline{\widehat{F}}_{(k)}\underline{\widehat{F}}'_{(k)})^{-1}$.
7: Obtain residuals $Y_{(k+1)} = Y_{(k)} - \widehat{\alpha}_{(k)}\underline{\widehat{F}}_{(k)}$ and $X_{(k+1)} = X_{(k)} - \widehat{\beta}_{(k)}\underline{\widehat{F}}_{(k)}$.
8: $k = k + 1$.
9: **until** $k = \widehat{K}$, where $\widehat{K}$ is chosen based on some proper stopping rule.
10: 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 $\widehat{\alpha}_w = \overline{Y}\underline{W}'(\underline{WW}')^{-1}$.
11: **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 $\widehat{\gamma}$.

in conventional marginal screening. Correspondingly, the algorithm terminates as soon as

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

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.

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 selection rules in (6) and (7) ensure that, prior to applying PCA, the selected subset contains at least qN predictors with non-negligible predictive strength. As a result, the leading eigenvalue of the corresponding covariance matrix is of the asymptotic order $qN^{1-\nu}$, ensuring the presence of at least one strong factor and enabling its consistent recovery through PCA.

Line 10 of the algorithm needs more explanations. Line 5 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, $\widehat{\alpha}_w = \overline{Y}\mathbb{M}_{\underline{\widehat{F}}'}\underline{W}'\left(\underline{W}\mathbb{M}_{\underline{\widehat{F}}'}\underline{W}'\right)^{-1} = \overline{Y}\underline{W}'(\underline{W}\underline{W}')^{-1}$, since $\mathbb{M}_{\underline{\widehat{F}}'}\underline{W}' = \underline{W}'$ by construction, which explains the formula for $\widehat{\alpha}_w$ in Line 10. Finally, with $\widehat{\alpha}$, $\widehat{\alpha}_w$, and $\widehat{f}_T$, it is sufficient to construct $\widehat{y}_{T+h}$ by combining $\widehat{\alpha}\widehat{f}_T$ with $\widehat{\alpha}_w w_T$, which in turn can be written as a projection on x_T and w_T.

3 Asymptotic Theory

We now turn to the asymptotic properties of SPCA. The analysis is involved owing to its iterative nature and the general weak factor setting under consideration.

3.1 Consistency in Prediction

To establish the consistency of SPCA for prediction, we first investigate the consistency of factor estimation. In the pervasive 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 non-random counterparts iteratively, for any given choice of c and q. This step is critical to characterize the exact factor space recovered by SPCA. 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 $\beta\Sigma_f^{1/2}$ and $\alpha\Sigma_f^{1/2}$ 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 $a_i^{(k)} := \left\|\beta_{[i]}\prod_{j<k}\mathbb{M}_{b_{(j)}}\alpha'\right\|_{\text{MAX}}$, $I_k := \{i|a_i^{(k)} \geq c_{qN}^{(k)}\}$, 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 $b_{(j)}$ are well defined and identifiable, we require the top two singular values of $\beta_{(k)}$ to be distinct at each stage k. We also need the constants $c_{qN}^{(k)}$ to be separated so that the sets I_k are uniquely determined. We say that two sequences a_N and b_N are *asymptotically distinct* if there exists $\delta > 0$ such that $|a_N - b_N| \geq \delta |b_N|$ for sufficiently large N. Motivated by these considerations, we impose the following condition:

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 mild and rules out only corner cases. Separation between $\sigma_1(\beta_{(k)})$ and $\sigma_2(\beta_{(k)})$ ensures identification of the leading eigenvector, which is essential for consistent factor recovery. Likewise, asymptotic separation of the $c_{qN}^{(k)}$ values ensures sufficient spacing between population covariance levels, helping to control sampling noise in the selection step. Excluding such cases is standard in the high-dimensional PCA literature; see, for example, 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 would be to select predictors according to a predetermined covariance threshold and terminate the selection process once $|I_k|$ falls below another threshold. By allowing the number of predictors to vary across iterations, this alternative procedure can be particularly useful for handling corner cases that are ruled out under the current form of Assumption 5.[5] Similar asymptotic results—analogous to those in Theorems 3.1–3.3—can be established, though at the expense of more intricate conditions on rates of convergence and related aspects. For

[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.

clarity and parsimony, however, we focus in the main text on the current version of SPCA, which yields cleaner theorems and exhibits superior performance in simulations.

We now are ready to present the consistency of the estimated factors by SPCA:

Theorem 3.1. *Suppose that x_t follows (1) and y_t satisfies (2), and that Assumptions 1-5 hold. Then for any tuning parameters c and q that satisfy*[6]

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

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 $1 \leq k \leq \tilde{K}$,

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

We make a few observations regarding this result. First, the assumptions in Theorem 3.1 do not guarantee a consistent estimate of the number of factors, K, because SPCA 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 and its dimension are 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 $\tilde{K} \leq K$.[7] 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.

[6]The existence of parameters c and q satisfying (8) is equivalent to the condition $\log(NT)N^{2\nu}(1/N_0 + 1/T) \to 0$. This, in turn, implies the usual requirement $N^{\nu}/N_0 + N^{\nu}/T \to 0$ needed for PCA consistency on the subset I_0; the latter coincides with Assumption A in Bai and Ng (2023) when $I_0 = \langle N \rangle$. The additional factor $\log(NT)N^{\nu}$ is due to the error in the extra selection and projection steps in our algorithm.

[7]The proof of Theorem 3.1 also demonstrates that the threshold $\hat{c}_{qN}^{(k)}$ used in the selection step consistently estimates its population counterpart $c_{qN}^{(k)}$ for any $1 \leq k \leq \tilde{K} + 1$.

$\|\widehat{\underline{F}}_{(k)}\|$ is a normalization constant that can be omitted if the estimated factors are normalized to have unit ℓ_2-norm in implementation.

Second, with respect to the tuning parameters, the condition (8) implies that $cN^{\nu/2} \to 0$, $c\sqrt{T/N^{\nu}} \to \infty$, and $c\sqrt{qN/N^{\nu}} \to \infty$. 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, 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 in Lemma S4.3 in the Appendix, the estimation error is of order $(qN^{1-\nu})^{-1/2} + T^{-1/2}N^{\nu/2}$, and the choice of c needs dominate it. 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 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, when $\nu = 0$, the estimation error of the factors is bounded by $q^{-1/2}N^{-1/2} + T^{-1}$. In the pervasive factor case, the factor space can be recovered at the rate of $N^{-1/2} + T^{-1}$; see, e.g., Bai (2003). In our setting, qN plays the same role as N in the pervasive case. Importantly, Assumption 2 does not require all factors to have equal strength. Some factors could, in principle, be recovered at faster rates if different number of predictors were selected for different factors. Indeed, the alternative choice of $\widehat{I}_k$ based on (3) permits varying subset sizes across stages because the threshold is fixed. While this may achieve a faster rate for relatively stronger factors, the overall prediction error is ultimately dictated by the weakest factor. Moreover, we find that the selection rule in (6) delivers more stable out-of-sample performance, whereas (3) can be sensitive to tuning choices. Since our primary objective is

prediction rather than factor recovery, we focus on the more stable approach.

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

Theorem 3.2. *Under the same assumptions as in Theorem 3.1, we have* $\|\widehat{\alpha}_w - \alpha_w\| \xrightarrow{\text{P}} 0$, $\|\widehat{\gamma}\beta - \alpha\| \xrightarrow{\text{P}} 0$, *and consequently,* $\|\widehat{y}_{T+h} - \text{E}_T(y_{T+h})\| \xrightarrow{\text{P}} 0$.

Theorem 3.2 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 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

$$\widehat{\gamma}\beta = \widehat{\alpha}H, \quad \text{so that by Theorem 3.2} \quad \|\widehat{\alpha}H - \alpha\| = o_{\text{P}}(1). \tag{10}$$

On the other hand, the proof of Theorem 3.1 also establishes that for $k \le \tilde{K}$:

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

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

The consistency result in Theorem 3.2 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 require normally distributed errors or the assumption that all factors have the same strength. The assumption on the relative sizes of N and T is also flexible, in contrast with the existing literature in which N cannot grow faster than a certain polynomial rate of T, e.g., Bai and Ng (2023), Huang et al. (2022).

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 3.2 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 nearly random noise. Consequently, screening will terminate right away as 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, this issue arises from the potential failure to recover all factors in the DGP. As long as all factors are found, the bias is negligible and the CLT holds regardless of the magnitude of α. To move beyond consistency and obtain valid inference, a stronger assumption is required to rule out such cases, in which higher order omitted factor bias undermines the CLT, even if consistency itself is unaffected. Specifically, if $\alpha \in \mathbb{R}^{D\times K}$ satisfies $\lambda_{\min}(\alpha'\alpha) \gtrsim 1$, then the omission of factors can be ruled out asymptotically. This condition implies, on the one hand, that the dimension of the target variables, D, must be no smaller than the dimension of

the factors, K, and that each factor is correlated with at least one target variable in y; together, these requirements ensure that no factor is omitted. On the other hand, our algorithm will not asymptotically select more factors than necessary, since the iteration terminates once all covariances vanish. With a consistent estimator of the number of factors, we can therefore recover the factor space and conduct inference on the prediction targets.

The inference theory on pervasive factor models also relies on a consistent estimator of the count of factors, e.g., Bai and Ng (2002). 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 the appendix, 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 3.3. *Under the same assumptions as Theorem 3.2, if we further have* $\lambda_{\min}(\alpha'\alpha) \gtrsim 1$, *then for any tuning parameters* c *and* q *in* (6) *and* (7) *satisfying* (8), *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* $\left\| \mathbb{P}_{\widehat{\underline{F}}'} - \mathbb{P}_{\underline{F}'} \right\| = O_{\mathrm{P}}\left(q^{-1/2}N^{-1/2+\nu/2} + T^{-1}N^{\nu}\right)$. *(iii) The estimator* $\widehat{\gamma}$ *constructed via Algorithm 1 satisfies* $\left\| \widehat{\gamma}\beta - \alpha - T^{-1}\overline{Z}\underline{F}'\Sigma_f^{-1} \right\| = O_{\mathrm{P}}(q^{-1}N^{-1+\nu} + T^{-1}N^{\nu})$.

Theorem 3.3 (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 3.1(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 when $\nu = 0$ 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 3.2, 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 $O_{\mathrm{P}}(q^{-1}N^{-1+\nu} + T^{-1}N^{\nu})$.

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 $(qN^{1-\nu})^{-1/2}$. In light of Theorem 3.3(iii), if $q^{-1}N^{-1+\nu}T \to 0$ and $T^{-1/2}N^{\nu} \to 0$, then the second term is asymptotically negligible (i.e., $o_{\mathrm{P}}(T^{-1/2})$) compared to the first term, and $\left\|(\widehat{\gamma}\beta - \alpha)f_T - T^{-1}\overline{Z}\underline{F}'\Sigma_f^{-1}f_T\right\| = O_{\mathrm{P}}(N^{\nu}/T) = o_{\mathrm{P}}(T^{-1/2})$, 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 magnitude of $qN^{1-\nu}$. 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 $(qN^{1-\nu})^{-1/2}\Psi u_T$ are jointly asymptotically normally distributed, satisfying:

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

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, their distributions are a direct result of a large-T CLT for mixing processes. For 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 $(qN^{1-\nu})^{1/2}$ for Ψu_T arises naturally because the factor loading satisfies $\left\|\beta_{[I_k]}\right\| \asymp (qN^{1-\nu})^{1/2}$.

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 3.4. *Suppose the same assumptions as in Theorem 3.3 hold. If in addition, Assumption 6 holds and $T^{-1/2}N^{\nu} \to 0$, we have $\Phi^{-1/2}(\widehat{y}_{T+h} - \mathrm{E}_T(y_{T+h})) \xrightarrow{d} \mathcal{N}(0, \mathbb{I}_D)$, where*

$$\Phi = T^{-1}\Phi_1 + q^{-1}N^{-1+\nu}\Phi_2,$$

and Φ_1 and Φ_2 are given by

$$\Phi_1 = \left((f'_T, w'_T)\Sigma_{f,w}^{-1} \otimes \mathbb{I}_D\right) \begin{pmatrix} \Pi_{11} & \Pi_{12} \\ \Pi'_{12} & \Pi_{22} \end{pmatrix} \left(\Sigma_{f,w}^{-1}(f'_T, w'_T)' \otimes \mathbb{I}_D\right),$$

$$\Phi_2 = \alpha B(\Lambda/qN^{1-\nu})^{-1}\Omega'\Pi_{33}\Omega(\Lambda/qN^{1-\nu})^{-1}B'\alpha'.$$

Here, 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 $qN^{1-\nu}$. For inference, we need construct estimators for Φ, which we discuss in Section S1 of the appendix.

3.4 Tuning Parameter Selection

Along with the gain in robustness to weak factors comes the cost of an extra tuning parameter. Implementing SPCA requires two tuning parameters, q and c. The parameter q dictates the size of the subset used for PCA, whereas c determines the stopping rule and thereby the number of factors, K. By comparison, PCA and PLS only require selecting K. As established in Theorem 3.3, K can be consistently recovered, 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. For q, larger values make SPCA's performance resemble PCA, reducing robustness to weak factors, while smaller values increase the risk of overfitting, as the selected predictors are more likely to overfit y. We suggest tuning $\lfloor qN \rfloor$ instead of q, as the former takes integer values and avoids redundancies where multiple q values map to the same $\lfloor qN \rfloor$.

In our applications, tuning parameters are selected using 3-fold cross-validation (CV) as follows. The entire sample is split into three consecutive folds to account for time series dependence, avoiding random splits. Each fold is used for validation in turn, while the other two serve as training sets. The optimal tuning parameters are determined based on the average R^2 across the validation folds. The model is then refitted using the full dataset with these parameters before making predictions. A detailed investigation of the impact of tuning on the finite sample performance of all procedures is conducted below.

4 Simulations

In this section, we evaluate the finite-sample performance of SPCA through Monte Carlo simulations. The predictors follow the two-factor model as specified in equation (1), with one strong factor, f_{1t}, and one potentially weak factor, f_{2t}. Their relative strength is controlled by a parameter a. We construct two disjoint index sets, I_1 and I_2, each of size $\lfloor aN \rfloor$, with $a \in \{50\%, 10\%, 2.5\%\}$, and assign loadings

$$\beta_{i,1} \sim \begin{cases} \text{Unif}[3,4], & \text{if } i \in I_1, \\ \text{Unif}[0,1], & \text{if } i \notin I_1, \end{cases} \qquad \beta_{i,2} \sim \begin{cases} \text{Unif}[1,2], & \text{if } i \in I_2, \\ 0, & \text{if } i \notin I_2. \end{cases}$$

Smaller a weakens the second factor by shrinking the share of predictors with nonzero loadings.

We first consider a DGP for y of the form (2), taking $\alpha_w = (0, 0.2)$ and $\alpha = (3, 1)$, yielding $y_{t+1} = 3f_{1t} + f_{2t} + 0.2y_t + z_{t+1}$. Both f_{it} and z_t are drawn independently from standard normal distributions. This specification satisfies the assumptions used in our theoretical analysis. Under this DGP, SPCA is expected to select I_1 in the first step and recover f_{1t}, and then select I_2 in the second iteration and recover f_{2t}. The idiosyncratic errors u_{it} in (1) are generated by first drawing $\epsilon_{it} \sim \mathcal{N}(0, 9)$, and then forming the noise matrix $U = \epsilon A$, where $A = S\Gamma$. Here,

S is diagonal with entries sampled from Unif$(0.5, 1.5)$, and Γ is a random rotation matrix drawn uniformly from the unit sphere. This construction introduces heteroskedasticity in u_{it}.

In this experiment, we evaluate prediction accuracy for y_{T+1} and evaluate the estimation errors for K and the sets I_k under varying values of T, while fixing $aN = 100$ so that the number of predictors loading on the weaker factor remains constant. Prediction performance is measured by the mean squared error (MSE) and the parameter estimation error $\|\widehat{\gamma}\beta - \alpha\|$. We also report the mean and standard deviation of the estimated number of factors $\widehat{K}$. For subset recovery, we set $q = a$ for convenience, so that $|\widehat{I}_k| = |I_k| = aN$, and measure accuracy using $|\widehat{I}_k \cap I_k|/|I_k|$. According to Theorem 3.2, the remaining tuning parameter is chosen as $c = c_0 \cdot \log(NT)^{1/2}(T^{-0.4} + (aN)^{-0.4})$, with c_0 selected via CV.

Table 1: Consistency in Prediction, Subset Selection, and Factor Number Estimation of SPCA. We evaluate the performance of SPCA in terms of prediction MSE, parameter estimation error $\|\widehat{\gamma}\beta - \alpha\|$, subset recovery accuracy for I_1 and I_2, and the mean and standard deviation of the estimated number of factors $\widehat{K}$. All results are based on 1,000 Monte Carlo repetitions.

	MSE	$\|\widehat{\gamma}\beta - \alpha\|$	I_1 Accuracy	I_2 Accuracy	mean$(\widehat{K})$	sd$(\widehat{K})$
Panel A: $N = 200$, $a = 0.5$						
$T = 120$	0.13	0.14	99.66%	92.27%	2.15	0.36
$T = 240$	0.13	0.10	99.99%	97.82%	2.09	0.28
Panel B: $N = 1000$, $a = 0.1$						
$T = 120$	0.18	0.16	99.58%	79.97%	2.06	0.30
$T = 240$	0.12	0.10	99.99%	93.36%	2.02	0.13
Panel C: $N = 4000$, $a = 0.025$						
$T = 120$	0.29	0.25	99.42%	73.84%	1.99	0.51
$T = 240$	0.14	0.13	99.99%	91.96%	1.99	0.23

Table 1 shows that SPCA consistently recovers the strong factor set I_1 and yields accurate estimates of the number of factors K. Recovery of the weaker factor set I_2 and overall prediction accuracy improve with larger T, even when the second factor is weak ($a = 0.025$). Both MSE and parameter estimation error decline steadily as T increases, confirming the consistency results in Theorem 3.2.

Next, we compare the finite-sample performance of SPCA, SPCA-NI (the non-iterative version of Bair et al. (2006) that performs selection only once), PCA, PLS, and Scaled PCA across different scenarios. In this experiment, q and K are tuned jointly via CV for SPCA and

SPCA-NI, while only K needs to be tuned for the rest. To ensure a fair comparison, we tune K for the former rather than using c_0 as in the previous experiment. We also assess robustness to factor number misspecification by reporting results under fixed values of $K = 1, 2, 3$.

Table 2: Finite-Sample Comparison of Predictors (Univariate y). We evaluate the performance of SPCA, SPCA-NI, Scaled PCA, PCA, and PLS in terms of prediction MSE and $\|\widehat{\gamma}\beta - \alpha\|$. For SPCA and SPCA-NI, we also report the average value of the selected tuning parameter q. All results are based on 1,000 Monte Carlo repetitions.

		MSE				$\|\widehat{\gamma}\beta - \alpha\|$			
	K	CV	1	2	3	CV	1	2	3
Panel A: $N = 4000$, $a = 0.025$									
T =120	SPCA	0.18	0.94	0.17	0.34	0.16	0.94	0.16	0.16
	q	*0.02*	*0.09*	*0.02*	*0.03*	*0.02*	*0.09*	*0.02*	*0.03*
	SPCA-NI	0.58	0.94	0.57	0.57	0.61	0.94	0.61	0.61
	q	*0.12*	*0.09*	*0.12*	*0.13*	*0.12*	*0.09*	*0.12*	*0.13*
	Scaled PCA	0.63	0.97	0.63	0.62	0.70	0.96	0.70	0.69
	PCA	0.86	1.02	0.88	0.86	0.88	0.98	0.90	0.88
	PLS	0.67	0.83	0.62	0.77	0.62	0.88	0.58	0.54
T =240	SPCA	0.12	0.85	0.12	0.29	0.11	0.89	0.11	0.13
	q	*0.03*	*0.25*	*0.03*	*0.06*	*0.03*	*0.25*	*0.03*	*0.06*
	SPCA-NI	0.23	0.78	0.23	0.23	0.34	0.89	0.34	0.34
	q	*0.18*	*0.25*	*0.18*	*0.19*	*0.18*	*0.25*	*0.18*	*0.19*
	Scaled PCA	0.29	0.95	0.29	0.29	0.41	0.94	0.41	0.41
	PCA	0.41	0.90	0.42	0.41	0.56	0.92	0.56	0.56
	PLS	0.30	0.74	0.30	0.50	0.34	0.83	0.34	0.35
Panel B: $N = 200$, $a = 0.5$									
T =120	SPCA	0.13	0.70	0.13	0.14	0.15	0.79	0.15	0.15
	q	*0.67*	*0.96*	*0.65*	*0.80*	*0.67*	*0.96*	*0.65*	*0.80*
	SPCA-NI	0.13	0.70	0.13	0.13	0.16	0.79	0.16	0.16
	q	*0.97*	*0.96*	*0.97*	*0.97*	*0.97*	*0.96*	*0.97*	*0.97*
	Scaled PCA	0.19	0.92	0.19	0.20	0.25	0.92	0.26	0.25
	PCA	0.12	0.70	0.12	0.13	0.15	0.79	0.15	0.15
	PLS	0.13	0.50	0.12	0.43	0.13	0.65	0.12	0.15
T =240	SPCA	0.13	0.70	0.13	0.13	0.11	0.79	0.11	0.11
	q	*0.66*	*0.99*	*0.67*	*0.80*	*0.66*	*0.99*	*0.67*	*0.80*
	SPCA-NI	0.13	0.70	0.13	0.13	0.11	0.79	0.11	0.11
	q	*0.98*	*0.99*	*0.98*	*0.97*	*0.98*	*0.99*	*0.98*	*0.97*
	Scaled PCA	0.17	0.93	0.16	0.17	0.16	0.93	0.16	0.16
	PCA	0.13	0.70	0.13	0.13	0.11	0.79	0.11	0.11
	PLS	0.13	0.50	0.13	0.38	0.09	0.65	0.09	0.10

Table 2 reports the results. In the weak-factor setting (Panel A), SPCA delivers the best prediction and estimation performance, reflecting its ability to isolate the weak factor. SPCA-NI, PLS, and Scaled PCA all perform noticeably worse, and PCA performs the worst overall. The poor performance of PLS and Scaled PCA stems from their overweighting of predictors

associated with the strong factor f_1, making it difficult to recover the weak factor f_2. Although performance improves for all methods as T increases, SPCA maintains a clear lead. SPCA-NI outperforms PCA, as the former nests the latter as a special case.

In the strong-factor setting (Panel B), all methods except Scaled PCA perform similarly, showing that tuning q does not harm SPCA or SPCA-NI when factors are strong. Joint tuning of K and q (column "CV") introduces some additional variability, but still yields errors smaller than those produced under misspecified values of K. Overshooting K is generally less harmful than undershooting. Finally, in the weak-factor case, CV selects values of q close to the true $a = 0.025$, whereas in the strong-factor case it tends to choose much larger values, often close to one, making SPCA behave more like PCA and improving its efficiency in that setting.

Table 3: Finite Sample Comparison of Predictors (Multivariate y). We evaluate the performance of SPCA, SPCA-NI, 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 $aN = 100$.

	$N = 200, a = 0.5$				$N = 1000, a = 0.1$				$N = 4000, a = 0.025$			
	SPCA	SPCA-NI	PCA	PLS	SPCA	SPCA-NI	PCA	PLS	SPCA	SPCA-NI	PCA	PLS
Panel A: $T = 120$												
$d(\widehat{F}, F)$	0.28	0.29	0.28	0.28	0.30	0.32	0.38	0.40	0.30	0.34	0.84	0.61
MSE_1	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.06
MSE_2	0.10	0.10	0.10	0.09	0.11	0.14	0.24	0.19	0.11	0.16	0.88	0.58
Panel B: $T = 240$												
$d(\widehat{F}, F)$	0.24	0.25	0.24	0.24	0.25	0.26	0.28	0.30	0.24	0.26	0.50	0.48
MSE_1	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.02
MSE_2	0.07	0.07	0.07	0.07	0.07	0.08	0.12	0.10	0.06	0.08	0.41	0.28

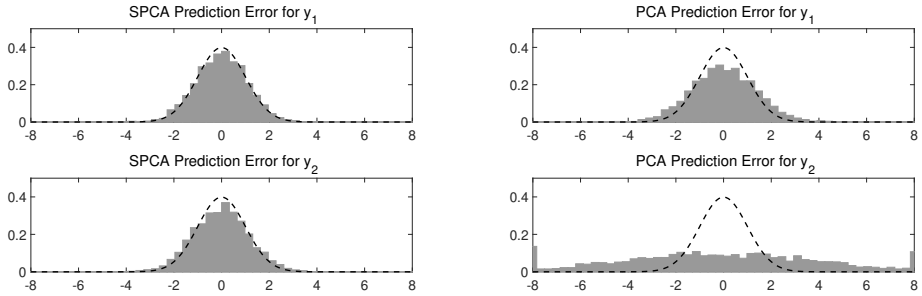
Further, we examine the quality of factor-space recovery and inference. We simulate a multivariate target with $\alpha = \mathbb{I}_2$ and $\alpha_w = (0_{2\times 1}, 0.2\mathbb{I}_2)$, yielding $y_{i,t+1} = f_{it} + 0.2y_{it} + z_{i,t+1}$ for $i = 1, 2$. Table 3 reports the performance of SPCA, SPCA-NI, PCA, and PLS for each component of the bivariate target y. Scaled PCA is omitted since its framework does not naturally extend to multi-target settings. As noted earlier, SPCA is expected to recover both factors because each factor loads on at least one component of y. Given that the true factor space is two-dimensional, we focus on results with $K = 2$ to evaluate each method's ability to recover the factors. We report the distance between $\widehat{F}$ and the true factor space, measured

by $d(\widehat{F}, F) = \left\|\mathbb{P}_{\widehat{F}'} - \mathbb{P}_{F'}\right\|$, along with the prediction errors MSE_i for $\widehat{y}_{i,T+1}$, $i = 1, 2$, where MSE_2 is affected by the potentially weak factor f_2.

The findings align with our theoretical predictions. As a decreases from 0.5 to 0.025, detecting f_{2t} becomes increasingly difficult, and SPCA-NI, PCA, and PLS exhibit larger distances $d(\widehat{F}, F)$ relative to SPCA. The resulting distortion in the recovered factor space leads to higher prediction errors for y_2, which relies on f_{2t}. In contrast, SPCA consistently maintains accurate factor-space recovery and low prediction errors across all values of a, demonstrating robustness to weak factors.

Finally, Figure 1 presents histograms of the standardized prediction errors based on the CLT in Theorem 3.4. The setup mirrors that of Table 3, with $a = 0.025$ and $T = 120$. For SPCA, the histograms align closely with the standard normal density, confirming the validity of the derived CLT. In contrast, PCA exhibits clear deviations from normality for y_2, reflecting the effect of the weak factor.

Figure 1: Histograms of the Standardized Prediction Errors. We provide histograms of standardized prediction errors for each entry of y using SPCA and PCA, respectively, based on 3,000 Monte Carlo repetitions. The dashed curve on each plot corresponds to the standard normal density.



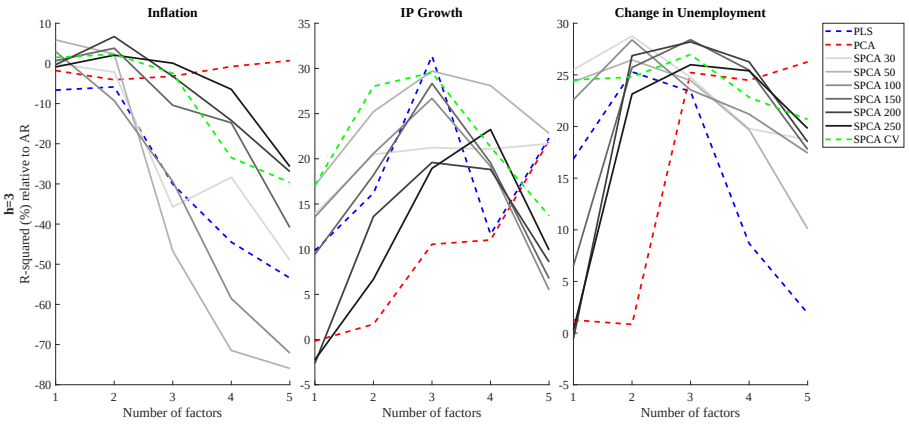
5 Empirical Analysis

Predicting macroeconomic variables such as output, unemployment and inflation often requires handling a large pool of potential predictors, motivating the use of dimension-reduction methods. SPCA provides a supervised approach that prioritizes predictors relevant for the forecasting target, making it well suited for settings with many noisy or weakly informative

variables. To evaluate its performance, we apply SPCA in an environment with a substantial number of predictors, combining a standard panel of macroeconomic variables with a large dataset of individual forecasts from professional forecasters (see the appendix for data details). The use of forecasts in macroeconomic prediction is well established, either through consensus forecasts (Faust and Wright (2013)) or optimal forecast combination (Genre et al. (2013)). In our application, SPCA automatically determines which individual forecasts, if any, complement the macroeconomic predictors, thereby providing a data-driven approach to forecast combination.

We focus on predicting at the quarterly horizon, a standard timeframe in the literature, with implementation details provided in Section S3.2 of the appendix. Figure 2 displays the out-of-sample R^2 of various forecasting methods relative to the autoregressive (AR) model benchmark for inflation (left panel), industrial production growth (center panel), and change in unemployment (right panel). In this exercise, the number of factors K is fixed. PCA (red line) and PLS (blue line) require no additional tuning parameters. For SPCA, the figure shows separate results for each K (grey lines) and for the value of $\lfloor qN \rfloor$ selected via CV (green line).

Figure 2: OOS Performance of SPCA, PCA and PLS (for different number of factors). Each panel reports the out-of-sample R^2 relative to the AR model for a different target, aggregated over 3 months. The three panels predict inflation, industrial production growth and change in unemployment rate, respectively. The green dashed line shows the performance of SPCA with 3-fold cross validation for the tuning parameter $\lfloor qN \rfloor$. The grey lines show the performance of SPCA with fixed number of predictors, $\lfloor qN \rfloor$. The blue dashed line uses PLS. The red dashed line uses PCA. Rolling window of 240 months is used. Sample covers 1993-2022.

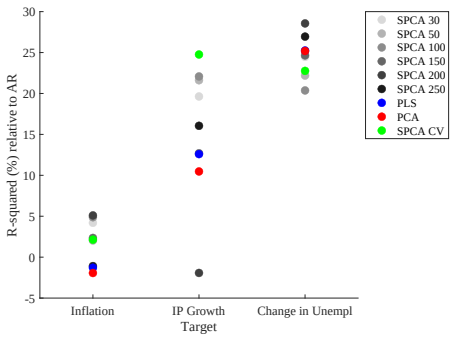


The figure reveals several interesting findings. First, predicting inflation beyond what an AR model achieves is challenging (see also Faust and Wright (2013)), with out-of-sample R^2 values generally near zero or negative. Among all methods, only SPCA delivers positive R^2 values, and it achieves this with a small number of factors. Predictability beyond the AR model is significantly higher for IP growth and unemployment. Second, SPCA generally outperforms PCA and PLS for most choices of the number of factors. Third, the performance of all methods is sensitive to the number of factors. Methods that use target information, such as PLS and SPCA with smaller $\lfloor qN \rfloor$, show significant performance declines as the number of factors increases, reflecting an potential issue of overfitting discussed in Appendix S2.

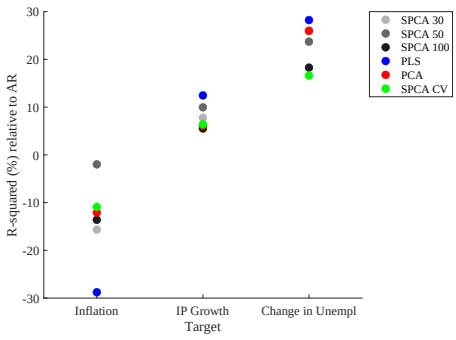
Given the critical role of the number of factors in determining out-of-sample performance, we use CV to select the number of factors for all three methods. For SPCA, this involves jointly selecting $\lfloor qN \rfloor$ and K. The results are presented in the left panel of Figure 3, where all three targets are shown together. The panel confirms that SPCA generally performs well, often outperforming the alternatives (for unemployment, several choices of the tuning parameter $\lfloor qN \rfloor$ outperform PCA and PLS, but not the one chosen by CV).

Figure 3: OOS Performance of SPCA, PCA and PLS (using CV to choose the number of factors). The left panel of this figure repeats the analysis of Figure 2, but chooses the number of factors via CV. The right panel performs the same analysis as the left panel, but using only Fred data.

(a) Fred+BlueChip



(b) Fred Only



To assess the impact of individual forecasts data on SPCA's performance, the right panel of Figure 3 presents results from the same analysis but using only the Fred dataset. The results indicate that SPCA's performance, while remaining broadly comparable to PCA and PLS,

deteriorates in this more limited setting. PLS, however, shows mixed results—performing well for industrial production growth and unemployment but failing for inflation. For reason of space, further results on SPCA-based forecasts are provided in the appendix.

Our analysis highlights two key findings: First, individual expert forecasts add value in predicting macroeconomic variables, corroborating the findings in Faust and Wright (2013). Second, SPCA is particularly effective when applied to a large, informative, but noisy set of individual forecasts, leveraging its ability to filter out noise while retaining valuable signals.

6 Conclusion

The availability of large datasets has spurred the development of methods to reduce predictor dimensionality, aiming to balance parsimony with improved out-of-sample predictions. SPCA introduces a key innovation: discarding uninformative predictors by using the target variable to guide selection. First proposed in Bair and Tibshirani (2004), this idea incorporates a screening step before factor extraction. However, the original SPCA approach assumes all factors can be extracted from the same subset of predictors, a restrictive condition rarely met in practice. We address this limitation by proposing a new SPCA methodology that iteratively combines selection, factor extraction, and projection.

Our theoretical framework highlights a key distinction in how methods handle weak factors. Unsupervised approaches tend to miss signals below a certain strength, making weak factors hard or impossible to recover. SPCA is able to overcome this limitation by iteratively leveraging the target variable to extract weak but relevant signals, while ignoring factors unrelated to prediction. This property makes it especially effective for forecasting, where excluding non-target-related factors does not harm accuracy. By contrast, other supervised approaches such as Scaled PCA or PLS do not succeed in this setting, underscoring that SPCA's design is crucial to its success.

References

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

Bai, J. and S. Ng (2023). Approximate factor models with weaker loadings. *Journal of Econometrics 235*(2), 1893–1916.

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

Bailey, N., G. Kapetanios, and M. H. Pesaran (2021). Measurement of factor strength: Theory and practice. *Journal of Applied Econometrics 36*(5), 587–613.

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.

Cai, T. T. and L. Wang (2011). Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Transactions on Information Theory 57*(7), 4680–4688.

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.

Choi, J. and M. Yuan (2025). High dimensional factor analysis with weak factors. *Journal of Econometrics 252*, 106086.

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.

Efron, B., T. Hastie, I. M. Johnstone, and R. Tibshirani (2004). Least angle regression. *Annals of Statistics 32*(2), 407–499.

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

Faust, J. and J. H. Wright (2013). Forecasting inflation. In *Handbook of economic forecasting*, Volume 2, pp. 2–56. Elsevier.

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

Gao, Z. and R. S. Tsay (2024). Supervised dynamic pca: Linear dynamic forecasting with many predictors. *Journal of the American Statistical Association*, 1–15.

Genre, V., G. Kenny, A. Meyler, and A. Timmermann (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting 29*(1), 108–121.

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 (2022). Scaled pca: A new approach to dimension reduction. *Management Science 68*(3), 1678–1695.

Huang, S.-C. and R. S. Tsay (2024). Time series forecasting with many predictors. *Mathematics 12*(15), 2336.

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.

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 (2022). Estimation of sparsity-induced weak factor models. *Journal of Business & Economic Statistics 41*(1), 213–227.

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.

Conflict of Interest Statement

The authors report there are no competing interests to declare.

Supplement to "Prediction When Factors are Weak"

Stefano Giglio
Yale University
Dacheng Xiu
University of Chicago
Dake Zhang
Shanghai Jiao Tong University

Abstract

This appendix presents the covariance-matrix estimator and its asymptotic justification, analyzes several alternative methods, provides additional empirical applications, and concludes with the mathematical proofs.

S1 Estimation of Φ_1 and Φ_2

Recall that from the outputs of Algorithm 1, we have defined $\widehat{\underline{F}}$, $\widehat{\beta}$, and $\widehat{\alpha}$. As a result, we can also estimate $\widehat{Z} = Y - \widehat{\alpha}\widehat{\underline{F}} - \widehat{\alpha}_w \underline{W}$ and $\widehat{\underline{U}} = \underline{X} - \widehat{\beta}\widehat{\underline{F}} - \widehat{\beta}_w \underline{W}$. Then we can construct Newey-West-type estimators for Π_{11}, Π_{12} and Π_{22}, given that each component of them can be estimated based on their sample analog constructed above. Estimators of Σ_f and Σ_w can be obtained by $\widehat{\Sigma}_f = T_h^{-1}\widehat{\underline{F}}\widehat{\underline{F}}'$ and $\widehat{\Sigma}_w = T_h^{-1}\underline{WW}'$. With $\widehat{f}_T = \widehat{\zeta}'(x_T - \widehat{\beta}_w w_T)$, $\widehat{\Phi}_1$ can be constructed as follows:

$$\widehat{\Phi}_1 = \left((\widehat{f}_T', w_T')\widehat{\Sigma}_{f,w}^{-1} \otimes \mathbb{I}_D\right) \begin{pmatrix} \widehat{\Pi}_{11} & \widehat{\Pi}_{12} \\ \widehat{\Pi}_{12}' & \widehat{\Pi}_{22} \end{pmatrix} \left(\widehat{\Sigma}_{f,w}^{-1}(\widehat{f}_T', w_T')' \otimes \mathbb{I}_D\right).$$

The above estimators are built as if the latent factors were observed. This is because any rotation matrix involved with latent factor estimates is canceled out, which eventually yields a consistent estimator of Φ_1. This part of the asymptotic variance is straightforward to implement, thanks to the fact that it does not involve estimation of high-dimensional quantities like Σ_u. The proof of consistency of $\widehat{\Phi}_1$ follows directly from Giglio and Xiu (2021) and is thus omitted here.

With respect to Φ_2, we may apply a thresholding estimator of $\Sigma_u = \mathrm{Cov}(u_t)$ following Fan et al. (2013). In detail, $\widehat{\Sigma}_u$ can be constructed by

$$(\widehat{\Sigma}_u)_{ij} = \begin{cases} (\widetilde{\Sigma}_u)_{ij}, & i = j \\ s_{ij}\left((\widetilde{\Sigma}_u)_{ij}\right), & i \neq j \end{cases}, \qquad \widetilde{\Sigma}_u = T_h^{-1}\widehat{\underline{U}}\widehat{\underline{U}}',$$

where $s_{ij}(\cdot)$ is a general thresholding function with an entry-dependent threshold τ_{ij} satisfying (i) $s_{ij}(z) = 0$ when $|z| \leq \tau_{ij}$ (ii) $|s_{ij}(z) - z| \leq \tau_{ij}$. The adaptive threshold can be chosen by $\tau_{ij} = C(\log NT)^{1/2}(T^{-1/2}N^{\nu/2} + q^{-1/2}N^{-1/2+\nu/2})\widehat{\theta}_{ij}^{1/2}$, where $C > 0$ is a sufficiently large constant and

$$\widehat{\theta}_{ij} = \frac{1}{T_h}\sum_{t \leq T_h}(\widehat{\underline{u}}_{it}\widehat{\underline{u}}_{jt} - (\widetilde{\Sigma}_u)_{ij})^2,$$

where $\widehat{\underline{u}}_{it}$ are the entries of $\widehat{\underline{U}}$. With $\widehat{\Sigma}_u$, Φ_2 can be estimated by $\widehat{\Phi}_2 = qN^{1-\nu}\widehat{\gamma}\widehat{\Sigma}_u\widehat{\gamma}'$.

The following theorem ensures the consistency of $\widehat{\Phi}_2$ under standard assumptions as in Fan et al. (2013).

Theorem S1.1. *Under the assumptions of Theorem 3.4, we further assume that*

(i) u_t *is stationary with* $\mathrm{E}(u_t) = 0$ *and* $\Sigma_u = \mathrm{Cov}(u_t)$ *satisfying* $C_1 > \lambda_1(\Sigma_u) \geq \lambda_N(\Sigma_u) > C_2$ *and* $\min_{i,j} \mathrm{Var}(u_{it}u_{jt}) > C_2$ *for some constant* $C_1, C_2 > 0$,

(ii) u_t *has exponential tail, i.e., there exist* $r_1 > 0$ *and* $C > 0$, *such that for any* $s > 0$ *and* $i \leq N$, $\mathrm{P}(|u_{it}| > s) \leq \exp(-(s/C)^{r_1})$.

(iii) u_t *is strong mixing, i.e., there exist positive constants* r_2 *and* C *such that for all* $t \in \mathbb{Z}^+$, $\alpha(t) \leq \exp(-Ct^{r_2})$, *where* $\alpha(T) = \sup_{A\in\mathcal{F}^0_{-\infty}, B\in\mathcal{F}^\infty_T} |\mathrm{P}(A)\mathrm{P}(B) - \mathrm{P}(AB)|$ *and* $\mathcal{F}^0_{-\infty}$, $\mathcal{F}^\infty_T$ *are* σ*-algebras generated by* $\{u_t\}_{-\infty\leq t\leq 0}$, $\{u_t\}_{T\leq t\leq\infty}$.

(iii) $(\log N)^{6(3r_1^{-1}+r_2^{-1}+1)} = o(T)$, $T = o(q^2N^{2-2\nu})$.

Then $\widehat{\Sigma}_u$ *satisfies* $\left\|\widehat{\Sigma}_u - \Sigma_u\right\| \lesssim_{\mathrm{P}} \mathcal{C}_q := m_{q,N}\left((\log NT)^{1/2}(T^{-1/2}N^{\nu/2} + q^{-1/2}N^{-1/2+\nu/2})\right)^{1-q}$, *where* $m_{q,N} = \max_{i\leq N}\sum_{j\leq N}|(\Sigma_u)_{ij}|^q$. *In addition, if* $\mathcal{C}_q = o(1)$, *then* $\widehat{\Phi}_2 \xrightarrow{\mathrm{P}} \Phi_2$.

S2 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 S2.1. 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 $\|A\| \lesssim 1$.

S2.1 Principal Component Regression

Formally, we present the algorithm below:

Algorithm S2.1 PCA Regression

1: **Inputs**: $\overline{Y}$, $\underline{X}$, $\underline{W}$, x_T, w_T and K.
2: **Initialization**: $Y_{(1)} := \overline{Y}\mathbb{M}_{\underline{W}'}$, $X_{(1)} := \underline{X}\mathbb{M}_{\underline{W}'}$.
3: 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 $\widehat{\alpha} = \overline{Y}\widehat{\underline{F}}'_{PCA}\left(\widehat{\underline{F}}_{PCA}\widehat{\underline{F}}'_{PCA}\right)^{-1}$.
4: Obtain $\widehat{\gamma} = \widehat{\alpha}\widehat{\varsigma}'$ and output the prediction $\widehat{y}^{PCA}_{T+h} = \widehat{\gamma}x_T + (\widehat{\alpha}_w - \widehat{\gamma}\widehat{\beta}_w)w_T$, where $\widehat{\alpha}_w = \overline{Y}\underline{W}'(\underline{WW}')^{-1}$ and $\widehat{\beta}_w = \underline{XW}'(\underline{WW}')^{-1}$.
5: **Outputs**: $\widehat{y}^{PCA}_{T+h}$, $\widehat{\underline{F}}_{PCA}$, $\widehat{\alpha}$, $\widehat{\alpha}_w$, $\widehat{\beta}_w$, and $\widehat{\gamma}$.

Proposition S2.1. *In Example S2.1, 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 S2.1 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 S2.1 is not consistent.

Proposition S2.2. *Under the same assumptions as in Proposition S2.1, if we further assume* $A_1' A_1 = \mathbb{I}_{T-h}$, *then we have* $\widehat{y}_{T+h}^{PCA} \xrightarrow{\text{P}} (1+\delta)^{-1}\text{E}_T(y_{T+h})$.

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, estiamtors of 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.

S2.2 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 in Algorithm S2.2. The PLS estimator has a closed-form formula if Y is a $1 \times T$ vector and a single factor model is estimated ($K = 1$):

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

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

Algorithm S2.2 PLS Regression

1: **Inputs**: $\overline{Y}$, $\underline{X}$, $\underline{W}$, x_T, w_T and K.
2: **Initialization**: $Y_{(1)} := \overline{Y}\mathbb{M}_{\underline{W}'}$, $X_{(1)} := \underline{X}\mathbb{M}_{\underline{W}'}$.
3: **for** $k = 1, 2, \cdots, K$ **do**
4: Obtain the weight vector $\widehat{\varsigma}_{(k)}$ from the largest left singular vector of $X_{(k)}Y'_{(k)}$.
5: Estimate the kth factor as $\widehat{\underline{F}}_{(k)} = \widehat{\varsigma}'_{(k)}X_{(k)}$.
6: Estimate coefficients $\widehat{\alpha}_{(k)} = Y_{(k)}\widehat{\underline{F}}'_{(k)}\left(\widehat{\underline{F}}_{(k)}\widehat{\underline{F}}'_{(k)}\right)^{-1}$ and $\widehat{\beta}_{(k)} = X_{(k)}\widehat{\underline{F}}'_{(k)}\left(\widehat{\underline{F}}_{(k)}\widehat{\underline{F}}'_{(k)}\right)^{-1}$.
7: 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 $Y_{(k+1)} = Y_{(k)} - \widehat{\alpha}_{(k)}\widehat{\underline{F}}_{(k)}$.
8: **end for**
9: 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 $\widehat{\alpha}_w = \overline{Y}\underline{W}'(\underline{WW}')^{-1}$ and $\widehat{\beta}_w = \underline{XW}'(\underline{WW}')^{-1}$.
10: **Outputs**: $\widehat{y}^{PLS}_{T+h}$, $\widehat{\underline{F}}_{PLS} := (\widehat{\underline{F}}'_{(1)}, \ldots, \widehat{\underline{F}}'_{(\widehat{K})})'$, $\widehat{\alpha}$, $\widehat{\alpha}_w$, $\widehat{\beta}_w$, and $\widehat{\gamma}$.

Proposition S2.3. *In Example S2.1, suppose that* $N/(N_0T) \to \delta \geq 0$ *and* $\|\beta\| \to \infty$, *then the estimated factor* $\widehat{\underline{F}}_{PLS}$ *satisfies*

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

where $\eta_{PLS} = (\mathbb{I}_{T-h} + \delta A'_1A_1)\underline{F}'$. *In the special case that* $A'_1A_1 = \mathbb{I}_{T-h}$, *it satisfies*

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

Proposition S2.4. *Under the assumptions of Proposition S2.3, if we further assume that* $A'_1A_1 = \mathbb{I}_{T-h}$, *then we have* $\widehat{y}^{PLS}_{T+h} \xrightarrow{\text{P}} (1+\delta)^{-1}\text{E}_T(y_{T+h})$.

Therefore, the consistency of the PLS factor also depends on the homoskedasticity assumption $A'_1A_1 = \mathbb{I}_{T-h}$ 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 S2.2. 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 S2.5. *In Example S2.2, if we use $\widehat{K} = 1$, then we have $\widehat{y}_{T+h}^{PLS} \gtrsim_P N^{3/2}T^{1/2}/(N^2 + T^2)$ while $\widehat{y}_{T+h}^{PCA} \lesssim_P 1/(N^{1/2} + T^{1/2})$. Specifically, in the case of $N \asymp T$, $\widehat{y}_{T+h}^{PLS} \gtrsim_P 1$ and $\widehat{y}_{T+h}^{PCA} \lesssim_P N^{-1/2}$.*

The conditional expectation of y_{T+h} is 0 in this example, but $\widehat{y}_{T+h}^{PLS}$ can be bounded away from 0 when using more factors than necessary. In contrast, $\widehat{y}_{T+h}^{PCA}$ 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 avoid overfitting.

S2.3 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 S2.1. 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 S2.1.

Formally, we present their algorithm in Algorithm S2.3. 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 S2.2. In other words, even though this is not true in general, $\widehat{y}_{T+h}^{SW}$

Algorithm S2.3 SW-PCA

1: **Inputs**: $\overline{Y}$, X, W and K.
2: 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: $\widehat{\alpha} = \overline{Y}\mathbb{M}_{\underline{W}'}\underline{\widehat{F}}'_{SW}\left(\underline{\widehat{F}}_{SW}\mathbb{M}_{\underline{W}'}\underline{\widehat{F}}'_{SW}\right)^{-1}$[1] and $\widehat{\alpha}_w = \overline{Y}M_{\underline{\widehat{F}}'_{SW}}\underline{W}'\left(\underline{W}\mathbb{M}_{\underline{\widehat{F}}'_{SW}}\underline{W}'\right)^{-1}$.
3: 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 $\widehat{\alpha}_w = \overline{Y}\underline{W}'(\underline{W}\underline{W}')^{-1}$.
4: **Outputs**: $\widehat{y}^{SW}_{T+h}$, $\widehat{F}_{SW}$, $\widehat{\alpha}$, and $\widehat{\alpha}_w$.

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.

Example S2.3. 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 S2.6. *In Example S2.3, 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 S2.3 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 S2.1 for robustness.

S3 Additional Empirical Details

In this section, we provide additional empirical details and further results to highlight the empirical advantages of our SPCA.

S3.1 Data

Our empirical exercise combines two datasets. First, we use the standard Fred-Md database (McCracken and Ng, 2016) that contains 127 monthly macroeconomic and financial series.[2]

[2]The series are grouped in the following categories: output and income; labor market; housing; consumption, orders and inventories; money and credit; interest and exchange rates; prices; stock market. The dataset applies a variety of transformations to the underlying series, which we follow in our analysis. We however make a few adjustments to the series' data transformations, to ensure that all series are stationary and based on economic reasoning. For the Effective Federal Funds Rate (FEDFUNDS), we keep its level (i.e., no transformation) instead of taking the first difference. We also compute the first difference of natural log instead of the second difference of natural log for the following series: M1 Money Stock (M1SL), M2 Money Stock (M2SL), Board of Governors Monetary Base (BOGMBASE; note: starting from the January 2020 (2020-01) vintage, BOGMBASE replaced the St. Louis Adjusted Monetary Base (AMBSL)), Total Reserves of Depository Institutions (TOTRESNS), Commercial and Industrial Loans (BUSLOANS), Real Estate Loans at All Commercial Banks (REALLN), Total Nonrevolving Credit (NONREVSL), Finished Goods (WPSFD49207), Finished Consumer Goods (WPSFD49502), Processed Goods for Intermediate Demand (WPSID61), Unprocessed Goods for Intermediate Demand (WPSID62; note: starting from the March 2016 (2016-03) vintage, PPI: Finished Goods (PPIFGS), PPI: Finished Consumer Goods (PPIFCG), PPI: Intermediate Materials (PPIITM), and PPI: Crude Materials (PPICRM) have been replaced with WPSFD49207, WPSFD49502, WPSID61, and WPSID62 respectively), Crude Oil, spliced WTI and Cushing (OILPRICEx), PPI: Metals and Metal Products (PPICMM), Consumer Price Index for All Urban Consumers (CPIAUCSL), CPI: Apparel (CPIAPPSL), CPI: Transportation (CPITRNSL), CPI: Medical Care (CPIMEDSL), CPI: Commodities (CUSR0000SAC), CPI: Durables (CUSR0000SAD), CPI: Services (CUSR0000SAS), CPI: All Items Less Food (CPIULFSL), CPI: All Items Less Shelter (CUSR0000SA0L2)[3], CPI: All Items Less Medical Care (CUSR0000SA0L5), Personal Cons. Exp: Chain Index (PCEPI), Personal Cons. Exp: Durable Goods (DDURRG3M086SBEA), Personal Cons. Exp: Nondurable Goods (DNDGRG3M086SBEA), Personal Cons. Exp: Services (DSERRG3M086SBEA), Avg Hourly Earnings: Goods-Producing (CES0600000008), Avg

The Fred-Md data spans the period March 1959 to February 2022. Second, we use individual forecasts from the Blue Chip Financial Forecasts data, which is a monthly survey of experts from various major financial institutions[4] and provides forecasts of interest rates and many other macroeconomic quantities[5] for each of the next six quarters (i.e., current quarter t through $t+5$), for a total of hundreds of forecasts every month. Our data covers the period February 1993 to February 2022 and we use all forecasts available (for all possible macroeconomic targets) as potential predictors. This gives us up to 18,053 different individual forecasts that could in theory be used as predictors (though, as discussed below, many of these forecasts are available for only a small number of periods, so they are not used in our analysis). Given that the Blue Chip forecast is only available since 1993, we conduct all of our analysis for the period February 1993 to February 2022.

S3.2 Out of Sample Forecast Evaluation

We forecast each of the three targets (inflation, industrial production growth, and change in the unemployment rate) using a rolling out of sample procedure. We evaluate the out of sample forecast of SPCA and compare it with two alternative forecasting methods, PCA and PLS. We choose these alternatives because each is a prominent example of a class of methods used in large-dimensional macroeconomic forecasting (respectively, unsupervised and supervised dimension reduction). Each of the three methods we evaluate (SPCA, PCA, PLS) is benchmarked to the forecast of an autoregressive model, whose number of lags is selected by the BIC criterion with a maximum lag of 12 lags, using a direct projection approach (Marcellino et al. (2006); Faust and Wright (2013)). We study forecast horizons of 1 to 12

Hourly Earnings: Construction (CES2000000008), Avg Hourly Earnings: Manufacturing (CES3000000008), Consumer Motor Vehicle Loans Outstanding (DTCOLNVHFNM), Total Consumer Loans and Leases Outstanding (DTCTHFNM) and Securities in Bank Credit at All Commercial Banks (INVEST).

[4]For instance, Bank of America, Goldman Sachs & Co. and J.P. MorganChase.

[5]For instance, the percentage changes in Real GDP, the GDP Chained Price Index, the Consumer Price Index and a set of interest rates (e.g., Federal Funds, 3-month Treasury, Aaa as well as Baa Corporate Bonds).

months.

All of the analysis is performed using a rolling estimation on a 240-months window. At every time t starting at the last month of the window, we predict the cumulated macroeconomic variables from t to $t+h$, where h is the forecast horizon, as in Huang et al. (2022). Within each window, we only keep predictors that have less than 10% missing data points. For those series that are included but do have some missing data (mostly Blue Chip forecasts) we forward fill the last non-missing value. About half of the total of around 40 forecasters from BlueChip available in the average month have sufficiently long series of forecasts to be included in our analysis. All predictors are standardized within each window. Then, a forecast is made for $t+1$ using the three different methods, and these forecasts are then joined over time to compute the out-of-sample R^2 (relative to the AR benchmark). When we use the Blue Chip data, we also include dummies for month of the quarter, to account for the fact that the Blue Chip data makes forecasts for calendar quarters irrespective of the month.[6]

Recall that the SPCA procedure presented in Giglio et al. (2022) relies on two tuning parameters, K and $\lfloor qN \rfloor$, whereas PCA and PLS only rely on tuning K. To demonstrate the effect of tuning parameters, we report three versions of the results. We first show the performance of the forecasting methods for different (fixed) number of factors K and different (fixed) choice of $\lfloor qN \rfloor$. In this case, no tuning is needed for SPCA. We then show the performance of SPCA for each K, with a single tuning parameter of SPCA that drives the selection step $\lfloor qN \rfloor$ chosen via 3-fold cross-validation (CV) separately in each time window. Next, we show the results when both the number of factors K (for SPCA, PCA and PLS) and the tuning parameter $\lfloor qN \rfloor$ (for SPCA) are jointly chosen via CV. We consider a range of $\lfloor qN \rfloor$ from 50 to 300.

[6]For example, in January, February and March, the "current quarter" forecast always refers to Q1.

S3.3 Additional Results

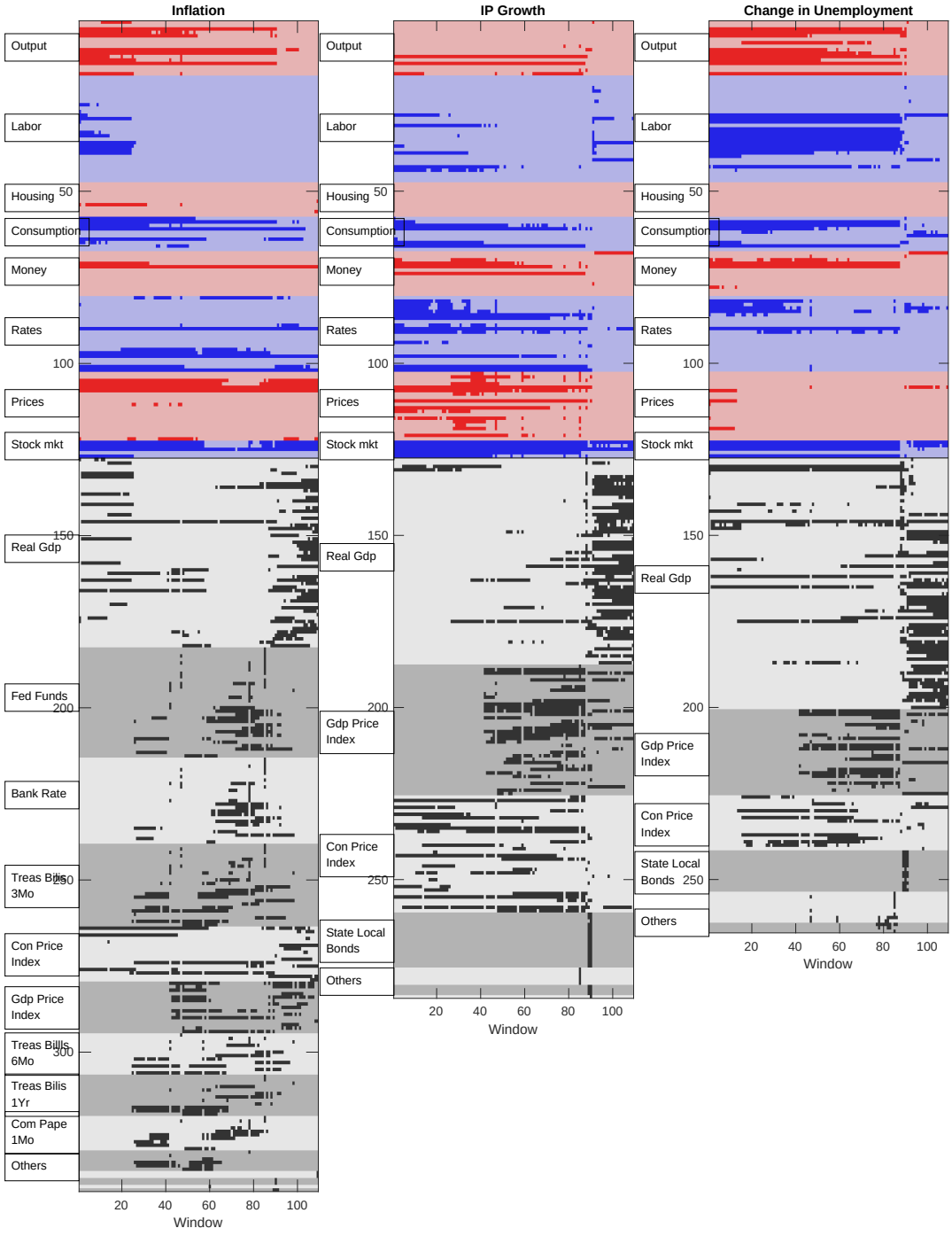
S3.3.1 Predictors Selected by SPCA

Building on the empirical analysis in the main text, we now examine in detail how SPCA selects predictors. Figure S1 shows which variables are chosen by SPCA to extract the first factor (focusing on the 50 with highest correlation with the target, for reasons of readability). For the three targets (one per column), the graph reports which variables were selected in each of the rolling windows in our sample. The top part of the graph collects the 127 Fred variables, grouped according to the standard Fred-Md categorization, in alternating blue and red colors. The bottom part corresponds to the BlueChip surveys, grouped by the target of the individual forecast (therefore, each row in this part of the graph is a forecast of a particular variable, at a particular horizon, by a particular expert). A darker color in this graph means that the variable is selected in that window.

Consider for example the inflation graph on the left. To extract a factor useful to predict inflation, SPCA selects a large number of variables from a few groups: output, consumption, rates, prices, and the stock market. Other groups are almost never selected. Rates are selected more for IP growth, and labor variables are selected more when predicting unemployment. Housing variables are rarely used for all three targets. Note that in many cases, the same predictors from each group are used, indicating that the predictive power of these macroeconomic variables is persistent.

To this macroeconomic set of predictors, SPCA adds a selection of individual forecasts from the BlueChip data as additional predictors. For reasons of space, the greyscale part of the graph shows a subset of these predictors: only those that are selected among the top 50 predictors at least in one window. The graph shows that different types of forecasts are used at different points in time, with some exceptions. Not surprisingly, to predict inflation, forecasts of the consumer price index are always included. To these forecasts, SPCA adds forecasts of GDP in the first and last part of the sample, and interest rates in the intermediate part of the

Figure S1: Top 50 Predictors Selected by SPCA. Under the same settings as Figure 2, each panel visualizes the top 50 predictors selected by SPCA across windows while predicting each target. The first set of variables (in red and blue) are Fred predictors, and the second set (in grey) corresponds to the BlueChip forecasts. For the latter set, only the predictors ever among the top 50 by correlation with the target are visualized.



sample. GDP forecasts are used throughout the sample to predict changes in unemployment, and become more dominant for all target variables toward the end of the sample, whereas inflation predictors tend to be more important beforehand. This switch is perhaps due to the fact that in the later part of the sample the zero lower bound was close or binding and inflation was low and not very volatile.

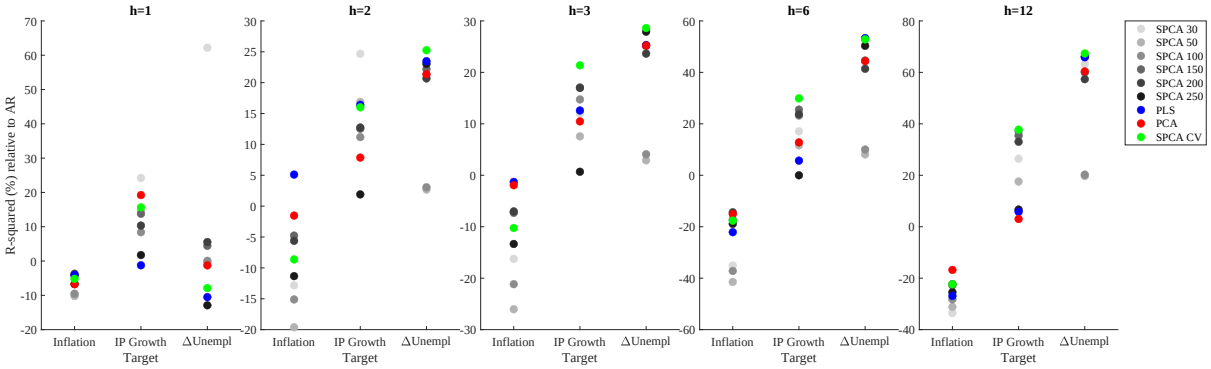
Finally, we note that not all Blue Chip forecasters are the same in terms of forecasting ability. Among the institutions whose forecasts are included in our analysis because they have a sufficiently long time series (each providing tens of forecasts, of different variables at different horizons), we find significant heterogeneity in the frequency with which their forecasts are selected by SPCA. For example, Nomura has its forecasts selected between 23% and 39% of the time at the first iteration (depending on the target). Swiss RE, on the other hand, has its forecasts selected only 0.1% of the time, for each target. This distribution is quite skewed: only 5 institutions have their forecasts selected more than 10% of the time for each target, out of the 20 included in our sample. Similar results hold when looking at selection at any iteration of SPCA.

S3.3.2 Joint Forecasts using Many Targets

Next, one special feature of SPCA is that it can operate the selection using a set of multiple targets jointly. In fact, using multiple targets is required by the theory (see Giglio et al. (2022)) to do inference, as long as there are more than one factors in the true DGP. We implement this here by predicting each target at horizons of 1, 2, 3, 6 and 12 months jointly. Figure S2 reports the out of sample R^2s on each horizon. There are two main results that this figure highlights. First, SPCA tends to do on average well at longer horizons (3, 6 and 12 months), whereas its performance is more uneven at shorter horizons. Second, comparing the middle panel (predicting one quarter ahead) with the left panel of Figure 3, which focused on the 3-month horizon only, we see that the use of other horizons to help select predictors has different effects for different targets. It significantly improves the forecasting ability for unemployment, but

reduces the forecasting ability for IP growth (mildly) and inflation (significantly so). Overall, the performance of SPCA remains on par with the other predictors when using multiple targets, especially at longer horizons.

Figure S2: OOS Performances - Different Targeted Horizons. Similar to Figure 3, but showing the out of sample R^2s at different horizons, and using all the horizons concurrently to estimate the factors in SPCA.



S3.3.3 Time Series of the Forecasts

Finally, we study the time series of our out-of-sample forecasts at different horizons, using the estimates obtained in Section S3.3.2, for horizons of 1, 2, 3, 6 and 12 months. Figure S3 reports SPCA's forecasts with asymptotic forecast standard errors at each maturity. In the figure, the blue dots represent the underlying time series that is the target of the forecast: log CPI, log IP, and unemployment, all scaled to start from 0 at the beginning of the sample. For readability, we show the forecasts every six months, each for horizons up to 12 months. Standard errors are obtained using the asymptotic distributions derived in Giglio et al. (2022), and are plotted in three shades (the 10th and 90th percentiles in the darkest shade, 5th and 95th in the middle shade, and 1st and 99th in the lightest shade).

Overall, SPCA does a good job forecasting the three series, with the forecasts often anticipating changes in the direction of the different variables. For example, IP forecasts predicted the increase starting in 2016, and the decrease that started in 2018. Of course, in other times the forecasts miss significantly, sometimes for several periods in the same direction. Two examples: first, forecasts do not fully anticipate the persistent decrease in unemployment

that occurred during 2013 and 2014. Second, all forecasts miss (as they should have) the unexpected and extraordinary events of the Covid pandemic (both the initial shock and the recovery). In that period, the point estimates change dramatically over a short period of time, and standard errors increase noticeably, demonstrating the large amount of uncertainty about the path of the economy during those times.

S4 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.

S4.1 Proof of Theorem 3.1

Proof. We start with the DGP without w_t first. Throughout the proof, we use $\widetilde{X}_{(k)} := \big(X_{(k)}\big)_{[\widehat{I}_k]}$ 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

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

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:

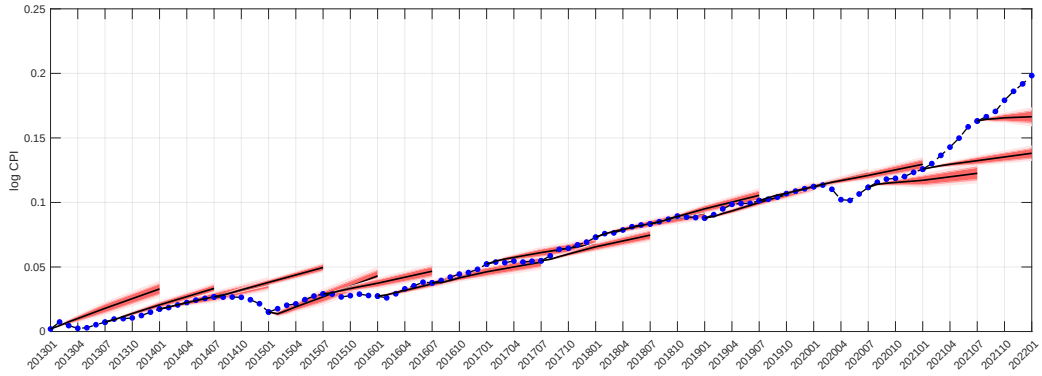
$$\widehat{\beta}_{(k)} = T_h^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}X_{(k)}\widehat{\xi}_{(k)} \quad \text{and} \quad \widehat{\alpha}_{(k)} = T_h^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}Y_{(k)}\widehat{\xi}_{(k)}. \tag{2}$$

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

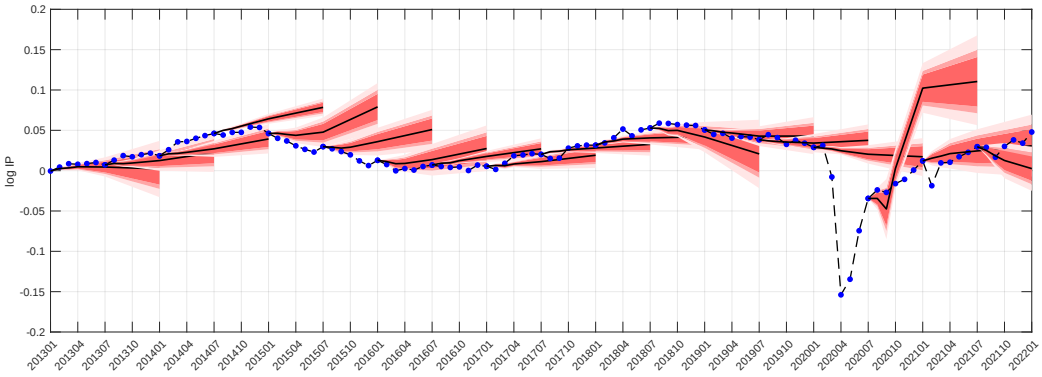
$$\widetilde{D}_{(k)} = (\mathbb{I}_N)_{[\widehat{I}_k]} - \sum_{i=1}^{k-1} T_h^{-1/2}\widehat{\lambda}_{(i)}^{-1/2}X_{[\widehat{I}_k]}\widehat{\xi}_{(i)}\widehat{\varsigma}'_{(i)}\widetilde{D}_{(i)},$$

Figure S3: Fan Charts. Using the same estimates as Figure S2, each panel shows the forecasts and confidence intervals for horizons up to 12 months. The forecasts are shown every 6 months, in alternating red and green colors (for readability). The blue dots are the cumulative targets of the forecasts.

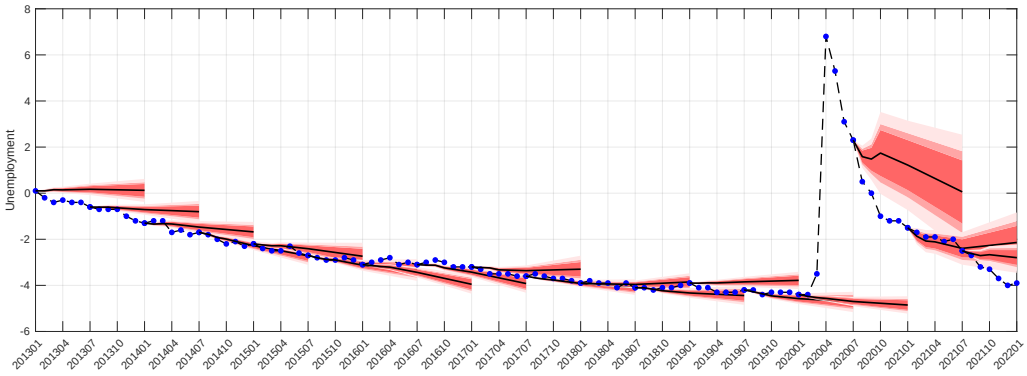
(a) Inflation



(b) IP Growth



(c) Change in Unemployment



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 S4.1, 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

$$\widetilde{X}_{(k)} = \left(X_{(k)}\right)_{[\widehat{I}_k]} = X_{[\widehat{I}_k]} \prod_{i=1}^{k-1} \mathbb{M}_{\widehat{F}'_{(i)}} = X_{[\widehat{I}_k]} \left(\mathbb{I}_{T_h} - \sum_{i=1}^{k-1} \widehat{\xi}_{(i)}\widehat{\xi}'_{(i)} \right), \tag{3}$$

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

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

Using (1), if $\widetilde{X}_{(i)} = \widetilde{D}_{(i)}X$ for any $i < k$ we can write (3) 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

$$\widehat{F}_{(k)} = \widehat{\varsigma}'_{(k)}\widetilde{X}_{(k)} = \widehat{\varsigma}'_{(k)}\widetilde{D}_{(k)}X, \tag{4}$$

for all k, and by definition, we have $\widehat{\zeta}_{(k)} = (\widehat{\varsigma}'_{(k)}\widetilde{D}_{(k)})'$. Moreover, using (2) the estimated coefficient $\widehat{\gamma}$ can be written as

$$\widehat{\gamma} = \sum_{k=1}^{\widehat{K}} \widehat{\alpha}_{(k)}\widehat{\varsigma}'_{(k)}\widetilde{D}_{(k)} = \sum_{k=1}^{\widehat{K}} T_h^{-1/2}\widehat{\lambda}_{(k)}^{-1/2} Y\widehat{\xi}_{(k)}\widehat{\varsigma}'_{(k)}\widetilde{D}_{(k)}. \tag{5}$$

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

$$\widetilde{X}_{(k)} = \widetilde{\beta}_{(k)}F + \widetilde{U}_{(k)}. \tag{6}$$

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

$$Y_{(k)} = Y\left(\mathbb{I}_{T_h} - \sum_{i=1}^{k-1}\widehat{\xi}_{(i)}\widehat{\xi}'_{(i)}\right) = \widetilde{\alpha}_{(k)}F + \widetilde{Z}_{(k)}, \tag{7}$$

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 S4.3, we have $\mathrm{P}(\widehat{I}_k = I_k) \to 1$ for $k \le \tilde{K}$ 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 3.1, using (6), 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 S4.5(i), $\left\|\widehat{F}_{(k)}\right\| = \sqrt{T_h\widehat{\lambda}_{(k)}}$, and $\|\mathbb{M}_{F'}\| \le 1$, we have

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

□

S4.2 Proof of Theorem 3.2

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 (7), 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 S4.1. 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 $T_h^{-1}\left\|X_{[i]}Y'_{(\tilde{K}+1)}\right\|_{\text{MAX}} \geq c$. Let S denote the set of these predictors. For $i \in S$, we have

$$\left\|T_h^{-1}X_{[i]}Y'_{(\tilde{K}+1)}\right\|_{\text{F}}^2 \lesssim \left\|T_h^{-1}X_{[i]}Y'_{(\tilde{K}+1)}\right\|_{\text{MAX}}^2 \lesssim_{\text{P}} N^{-\nu}, \tag{8}$$

where we use $\|\beta\|_{\text{MAX}} \lesssim N^{-\nu}$ from Assumption 2 and Lemma S4.3(vi) in the last step. On the other hand, in light of the set I_0 in Assumption 2, we have

$$\begin{aligned}\sum_{i\in I_0}\left\|T_h^{-1}X_{[i]}Y'_{(\tilde{K}+1)}\right\|_{\text{F}}^2 &= \sum_{i\in I_0\cap S}\left\|T_h^{-1}X_{[i]}Y'_{(\tilde{K}+1)}\right\|_{\text{F}}^2 + \sum_{i\in I_0\cap S^c}\left\|T_h^{-1}X_{[i]}Y'_{(\tilde{K}+1)}\right\|_{\text{F}}^2 \\ &\lesssim_{\text{P}} |I_0\cap S|N^{-\nu} + |I_0\cap S^c|c^2 \leq qN^{1-\nu} + c^2N_0 = o(N_0N^{-\nu}),\end{aligned} \tag{9}$$

where we use (8), $|S| \leq qN-1$, $cN^{\nu/2}\to 0$, and $qN/N_0 \to 0$. Consequently, (9) leads to $\left\|Y_{(\tilde{K}+1)}X'_{[I_0]}\right\| = o_{\text{P}}(N_0^{1/2}N^{-\nu/2}T)$. Moreover, using (7) and that $X = \beta F + U$, we can decompose

$$Y_{(\tilde{K}+1)}X'_{[I_0]} = \widetilde{\alpha}_{(\tilde{K}+1)}FF'\beta'_{[I_0]} + \widetilde{\alpha}_{(\tilde{K}+1)}FU'_{[I_0]} + \widetilde{Z}_{(\tilde{K}+1)}F'\beta'_{[I_0]} + \widetilde{Z}_{(\tilde{K}+1)}U'_{[I_0]}. \tag{10}$$

Using (9), (10), Lemma S4.8(i)(ii), and the fact that $\left\|\beta_{[I_0]}\right\| \lesssim N_0^{1/2}N^{-\nu/2}$, we have

$$\left\|\widetilde{\alpha}_{(\tilde{K}+1)}\left(FF'\beta'_{[I_0]} + FU'_{[I_0]}\right)\right\| = o_{\text{P}}\left(N_0^{1/2}N^{-\nu/2}T\right). \tag{11}$$

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

$$\begin{aligned}|\sigma_K(FF'\beta'_{[I_0]} + FU'_{[I_0]}) - \sigma_K(T_h\beta_{[I_0]})| &\leq \left\|FU'_{[I_0]}\right\| + T_h\left\|T_h^{-1}FF' - \mathbb{I}_K\right\|\left\|\beta_{[I_0]}\right\| \\ &\lesssim_{\text{P}} N_0^{1/2}T^{1/2}.\end{aligned} \tag{12}$$

Since Assumption 2 implies that $\sigma_K(\beta_{[I_0]}) \asymp N_0^{1/2}N^{-\nu/2}$, we have $\sigma_K(FF'\beta'_{[I_0]} + FU'_{[I_0]}) \asymp N_0^{1/2}N^{-\nu/2}T$. Using this result, (11) 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 $\left\|\widetilde{\alpha}_{(\tilde{K}+1)}\right\| \xrightarrow{\text{P}} 0$. That is, by definition of $\tilde{\alpha}_{(\tilde{K}+1)}$ in (7),

$$\left\|\alpha - \sum_{i=1}^{\tilde{K}} Y\widehat{\xi}_{(i)}\frac{\widehat{\varsigma}'_{(i)}\widetilde{\beta}_{(i)}}{\sqrt{T_h\widehat{\lambda}_{(i)}}}\right\| = o_{\text{P}}(1). \tag{13}$$

Next, (5) and $\widetilde{\beta}_{(k)} = \widetilde{D}_{(k)}\beta$ imply that

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

Therefore, (13) is equivalent to $\|\widehat{\gamma}\beta - \alpha\| = o_{\mathrm{P}}(1)$.

As shown in Lemma S4.11, 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

$$\begin{aligned}&\widehat{y}_{T+h} - \mathrm{E}_T(y_{T+h}) = \widehat{\gamma}x_T + (\widehat{\alpha}_w - \widehat{\gamma}\widehat{\beta}_w)w_T - \alpha f_T - \alpha_w w_T \\ =&(\widehat{\gamma}\beta - \alpha)\left(f_T - FW'(WW')^{-1}w_T\right) + \widehat{\gamma}(u_T - UW'(WW')^{-1}w_T) + ZW'(WW')^{-1}w_T.\end{aligned} \tag{14}$$

Using (5) and $\|Y\| \leq \|\alpha F\| + \|Z\| \lesssim_{\mathrm{P}} T^{1/2}$ by Assumption 1, we have

$$\|\widehat{\gamma}u_T\| \leq \sum_{k\leq\tilde{K}} T_h^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}\|Y\|\left\|\widehat{\xi}_{(k)}\right\||\widetilde{\varsigma}_{(k)}'\widetilde{D}_{(k)}u_T| \lesssim_{\mathrm{P}} \sum_{k\leq\tilde{K}}\widehat{\lambda}_{(k)}^{-1/2}|\widetilde{\varsigma}_{(k)}'\widetilde{D}_{(k)}u_T|, \tag{15}$$

and

$$\|\widehat{\gamma}UW'\| \leq \sum_{k\leq\tilde{K}} T_h^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}\|Y\|\left\|\widehat{\xi}_{(k)}\right\|\left\|\widetilde{\varsigma}_{(k)}'\widetilde{D}_{(k)}UW'\right\| \lesssim_{\mathrm{P}} \sum_{k\leq\tilde{K}}\widehat{\lambda}_{(k)}^{-1/2}\left\|\widetilde{\varsigma}_{(k)}'\widetilde{U}_{(k)}W'\right\|. \tag{16}$$

Together with $\|(WW')^{-1}\| \lesssim_{\mathrm{P}} T^{-1}$ from Assumption 1, $\widehat{\lambda}_{(k)} \asymp_{\mathrm{P}} qN^{1-\nu}$ from Lemma S4.3 and Lemma S4.5(ii)(iv), we have $\|\widehat{\gamma}u_T\| = o_{\mathrm{P}}(1)$ and $\|\widehat{\gamma}UW'(WW')^{-1}\| = o_{\mathrm{P}}(1)$. In addition, using $\|FW'\| \lesssim_{\mathrm{P}} T^{1/2}$, $\|ZW'\| \lesssim_{\mathrm{P}} T^{1/2}$ from Assumption 1 and $\|\widehat{\gamma}\beta - \alpha\| = o_{\mathrm{P}}(1)$, we show that each term of (14) vanishes, and hence $\widehat{y}_{T+h} - \mathrm{E}_T[y_{T+h}] \xrightarrow{\mathrm{P}} 0$. □

S4.3 Proof of Theorem 3.3

Proof. As in the proof of Theorem 3.1, we impose that $\widehat{K} = \tilde{K}$ and $\widehat{I}_k = I_k$, since Lemma S4.3 shows that both events occur with probability approaching 1. As shown in Lemma S4.2(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 3.3. Below we directly impose that $\widehat{K} = K$.

Again, following the same argument above (14), 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 3.1 implies $\left\|\widehat{\xi}'_{(k)}\mathbb{M}_{F'}\right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2+\nu/2} + T^{-1}N^{\nu}$ for $k \leq K$. Let v denote $F'(FF')^{-1/2}$, we have

$$\left\|\widehat{\xi} - \mathbb{P}_{F'}\widehat{\xi}\right\| = \left\|\widehat{\xi} - vv'\widehat{\xi}\right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2+\nu/2} + T^{-1}N^{\nu}, \tag{17}$$

where $\widehat{\xi}$ is a $T \times K$ matrix with each column equal to $\widehat{\xi}_{(k)}$. (17) implies that $\left\|\widehat{\xi}'vv'\widehat{\xi} - \mathbb{I}_K\right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2+\nu/2} + T^{-1}N^{\nu}$. By Weyl's inequality, $|\sigma_i(\widehat{\xi}'v) - 1| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2+\nu/2} + T^{-1}N^{\nu}$, for $1 \leq i \leq K$, and thus

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

Then, using this, (17), and the fact that $\|v\| = 1$ and $\left\|\widehat{\xi}\right\| = 1$, we have

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

Next, we need a more intricate analysis of $\widehat{\gamma}$. Recall from the proof of Theorem 3.2 that

$$\widehat{\gamma}\beta = \sum_{k=1}^{\tilde{K}} T_h^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}Y\widehat{\xi}_{(k)}\widehat{\varsigma}'_{(k)}\widetilde{\beta}_{(k)}. \tag{18}$$

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

$$b_{k1} = T^{-1/2}F\widehat{\xi}_{(k)}, \quad b_{k2} = \widehat{\lambda}_{(k)}^{-1/2}\widetilde{\beta}'_{(k)}\widehat{\varsigma}_{(k)}. \tag{19}$$

By Lemma S4.6,

$$\left\|T_h^{-1/2}Z\widehat{\xi}_{(k)} - T_h^{-1}ZF'b_{k2}\right\| \lesssim_{\mathrm{P}} T^{-1}N^{\nu} + q^{-1}N^{-1+\nu}. \tag{20}$$

As we impose that $\widehat{K} = \tilde{K} = K$, combining (18), (19) and (20), with $\|B_1\| \lesssim_{\mathrm{P}} 1$, $\|B_2\| \lesssim_{\mathrm{P}} 1$ from Lemma S4.9, we have

$$\left\|\widehat{\gamma}\beta - \alpha B_1B_2' - T_h^{-1}ZF'B_2B_2'\right\| \lesssim_{\mathrm{P}} T^{-1}N^{\nu} + q^{-1}N^{-1+\nu}. \tag{21}$$

Using Lemma S4.9(iv)(v), we obtain $\left\|\widehat{\gamma}\beta - \alpha - T_h^{-1}ZF'\right\| \lesssim_{\mathrm{P}} T^{-1}N^{\nu} + q^{-1}N^{-1+\nu}$. □

S4.4 Proof of Theorem 3.4

Proof. As in the proof of Theorem 3.2, we have $\|FW'(WW')^{-1}\| \lesssim_{\mathrm{P}} T^{-1/2}$ from Assumption 1 and $\|\widehat{\gamma}UW'(WW')^{-1}\| \lesssim_{\mathrm{P}} T^{-1} + q^{-1}N^{-1}$ as shown in (16). Together with $\left\|\widehat{\gamma}\beta - \alpha - T_h^{-1}ZF'\right\| \lesssim_{\mathrm{P}} T^{-1} + q^{-1}N^{-1}$, we can derive from (14) that:

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

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

$$\begin{aligned}
&\left\|ZW'(WW')^{-1}w_T - T_h^{-1}ZW'\Sigma_w^{-1}w_T\right\| \\
\leq& T_h^{-1}\left\|ZW'\right\|\left\|(T_h^{-1}WW')^{-1} - \Sigma_w^{-1}\right\|\left\|w_T\right\| \\
\lesssim_{\mathrm{P}}& T_h^{-1/2}\left\|T_h^{-1}\Sigma_w^{-1/2}WW'\Sigma_w^{-1/2} - \mathbb{I}_D\right\| = T_h^{-1/2}\max_{i\leq D}|\lambda_i\left(T_h^{-1}\Sigma_w^{-1/2}WW'\Sigma_w^{-1/2}\right)^{-1} - 1| \\
\lesssim_{\mathrm{P}}& T^{-1}.
\end{aligned} \tag{22}$$

For $\widehat{\gamma}u_T$, by (5), we have $\widehat{\gamma}u_T = \sum_{k=1}^{K}\widehat{\alpha}_{(k)}\widehat{\varsigma}'_{(k)}\widetilde{D}_{(k)}u_T$ and thus

$$\left\|\widehat{\gamma}u_T - \sum_{k=1}^{K}\lambda_{(k)}^{-1/2}\alpha b_{(k)}\varsigma'_{(k)}D_{(k)}u_T\right\| \leq \sum_{k=1}^{K}\left\|\widehat{\alpha}_{(k)}\widehat{\varsigma}'_{(k)}\widetilde{D}_{(k)}u_T - \lambda_{(k)}^{-1/2}\alpha b_{(k)}\varsigma'_{(k)}D_{(k)}u_T\right\|. \tag{23}$$

Lemma S4.7(ii) and Assumption 4 gives

$$|\widehat{\varsigma}'_{(k)}\widetilde{D}_{(k)}u_T - \varsigma'_{(k)}D_{(k)}u_T| \lesssim_{\mathrm{P}} T^{-1/2}N^{\nu/2} + q^{-1/2}N^{-1/2+\nu/2}. \tag{24}$$

In addition, (2) and Lemma S4.1 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 (20), $\|ZF'\| \lesssim_{\mathrm{P}} T^{1/2}$ and $\|b_{k2}\| \lesssim_{\mathrm{P}} 1$ from Lemma S4.9(i), this equation leads to

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

Using $\left\|b_{k2} - b_{(k)}\right\| \lesssim_{\mathrm{P}} T^{-1/2}N^{\nu/2} + q^{-1/2}N^{-1/2+\nu/2}$ implied by Lemma S4.9(iii) and $\widehat{\lambda}_{(k)} \asymp_{\mathrm{P}} qN^{1-\nu}$ from Lemma S4.3(iii), we have

$$\begin{aligned}
\left\|\widehat{\alpha}_{(k)} - \widehat{\lambda}_{(k)}^{-1/2}\alpha b_{(k)}\right\| &\leq \left\|\widehat{\alpha}_{(k)} - \widehat{\lambda}_{(k)}^{-1/2}\alpha b_{k2}\right\| + \left\|\widehat{\lambda}_{(k)}^{-1/2}\alpha(b_{(k)} - b_{k2})\right\| \\
&\lesssim_{\mathrm{P}} T^{-1/2}q^{-1/2}N^{-1/2+\nu} + q^{-1}N^{-1+\nu}.
\end{aligned} \tag{25}$$

Also, with Lemma S4.3(iii), we have

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

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

$$\left\|\widehat{\alpha}_{(k)} - \lambda_{(k)}^{-1/2} \alpha b_{(k)}\right\| \leq T^{-1/2} q^{-1/2} N^{-1/2+\nu} + q^{-1} N^{-1+\nu}. \tag{26}$$

For each term in the summation of (23), we have

$$\begin{aligned} &\left\|\widehat{\alpha}_{(k)} \widehat{\varsigma}'_{(k)} \widetilde{D}_{(k)} u_T - \lambda_{(k)}^{-1/2} \alpha b_{(k)} \varsigma'_{(k)} D_{(k)} u_T\right\| \\ \leq &\left\|\widehat{\alpha}_{(k)} (\widehat{\varsigma}'_{(k)} \widetilde{D}_{(k)} u_T - \varsigma'_{(k)} D_{(k)} u_T)\right\| + \left\|(\widehat{\alpha}_{(k)} - \lambda_{(k)}^{-1/2} \alpha b_{(k)}) \varsigma'_{(k)} D_{(k)} u_T\right\|. \end{aligned} \tag{27}$$

Note that (25) also implies $\left\|\widehat{\alpha}_{(k)}\right\| \lesssim_{\mathrm{P}} q^{-1/2} N^{-1/2+\nu/2}$ as $\widehat{\lambda}_{(k)} \asymp q N^{1-\nu}$, and that (24) implies the first term in (27) is $O_{\mathrm{P}}(T^{-1/2} q^{-1/2} N^{-1/2+\nu} + q^{-1} N^{-1+\nu})$. Furthermore, $|\varsigma'_{(k)} D_{(k)} u_T| \lesssim_{\mathrm{P}} 1$ from Lemma S4.5(iv) and (26) show that the second term in (27) is also $O_{\mathrm{P}}(T^{-1/2} q^{-1/2} N^{-1/2+\nu} + q^{-1} N^{-1+\nu})$. Given this, (23) becomes

$$\left\|\widehat{\gamma} u_T - \sum_{k=1}^{K} \lambda_{(k)}^{-1/2} \alpha b_{(k)} \varsigma'_{(k)} D_{(k)} u_T\right\| \lesssim_{\mathrm{P}} T^{-1/2} q^{-1/2} N^{-1/2+\nu} + q^{-1} N^{-1+\nu}. \tag{28}$$

To sum up, we have established that

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

In the general case that Σ_f may not be $\mathbb{I}_K$, the first term becomes $T_h^{-1} Z F' \Sigma_f^{-1} f_T$. 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 3.4. Using Delta method and Assumption 6, it is straightforward to obtain the desired CLT. □

S4.5 Proof of Theorem S1.1

Proof. Using Theorem 5 in Fan et al. (2013), to establish the error bound $\left\|\widehat{\Sigma}_u - \Sigma_u\right\|$, it is sufficient to show that $\left\|\widehat{U} - U\right\|_{\text{MAX}} = o_{\text{P}}(1)$ and

$$\max_{i\leq N} T_h^{-1}\sum_t |u_{it} - \widehat{u}_{it}|^2 \lesssim_{\text{P}} (\log NT)(T^{-1}N^{\nu} + q^{-1}N^{-1+\nu}).$$

These two estimates have been shown by Lemma S4.10(iii)(iv). If

$$m_{q,N}\left((\log NT)^{1/2}(T^{-1/2}N^{\nu/2} + q^{-1/2}N^{-1/2+\nu/2})\right)^{1-q} = o(1),$$

then $\left\|\widehat{\Sigma}_u - \Sigma_u\right\| = o_{\text{P}}(1)$. With $\left\|\widehat{\varsigma}_{(k)}'\widetilde{D}_{(k)} - \varsigma_{(k)}'D_{(k)}\right\| \lesssim_{\text{P}} T^{-1/2}N^{\nu/2} + q^{-1/2}N^{-1/2+\nu/2}$ from Lemma S4.7(ii) and $\widehat{\gamma} = \sum_{k\leq K}\widehat{\alpha}_{(k)}\widehat{\varsigma}_{(k)}'\widetilde{D}_{(k)}$, rewrite the proof of (28), we have

$$\left\|\widehat{\gamma} - \sum_{k\leq K}\lambda_{(k)}^{-1/2}\alpha b_{(k)}\varsigma_{(k)}'D_{(k)}\right\| \lesssim_{\text{P}} T^{-1/2}q^{-1/2}N^{-1/2+\nu} + q^{-1}N^{-1+\nu}. \tag{29}$$

Recall that $\lambda_{(k)}^{-1/2}\varsigma_{(k)}'D_{(k)}$ is exactly the kth row of $\Lambda^{-1}\Omega'\Psi$, the left hand side of (29) is equivalent to $\|\widehat{\gamma} - \alpha B\Lambda^{-1}\Omega'\Psi\|$. In addition, under the assumption $\text{Cov}(u_t) = \Sigma_u$, Π_{33} equals to $(qN^{1-\nu})^{-1}\Psi\Sigma_u\Psi'$. Let $\tilde{\gamma}$ denote $\alpha B\Lambda^{-1}\Omega'\Psi$, then we have

$$\widehat{\Phi}_2 - \Phi_2 = qN^{1-\nu}\left(\widehat{\gamma}\widehat{\Sigma}_u\widehat{\gamma}' - \tilde{\gamma}\Sigma_u\tilde{\gamma}'\right).$$

Consequently, we have

$$\left\|\widehat{\Phi}_2 - \Phi_2\right\| \leq qN^{1-\nu}\left\|\widehat{\gamma}(\widehat{\Sigma}_u - \Sigma_u)\widehat{\gamma}'\right\| + qN^{1-\nu}\left\|(\widehat{\gamma} - \tilde{\gamma})\Sigma_u\widehat{\gamma}'\right\| + qN^{1-\nu}\left\|\tilde{\gamma}\Sigma_u(\widehat{\gamma} - \tilde{\gamma})'\right\|. \tag{30}$$

Using the definition of $D_{(k)}$, $\left\|\beta_{[I_k]}\right\| \lesssim (qN^{1-\nu})^{1/2}$, and $\lambda_{(k)} \asymp qN^{1-\nu}$, we have $\|\widehat{\gamma}\| \lesssim q^{-1/2}N^{-1/2+\nu/2}$. Using $\left\|\widehat{\Sigma}_u - \Sigma_u\right\| = o_{\text{P}}(1)$, (29), $\|\Sigma_u\| \lesssim 1$ from the assumption and $\|\widehat{\gamma}\| \lesssim q^{-1/2}N^{-1/2+\nu/2}$, all three terms in (30) are $o_{\text{P}}(1)$. □

S4.6 Proof of Propositions S2.1 and S2.2

Proof. Note that for any orthogonal matrix $\Gamma \in \mathbb{R}^{N\times N}$, the estimators based on PCA and PLS on ΓR are the same as those based on R. Thus, without loss of generality, we can assume $\beta = (\lambda^{1/2}, 0, \cdots, 0)'$, where $\lambda = \|\beta\|^2$ and it will not affect A.

We can then write X in the following form:

$$X = \beta F + U = \beta F + \epsilon A_1 = \begin{pmatrix} \sqrt{\lambda} F + \epsilon_1 A_1 \\ \epsilon_2 A_1 \end{pmatrix}, \tag{31}$$

where ϵ_1 is the first row of ϵ and ϵ_2 contains the remaining rows. Correspondingly, we write the first left singular vector of X as $\widehat{\varsigma} = (\widehat{\varsigma}_1, \widehat{\varsigma}_2')'$, where $\widehat{\varsigma}_1$ is the first element of $\widehat{\varsigma}$ and $\widehat{\varsigma}_2$ is a vector of the remaining $N-1$ entries of $\widehat{\varsigma}$, write $\widehat{\xi}$ as the first right singular vector of X, and denote the first singular value as $\sqrt{T\widehat{\lambda}}$. By simple algebra we have

$$\widehat{\varsigma}_1 = \frac{(\sqrt{\lambda} F + \epsilon_1 A_1)\widehat{\xi}}{\sqrt{T\widehat{\lambda}}}, \quad \widehat{\varsigma}_2 = \frac{\epsilon_2 A_1 \widehat{\xi}}{\sqrt{T\widehat{\lambda}}}. \tag{32}$$

Since the entries of F are i.i.d. $\mathcal{N}(0,1)$, we have large deviation inequality $|T_h^{-1} F F' - 1| \lesssim_{\mathrm{P}} T^{-1/2}$. This also implies that $\|F\| - T_h^{-1/2} \lesssim_{\mathrm{P}} 1$ by Weyl's inequality.

Similarly, we can get $|T_h^{-1}\epsilon_1\epsilon_1' - 1| \lesssim_{\mathrm{P}} T^{-1/2}$ and $\|\epsilon_1\| - T_h^{-1/2} \lesssim_{\mathrm{P}} 1$. In addition, by Lemma A.1 in Wang and Fan (2017), we have $\|N^{-1}U'U - A_1'A_1\| \leq \|A_1\|^2 \|N^{-1}\epsilon'\epsilon - \mathbb{I}_{T_h}\| \lesssim_{\mathrm{P}} \sqrt{T/N}$. Next, by direct calculation using the previous inequalities we obtain

$$\left\| \frac{F'\epsilon_1 A_1 + A_1'\epsilon_1' F}{T_h\sqrt{\lambda}} + \frac{U'U - N A_1' A_1}{T_h \lambda} \right\| \lesssim_{\mathrm{P}} \frac{1}{\sqrt{\lambda}} + \frac{\sqrt{NT}}{T\lambda} \lesssim_{\mathrm{P}} \frac{1}{\sqrt{\lambda}}.$$

Together with (31), we have

$$\left\| \frac{X'X}{T_h\lambda} - \frac{F'F}{T_h} - \frac{N A_1' A_1}{T_h \lambda} \right\| \lesssim_{\mathrm{P}} \frac{1}{\sqrt{\lambda}}. \tag{33}$$

Let η denote the first eigenvector of the matrix $M := T_h^{-1}F'F + \delta A_1'A_1$. With the assumption that $N/(T\lambda) \to \delta$, $(\lambda_1(M) - \lambda_2(M))/\lambda_1(M) \gtrsim_{\mathrm{P}} 1$ and (33), by the sin-theta theorem in Davis and Kahan (1970), we have $\left\| \mathbb{P}_\eta - \mathbb{P}_{\widehat{\xi}} \right\| = \left\| \mathbb{P}_\eta - \mathbb{P}_{\widehat{F}'} \right\| = o_{\mathrm{P}}(1)$.

In the case that $A_1'A_1 = \mathbb{I}_{T_h}$, the eigenvalues of M are given by

$$\lambda_i = \begin{cases} T_h^{-1} F F' + \delta & i = 1; \\ \delta & i \geq 2. \end{cases} \tag{34}$$

and the first eigenvector is $F'/\|F\|$. Since the largest eigenvalue of $X'X/(T_h\lambda)$ is $\widehat{\lambda}/\lambda$ with its corresponding eigenvector $\widehat{\xi}$, (33) and Weyl's theorem yield that

$$\frac{\widehat{\lambda}}{\lambda} = \frac{FF'}{T_h} + \frac{N}{T_h\lambda} + O_{\mathrm{P}}\left(\frac{1}{\sqrt{\lambda}}\right) = 1 + \delta + o_{\mathrm{P}}(1), \tag{35}$$

and the sin-theta theorem implies that

$$\left\|\mathbb{P}_{F'} - \mathbb{P}_{\widehat{\xi}}\right\| = \left\|F'(FF')^{-1}F - \widehat{\xi}\widehat{\xi}'\right\| = o_{\mathrm{P}}(1). \tag{36}$$

Furthermore, (36) implies that $(FF)^{-1}(F\widehat{\xi})^2 = \widehat{\xi}'F'(FF)^{-1}F\widehat{\xi} = 1 + o_{\mathrm{P}}(1)$. Together with $|T_h^{-1}FF' - 1| \lesssim T^{-1/2}$, and the fact that the sign of $\widehat{\xi}$ plays no role in the estimator $\widehat{y}_{T+h}$, we can choose $\widehat{\xi}$ such that

$$\frac{F\widehat{\xi}}{\sqrt{T_h}} - 1 = o_{\mathrm{P}}(1). \tag{37}$$

Therefore, we have

$$\widehat{y}_{t+h} = \widehat{\alpha}\widehat{\varsigma}'x_T = \frac{Y\widehat{\xi}\widehat{\varsigma}'x_T}{\sqrt{T_h\widehat{\lambda}}} = \alpha\frac{F\widehat{\xi}\widehat{\varsigma}'x_T}{\sqrt{T_h\widehat{\lambda}}} = \alpha\frac{\widehat{\varsigma}'\beta f_T + \widehat{\varsigma}'u_T}{\sqrt{\widehat{\lambda}}}(1 + o_{\mathrm{P}}(1)). \tag{38}$$

Using (32), we have

$$\frac{\widehat{\varsigma}'\beta}{\sqrt{\widehat{\lambda}}} = \frac{\sqrt{\lambda}\widehat{\varsigma}_1}{\sqrt{\widehat{\lambda}}} = \frac{\lambda}{\widehat{\lambda}}\frac{(F + \lambda^{-1/2}\epsilon_1 A_1)\widehat{\xi}}{\sqrt{T_h}} = \frac{\lambda}{\widehat{\lambda}}\left(\frac{F\widehat{\xi}}{\sqrt{T_h}} + \frac{\epsilon_1 A_1\widehat{\xi}}{\sqrt{T_h\lambda}}\right).$$

Using (35), (37), $\|A_1\| \leq 1$, and $\|\epsilon_1\| \lesssim_{\mathrm{P}} \sqrt{T}$, it follows that

$$\frac{\widehat{\varsigma}'\beta}{\sqrt{\widehat{\lambda}}} \xrightarrow{\mathrm{P}} \frac{1}{1+\delta}. \tag{39}$$

In addition, as $\mathrm{Cov}(u_s, u_t) = 0$ for $s \neq t$, u_T is independent of $\widehat{\varsigma}$ and thus $\widehat{\varsigma}'u_T = O_{\mathrm{P}}(1)$. Combined with (38) and (39), we have $\widehat{y}_{T+h} \xrightarrow{\mathrm{P}} \frac{\alpha f_T}{1+\delta} = (1+\delta)^{-1}\mathrm{E}_T(y_{T+h})$. □

S4.7 Proof of Propositions S2.3 and S2.4

Proof. In the case $d = K = 1$ and $z_t = 0$, the PLS estimate of the factor is $\widehat{F} = FX'X$. With (33) and $T_h^{-1}FF' - 1 = o_{\mathrm{P}}(T^{-1/2})$, we have

$$\left\|T_h^{-1}\lambda^{-1}\widehat{F} - F\left(\mathbb{I}_{T_h} + \delta A_1'A_1\right)\right\| = o_{\mathrm{P}}\left(T^{1/2}\right). \tag{40}$$

Let $\eta = F\left(\mathbb{I}_{T_h} + \delta A_1' A_1\right)$, and $\widehat{\xi}_1 = \widehat{F}/\left\|\widehat{F}\right\|$, $\widehat{\xi}_2 = \eta/\|\eta\|$, with $\|\eta\| \asymp T^{1/2}$ and $\left\|T_h^{-1}\lambda^{-1}\widehat{F}\right\| - \|\eta\| = o_{\mathrm{P}}(T^{1/2})$ implied by (40), we have $\left\|\widehat{\xi}_1 - \widehat{\xi}_2\right\| \xrightarrow{\mathrm{P}} 0$ and thus

$$\left\|\mathbb{P}_{\widehat{F}'} - \mathbb{P}_{\eta'}\right\| = \left\|\widehat{\xi}_1'\widehat{\xi}_1 - \widehat{\xi}_2'\widehat{\xi}_2\right\| \leq 2\left\|\widehat{\xi}_1 - \widehat{\xi}_2\right\| \xrightarrow{\mathrm{P}} 0.$$

This completes the proof of Proposition S2.3. In the specical case $A_1'A_1 = \mathbb{I}_{T_h}$, as in Section S2.2, we can write

$$\widehat{y}_{T+h} = \|YX'X\|^{-2} YX'XY'YX'x_T = \alpha\|FX'X\|^{-2} FX'XF'FX'x_T. \tag{41}$$

We now analyze $\|FX'X\|$, $FX'XF'$, and $FX'x_T$, respectively. Recall that from (33), we have $\left\|\frac{X'X}{T_h\lambda} - \frac{F'F}{T_h} - \delta\mathbb{I}_{T_h}\right\| = o_{\mathrm{P}}(1)$. Along with $|T_h^{-1}FF' - 1| \lesssim_{\mathrm{P}} T^{-1/2}$, we have

$$\frac{1}{T_h^{3/2}\lambda}\|FX'X\| = \frac{1}{\sqrt{T_h}}\left\|F\left(\frac{F'F}{T_h} + \delta\mathbb{I}_{T_h}\right)\right\| + o_{\mathrm{P}}(1) = 1 + \delta + o_{\mathrm{P}}(1). \tag{42}$$

For the same reason, by direct calculation we have

$$\frac{1}{T_h^2\lambda}FX'XF' = \frac{1}{T_h}F\left(\frac{F'F}{T_h} + \delta\mathbb{I}_{T_h}\right)F' + o_{\mathrm{P}}(1) \xrightarrow{p} 1 + \delta. \tag{43}$$

Next, write X in the form of (31) as in the proof of Proposition S2.1. Then, using $\|\epsilon_1\| \lesssim_{\mathrm{P}} \sqrt{T}$, we have

$$\frac{1}{T_h\lambda}FX'\beta = \frac{FF'}{T_h} + \frac{FA_1'\epsilon_1'}{T\sqrt{\lambda}} \xrightarrow{\mathrm{P}} 1. \tag{44}$$

In addition, as u_T is independent of f_t and x_t for $t < T$, and (43), we have

$$\frac{1}{T_h\lambda}\|FX'u_T\| \lesssim_{\mathrm{P}} \frac{1}{T_h\lambda}\|FX'\| \xrightarrow{\mathrm{P}} 0. \tag{45}$$

In light of (42), (43), (44), (45) and (41), we have concluded the proof. □

S4.8 Proof of Proposition S2.5

Proof. The explicit form of the PLS estimator in this case is

$$\widehat{y}_{T+h}^{PLS} = \|YX'X\|^{-2} YX'XY'YX'x_T = \|ZU'U\|^{-2} ZU'UZ'ZU'u_T.$$

Recall that $U=(u_1,\cdots,u_{T-h})$, u_T is independent of U and Z. Therefore,

$$\mathrm{Var}(\widehat{y}_{T+h}^{PLS}) = \|ZU'U\|^{-4}\|ZU'\|^6.$$

As z_i and u_i are generated from independent standard normal distribution, we have $\|ZU'\| \asymp_{\mathrm{P}} T^{1/2}N^{1/2}$ and $\|U\| \asymp_{\mathrm{P}} N^{1/2}+T^{1/2}$. Thus, $\mathrm{Var}(\widehat{y}_{T+h}^{PLS}) \gtrsim_{\mathrm{P}} N^3T/(N^4+T^4)$. On the other hand, the PCA estimator is $\widehat{y}_{T+h}^{PCA} = \|U\|^{-1}Z\widehat{\xi}\widehat{\varsigma}'u_T$, where $\widehat{\varsigma}$ and $\widehat{\xi}$ are the first left and right singular vectors of U. Note that Z is independent of $\widehat{\xi}$ and u_T is independent of $\widehat{\varsigma}$, we have $\left\|Z\widehat{\xi}\right\| \lesssim_{\mathrm{P}} 1$ and $\|\widehat{\varsigma}'u_T\| \lesssim_{\mathrm{P}} 1$. Along with the fact that $\|U\| \gtrsim_{\mathrm{P}} N^{1/2}+T^{1/2}$, we have $\widehat{y}_{T+h}^{PCA} \lesssim_{\mathrm{P}} 1/(N^{1/2}+T^{1/2})$. □

S4.9 Proof of Proposition S2.6

Proof. The estimated factor $\widehat{F}$ is the first eigenvector of $X'X = U'U$. By Lemma A.1 in Wang and Fan (2017), we have $\|N^{-1}U'U - \mathrm{diag}(1,\ldots,1,1+\epsilon)\| \lesssim_{\mathrm{P}} \sqrt{T/N}$. Note that the first eigenvector of $\mathrm{diag}(1,\ldots,1,1+\epsilon)$ is $(0,0,\ldots,1)$, sin-theta theorem implies that $|\widehat{f}_T|/\left\|\widehat{F}\right\| \xrightarrow{\mathrm{P}} 1$. As $\left\|\underline{\widehat{F}}\right\|^2 + \widehat{f}_T^2 = \left\|\widehat{F}\right\|^2$, we have $\left\|\underline{\widehat{F}}\right\|/\widehat{f}_T \xrightarrow{\mathrm{P}} 0$. As $\overline{Z}$ is independent of U, conditioning on U, the estimated coefficient $\widehat{\alpha} = \overline{Z}\underline{\widehat{F}}'\left(\underline{\widehat{F}}\underline{\widehat{F}}'\right)^{-1}$ follows a normal distribution with mean 0 and variance $\left\|\underline{\widehat{F}}\right\|^{-2}$. Consequently,

$$\mathrm{Var}(\widehat{f}_{T+h}^{SW}|U) = \mathrm{Var}(\widehat{\alpha}\widehat{f}_T|U) = \left(\widehat{f}_T/\left\|\underline{\widehat{F}}\right\|\right)^2 \xrightarrow{\mathrm{P}} \infty,$$

which in turn implies that $\mathrm{Var}(\widehat{f}_{T+h}^{SW}) \to \infty$. On the other hand, in our PCA algorithm, let $\widehat{\varsigma}$ and $\widehat{\xi}$ denote the first left and right singular vectors of $\underline{X} = \underline{U}$, then $\widehat{y}_{T+h}^{PCA} = \|\underline{U}\|^{-1}\overline{Z}\widehat{\xi}\widehat{\varsigma}'u_T$. Note that $\overline{Z}$ is independent of $\widehat{\xi}$ and u_T is independent of $\widehat{\varsigma}$, we have $\left\|\overline{Z}\widehat{\xi}\right\| \lesssim_{\mathrm{P}} 1$ and $\|\widehat{\varsigma}'u_T\| \lesssim_{\mathrm{P}} 1$. Along with the fact that $\|\underline{U}\| \gtrsim_{\mathrm{P}} N^{1/2}+T^{1/2}$, we have $\widehat{y}_{T+h}^{PCA} \xrightarrow{\mathrm{P}} 0$. □

S4.10 Technical Lemmas and Their Proofs

Without loss of generality, we assume that $\Sigma_f = \mathbb{I}_K$ in the following lemmas. Also, except for Lemma S4.3, we assume that $\widehat{K} = \tilde{K}$ and $\widehat{I}_k = I_k$ for $k \le \tilde{K}$, which hold with probability

approaching one as we will show in Lemma S4.3.

Lemma S4.1. *The singular vectors $\widehat{\xi}_{(k)}$s in Algorithm 1 satisfy $\widehat{\xi}'_{(j)}\widehat{\xi}_{(k)} = \delta_{jk}$ for $j, k \leq \widehat{K}$.*

Proof. If $j = k$, this result holds from the definition of $\widehat{\xi}_{(k)}$. If $j < k$, recall that $\widetilde{X}_{(k)}$ is defined in (3) and $\widehat{\xi}_{(k)}$ is the first right singular vector of $\widetilde{X}_{(k)}$, we have

$$\widetilde{X}_{(k)} = X_{[I_k]} \prod_{i<k} \left(\mathbb{I}_T - \widehat{\xi}_{(i)}\widehat{\xi}'_{(i)} \right) \quad \text{and} \quad \widehat{\xi}_{(k)} = \arg\max_{v \in \mathbb{R}^T} \frac{\left\| \widetilde{X}_{(k)} v \right\|}{\|v\|}.$$

If $\widehat{\xi}'_{(k)}\widehat{\xi}_{(j)} = c_0 \neq 0$ for some $j < k$, then

$$\left\| \widetilde{X}_{(k)}(\widehat{\xi}_{(k)} - c_0\widehat{\xi}_{(j)}) \right\| = \left\| \widetilde{X}_{(k)}\widehat{\xi}_{(k)} - c_0\widetilde{X}_{(k)}\widehat{\xi}_{(j)} \right\| = \left\| \widetilde{X}_{(k)}\widehat{\xi}_{(k)} \right\|, \tag{46}$$

since the definition of $\widetilde{X}_{(k)}$ implies that $\widetilde{X}_{(k)}\widehat{\xi}_{(j)} = 0$ for $j < k$. On the other hand, since $\widehat{\xi}'_{(k)}\widehat{\xi}_{(j)} = c_0 \neq 0$, we have $(\widehat{\xi}_{(k)} - c_0\widehat{\xi}_{(j)})'\widehat{\xi}_{(j)} = 0$, and consequently,

$$\left\| \widehat{\xi}_{(k)} \right\|^2 = \left\| \widehat{\xi}_{(k)} - c_0\widehat{\xi}_{(j)} \right\|^2 + \left\| c_0\widehat{\xi}_{(j)} \right\|^2 > \left\| \widehat{\xi}_{(k)} - c_0\widehat{\xi}_{(j)} \right\|^2. \tag{47}$$

Apparently, if $\left\| \widetilde{X}_{(k)} \right\| = 0$, the SPCA procedure will terminate so we have $\left\| \widetilde{X}_{(k)} \right\| > 0$ for $k \leq \widehat{K}$. Together with (46) and (47), we have

$$\left\| \widetilde{X}_{(k)} \right\| = \frac{\left\| \widetilde{X}_{(k)}\widehat{\xi}_{(k)} \right\|}{\left\| \widehat{\xi}_{(k)} \right\|} \leq \frac{\left\| \widetilde{X}_{(k)}(\widehat{\xi}_{(k)} - c_0\widehat{\xi}_{(j)}) \right\|}{\left\| \widehat{\xi}_{(k)} - c_0\widehat{\xi}_{(j)} \right\|},$$

which contradicts with the definition of $\widehat{\xi}_{(k)}$. Therefore, $\widehat{\xi}'_{(k)}\widehat{\xi}_{(j)} = 0$ for any $j < k \leq \widehat{K}$. □

Lemma S4.2. *Under assumptions of Theorem 3.1, $b_{(k)}$, $\beta_{(k)}$ and $\tilde{K}$ in Section 3.1 satisfy*

(i) $b'_{(j)}b_{(k)} = \delta_{jk}$ *for* $j \leq k \leq \tilde{K}$.

(ii) $\lambda_{(k)} = \left\| \beta_{(k)} \right\|^2 \asymp qN^{1-\nu}$.

(iii) $\tilde{K} \leq K$.

(iv) $\tilde{K} = K$, *if we further have* $\lambda_K(\alpha'\alpha) \gtrsim 1$.

Proof. (i) Recall that $b_{(k)}$ is the first right singular vector of $\beta_{(k)}$ and $\beta_{(k)} = \beta_{[I_k]} \prod_{j<k} \mathbb{M}_{b_{(j)}}$. Using the same argument as in the proof of Lemma S4.1, we have $b'_{(j)} b_{(k)} = \delta_{jk}$ for $j, k \leq \tilde{K}$.

(ii) The selection rule at kth step implies that

$$|I_k|^{-1} \sum_{i \in I_k} \left\| \beta_{[i]} \prod_{j<k} \mathbb{M}_{b_{(j)}} \alpha' \right\|_{\text{MAX}}^2 \geq N_0^{-1} \sum_{i \in I_0} \left\| \beta_{[i]} \prod_{j<k} \mathbb{M}_{b_{(j)}} \alpha' \right\|_{\text{MAX}}^2 . \tag{48}$$

For any matrix $A \in \mathbb{R}^{N \times D}$ and set $I \subset [N]$, we have

$$\sum_{i \in I} \left\| A_{[i]} \right\|_{\text{MAX}}^2 \leq \left\| A_{[I]} \right\|_{\text{F}}^2 \leq D \sum_{i \in I} \left\| A_{[i]} \right\|_{\text{MAX}}^2 ,$$

and $\left\| A_{[I]} \right\|^2 \leq \left\| A_{[I]} \right\|_{\text{F}}^2 \leq D \left\| A_{[I]} \right\|^2$. We thereby have

$$\left\| A_{[I]} \right\|^2 \asymp \sum_{i \in I} \left\| A_{[i]} \right\|_{\text{MAX}}^2 . \tag{49}$$

Using this result, (48) becomes

$$|I_k|^{-1} \left\| \beta_{[I_k]} \prod_{j<k} \mathbb{M}_{b_{(j)}} \alpha' \right\|^2 \gtrsim N_0^{-1} \left\| \beta_{[I_0]} \prod_{j<k} \mathbb{M}_{b_{(j)}} \alpha' \right\|^2 .$$

Then, we have

$$\begin{aligned} &\frac{\left\| \beta_{(k)} \right\|}{\sqrt{|I_k|}} \left\| \prod_{j<k} \mathbb{M}_{b_{(j)}} \alpha' \right\| \geq \frac{1}{\sqrt{|I_k|}} \left\| \beta_{[I_k]} \prod_{j<k} \mathbb{M}_{b_{(j)}} \alpha' \right\| \gtrsim \frac{1}{\sqrt{N_0}} \left\| \beta_{[I_0]} \prod_{j<k} \mathbb{M}_{b_{(j)}} \alpha' \right\| \\ \geq &\frac{\sigma_K \left(\beta_{[I_0]} \right)}{\sqrt{N_0}} \left\| \prod_{j<k} \mathbb{M}_{b_{(j)}} \alpha' \right\| , \end{aligned} \tag{50}$$

where we use $\beta_{[I_k]} \prod_{j<k} \mathbb{M}_{b_{(j)}} \alpha' = \beta_{[I_k]} (\prod_{j<k} \mathbb{M}_{b_{(j)}})^2 \alpha' = \beta_{(k)} \prod_{j<k} \mathbb{M}_{b_{(j)}} \alpha'$ in the first inequality. With $\sigma_K(\beta_{[I_0]}) \gtrsim \sqrt{N_0 / N^\nu}$ from Assumption 2, (50) leads to $\left\| \beta_{(k)} \right\| \gtrsim \sqrt{|I_k| / N^\nu}$. In addition, Assumption 2 leads to $\left\| \beta_{(k)} \right\| \lesssim \sqrt{|I_k| / N^\nu}$. Therefore, we have $\left\| \beta_{(k)} \right\|^2 \asymp |I_k| / N^\nu \asymp qN^{1-\nu}$ as $|I_k| = qN$.

(iii) From (i), we have shown that $b_{(k)}$'s are pairwise orthogonal for $k \leq \tilde{K}$. It is impossible to have more than K pairwise orthogonal K dimensional vectors. Thus, $\tilde{K} \leq K$ holds directly.

(iv) Recall that $\tilde{K}$ is defined in Section 3.1. Since the SPCA procedure stops at $\tilde{K}+1$, we have at most $qN - 1$ rows of β satisfying $\left\| \beta_{[i]} \prod_{j \leq \tilde{K}} \mathbb{M}_{b_{(j)}} \alpha' \right\|_{\text{MAX}} \geq c$, which implies

$$\left\| \beta_{[I_0]} \prod_{j \leq \tilde{K}} \mathbb{M}_{b_{(j)}} \alpha' \right\|^2 \lesssim qN / N^\nu + (N_0 - qN) c^2 = o(N_0 / N^\nu),$$

where we use (49) and the assumptions $cN^{\nu/2} \to 0$, $qN/N_0 \to 0$, and a similar argument for the proof of (9). With $\sigma_K(\beta_{[I_0]}) \gtrsim \sqrt{N_0/N^\nu}$ from Assumption 2, we have

$$\left\| \alpha \prod_{j \le \tilde{K}} \mathbb{M}_{b_{(j)}} \right\| \le \sigma_K(\beta_{[I_0]})^{-1} \left\| \beta_{[I_0]} \prod_{j \le \tilde{K}} \mathbb{M}_{b_{(j)}} \alpha' \right\| = o(1). \tag{51}$$

If $\tilde{K} \le K-1$, using (i), we have $\alpha \prod_{j \le \tilde{K}} \mathbb{M}_{b_{(j)}} = \alpha - \alpha \sum_{j \le \tilde{K}} b_{(j)} b'_{(j)}$, so that

$$\sigma_K(\alpha) \le \sigma_1 \left(\alpha \prod_{j \le \tilde{K}} \mathbb{M}_{b_{(j)}} \right) + \sigma_K \left(\alpha \sum_{j \le \tilde{K}} b_{(j)} b'_{(j)} \right). \tag{52}$$

Since

$$\text{Rank} \left(\alpha \sum_{j \le \tilde{K}} b_{(j)} b'_{(j)} \right) \le \tilde{K} \le K-1, \tag{53}$$

we have $\sigma_K \left(\alpha \sum_{j \le \tilde{K}} b_{(j)} b'_{(j)} \right) = 0$. Therefore, by (51) and (52), we further have $\sigma_K(\alpha) \lesssim \sigma_1 \left(\alpha \prod_{j \le \tilde{K}} \mathbb{M}_{b_{(j)}} \right) \to 0$. This contradicts with the assumption that $\lambda_K(\alpha'\alpha) \gtrsim 1$. Therefore, we have established that $\tilde{K} \ge K$. Together with (iii), we have $\tilde{K} = K$. □

Lemma S4.3. *Under assumptions of Theorem 3.1, for $k \le \tilde{K}$, I_k, $\tilde{K}$ and $\beta_{(k)}$ satisfy*

(i) $\mathrm{P}(\widehat{I}_k = I_k) \to 1$.

(ii) $\left\| \widetilde{X}_{(k)} - \beta_{(k)} F \right\| \lesssim_{\mathrm{P}} q^{1/2} N^{1/2} + T^{1/2}$.

(iii) $|\widehat{\lambda}_{(k)}^{1/2} / \lambda_{(k)}^{1/2} - 1| \lesssim_{\mathrm{P}} q^{-1/2} N^{-1/2+\nu/2} + T^{-1/2} N^{\nu/2}$, *and* $\widehat{\lambda}_{(k)} \asymp_{\mathrm{P}} \lambda_{(k)} \asymp qN^{1-\nu}$.

(iv) $\left\| T_h^{-1/2} F \widehat{\xi}_{(k)} - b_{(k)} \right\| \asymp \left\| \mathbb{P}_{\widehat{F}'_{(k)}} - T_h^{-1} F' \mathbb{P}_{b_{(k)}} F \right\| \lesssim_{\mathrm{P}} q^{-1/2} N^{-1/2+\nu/2} + T^{-1/2} N^{\nu/2}$.

(v) $\mathrm{P}(\widehat{K} = \tilde{K}) \to 1$.

For $k \le \tilde{K}+1$, we have

(vi) $\left\| T_h^{-1} X Y'_{(k)} - \beta \prod_{j=1}^{k-1} \mathbb{M}_{b_{(j)}} \alpha' \right\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} (\log NT)^{1/2} \left(q^{-1/2} N^{-1/2+\nu/2} + T^{-1/2} N^{\nu/2} \right)$.

Proof. We prove (i)-(iv) by induction. First, we show that (i)-(iv) hold when $k = 1$:

(i) Recall that $\widehat{I}_1$ is selected based on $T_h^{-1}XY'$ and I_1 is selected based on $\beta\alpha'$. With simple algebra, we have

$$T_h^{-1}XY' - \beta\alpha' = \beta\left(T_h^{-1}FF' - \mathbb{I}_K\right)\alpha' + T_h^{-1}UF'\alpha' + T_h^{-1}\beta FZ' + T_h^{-1}UZ'.$$

With Assumptions 1, 2 and 4, we have

$$\begin{aligned}\left\|T_h^{-1}XY' - \beta\alpha'\right\|_{\max} \lesssim{}& \|\beta\|_{\max}\left\|T_h^{-1}FF' - \mathbb{I}_K\right\|\|\alpha\| + T_h^{-1}\|UF'\|_{\max}\|\alpha\| \\ &+ T_h^{-1}\|\beta\|_{\max}\|FZ'\| + T_h^{-1}\|UZ'\|_{\max} \lesssim_{\mathrm{P}} (\log N)^{1/2}T^{-1/2}.\end{aligned}$$

From Assumption 5, we have $c_{qN}^{(1)} - c_{qN+1}^{(1)} \gtrsim c_{qN}^{(1)}$ and the definition of $\tilde{K}$ implies that $c_{qN}^{(k)} \geq c$ for $k \leq \tilde{K}$. Thus, we have $c_{qN}^{(1)} - c_{qN+1}^{(1)} \gtrsim c$. Define the events

$$\begin{aligned}A_1 :={}& \left\{ \left\|T_h^{-1}X_{[i]}Y'\right\|_{\max} > (c_{qN}^{(1)} + c_{qN+1}^{(1)})/2 \text{ for all } i \in I_1\right\}, \\ A_2 :={}& \left\{ \left\|T_h^{-1}X_{[i]}Y'\right\|_{\max} < (c_{qN}^{(1)} + c_{qN+1}^{(1)})/2 \text{ for all } i \in I_1^c\right\}, \\ A_3 :={}& \left\{ \left\|T_h^{-1}X_{[i]}Y' - \beta_{[i]}\alpha'\right\|_{\max} \geq (c_{qN}^{(1)} - c_{qN+1}^{(1)})/2 \text{ for some } i \in [N]\right\}.\end{aligned} \tag{54}$$

It is easy to observe that $\{\widehat{I}_1 = I_1\} \supset A_1 \cap A_2$. In addition, from the definition of I_1, we have $\left\|\beta_{[i]}\alpha'\right\|_{\max} \geq c_{qN}^{(1)}$ for all $i \in I_1$ and $\left\|\beta_{[i]}\alpha'\right\|_{\max} \leq c_{qN+1}^{(1)}$ for all $i \in I_1^c$. Therefore, if A_1^c occurs, we have $\left\|T_h^{-1}X_{[i]}Y' - \beta_{[i]}\alpha'\right\|_{\max} \geq (c_{qN}^{(1)} - c_{qN+1}^{(1)})/2$, for some $i \in I_1$, which implies $A_1^c \subset A_3$. Similarly, we have $A_2^c \subset A_3$. Using $\{\widehat{I}_1 = I_1\} \supset A_1 \cap A_2$ and $A_1^c \cup A_2^c \subset A_3$, we have

$$\mathrm{P}(\widehat{I}_1 = I_1) \geq \mathrm{P}(A_1 \cap A_2) = 1 - \mathrm{P}(A_1^c \cup A_2^c) \geq 1 - \mathrm{P}(A_3). \tag{55}$$

Using $c^{-1}(\log N)^{1/2}T^{-1/2} \to 0$ and $c_{qN}^{(1)} - c_{qN+1}^{(1)} \gtrsim c$, we have $\mathrm{P}(A_3) \to 0$ and consequently, $\mathrm{P}(\widehat{I}_1 = I_1) \to 1$.

(ii) Since $\widehat{I}_1 = I_1$ with probability approaching one, we impose $\widehat{I}_1 = I_1$ below. Then, we have $\widetilde{X}_{(1)} = X_{[I_1]}$ by (3) and Assumption 3 gives $\left\|\widetilde{X}_{(1)} - \beta_{(1)}F\right\| = \left\|U_{[I_1]}\right\| \lesssim_{\mathrm{P}} q^{1/2}N^{1/2} + T^{1/2}$.

(iii) From Lemma S4.12, we have $\sigma_j(\beta_{(1)}F)/\sigma_j(\beta_{(1)}) = T_h^{1/2} + O_{\mathrm{P}}(1)$, which leads to

$$\left|\left\|\beta_{(1)}F\right\| - T_h^{1/2}\lambda_{(1)}^{1/2}\right| = \left|\sigma_1(\beta_{(1)}F) - T_h^{1/2}\sigma_1(\beta_{(1)})\right| \lesssim_{\mathrm{P}} q^{1/2}N^{1/2-\nu/2}, \tag{56}$$

where we use $\lambda_{(1)}^{1/2} = \left\|\beta_{(1)}\right\| \asymp q^{1/2}N^{1/2-\nu/2}$ from Lemma S4.2 in the last step. In addition, the result in (ii) implies that

$$\left|\left\|\widetilde{X}_{(1)}\right\| - \left\|\beta_{(1)}F\right\|\right| \leq \left\|\widetilde{X}_{(1)} - \beta_{(1)}F\right\| \lesssim_{\mathrm{P}} q^{1/2}N^{1/2} + T^{1/2}. \tag{57}$$

Using (56), (57) and $\lambda_{(1)} \asymp qN^{1-\nu}$, we have

$$\left|\frac{\widehat{\lambda}_{(1)}^{1/2}}{\lambda_{(1)}^{1/2}} - 1\right| = \left|\frac{\left\|\widetilde{X}_{(1)}\right\|}{T_h^{1/2}\lambda_{(1)}^{1/2}} - 1\right| \leq \frac{\left|\left\|\widetilde{X}_{(1)}\right\| - \left\|\beta_{(1)}F\right\|\right|}{T_h^{1/2}\lambda_{(1)}^{1/2}} + \frac{\left|\left\|\beta_{(1)}F\right\| - T_h^{1/2}\lambda_{(1)}^{1/2}\right|}{T_h^{1/2}\lambda_{(1)}^{1/2}}$$
$$\lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2+\nu/2} + T^{-1/2}N^{\nu/2}.$$

and thus $\widehat{\lambda}_{(1)} \asymp_{\mathrm{P}} qN^{1-\nu}$.

(iv) Let $\tilde{\xi}_{(1)} \in \mathbb{R}^{T_h \times 1}$ denote the first right singular vector of $\beta_{(1)}F$. Lemma S4.12 yields

$$\left\|\mathbb{P}_{\tilde{\xi}_{(1)}} - T_h^{-1}F'\mathbb{P}_{b_{(1)}}F\right\| \lesssim_{\mathrm{P}} T^{-1/2} \tag{58}$$

and $\sigma_j(\beta_{(1)}F)/\sigma_j(\beta_{(1)}) = T_h^{1/2} + O_{\mathrm{P}}(1)$. The latter further leads to

$$\sigma_1(\beta_{(1)}F) - \sigma_2(\beta_{(1)}F) = T_h^{1/2}(\sigma_1(\beta_{(1)}) - \sigma_2(\beta_{(1)})) + O_{\mathrm{P}}(\sigma_1(\beta_{(1)})) \asymp_p T^{1/2}\sigma_1(\beta_{(1)}), \tag{59}$$

where we use the assumption that $\sigma_2(\beta_{(1)}) \leq (1+\delta)^{-1}\sigma_1(\beta_{(1)})$ in the last equation.

Using $\left\|\widetilde{X}_{(1)} - \beta_{(1)}F\right\| \lesssim_{\mathrm{P}} q^{1/2}N^{1/2} + T^{1/2}$ as proved in (ii), (59), Lemma S4.2 and Wedin (1972)'s sin-theta theorem for singular vectors, we have

$$\left\|\mathbb{P}_{\widehat{F}'_{(1)}} - \mathbb{P}_{\tilde{\xi}_{(1)}}\right\| \lesssim_{\mathrm{P}} \frac{q^{1/2}N^{1/2} + T^{1/2}}{\sigma_1(\beta_{(1)}F) - \sigma_2(\beta_{(1)}F)} \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2+\nu/2} + T^{-1/2}N^{\nu/2}. \tag{60}$$

In light of (58) and (60),the first equation in (iv) holds for $k = 1$. As $\mathbb{P}_{\widehat{F}'_{(k)}} = \widehat{\xi}_{(k)}\widehat{\xi}'_{(k)}$, left and right multiplying this equation by $\widehat{\xi}'_{(1)}$ and $\widehat{\xi}_{(1)}$, we have

$$|1 - T_h^{-1}(b'_{(1)}F\widehat{\xi}_{(1)})^2| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2+\nu/2} + T^{-1/2}N^{\nu/2},$$

which leads to $|1 - T_h^{-1/2}b'_{(1)}F\widehat{\xi}_{(1)}| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2+\nu/2} + T^{-1/2}N^{\nu/2}$. Left-multiplying it by $b_{(1)}$ gives the second equation in (iv).

So far, we have proved that (i)-(iv) hold for $k = 1$. Now, assuming that (i)-(iv) hold for $j \leq k-1$, we will show that (i)-(iv) continue to hold for $j = k$.

(i) Again, we show the difference between the sample covariances and their population counterparts introduced in the SPCA procedure is tiny. At the kth step, the difference can be written as

$$\begin{aligned}
&\left\| \beta \prod_{j=1}^{k-1} \mathbb{M}_{b_{(j)}} \alpha' - T_h^{-1}(\beta F + U) \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{F}'_{(j)}} (\alpha F + Z)' \right\|_{\text{MAX}} \\
\leq &\left\| \beta \prod_{j=1}^{k-1} \mathbb{M}_{b_{(j)}} \alpha' - T_h^{-1} \beta F \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{F}'_{(j)}} F' \alpha' \right\|_{\text{MAX}} + T_h^{-1} \left\| \beta F \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{F}'_{(j)}} Z' \right\|_{\text{MAX}} \\
&+ T_h^{-1} \left\| U \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{F}'_{(j)}} F' \alpha' \right\|_{\text{MAX}} + T_h^{-1} \left\| U \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{F}'_{(j)}} Z' \right\|_{\text{MAX}}.
\end{aligned} \tag{61}$$

Since (iv) holds for $j \leq k-1$, we have

$$\begin{aligned}
\left\| \sum_{j=1}^{k-1} \mathbb{P}_{\widehat{F}'_{(j)}} - T_h^{-1} F' \sum_{j=1}^{k-1} \mathbb{P}_{b_{(j)}} F \right\| &= \left\| \sum_{j=1}^{k-1} \left(\mathbb{P}_{\widehat{F}'_{(j)}} - T_h^{-1} F' \mathbb{P}_{b_{(j)}} F \right) \right\| \\
&\lesssim_{\text{P}} q^{-1/2} N^{-1/2+\nu/2} + T^{-1/2} N^{\nu/2}.
\end{aligned} \tag{62}$$

Using Lemma S4.1 and Lemma S4.2(i), we have

$$\prod_{j=1}^{k-1} \mathbb{M}_{b_{(j)}} = \mathbb{I}_K - \sum_{j=1}^{k-1} \mathbb{P}_{b_{(j)}}, \quad \text{and} \quad \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{F}'_{(j)}} = \mathbb{I}_{T_h} - \sum_{j=1}^{k-1} \mathbb{P}_{\widehat{F}'_{(j)}}.$$

Using the above equations, (62), and $\left\| T_h^{-1} F F' - \mathbb{I}_K \right\| \lesssim_{\text{P}} T^{-1/2}$, we have

$$\begin{aligned}
T_h^{-1/2} \left\| F \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{F}'_{(j)}} - \prod_{j=1}^{k-1} \mathbb{M}_{b_{(j)}} F \right\| &= T_h^{-1/2} \left\| F \sum_{j=1}^{k-1} \mathbb{P}_{\widehat{F}'_{(j)}} - \sum_{j=1}^{k-1} \mathbb{P}_{b_{(j)}} F \right\| \\
&\lesssim_{\text{P}} q^{-1/2} N^{-1/2+\nu/2} + T^{-1/2} N^{\nu/2}.
\end{aligned} \tag{63}$$

Similarly, right multiplying F' to the term inside the $\|\cdot\|$ of (63), we have

$$\left\| T_h^{-1} F \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{F}'_{(j)}} F' - \prod_{j=1}^{k-1} \mathbb{M}_{b_{(j)}} \right\| \lesssim_{\text{P}} q^{-1/2} N^{-1/2+\nu/2} + T^{-1/2} N^{\nu/2}. \tag{64}$$

Next, we analyze the four terms in (61) one by one. For the first term, using (64) and Assumption 2, we have

$$\left\| \beta \prod_{j=1}^{k-1} \mathbb{M}_{b_{(j)}} \alpha' - T_h^{-1} \beta F \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{F}'_{(j)}} F' \alpha' \right\|_{\text{MAX}}$$

$$\lesssim \|\beta\|_{\text{MAX}} \left\| \prod_{j=1}^{k-1} \mathbb{M}_{b_{(j)}} - T_h^{-1} F \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{F}'_{(j)}} F' \right\| \|\alpha\| \lesssim_p q^{-1/2} N^{-1/2} + T^{-1/2}.$$

For the second term, using (63), Assumption 2, we have

$$T_h^{-1} \left\| \beta F \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{F}'_{(j)}} Z' \right\|_{\text{MAX}} \lesssim T_h^{-1} \|\beta\|_{\text{MAX}} \left\| \prod_{j=1}^{k-1} \mathbb{M}_{b_{(j)}} \right\| \|FZ'\|$$
$$+ T_h^{-1} \|\beta\|_{\text{MAX}} \left\| F \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{F}'_{(j)}} - \prod_{j=1}^{k-1} \mathbb{M}_{b_{(j)}} F \right\| \|Z\| \lesssim_p q^{-1/2} N^{-1/2} + T^{-1/2}.$$

For the third term, using (63), we have

$$T_h^{-1} \left\| U \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{F}'_{(j)}} F' \alpha' \right\|_{\text{MAX}} \lesssim T_h^{-1} \|U\|_{\text{MAX}} T_h^{1/2} \left\| F \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{F}'_{(j)}} - \prod_{j=1}^{k-1} \mathbb{M}_{b_{(j)}} F \right\| \|\alpha\|$$
$$+ T_h^{-1} \|UF'\|_{\text{MAX}} \left\| \prod_{j=1}^{k-1} \mathbb{M}_{b_{(j)}} \right\| \|\alpha\| \lesssim_p (\log NT)^{1/2} \left(q^{-1/2} N^{-1/2+\nu/2} + T^{-1/2} N^{\nu/2} \right).$$

For the forth term, using (62), we have

$$T_h^{-1} \left\| U \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{F}'_{(j)}} Z' \right\|_{\text{MAX}} \lesssim T_h^{-1} \|UZ'\|_{\text{MAX}} + T_h^{-2} \|UF'\|_{\text{MAX}} \left\| \sum_{j=1}^{k-1} \mathbb{P}_{b_{(j)}} \right\| \|FZ'\| +$$
$$T_h^{-1/2} \|U\|_{\text{MAX}} \left\| T_h^{-1} F' \sum_{j=1}^{k-1} \mathbb{P}_{b_{(j)}} F - \sum_{j=1}^{k-1} \mathbb{P}_{\widehat{F}'_{(j)}} \right\| \|Z\| \lesssim_p (\log NT)^{1/2} \left(q^{-1/2} N^{-1/2+\nu/2} + T^{-1/2} N^{\nu/2} \right).$$

Hence, we have

$$\left\| T_h^{-1} X \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{F}'_{(j)}} Y' - \beta \prod_{j=1}^{k-1} \mathbb{M}_{b_{(j)}} \alpha' \right\|_{\text{MAX}} \lesssim_{\text{P}} (\log NT)^{1/2} \left(q^{-1/2} N^{-1/2+\nu/2} + T^{-1/2} N^{\nu/2} \right). \quad (65)$$

As in the case of $k = 1$, with the assumption that

$$c^{-1} (\log NT)^{1/2} \left(q^{-1/2} N^{-1/2+\nu/2} + T^{-1/2} N^{\nu/2} \right) \to 0,$$

and Assumption 5, we can reuse the arguments for (54) and (55) in the case of $k = 1$ and obtain $\text{P}(\widehat{I}_k = I_k) \to 1$.

(ii) We impose $\widehat{I}_k = I_k$ below. Then, we have $\widetilde{X}_{(k)} = X_{[I_k]} \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{F}'_{(j)}}$ and thus

$$\widetilde{X}_{(k)} - \beta_{(k)} F = X_{[I_k]} \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{F}'_{(j)}} - \beta_{(k)} F = \beta_{[I_k]} \left(F \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{F}'_{(j)}} - \prod_{j=1}^{k-1} \mathbb{M}_{b_{(j)}} F \right) + U_{[I_k]} \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{F}'_{(j)}}.$$

Hence, using Assumption 2 and (63), we have

$$\left\|\widetilde{X}_{(k)}-\beta_{(k)}F\right\| \leq \left\|\beta_{[I_k]}\right\| \left\|F\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{F}'_{(j)}}-\prod_{j=1}^{k-1}\mathbb{M}_{b_{(j)}}F\right\| + \left\|U_{[I_k]}\right\| \left\|\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{F}'_{(j)}}\right\|$$
$$\lesssim_{\mathrm{P}} q^{1/2}N^{1/2}+T^{1/2}.$$

(iii)(iv) The proofs of (iii) and (iv) are analogous to the case $k=1$.

To sum up, by induction, we have shown that (i)-(iv) hold for $k \leq \tilde{K}$.

(v) Recall that $\tilde{K}$ is determined by $\beta_{[i]}\prod_{j<k}\mathbb{M}_{b_{(j)}}\alpha'$ whereas $\widehat{K}$ is determined by $T_h^{-1}X_{[i]}\prod_{j<k}\mathbb{M}_{\widehat{F}'_{(j)}}Y'$. Since (iv) holds for $j \leq \tilde{K}$ as shown above, using the same proof for (65), we have

$$\left\|T_h^{-1}X\prod_{j=1}^{\tilde{K}}\mathbb{M}_{\widehat{F}'_{(j)}}Y'-\beta\prod_{j=1}^{\tilde{K}}\mathbb{M}_{b_{(j)}}\alpha'\right\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} (\log NT)^{1/2}\left(q^{-1/2}N^{-1/2+\nu/2}+T^{-1/2}N^{\nu/2}\right). \quad (66)$$

The assumption $c_{qN}^{(\tilde{K}+1)} \leq (1+\delta)^{-1}c$ in Assumption 5 implies that $c-c_{qN}^{(\tilde{K}+1)} \asymp c$. Then, we can reuse the arguments for (54) and (55) with events

$$B_1=\left\{\left\|T_h^{-1}X_{[i]}\prod_{j=1}^{\tilde{K}}\mathbb{M}_{\widehat{F}'_{(j)}}Y'\right\|_{\mathrm{MAX}} > (c+c_{qN}^{(\tilde{K}+1)})/2 \text{ for at most } qN-1 \text{ different } i\text{s in } [N]\right\},$$
$$B_2=\left\{\left\|T_h^{-1}X_{[i]}\prod_{j=1}^{\tilde{K}}\mathbb{M}_{\widehat{F}'_{(j)}}Y'-\beta_{[i]}\prod_{j=1}^{\tilde{K}}\mathbb{M}_{b_{(j)}}\alpha'\right\|_{\mathrm{MAX}} \geq (c-c_{qN}^{(\tilde{K}+1)})/2 \text{ for some } i\in[N]\right\},$$

to obtain $\mathrm{P}(\widehat{K}=\tilde{K}) \geq \mathrm{P}(B_1) = 1-\mathrm{P}(B_1^c) \geq 1-\mathrm{P}(B_2) \to 1$.

(vi) This result comes directly from (65) and (66). □

Lemma S4.4. *Under assumptions of Theorem 3.1, for $k \leq \tilde{K}$, we have*

(i) $\left\|U'_{[I_k]}\widehat{\varsigma}_{(k)}\right\| \lesssim_{\mathrm{P}} T^{1/2}+T^{-1/2}q^{1/2}N^{1/2+\nu/2}$.

(ii) $\left\|AU'_{[I_k]}\widehat{\varsigma}_{(k)}\right\| \lesssim_{\mathrm{P}} q^{1/2}N^{1/2+\nu/2}+T^{1/2}N^{\nu/2}$, *for* $A=F,Z,W$.

(iii) $|\widehat{\varsigma}'_{(k)}(u_T)_{[I_k]}| \lesssim_{\mathrm{P}} 1$.

Proof. (i) Using Lemma S4.1, we have

$$T_h^{1/2}\widehat{\lambda}_{(k)}^{1/2}\widehat{\varsigma}_{(k)} = X_{[I_k]}\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{\xi}_{(j)}}\widehat{\xi}_{(k)} = X_{[I_k]}\widehat{\xi}_{(k)} = \beta_{[I_k]}F\widehat{\xi}_{(k)} + U_{[I_k]}\widehat{\xi}_{(k)}. \tag{67}$$

Therefore, along with Assumption 1, Assumption 3 and Assumption 4(ii), we obtain

$$\begin{aligned} T_h^{1/2}\widehat{\lambda}_{(k)}^{1/2}\left\|\widehat{\varsigma}_{(k)}'U_{[I_k]}\right\| &\le \left\|\widehat{\xi}_{(k)}'F'\beta_{[I_k]}'U_{[I_k]}\right\| + \left\|\widehat{\xi}_{(k)}'U_{[I_k]}'U_{[I_k]}\right\| \\ &\le \|F\|\left\|\beta_{[I_k]}'U_{[I_k]}\right\| + \left\|U_{[I_k]}'U_{[I_k]}\right\| \lesssim_{\mathrm{P}} q^{1/2}N^{1/2-\nu/2}T + qN. \end{aligned} \tag{68}$$

Together with $\widehat{\lambda}_{(k)} \asymp_{\mathrm{P}} qN^{1-\nu}$, we have the desired result.

(ii) Similarly, by Assumption 1, Assumption 3 and Assumption 4(i)(ii), we have

$$\begin{aligned} T^{1/2}\widehat{\lambda}_{(k)}^{1/2}\left\|\widehat{\varsigma}_{(k)}'U_{[I_k]}F'\right\| &\le \left\|\widehat{\xi}_{(k)}'F'\beta_{[I_k]}'U_{[I_k]}F'\right\| + \left\|\widehat{\xi}_{(k)}'U_{[I_k]}'U_{[I_k]}F'\right\| \\ \le \|F\|\left\|\beta_{[I_k]}'U_{[I_k]}F'\right\| + \left\|U_{[I_k]}\right\|\left\|U_{[I_k]}F'\right\| &\lesssim_{\mathrm{P}} q^{1/2}N^{1/2}T + qNT^{1/2}. \end{aligned} \tag{69}$$

Together with $\widehat{\lambda}_{(k)} \asymp_{\mathrm{P}} qN^{1-\nu}$, we have the desired result. In addition, replacing F by Z and W, we have the second and third equations in (ii).

(iii) Using Assumption 4(iii) and $\left\|\widehat{\varsigma}_{(k)}\right\| = 1$, we have (iii) directly. □

Lemma S4.5. *Under assumptions of Theorem 3.1, for $k, l \le \tilde{K}$, we have*

(i) $\left\|\frac{\widetilde{U}_{(k)}'\widehat{\varsigma}_{(k)}}{\sqrt{T_h\widehat{\lambda}_{(k)}}}\right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2+\nu/2} + T^{-1}N^{\nu}$, $\left\|\frac{\widetilde{U}_{(k)}}{\sqrt{T_h\widehat{\lambda}_{(k)}}}\right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2+\nu/2} + T^{-1/2}N^{\nu/2}$.

(ii) $\left\|\frac{A\widetilde{U}_{(k)}'\widehat{\varsigma}_{(k)}}{T_h\sqrt{\widehat{\lambda}_{(k)}}}\right\| \lesssim_{\mathrm{P}} q^{-1}N^{-1+\nu} + T^{-1}N^{\nu}$, *for $A = F, Z,$ and W.*

(iii) $|\frac{\widehat{\xi}_{(l)}'U_{[I_k]}'\widehat{\varsigma}_{(k)}}{\sqrt{T_h\widehat{\lambda}_{(k)}}}| \lesssim_{\mathrm{P}} q^{-1}N^{-1+\nu} + T^{-1}N^{\nu}$, $|\frac{\widehat{\xi}_{(l)}'\widetilde{U}_{(k)}'\widehat{\varsigma}_{(k)}}{\sqrt{T_h\widehat{\lambda}_{(k)}}}| \lesssim_{\mathrm{P}} q^{-1}N^{-1+\nu} + T^{-1}N^{\nu}$.

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

Proof. (i) Recall that from the definition of $U_{(k)}$ (below (5)), we have

$$\widetilde{U}_{(k)} = U_{[I_k]} - \sum_{i=1}^{k-1}\frac{X_{[I_k]}\widehat{\xi}_{(i)}}{\sqrt{T_h}}\frac{\widehat{\varsigma}_{(i)}'\widetilde{U}_{(i)}}{\sqrt{\widehat{\lambda}_{(i)}}}. \tag{70}$$

Then, a direct multiplication of $\widehat{\varsigma}_{(k)}'/\sqrt{T_h\widehat{\lambda}_{(k)}}$ from the left side of (70) leads to

$$\frac{\widehat{\varsigma}_{(k)}'\widetilde{U}_{(k)}}{\sqrt{T_h\widehat{\lambda}_{(k)}}} = \frac{\widehat{\varsigma}_{(k)}'U_{[I_k]}}{\sqrt{T_h\widehat{\lambda}_{(k)}}} - \sum_{i=1}^{k-1}\frac{\widehat{\varsigma}_{(k)}'X_{[I_k]}\widehat{\xi}_{(i)}}{\sqrt{T_h\widehat{\lambda}_{(k)}}}\frac{\widehat{\varsigma}_{(i)}'\widetilde{U}_{(i)}}{\sqrt{T_h\widehat{\lambda}_{(i)}}}.$$

Consequently, using $\left\|X_{[I_k]}\right\| \leq \left\|\beta_{[I_k]}\right\| \|F\| + \left\|U_{[I_k]}\right\| \lesssim_{\mathrm{P}} q^{1/2}N^{1/2-\nu/2}T^{1/2}$, $\widehat{\lambda}_{(k)} \asymp_{\mathrm{P}} qN^{1-\nu}$ and Lemma S4.4(i) we have

$$\begin{aligned}\left\|\frac{\widehat{\varsigma}_{(k)}\widetilde{U}_{(k)}}{\sqrt{T_h\widehat{\lambda}_{(k)}}}\right\| &\leq \left\|\frac{\widehat{\varsigma}_{(k)}U_{[I_k]}}{\sqrt{T_h\widehat{\lambda}_{(k)}}}\right\| + \sum_{i=1}^{k-1}\left\|\frac{X_{[I_k]}}{\sqrt{T_h\widehat{\lambda}_{(k)}}}\right\|\left\|\frac{\widehat{\varsigma}_{(i)}\widetilde{U}_{(i)}}{\sqrt{T_h\widehat{\lambda}_{(i)}}}\right\| \\ &\lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2+\nu/2} + T^{-1}N^{\nu} + \sum_{i=1}^{k-1}\left\|\frac{\widehat{\varsigma}_{(i)}\widetilde{U}_{(i)}}{\sqrt{T_h\widehat{\lambda}_{(i)}}}\right\|.\end{aligned} \tag{71}$$

If $\left\|T_h^{-1/2}\widehat{\lambda}_{(i)}^{-1/2}\widehat{\varsigma}_{(i)}\widetilde{U}_{(i)}\right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2+\nu/2} + T^{-1}N^{\nu}$ holds for $i \leq k-1$, then (71) implies that this inequality also holds for k. In addition, when $k = 1$, $\widetilde{U}_{(1)} = U_{[I_1]}$ and this equation is implied from Lemma S4.4(i). Therefore, we have (i) holds for $k \leq \tilde{K}$ by induction.

Using (70) again, with Assumption 3, we have

$$\begin{aligned}\left\|\frac{\widetilde{U}_{(k)}}{\sqrt{T_h\widehat{\lambda}_{(k)}}}\right\| &\leq \left\|\frac{U_{[I_k]}}{\sqrt{T_h\widehat{\lambda}_{(k)}}}\right\| + \sum_{i=1}^{k-1}\left\|\frac{X_{[I_k]}}{\sqrt{T_h\widehat{\lambda}_{(k)}}}\right\|\left\|\frac{\widetilde{U}_{(i)}}{\sqrt{T_h\widehat{\lambda}_{(i)}}}\right\| \\ &\lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2+\nu/2} + T^{-1/2}N^{\nu/2} + \sum_{i=1}^{k-1}\left\|\frac{\widetilde{U}_{(i)}}{\sqrt{T_h\widehat{\lambda}_{(i)}}}\right\|.\end{aligned} \tag{72}$$

When $k = 1$, Assumption 3 implies $\left\|T_h^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}\widetilde{U}_{(k)}\right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2+\nu/2} + T^{-1/2}N^{\nu/2}$. Then, using the same induction argument with (72), we have this inequality holds for $k \leq \tilde{K}$.

(ii) Similarly, by simple multiplication of F' from the right side of (70), we have

$$\frac{\widehat{\varsigma}_{(k)}\widetilde{U}_{(k)}F'}{T_h\sqrt{\widehat{\lambda}_{(k)}}} = \frac{\widehat{\varsigma}_{(k)}U_{[I_k]}F'}{T_h\sqrt{\widehat{\lambda}_{(k)}}} - \sum_{i=1}^{k-1}\frac{\widehat{\varsigma}_{(k)}X_{[I_k]}\widehat{\xi}_{(i)}}{\sqrt{T_h\widehat{\lambda}_{(k)}}}\frac{\widehat{\varsigma}_{(i)}\widetilde{U}_{(i)}F'}{T_h\sqrt{\widehat{\lambda}_{(i)}}}.$$

Consequently, we have

$$\begin{aligned}\left\|\frac{\widehat{\varsigma}_{(k)}\widetilde{U}_{(k)}F'}{T_h\sqrt{\widehat{\lambda}_{(k)}}}\right\| &\leq \left\|\frac{\widehat{\varsigma}_{(k)}U_{[I_k]}F'}{T_h\sqrt{\widehat{\lambda}_{(k)}}}\right\| + \sum_{i=1}^{k-1}\left\|\frac{X_{[I_k]}}{\sqrt{T_h\widehat{\lambda}_{(k)}}}\right\|\left\|\frac{\widehat{\varsigma}_{(i)}\widetilde{U}_{(i)}F'}{T_h\sqrt{\widehat{\lambda}_{(i)}}}\right\| \\ &\lesssim_{\mathrm{P}} q^{-1}N^{-1+\nu} + T^{-1}N^{\nu} + \sum_{i=1}^{k-1}\left\|\frac{\widehat{\varsigma}_{(i)}\widetilde{U}_{(i)}F'}{\sqrt{T_h\widehat{\lambda}_{(i)}}}\right\|.\end{aligned} \tag{73}$$

When $k = 1$, $\left\|T_h^{-1}\widehat{\lambda}_{(k)}^{-1/2}\widehat{\varsigma}_{(k)}\widetilde{U}_{(k)}F'\right\| \lesssim_{\mathrm{P}} q^{-1}N^{-1+\nu} + T^{-1}N^{\nu}$ is a result of Lemma S4.4(ii). Then, a direct induction argument using (73) leads to this inequality for $k \leq \tilde{K}$. Replacing F by Z and W in the above proof, we have (ii).

(iii) Recall that $\widetilde{X}_{(k)} = \widetilde{\beta}_{(k)}F + \widetilde{U}_{(k)}$ as defined in (3), we have

$$|\widehat{\varsigma}'_{(l)}\widetilde{X}_{(l)}U'_{[I_k]}\widehat{\varsigma}_{(k)}| \le \left\|\widehat{\varsigma}'_{(l)}\widetilde{\beta}_{(l)}\right\| \left\|FU'_{[I_k]}\widehat{\varsigma}_{(k)}\right\| + \left\|\widehat{\varsigma}'_{(l)}\widetilde{U}_{(l)}\right\| \left\|U'_{[I_k]}\widehat{\varsigma}_{(k)}\right\|.$$

Along with (1), we have

$$|\frac{\widehat{\xi}'_{(l)}U'_{[I_k]}\widehat{\varsigma}_{(k)}}{\sqrt{T_h\widehat{\lambda}_{(k)}}}| \le \left\|\frac{\widehat{\varsigma}'_{(l)}\widetilde{\beta}_{(l)}}{\sqrt{\widehat{\lambda}_{(l)}}}\right\| \left\|\frac{FU'_{[I_k]}\widehat{\varsigma}_{(k)}}{T_h\sqrt{\widehat{\lambda}_{(k)}}}\right\| + \left\|\frac{U'_{[I_k]}\widehat{\varsigma}_{(k)}}{\sqrt{T_h\widehat{\lambda}_{(k)}}}\right\| \left\|\frac{\widetilde{U}'_{(l)}\widehat{\varsigma}_{(l)}}{\sqrt{T_h\widehat{\lambda}_{(l)}}}\right\|. \tag{74}$$

Using $\left\|\widehat{\lambda}_{(k)}^{-1/2}\widehat{\varsigma}'_{(k)}\widetilde{\beta}_{(k)}\right\| \lesssim_{\mathrm{P}} 1$ from Lemma S4.9, results of (i)(ii) and Lemma S4.4(i) completes the proof. Replacing $U_{[I_k]}$ by $\widetilde{U}_{(k)}$ above and using the inequality that

$$|\widehat{\varsigma}'_{(l)}\widetilde{X}_{(l)}\widetilde{U}'_{(k)}\widehat{\varsigma}_{(k)}| \le \left\|\widehat{\varsigma}'_{(l)}\widetilde{\beta}_{(l)}\right\| \left\|F\widetilde{U}'_{(k)}\widehat{\varsigma}_{(k)}\right\| + \left\|\widehat{\varsigma}'_{(l)}\widetilde{U}_{(l)}\right\| \left\|\widetilde{U}'_{(k)}\widehat{\varsigma}_{(k)}\right\|$$

and (1), we obtain the second equation in (iii).

(iv) Similar to (ii), by induction, we have

$$\left\|\widetilde{D}_{(k)}\right\| \le 1 + \sum_{i<k}\left\|\frac{X_{[I_k]}}{\sqrt{T_h\widehat{\lambda}_{(i)}}}\right\| \left\|\widetilde{D}_{(i)}\right\| \lesssim_{\mathrm{P}} 1.$$

and

$$\left\|D_{(k)}\right\| \le 1 + \sum_{i<k}\lambda_{(i)}^{-1/2}\left\|\beta_{[I_k]}\right\| \left\|D_{(i)}\right\| \lesssim 1$$

Together with Assumption 4(iii), we have (iv). □

Lemma S4.6. *Under assumptions of Theorem 3.1, for $k \le \tilde{K}$, we have*

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

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

Proof. (i) By the definitions of b_{k2} and $\widehat{\xi}_{(k)}$, $\widetilde{X}_{(k)} = \widetilde{\beta}_{(k)}F + \widetilde{U}_{(k)}$, we have

$$\widehat{\xi}_{(k)} - T_h^{-1/2}F'b_{k2} = \frac{\widetilde{U}'_{(k)}\widehat{\varsigma}_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}}. \tag{75}$$

Then, Lemma S4.5(i) leads to (i) directly.

(ii) Similarly, Lemma S4.5(ii) yields (ii). □

Lemma S4.7. *Under assumptions of Theorem 3.1, for $k \leq \tilde{K}$, we have*

(i) $\left\|\widehat{\varsigma}_{(k)} - \varsigma_{(k)}\right\| \lesssim_{\mathrm{P}} T^{-1/2}N^{\nu/2} + q^{-1/2}N^{-1/2+\nu/2}$.

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

Proof. We prove (i) and (ii) by induction. Consider the $k = 1$ case. The definitions of $\widehat{\varsigma}_{(k)}$ in (1) and $\varsigma_{(k)}$ in Section 3.1 lead to

$$\widehat{\varsigma}_{(k)} - \varsigma_{(k)} = T_h^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}(\widetilde{D}_{(k)}\beta F\widehat{\xi}_{(k)} + \widetilde{D}_{(k)}U\widehat{\xi}_{(k)}) - \lambda_{(k)}^{-1/2}D_{(k)}\beta b_{(k)}, \tag{76}$$

when $k = 1$, as $\widetilde{D}_{(1)} = D_{(1)} = (\mathbb{I}_N)_{[I_1]}$, (76) becomes

$$\widehat{\varsigma}_{(1)} - \varsigma_{(1)} = (T_h^{-1/2}\widehat{\lambda}_{(1)}^{-1/2}\beta_{(1)}F\widehat{\xi}_{(1)} - \lambda_{(1)}^{-1/2}\beta_{(1)}b_{(1)}) + T_h^{-1/2}\widehat{\lambda}_{(1)}^{-1/2}U_{[I_1]}\widehat{\xi}_{(1)}. \tag{77}$$

As Lemma S4.3(iii) and (iv) imply that $\left\|T_h^{-1/2}\widehat{\lambda}_{(1)}^{-1/2}F\widehat{\xi}_{(1)} - \lambda_{(1)}^{-1/2}b_{(1)}\right\| \lesssim_{\mathrm{P}} q^{-1}N^{-1+\nu} + T^{-1/2}q^{-1/2}N^{-1/2+\nu}$ and $\left\|\beta_{(1)}\right\| \lesssim q^{1/2}N^{1/2-\nu/2}$, to prove (i) it is sufficient to show that $\left\|U_{[I_1]}\right\| \lesssim_{\mathrm{P}} T^{1/2} + q^{1/2}N^{1/2}$, which is given by Assumption 3.

(ii) is equivalent to (i) when $k = 1$ as $\widetilde{D}_{(1)} = D_{(1)} = (\mathbb{I}_N)_{[I_1]}$.

Then, we assume that (i) and (ii) hold for $i < k$ and prove (i) and (ii) also hold for k.

(i) Similar to the $k = 1$ case, using (76) and Lemma S4.3(iii)(iv), it is sufficient to show that $\left\|\widetilde{U}_{(k)}\right\| \lesssim_{\mathrm{P}} T^{1/2} + q^{1/2}N^{1/2}$ and $\left\|(\widetilde{D}_{(k)} - D_{(k)})\beta\right\| \lesssim_{\mathrm{P}} 1 + q^{1/2}N^{1/2}T^{-1/2}$. The first inequality is the same as the $k = 1$ case, which is implied by Lemma S4.5. As to the second inequality, write

$$(\widetilde{D}_{(k)} - D_{(k)})\beta = \sum_{i<k}\left(\frac{\beta_{[I_k]}b_{(i)}}{\sqrt{\lambda_{(i)}}}\varsigma'_{(i)}D_{(i)}\beta - \frac{X_{[I_k]}\widehat{\xi}_{(i)}}{\sqrt{T_h\widehat{\lambda}_{(i)}}}\widehat{\varsigma}'_{(i)}\widetilde{D}_{(i)}\beta\right).$$

As (ii) holds for $i < k$ and $\|\beta\| \lesssim q^{1/2}N^{1/2-\nu/2}$, it is sufficient to show that

$$\left\|\frac{\beta_{[I_k]}b_{(i)}}{\sqrt{\lambda_{(i)}}} - \frac{X_{[I_k]}\widehat{\xi}_{(i)}}{\sqrt{T_h\widehat{\lambda}_{(i)}}}\right\| \lesssim_{\mathrm{P}} T^{-1/2}N^{\nu/2} + q^{-1/2}N^{-1/2+\nu/2}. \tag{78}$$

Plugging $X_{[I_k]} = \beta_{[I_k]}F + U_{[I_k]}$ into (78) and using Lemma S4.3(iii)(iv) again, we only need to show that $\left\|U_{[I_k]}\widehat{\xi}_{(i)}\right\| \lesssim_{\mathrm{P}} q^{1/2}N^{1/2} + T^{1/2}$, which holds by Assumption 3 and $\left\|\widehat{\xi}_{(i)}\right\| = 1$.

(ii) By simple algebra, we have

$$\widehat{\varsigma}'_{(k)}\widetilde{D}_{(k)} - \varsigma'_{(k)}D_{(k)} = (\widehat{\varsigma}'_{(k)} - \varsigma'_{(k)})(\mathbb{I}_N)_{[I_k]} + \sum_{i<k}\left(\frac{\varsigma'_{(k)}\beta_{[I_k]}b_{(i)}}{\sqrt{\lambda_{(i)}}}\varsigma'_{(i)}D_{(i)} - \frac{\widehat{\varsigma}'_{(k)}X_{[I_k]}\widehat{\xi}_{(i)}}{\sqrt{T_h\widehat{\lambda}_{(i)}}}\widehat{\varsigma}'_{(i)}\widetilde{D}_{(i)}\right).$$

Using the fact that (i) holds or $i < k$ and (78), the proof is completed. □

Lemma S4.8. *Under assumptions of Theorem 3.1, for $k \le \tilde{K}+1$, we have*

(i) $\left\|\widetilde{Z}_{(k)}F'\right\| \lesssim_{\mathrm{P}} T^{1/2} + Tq^{-1}N^{-1+\nu}$.

(ii) $\left\|\widetilde{Z}_{(k)}U'_{[I_0]}\right\| \lesssim_{\mathrm{P}} N_0^{1/2}T^{1/2} + Tq^{-1/2}N^{-1/2+\nu/2}$.

Proof. (i) From the definition (7) of $\widetilde{Z}_{(k)}$, we have

$$\widetilde{Z}_{(k)}F' = ZF' - \sum_{i=1}^{k-1} Y\widehat{\xi}_{(i)}\frac{\widehat{\varsigma}'_{(i)}\widetilde{U}_{(i)}F'}{\sqrt{T_h\widehat{\lambda}_{(i)}}}.$$

Then along with Lemma S4.5(ii), we have

$$\left\|\widetilde{Z}_{(k)}F'\right\| \le \|ZF'\| + \sum_{i=1}^{k-1}\left\|Y\widehat{\xi}_{(i)}\right\|\left\|\frac{\widehat{\varsigma}'_{(i)}\widetilde{U}_{(i)}F'}{\sqrt{T_h\widehat{\lambda}_{(i)}}}\right\| \lesssim_{\mathrm{P}} T^{1/2} + Tq^{-1}N^{-1+\nu}.$$

(ii) With (7) again, we have

$$\widetilde{Z}_{(k)}U'_{[I_0]} = ZU'_{[I_0]} - \sum_{i=1}^{k-1} Y\widehat{\xi}_{(i)}\frac{\widehat{\varsigma}'_{(i)}\widetilde{U}_{(i)}U'_{[I_0]}}{\sqrt{T_h\widehat{\lambda}_{(i)}}},$$

which, along with Lemma S4.5(i) and the assumptions on q, lead to

$$\begin{aligned}\left\|\widetilde{Z}_{(k)}U'_{[I_0]}\right\| &\le \left\|ZU'_{[I_0]}\right\| + \sum_{i=1}^{k-1}\left\|Y\widehat{\xi}_{(i)}\right\|\left\|\frac{\widehat{\varsigma}'_{(i)}\widetilde{U}_{(i)}}{\sqrt{T_h\widehat{\lambda}_{(i)}}}\right\|\left\|U_{[I_0]}\right\| \\ &\lesssim_{\mathrm{P}} N_0^{1/2}T^{1/2} + \left(q^{-1/2}N^{-1/2+\nu/2} + T^{-1}N^{\nu}\right)\left(N_0^{1/2}T^{1/2} + T\right) \\ &\lesssim_{\mathrm{P}} N_0^{1/2}T^{1/2} + Tq^{-1/2}N^{-1/2+\nu/2}.\end{aligned}$$

□

Lemma S4.9. *Under assumptions of Theorem 3.1, B_1, B_2 defined by (19) satisfy*

(i) $\|B_1\| \lesssim_{\mathrm{P}} 1$, $\|B_2\| \lesssim_{\mathrm{P}} 1$.

(ii) $\left\|B_1'B_2 - \mathbb{I}_{\widetilde{K}}\right\| \lesssim_{\mathrm{P}} q^{-1}N^{-1+\nu} + T^{-1}N^{\nu}$.

(iii) $\|B_1 - B_2\| \lesssim_{\mathrm{P}} T^{-1/2} + T^{-1}N^{\nu} + q^{-1}N^{-1+\nu}$, $\|B_1 - B\| \lesssim_{\mathrm{P}} T^{-1/2}N^{\nu/2} + q^{-1/2}N^{-1/2+\nu/2}$.

(iv) $\|B_2B_2' - \mathbb{I}_K\| \lesssim_{\mathrm{P}} T^{-1/2} + T^{-1}N^{\nu} + q^{-1}N^{-1+\nu}$ *when* $\tilde{K} = K$.

(v) $\|B_1B_2' - \mathbb{I}_K\| \lesssim_{\mathrm{P}} T^{-1}N^{\nu} + q^{-1}N^{-1+\nu}$ *when* $\tilde{K} = K$.

Proof. (i) Using the definition (19) of B_1 and Assumption 1, we have

$$\|b_{k1}\| = \left\|\frac{F\widehat{\xi}_{(k)}}{\sqrt{T_h}}\right\| \lesssim_{\mathrm{P}} 1,$$

which leads to $\|B_1\| \lesssim_{\mathrm{P}} 1$. Using the definition (19) of B_2, we have

$$\|b_{k2}\| = \left\|\frac{\widetilde{\beta}_{(k)}'\widehat{\varsigma}_{(k)}}{\sqrt{\widehat{\lambda}_{(k)}}}\right\| \leq q^{-1/2}N^{-1/2+\nu/2}\left\|\widetilde{\beta}_{(k)}\right\|. \tag{79}$$

Note that

$$\left\|\widetilde{\beta}_{(k)}\right\| \leq \left\|\beta_{[I_k]}\right\| + \sum_{i=1}^{k-1}\left\|\frac{X_{[I_k]}\widehat{\xi}_{(i)}}{\sqrt{T_h\widehat{\lambda}_{(i)}}}\right\|\left\|\widehat{\varsigma}_{(i)}'\widetilde{\beta}_{(i)}\right\| \lesssim_{\mathrm{P}} q^{1/2}N^{1/2-\nu/2} + \sum_{i=1}^{k-1}\left\|\widetilde{\beta}_{(i)}\right\| \tag{80}$$

and $\left\|\widetilde{\beta}_{(1)}\right\| = \left\|\widetilde{\beta}_{[I_1]}\right\| \lesssim q^{1/2}N^{1/2-\nu/2}$, we have $\left\|\widetilde{\beta}_{(k)}\right\| \lesssim q^{1/2}N^{1/2-\nu/2}$ by induction. Together with (79), we have $\|b_{k2}\| \lesssim_{\mathrm{P}} 1$ and thus $\|B_2\| \lesssim_{\mathrm{P}} 1$.

(ii) By (1) and Lemma S4.1, we have

$$\delta_{lk} = \widehat{\xi}_{(l)}'\widehat{\xi}_{(k)} = \frac{\widehat{\xi}_{(l)}'F'\widetilde{\beta}_{(k)}'\widehat{\varsigma}_{(k)}}{\sqrt{T_h\widehat{\lambda}_{(k)}}} + \frac{\widehat{\xi}_{(l)}'\widetilde{U}_{(k)}'\widehat{\varsigma}_{(k)}}{\sqrt{T_h\widehat{\lambda}_{(k)}}} = b_{l1}'b_{k2} + \frac{\widehat{\xi}_{(l)}'\widetilde{U}_{(k)}'\widehat{\varsigma}_{(k)}}{\sqrt{T_h\widehat{\lambda}_{(k)}}}.$$

By Lemma S4.5(iii), we have $|b_{l1}'b_{k2} - \delta_{lk}| \lesssim_{\mathrm{P}} q^{-1}N^{-1+\nu} + T^{-1}N^{\nu}$, and thus $\left\|B_1'B_2 - \mathbb{I}_{\widetilde{K}}\right\| \lesssim_{\mathrm{P}}$ $q^{-1}N^{-1+\nu} + T^{-1}N^{\nu}$.

(iii) Using (1) and $\widetilde{X}_{(k)} = \widetilde{\beta}_{(k)}F + \widetilde{U}_{(k)}$, we have

$$F\widehat{\xi}_{(k)} = \frac{FF'\widetilde{\beta}_{(k)}'}{\sqrt{T_h\widehat{\lambda}_{(k)}}}\widehat{\varsigma}_{(k)} + \frac{F\widetilde{U}_{(k)}'\widehat{\varsigma}_{(k)}}{\sqrt{T_h\widehat{\lambda}_{(k)}}}.$$

By the definitions of b_{k1} and b_{k2}, it becomes

$$b_{k1} = \frac{FF'}{T_h} b_{k2} + \frac{F\widetilde{U}'_{(k)}\widehat{\varsigma}_{(k)}}{T_h\sqrt{\widehat{\lambda}_{(k)}}}. \tag{81}$$

With $\|B_2\| \lesssim_{\mathrm{P}} 1$, Assumption 1 and Lemma S4.5(ii), (81) leads to

$$b_{k1} - b_{k2} \lesssim_{\mathrm{P}} T^{-1/2} + T^{-1}N^{\nu} + q^{-1}N^{-1+\nu}.$$

This completes the proof. The second inequality of (iii) comes from Lemma S4.3(iv) directly.

(iv) When $\tilde{K} = K$, $B_2'B_2$ is a $K \times K$ matrix. By (i), (ii) and (iii), we have

$$\|B_2'B_2 - \mathbb{I}_K\| \leq \|B_1'B_2 - \mathbb{I}_K\| + \|B_1 - B_2\|\,\|B_2\| \lesssim_{\mathrm{P}} T^{-1/2} + T^{-1}N^{\nu} + q^{-1}N^{-1+\nu}.$$

Since B_2 is a $K \times K$ matrix, we have

$$\|B_2B_2' - \mathbb{I}_K\| = \max_{1\leq i\leq K} |\lambda_i(B_2'B_2) - 1| = \|B_2'B_2 - \mathbb{I}_K\| \lesssim_{\mathrm{P}} T^{-1/2} + T^{-1}N^{\nu} + q^{-1}N^{-1+\nu}.$$

(v) With respect to B_1B_2', we have

$$\sigma_K(B_2)\,\|B_2B_1' - \mathbb{I}_K\| \leq \|(B_2B_1' - \mathbb{I}_K)B_2\| = \|B_2(B_1'B_2 - \mathbb{I}_K)\| \leq \sigma_1(B_2)\,\|B_1'B_2 - \mathbb{I}_K\|. \tag{82}$$

Since (iv) implies that $\sigma_1(B_2)/\sigma_K(B_2) \lesssim_{\mathrm{P}} 1$ when $\tilde{K} = K$, (ii) and (82) yield

$$\|B_1B_2' - \mathbb{I}_K\| = \|B_2B_1' - \mathbb{I}_K\| \leq \frac{\sigma_1(B_2)}{\sigma_K(B_2)}\,\|B_1'B_2 - \mathbb{I}_K\| \lesssim_{\mathrm{P}} T^{-1}N^{\nu} + q^{-1}N^{-1+\nu}. \tag{83}$$

□

Lemma S4.10. *Under Assumptions 1-5, we have*

(i) $\left\|T_h^{1/2}\widehat{\xi}'_{(k)} - b'_{k2}F\right\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2+\nu/2}(\log T)^{1/2} + T^{-1/2}N^{\nu} + q^{-1}N^{-1+\nu}T^{1/2}.$

(ii) $\left\|\widehat{\lambda}_{(k)}^{1/2}\widehat{\beta}_{(k)} - \beta b_{k1}\right\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} (\log NT)^{1/2}(T^{-1/2} + q^{-1/2}N^{-1/2+\nu/2} + T^{-1}N^{\nu}).$

(iii) $\left\|\widehat{U} - U\right\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} \log(NT)(q^{-1/2}N^{-1/2+\nu/2} + T^{-1/2}N^{\nu} + q^{-1}N^{-1+\nu}T^{1/2}).$

(iv) $\max_{i\leq N} T_h^{-1/2}\left\|\widehat{U}_{[i]} - U_{[i]}\right\| \lesssim_{\mathrm{P}} (\log NT)^{1/2}(T^{-1/2}N^{\nu/2} + q^{-1/2}N^{-1/2+\nu/2}).$

Proof. (i) Recall that by (75), (1), and (3), we have $T_h^{1/2}\widehat{\xi}'_{(k)} - b'_{k2}F = \widehat{\lambda}_{(k)}^{-1/2}\widehat{\varsigma}'_{(k)}\widetilde{U}_{(k)}$, and $\widehat{\varsigma}_{(k)} = T^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}\widetilde{X}_{(k)}\widehat{\xi}_{(k)} = T^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}X_{[I_k]}\widehat{\xi}_{(k)}$. Therefore, we have

$$\left\|T_h^{1/2}\widehat{\xi}_{(k)} - b'_{k2}F\right\|_{\text{MAX}} \lesssim_{\text{P}} q^{-1}N^{-1+\nu}T^{-1/2}\left(\left\|\widehat{\xi}'_{(k)}F'\beta'_{[I_k]}\widetilde{U}_{(k)}\right\|_{\text{MAX}} + \left\|\widehat{\xi}'_{(k)}U'_{[I_k]}\widetilde{U}_{(k)}\right\|_{\text{MAX}}\right). \tag{84}$$

When $k = 1$, $\widetilde{U}_{(1)} = U_{[I_1]}$, with $\left\|\beta'_{[I_1]}\widetilde{U}_{(1)}\right\|_{\text{MAX}} \lesssim_{\text{P}} q^{1/2}N^{1/2-\nu/2}(\log T)^{1/2}$ from Assumption 4 and $\left\|U_{[I_1]}\right\| \lesssim_{\text{P}} q^{1/2}N^{1/2} + T^{1/2}$ from Assumption 3, we have

$$\begin{aligned}\left\|T_h^{1/2}\widehat{\xi}_{(1)} - b'_{12}F\right\|_{\text{MAX}} &\lesssim_{\text{P}} q^{-1}N^{-1+\nu}\left\|\beta'_{[I_1]}U_{[I_1]}\right\|_{\text{MAX}} + q^{-1}N^{-1+\nu}T^{-1/2}\left\|U'_{[I_1]}U_{[I_1]}\right\| \\ &\lesssim_{\text{P}} q^{-1/2}N^{-1/2+\nu/2}(\log T)^{1/2} + T^{-1/2}N^{\nu} + q^{-1}N^{-1+\nu}T^{1/2}.\end{aligned}$$

Now suppose that this property holds for $i < k$, then for the first term in (84), we have

$$\left\|\beta'_{[I_k]}\widetilde{U}_{(k)}\right\|_{\text{MAX}} \lesssim \left\|\beta'_{[I_k]}U_{[I_k]}\right\|_{\text{MAX}} + \sum_{i<k}\left\|\beta_{[I_k]}\right\|\left\|T_h^{-1/2}\widehat{\lambda}_{(i)}^{-1/2}X_{[I_k]}\widehat{\xi}_{(i)}\right\|\left\|\widehat{\varsigma}'_{(i)}\widetilde{U}_{(i)}\right\|_{\text{MAX}}.$$

The assumption that (i) holds for $i < k$ implies that

$$\left\|\widehat{\varsigma}'_{(i)}\widetilde{U}_{(i)}\right\|_{\text{MAX}} = \widehat{\lambda}_{(k)}^{1/2}\left\|T_h^{1/2}\widehat{\xi}_{(k)} - b'_{k2}F\right\|_{\text{MAX}} \lesssim_{\text{P}} (\log T)^{1/2} + q^{1/2}N^{1/2+\nu/2}T^{-1/2} + q^{-1/2}N^{-1/2+\nu/2}.$$

With $\left\|\beta_{[I_k]}\right\| \lesssim q^{1/2}N^{1/2-\nu/2}$ and $\left\|X_{[I_k]}\right\| \leq \left\|\beta_{[I_k]}\right\|\|F\| + \left\|U_{[I_k]}\right\| \lesssim_{\text{P}} q^{1/2}N^{1/2-\nu/2}T^{1/2}$ and Assumption 4(ii), we have the first term in (84) satisfies

$$\begin{aligned}&q^{-1}N^{-1+\nu}T^{-1/2}\left\|\widehat{\xi}'_{(k)}F'\beta'_{[I_k]}\widetilde{U}_{(k)}\right\|_{\text{MAX}} \lesssim q^{-1}N^{-1+\nu}T^{-1/2}\left\|\widehat{\xi}'_{(k)}F'\right\|\left\|\beta'_{[I_k]}\widetilde{U}_{(k)}\right\|_{\text{MAX}} \\ &\lesssim_{\text{P}} q^{-1}N^{-1+\nu}\left\|\beta'_{[I_k]}U_{[I_k]}\right\|_{\text{MAX}} + \sum_{i<k}q^{-1/2}N^{-1/2+\nu/2}\left\|\widehat{\varsigma}'_{(i)}\widetilde{U}_{(i)}\right\|_{\text{MAX}} \\ &\lesssim_{\text{P}} q^{-1/2}N^{-1/2+\nu/2}(\log T)^{1/2} + T^{-1/2}N^{\nu} + q^{-1}N^{-1+\nu}T^{1/2}.\end{aligned}$$

For the second term in (84), we have

$$\left\|\widehat{\xi}'_{(k)}U'_{[I_k]}\widetilde{U}_{(k)}\right\|_{\text{MAX}} \leq \left\|\widehat{\xi}'_{(k)}U'_{[I_k]}\widetilde{U}_{(k)}\right\| \leq \left\|U_{[I_k]}\right\|\left\|\widetilde{U}_{(k)}\right\| \lesssim q^{1/2}N^{1/2} + T^{1/2},$$

where we use Assumption 3 and Lemma S4.5(i) in the last step. Consequently, (i) also holds for k and this concludes the proof by induction.

(ii) By simple algebra, $\widehat{\lambda}_{(k)}^{1/2}\widehat{\beta}_{(k)} = T_h^{-1/2}X\widehat{\xi}_{(k)} = \beta b_{k1} + T_h^{-1/2}U\widehat{\xi}_{(k)}$, which leads to

$$\begin{aligned}\left\|\widehat{\lambda}_{(k)}^{1/2}\widehat{\beta}_{(k)} - \beta b_{k1}\right\|_{\text{MAX}} \lesssim & T_h^{-1}\left\|UF'b_{k2}\right\|_{\text{MAX}} + T^{-1}\left\|U(T_h^{1/2}\widehat{\xi}_{(k)} - F'b_{k2})\right\|_{\text{MAX}} \\ \lesssim_{\text{P}} & T^{-1}\left\|UF'\right\|_{\text{MAX}} + T^{-1/2}\left\|U\right\|_{\text{MAX}}\left\|T_h^{1/2}\widehat{\xi}_{(k)} - F'b_{k2}\right\| \\ \lesssim_{\text{P}} & (\log NT)^{1/2}(T^{-1/2} + q^{-1/2}N^{-1/2+\nu/2} + T^{-1}N^{\nu}),\end{aligned}$$

where we use Assumptions 3, 4, and Lemma S4.6.

(iii) By triangle inequality, we have

$$\left\|\widehat{U} - U\right\|_{\text{MAX}} \leq \left\|\beta\left(\sum_{k\leq K} b_{k1}b_{k2}' - \mathbb{I}_K\right)F\right\|_{\text{MAX}} + \sum_{k\leq K}\left\|\widehat{\beta}_{(k)}\widehat{F} - \beta b_{k1}b_{k2}'F\right\|_{\text{MAX}}.$$

By Assumptions 1, 2 and Lemma S4.9, the first term satisfies

$$\begin{aligned}\left\|\beta\left(\sum_{k\leq K} b_{k1}b_{k2}' - \mathbb{I}_K\right)F\right\|_{\text{MAX}} \lesssim & \left\|\beta\right\|_{\text{MAX}}\left\|F\right\|_{\text{MAX}}\left\|B_1B_2' - \mathbb{I}_K\right\| \\ \lesssim_p & (\log T)^{1/2}(T^{-1}N^{\nu} + q^{-1}N^{-1+\nu}).\end{aligned}$$

For the second term, note that by triangle inequality we have

$$\begin{aligned}&\left\|\widehat{\beta}_{(k)}\widehat{F} - \beta b_{k1}b_{k2}'F\right\|_{\text{MAX}} \leq \left\|\widehat{\lambda}_{(k)}^{1/2}\widehat{\beta} - \beta b_{k1}\right\|_{\text{MAX}}\left\|b_{k2}'F\right\|_{\text{MAX}} \\ &+ \left\|\beta b_{k1}\right\|_{\text{MAX}}\left\|b_{k2}'F - \widehat{\lambda}_{(k)}^{-1/2}\widehat{F}\right\|_{\text{MAX}} + \left\|\widehat{\lambda}_{(k)}^{1/2}\widehat{\beta} - \beta b_{k1}\right\|_{\text{MAX}}\left\|b_{k2}'F - \widehat{\lambda}_{(k)}^{-1/2}\widehat{F}\right\|_{\text{MAX}},\end{aligned}$$

which, together with (i)(ii), concludes the proof.

(iv) Because we have $T_h^{-1/2}\left\|\widehat{U}_{[i]} - U_{[i]}\right\| = T_h^{-1/2}\left\|\widehat{\beta}_{[i]}\widehat{F} - \beta_{[i]}F\right\|$, it then follows from triangle inequality that

$$\begin{aligned}\left\|\widehat{\beta}_{[i]}\widehat{F} - \beta_{[i]}F\right\| \leq & \left\|\beta_{[i]}\left(B_1B_2' - \mathbb{I}_K\right)F\right\| + \left\|\beta_{[i]}B_1\right\|\left\|\widehat{\Lambda}^{-1/2}\widehat{F} - B_2'F\right\| \\ & + \left\|\beta_{[i]}B_1 - \widehat{\beta}_{[i]}\widehat{\Lambda}^{1/2}\right\|\left\|\widehat{\Lambda}^{-1/2}\widehat{F}\right\|.\end{aligned}$$

We analyze the three terms on the right-hand side one by one. With $T^{-1/2}\left\|\widehat{\Lambda}^{-1/2}\widehat{F} - B_2'F\right\| \lesssim_{\text{P}} T^{-1}N^{\nu} + q^{-1/2}N^{-1/2+\nu/2}$ from Lemma S4.6, $\left\|\beta\right\|_{\text{MAX}} \lesssim N^{-\nu/2}$, Lemma S4.9 and (ii), we have

$$\max_i T_h^{-1/2}\left\|\beta_{[i]}\left(B_1B_2' - \mathbb{I}_K\right)F\right\| \lesssim T_h^{-1/2}\left\|\beta\right\|_{\text{MAX}}\left\|B_1B_2' - \mathbb{I}_K\right\|\left\|F\right\| \lesssim_{\text{P}} q^{-1}N^{-1+\nu/2} + T^{-1}N^{\nu/2},$$

$$\max_i T_h^{-1/2} \left\|\beta_{[i]} B_1\right\| \left\|\widehat{\Lambda}^{-1/2}\widehat{F} - B_2' F\right\| \lesssim T_h^{-1/2} \left\|\beta\right\|_{\text{MAX}} \left\|\widehat{\Lambda}^{-1/2}\widehat{F} - B_2' F\right\| \lesssim_{\text{P}} q^{-1/2}N^{-1/2} + T^{-1}N^{\nu/2},$$

$$\max_i T_h^{-1/2} \left\|\beta_{[i]} B_1 - \widehat{\beta}_{[i]}\widehat{\Lambda}^{1/2}\right\| \left\|\widehat{\Lambda}^{-1/2}\widehat{F}\right\| \lesssim T_h^{-1/2} \left\|\beta B_1 - \widehat{\beta}\widehat{\Lambda}^{1/2}\right\|_{\text{MAX}} \|F\|$$

$$\lesssim_{\text{P}} (\log NT)^{1/2}(T^{-1/2} + q^{-1/2}N^{-1/2+\nu/2} + T^{-1}N^{\nu}).$$

Consequently, we have the desired bound. □

Lemma S4.11. *Under Assumptions 1-4, for any $I \subset [N]$, we have the following results:*

(i) $\left\|T_h^{-1} F\mathbb{M}_{W'}F' - \mathbb{I}_K\right\| \lesssim_{\text{P}} T^{-1/2}$, $\|Z\mathbb{M}_{W'}\| \lesssim_{\text{P}} T^{1/2}$.

(ii) $\|F\mathbb{M}_{W'}\|_{\text{MAX}} \lesssim_{\text{P}} (\log T)^{1/2}$, $\|Z\mathbb{M}_{W'}\|_{\text{MAX}} \lesssim_{\text{P}} (\log T)^{1/2}$.

(iii) $\left\|\breve{\beta}_{[I]}' U_{[I]} M_{W'}\right\| \lesssim_{\text{P}} T^{1/2}$, $\left\|\breve{\beta}_{[I]}' U_{[I]} M_{W'}\right\|_{\text{MAX}} \lesssim_{\text{P}} (\log T)^{1/2}$.

(iv) $\left\|\breve{\beta}_{[I]}' U_{[I]}\mathbb{M}_{W'}F'\right\| \lesssim_{\text{P}} T^{1/2}$, $\left\|\breve{\beta}_{[I]}' U_{[I]}\mathbb{M}_{W'}Z'\right\| \lesssim_{\text{P}} T^{1/2}$.

(v) $\|U\mathbb{M}_{W'}\|_{\text{MAX}} \lesssim_{\text{P}} (\log NT)^{1/2}$.

(vi) $\|U\mathbb{M}_{W'}F'\|_{\text{MAX}} \lesssim_{\text{P}} (\log N)^{1/2}T^{1/2}$,$\|U\mathbb{M}_{W'}Z'\|_{\text{MAX}} \lesssim_{\text{P}} (\log N)^{1/2}T^{1/2}$.

(vii) $\left\|U_{[I]}\mathbb{M}_{W'}\right\| \lesssim_{\text{P}} |I|^{1/2} + T^{1/2}$, $\left\|U_{[I]}\mathbb{M}_{W'}A'\right\| \lesssim_{\text{P}} |I|^{1/2}T^{1/2}$, *for* $A = F, Z$.

(viii) $\|F\mathbb{M}_{W'}Z'\| \lesssim_{\text{P}} T^{1/2}$, $\|F\mathbb{M}_{W'}Z' - FZ'\| \lesssim_{\text{P}} 1$.

Proof. (i) With $\|(WW')^{-1}\| \lesssim_{\text{P}} T^{-1}$ from Assumption 1, $\|WF'\| \lesssim_{\text{P}} T^{1/2}$, we have

$$\left\|T_h^{-1} F\mathbb{M}_{W'}F' - \mathbb{I}_K\right\| \le \left\|T_h^{-1}FF' - \mathbb{I}_K\right\| + T_h^{-1}\left\|FW'\right\|^2 \left\|(WW')^{-1}\right\| \lesssim_{\text{P}} T^{-1/2}.$$

(ii) Using the bound on $\|F\|_{\text{MAX}}$ and that $\|W\| \lesssim T^{1/2}$ by Assumption 1, we have

$$\|F\mathbb{M}_{W'}\|_{\text{MAX}} \le \|F\|_{\text{MAX}} + \|FW'\| \left\|(WW')^{-1}\right\| \|W\| \lesssim_{\text{P}} (\log T)^{1/2}$$

Replacing F by Z in the above proof, we obtain the second inequality.

(iii) Using Assumption 4(ii) and $\|\mathbb{M}_{W'}\| \le 1$, the first equation holds directly. Also,

$$\left\|\breve{\beta}_{[I]}' U_{[I]}\mathbb{M}_{W'}\right\|_{\text{MAX}} \lesssim \left\|\breve{\beta}_{[I]}' U_{[I]}\right\|_{\text{MAX}} + \left\|\breve{\beta}_{[I]}' U_{[I]}W'\right\| \left\|(WW')^{-1}\right\| \|W\| \lesssim_{\text{P}} (\log T)^{1/2}$$

where we use Assumption 1 and Assumption 4(ii) in the last equality.

(iv) With $\|(WW')^{-1}\| \lesssim_{\mathrm{P}} T^{-1}$, $\|WF'\| \lesssim_{\mathrm{P}} T^{1/2}$, and by Assumption 4, we have

$$\left\|\breve{\beta}'_{[I]}U_{[I]}\mathbb{M}_{W'}F'\right\| \leq \left\|\breve{\beta}'_{[I]}U_{[I]}F'\right\| + \left\|\breve{\beta}'_{[I]}U_{[I]}W'\right\| \left\|(WW')^{-1}\right\| \|WF'\| \lesssim_{\mathrm{P}} T^{1/2}.$$

Replacing F by Z in the above proof, we have the second inequality in (iv).

(v) Similar to (ii), using Assumption 3 and 4, we have

$$\|U\mathbb{M}_{W'}\|_{\mathrm{MAX}} \lesssim \|U\|_{\mathrm{MAX}} + \|UW'\|_{\mathrm{MAX}} \left\|(WW')^{-1}W\right\| \lesssim_{\mathrm{P}} (\log N)^{1/2} + (\log T)^{1/2}.$$

(vi) Similar to (iv), by Assumptions 1 and 3, we have

$$\|U\mathbb{M}_{W'}F'\|_{\mathrm{MAX}} \lesssim \|UF'\|_{\mathrm{MAX}} + \|UW'\|_{\mathrm{MAX}} \left\|(WW')^{-1}WF'\right\| \lesssim_{\mathrm{P}} (\log N)^{1/2}T^{1/2}.$$

Replacing F by Z in the above inequality, we also have $\|U\mathbb{M}_{W'}Z'\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} (\log N)^{1/2}T^{1/2}$.

(vii) With Assumption 3 and $\|\mathbb{M}_{W'}\| \leq 1$, the first inequality holds directly. By Assumptions 1 and 4, we have

$$\left\|U_{[I]}\mathbb{M}_{W'}F'\right\| \leq \left\|U_{[I]}F'\right\| + \left\|U_{[I]}W'\right\| \left\|(WW')^{-1}\right\| \|WF'\| \lesssim_{\mathrm{P}} |I|^{1/2}T^{1/2}.$$

Replacing F by Z in the above proof, we also have the third inequality.

(viii) Using Assumption 1 and $\|\mathbb{M}_{W'}\| \leq 1$, we have $\|Z\mathbb{M}_{W'}\| \lesssim_{\mathrm{P}} T^{1/2}$. Also,

$$\|F\mathbb{M}_{W'}Z' - FZ'\| = \|F\mathbb{P}_{W'}Z'\| \leq \|FW'\| \left\|(WW')^{-1}\right\| \|WZ'\| \lesssim_{\mathrm{P}} 1.$$

Consequently, $\|F\mathbb{M}_{W'}Z'\| \leq \|F\mathbb{M}_{W'}Z' - FZ'\| + \|FZ'\| \lesssim_{\mathrm{P}} T^{1/2}$ as we have $\|FZ'\| \lesssim_{\mathrm{P}} T^{1/2}$ from Assumption 1.

□

Lemma S4.12. *For any $N \times K$ matrix β, if $\left\|T_h^{-1}FF' - \mathbb{I}_K\right\| \lesssim_{\mathrm{P}} T^{-1/2}$, we have*

(i) $\sigma_j(\beta F)/\sigma_j(\beta) = T_h^{1/2} + O_{\mathrm{P}}(1)$ for $j \leq K$.

(ii) If $\sigma_1(\beta) - \sigma_2(\beta) \asymp \sigma_1(\beta)$, then $\left\|\mathbb{P}_{\tilde{\xi}} - T_h^{-1}F'\mathbb{P}_b F\right\| \lesssim_{\mathrm{P}} T^{-1/2}$, where b and $\tilde{\xi}$ are the first right singular vectors of β and βF, respectively.

Proof. (i) For $j \leq K$, $\sigma_j(\beta F)^2 = \lambda_j(\beta F F'\beta') = \lambda_j(\beta'\beta F F')$, which implies $\lambda_j(\beta'\beta)$ $\lambda_p(FF') \leq \sigma_j(\beta F)^2 \leq \lambda_j(\beta'\beta)\lambda_1(FF')$. With the assumption $\left\|T_h^{-1}FF - \mathbb{I}_K\right\| \lesssim_{\mathrm{P}} T^{-1/2}$, we have $T_h^{-1/2}\sigma_j(\beta F)/\sigma_j(\beta) = 1 + O_{\mathrm{P}}\left(T^{-1/2}\right)$ by Weyl's inequality.

(ii) Let ς and $\tilde{\varsigma}$ be the first left singular vectors of β and βF, respectively. Equivalently, ς and $\tilde{\varsigma}$ are the eigenvectors of $\beta\beta'$ and $T_h^{-1}\beta F F'\beta'$. As $\left\|\beta\beta' - T_h^{-1}\beta F F'\beta'\right\| \leq \|\beta\|^2 \left\|T_h^{-1}FF' - \mathbb{I}_K\right\| \lesssim_{\mathrm{P}} \sigma_1(\beta)^2 T^{-1/2}$ and $\sigma_1(\beta) - \sigma_2(\beta) \asymp \sigma_1(\beta)$, by sin-theta theorem

$$\|\varsigma\varsigma' - \tilde{\varsigma}\tilde{\varsigma}'\| \lesssim \frac{\left\|\beta\beta' - T_h^{-1}\beta F F'\beta'\right\|}{\sigma_1(\beta)^2 - \sigma_2(\beta)^2 - O(\left\|\beta\beta' - T_h^{-1}\beta F F'\beta'\right\|)} \lesssim_{\mathrm{P}} T^{-1/2}.$$

Using the relationship between left and right singular vectors, we have $b' = \varsigma'\beta/\sigma_1(\beta)$ and $\tilde{\xi}' = \tilde{\varsigma}'\beta F/\|\beta F\|$. Therefore,

$$\left\|\mathbb{P}_{\tilde{\xi}} - \frac{\sigma_1(\beta)^2}{\|\beta F\|^2}F'\mathbb{P}_b F\right\| = \left\|\tilde{\xi}\tilde{\xi}' - \frac{F'\beta'\varsigma\varsigma'\beta F}{\|\beta F\|^2}\right\| = \left\|\frac{F'\beta'\tilde{\varsigma}\tilde{\varsigma}'\beta F}{\|\beta F\|^2} - \frac{F'\beta'\varsigma\varsigma'\beta F}{\|\beta F\|^2}\right\| \lesssim_{\mathrm{P}} T^{-1/2}. \quad (85)$$

By Weyl's inequality, $T_h^{-1}\|\beta F\|^2 = \lambda_1(T^{-1}\beta F F'\beta') = \sigma_1(\beta)^2 + O_{\mathrm{P}}(\sigma_1(\beta)^2 T^{-1/2})$. In light of (85), we have $\left\|\mathbb{P}_{\tilde{\xi}} - T_h^{-1}F'\mathbb{P}_b F\right\| \lesssim_{\mathrm{P}} T^{-1/2}$. □

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.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica 71*(1), 135–171.

Davis, C. and W. M. Kahan (1970). The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis 7*(1), 1–46.

Fan, J., Y. Liao, and M. Mincheva (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society, B 75*, 603–680.

Faust, J. and J. H. Wright (2013). Forecasting inflation. In *Handbook of economic forecasting*, Volume 2, pp. 2–56. Elsevier.

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

Giglio, S. W. and D. Xiu (2021). Asset pricing with omitted factors. *Journal of Political Economy 129*(7), 1947–1990.

Huang, D., F. Jiang, K. Li, G. Tong, and G. Zhou (2022). Scaled pca: A new approach to dimension reduction. *Management Science 68*(3), 1678–1695.

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

Marcellino, M., J. H. Stock, and M. W. Watson (2006). A comparison of direct and iterated multistep ar methods for forecasting macroeconomic time series. *Journal of econometrics 135*(1-2), 499–526.

McCracken, M. W. and S. Ng (2016). Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics 34*(4), 574–589.

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.

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

Wedin, P.-Å. (1972). Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics 12*(1), 99–111.